ETH zürich

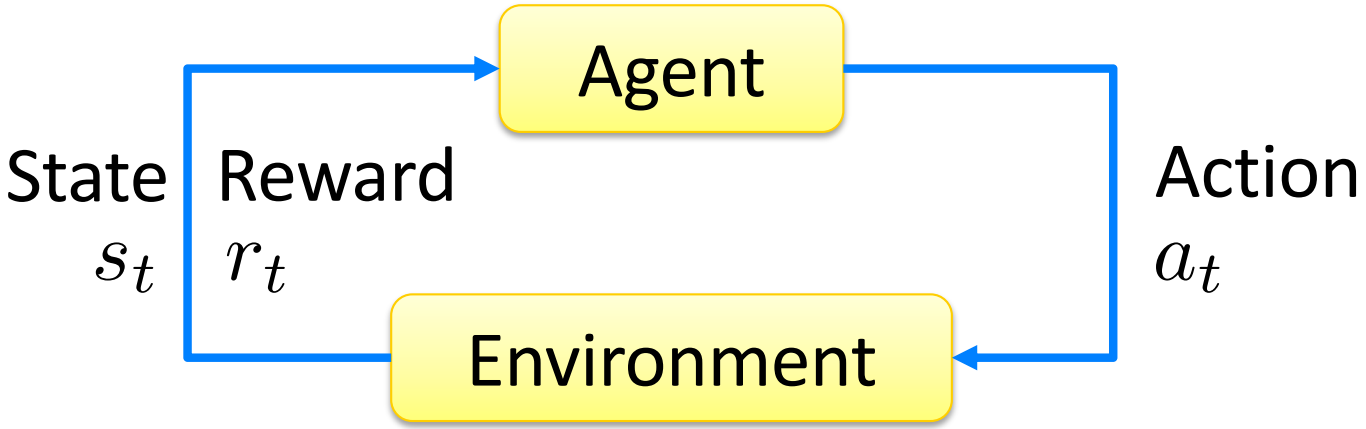
inf | Informatik Computer Science

Safe and Efficient Exploration in Reinforcement Learning

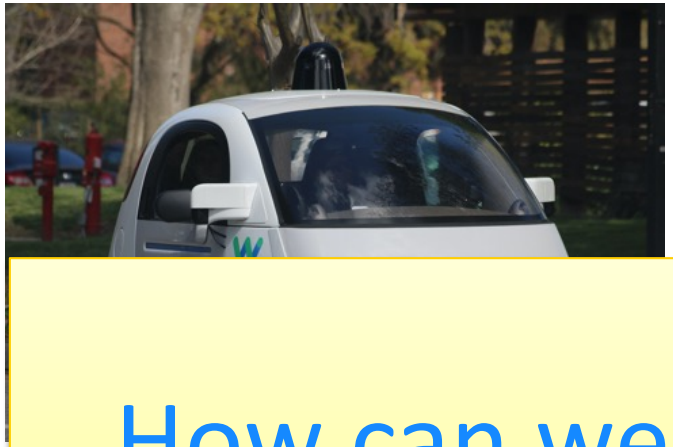
Andreas Krause

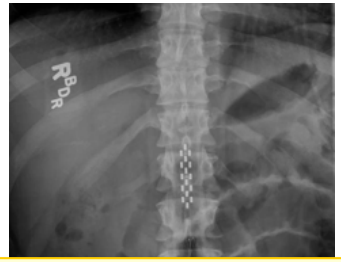
National University of Singapore 01/2021

Reinforcement Learning

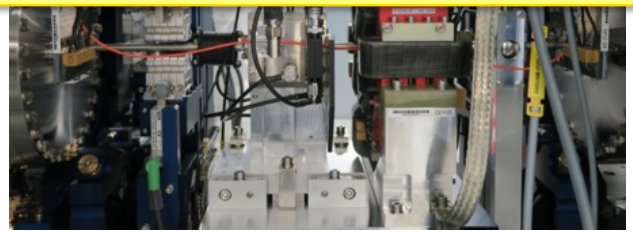


Beyond Approximate Dynamic Programming





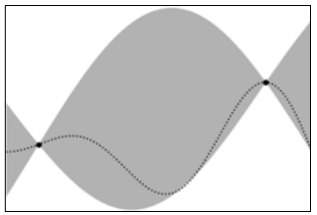
How can we *efficiently learn* to act *safely* in *unknown* environments?



. . .

Overview & Approach

Nonparametric confidence bounds



U_x

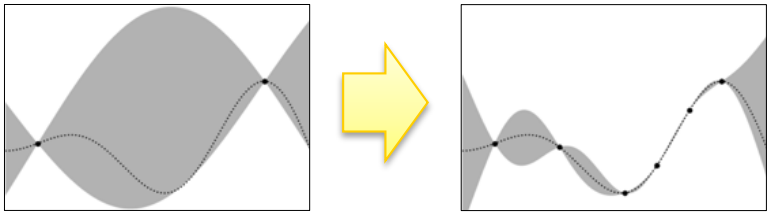
+

Robust optimization / verification

$$\min_{\mathbf{x}} \max_{\delta \in U_{\mathbf{x}}} f(\mathbf{x}, \delta)$$

+

Safe and efficient exploration



=

Learning-based performance gains with safety guarantees

Approaches towards RL

Model-*Based*

$$[s_{t+1}, r_t] \sim P(\cdot \mid s_t, a_t; \theta)$$

Estimate/identify,
then plan/control

Model-*Free*

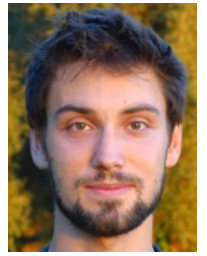
$$a_t = \pi(s_t, \theta)$$

Estimate value $J(\boldsymbol{\theta})$
and optimize

Safe Bayesian Optimization

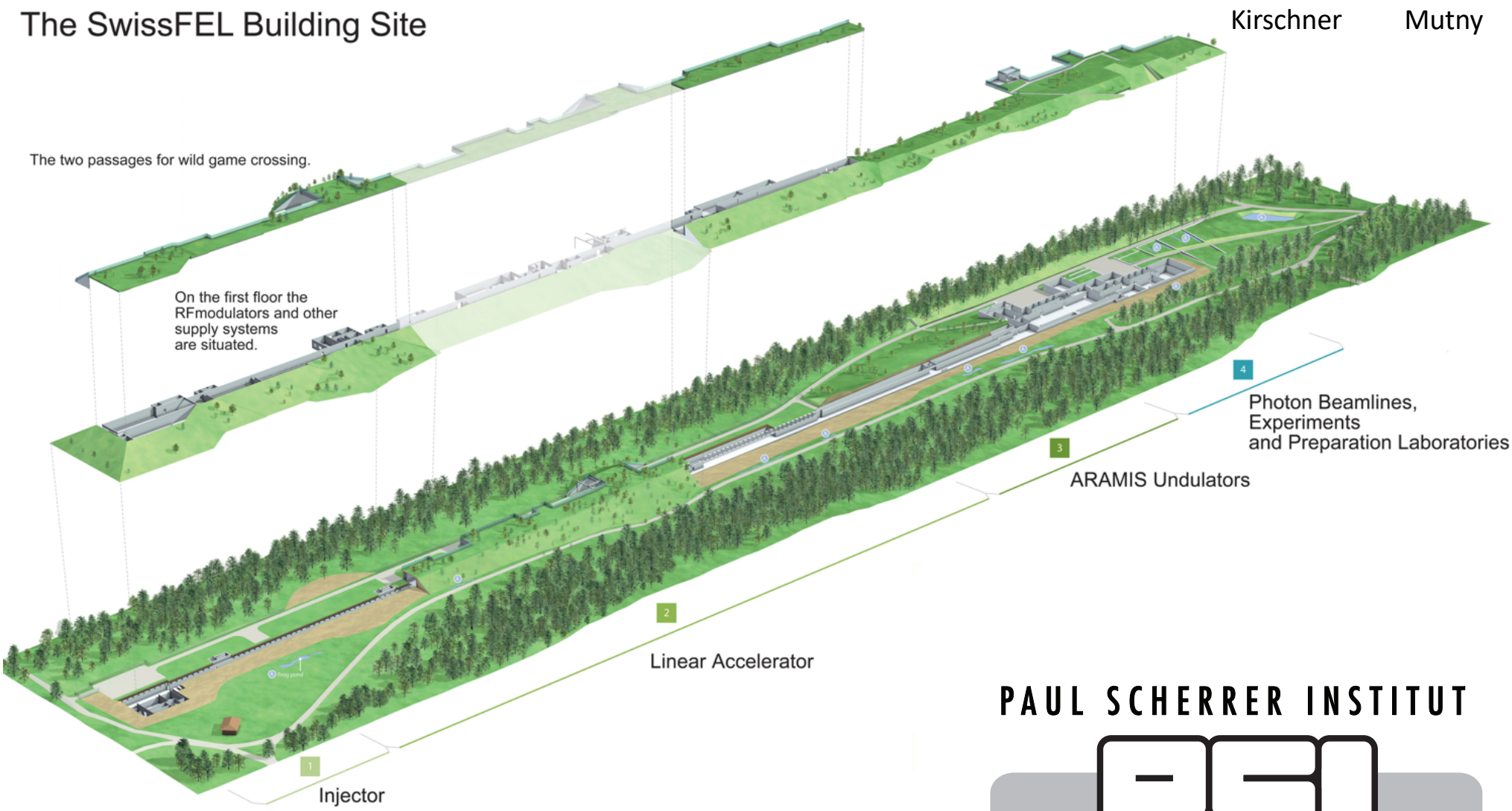
Tuning the Swiss Free Electron Laser

[with Kirschner, Mutny, Ischebeck et al ICML '19]



Johannes Kirschner

Mojmir Mutny

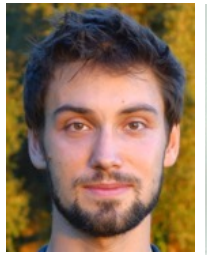


PAUL SCHERRER INSTITUT

PSI

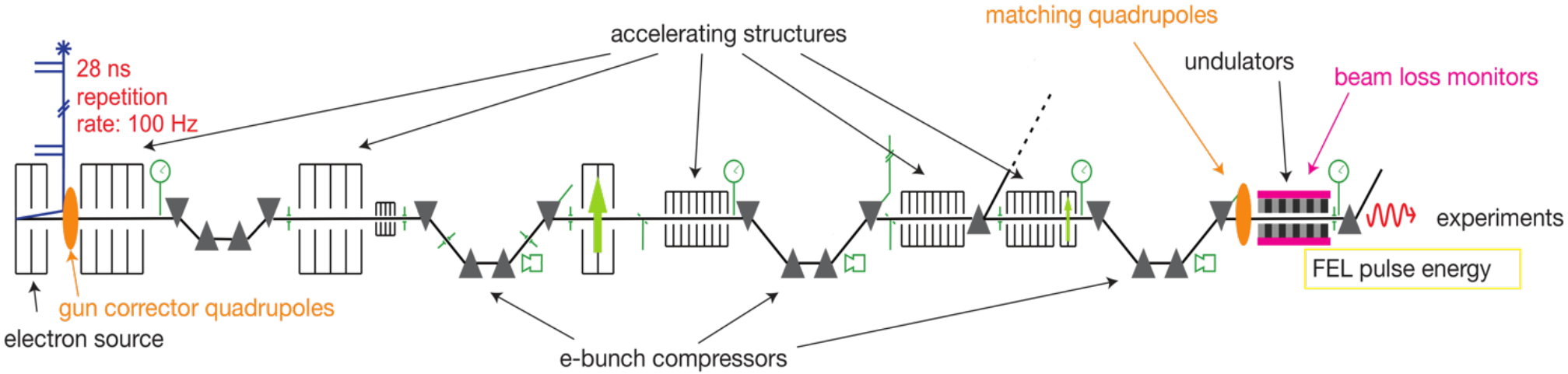
Tuning SwissFEL

[w Kirschner, Mutny, Hiller, Ischebeck et al ’19]



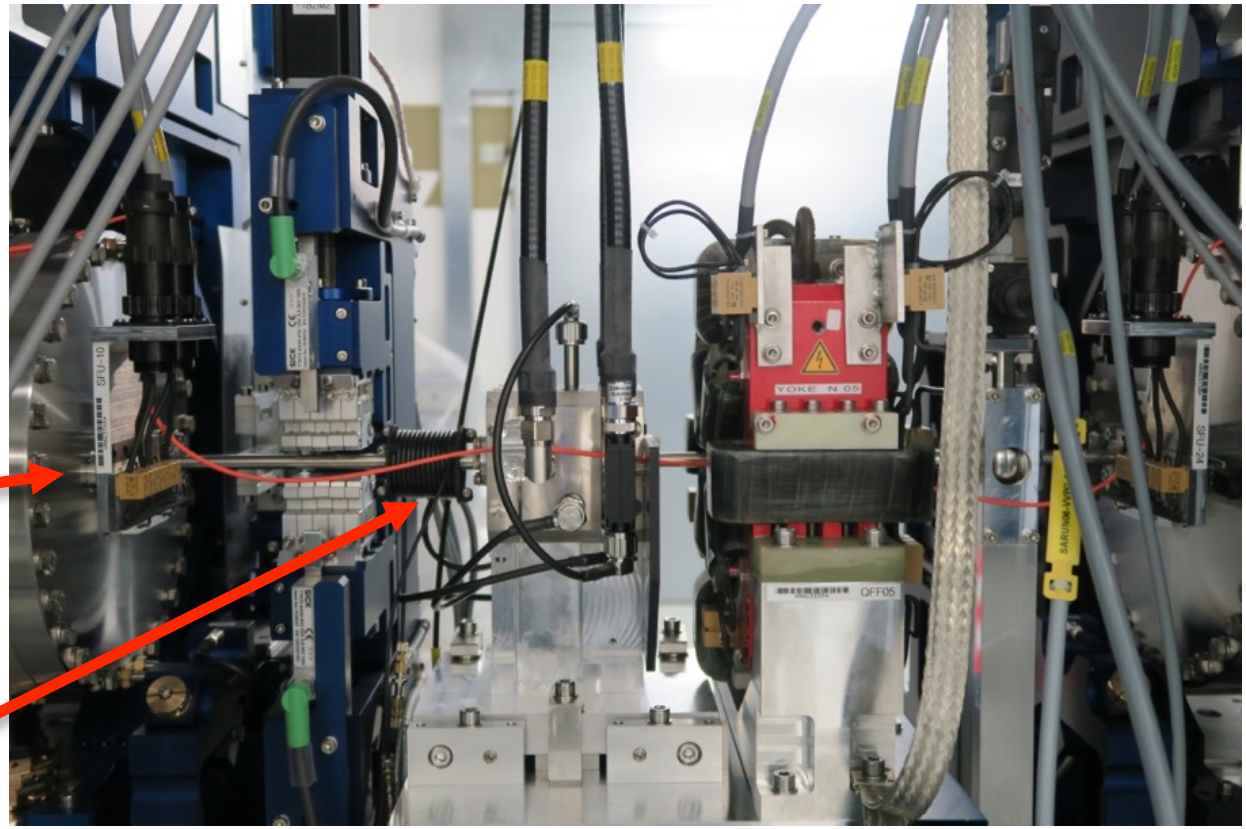
Johannes Kirschner

Mojmir Mutny



[c.f., McIntire, Ratner, Ermon ’16]

Challenge: Safety Constraints



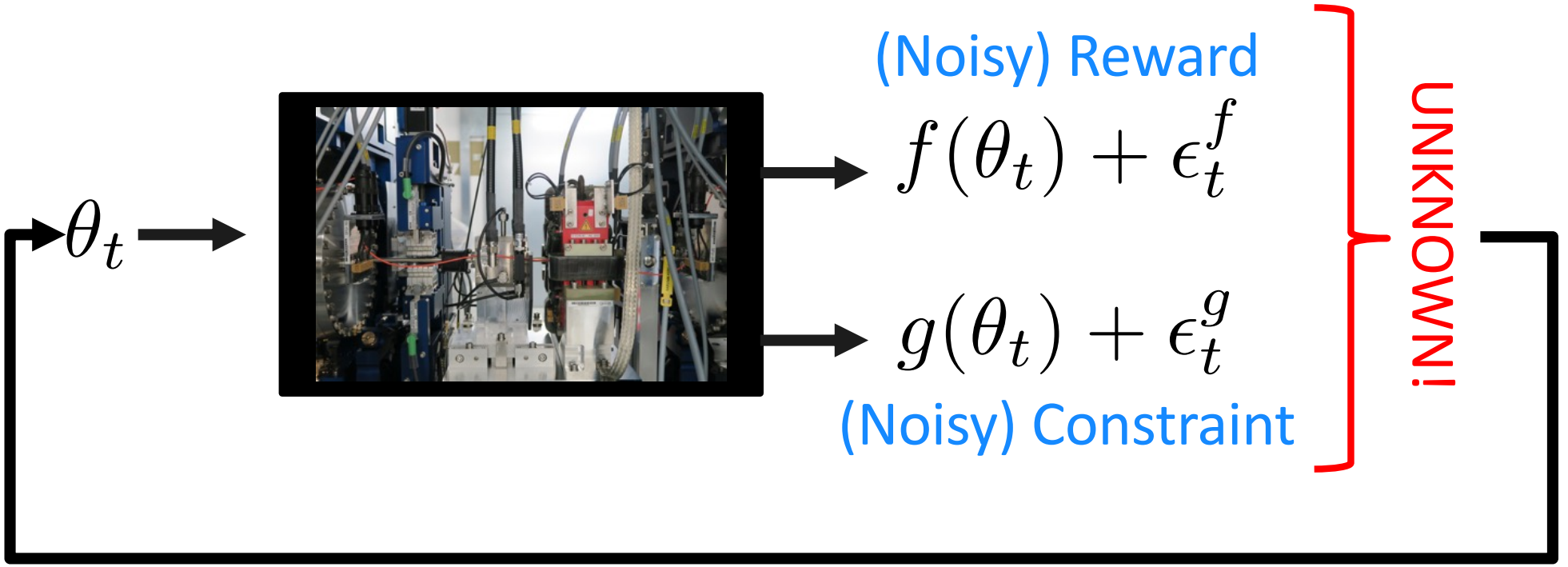
Vacuum Chamber of Undulator Module

Beam Loss Monitor

Radiation damage leads to loss of the magnetization

➔ Undulators need to be replaced

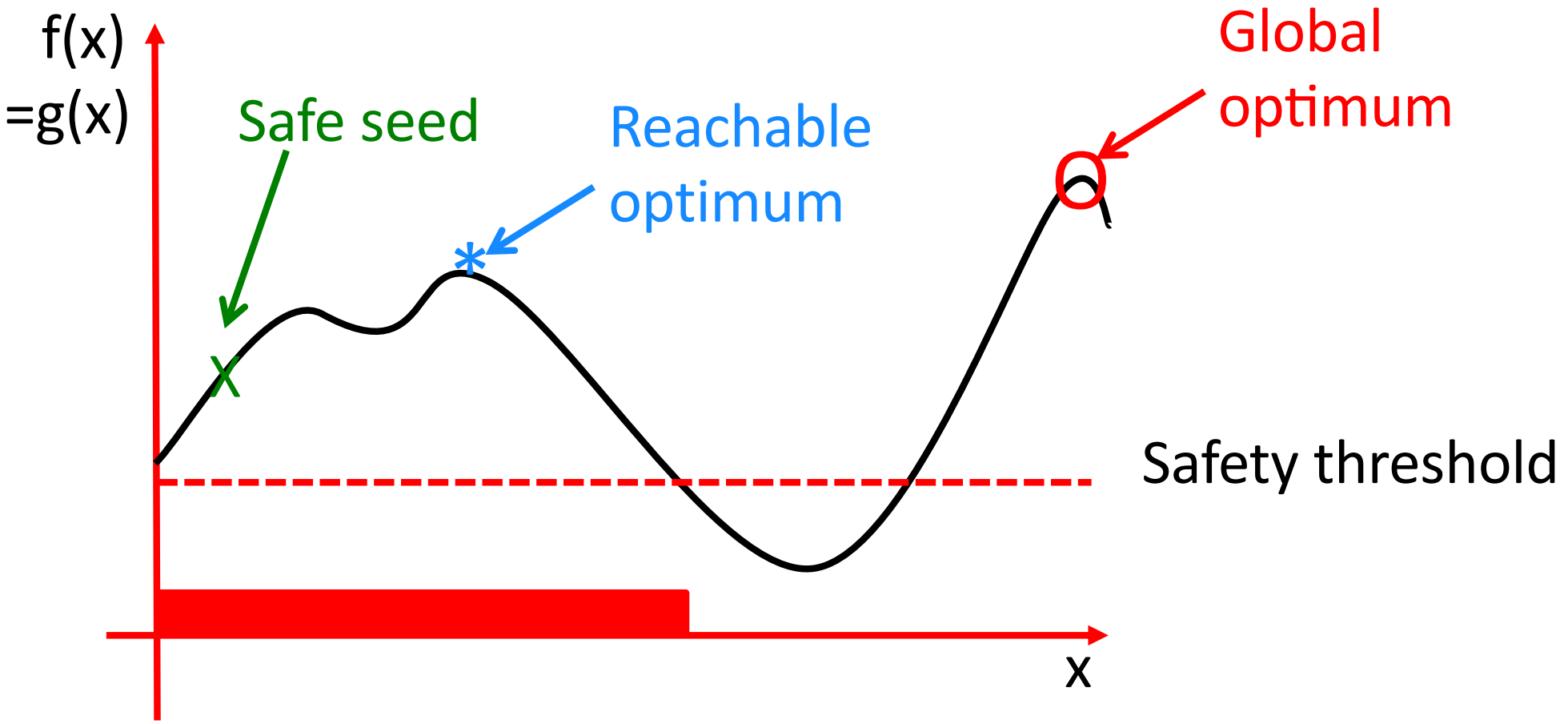
Safe optimization



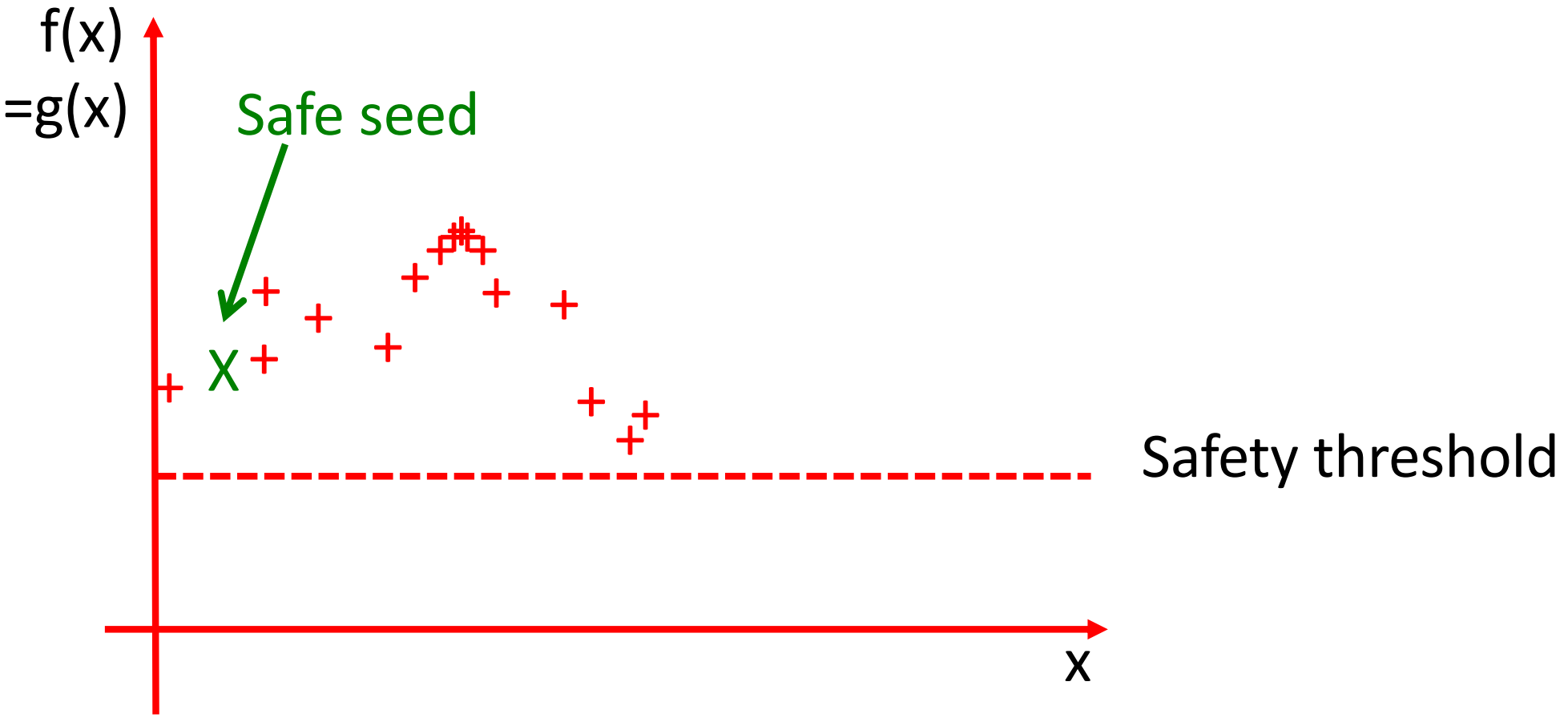
Goal: $\max_\theta f(\theta) \text{ s.t. } g(\theta) \geq 0$

Safety: $g(\theta_t) \geq 0 \text{ for all } t$

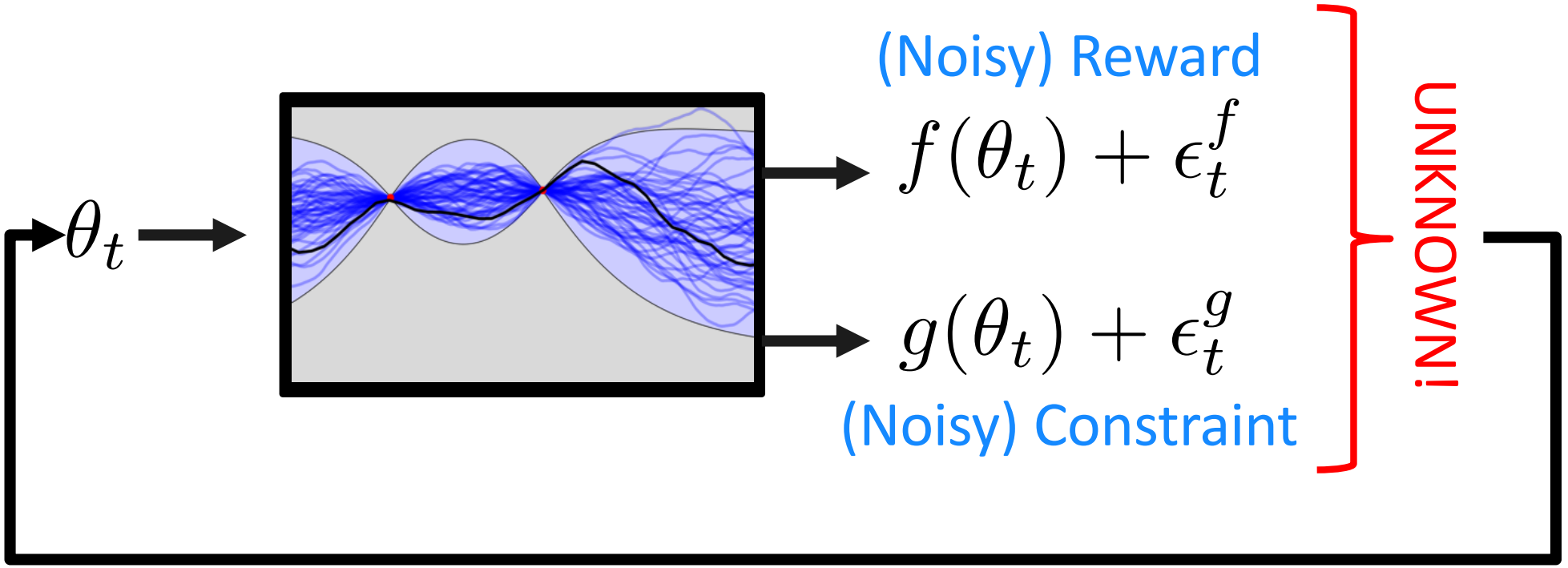
Safe optimization



Safe optimization



Safe *Bayesian* optimization



Goal: $\max_\theta f(\theta) \text{ s.t. } g(\theta) \geq 0$

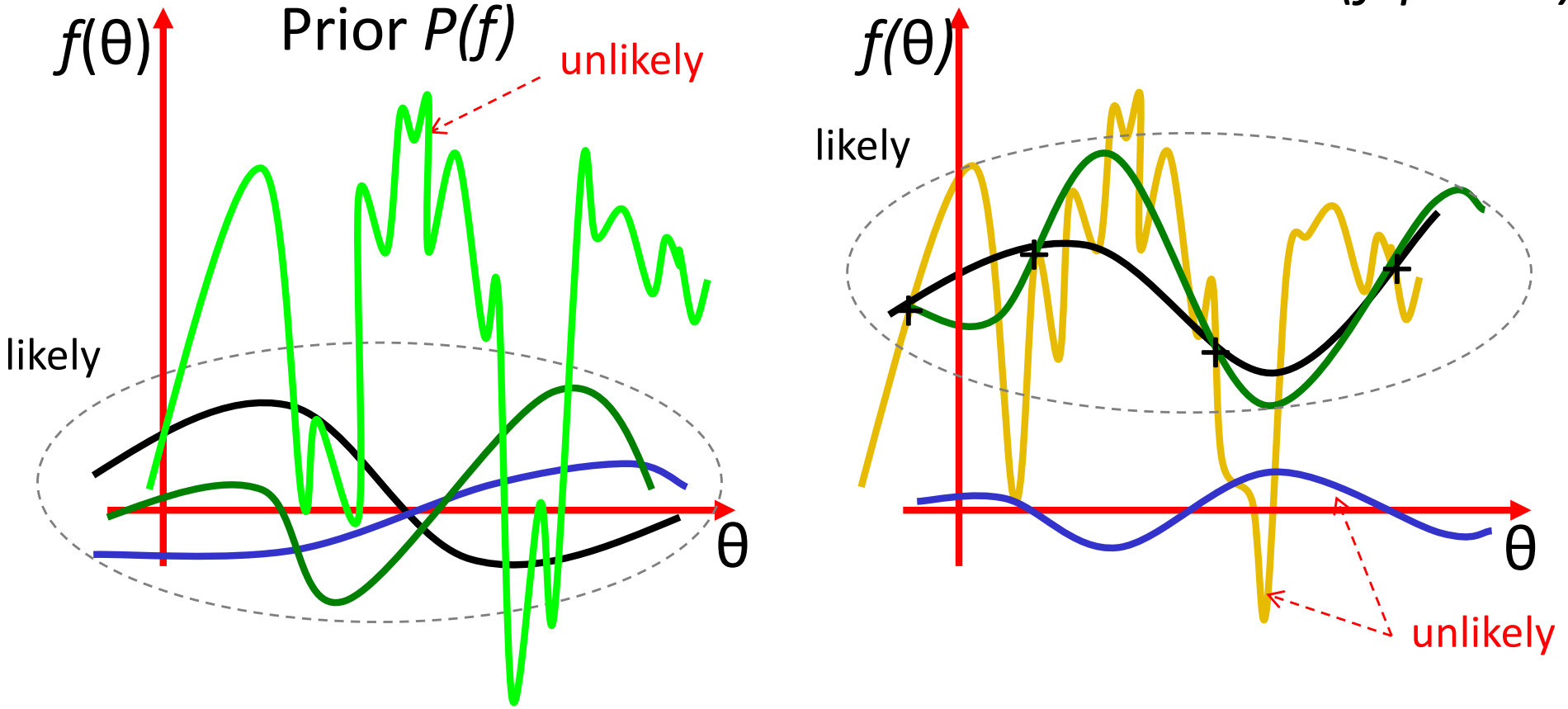
Safety: $g(\theta_t) \geq 0 \text{ for all } t$

Gaussian processes

[*c.f.* Rasmussen & Williams 2006]

Likelihood: *P(data | f)*

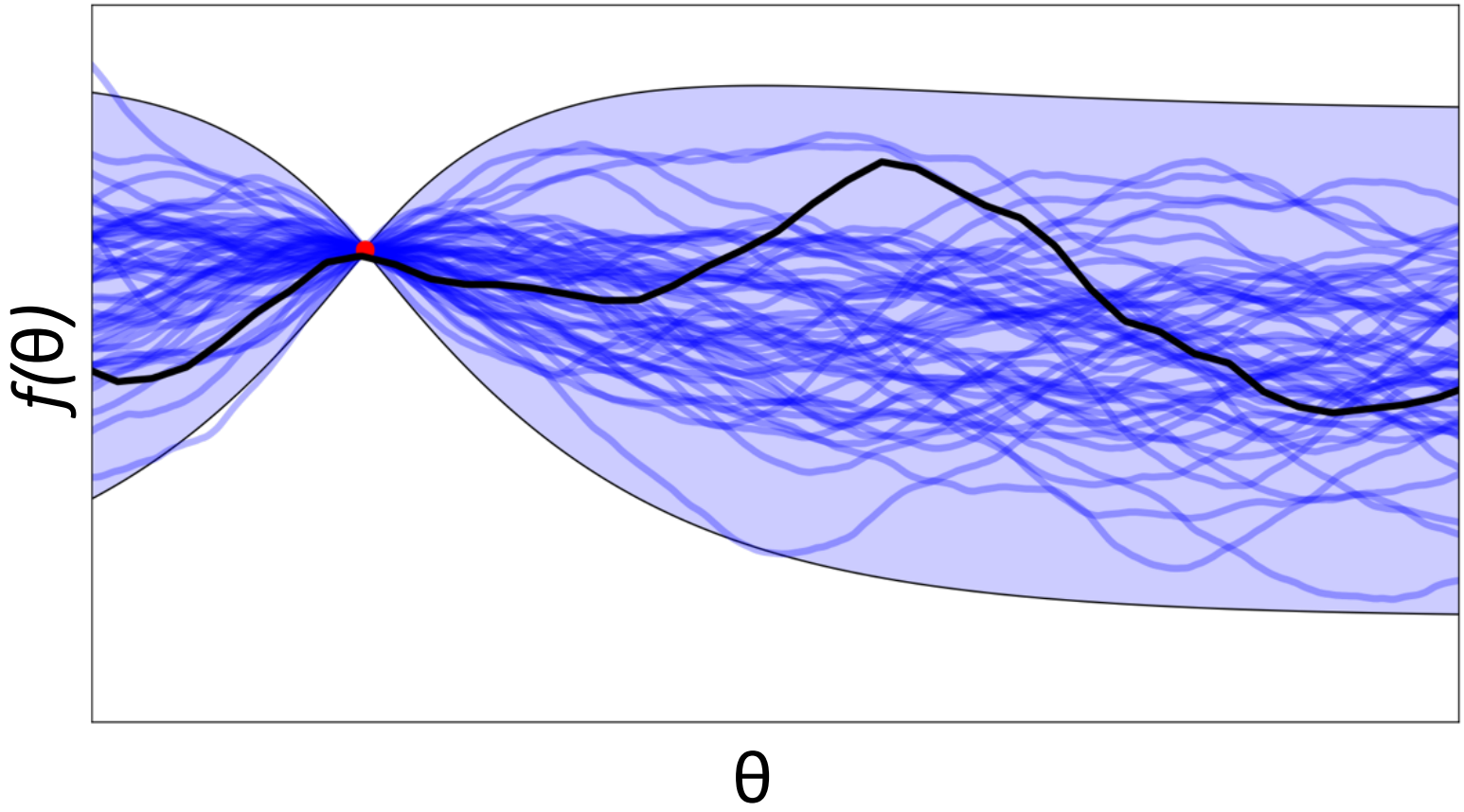
➔ Posterior: *P(f | data)*



Expressive + predictive uncertainty + tractable inference

Illustration of Gaussian Process Inference

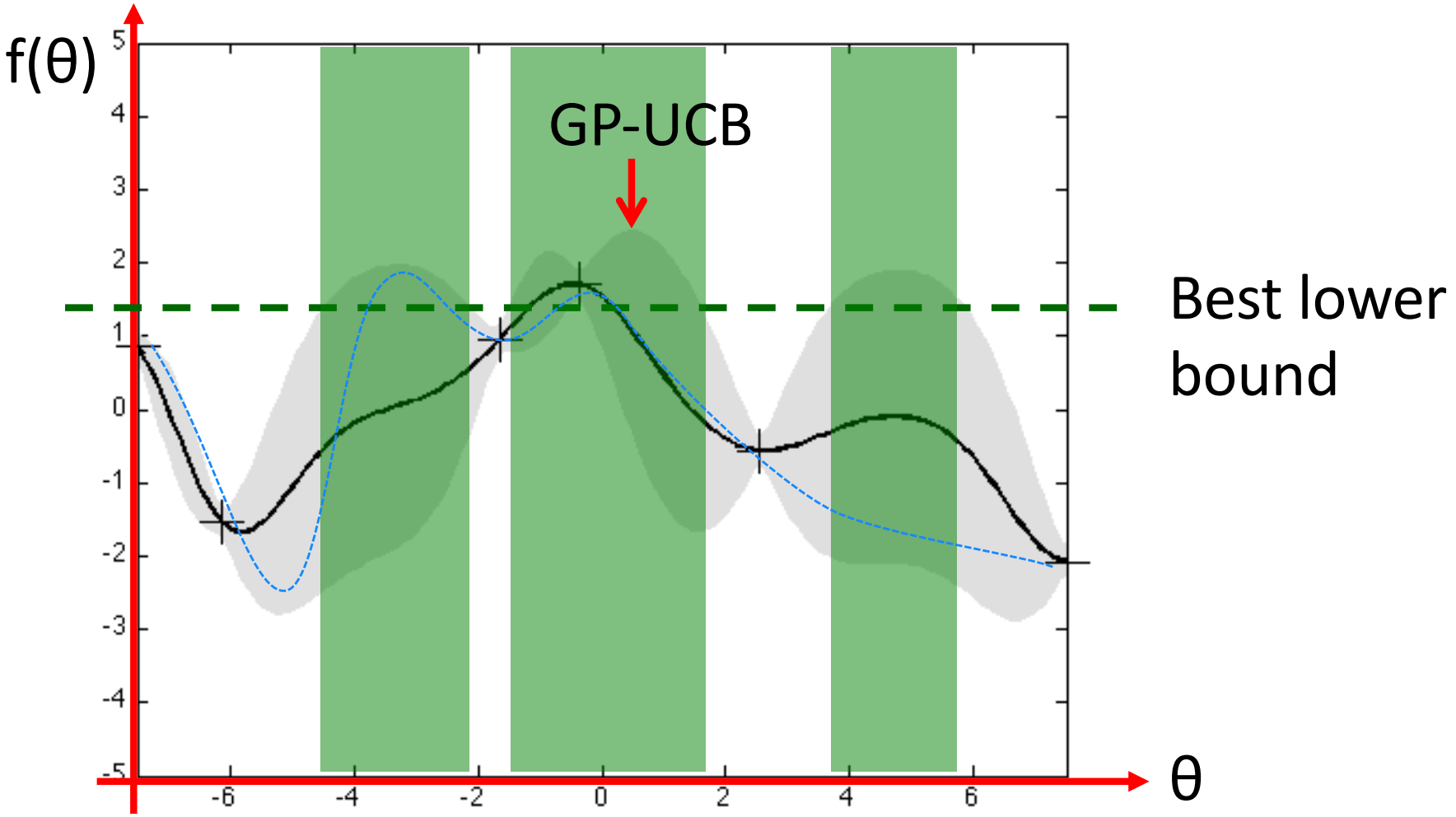
[cf, Rasmussen & Williams 2006]



Smoothness characterized via covariance function

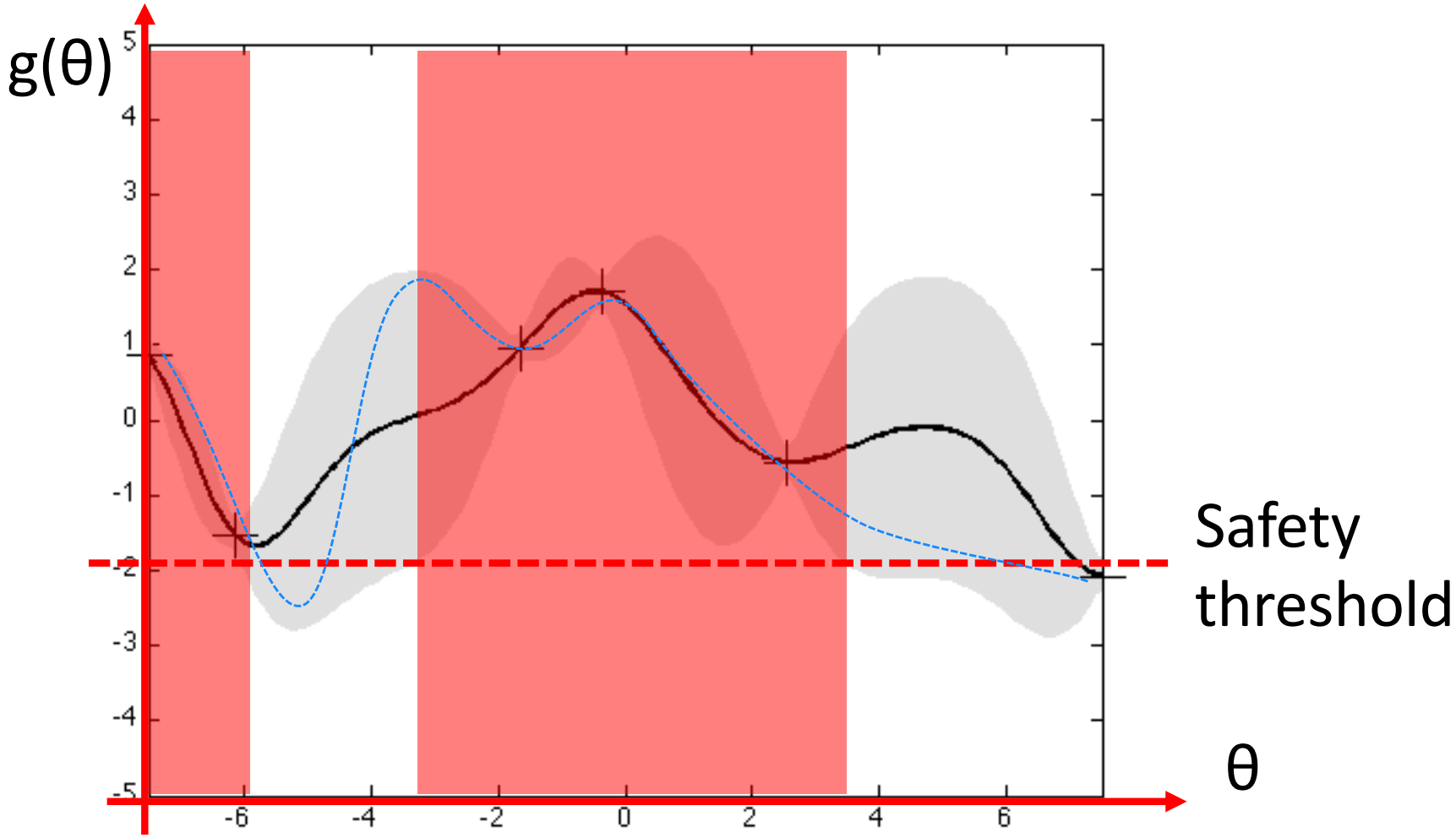
$$k(\theta, \theta') = \mathrm{Cov}\Big(f(\theta), f(\theta')\Big)$$

Plausible maximizers



Focus exploration where
upper confidence bound ≥ best lower bound!

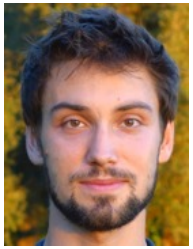
Certifying Safety

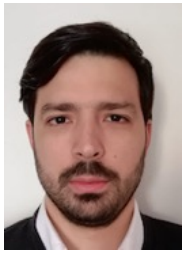


Statistically certify safety where lower bound > threshold!

Confidence intervals for GPs?


Niranjan Srinivas

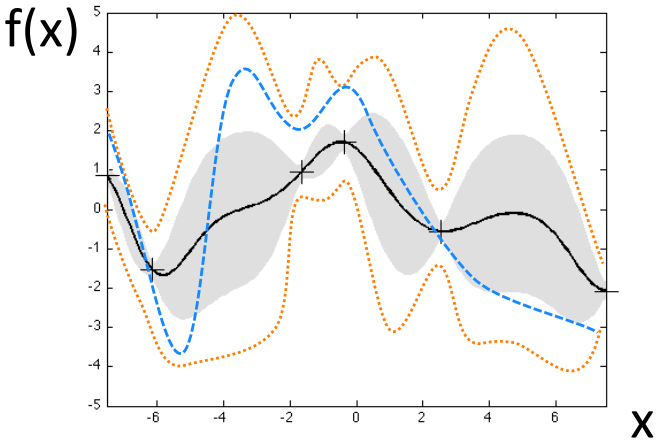

Johannes Kirschner


Ilija Bogunovic

Theorem: [w Srinivas, Kakade, Seeger'12; w Kirschner'18]

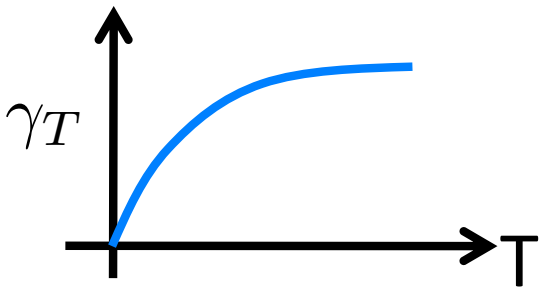
$$\Pr\Big(\forall x, t : f(x) \in [\mu_t(x) \pm \beta_t \sigma_t(x)]\Big) \geq 1 - \delta$$

$$\tilde{O}(\|f\|_k + \sqrt{\gamma_t})$$



"Complexity" of *f*

$$\gamma_T = \max_{|A| \leq T} I(f; y_A)$$

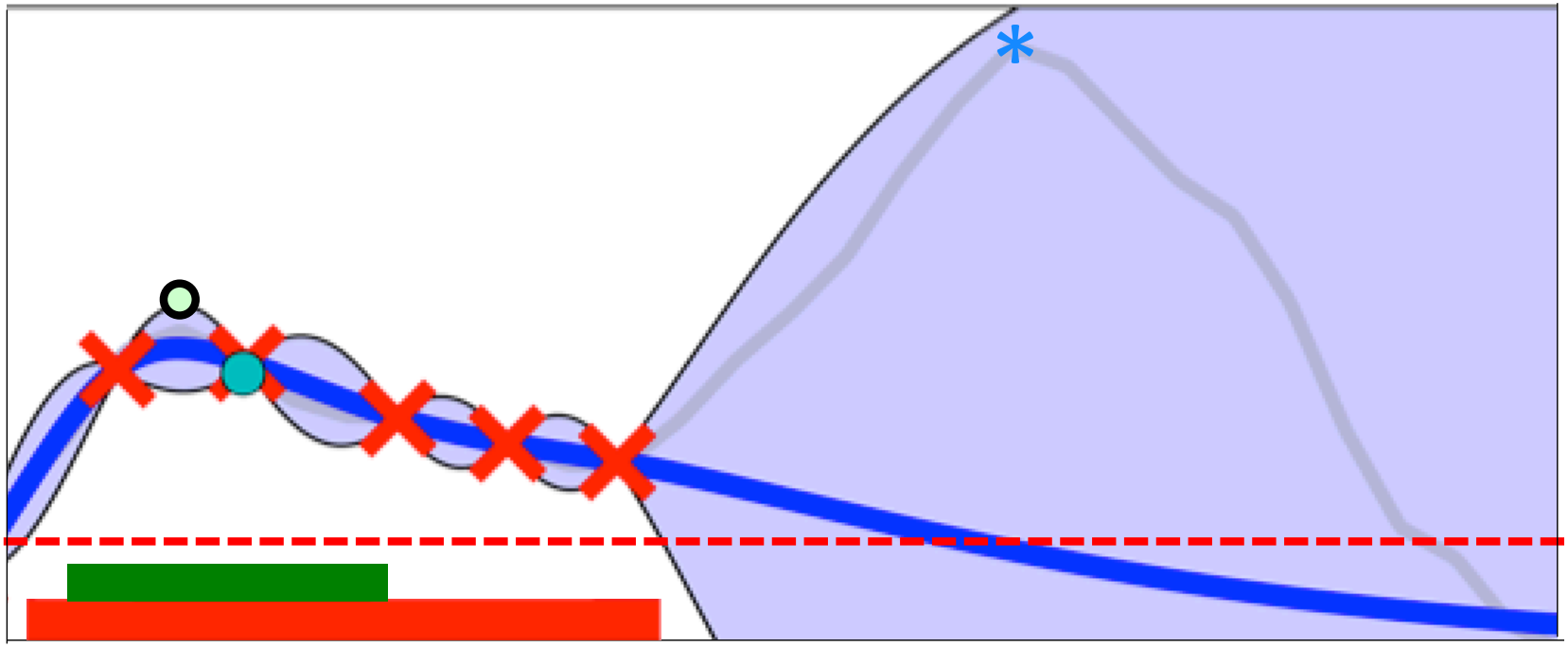


Maximum #bits about *f*

Can bound via *submodular* analysis

Can even handle *adversarial* corruptions! [AISTATS 2020]

First Attempt: SafeUCB

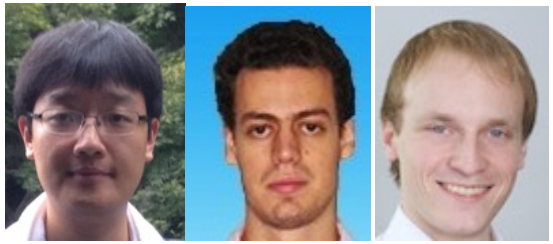


Maximize acquisition function (GP-UCB, EI, …)
over certified safe domain

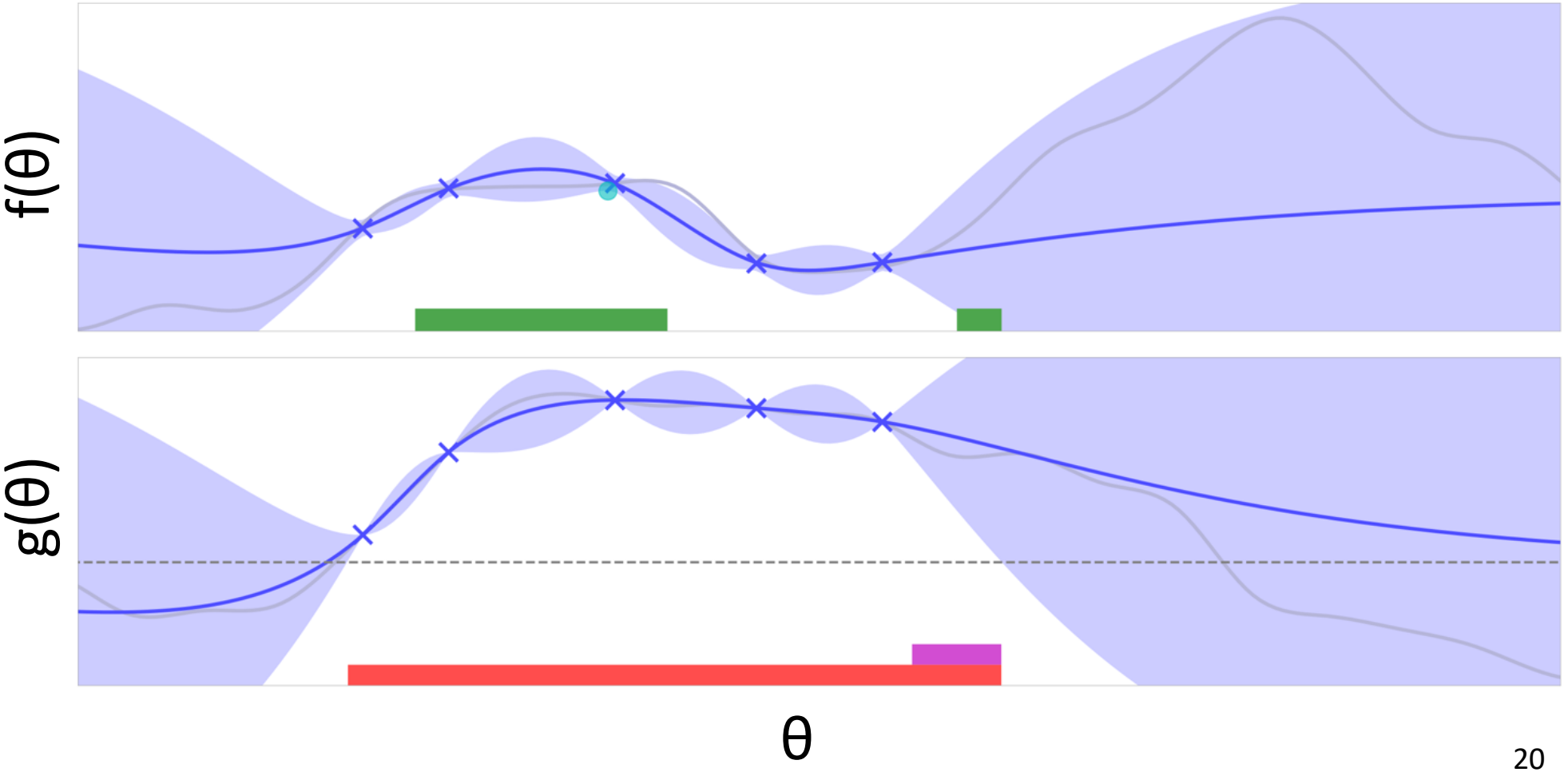
➔ Gets stuck in local optima!

SAFEOPT

[Sui, Gotovos, Burdick, K ICML'15],
[Berkenkamp, Schoellig K'16]



Yanan Sui | Alkis Gotovos | Felix Berkenkamp



SAFEOPT Guarantees

[Sui, Gotovos, Burdick, K ICML ’15; Berkenkamp Schoellig, K’16]

Yanan Sui | Alkis Gotovos | Felix Berkenkamp

Theorem (informal):

Under suitable conditions on the kernel and on *f*,*g*, there exists a function *T(ε,δ)* such that for any *ε>0* and *δ>0*, it holds with probability at least *1-δ* that

1) SAFEOPT never makes an unsafe decision

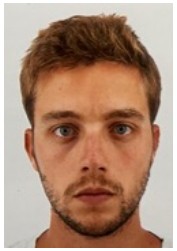
2) After at most *T(ε,δ)* iterations, it found an ε-optimal reachable point

For Gaussian kernel: (fixed domain & dim.)

$$T(\varepsilon,\delta) \in \tilde{O}\left((||f||_k + ||g||_k) \frac{\log^3 1/\delta}{\varepsilon^2} \right)$$

Goal-directed Safe Exploration

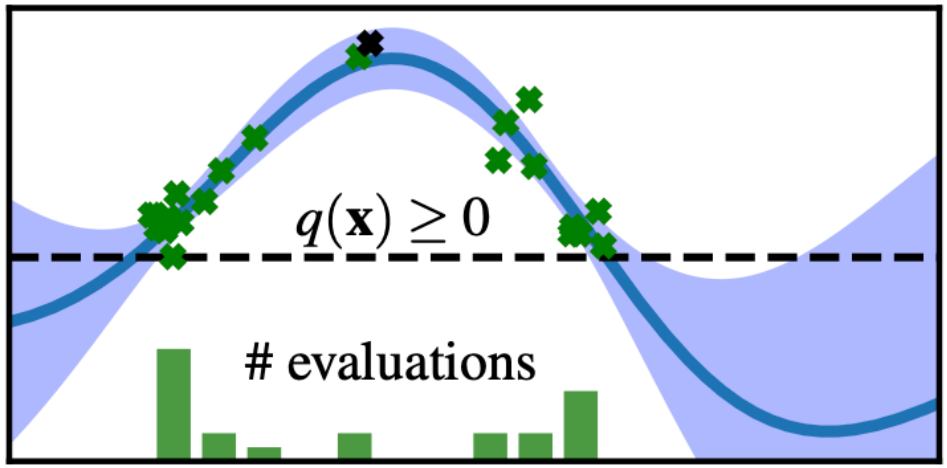
[M. Turchetta, F. Berkenkamp, A. Krause, NeurIPS 2019]



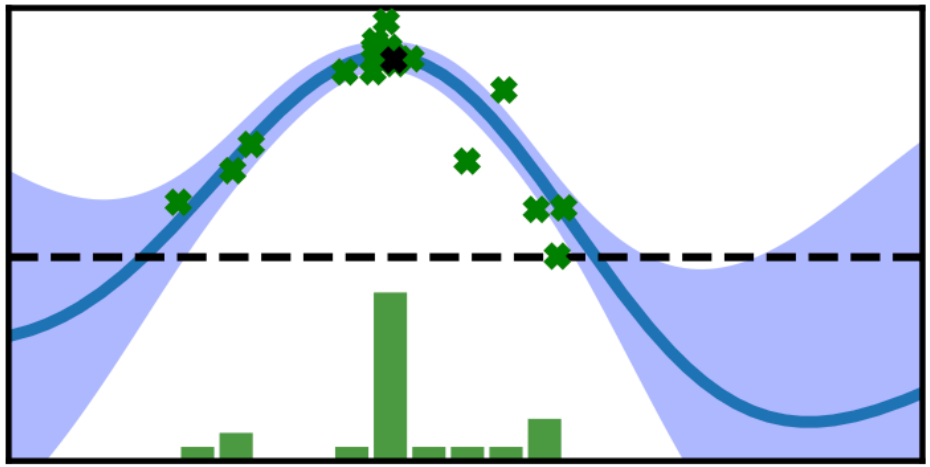
Matteo Turchetta

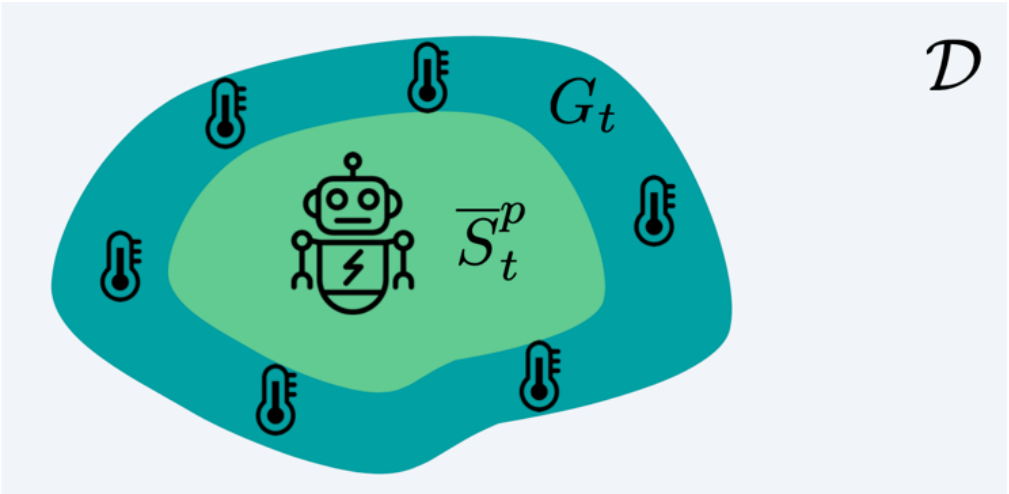
Felix Berkenkamp

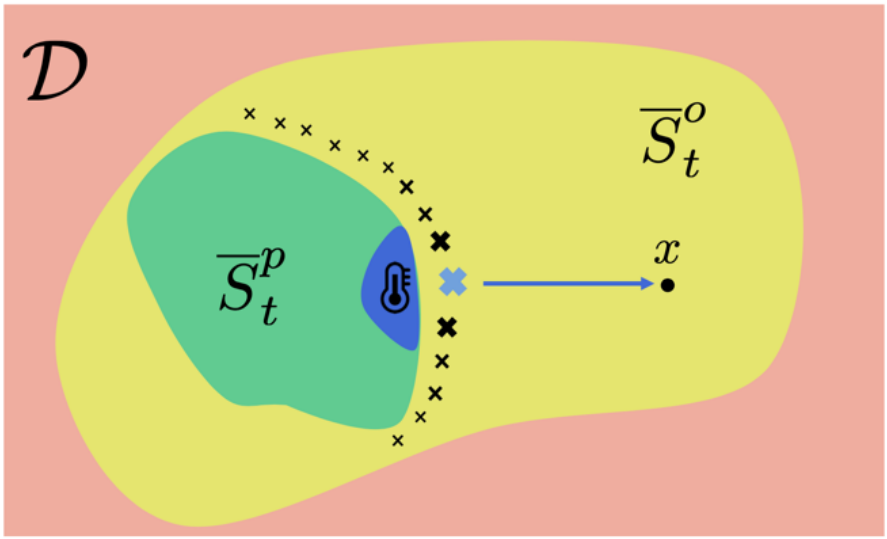
SafeOPT



GoOSE + GP-UCB



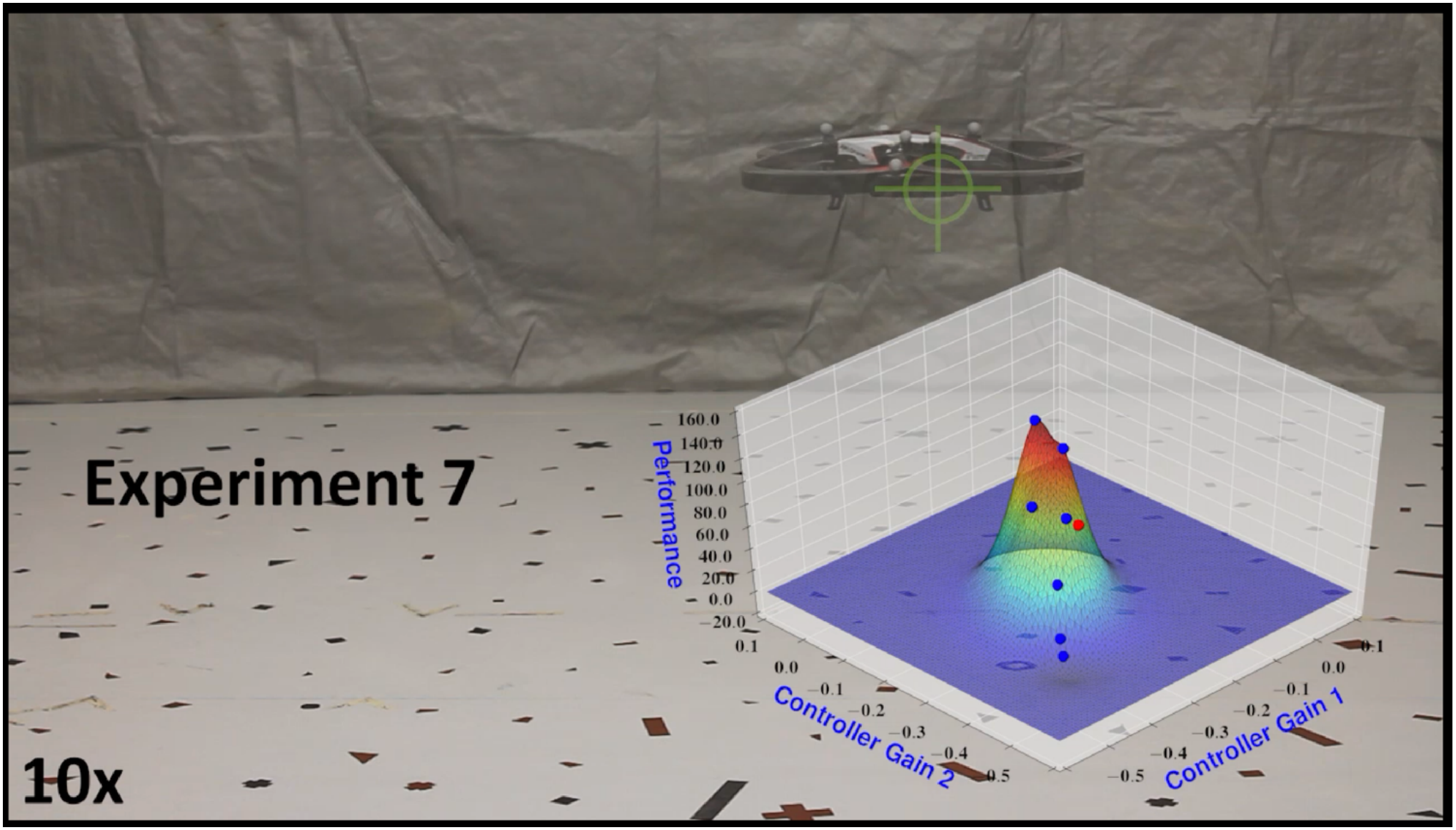




Safe controller tuning

[Berkenkamp, Schoellig, K, ICRA ’16]

Felix Berkenkamp



Challenge: High Dimensions

[with Mutny, Kirschner, Nonnenmacher, Hiller, Ischebeck '19]

- Basic SAFEOPT algorithm relies on discretization
- This does not scale to high dimensions

- Can extend to higher dimensions: LINEBO

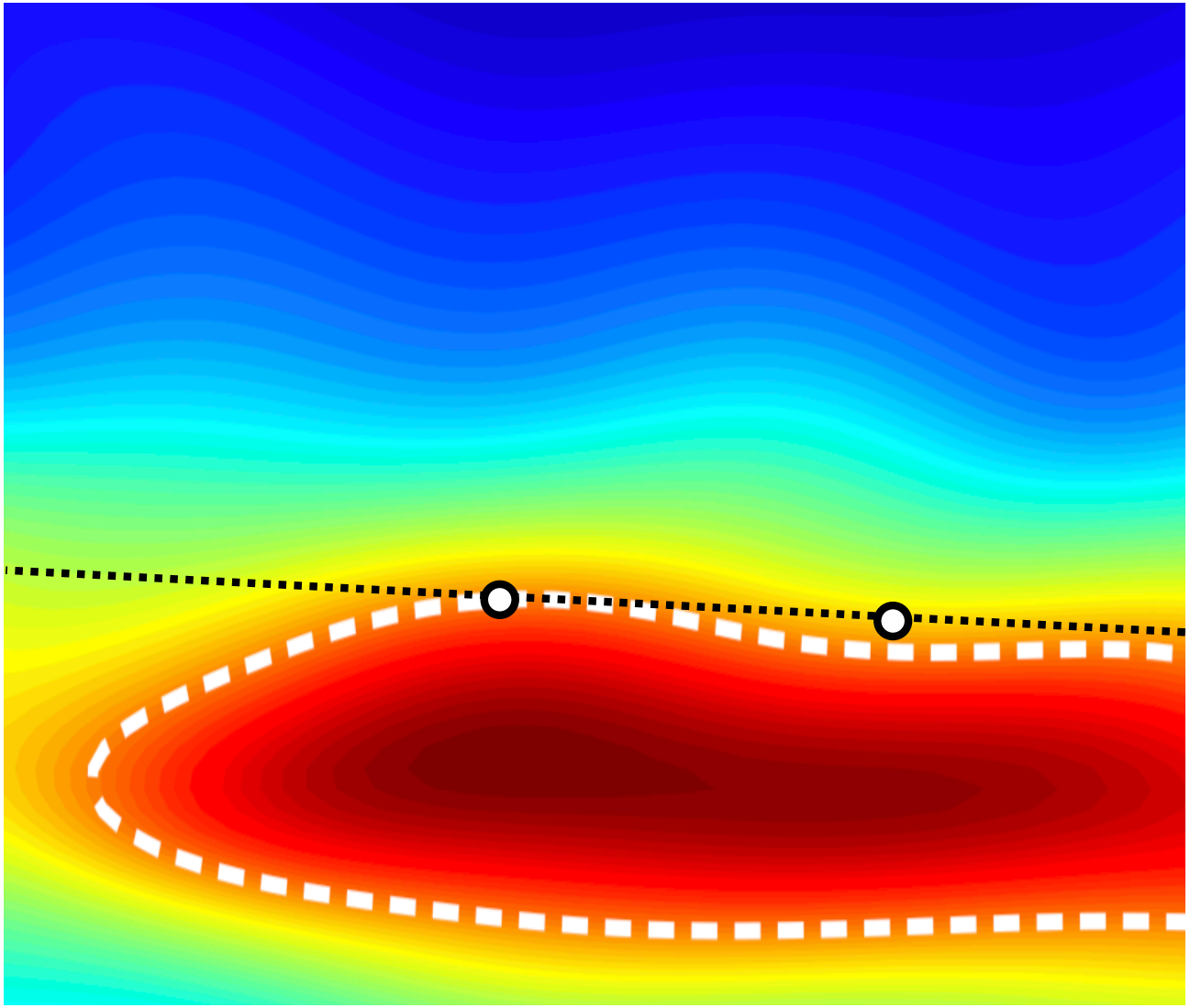
  Solve a sequence of one-dimensional Bayesian Optimization problems on one dimensional subspaces

  ➔ Each subproblem is efficient

  ➔ Global model allows to share observations

  ➔ Flexible choice of acquisition function (e.g., SafeOPT)

LineBO

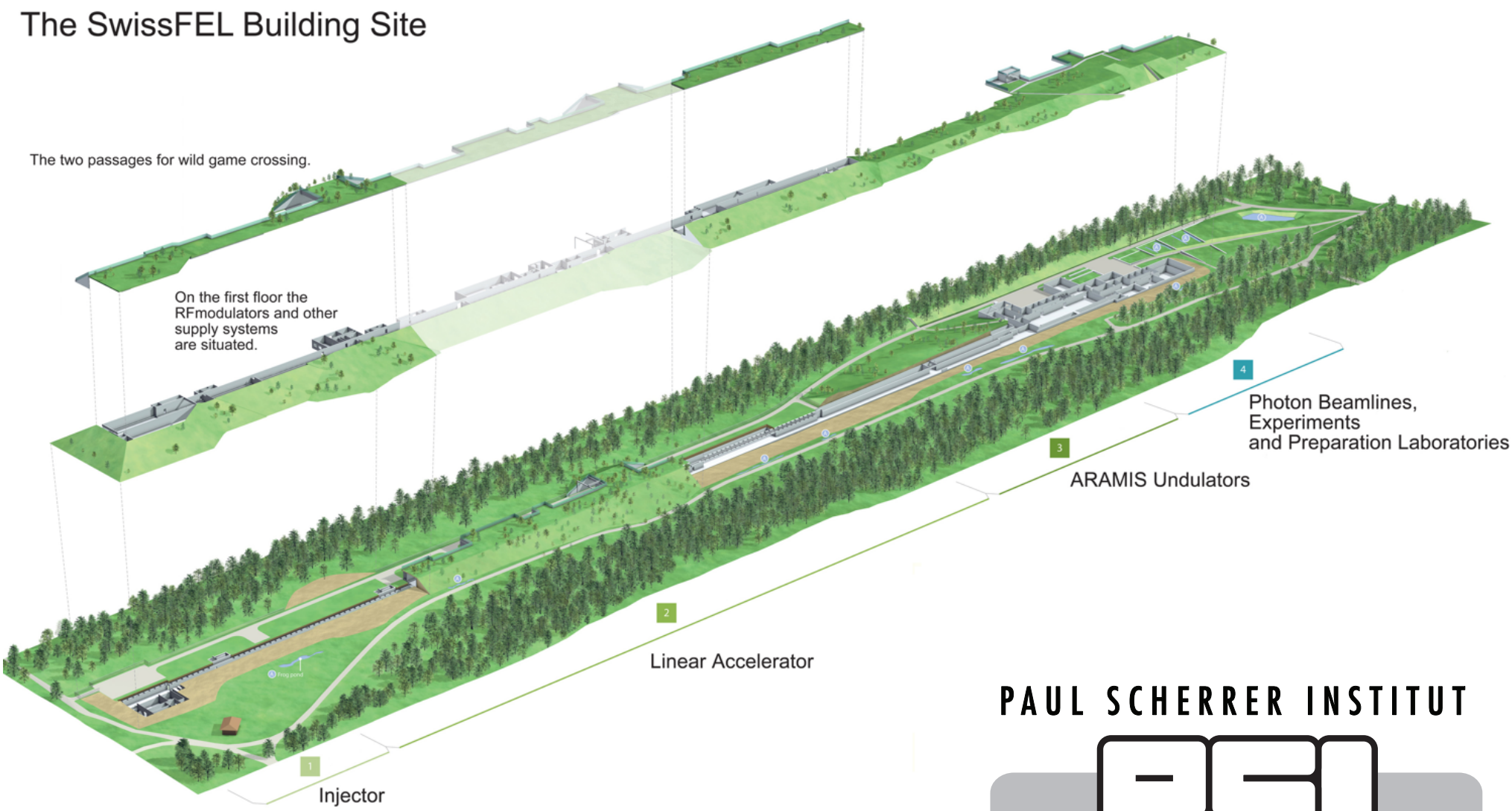


Guarantees in high dimensions

[with Mutny, Kirschner, Hiller, Ischebeck '19]

- Safe Bayesian Optimization in high dim.: LINEBO
  - Solve a sequence of one-dimensional Bayesian optimization problems on one dimensional subspaces
- For random subspaces, can guarantee simultaneously
  - Global convergence (at Lipschitz rates, automatically *adapting to intrinsic dimension*)
  - Local convergence (at *fast rates* in case of locally strongly convex functions)
- Can also (heuristically) use more informed directions

The Swiss Free Electron Laser

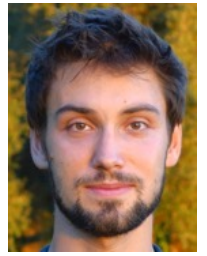


PAUL SCHERRER INSTITUT

PSI

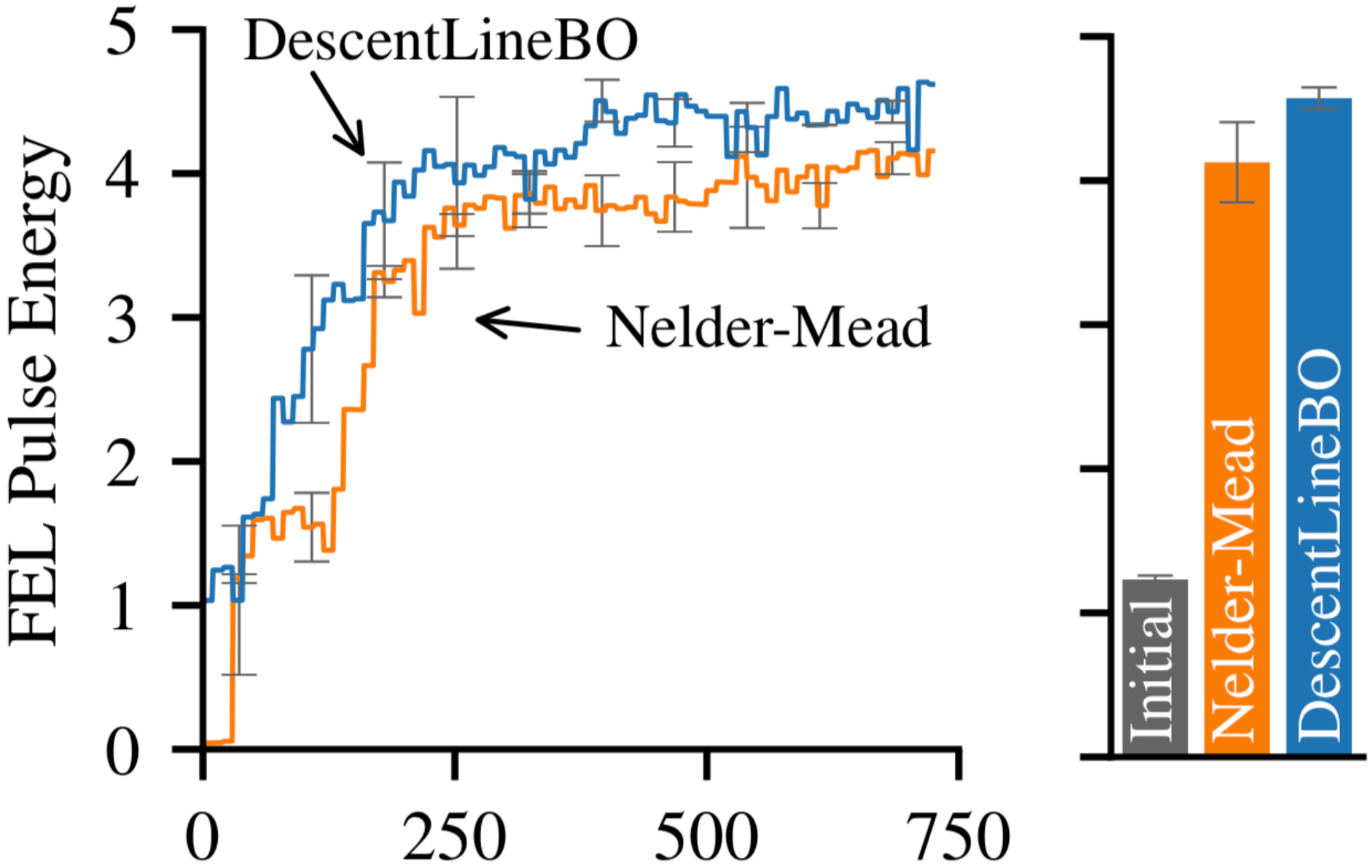
Online tuning of 24 parameters

[Kirschner, Mutny, Hiller, Ischebeck, K ICML 2019]



Johannes Kirschner

Mojmir Mutny



Approaches towards RL

Model-*Based*

$$[s_{t+1}, r_t] \sim P(\cdot \mid s_t, a_t; \theta)$$

Estimate/identify,
then plan/control

Model-*Free*

$$a_t = \pi(s_t, \theta)$$

Estimate value $J(\boldsymbol{\theta})$
and optimize

Safe Learning-based Model Predictive Control via Confidence Bounds

Safe learning for dynamical systems

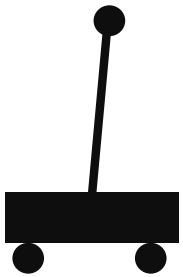
[w Koller, Berkenkamp, Turchetta CDC '18, arXiv '19]

Torsten Koller | Felix Berkenkamp | Matteo Turchetta

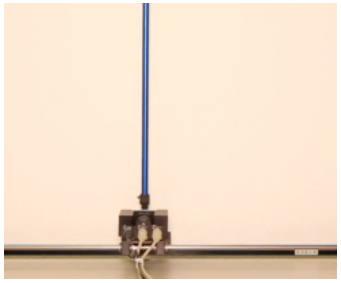
$$s_{t+1} = \underbrace{f(s_t, a_t)}_{\text{a priori model}} \quad \underbrace{\qquad}_{\text{disturbance}}$$

Learnt with a nonparametric (GP) model

[c.f. Wang+'05, Deisenroth & Rasmussen '11, Akametalu+ '14, …]



vs



vs

Stylized task

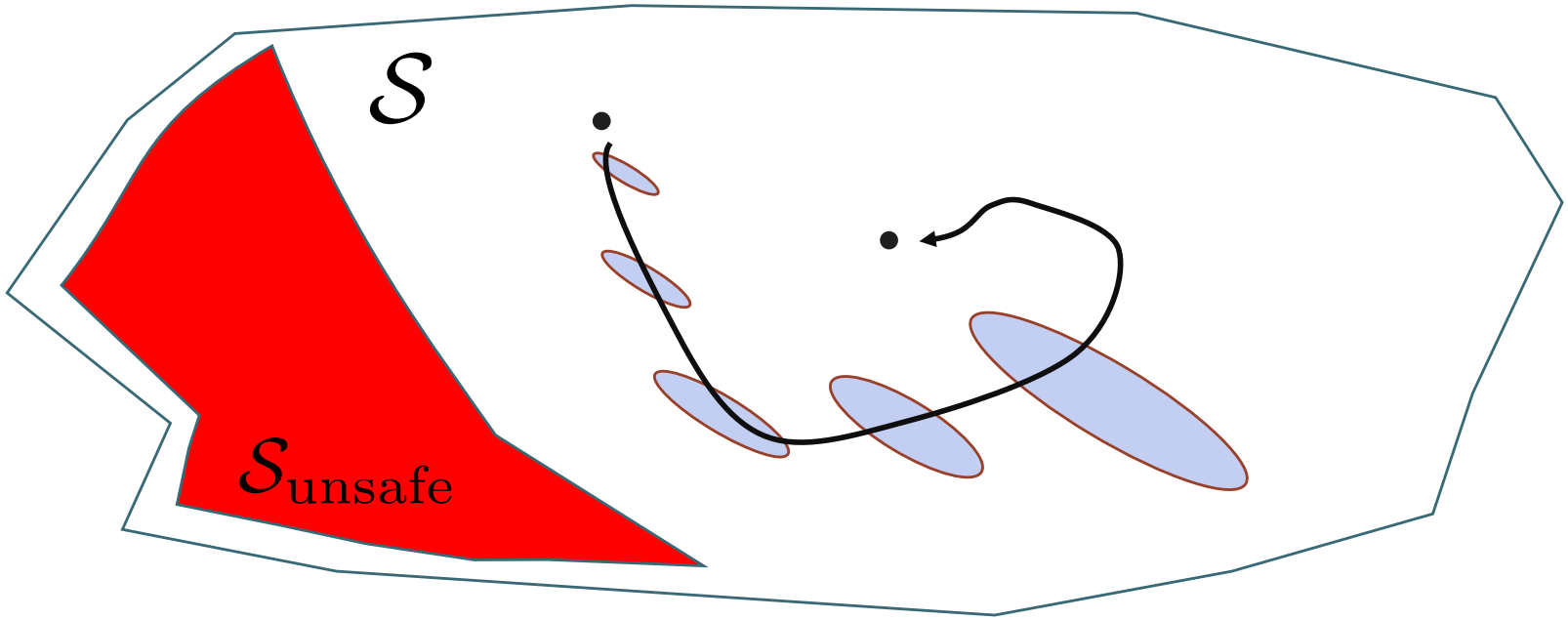
Want to reach target (quickly) while balancing pole

Target location

Planning with confidence bounds

[w Koller, Berkenkamp, Turchetta CDC '18, arXiv '19]

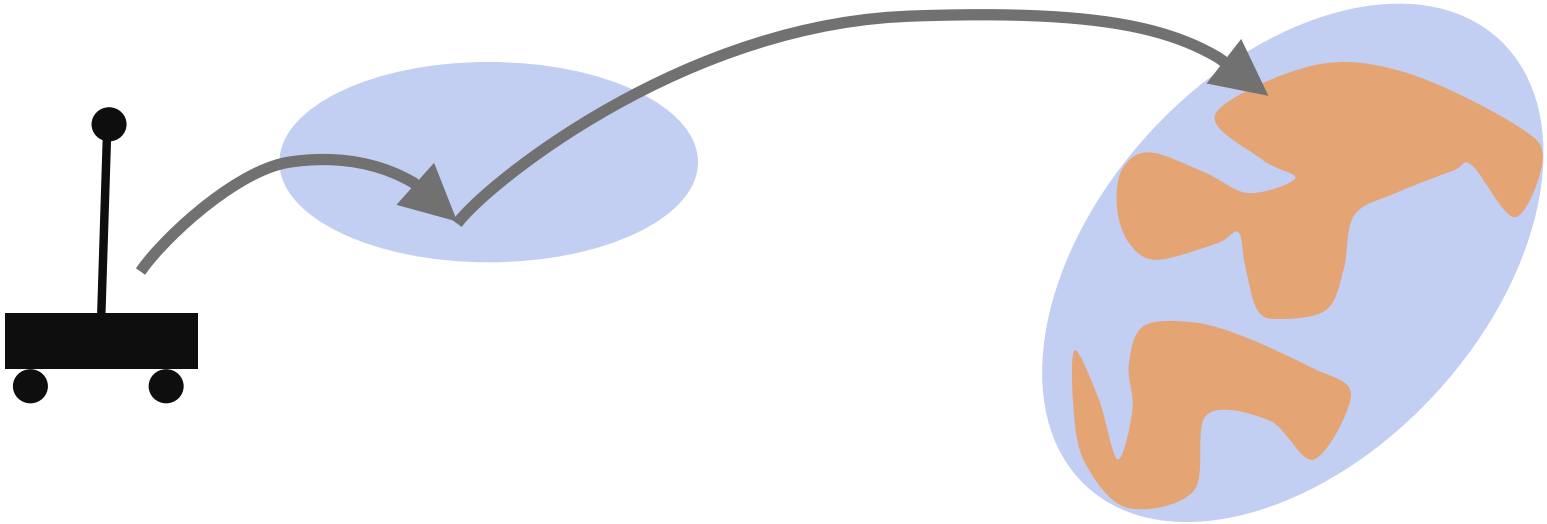
Torsten Koller | Felix Berkenkamp | Matteo Turchetta



Forwards-propagating uncertain, nonlinear dynamics

[w Koller, Berkenkamp, Turchetta CDC '18]

Torsten Koller | Felix Berkenkamp | Matteo Turchetta



Thm: Outer approximation contains true dynamics for all time steps with probability at least $1 - \delta$

Challenges with long-term action dependencies

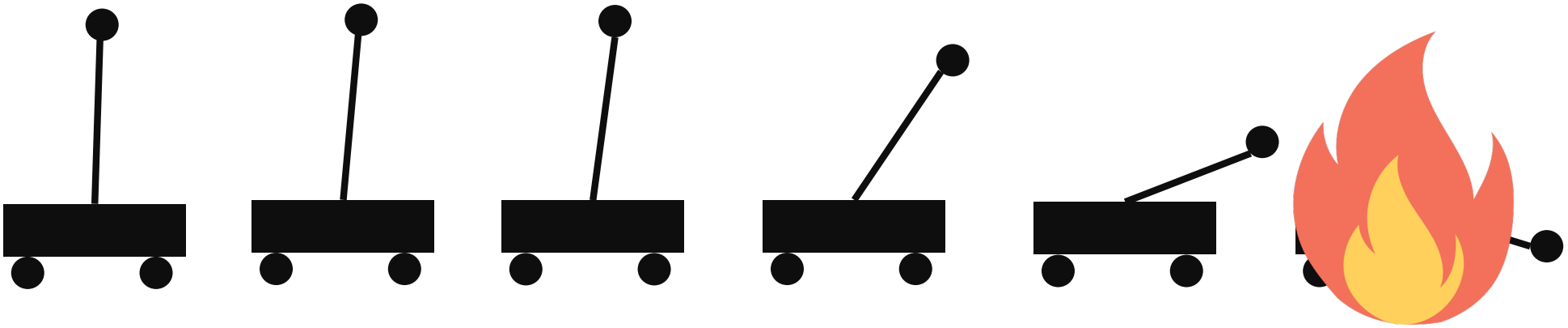
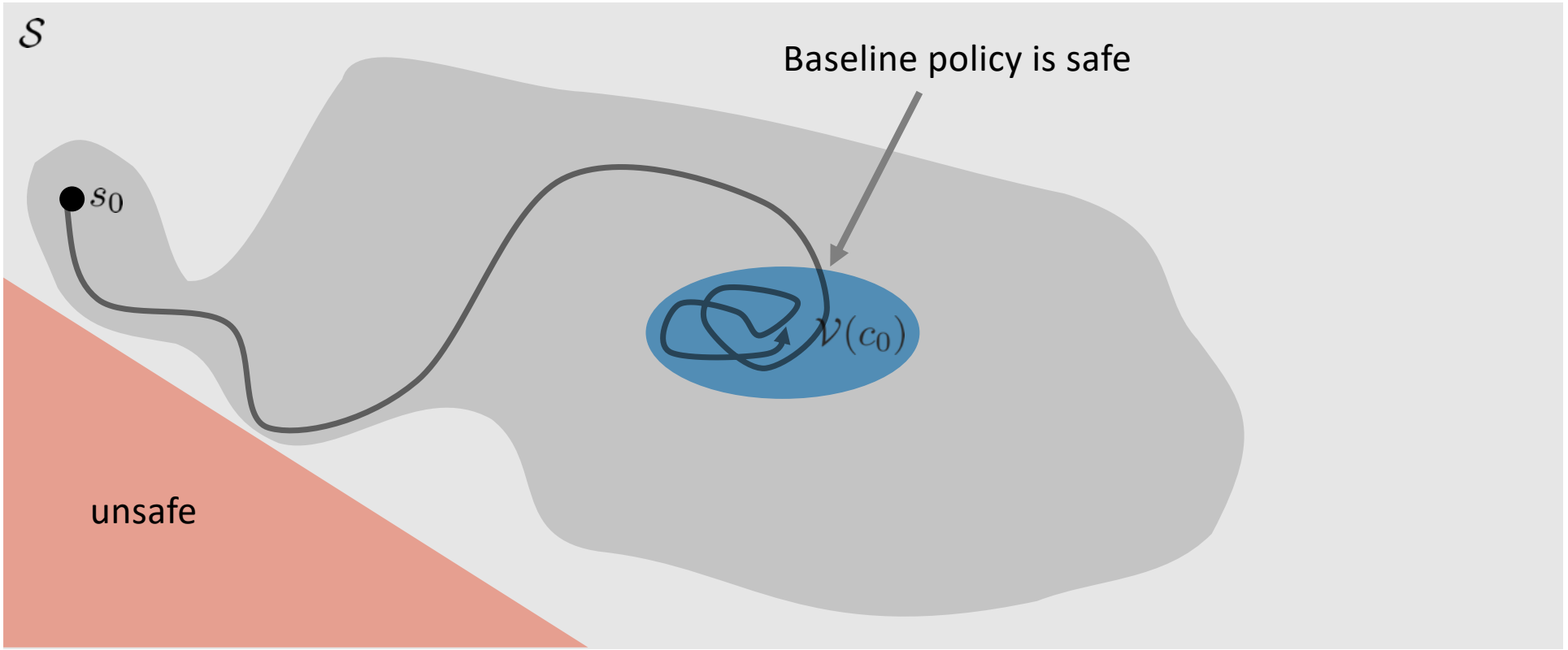


Image: Plainicon, VectorsMarket, https://flaticon.com

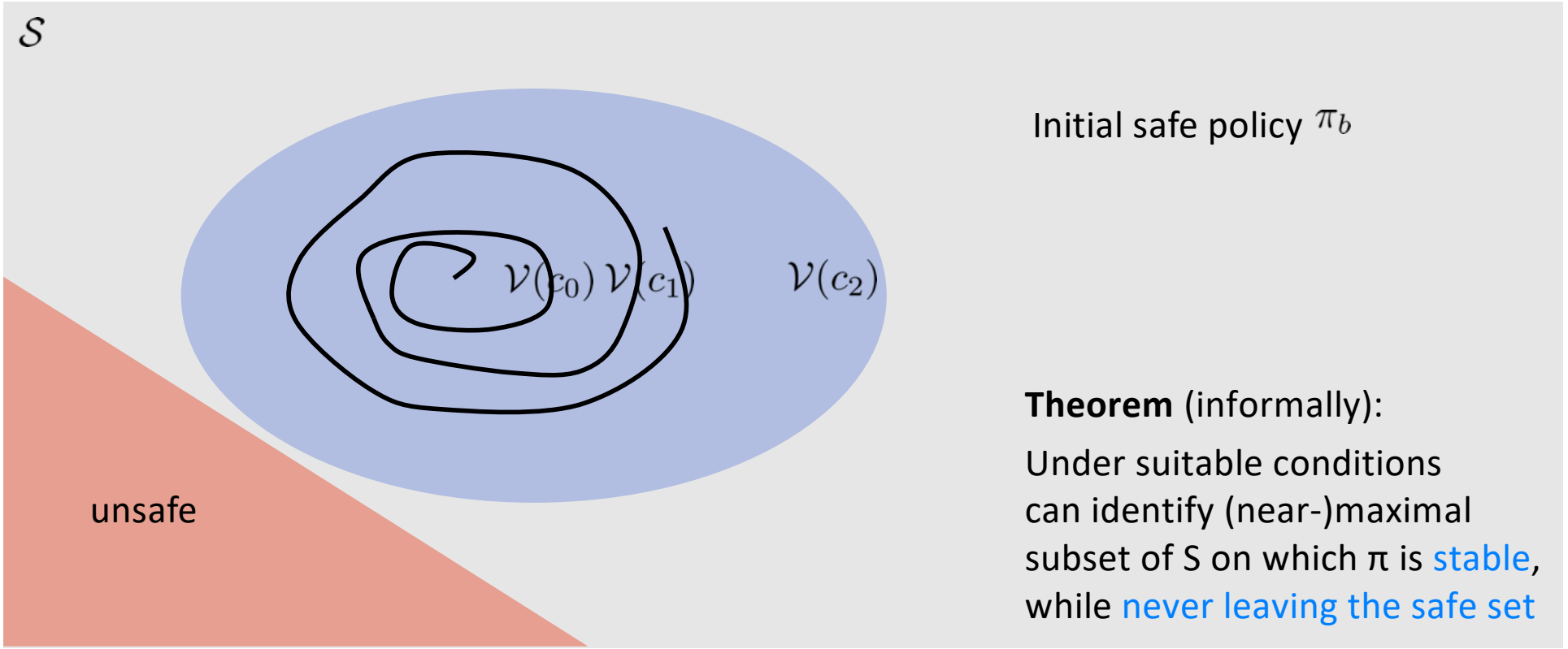
No return possible!

Can use confidence bounds for certifying long-term safety!

Region of attraction



Region of attraction



Safe Model-based Reinforcement Learning with Stability Guarantees
F. Berkenkamp, M. Turchetta, A.P. Schoellig, A. Krause, NIPS, 2017

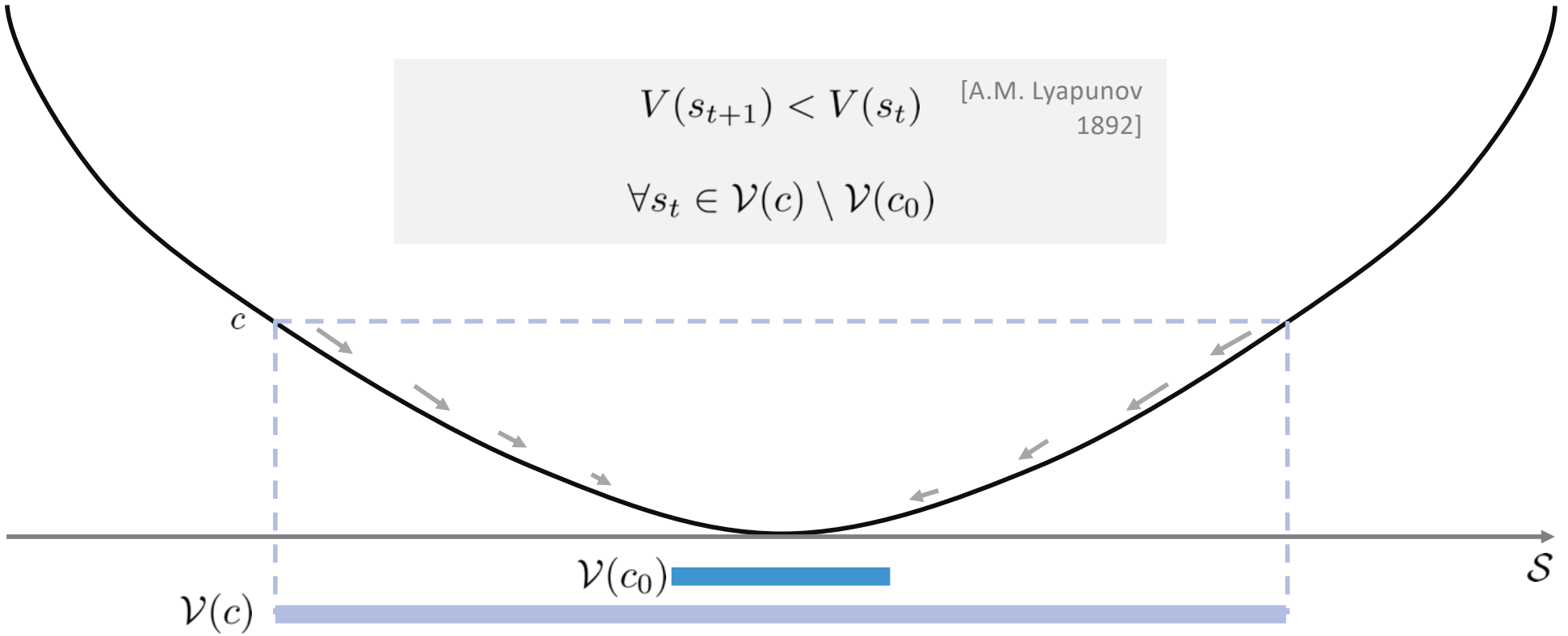
Lyapunov functions

$s_{t+1} = f(s_t, \pi(s, \theta))$

$V(s)$

$$V(s_{t+1}) < V(s_t) \qquad \text{[A.M. Lyapunov 1892]}$$

$$\forall s_t \in \mathcal{V}(c) \setminus \mathcal{V}(c_0)$$



Confidence-based Lyapunov analysis

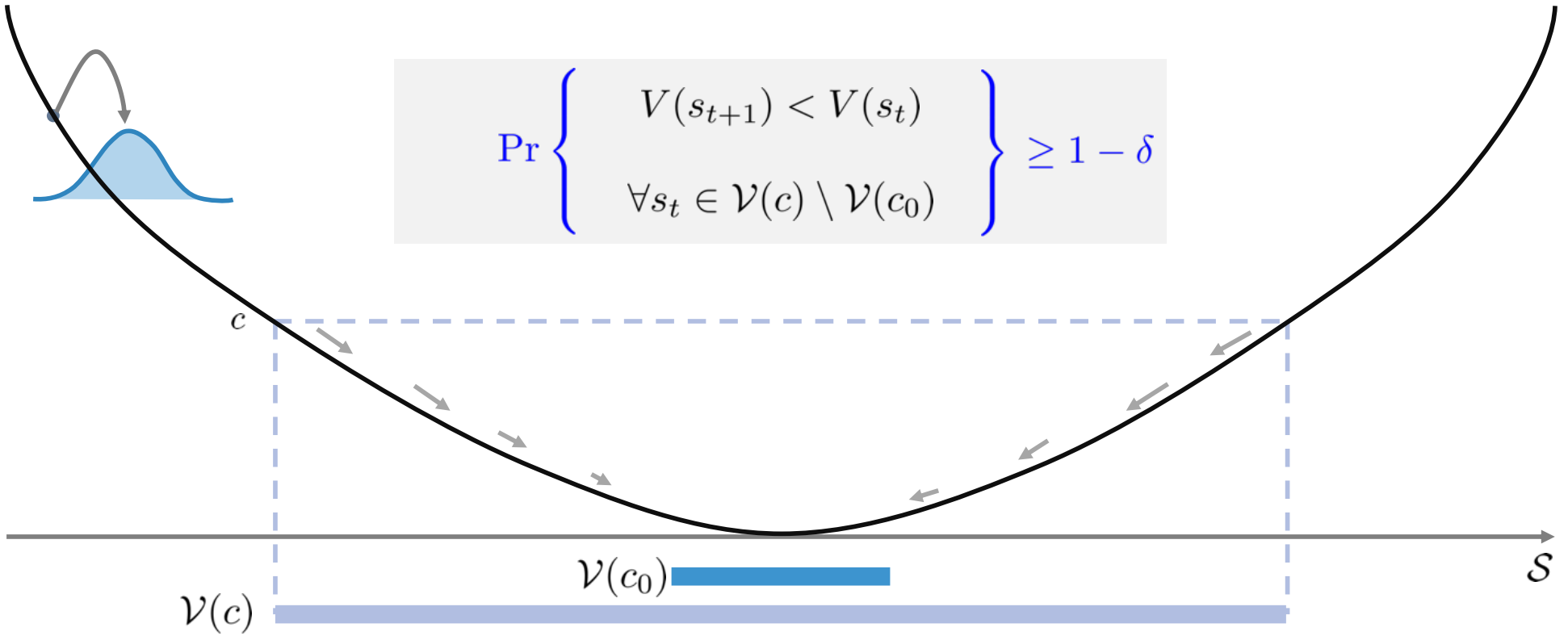
[Berkenkamp, Turchetta, Schoellig, K, NeurIPS 2017]

Felix Berkenkamp

Matteo Turchetta

$$s_{t+1} = f(s_t, \pi(s, \theta)) \; + \; g(s_t, \pi(s, \theta))$$

$$\Pr\left\{ \begin{array}{c} V(s_{t+1}) < V(s_t) \\ \\ \forall s_t \in \mathcal{V}(c) \setminus \mathcal{V}(c_0) \end{array} \right\} \geq 1 - \delta$$



Can also learn Lyapunov candidates via neural networks via reduction to classification

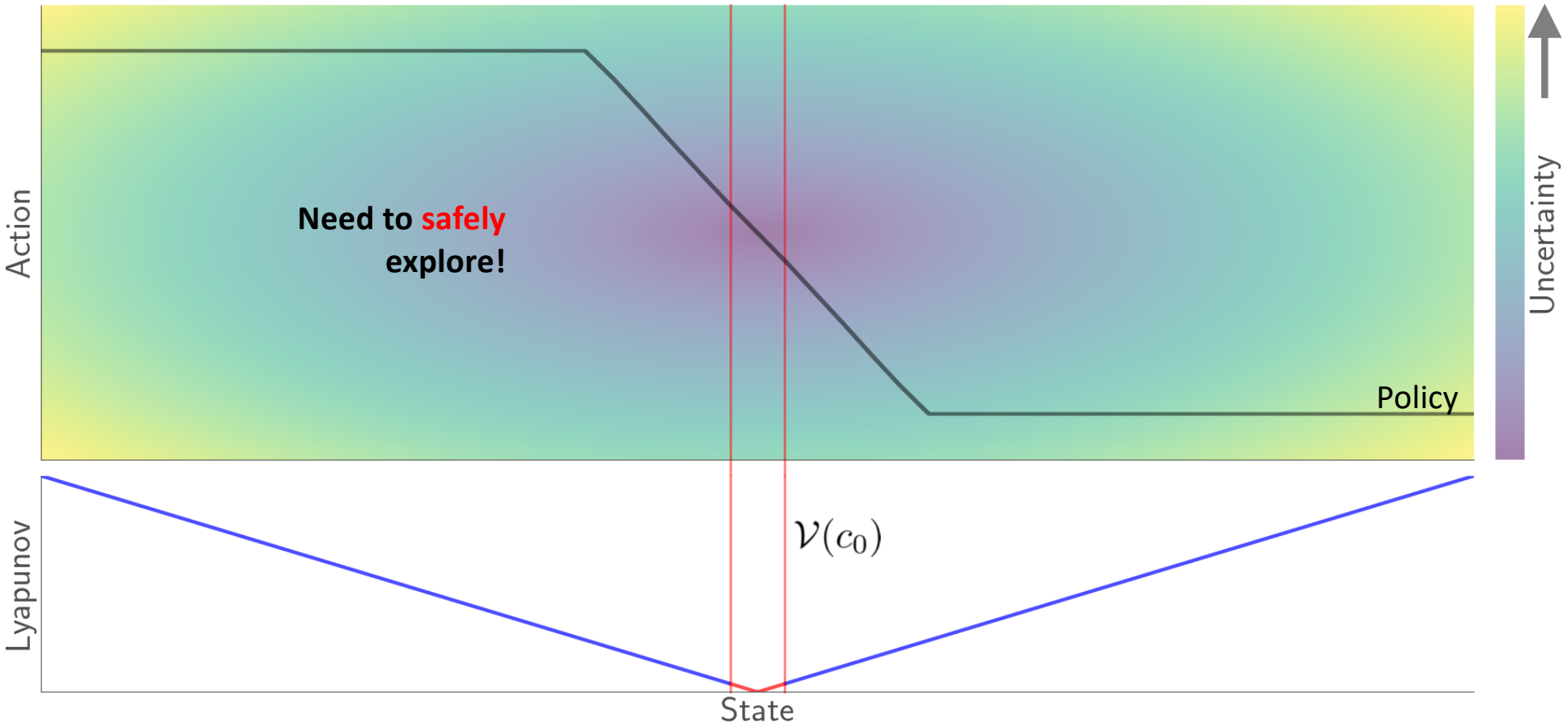
[Richards, Berkenkamp, K, CoRL '18]

Illustration of safe learning

[Berkenkamp, Turchetta, Schoellig, K, NeurIPS 2017]

Felix Berkenkamp

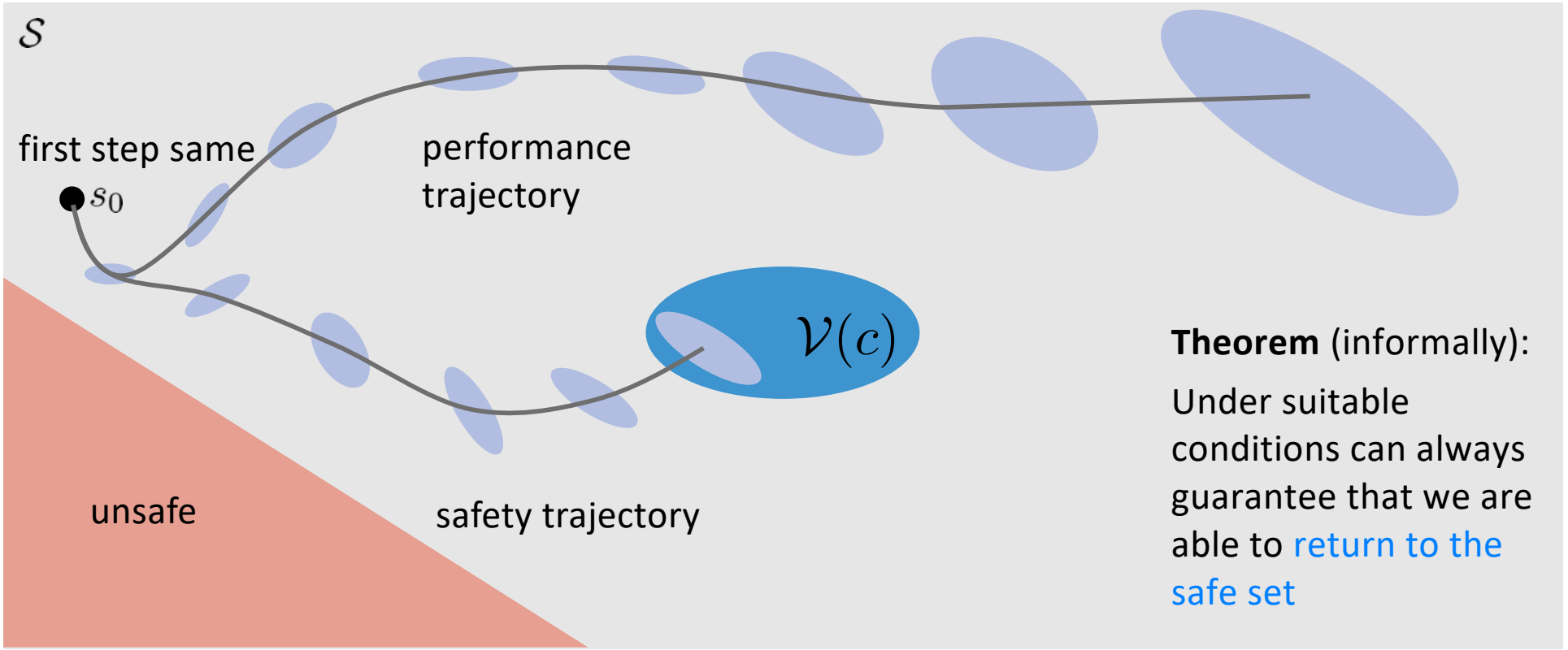
Matteo Turchetta



Safe learning-based MPC

[Koller, Berkenkamp, Turchetta, K CDC '18,'19]

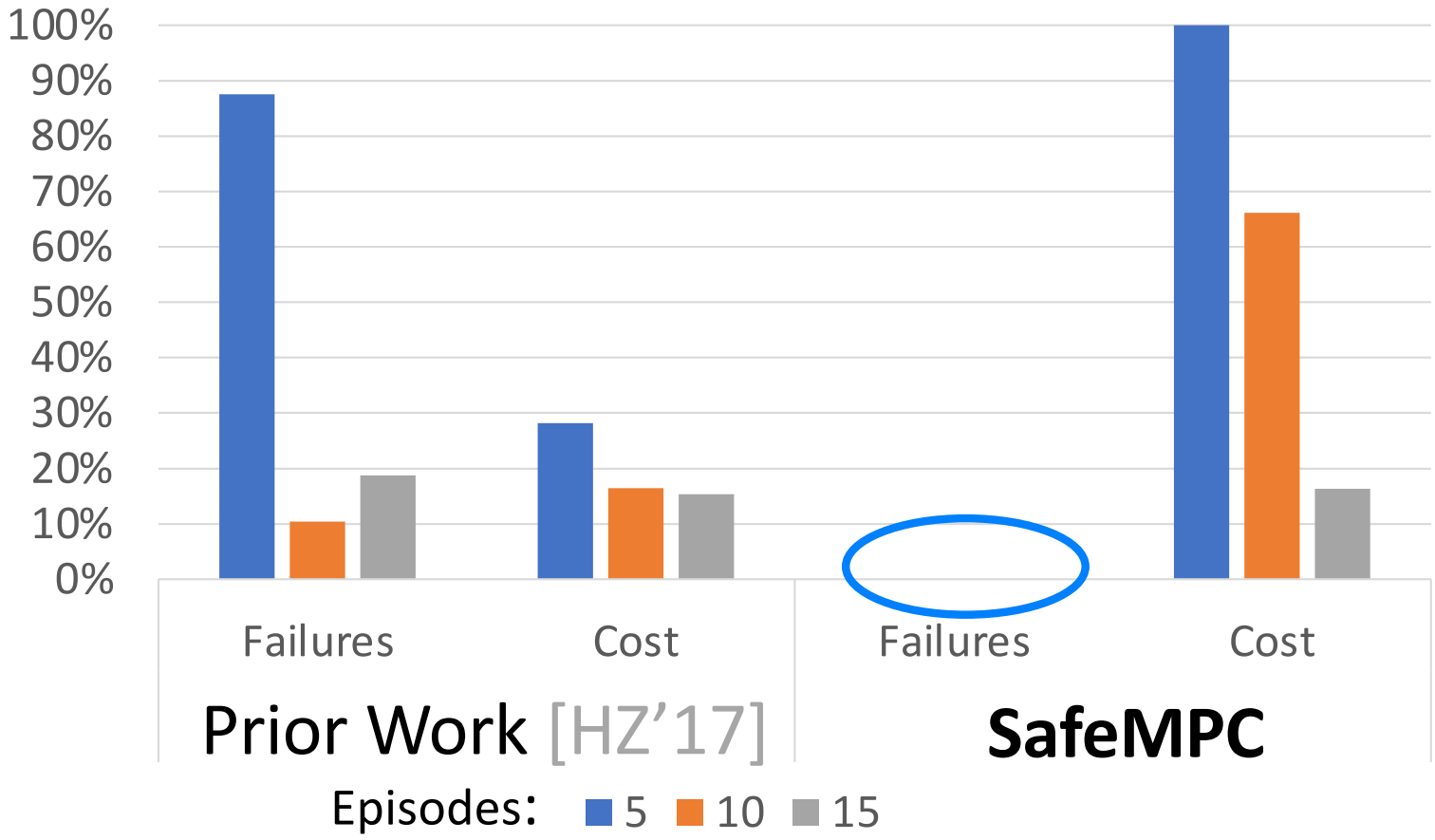
Torsten Koller | Felix Berkenkamp | Matteo Turchetta

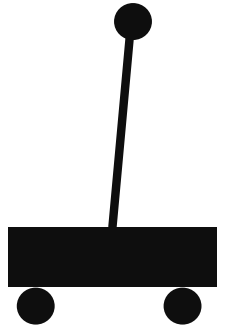


[c.f. Wabersich & Zeilinger '18]

Experiments

[Koller, Berkenkamp, Turchetta, K CDC '18, '19]





Scaling up: Efficient Optimistic Exploration in Deep Model-based Reinforcement Learning

Exploration in Model-based Deep RL

Hard to implement optimistic exploration in model-based deep reinforcement learning:

$$\max_{\pi} \max_{\tilde{f} \in \mathcal{M}_t} J(\tilde{f}, \pi)$$

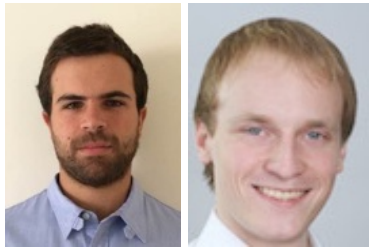
State-of-the-art approaches greedily optimize

$$\max_{\pi} \mathbb{E}_{\tilde{f}_t} J(\tilde{f}_t, \pi)$$

(e.g., PILCO, PETS, …)

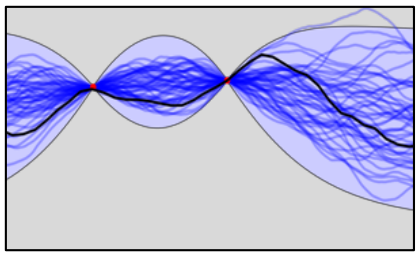
Deep Model-based RL with Confidence: H-UCRL

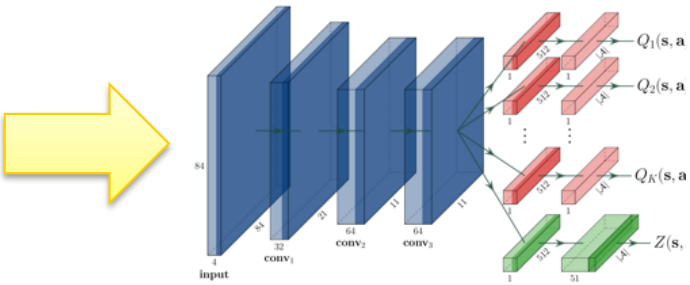
[Curi, Berkenkamp, K, NeurIPS 2020]

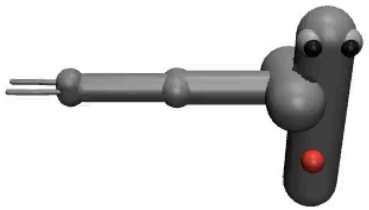


Sebastian Curi Felix Berkenkamp

- Hallucinate control authority bounded by model uncertainty
- Allows to use highly efficient policy search methods
- For GPs: Can prove sublinear regret bounds
- For high-dim apps: Use neural net ensembles for modeling

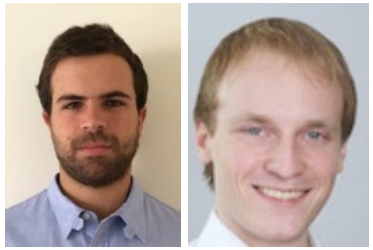




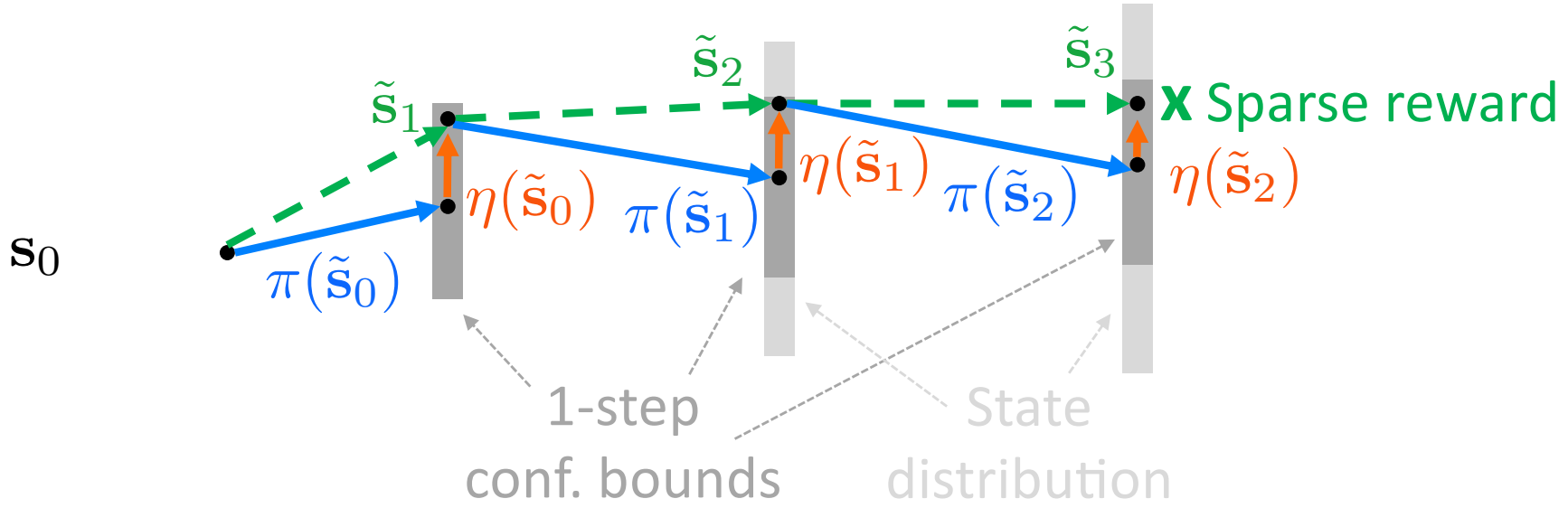


Deep Model-based RL with Confidence: H-UCRL

[Curi, Berkenkamp, K, NeurIPS 2020]



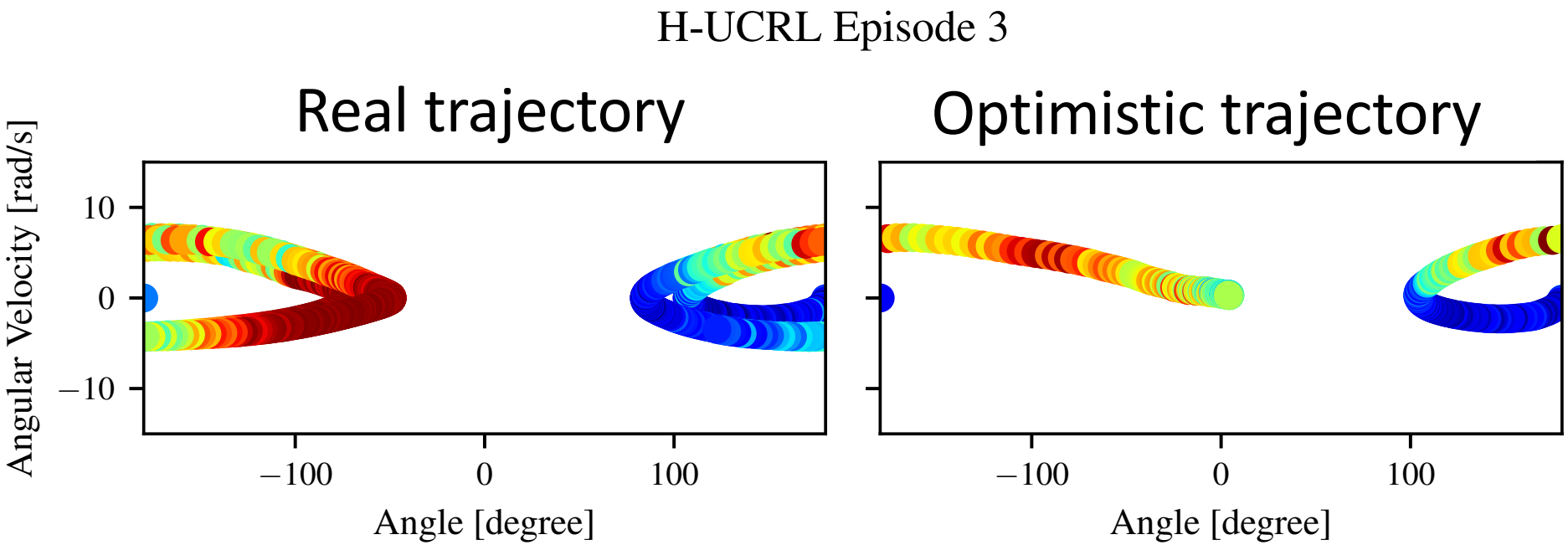
Sebastian Curi Felix Berkenkamp



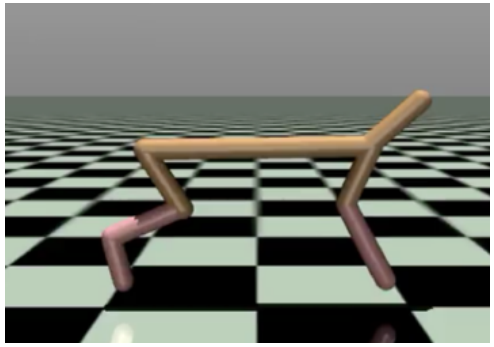
$$\pi_t^{\text{H-UCRL}} = \underset{\pi(\cdot)}{\operatorname{argmax}} \max_{\eta(\cdot)\in[-1,1]^p} J(\tilde{f}, \pi)$$

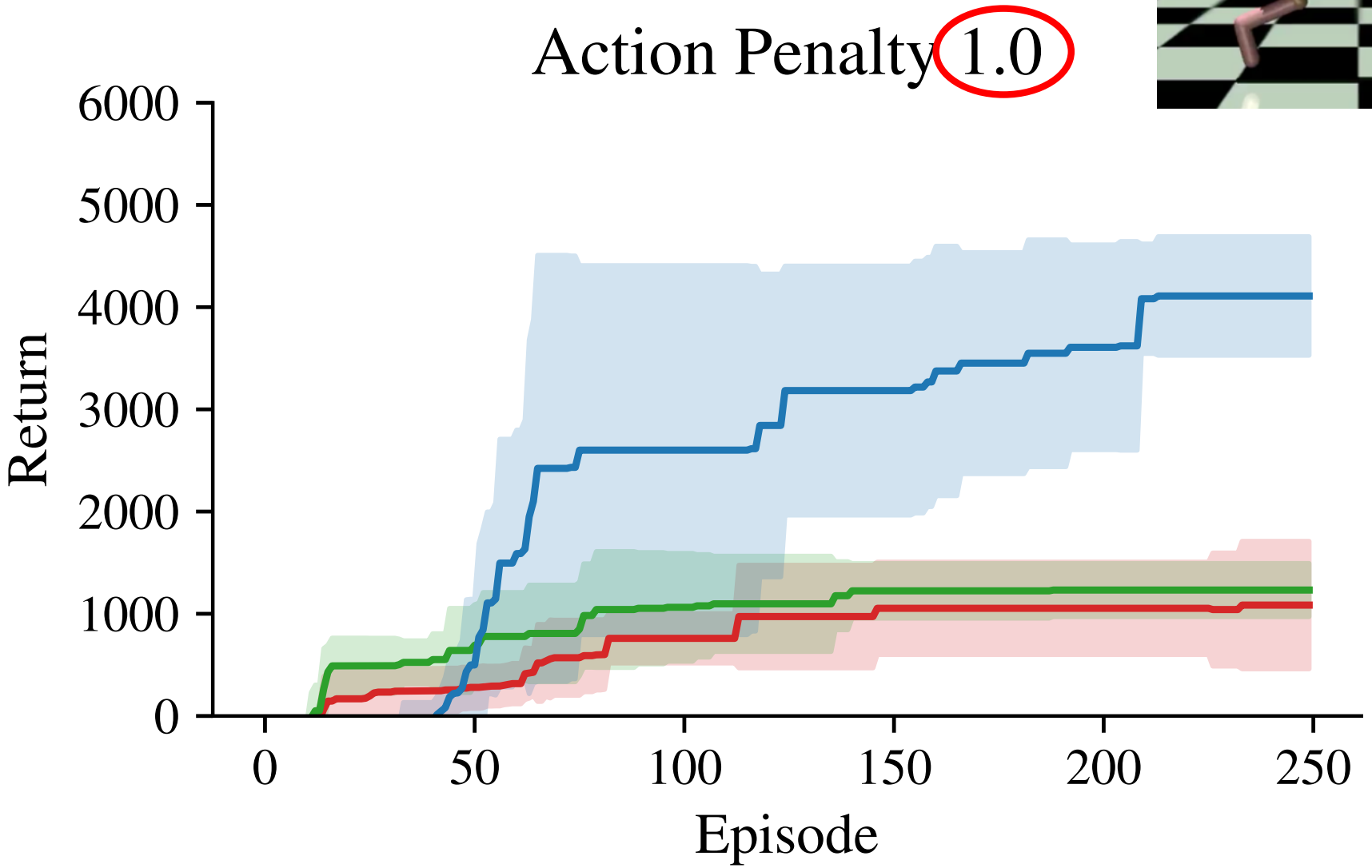
$$\text{s.t. } \tilde{f}(\mathbf{s},\mathbf{a}) = \mu_{t-1}(\mathbf{s},\mathbf{a}) + \beta_{t-1}\Sigma_{t-1}(\mathbf{s},\mathbf{a})\eta(\mathbf{s},\mathbf{a})$$

Illustration on Inverted Pendulum



Deep RL: Mujoco Half-Cheetah

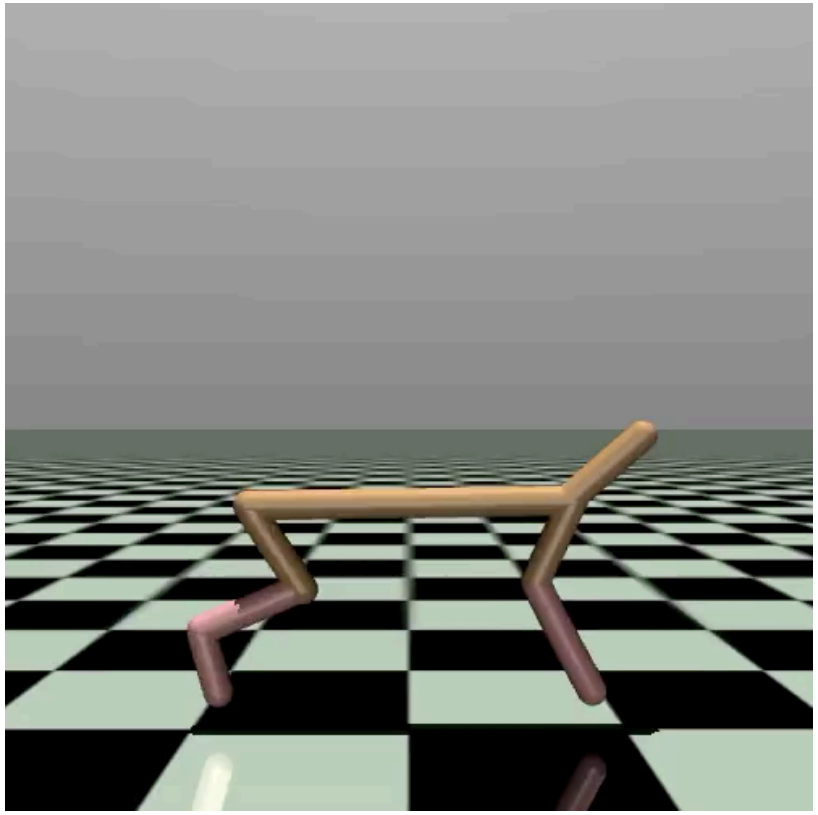


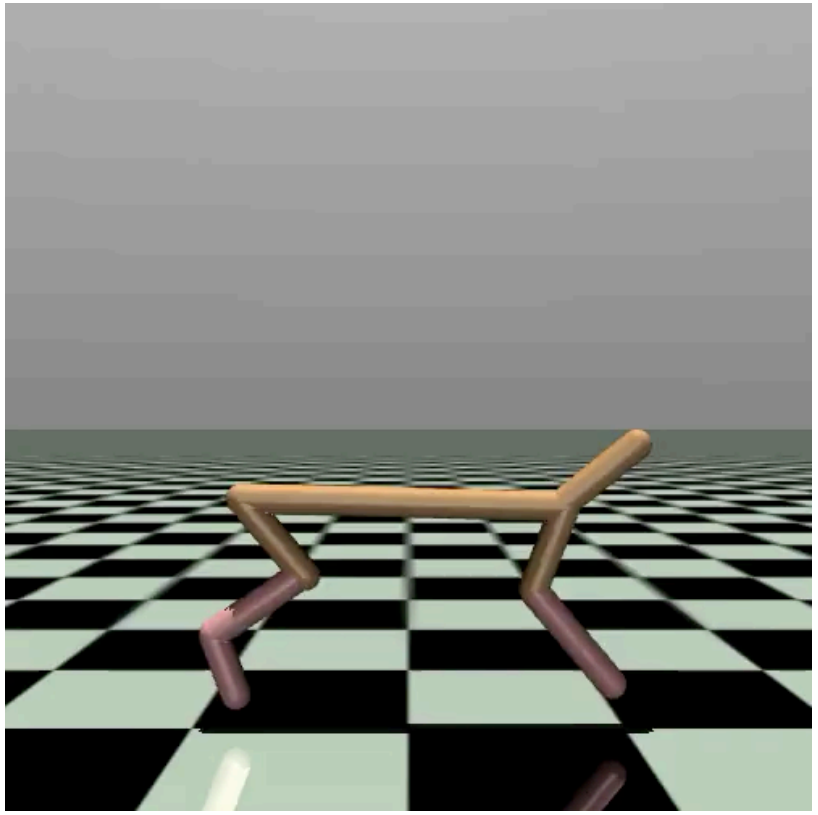


H-UCRL outperforms Greedy & Thompson sampling

Stronger effect for harder exploration tasks

Action penalty effect





Small action penalty:

➔ Unrealistic behaviors allowed

➔ Exploration easy

➔ Existing approaches work fine

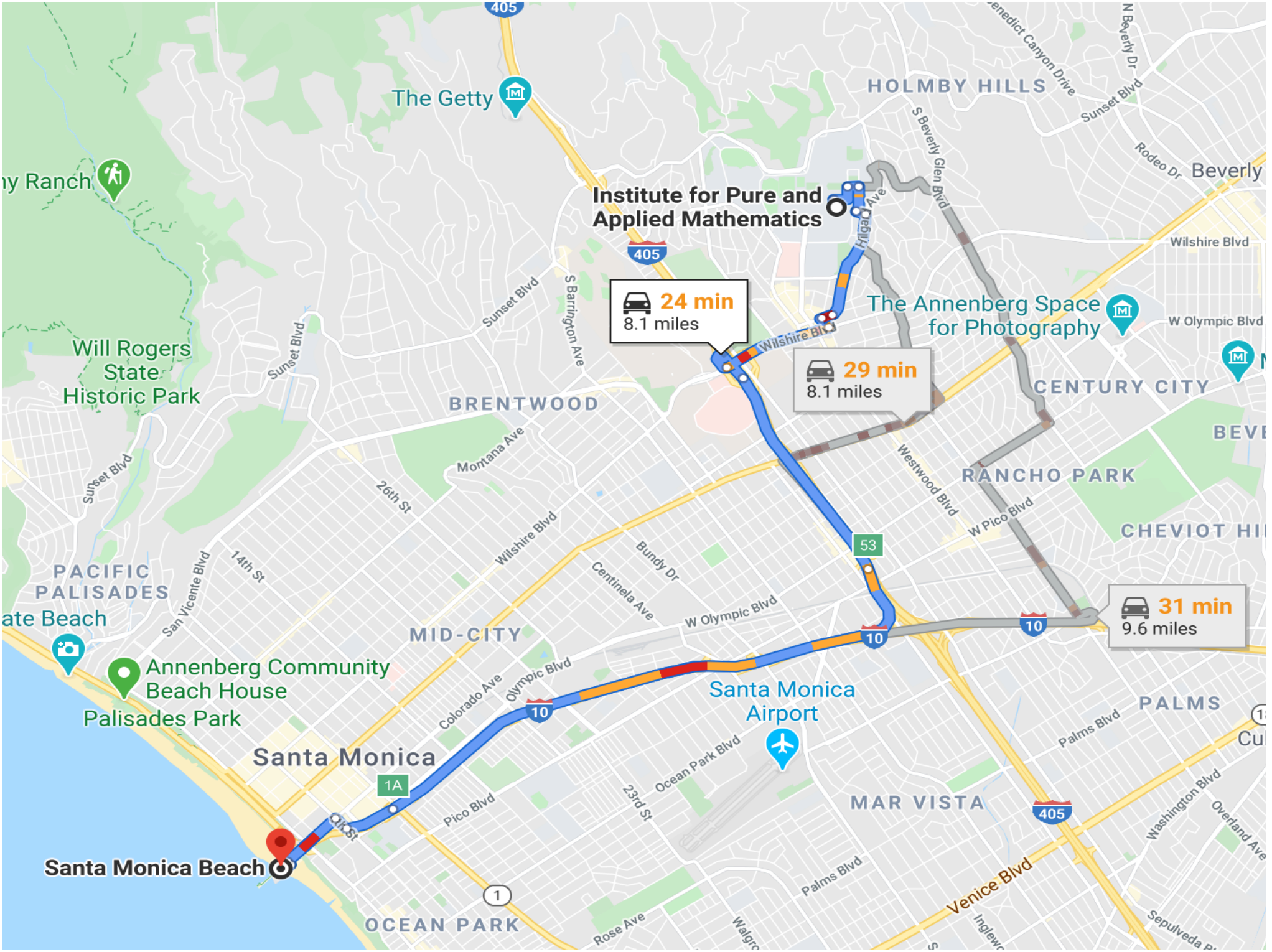
Large action penalty:

➔ Avoids aggressive controls

➔ Exploration hard

➔ H-UCRL still finds good policies

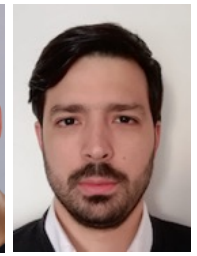
Efficient Learning in Games via “Optimistic Hallucination”


Institute for Pure and Applied Mathematics
24 min
8.1 miles
29 min
8.1 miles
31 min
9.6 miles
The Getty
HOLMBY HILLS
Beverly
Wilshire Blvd
The Annenberg Space for Photography
W Olympic Blvd
CENTURY CITY
Will Rogers State Historic Park
BRENTWOOD
RANCHO PARK
CHEVIOT HI
PACIFIC PALISADES
MID-CITY
Annenberg Community Beach House
Palisades Park
Santa Monica
Santa Monica Airport
PALMS
MAR VISTA
Santa Monica Beach
OCEAN PARK
Venice Blvd
405
10
53
1A
1
Sunset Blvd
Montana Ave
26th St
Wilshire Blvd
Bundy Dr
Centinela Ave
W Olympic Blvd
Olympic Blvd
Colorado Ave
Pico Blvd
Ocean Park Blvd
23rd St
Palms Blvd
Rose Ave
Westwood Blvd
W Pico Blvd
S Beverly Glen Blvd
Sunset Blvd
Rodeo Dr
S Barrington Ave
San Vicente Blvd
14th St
Washington Blvd
Overland Ave
Sepulveda

Sample-Efficient Multi-Agent Learning

[w P. Sessa, I. Bogunovic, M. Kamgarpour, NeurIPS 2019]


Pier G. Sessa

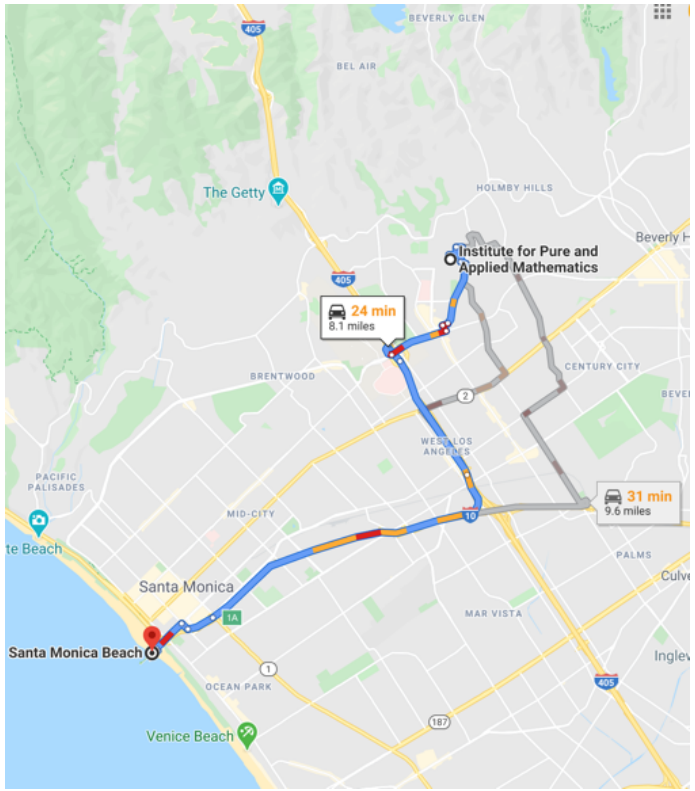

Ilija Bogunovic

Suppose agent *i* wants to learn to get to the beach quickly

Every day *t*, pick a route a_t^i

Reward depends on all agents' actions

$$r_t^i = f(a_t^i, a_t^{-i})$$



Want to minimize our regret

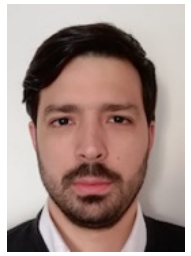
$$R_T = \max_a \sum_{t=1}^{T} r(a, a_t^{-i}) - \sum_{t=1}^{T} r(a_t^i, a_t^{-i})$$

Max in hindsight — What we got

Sample-Efficient Multi-Agent Learning

[w P. Sessa, I. Bogunovic, M. Kamgarpour, NeurIPS 2019]


Pier G. Sessa


Ilija Bogunovic

At every round of the game, agent

i : 1. Obtains reward $r_t^i = f(a_t^i, a_t^{-i})$

2. Updates her strategy based on **feedback**

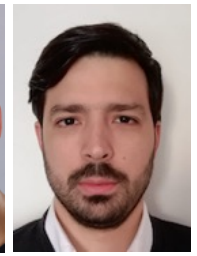

agent-i


rest of the agents

	Bandit Feedback	**Full-Info. Feedback**	**Proposed feedback model**
Observed:	r_t^i	$[f(a_1, a_t^{-i}), \cdots, f(a_K, a_t^{-i})]$	$r_t^i + \text{noise}$, a_t^{-i}
Regret:	$O(\sqrt{TK \log K})$	$O(\sqrt{T \log K})$	$O(\sqrt{T \log K} + \gamma_T \sqrt{T})$
	EXP3 [Auer et al. '02]	MW [Freund and Schapire '97]	
	Scales poorly with K # of available actions ☹	Not realistic, since $f(\cdot,\cdot)$ is **unknown** ☹	Some aggregate info about a_t^{-i} can also be enough ☺

Key Insights


Pier G. Sessa

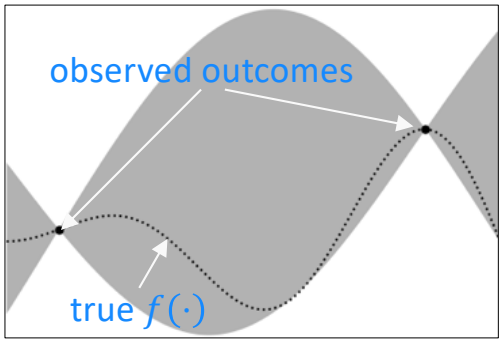

Ilija Bogunovic

1. Often similar game outcomes produce similar rewards

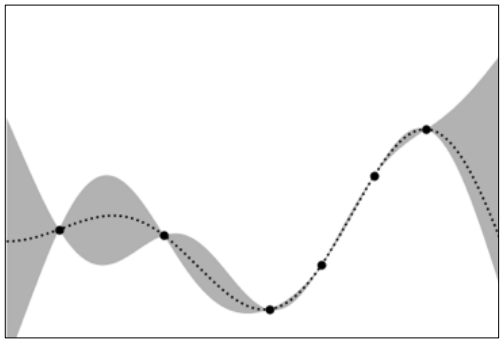
   → We assume $f(\cdot)$ has a **bounded RKHS norm** w.r.t. a kernel function k

2. Then agent-i can use the history of play $\{r_1^i, a_1^i, a_1^{-i}, \dots, r_{t-1}^i, a_{t-1}^i a_{t-1}^{-i}\}$ to build **shrinking confidence bounds** on $f(\cdot)$:

   - t = 3:



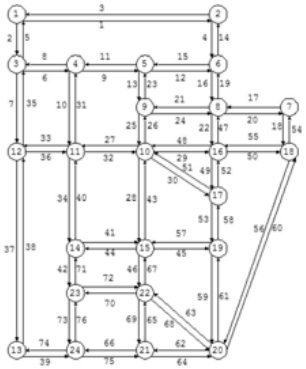
   - t = 7:



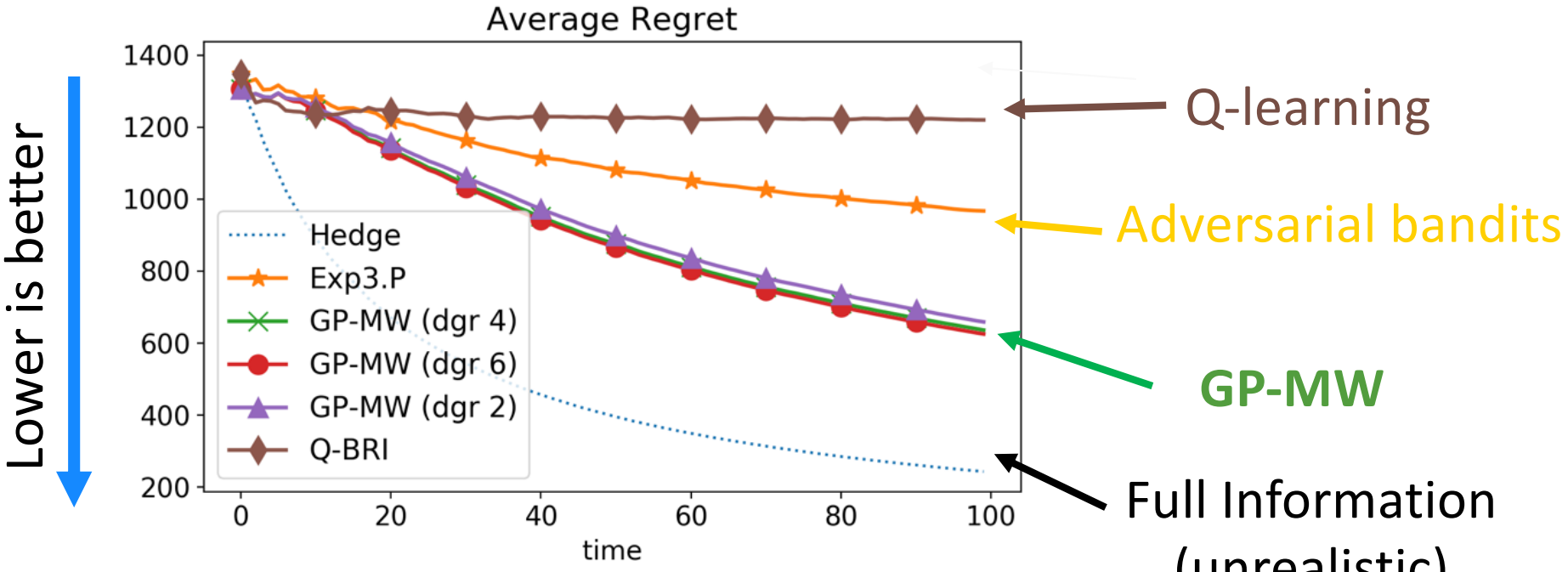
3. The proposed algorithm GP-MW, **emulates the full-info. feedback** using an **Upper Confidence Bound** on $f(\cdot)$

Learning to route

- $N = 100$ 'learning' agents



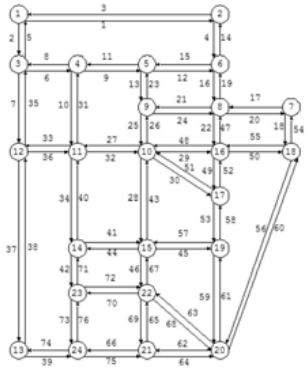
Sioux-Falls Network



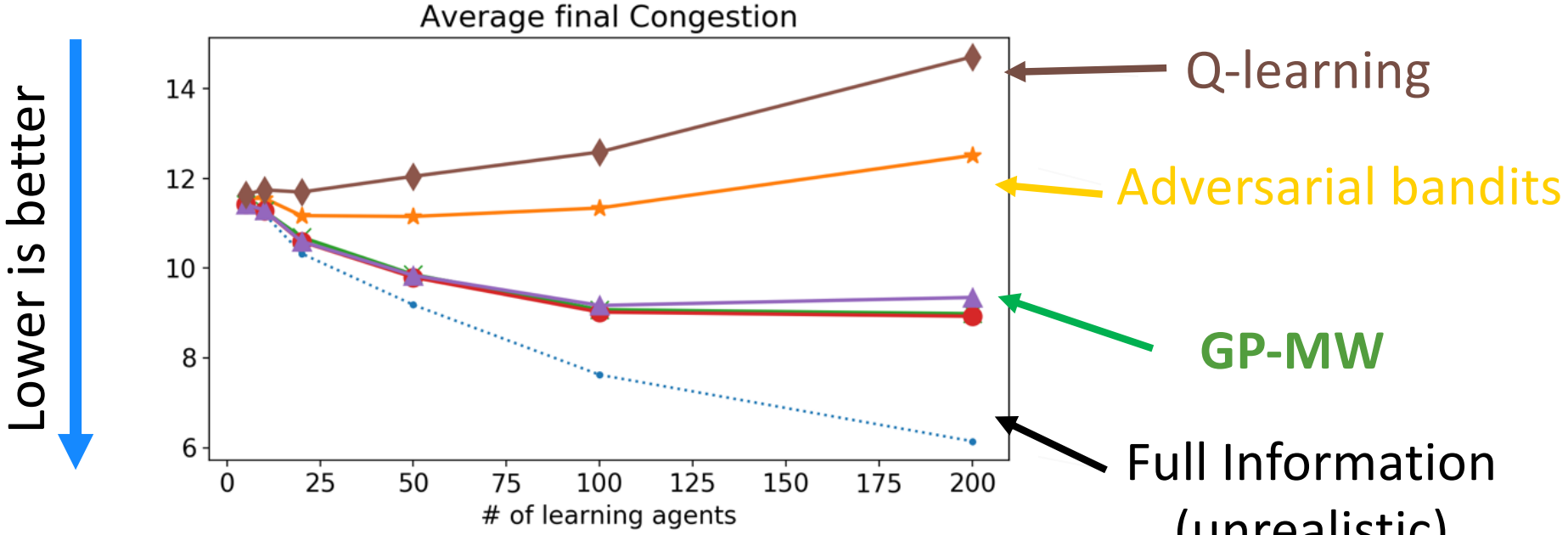
GP-MW leads to smaller average regret,
(averaged over the learning agents)

Learning to route

- $N = 100$ 'learning' agents



Sioux-Falls Network



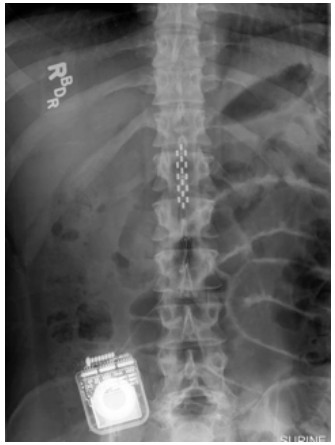
GP-MW reduces the congestion in the network, (averaged over the network edges)

Results extend to the contextual setting, sequential games, … [NeurIPS '20]

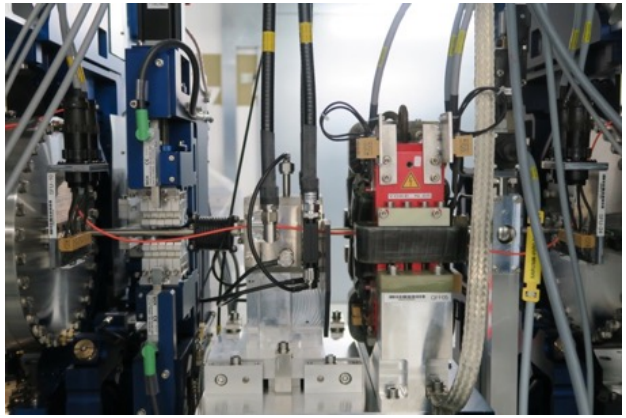
Towards RL in Real World Applications

[Grendelkhan]

[AleSpa]



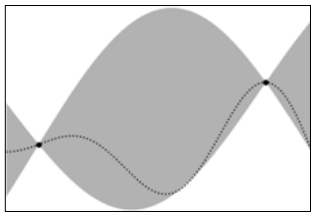
[S. Harkema, The Lancet]



. . .

Summary

Nonparametric confidence bounds



U_x

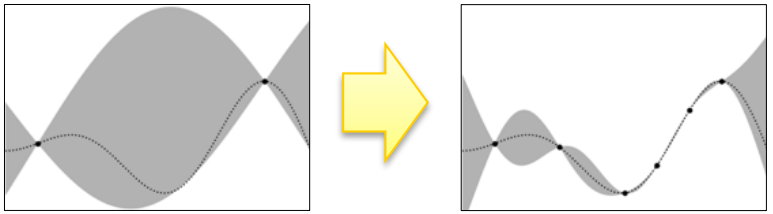
+

Robust optimization / verification

$$\min_{\mathbf{x}} \max_{\delta \in U_{\mathbf{x}}} f(\mathbf{x}, \delta)$$

+

Safe exploration / identification



=

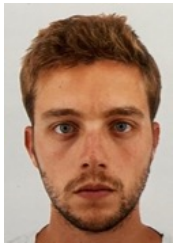
Learning-based performance gains with safety guarantees

More Directions & Open Questions

- How to deal with partial observability?
  - Bandits with context distributions [→ NeurIPS 2019]
  - Linear / kernelized partial monitoring [→ COLT 2020]
- How to deal with adversarial perturbations / misspecification and distribution shifts?
  - Corrupted kernelized bandits [→ AISTATS 2020]
  - Distributionally robust Bayesian Optimization [→ AISTATS'20]
- How to select the prior?
  - SRM-based online model tuning [→ JMLR 2019]
  - Meta-learning [→ arXiv 2020]

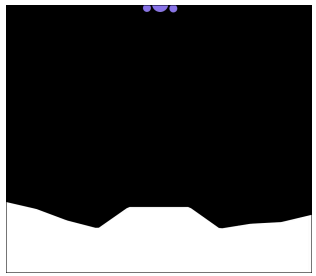
Curriculum Induction for Safe Reinforcement Learning (CISR)

[w Turchetta, Kolobov, Shah, Agarwal, NeurIPS'20]

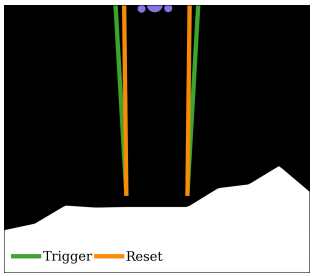


Matteo Turchetta

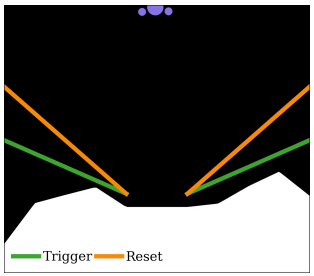
Original



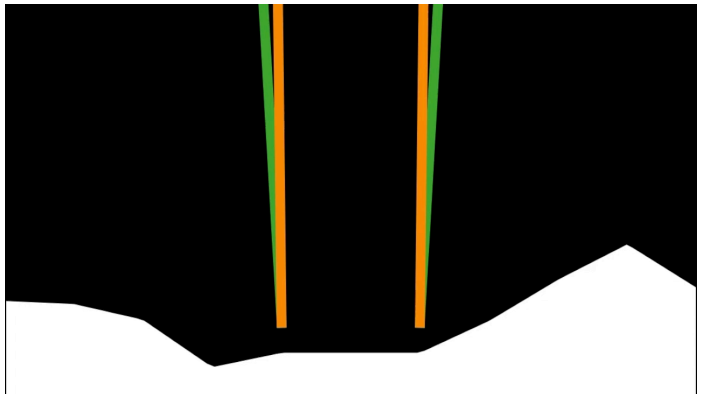
Narrow



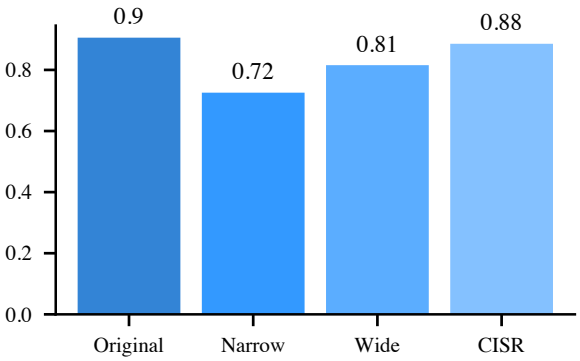
Wide



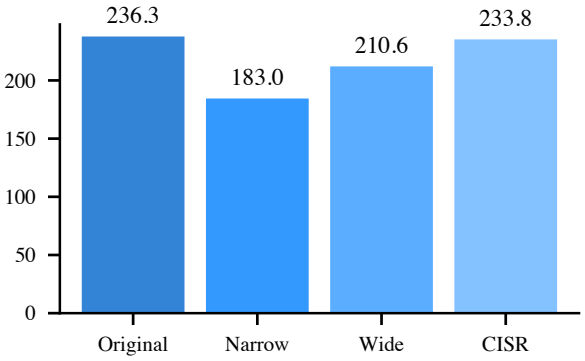
Training with optimized curriculum



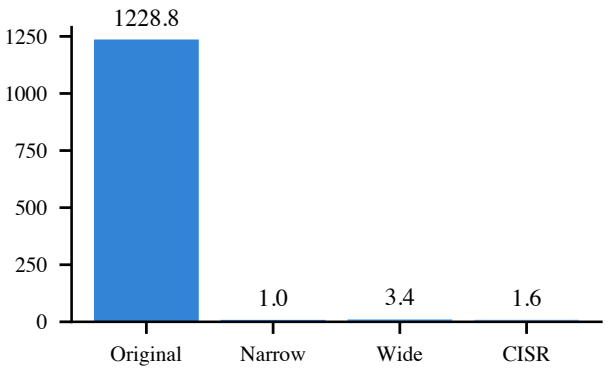
Successes



Returns



Safety violations

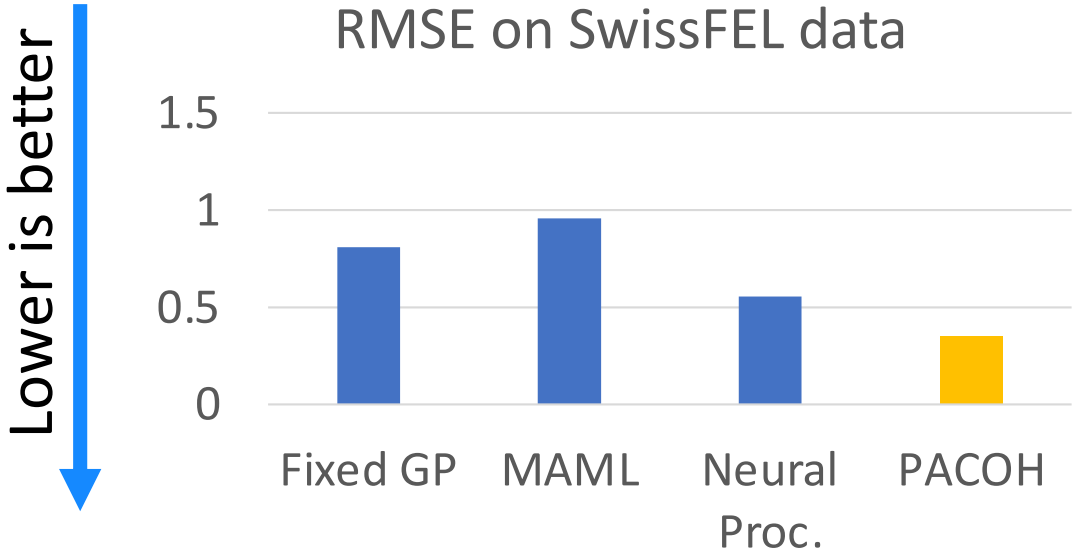


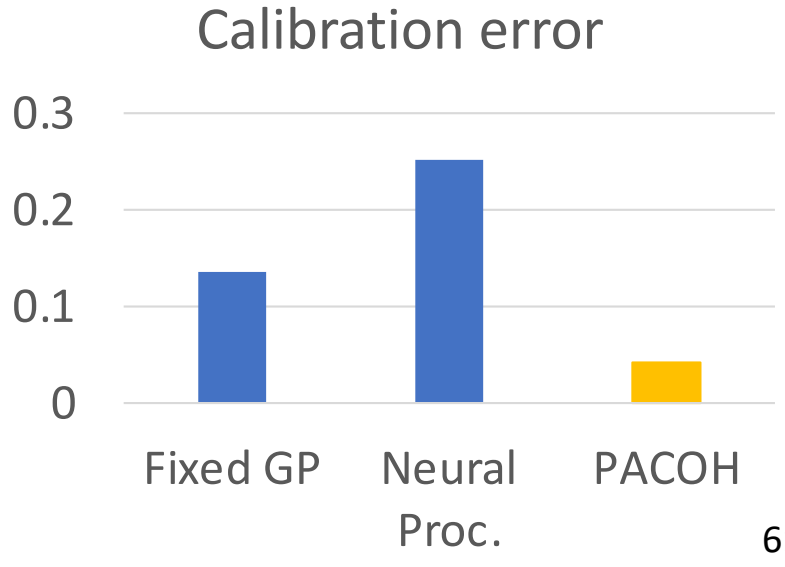
PAC-Bayesian Meta Learning

[Rothfuss, Fortuin, K, arXiv 2020]

Jonas Rothfuss | Vincent Fortuin

- Meta-learn (Gaussian process) priors from related tasks
- Parametrize mean and covariance via neural network
- Minimize PAC-Bayesian bound to avoid (meta)-overfitting and obtain PAC Optimal Hyperposterior (PACOH)





- N. Srinivas, A. Krause, S. Kakade, M. Seeger. Information-Theoretic Regret Bounds for Gaussian Process Optimization in the Bandit Setting. *IEEE Transactions on Information Theory, 2012*
- F. Berkenkamp, A. P. Schoellig, A. Krause. Safe Controller Optimization for Quadrotors with Gaussian Processes, *ICRA 2016*
- F. Berkenkamp, M. Turchetta, A. P. Schoellig, A. Krause. Safe Model-based Reinforcement Learning with Stability Guarantees, *NeurIPS 2017*
- T. Koller, F. Berkenkamp, M. Turchetta, A. Krause. Learning-based Model Predictive Control for Safe Exploration and Reinforcement Learning, *CDC, 2018; arXiv 2019*
- S. M. Richards, F. Berkenkamp, A. Krause. The Lyapunov Neural Network: Adaptive Stability Certification for Safe Learning of Dynamical Systems, *CoRL 2018*
- M. Fiducioso, S. Curi, A. Krause, M. Gwerder, B. Schumacher. Safe Contextual Bayesian Optimization for Sustainable Room Temperature PID Control Tuning, *IJCAI 2019*
- J. Kirschner, M. Mutny, N. Hiller, R. Ischebeck, A. Krause. Adaptive and Safe Bayesian Optimization in High Dimensions via One-Dimensional Subspaces, *ICML* 2019
- I. Bogunovic, A. Krause, J. Scarlett. Corruption-Tolerant Gaussian Process Bandit Optimization. *AISTATS '20*
- P. G. Sessa, I. Bogunovic, A. Krause, M. Kamgarpour. Contextual Games: Multi-Agent Learning with Side Information, *NeurIPS 2020*
- P. G. Sessa, I. Bogunovic, M. Kamgarpour, A. Krause. No-Regret Learning in Unknown Games with Correlated Payoffs. *NeurIPS 2020*
- S. Curi, F. Berkenkamp, A. Krause. Efficient Model-based Reinforcement Learning through Optimistic Policy Search and Planning. *NeurIPS 2020*
- J. Rothfuss, V. Fortuin, A. Krause PACOH: Bayes-Optimal Meta-Learning with PAC-Guarantees, *arXiv '20*
- M. Turchetta, A. Kolobov, S. Shah, A. Krause, A. Agarwal: Safe Reinforcement Learning via Curriculum Induction, NeurIPS 2020

➔ Tutorial at CoRL 2020

Acknowledgments

Senior Collaborators: Joel Burdick, Nicole Hiller, Rasmus Ischebeck, Sham Kakade, Andrei Kolobov, Jonathan Scarlett, Angela Schoellig, Matthias Seeger, Shital Shah, Alekh Agarwal

Funding