

Tutorial on Safe Reinforcement Learning

Felix Berkenkamp, Andreas Krause

@EWRL, October 1 2018

Reinforcement Learning (RL)

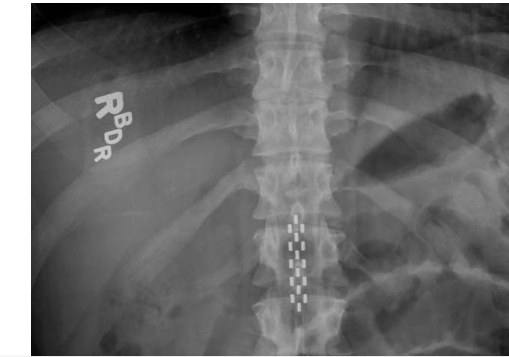


Need to trade exploration & exploitation

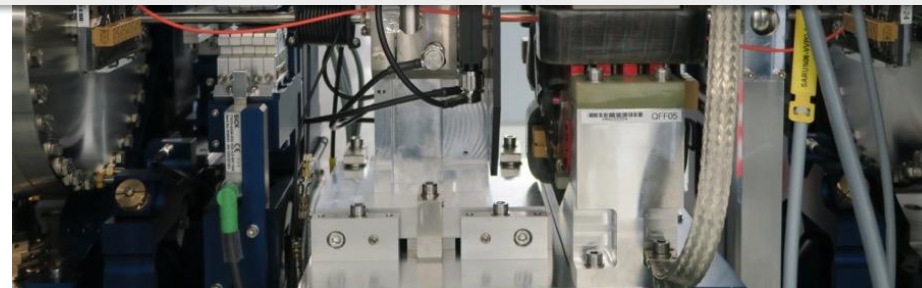
Reinforcement Learning: An Introduction

R. Sutton, A.G. Barto, 1998





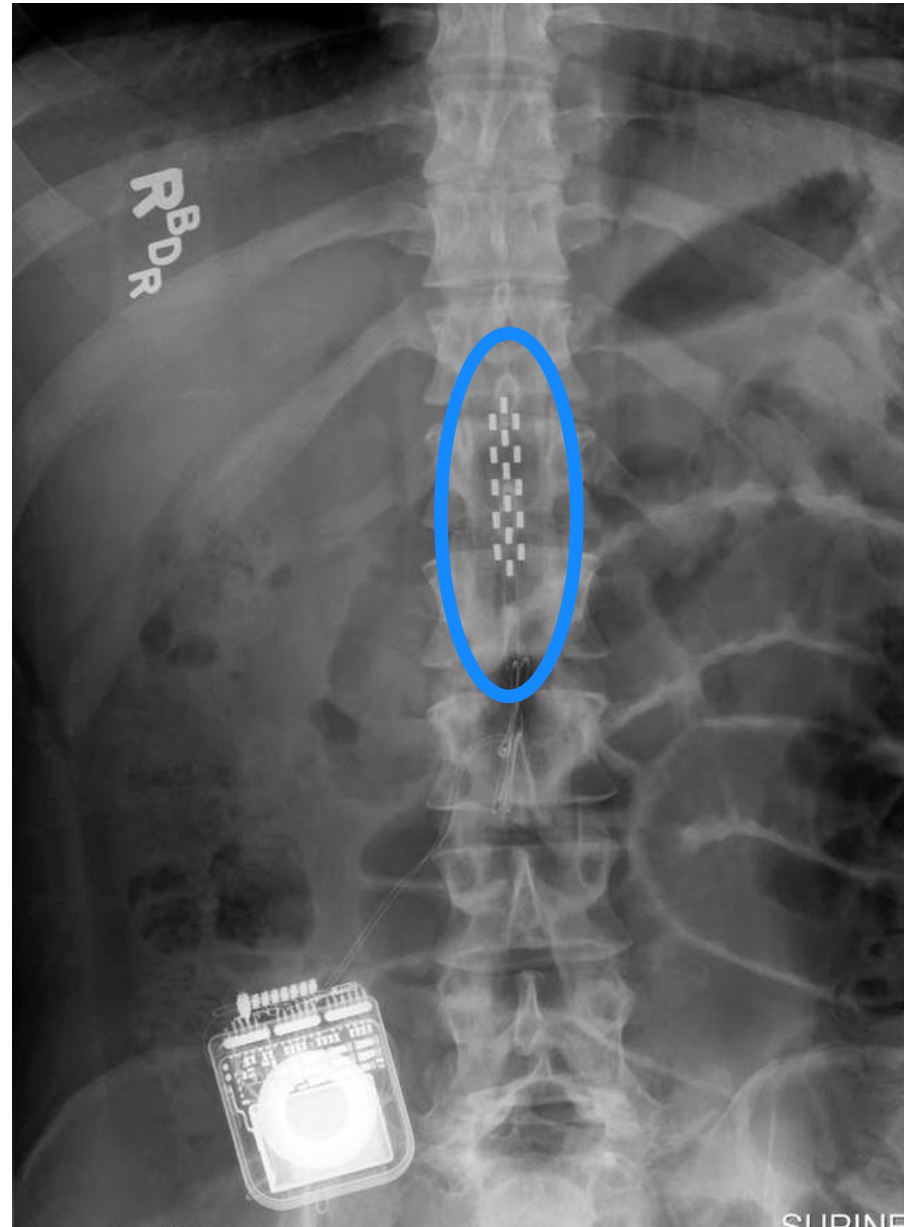
How can we *learn* to act
safely in unknown environments?



Therapeutic Spinal Cord Stimulation



girardgibbs.com



S. Harkema, The Lancet, Elsevier

Safe Exploration for Optimization with Gaussian Processes

Y. Sui, A. Gotovos, J. W. Burdick, A. Krause

Stagewise Safe Bayesian Optimization with Gaussian Processes

Y. Sui, V. Zhuang, J. W. Burdick, Y. Yue



Safe Controller Optimization for Quadrotors
with Gaussian Processes
F. Berkenkamp, A. P. Schoellig, A. Krause, ICRA 2016

Outline

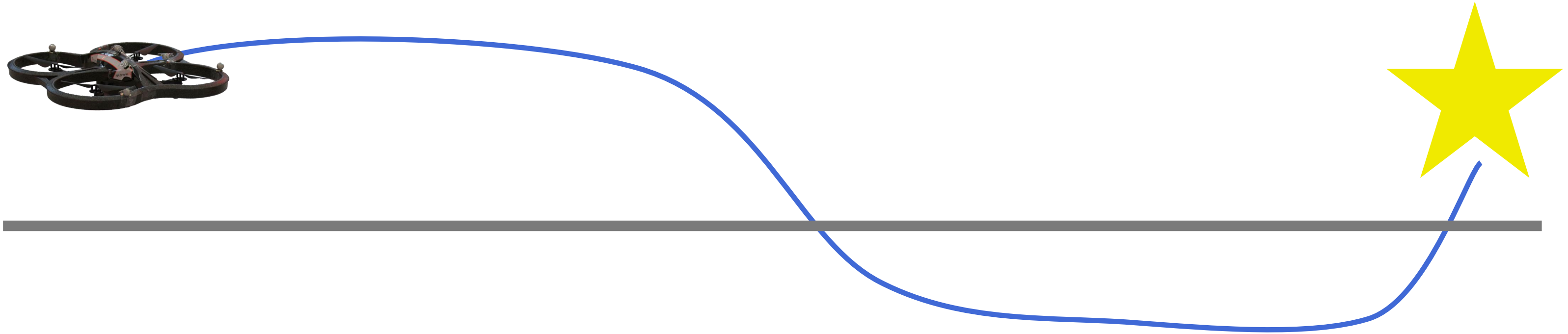
Specifying safety requirements and quantify risk

Acting safely in *known* environments

Acting safely in *unknown* environments

Safe exploration (model-free and model-based)

Specifying safe behavior



Is this trajectory safe?

$$g(\{s_t, a_t\}_{t=0}^N) = g(\tau) > 0$$

e.g. $g(\tau) = \min_{t=1:N} \Delta(s_t, a_t)$

Monitoring temporal properties of continuous signals

O. Maler, D. Nickovic, FT, 2004

Safe Control under Uncertainty

D. Sadigh, A. Kapoor, RSS, 2016

What does it mean to be safe?

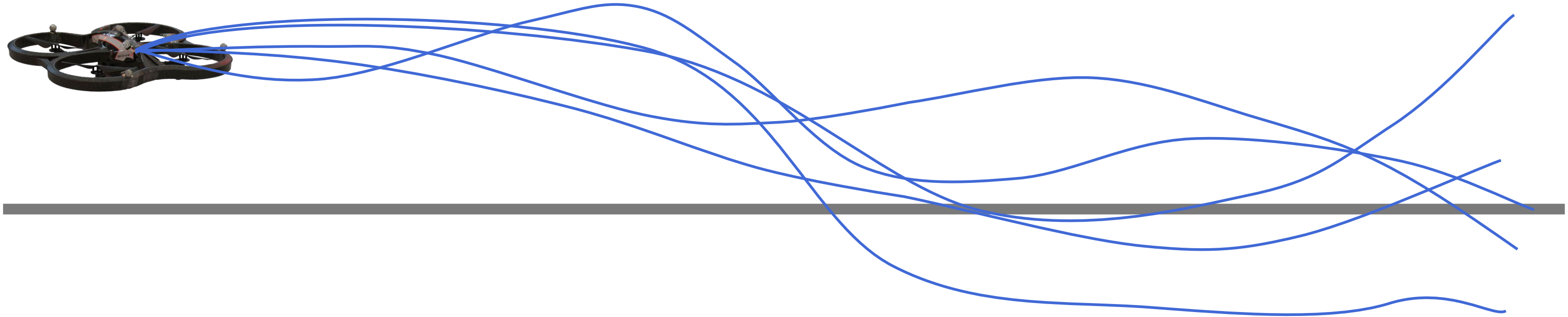


Safety \cong avoid bad trajectories (states/actions) $g(\{s_t, a_t\}_{t=0}^N) > 0$

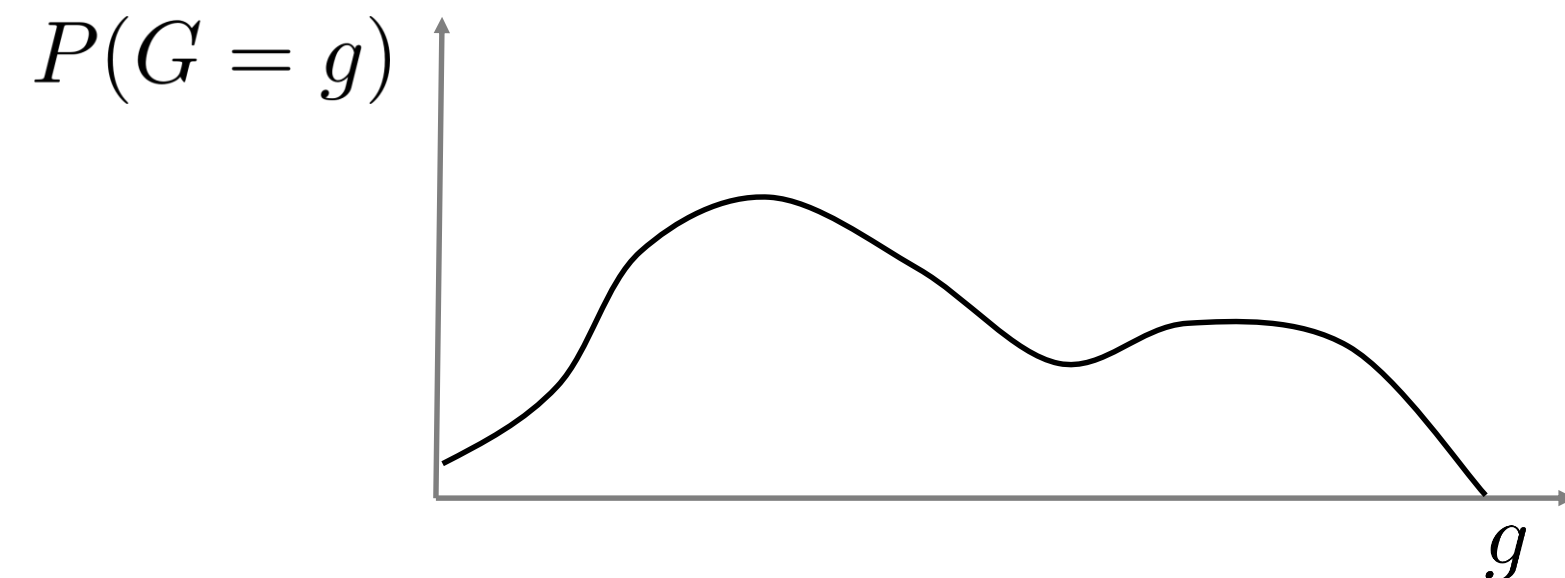
Fix a policy $a_t = \pi(s_t, \theta)$

How do I quantify uncertainty and risk?

Stochastic environment / policy



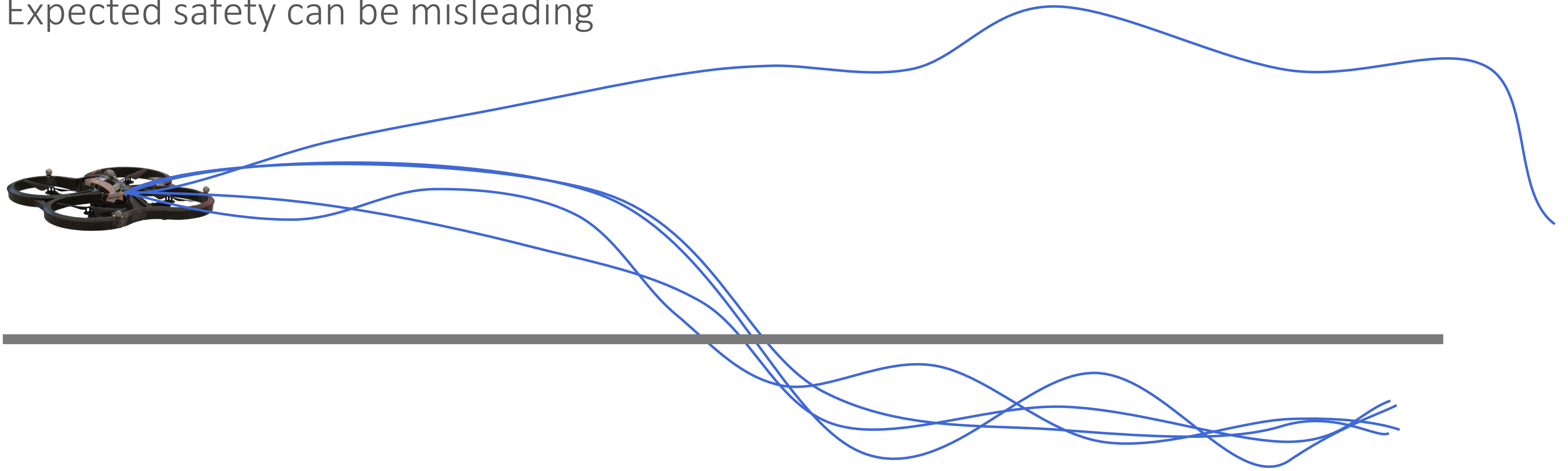
Safety function $g(\tau) \geq 0$ is now a random variable G



Expected safety

$$E[G]$$

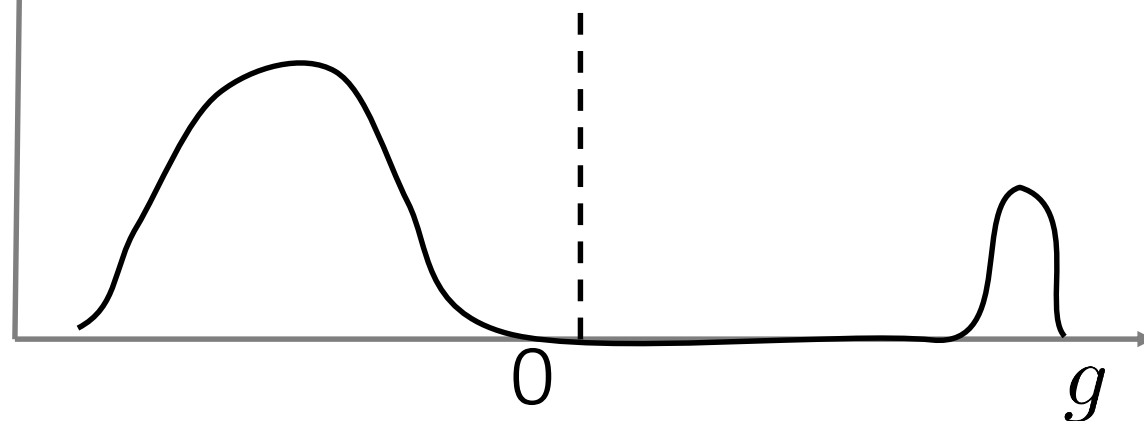
Expected safety can be misleading



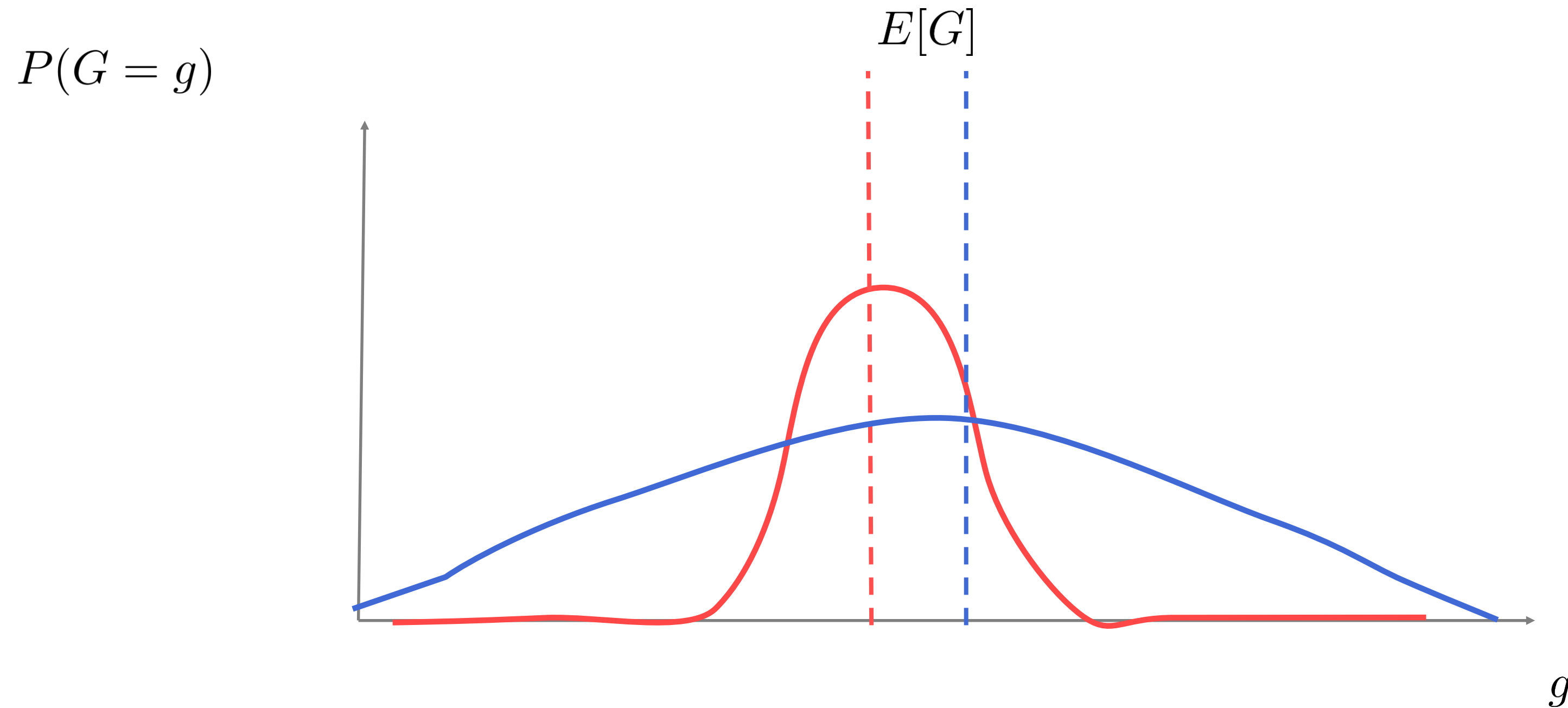
$P(G = g)$

$E[G]$

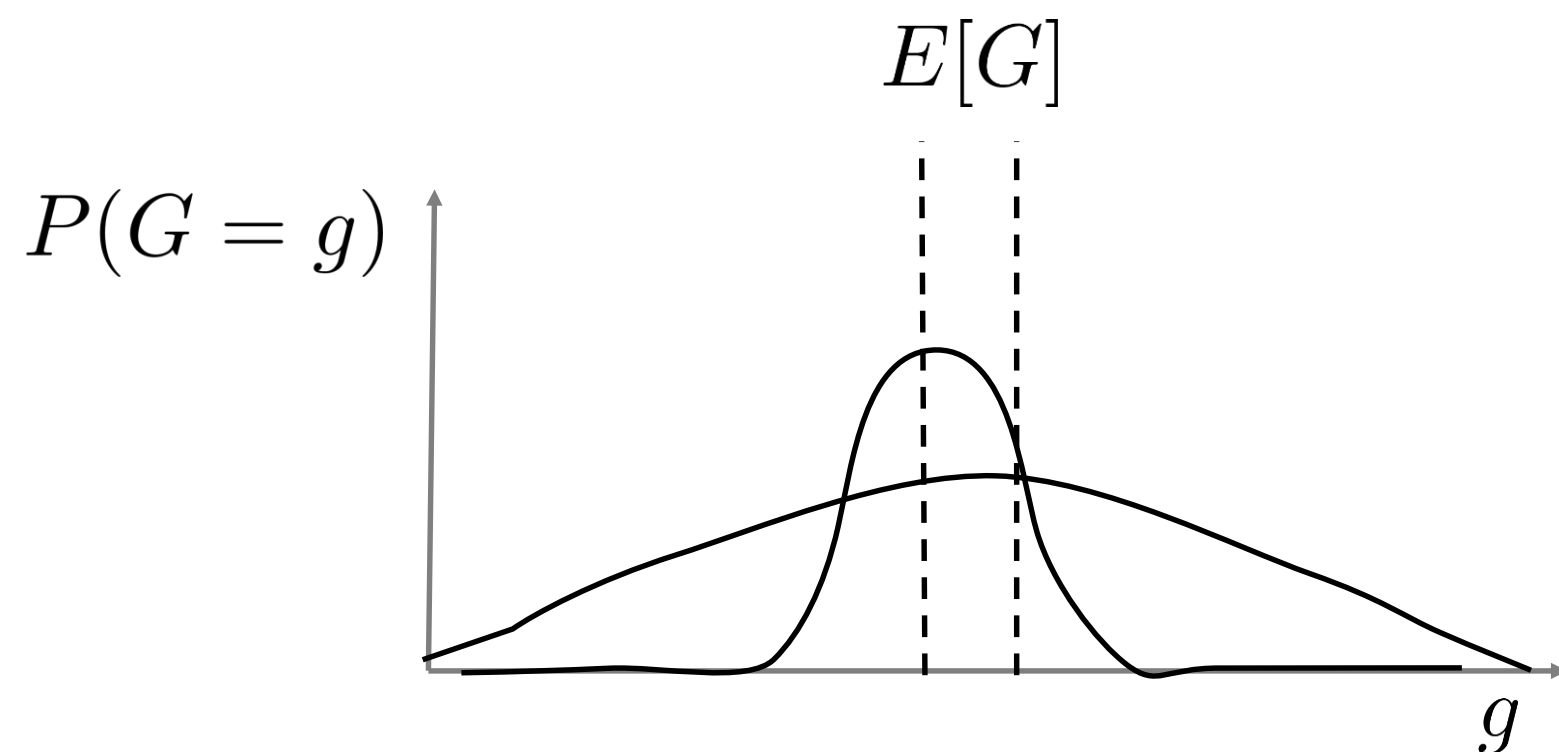
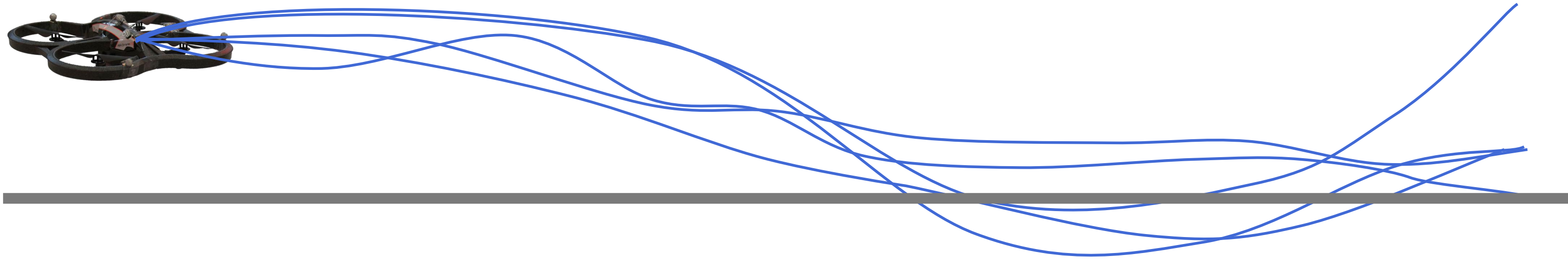
Safe in expectation!



Expected safety and variance



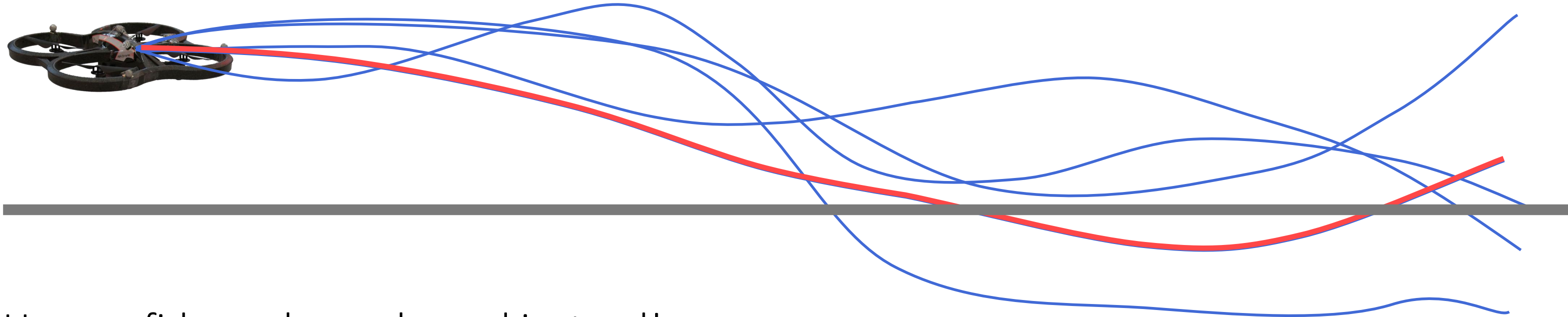
Risk sensitivity



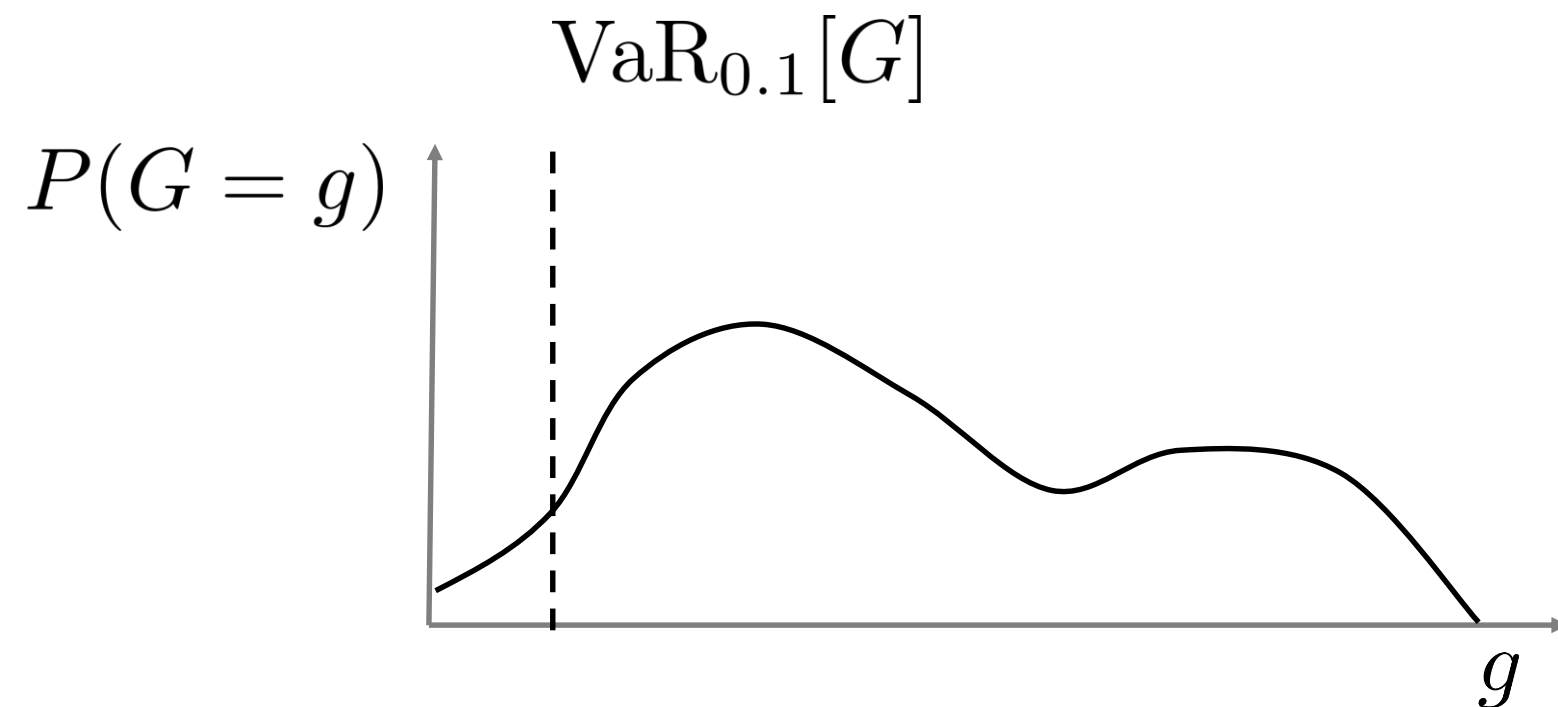
$$\mathbb{E} [e^{\tau G}] \propto \mathbb{E}[G] + \tau \mathbb{E}[G^2] + \mathcal{O}(\tau^2)$$

Even at low variance, a significant amount of trajectories may still be unsafe.

Value at Risk

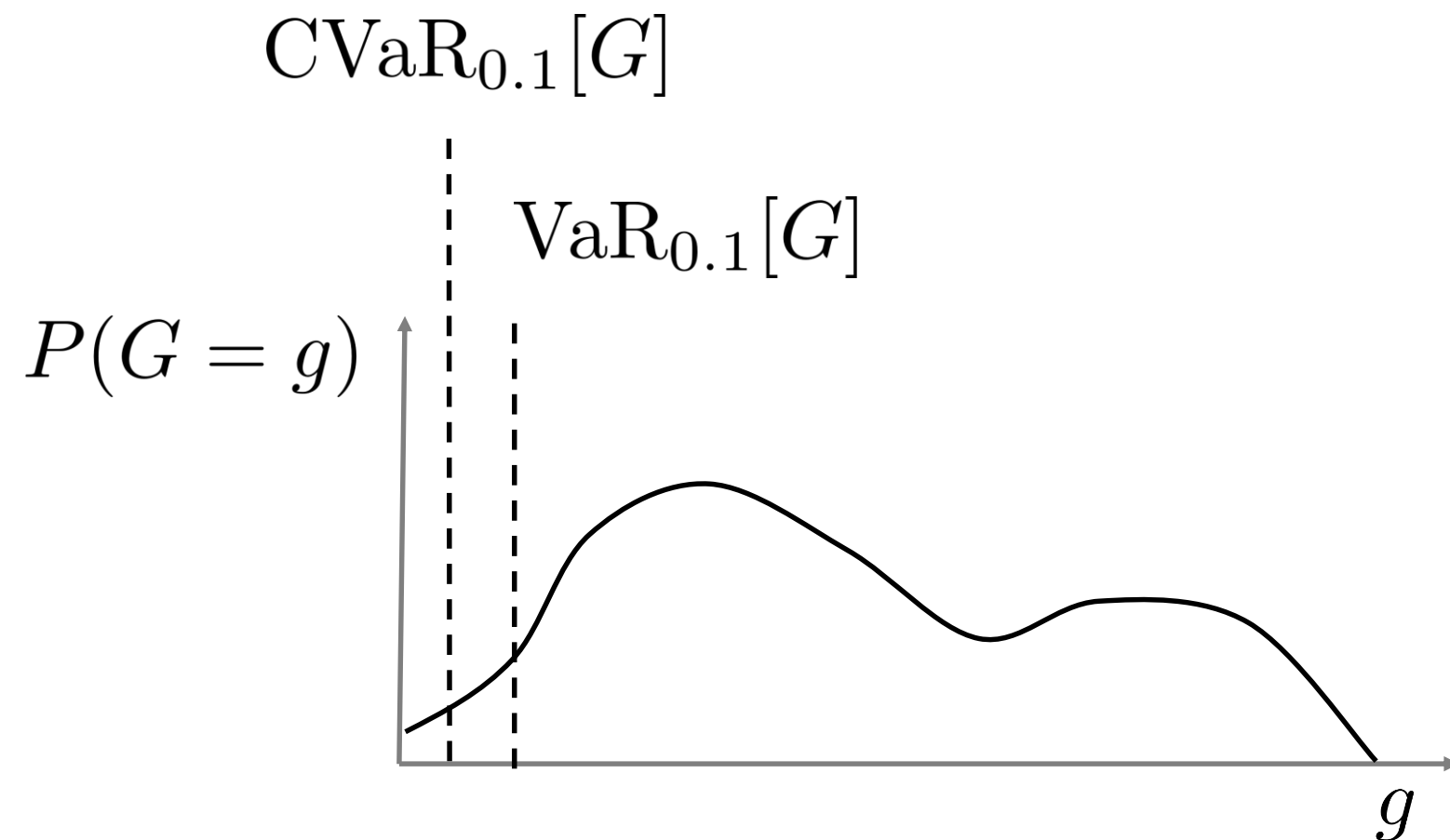
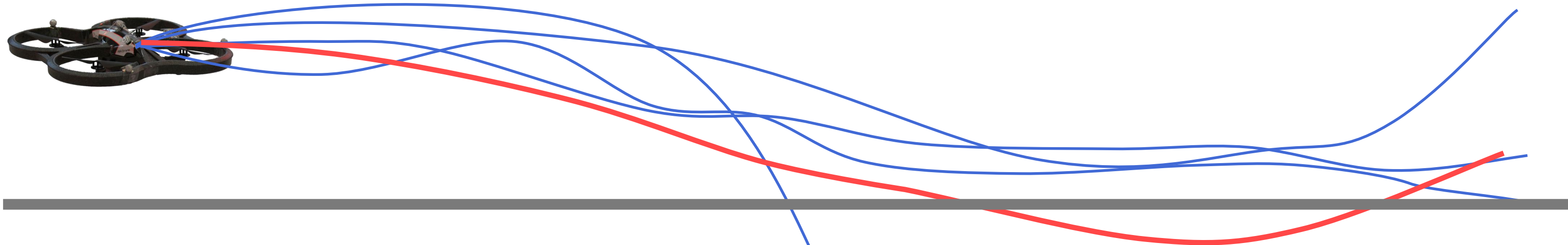


Use confidence lower-bound instead!



$$\text{VaR}_{\delta}[G] = \inf\{\epsilon \in \mathbb{R} : P(G \leq \epsilon)\} \geq \delta$$

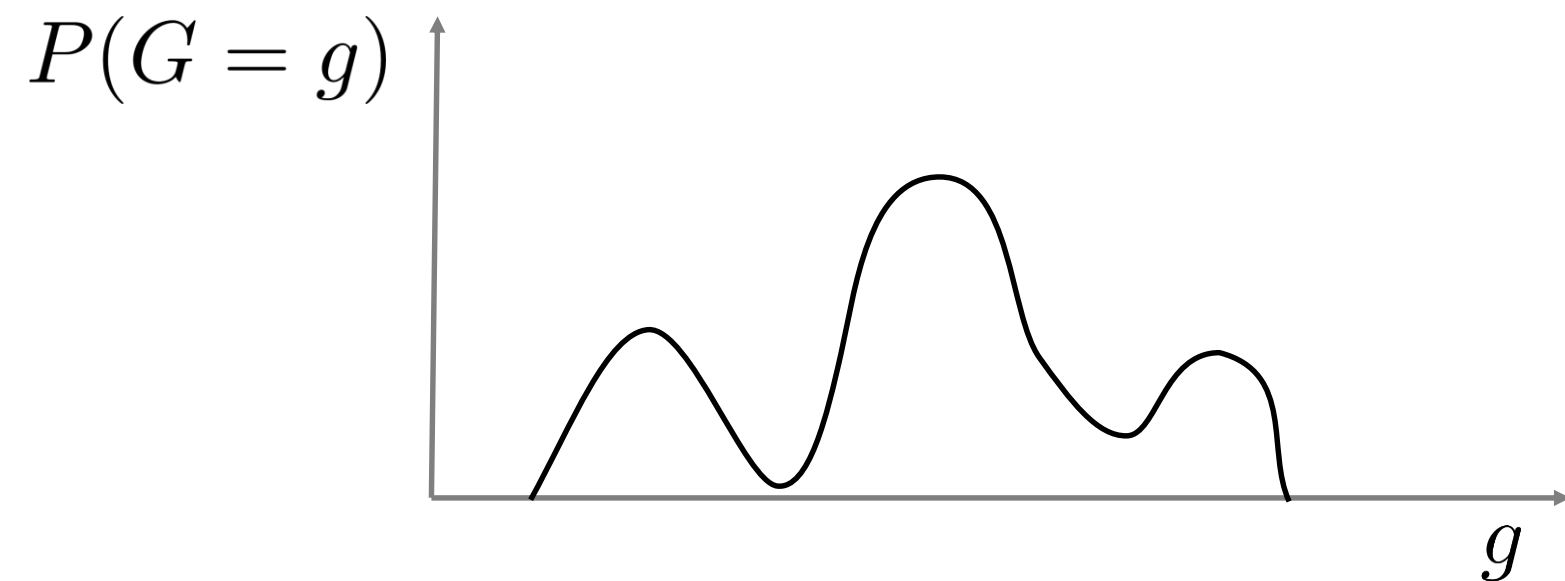
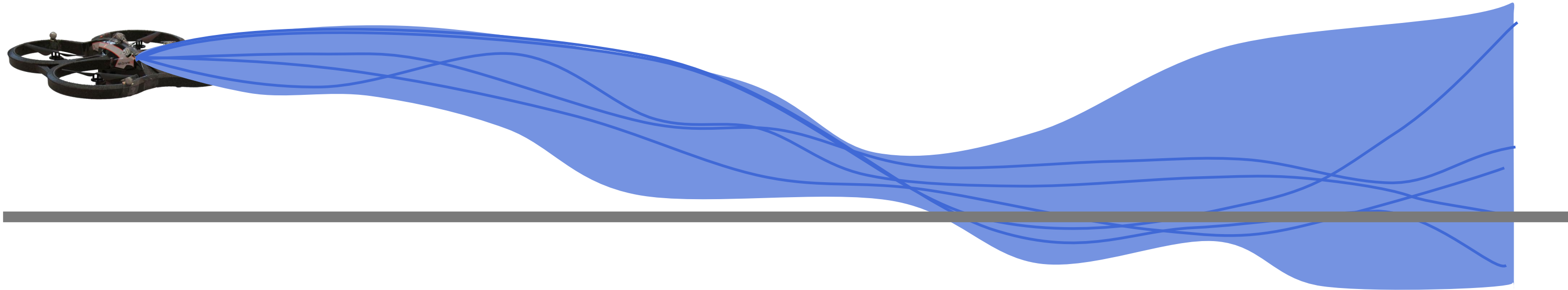
Conditional Value at Risk



$$\text{VaR}_\delta[G] = \inf \{ \epsilon \in \mathbb{R} : P(G \leq \epsilon) \geq \delta \}$$

$$\text{CVaR}_\delta[G] = \frac{1}{\delta} \int_0^\delta \text{VaR}_\alpha[G] d\alpha$$

Worst-case



$$P(G > 0) = 1$$

$$\text{or } g(\tau) > 0 \quad \forall \tau \in \Gamma$$

Notions of safety

Stochastic

Expected risk $\mathbb{E}[G]$

Moment penalized $\mathbb{E}[e^{\tau G}]$

Value at risk $\text{VaR}_\delta[G] = \inf\{\epsilon \in \mathbb{R} : P(G \leq \epsilon)\} \geq \delta$

Conditional value at risk $\text{CVaR}_\delta[G] = \frac{1}{\delta} \int_0^\delta \text{VaR}_\alpha[G] d\alpha$

Worst--case

$g(\tau) > 0 \quad \forall \tau \in \Gamma$

→ Robust Control

→ Formal verification

Acting in *known* model with safety constraints



Constrained Markov decision processes

Eitan Altman, CRC Press, 1999

Essentials of robust control

Kemin Zhou, John C. Doyle, PH, 1998

Robust control of Markov decision processes with uncertain transition matrices

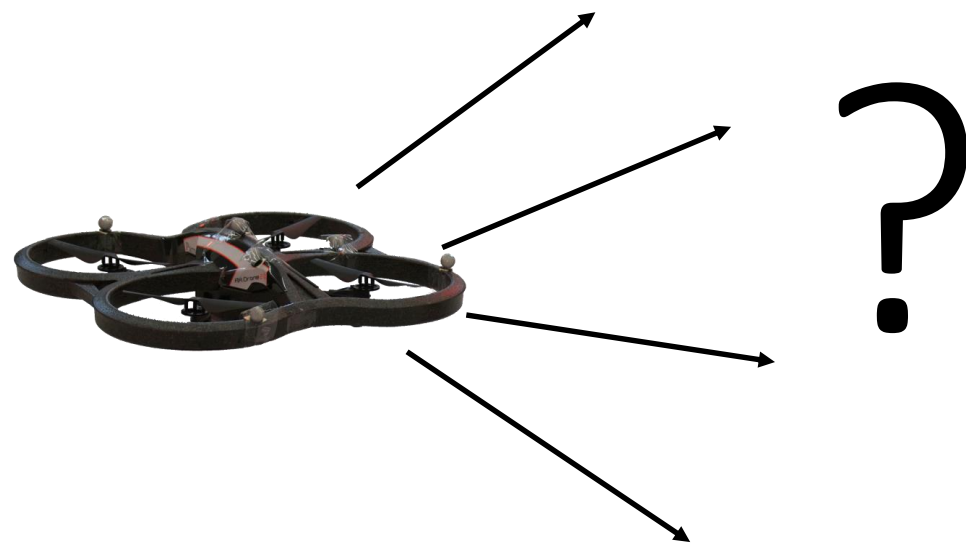
Arnab Nilim, Laurent El Ghaoui, OR, 2005

Reinforcement Learning



Key challenge: Don't know the consequences of actions!

How to start acting safely?



No knowledge!
Now what?

Imitation learning

$$a_t = \text{[head with gears icon]} = \pi^*(s_t)$$

Dataset $\mathcal{D} = \{s_t, \pi^*(s_t)\}_{t=0}^N$



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

$$\theta = \operatorname{argmin}_{\theta} \sum_{s, a \in \mathcal{D}} \|\pi(s, \theta) - a_t\|$$

No experience gained here!

Imitation learning algorithms

Data aggregation

Generate state sequence with policy $\pi(s_t, \theta)$

$$D \leftarrow D \cup \{(s_t, \pi^*(s_t))\}_{t=1}^T$$

$$\theta = \operatorname{argmin}_{\theta} \sum_{s, a \in \mathcal{D}} \|\pi(s, \theta) - a_t\|$$

Search-based Structured Prediction

Hal Daume III, John Langford, Daniel Marcu, ML, 2009

Efficient Reductions for Imitation Learning

Stephan Ross, Drew Bagnell, AISTATS 2010

Policy aggregation

$$\pi_0 = \pi^*$$

Generate state sequence \mathcal{D} with policy π_i

$$\theta_{i+1} = \operatorname{argmin}_{\theta} \sum_{s, a \in \mathcal{D}} \|\pi(s, \theta) - a_t\|$$

$$\pi_{i+1}(s) = \alpha_0 \pi^*(s) + \sum_{j=1}^{i+1} \alpha_j \pi(s, \theta_j)$$

A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning

Stephane Ross, Geoffrey J. Gordon, J. Andrew Bagnell, 2011

Safe Imitation learning

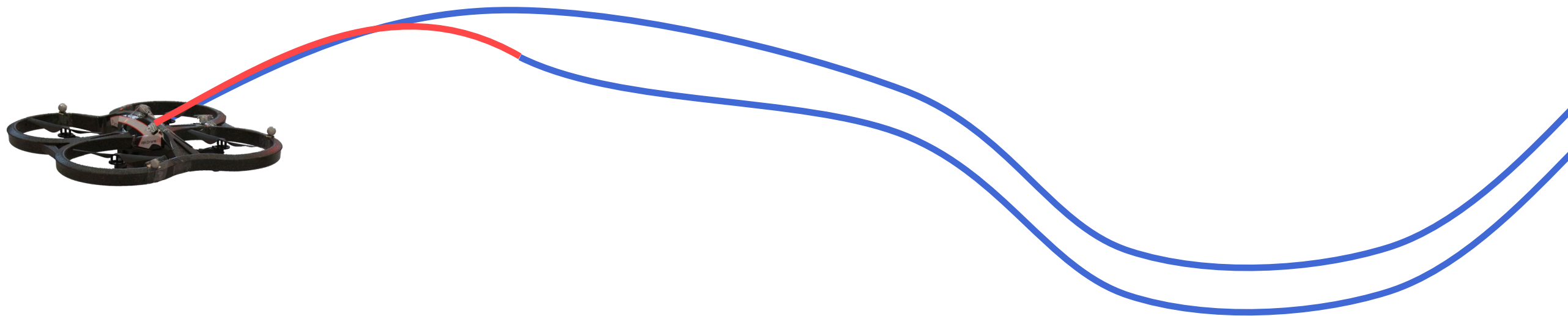
Only apply action from learned policy when

$$\|\pi(s_t, \theta) - \pi^*(s_t)\| \leq \epsilon$$

Query-Efficient Imitation Learning for End-to-End Autonomous Driving
Jiakai Zhang, Kyunghyun Cho, AAAI, 2017

$$\text{Var}[\pi(s_t, \theta)] \leq \gamma$$

EnsembleDAgger: A Bayesian Approach to Safe Imitation Learning
Kunal Menda, Katherine Driggs-Campbell, Mykel J. Kochenderfer, arXiv2018



What to do with this initial policy?

Can find an initial, safe policy based on domain knowledge.

How to improve?



Prior knowledge as backup for learning



Safety controller takes over

Learner is seen as a disturbance

Know what is safe

Provably safe and robust learning-based model predictive control

A. Aswani, H. Gonzalez, S.S. Satry, C. Tomlin, Automatica, 2013

Safe Exploration of State and Action Spaces in Reinforcement Learning

J. Garcia, F. Fernandez, JAIR, 2012

Safe Reinforcement Learning via Shielding

M. Alshiekh, R. Bloem, R. Ehlers, B. Könighofer, S. Nickum, U. Topcu, AAI, 2018

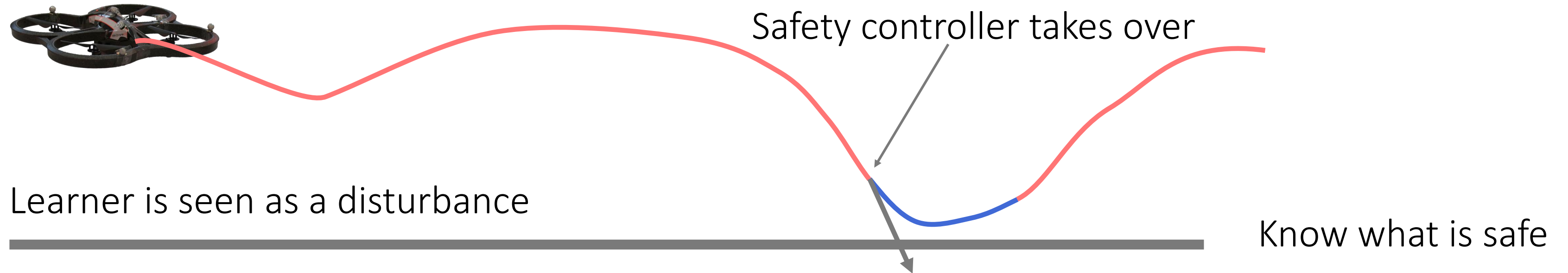
Safe Exploration in Continuous Action Spaces

G. Dalai, K. Dvijotham, M. Veccerik, T. Hester, C. Paduraru, Y. Tassa, arXiv, 2018

Linear Model Predictive Safety Certification for Learning-based Control

K.P. Wabersich, M.N. Zeilinger, CDC, 2018

Prior knowledge as backup for learning



Need to know what is unsafe in advance.

Without learning, need significant prior knowledge.

The learner does not know what's happening!

Safety as improvement in performance (Expected safety)

Initial, stochastic policy $\pi(s, \theta_b)$

Performance
$$J(\theta) = \mathbb{E}_{s_t \sim \rho(\theta)} \left[\sum_{t=1}^T \gamma^t r_t(s_t) \right] = \mathbb{E}_{\tau \sim \rho(\theta)} \left[g(\tau) \right]$$

Safety constraint
$$\Pr(J(\theta) \geq J(\theta_b)) \geq 1 - \delta$$

Need to estimate $J(\theta)$ based only on data from $\pi(s, \theta_b)$

Off-Policy Policy Evaluation

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



What does this tell me about a different policy $\pi(s, \theta)$?

Importance sampling:

$$\mathbb{E}_{\tau \sim \rho(\theta)} [g(\tau)] = \mathbb{E}_{\tau \sim \rho(\theta_b)} \left[\underbrace{\frac{p(\tau|\theta)}{p(\tau|\theta_b)}}_{\prod_{(s_t, a_t) \in \tau} \frac{p(a_t|s_t, \theta)}{p(a_t|s_t, \theta_b)}} g(\tau) \right]$$

(there are better ways to do this)

Eligibility Traces for Off-Policy Policy Evaluation
Doina Precup, Richard S. Sutton, S. Singh

Guaranteeing improvement

Unbiased estimate of $J(\theta)$.

What about $\Pr(J(\theta) \geq J(\theta_b)) \geq 1 - \delta$?

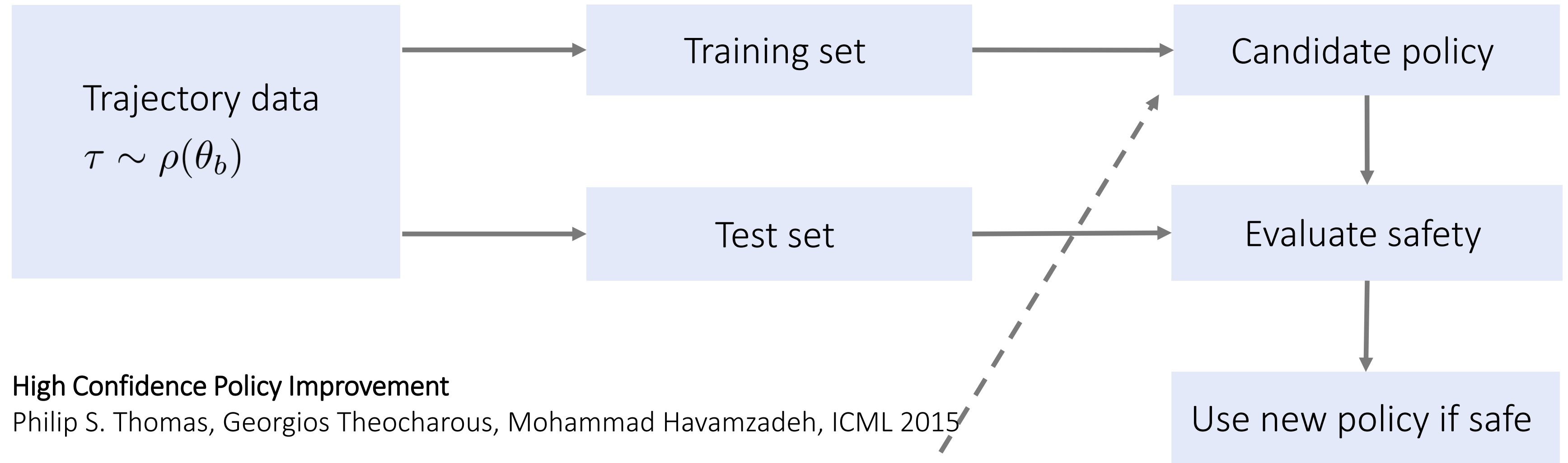
Generate trajectories using $\pi(s, \theta_b)$, $\tau \sim \rho(\theta_b)$

Use concentration inequality to obtain confidence intervals

With probability at least $1 - \delta$:

$$J(\theta) = \mathbb{E}_{\tau \sim \rho(\theta)} [g(\tau)] \geq \sum_{i=1}^N \frac{p(\tau_i | \theta)}{p(\tau_i | \theta_b)} g(\tau_i) - c(N, \delta)$$

Overview of expected safety pipeline



High Confidence Policy Improvement

Philip S. Thomas, Georgios Theodorou, Mohammad Havamzadeh, ICML 2015

Safe and efficient off-policy reinforcement learning

Remi Munos, Thomas Stepleton, Anna Harutyunyan, Marc G. Bellemare, NIPS, 2016

Constrained Policy Optimization

Joshua Achiam, David Held, Aviv Tamar, Pieter Abbeel, ICML, 2017

Summary part one

Reviewed safety definitions

Saw how to obtain a first, safe policy

Reviewed a first method for safe learning in expectation

Stochastic

- Expected risk
- Moment penalized
- VaR / CVaR

Worst-case

- Formal verification
- Robust optimization

Second half: Explicit safe exploration

More model-free safe exploration

Model-based safe exploration without ergodicity

Reinforcement learning (recap)

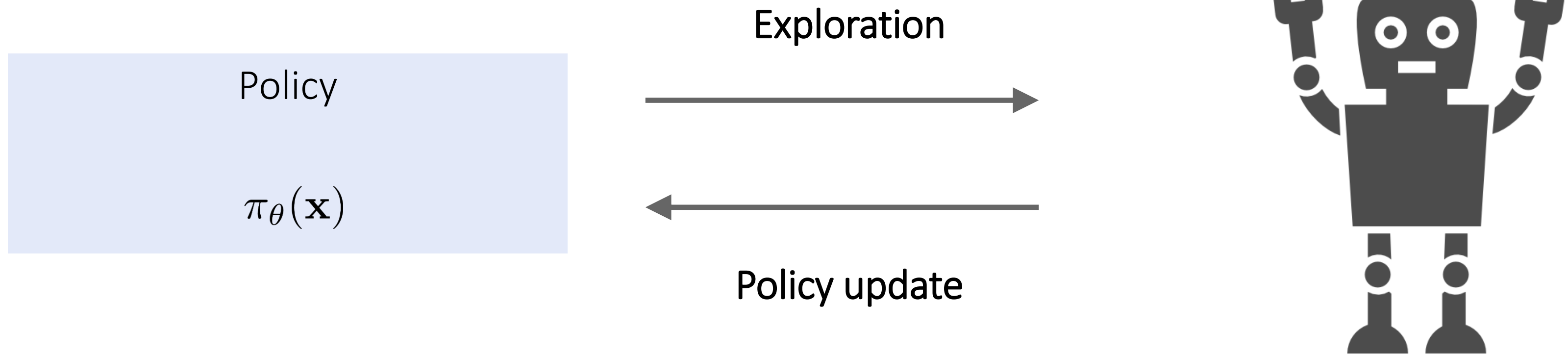


Image: Plainicon, <https://flaticon.com>

Statistical models to guarantee safety

Model-free

Model-based

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

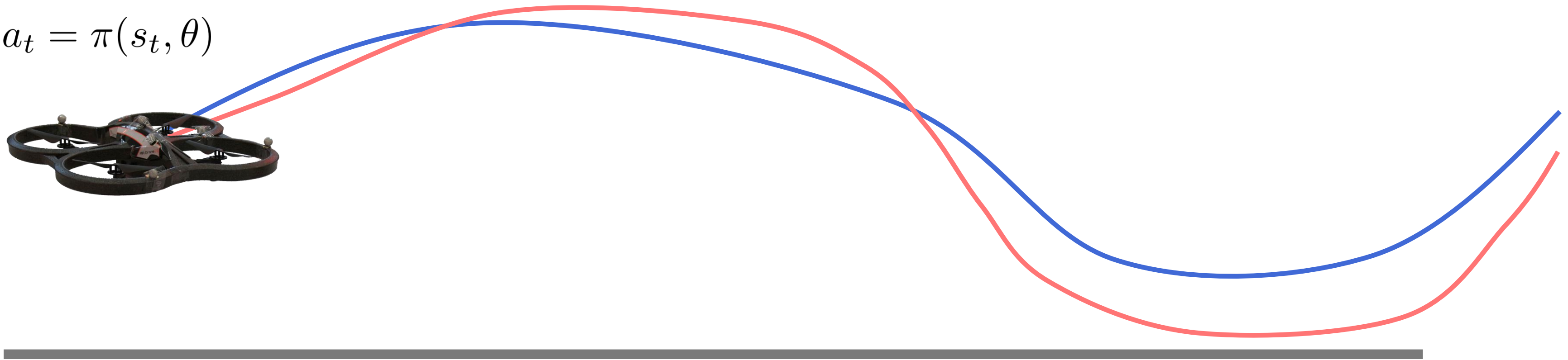
Estimate $J(\theta)$
and optimize

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

Estimate/identify,
then plan/control

Model-free reinforcement learning

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



Tracking performance

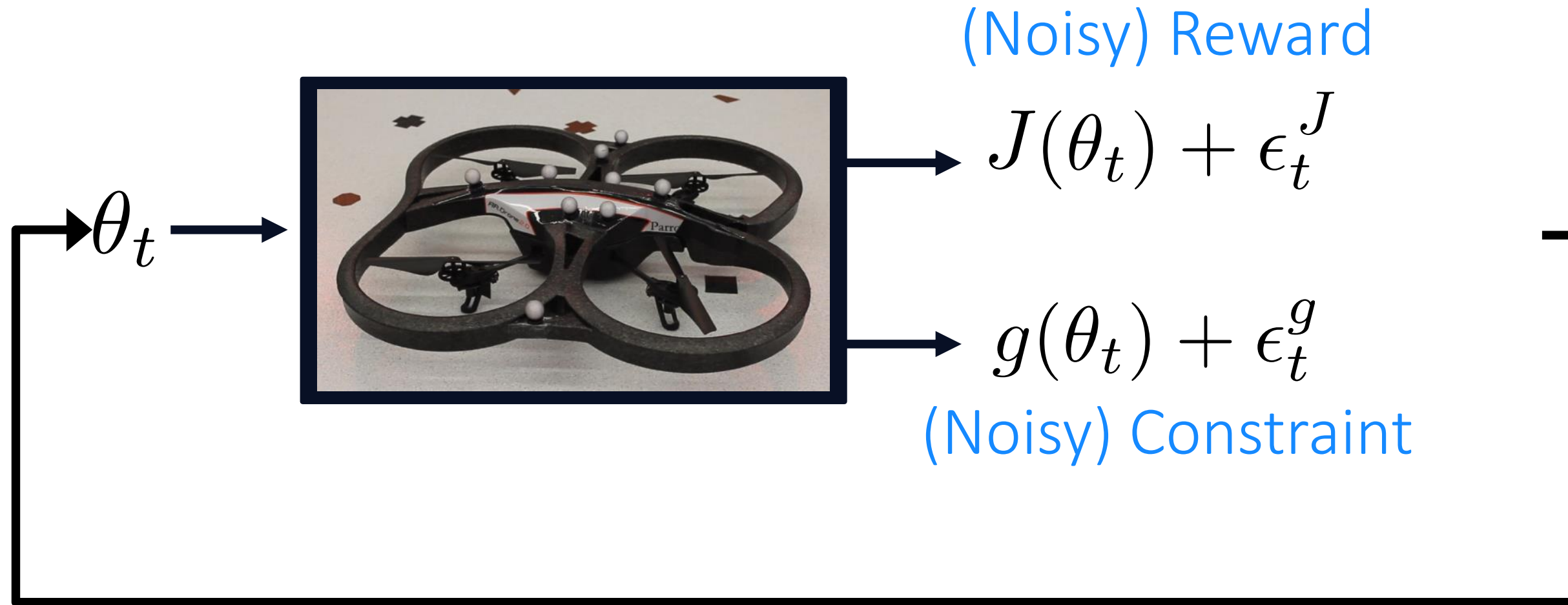
$$\max_{\theta} J(\theta)$$

Few, noisy experiments

Safety constraint

$$g(\theta) \geq 0$$

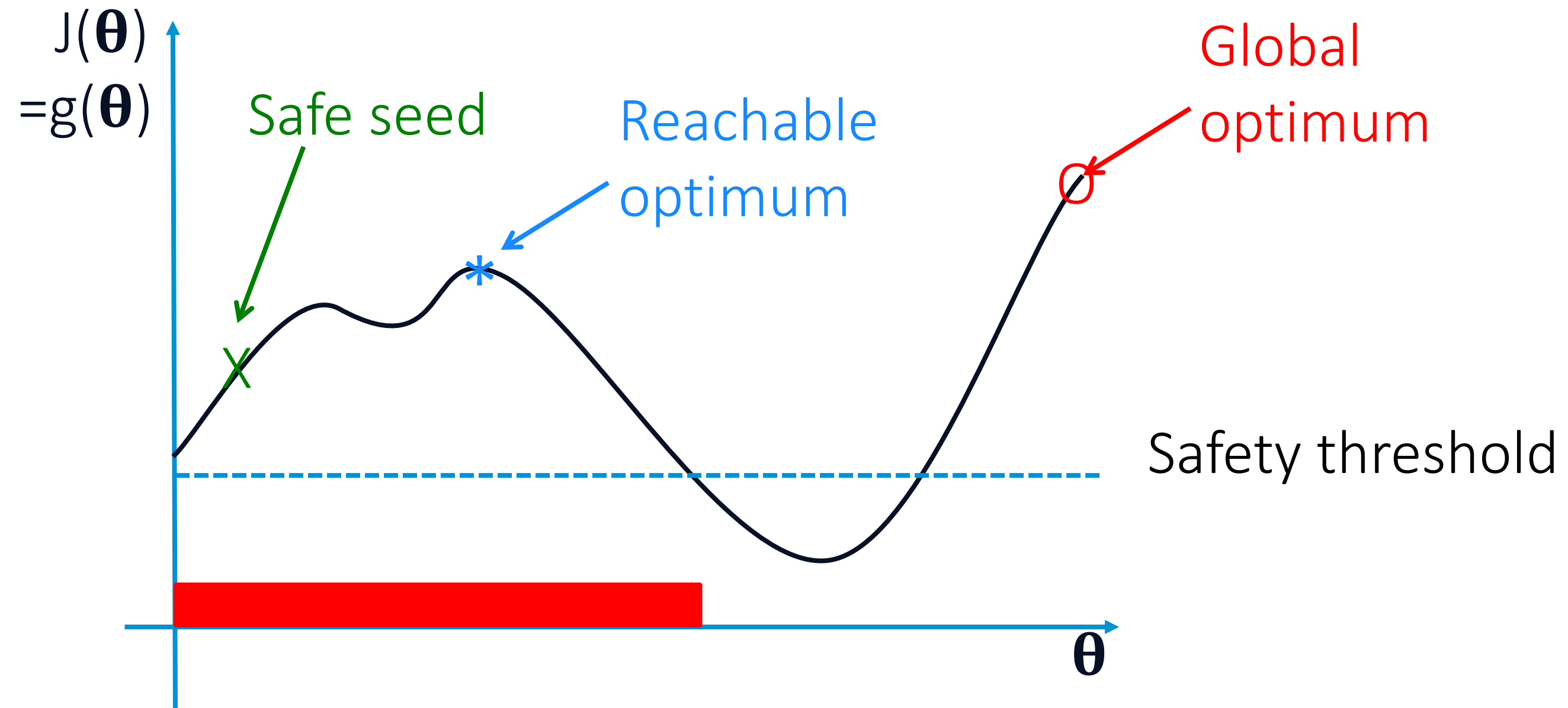
Safety for all experiments



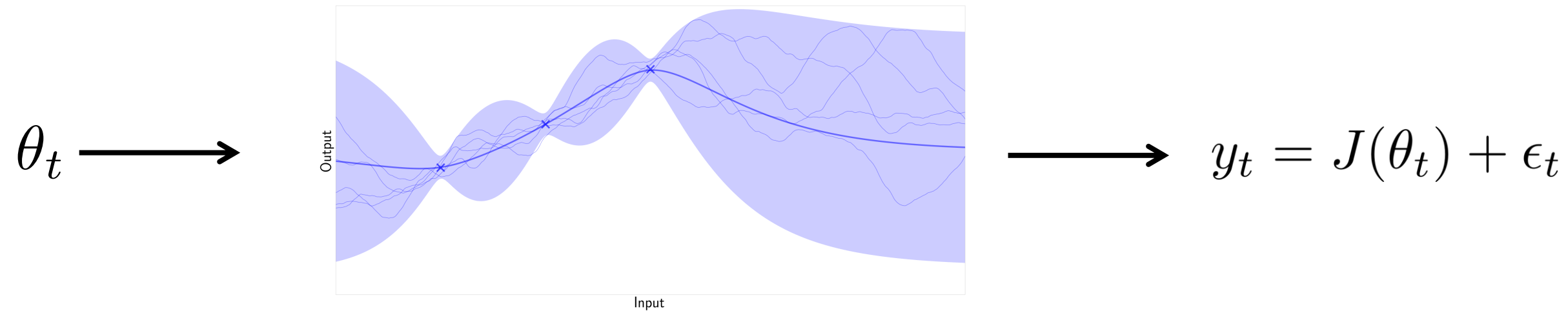
Goal: $\max_{\theta} J(\theta)$ s.t. $g(\theta) \geq 0$

Safety: $g(\theta_t) \geq 0$ for all t with probability $\geq 1-\delta$

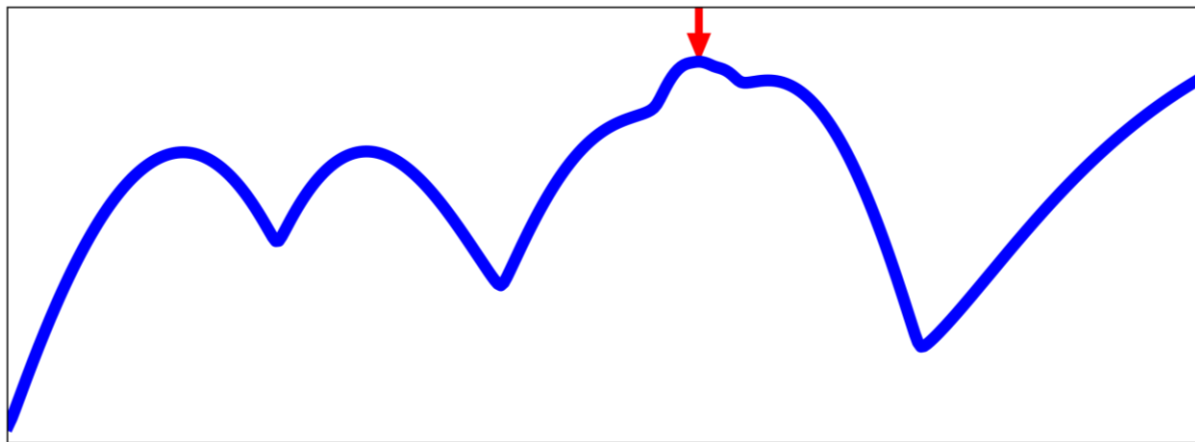
Safe policy optimization illustration



Starting Point: Bayesian Optimization



Acquisition function



Expected/most prob. improvement [Moćkus *et al.* '78,'89]

Information gain about maximum [Villemonteix *et al.* '09]

Knowledge gradient [Powell *et al.* '10]

Predictive Entropy Search [Hernández-Lobato *et al.* '14]

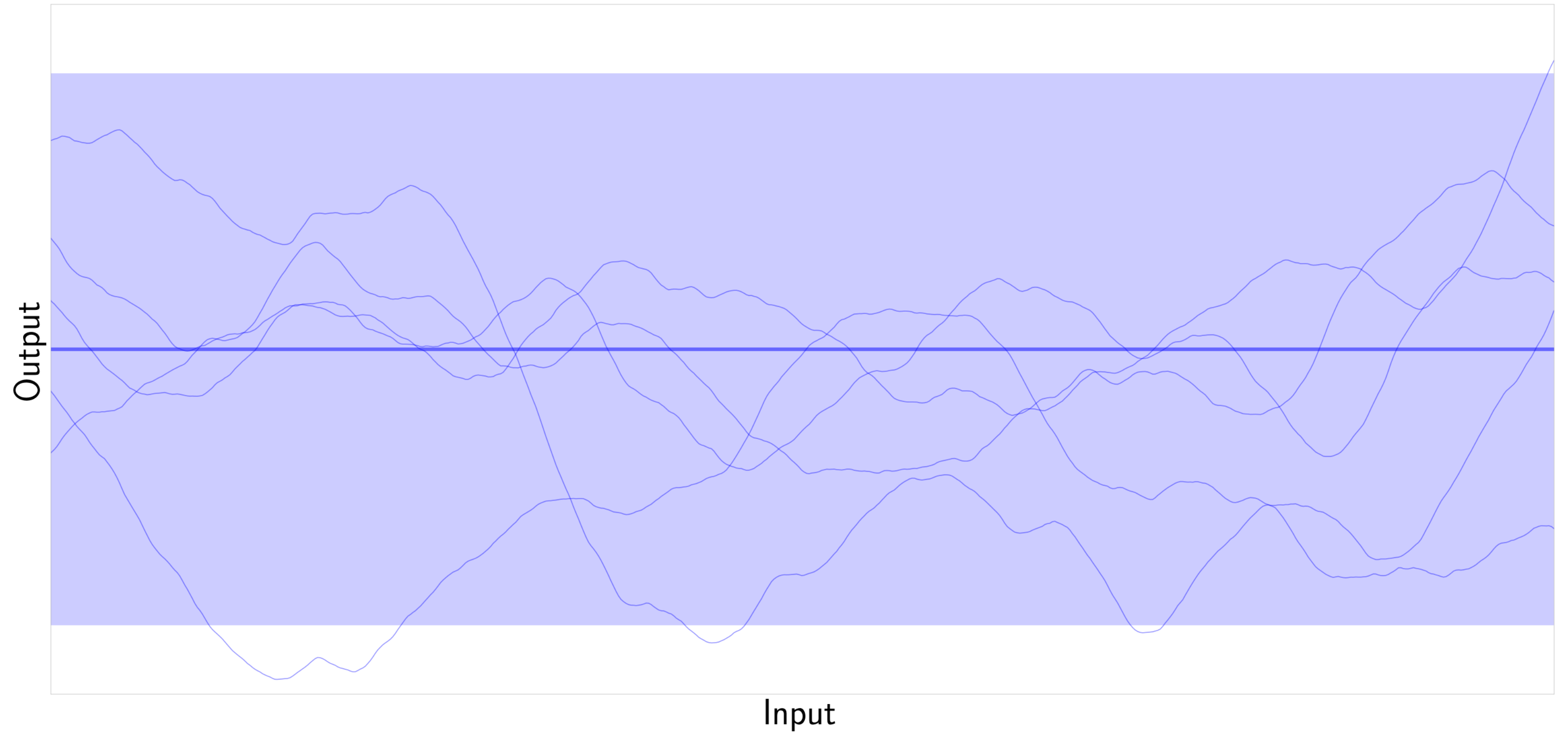
TruVaR [Bogunovic *et al.* '17]

Max Value Entropy Search [Wang *et al.* '17]

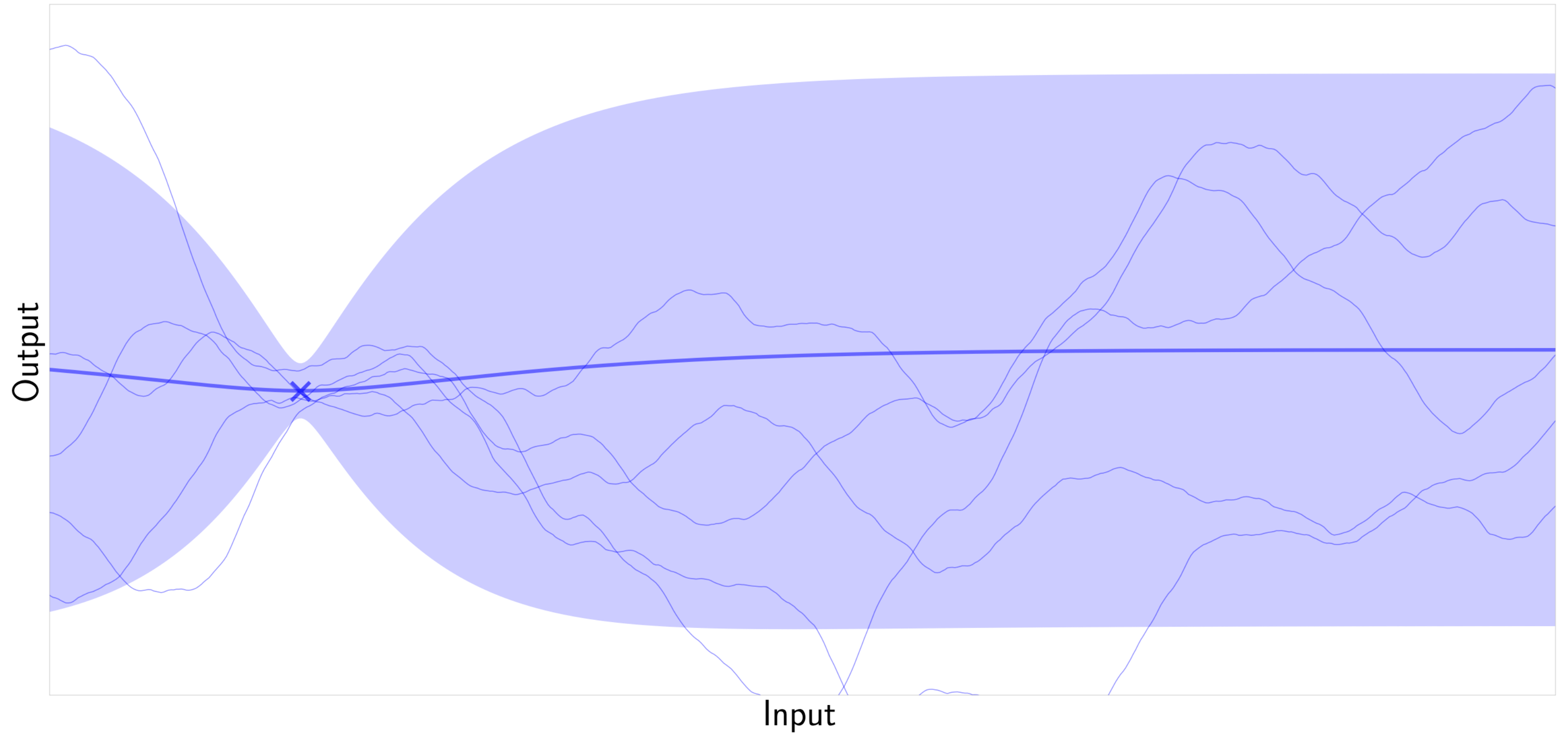
Constraints/Multiple Objectives

[Snoek *et al.* '13, Gelbart *et al.* '14, Gardner *et al.* '14, Zuluaga *et al.* '16]

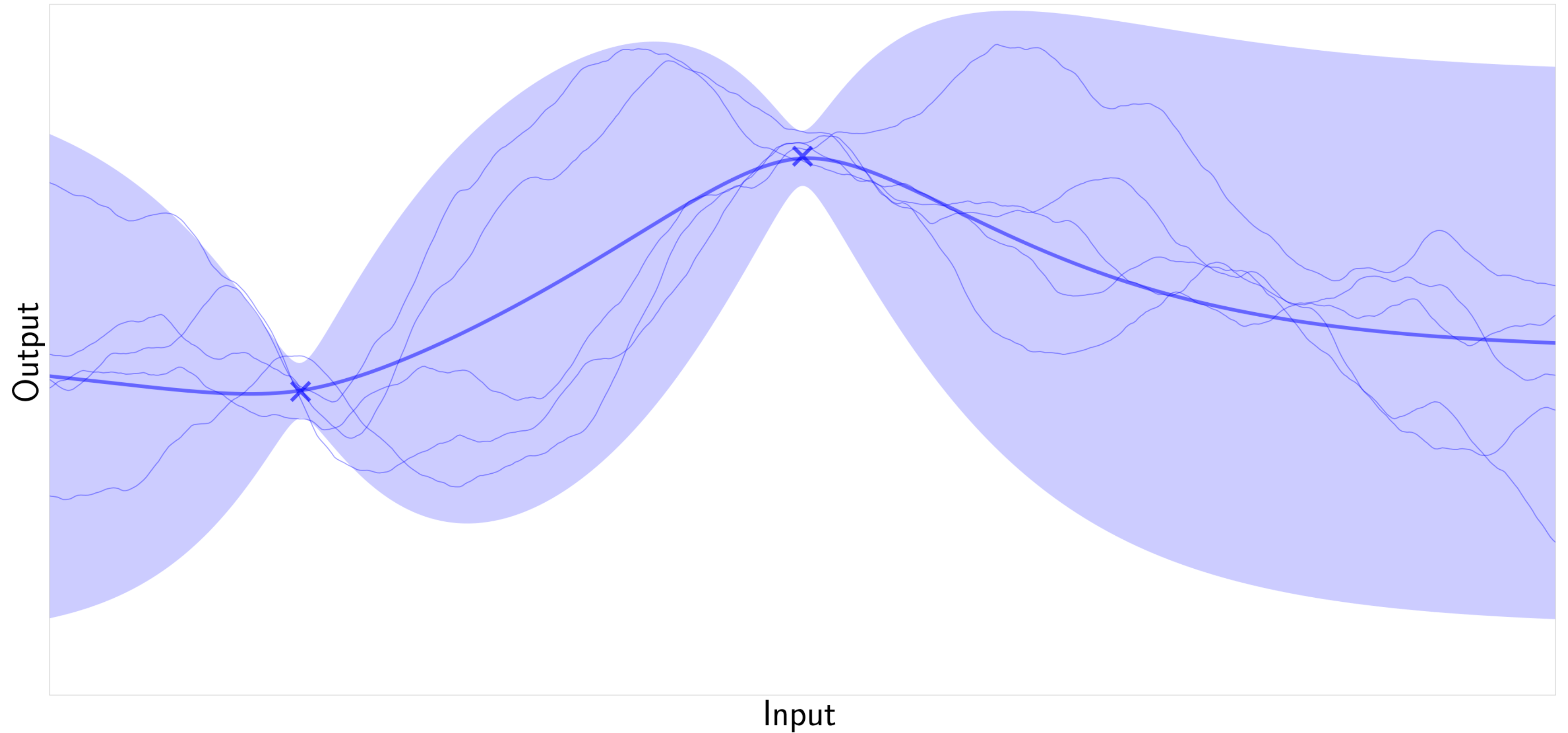
Gaussian process



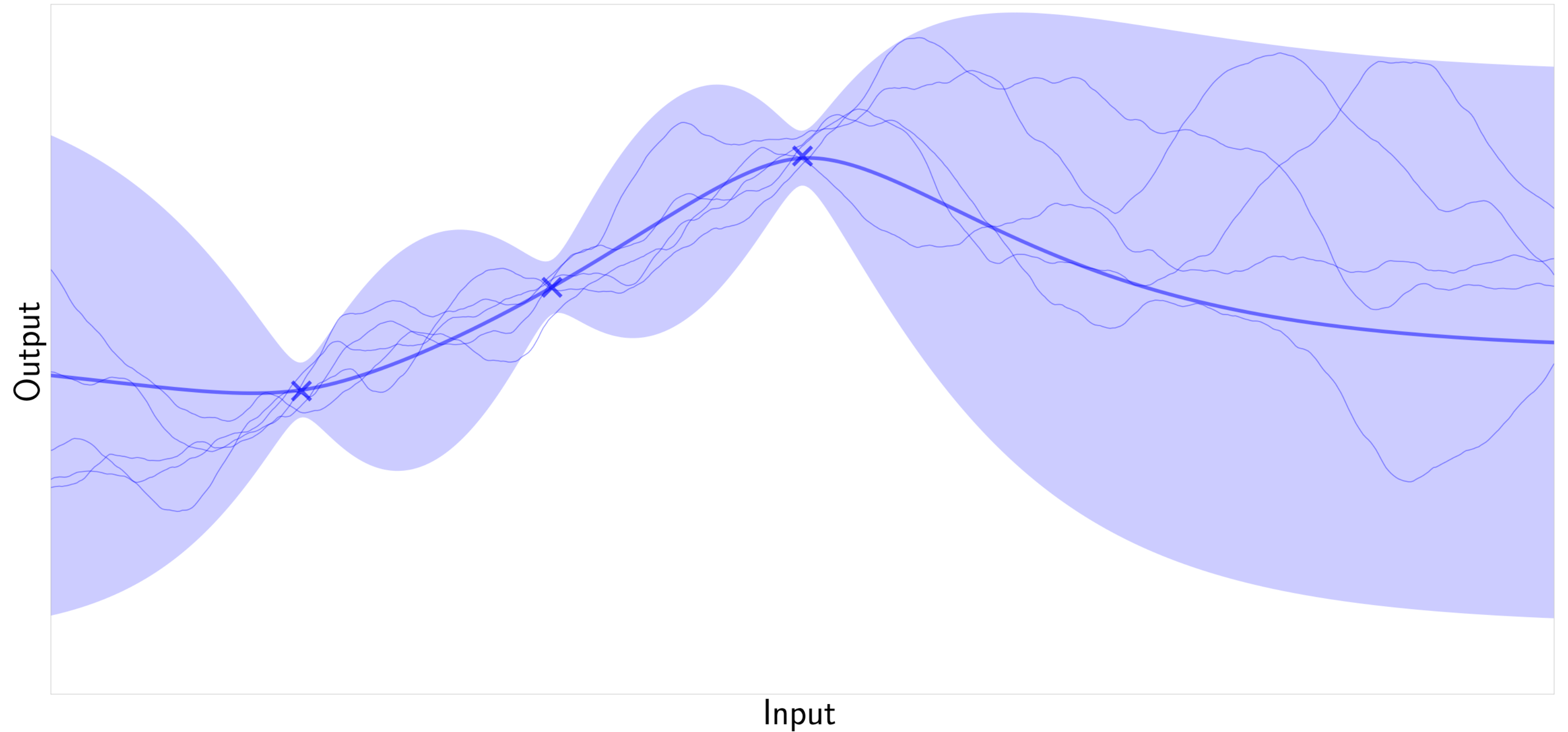
Gaussian process



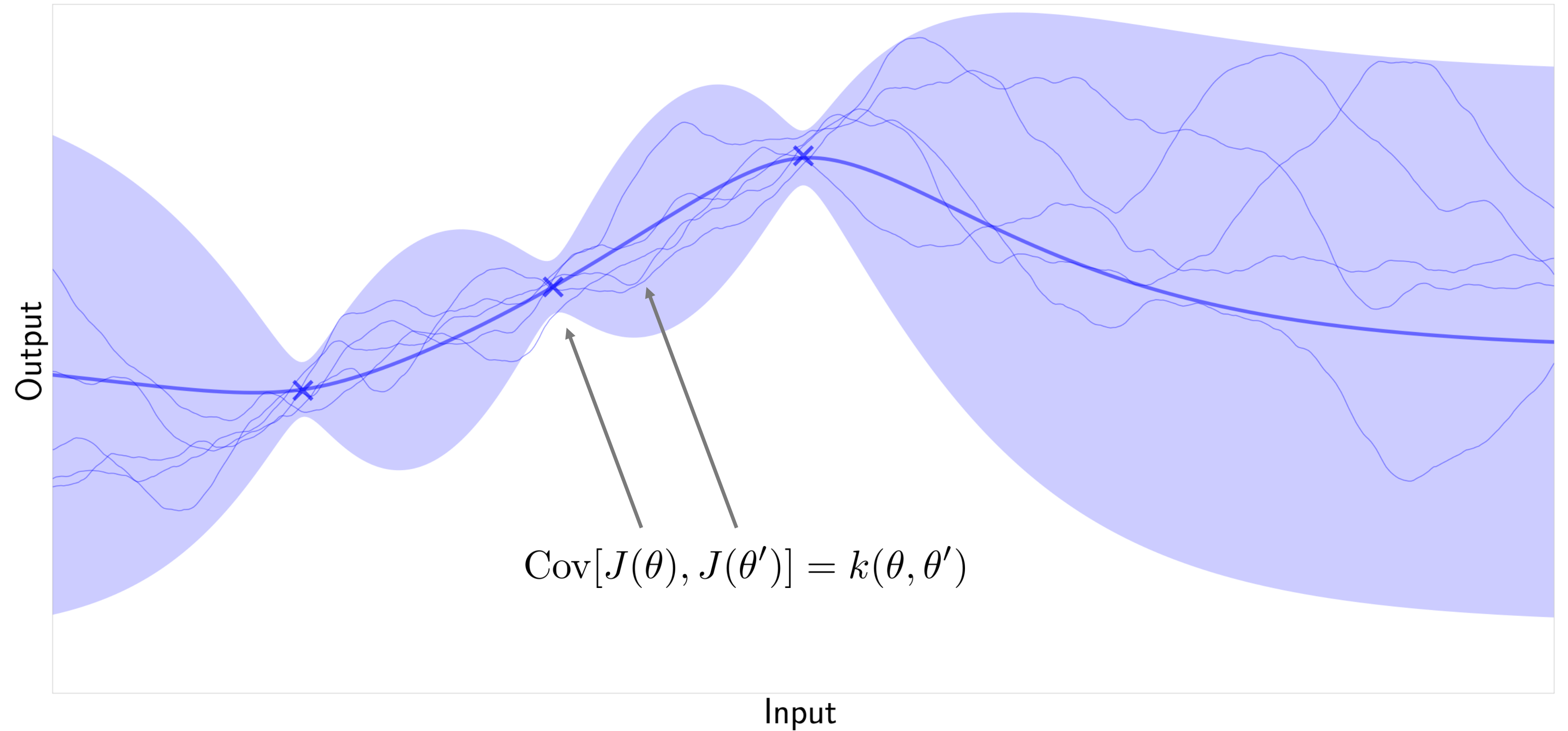
Gaussian process



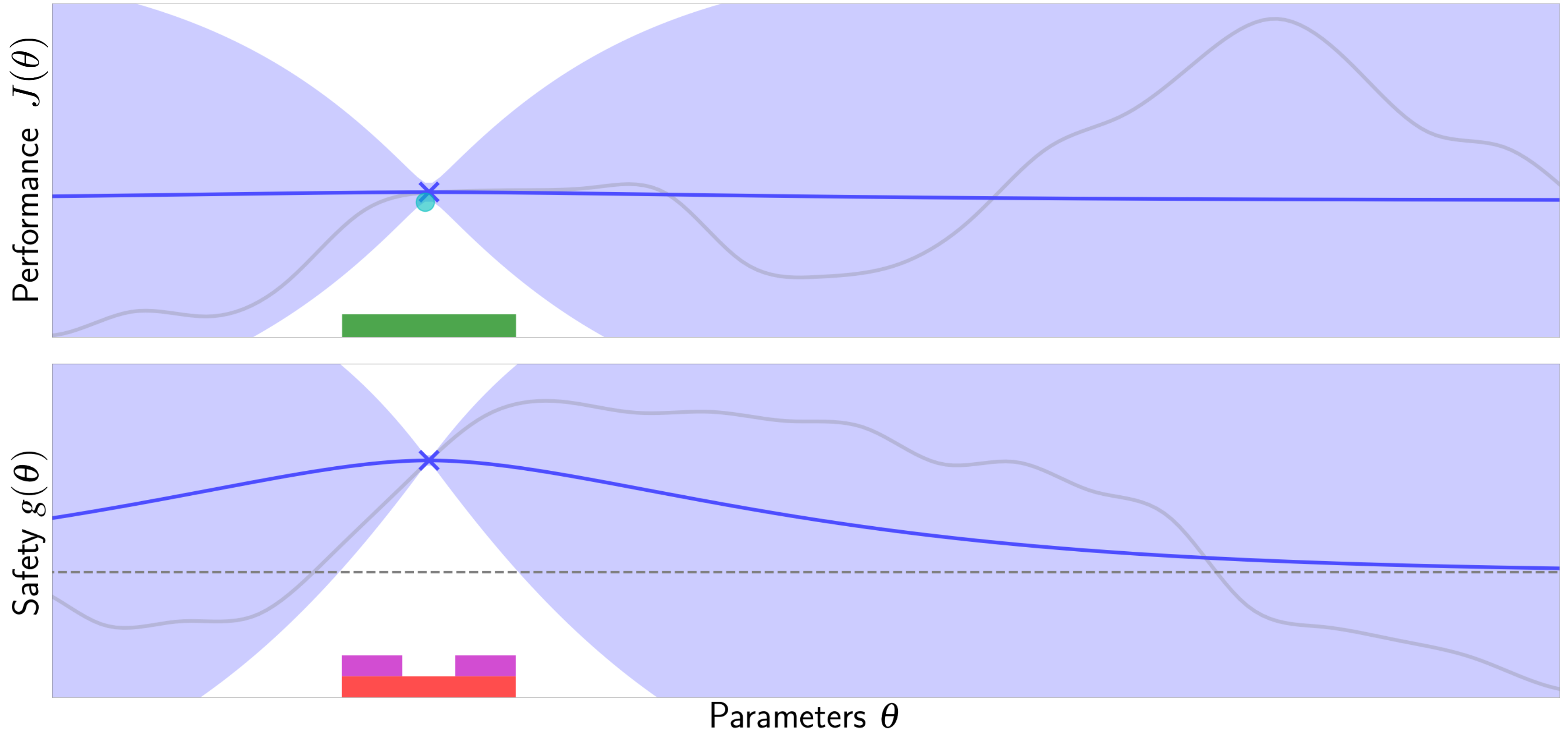
Gaussian process



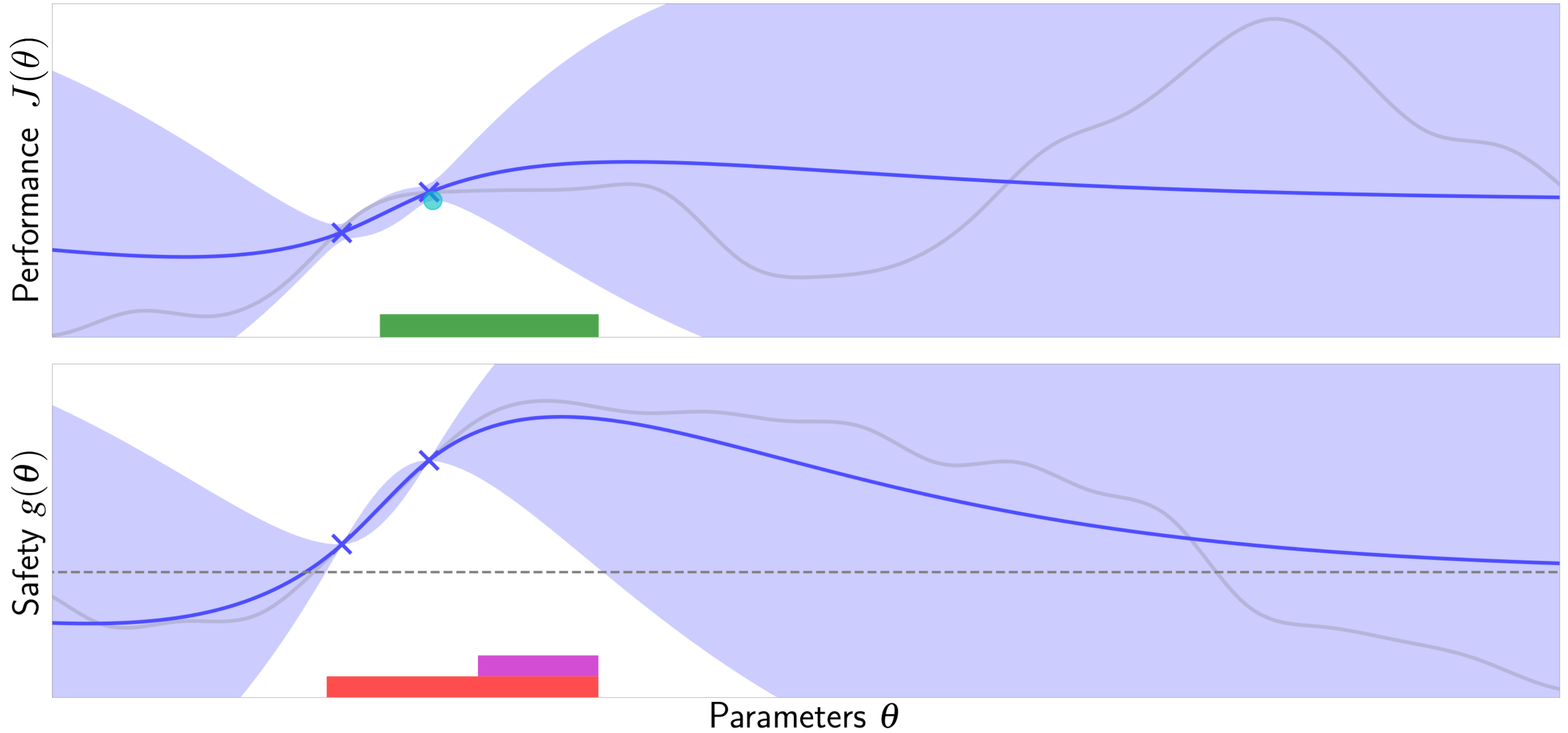
Gaussian process



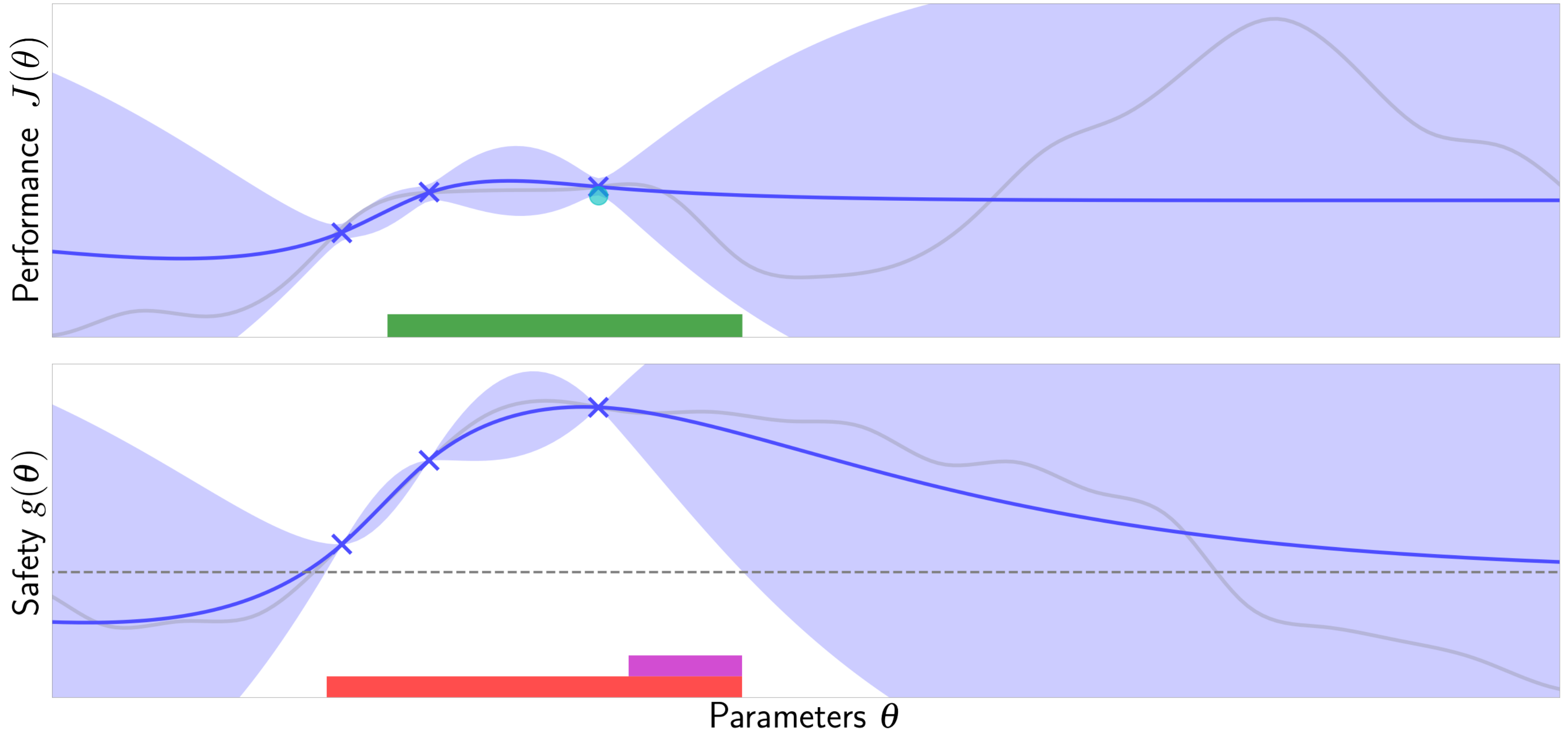
SafeOPT: Constrained Bayesian optimization



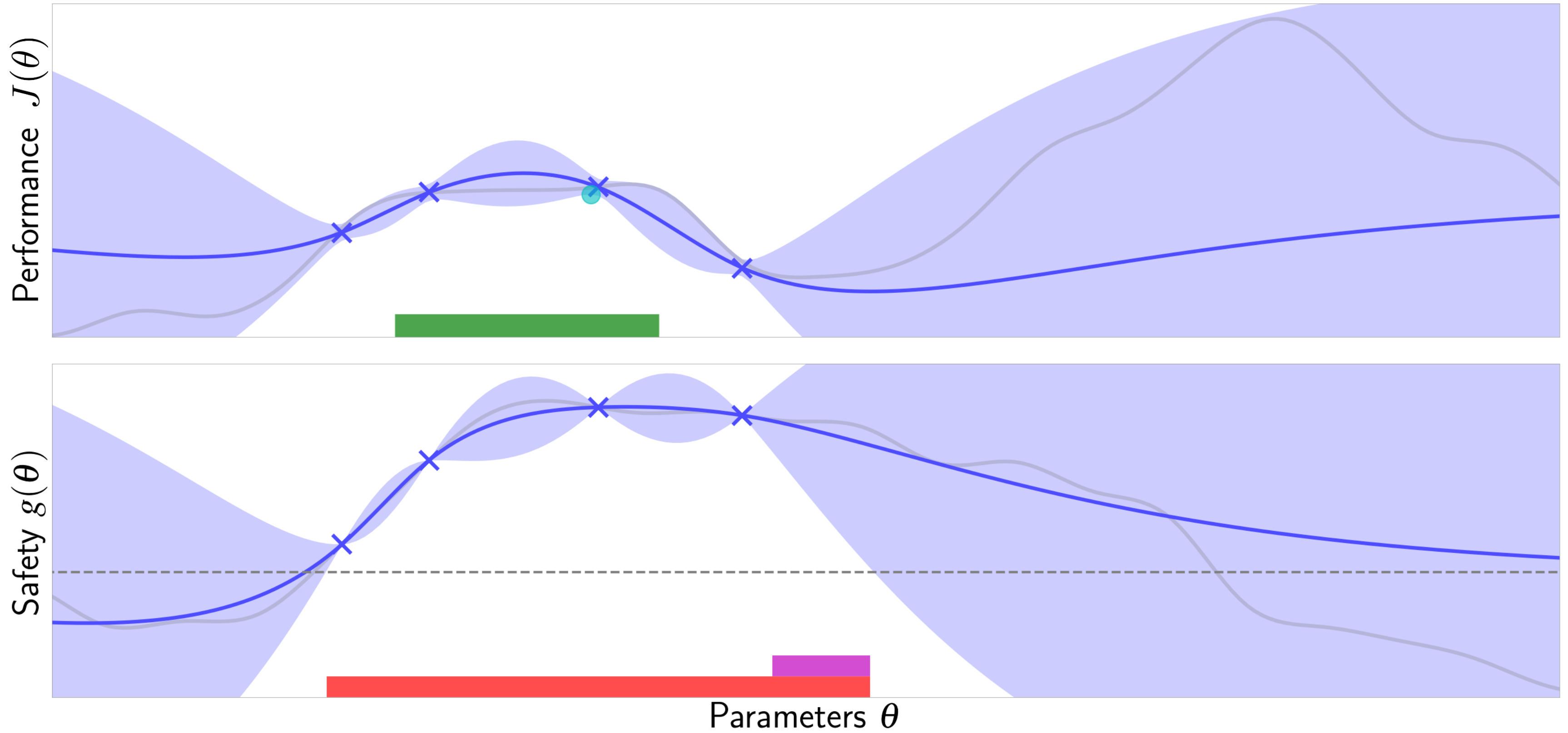
SafeOPT: Constrained Bayesian optimization



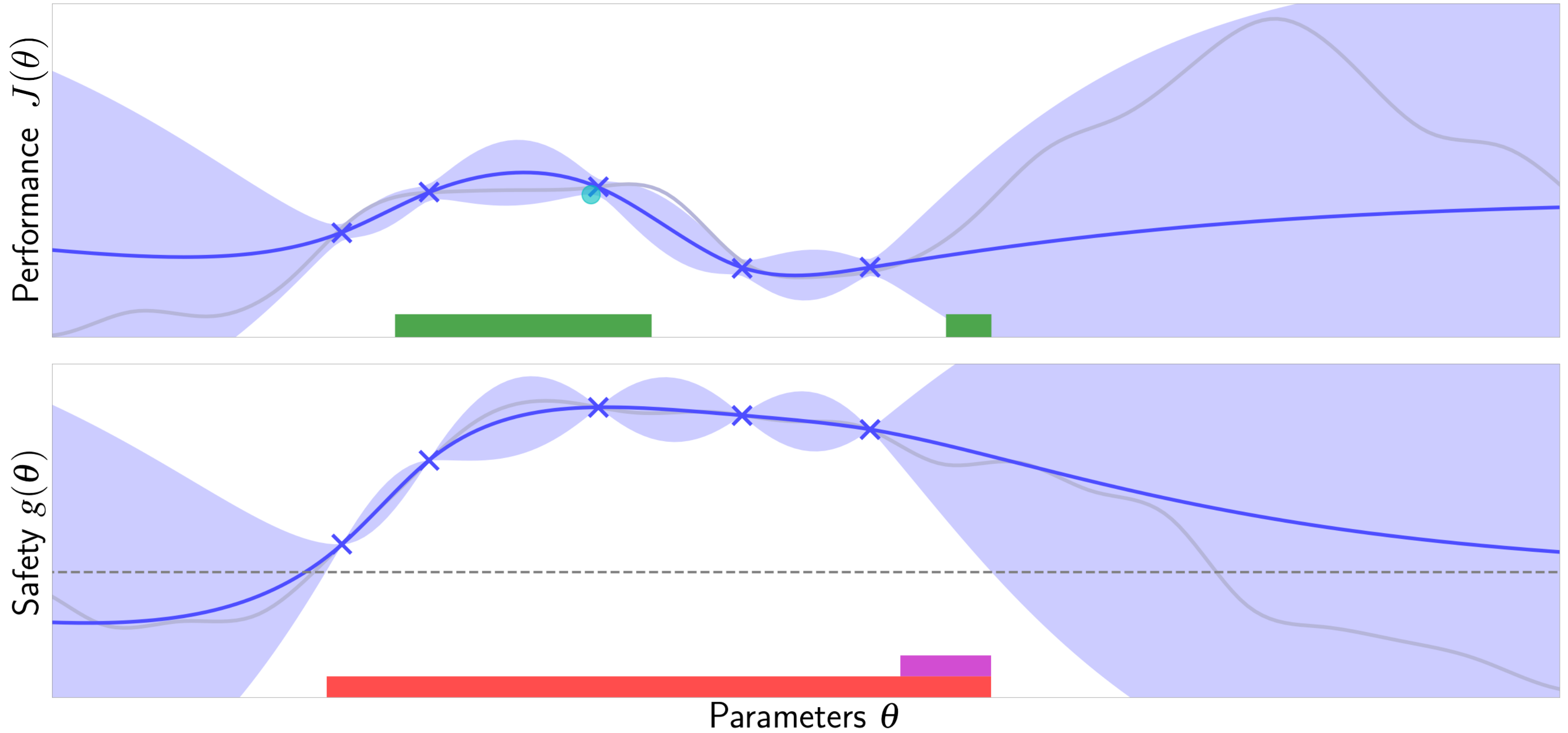
SafeOPT: Constrained Bayesian optimization



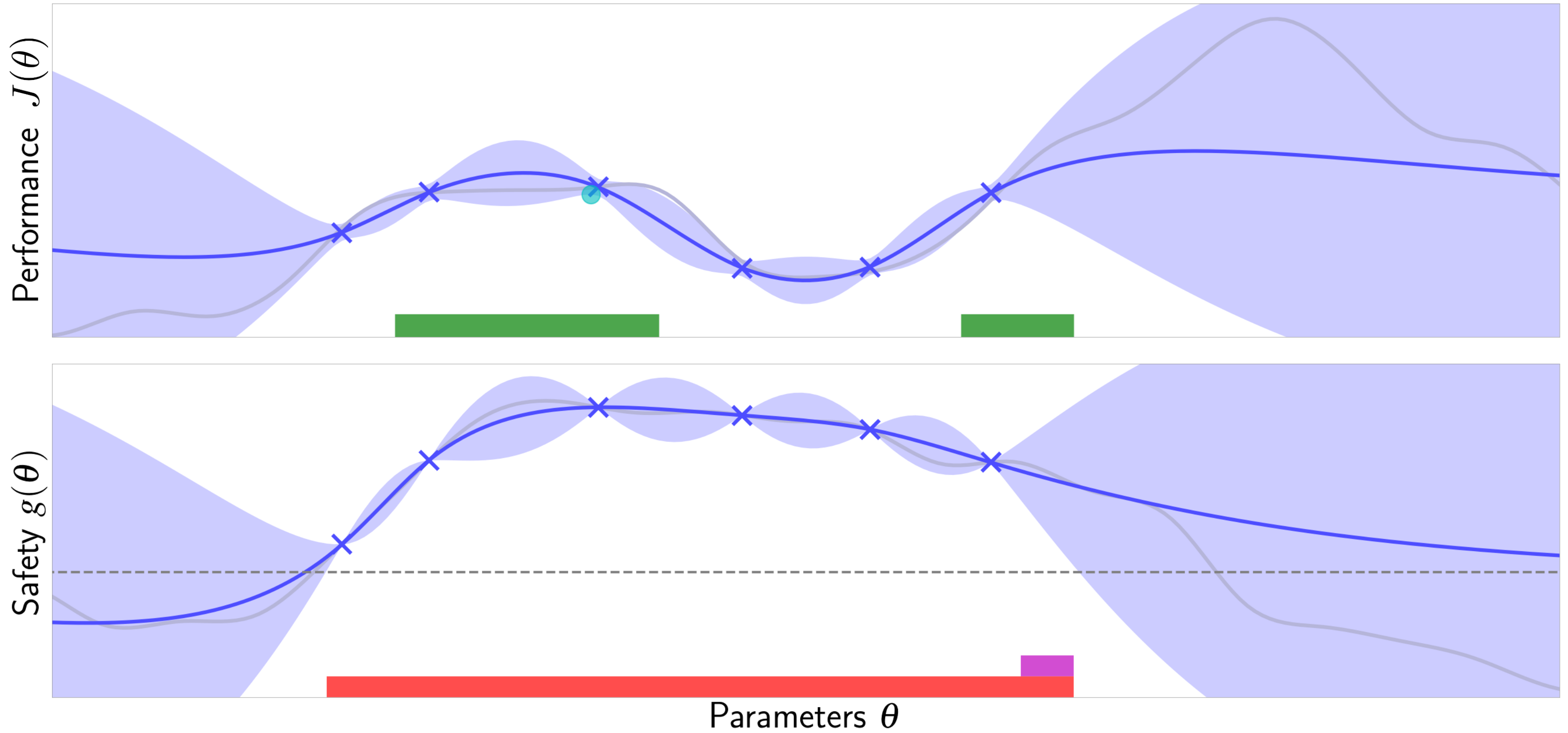
SafeOPT: Constrained Bayesian optimization



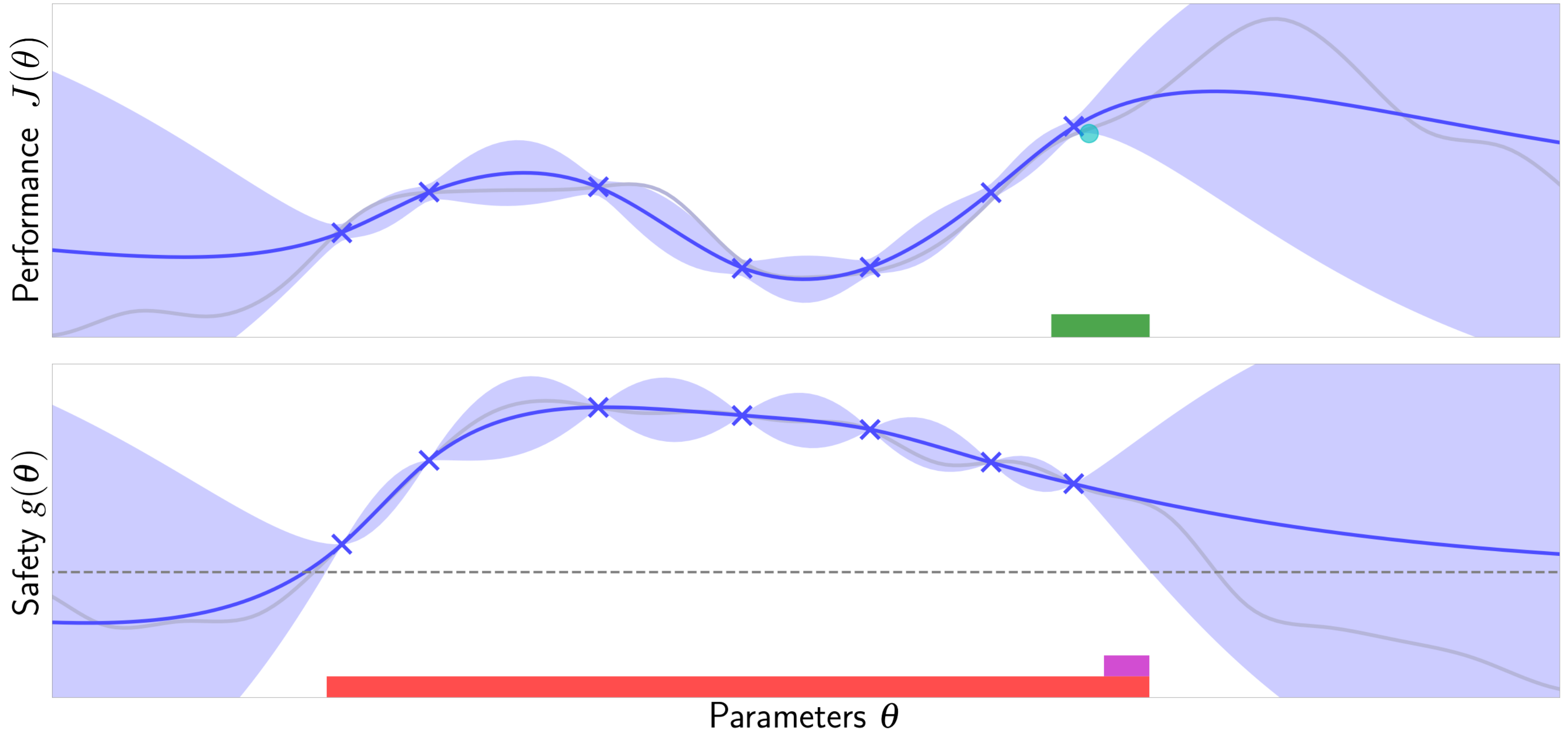
SafeOPT: Constrained Bayesian optimization



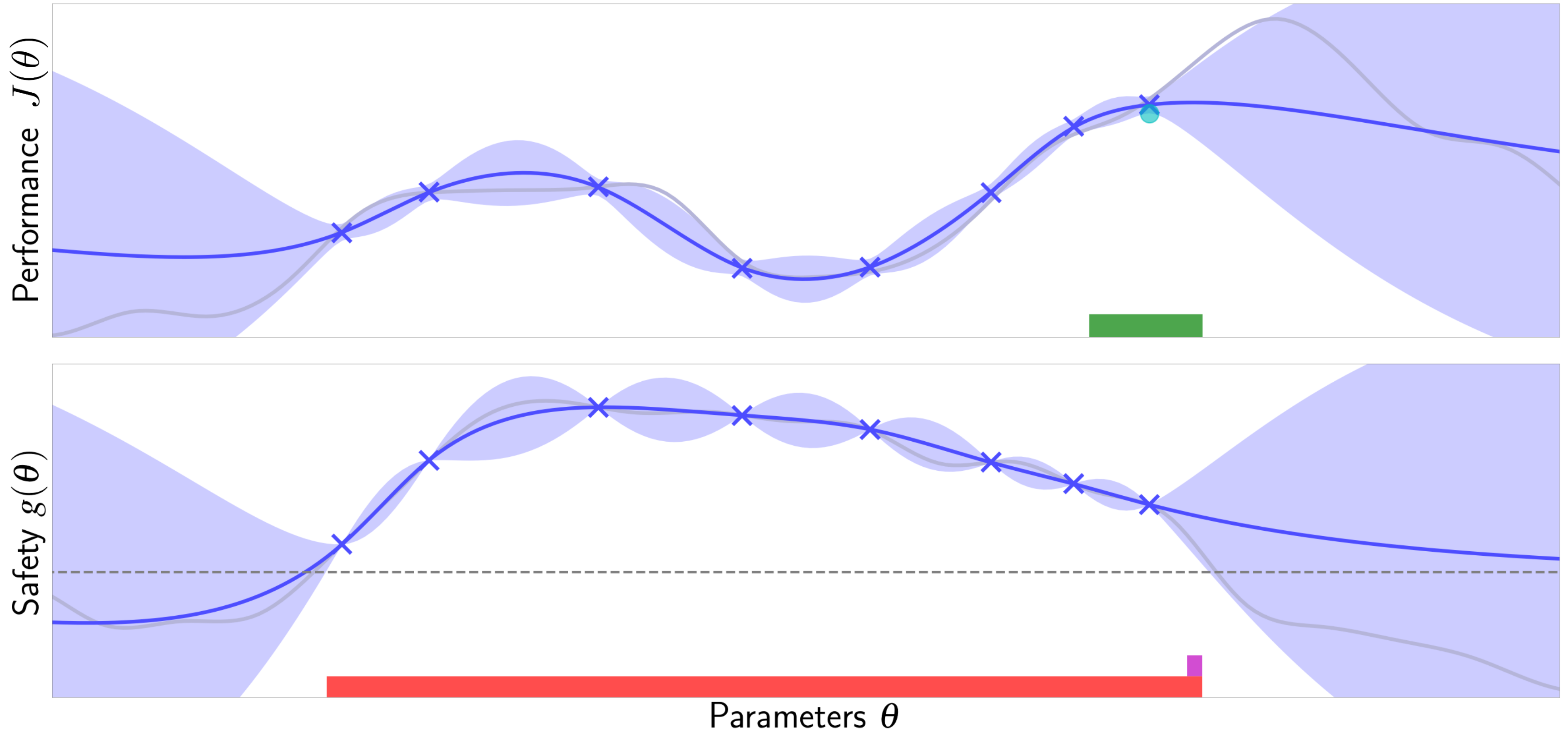
SafeOPT: Constrained Bayesian optimization



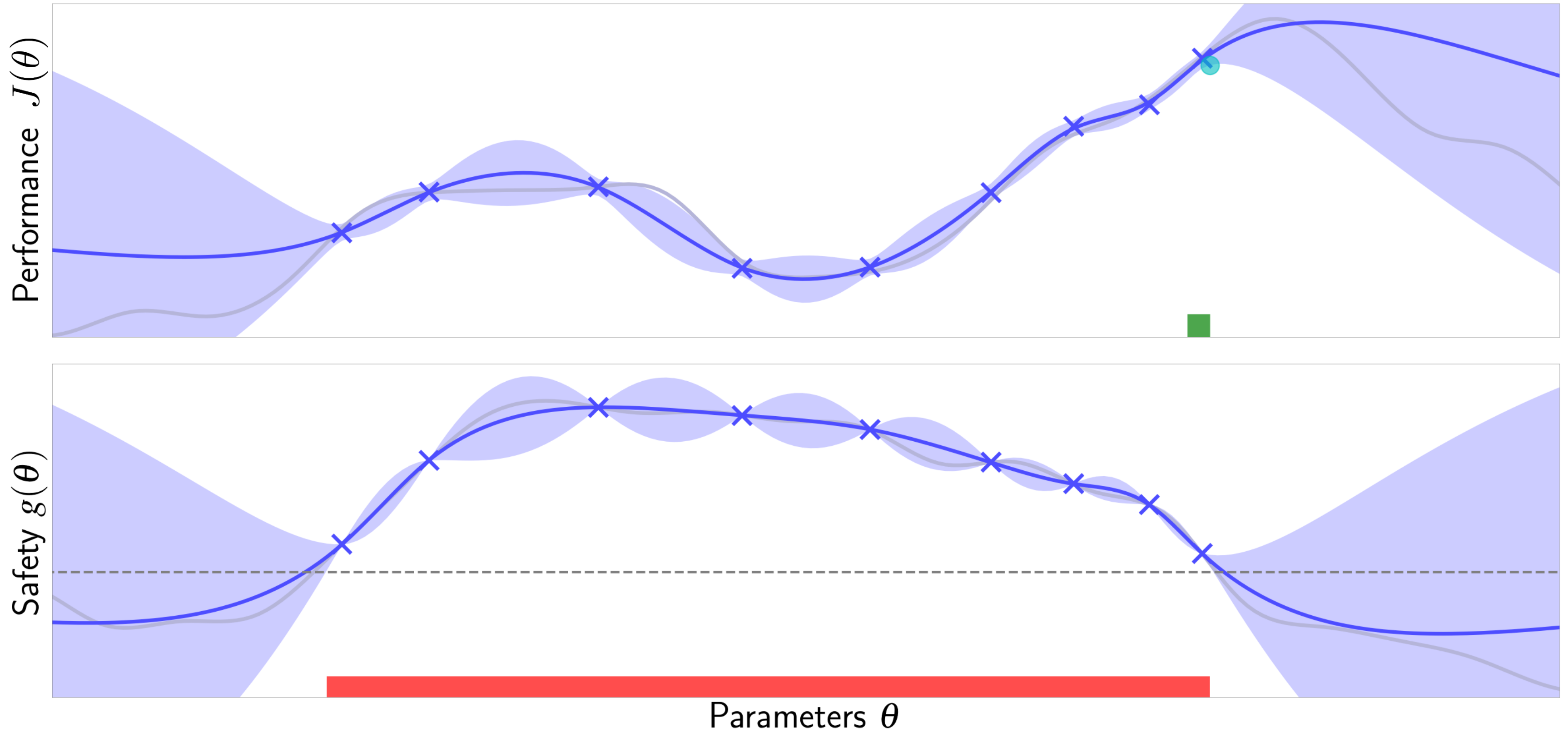
SafeOPT: Constrained Bayesian optimization



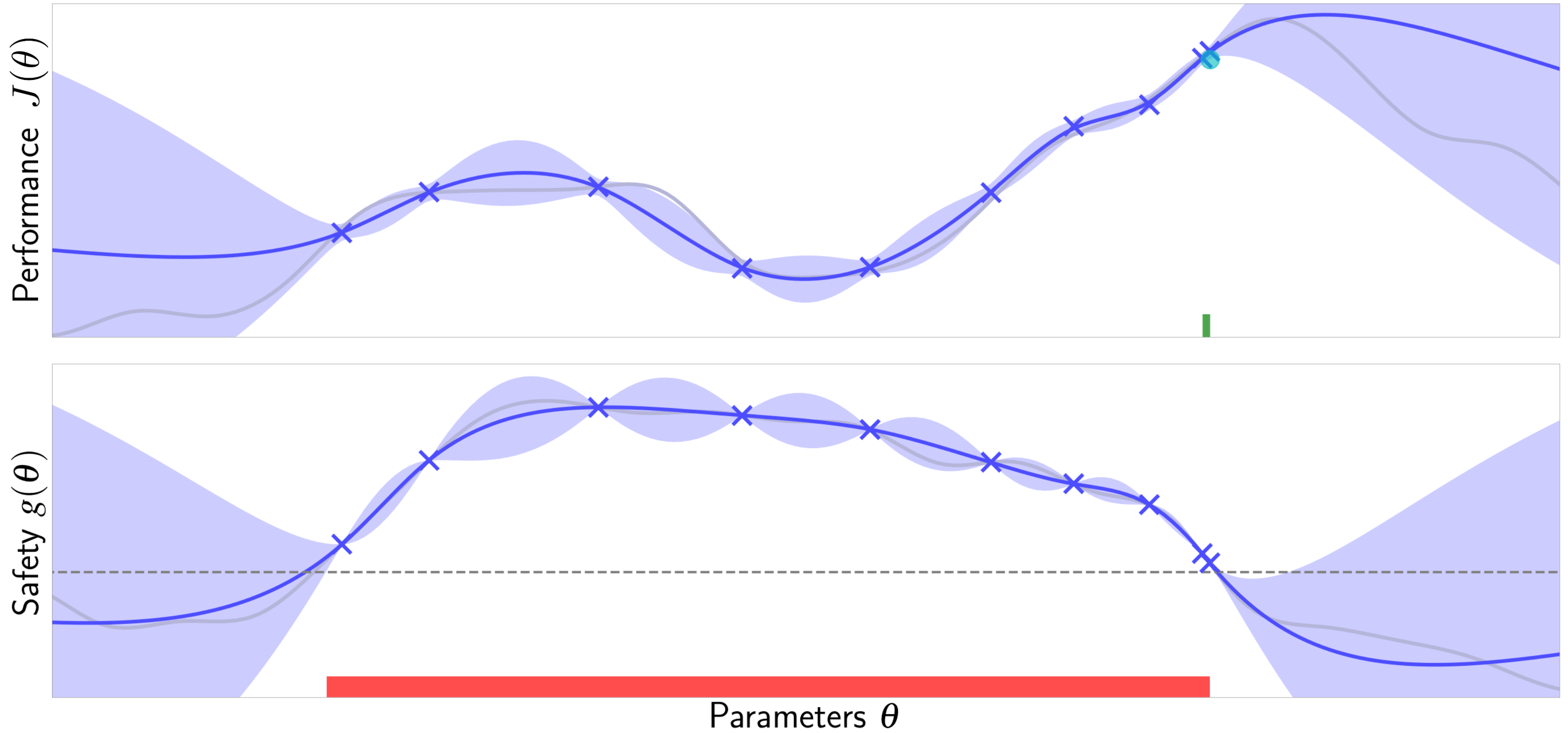
SafeOPT: Constrained Bayesian optimization



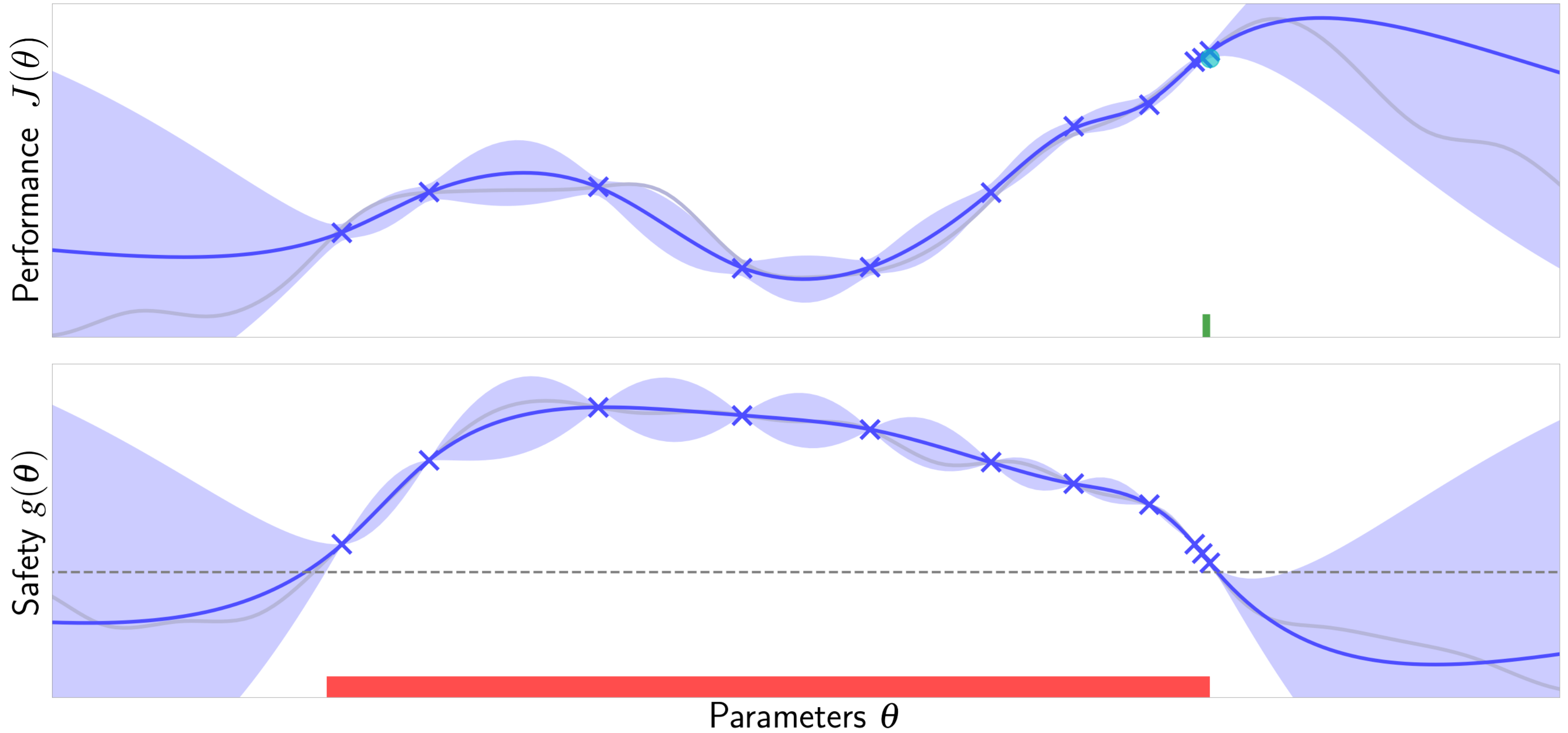
SafeOPT: Constrained Bayesian optimization



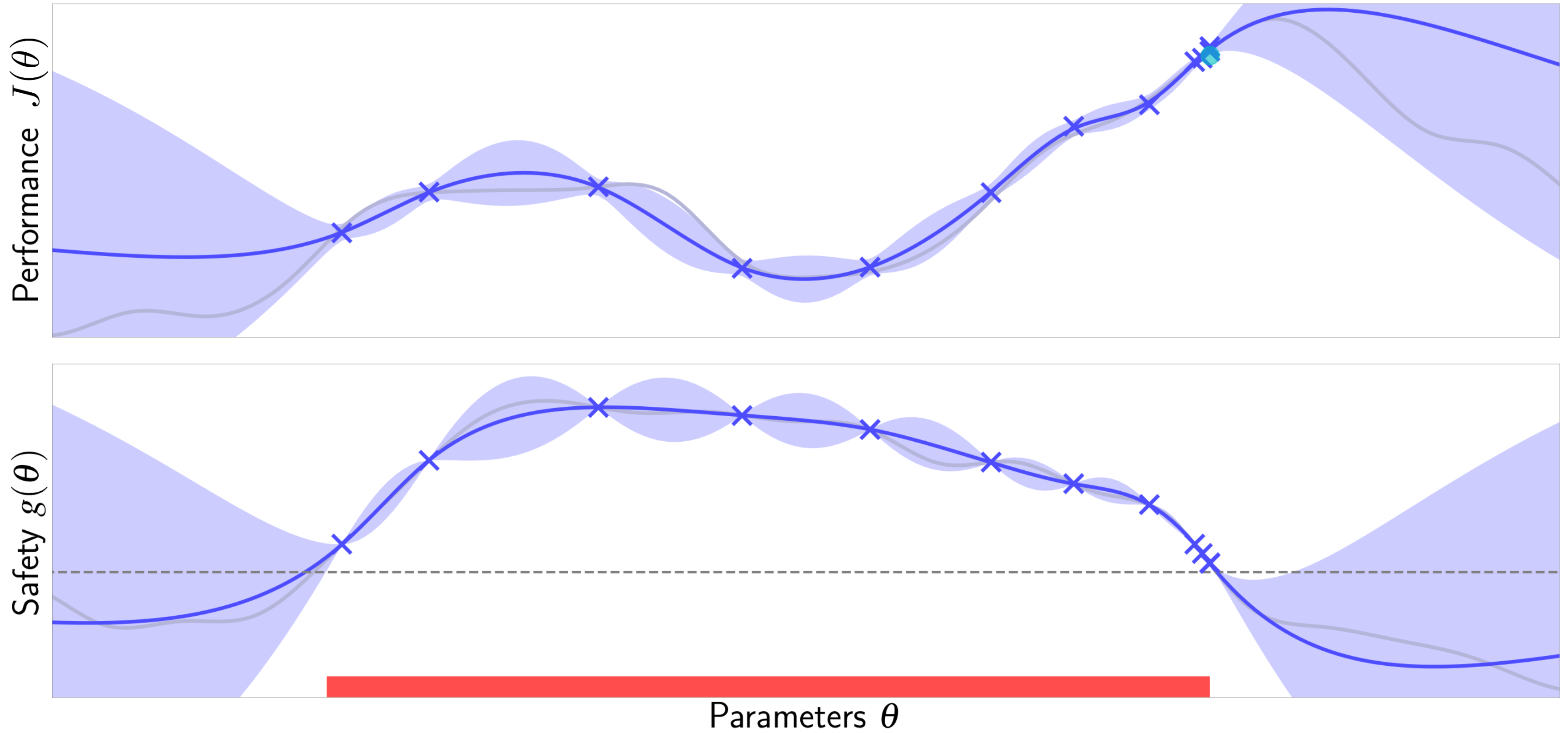
SafeOPT: Constrained Bayesian optimization



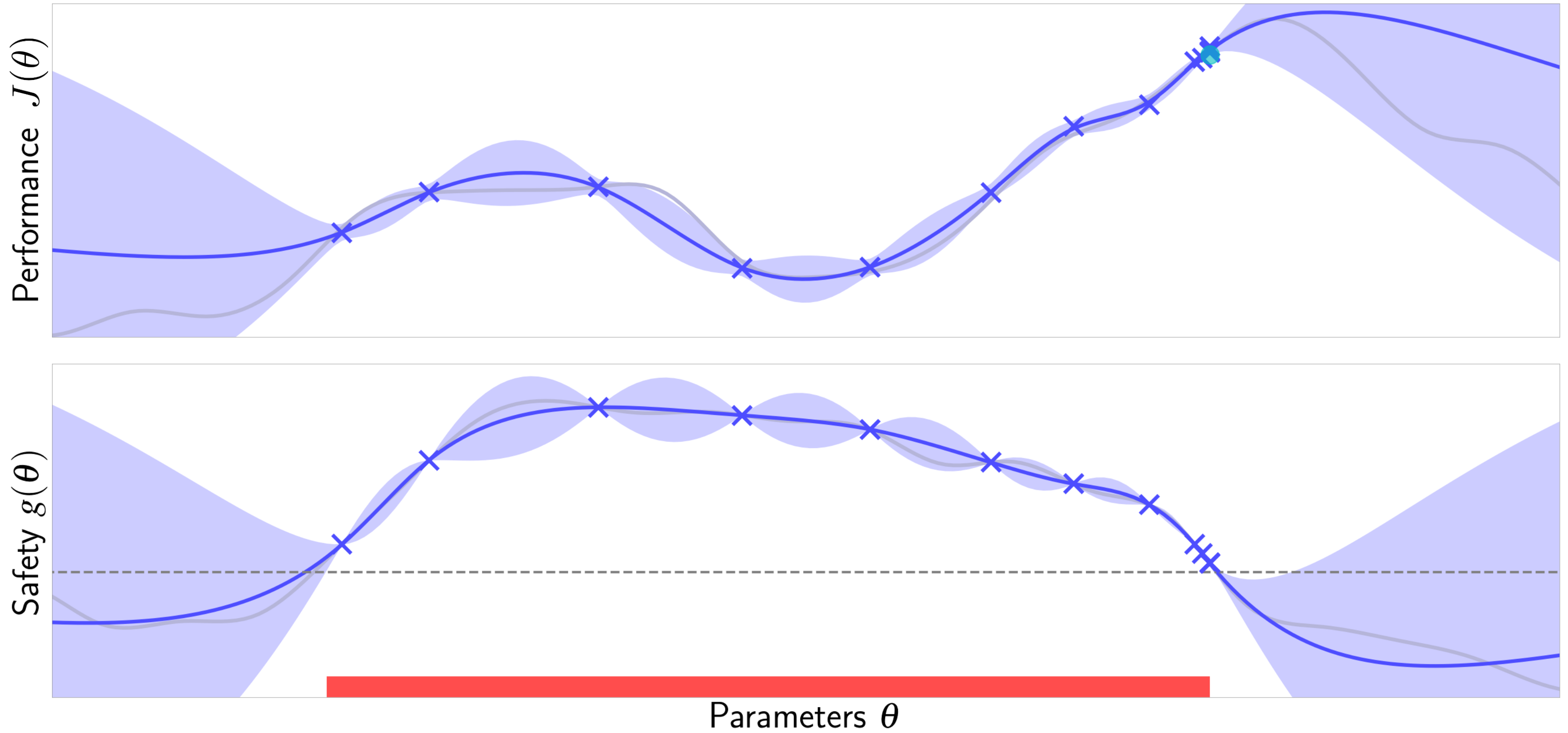
SafeOPT: Constrained Bayesian optimization



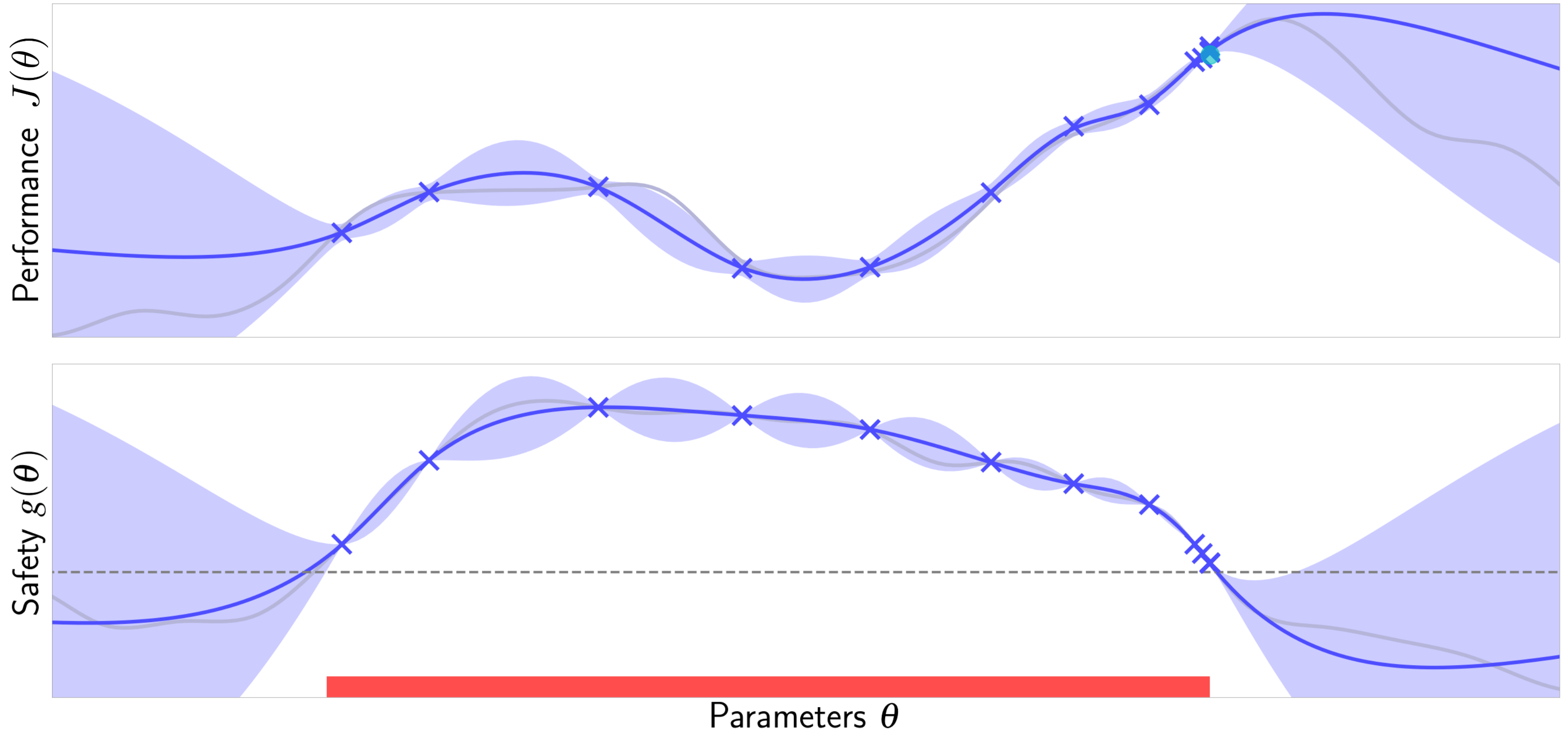
SafeOPT: Constrained Bayesian optimization



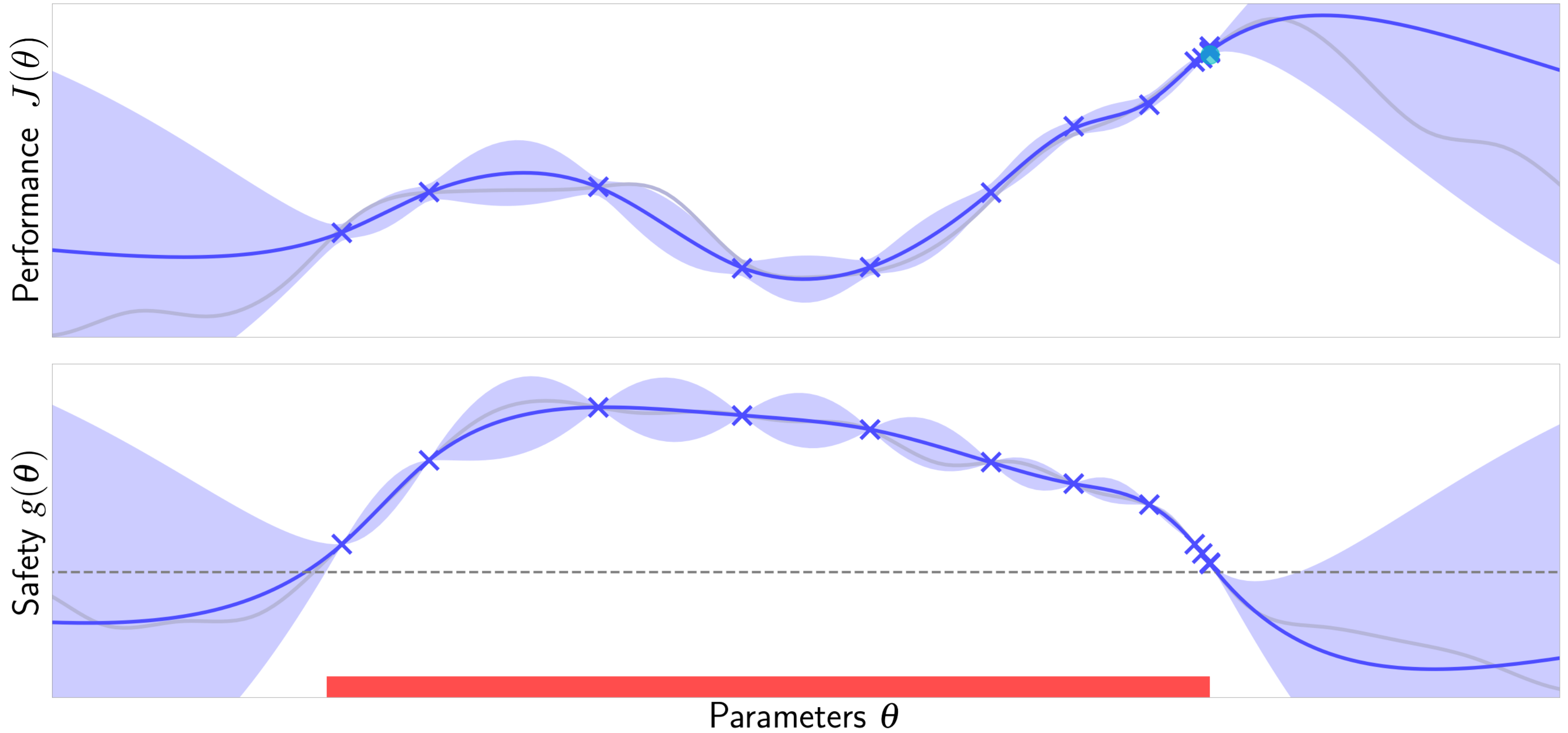
SafeOPT: Constrained Bayesian optimization



SafeOPT: Constrained Bayesian optimization



SafeOPT: Constrained Bayesian optimization



Theorem (informal):

Under suitable conditions on the kernel and on J, g , there exists a function $T(\epsilon, \delta)$ such that for any $\epsilon > 0$ and $\delta > 0$, it holds with probability at least $1 - \delta$ that

- 1) SAFEOPT **never makes an unsafe** decision
- 2) After at most $T(\epsilon, \delta)$ iterations, it found an **ϵ -optimal reachable** point

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

Safe Exploration for Optimization with Gaussian Processes

Y. Sui, A. Gotovos, J.W. Burdick, A. Krause

Safe Exploration for Active Learning with Gaussian Processes

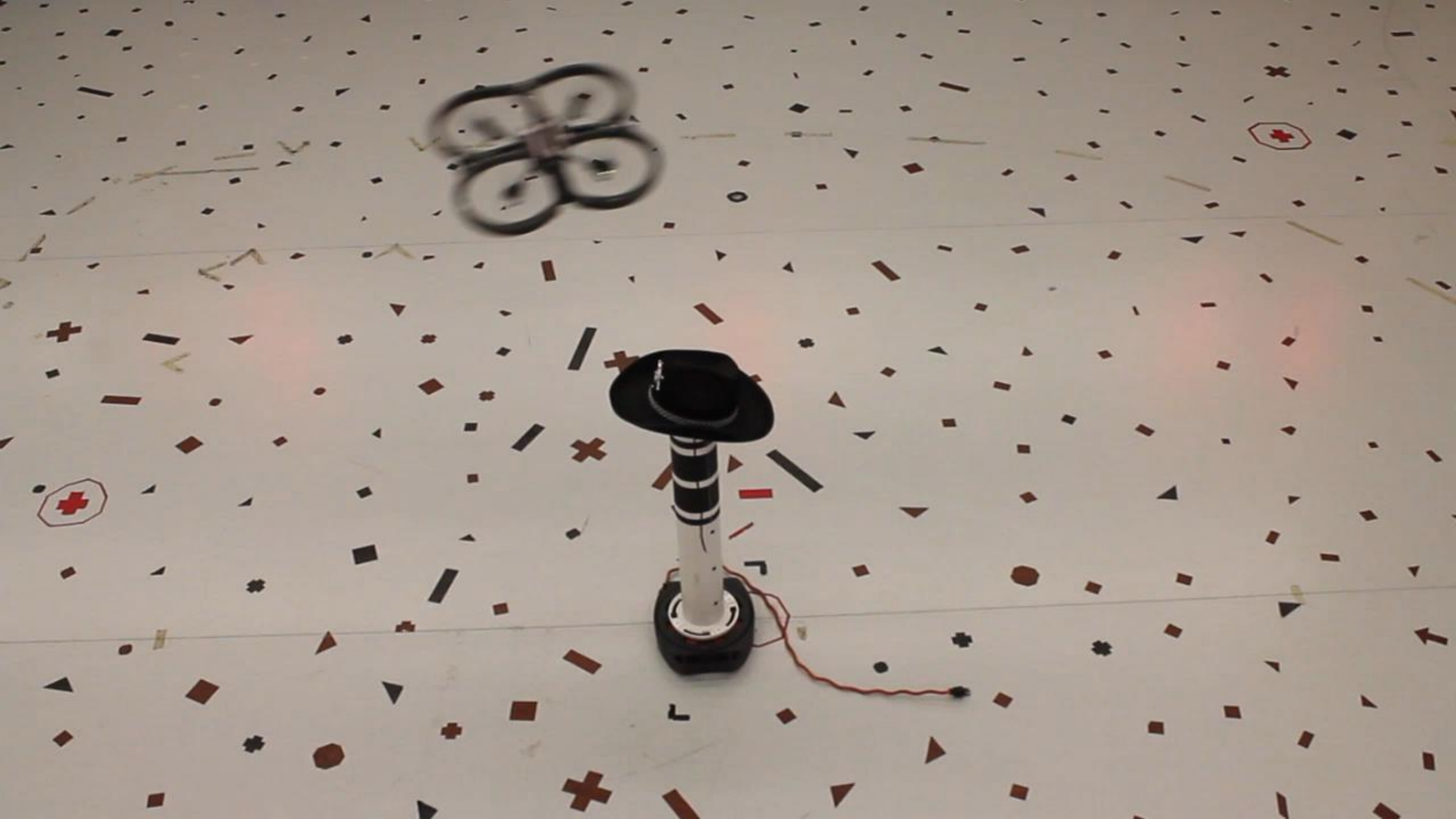
J. Schreiter, D. Nguyen-Tuong, M. Eberts, B. Bischoff,
H. Markert, M. Toussaint

Bayesian Optimization with Safety Constraints: Safe and Automatic
Parameter Tuning in Robotics

F. Berkenkamp, A.P. Schoellig, A. Krause

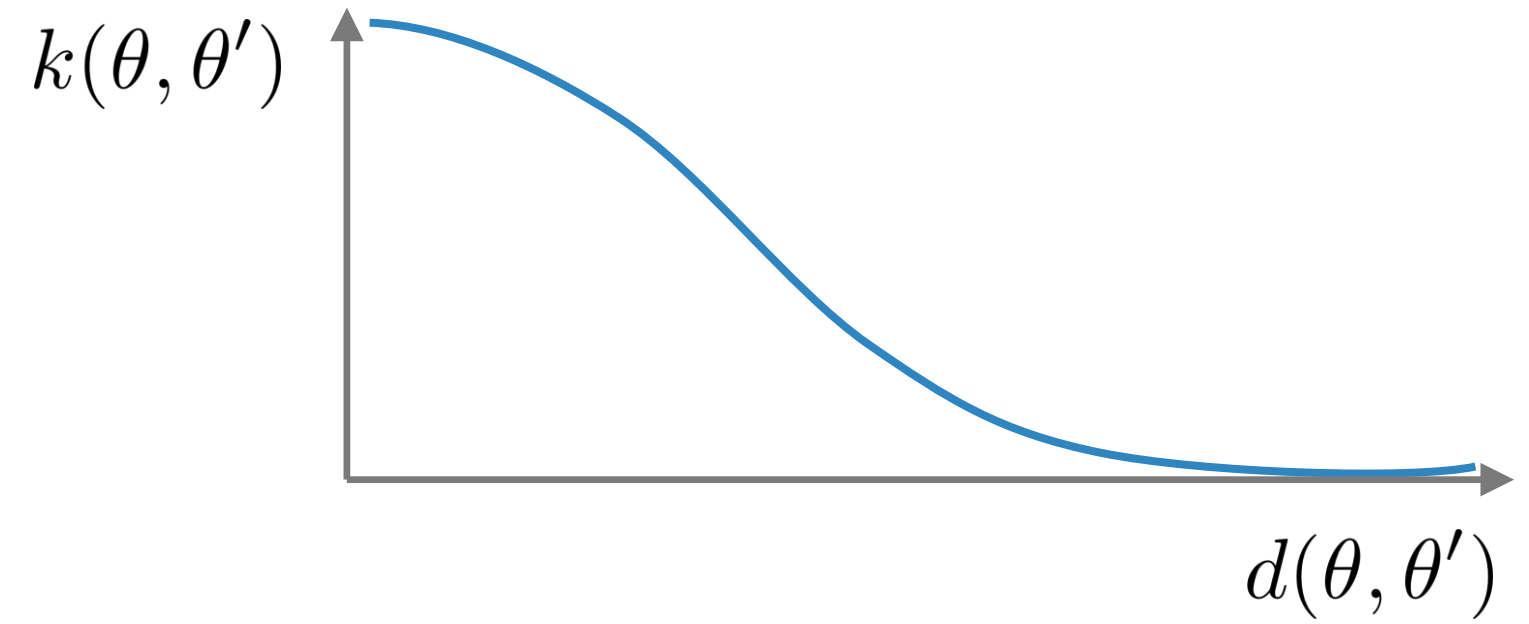


Video link: http://tiny.cc/icra16_video



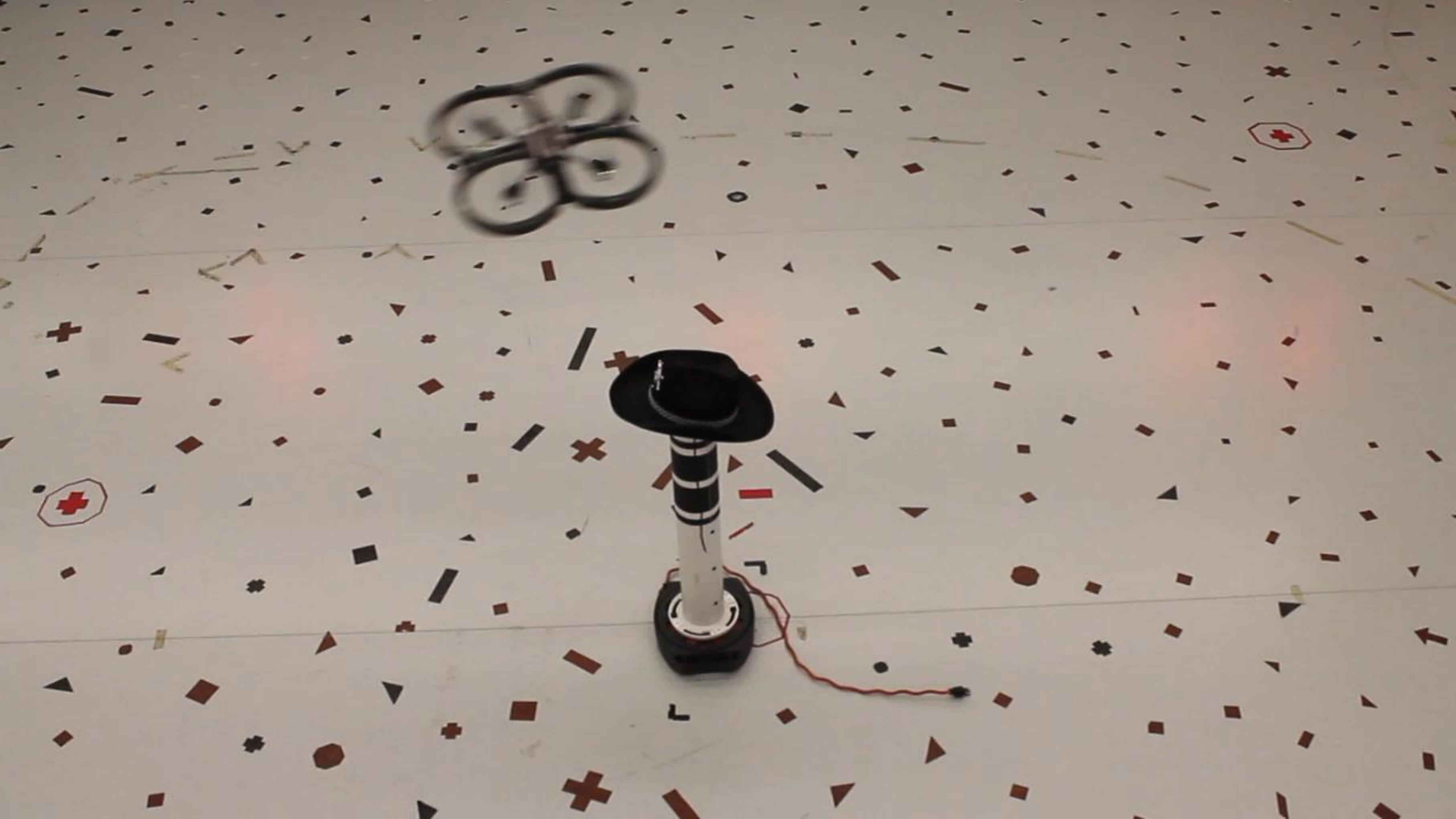
Modelling context

$$\text{Cov}[J(\theta), J(\theta')] = k(\theta, \theta')$$

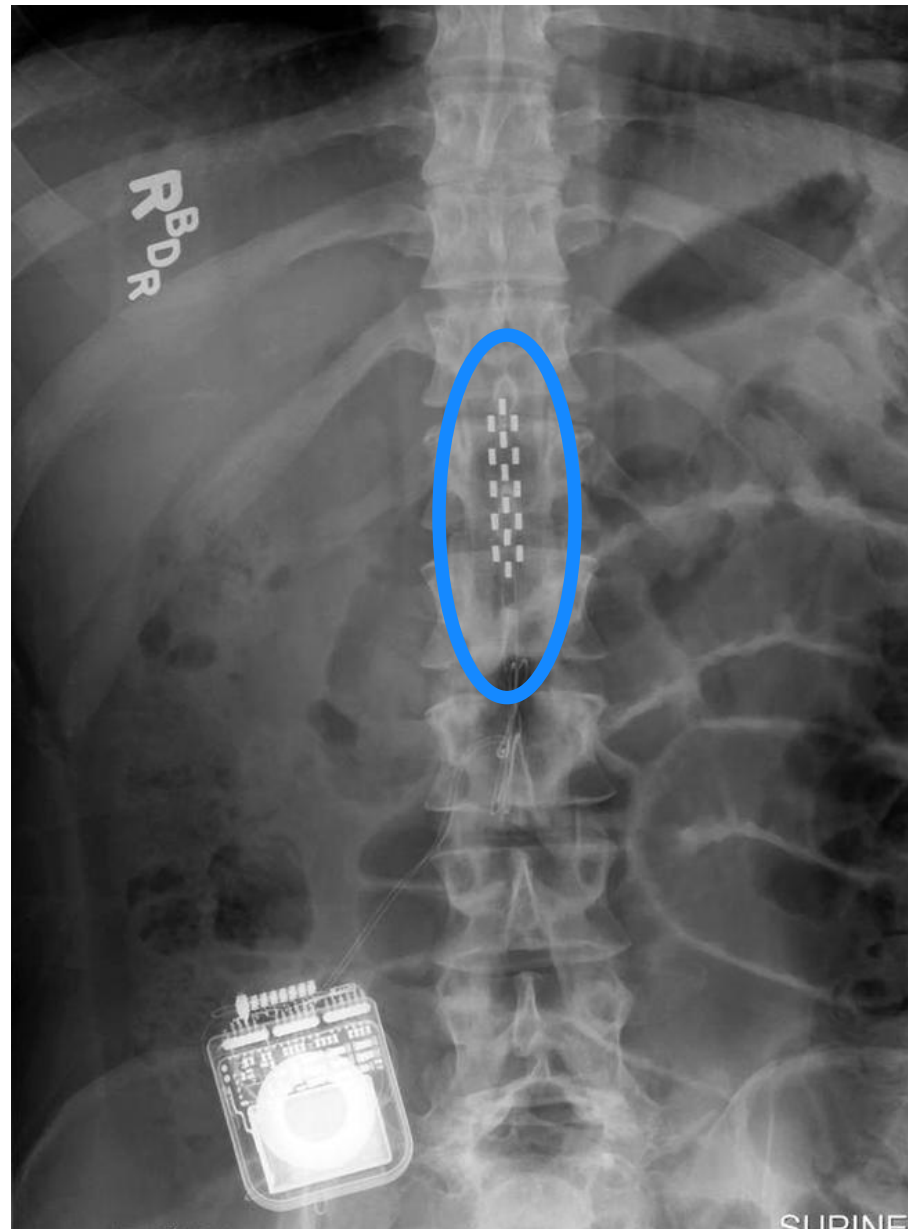


Additional parameters

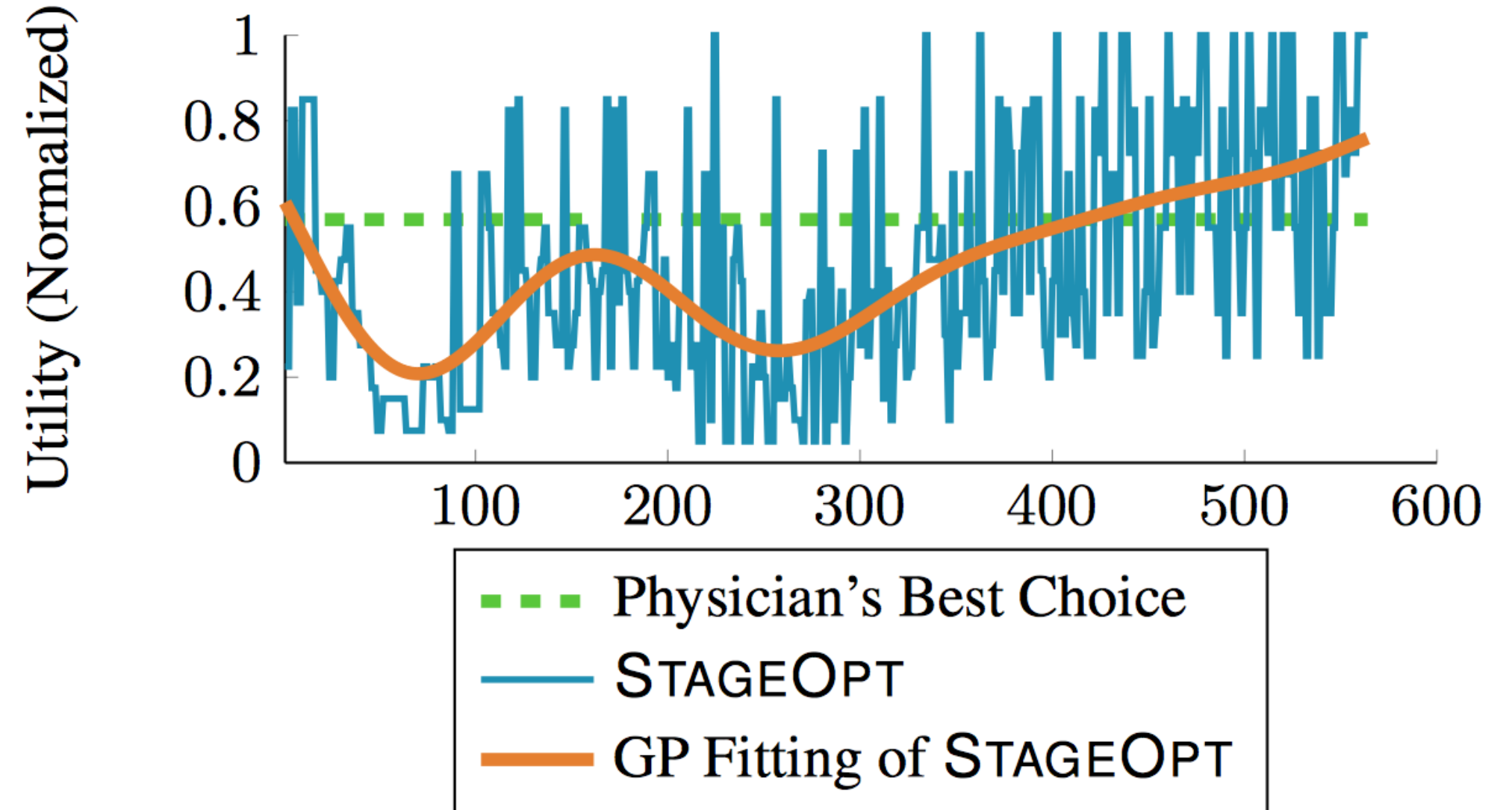
$$\text{Cov}[J(\theta, \mathbf{z}), J(\theta', \mathbf{z}')] = k(\theta, \theta') * k(\mathbf{z}, \mathbf{z}')$$



Therapeutic Spinal Cord Stimulation



S. Harkema, The Lancet, Elsevier



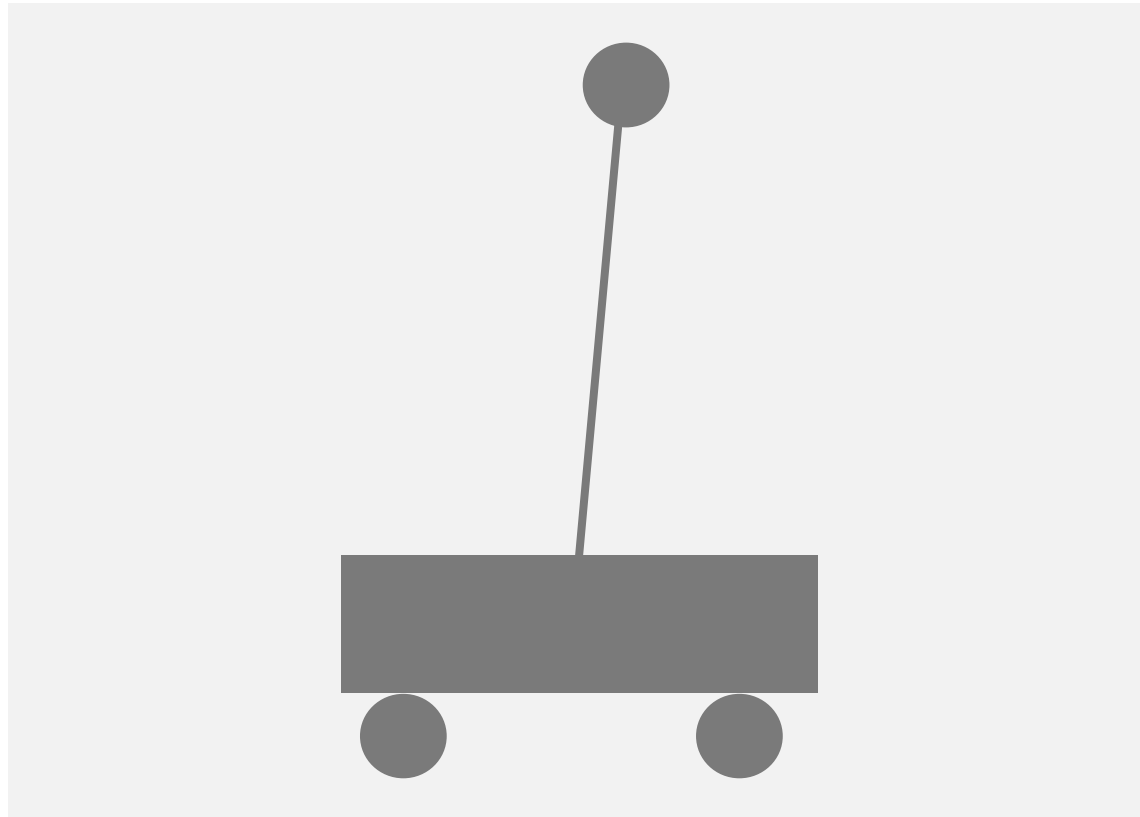
Stagewise Safe Bayesian Optimization with Gaussian Processes

Y. Sui, V. Zhuang, J. W. Burdick, Y. Yue

Multiple sources of information

Automatic tradeoff

cheap, inaccurate



expensive, accurate

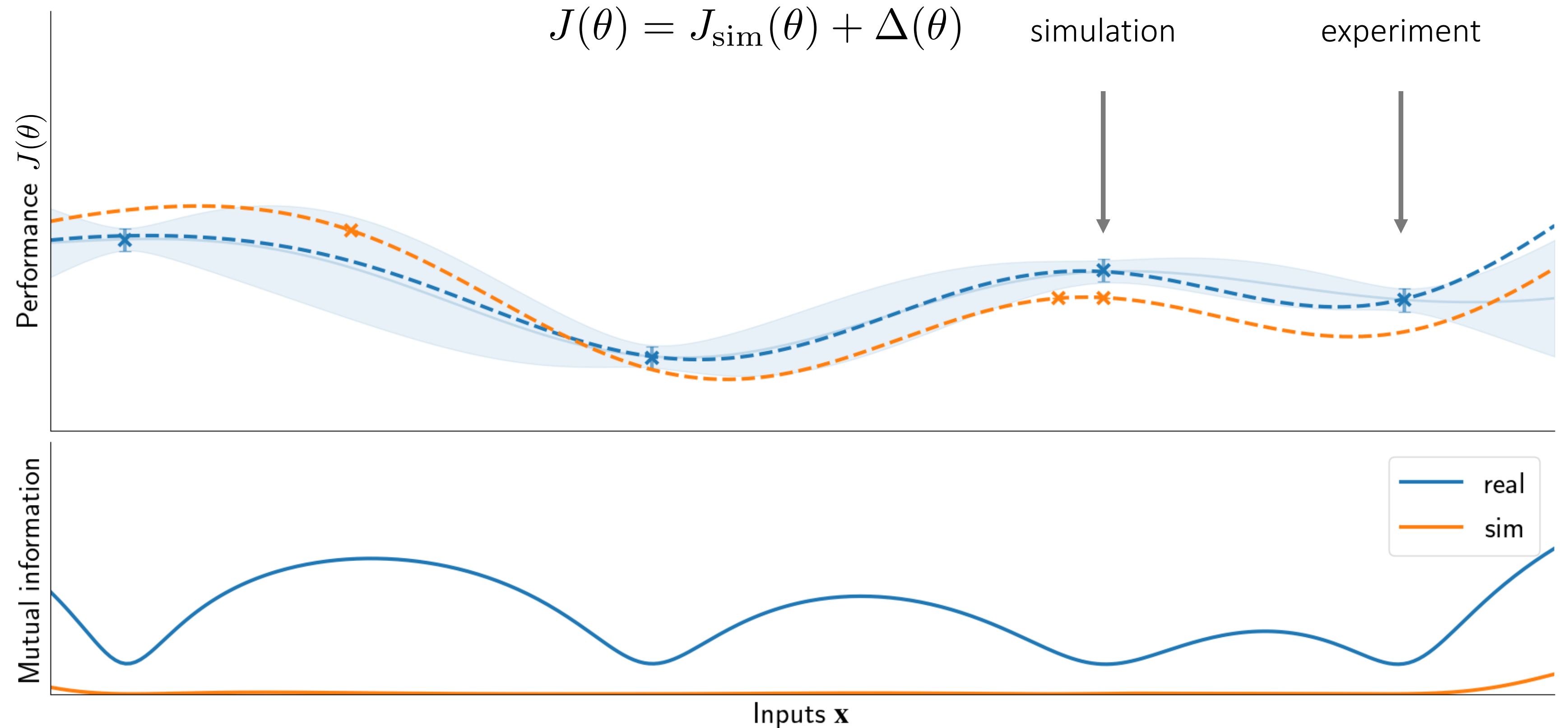


$$J(\theta) = J_{\text{sim}}(\theta) + \Delta(\theta)$$

Virtual vs. Real: Trading Off Simulations and Physical Experiments in Reinforcement Learning with Bayesian Optimization

A. Marco, F. Berkenkamp, P. Hennig, A. Schöllig, A. Krause, S. Schaal, S. Trimpe, ICRA'17

Modeling this in a Gaussian process



Performance improvement

Starting controller

Learned controller

Video at <https://youtu.be/oq9Qgq1lpp8>



Safe reinforcement learning

Statistical models to guarantee safety

Model-free

Model-based

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

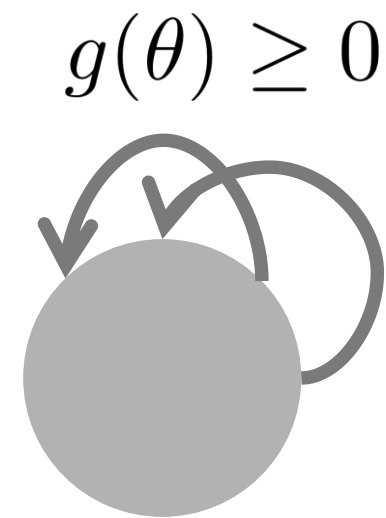
$$[s_{t+1}, r_t] \sim P(\cdot \mid s_t, a_t; \theta)$$

Estimate $J(\theta)$
and optimize

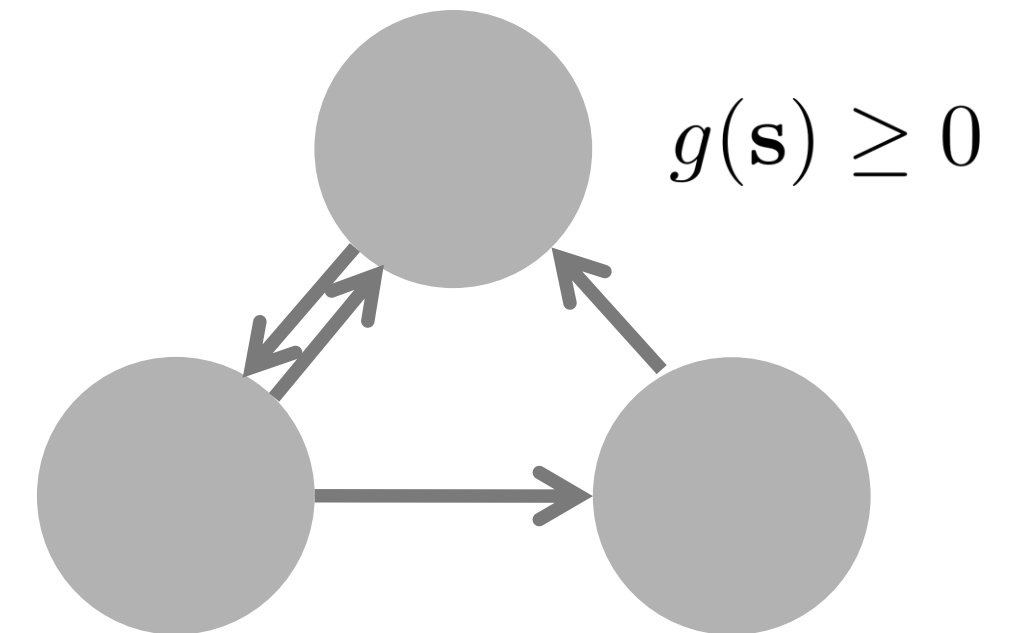
Estimate/identify,
then plan/control

From bandits to Markov decision processes

Bandit



Markov Decision Process



Can use the same Bayesian model to determine safety of states

Challenges with long-term action dependencies

Non-ergodic MDP

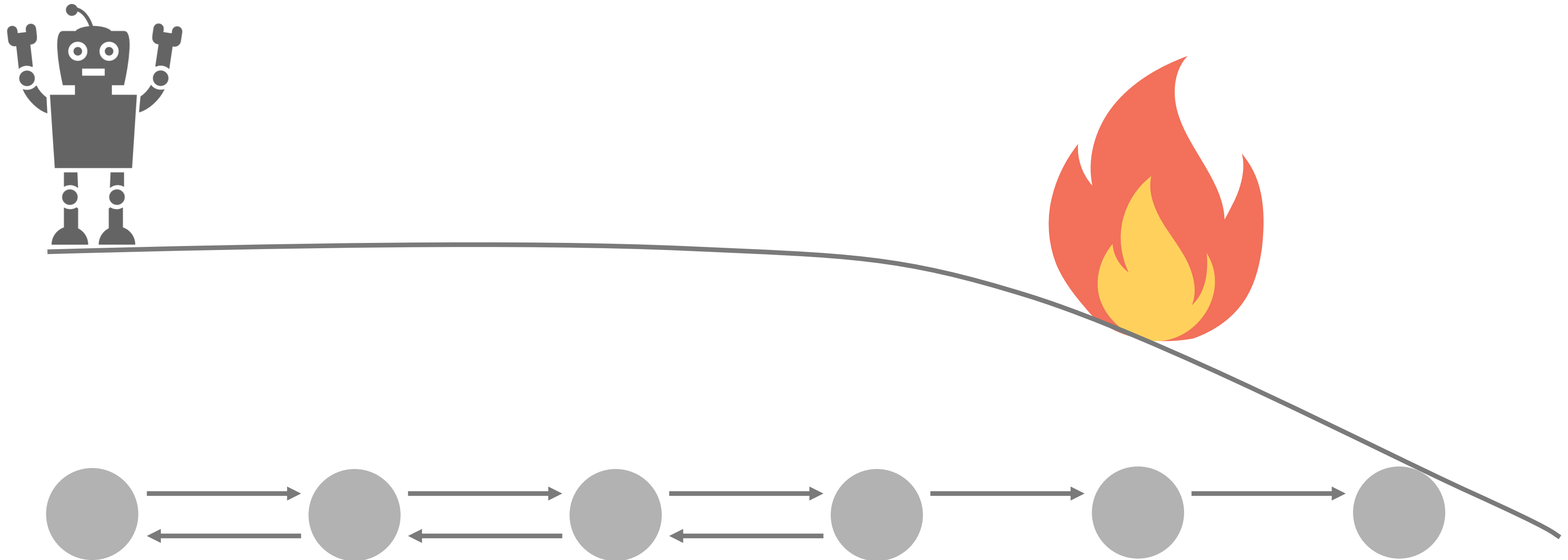
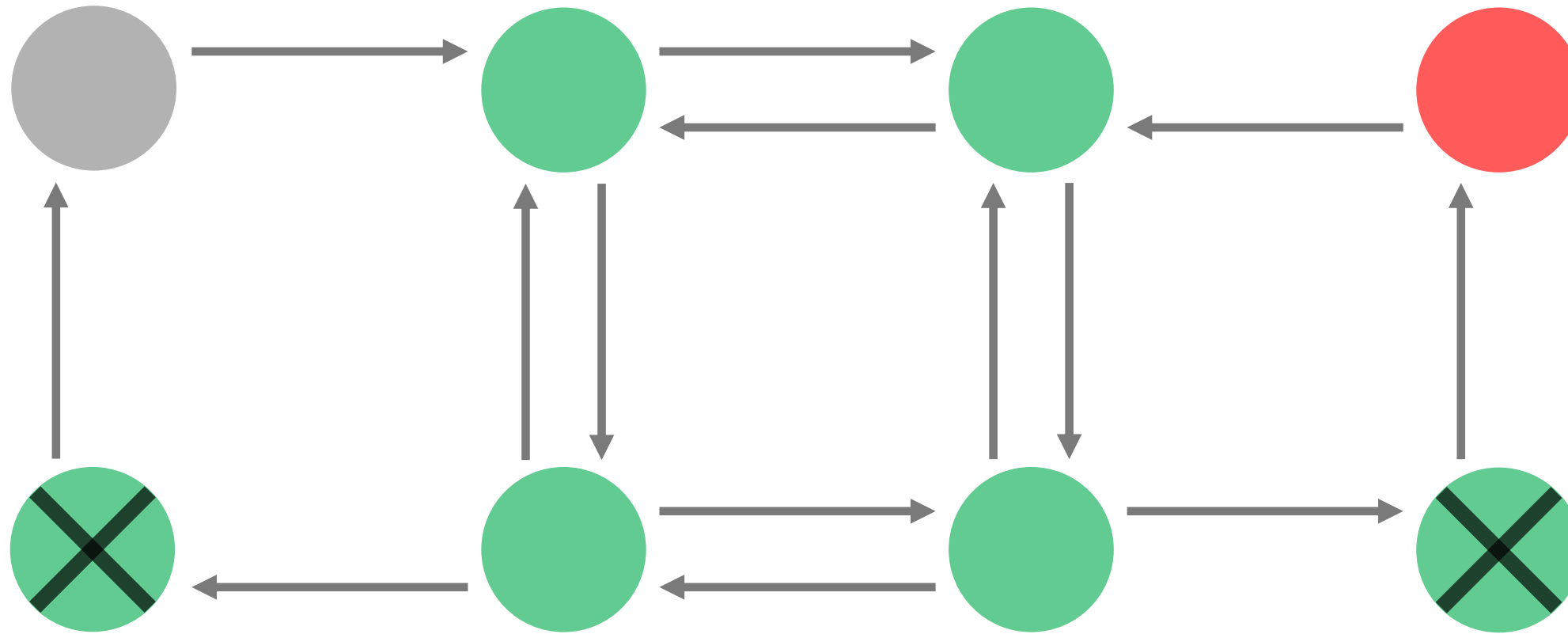


Image: Plainicon, VectorsMarket, <https://flaticon.com>

Rendering exploration safe



Exploration:

Reduce model uncertainty

Only visit states from which
the agent can recover safely

Safe Exploration in Markov Decision Processes

T.M. Moldovan, P. Abbeel, ICML, 2012

Safe Exploration in Finite Markov Decision Processes with Gaussian Processes

M. Turchetta, F. Berkenkamp, A. Krause, NIPS, 2016

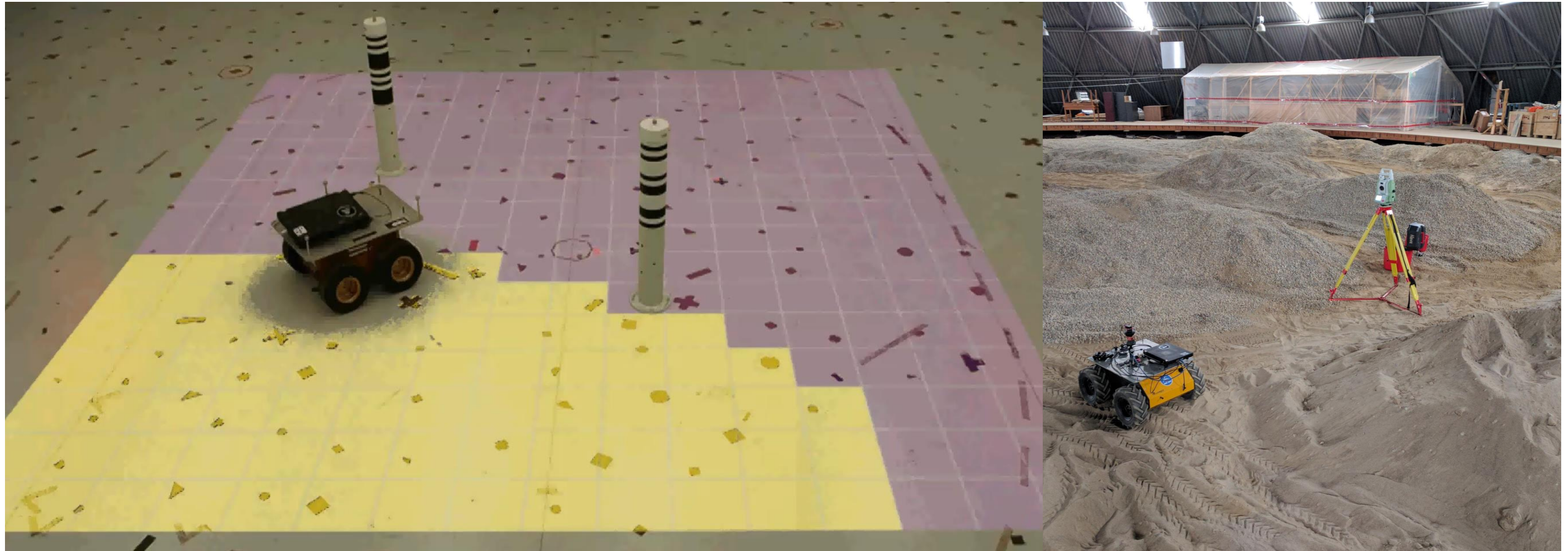
Safe Exploration and Optimization of Constrained MDPs using Gaussian Processes

Akifumi Wachi, Yanan Sui, Yisong Yue, Masahiro Ono, AAAI, 2018

Safe Control under Uncertainty

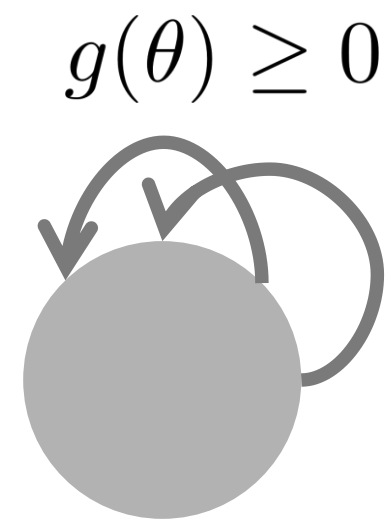
D. Sadigh, A. Kapoor, RSS, 2016

On a real robot

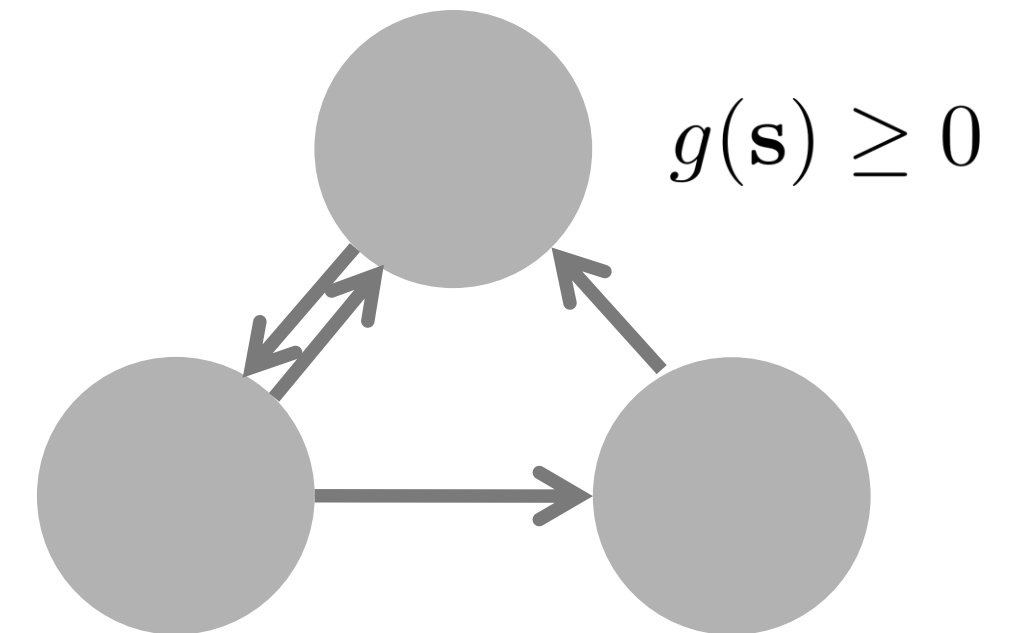


From bandits to Markov decision processes

Bandit



Markov Decision Process



Next: model-based reinforcement learning

Reinforcement learning (recap)

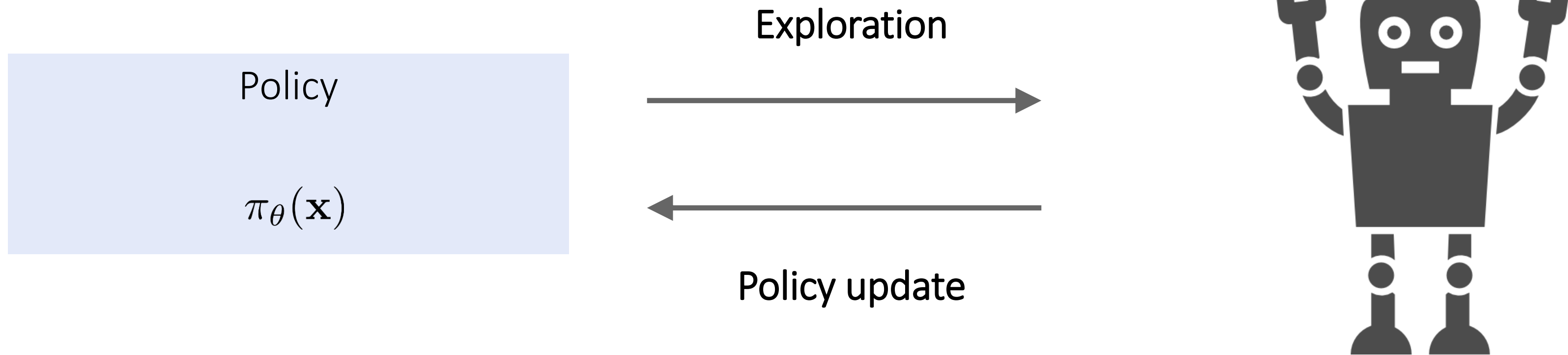


Image: Plainicon, <https://flaticon.com>

Model-based reinforcement learning

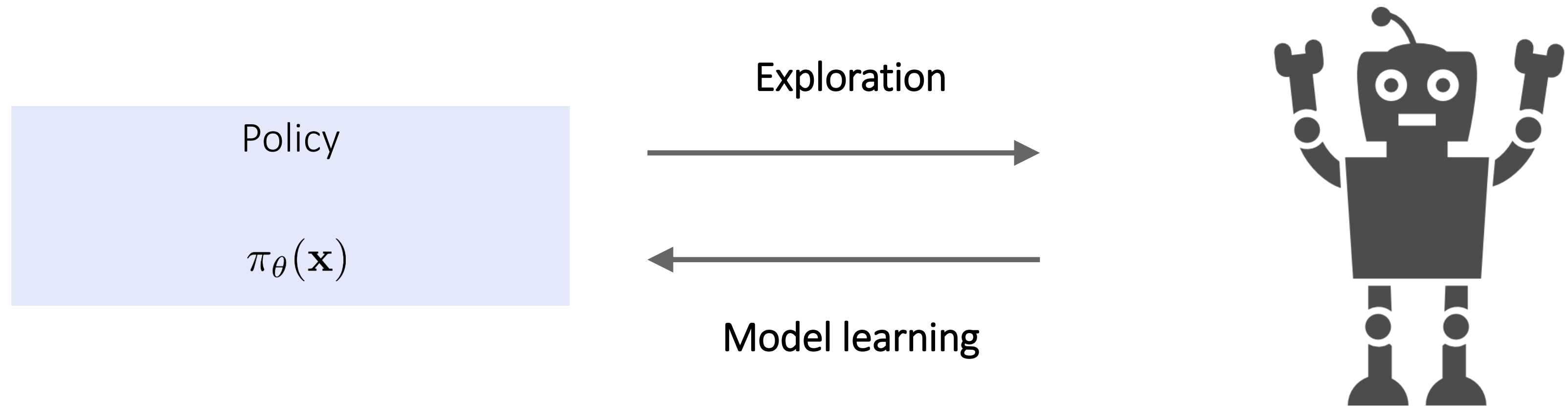


Image: Plainicon, <https://flaticon.com>

Safe model-based reinforcement learning

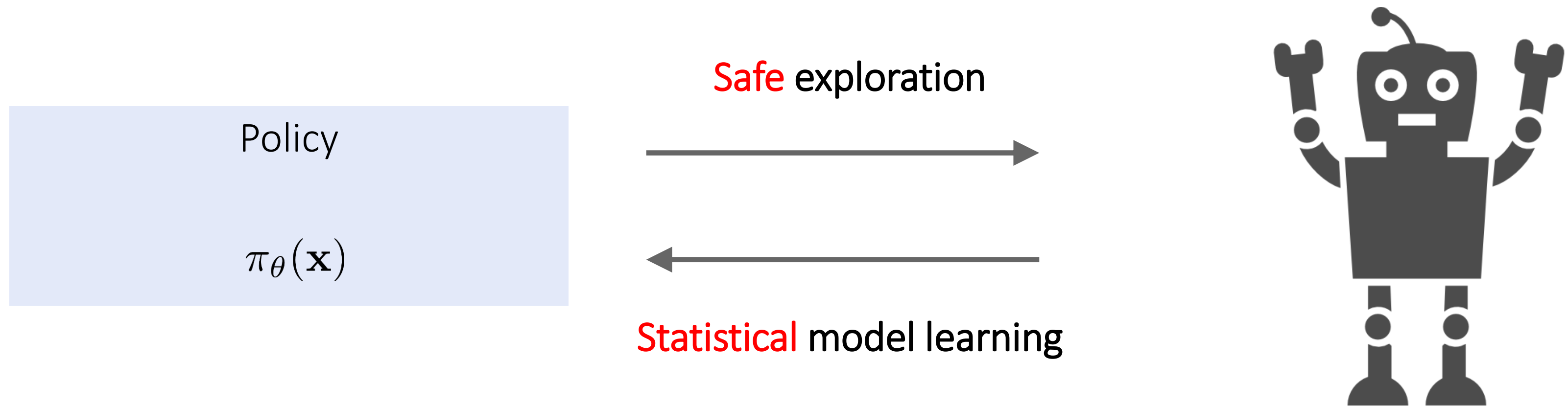


Image: Plainicon, <https://flaticon.com>

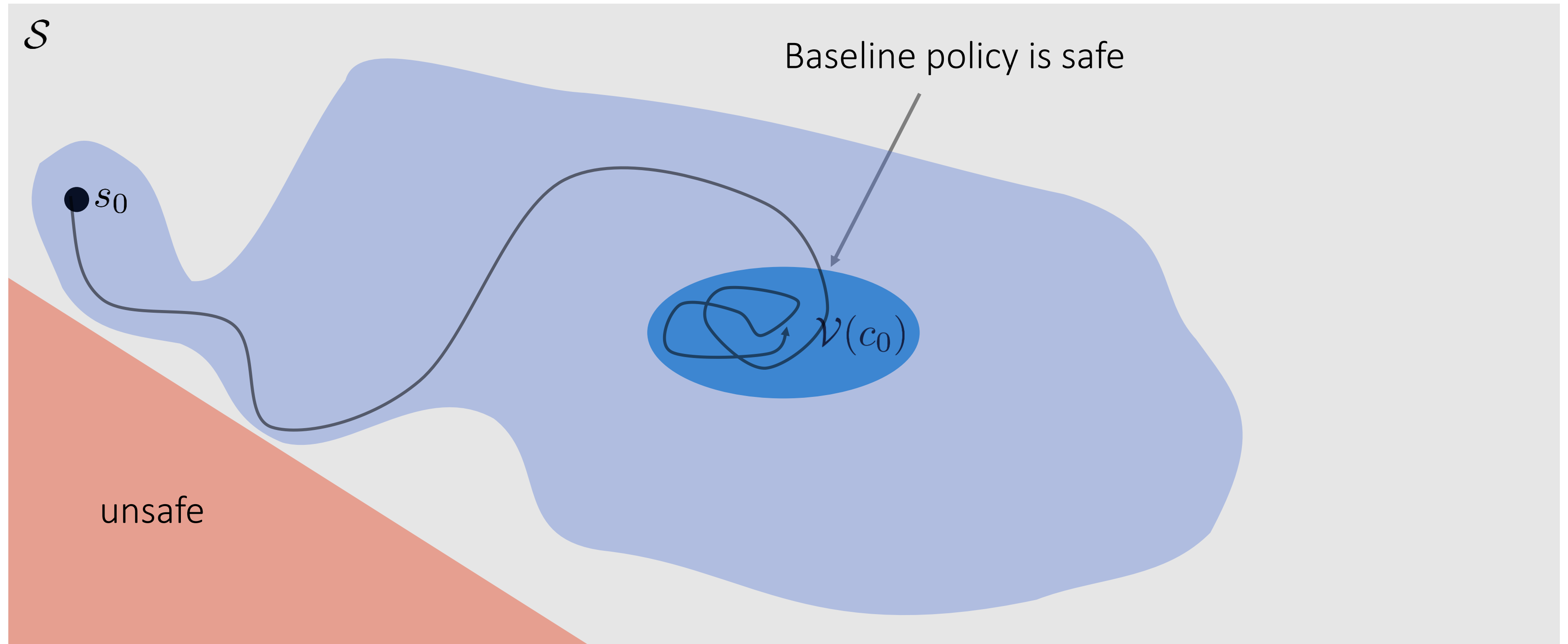
A Bayesian dynamics model

Dynamics

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



Region of attraction



Linear case

$$s_{t+1} = \mathbf{A} s_t + \mathbf{B} a_t$$



Uncertainty about entries

Designing safe controllers for quadratic costs is a convex optimization problem

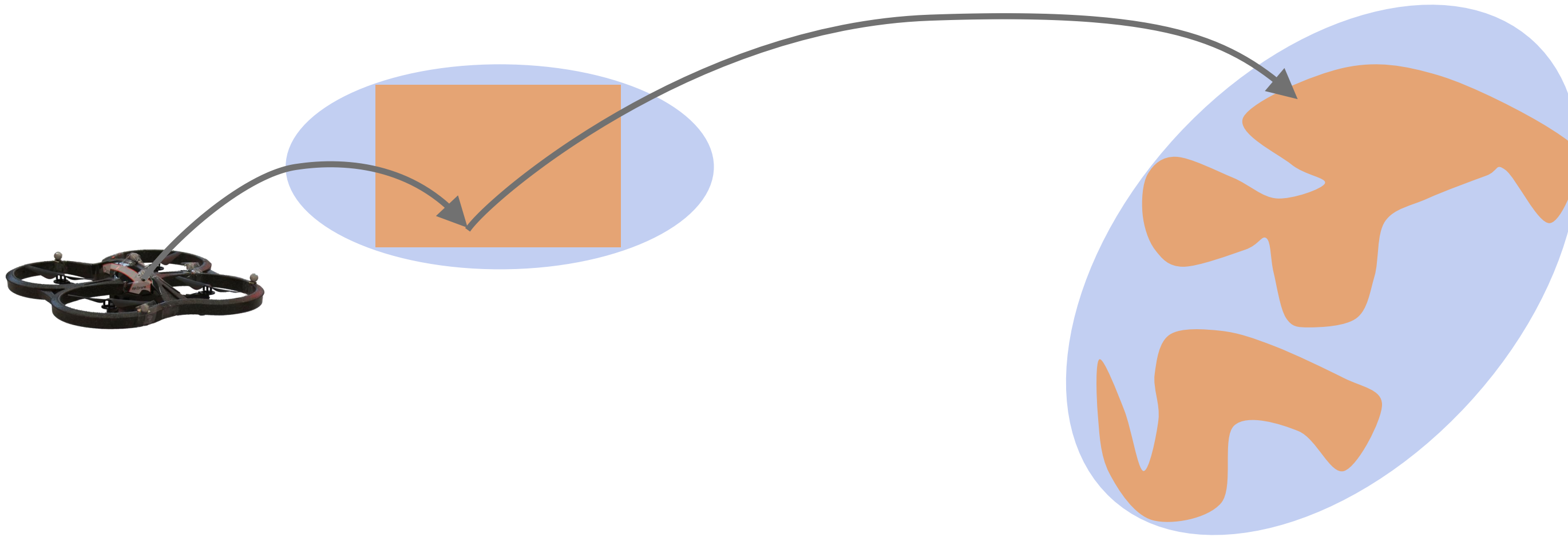
Safe and Robust Learning Control with Gaussian Processes

F. Berkenkamp, A.P. Schoellig, ECC, 2015

Regret Bounds for Robust Adaptive Control of the Linear Quadratic Regulator

S. Dean, H. Mania, N. Matni, B. Recht, S. Tu, arXiv, 2018

Forwards-propagating uncertain, nonlinear dynamics

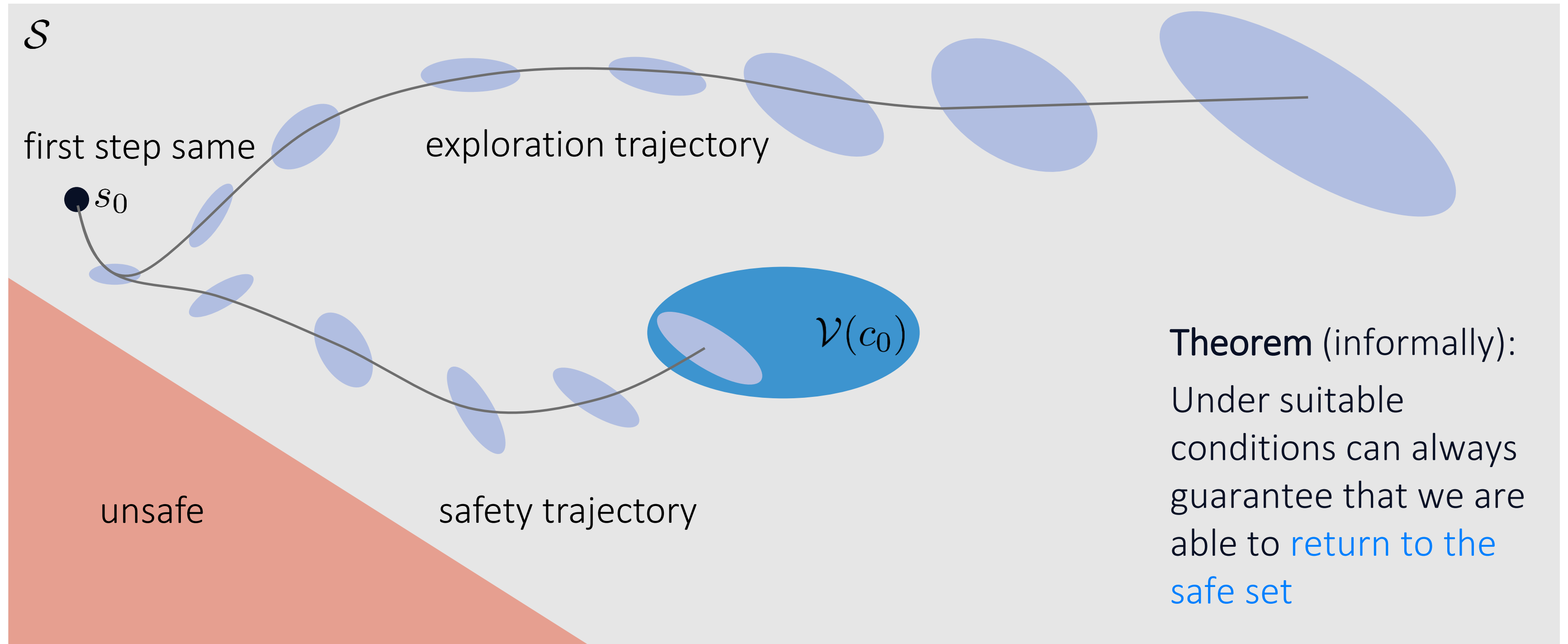


Outer approximation contains true dynamics for all time steps with probability at least $1 - \delta$

Learning-based Model Predictive Control for Safe Exploration

T. Koller, F. Berkenkamp, M. Turchetta, A. Krause, CDC, 2018

Region of attraction



Theorem (informally):
Under suitable
conditions can always
guarantee that we are
able to [return to the
safe set](#)

Model predictive control references

Learning-based Model Predictive Control for Safe Exploration

T. Koller, F. Berkenkamp, M. Turchetta, A. Krause, CDC, 2018

Reachability-Based Safe Learning with Gaussian Processes

A.K. Akametalu, J.F. Fisac, J.H. Gillula, S. Kaynama, M.N. Zeilinger, C.J. Tomlin, CDC, 2014

Robust constrained learning-based NMPC enabling reliable mobile robot path tracking

C.J. Ostafew, A.P. Schoellig, T.D. Barfoot, IJRR, 2016

Data-Efficient Reinforcement Learning with Probabilistic Model Predictive Control

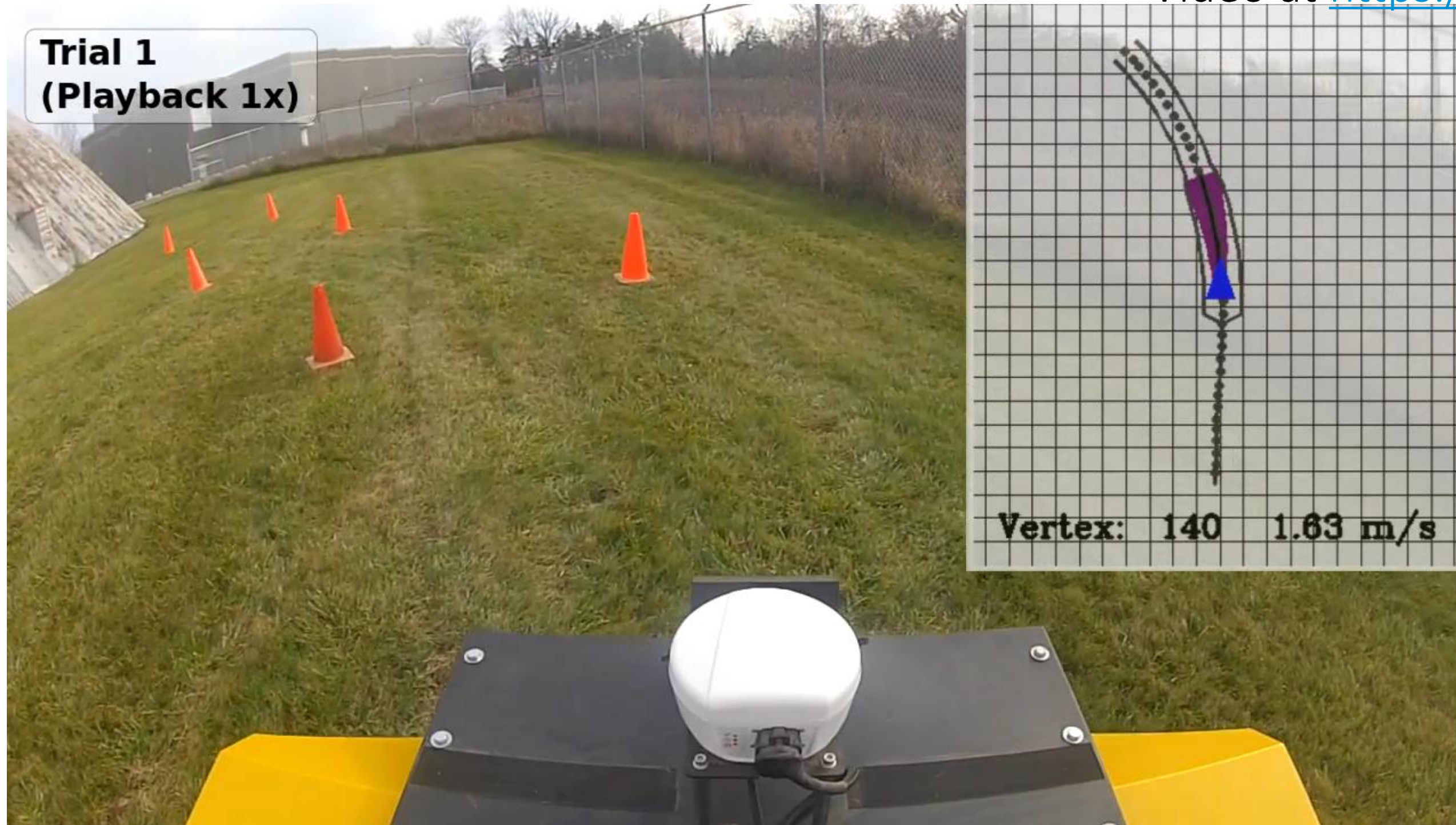
S. Kamthe, M.P. Deisenroth, AISTATS, 2018

Chance Constrained Model Predictive Control

A.T. Schwarm, M. Nikolaou, AIChE, 1999

Example

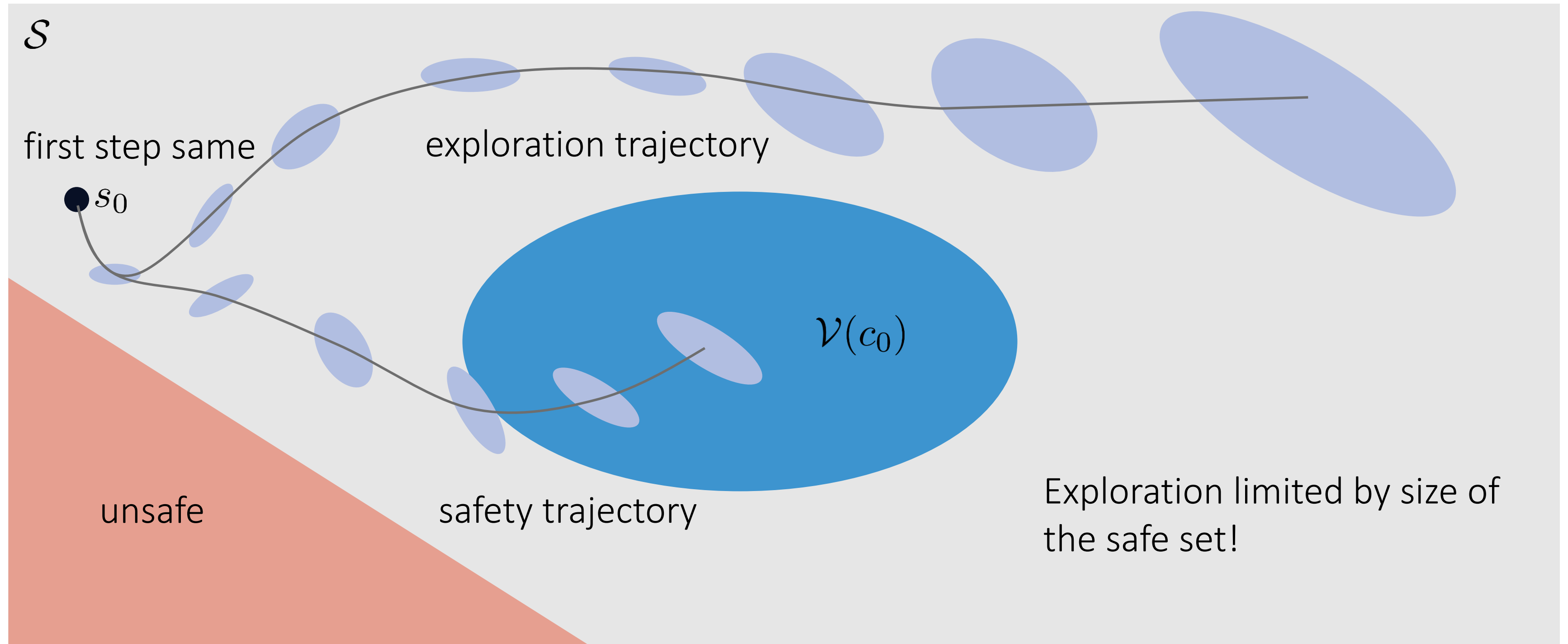
Video at <https://youtu.be/3xRNmNv5Efk>

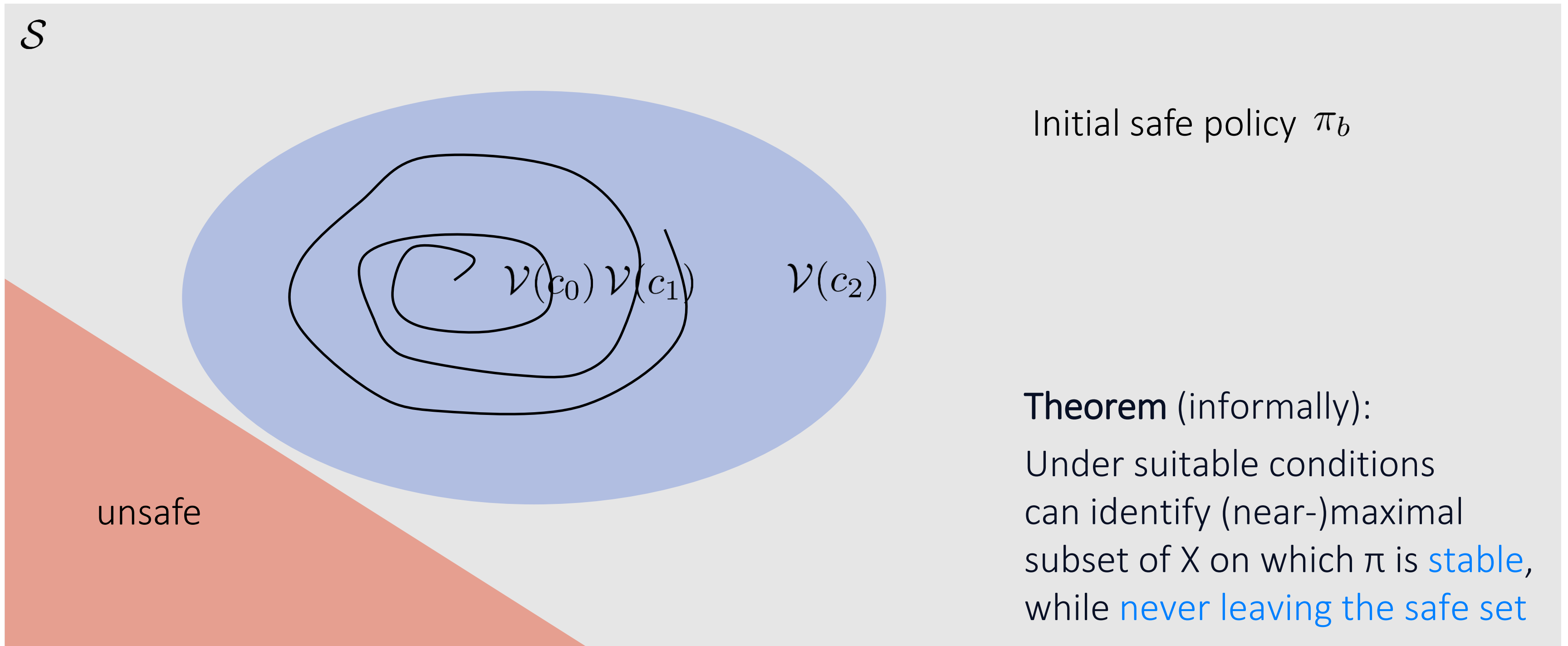


Robust constrained learning-based NMPC enabling reliable mobile robot path tracking

C.J. Ostafew, A.P. Schoellig, T.D. Barfoot, IJRR, 2016

Region of attraction





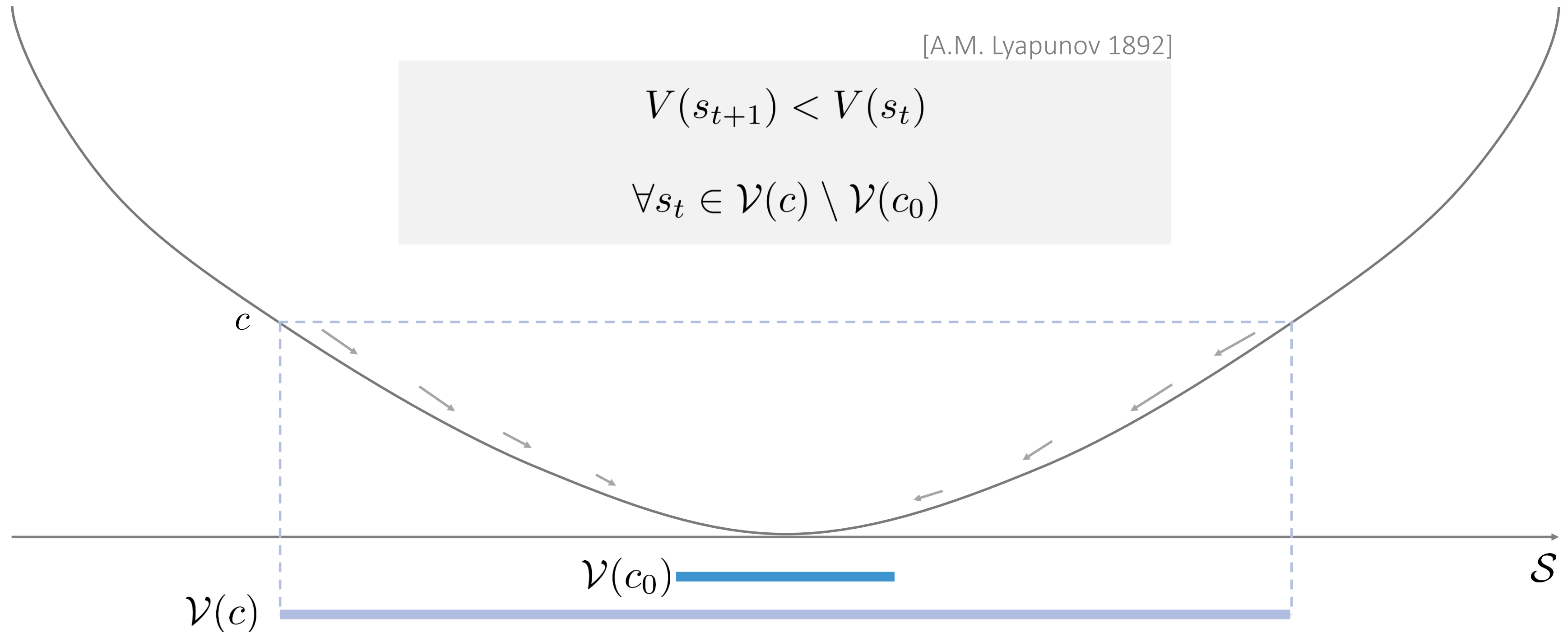
$$s_{t+1} = f(s_t, \pi(s, \theta))$$

$$V(s)$$

[A.M. Lyapunov 1892]

$$V(s_{t+1}) < V(s_t)$$

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



Lyapunov functions

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

$V(s)$

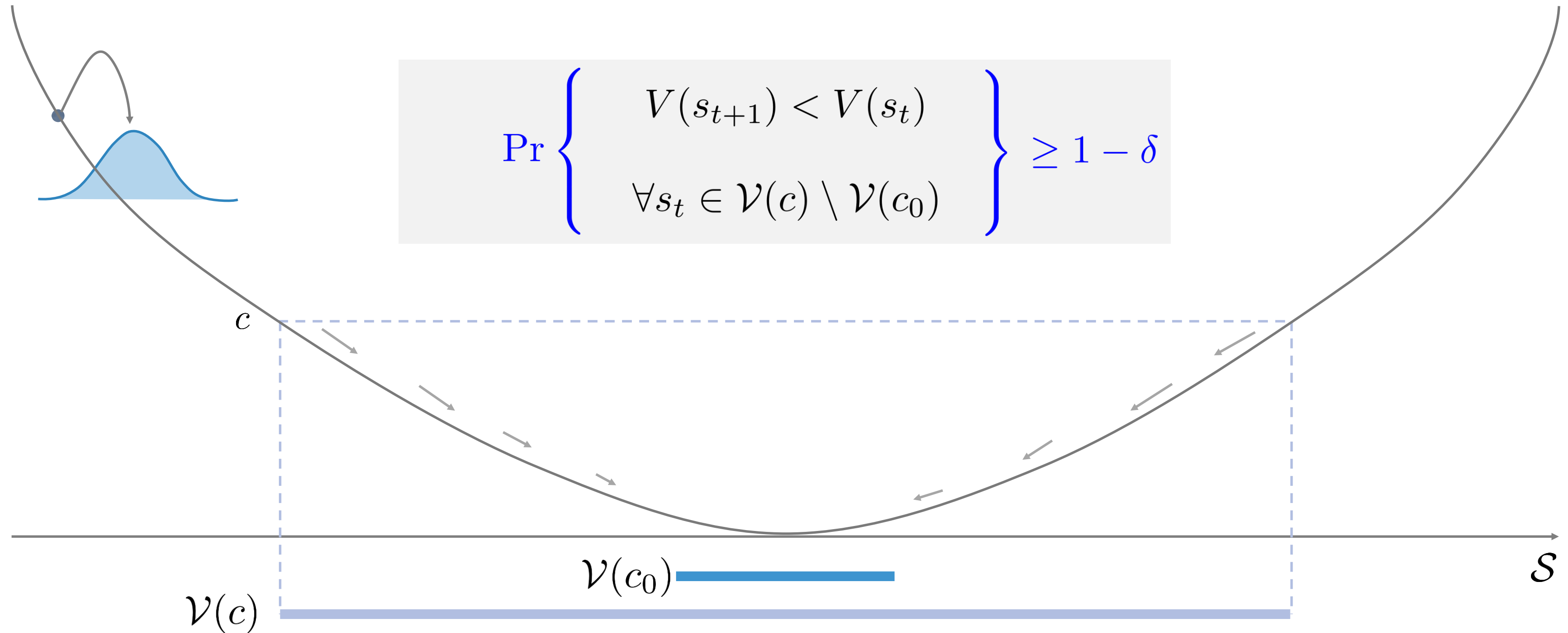
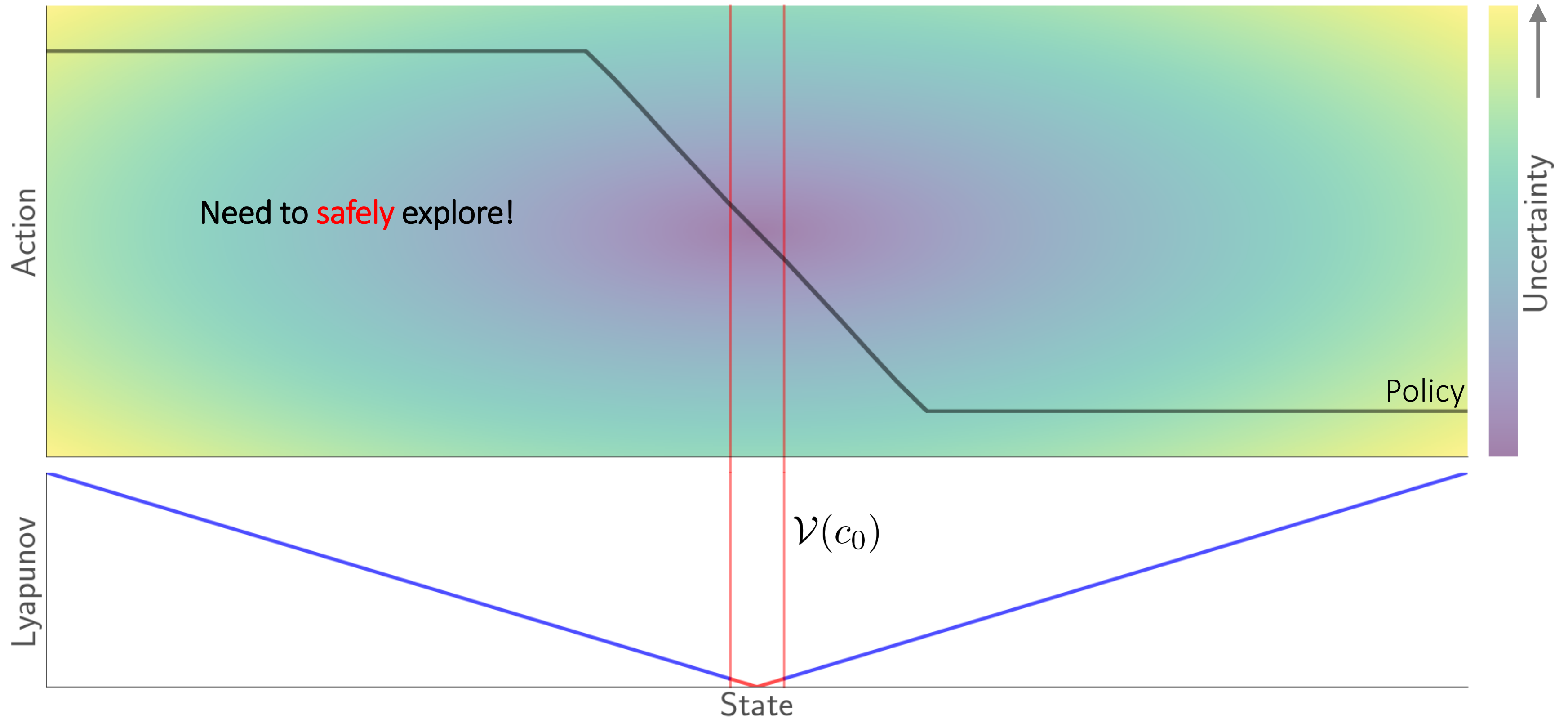


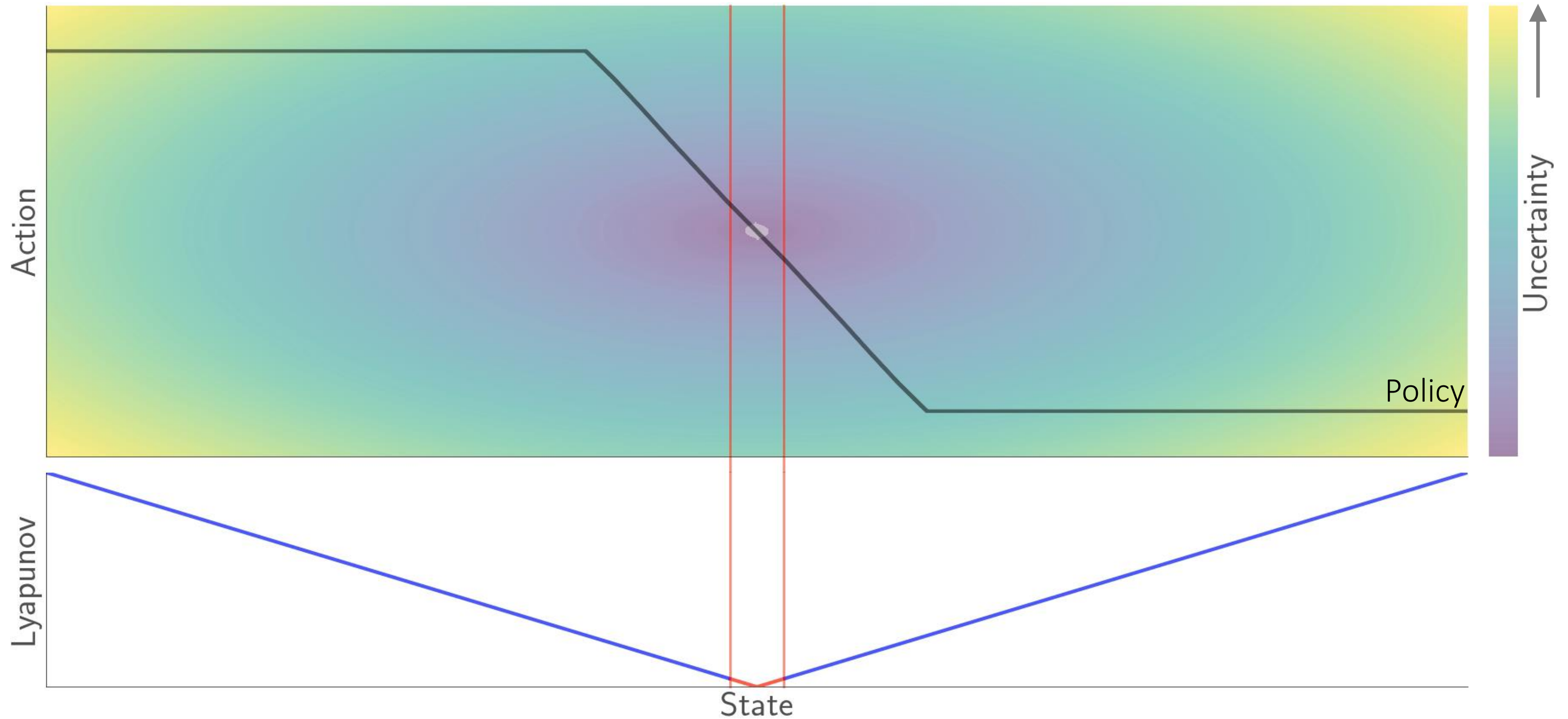
Illustration of safe learning



Safe Model-based Reinforcement Learning with Stability Guarantees

F. Berkenkamp, M. Turchetta, A.P. Schoellig, A. Krause, NIPS, 2017

Illustration of safe learning

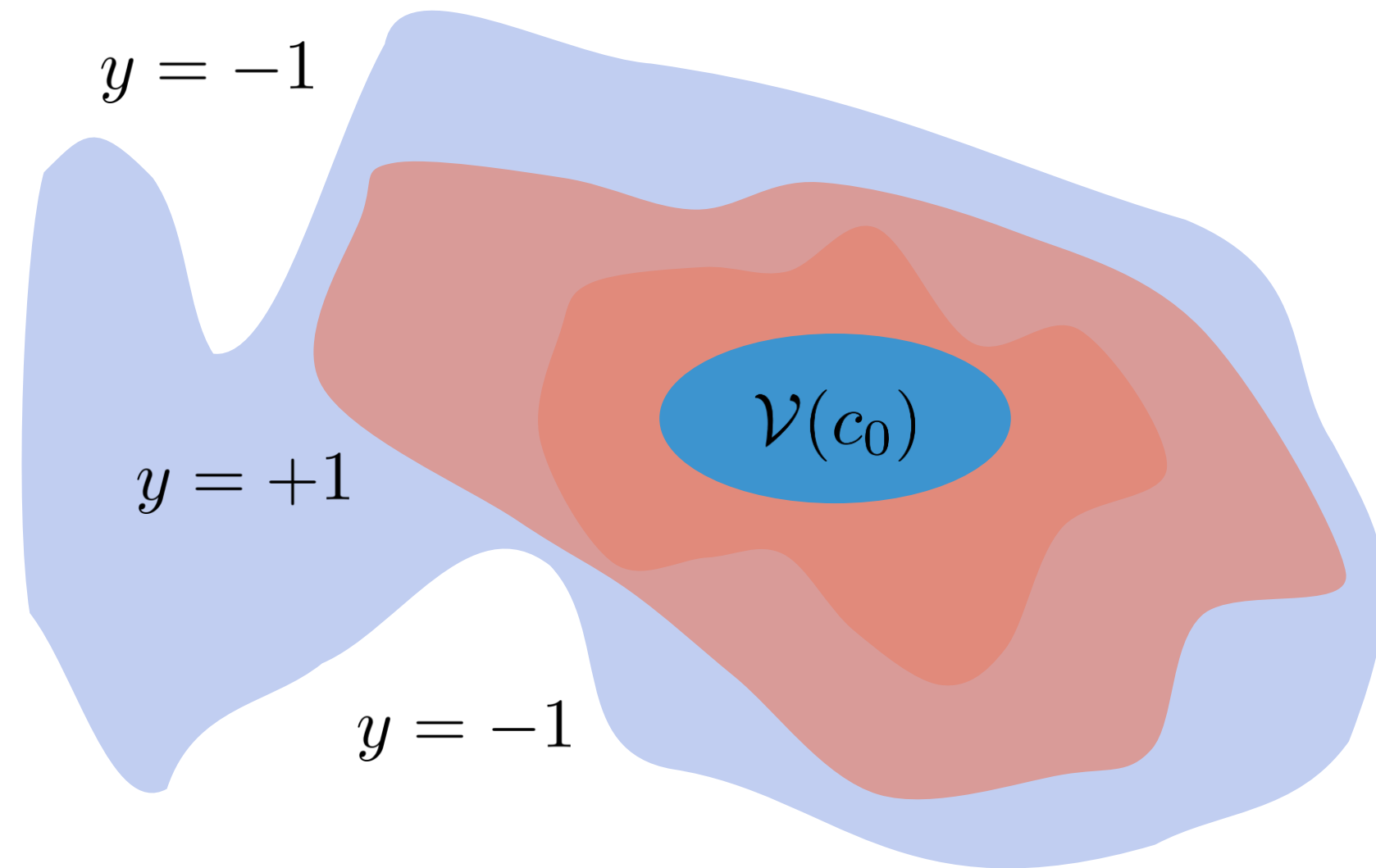


Safe Model-based Reinforcement Learning with Stability Guarantees

F. Berkenkamp, M. Turchetta, A.P. Schoellig, A. Krause, NIPS, 2017

Lyapunov function

Finding the right Lyapunov function is difficult!



$$V(s) = \phi_{\theta}(s)^T \phi_{\theta}(s)$$

Weights - positive-definite

Nonlinearities - trivial nullspace

Decision boundary $V(s) = 1$

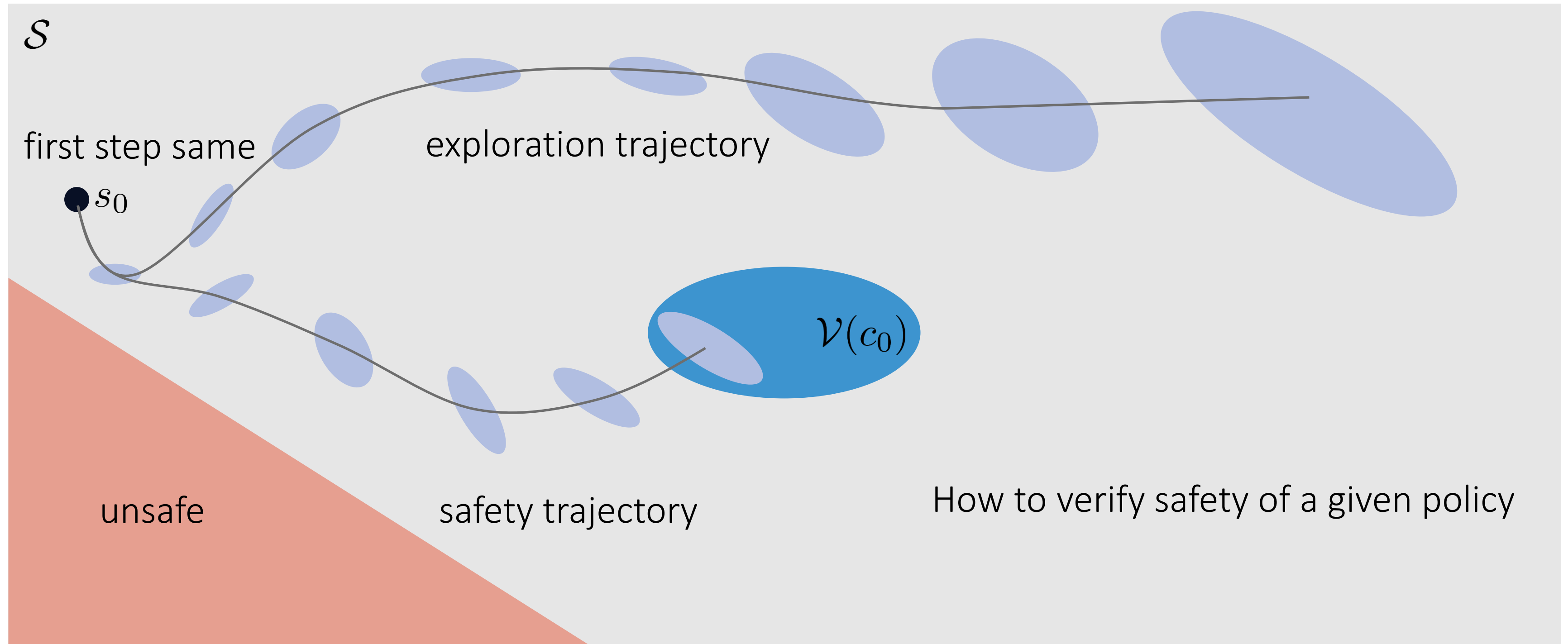
$$V(s_{t+1}) < V(s_t)$$

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

The Lyapunov Neural Network: Adaptive Stability Certification for Safe Learning of Dynamic Systems

S.M. Richards, F. Berkenkamp, A. Krause

Towards safe reinforcement learning



Summary

Reviewed safety definitions

Saw how to obtain a first, safe policy

Reviewed a first method for safe learning in expectation

Safe Bayesian optimization for safe exploration

How to transfer this intuition to the safe exploration in MDPs

Model-based methods (reachability=safety, certification, exploration)

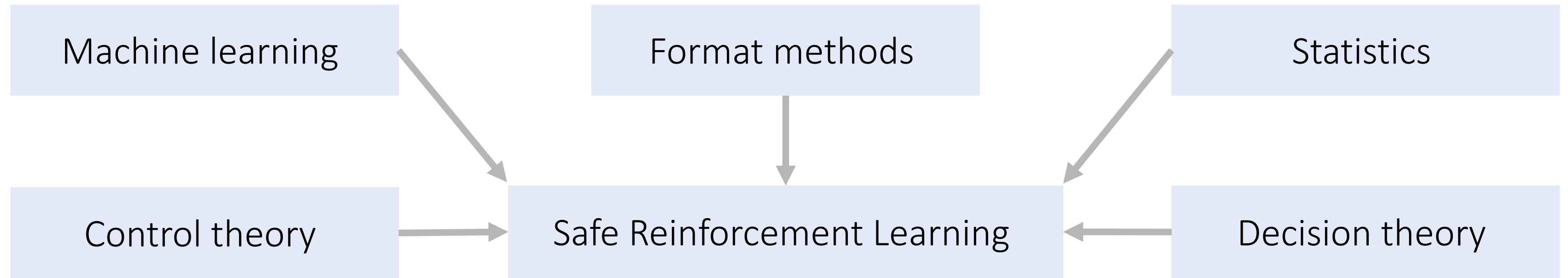
Stochastic

- Expected risk
- Moment penalized
- VaR / CVaR

Worst-case

- Formal verification
- Robust optimization

Where to go from here?



Scalability (computational & statistical)

Safe imitation learning

Tradeoff safety and performance (theory & practice)

Lower bounds; define function classes that are safely learnable