Poster #50

Fingerprint Policy Optimisation for Robust Reinforcement Learning

Supratik Paul, Michael A. Osborne, Shimon Whiteson

This project has received funding from the European Research Council (ERC) under the
European Union’s Horizon 2020 research and innovation programme (grant agreements \#637713)

Motivation

Motivation

Motivation

- Environment variable (EV)
  - E.g. wind conditions
  - Controllable during learning but not during execution

Motivation

- Environment variable (EV)
    - E.g. wind conditions
    - Controllable during learning but not during execution

- Objective: Find $\pi^* = argmax_\pi \, J(\pi) = argmax_\pi \mathbb{E}_{EV \sim p(EV)}[R(\pi)]$

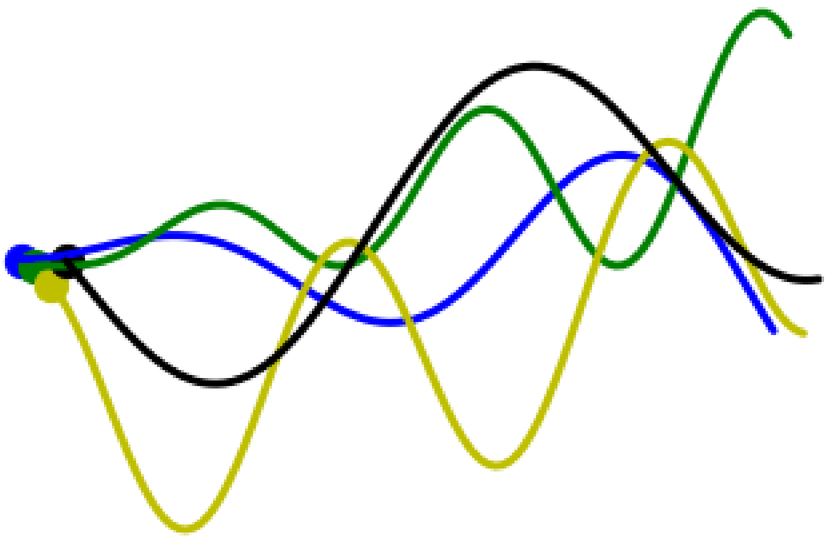
Motivation

- Environment variable (EV)
  - E.g. wind conditions
  - Controllable during learning but not during execution

- Objective: Find $\pi^* = argmax_\pi \, J(\pi) = argmax_\pi \mathbb{E}_{EV \sim p(EV)}[R(\pi)]$
- Need to account for rare events
  - E.g. rare wind conditions leading to a crash

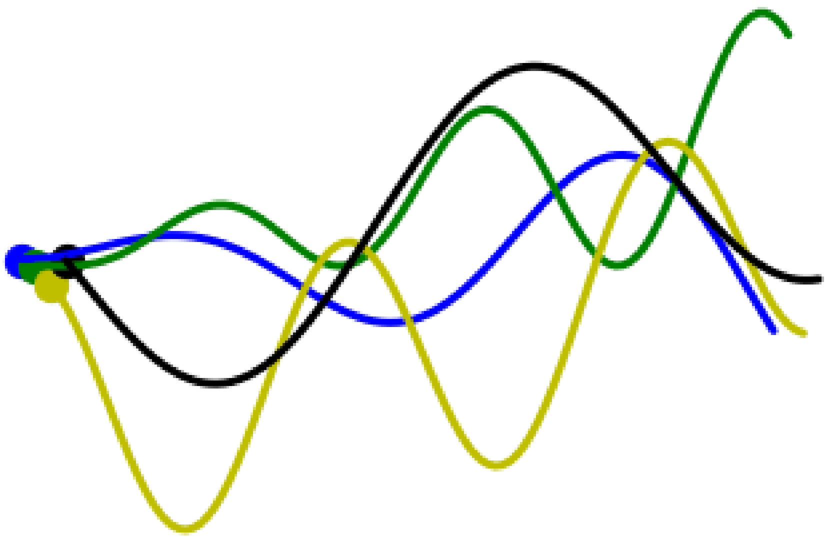
Naïve application of policy gradients

Naïve application of policy gradients

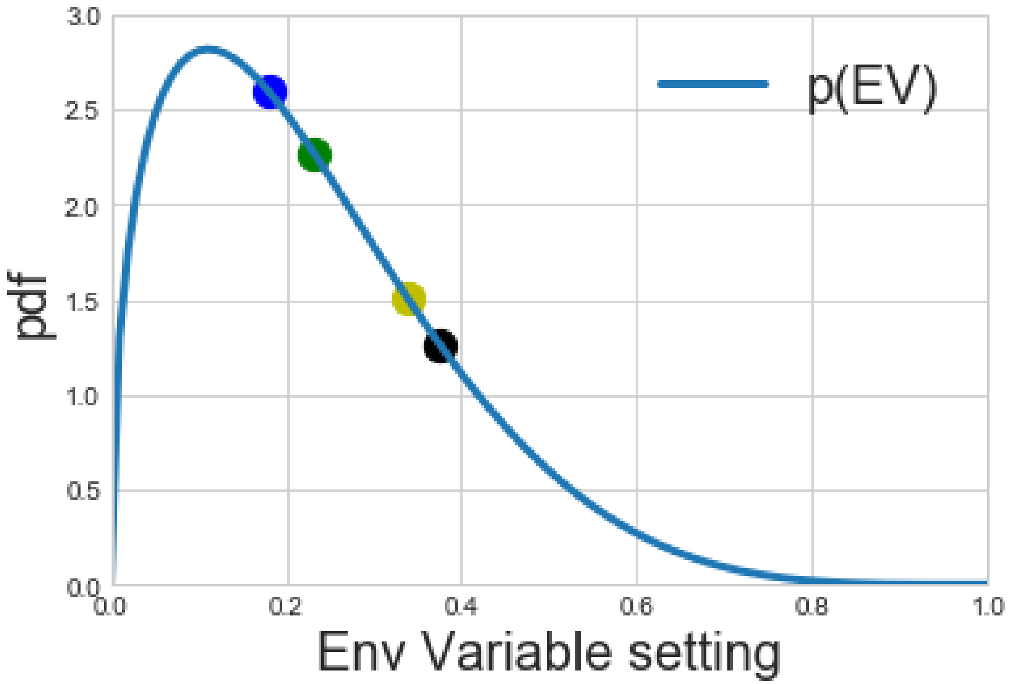


Trajectories $\sim \pi$

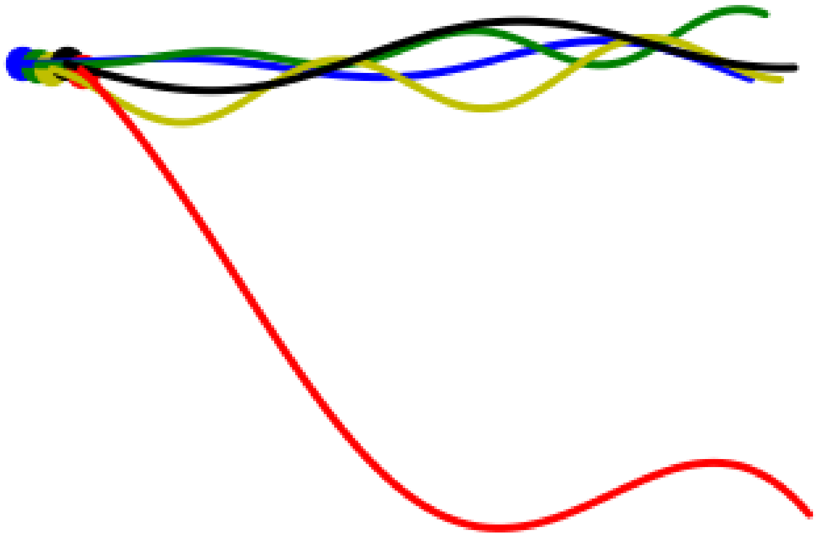
Naïve application of policy gradients



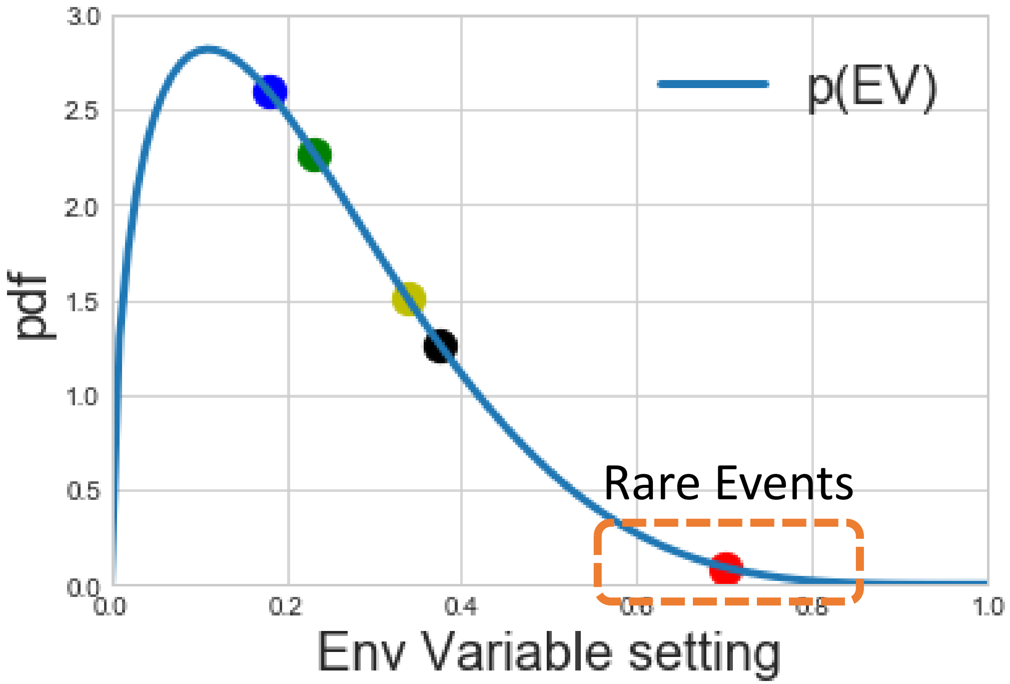
Trajectories $\sim \pi$



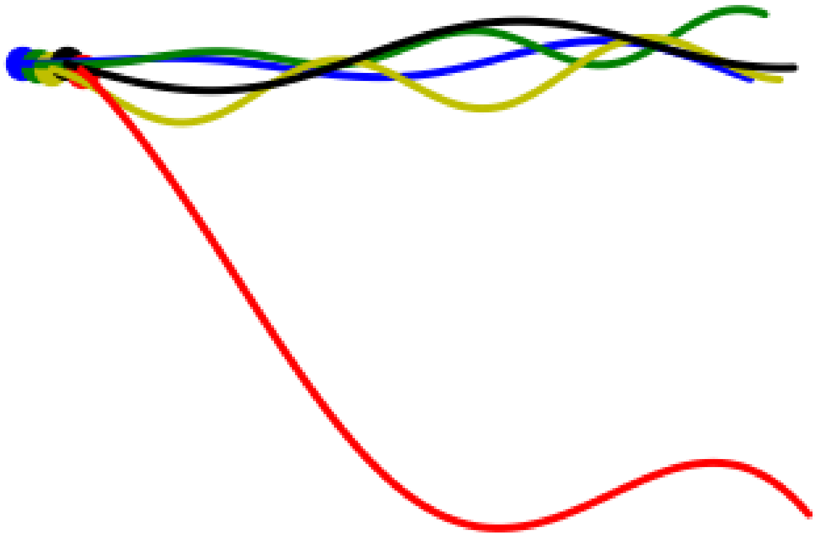
Naïve application of policy gradients



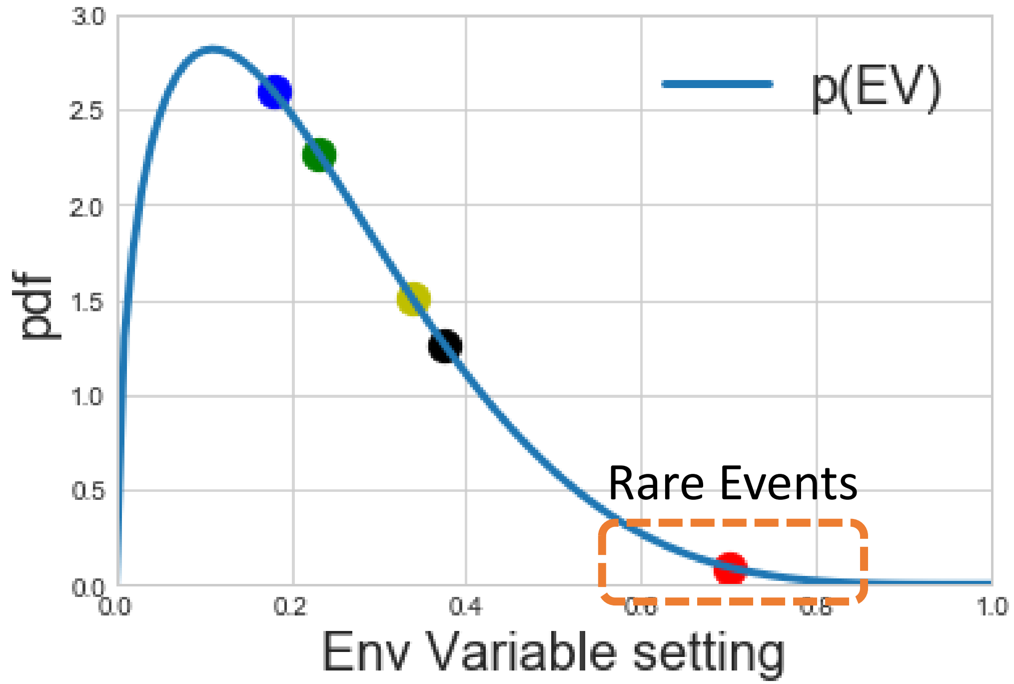
Trajectories $\sim \pi$



Naïve application of policy gradients



Trajectories $\sim \pi$

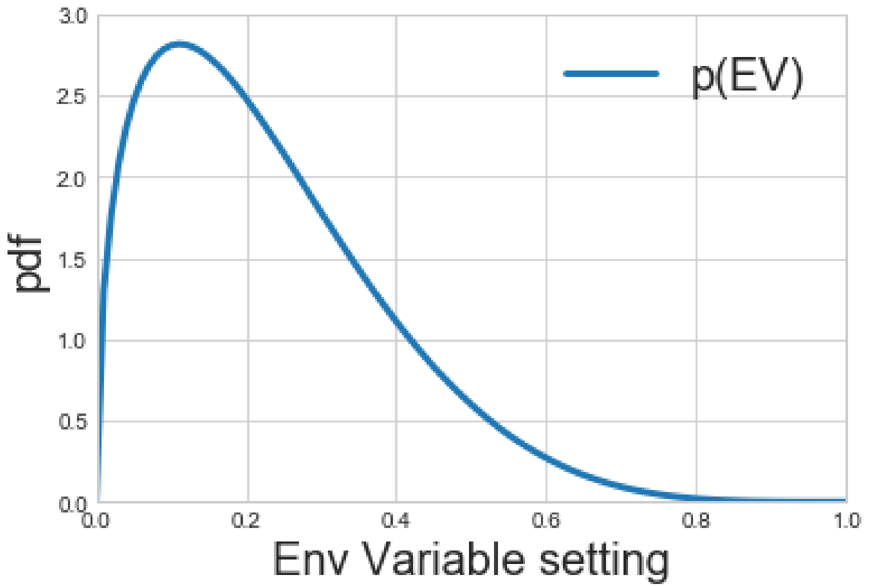


- Monte Carlo estimate of the Policy Gradient has very high variance

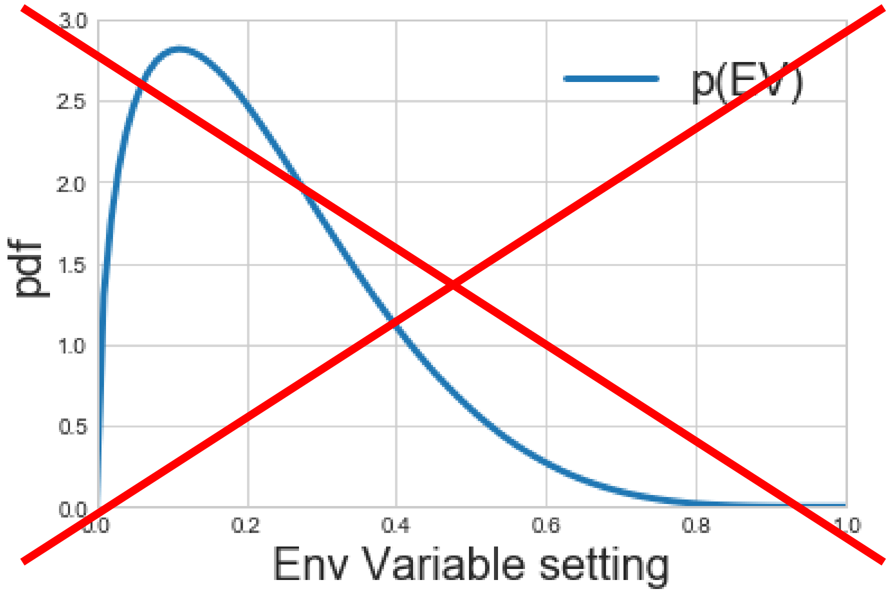
  $\Longrightarrow$ Doomed to failure

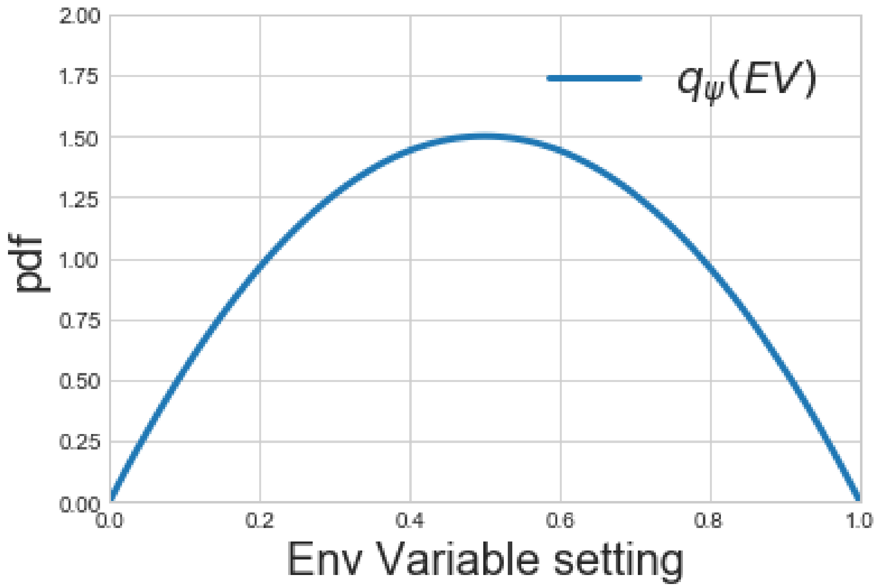
Fingerprint Policy Optimisation (FPO)

Fingerprint Policy Optimisation (FPO)

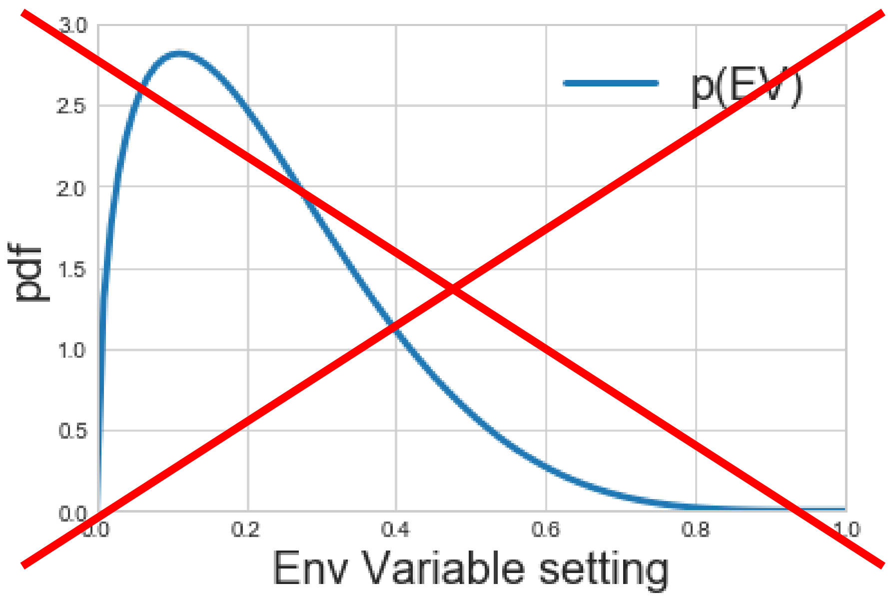


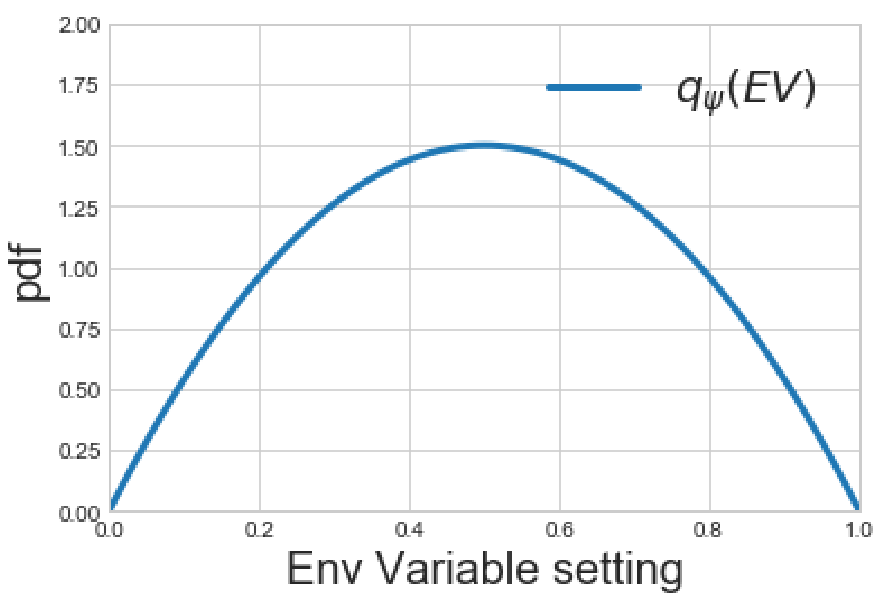
Fingerprint Policy Optimisation (FPO)





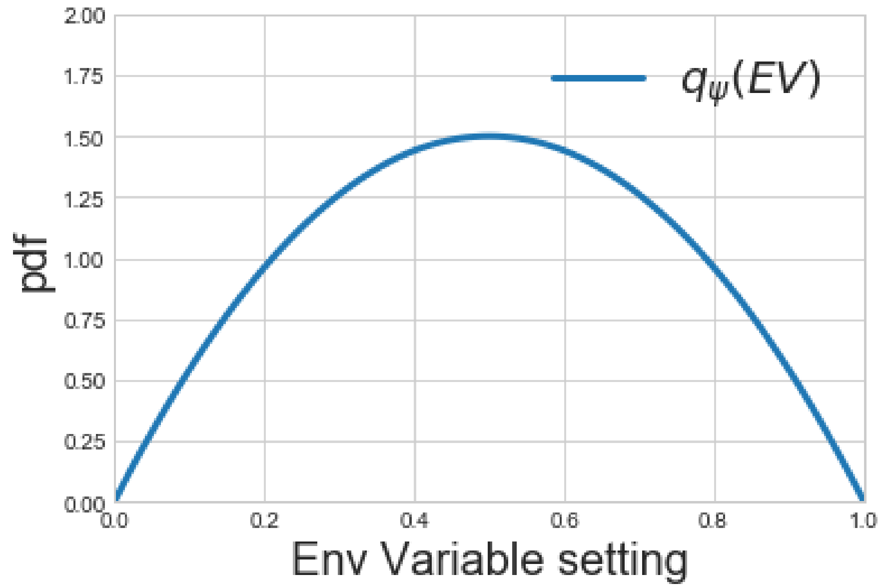
Fingerprint Policy Optimisation (FPO)





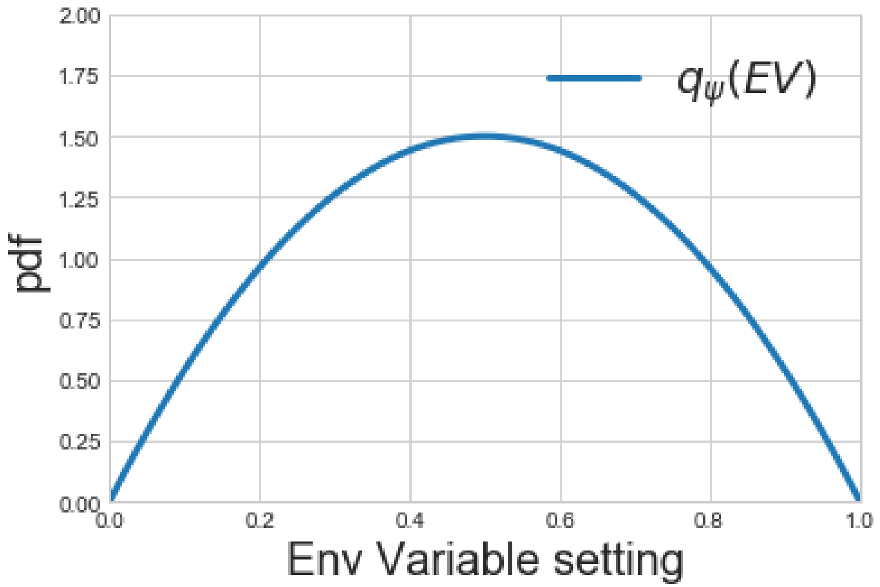
At each iteration, select parameters ψ of $q_\psi(EV)$ such that it maximises one-step expected return

Fingerprint Policy Optimisation (FPO)



Fingerprint Policy Optimisation (FPO)

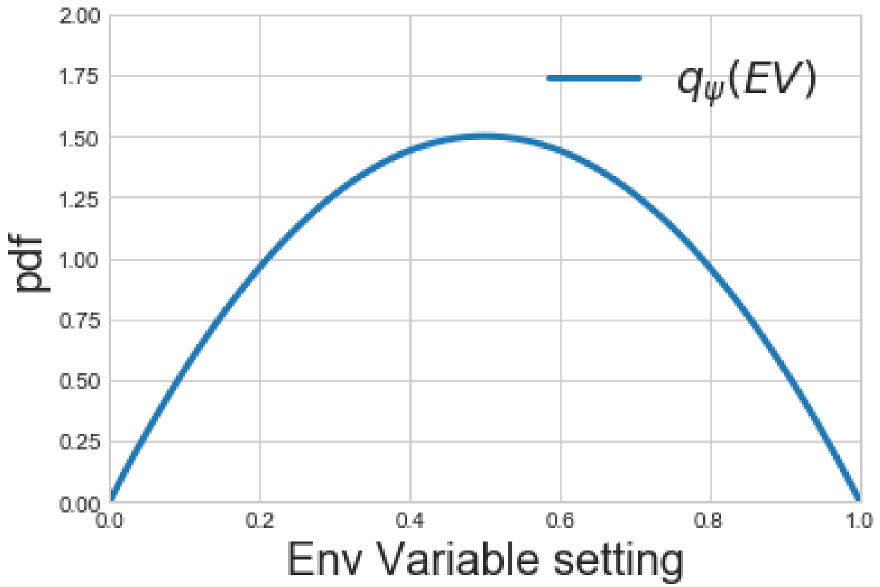
- $\pi' = \pi + \alpha \nabla J(\pi)$
- $J(\pi') = \mathrm{f}(\pi, \psi)$



Fingerprint Policy Optimisation (FPO)

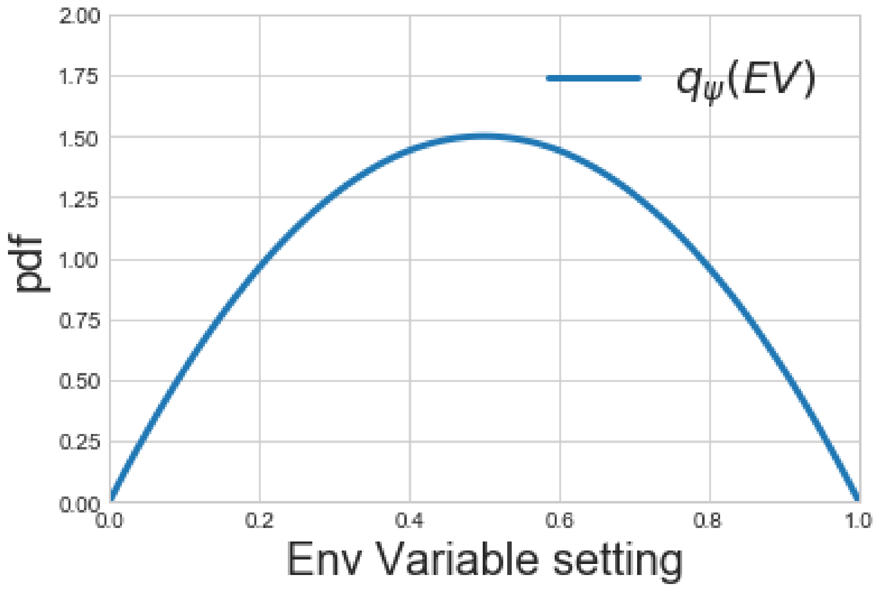
- $\pi' = \pi + \alpha \nabla J(\pi)$
- $J(\pi') = \mathrm{f}(\pi, \psi)$

- Model $J(\pi')$ as a Gaussian Process with inputs (π, ψ)
- Use Bayesian Optimisation to select $\psi | \pi = \mathrm{argmax}_{\psi} \mathrm{f}(\pi, \psi)$



Fingerprint Policy Optimisation (FPO)

- $\pi' = \pi + \alpha \nabla J(\pi)$
- $J(\pi') = \mathrm{f}(\pi, \psi)$



- Model $J(\pi')$ as a Gaussian Process with inputs (π, ψ)
- Use Bayesian Optimisation to select $\psi|\pi = \mathrm{argmax}_\psi \mathrm{f}(\pi, \psi)$

π is high dimensional → Low dimensional representation "Fingerprint"

Policy fingerprints

Policy fingerprints

- Disambiguation, not accurate representation

Policy fingerprints

- Disambiguation, not accurate representation

- State/Action fingerprints: Gaussians fitted to the stationary state/action distribution induced by π

Policy fingerprints

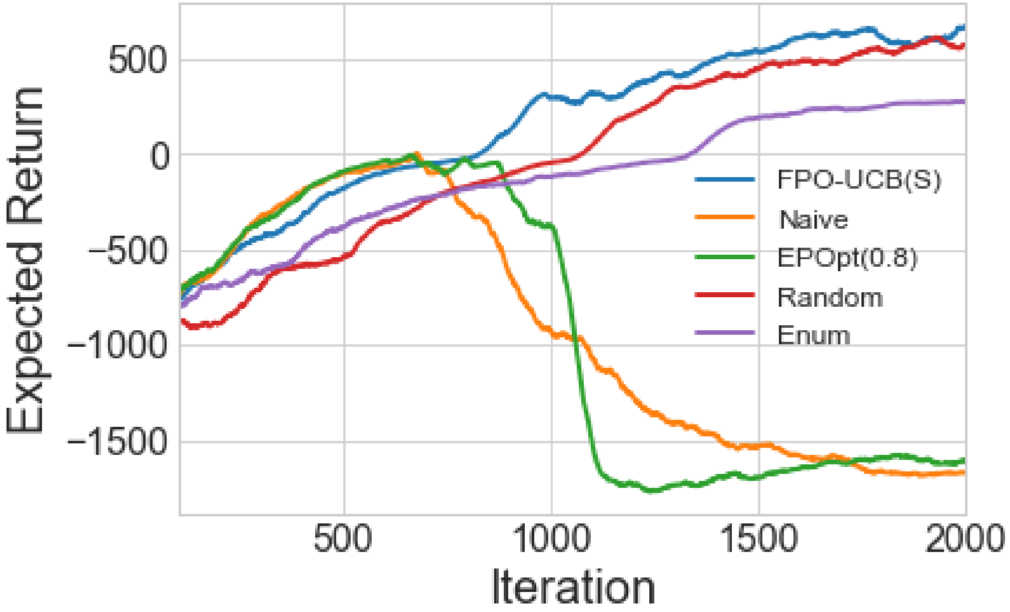
- Disambiguation, not accurate representation

- State/Action fingerprints: Gaussians fitted to the stationary state/action distribution induced by π

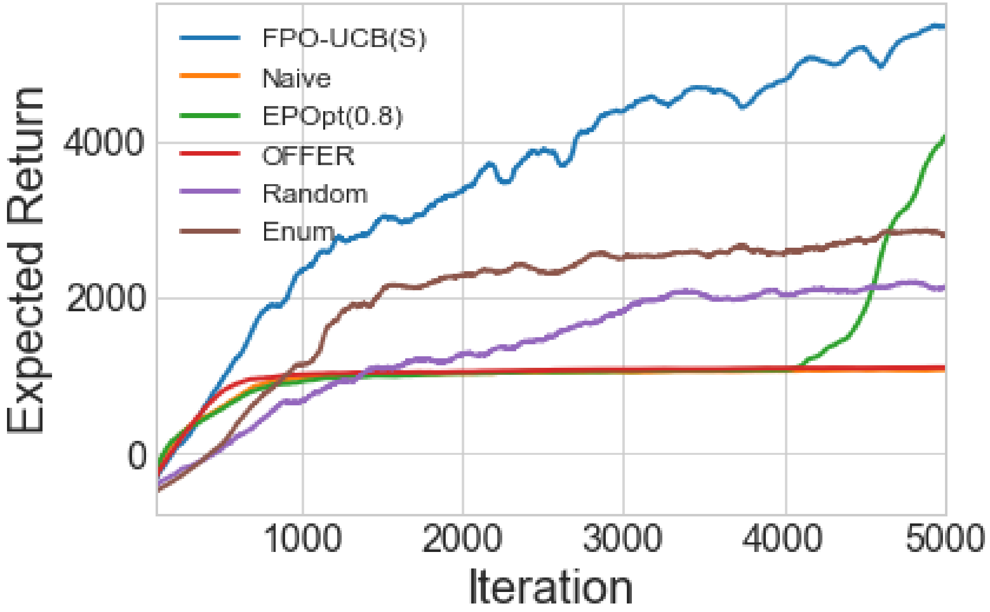
- Gross simplification, but good at disambiguating between policies

Results

Ant



Half Cheetah



- Reward proportional to velocity
- 5% chance that velocity > 2 leads to joint damage with large negative reward

- Velocity target = 2 with probability 98% and 'normal' reward
- Velocity target = 4 with probability 2% with significantly high reward

Poster #50

Fingerprint Policy Optimisation for Robust Reinforcement Learning

Supratik Paul, Michael A. Osborne, Shimon Whiteson

This project has received funding from the European Research Council (ERC) under the
European Union’s Horizon 2020 research and innovation programme (grant agreements \#637713)