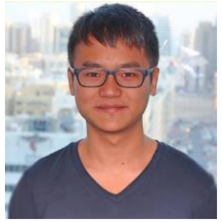
Microsoft

Greed is Bad or
Better Exploration with Optimistic Actor Critic

Kamil Ciosek
Microsoft Research Cambridge


Kamil Ciosek
MSR Cambridge


Quan Vuong
PhD student, UCSD


Robert Loftin
MSR Cambridge

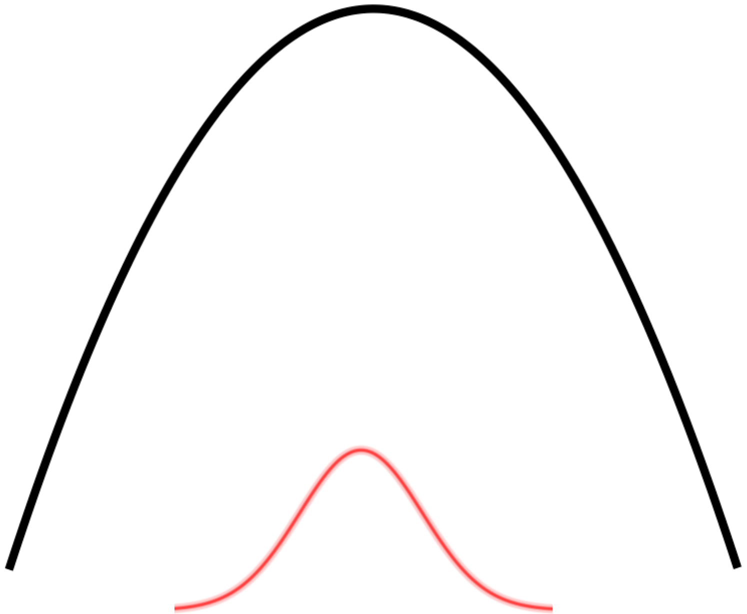

Katja Hofmann
MSR Cambridge

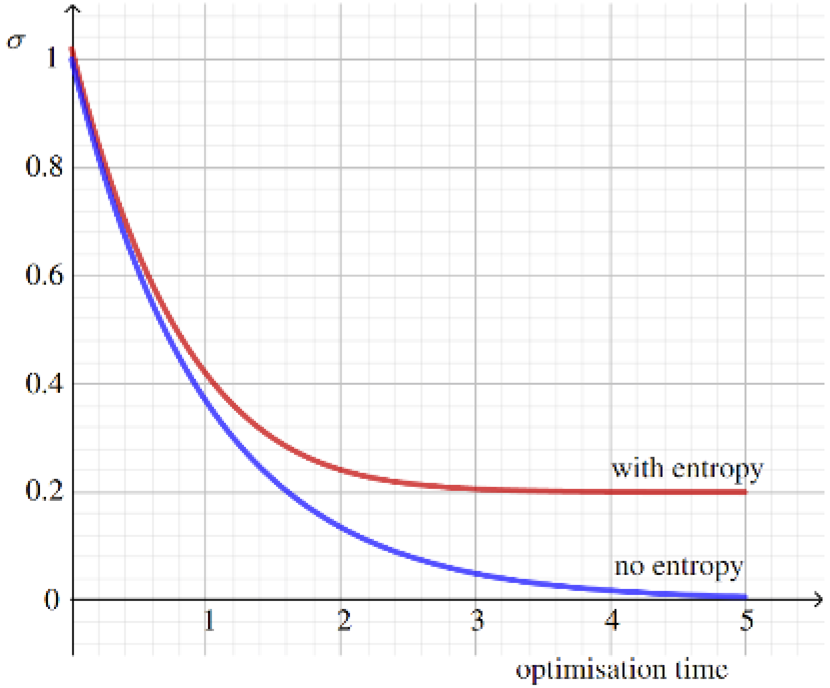
Policy Gradients are greedy

Policy gradients are *greedy*

Maximise $\nabla_\theta J$

What happens to the policy standard deviation?

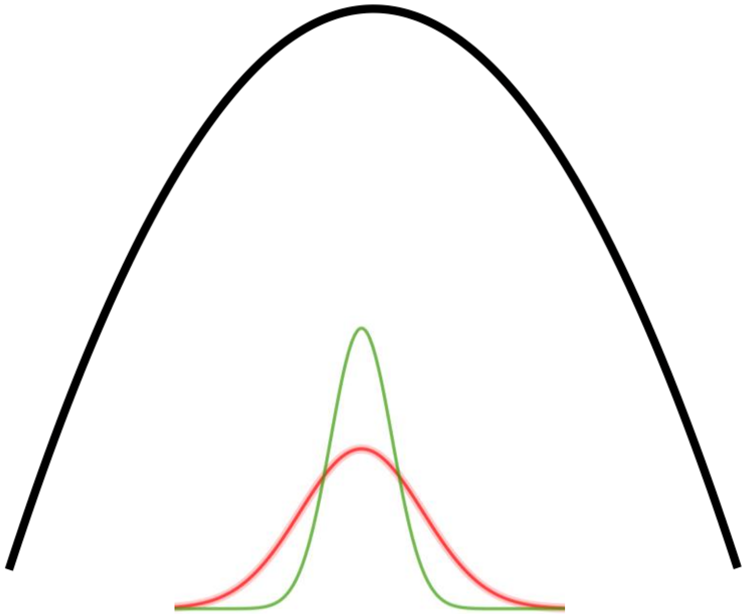




Consider a bandit with quadratic reward.

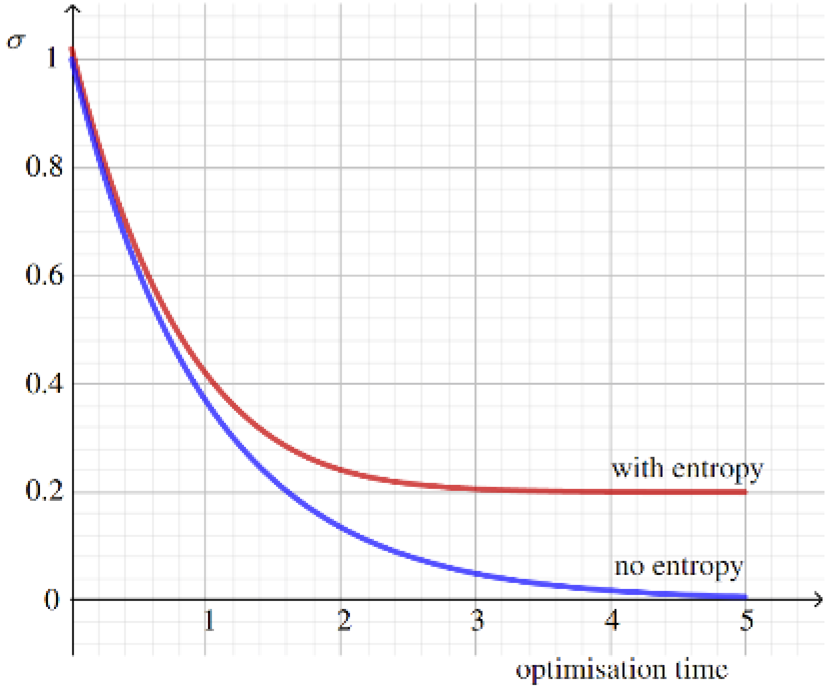
Policy gradients are *greedy*

Maximise $\nabla_\theta J$



Consider a bandit with quadratic reward.

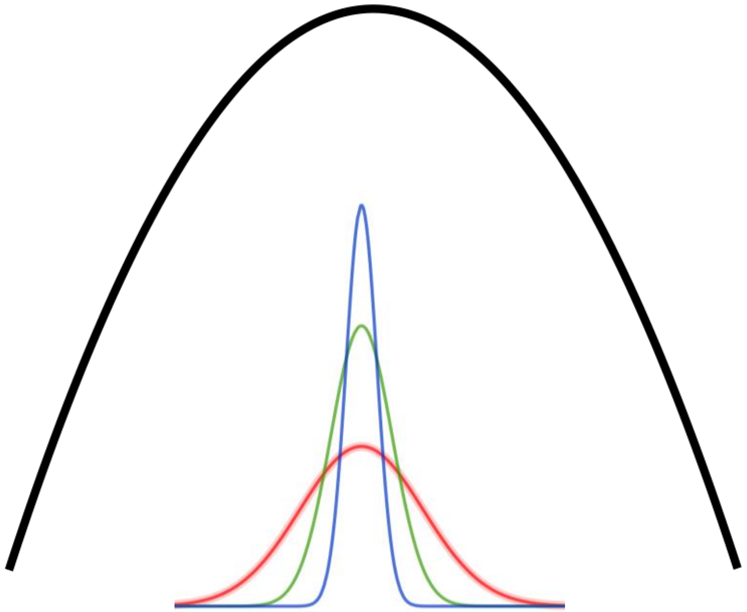
What happens to the policy standard deviation?

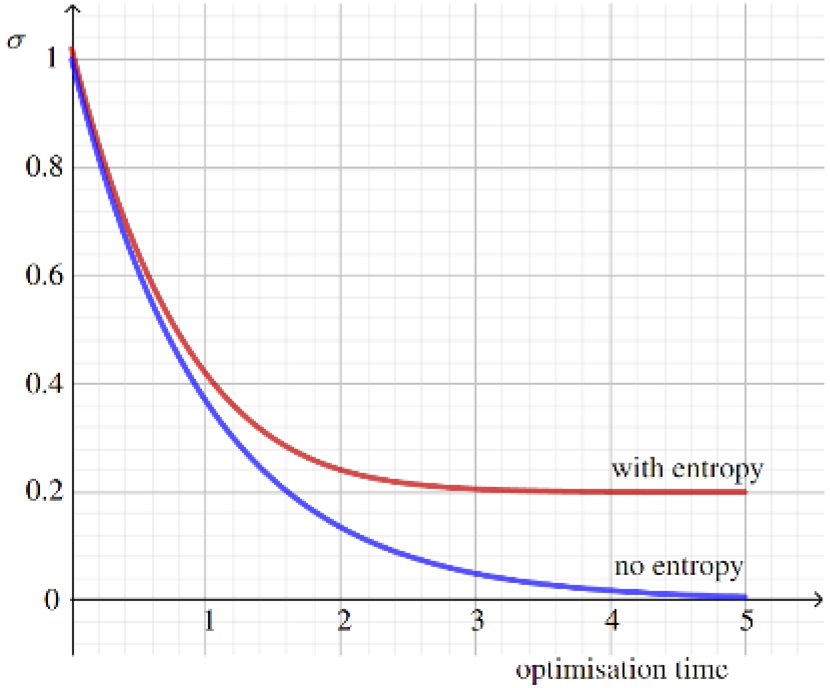


Policy gradients are *greedy*

Maximise $\nabla_\theta J$

What happens to the policy standard deviation?





Consider a bandit with quadratic reward.

Modern Policy Gradient Methods use a Lower bound

Lower bound on critic

if this is too large…

$$\hat{Q}(s_t, a_t) \leftarrow R(s_t, a_t) + \gamma \breve{Q}(s_{t+1}, a) \quad a \sim \pi_T(\cdot|s_{t+1})$$

…this becomes too large

+ effect amplified by policy optimisation

Lower bound on critic

if this is too large...

$$\hat{Q}(s_t, a_t) \leftarrow R(s_t, a_t) + \gamma \breve{Q}(s_{t+1}, a) \quad a \sim \pi_T(\cdot|s_{t+1})$$

...this becomes too large

+ effect amplified by policy optimisation

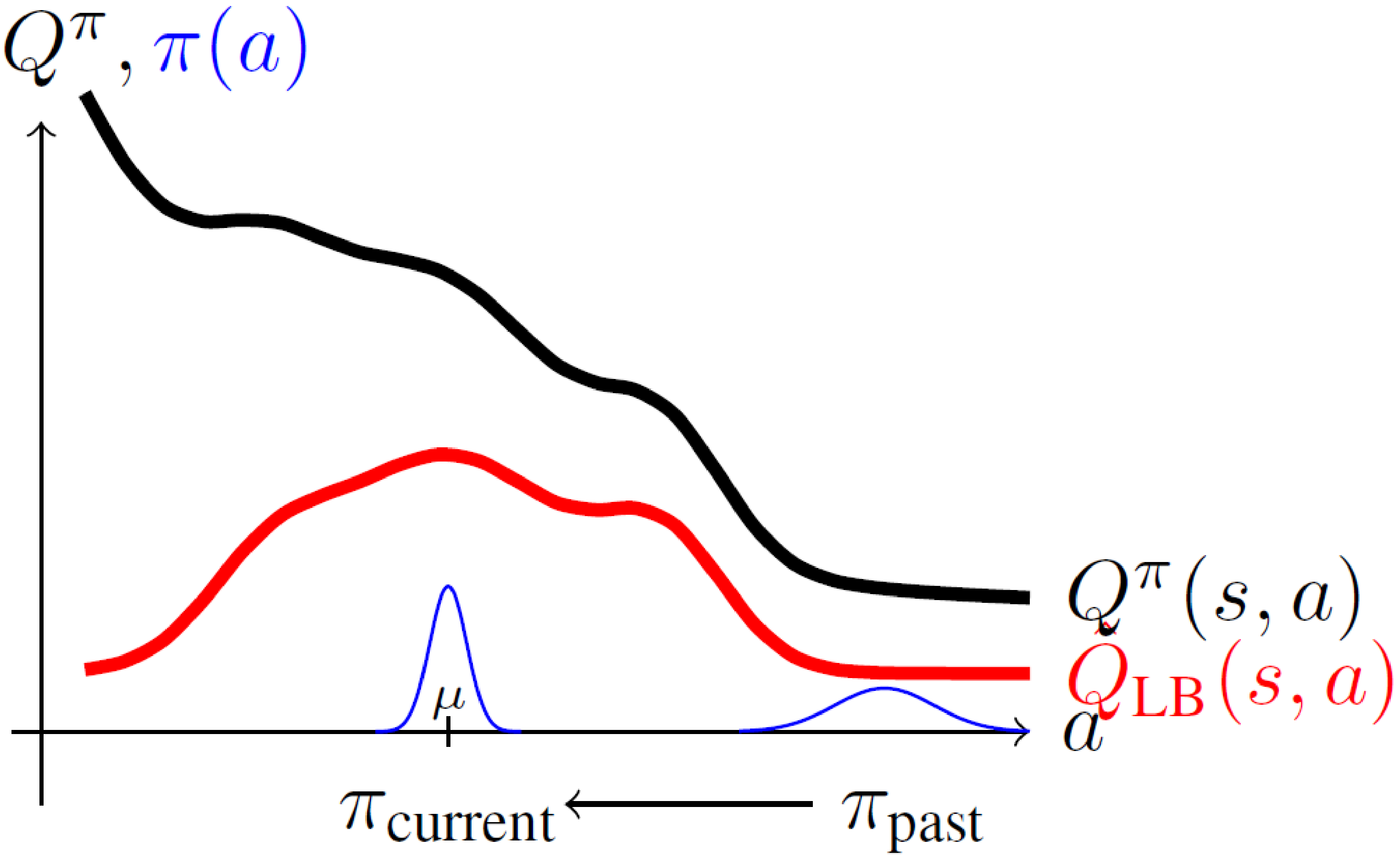
$$\hat{Q}^{\{1,2\}}_{\text{LB}}(s_t, a_t) \leftarrow R(s_t, a_t) + \gamma \min(\breve{Q}^1_{\text{LB}}(s_{t+1}, a), \breve{Q}^2_{\text{LB}}(s_{t+1}, a))$$

conservative update reduces overestimation

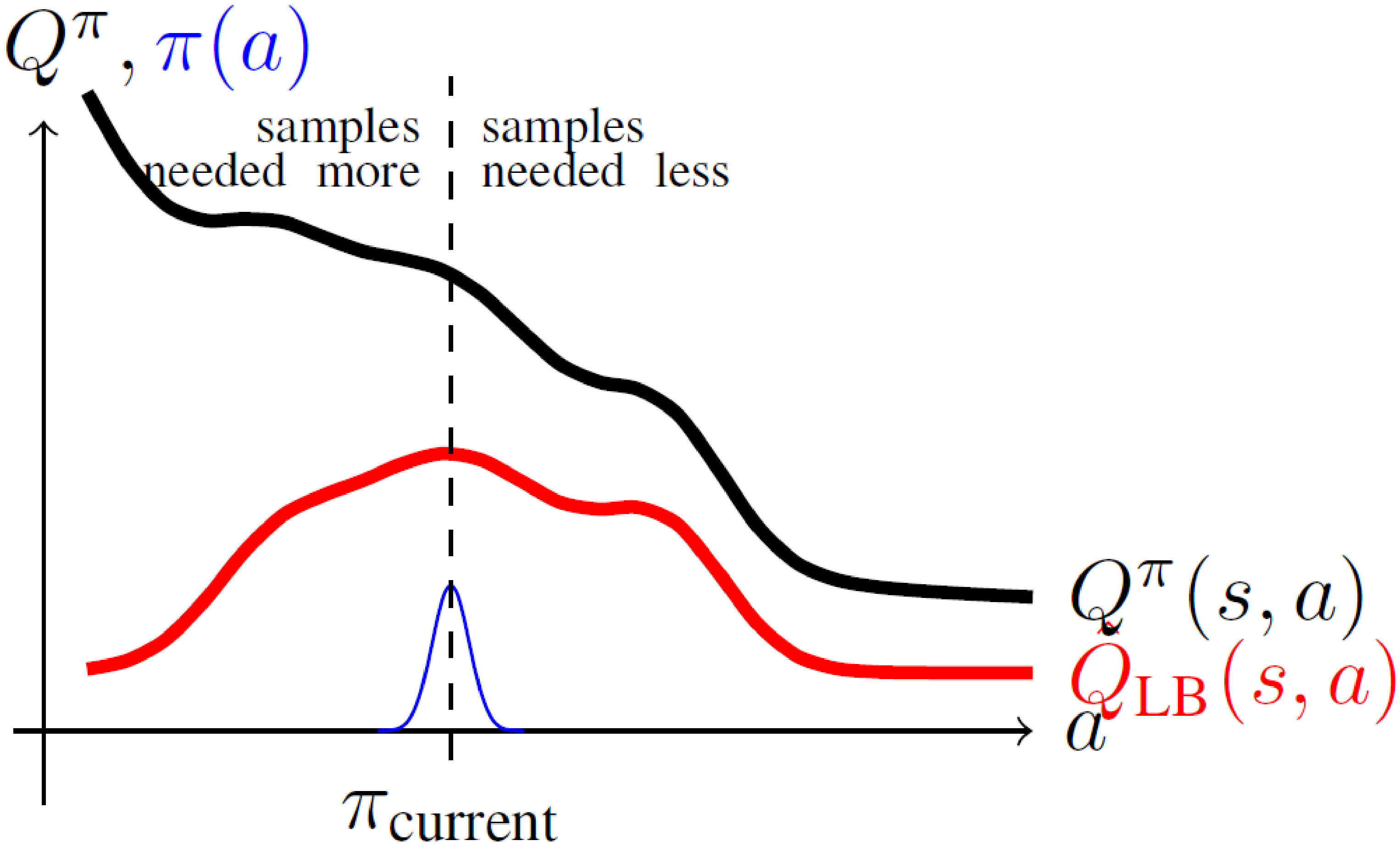
van Hasselt (2010, 2016), Fujimoto et al (2018)

Greediness + Lower Bound Lead To Problems

First problem: pessimistic underexploration



Second problem: directional uninformedness



Solve these Problems by Exploring with Upper Bound

Use the bootstrap to make an upper bound.

$$\mu_Q(s,a) = \frac{1}{2}\left(\hat{Q}^1_{\mathrm{LB}}(s,a) + \hat{Q}^2_{\mathrm{LB}}(s,a)\right)$$

$$\sigma_Q(s,a) = \sqrt{\sum_{i\in\{1,2\}} \frac{1}{2}\left(\hat{Q}^i_{\mathrm{LB}}(s,a) - \mu_Q(s,a)\right)^2}$$

level of optimism

$$\hat{Q}_{\mathrm{UB}}(s,a) = \mu_Q(s,a) + \beta_{\mathrm{UB}}\sigma_Q(s,a)$$

How to choose the exploration policy

We want a policy that:

- Is close to target policy.
- Maximises the critic upper bound.

$$\mu_e, \Sigma_E = \underset{\substack{\mu,\Sigma: \\ \mathrm{KL}(\mathcal{N}(\mu,\Sigma), \mathcal{N}(\mu_T, \Sigma_T)) \leq \delta}}{\arg\max} E_{a\sim\mathcal{N}(\mu,\Sigma)} \left[\hat{Q}_{\mathrm{UB}}(s, a)\right]$$

How to choose the exploration policy

We want a policy that:

- Is close to target policy.
- Maximises the critic upper bound.

$$\mu_e, \Sigma_E = \underset{\substack{\mu, \Sigma: \\ \mathrm{KL}(\mathcal{N}(\mu,\Sigma), \mathcal{N}(\mu_T, \Sigma_T)) \leq \delta}}{\arg\max} E_{a \sim \mathcal{N}(\mu,\Sigma)} \left[\bar{Q}_{\mathrm{UB}}(s, a) \right]$$

Linearize!

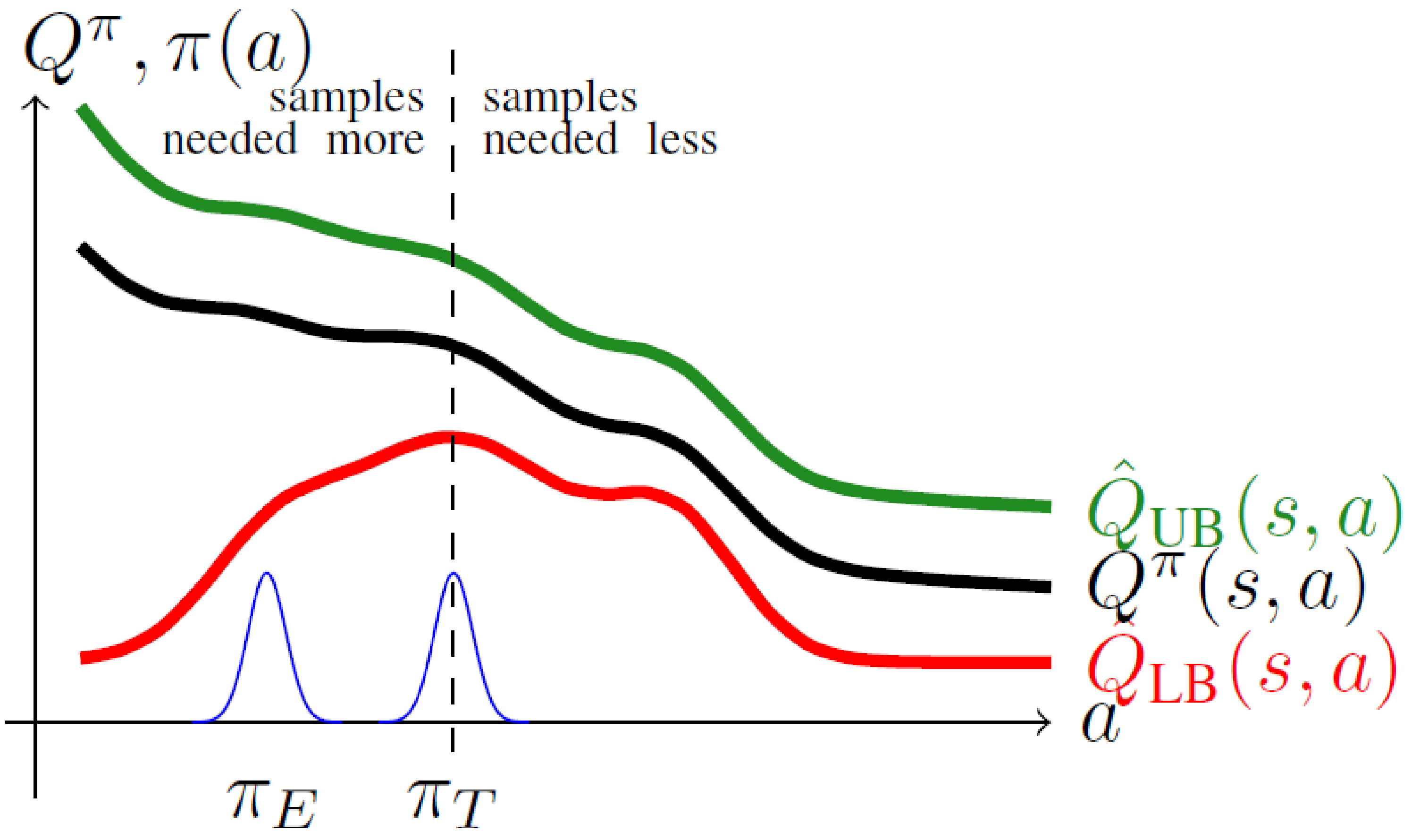
The OAC exploration policy (interpretation)

$$\pi_E = \mathcal{N}(\mu_E, \Sigma_E), \quad \mu_E = \mu_T + \underbrace{\frac{\sqrt{2\delta}}{\left\| \left[\nabla_a \hat{Q}_{UB}(s,a)\right]_{a=\mu_T} \right\|_{\Sigma_T}} \Sigma_T \left[\nabla_a \hat{Q}_{UB}(s,a)\right]_{a=\mu_T}}_{\text{shift}} \quad and \quad \Sigma_E = \Sigma_T.$$

OAC explores with a shifted policy!

shift in the direction given by upper bound.

OAC explores efficiently

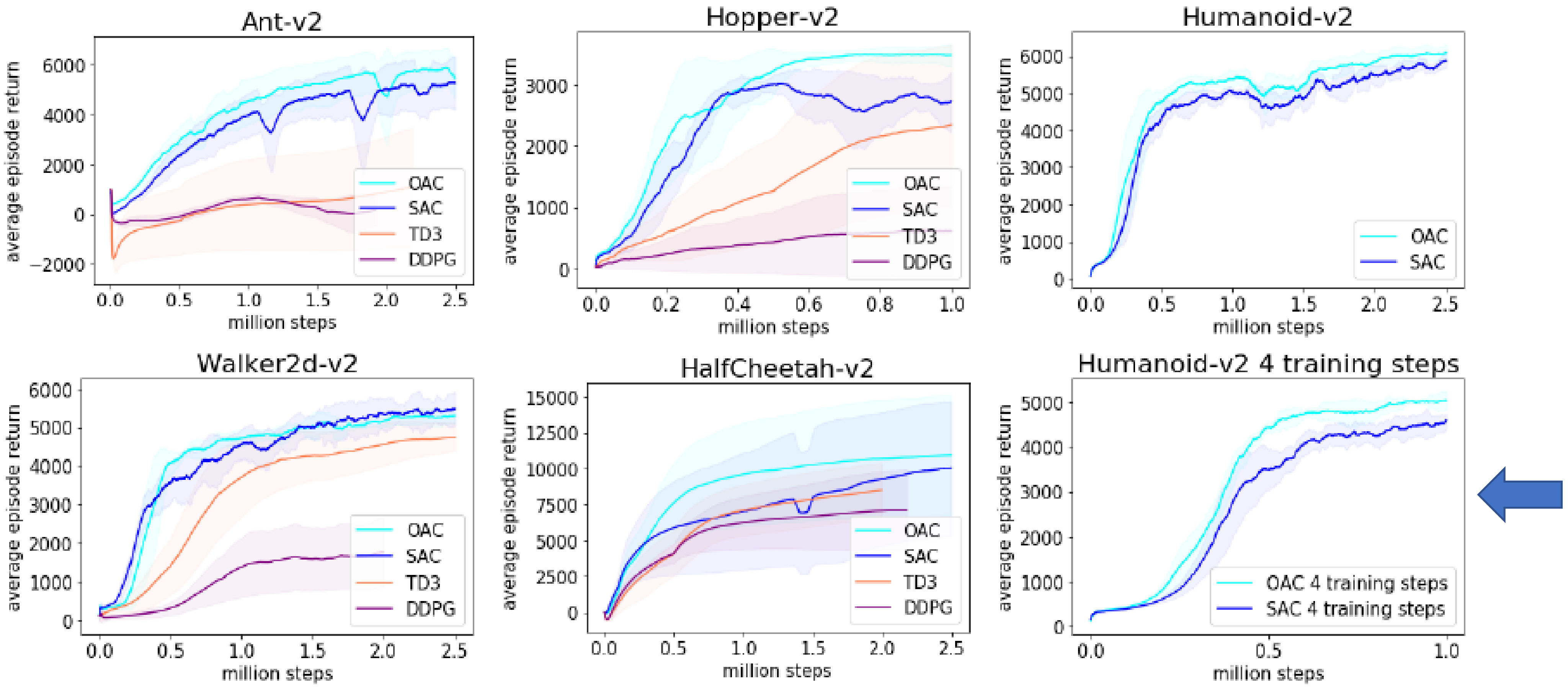


OAC avoids spurious maximum

OAC is directionally informed

It works!

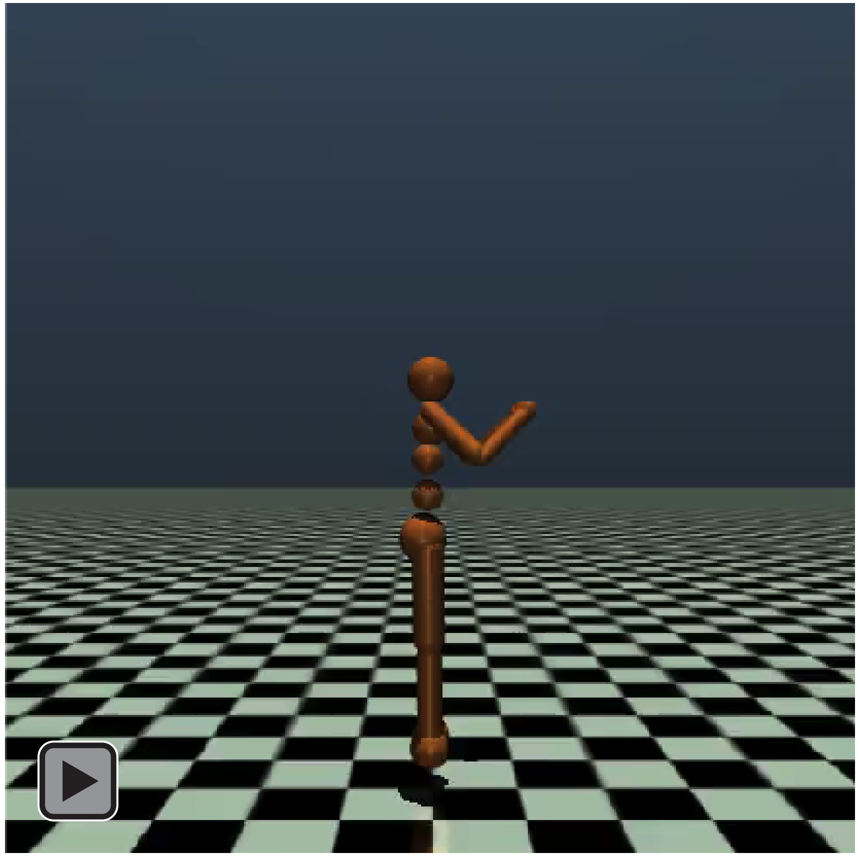
It works!



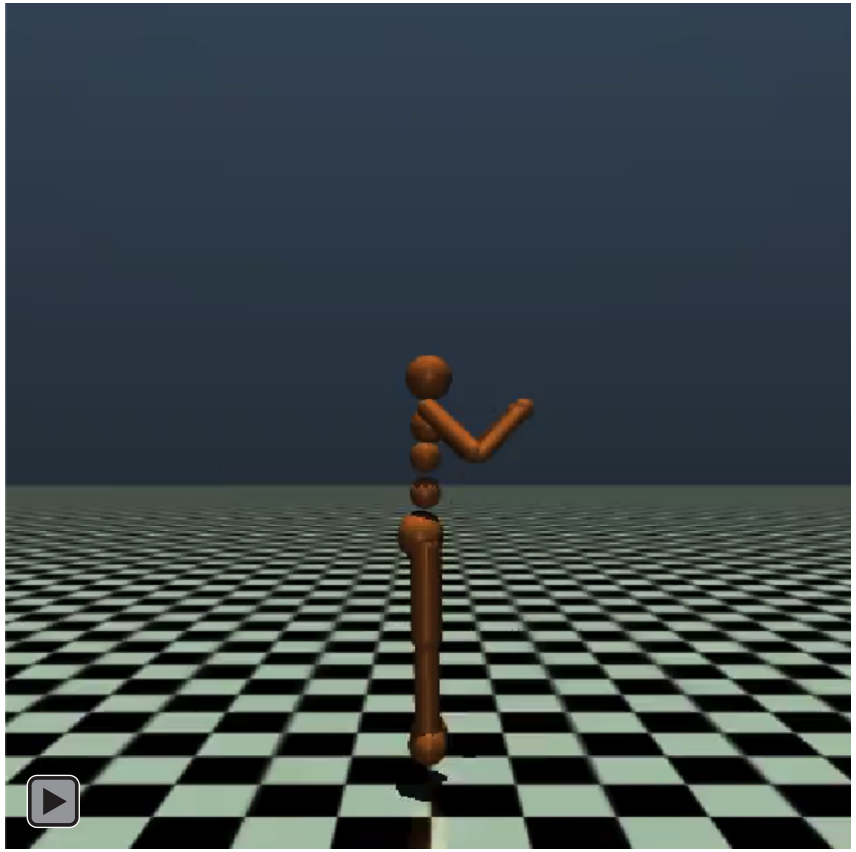
No hyperparameters were tuned on Humanoid!

Visual Comparison

SAC (previous state of the art)



OAC (our approach)



Microsoft

We have openings for interns, post-docs, researchers.

Kamil.Ciosek@Microsoft.com

Talk to me at the poster session!

Poster #179, starting at 5:30PM, East Exhibition Hall B+C