



**Whiteson
Research
Lab**



UNIVERSITY OF
OXFORD

VIREL: A Variational Inference Framework for Reinforcement Learning



Mattie Fellows



Anuj Mahajan



Tim Rudner



Shimon Whiteson

Reinforcement Learning as Inference

AIM: To cast the reinforcement learning problem into one of probabilistic inference

Reinforcement Learning as Inference

AIM: To cast the reinforcement learning problem into one of probabilistic inference

MOTIVATION: Powerful algorithms from variational inference can be applied to reinforcement learning

Reinforcement Learning as Inference

AIM: To cast the reinforcement learning problem into one of probabilistic inference

MOTIVATION: Powerful algorithms from variational inference can be applied to reinforcement learning

Existing methods present several theoretical and practical barriers

Pseudo-Likelihood Methods

Pseudo-Likelihood Methods

Recast the RL objective as marginal likelihood
then maximise a tractable bound:

Pseudo-Likelihood Methods

Recast the RL objective as marginal likelihood
then maximise a tractable bound:


Pseudo-likelihood Objective : $KL(q(\tau) || p_{\theta}(\tau | \mathcal{O}))$

Pseudo-Likelihood Methods

Recast the RL objective as marginal likelihood
then maximise a tractable bound:

Pseudo-likelihood Objective : $KL(q(\tau) || p_{\theta}(\tau | \mathcal{O}))$

Target distribution,
proportional to return

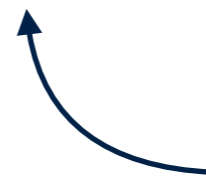


Pseudo-Likelihood Methods

Recast the RL objective as marginal likelihood
then maximise a tractable bound:

Pseudo-likelihood Objective : $KL(q(\tau) || p_{\theta}(\tau | \mathcal{O}))$

Target distribution,
proportional to return



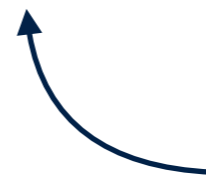
Distribution containing
policy to be improved

Pseudo-Likelihood Methods

Recast the RL objective as marginal likelihood
then maximise a tractable bound:

Pseudo-likelihood Objective : $KL(q(\tau) || p_{\theta}(\tau | \mathcal{O}))$

Target distribution,
proportional to return



Distribution containing
policy to be improved

Classic RL optimises the reverse (mode-seeking) form of KL divergence:

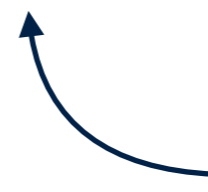
Classic RL Objective: $KL(p_{\theta}(\tau | \mathcal{O}) || q(\tau))$

Pseudo-Likelihood Methods

Recast the RL objective as marginal likelihood
then maximise a tractable bound:

Pseudo-likelihood Objective : $KL(q(\tau) || p_{\theta}(\tau | \mathcal{O}))$

Target distribution,
proportional to return



Distribution containing
policy to be improved

Classic RL optimises the reverse (mode-seeking) form of KL divergence:

Classic RL Objective: $KL(p_{\theta}(\tau | \mathcal{O}) || q(\tau))$

Pseudo-likelihood promotes risk-seeking behaviour


Maximum Entropy RL Objective

Maximum Entropy RL Objective

$$\mathcal{L}(\theta) = \mathbb{E}_{q_{\theta}(\tau)} \left[\sum_{i=0}^{N-1} (r_i - c \log \pi_{\theta}(a_i | s_i)) \right]$$

Maximum Entropy RL Objective

Variational
distribution
containing policy

$$\mathcal{L}(\theta) = \mathbb{E}_{q_{\theta}(\tau)} \left[\sum_{i=0}^{N-1} (r_i - c \log \pi_{\theta}(a_i | s_i)) \right]$$


Maximum Entropy RL Objective

$$\mathcal{L}(\theta) = \mathbb{E}_{q_{\theta}(\tau)} \left[\sum_{i=0}^{N-1} (r_i - c \log \pi_{\theta}(a_i | s_i)) \right]$$

Variational distribution containing policy

Temperature parameter

Maximum Entropy RL Objective

$$\mathcal{L}(\theta) = \mathbb{E}_{q_{\theta}(\tau)} \left[\sum_{i=0}^{N-1} (r_i - c \log \pi_{\theta}(a_i | s_i)) \right]$$

Variational distribution containing policy

Temperature parameter

Canonical algorithm: Soft Actor Critic (Haarnoja et al. 18)

Maximum Entropy RL Objective

$$\mathcal{L}(\theta) = \mathbb{E}_{q_{\theta}(\tau)} \left[\sum_{i=0}^{N-1} (r_i - c \log \pi_{\theta}(a_i | s_i)) \right]$$

Variational distribution containing policy

Temperature parameter

Canonical algorithm: Soft Actor Critic (Haarnoja et al. 18)

Optimal deterministic policies are not learned

Maximum Entropy RL Objective

$$\mathcal{L}(\theta) = \mathbb{E}_{q_{\theta}(\tau)} \left[\sum_{i=0}^{N-1} (r_i - c \log \pi_{\theta}(a_i | s_i)) \right]$$

Variational distribution containing policy

Temperature parameter

Canonical algorithm: Soft Actor Critic (Haarnoja et al. 18)

Optimal deterministic policies are not learned

Counterexamples show several cases when optimal RL policy can't be recovered

Maximum Entropy RL Objective

$$\mathcal{L}(\theta) = \mathbb{E}_{q_{\theta}(\tau)} \left[\sum_{i=0}^{N-1} (r_i - c \log \pi_{\theta}(a_i | s_i)) \right]$$

Variational distribution containing policy

Temperature parameter

Canonical algorithm: Soft Actor Critic (Haarnoja et al. 18)

Optimal deterministic policies are not learned

Counterexamples show several cases when optimal RL policy can't be recovered

Optimality of solution sensitive to temperature

Desiderata for an Inference Framework

Desiderata for an Inference Framework

Naturally learns optimal
deterministic policies

Desiderata for an Inference Framework

Naturally learns optimal deterministic policies

Temperature not a hyperparameter

Desiderata for an Inference Framework

Naturally learns optimal deterministic policies

Temperature not a hyperparameter

Function approximators explicitly used

Desiderata for an Inference Framework

Naturally learns optimal deterministic policies

Temperature not a hyperparameter

Function approximators explicitly used

Stochastic policies used for learning

Desiderata for an Inference Framework

Naturally learns optimal deterministic policies

Temperature not a hyperparameter

Discounting easily incorporated

Function approximators explicitly used

Stochastic policies used for learning

Desiderata for an Inference Framework

Naturally learns optimal deterministic policies

Optimises the reverse form of KL divergence

Temperature not a hyperparameter

Discounting easily incorporated

Function approximators explicitly used

Stochastic policies used for learning

Desiderata for an Inference Framework

Naturally learns optimal deterministic policies

Optimises the reverse form of KL divergence

Temperature not a hyperparameter

VIREL

Variational Inference
for Reinforcement Learning

Discounting easily incorporated

Function approximators explicitly used

Stochastic policies used for learning

VIREL Framework

VIREL Framework


Introduce a Boltzmann Policy

$$\pi_{\omega}(a | s) = \frac{\frac{\exp(\hat{Q}_{\omega}(s, a))}{\varepsilon_{\omega}}}{\int \frac{\exp(\hat{Q}_{\omega}(s, a))}{\varepsilon_{\omega}} dh}$$

VIREL Framework

Introduce a Boltzmann Policy $\pi_{\omega}(a | s) = \frac{\frac{\exp(\hat{Q}_{\omega}(s, a))}{\epsilon_{\omega}}}{\int \frac{\exp(\hat{Q}_{\omega}(s, a))}{\epsilon_{\omega}} dh}$

Any bounded,
smooth function
approximator



VIREL Framework

Introduce a Boltzmann Policy $\pi_{\omega}(a | s) = \frac{\exp(\hat{Q}_{\omega}(s, a))}{\int \frac{\exp(\hat{Q}_{\omega}(s, a))}{\varepsilon_{\omega}} dh}$

Any bounded,
smooth function
approximator

Temperature is mean-squared Bellman error: $\varepsilon_{\omega} = \|\mathcal{T}^* \hat{Q}_{\omega}(\cdot) - \hat{Q}_{\omega}(\cdot)\|_{d(h)}^2$

VIREL Framework

Introduce a Boltzmann Policy $\pi_{\omega}(a | s) = \frac{\exp(\hat{Q}_{\omega}(s, a))}{\int \frac{\exp(\hat{Q}_{\omega}(s, a))}{\varepsilon_{\omega}} dh}$

Any bounded,
smooth function
approximator

Temperature is mean-squared Bellman error: $\varepsilon_{\omega} = \|\mathcal{T}^* \hat{Q}_{\omega}(\cdot) - \hat{Q}_{\omega}(\cdot)\|_{d(h)}^2$

$$\varepsilon_{\omega} = 0 \implies \hat{Q}_{\omega^*}(\cdot) = Q^*(\cdot) \implies \pi_{\omega^*} = \pi^*$$

VIREL Framework

Introduce a Boltzmann Policy $\pi_{\omega}(a | s) = \frac{\frac{\exp(\hat{Q}_{\omega}(s, a))}{\epsilon_{\omega}}}{\int \frac{\exp(\hat{Q}_{\omega}(s, a))}{\epsilon_{\omega}} dh}$

Any bounded,
smooth function
approximator

Temperature is mean-squared Bellman error: $\epsilon_{\omega} = \|\mathcal{T}^* \hat{Q}_{\omega}(\cdot) - \hat{Q}_{\omega}(\cdot)\|_{d(h)}^2$

$$\epsilon_{\omega} = 0 \implies \hat{Q}_{\omega^*}(\cdot) = Q^*(\cdot) \implies \pi_{\omega^*} = \pi^*$$

Optimal
deterministic policy
learnt

VIREL Framework

Introduce a Boltzmann Policy $\pi_{\omega}(a | s) = \frac{\frac{\exp(\hat{Q}_{\omega}(s, a))}{\epsilon_{\omega}}}{\int \frac{\exp(\hat{Q}_{\omega}(s, a))}{\epsilon_{\omega}} dh}$

Any bounded,
smooth function
approximator

Temperature is mean-squared Bellman error: $\epsilon_{\omega} = \|\mathcal{T}^* \hat{Q}_{\omega}(\cdot) - \hat{Q}_{\omega}(\cdot)\|_{d(h)}^2$

$$\epsilon_{\omega} = 0 \implies \hat{Q}_{\omega^*}(\cdot) = Q^*(\cdot) \implies \pi_{\omega^*} = \pi^*$$

AIM: Find $\omega^* = \arg_{\omega} \min \epsilon_{\omega}$ and infer π_{ω^*}

Optimal
deterministic policy
learnt

VIREL Framework

Introduce a Boltzmann Policy $\pi_{\omega}(a | s) = \frac{\frac{\exp(\hat{Q}_{\omega}(s, a))}{\varepsilon_{\omega}}}{\int \frac{\exp(\hat{Q}_{\omega}(s, a))}{\varepsilon_{\omega}} dh}$

Any bounded,
smooth function
approximator

Temperature is mean-squared Bellman error: $\varepsilon_{\omega} = \|\mathcal{T}^* \hat{Q}_{\omega}(\cdot) - \hat{Q}_{\omega}(\cdot)\|_{d(h)}^2$

$$\varepsilon_{\omega} = 0 \implies \hat{Q}_{\omega^*}(\cdot) = Q^*(\cdot) \implies \pi_{\omega^*} = \pi^*$$

AIM: Find $\omega^* = \arg_{\omega} \min \varepsilon_{\omega}$ and infer π_{ω^*}

Optimal
deterministic policy
learnt

Introduce variational policy $\pi_{\theta}(a | s) \approx \pi_{\omega}(a | s)$

VIREL Framework

Introduce a Boltzmann Policy $\pi_{\omega}(a | s) = \frac{\frac{\exp(\hat{Q}_{\omega}(s, a))}{\varepsilon_{\omega}}}{\int \frac{\exp(\hat{Q}_{\omega}(s, a))}{\varepsilon_{\omega}} dh}$

Any bounded,
smooth function
approximator

Temperature is mean-squared Bellman error: $\varepsilon_{\omega} = \|\mathcal{T}^* \hat{Q}_{\omega}(\cdot) - \hat{Q}_{\omega}(\cdot)\|_{d(h)}^2$

$$\varepsilon_{\omega} = 0 \implies \hat{Q}_{\omega^*}(\cdot) = Q^*(\cdot) \implies \pi_{\omega^*} = \pi^*$$

AIM: Find $\omega^* = \arg_{\omega} \min \varepsilon_{\omega}$ and infer π_{ω^*}

Optimal
deterministic policy
learnt

Introduce variational policy $\pi_{\theta}(a | s) \approx \pi_{\omega}(a | s)$

Find $\theta^* = \arg_{\theta} \min KL(\pi_{\theta} || \pi_{\omega})$

VIREL Framework

Introduce a Boltzmann Policy $\pi_{\omega}(a | s) = \frac{\frac{\exp(\hat{Q}_{\omega}(s, a))}{\epsilon_{\omega}}}{\int \frac{\exp(\hat{Q}_{\omega}(s, a))}{\epsilon_{\omega}} dh}$

Any bounded,
smooth function
approximator

Temperature is mean-squared Bellman error: $\epsilon_{\omega} = \|\mathcal{T}^* \hat{Q}_{\omega}(\cdot) - \hat{Q}_{\omega}(\cdot)\|_{d(h)}^2$

$$\epsilon_{\omega} = 0 \implies \hat{Q}_{\omega^*}(\cdot) = Q^*(\cdot) \implies \pi_{\omega^*} = \pi^*$$

AIM: Find $\omega^* = \arg_{\omega} \min \epsilon_{\omega}$ and infer π_{ω^*}

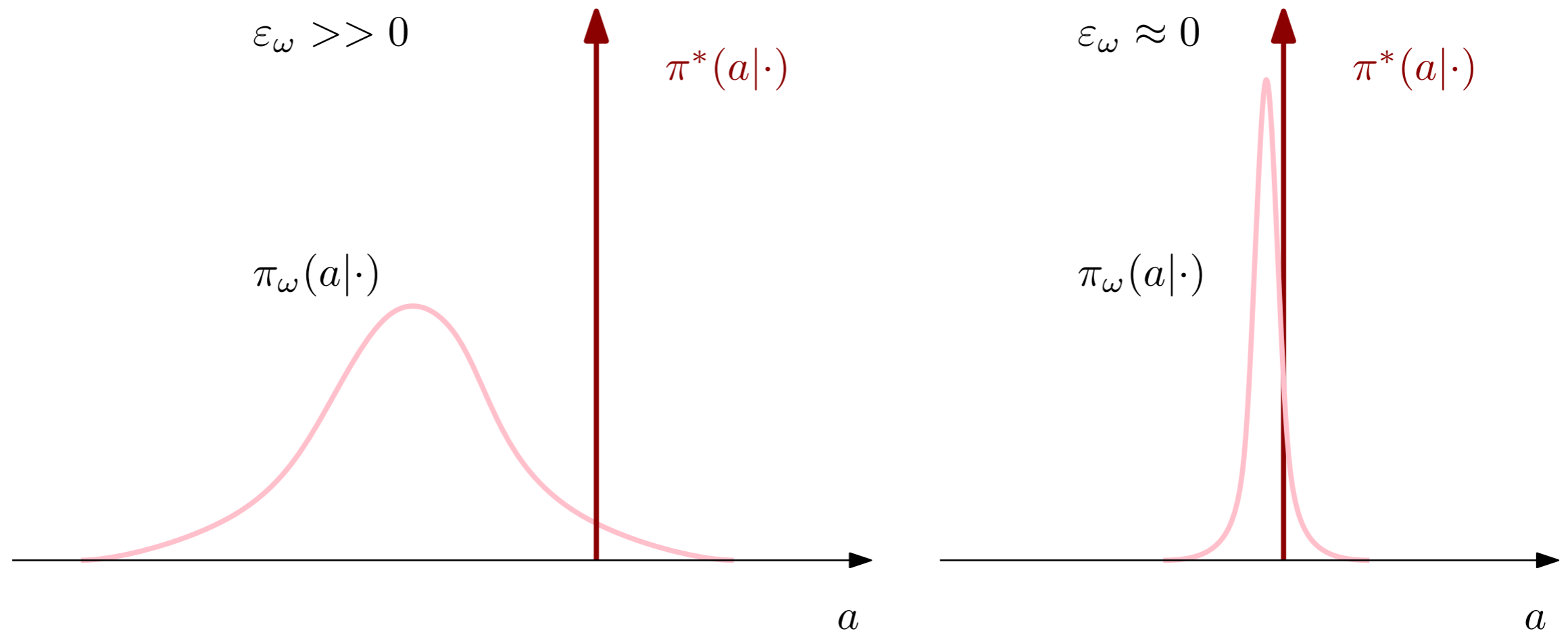
Optimal
deterministic policy
learnt

Introduce variational policy $\pi_{\theta}(a | s) \approx \pi_{\omega}(a | s)$

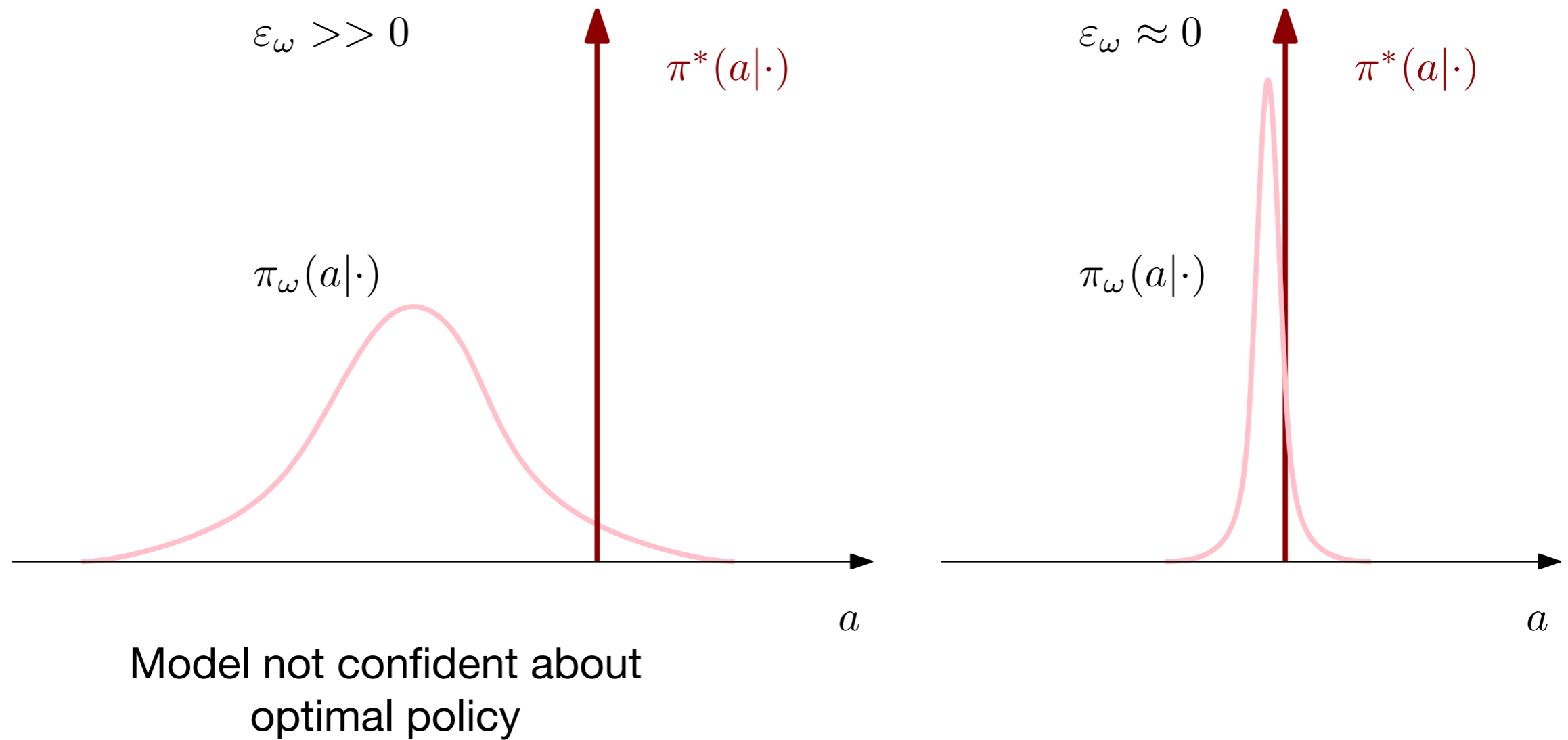
Find $\theta^* = \arg_{\theta} \min KL(\pi_{\theta} || \pi_{\omega})$

Minimises reverse
KL divergence

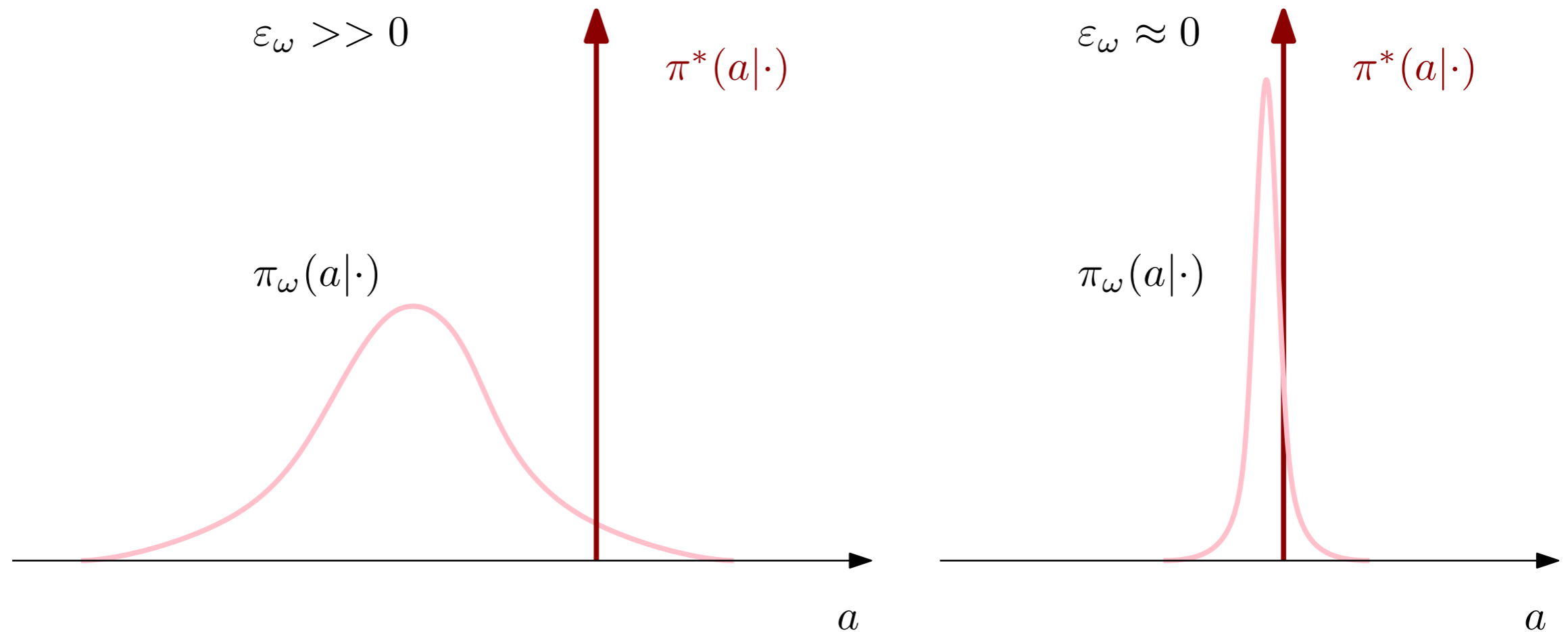
Naturally Balanced Exploration/Exploitation:



Naturally Balanced Exploration/Exploitation:



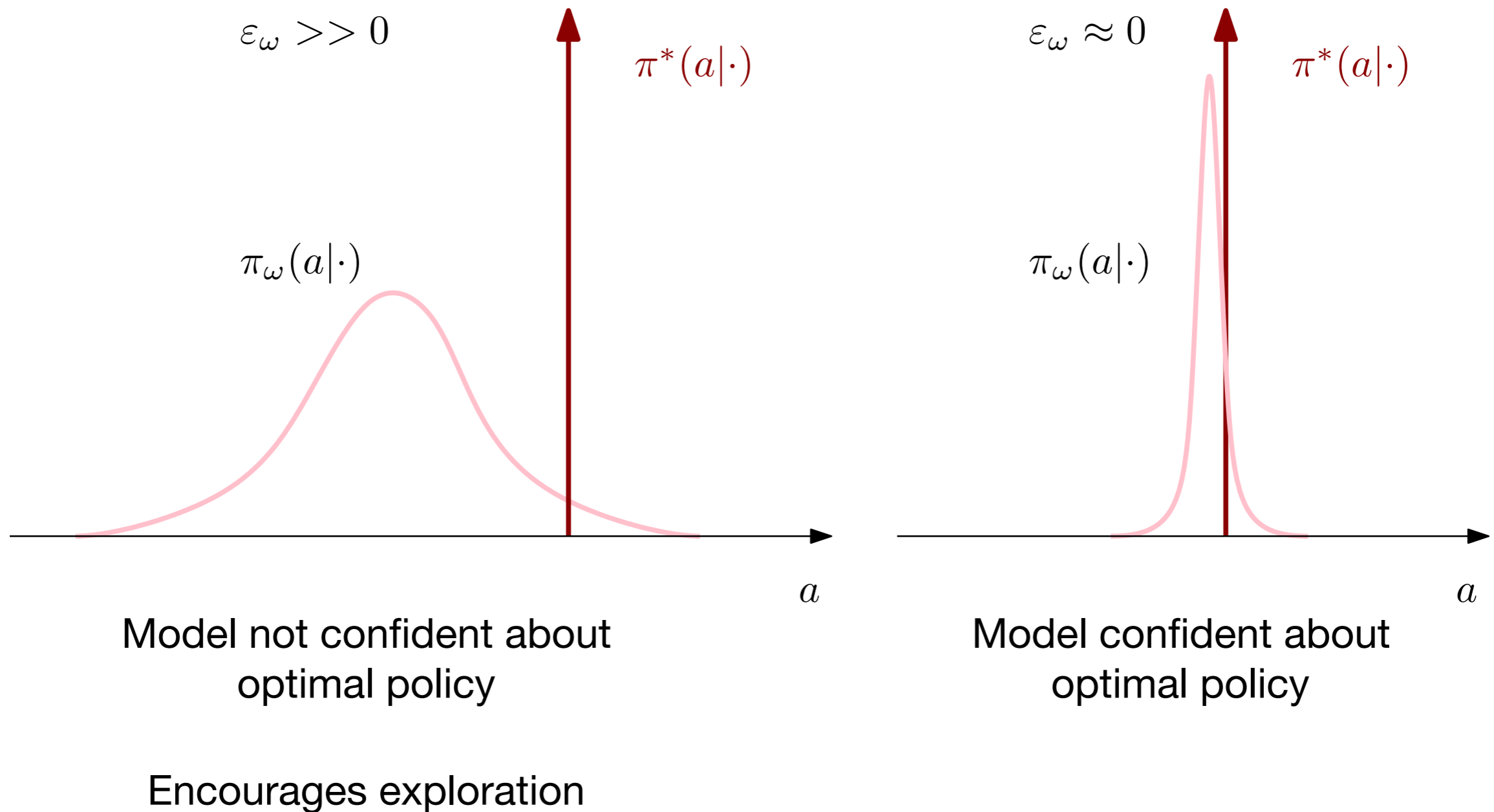
Naturally Balanced Exploration/Exploitation:



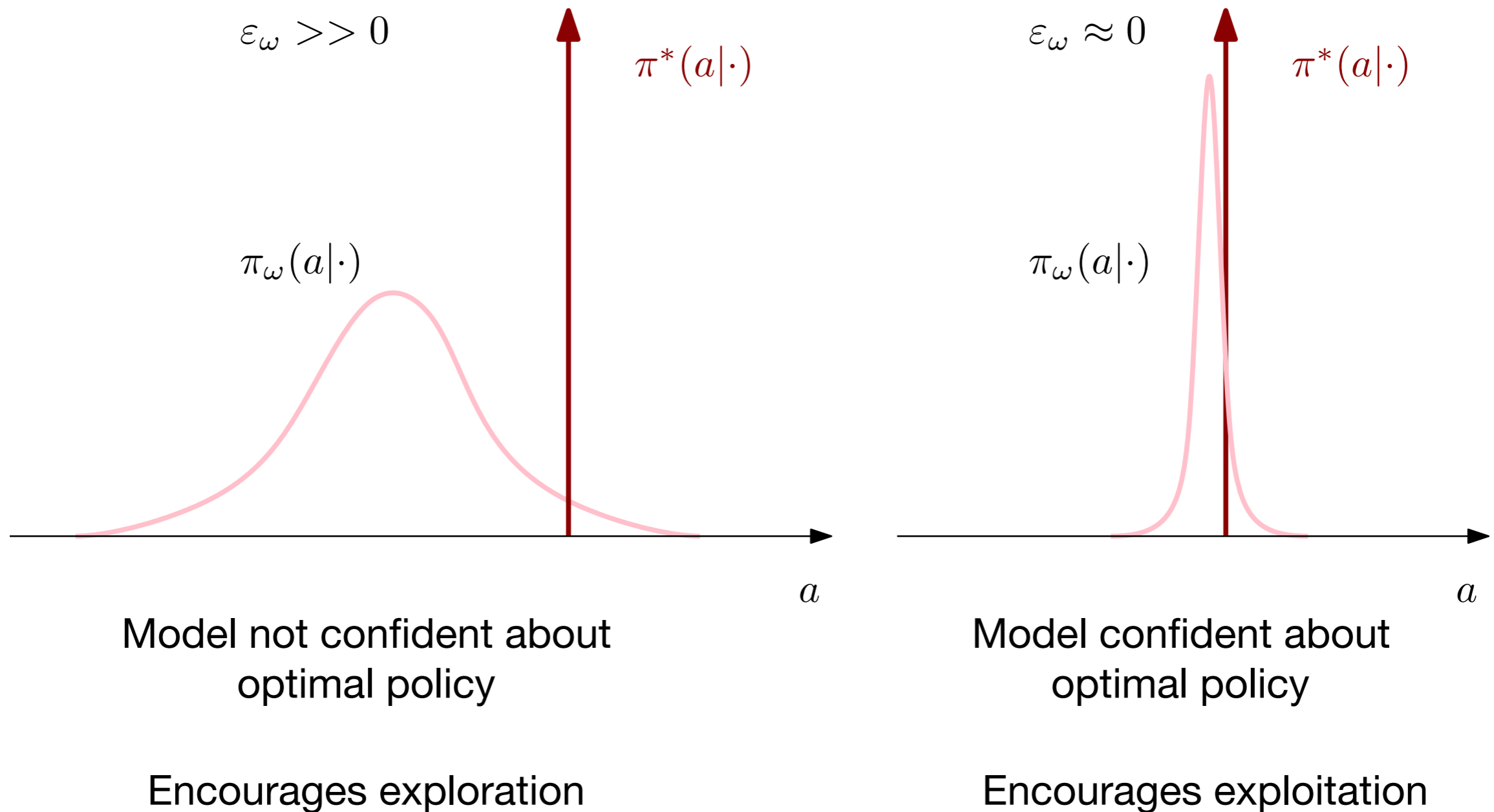
Model not confident about
optimal policy

Encourages exploration

Naturally Balanced Exploration/Exploitation:



Naturally Balanced Exploration/Exploitation:



Actor-Critic Algorithms

Actor-Critic Algorithms

KL divergence intractable, so maximise evidence lower bound instead

Actor-Critic Algorithms

KL divergence intractable, so maximise evidence lower bound instead

Using expectation maximisation (EM) with VIREL framework yields an actor-critic algorithm

Actor-Critic Algorithms

KL divergence intractable, so maximise evidence lower bound instead

Using expectation maximisation (EM) with VIREL framework yields an actor-critic algorithm

E-step = Entropy regularised policy improvement (**actor**)

Actor-Critic Algorithms

KL divergence intractable, so maximise evidence lower bound instead

Using expectation maximisation (EM) with VIREL framework yields an actor-critic algorithm

E-step = Entropy regularised policy improvement (**actor**)

M-step = Policy evaluation (**critic**)

Actor-Critic Algorithms

KL divergence intractable, so maximise evidence lower bound instead

Using expectation maximisation (EM) with VIREL framework yields an actor-critic algorithm

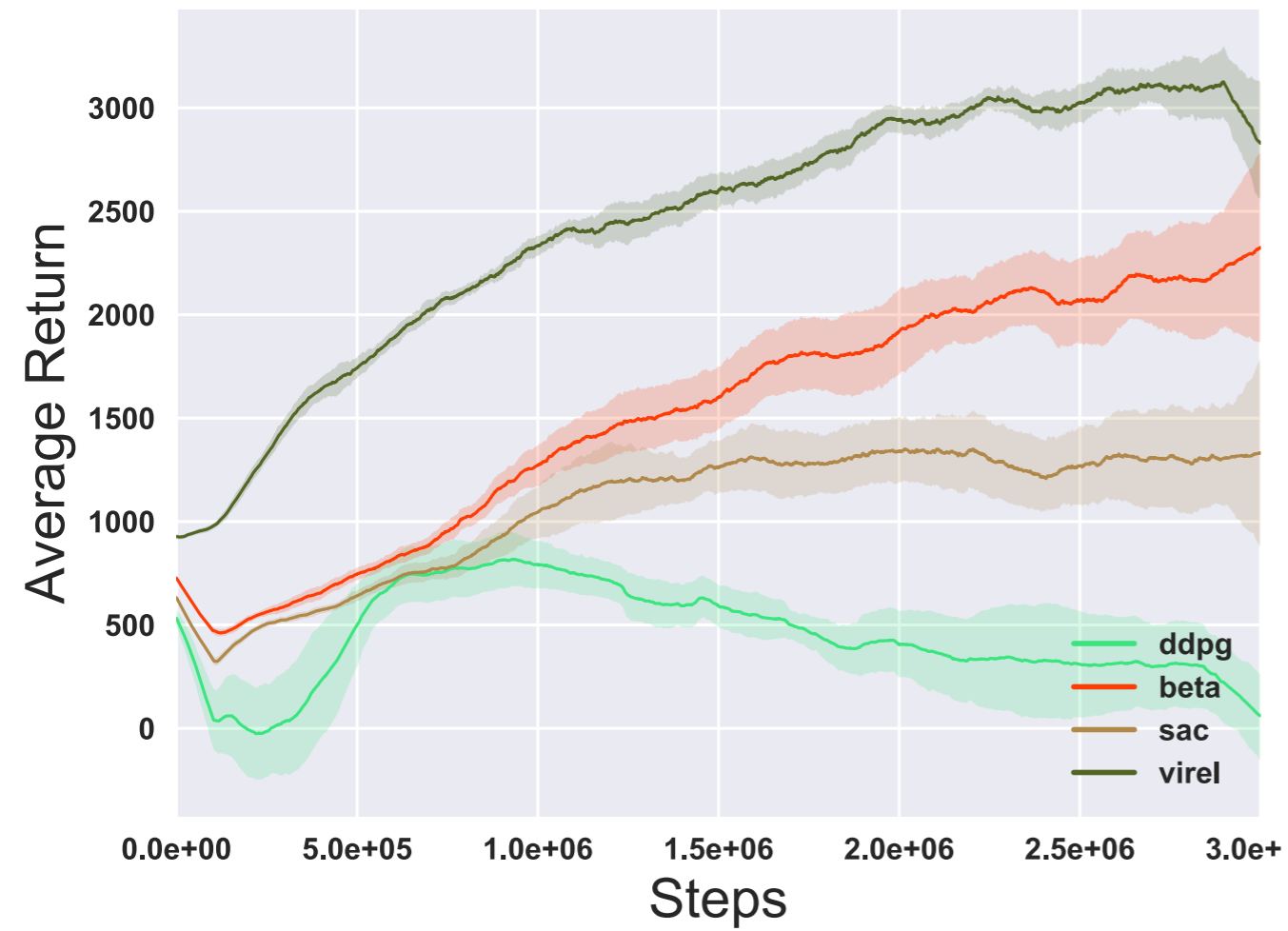
E-step = Entropy regularised policy improvement (**actor**)

M-step = Policy evaluation (**critic**)

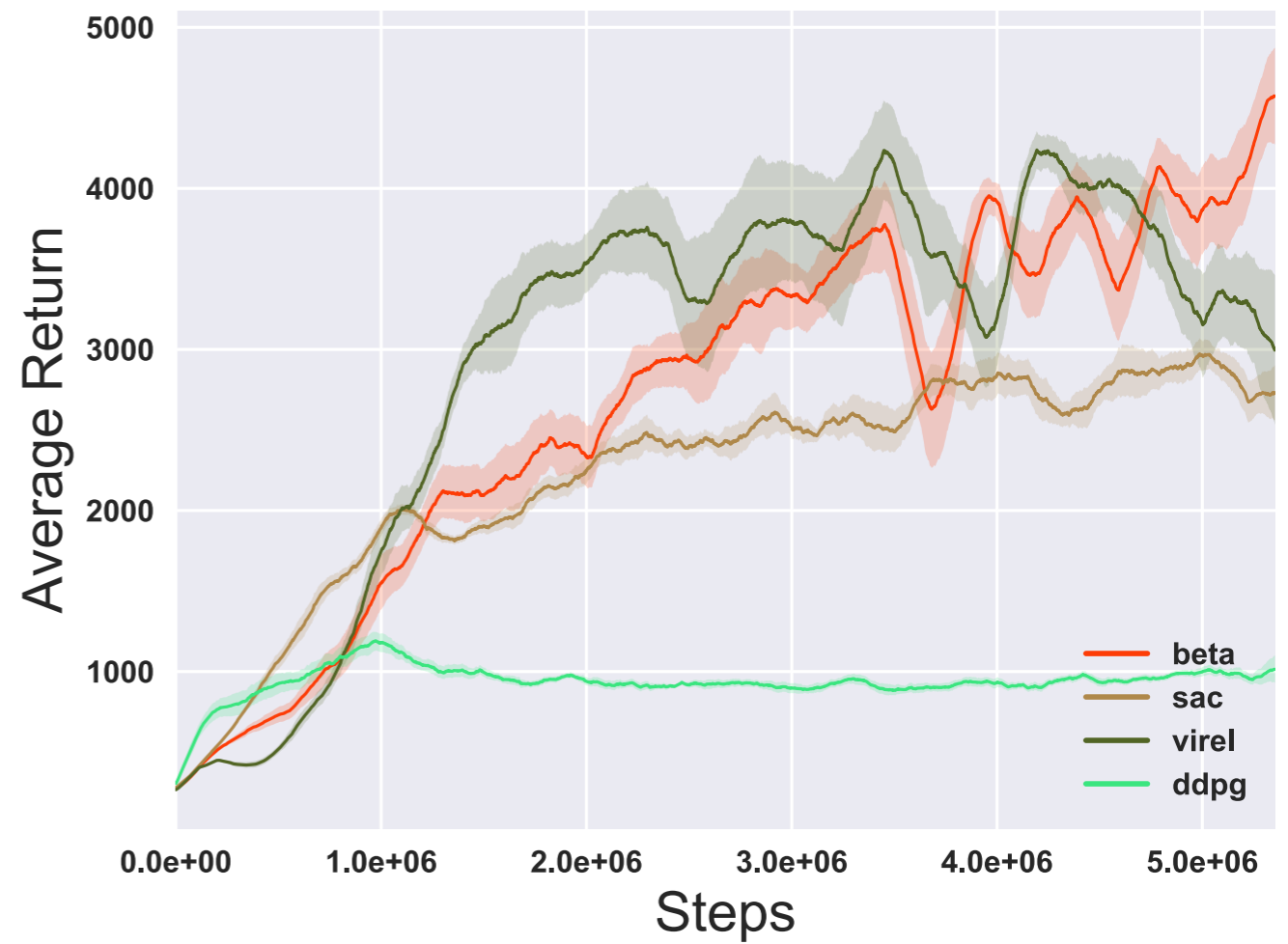
Framework very general: Maximum a Posteriori Policy Optimisation (Abdolmaleki et al. 18) easily derived from VIREL **without simplifying approximations**

Results

Ant-v2



Humanoid-v2





**Whiteson
Research
Lab**



UNIVERSITY OF
OXFORD

Thank you for listening

Please visit our poster

**East Exhibition Hall B + C
Number 214**