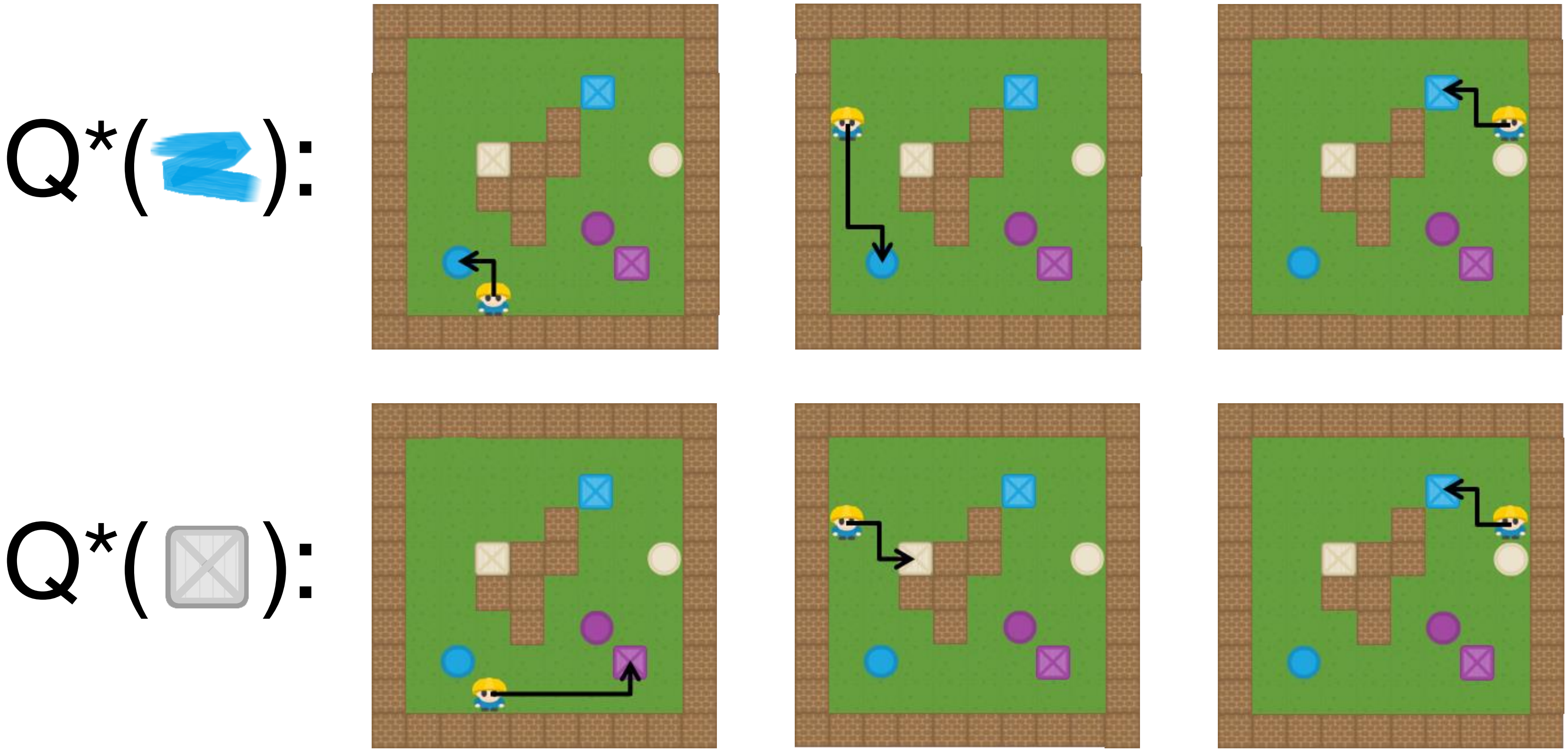
Composing Value Functions in Reinforcement Learning

Benjamin van Niekerk*, Steven James*,
Adam Earle, Benjamin Rosman
University of the Witwatersrand, Johannesburg, South Africa

Poster #251

WITS UNIVERSITY

R A I L
L A B



Can we **blend** these value functions to solve interesting **combinations** of the tasks without further learning?

WITS UNIVERSITY

Poster #251

RAIL LAB

Skill Composition

- In general RL, for two skills Q_1 and Q_2, we have that

WITS UNIVERSITY

RAIL LAB

Skill Composition

- In general RL, for two skills Q_1 and Q_2, we have that

$Q_1 \oplus Q_2 =$ 💩

Entropy Regularised RL

- Augment reward function with penalty term:

$$r_{ent}(s,a) = r(s,a) - \tau \mathrm{KL}[\pi_s \,||\, \bar{\pi}_s]$$

$$V_\pi(s) = \mathbb{E}_s^\pi \left[\sum_{t=0}^{\infty} r(s_t, a_t) - \tau \mathrm{KL}[\pi_{s_t} || \bar{\pi}_{s_t}] \right]$$

WITS UNIVERSITY

RAIL LAB

Entropy Regularised RL

- Augment reward function with penalty term:

$$r_{ent}(s,a) = r(s,a) - \tau \mathrm{KL}[\pi_s \,||\, \bar{\pi}_s]$$

Being different to "reference" policy $\bar{\pi}$

$$V_\pi(s) = \mathbb{E}_s^\pi\left[\sum_{t=0}^{\infty} r(s_t,a_t) - \tau \mathrm{KL}[\pi_{s_t} || \bar{\pi}_{s_t}]\right]$$

WITS UNIVERSITY

RAIL LAB

Entropy Regularised RL

- Augment reward function with penalty term:

$$r_{ent}(s,a) = r(s,a) - \tau \mathrm{KL}[\pi_s \,||\, \bar{\pi}_s]$$

Temperature

Being different to "reference" policy $\bar{\pi}$

$$V_\pi(s) = \mathbb{E}_s^\pi \left[\sum_{t=0}^{\infty} r(s_t, a_t) - \tau \mathrm{KL}[\pi_{s_t} \,||\, \bar{\pi}_{s_t}] \right]$$

Skill Composition

- In general RL, for two skills Q_1 and Q_2, we have that

$$Q_1 \oplus Q_2 = \text{💩}$$

Skill Composition

- In general RL, for two skills Q_1 and Q_2, we have that

$Q_1 \oplus Q_2 =$ 💩

- With entropy regularisation, we show that

$Q_1 \oplus Q_2 =$ 🥰

Provably

Skill Composition

- In general RL, for two skills Q_1 and Q_2, we have that

$$Q_1 \oplus Q_2 = 💩$$

- With entropy regularisation, we show that

$$Q_1 \oplus Q_2 = 🥰$$

Deterministic transitions only

Provably

WITS UNIVERSITY

Poster #251

RAIL LAB

-OR- Task Composition

We can optimally compose $Q(\boxtimes)$ and $Q(\bullet)$ to solve the task of collecting $\boxtimes$ or $\bullet$:

$$Q(\boxtimes \text{ or } \bullet) =$$

$$\tau \log(w_{\bullet} \exp(Q(\bullet)/\tau) + w_{\boxtimes} \exp(Q(\boxtimes)/\tau))$$

-OR- Task Composition

We can optimally compose $Q(\boxtimes)$ and $Q(\bullet)$ to solve the task of collecting $\boxtimes$ or $\bullet$:

$$Q(\boxtimes \text{ or } \bullet) =$$

LogSumExp = "soft" maximum

$$\tau \log(w_{\bullet} \exp(Q(\bullet)/\tau) + w_{\boxtimes} \exp(Q(\boxtimes)/\tau))$$

WITS UNIVERSITY

RAIL LAB

-OR- Task Composition

We can optimally compose $Q(\quad)$ and $Q(\quad)$ to solve the task of collecting or :

$$Q(\quad \text{or} \quad) =$$

LogSumExp = "soft" maximum

$$\tau \log(w \exp(Q(\quad)/\tau) + w \exp(Q(\quad)/\tau))$$

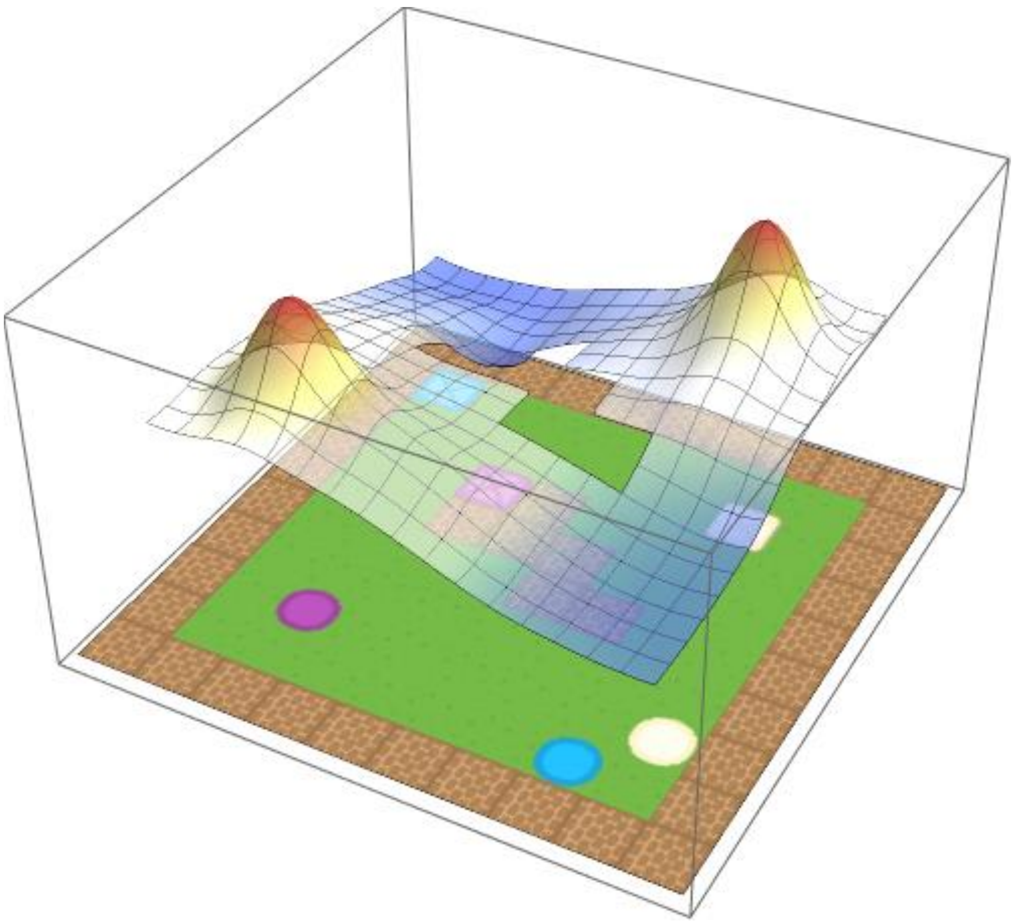
Use these to weight

WITS UNIVERSITY

Poster #251

RAIL LAB

Experiment -OR-



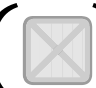
Poster #251

WITS UNIVERSITY

RAIL LAB

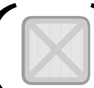
A Connection to Standard RL

<u>Corollary</u>: In the limit as $\tau \downarrow 0$ i.e. in the standard RL setting, we prove that:

$Q($ or $) = \max\{Q($$), Q($$)\}$

A Connection to Standard RL

<u>Corollary</u>: In the limit as $\tau \downarrow 0$ i.e. in the standard RL setting, we prove that:

$Q($  or  $) = \max\{Q($  $), Q($  $)\}$

regular max

A Connection to Standard RL

<u>Corollary</u>: In the limit as $\tau \downarrow 0$ i.e. in the standard RL setting, we prove that:

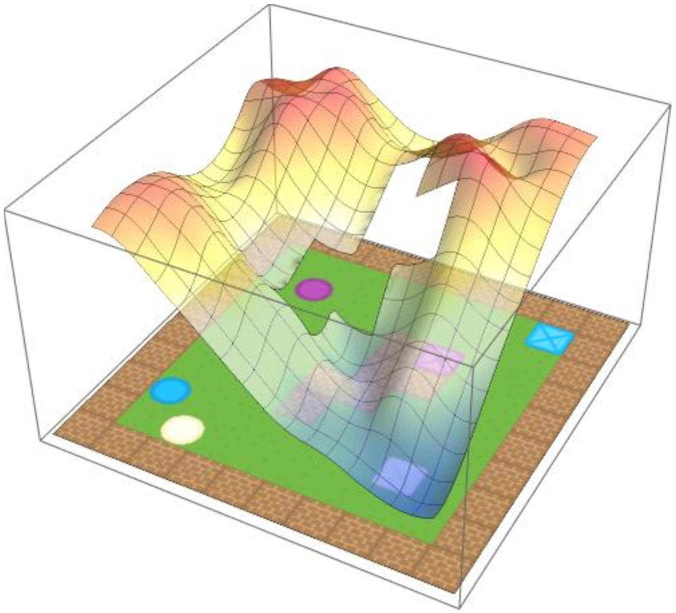
$$Q(\quad \mathit{or} \quad) = \max\{Q(\quad), Q(\quad)\}$$

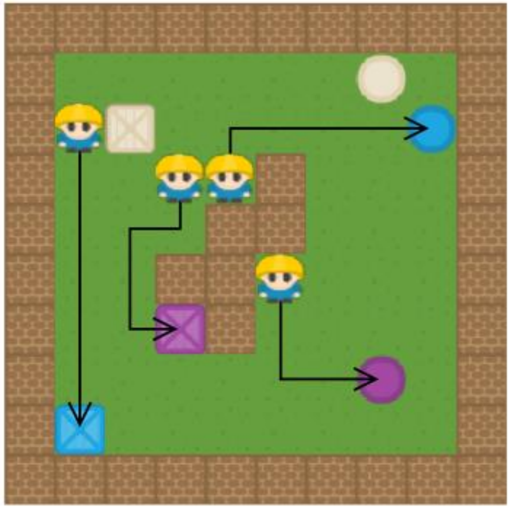
regular max

No weights! ☹

Poster #251

Experiment -OR-





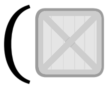
WITS UNIVERSITY

Poster #251

RAIL LAB

-AND- Task Composition

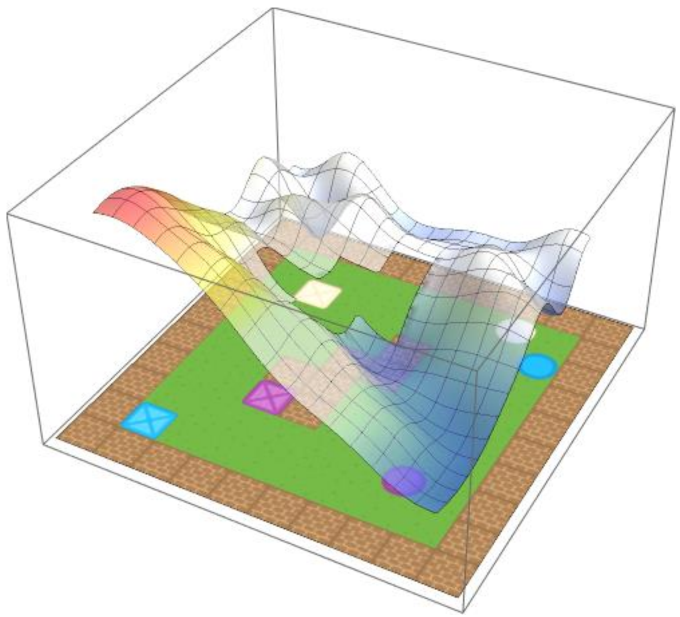
- Previous work in entropy regularised RL [Haarnoja et al, 2018] shows that -AND- task is approximately

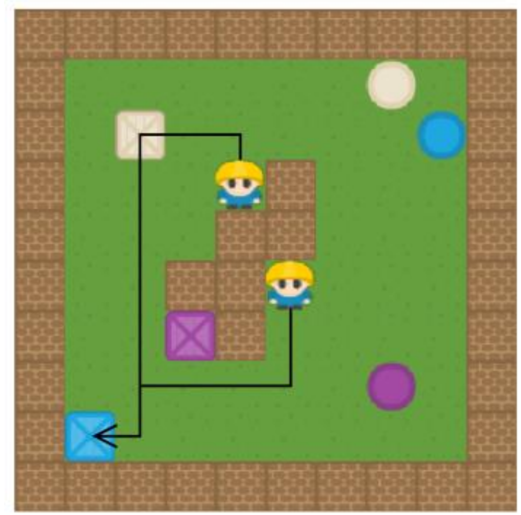
$Q($  and  $) \approx (Q($  $) + Q($  $))/2$

- Conjecture this holds in low-temperature limit

Haarnoja, T., Pong, V., Zhou, A., Dalal, M., Abbeel, P., and Levine, S.
Composable deep reinforcement learning for robotic manipulation. *arXiv preprint arXiv:1803.06773, 2018.*

Experiment -AND-

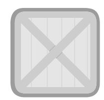




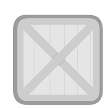
Poster #251

Conclusion

- We can do zero-shot composition to provably find

Q*( or ) (with different priorities)

- We provide empirical evidence that averaging Q-values approximates:

Q*( and )

Poster #251

WITS UNIVERSITY

R A I L
L A B