Composition and Zero-Shot Transfer with Lattice Structures in Reinforcement Learning

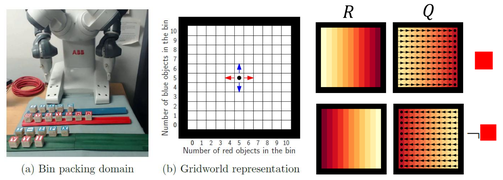
Geraud Nangue Tasse*, Steven James and Benjamin Rosman

University of the Witwatersrand, Johannesburg, South Africa

We can provably: specify tasks with arbitrary rewards using logics, construct a basis for arbitrary rewards bounds, and solve tasks zero-shot by leveraging World Value Functions.

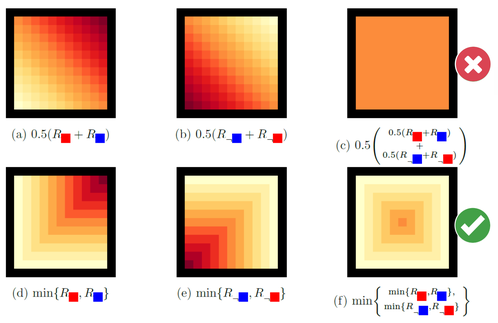
Motivation

We want **instructable agents** (e.g. via language) that can **solve tasks beyond Boolean rewards** [1,2], and **generalise compositionally** new tasks.



(a) Bin packing domain (b) Gridworld representation

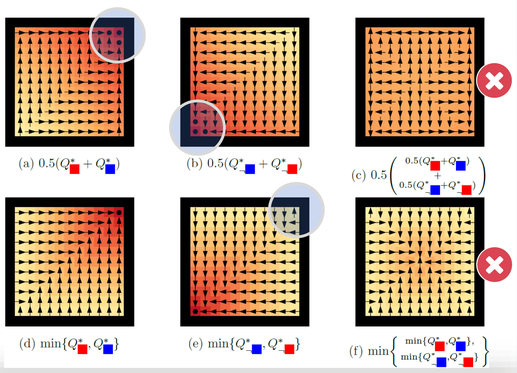
Task/Reward specification is hard!
Arbitrary reward compositions maybe not result in the desired tasks



(a) $0.5(R_{\blacksquare}+R_{\blacksquare})$ (b) $0.5(R_{\blacksquare}+R_{\blacksquare})$ (c) $0.5\left(\frac{0.5(R_{\blacksquare}+R_{\blacksquare})}{0.5(R_{\blacksquare}+R_{\blacksquare})}\right)$

(d) $\min\{R_{\blacksquare},R_{\blacksquare}\}$ (e) $\min\{R_{\blacksquare},R_{\blacksquare}\}$ (f) $\min\left\{\begin{matrix}\min\{R_{\blacksquare},R_{\blacksquare}\},\\ \min\{R_{\blacksquare},R_{\blacksquare}\}\end{matrix}\right\}$

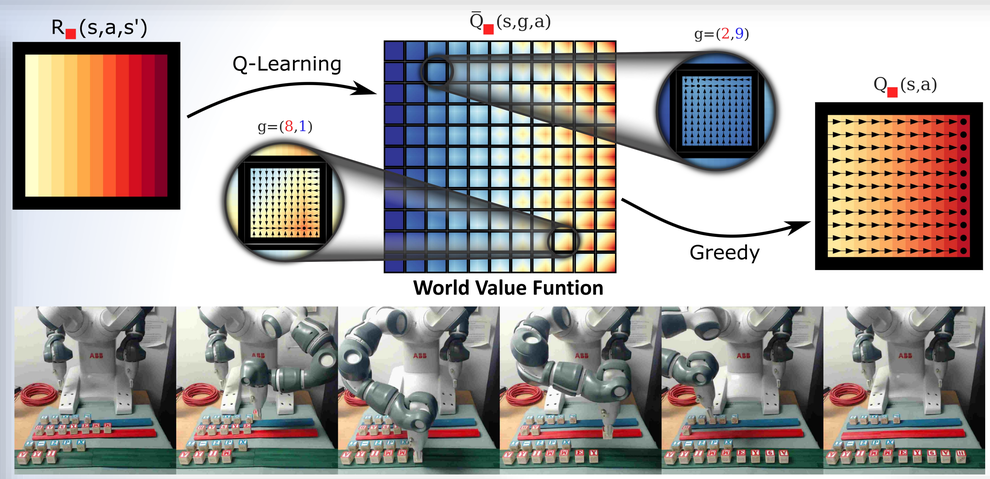
Skill/Value function/Policy learning is hard!
Arbitrary skill compositions maybe not result in the desired solutions

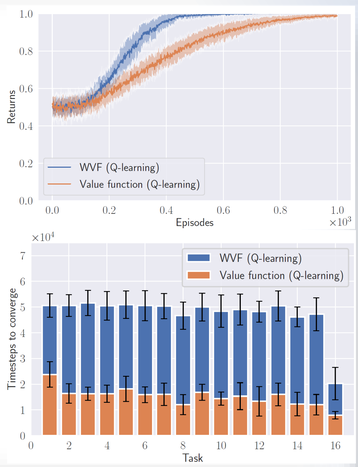


(a) $0.5(Q^*_{\blacksquare}+Q^*_{\blacksquare})$ (b) $0.5(Q^*_{\blacksquare}+Q^*_{\blacksquare})$ (c) $0.5\left(\frac{0.5(Q^*_{\blacksquare}+Q^*_{\blacksquare})}{0.5(Q^*_{\blacksquare}+Q^*_{\blacksquare})}\right)$

(d) $\min\{Q^*_{\blacksquare},Q^*_{\blacksquare}\}$ (e) $\min\{Q^*_{\blacksquare},Q^*_{\blacksquare}\}$ (f) $\min\left\{\begin{matrix}\min\{Q^*_{\blacksquare},Q^*_{\blacksquare}\},\\ \min\{Q^*_{\blacksquare},Q^*_{\blacksquare}\}\end{matrix}\right\}$

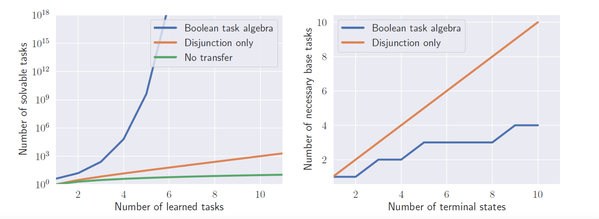
[1] G. Nangue Tasse et al., "A Boolean task algebra for reinforcement learning" NeurIPS, 2020.
[2] R. T. Icarte et al., "Using reward machines for high-level task specification and decomposition in reinforcement learning," ICML, 2018





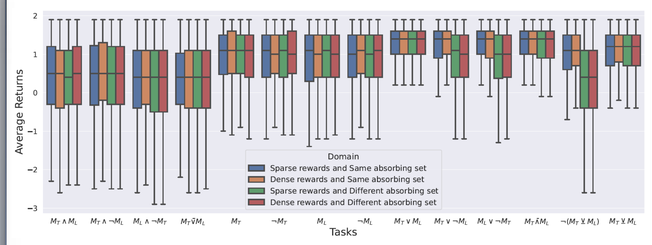
Blessing of Dimensionality

Exponential explosion of skills from a **logarithmic amount of learning**



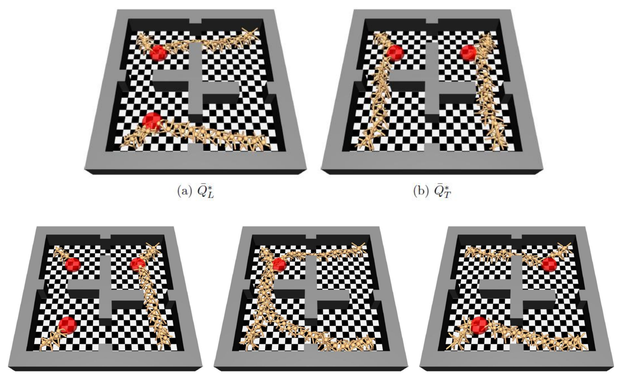
Empirically works even when theoretical assumptions do not hold

Assumptions:
- *Deterministic dynamics,*
- *Same dynamics (i.e. even termination must be the same)*
- *Same non-terminal rewards (i.e. goal-reaching tasks)*



Trained and evaluated with noisy transitions (0.3% noise)

Empirically works even in continuous control with function approximation



(a) $\bar{Q}^*_L$ (b) $\bar{Q}^*_T$

(a) $\bar{Q}^*_L \vee \bar{Q}^*_T$ (b) $\bar{Q}^*_L \wedge \bar{Q}^*_T$ (c) $\bar{Q}^*_L \veebar \bar{Q}^*_T$

Task (Rewards) composition

Logic Operators
- ***Disjunction (OR):*** $A \vee B := \max\{R_A(s,a,s'), R_B(s,a,s')\}$
- ***Conjunction (AND):*** $A \wedge B := \min\{R_A(s,a,s'), R_B(s,a,s')\}$
- ***Negation (NOT):*** $\neg A := \big(R_{MAX}(s,a,s') + R_{MIN}(s,a,s')\big) - R_A(s,a,s')$

Task space bounds $\{R_{MAX}, R_{MIN}\}$ change semantics: $\neg\{\square,\blacksquare\} = \{\square,\blacksquare\}$, $\neg\{\square,\blacksquare\} = \{\blacksquare,\blacksquare\}$

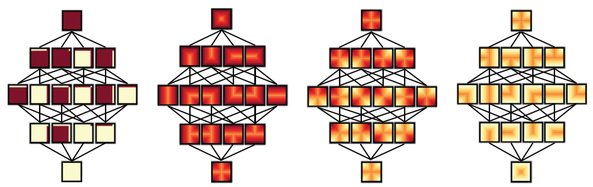


Figure 9: Examples of Boolean task sub-algebra with basis $\{\square,\square\}$, $\{\blacksquare,\blacksquare\}$, $\{\square,\blacksquare\}$, $\{\square,\square\}$ and task space bounds $\{\square,\blacksquare\}$, $\{\blacksquare,\blacksquare\}$, $\{\square,\blacksquare\}$, $\{\square,\square\}$ respectively.

Constructing a Basis

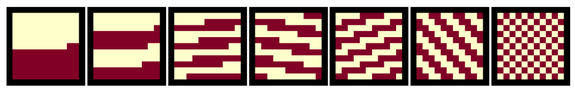
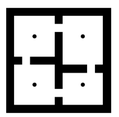


Figure 11: Basis (minimal generators) for the Boolean task algebra bounded by $\{\square,\blacksquare\}$.



(a) Four Rooms domain. The circles indicate goals the agent must reach ($\mathcal{G}$).

Goals ($\mathcal{G}$)	M_D	M_R
top-left	0	0
top-right	0	1
bottom-left	1	0
bottom-right	1	1

(b) Goals labeled by the well order $\preceq$ given by: `top-left` $\preceq$ `top-right` $\preceq$ `bottom-left` $\preceq$ `bottom-right`.

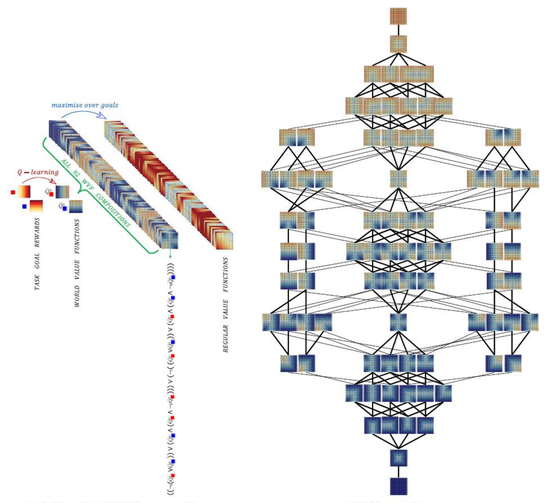
Goals ($\mathcal{G}$)	M_T	M_L
bottom-right	0	0
bottom-left	0	1
top-right	1	0
top-left	1	1

(c) Goals labeled by the well order $\preceq$ given by: `bottom-right` $\preceq$ `bottom-left` $\preceq$ `top-right` $\preceq$ `top-left`.

Skill (WVF) composition

Logic Operators
- ***Disjunction (OR):*** $Q_A \vee Q_B := \max\{Q_A(s,a,g), Q_B(s,a,g)\}$
- ***Conjunction (AND):*** $Q_A \wedge Q_B := \min\{Q_A(s,a,g), Q_B(s,a,g)\}$
- ***Negation (NOT):*** $\neg Q_A := \big(Q_{MAX}(s,a,g) + Q_{MIN}(s,a,g)\big) - Q_A(s,a,g)$

Skill space bounds $\{Q_{MAX}, Q_{MIN}\}$ similarly change semantics.



(a) Zero-shot WVF compositions (b) Hasse diagram

Illustration of the De Morgan sub-lattice generated by composing $\bar{Q}^*_{\blacksquare}$ and $\bar{Q}^*_{\blacksquare}$