Improving Action Selection in MDP's via Knowledge Transfer

Alexander Sherstov and **Peter Stone**

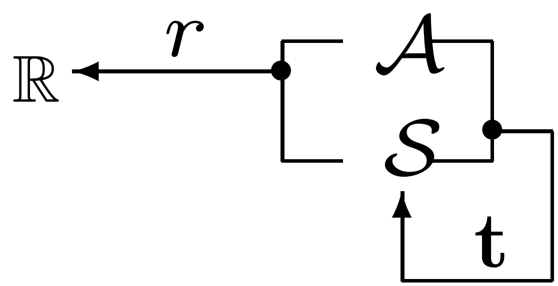
Department of Computer Sciences
The University of Texas at Austin

Overview

- Learning setting
  - Must learn ***multiple*** tasks in the same domain
  - Actions ***not uniformly*** relevant
  - Designed for ***large*** action sets

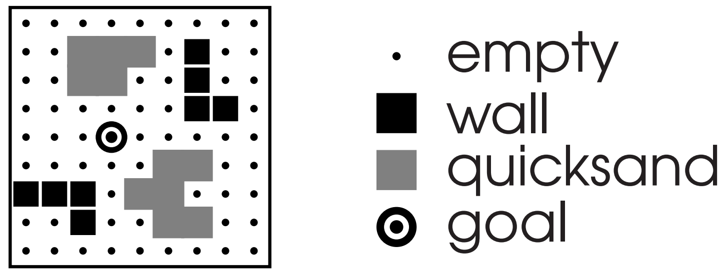
- Solution: action transfer
  - Usually beneficial (pastry chef, driver, bidding agent)
  - Formalism + analysis of action transfer
  - Enhancement: ***randomized task perturbation***
  - Empirical validation

MDP Formalism



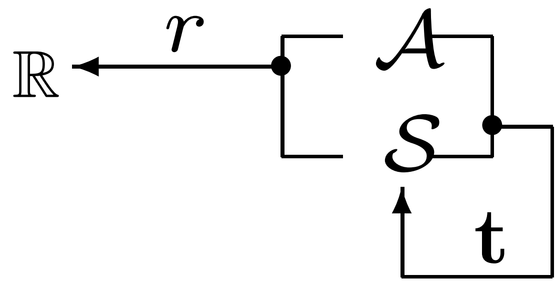
- $\mathcal{S}$ is a set of ***states***, $\mathcal{A}$ is a set of ***actions***
- $\mathbf{t} : \mathcal{S} \times \mathcal{A} \to \Pr(\mathcal{S})$ is a ***transition function***
- $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is a ***reward function***
- ***policy*** is any mapping $\mathcal{S} \to \mathcal{A}$; want policy with maximum expected return in all states

Running Example: Grid World Domain



- states = cells,
  actions = $\{(d,p) : d \in \{\uparrow, \downarrow, \rightarrow, \leftarrow\}, p \in [0.5, 0.9]\}$
- move succeeds w/ prob. proportional to p, random o/w
- reward: $-1/2$ in quicksand, $1/2$ in goal, $-p^2$ o/w
- optimal actions have $p \in [0.5, 0.6]$

Need for Related-Task Formalism



- $r, \mathbf{t}$ defined in terms of state space $\mathcal{S}$
- $r, \mathbf{t}$ incomparable across tasks
- But: want to exploit ***abstract*** transition/reward dynamics

Eliminating $\mathcal{S}$: Outcomes

- Original definition:

$$\mathbf{t} : \mathcal{S} \times \mathcal{A} \to \Pr(\mathcal{S})$$
$$r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$$

- Definition using ***outcomes*** $\mathcal{O}$:

$$\mathbf{t} : \mathcal{S} \times \mathcal{A} \to \Pr(\mathcal{O})$$
$$r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$$

- Example in grid world: $\mathcal{O} = \{\uparrow, \downarrow, \rightarrow, \leftarrow, \text{STAY}\}$.

Eliminating $\mathcal{S}$: Classes

- Previous definition:

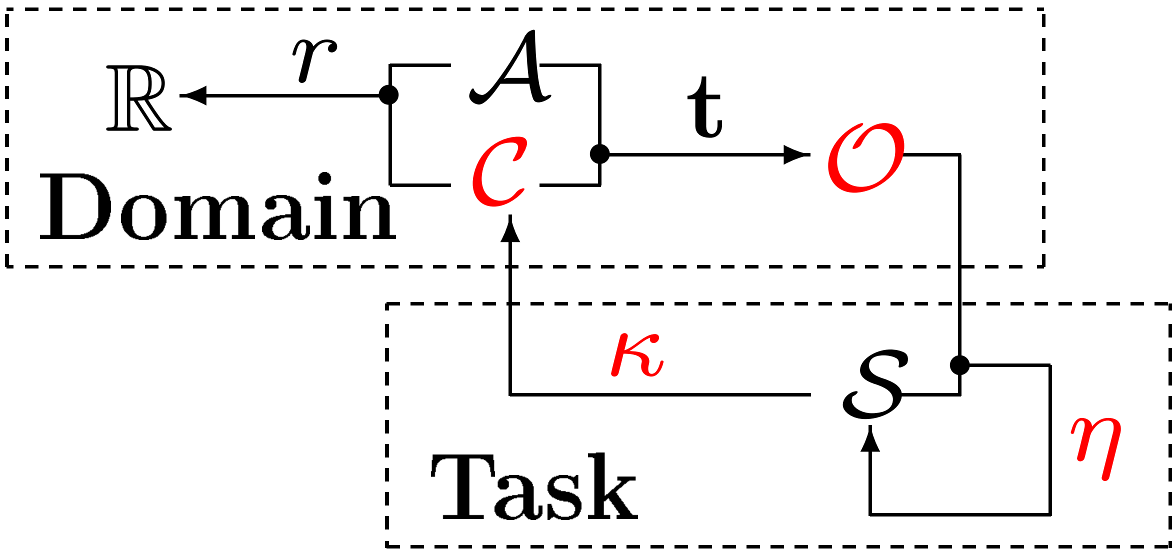
$$\mathbf{t} : \mathcal{S} \times \mathcal{A} \rightarrow \Pr(\mathcal{O})$$
$$r : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$$

- Final definition, using outcomes $\mathcal{O}$ and ***classes*** $\mathcal{C}$:

$$\mathbf{t} : \mathcal{C} \times \mathcal{A} \rightarrow \Pr(\mathcal{O})$$
$$r : \mathcal{C} \times \mathcal{A} \rightarrow \mathbb{R}$$

- Example in grid world: $\mathcal{C} = \{$ EMPTY, GOAL, QUICKSAND $\}$.

Complete Formalism



$$\mathbf{t} : \mathcal{C} \times \mathcal{A} \to \Pr(\mathcal{O})$$

$$r : \mathcal{C} \times \mathcal{A} \to \mathbb{R}$$

Dissimilarity Metric

- Similarities across tasks are implicit, unstated.

- Dissimilarity metric $\Delta(U, \tilde{U})$ ***re-expresses*** these high-level similarities in a ***precise, analytically tractable*** geometric quantity.

- $\Delta(U, \tilde{U})$ expressed i.n.o. the new formalism and V^*

Bound Based on Task Similarity

Transfer results in a value drop of at most

$$\Delta(U, \tilde{U}) \cdot \sqrt{2}\gamma/(1-\gamma)$$

at each state.

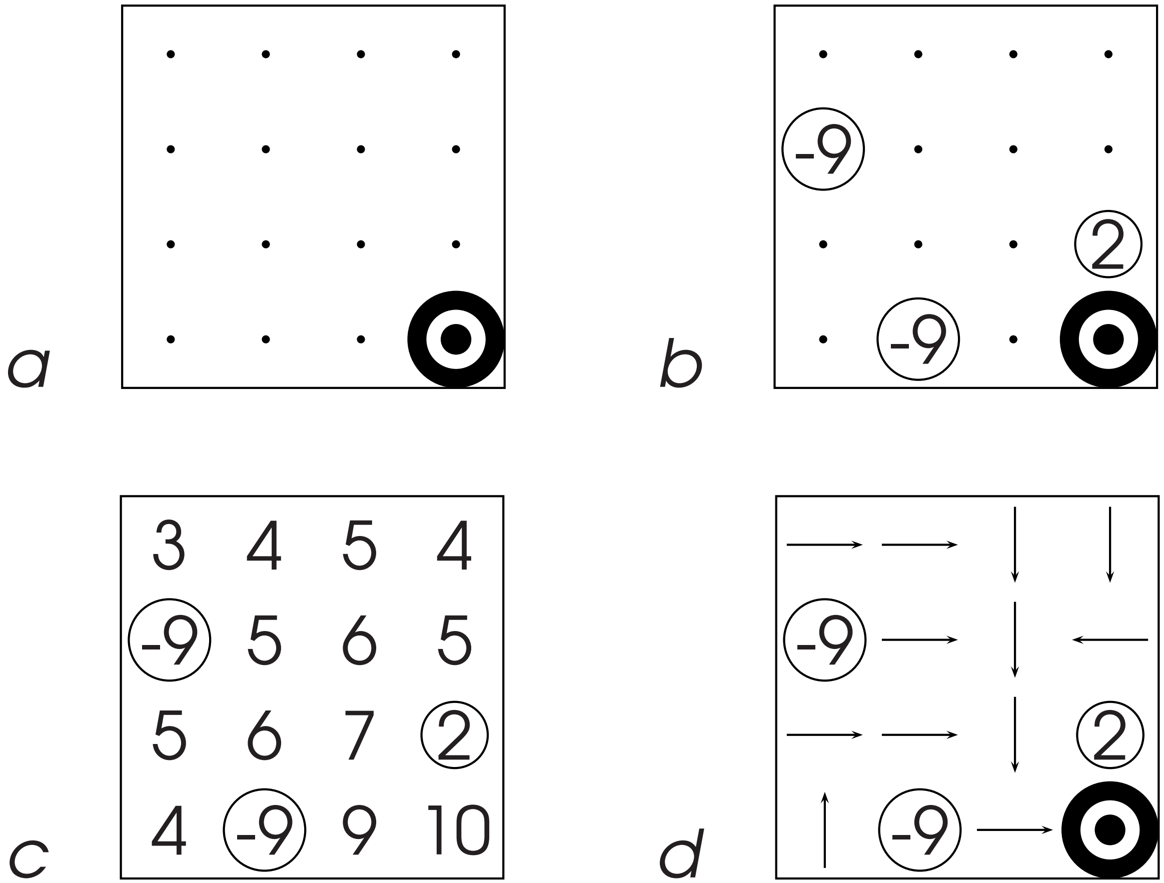
Reducing Task Dissimilarity

- Optimization possible if $\mathcal{O}$, $\mathcal{C}$, κ, and η known
- Alternative: ***uniform sampling*** of value space
  Complexity:

$$\Omega\left(\frac{(v_{\max} - v_{\min})^n}{\epsilon^n}\right) \text{ draws.}$$

- More informed search: ***randomized task perturbation*** (RTP)

RTP Action Transfer at Work



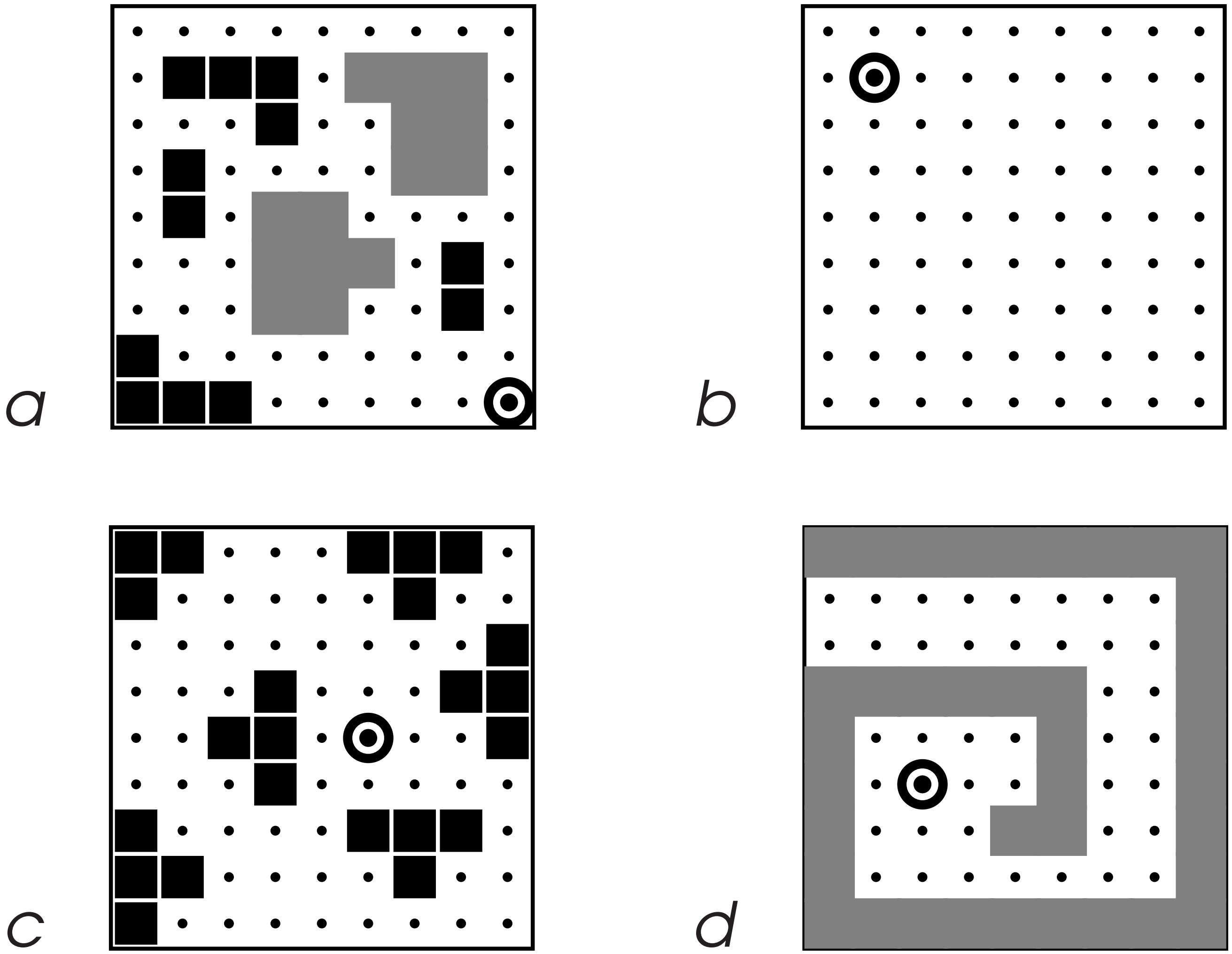
RTP Transfer in Pseudocode (Q-Learning)

new → Add each $s \in \mathcal{S}$ to $\mathcal{F}$ with probability ϕ
→ **foreach** $s \in \mathcal{F}$
→ **do** $\mathit{random\text{-}value} \leftarrow \mathrm{rand}(v_{\min}, v_{\max})$
→ $Q^{+}(s,a) \leftarrow \mathit{random\text{-}value}$ for all $a \in \mathcal{A}$
repeat $s \leftarrow$ current state, $a \leftarrow \pi(s)$
Take action a, observe reward r, state s'
$Q(s,a) \overset{\alpha}{\leftarrow} r + \gamma \max_{a' \in \mathcal{A}} Q(s', a')$
→ **if** $s \in \mathcal{S} \setminus \mathcal{F}$
then $Q^{+}(s,a) \overset{\alpha}{\leftarrow} r + \gamma \max_{a' \in \mathcal{A}} Q^{+}(s', a')$
until converged
→ $\mathcal{A}^{*} = \cup_{s \in \mathcal{S}} \{\arg\max_{a \in \mathcal{A}} Q(s,a)\}$
→ $\mathcal{A}^{+} = \cup_{s \in \mathcal{S} \setminus \mathcal{F}} \{\arg\max_{a \in \mathcal{A}} Q^{+}(s,a)\}$
→ **return** $\mathcal{A}^{*} \cup \mathcal{A}^{+}$

Relevance-weighted action selection

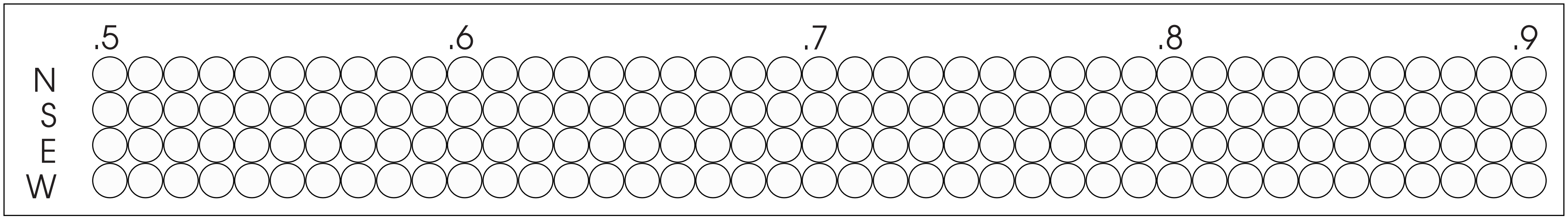
- $\text{RELEVANCE}(a) = |\{s \in \mathcal{S} : \pi^*(s) = a\}|/|\mathcal{S}|.$
- Choice probability on ***exploratory*** moves proportional to relevance
- Major performance gains

Empirical Results: Grid World

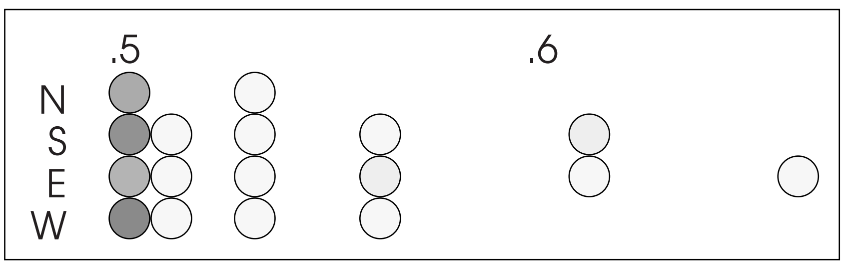


Action Sets

- Full:

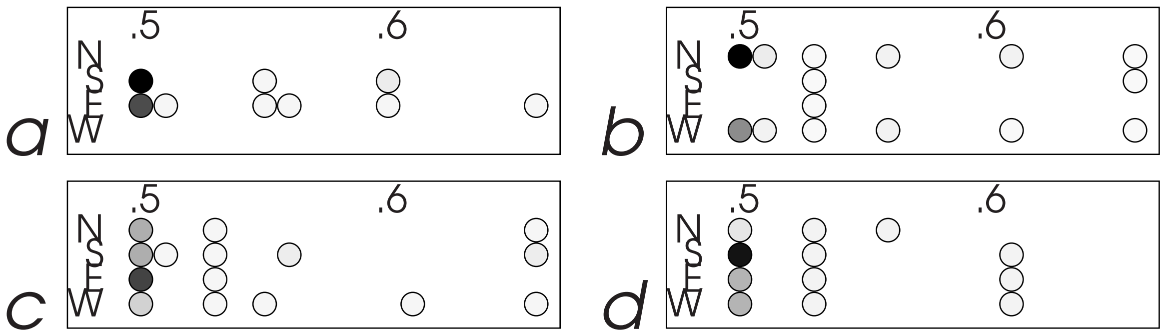


- Optimal (value iteration):

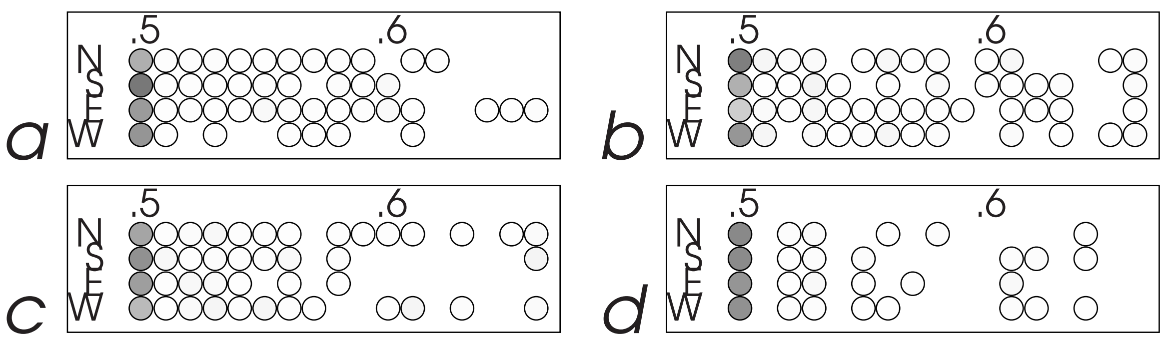


Action Sets, cont.

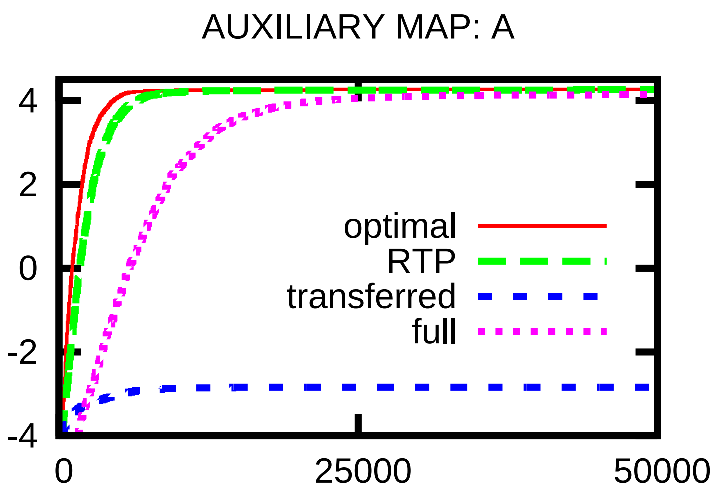
- Transferred:

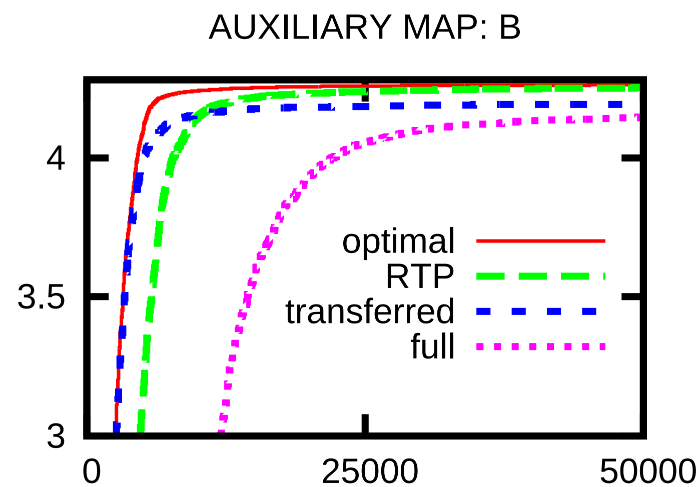


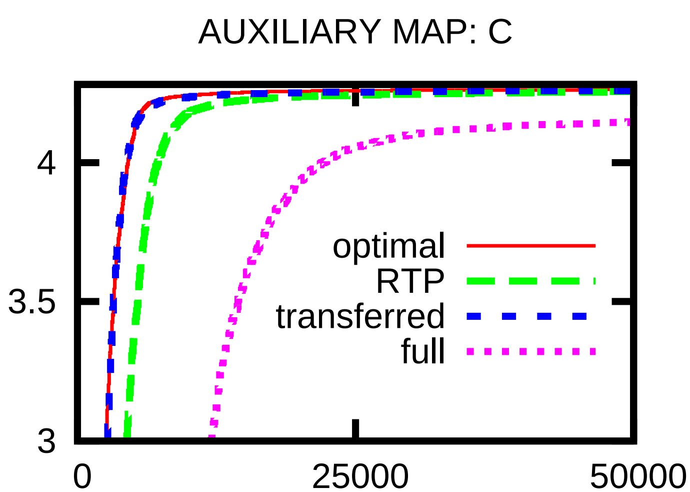
- RTP-transferred (sample run):

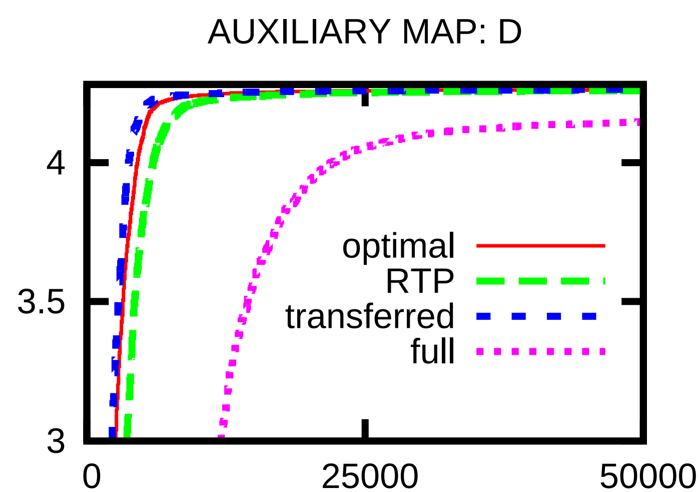


Performance









Related Work

- Transfer in MDP's:
  - hierarchical (Hauskrecht et al., 1998, Dietterich, 2000),
  - first-order (Boutilier et al., 2001),
  - factored (Guestrin et al., 2003)

- Limitation: reliance on description of similarities

- RTP: no guidance, robust to noise, focuses on ***actions***

Summary

Contributions:

- Theoretical abstraction for transfer learning
- Formal analysis + transfer quality guarantees
- Empirical validation

Future work:

- Combine with non-transfer approaches to action selection
- Enable fully continuous learning

References

(Boutilier et al., 2001) Boutilier, C., Reiter, R., and Price, B. (2001). Symbolic dynamic programming for first-order MDPs. In *Proc. 17th International Joint Conference on Artificial Intelligence (IJCAI-01)*, pages 690–697, Seattle, WA.

(Dietterich, 2000) Dietterich, T. G. (2000). Hierarchical reinforcement learning with the MAXQ value function decomposition. *Journal of Artificial Intelligence Research*, 13:227–303.

(Guestrin et al., 2003) Guestrin, C., Koller, D., Gearhart, C., and Kanodia, N. (2003). Generalizing plans to new environments in relational MDPs. In *Proc. 18th International Joint Conference on Artificial Intelligence (IJCAI-03)*.

(Hauskrecht et al., 1998) Hauskrecht, M., Meuleau, N., Kaelbling, L. P., Dean, T., and Boutilier, C. (1998). Hierarchical solution of Markov decision processes using macro-actions. In *Proc. Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI-98)*, pages 220–229.

Formalism for Related Tasks

Definition 1. *A **domain** is a quintuple* $\langle \mathcal{A}, \mathcal{C}, \mathcal{O}, \mathbf{t}, r \rangle$*, where* $\mathcal{A}$ *is a set of actions;* $\mathcal{C}$ *is a set of state classes;* $\mathcal{O}$ *is a set of action outcomes;* $\mathbf{t} : \mathcal{C} \times \mathcal{A} \longrightarrow \Pr(\mathcal{O})$ *is a transition function; and* $r : \mathcal{C} \times \mathcal{A} \longrightarrow \mathbb{R}$ *is a reward function.*

Definition 2. *A **task** within the domain* $\langle \mathcal{A}, \mathcal{C}, \mathcal{O}, \mathbf{t}, r \rangle$ *is a triple* $\langle \mathcal{S}, \kappa, \eta \rangle$*, where* $\mathcal{S}$ *is a set of states;* $\kappa : \mathcal{S} \longrightarrow \mathcal{C}$ *is a state classification function; and* $\eta : \mathcal{S} \times \mathcal{O} \longrightarrow \mathcal{S}$ *is a next-state function.*

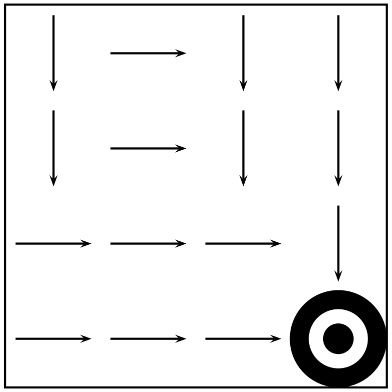
Definitions for Suboptimality Analysis

Definition 3. *The **outcome value vector** of state s in the task $\langle \mathcal{S}, \kappa, \eta \rangle$ within the domain $\langle \mathcal{A}, \mathcal{C}, \mathcal{O}, \mathbf{t}, r \rangle$ is the vector $[V^*(s_1) \quad V^*(s_2) \quad \ldots \quad V^*(s_{|\mathcal{O}|})]^{\mathrm{T}}$, where $V^* : \mathcal{S} \rightarrow \mathbb{R}$ is the optimal value function of the task, and each $s_i = \eta(s, o_i)$ is a successor state of s upon outcome $o_i \in \mathcal{O}$.*

Definition 4. *Let $U = \langle U_{c_1}, \ldots, U_{c_{|\mathcal{C}|}} \rangle$ and $\tilde{U} = \langle \tilde{U}_{c_1}, \ldots, \tilde{U}_{c_{|\mathcal{C}|}} \rangle$ be the OVV sets of the primary and auxiliary tasks, respectively. The **dissimilarity** of the primary and auxiliary tasks, denoted $\Delta(U, \tilde{U})$, is:*

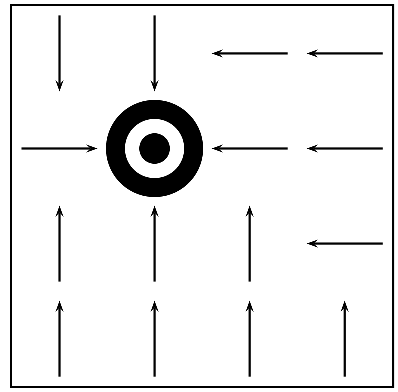
$$\Delta(U, \tilde{U}) \overset{def}{=} \max_{c \in \mathcal{C}} \max_{\mathbf{u} \in U_c} \left\{ \min_{\tilde{\mathbf{u}} \in \tilde{U}_c} ||\mathbf{u} - \tilde{\mathbf{u}}||_2 \right\}.$$

Detrimental Action Transfer



3	4	5	6
4	5	6	7
5	6	7	9
6	7	9	10

Auxiliary task



7	9	7	6
9	10	9	7
7	9	7	6
6	7	6	5

Primary task

OVV Sets

Auxiliary Task:

$[\,3\ 4\ 4\ 3\ 3\,]^{\mathrm{T}}$	$[\,4\ 5\ 5\ 3\ 4\,]^{\mathrm{T}}$	$[\,5\ 6\ 6\ 4\ 5\,]^{\mathrm{T}}$	$[\,6\ 7\ 6\ 5\ 6\,]^{\mathrm{T}}$
$[\,3\ 5\ 5\ 4\ 4\,]^{\mathrm{T}}$	$[\,4\ 6\ 6\ 4\ 5\,]^{\mathrm{T}}$	$[\,5\ 7\ 7\ 5\ 6\,]^{\mathrm{T}}$	$[\,6\ 9\ 7\ 6\ 7\,]^{\mathrm{T}}$
$[\,4\ 6\ 6\ 5\ 5\,]^{\mathrm{T}}$	$[\,5\ 7\ 7\ 5\ 6\,]^{\mathrm{T}}$	$[\,6\ 9\ 9\ 6\ 7\,]^{\mathrm{T}}$	$[\,7\ 10\ 9\ 7\ 9\,]^{\mathrm{T}}$
$[\,5\ 6\ 7\ 6\ 6\,]^{\mathrm{T}}$	$[\,6\ 7\ 9\ 6\ 7\,]^{\mathrm{T}}$	$[\,7\ 9\ 10\ 7\ 9\,]^{\mathrm{T}}$	$[\,10\ 10\ 10\ 10\ 10\,]^{\mathrm{T}}$

Primary Task:

$[\,7\ 9\ 9\ 7\ 7\,]^{\mathrm{T}}$	$[\,9\ 10\ 7\ 7\ 9\,]^{\mathrm{T}}$	$[\,7\ 9\ 6\ 9\ 7\,]^{\mathrm{T}}$	$[\,6\ 7\ 6\ 7\ 6\,]^{\mathrm{T}}$
$[\,7\ 7\ 10\ 9\ 9\,]^{\mathrm{T}}$	$[\,10\ 10\ 10\ 10\ 10\,]^{\mathrm{T}}$	$[\,7\ 7\ 7\ 10\ 9\,]^{\mathrm{T}}$	$[\,6\ 6\ 7\ 9\ 7\,]^{\mathrm{T}}$
$[\,9\ 6\ 9\ 7\ 7\,]^{\mathrm{T}}$	$[\,10\ 7\ 7\ 7\ 9\,]^{\mathrm{T}}$	$[\,9\ 6\ 6\ 9\ 7\,]^{\mathrm{T}}$	$[\,7\ 5\ 6\ 7\ 6\,]^{\mathrm{T}}$
$[\,7\ 6\ 7\ 6\ 6\,]^{\mathrm{T}}$	$[\,9\ 7\ 6\ 6\ 7\,]^{\mathrm{T}}$	$[\,7\ 6\ 5\ 7\ 6\,]^{\mathrm{T}}$	$[\,6\ 5\ 5\ 6\ 5\,]^{\mathrm{T}}$