RL 16: Multi-Objective Reinforcement Learning MORL

Michael Herrmann

University of Edinburgh, School of Informatics

11/03/2016

Multiobjective Reinforcement Learning

- RL is sequential decision making under uncertainties based on a scalar evaluation signal
- Defining a single reward signal is often the result of a complex design process. Typically several reward signals are available to the agent.
- How can an agent solve several tasks with different rewards simultaneously?
- Does not annihilate information by summing the rewards (which may not be comparable)
- Does the problem become easier or harder for multiple values?
- Robot example: Reach goal(s), avoid wear, keep track of position, avoid getting to close to a human, avoid running out of energy, help other agent that are met on the way ...
- Two main strategies:
  - Scalar combination of the reward signals (single policy)
  - Pareto optimisation (multiple policies)

Multiobjective Reinforcement Learning

- Scalar (e.g. weighted linear) combination of the reward signals (one policy for each combination)
- Threshold-based strategies (one policy for each threshold and ranking):
  - Rank goals by importance
  - Follow first one goal until reaching a threshold
  - continue with other goals
- Pareto optimisation (multiple policies)

MORL is different from Multiple goal RL

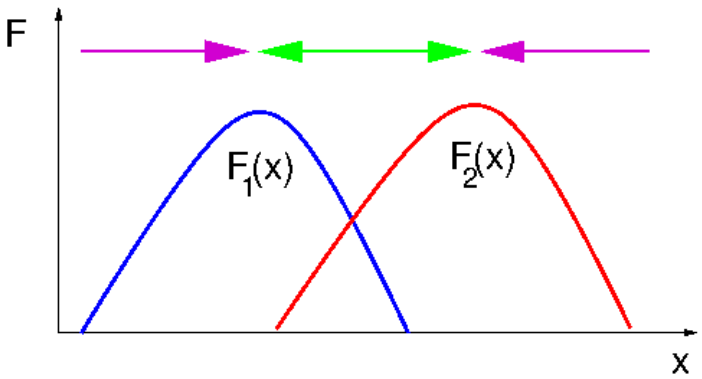
- Multiple goal states, but reward is given only at at specific target (which can change)
- Example: Grazing animals which must selected among several pastures, moving continually to the next best one.
- Our taxi problem is another example.
- At most one reward will be non-zero at any point in time

MORL instead aims compromises between the goals.

Crabbe, F. L. (2001). Multiple goal Q-learning: Issues and functions. In Proceedings of the international conference on computational intelligence for modelling control and automation (CIMCA). San Mateo: Morgan Kaufmann.

Multi-objective Optimisation

Example: A machine is characterised by power and torque. A machine is better if – at equal torque – its power is higher.



Combination of utility functions, e.g.

$f(x) = |f_1(x)|^\alpha + |f_2(x)|^\alpha$

$f(x) = \alpha f_1(x) + (1-\alpha) f_2(x)$

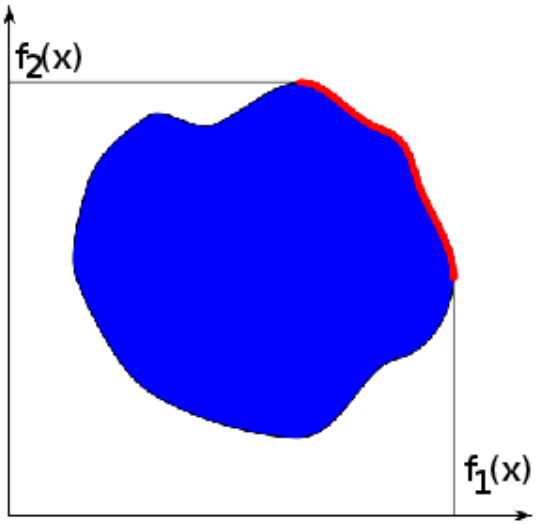
How to set α?

If α is not implied by the problem, any value in between the two maxima is equally good.

If a comparison between the two quantities is not possible, a set of solutions should be considered as optimal (Pareto-optimal).

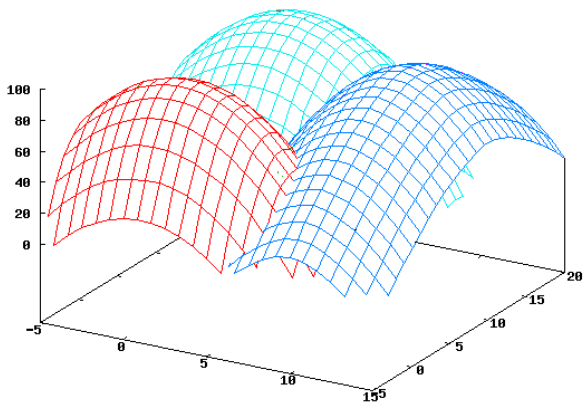
How to optimise one criterion without loosing on other criteria?

Multi-objective Optimisation

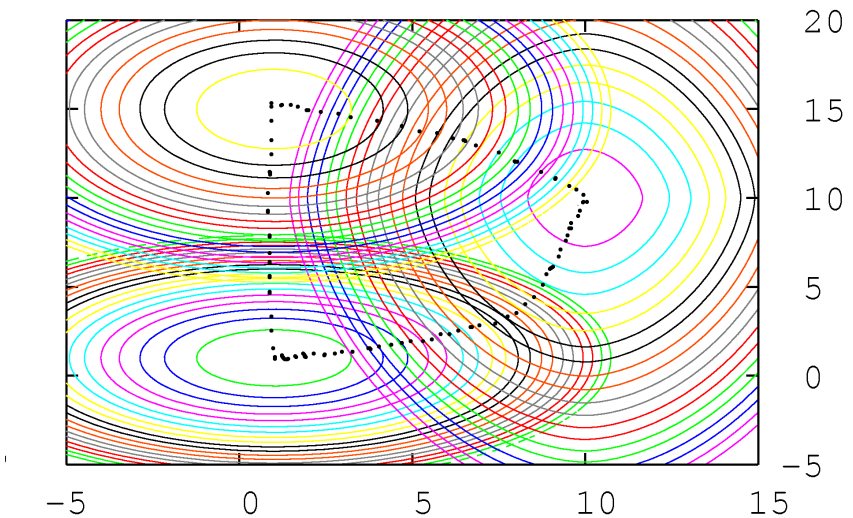


x^* is Pareto optimal for a class of fitness functions $\{f_i\}$ if there exists **no** $x \neq x^*$ with $f_i(x) \geq f_i(x^*)$ for all i

or, equivalently, x^* is not **dominated** by any other x : $\sim\exists x \succ x^*$
(more specifically $\sim\exists x \succ_{\{f_i\}} x^*$)



Example with three fitness functions



Same example: Pareto area spanned by maxima in a shape-dependent way

Multi-objective RL

Two strategies: ... and questions

- Scalarised approach: Find a single policy that optimises a combinations of the rewards
  - Which reward combination is preferable at which state?
  - Although a weighted sum of rewards might be an option, usually a weighted sum of values is considered to more relevant of the actions choice
- Pareto approach
  - Find multiple policies that cover the Pareto front: Sampling in a high-dimensional case
  - In principle, collective search required for sampling the Pareto set
  - What is a good approximation/representation of the Pareto front?

Relation between Scalarisation and Pareto

- A parametrised combination of multiple reward signals is used with different parameters in different runs to address different points along the Pareto front. The set of all solutions obtained in this way contains the Pareto front (e.g. in case of a non-connected Pareto front also non-Pareto optimal solutions may be found)
- The agent may change the parametrisation according to progress on each of the goals

Evaluation of MORL algorithms: $|HV \cap PF| \,/\, |HV|$

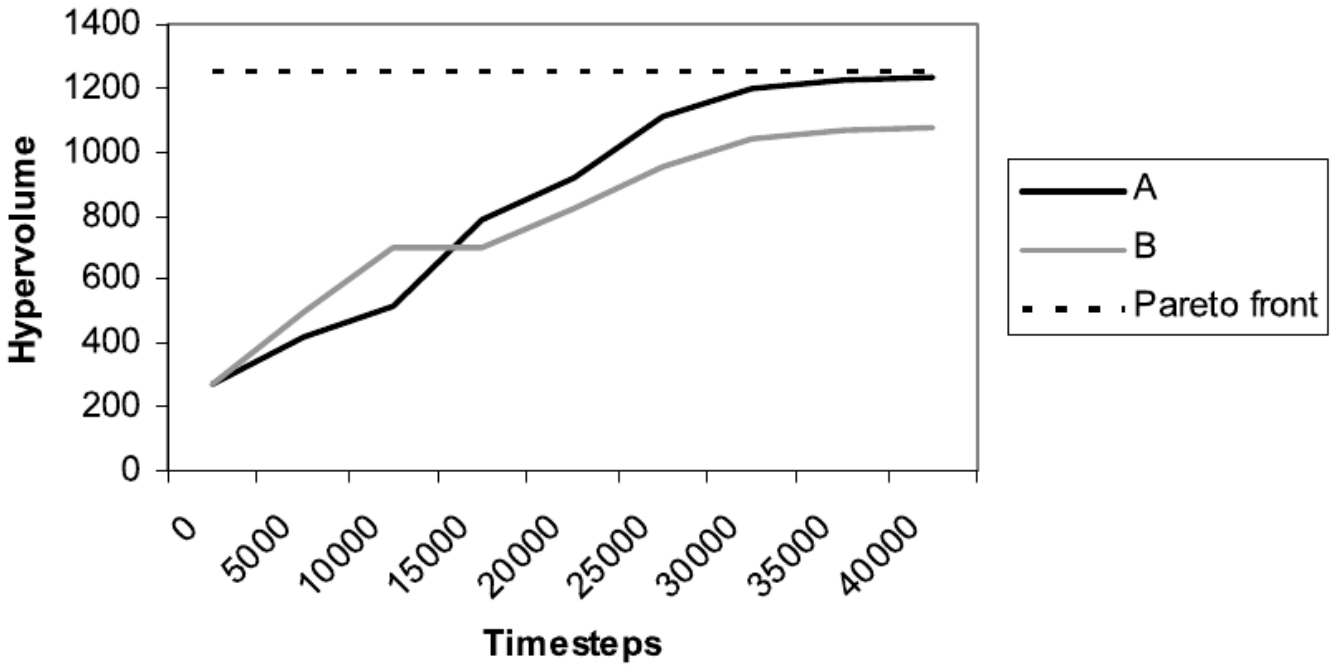


Fig. 5 A simulated example of the performance of multiple-policy algorithms: The relative performance of Algorithms *A* and *B* varies depending on the point in learning at which the results are compared—as discussed above the aim is to maximise the hypervolume metric, which is measured in test periods occurring at fixed intervals during learning. The hypervolume of the Pareto front provides a reference point for the absolute performance of the algorithms

Vamplew, P., Dazeley, R., Berry, A., Issabekov, R. and Dekker, E., 2011. Empirical evaluation methods for multiobjective reinforcement learning algorithms. Machine Learning, 84(1-2), pp.51-80.

Example: Resource Gathering

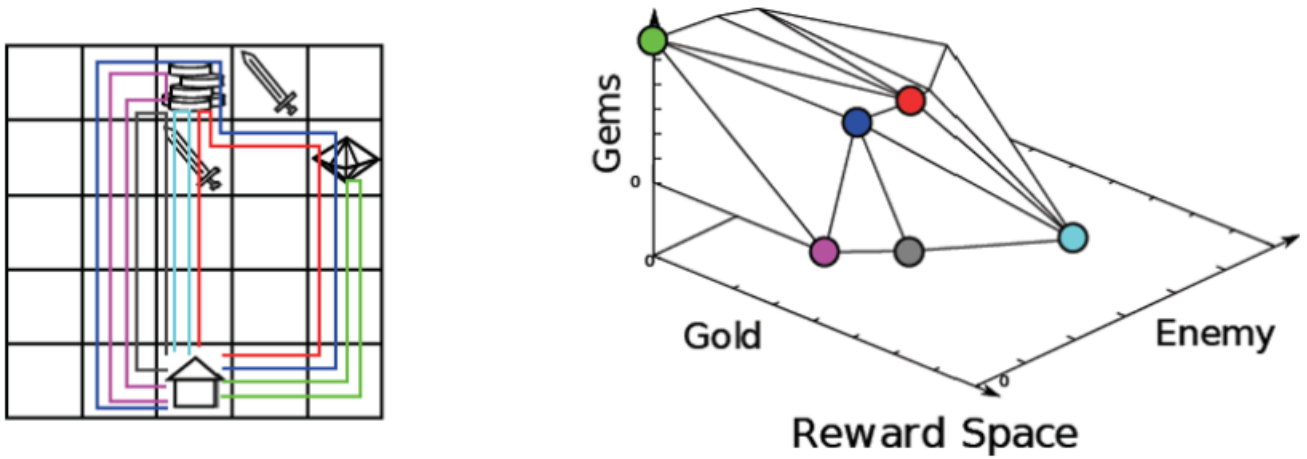


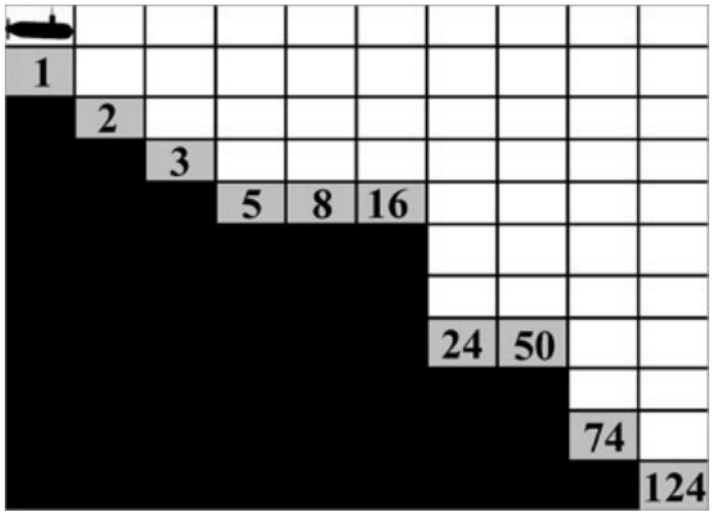
Fig. 16 The policies for Resource Gathering found by CHVI with a discounting factor of 0.9 (*left*), and the hull formed in objective-space by these policies (*right*) (Barrett and Narayanan 2008)

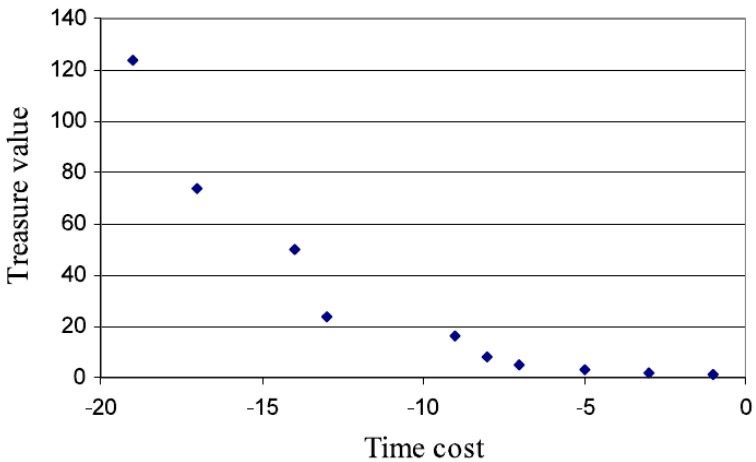
Rewards: [−1, 0, 0] in case of an enemy attack (occurs with 10% probability); [0, 1, 0] for returning home with gold but no gems; [0, 0, 1] for returning home with gems but no gold; [0, 1, 1] for returning home with both gold and gems. [see Vamplew ea]

Deep-sea treasure problem

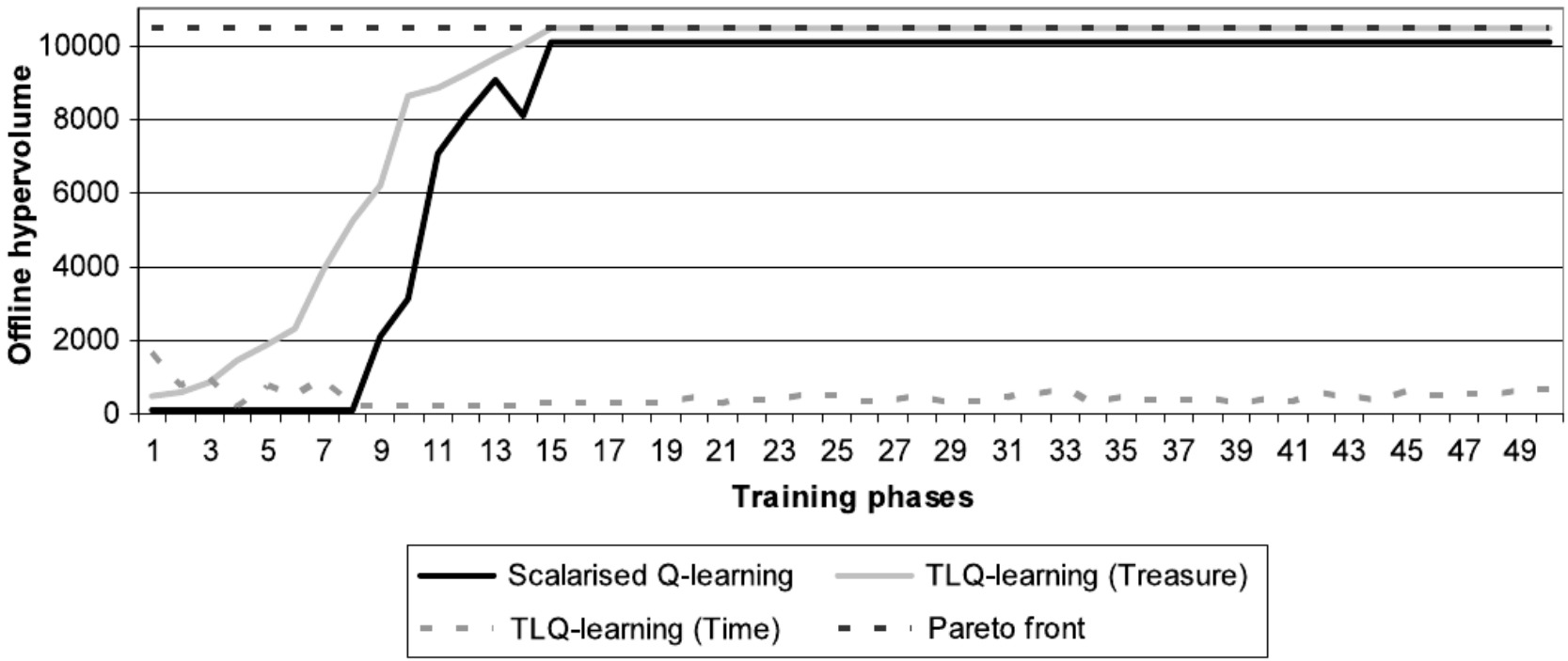
Black cells indicate the sea-floor; grey cells indicate a treasure location.
$R_1 = -1$ (per step)
$R_2 =$ treasure (as indicated)

- Scalarised algorithms may have problems finding all points on the PF
- Threshold-based algorithms may become trapped.





Deep-sea treasure problem: Results



Approaches to MORL

MORL Approaches		Basic Principle
Single-policy approaches	The weighted sum approach	A linear weighted sum of Q-values is computed as the synthetic objective function.
	The W-learning approach	Each objective has its own recommendation for action selection and the final decision is based on the objective with the largest value.
	The AHP approach	The analytic hierarchy process (AHP) is employed to derive a synthetic objective function.
	The ranking approach	"Partial policies" are used as the synthetic objective function.
	The geometric approach	A target set satisfying certain geometric conditions in multi-dimensional objective space is used as the synthetic objective function.
Multiple-policy approaches	The convex hull approach	Learn optimal value functions or policies for all linear preference settings in the objective space.
	The varying parameter approach	Performing any single-policy algorithm for multiple runs with different parameters, objective threshold values and orderings.

(C. Liu et al., 2013)

$\mathcal{W}$-learning

The Top-$\mathcal{Q}$ algorithm chooses simply

$$a_t = \max_i \mathcal{W}_i = \max_i \max_a \mathcal{Q}_i(s_t, a)$$

The result depends usually on the scales of the reward signals.

$\mathcal{W}$-learning: Define a principal value function ℓ and choose

$$a_\ell(t) = \max_a \mathcal{Q}_\ell(s(t), a)$$

Calculate $\mathcal{W}$-values by

$$\mathcal{W}_i = \max_a \mathcal{Q}_i(s(t), a) - \mathcal{Q}_i(s(t), a_\ell)$$

or (to avoid oscillations)

$$\begin{aligned} \mathcal{W}_i(s) &= (1-\alpha)\mathcal{W}_i(s) + \alpha P_i(s) \\ P_i(s) &= \max_a \mathcal{Q}_i(s, a) - \left(r_i + \gamma \max_b \mathcal{Q}_i\left(s', b\right)\right) \end{aligned}$$

set new $\ell = \arg\max_i \mathcal{W}_i$

Approaches to MORL

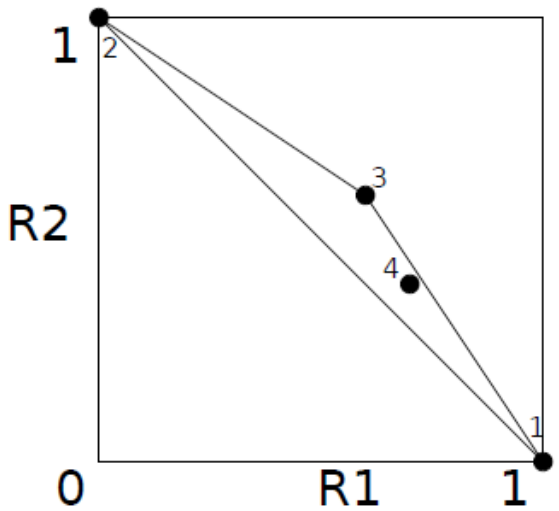
- AHP: Choose action a if is superior for L out of N objectives with a total improvement over the next best action of at least $\Delta\mathcal{Q}$. Combine L and $\Delta\mathcal{Q}$ using a fuzzy system.
- Ranking: Define an ordering of rewards, and check low-priority rewards only if decision is not possible by high-priority rewards.

Advantages of multi-policy approaches

- Agent remains flexible to decide about goals after learning
- Constraints can be expressed by rewards
- "being dominated by" denotes a partial order which is sufficient for many RL approaches
- Non-domination instead of (greedy) maximisation
- Exploration along and across the non-dominated front
- Use several agents (could be represented by the same robot)

Multiple policies

- Convex hull
  Barrett, L., & Narayanan, S. (2008). Learning all optimal policies with multiple criteria. In: 25th ICML, 41-47.
- Varying parameter approach:
  Finding a Nash-equilibrium of the returns
  C.R. Shelton (2001) Balancing multiple sources of reward in reinforcement learning. NIPS.



convex hull approach

MORL with Policy gradient

- Policy gradient techniques to approximate the Pareto frontier
- How can gradient information be derived from multi-objective sequential decision problems?
- Different MORL approaches based on MO policy gradient
  - radial
  - Pareto following
- see next three slides

Parisi, S., Pirotta, M., Smacchia, N., Bascetta, L., & Restelli, M. (2014) Policy gradient approaches for multi-objective sequential decision making. In: IJCNN, 2323-2330). IEEE.

Slides and source code at: http://home.dei.polimi.it/pirotta

Multi-Objective Policy Gradient (Parisi et al. 2014)

- Half Spaces
- Ascent Cone

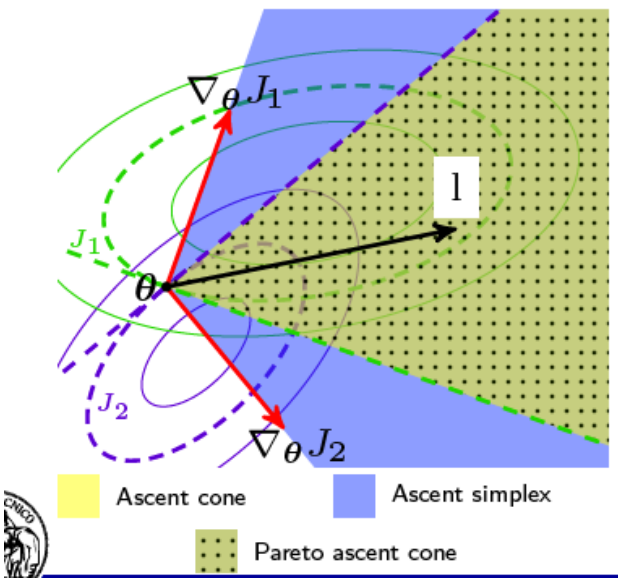
$$C(\boldsymbol{\theta}) = \{\boldsymbol{l} : \boldsymbol{l} \cdot \nabla_{\boldsymbol{\theta}} J_i(\boldsymbol{\theta}) \geq 0\}$$

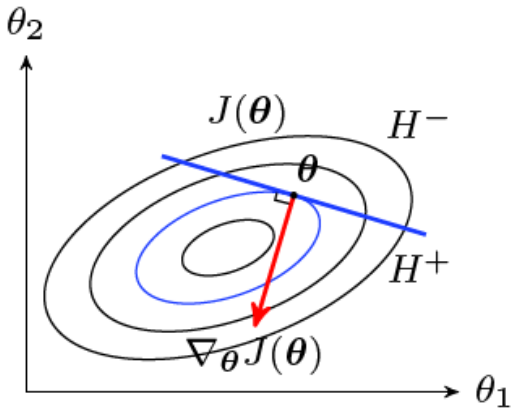
- Ascent Simplex

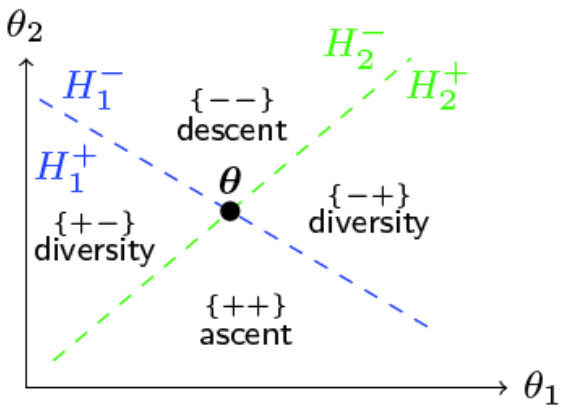
$$S(\boldsymbol{\lambda}, \boldsymbol{\theta}) = \sum_{i=1}^{q} \lambda_i \nabla_{\boldsymbol{\theta}} J_i(\boldsymbol{\theta})$$

- **Pareto–Ascent Cone**

$$S(\boldsymbol{\lambda}, \boldsymbol{\theta}) \cap C(\boldsymbol{\theta})$$





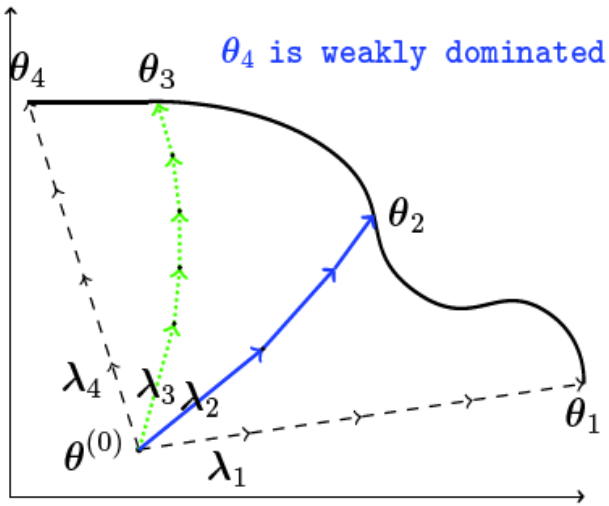


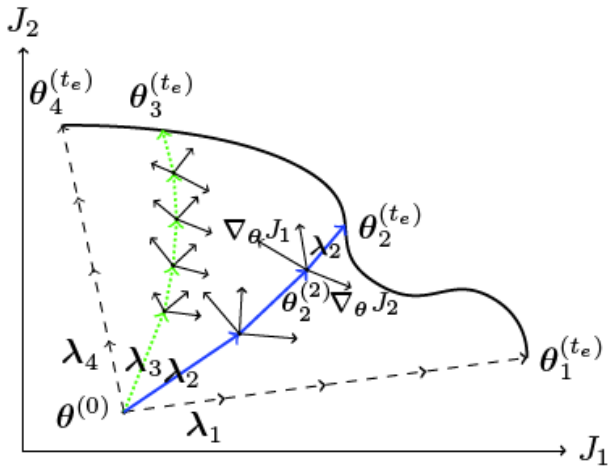
Null Pareto–Ascent Cone ⇒ (local) optimal solution

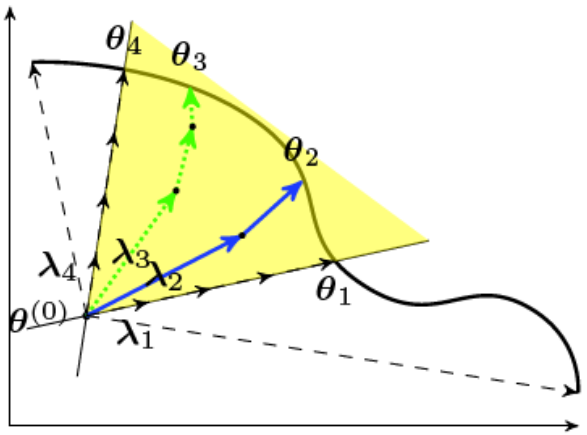
Radial Algorithm (Parisi et al. 2014)

Idea: p gradient ascent searches are performed, each one following a different, *uniformly spaced* direction in the **ascent simplex**

Problem: **weak optimal** solutions







Pareto Following Algorithm (Parisi et al. 2014)

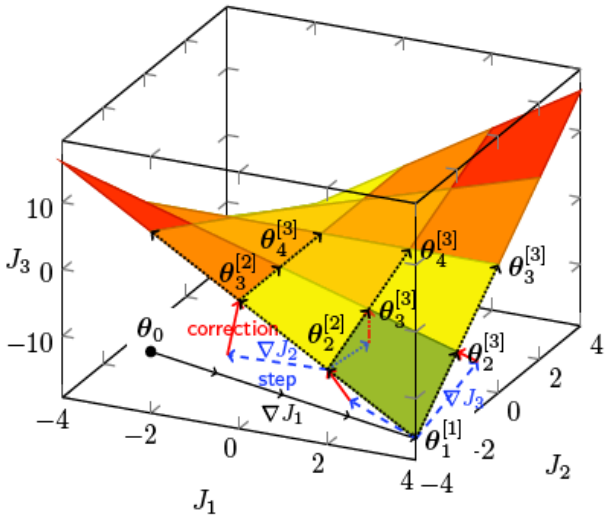
Phase 1: A solution on the Pareto frontier is reached by considering a single objective

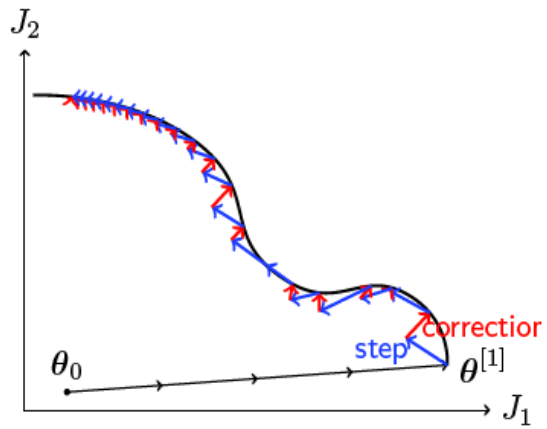
Phase 2: Exploration

- Improvement step: move the solution toward **one** objective at a time
- Correction step: improvement may lead the point outside the frontier. Correction **moves** the point again **on the frontier**

Problems:

- Can reach deterministic policies
- Need to **reintroduce stochasticity** (e.g., based on the entropy)
- Tuning of learning rate





Further Work

- Combination with IRL: Reconstructed reward is a second objective in addition to direct reward (A. Agarwal et al. JMLR 2014)
- Chebyshev scalarization (i.e. using an L_p-norm) can find all points in a non-convex Pareto set by introducing another parameter (A. Nove et al, 2013)
- Representation of the PF
  - function approximation
  - linear interpolation

Discussion

- Remain flexible
- Scalarisation is in some benchmarks a bit slower, as the simpler partial goals are more easily learnable.
- Applications, e.g.: Traffic control, Quality of medical service in mobile health care, robot control, network routing, grid computing.
- MARL: In Multi-Agent systems different agents may have different objectives. Different equilibria are possible, differently from the discussed approaches to MORL.

Acknowledgements & References

Liu, C., Xu, X., & Hu, D. (2013). Multiobjective reinforcement learning: A comprehensive overview. *IEEE TA Systems, Man, and Cybern.* **45**:3, 385-398.

Roijers, D. M., Vamplew, P., Whiteson, S., & Dazeley, R. (2014). A survey of multi-objective sequential decision-making. *J. Artific. Intellig. Res.* **48**, 67-113.

Parisi, S., Pirotta, M., Smacchia, N., Bascetta, L., & Restelli, M. (2014) Policy gradient approaches for multi-objective sequential decision making. In: IJCNN, 2323-2330). IEEE.

See also:

See also: http://umichrl.pbworks.com/w/page/7597585/Myths of Reinforcement Learning