**Multiobjective Reinforcement Learning: A Comprehensive Overview, 2015**

Chunming Liu, Xin Xu, Dewen Hu

National University of Defense Technology

IV. REPRESENTATIVE APPROACHES TO MORL

Weighted Sum Approach [28]

[28] J. Karlsson, “Learning to solve multiple goals,” Ph.D. dissertation, Dept. Comput. Sci., Univ. Rochester, Rochester, NY, USA, 1997.

GM-Sarsa(0) [11]

[11] N. Sprague and D. Ballard, “Multiple-goal reinforcement learning with modular Sarsa(0),” in Proc. 18th Int. Joint Conf. Artif. Intell., 2003, pp. 1445–1447.

1. Learning Approach

winner-take-all methods for MORL were studied in [32]

One simple method to compute W values is called Top-Q [11] [32]

[32] M. Humphrys, “Action selection methods using reinforcement learning,” in From Animals to Animats 4, P. Maes, M. Mataric, J.-A. Meyer, J. Pollack, and S. W. Wilson, Eds. Cambridge, MA, USA: MIT Press, 1996, pp. 134–144.

AHP Approach [34]

the MORL method proposed in [34]

[34] Y. Zhao, Q. W. Chen, and W. L. Hu, “Multi-objective reinforcement learning algorithm for MOSDMP in unknown environment,” in Proc. 8th World Congr. Int. Control Autom., 2010, pp. 3190–3194.

Ranking Approach

The idea of using ordinal relations in optimal decision making was studied in the early research work by Mitten [35] and Sobel [36].

an ordering of multiple objectives was established in [37] for MORL where threshold values were speciﬁed for some objectives in order to put the constraints on the objectives.

1. L. G. Mitten, “Composition principles for synthesis of optimum multistage processes,” Oper. Res., vol. 12, pp. 610–619, Aug. 1964.
2. M. J. Sobel, “Ordinal dynamic programming,” Manage. Sci., vol. 21, pp. 967–975, May 1975.

[37] Z. Gabor, Z. Kalmar, and C. Szepesvari, “Multi-criteria reinforcement learning,” in Proc. 15th Int. Conf. Mach. Learn., 1998?2004, pp. 197–205.

Geometric Approach

Mannor and Shimkin [40] proposed a geometric approach to MORL.

[40] S. Mannor and N. Shimkin, “A geometric approach to multi-criterion reinforcement learning,” J. Mach. Learn. Res., vol. 5, pp. 325–360, Jan. 2004.

Convex Hull Approach

Barrett and Narayanan [42] presented a multiple-policy algorithm to MORL, which can simultaneously learn optimal policies for all linear preference assignments in the objective space.

Barrett and Narayanan [42] presented this algorithm and proved that the solution can ﬁnd the optimal policy for any linear preference function.

[42] L. Barrett and S. Narayanan, “Learning all optimal policies with multiple criteria,” in Proc. 25th Int. Conf. Mach. Learn., 2008, pp. 41–47.

Varying Parameter Approach

For example, as indicated in [12] and [27], scalarized Q-learning can be used in a multiple-policy manner by executing repeated runs of the Q-learning algorithm using different parameters.

Shelton [43] applied policy gradient methods and the idea of varying parameters to the MORL domain.

1. A. Castelletti, G. Corani, A. Rizzolli, R. Soncinie-Sessa, and E. Weber, “Reinforcement learning in the operational management of a water system,” in Proc. IFAC Workshop Model. Control Environ. Issues, Yokohama, Japan, 2002, pp. 325–330.

[43] C. R. Shelton, “Balancing multiple sources of reward in reinforcement learning,” in Proc. Adv. Neural Inf. Process. Syst., 2000, pp. 1082–1088.

Addition

In addition to these approaches, there are some other MORL approaches proposed recently [44], [45].

multiobjective ﬁtted Q-iteration (FQI) [54] that can ﬁnd control policies for all the linear combinations of preferences assigned to the objectives in a single training procedure.

1. A. Castelletti, F. Pianosi, and M. Restelli, “Tree-based ﬁtted Q-iteration for multi-objective Markov decision problems,” in Proc. Int. Joint Conf. Neural Netw., 2012, pp. 1–8.
2. H. L. Liu and Q. H. Wu, “Multi-objective optimization by reinforcement learning,” in Proc. IEEE Congr. Evol. Comput., 2010, pp. 1–8.
3. A. Castelletti, F. Pianosi, and M. Restelli, “Multi-objective ﬁtted Q-iteration: Pareto frontier approximation in one single run,” in Proc. IEEE Int. Conf. Netw. Sens. Control, Delft, The Netherlands, 2011, pp. 260–265.

Note on references:

**Learning to Solve Multiple Goals, 1997, PhD thesis**

Jonas Karlsson

The University of Rochester

The greatest mass approximation algorithm, p41, p21

The nearest neighbor strategy, p42, p21

Some application of greatest mass approximation, maybe contains some works related to MORL, p22

**Multiple-Goal Reinforcement Learning with Modular Sarsa(0), 2003**

Nathan Sprague, Dana Ballard

The University of Rochester

New algorithm, GM-Sarsa(0), for finding approximate solutions to multiple-goal reinforcement learning problems that are modeled as composite Markov decision processes.

1 Introduction

Naive approach,

create a state space, all of the information that is relevant to each sub-task,

agent learn in joint space,

receive reward when any of the sub-goals are accomplished,

the curse of dimensionality

1. Learning,

To train one learning module to handle each of the sub-goals

Existing algorithms learn component polices, highly sub-optimal,

The component modules are forced to share control

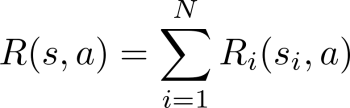
Fix, replacing the Q-Learning with the closely related Sarsa(0)

2 The problem formalized

Consider the problem of discovering a joint policy for a set of N MDP’s qt_temp

Distinct state space, common action space, required to execute the same action on each time step, to find the optimal policy for this composite MDP,

Optimal composite policy, maximizes summed discounted reward across the component MDPs,

Composite reward function, 

Traditional Q-learning, the composite state space grow exponentially

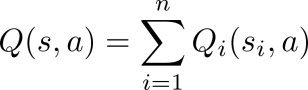
3 Modular Q-Learning

Humphrys, 1996; Karlsson, 1997

A separate learning module, each component MDP, each module is a Q-learning

Q-values indicates the degree of preference for different actions,

Several different ways to select a compromise action using these Q-values,

Karlsson, also above PhD thesis, greatest mass, , GM-Q, greatest mass Q-learning

Humphrys, raises the objection that the action with highest sum may not be particularly good for any of the modules, with the result that no module is able to reach its goal,

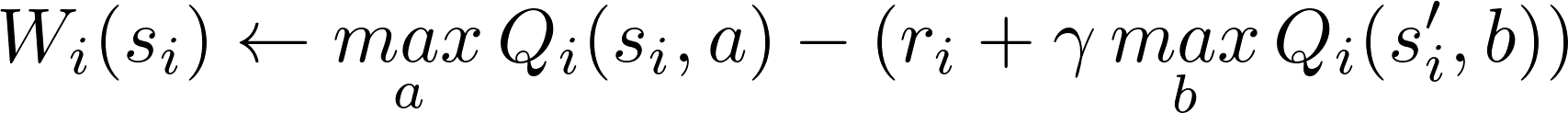
Improvement, winner-take-all alternatives, constrain the chosen action to be optimal for at least one module,

For a given state s each of the N modules promotes its own action with a value qt_temp, the module with the largest W value is then allow to execute its preferred action.

Top-Q, generate the W-values, 

Drawback, this method suffers from the drawback that the module with the highest Q-value may have no preference over what action is chosen, while another module stands to lose a great deal if its action is not selected. The method sometimes exhibits reasonable performance, but this is strongly dependent on the structure of the reward functions.

A better alternative, W-Learning, learn W-values,



4 Modular Sarsa(0)

Sarsa(0) with greatest mass, GM-Sarsa(0)

5 Convergence

Nothing.

6 Examples

Test on T-juncture and game of gathering stationary food while avoiding a predator in a 5x5 grid.

Algorithms, GM-Sarsa(0), Negotiated-W, GM-Q, Top-Q

7 Related work

Much previous work, large reinforcement learning problems, hierarchically structured goals and sub-goals,

Related to work in the area of multi-agent reinforcement learning, surveyed in [Stone and Veloso, 2000],

Others, neither of these approaches address the reinforcement learning problem, the curse of dimensionality

8 Conclusion

A method for learning approximately optimal policies for a certain class of composite Markov decision process,

这篇文章对多目标收益和Q值的处理是简单的，利用价值不大，但是根据目标不同，将状态分别表示成独立的变量，然后合成考虑的思想可以借鉴，可以将多维状态变量合成一个状态向量，甚至再加上时序的考虑，组成一个状态矩阵，作为一个时刻的状态输入。

**Action Selection methods using Reinforcement Learning, 1996, PhD thesis**

Mark Humphrys, the University of Cambridge

Abstract

Action selection problem, run-time choice between conflicting and heterogenous goals,

Considering a decentralised model of mind with internal tension and competition between selfish behaviors,

W-learning,

Others

Neural network implementation, generalization,

break the state-action space up into one network per action,

Each takes a vector input x, produces a float point output Q\_a(x),

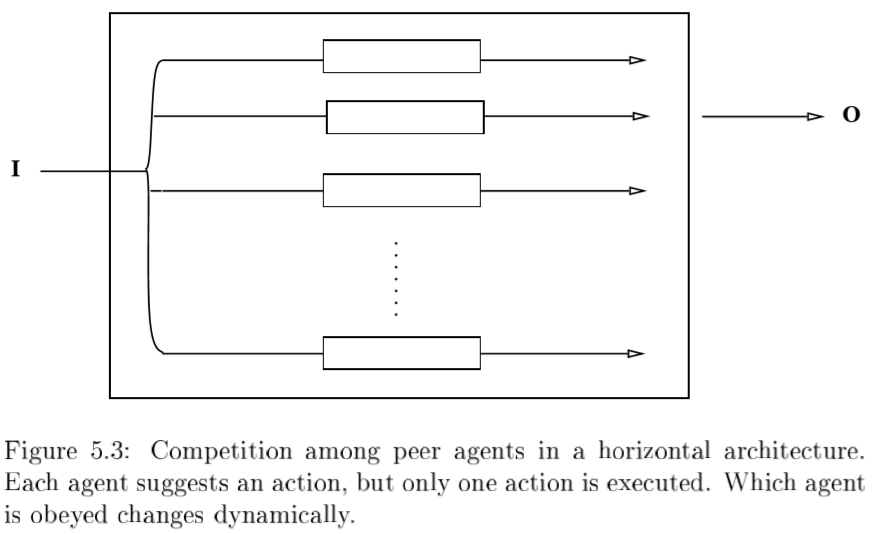
[Lin, 1992]

Decentralised model of mind

How agents might organize themselves sensibly in the absence of a global reward,

Looking for more biologically-plausible action selection,

The starting point for this exploration of decentralised minds is Rodney Brooks[Brooks, 1986],



Lin’s Hierarchical Q-learning, the switch is complex and what is sent is simple,

???if we can make the switch simple, and what is sent more complex,

Can the agents organize themselves off a dumb switch,

The weight W

From an agent’s point of view,

The basic model is that when the creature observes state x, the agents suggest their actions with numeric strengths or Weights W\_i(x) (W-values),

这一观点有点像，一个人具有多个分离的人格，每个人格有一个明确的目标，根据环境的不同，激发不同人格层面的决策。

The switch becomes a simple gate letting through the highest W-value,



W表示根据当前条件，一个agent所维护的目标的重要程度，可视为当前环境下一个神经元的兴奋程度，引入到多目标问题，多个agent维护不同的目标，假设，其中一个agent A强调随机探索周边的环境，另一个agent B强调在当前环境下选择最优的改进方向，现在考虑这样的问题，如果当前环境中有明确的最优改进的方向，那么B兴奋，A被抑制，W\_B > W\_A，

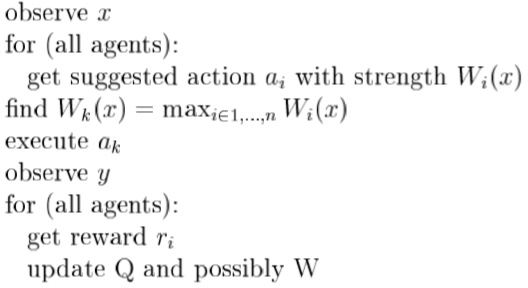
如果当前环境中没有明显的最优改进方向，那么A兴奋，B被抑制。原文中举例为，有两个agent，A目标是躲避捕食者，B的目标是漫游寻找食物，当环境中有捕食者的时候，A兴奋，B抑制，反之亦然。

W = Drive Strength

Drives model, ethology (Hull’s work in the 1940s),

‘drive strength’ or ‘importance variable’ [Tyrrell, 1993], equivalent to the W-value,

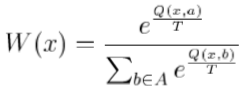
RL, build up value functions for actions, an agent not only knows ‘what’ it wants to do, it also knows ‘how much’ it wants to do it,



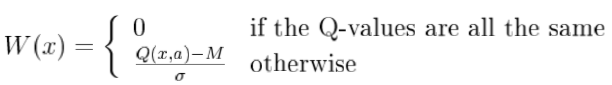
Static W = Q,

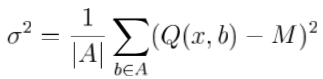
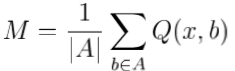


Static W = importance,

, or,,

Static W = standard deviation,

,

,

Dynamic(learned) W-values,

**Multi-objective Reinforcement Learning Algorithm for MOSDMP in Unknown Environment, 2010**

Yun Zhao, Qingwei Chen, Weili Hu,

Department of Automation, University of Science and Technology

Abstract,

MORL, multi-objective sequential decision making problems(MOSDMP), unknown environment,

Salient characters,

1. 引入决策者的客观偏好来引导学习方向
2. 定义了一种新的基于模糊推理系统，比较多目标下的动作决策
3. 实现快速学习

Introduction,

Incommensurability and confliction, MOSDMP and SOSDMP,

Incommensurability, there is no uniform measure standard for all objectives,

Weighted sum approach,

geometric approach, multi-criteria RL, Mannor and Shimkin, binary zero-sum stochastic games,

MDQL, distributed Q-learning for multi-objective optimization problems, cost much resources,

New model, under MOMDP,

improve AHP(analytic hierarchy process) for quantifying DM(Decision Maker) preference information to guide learning direction,

A new ‘comprehensive goodness’ of action, selection of the best action, fuzzy inference systems,

Action-selection mechanism, value-function updating method,

Multiple Objectives Markov Decision Process,

D.J.White, MDP --> MOMDP,

自从MOMDP提出后，产生了很多解决这一问题的算法，但是这些技术没有处理不确定环境下的问题,

Obviously, in general MOMDP, a policy which can maximize all of the expected optimization values dose not exist.

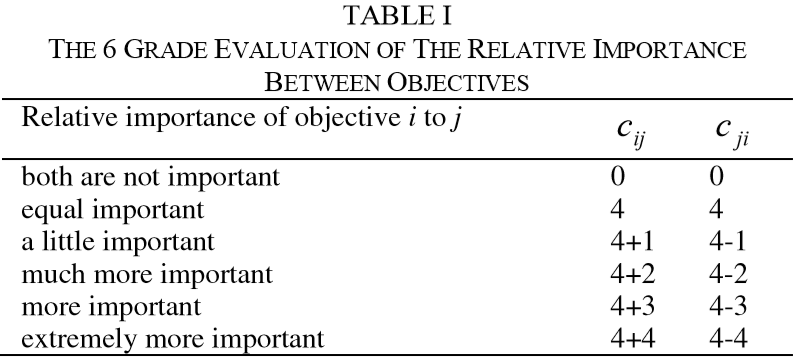
Here we consider a specialization in which decision maker has different preference on each objective.

利用偏好来定义动作选择的标准

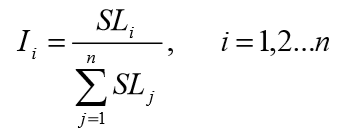
Preference Information Quantifying,

本文的AHP方法是文献[11]的改进，改进后的方法可以处理目标数变化的问题。

The relative importance degree between pair of objectives is divided into L grades,



The relative importance matrix of objectives qt_temp,

The importance factor ,

Multi-Objective Reinforcement Learning Algorithm Incorporating Preference Information,

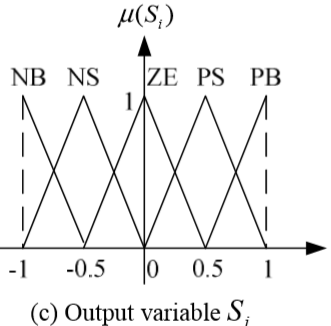
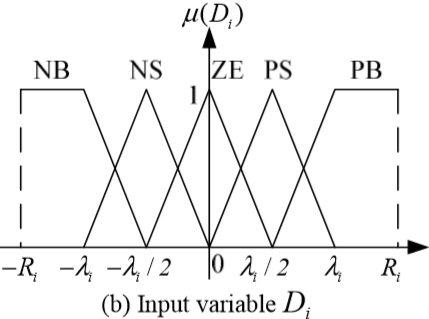
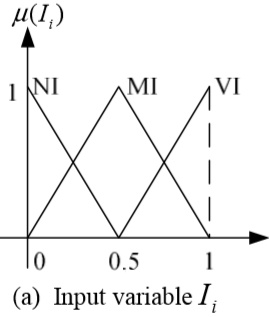
Relative goodness,

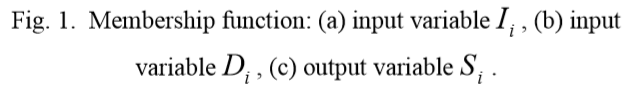
A Mamdani fuzzy inference system, for each objective,

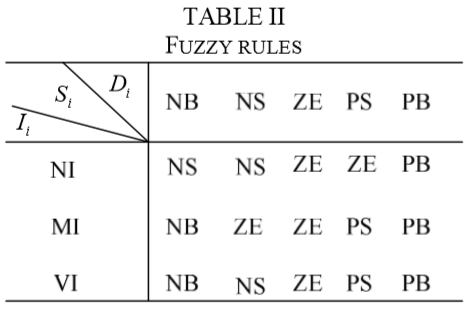
Input of fuzzy system, comparing two action decision ap and aq for objective i, input the importance factor qt_temp and the improvement ,

Output of fuzzy system, the ‘relative goodness’ qt_temp of ap to aq on objective i,

Both of inputs and output adopt triangular membership function,

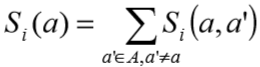






Absolute Goodness,

The absolute goodness of action a on objective i,

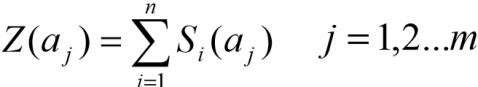


Commensurable decision matrix,



Comprehensive Goodness,

The comprehensive goodness of action aj is,



Action Selection Mechanism,

A new action selection mechanism based on epsilon-Greedy strategy borrowing ‘comprehensive goodness’ of actions,

Action Value Function Updating,

价值函数的更新与普通Q-learning相同，但是因为单步收益是一个n维向量，因此Q值需要针对n个目标分别更新各自的值。

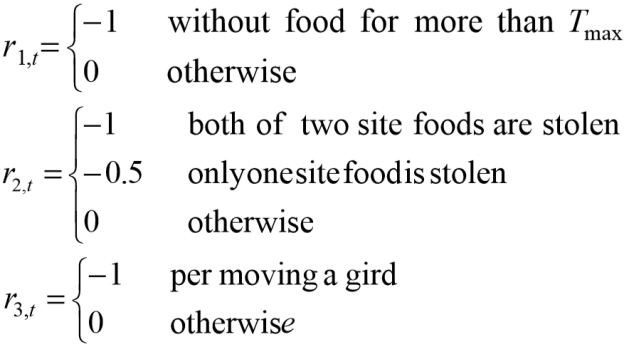
Experimental Results,

Example, Buridan’ass problem, MOMDP, conflicting goals,

3 costs: hunger, food-stolen, walking,

Actions: up, down, left, right and stay,

Reward:

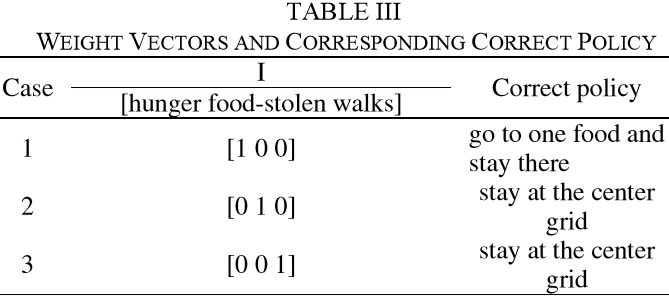


Experimental Setup,

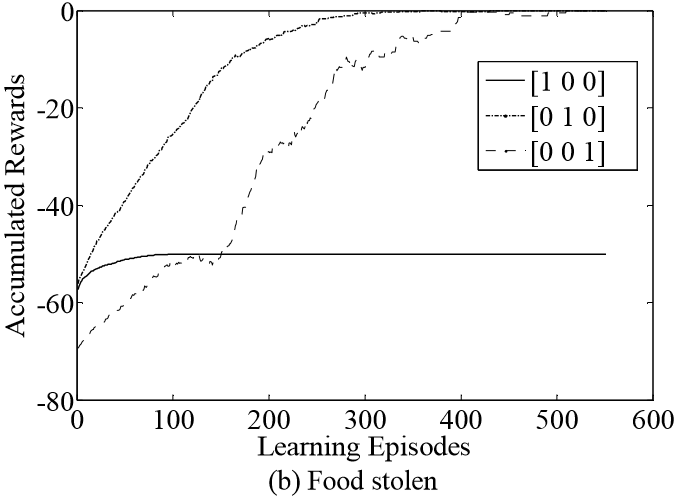
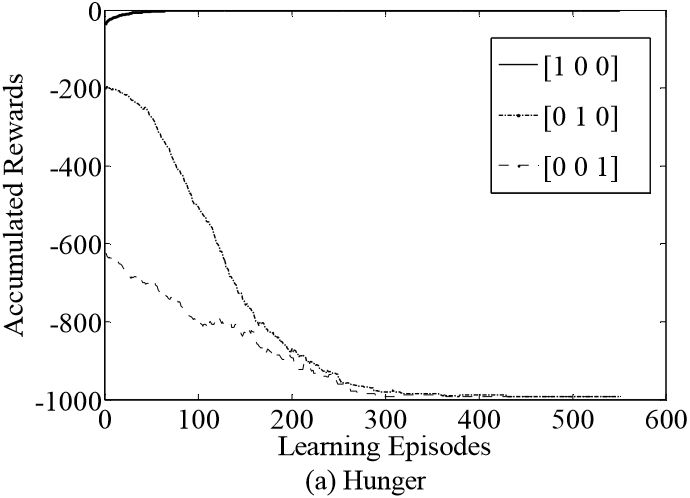
Each preference case, 600 episodes, maximum time steps 1000,

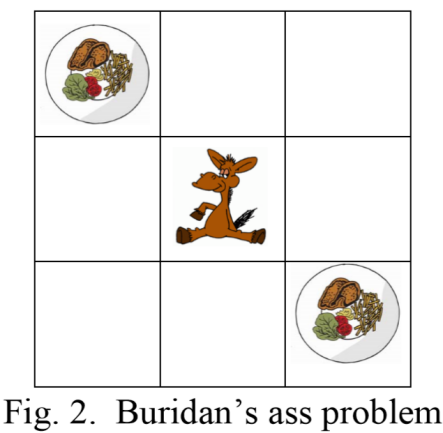
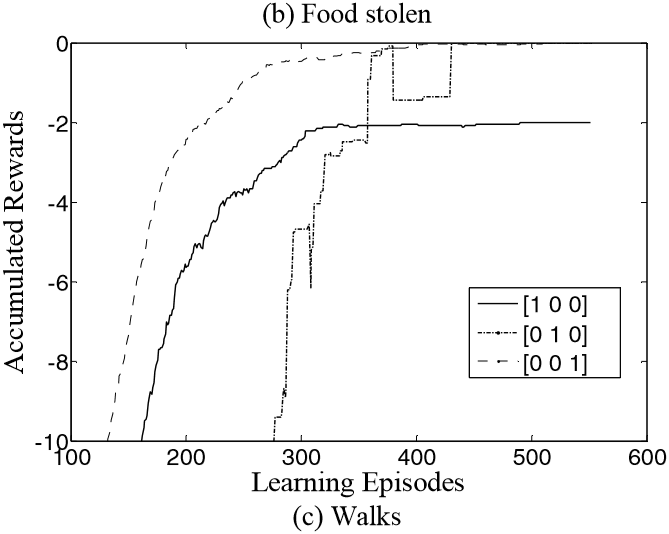
Experimental Results,

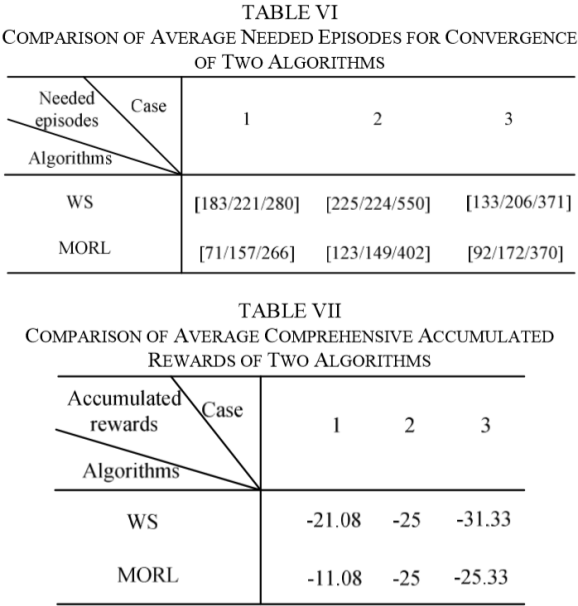
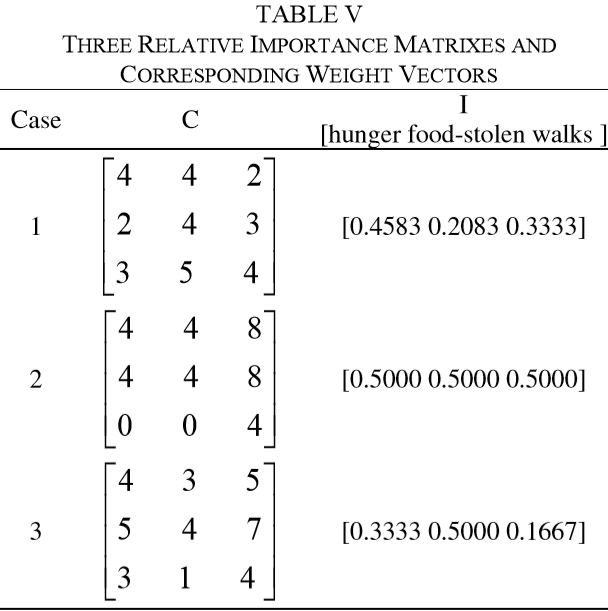
1. ability to learn correct policy, 3 extreme cases,



Results:







总评，方法普通，加入模糊系统用于偏好量化算是个亮点，其他的都是常用的。

**Multi-criteria Reinforcement Learning, 2004**

Zoltan Gabor, Zsolt Kalmar and Csaba Szcpcsvari

Associative Computing Ltd.(联想电脑公司?!)

Abstract,

Multi-criteria sequential decision making problems, vector-valued evaluation, fixed total ordering

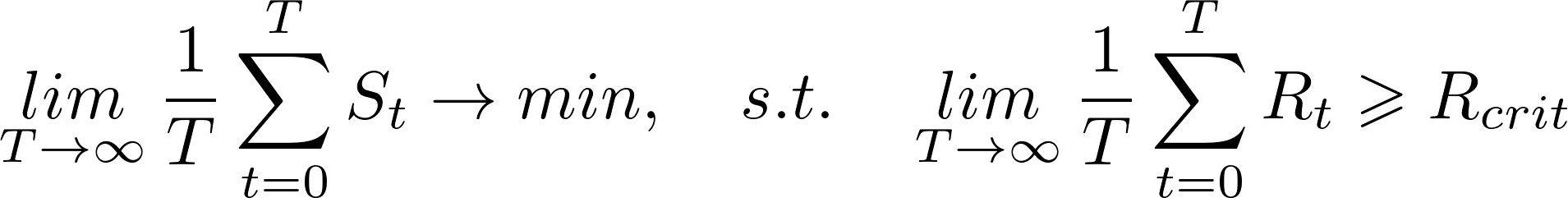
Optimality conditions of stationary policies, Bellman optimality equation, problem the evaluation of policies can be computed for the criteria independently of each other,

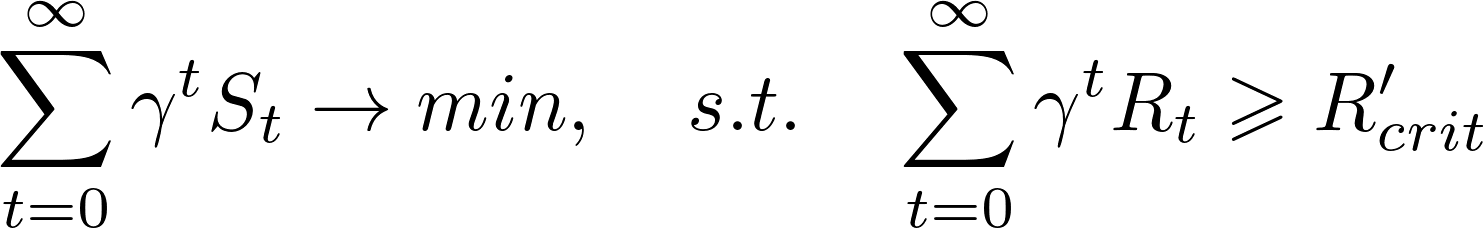
Special care for analysis, the topology introduced by pointwise convergence, the order-topology introduced by the preference order, in general incompatible,

Introduction,

The dilemma of Leibniz’s ass, eat to alive, avoid food stolen,

The watchmen’s compromise, to minimize the number of dishes stolen per unit time such that the ass manages to stay alive:





Another compromise, weighted sum,

Vector-valued immediate reinforcement, long-term reinforcement, evaluation of policies,

Comparison of policies, compare any pairs of policies, transitive and reflexive comparison operator,

Ordinal multi-criteria decision problems, [Mitten 1964], [Sobel 1975], partial policies,

先比较主要目标，再比较次要目标，默认目标是有重要度的排序,

Another comparison approach, weighted criterion,

作者认为Pareto最优性通常用于研究特定形式的优化策略的存在性条件。

Earliest result for dynamic vector-valued models, [Brown & Strauch 1965],

Abstract ordinal dynamic programming,

ADP(abstract dynamic programming), given as a 5-tuple (R, X, A, A~, Q),

Computer Simulations,

在井字棋上做的测试，两个目标函数，第一个目标函数表示对弈输赢结果，第二个目标函数表示对弈持续的步数。最优情况是用最少的步数赢得对弈，首先保证结果是赢，然后最小化对弈的步数。用了两种算法测试，一种只考虑单目标，ARTDP，另一种是它的上述双目标版本MC-ARTDP，结果显示双目标的结构有利于寻优。

评述，本文是基于ADP抽象动态规划对多目标规划问题进行建模求解，对多目标的偏序性质，采用了分层比较的方式，确定主要目标，比较主要目标后再比较次一级的目标。方法上没有什么新意，在2004年的历史条件下，应用强化学习解决这一问题算是一个创新，同时文章两点在于给出了ADP数学概念，对强化学习的求解过程给出了简单的证明，这在当时算是比较前沿的工作。

**A Geometric Approach to Multi-Criterion Reinforcement Learning, 2004**

Shie Mannor, Nahum Shimkin,

MIT, Israel Institute of Technology,

Abstract,

Reinforcement learning problem, controlled Markov environment with multiple objective functions of the long-term average reward,

Environment unknown initially, furthermore can be affected but can not be predicted,

Capture this situation, stochastic game model, an adversary whose policy is arbitrary and unknown, vector-valued reward function,

A desired target set is specified in the vector reward space, the objective of the learning is to approach the target set,

Learning algorithm, use multiple reinforcement learning algorithms for the standard scalar reward problem, which are combined using the geometric insight from the theory of approachability for vector-valued stochastic games,

Stochastic games with average reward constraints,

Constrained Markov decision process.

Introduction,

Address the problem of RL in a dynamic environment,

Agent’s goals are formulated in terms of multiple objective functions,

Each objective function corresponding to a long-term average reward functional,

Stochastic Games, a flexible model of dynamic conflict situation,

Existence of state-independent value and stationary optimal policies,

Stochastic games provide a natural generalization of the single-controller Markov decision problem to the multi-agent setting,

Conclusion,

这篇文章很长，最后得到的结果只是寻找一个满足要求的解，并非求解Pareto前沿，但是这篇文章的思路是与众不同的。首先给出预设目标在目标空间中的一个目标几何体，然后算法的目标就是在未知动态环境下，找到几何体内的点。作者建立了一套理论，这套理论主要基于三块基础：可达性理论(the theory of approachability)，强化学习以及随机博弈(stochastic game)。由于文章并非实现学习寻找Pareto最优解集，因此暂时没有仔细研读，后续应该对这篇有价值的、想法与众不同的文章进行研读。

**Learning All Optimal Policies with Multiple Criteria, 2008**

Leon Barrett, Srini Narayanan

Berkeley,

Abstract,

An algorithm for learning in the presence of multiple criteria,

Learn optimal policies for all linear preference assignments over the multiple reward criteria at once,

Back up the set of expected rewards that are maximal for set of liner preferences (given by a weight vector, w),

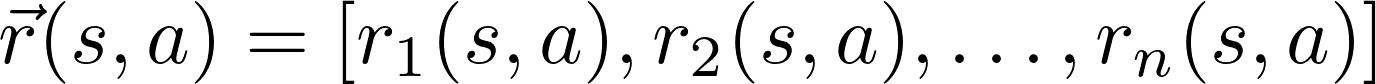
The solution reduces to the standard value iteration algorithm for a specific weight vector, w,

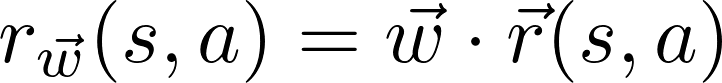
Introduction,

The resulting policies depend heavily on the preferences over these rewards, and they may change swiftly as preferences vary.

Present both algorithm for the general case of learning all optimal policies under all assignments of linear priorities for the reward components, and a proof showing the correctness of algorithm.

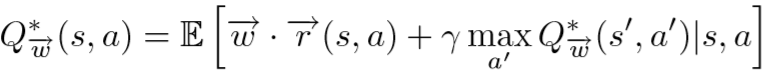
Explanation and Motivating Example,

A reward vector, ,

For every fixed weight vector w, 

Our method learns the set of optimal policies for all w at the same time,

For a fixed priority scheme (fixed w),

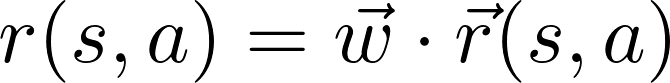


In the general case, weight vector w is unknown,

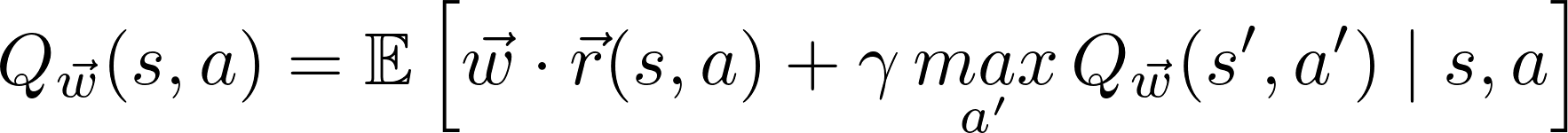
The extrema of the set of Q-values vectors is the same as the convex hull of the Q-value vectors,

Convex Hull Value Iteration,

Approach: Convex Hulls,

Given some qt_temp, the resulting reward, ,

Recurrence for optimal Q-values,



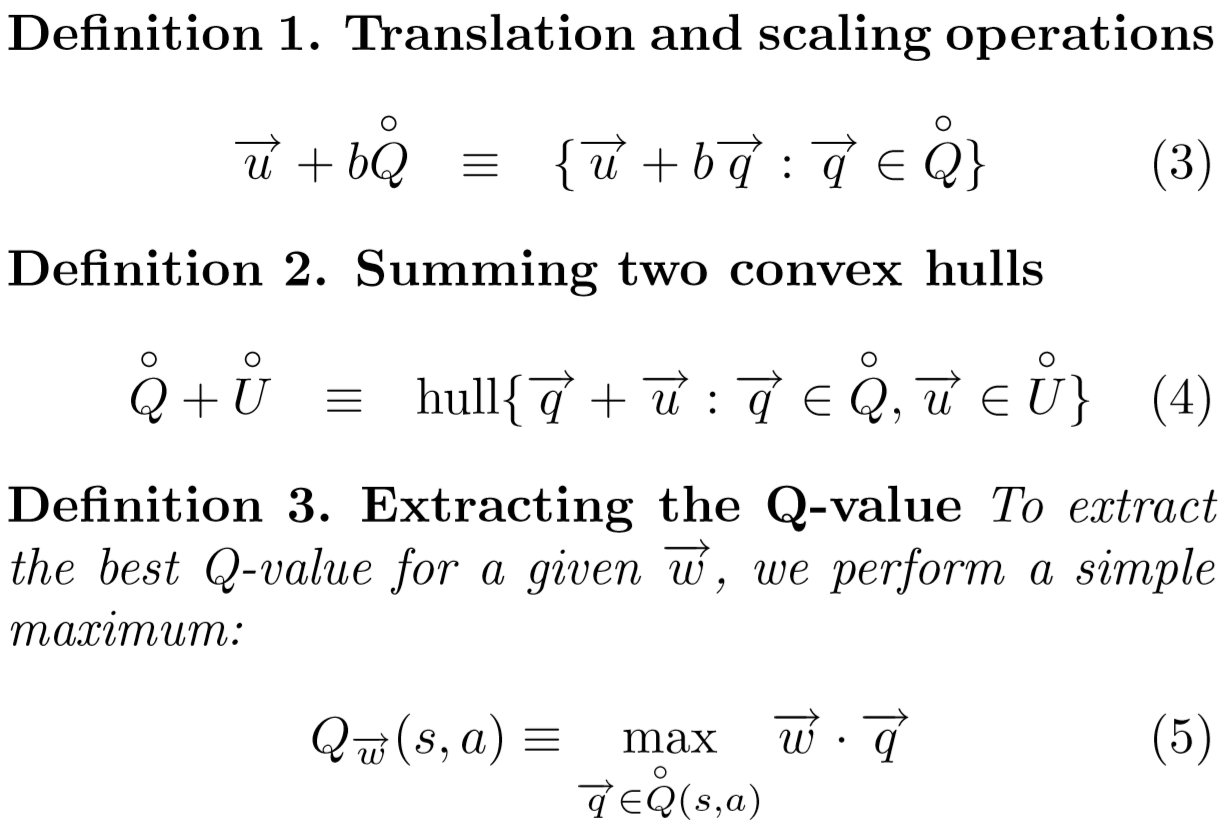
For a fixed qt_temp, only one such qt_temp can be optimal,

A set of any qt_temps that are maximal for some qt_temp, this set of Q-values is the convex hull of the Q-values,

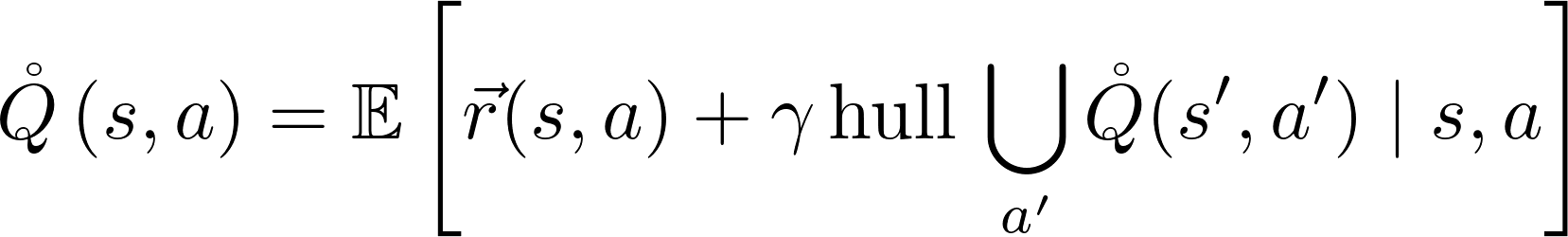
To use standard convex hull operations to pare down the set of points and rise to a proposition,



, the vertices of the convex hull of possible Q-value vectors for taking action a at state s,



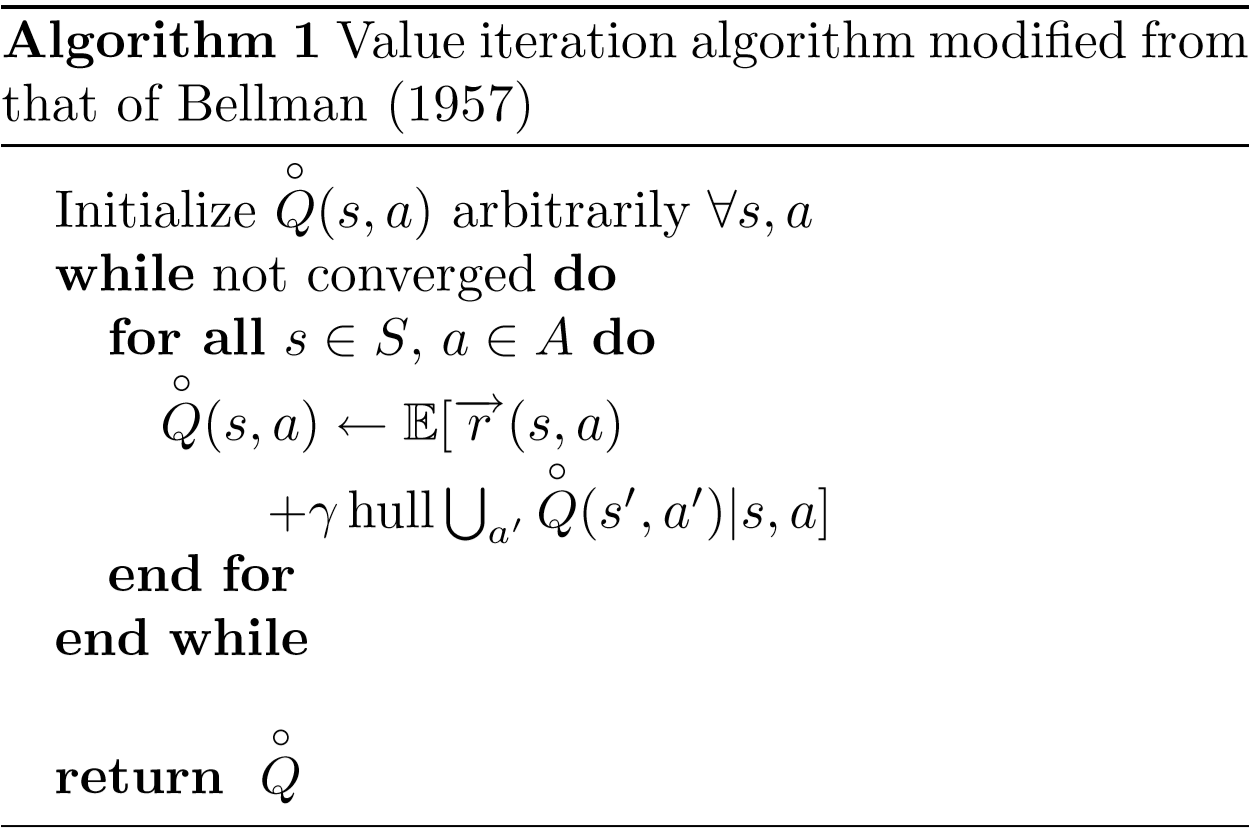
Convex Hull Value Iteration Algorithm,



Instead of repeatedly backing up maximal expected rewards, we back up the set of expected rewards that are maximal for some qt_temp,

The expectation over hulls is the natural equivalent of an expectation of maxima,

In the usual way, into the scalings and sums we have already defined,



Related Work,

A fixed ordering between different rewards, [Gabor et al., 1998),

Formulating sum of the discounted total rewards for multiple reward types, [Feinberg & Schwartz, 1995],

Decompose the reward function into multiple components which are learned independently(with a single policy) [Russell & Zimdars, 2003],

Above, preferences, fixed and time-invariant,

A slightly more flexible formulation, by geometric method, learn to attain a expected particular region of objective space, [Mannor & Shimkin, 2004],

Formulate the multiple reward RL problem as this paper do, [Natarajan & Tadepalli, 2005],

Relation to POMDPs,

POMDP, partially observable Markov decision process,

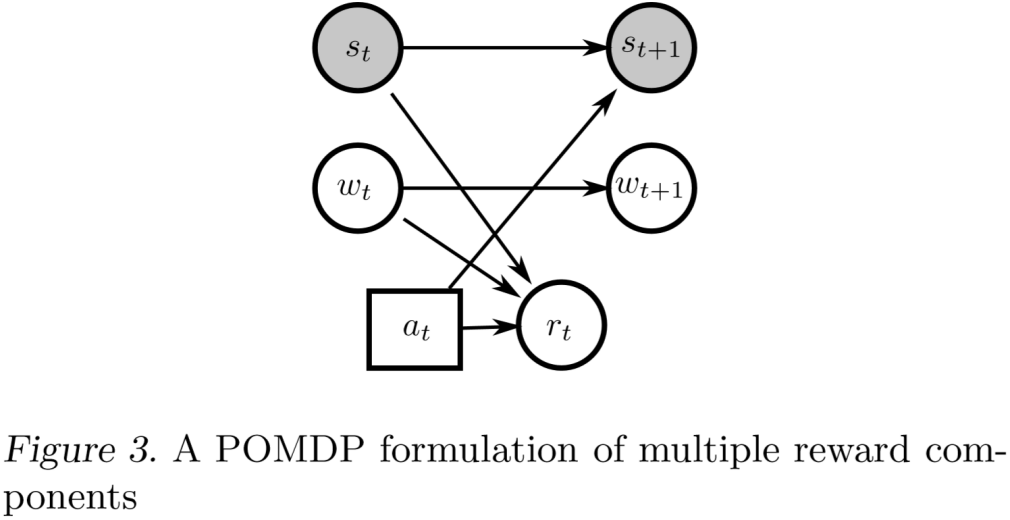
In a POMDP, a model of both observed and unobserved variables,

Use Bayesian reasoning to infer a joint distribution over the hidden variables,

Choose optimal action based on observed state and the continuous beliefs, [Kaelbling et al.1998],

In figure 3 shows a POMDP,

这部分讨论了作者解决的问题与部分可观测马尔可夫决策问题之间的关系。在图示的情况下，奖励r可以看作权重w的多项式，如果权重的分布是时不变的，那么分布就是稳定的，那么这种问题就是个线性加权的问题，权重分布是先验的，这时候，一对POMDP的最大超平面算法(maximum-hyperplane)与奖励部分的凸包操作相关。作者认为这与他们依赖于凸包的算法是相似的，他们的问题同样可以形式化成一个POMDP问题，这将有助于推广他们的算法，但是还要看实际效果而行。



Complexity,

4 convex hull operations,

n, the number of points on the hull, in the limit, this number converges to the number of optimal policies in the environment,

这里是对几个凸包操作复杂度的介绍，就不记录了。

下面作者还对降低复杂度的方式做了讨论。

最简单的是严格限制奖励向量的维度。

当不得不处理高维问题时，可以对我们优化的权重向量施加一些约束来降低复杂度。

由于Q值迭代更新，所以会进行大量的相似的凸包运算，因此对这些凸包中点记录一些附加信息，可以在下一次迭代计算的时候，利用上一次记录的信息简化计算。

Example Application: Resource Gathering,

Tested on a resource-collecting problem,

H, home base, the goal is to gather resources and take them back to the home base,

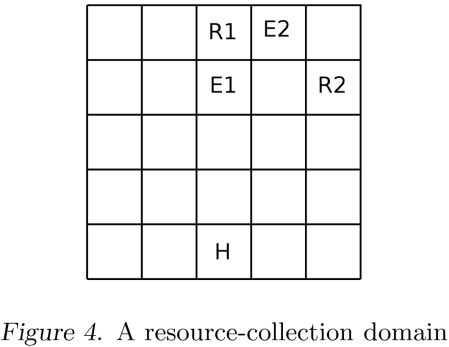
R1, R2, resource 1 and resource 2, can carry both of them at the same time,

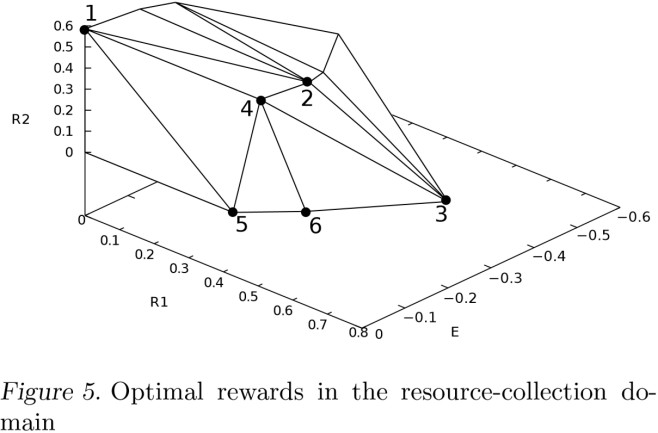
Reward, when get resource back to home, get reward for each of the resource,

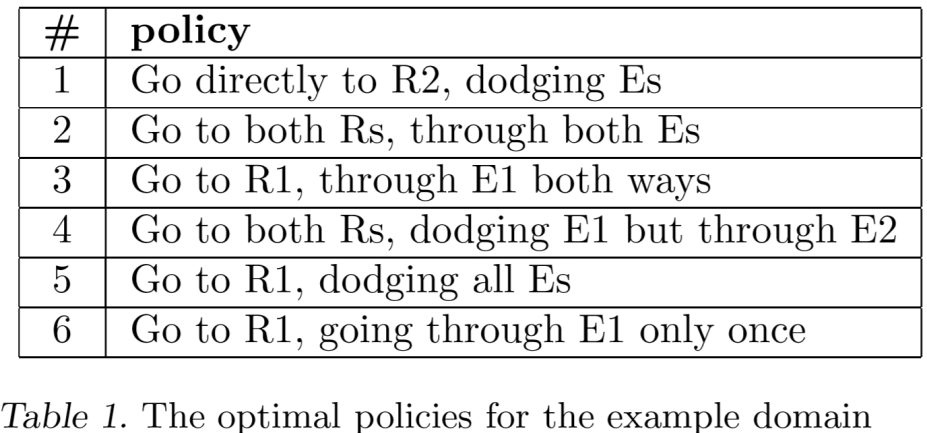
E1, E2, enemy space 1 and 2, with a 10% probability the agent will be attacked, receiving a penalty, resetting to the home space and losing all ti carries,

Reward space, ,

Penalty, (-1, 0, 0), R1, (0, 1, 0), R2, (0, 0, 1), gamma = 0.9,







Extensions and Current Work,

For temporal difference learning algorithm,

Inverse reinforcement learning problem, to use observed behavior to infer weights from a user,

Examine reward at different time scales, find what policies are optimal for a whole range of discounting rates,

**Multi-Objective Optimisation by Reinforcement Learning, 2010,**

H.L. Liao, Q.H. Wu,(Senior Member, IEEE)

University of Liverpool,Huilian Liao,<https://www.researchgate.net/profile/Huilian_Liao>

Abstract,

MORL, particular high-dimensional space,

Search on individual dimension in a high-dimensional space via a path selected by an estimated path value,

Path value, estimated by weighting the state values on the selected path,

State values, the potentiality of finding a better solution if search on the paths and memorize the quality of previously visited states,

Immediate rewards, comparing the objective vector of current state with those of the Pareto optimal solutions found previously,

Pareto optimal solutions, elite list,

Compared with a promising multi-objective evolutionary algorithm (MOEA/D) based on decomposition on 4 benchmark functions,

Introduction,

Group search optimiser with multiple producers, [2][3],

As the dimensionality of problems increases, the performance of the evolutionary algorithms, including genetic algorithms and particle swarm optimisers, degrades with respect to the range of the Pareto fronts,

Kamioka Takumi et al. [13] have applied RL algorithms in solving multi-objective problems by updating a separate state value function for each objective,

Mariano et al. [14][15] have applied distributed Q-learning, which assigns a family of agents to each objective, for finding the solution to multi-objective optimisation problems,

The method has only been tested on low-dimensional functions,

Path values, the difference of MORL and RL,

Path values play an important role in the evaluation of state values and action selection,

Problems Formulation,

主要写了多目标优化和强化学习的概念。

用一句话总结了多目标优化，挺有意思：

Multi-objective optimisation problems aim to find a family of solutions that is composed of all those elements of the search space which are such that the components of the corresponding objective vectors cannot be all simultaneously improved.

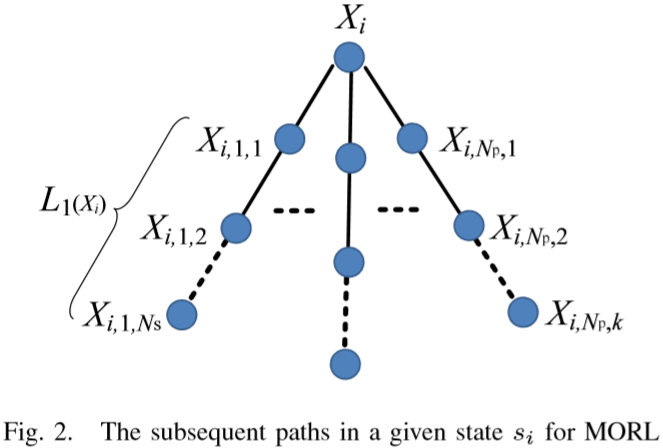
RL stems from the concept of trial-and-error learning,....

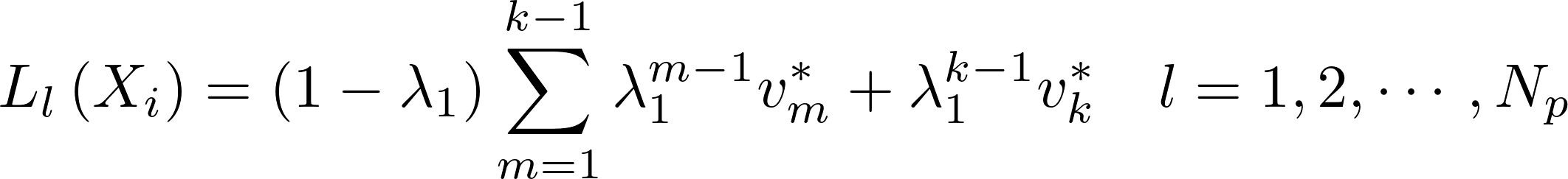
Originated in the study of animal behaviour,

Of several responses made to the same situation, those which are associated with satisfaction to the animal will, other things being equal, be more likely to recur[10].

The MORL Algorithm,

1. Reinforcement Learning for Function Optimisation,
2. path values, qt_temp, at state qt_temp, on the l-th path,



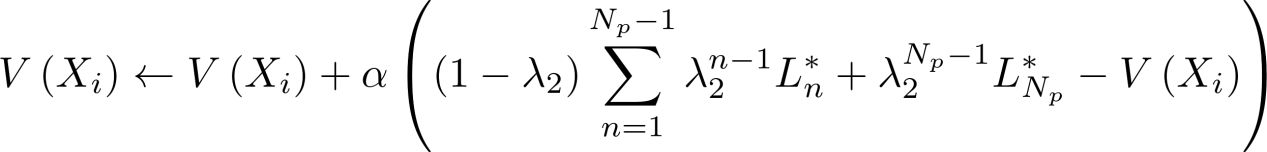


, denotes the m-th element of the vector reordered in descending k state values and k is a pre-set integer,

, denotes a weight introduced for the values of the states on the l-th path,

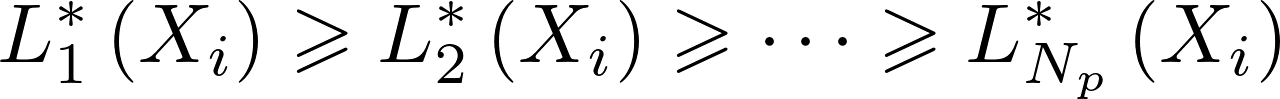
1. Evaluation of state values:

The update of the value of state Xi,

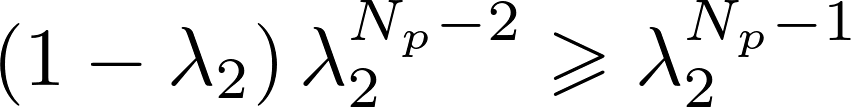


, denotes the n-th element of the vector reordered in descending of the  estimates C:/Users/71903/AppData/Local/Temp/qt_temp.it5164qt_temp,

To make:

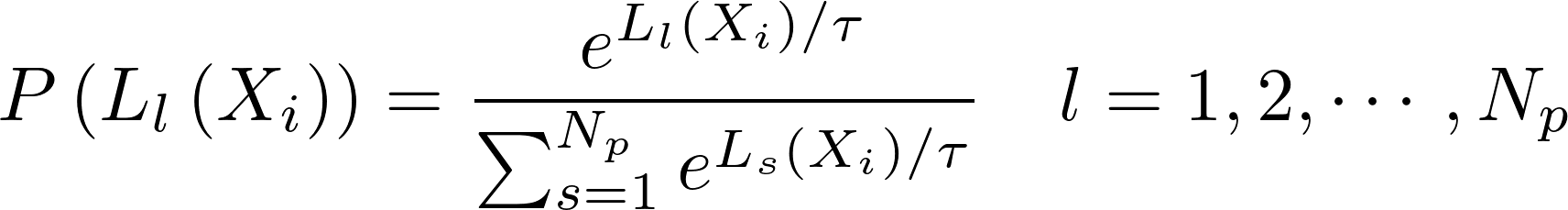


 should be given such that



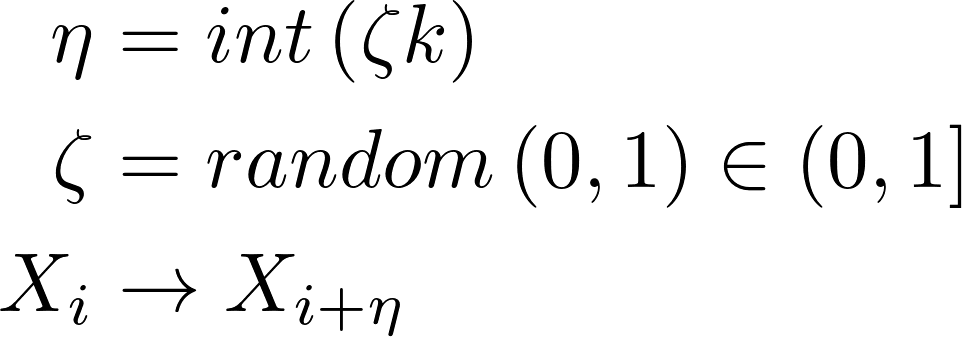
1. Action selection:

The l-th path is chosen with a probability given as follows,



Temperature is a parameter to balance the exploration and exploitation,

Once the direction is selected, an action is taken by moving from the current state to another on the same dimension with a step length η，



1. Implementation For Multi-Objective Function Optimisation

To solve optimization problems with continuous functions by RL, need to:

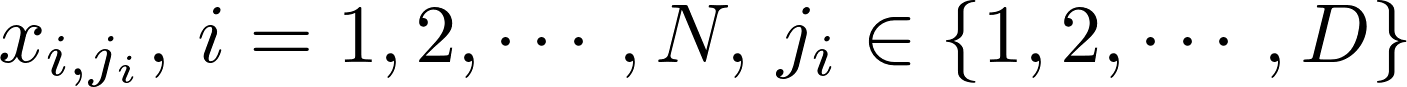
1. discretize the space into a set of states,
2. Decide how to traverse(遍历) such space,
3. Establish a criterion for assigning immediate rewards.

The curse of dimensionality,

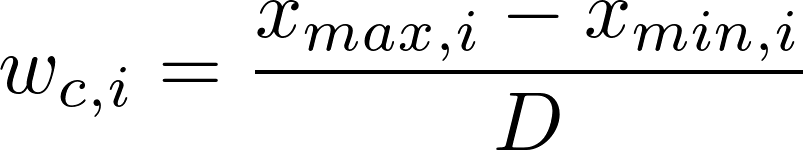
MORL adopts a dimensional search strategy and also a mechanism of dividing each dimension into cells so that a search action is taken by moving a state from one cell to another to avoid unmanageable memory of a infinite number of state.

1. Dimensional states and search in a dimension:

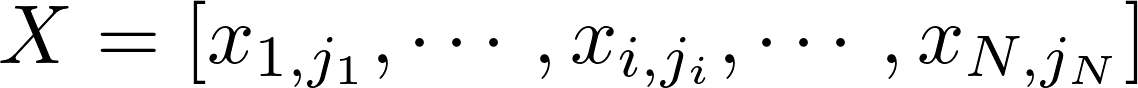
Each dimension (N dimensions) is divided into D cells,

, a dimensional state, i-th dimension, j\_{i}-th cell,

A cell width,



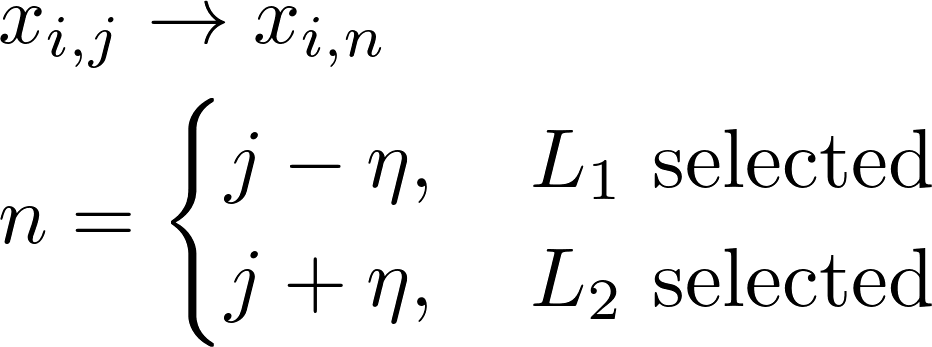
A state in the space,



As a search action of MORL is undertaken along a single dimension in each episode, a state X is also denoted as  which implies that all the elements of X are unchanged except for qt_temp.

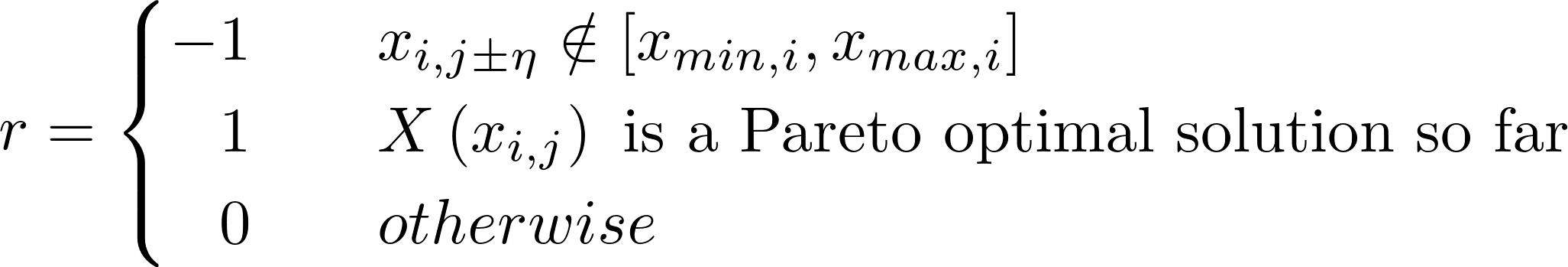
At dimensional state C:/Users/71903/AppData/Local/Temp/qt_temp.QJ5164qt_temp, there are two possible directions: moving to left or right.

There are 2 possible subsequent paths, C:/Users/71903/AppData/Local/Temp/qt_temp.cW5164qt_temp



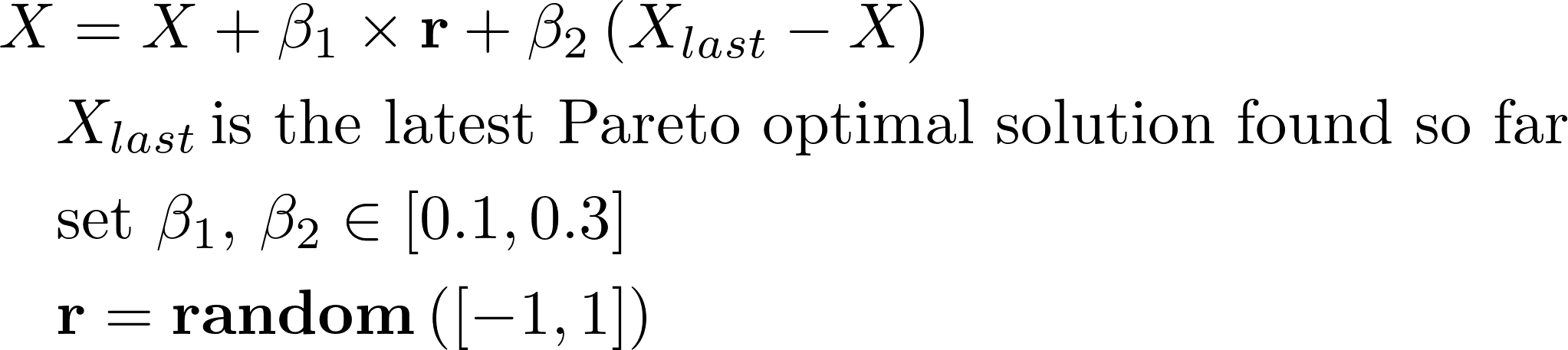
1. Immediate reward:

The immediate reward assigned to  is determined by applying  in the multi-objective functions and comparing its fitness values with other non-dominated objective vectors which correspond to the Pareto optimal solutions found so far, according to the following rule:



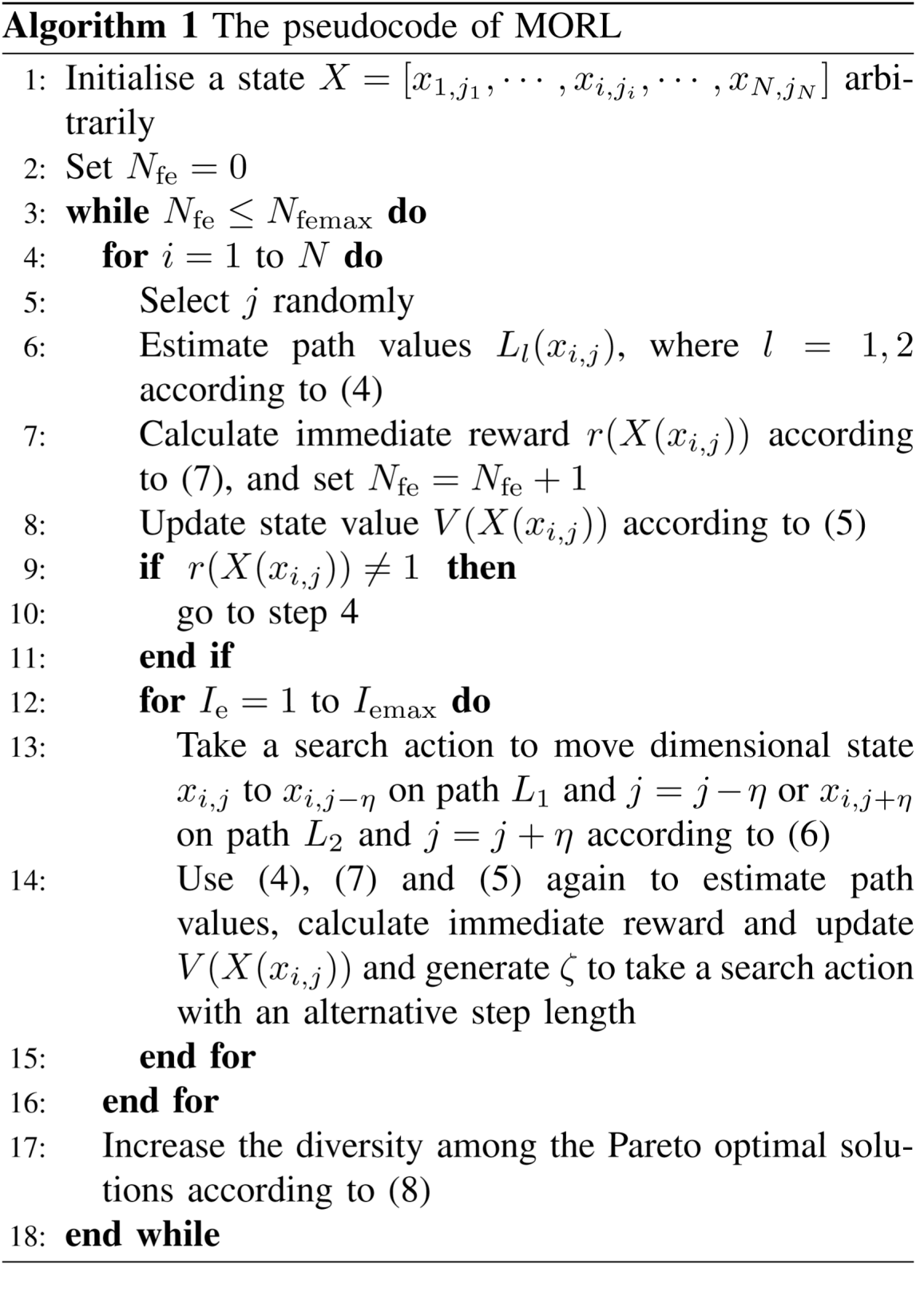
if  is a Pareto optimal solution, then put  into the elite set and delete the solutions whose objective vectors are dominated by the newly found non-dominated objective vector,

To increase the diversity of the solutions, a small portion of the elite list are selected,

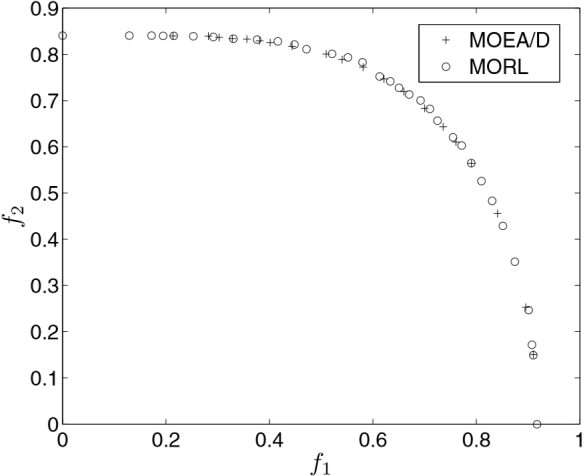
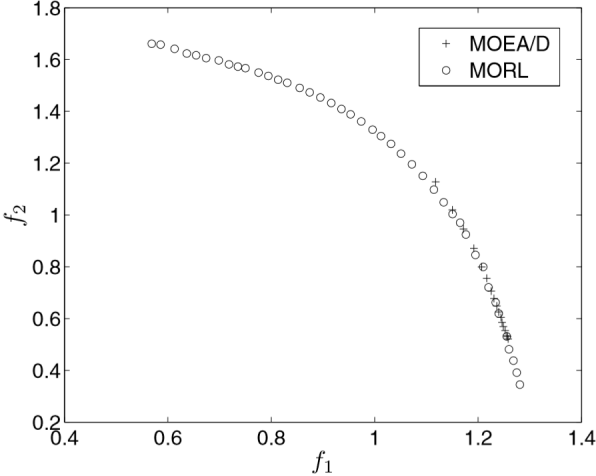
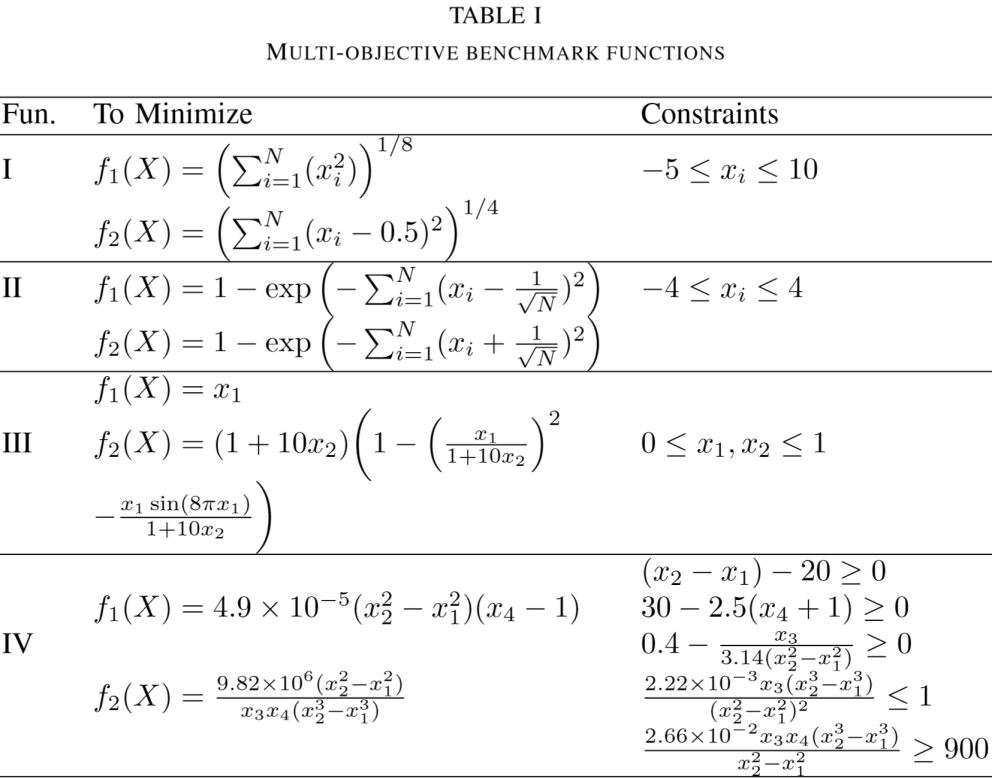


Create the diversity in the solutions, and to guarantee that the algorithm can reach every area of the search space.

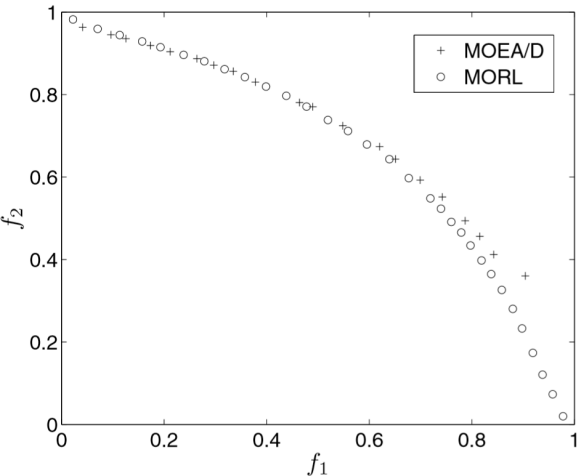
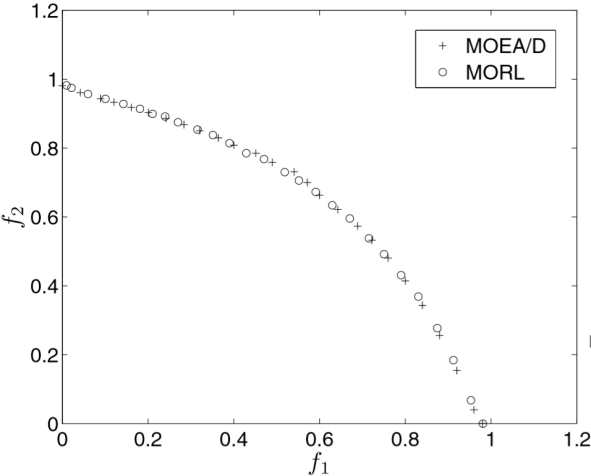
1. The pseudo code of MORL:



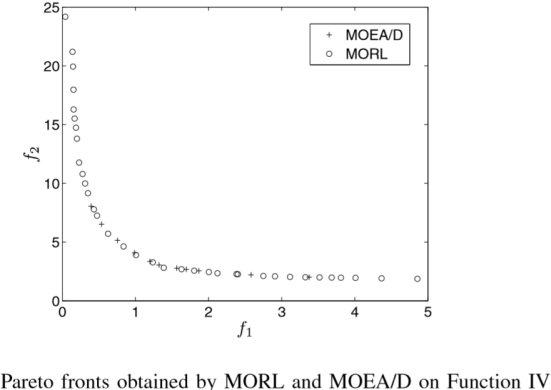
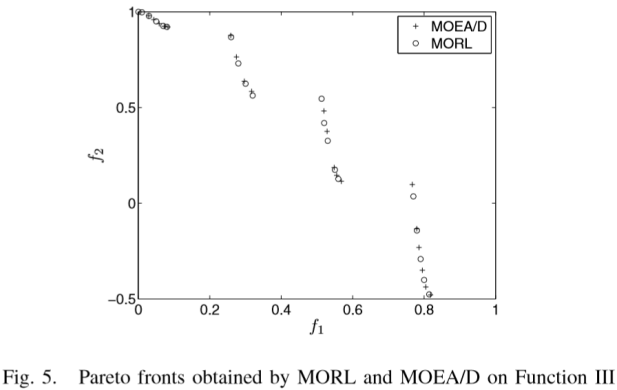
Experiment Results,











作者的实验对比了两种算法的效果，但是有一点，作者是在保证计算目标函数次数相同的条件下比较的，他们记作FEs，但是就这两种算法来说，这种方式可能对进化算法时不公平的。因为对于一个RL算法，除了计算目标函数以外，还要计算路径价值和状态价值，然后分别更新他们，这些操作的计算也是需要花费相当的资源，但是对于EA算法来说，除目标函数的计算之外，其他的操作所占资源可能是低于RL算法的其余操作的。如果考虑他们各自最好的收敛结果，可能对比更加具有说服力，或者限定在相同的运行时间内来讨论。

Conclusion:

Path Values, State Values,

Pareto dominant for reward function,

Non-dominant elite list,

Compared with MOEA/D,

这篇文章的算法非常具有方法行，具有很强的启发性，可以尝试推广到深度强化学习。

**High-dimensional Function Optimisation by Reinforcement Learning, 2010**

Q.H. Wu, H.L. Liao

Abstract,

FORL, function optimisation by Reinforcement Learning,

Large-scale and complex function optimisation problems,

FORL undertakes the dimensional search in sequence, EA based on the population-based search,

RL algorithm have a memory of the visited states,

EA algorithm, aggregate the individuals of a population towards the best selected in a current population,

Introduction,

Real word optimisation problems are complex,

Demand optimisation algorithm solve these problems accurately with a smaller number of FEs (objective function evaluation),

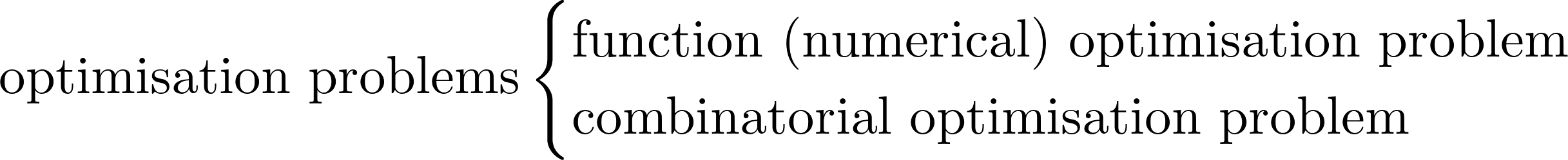
EAs stem from the study of adaptation in natural and artificial systems, and incorporate the major behaviours of a biological evolutionary process and a principle of ‘the survival of the fittest’ into an algorithmic framework to solve optimisation problems.

Large scale optimisation, EAs are notorious for their heavy computation load and slow convergence,

Inefficient search capability to achieve an accurate solution,

These problems result from using a population-based search approach, with its high level of randomness, to find a global optimum.

FORL, large-scale and complex function optimisation problems,



Essential differences between FORL and EAs:

1. FORL undertakes the dimensional search in sequence, but most EAs are population-based search;
2. FORL has a memory of history incorporated via estimating and updating of the values that have been visited, whereas EAs incorporate the behaviour of aggregating individuals of population towards the fittest selected in a current population.

Reinforcement Learning,

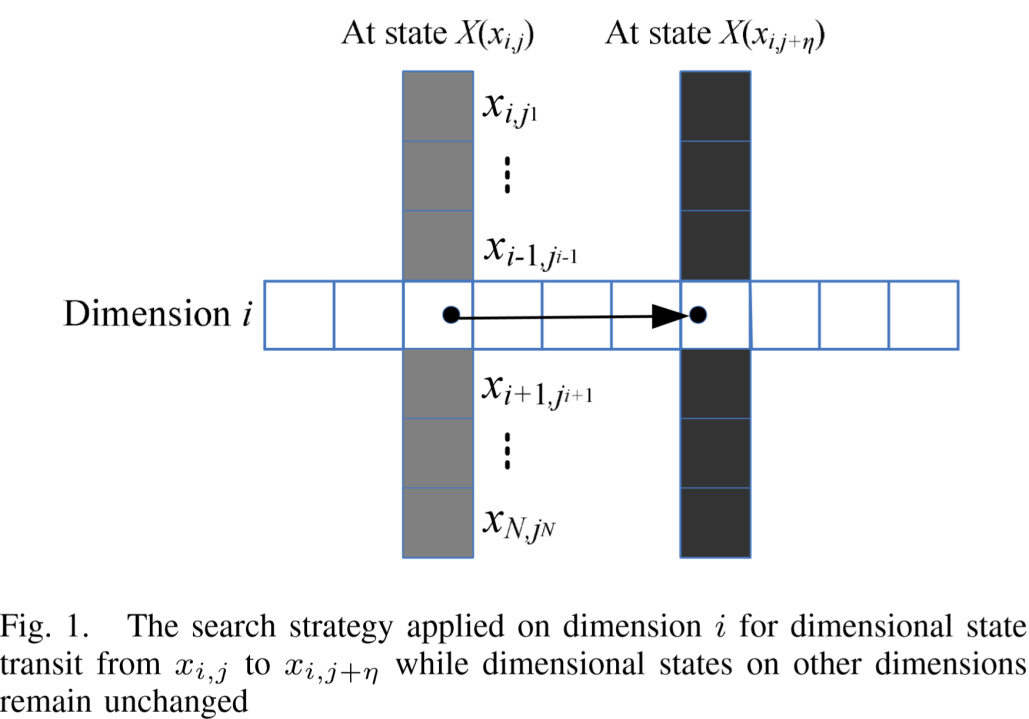
Balabalabala........

The FORL Method,



Traditional RL methods, the curse of dimensionality,

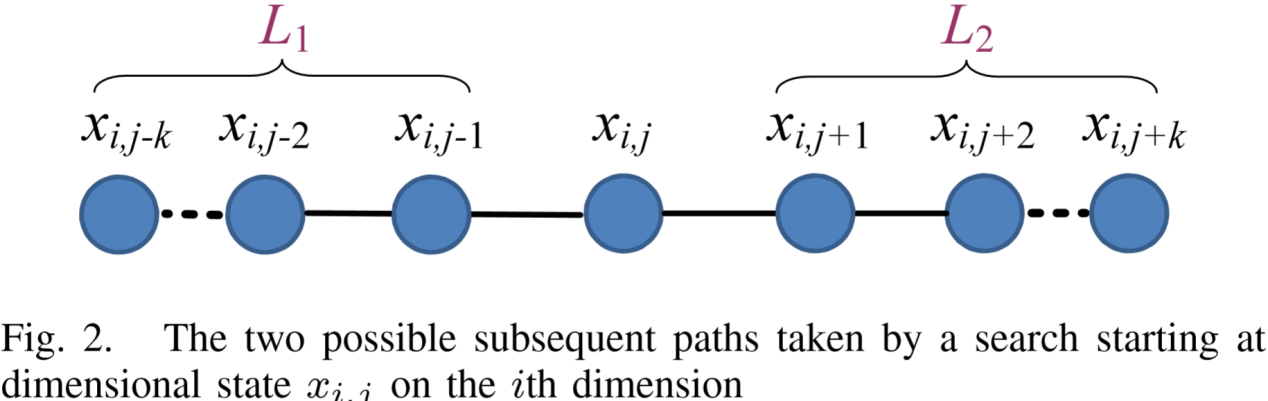
FORL, a dimensional search strategy, and also a mechanism of dividing each dimension into cells so that a search action is taken by moving a state from one cell to another to avoid unmanageable memory of a infinite number of states.



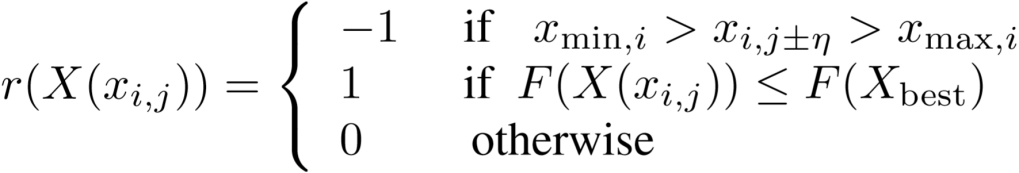
1. Dimensional states and search in a dimension

N dimension, each dimension D cells,

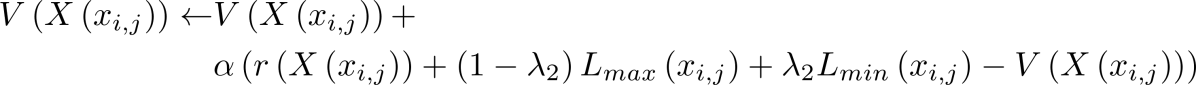
1. Search on possible subsequent paths



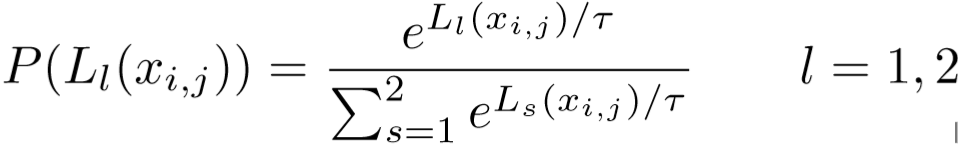
1. Immediate reward



1. Evaluation of state values



1. Taking a search action

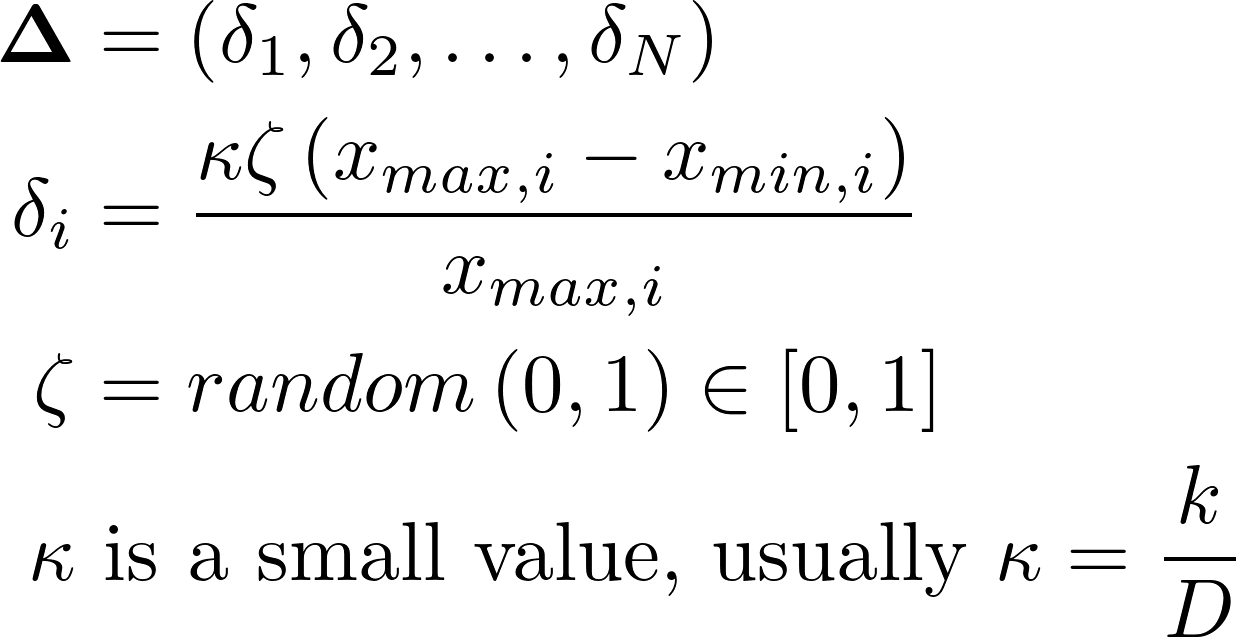


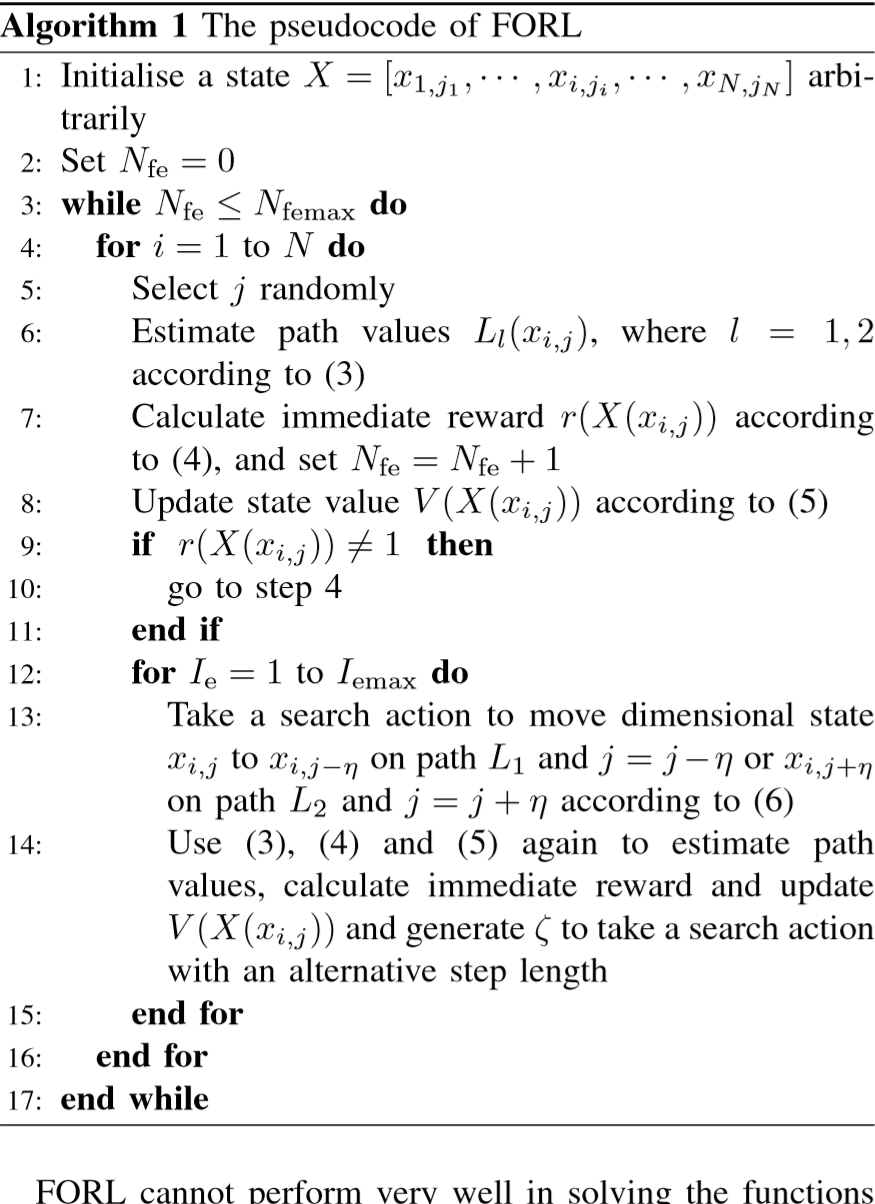
1. The pseudocode of FORL

FORL cannot perform very well in solving the functions which have flatter function hyper planes with minima located in small regions,

An alternative approach,

A perturbation vector, Δ， added to current state X, takes place after N episodes are undertaken





两位作者用基本同样的方法，分别做了大规模单目标优化问题和多目标优化问题，分别写了两篇基本一样的文章，这篇文章中对上一篇中状态空间的离散方式做了更详细的解释。这些方法的处理手段，从本质上来讲，更像是将这个问题转换成了一个组合优化的问题来处理。可能强化学习在组合优化方面的确又不错的效果吧。

**Multi-Objective Fitted Q-Iteration: Pareto frontier approximation in on single run, 2011**

Andrea Castelletti, Francesca Pianosi, Marcello Restelli,

Politecnico di Milano, Milano, Italy

Abstract,

Batch-mode Reinforcement Learning, optimal controllers in the presence of multiple objectives,

An extension of Fitted Q-iteration (FQI), for all the linear combinations of preferences (weights) assigned to the objectives in a single run.

MOFQI (multi-objective FQI), key idea, enlarge the continuous approximation of the value function, which is performed by single-objective FQI over the state-control space, also to the weight space.

MOFQI, compared with FQI for different weight values, when more than five points on the PF,

Tested on water resources systems, water reservoir,

Introduction,

Some standard single-objective control design method, stochastic dynamic programming, reinforcement learning algorithm,

MOFQI, extension of the single-objective Fitted Q-Iteration (FQI) proposed in [7],

Key idea, enlarge the continuous approximation of the value function(FQI performs over the state-control space), to the weight space (including a new variable (the weight) within the arguments of the value function),

核心思想就是，将FQI在状态空间上运行的值函数连续近似的这种方式，延申到权重空间，通过在值函数中考虑一个新的变量——权重。

Problem Formulation,

A discrete-time dynamic system,

 (1)

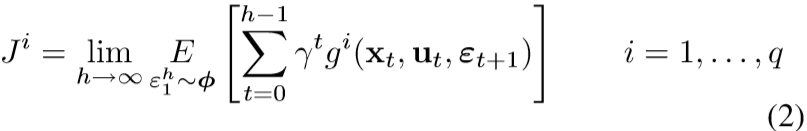
State, 

Control, 

The disturbance acting in the time interval [t, t+1), 

At time t, a feed-back stationary control law, 

The performance of the controlled system, q objective functions,



, is a scalar function that expresses the immediate cost associated to the system transition,

Problem P1,



Problem P2,



Multi-Objective Fitted Q-Iteration,

SDP(Stochastic Dynamic Programming) is the most flexible single-objective approach to solve P2,

A dual curse, dimensionality [8], modelling [9],

Approximate dynamic programming methods [10],

Batch-mode Reinforcement Learning [11],

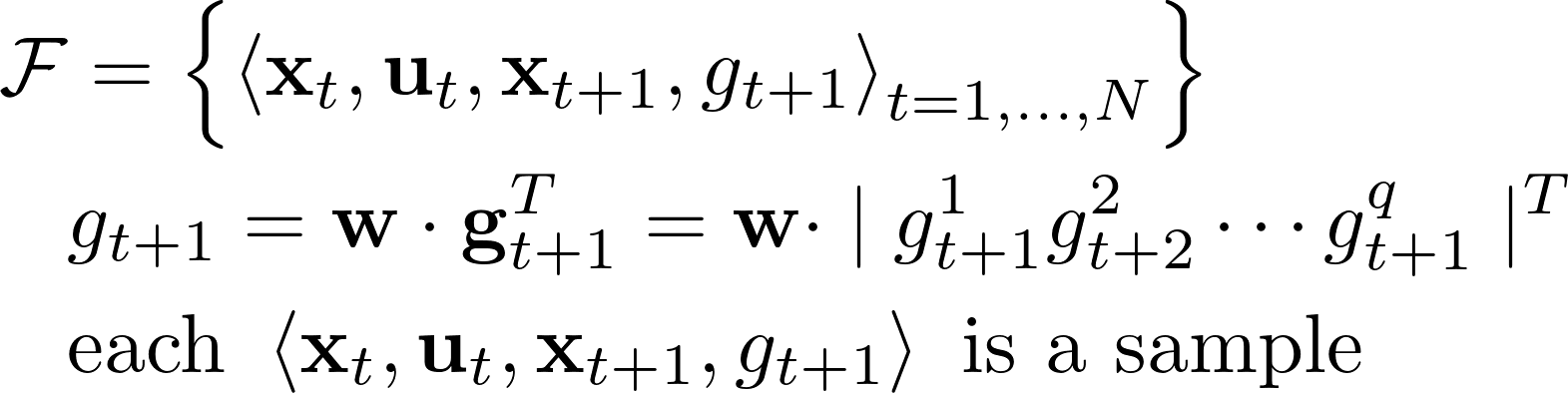
Fitted Q-Iteration (FQI) [7], [12], [13], early work [14],

Continuous approximation of the value function developed for large-scale dynamic programming [9],

FQI not require explicit modeling of the system,

The control law is determined by learning from experience,

Experience is represented as a finite data set of 4-tuples of the form:



Using the learning data set to make a regression and calculate the Q value, stopping condition and convergence properties, [7],

Once the approximation optimal Q-function is known, the control law is,

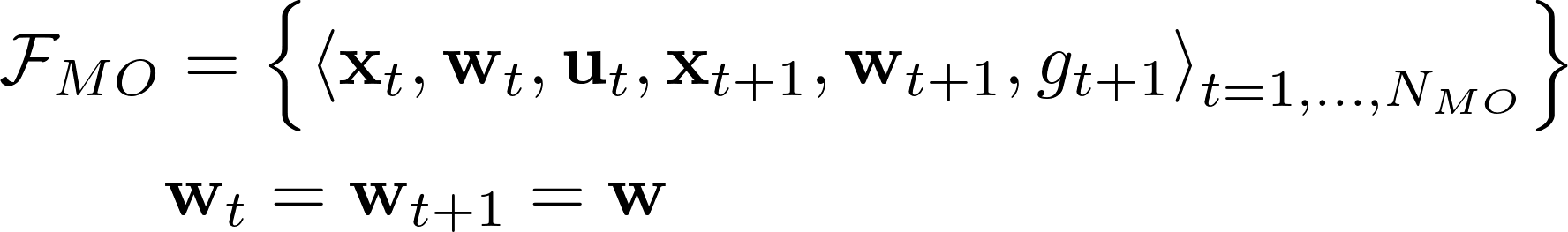
一般来说，FQI不一定能够通过有限的转移样本得到最优控制律，它返回一个问题P2的最优Q函数的近似值。利用这个近似值，可以得到最优控制律。

C:/Users/71903/AppData/Local/Temp/qt_temp.B30868qt_temp

学习数据集F是惟一需要的数据，得到它的方式是无关紧要的。

MOFQI,

The learning data set,



Optimal control law,

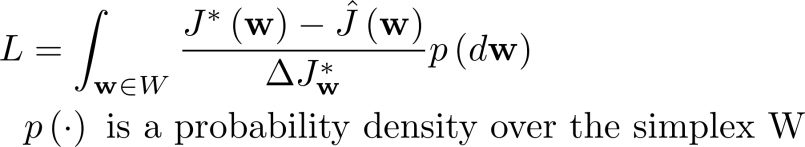
C:/Users/71903/AppData/Local/Temp/qt_temp.H30868qt_temp

对于MOFQI，由于学习样本加入了参数w，状态空间比原先大得多，因此需要更多的样本量来学习。产生学习样本集的策略主要有两种：

一种是利用之前的学习样本集F，任意随机大量的权重，然后由F计算新的样本集。

另一种就是直接在状态-权重-控制空间随机生成样本。利用一些采样技术来降低样本数量。

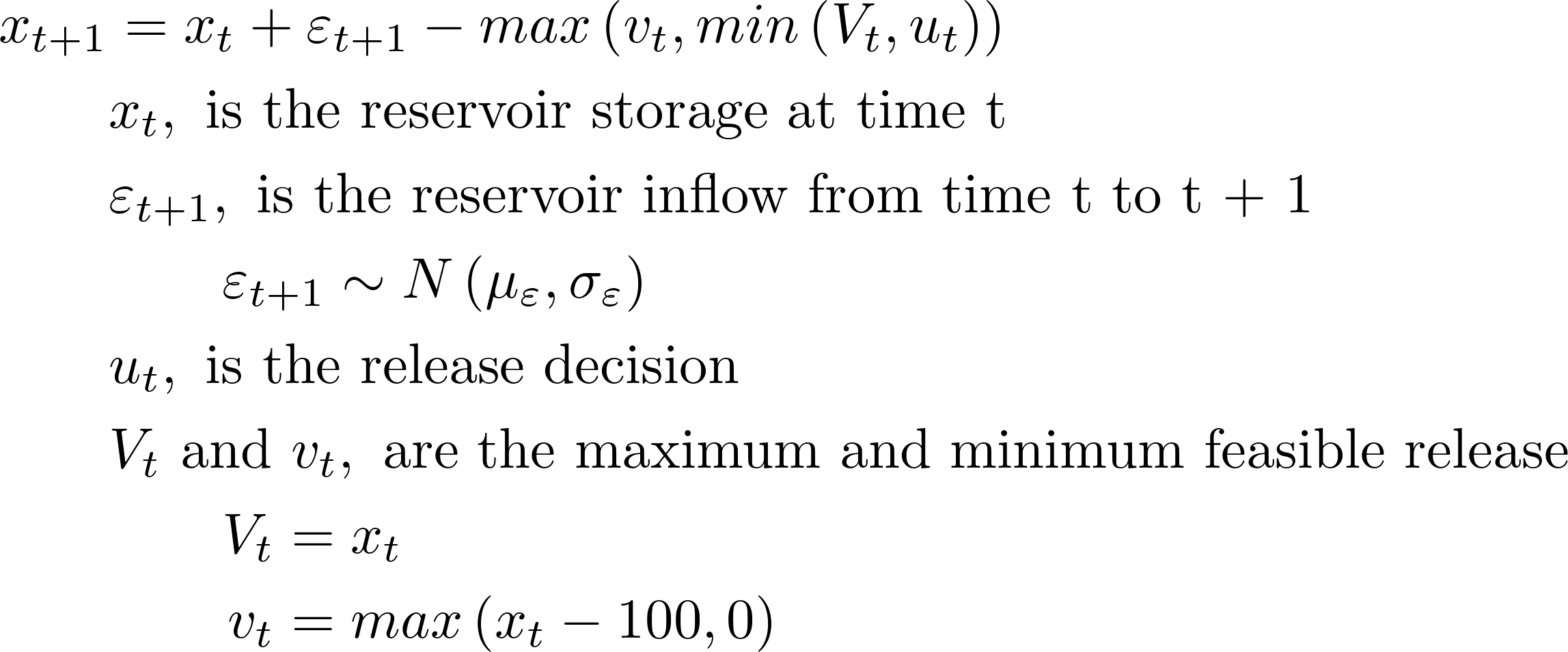
Performance Evaluation,



Numerical Experiments and Results,

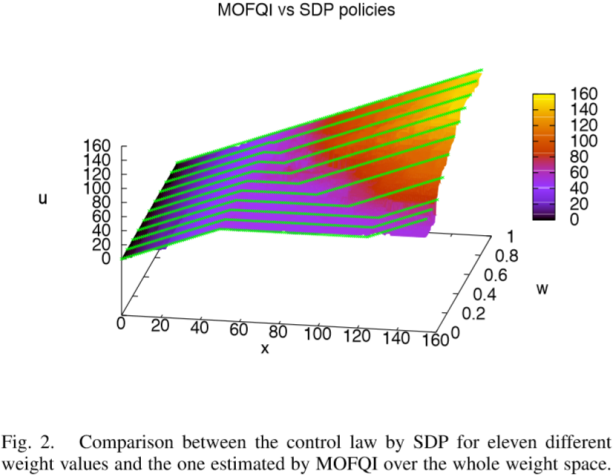
Simple water system composed by a multipurpose reservoir,

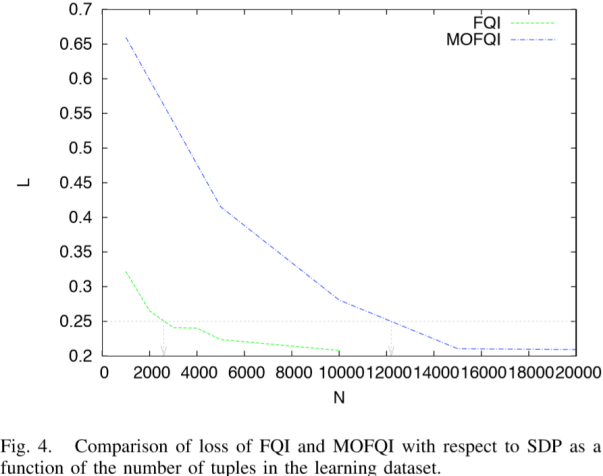
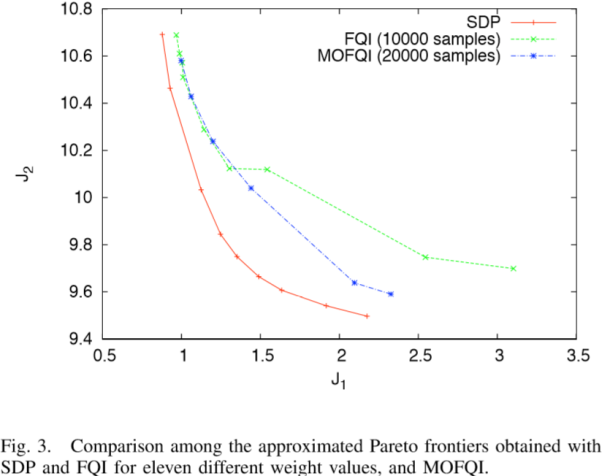
The system transition function is the mass balance equation,



Two objectives are considered, flooding and irrigation,







在实验中，用SDP方法来构造近似最优的Pareto前沿，这是一种求解Bellman方程的迭代方法，而且求解空间非常大。通过对比，显示MOFQI比多次运行固定权重的FQI得到的Pareto前沿更接近于理论前沿。他们的学习数据集都是均匀随机产生的。

**Balancing Multiple Sources of Reward in Reinforcement Learning, 2000**

Christian R. Shelton,

Artificial Intelligence Lab, MIT,

Abstract,

Multiple scalar reward, multiple reward source problem,

A single reward value by combining the multiple components can throw away vital information and can lead to incorrect solutions.

Introduction,

Multiple users, multiple goals,

Multiagent reinforcement learning, [1, 4, 5, 7],

这篇文章的术语有点奇怪，主要从信息源的角度讨论了多目标，多个信息源和多个任务目标这两类问题，用了策略梯度的方法。

Conclusion,

The biggest improvement in the algorithm will come from changing the form of the qt_temp estimator

这篇文章需要酌情重读，感觉启发性不强，但是叙述角度不常见，可能是没有读懂。

**A Multi-Objective Deep Reinforcement Learning Framework, 2020**

Thanh Thi Nguyen1, Ngoc Duy Nguyen2, Peter Vamplew3, Saeid Nahavandi2, Richard Dazeley1, Chee Peng Lim2

1School of Information Technology, Deakin University, Victoria, Australia

2Institute for Intelligent Systems Research and Innovation, Deakin University, Victoria, Australia

3School of Science, Engineering and Information Technology, Federation University, Australia

Abstract,

Scalable multi-objective reinforcement learning (MODRL) framework based on deep Q-networks,

Both single-policy and multi-policy, both linear and non-linear approaches to action selection,

2 benchmark problems, 2-objective deep sea treasure environment and 3-objective Mountain Car problem,

Introduction,

Most MORL, on relatively simple gridworld tasks,

Deep Q-network, overcomes a high degree of memory usage,

MORL, solve real-world problems, for example,

Molecule deep Q-network (2019), multi-objective DQN autonomous driving,

Multi-objective environment problem (2020),

The framework is generic and highly modularized so that it can accommodate any DRL method,

The framework flexibly accepts any state representation (scalar, vector, and graphical data) and plug-and-play, by using the network configuration,

The framework uses multithreading to significantly reduce the training time.

Contributions, one of the first frameworks to facilitate the development of MODRL algorithm. In addition, the framework proposes the use of hypervolume computation for evaluating different MODRL algorithms and a set of graphical environments.

MORL Methods and Deep Learning Extensions,

The reward signal of MORL is not a scalar but a vector where each element corresponds to an objective.

The main disadvantage of generating multiple policies is the high computational cost.

Implement two approaches, the linear weighted sum and the nonlinear TLO method,

TLO, thresholded lexicographic ordering,

MODRL, [2016 Mossalam et al.], single-policy linear MORL.

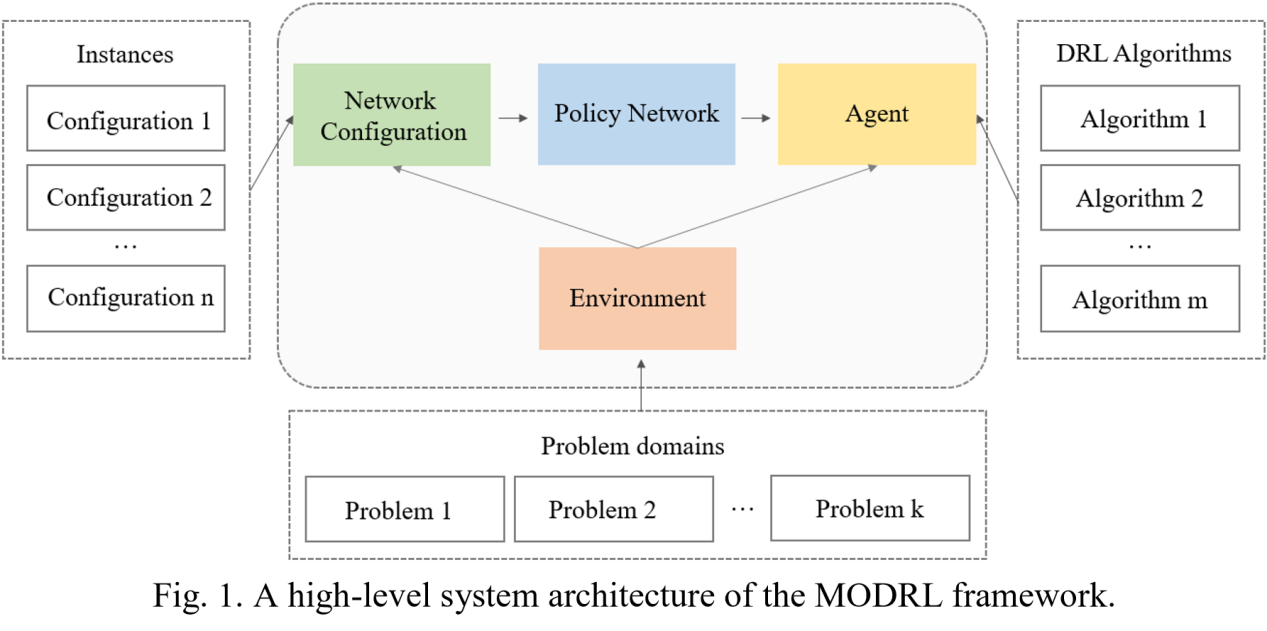
[2017 Tajmajer], DQN, non-linear action selection approach.

Recently, diverse experience replay, overcomes the inherent non-stationary problem.

[2018 Tajmajer] a modular architecture such that multiple DQNs control the agent’s behaviour in real time.

[2019 Wang et al.] an end-to-end multi-objective workflow, infrastructures-as-a-serve clouds, multiple agents.

MODRL Framework Development,

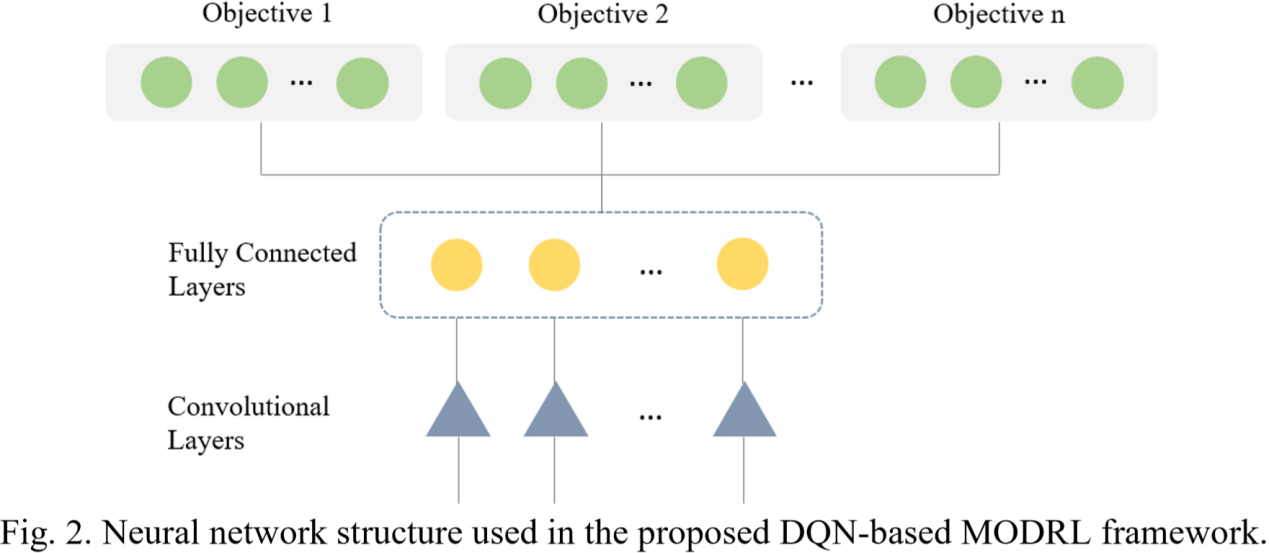


Easily apply different DRL algorithm to a variety of problem domains.

设计这个框架是为了便于将不同的DRL（深度强化学习）方法应用于不同领域的问题。

为了实现这个目的，框架的设计必须模块化使得三个统一的核心要素分离开来：神经网络，DRL 算法，以及环境。Neural networks, DRL algorithms, and environments.

Details, <http://www.deakin.edu.au/~thanhthi/drl.htm>



Finally, the output layer includes multiple groups of nodes where the number of groups is equal to the number of the objectives. Each group consists of a number of nodes corresponding to the number of possible actions.

输出根据目标函数的数量分组，每组的输出节点的数目与可能动作的数目一致。

Experiment Setting and Evaluation,

Deep Sea Treasure, DST.

MO Mountain Car.

这是个值得关注的概念:

The Intrinsic(内在的，本质的，固有的) Reward takes a non-zero signal most of the time.

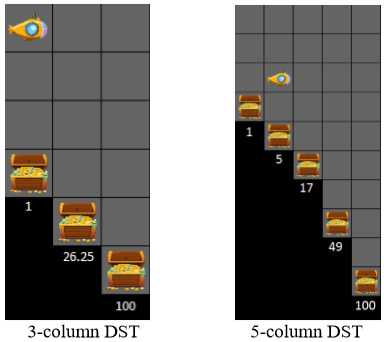
The Extrinsic(外在的，外来的，非本质的) Reward gets non-zero signal at specific time.

Several metrics: hypervolume indicator, accumulated reward, regret metric, user-based testing or simulated user testing.

The Deep Sea Treasure (DST) Problem,

Predefined Pareto solutions, normative multi-objective environment.

Two objective: maximize the treasure values and minimize the search time.



Ends when it finds a treasure location, or the predefined maximum number of actions is reached.

线性权值的方式不是通用的，这种方式无法得到(26.25, -5)这个解，如果限制权值都是正数。

但是这里有点不太严密，严格来说，这个是因为定义的奖励方式和目标函数造成的，实可以人为消除的，不是问题的本质造成的。

通过对比，非线性的TLO算法要优于线性权值的算法。

The MO Mountain Car Problem,

Input, the car’s current position and velocity.

The action sets, forward acceleration, backward acceleration, and zero throttle (null action).

Objectives, minimize the number of steps , minimizing the number of backward and forward acceleration actions.

The TLO can only be effective for the problem with no more than one intrinsic objective.

Conclusion and Further Work,

这篇文章的实际内容其实并不多，我没有仔细去核对他们提交的代码，不知道是否做了一个完备的框架，如果实现了一个框架，还是有相当的工作量的，主要是据我所知，这是首次提出做多目标强化学习的通用框架的文章。文章的设想具有一定的科学性，对于三个部分的拆分可以借鉴，试验是比较简单的，也是通用的例子，对比很清晰。

**Reinforcement Learning of Pareto-Optimal Multiobjective Policies Using Steering, 2015**

Peter Vamplew, Rustam Issabekov, Richard Dazeley, Cameron Foale

Abstract,

**Steering Approaches to Pareto-Optimal Multiobjective Reinforcement Learning, 2016**

Peter Vamplew, Rustam Issabekov, et. al.

Abstract,

