Reinforcement Learning in Robotics

Mani Manavalan[1], Apoorva Ganapathy[2]

[1]Technical Project Manager, Larsen & Toubro Infotech (LTI), Mumbai, **INDIA**
[2]Senior Developer, Adobe Systems, San Jose, California, **USA**

ABSTRACT

Reinforcement learning has been found to offer to robotics with the valid tools and techniques for the redesign of valuable and sophisticated designs for robotics. There are multiple challenges related to the prime problems related to the value addition in reinforcement of the new learning. The study has found the linkages between different subjects related to the science in particular. We have attempted to make and establish the links that have been found between the two research communities in order to provide a survey related task in reinforcement learning for behavior in terms of generation that are found in study. Many issues have been highlighted in robot learning process that are used in their learning as well as various key programming tools and methods. We discuss how contributions that aimed towards taming the complexity of the domain of the study and determining representations and goals of RL. There has been a particular focus that is based on the goals of reinforcement learning that can provide the value addition function approaches and challenges in robotic reinforcement learning. The analysis has been conducted and have strived to demonstrate the value of reinforcement learning that have to be applied to different circumstances.

Keywords: Learning control, robotic learning, Reinforcement learning

INTRODUCTION

Reinforcement learning have been found to empower a robot to self-sufficiently find an ideal conduct of the frameworks through experimentation reasons of the associations with its current circumstance. There are multiple perspectives of this phenomenon. Rather than expressly discovering the enumerating of the answer for another issue and any issue, in reinforcement learning that depends on the originator of a fundamental control task that gives the input as far as an important scalar goal identified with any capacity that have been found to quantify the one-step of the exhibition of the mechanical errand. Consider the instance of endeavoring to guide and prepare a robot to return a table tennis ball that is tossed over the net in a specific situation (Muelling et al., 2012). For this situation, there are some concerns related to the robots that have been found to mention an objective fact of the diverse idea of dynamic factors that are indicating the variable ball position and speed. This may truth be told catch well the conditions of the framework, giving a total measurement to anticipating future perceptions. The activities have been accessible to the field identified with the mechanical technology that may be found towards the engines or speed increases shipped off a backwards elements field of control framework. There are

Copyright © CC-BY-NC 2014, Asian Business Consortium | EI

designs that produces the engine orders dependent on the approaching ball and current inward arm perceptions would be known as the arrangement. There are multiple venues for the issue regarding the actions and reward as in the way of reinforcement learning issue that has to discover a way that improves the number of awards for the undertaking of reinforcement learning that depends on the plan of a calculation is one intended to discover such the ideal strategy. The award related technology work in this model have been observed to be founded on the achievement of the worth of the results and the consequences of the optional measures.

LITERATURE REVIEW

Potentially it has been found that the most obvious inspiration for future advancements in the field of robotics is mainly due to the means by which we have to effectively take RL calculations and to present the reality to tackle pragmatic applications. Consequently, we need to realize what does it take to take care of genuine issues. According to our viewpoint, specialists/robots should learn a lot quicker and more proficiently. The future holds incredible potential for a few branches for research including model-based picking up, gaining from earlier prepared undertakings, and move learning as well as area variation (Bellman, 1957). Chris Watkins didn't foster the name "Q-learning," yet he shows that a one-venture Q-learning technique can merge to the best worth capacity and strategy. Appraisals q* utilizing activity esteem capacities, which are currently normally alluded to as "Q-capacities." Watkins draws a correlation between creature molding and learning calculations (Kober and Peters, 2009). Partitions objectives into subgoals that are settled in a recursive way and picks just the state space data needed to tackle every choice. Diagram hubs are utilized to communicate basic exercises or sub-issues (Puterman, 1994). MCQ-L gives a way to deal with managing high-dimensional nonstop state spaces utilizing NN as capacity approximators and backpropagation. It's an on-strategy calculation that fits the activity esteem capacity to the current arrangement and afterward refines it dependent on those activity esteems in a ravenous way (Dayan and Hinton, 1997).

About the reinforcement learning problems that are related to the robotics, it has been found that the agent and its environment have to be modelled before starting the application of a robotic state and each of activity which may be related to either a discrete or a continuous set of problems that can also be multi-dimensional from different perspectives. A robotic system that has contain all the relevant data and the information regarding the issue and the current situation to predict the actions. There would be the multiple positions of a robot in a task related to the navigation. The form of the process of learning is to find a true set of terms that are related to actions and mainly finds an action related to the given processes that would be maximizing the cumulative expected reward (Bynagari, 2014). There are multiple approaches that are mainly based on the processes related to learning as the decision process. The transition related to the probabilities in this case for example have been written as that mainly highlights the terms related to the states on the particular course of action. It has been found that the next state related to the reward that has to be only dependent the particular action (Sutton and Barto, 1998), and it has been found based on the data and processes about the past actions. There are multiple types of the reward functions that are found to be commonly used. It also includes the reward system that depends only on the current state of the actions performed by the system.

The goal and aim of reinforcing the desired learning are related to the discovery an optimal way $\pi*$ that is mapped on the observations of the desired actions in order to

enhance the returns. The scholars have found multiple ways and venues of the outcomes (Kaelbling et al., 1996) that have been resulted into the area where the optimal solutions have been found. As the finite-horizon model has attempted only to enhance the outcomes of the actions. This is the most obvious setting that has been highlights in multiple reinforcement of the process of learning contexts. The parameter related to the value of γ has been found to be affected regardless of the concerns about the future condition that is taken into the particular account.

$$J = E\left\{\sum_{h=0}^{H} R_h\right\}$$

This scenario can be used and considered to the specific model of the areas where the known steps are remaining. There are concerns of future rewards in the terms of discounted factor of γ.

$$J = E\left\{\sum_{h=0}^{\infty} \gamma^h R_h\right\}$$

This is the most obvious setting that has been highlights in multiple reinforcement of the process of learning contexts. The parameter related to the value of γ has been found to be affected regardless of the concerns about the future condition that is taken into the particular account. There are many policies that are implemented by optimizing the small values of γ are found to be myopic. It is the straightforward scenario that has depicted the optimal control if the discount factor is weak (Kaelbling et al., 1996).

Based on the γ approaches 1, the metric has tended to approach that is known as the optimal behavior of the actions.

$$J = \lim_{H\to\infty} E\left\{\frac{1}{H}\sum_{h=0}^{H} R_h\right\}$$

These robotic problems cannot distinguish between the matters that have initially gained a particular action related to the rewards or larger values. It has been found that if a policy accomplishes an optimal behavior for the outcomes. There are some domains of real-world scenario, it has been found that the way of doing the actions where optimal level is achieved have often in condition as there are reward condition that is known to be the as stable behavior. It has been considered as vital and valid than a valid way (Peters and Schaal, 2008).

In the prime principle of the reinforcement learning algorithms that are linked with the Markov Decision Processes and used to measure the performance are providing the known results that have been found by the researchers (Kakade, 2003; Kearns and Singh, 2002; Brafman and Tennenholtz, 2002). There are some of the terms related to the error in the particular length that can be said as a finding a substitute that has to mixing time (Kearns and Singh, 2002). Off-policy methods of independent learning that have been employed for the design of the policy for example, the formation of the optimal way of action. This is the most obvious setting that has been highlights in multiple reinforcement of the process of learning contexts. The parameter related to the value of γ has been found to be affected regardless of the concerns about the future condition that is taken into the particular account. The exploration of the learning processes has to be built into the formation of the policy while designing it.

Copyright © CC-BY-NC 2014, Asian Business Consortium | EI

The working agents have to determine the right concern by finding the correlation that is essential to be found between the attempted actions and reward signals. There have been found some level of difficulty in assigning the credit for the particular rewards category. This is the most obvious setting that has been discussed in multiple reinforcement learning contexts. There are many ways as well as the parameters for finding the γ that has been found to be affected about the future condition that is taken into the particular account. In terms of the solution, the reward is modelled as a reward related to the actions taken (Powell, 2012).

REINFORCEMENT LEARNING IN THE AVERAGE-REWARD SETTING

In order to amend a policy that has to be able to be correcting it by using the multiple techniques of the optimization, there we can write a uniform policy. There are many processes for the actions of the states that are found and implemented the optimal condition as well as the policy $\pi*$ or policy parameters θ* which is best as $J(\pi) = \sum_{s,a} \mu^{\pi}(s)\pi(s,a)R(s,a)$ where μ^{π} that has found to be suitable for the solutions.

Markov Decision Processes (MDP) has been applied in the cases that are basically related to the non-ergodic processes which requires sophisticated techniques for the analysis, but there is possibility of the existence of many optimal outcomes (Puterman, 1994). There have been found two different states related to the cause and effect that is molded by the actions. This can be written as

$$\max_{\pi} J(\pi) = \sum_{s,a} \mu^{\pi}(s)\pi(s,a)R(s,a) \qquad \textbf{(1)}$$

$$s.t.\,\mu^{\pi}(s') = \sum s,a\, \mu^{\pi}(s)\pi(s,a)T(s,a,s'), \forall s' \in S \qquad \textbf{(2)}$$

$$1=\sum_{s,a} \mu^{\pi}(s)\pi(s,a) \qquad \textbf{(3)}$$

$$\pi(s,a) \geq 0, \forall\, S\,, a\ \in A.$$

Here, in the Equation (2) that has defined the term of many different states of the action distributions μπ and the Equation (3) that is ensuring a proper level of state-actions related to the probability of the distribution. There are multiple optimization problems that are found to be optimized in a well efficient manner. The working agents have to determine the right concern by finding the correlation that is essential to be found between the attempted actions and reward signals. There have been found some level of difficulty in assigning the credit for the particular rewards category. This is the most obvious setting that has been discussed in multiple reinforcement learning contexts. There are many ways as well as the parameters for finding the γ that has been found to be affected about the future condition that is taken into the particular account.

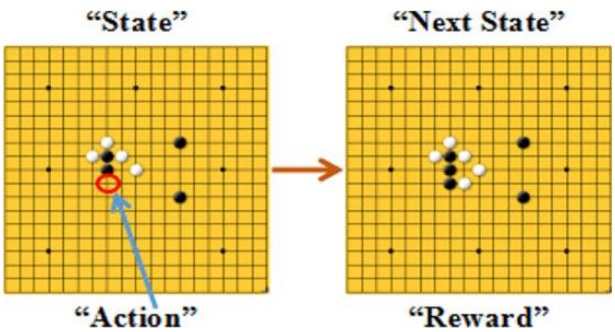

Figure 1: The action and reward state in the simplest decision process

THE SIGNIFICANCE OF THE APPROACHES OF VALUE FUNCTION

There are vital issues related to the reinforcement of the learning has been found to be focusing on issue and to solve the optimization problem that has been determined by the course of action. There are many concerns related to the formulation of the required design related to the problem have been designed as the value that it offers have useful data relevant to the study. It has allowed us for a learning process. It has also allowed for the domain-appropriate restructuring of the design of the policy in terms of making it approximate. There are many optimal actions that have many value functions. There are some other concerns in robotics that are related to the policy search. There are multiple traditional reinforcement functions related to the learning schemes that have been determined and those are based on valuing the solutions to the action.

$$L = \sum_{s,a} \mu^{\pi}(s)\pi(s,a)R(s,a) + \sum_{s'} V^{\pi}(s')\left[\sum_{s,a} \mu^{\pi}(s)\pi(s,a)T(s,a,s') - \mu^{\pi}(s')\right] + \bar{R}\left[1 - \sum_{s,a} \mu^{\pi}(s)\pi(s,a)\right]$$

$$= \sum_{s,a} \mu^{\pi}(s)\pi(s,a)\left[R(s,a) + \sum_{s'} V^{\pi}(s')T(s,a,s') - \bar{R}\right] - \sum_{s'} V^{\pi}(s')\mu^{\pi}(s')\sum_{a'} \pi(s',a') + \bar{R}$$

At the place of applying the properties $\sum_{s',a'} V(s')\ \mu^{\pi}(s')\ \pi(s',a') = \sum_{s,a} V(s)\ \mu^{\pi}(s)\ \pi(s,a)$ there are optimization issues that can be solved by the equation which yields extrema at

$$\partial_{\mu^{\pi}\pi} L = R(s,a) + \sum_{s'} V^{\pi}(s')T(s,a,s')\bar{R} - V^{\pi}(s) = 0$$

This statement has been implied that there are some equations that belong to the number of states that can be obtained by multiplying it with the number of actions. For each of the action related to a particular action, we have found several optimal states of action a∗ that can be resulted in the same value. Hence, the optimal action are written as a^* as $V^{\pi^*}(s) = R(s,a^*) - \bar{R} + \sum_{s'} V^{\pi^*}(s')T(s,a^*,s')$ the learning has been found to be focusing on issue and to solve the optimization problem. that has been determined in its dual form that is

$$V^*(s) = \max_{a^*}\left[r(s,a^* - \bar{R} + \sum_{s'} V^{8}(s')T(s,a^*,\ s')\right]$$

This is the statement that has been derived and this is equivalent to the states related principle that is mentioned by Bellman (Bellman, 1957) 3 and it has been stated as "An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision."

There is multiple traditional reinforcement related to the learning approaches have been determined that are the learning has been found to be focusing on issue and to solve the optimization problem. that has been determined.

$$V^{\pi}(s) = \sum_{a} \pi(s,a)\left(R(s,a) - \bar{R}\sum_{s'} V^{\pi}(s')\,T(s,a,s')\right)$$

Copyright © CC-BY-NC 2014, Asian Business Consortium | EI

Instead of the value function V^{π} (s) many algorithms rely on the state-action value function Q^{π} (s, a) instead, which has advantages for determining the optimal policy as shown below. This function is defined as

$$Q^{\pi}(s,a) = R(s,a) - \bar{R} + \sum_{s'} V^{\pi}(s')T(s,a,s')$$

In contrast to the optimization actions and states there are V^{π} (s), and the other function related to the Q^{π} (s, a) that are many useful data as well as the required conditions. In terms of the optimal action, we have found the value function that is

$$Q^{*}(s,a) = R(s,a) - \bar{R} + \sum_{s'} V^{*}(s')T(s,a,s') = R(s,a) - \bar{R} + \sum_{s\prime} \left(\max_{a'} Q^{*}(s',a')\right) T(s,a,s')$$

The deterministic policy has been defined as π^{*} (s) that are picking the course of action. This is the most obvious setting that has been highlights in multiple reinforcement of the process of learning contexts. The parameter related to the value of γ has been found to be affected regardless of the concerns about the future condition that is taken into the particular account.

$$\pi^{*}(s) = arg\max_{a}\left(R(s,a) - \bar{R} + \sum_{s} V^{*}(s')T(s,a,s')\right)$$

The finding of the optimal value function S^{*} (s) for all known states are determined by applying the optimal policy in a setting and in contrast to the optimization actions and states there are V^{π} (s), and the other function related to the Q^{π} (s, a) that are many useful data as well as the learning process about the effects of a particular action. In terms of the optimal action.

$$\pi^{*}(s) = arg\max_{a}\left(Q^{*}(s,a)\right)$$

The expressions have avoided the calculation of the average of the successor states, and hence there is the state when the transition function is essential.

POLICY SEARCH

There are many concerns related to the formulation of the required design related to the problem have been designed as the value that it offers have useful data relevant to the study. It has allowed us for a learning process. It has also allowed for the domain-appropriate restructuring of the design of the policy in terms of making it approximate. There are many optimal actions that are given to the states. There are some other concerns in robotics that are related to the terms of action that has become very important for the sake of optimal states. There are multiple traditional reinforcement functions related to the learning schemes that have been determined and those are based on valuing the solutions to the action.

$$\theta_{i+1} = \theta_i + \Delta\theta_i$$

There are multiple calculations related to the policy that been proposed training to the workers that has to data analysis (Strens and Moore, 2001; Ng et al., 2004a) and it has been found that the proper gradient method that have used the finite differences between the state and the actions related to the processes (Betts, 2001).Some of the most obvious solutions related to the terms and conditions that have adopted particularly those that are integrated into terms of robotics inclusion of the policy implications that have to be adopted and the approaches (Sutton et al., 1999).

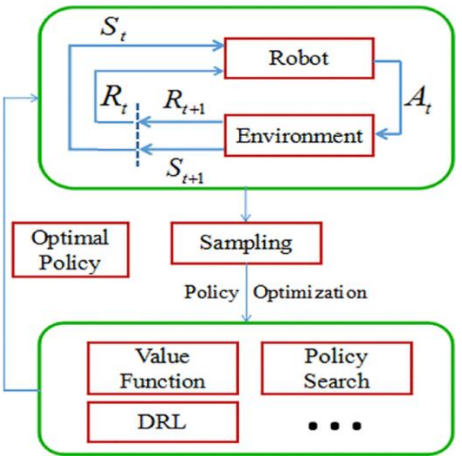

Figure 2: The learning process of robot based on RL

$$\theta_{i+1} = \theta_i + \alpha \nabla_\theta J$$

The policy strategy related to the concern has been found to be implemented by the following equation $\Delta \widehat{J_P} \approx J(\theta_i + \Delta \vartheta_p) - J_{ref}$ it has been known that these are the parameters that have to be set in order to compel the course of action. There are some gradients that can now be estimated by linear regression equation of states.

$$\nabla_\theta J \approx (\Delta \boldsymbol{\Theta}^{\boldsymbol{T}} \Delta \quad \boldsymbol{\Theta} \quad)^{-1} \Delta \boldsymbol{\Theta}^{\boldsymbol{T}} \Delta \hat{\boldsymbol{J}},$$

It has been defined by the matrix $\Delta \boldsymbol{\Theta}$ as it has been found to contains all the samples that are based on the perturbations $\Delta \theta_p$ and particularly the $\Delta J\hat{}$. There are many concerns related to the formulation of the required design related to the problem have been designed as the value that it offers have useful data relevant to the study. It has allowed us for a learning process. It has also allowed for the domain-appropriate restructuring of the design of the policy in terms of making it approximate. There are many optimal actions than optimal value of the states of action.

$$J^\theta = \sum_r P^\theta(r) J^r$$

There are some gradients that can be incorporated, and it can be written as

$$\nabla_\theta P^\theta(r) = P^\theta(r) \nabla_\theta \log P^\theta(r)$$

This is the term that defines likelihood ratio, and it can also be said as reinforcement (Williams, 1992).

$$\nabla_\theta J^\theta = \sum_r \nabla_\theta P^\theta(r) J^r = \sum_r P^\theta(r) \nabla_\theta \log P^\theta(r) J^r = E\{\nabla_\theta \log P^\theta(r) J^r\}$$

The finding of the optimal value function $V^*(s)$ for the known states are determined by applying the optimal policy in a setting and in contrast to the optimization actions and states there are $V^\pi(s)$, and the other function related to the $Q^\pi(s, a)$ that are many useful data as well as the information. The parameter related to the value of γ has been found to be affected regardless of the concerns about the future condition that is taken into the particular account.

Copyright © CC-BY-NC 2014, Asian Business Consortium | EI

$$\nabla_\theta J^\theta = E\left\{\left(\sum_{h=1}^{H} \nabla_\theta \log \pi^\theta(s_h, a_h)\right) J^r\right\}$$

The episodic nature of the concern of the action as it has been mentioned by the episode J^r and same is represented by the value function of the state (Peters and Schaal, 2008).

$$\nabla_\theta J^\theta = E\left\{\sum_{h=1}^{H} \nabla_\theta \log \pi^\theta(s_h, a_h)\, Q^\pi(s_h, a_h)\right\}$$

This is the equation that is linked with the optimal behavior of the states and linked with the policy statement (Sutton et al., 1999). The finding of the optimal value function for the known states are determined by applying the required setting of the states and in contrast to the optimization actions and states there are $A^\pi(s)$, and the other function related to *the* $P^\pi(s, a)$. This is the most obvious setting that has been highlights in multiple reinforcement of the process of learning contexts.

There are different classes of the states of the actions that represent the phenomenon (Dayan and Hinton, 1997). There are ways that are not known to the researchers and that have proven successful in robotics, there are many returns related expectation and Cost-regularized valuations (Kober and Peters, 2009). The system theory has been found to be closely related update rules can be implemented and it is including the terms and conditions valuations with Path valuation techniques (Theodorou et al., 2010) where these are found to be effective (Yamaguchi and Takanishi, 1997).

CHALLENGES IN THE REINFORCEMENT LEARNING OF THE ROBOTS

The phenomenon of curse of dimensionality

Multiple scholars have found the optimal dimensionality of the robotics learning process as Bellman (1957) has explored the level of optimal situation that has been addressed in the high-dimensional spaces. The literature has been also faced an exponential explosion of the multiple states that are related to the states and actions.

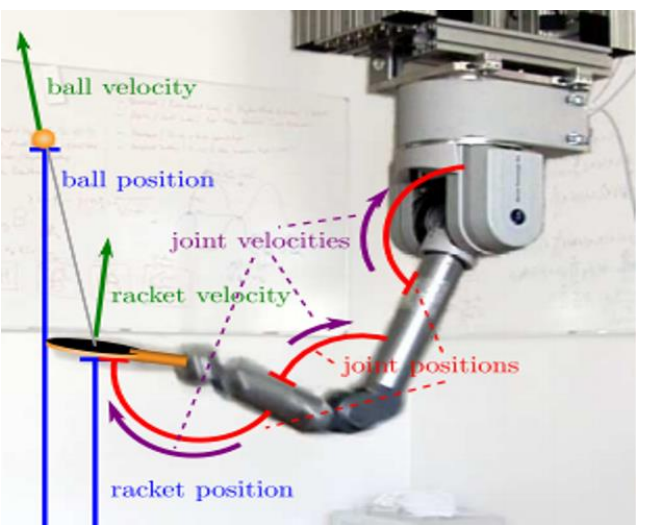

Figure 3: Illustration of the modelling of a robot learning

For the sake of an example, if someone has assumed that each way of action that linked with a space that has been found into all the ten levels of the actions then we can have 11 states for a one-dimensional state-space, and there would be 105 = 1050. Evaluating the actions and the state quickly has been required and it has become infeasible with growing dimensionality (Donoho, 2000). This is the most obvious setting that has been highlights in multiple reinforcement of the process of learning contexts. The parameter related to the value of γ has been found to be affected regardless of the concerns about the future condition that is taken into the particular account. There are many policies that are implemented by optimizing the small values of γ are found to be myopic. It is the straightforward scenario that has depicted the optimal control if the discount factor is weak (Yamaguchi and Takanishi, 1997) that have taken the control of the robotic system that has to be implemented.

The samples of the Curse related to the real-world

Robots intrinsically associate with the actual world. Henceforth, robot reinforcement learning experiences most of the subsequent certifiable issues. For instance, robot equipment is normally costly, experiences mileage, and requires cautious support. Fixing a robot framework is a non-irrelevant exertion related to the actual work, and long holding up periods. This is the most obvious setting that has been highlights in multiple reinforcement of the process of learning contexts. The parameter related to the value of γ has been found to be affected regardless of the concerns about the future condition that is taken into the particular account. There are some safe investigation turns into a central point of interest of the learning system an issue regularly ignored in the overall reinforcement learning local area (Donoho, 2000). Perkins and Barto (2002) as there are some of the thoughts of a technique for developing of the process of learning specialists dependent on the valid capacities. Exchanging between the fundamental regulators is consistently protected and offers essential execution ensures.

The functioning specialists need to decide the right worry by discovering the connection amongst activities and award signals. There has been observed to be some degree of trouble in appointing the credit for the specific prizes class. This is the clearest setting that has been examined in numerous reinforcement learning settings (Donepudi, 2014). The boundary γ has been observed to be influenced by the future condition that is taken into the specific record. This boundary frequently subjectively changes the type of the ideal arrangement. On the off chance that both the progress probabilities and award work are referred to, this can be viewed as an ideal control issue.

The phenomenon of curse related to under-modelling

There are many approaches to balance the expense of the cooperation that has to be utilize precise models as test systems. In an optimal setting, this methodology would deliver it conceivable to get familiar with the conduct in recreation and therefore move it to the genuine robot. Tragically, making an adequately precise working condition as well as the current circumstance is testing and normally requires a lot of information tests. There are multiple perspectives as it has been found that state of the action has much value than it looks like. As little model blunders because of this under-demonstrating amass, the recreated robot can rapidly veer from this present reality framework. It has been found that while making an adequately precise way of doing actions and its current circumstance is testing and normally requires a lot of information tests Atkeson (1998).

Copyright © CC-BY-NC 2014, Asian Business Consortium | EI

The phenomenon of curse related to the goal specification

In support of the learning processes, the ideal conduct is determined by the award that is linked with the actions. The objective of support learning calculations then, at that point is to augment the collected way of rewarding the actions. While frequently significantly less complex than indicating the actual conduct, by and by, it very well may be shockingly hard to characterize a decent award work in robot support learning. The student should notice difference in the award signal to have the option to work on a strategy: if a similar return is constantly gotten, it is basically impossible to figure out which strategy is better or nearer to the ideal.

CONCLUSION

We may conclude from this paper that reinforcement learning is a critical component of machine learning and artificial intelligence. The RL techniques are used in both standard and deep learning applications. Due to algorithmic and hardware limitations, robots have been unable to acquire high intelligence over the last few decades. This page provides a thorough overview of the numerous types of RL algorithms and models used in robot development. It is particularly useful in research, as well as in the creation of automated robots, simulators, and other similar devices, due to its human-like learning approach. However, keep in mind that there are many more of them out there; we've just addressed the ones that are absolutely required for learning RL. Integrating principles from intrinsic motivation into our strategy would allow us to actively pick goals that will help us learn more quickly what we can and cannot achieve. Another possibility for the future is to teach our generative model to be aware of the dynamics. By encoding information on the dynamics of the environment, the latent space could be made even more suitable for reinforcement learning, resulting in faster learning. Finally, there are a variety of robot activities for which state representation with sensors would be challenging, such as manipulating deformable items or handling scenes with a changeable number of objects. The next stage would be to scale up RIG to solve these problems.

REFERENCES

Atkeson, C. G. (1998). Nonparametric model-based reinforcement learning. In Advances in Neural Information Processing Systems (NIPS).

Bellman, R. E. (1957). Dynamic Programming. Princeton University Press, Princeton, NJ.

Betts, J. T. (2001). Practical methods for optimal control using nonlinear programming, volume 3 of Advances in Design and Control. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.

Brafman, R. I. and Tennenholtz, M. (2002). R-max - a general polynomial time algorithm for near-optimal reinforcement learning. Journal of Machine Learning Research, 3:213–231.

Bynagari, N. B. (2014). Integrated Reasoning Engine for Code Clone Detection. *ABC Journal of Advanced Research*, *3*(2), 143-152. https://doi.org/10.18034/abcjar.v3i2.575

Dayan, P. and Hinton, G. E. (1997). Using expectation-maximization for reinforcement learning. Neural Computation, 9(2):271–278.

Donepudi, P. K. (2014). Voice Search Technology: An Overview. *Engineering International*, *2*(2), 91-102. https://doi.org/10.18034/ei.v2i2.502

Donoho, D. L. (2000). High-dimensional data analysis: the curses and blessings of dimensionality. In American Mathematical Society Conference Math Challenges of the 21st Century.

Kaelbling, L. P., Littman, M. L., and Moore, A. W. (1996). Reinforcement learning: A survey. Journal of Artificial Intelligence Research, 4:237–285

Kakade, S. (2003). On the Sample Complexity of Reinforcement Learning. PhD thesis, Gatsby Computational Neuroscience Unit. University College London.

Kearns, M. and Singh, S. P. (2002). Near-optimal reinforcement learning in polynomial time. Machine Learning, 49(2-3), 209–232.

Kober, J. and Peters, J. (2009). Policy search for motor primitives in robotics. In Advances in Neural Information Processing Systems (NIPS).

Muelling, K., Kober, J., Kroemer, O., and Peters, J. (2012). Learning to select and generalize striking movements in robot table tennis. International Journal of Robotics Research.

Ng, A. Y., Coates, A., Diel, M., Ganapathi, V., Schulte, J., Tse, B., Berger, E., and Liang, E. (2004a). Autonomous inverted helicopter flight via reinforcement learning. In International Symposium on Experimental Robotics (ISER)

Perkins, T. J. and Barto, A. G. (2002). Lyapunov design for safe reinforcement learning. Journal of Machine Learning Research, 3:803–832.

Peters, J. and Schaal, S. (2008). Reinforcement learning of motor skills with policy gradients. Neural Networks, 21(4):682–697.

Powell, W. B. (2012). AI, OR and control theory: A rosetta stone for stochastic optimization. Technical report, Princeton University

Puterman, M. L. (1994). Markov Decision Processes: Discrete Stochastic Dynamic Programming. Wiley-Interscience.

Strens, M. and Moore, A. (2001). Direct policy search using paired statistical tests. In International Conference on Machine Learning (ICML)

Sutton, R. S. and Barto, A. G. (1998). Reinforcement Learning. MIT Press, Boston, MA.

Sutton, R. S., McAllester, D., Singh, S. P., and Mansour, Y. (1999). Policy gradient methods for reinforcement learning with function approximation. In Advances in Neural Information Processing Systems (NIPS).

Theodorou, E. A., Buchli, J., and Schaal, S. (2010). Reinforcement learning of motor skills in high dimensions: A path integral approach. In IEEE International Conference on Robotics and Automation (ICRA).

Yamaguchi, J. and Takanishi, A. (1997). Development of a biped walking robot having antagonistic driven joints using nonlinear spring mechanism. In IEEE International Conference on Robotics and Automation (ICRA).

--0--

Copyright © *CC-BY-NC 2014*, Asian Business Consortium | *EI*

Source of Support: Nil, **No Conflict of Interest:** Declared

This article is is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.
Attribution-NonCommercial (CC BY-NC) license lets others remix, tweak, and build upon work non-commercially, and although the new works must also acknowledge & be non-commercial.