Reinforcement Learning for Constrained Control Problems with High-Dimensional Continuous Action Spaces

Authors[1*]
[1]Affiliation
{email}@affiliation

1 Related work

We classify prior literature related to this work into various categories, including the problem domain, traditional approaches for control of such systems, data-driven approaches such as adaptive control, approximate dynamic programming (ADP) and imitation learning (IL), control algorithms for supply chain and inventory management, and the use of reinforcement learning in related problem areas.

Problem forms: Dynamic systems are typically distinguished based on the form of state that they aim to control. The simpler form is a scalar state such as the rate of a chemical reaction [Gustafsson and Waller, 1983], angle of pitch of an aircraft [Levant *et al.*, 2000], or temperature of a boiler [Liu and Chan, 2006]. A more complex version is the multivariable control problem, involving a vector of states which may or may not be directly dependent on each other. The canonical version of a multivariable system with dynamics, is to represent the state by a vector of variables, and their evolution over time using matrix differential equations [Ogata and Yang, 2002]. Examples of multivariable problems with directly dependent state are the position and its derivatives (velocity and acceleration) of an inverted pendulum [Pathak *et al.*, 2005], while indirectly dependent states are found in transportation networks [De Oliveira and Camponogara, 2010] and inventory management in supply chains [Van der Laan and Salomon, 1997].

Traditional control: The preferred control approach for multivariable problems is to use a linear time invariant (LTI) model of the system and to design a controller using classical methods in the frequency domain or using state feedback [Bouabdallah *et al.*, 2004; Golnaraghi and Kuo, 2010]. Nonlinear versions of the problem are solved using techniques such as sliding mode [Utkin *et al.*, 2009] and model predictive control [Mayne *et al.*, 2000], while robust control under stochastic disturbances is handled using techniques such as H_∞ [Doyle *et al.*, 1989]. The key takeaway from these techniques is that they address one complex aspect of the problem in isolation. For example, frequency domain control synthesis addresses stability and robustness when the system dynamics are linear and known, while state feedback control works with high dimensionality under the same assumptions. Sliding mode and model predictive control also assume that the system dynamics are known, and typically are only applicable in low dimensional problems. Robust control techniques allow the noise or disturbance to be arbitrary, but the system dynamics are known. A separate class of techniques addresses the problem of system identification, which focuses on models with unknown dynamics [Ljung, 1998; Bestle and Zeitz, 1983]. However, these methods do not include control design with this function. In fact, there are results in many classes of problems where a simple closed loop with system identification followed by a traditional controller has no stability guarantees whatsoever [Doyle, 1978].

Adaptive control, ADP, imitation learning: There is thus a clear motivation for using algorithms that are developed for the generic problem but can adapt to specific problem instances. There exists a large volume of existing literature on adaptive control [Ioannou and Sun, 1996; Åström and Wittenmark, 2013], where the control law is defined as a functional relationship while the parameters are computed using empirical data. However, adaptive control typically requires analytical models of the control and adaptation laws. Approximate Dynamic Programming (ADP) [Bertsekas, 2005; Powell, 2007] has a similar dependence on analytical forms of the value function, at least as a weighted sum of basis functions for solving the projected Bellman equation. It also requires models of the state transition probabilities and stage costs, which may not be available in the current context, since the noise and system dynamics are assumed to be unknown. The closest form of ADP for problems of the current type is the literature on Adaptive Critics [Si *et al.*, 2004], which has considerable overlap with reinforcement learning.

ADP in the policy space solves the problem by using policy gradients to compute the optimal policy parameters, each of which defines a stationary policy. Actor-critic based reinforcement learning could be viewed as an extension of this approach, where the 'policy parameters' are actually the parameters of the critic and actor networks and are computed using simulation (or equivalently, *trained*). ADP has been used in prior literature for relatively large task allocation problems in transportation networks [Godfrey and Powell, 2002; Topaloglu and Powell, 2006]. These studies use non-linear approximations of the value function, but the forms are still analytically described. Furthermore, they require at least a one-step rollout of the policy. This may not be feasible in the current context, since the dimensionality is high and each ac-

*Contact Author

tion is continuous (or at least finely quantised), and the goal is not to track some reference signal as in the standard linear quadratic regulator (LQR) [Ogata and Yang, 2002].

Imitation learning (IL) is a well-known approach for learning from expert behaviour without having any need of a reward signal and with the simplicity of a supervised learning. This approach assumes that expert decisions have considered all the constraints of the system in order to accomplish the objective. IL has been used in variety of problems including games [Ross and Bagnell, 2010], 3D games [Harmer *et al.*, 2018], and robotics [Duan *et al.*, 2017]. The inherent problems of design complexity and performance limitations apply here as well; to the definition of the expert policy rather than to the IL algorithm. Additionally, the general form of the problem may not admit an obvious expert policy to train with.

Reinforcement learning in related areas: The majority of research in reinforcement learning has been on the use of Deep RL for computing actions in games such as Atari and chess. Deep Q-Network [Mnih *et al.*, 2015] was used to achieve superhuman performance on Atari using raw pixel inputs. Subsequent modifications were proposed to stabilize the training and making it more sample efficient [Van Hasselt *et al.*, 2016; Schaul *et al.*, 2015]. Recently, policy gradient methods have increasingly become the state-of-the-art methods for handling continuous action spaces [Islam *et al.*, 2017]. Deep Deterministic Policy Gradients (DDPG) [Lillicrap *et al.*, 2015] improves upon the basic advantage actor critic [Konda and Tsitsiklis, 2000] by using the actions as inputs to the critic, and using sampled gradients from the critic to update the actor policy. DDPG has been shown to work well on continuous action spaces, although results so far are limited to a few (less than 10) action outputs. Trust Region Policy Optimization [Schulman *et al.*, 2015] and Proximal Policy Optimization [Schulman *et al.*, 2017] have also proven effective for optimal control using RL, but these are on-policy computationally expensive algorithms and are difficult to apply where episodes are not naturally finite-horizon. Approaches for multiple continuous actions such as Branching DQN [Tavakoli *et al.*, 2018] still have significant growth in size with the state space.

In the area of system dynamics, there is significant work in the computation of torque commands for robotic applications [Powell, 2012; Kober *et al.*, 2013] including locomotion [Kohl and Stone, 2004] and manipulation [Theodorou *et al.*, 2010]. A number of these methods are model-based [Nagabandi *et al.*, 2018], because of the availability of accurate dynamic models of the robots. The action spaces are naturally continuous and are either discretised for tractability, or are represented by function approximations. Alternatively, the policy is parameterised for simplicity [Theodorou *et al.*, 2010]. The key point of complexity is the curse of dimensionality, which is much more acute in the current context than in typical robotic applications with fewer than 10 degrees of freedom. Applying methods such as DDPG to problems with hundreds of degrees of freedom is difficult even with recent sample-efficient techniques [Gu *et al.*, 2017]. A recent approach for exploration in large state-action spaces is learning by demonstration [Nair *et al.*, 2018]. However, this requires the equivalent of an expert policy for imitation learning.

Intelligent transportation systems [Bazzan and Klügl, 2013] also require online decisions for managing transportation network operations for maximizing safety, throughput, and efficiency. Adaptive traffic signal control has been a major challenge in transportation systems. In literature, it has been solved by modeling it as a multiple player stochastic game, and solve it with the approach of multi-agent RL [Shoham *et al.*, 2007; Busoniu *et al.*, 2008; El-Tantawy *et al.*, 2013]. However, these approaches are difficult to scale. Other approaches [Khadilkar, 2018] tackle the scalability issue by dividing the global decision-making problem into smaller pieces, with both local and global performance affecting the reward. This is analogous to the current context, since one may choose to decompose the action space for tractability while retaining the objective of maximising global reward.

Inventory management problems: We finally summarise existing literature on supply chain control and inventory management, which is introduced in Section 3 as a specific instance of the problem. The operations research community has addressed this problem in detail because of its commercial implications [Silver, 1981]. Instances at relatively small scale are solved as joint assortment-stocking problems using mixed-integer linear programming [Smith and Agrawal, 2000; Caro and Gallien, 2010]. Implementations at practical business scales typically operate with simple heuristics such as threshold-based policies [Condea *et al.*, 2012]. Adaptive critic [Shervais *et al.*, 2003] and reinforcement learning [Giannoccaro and Pontrandolfo, 2002; Jiang and Sheng, 2009; Mortazavi *et al.*, 2015] approaches are also reported in literature, but tend to focus on single-product problems.

2 Results

For an additional validation of the methodology, we perform an analogous training and testing study on a set of 100 products independent of the 220 products in the main manuscript. These products have their own order rates and meta-data. As shown in Figure 1, the learning trend between A2C and the proportional heuristic is of a similar nature. Other methods were not implemented because of a lack of time, but these can be populated in the final version of the manuscript.

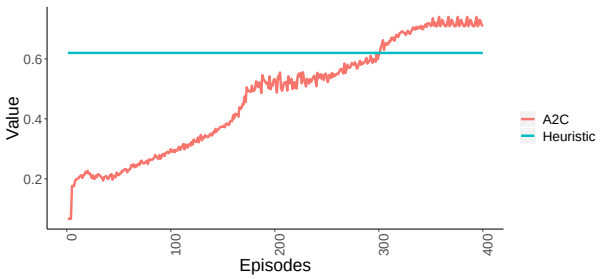

Figure 1: Mean rewards over the course of training for 100 products

Figure 2 characterises the A2C learning trends in more detail. We see that the internal reward approaches the actual business reward as training progresses, and the algorithm

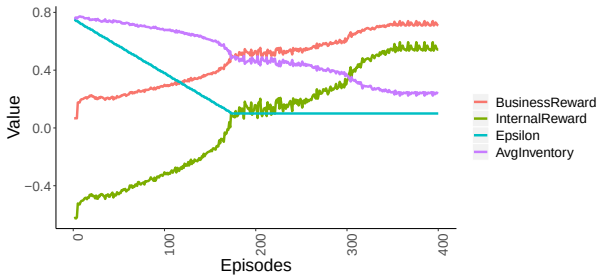

Figure 2: Components of A2C performance during training with 100 products.

learns to minimise the additional penalties imposed by (14) in addition to the actual reward in (12).

References

[Åström and Wittenmark, 2013] Karl J Åström and Björn Wittenmark. *Adaptive control*. Courier Corporation, 2013.

[Bazzan and Klügl, 2013] Ana LC Bazzan and Franziska Klügl. Introduction to intelligent systems in traffic and transportation. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 7(3):1–137, 2013.

[Bertsekas, 2005] Dimitri P Bertsekas. *Dynamic programming and optimal control, Chapter 6*, volume 1. Athena scientific Belmont, MA, 2005.

[Bestle and Zeitz, 1983] David Bestle and M Zeitz. Canonical form observer design for non-linear time-variable systems. *International Journal of control*, 38(2):419–431, 1983.

[Bouabdallah *et al.*, 2004] Samir Bouabdallah, Andre Noth, and Roland Siegwart. Pid vs lq control techniques applied to an indoor micro quadrotor. In *Proc. of The IEEE International Conference on Intelligent Robots and Systems (IROS)*, pages 2451–2456. IEEE, 2004.

[Busoniu *et al.*, 2008] Lucian Busoniu, Robert Babuska, and Bart De Schutter. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, And Cybernetics-Part C: Applications and Reviews, 38 (2), 2008*, 2008.

[Caro and Gallien, 2010] Felipe Caro and Jérémie Gallien. Inventory management of a fast-fashion retail network. *Operations Research*, 58(2):257–273, 2010.

[Condea *et al.*, 2012] Cosmin Condea, Frédéric Thiesse, and Elgar Fleisch. Rfid-enabled shelf replenishment with backroom monitoring in retail stores. *Decision Support Systems*, 52(4):839–849, 2012.

[De Oliveira and Camponogara, 2010] Lucas Barcelos De Oliveira and Eduardo Camponogara. Multi-agent model predictive control of signaling split in urban traffic networks. *Transportation Research Part C: Emerging Technologies*, 18(1):120–139, 2010.

[Doyle *et al.*, 1989] John C Doyle, Keith Glover, Pramod P Khargonekar, and Bruce A Francis. State-space solutions to standard h/sub 2/and h/sub infinity/control problems. *IEEE Trans. on Automatic control*, 34(8):831–847, 1989.

[Doyle, 1978] John Doyle. Guaranteed margins for lqg regulators. *IEEE Trans. on Automatic Control*, 23(4):756–757, 1978.

[Duan *et al.*, 2017] Yan Duan, Marcin Andrychowicz, Bradly Stadie, Jonathan Ho, Jonas Schneider, Ilya Sutskever, Pieter Abbeel, and Wojciech Zaremba. One-shot imitation learning. In *NIPS*, volume 31, 2017.

[El-Tantawy *et al.*, 2013] Samah El-Tantawy, Baher Abdulhai, and Hossam Abdelgawad. Multiagent reinforcement learning for integrated network of adaptive traffic signal controllers (MARLIN-ATSC). *IEEE Transactions on Intelligent Transportation Systems*, 14(3):1140–1150, 2013.

[Giannoccaro and Pontrandolfo, 2002] Ilaria Giannoccaro and Pierpaolo Pontrandolfo. Inventory management in supply chains: a reinforcement learning approach. *International Journal of Production Economics*, 78(2):153–161, 2002.

[Godfrey and Powell, 2002] Gregory A Godfrey and Warren B Powell. An adaptive dynamic programming algorithm for dynamic fleet management, i: Single period travel times. *Transportation Science*, 36(1):21–39, 2002.

[Golnaraghi and Kuo, 2010] F Golnaraghi and BC Kuo. Automatic control systems. *Complex Variables*, 2:1–1, 2010.

[Gu *et al.*, 2017] Shixiang Gu, Ethan Holly, Timothy Lillicrap, and Sergey Levine. Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on*, pages 3389–3396. IEEE, 2017.

[Gustafsson and Waller, 1983] Tore K Gustafsson and Kurt V Waller. Dynamic modeling and reaction invariant control of ph. *Chemical Engineering Science*, 38(3):389–398, 1983.

[Harmer *et al.*, 2018] Jack Harmer, Linus Gisslen, Jorge del Val, Henrik Holst, Joakim Bergdahl, Tom Olsson, Kristoffer Sjoo, and Magnus Nordin. Imitation learning with concurrent actions in 3d games. *arXiv preprint arXiv:1803.05402*, 2018.

[Ioannou and Sun, 1996] Petros A Ioannou and Jing Sun. *Robust adaptive control*, volume 1. PTR Prentice-Hall Upper Saddle River, NJ, 1996.

[Islam *et al.*, 2017] Riashat Islam, Peter Henderson, Maziar Gomrokchi, and Doina Precup. Reproducibility of benchmarked deep reinforcement learning tasks for continuous control. *arXiv preprint arXiv:1708.04133*, 2017.

[Jiang and Sheng, 2009] Chengzhi Jiang and Zhaohan Sheng. Case-based reinforcement learning for dynamic inventory control in a multi-agent supply-chain system. *Expert Systems with Applications*, 36(3):6520–6526, 2009.

[Khadilkar, 2018] Harshad Khadilkar. A scalable reinforcement learning algorithm for scheduling railway lines. *IEEE Transactions on Intelligent Transportation Systems*, 2018.

[Kober *et al.*, 2013] Jens Kober, J Andrew Bagnell, and Jan Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.

[Kohl and Stone, 2004] Nate Kohl and Peter Stone. Policy gradient reinforcement learning for fast quadrupedal locomotion. In *Robotics and Automation, 2004. Proceedings. ICRA'04. 2004 IEEE International Conference on*, volume 3, pages 2619–2624. IEEE, 2004.

[Konda and Tsitsiklis, 2000] V Konda and J Tsitsiklis. Actor-critic algorithms. In *Advances in Neural Information Processing Systems*, pages 1008–1014, 2000.

[Levant *et al.*, 2000] A Levant, A Pridor, R Gitizadeh, I Yaesh, and JZ Ben-Asher. Aircraft pitch control via second-order sliding technique. *Journal of Guidance, Control, and Dynamics*, 23(4):586–594, 2000.

[Lillicrap *et al.*, 2015] Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *CoRR*, abs/1509.02971, 2015.

[Liu and Chan, 2006] X-J Liu and CW Chan. Neuro-fuzzy generalized predictive control of boiler steam temperature. *IEEE Transactions on energy conversion*, 21(4):900–908, 2006.

[Ljung, 1998] Lennart Ljung. System identification. In *Signal analysis and prediction*, pages 163–173. Springer, 1998.

[Mayne *et al.*, 2000] David Q Mayne, James B Rawlings, Christopher V Rao, and Pierre OM Scokaert. Constrained model predictive control: Stability and optimality. *Automatica*, 36(6):789–814, 2000.

[Mnih *et al.*, 2015] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep RL. *Nature*, 518(7540):529, 2015.

[Mortazavi *et al.*, 2015] Ahmad Mortazavi, Alireza Arshadi Khamseh, and Parham Azimi. Designing of an intelligent self-adaptive model for supply chain ordering management system. *Engineering Applications of Artificial Intelligence*, 37:207–220, 2015.

[Nagabandi *et al.*, 2018] Anusha Nagabandi, Gregory Kahn, Ronald S Fearing, and Sergey Levine. Neural network dynamics for model-based deep rl with model-free fine-tuning. In *ICRA*, pages 7559–7566. IEEE, 2018.

[Nair *et al.*, 2018] Ashvin Nair, Bob McGrew, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Overcoming exploration in reinforcement learning with demonstrations. In *ICRA*, pages 6292–6299. IEEE, 2018.

[Ogata and Yang, 2002] K Ogata and Y Yang. *Modern control engineering*, volume 4. Prentice Hall, 2002.

[Pathak *et al.*, 2005] Kaustubh Pathak, Jaume Franch, and Sunil Kumar Agrawal. Velocity and position control of a wheeled inverted pendulum by partial feedback linearization. *IEEE Transactions on robotics*, 21(3):505–513, 2005.

[Powell, 2007] Warren B Powell. *Approximate Dynamic Programming: Solving the curses of dimensionality*, volume 703. John Wiley & Sons, 2007.

[Powell, 2012] Warren B Powell. Ai, or and control theory: A rosetta stone for stochastic optimization. *Princeton University*, 2012.

[Ross and Bagnell, 2010] Stéphane Ross and J. Andrew Bagnell. Efficient reductions for imitation learning. In *Proc. of The International Conference Artificial Intelligence and Statistics (AISTATS)*, 2010.

[Schaul *et al.*, 2015] Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. *arXiv preprint arXiv:1511.05952*, 2015.

[Schulman *et al.*, 2015] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International Conference on Machine Learning*, pages 1889–1897, 2015.

[Schulman *et al.*, 2017] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[Shervais *et al.*, 2003] Stephen Shervais, Thaddeus T Shannon, and George G Lendaris. Intelligent supply chain management using adaptive critic learning. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 33(2):235–244, 2003.

[Shoham *et al.*, 2007] Yoav Shoham, Rob Powers, Trond Grenager, et al. If multi-agent learning is the answer, what is the question? *Artificial Intelligence*, 171(7):365–377, 2007.

[Si *et al.*, 2004] Jennie Si, Andrew G Barto, Warren B Powell, and Don Wunsch. *Handbook of learning and approximate dynamic programming*, volume 2. John Wiley & Sons, 2004.

[Silver, 1981] Edward A Silver. Operations research in inventory management: A review and critique. *Operations Research*, 29(4):628–645, 1981.

[Smith and Agrawal, 2000] Stephen A Smith and Narendra Agrawal. Management of multi-item retail inventory systems with demand substitution. *Operations Research*, 48(1):50–64, 2000.

[Tavakoli *et al.*, 2018] Arash Tavakoli, Fabio Pardo, and Petar Kormushev. Action branching architectures for deep reinforcement learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[Theodorou *et al.*, 2010] Evangelos Theodorou, Jonas Buchli, and Stefan Schaal. Reinforcement learning of motor skills in high dimensions: A path integral approach. In *Robotics and Automation (ICRA)*, pages 2397–2403. IEEE, 2010.

[Topaloglu and Powell, 2006] H Topaloglu and W Powell. Dynamic-programming approximations for stochastic time-staged integer multicommodity-flow problems. *INFORMS Journal on Computing*, 18(1):31–42, 2006.

[Utkin *et al.*, 2009] Vadim Utkin, Jürgen Guldner, and Jingxin Shi. *Sliding mode control in electro-mechanical systems*. CRC press, 2009.

[Van der Laan and Salomon, 1997] Erwin Van der Laan and Marc Salomon. Production planning and inventory control with remanufacturing and disposal. *European Journal of Operational Research*, pages 264–278, 1997.

[Van Hasselt *et al.*, 2016] Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. In *AAAI*, volume 2, page 5. Phoenix, AZ, 2016.