
The Potential of Reinforcement Learning for Live Musical Agents

Keywords: interactive musical agents, reinforcement learning, computer music

Nick Collins

N.COLLINS@SUSSEX.AC.UK

Department of Informatics, University of Sussex, Falmer, Brighton, BN1 9QJ, UK

Abstract

Reinforcement learning has great potential applicability in computer music, particularly for interactive scenarios and the production of effective control policies for systems. This paper considers in particular the case of interactive music systems, where a software agent can be trained online during a rehearsal session with a musician. A small-scale system is described which uses the Sarsa(λ) algorithm to update state-action values to determine a policy for output behaviours over a modest feature space. It is hoped that issues arising can be discussed fruitfully in a workshop context.

1. Introduction

One holy grail of computer music is the creation of a live musical agent whose behaviours are fully commensurate with human music making, but whose architecture is not of flesh and blood (Rowe, 2001; Collins, 2007). Whilst many projects have been undertaken, most have sought out the stimulating but inhuman properties of certain algorithms for their own sake, or have otherwise failed to live up to stricter definitions of autonomous agency (Collins, 2006). The most promising research to date is probably that which utilises machine learning techniques (Thom, 2003; Hamanaka et al., 2003; Assayag et al., 2006). This paper considers the potential in particular of reinforcement learning (RL) (Sutton & Barto, 1998), as suitably equipped for the situation of an agent exploring a musical environment in realtime. Whilst the research presented here is hardly conclusive, it may stimulate consideration of the potential benefits and pitfalls of RL approaches.

In prior work, Franklin and Manfredi (Franklin & Manfredi, 2002) studied actor-critic reinforcement learning in a jazz improvisation context, generating notes and rating actions with respect to a basic hard-coded jazz music theory. RL has also been explored in the OMax research project at IRCAM ([\[recherche.ircam.fr/equipes/repmus/OMax/\]\(http://recherche.ircam.fr/equipes/repmus/OMax/\)\). In the closing stages of a paper, Assayag et al. \(2006\) weight links in the Factor Oracle algorithm using discrete state-action RL; the reward signal in online interaction is heightened performer attention to particular materials. An offline mode feeds back the system's generated material to itself with negative reinforcement to increase variety in productions.](http://</p></div><div data-bbox=)

To effect learning in RL, a reward signal is necessary to clarify the success of any action choice, given the state. Various RL algorithms exist; the Sarsa(λ) variant utilised herein allows the back-propagation of reward values along the set of recent state-action pairs with falloff of influence controlled by λ . Salient notions of reward for interactive music systems might include quality of anticipation, influence on other parties (perhaps in take-up of material), and direct feedback from participants or observers on their experience. Human or objective fitness functions can also be used, analogously to interactive genetic algorithm's human assessment bottleneck as against programmed criteria. Human feedback is hard to solicit, however; whilst physiological sensors such as GSR or EEG may provide continuous measures of emotional state, their trustworthiness remains to be proven in future investigation. Whilst in other recent work I have explored signals based on 'prediction' ability and a degree of 'consequence', this article considers direct user entry of quality ratings, despite potential concerns on this task being a distractor from playing.

2. A test case for online tuition using Sarsa(λ)

Whilst experiments with rather more complicated state-action spaces and alternative reward signals have been carried out with *Improvagent*¹, a simpler study is presented here in order to provide a stimulus for discussion. In this symbolic (MIDI) system, a user plays the keyboard, and various features are abstracted within a time window of three seconds modelling the

¹Paper to be presented at ICMC 2008

perceptual present; the state is updated once per second. This external data is the current environment state to which the system must respond with an action. To keep the setting manageable, three features are extracted (log density, pitch variance and log latest event time) and five bands allowed to partition their values. Five behaviours are defined (0=silence, 1-4 are small imitative/processing algorithms exploiting collected note events from the current window). The state-action space thus has 625 (5^4) values to determine for a performance policy. The performer enters rewards at any time on the computer keyboard, using one key for positive reward and one for negative reward, reminiscent of Al Biles' online IGA for GenJam.

Testing the system revealed how closely the coverage of states depended on fine representational decisions; for instance, the values particular features take on can be decidedly non-uniform. Even after scaling adjustments, only 20% of states were touched in a typical ten minute interaction session. In Sarsa(λ) training, a ten minute session would give only 600 cases; an offline process of re-running these cases in internal rehearsal can be used to refine estimates to speed up convergence. Without higher level musical goals and extensive long-term training, it was hard to tell how long-term interaction quality was affected; but it was possible to stamp out certain aberrant behaviours, and encourage favoured modes, by feeding back negative and positive reward.

3. Discussion

Reinforcement learning experts have debunked some common myths concerning RL schemes, such as the slowness of training, the divergence under function approximation and that a strict MDP is required (<http://neuromancer.eecs.umich.edu/cgi-bin/twiki/view/Main/MythsofRL>). It must be admitted, however, that human performers can be wilfully non-deterministic, taking a different choice given the same situation; it is hoped that the statistics of their own policies can be tractably modelled. Whilst as noted by Belinda Thom (Thom, 2003) any given musical interaction may offer an extreme case of sparse data, the extensive practice regimes of thousands of hours carried out by expert human musicians may be offered as a counter-example to illustrate the amount of training data potentially required (Deliège & Sloboda, 1996). A pertinent question is how best to scale up small systems into larger systems worthy of the investment of such time, and how to train in non-interactive situations (from MIDI file data, say) in readiness for real interaction.

The choice of representation is critical as ever, and it may be productive to consider more closely how human musicians manage to encode and recall musical data, that is, to employ psychological data on human learning and memory. Chunking of data (storage as manageable chunks of around three notes at a time with context in a memory record) may be an ecologically valid tactic for keeping dimensionality low.

4. Conclusions

Reinforcement learning has potential for learning in interactive music systems, but with many outstanding issues concerning nature (bootstrapping) and nurture (learning in an environment) and associated representations and modelling. There are also questions of evaluation for the systems themselves, which can be approached on a number of levels, from participant experience, to audience and critical response, alongside more objective technical criteria.

References

- Assayag, G., Bloch, G., Chemillier, M., Cont, A., & Dubnov, S. (2006). OMax brothers: a dynamic topology of agents for improvisation learning. *AMCMM '06: Proceedings of the 1st ACM workshop on audio and music computing multimedia* (pp. 125–132).
- Collins, N. (2006). *Towards autonomous agents for live computer music: Realtime machine listening and interactive music systems*. Doctoral dissertation, University of Cambridge.
- Collins, N. (2007). Musical robots and listening machines. In N. Collins and J. d'Esquivan (Eds.), *Cambridge companion to electronic music*, 171–84. Cambridge: Cambridge University Press.
- Deliège, I., & Sloboda, J. (1996). *Musical beginnings: Origins and development of musical competence*. New York: Oxford University Press.
- Franklin, J. A., & Manfredi, V. U. (2002). Nonlinear credit assignment for musical sequences. *Second international workshop on Intelligent systems design and application* (pp. 245–250).
- Hamanaka, M., Goto, M., Asoh, H., & Otsu, N. (2003). A learning-based jam session system that imitates a player's personality model. *IJCAI: International Joint Conference on Artificial Intelligence* (pp. 51–58).
- Rowe, R. (2001). *Machine musicianship*. Cambs, MA: MIT Press.
- Sutton, R., & Barto, A. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.
- Thom, B. (2003). Interactive improvisational music companionship: A user-modeling approach. *User Modeling and User-Adapted Interaction Journal*, 13, 133–77.