On-line policy optimisation of spoken dialogue systems via live interaction with human subjects

Milica Gašić, Filip Jurčiček, Blaise Thomson, Kai Yu and Steve Young
Cambridge University Engineering Department
{mg436,fj228,brntt2,ky219,sfy}@eng.cam.ac.uk

1 INTRODUCTION

- Hidden Information State system
  - POMDP-based dialogue manager that maintains a distribution over possible states at every dialogue turn
  - Optimises the policy in a smaller summary space
  - Requires a large number of dialogues to train a policy
  - Relies on the use of a user simulator
- Gaussian process reinforcement learning
  - POMDP dialogue policy maps (summary) states b into actions a so that the total reward is maximal:
    \[ Q(b, a) = \max_a E_r \sum_{t=1}^{T} \gamma^{t-1} r_t | b_0 = b, a_0 = a \]
  - Q-function can be modelled as a Gaussian process (GP), which for every summary state-action pair \((b, a)\) gives a Gaussian distribution \(N(Q(b, a), \text{cov}((b, a), (b, a)))\)
  - This enables faster policy optimisation.
- We investigate policy optimisation directly from human interaction
  - Using a low risk learning technique based on GPs,
  - Via Amazon Mechanical Turk service,
  - To replace the need for a user simulator.

2 LOW-RISK POLICY MODEL

On-line learning requires manual balancing of exploitation of current estimate of the Q-function and exploration of unexplored actions.

- We propose a stochastic policy model which
  - Automatically balances exploration/exploitation via sampling from Gaussian distributions for \(Q(b, a)\) for every action \(a\) and taking the action which has the highest sampled Q-value:
    \[ Q(b, a) \sim N(Q(b, a), \text{cov}((b, a), (b, a))) \]
    \[ a = \arg \max_a Q(b, a) \]
    - This reduces the risk of taking bad actions during learning.

3 ON-LINE LEARNING

- Experimental Set-up
  - 252 users were recruited via Amazon Mechanical Turk and provided with a dialogue task.
  - They called a telephone-based dialogue system for Cambridge restaurant domain.
  - 2960 dialogues were collected.
  - Users gave a binary feedback at the end of every dialogue for on-line learning reward.
- Initial training
  - Performance of the policy that is learning on-line during initial 680 dialogues:
    - Partial task completion – the system partially completed the dialogue task
    - Exhibits
      - Variability in performance
      - Divergence between objective and subjective measures

4 CONCLUSION

- Policy trained on-line reached the performance of a policy trained on a simulated user.
- Once the policy reached reasonable behaviour it is difficult for the users to estimate the reward accurately.
- While the framework deals well with noisy inputs from the recogniser, inaccuracy of the assigned reward can lead to variability in performance.