Q-Learning in Continuous State Action Spaces

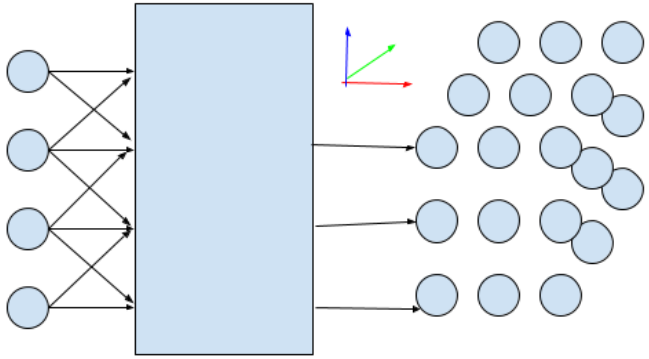
Alex Irpan

Motivation

- New trend of using deep neural nets to represent policies in MDPs
- Deep reinforcement learning success, but small discrete action space.
- Want to apply similar ideas to continuous problems.

Discretization

- Bucket range into discrete options
- Problem: exponentially large
- Solution: add conditional independence



Independence Assumption

- Similar to directed graphical models.
- Represent joint action space compactly
- For Q-learning, need easily computable max, suggests additive basis functions.
- Best case, N^D outputs -> ND ouputs

Modifications to Algorithm

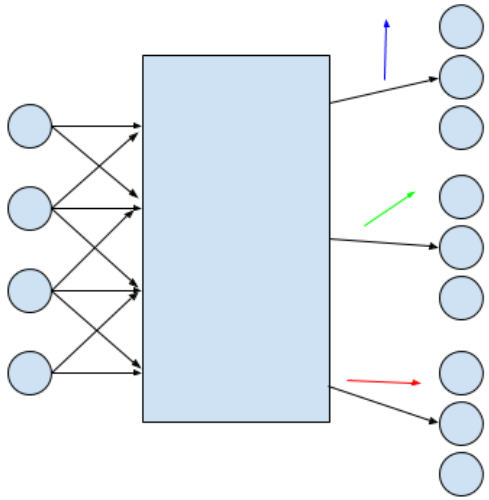
$$\ell^{(n)}(\theta) = \frac{1}{2}((R_t + \gamma \max_a Q_{\theta^{(n)}}(s_{t+1}, a)) - Q_\theta(s_t, a_t))^2$$

$$\theta^{(n+1)} \leftarrow \theta^{(n)} - \eta \nabla_\theta \ell^{(n)}(\theta^{(n)})$$

$$\text{If } Q(s,a) = \sum_i Q_i(s, a(i))$$

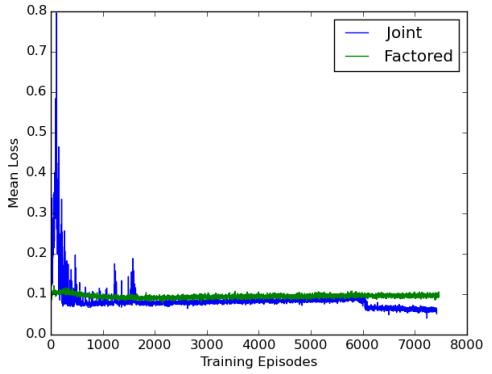
$$\text{Then } \max_a Q(s,a) = \sum_i \max_a Q_i(s, a(i))$$

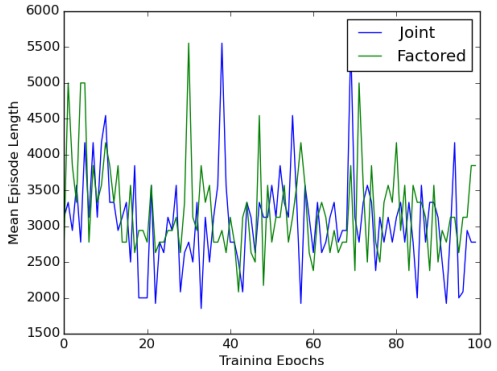
$$\nabla_\theta Q_\theta(s,a) = \sum_i \nabla_\theta Q_i(s, a(i))$$

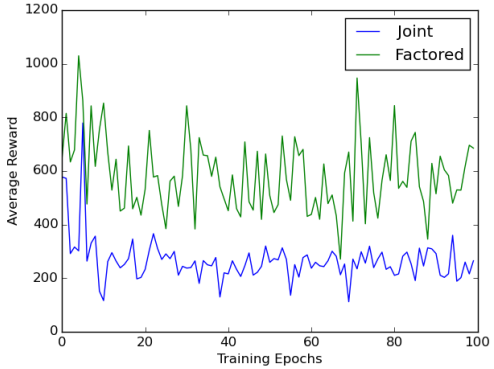


Extendable to conditionally independent actions, with analogous max and gradient computation. See [1]

Experimental Results







Further Work

- Policy gradient methods
  - Stochastic policy, actions form joint probability distribution
  - Represent with directed model

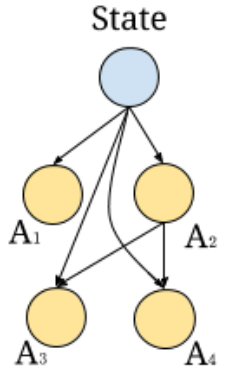
$$\text{For trajectory } \tau = (s_0, a_0, s_1, a_1, \cdots)$$

$$\ell(\tau) = E[R(\tau)]$$

$$\nabla_\theta \ell(\tau) = E\left[R(\tau) \sum_t \nabla_\theta \log \pi(a_t|s_t)\right]$$

$$\text{If } \pi(a|s) = \prod_i \pi_i(a(i)|\text{ parents}(a(i)), s)$$

$$\text{Then } \nabla_\theta \log \pi(a|s) = \sum_i \nabla_\theta \log \pi_i(a(i)|\text{ parents}(a(i)), s)$$



Also looking into wire fitting methods

- Directly exploits continuous nature
- Wires as guides, interpolation to generalize.
- Factorization may also apply to this method
- See [2] for more details

References

[1] Guestrin, Carlos, et al. "Efficient solution algorithms for factored MDPs." *Journal of Artificial Intelligence Research* (2003): 399-468.

[2] Gaskett, Chris, David Wettergreen, and Alexander Zelinsky. "Q-learning in continuous state and action spaces." *Australian Joint Conference on Artificial Intelligence*. 1999.