


Trust Region Policy Optimization

$q(\cdot/S_n)$ with a support that includes the support of $q_i(\cdot/S_n)$ will produce a consistent estimator. In practice, we found that q

Joint angles and kinematics



Trust Region Policy Optimization

| | <i>B. Rider</i> | <i>Breakout</i> | <i>Enduro</i> | <i>Pong</i> | <i>Q*bert</i> | <i>Seaquest</i> | <i>S. Invaders</i> |
|-------------------------------------|-----------------|-----------------|---------------|-------------|---------------|-----------------|--------------------|
| Random | 354 | 1.2 | 0 | - 20.4 | 157 | 110 | 179 |
| Human (Mnih et al., 2013) | 7456 | 31.0 | 368 | - 3.0 | 18900 | 28010 | 3690 |
| Deep Q Learning (Mnih et al., 2013) | 4092 | 168.0 | 470 | 20.0 | 1952 | 1705 | 581 |
| UCC-I (Guo et al., 2014) | 5702 | 380 | 741 | 21 | 20025 | 2995 | 692 |
| TRPO - single path | 1425.2 | 10.8 | 534.6 | 20.9 | 1973.5 | 1908.6 | 568.4 |

References

- Bagnell, J. A. and Schneider, J. Covariant policy search. IJCAI, 2003.
- Bartlett, P. L. and Baxter, J. Infinite-horizon policy-gradient estimation. *arXiv preprint arXiv:1106.0665*