2nd International Symposium on
Information Geometry and its Applications
December 12-16, 2005, Tokyo

UTILIZING NATURAL GRADIENT IN TEMPORAL DIFFERENCE REINFORCEMENT LEARNING WITH ELIGIBILITY TRACES

TETSURO MORIMURA[1,2], EIJI UCHIBE[2], AND KENJI DOYA[1,2,3]

1:Nara Institute of Science and Technology

2:Okinawa Institute of Science and Technology

3:ATR Computational Neuroscience Laboratories

The method of policy gradient is a promising approach in reinforcement learning (RL) for optimizing action policy parameters to maximize cumulative rewards. The natural gradient method has been applied to policy gradient RL, but the original methods suffered from numerical instability by matrix inversion. Here we show that the natural policy gradient can be estimated without matrix inversion by regressing the temporal difference (TD) reward prediction errors by a set of basis functions given by the parameterization of the policy. We further demonstrate that the bias in the estimate can be reduced by the use of 'eligibility traces' of parameters. The proposed method, the natural temporal difference (NTD) algorithm, is applied to two simple Markov decision problems and a more challenging non-linear pendulum control problem to show its effectiveness.