Entropy annealing for policy mirror descent in continuous time and space

David Siska

Entropy regularization has been extensively used in policy optimization algorithms to enhance exploration and the robustness of the optimal control, however it introduces an additional regularization bias. This work quantifies the impact of entropy regularization on the convergence of policy gradient methods for stochastic exit time control problems. We analyze a continuous-time policy mirror descent dynamics, which updates the policy based on the gradient of an entropy-regularized value function and adjusts the strength of entropy regularization as the algorithm progresses. This leads to a gradient flow over the infinite dimensional space of Markov kernels.
We prove that with a fixed entropy level, the dynamics converges exponentially to the optimal solution of the regularized problem.
We further show that when the entropy level decays at suitable polynomial rates, the annealed flow converges to the solution of the unregularized problem at a rate of $\mathcal O(1/S)$ for discrete action spaces and, under suitable conditions, at a rate of $\mathcal O(1/\sqrt{S})$ for general action spaces, with S being the gradient flow time. This paper explains how entropy regularization improves policy optimization, even with the true gradient, from the perspective of convergence rate.
This is joint work with D. Sethi (Edinburgh) and Y. Zhang (Imperial).