The conditioning effect of stochastic dynamics in continuous reinforcement learning

Yuval Tassa,
Hebrew University, Jerusalem
tassa@alice.nc.huji.ac.il

Teaching robots to perform complex behaviors using reinforcement learning (RL) algorithms is a long-term goal of the machine learning community. This goal is underwritten by our familiarity with nervous systems which evidently implement some form of RL, while simultaneously setting the highest benchmark for motor control. When attempting to learn optimal controllers, *stochasticity* is often considered a hurdle to be overcome. Methods which deal with uncertainty, whether process noise, measurement noise or modeling uncertainty are seen as extensions of the observable deterministic case, as exemplified by the stochastic and robust extensions to linear-quadratic control.

Perhaps surprisingly, the assumption of stochastic dynamics can also *simplify* learning algorithms. Intuitively, the smoothing of our beliefs about the world into a distribution function entails a limiting of the spatial bandwidth of our predictions and therefore an effective reduction of dimensionality, which can consequently be exploited.

In order to benefit from this low-pass effect, our model and algorithm must be spatially continuous. An important class of algorithms which take advantage of continuity and smoothness are those which construct second-order approximations of their parameters, as epitomized by the classic *Newton's method*. When implementing second-order algorithms, stochasticity of the dynamics is shown to reduce the condition number of Hessian matrices, a chief measure of convergence quality. We present two recent results where such phenomena occur in two very different settings.

Restricting ourselves to the dynamic programming (DP) framework, we consider algorithms which learn the *value function*, or solve the Hamilton Jacobi Bellman (HJB) equation. Using a general-purpose feedforward neural network to approximate the value function, we apply a Levenberg-Marquardt algorithm to minimize the squared HJB residual [1]. When a Brownian noise term is introduced into the dynamics, an additional second-derivative term is added to the HJB equation. We show how the addition of this term to the residual, *without injecting any actual noise*, smoothes the value function and reduces the condition number of the Hessian by orders of magnitude.

In the second example we use Differential Dynamic Programming (DDP), a method which iteratively computes an explicit second-order model of the value function along a trajectory [2]. At points where small changes in the policy have large effects on the value, the local Hessian matrices can become ill-conditioned and lead to divergence or slow convergence. We show how the introduction of worst-case minimax noise in the H^{∞} control framework also has a conditioning effect on these local matrices.

[1] Tassa & Erez (2007). IEEE Transactions on Neural Networks, 18(4):1031-1041
[2] Tassa, Erez & Smart (2007). NIPS 2007 (accepted)