R8/Ishank Juneja/16D070012

February 15, 2020

Controlling legged robots is a challenging problem due to the high dimensional state spaces arising from numerous degrees of freedom. In addition, there is also the issue of maintaining coordination between actions for the purpose of balance and stability. The paper by Kohl and Stone tackles the problem of optimizing the gait of a robotic quadruped. In particular, they maximize the forward velocity of the quadruped using policy-gradient methods.

The robot used in this work is the Sony Aibo platform. The robots gait (manner of walk) is determined by the positions of three joints located in every leg. A certain strategy for controlling the extent of rotation in these joints would effect the locus of the centre of the robot's foot among other things. The paper explains that the quadruped's gait can be parametrized effectively in terms of the loci of all its feet. In this paper the author's have chosen a half elliptical locus as their choice of gait parametrization. To characterize this gait, four parameters are needed, the length of the ellipse, its peak height above the ground and the position of the ellipse on the $x - y$ plane. On introducing additional parameters such as time taken to complete the locus and the fraction of this time spent by the foot making contact with the ground, the paper makes the problem boil down to a total of 12 parameters. The parameter set implicitly models constraints for considerations such as straight line motion and stability.

The gait policy parametrization described above reduces the task to a 12-dimensional optimization problem where the objective is to maximize forward velocity. A good solution to learning an optimal policy for this task would be a policy gradient approach, however, the major obstacle to this is the unknown functional form of the policy (in terms of the 12 parameters), which makes evaluating its gradient vector a challenge. The usual work around to this obstacle would be the numerical estimation of the gradient. However, this approach is not applicable due to the unavailability of an accurate simulator for the Aibo. To estimate the gradient for a certain policy π parametrised by θ, the authors evaluate policies which are randomly perturbed versions of π. Specifically, each component of θ, is perturbed by either $\epsilon, -\epsilon$ or 0 each with equal probability. Once all the randomly perturbed policies are evaluated, the partial derivatives with respect to each parameter θ_n are estimated as the average of the score associated with policies perturbed in one of two directions (ϵ or $-\epsilon$). After estimating the partial derivatives, an *adjustment vector* A is computed and is used to perform policy updates.

Since all policy evaluations have to occur on actual robots, the authors parallelised evaluation by deploying three Aibo quadrupeds onto the learning task. Further, a significantly large step size of $\eta = 2$ was used by the team. This was likely done with the hope of reaching an optimum faster. On performing the learning process, the robot's velocity plateaus in about three hours of learn time with a final policy that makes the robot move at 291 mm/s. The paper's method learns a gait that is a significant improvement over the previous best of 270 mm/s. Overall their method is a promising way to train quadrupeds (such as the Sony Aibo) and even other types of robots, to learn good gait policies with minimal human intervention.

A question I have is -

- The authors haven't explained their choice of random perturbation ϵ in estimating the policy gradient. Did it have to do with resolution with which we could control the robot, since a perturbation of 0.35cm when the initial value is 4.893 (for ellipse locus length) seems a bit too large to me.