The Hannan Penalty

June 2, 2020

The standard ES update for n models indexed by i is given as-

$$\theta \leftarrow \theta + \frac{\alpha}{n\sigma} \sum_{i=1}^{n} F^{(i)} N^{(i)}$$

And for a gradient-based approach, we defined the overall loss as-

$$L = -A_\pi + 0.5 * C_V - 0.01 * E$$

Here A_π is the Actor's loss under policy $\pi(a_t|s_t)$, C_V is the loss of critic given the Value Function V_π and E is the entropy of the action distribution.

ES and gradient-based updates should be combined in a manner so that both policies move towards monotonic improvement, i.e- both set of parameters should move towards a global minimal objective in the weight space. Since, ES is exploration-based, penalizing the agent during evolutions would indicate limited exploration and local convergence. On the other hand, a gradient updates have stronger reward signals. Thus, penalizing these for deviating behaviour seems suitable. Let r_{es} be the maximum reward obtained by the population during ES update steps and r_{grad} be the average reward obtained by the gradient model during its experiences. Then, we can penalize the gradient-based agent using the following-

$$L = -A_\pi + 0.5 * C_V - 0.01 * E + \frac{1}{2}(r_{es} - r_{grad})^2$$

$$L = -A_\pi + 0.5 * C_V - 0.01 * E + \frac{1}{2}(\sum_{t=1}^{T} r_{es}(s_t, a_t) - \sum_{t=1}^{T} r_{grad}(s_t, a_t))^2$$

$$L = -A_\pi + 0.5 * C_V - 0.01 * E + \frac{1}{2}(\sum_{t=1}^{T} \underset{i}{max} r_{es}^{(i)}(s_t, a_t) - \frac{1}{M} \sum_{m=1}^{M} \sum_{t=1}^{T} r_{grad}(s_t, a_t))^2$$

Here M is the number of episodes played by the gradient-based agent. Penalizing the objective in such a manner forces the gradient updates to be in accordance with the ES exploration scheme. We introduce this added penalty as the **Hannan Penalty** which is based on Hannan consistency. The penalty is equivalent to regularization in the sense that it penalizes the agent for obtaining too less or high rewards in comparison to the ES update, hence prserving the global direction.