

Evolve To Control: Evolution-based Soft Actor-Critic for Scalable Reinforcement Learning

Karush Suri

University of Toronto
karush.suri@mail.utoronto.ca

Xiao Qi Shi

RBC Capital Markets
xiaoqi.shi@borealisai.ca

Konstantinos N. Plataniotis

Multimedia Laboratory
University of Toronto
kostas@ece.utoronto.ca

Yuri A. Lawryshyn

CMTE
University of Toronto
yuri.lawryshyn@utoronto.ca

Abstract: Advances in Reinforcement Learning (RL) have successfully tackled sample efficiency and overestimation bias. However, these methods often fall short of scalable performance. On the other hand, genetic methods provide scalability but depict hyperparameter sensitivity to evolutionary operations. We present the Evolution-based Soft Actor-Critic (ESAC), a scalable RL algorithm. Our contributions are threefold; ESAC (1) abstracts exploration from exploitation by combining Evolution Strategies (ES) with Soft Actor-Critic (SAC), (2) provides dominant skill transfer between offsprings by making use of soft winner selections and genetic crossovers in hindsight and (3) improves hyperparameter sensitivity in evolutions using Automatic Mutation Tuning (AMT). AMT gradually replaces the entropy framework of SAC allowing the population to succeed at the task while acting as randomly as possible, without making use of backpropagation updates. On a range of challenging control tasks consisting of high-dimensional action spaces and sparse rewards, ESAC demonstrates state-of-the-art performance and sample efficiency equivalent to SAC. ESAC demonstrates scalability comparable to ES on the basis of hardware resources and algorithm overhead. A complete implementation of ESAC with notes on reproducibility and videos can be found at the project website <https://karush17.github.io/esac-web/>.

Keywords: ESAC, mutation, AMT, policy

1 Introduction

Concepts and applications of Reinforcement Learning (RL) have seen a tremendous growth over the past decade [1]. These consist of applications in arcade games [1], board games [2] and lately, robot control tasks [3]. A primary reason for this growth is the usage of computationally efficient function approximators such as neural networks [4]. Modern-day RL algorithms make use of parallelization to reduce training times [5] and boost agent's performance through effective exploration giving rise to scalable methods [6, 7, 8]. However, a number of open problems such as approximation bias, lack of scalability in the case of long time horizons and lack of diverse exploration restrict the application of scalability to complex control and robot tasks.

State-of-the-art RL algorithms such as Soft Actor-Critic (SAC) [9] maximize entropy which is indicative of continued exploration. However, using a computationally expensive framework limits scalability as it increases the number of backpropagation [10] updates of the overall algorithm. Moreover, tasks consisting of long time horizons have higher computational overhead as a result of long trajectory lengths. For instance, obtaining accurate position estimates [6] over longer horizons require additional computation times which varies linearly with the hardware requirement. Such a variation calls for increased scalability in the RL domain. Lastly, diverse exploration strategies are essential for the agent to navigate its way in the environment and comprehend intricate aspects of

理解 复杂的, 错综的

less visited states[11]. Various modern-day RL methods lack significant exploration [5, 1] which is addressed by making use of meta-controller[12] and curiosity-driven [13] strategies at the cost of sample efficiency and scalability.

Recent advances in RL have leveraged evolutionary computing for effective exploration and scalability [14, 15, 16, 17, 18]. However, these methods often fall short of state-of-the-art performance and depict sensitivity towards their hyperparameters. A common alternative for boosting performance is to combine gradient-based objectives with evolutionary methods [16]. These algorithms allow a population of learners to gain dominant skills [19] from modern-day RL methods and depict state-of-the-art control while demonstrating scalability. However, their applications do not extend to robot control tasks as a result of sensitivity to mutational hyperparameters which still remains an open problem.

We introduce the Evolution-based Soft Actor Critic (ESAC), an algorithm combining ES with SAC for state-of-the-performance equivalent to SAC and scalability comparable to ES. Our contributions are threefold; (1) ESAC abstracts exploration from exploitation by exploring policies in weight space using evolutions and exploiting gradient-based knowledge using the SAC framework. (2) ESAC makes use of soft winner selection function which, unlike the tournament selection used in ERL [16], does not shield winners from mutation. ESAC carries out genetic crossovers in hindsight resulting in dominant skill transfer between members of the population. Lastly, (3) ESAC introduces the Automatic Mutation Tuning (AMT) which maximizes the mutation rate of ES in a small clipped region and provides significant hyperparameter robustness without making use of backpropagation updates.

2 Related Work

Recent advances in RL have been successful in tackling sample-efficiency [9] and approximation bias (also known as overestimation bias) which stems from value of estimates approximated by the function approximator. Overestimation bias is a common phenomenon occurring in value-based methods [20, 21, 22] and can be addressed by making use of multiple critics [23]. This in turn limits scalability of algorithms [6] by increasing the number of gradient-based updates. Moreover, memory complexity of efficient RL methods increases linearly with the expressive power of approximators [24], thus hindering scalability of RL to complex control tasks.

Intersection of RL and Evolutionary methods for scalability has for long been studied in literature [15, 17, 18, 25, 26, 27]. [14] presents the large-scale parallelizable nature of Evolution Strategies (ES). Performance of ES on continuous robot control tasks in comparison to various gradient-based frameworks such as Trust Region Policy Optimization (TRPO) [28] and Proximal Policy Optimization (PPO) [29] have been found comparable. On the other hand, ES falls short of state-of-the-art performance resulting in local convergence and is extremely sensitive to mutation hyperparameters. An alternative to a pure evolution-based approach is a suitable combination of an evolutionary algorithm with a gradient-based method [30], commonly referred to as Evolutionary Reinforcement Learning (ERL) [16]. ERL makes use of selective mutations and genetic crossovers which allow weak learners of the population to inherit skills from strong learners while exploring. ERL methods are scalable to high-dimensional control problems including multi-agent settings [19]. Such an approach is a suitable trade-off between sample efficiency and scalability but does not necessarily introduce mutation robustness. Other methods in literature [15] follow a similar approach but are often limited to directional control tasks which require little mutation. Thus, addressing scalability and exploration while preserving state-of-the-art performance and mutation robustness requires attention from a critical standpoint.

3 Background

3.1 Reinforcement Learning and Soft Actor-Critic

We review the RL setup wherein an agent interacts with the environment in order to transition to new states and observe rewards by following a sequence of actions. The problem is modeled as a finite-horizon Markov Decision Process(MDP) [31] defined by the tuple $(\mathcal{S}, \mathcal{A}, r, P, \gamma)$ where the state space \mathcal{S} and action space \mathcal{A} are continuous, r presents the reward observed by agent such that $r : \mathcal{S} \times \mathcal{A} \rightarrow [r_{min}, r_{max}]$, $P : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \rightarrow [0, \infty)$ presents the unknown transition model consisting

of the transition probability to the next state $s_{t+1} \in \mathcal{S}$ given the current state $s_t \in \mathcal{S}$ and action $a_t \in \mathcal{A}$ at time step t and γ is the discount factor. We consider a policy $\pi_\theta(a_t|s_t)$ as a function of model parameters θ . Standard RL defines the agent’s objective to maximize the expected discounted reward $\mathbb{E}_{\pi_\theta}[\sum_{t=0}^T \gamma^t r(s_t, a_t)]$ as a function of the parameters θ . SAC [9] defines an entropy-based objective [32] $J(\pi_\theta) = \sum_{t=0}^T \gamma^t [r(s_t, a_t) + \lambda \mathcal{H}(\pi_\theta(\cdot|s_t))]$ wherein λ is the temperature coefficient and $\mathcal{H}(\pi_\theta(\cdot|s_t))$ is the entropy exhibited by the policy $\pi(\cdot|s_t)$ in s_t . For a fixed policy, the soft Q-value function can be computed iteratively, starting from any function $Q : \mathcal{S} \times \mathcal{A}$ and repeatedly applying a modified Bellman backup operator \mathcal{T}^π given by $\mathcal{T}^\pi Q(s_t, a_t) = r(s_t, a_t) + \gamma \mathbb{E}_{s_{t+1} \sim P}[V(s_{t+1})]$ where $V(s_t) = \mathbb{E}_{a_t \sim \pi}[Q(s_t, a_t) - \log(\pi(a_t|s_t))]$ is the soft state value function. We consider a parameterized state value function $V_\psi(s_t)$, a soft Q-function $Q_\phi(s_t, a_t)$ and a policy $\pi_\theta(a_t|s_t)$ which can be represented with neural networks with ψ, ϕ and θ being the parameters of these networks.

3.2 Evolution Strategies

We review the Evolution Strategies [14] framework which is motivated by natural evolution. ES is a heuristic search procedure in which a population of offsprings is mutated using random perturbations. Upon mutation, the fitness objective corresponding to each member of the population is evaluated and offsprings with greater scores are recombined to form the population for the next generation. Let n be the number of offsprings in the population. The parameter vectors of the model can then be represented as $\theta_{es,(i)}$ such that $i = 1, 2, \dots, n$. A total of n random perturbations $\epsilon_{(i)}$, $i = 1, 2, \dots, n$ are sampled from a Gaussian distribution $\mathcal{N}(0, 1)$ in order to mutate $\theta_{es,(i)}$ and evaluate the fitness objective $\mathbb{E}_{\epsilon_{(i)} \sim \mathcal{N}(0,1)}[F(\theta_{es,(i)} + \sigma \epsilon_{(i)})] = \frac{1}{\sigma} \mathbb{E}_{\epsilon_{(i)} \sim \mathcal{N}(0,1)}[F(\theta_{es,(i)} + \sigma \epsilon_{(i)}) \epsilon_{(i)}]$. Here, σ is the mutation rate which controls the extent of mutation. In the case of RL, the fitness $F(\theta_{es,(i)})$ are the episodic rewards observed by members of the population.

4 A Motivating Example: The Discrete Cyclic MDP

We consider a long-horizon discrete cyclic MDP as our motivation for the work. The MDP has a state space \mathcal{S}^3 consisting of 3 states- S_0, S_1 and S_2 and a discrete action space \mathcal{A}^3 consisting of 3 actions- clockwise, anticlockwise and stay. The agent starts in state S_0 . The reward function $r : \mathcal{S}^3 \times \mathcal{A}^3$ assigns a reward of +1 for moving clockwise and -1 otherwise. Each episode lasts 2000 timesteps and terminates if the agent reaches the end of horizon or incurs a negative reward. Figure 1 presents the long-horizon discrete cyclic MDP.

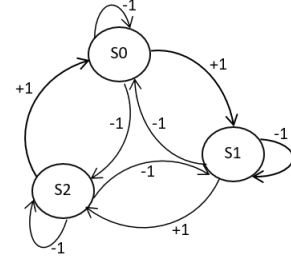


Figure 1: The long-horizon discrete Cyclic MDP. The agent observes a reward of +1 for moving clockwise and -1 otherwise.

The cyclic MDP, being a long-horizon problem, serves as a suitable benchmark for agent’s behavior consisting of minimum computational overhead and is a small-scale replication of policy-search for policy-based agents. The environment consists of a global objective which the agents can achieve if they solve the environment by obtaining the maximum reward of +2000. In order to assess evolution-based behavior, we compare the performance of a population of 50 offsprings utilizing ES with PPO [29] and Deep Deterministic Policy Gradient (DDPG) [3], an efficient off-policy RL method. Although DDPG is primarily a continuous control algorithm, we employ it as a result of the minimal nature of the problem. Figure 2 (left) presents the performance of ES in comparison to gradient-based agents in the cyclic MDP. The ES population presents sample efficiency by solving the task within the first 100 episodes. DDPG, on the other hand, starts solving the task in a lesser sample-efficient manner. The use of a deterministic policy allows DDPG to continuously move left whereas in the case of ES, the population carries out exploration in the weight space and moves along the direction of the strong learners. Lastly, PPO finds a local solution and does not converge towards solving the task. Driven by clipped updates, PPO restricts the search horizon in policy space leading to a sub-optimal policy.

ES has proven to be scalable to large-scale and high dimensional control tasks [14]. We assess this property of ES in the cyclic MDP by varying the operational hardware (number of CPUs) [33] and algorithm overhead (population size). We measure the average wall-clock time per episode [34]. As shown in Figure 2 (center), ES is parallelizable in nature and can be scaled up to larger population sizes by reducing the computation time. The large-scale readily parallelizable nature of ES is a

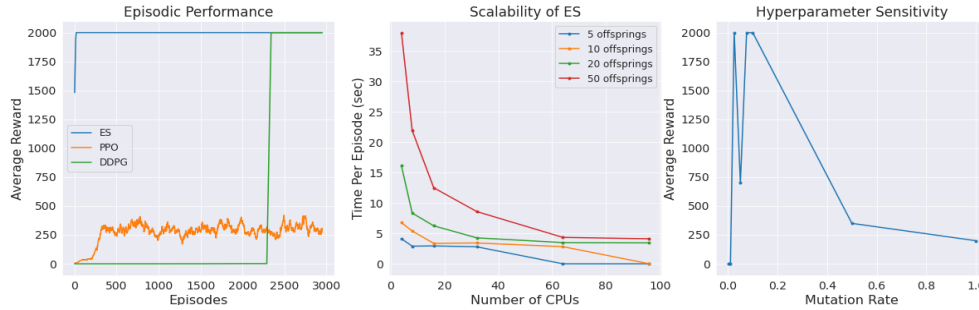


Figure 2: Comparison of ES, DDPG and PPO in the discrete cyclic MDP. ES depicts sample-efficient behavior (left) due to the presence of strong-learners in the population. Scalable nature of ES (center) with readily available computational resources allows in reduction of average episode execution time. However, ES is sensitive to hyperparameters (right) which results inconsistency across different seeds and rigorous fine-tuning.

convincing characteristic for utilizing CPU-based hardware. However, ES relies on excessively sensitive hyperparameters such as mutation rate. Figure 2 (right) presents the sensitivity of ES to mutation rate within a small range with a constant population size of 50. Varying population size does not present a trend in sensitivity indicating that mutation rate is the dominant hyperparameter governing policy behavior among offsprings. Hyperparameter sensitivity requires attention in the case of RL applications such as for real-world robot control [35]. These include excessive tuning of parameters and detailed ablation studies. The cyclic MDP highlights this sensitive nature of ES and serves as a motivating example for tackling sensitivity while preserving state-of-the-art performance in a scalable manner.

5 Evolution-based Continuous Control

The motivation behind ESAC stems from translating the scalability and tackling the mutation sensitivity of ES observed in discrete cyclic MDP to continuous control tasks. ESAC combines the scalable nature of ES with the limited approximation bias of SAC to yield a CPU-friendly state-of-the-art equivalent algorithm.

5.1 Evolution-based Soft Actor-Critic (ESAC)

Overview: Figure 3 provides a high-level schematic of the ESAC algorithm and its components. The population is evaluated in the environment with the fitness metric as episodic rewards obtained by each offspring. Top w winners are then segregated for mutation consisting of ES update followed by crossovers between perturbed offsprings and winners. The new population is formed using crossed-over offsprings and SAC agent. The SAC agent plays its own episodes at fixed timesteps and stores these experiences in a dedicated replay-buffer following policy update. During the SAC update timesteps, ESAC utilizes AMT which maximizes the mutation rate in a clipped region. SAC update timesteps are exponentially annealed to reduce entropy noise and abstract exploration in weight space.

Algorithm: Algorithm 1 presents the ESAC algorithm. We begin by initializing $\psi, \theta, \theta_{es}, \phi$ being the parameters of state-value function, SAC policy, ES policy and Q-function respectively. We then initialize learning rate for SAC agent α , learning rate of ES population α_{es} , mutation rate σ, p_{sac} which is the probability of SAC updates, $\hat{\psi}$ is the parameter vector of the target value function, ζ is the clip parameter, τ is the target smoothing coefficient, e is the fraction of winners and g is the gradient interval. A population of n

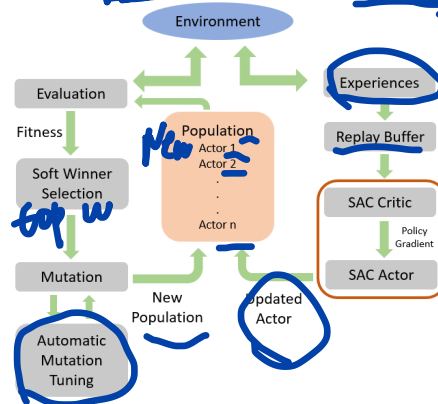


Figure 3: Workflow of ESAC combining ES with SAC. ESAC makes use of soft winner selections, hindsight crossovers and AMT for scalable performance.

Handwritten notes in purple and blue:

- RL \rightarrow heuristics
- optimize the
- choose which is opt or prob
- one
- to optimize
- Pop 1
- Pop 2
- PE1
- PE2

opt prob.

by adjusting heuristics

by RL
ERL

State 1: non-dominated
State 2: dominated by the same pop.

State 3

actors pop_n is initialized along with an empty replay buffer R . Following the main loop, for each offspring i in the population, we draw a noise vector ϵ_i from $\mathcal{N}(0, 1)$ and perturb the ES policy vector θ_{es} to yield the perturbed parameter vector $\theta_{es,(i)}$ as per the expression $\theta_{es,(i)} + \sigma \epsilon_{(i)}$. $\theta_{es,(i)}$ is then evaluated to yield the fitness $F_{(i)}$ as episodic rewards. These are collected in a set F which is normalized and ranked. We now execute soft winner selection wherein the first $w = (n * e)$ offsprings from F are selected for crossovers by forming the set W . The soft winner selection allows dominant skill transfer between winners and next generation offsprings. Mutation is carried out using the ES update [14]. SAC gradient updates are executed at selective gradient intervals during the training process. p_{sac} is exponentially annealed to reduce entropy noise and direct exploration in the weight space. During each g , the agent plays its own episodes by sampling $a_t \sim \pi_\theta(a_t|s_t)$, observing $r(s_t|a_t)$ and s_{t+1} and pushing these experiences to R as a tuple $(s_t, a_t, r(s_t, a_t), s_{t+1})$. Following the collection of experiences, we update the parameter vectors ψ , ϕ and θ by computing $\nabla_\psi J_V(\psi)$, $\nabla_\phi J_Q(\phi_{(i)})$ and $\nabla_\theta J_\pi(\theta)$ where $J_V(\psi)$, $J_Q(\phi_{(i)})$ and $J_\pi(\theta)$ are the objectives of the state-value function, each of the two Q-functions $i \in \{1, 2\}$ and policy as presented in [9] respectively. Gradient updates are followed by AMT update (subsection 5.2) which leads to hindsight crossovers between winners in W and ES policy parameter vector θ_{es} . Finally, the new population is formed using θ , θ_{es} and W .

Algorithm 1 Evolution-based Soft Actor-Critic (ESAC)

- 1: Initialize parameter vectors $\psi, \bar{\psi}, \theta, \theta_{es}, \phi$
- 2: Initialize $\alpha, \alpha_{es}, \sigma, \zeta, \tau, e, g, p_{sac}$
- 3: Initialize a population of n actors pop_n and an empty replay buffer R
- 4: **for** generation=1, ∞ **do**
- 5: **for** $i \in pop_n$ **do**
- 6: sample $\epsilon_{(i)} \sim \mathcal{N}(0, 1)$
- 7: $F_{(i)} \leftarrow$ evaluate $(\theta_{es,(i)} + \sigma \epsilon_{(i)})$ in the environment
- 8: **end for**
- 9: normalize and rank $F_{(i)} \in F$
- 10: select the first $w = (n * e)$ offsprings from F to form the set of winners W
- 11: set $\theta_{es} \leftarrow \theta_{es} + \frac{\alpha_{es}}{n\sigma} \sum_{i=1}^F F_{(i)} \epsilon_{(i)}$
- 12: **if** generation mod $g == 0$ & $\mu \sim \mathcal{N}(0, 1) < p_{sac}$ **then**
- 13: **for** each environment step **do**
- 14: $a_t \sim \pi_\theta(a_t|s_t)$
- 15: observe $r(s_t|a_t)$ and $s_{t+1} \sim P$
- 16: $R \leftarrow R \cup (s_t, a_t, r(s_t, a_t), s_{t+1})$
- 17: **end for**
- 18: **for** each gradient step **do**
- 19: $\psi \leftarrow \psi - \alpha \nabla_\psi J_V(\psi)$
- 20: $\phi \leftarrow \phi - \alpha \nabla_\phi J_Q(\phi_{(i)})$ for $i \in \{1, 2\}$
- 21: $\theta \leftarrow \theta - \alpha \nabla_\theta J_\pi(\theta)$
- 22: $\bar{\psi} \leftarrow \tau \psi + (1 - \tau) \bar{\psi}$
- 23: **end for**
- 24: $\sigma \leftarrow \sigma + clip(\frac{\alpha_{es}}{n\sigma} SmoothL1(F_{max}, F_{avg}), 0, \zeta)$
- 25: **end if**
- 26: crossover between $\theta_{es,(i)}$ and θ_{es} for $i = 1, 2, \dots, w$
- 27: Form new population pop_n using θ, θ_{es}, W
- 28: **end for**

S₁
不被自己POP支配

S₂
不被自己POP支配

S₃
不被自己POP支配

S₄
不被自己POP支配

5.2 Automatic Mutation Tuning (AMT)

Instead of maximizing randomness in the policy space, we maximize it in the weight space. We automatically tune σ with the initial value $\sigma_{(1)}$. σ is updated at fixed timesteps in a gradient-ascent manner without making use of backprop updates. AMT motivates guided exploration towards the objective by expanding the search horizon of population and enabling the agent to maximize rewards as randomly as possible. AMT makes use of the SmoothL1 (Huber) [36] loss function provided in Equation 1 and the update rule is mathematically expressed in Equation 2.

$$SmoothL1(x_i, y_i) = \begin{cases} 0.5(x_i - y_i)^2, & \text{if } |x_i - y_i| < 1 \\ |x_i - y_i| - 0.5, & \text{otherwise} \end{cases} \quad (1)$$

POP1 = {x₁¹, x₂¹, ..., x_N¹}
POP2 = {x₁², x₂², ..., x_N²}

Input: $\boxed{\text{Pop}_1} \quad \boxed{\text{Pop}_2}$

$\boxed{\text{Pop}_1/\text{Pop}_2}$ $\boxed{\text{AGENT}}$

$$\sigma_{(t+1)} \leftarrow \sigma_{(t)} + \frac{\alpha_{es}}{n\sigma_{(t)}} \text{SmoothL1}(R_{max}, R_{avg}) \quad (2)$$

$\boxed{\text{Pop}_1} \quad \boxed{\text{Pop}_2}$

Here, R_{max} is the reward observed by winner offspring, R_{avg} is the mean reward of the population with $\sigma_{(t)}$ and $\sigma_{(t+1)}$ the mutation rates at timesteps t and $t+1$ respectively. While exploring in weight space, the SmoothL1 loss tends to take up large values. This is indicative of the fact that the deviation between the winner and other learners of the population is significantly high. In order to reduce excessive noise from weight perturbations ϵ_i , we clip the update in a small region parameterized by the new clip parameter ζ . Suitable values for ζ range between 10^{-6} to 10^{-2} . The clipped update is mathematically expressed in Equation 3.

$$\sigma_{(t+1)} \leftarrow \sigma_{(t)} + \text{clip}\left(\frac{\alpha_{es}}{n\sigma_{(t)}} \text{SmoothL1}(R_{max}, R_{avg}), 0, \zeta\right) \quad (3)$$

The update can be unfolded recursively and written in terms of the initial mutation rate $\sigma_{(1)}$ as expressed in Equation 4 (derived in Appendix A).

$$\sigma_{(t+1)} \leftarrow \theta_{(t)} + \frac{\alpha_{es}}{n\sigma_{(1)}\hat{\Lambda}} \text{SmoothL1}(R_{max,(t)}, R_{avg,(t)}) \quad (4)$$

Here, $\hat{\Lambda}$ is defined as the Tuning Multiplier and can be mathematically expressed as $\hat{\Lambda} = \prod_{t'=1}^{t-1} \Lambda_{(t')} = \prod_{t'=1}^{t-1} \left(1 + \frac{\alpha_{es}}{n\sigma_{(t')}} \text{SmoothL1}(R_{max,(t')}, R_{avg,(t')})\right)$ (derived in Appendix A). We can additionally show that AMT, when combined with soft winner selection, leads to policy improvement among the set of winners (refer to Appendix B).

6 Experiments

Our experiments aim to evaluate performance, sample efficiency, scalability and mutation sensitivity of ESAC. Specifically, we aim to answer the following questions- (1) How does the algorithm compare to modern-day RL methods for complex tasks?, (2) How do evolutionary operations impact scalability in the presence of gradients? and (3) Which components of the method contribute to sensitivity and scalability?

6.1 Performance

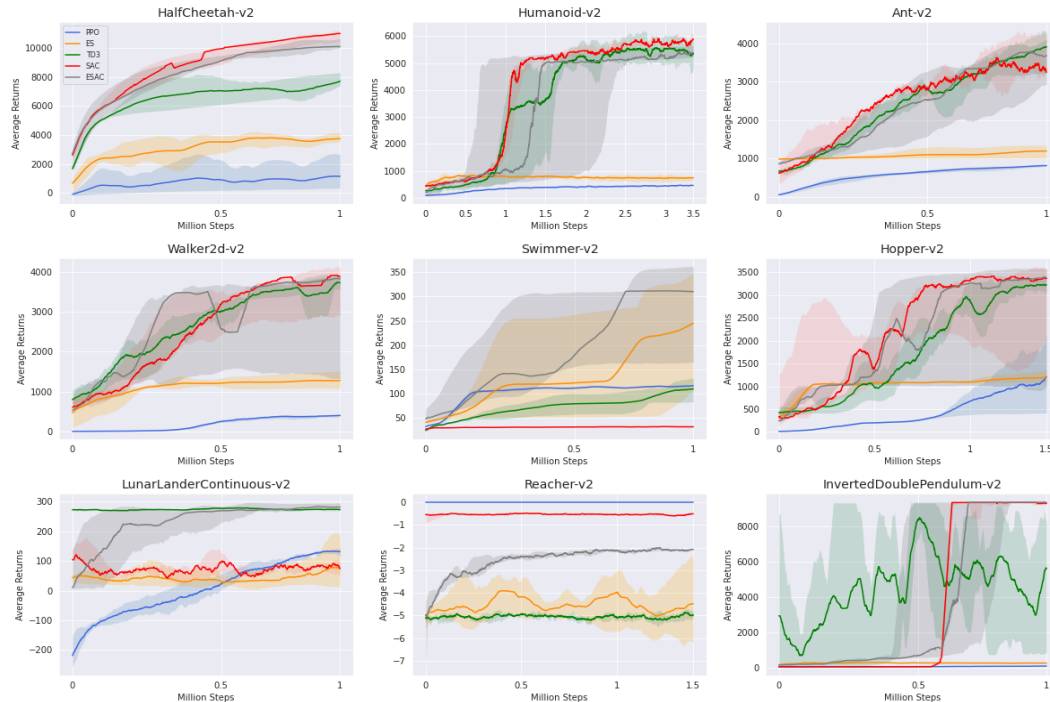


Figure 4: Average Returns on MuJoCo and OpenAI Gym's Lunar Lander continuous control tasks

We assess performance and sample efficiency of ESAC with state-of-the-art RL techniques including on-policy and off-policy algorithms. We compare our method to ES [14]; SAC [9]; Twin-Delayed Deep Deterministic Policy Gradient (TD3)[23] and PPO [29] on MuJoCo [37] control tasks and DeepMind Control Suite [38] (results for DeepMind Control Suite tasks are presented in [subsection C.1](#)). Additionally, we consider the LunarLander continuous task from OpenAI’s Gym suite [39] as a result of its narrow basin of learning. All methods were implemented using author-provided implementations except for ES in which Virtual Batch Normalization [40] was omitted as it did not provide significant performance boosts and hindered scalability. Agents were trained in OpenAI’s Gym environments [39] framework for a total of 5 random seeds. Training steps were interleaved with validation over 10 episodes. For all agents a multilayer architecture consisting of 2 hidden layers of 256 hidden units each activated with ReLU [41] nonlinearity and an output layer with tanh activation was used. We use this architecture as a result of its consistency in baseline implementations. We use Adam [42] as the optimizer (refer to [Appendix D](#) for hyperparameters). For ESAC and SAC, we use a Gaussian policy without automatic entropy tuning for a fair comparison. Training of gradient-based methods was conducted on 4 NVIDIA RTX2070 GPUs whereas for ES and ESAC, a total of 64 AMD Ryzen 2990WX CPUs were used.

[Figure 4](#) presents the total average returns of agents in the environment. The solid lines correspond to the average with the shaded area presenting standard deviation over random seeds. ESAC demonstrates comparable performance and sample-efficiency as state-of-the-art SAC and TD3. ESAC makes use of evolution-based weight-space exploration to converge to robust policies in tasks where SAC often learns a sub-optimal policy. For instance, ESAC demonstrates state-of-the-art performance in the Swimmer-v2 and LunarLanderContinuous-v2 environments which consist of local basins.

6.2 Scalability

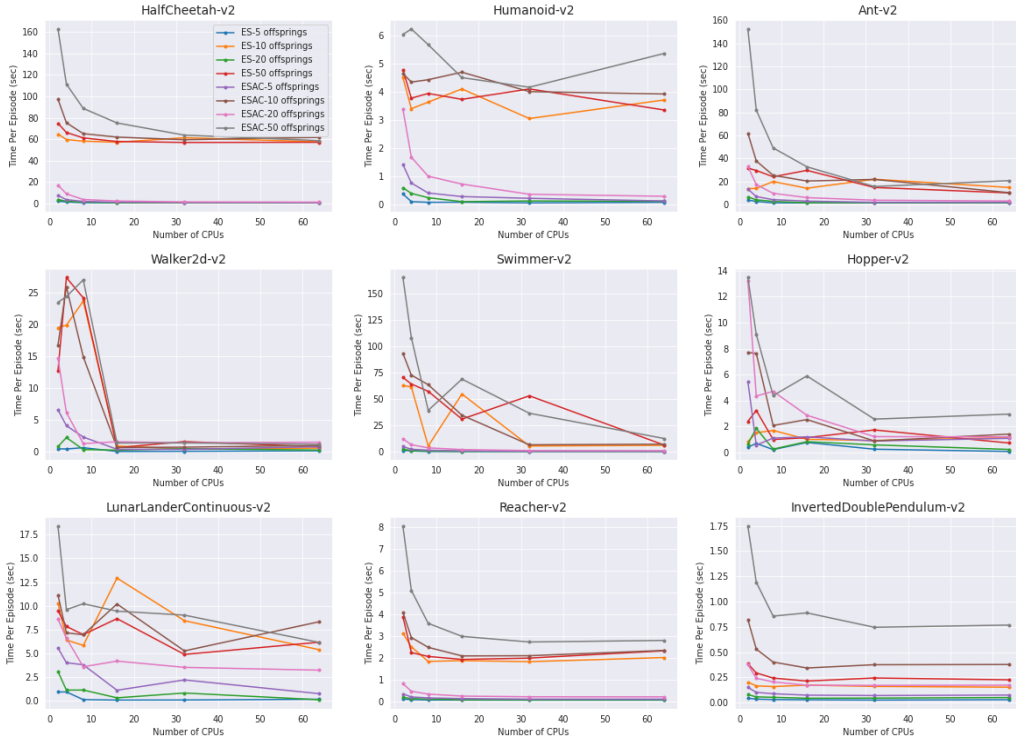


Figure 5: Variation of average time per episode (in seconds) with the number of operational CPUs and population size for MuJoCo control and LunarLanderContinuous tasks. ESAC demonstrates equivalent scalability as ES by providing an approximate 60% reduction in episode execution time.

We assess scalability of our method with ES on the basis of hardware resources and algorithm overhead. We vary the number of CPUs by keeping other training parameters constant. Parallelization on multiple CPU-based resources is readily available and cost-efficient in comparison to a single

efficient GPU resource. We also vary number of offsprings in the population by fixing CPU resources. Out of mutation rate σ and population size n , n governs the computational complexity of ES with σ being a scalar value. Thus, assessing variation w.r.t n provides a better understanding of resource utility and duration. For both experiments, we train the population for 10^6 steps and average the wall-clock time per episode. Another effective way to demonstrate scalability is by monitoring overall time taken to complete the training job [14]. However, this often tends to vary as initial learning periods have smaller episode lengths which does not compensate for fixed horizons of 1000 steps in MuJoCo.

Figure 5 presents the scalable nature of ESAC equivalent to ES on the MuJoCo and LunarLander-Continuous tasks. Average wall-clock time per episode is reduced utilizing CPU resources which is found to be favourable for evolution-based methods. Moreover, the variation depicts consistency with the increasing number of members in the population indicating large-scale utility of the proposed method. A notable finding here is that although ESAC incorporates gradient-based backpropagation updates, it is able to preserve its scalable nature by making use of evolutions as dominant operations during the learning process. Reduction in the number of SAC updates by exponentially annealing the gradient interval allows ESAC to reduce computation times and simultaneously explore using AMT.

6.3 Ablation Study

Mutation Sensitivity ES presents sensitivity to σ which is addressed by making use of AMT in ESAC. The AMT update gradually increases mutation rate σ using clip parameter ζ as learning progresses. Figure 6 presents the variation of average normalized rewards with different values of the new hyperparameter ζ for HalfCheetah-v2 and Ant-v2 tasks. Each experiment was run for 1 million steps. The population presents robustness and performance improvement for small values with the optimal range being 10^{-4} to 10^{-2} . On the other hand, sensitivity is observed in the 10^{-1} to 1 region which accounts for larger updates with high variance. The sensitivity is limited to a smaller region, unlike varying σ in the ES update. Lastly, offsprings remain robust to significantly large values of ζ due to early convergence of the population at the cost of poor performance among weak learners of the population.

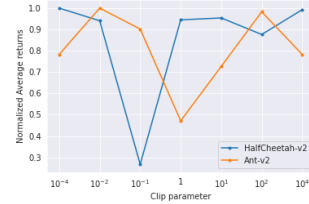


Figure 6: Mutation sensitivity of ESAC

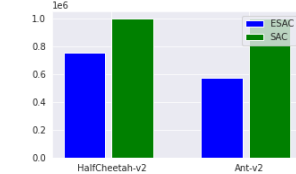


Figure 7: Number of backprop updates in ESAC compared to SAC

Number of Updates The main computation bottleneck in SAC arises from the number of backprop updates. This is tackled by exponentially annealing these updates and increasing winner-based evolutions and crossovers for transferring skills between SAC agent and ES offsprings. Figure 7 presents a comparison between the number of backprop updates carried out using SAC and ESAC during the training phase of HalfCheetah-v2 and Ant-v2 tasks. Results for the updates are averaged over 3 random seeds. ESAC executes lesser number of updates highlighting its computationally efficient nature and low dependency on a gradient-based scheme for monotonic improvement. Complete results can be found in subsection C.2.

7 Conclusions

In this paper, we introduce ESAC which combines the scalable nature of ES with low approximation bias and state-of-the-art performance of SAC. ESAC addresses the problem of mutation-sensitive evolutions by introducing AMT which maximizes the mutation rate of evolutions in a small clipped region as the SAC updates are exponentially decayed. ESAC demonstrates state-of-the-art performance equivalent to modern day gradient-based RL methods such as SAC and TD3 on MuJoCo and DeepMind Control Suite tasks. Additionally, ESAC presents scalability comparable to ES, reducing the average wall-clock time per episode by approximately 60%, hence depicting its suitability for large-scale RL tasks involving continuous control. On the other hand, high variance of the SAC agent under sparse rewards requiring consistent optimal behavior hurts the performance of ESAC. This can be addressed by combining the framework with a meta-controller or using a more sophisticated architecture such as a master-slave framework [43]. We leave this for future work.

Acknowledgments

We would like to thank the anonymous reviewers for providing valuable feedback on our work. We acknowledge Shauharda Khadka, Aravind Varier and Shashank Saurav for helpful discussions and Kamyar Ghasemipour for feedback on the initial draft. We appreciate the computing platform provided by the Department of Computer Science (DCS), University of Toronto. This work is supported by RBC Capital Markets, RBC Innovation Lab and the Center for Management of Technology and Entrepreneurship (CMTE).

References

- [1] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. A. Riedmiller. Playing atari with deep reinforcement learning. *CoRR*, abs/1312.5602, 2013.
- [2] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, Jan. 2016.
- [3] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. M. O. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra. Continuous control with deep reinforcement learning. *CoRR*, abs/1509.02971, 2015.
- [4] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105. Curran Associates, Inc., 2012.
- [5] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. P. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu. Asynchronous methods for deep reinforcement learning. *CoRR*, abs/1602.01783, 2016.
- [6] D. Kalashnikov, A. Irpan, P. Pastor, J. Ibarz, A. Herzog, E. Jang, D. Quillen, E. Holly, M. Kalakrishnan, V. Vanhoucke, and S. Levine. Scalable deep reinforcement learning for vision-based robotic manipulation. In *CoRL*, volume 87 of *PMLR*, pages 651–673. PMLR, 29–31 Oct 2018.
- [7] D. Hafner, T. Lillicrap, J. Ba, and M. Norouzi. Dream to control: Learning behaviors by latent imagination, 2019.
- [8] Y. Wu, E. Mansimov, S. Liao, R. Grosse, and J. Ba. Scalable trust-region method for deep reinforcement learning using kronecker-factored approximation, 2017.
- [9] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *CoRR*, abs/1801.01290, 2018.
- [10] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning Representations by Back-propagating Errors. *Nature*, 323(6088):533–536, 1986.
- [11] P. Malekzadeh, M. Salimibeni, A. Mohammadi, A. Assa, and K. N. Plataniotis. Mm-ktd: Multiple model kalman temporal differences for reinforcement learning, 2020.
- [12] T. D. Kulkarni, K. R. Narasimhan, A. Saeedi, and J. B. Tenenbaum. Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation. In *NIPS*, page 36823690, 2016.
- [13] Y. Burda, H. Edwards, A. Storkey, and O. Klimov. Exploration by random network distillation. In *ICLR*, 2019.
- [14] T. Salimans, J. Ho, X. Chen, S. Sidor, and I. Sutskever. Evolution strategies as a scalable alternative to reinforcement learning, 2017.
- [15] R. Houthooft, Y. Chen, P. Isola, B. Stadie, F. Wolski, J. Ho, and P. Abbeel. Evolved policy gradients. In *NIPS*, pages 5400–5409. 2018.
- [16] S. Khadka and K. Tumer. Evolution-guided policy gradient in reinforcement learning. In *NIPS*, NIPS18, page 11961208, Red Hook, NY, USA, 2018. Curran Associates Inc.
- [17] K. O. Stanley and R. Miikkulainen. Evolving neural networks through augmenting topologies. *Evolutionary Computing*, 10(2):99127, June 2002.
- [18] T. Miconi, A. Rawal, J. Clune, and K. O. Stanley. Backpropamine: training self-modifying neural networks with differentiable neuromodulated plasticity, 2020.

- [19] G. Rockefeller, S. Khadka, and K. Tumer. Multi-level fitness critics for cooperative coevolution. In *ICAAMS*, 2020.
- [20] H. V. Hasselt. Double q-learning. In *NIPS 23*. 2010.
- [21] H. van Hasselt, A. Guez, and D. Silver. Deep reinforcement learning with double q-learning. 2015. cite arxiv:1509.06461Comment: AAAI 2016.
- [22] Q. Lan, Y. Pan, A. Fyshe, and M. White. Maxmin q-learning: Controlling the estimation bias of q-learning. In *International Conference on Learning Representations*, 2020.
- [23] S. Fujimoto, H. van Hoof, and D. Meger. Addressing function approximation error in actor-critic methods. *CoRR*, abs/1802.09477, 2018.
- [24] S. Khadka, C. Yates, and K. Tumer. A memory-based multiagent framework for adaptive decision making. In *17th ICAAMS*, 2018.
- [25] D. E. Moriarty and R. Mikkulainen. Efficient reinforcement learning through symbiotic evolution. *Machine Learning*, 22(1-3):11–32, 1996.
- [26] P. Pagliuca, N. Milano, and S. Nolfi. Efficacy of modern neuro-evolutionary strategies for continuous control optimization, 2019.
- [27] A. Pourchot and O. Sigaud. Cem-rl: Combining evolutionary and gradient-based methods for policy search, 2018.
- [28] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz. Trust region policy optimization. In F. Bach and D. Blei, editors, *ICML*, volume 37 of *PMLR*, pages 1889–1897, Lille, France, 07–09 Jul 2015. PMLR.
- [29] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017.
- [30] P. Hmlinen, A. Babadi, X. Ma, and J. Lehtinen. Ppo-cma: Proximal policy optimization with covariance matrix adaptation. *CoRR*, abs/1810.02541, 2018.
- [31] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. 2018.
- [32] B. D. Ziebart. *Modeling Purposeful Adaptive Behavior with the Principle of Maximum Causal Entropy*. PhD thesis, USA, 2010.
- [33] L. Espeholt, R. Marinier, P. Stanczyk, K. Wang, and M. Michalski. Seed rl: Scalable and efficient deep-rl with accelerated central inference. In *ICLR*, 2020.
- [34] P. Balaprakash, R. Egele, M. Salim, S. Wild, V. Vishwanath, F. Xia, T. Bretin, and R. Stevens. Scalable reinforcement-learning-based neural architecture search for cancer deep learning research. *Proceedings of the SC*, Nov 2019.
- [35] A. R. Mahmood, D. Korenkevych, G. Vasani, W. Ma, and J. Bergstra. Benchmarking reinforcement learning algorithms on real-world robots. *CoRR*, abs/1809.07731, 2018.
- [36] P. J. Huber. Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35(1):73–101, Mar. 1964.
- [37] E. Todorov, T. Erez, and Y. Tassa. Mujoco: A physics engine for model-based control. In *IROS, 2012 IEEE/RSJ International Conference on*, pages 5026–5033. IEEE, 2012.
- [38] Y. Tassa, Y. Doron, A. Muldal, T. Erez, Y. Li, D. de Las Casas, D. Budden, A. Abdolmaleki, J. Merel, A. Lefrancq, T. P. Lillicrap, and M. A. Riedmiller. Deepmind control suite. *CoRR*, abs/1801.00690, 2018.
- [39] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba. Openai gym, 2016. cite arxiv:1606.01540.
- [40] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, and X. Chen. Improved techniques for training gans. In *29th NIPS*. 2016.
- [41] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In J. Frnkranz and T. Joachims, editors, *ICML*, pages 807–814, 2010.
- [42] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. 3rd International Conference for Learning Representations, San Diego, 2015.
- [43] K. Suri and R. Gupta. Transfer learning for semg-based hand gesture classification using deep learning in a master- slave architecture. *IC3I*, Oct 2018.

A Derivation

The AMT and ES update rules are given as

$$\begin{aligned}\sigma_{(t)} &\leftarrow \sigma_{(t-1)} + \frac{\alpha_{es}}{n\sigma_{(t-1)}} \text{SmoothL1}(R_{max,(t-1)}, R_{avg,(t-1)}) \\ \theta_{(t+1)} &\leftarrow \theta_{(t)} + \frac{\alpha_{es}}{n\sigma_{(t)}} \sum_{i=1}^N R_i \epsilon_i\end{aligned}$$

Using the expression for $\sigma_{(t)}$ in the ES update yields the following

$$\begin{aligned}\theta_{(t+1)} &\leftarrow \theta_{(t)} + \frac{\alpha_{es}}{n(\sigma_{(t-1)} + \frac{\alpha_{es}}{n\sigma_{(t-1)}} \text{SmoothL1}(R_{max,(t-1)}, R_{avg,(t-1)}))} \sum_{i=1}^n R_i \epsilon_i \\ &= \theta_{(t+1)} \leftarrow \theta_{(t)} + \frac{\alpha_{es}}{n\sigma_{(t-1)}(1 + \frac{\alpha_{es}}{n\sigma_{(t-1)}^2} \text{SmoothL1}(R_{max,(t-1)}, R_{avg,(t-1)}))} \sum_{i=1}^n R_i \epsilon_i \\ &= \theta_{(t+1)} \leftarrow \theta_{(t)} + \frac{\alpha_{es}}{n\sigma_{(t-1)}\Lambda_{(t-1)}} \sum_{i=1}^n R_i \epsilon_i\end{aligned}$$

where $\Lambda_{(t-1)} = 1 + \frac{\alpha_{es}}{n\sigma_{(t-1)}^2} \text{SmoothL1}(R_{max,(t-1)}, R_{avg,(t-1)})$. Expanding $\sigma_{(t-1)}$ using the AMT update rule gives us

$$\begin{aligned}\theta_{(t+1)} &\leftarrow \theta_{(t)} + \frac{\alpha_{es}}{n\Lambda_{(t-1)}(\sigma_{(t-2)} + \frac{\alpha_{es}}{n\sigma_{(t-2)}} \text{SmoothL1}(R_{max,(t-2)}, R_{avg,(t-2)}))} \sum_{i=1}^n R_i \epsilon_i \\ &= \theta_{(t+1)} \leftarrow \theta_{(t)} + \frac{\alpha_{es}}{n\Lambda_{(t-1)}\sigma_{(t-2)}(1 + \frac{\alpha_{es}}{n\sigma_{(t-2)}^2} \text{SmoothL1}(R_{max,(t-2)}, R_{avg,(t-2)}))} \sum_{i=1}^n R_i \epsilon_i \\ &= \theta_{(t+1)} \leftarrow \theta_{(t)} + \frac{\alpha_{es}}{n\Lambda_{(t-1)}\sigma_{(t-2)}\Lambda_{(t-2)}} \sum_{i=1}^n R_i \epsilon_i\end{aligned}$$

where $\Lambda_{(t-2)} = 1 + \frac{\alpha_{es}}{n\sigma_{(t-2)}^2} \text{SmoothL1}(R_{max,(t-2)}, R_{avg,(t-2)})$. Expanding this recursively gives us the following

$$\begin{aligned}\theta_{(t+1)} &\leftarrow \theta_{(t)} + \frac{\alpha_{es}}{n\sigma_{(1)}\Lambda_{(t-1)}\Lambda_{(t-2)}\dots\Lambda_{(1)}} \sum_{i=1}^n R_i \epsilon_i \\ &= \theta_{(t+1)} \leftarrow \theta_{(t)} + \frac{\alpha_{es}}{n\sigma_{(1)} \prod_{t'=1}^{t-1} \Lambda_{(t')}} \sum_{i=1}^n R_i \epsilon_i \\ &= \theta_{(t+1)} \leftarrow \theta_{(t)} + \frac{\alpha_{es}}{n\sigma_{(1)}\hat{\Lambda}} \sum_{i=1}^n R_i \epsilon_i\end{aligned}$$

Hence, yielding the AMT update in the form of initial mutation rate $\sigma_{(1)}$.

B Policy Improvement

We will now show that AMT improves the policy of winners in the population. Let us consider two successive gradient intervals g indexed by (l) and $(l-1)$. Let $p_{(l)}$ and $p_{(l-1)}$ be the probabilities of convergence to the optimal policy $\pi_{\theta_{es}}^*(a_t|s_t)$ in the weight space at (l) and $(l-1)$ respectively.

We start by evaluating the mutation rates at (l) and $(l-1)$ which are given as $\sigma_{(l)} > \sigma_{(l-1)}$. We can now evaluate the probabilities of convergence to $\pi_{\theta_{es}}^*(a_t|s_t)$ as

$$p_{(l)} \geq p_{(l-1)}$$

Using this fact, we can evaluate the winners (indexed by q) in the sorted reward population F .

$$\begin{aligned}
\sum_{q=1}^w p_{(l)}^{(q)} F_{(l)}^{(q)} &\geq \sum_{q=1}^w p_{(l-1)}^{(q)} F_{(l-1)}^{(q)} \\
&= \mathbf{E}_{F_{(l)}^{(q)} \sim F_{(l)}} [F_{(l)}^{(q)}] \geq \mathbf{E}_{F_{(l-1)}^{(q)} \sim F_{(l-1)}} [F_{(l-1)}^{(q)}] \\
&= \mathbf{E}[W_{(l)}] \geq \mathbf{E}[W_{(l-1)}]
\end{aligned}$$

Here, $p_{(l)}^{(q)}$ is the probability of convergence of actor q (having observed reward $F_{(l)}^{(q)}$) to its optimal policy $\pi_{\theta_{es}}^{(q),*}(a_t|s_t)$ at interval (l) . $W_{(l)}$ represents the set of winners at (l) . The mathematical expression obtained represents that the set of winners $W_{(l)}$ formed at the next gradient interval (l) is at least as good as the previous set of winners $W_{(l-1)}$, i.e.- $\pi_{\theta_{es},(l)}^{(q)}(a_t|s_t) \geq \pi_{\theta_{es},(l-1)}^{(q)}(a_t|s_t)$. This guarantees policy improvement among winners of the population.

C Additional Results

C.1 DeepMind Control Suite

We evaluate the performance and sample-efficiency of ESAC on the DeepMind Control Suite [38]. We compare the learning of ESAC to ES [14], TD3 [23] and PPO [29]. Unlike the MuJoCo benchmarks, we omit SAC as a baseline as a result of high variance and inconsistent performance across different seeds. We observed the reproducibility of SAC as a common problem [7, 38] irrespective of a rigorous hyperparameter search. Figure 8 presents average returns on DeepMind Control Suite for a total of 1 million steps. Training setup for all agents was kept same as in the MuJoCo tasks with different values of hyperparameters (presented in subsection D.2). ESAC demonstrates comparable performance to state-of-the-art TD3 and ES. In tasks such as CheetahRun, WalkerWalk and WalkerRun, ESAC converges to a sub-optimal yet robust policy as a result of successive evolutions.

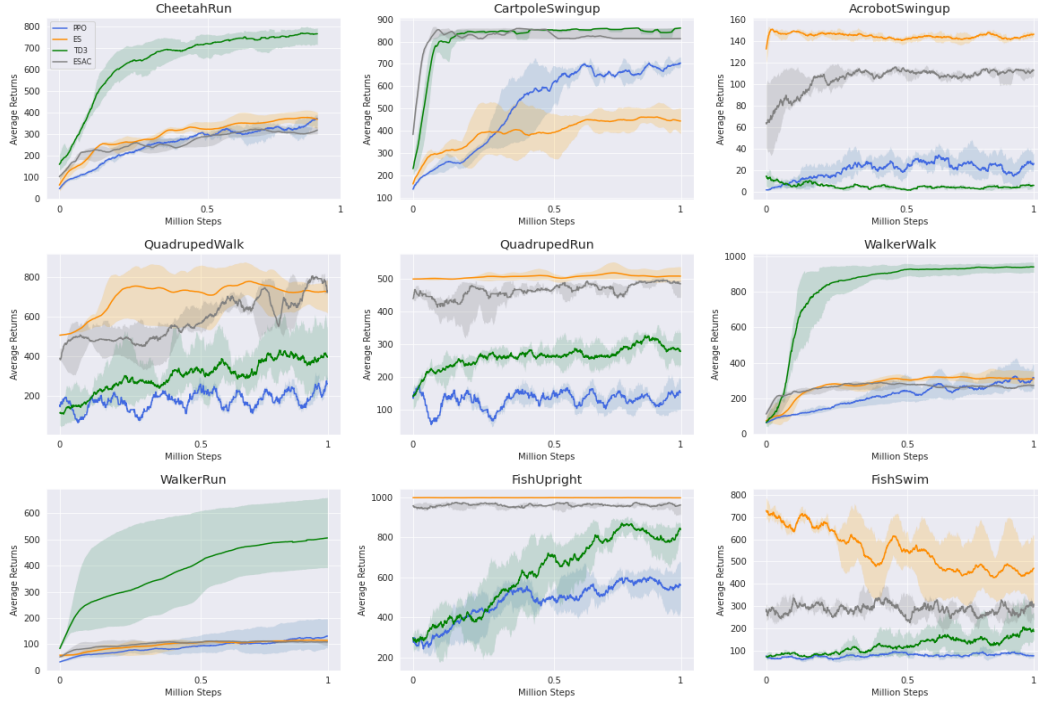


Figure 8: Average Returns on DeepMind Control Suite tasks. ESAC’s performance is comparable to state-of-the-art TD3 on majority of the tasks. In our experiments on the control suite, SAC did not present any learning behavior due to high variance under sparse rewards. This is found to be consistent with previous methods [7] and benchmarks [38] trying to use SAC as a baseline. ESAC’s performance suffers in this aspect on complex tasks such as CheetahRun, WalkerWalk and WalkerRun which require consistent optimal behavior from the agent.

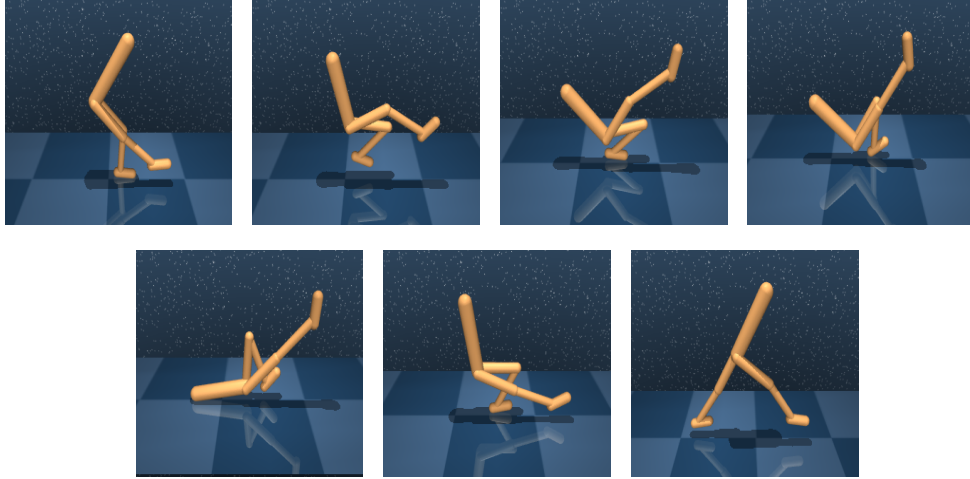


Figure 9: Robust behavior of ESAC observed on the WalkerWalk task. The ESAC policy prevents the walking robot from falling down when the robot loses its balance while walking. Although ESAC falls short of optimal performance as a result of sparse rewards, it exhibits robust policies on complex tasks due to successive evolutions.

C.2 Scalability

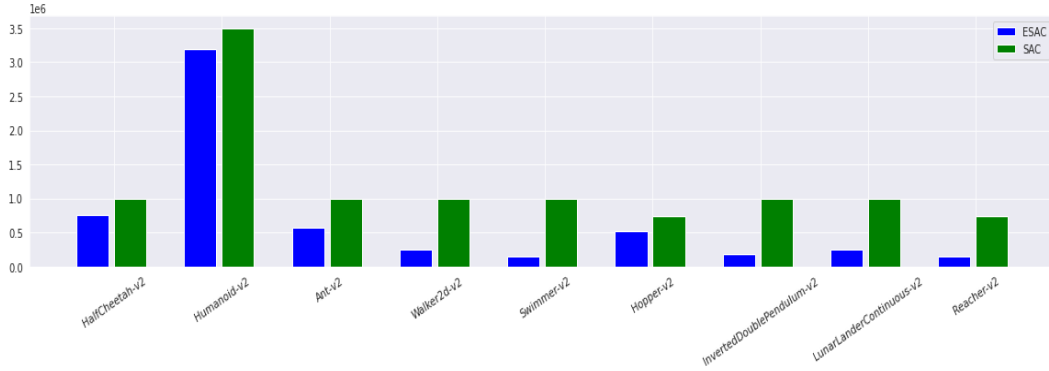


Figure 10: Complete results on the number of backprop updates for MuJoCo control tasks and LunarLanderContinuous environment from OpenAI’s Gym suite. ESAC exponentially anneals gradient-based SAC updates and leverages winner selection and crossovers for computational efficiency and performance improvement.

D Hyperparameters

D.1 MuJoCo

Hyperparameter values for our experiments are adjusted on the basis of complexity and reward functions of tasks. In the case of MuJoCo control tasks, training-based hyperparameters are kept mostly the same with the number of SAC episodes varying as per the complexity of task. Tuning was carried out on HalfCheetah-v2, Ant-v2, Hopper-v2 and Walker2d-v2 tasks. Out of these, Ant-v2 presented high variance indicating the requirement of a lower learning rate. Number of SAC updates in SAC and ESAC implementations were kept 1 for a fair comparison with TD3. All tasks have a common discount factor $\gamma = 0.99$, SAC learning rate $\alpha = 3 \times 10^{-4}$, population size $n = 50$, mutation rate $\sigma = 5 \times 10^{-3}$ and winner fraction $e = 0.4$. ES learning rate α_{es} was kept fixed at 5×10^{-3} for all tasks except Ant-v2 having $\alpha_{es} = 1 \times 10^{-4}$.

The only variable hyperparameter in our experiments is number of SAC episodes executed by the SAC agent. Although the ESAC population in general is robust to its hyperparameters, the SAC agent is sensitive to the number of episodes. During the tuning process, the number of episodes were kept constant to a value of 10 for all the tasks. However, this led to inconsistent results on some of the environments when compared to SAC

baseline. As a result, tuning was carried out around this value to obtain optimal results corresponding to each task. The SAC agent executed a total of 10 episodes for each gradient interval g for Ant-v2, Walker2d-v2, Hopper-v2 and Humanoid-v2 tasks; 5 episodes for HalfCheetah-v2, LunarLanderContinuous-v2, Reacher-v2 and InvertedPendulum-v2 tasks; and 1 episode for Swimmer-v2 task.

D.2 DeepMind Control Suite

The DeepMind control suite presents a range of tasks with sparse rewards and varying complexity for the same domain. Hyperparameter values for these tasks are different from that of MuJoCo control tasks. Tuning was carried on the Cheetah, Quadraped and Walker tasks in order to obtain sample-efficient convergence. In the case of SAC, granularity of hyperparameter search was refined in order to observe consistent behavior. However, different values of temperature parameter produced varying performance resulting in inconsistent performance across different seeds. As in the MuJoCo case, we kept the number of updates fixed to 1 in order to yield a fair comparison with TD3. However, SAC did not present any learning behavior in the case of sparse rewards.

All tasks have a common discount factor $\gamma = 0.99$, SAC learning rate $\alpha = 3 \times 10^{-4}$, population size $n = 50$, mutation rate $\sigma = 1 \times 10^{-2}$, winner fraction $e = 0.4$ and ES learning rate 1×10^{-2} . As in the case of MuJoCo tasks, the SAC is found to be sensitive to the number of episodes during the gradient interval g . These were kept constant at 5 and then tuned around this value for optimal performance. Final values of the SAC episodes were 5 for CartpoleSwingup, WalkerWalk and WalkerRun tasks and 1 for the remaining tasks.