SDSRA: A SKILL-DRIVEN SKILL-RECOMBINATION ALGORITHM FOR EFFICIENT POLICY LEARNING

Eric H. Jiang and Andrew Lizarraga
Department of Statistics and Data Science
University of California, Los Angeles
Los Angeles, CA 90095-1554
{ericjiang0318, andrewlizarraga}@g.ucla.edu

ABSTRACT

In this paper we introduce a novel algorithm-the Skill-Driven Skill Recombination Algorithm (SDSRA)—an innovative framework that significantly enhances the efficiency of achieving maximum entropy in reinforcement learning tasks. We find that SDSRA achieves faster convergence compared to the traditional Soft Actor-Critic (SAC) algorithm and produces improved policies. By integrating skill-based strategies within the robust Actor-Critic framework, SDSRA demonstrates remarkable adaptability and performance across a wide array of complex and diverse benchmarks.

1 INTRODUCTION

Reinforcement Learning (RL) has significantly advanced, with the Soft Actor-Critic (SAC) algorithm, introduced by Haarnoja et al. (2018), standing out for efficient exploration in complex tasks. Despite its strengths, SAC, like other RL methods, faces challenges in more intricate environments. To address these issues, recent research, such as goal-enforced hierarchical learning Chane-Sane et al. (2021) and intrinsically motivated RL with skill selection Singh et al. (2004), focuses on enhancing RL frameworks. In this paper we address these issues and make the following contributions:

- **Innovative Framework:** We introduce SDSRA a novel approach that surpasses SAC methods.
- **Integration of Intrinsic Motivation:** SDSRA incorporates intrinsically motivated learning within a hierarchical structure, enhancing self-directed exploration and skill development which is lacking in SAC.
- **Enhanced Skill Acquisition and Dynamic Selection:** Our method excels in acquiring and dynamically selecting a wide range of skills suitable for varying environmental conditions, offering greater adaptability.
- **Superior Performance and Learning Rate:** We demonstrate faster performance and a quicker learning rate compared to conventional SAC methods, leading to improved rewards in various benchmarks.

1.1 RELATED WORK

Reinforcement learning research is expanding, particularly in hierarchical structures and intrinsic motivation. Tang et al. (2021) developed a hierarchical SAC variant with sub-goals, yet lacks public code and detailed results. Ma et al. (2022) proposed ELIGN for predicting agent cooperation using intrinsic rewards, while Aubret et al. (2019) surveyed RL algorithms with intrinsic motivation. Other notable works include Laskin et al. (2022)'s skill learning algorithm combining intrinsic rewards and representation learning, Sharma et al. (2019)'s skill discovery algorithm, Bagaria & Konidaris (2020)'s skill discovery algorithm, and Zheng et al. (2018)'s intrinsic reward mechanism for Policy Gradient and PPO algorithm. Despite progress, a gap persists in skill-driven recombination algorithms using intrinsic rewards in Actor-Critic frameworks, particularly in physical environments like MuJoCo Gym. Our SDSRA work addresses this, blending skill-driven learning with Actor-Critic methods, proving effective in complex simulations.

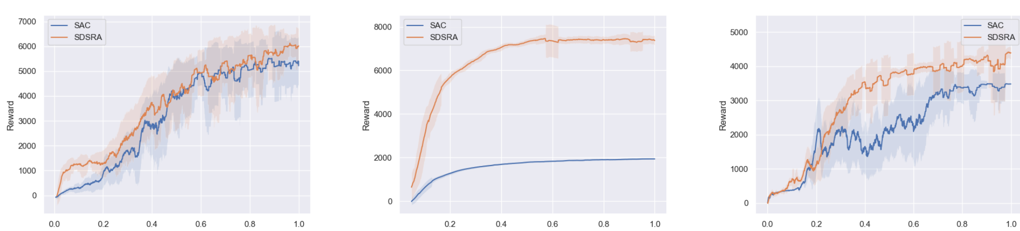
2 MOTIVATION FOR SDSRA

The SDSRA algorithm adapts the SAC framework, retaining its integration of rewards and entropy maximization, and using actor and critic networks for action selection and evaluation. While SAC emphasizes entropy for diverse exploration, SDSRA introduces a novel selection scheme for enhanced performance in complex environments. SDSRA defines a set of Gaussian Policy skills $S = \{\pi_1, \pi_2, \ldots, \pi_N\}$ with parameters θ_i representing mean $\mu_{\theta_i}(s)$ and covariance $\Sigma_{\theta_i}(s)$. Each skill π_i is formulated as: $\pi_i(\theta_i) = \mathcal{N}(\mu_{\theta_i}(s), \Sigma_{\theta_i}(s))$. Skills initially have a relevance score $r_i = c$, and skill selection is probabilistic, based on softmax distribution of relevance scores: $P(i|s) = \frac{e^{r_i}}{\sum_{j=1}^{N} e^{r_j}}$. Skill optimization in SDSRA involves minimizing a loss function $\text{loss}_i = \varepsilon_i + \beta \cdot \mathcal{H}(\pi_i)$, combining prediction error $\varepsilon_i = \frac{1}{M}\sum_{m=1}^{M}(\hat{a}_{i,m} - a_m)^2$ and policy entropy $\mathcal{H}(\pi_i) = -\int \pi_i(a|s)\log(\pi_i(a|s))\, da$. Precise parameter updates and implementations details are discussed in Appendix. B. SDSRA's decision-making involves selecting and executing actions based on skill selection and continuous skill refinement, enabling adaptive and effective decision-making in diverse environments. In the integrated framework, the SAC objective function is modified to incorporate the dynamic skill selection process. The new objective function aims to maximize not just the expected return, but also the individual entropy of each skill. The modified objective function is expressed as:

$$J_{\text{integrated}}(\pi) = \sum_{i=1}^{N} P(i|s)\left(\mathbb{E}_{(s_t,a_t)\sim\pi_i}\left[Q(s_t,a_t) + \alpha\mathcal{H}(\pi_i(\cdot|s_t))\right]\right), \tag{1}$$

where $Q(s_t, a_t)$ denotes the action-value function estimated by SAC's critic networks, and α is a coefficient that balances the importance of the entropy term for each skill π_i.Under this proposal we find that SDSRA converges to an improved policy, see Appendix. A.1 and Appendix. A.2. Moreso, we find that experiments ran on a commonly tested dataset for SAC algorithms demonstrates significant improvements in SDSRA over SAC.

3 EXPERIMENTS



(a) Ant-v2 (b) Half-Cheetah-v2 (c) Hopper-v2

Figure 1: Performance comparison of SDSRA and SAC in the MuJoCo Ant, Half-Cheetah, and Hopper environment.

We evaluated the Skill-Driven Skill Recombination Algorithm (SDSRA) in MuJoCo gym locomotion tasks Brockman et al. (2016) , and compared its performance with the Soft Actor-Critic (SAC) algorithm. To ensure consistency and enable a comparative analysis across all test environments, we standardized the number of skills to five in each task. Our tests demonstrated that SDSRA consistently outperformed SAC, achieving faster reward convergence in fewer steps.

4 CONCLUSION

In this paper, we introduced the Skill-Driven Skill Recombination Algorithm (SDSRA) outperforms the traditional Soft Actor-Critic in reinforcement learning, particularly in the MuJoCo environment. Its skill-based approach leads to faster convergence and higher rewards, showing great potential for complex tasks requiring quick adaptability and learning efficiency.

REFERENCES

A. Aubret, L. Matignon, and S. Hassas. A survey on intrinsic motivation in reinforcement learning. *arXiv preprint arXiv:1908.06976*, 2019. URL https://arxiv.org/abs/1908.06976.

Akjil Bagaria and George Konidaris. Option discovery using deep skill chaining. *ICLR 2020*, 2020. URL https://openreview.net/pdf?id=B1gqipNYwH.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, J. Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016. URL https://arxiv.org/abs/1606.01540.

Elliot Chane-Sane, Cordelia Schmid, and Ivan Laptev. Goal-conditioned reinforcement learning with imagined subgoals. *Proceeding in Machine Learning*, 2021. URL http://proceedings.mlr.press/v139/chane-sane21a/chane-sane21a.pdf.

Tuomas Haarnoja, Aurick Zhou, P. Abbeel, and S. Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *arXiv preprint arXiv:1801.01290*, 2018. URL https://arxiv.org/abs/1801.01290.

M. Laskin, Hao Liu, X. B. Peng, Denis Yarats, A. Rajeswaran, and P. Abbeel. Cic: Contrastive intrinsic control for unsupervised skill discovery. *arXiv preprint arXiv:2202.00161*, 2022. URL https://arxiv.org/abs/2202.00161.

Zixian Ma, Rose Wang, Li Fei-Fei, Michael Bernstein, and Ranjay Krishna. Elign: Expectation alignment as a multi-agent intrinsic reward. *NeurIPS 2022*, 2022. URL https://arxiv.org/pdf/2210.04365.pdf.

Archit Sharma, Shixiang Shane Gu, Sergey Levine, Vikash Kumar, and Karol Hausman. Dynamics-aware unsupervised discovery of skills. *ArXiv*, abs/1907.01657, 2019. URL https://api.semanticscholar.org/CorpusID:195791369.

Satinder Singh, Andrew G. Barto, and Nuttapong Chentanez. Intrinsically motivated reinforcement learning. *NeurIPS 2004*, 2004. URL https://www.cs.cornell.edu/~helou/IMRL.pdf.

Hengliang Tang, Anqi Wang, Fei Xue, Jiaxin Yang, and Yang Cao. A novel hierarchical soft actor-critic algorithm for multi-logistics robots task allocation. *IEEE Access*, 9:42568–42582, 2021. doi: 10.1109/ACCESS.2021.3062457.

Zeyu Zheng, Junhyuk Oh, and Satinder Singh. On learning intrinsic rewards for policy gradient methods. *NeurIPS 2004*, 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/51de85ddd068f0bc787691d356176df9-Paper.pdf.

URM STATEMENT

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2024 Tiny Papers Track.

A PROOFS OF MAIN RESULTS

A.1 POLICY IMPROVEMENT GUARANTEE WITH SKILLS

Lemma A.1 (Policy Improvement Guarantee with Skills). *Given a set of skills $\{\pi_i\}$ and a composite policy π, if the soft Q-values Q^{π_i} for each skill are updated according to the soft Bellman backup operator, then the policy π', which acts greedily with respect to the aggregated Q-values of the skills, achieves an equal or greater expected return than π.*

Proof. Let π be any composite policy as a function of individual skills $\{\pi_i\}$ and π' be the policy that is greedy with respect to the aggregated soft Q-values Q^{π_i}. For all states $s \in \mathcal{S}$, we have:

$$\pi'(a|s) = \arg\max_{a'} \sum_i \left(Q^{\pi_i}(s, a') + \alpha \mathcal{H}(\pi_i(\cdot|s))\right). \tag{2}$$

Now consider the soft value function V^π for the composite policy, which is given by:

$$V^\pi(s) = \mathbb{E}_{a\sim\pi}\left[\sum_i Q^{\pi_i}(s, a) - \alpha \log \pi(a|s)\right]. \tag{3}$$

Using the soft Bellman optimality equation for $Q^{\pi'}$, we get:

$$Q^{\pi'}(s, a) = \mathbb{E}_{s'\sim P}[r(s, a) + \gamma V^{\pi'}(s')]. \tag{4}$$

Substituting the expression for $V^{\pi'}$ into the above, we have:

$$Q^{\pi'}(s, a) = \mathbb{E}_{s'\sim P,a'\sim\pi'}[r(s, a) + \gamma(Q^{\pi'}(s', a') - \alpha \log \pi'(a'|s'))]. \tag{5}$$

Since π' is greedy with respect to the aggregated Q^{π_i}, it follows that $Q^{\pi'}(s, a) \geq Q^\pi(s, a)$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$.

Thus, we have shown that acting greedily with respect to the aggregated soft Q-values under the composite policy π results in a policy π' that has greater or equal Q-value for all state-action pairs, which completes the proof. □

A.2 THEOREM ??: CONVERGENCE TO OPTIMAL POLICY WITH SKILLS

Theorem A.2 (Convergence to Optimal Policy with Skills). *Repeated application of soft policy evaluation and improvement using a set of skills $\{\pi_i\}$ from any initial composite policy $\pi \in \Pi$ converges to a policy π^* that integrates these skills and such that $Q^{\pi^*}(s_t, a_t) \geq Q^\pi(s_t, a_t)$ for all $\pi \in \Pi$ and $(s_t, a_t) \in \mathcal{S} \times \mathcal{A}$, assuming $|\mathcal{A}| < \infty$.*

Proof. The soft Bellman backup operator for policy evaluation under the composite policy π considering skills is given by:

$$T^\pi Q(s, a) = \mathbb{E}_{s'\sim P,a'\sim\pi}\left[\sum_i r_i(s, a) + \gamma\left(Q(s', a') - \alpha \log \pi(a'|s')\right)\right]. \tag{6}$$

This operator is a contraction mapping in the supremum norm, ensuring convergence to a unique fixed point Q^π that satisfies the soft Bellman equation for the composite policy π.

Now, define the soft Bellman optimality operator T^* for the composite policy as:

$$T^* Q(s, a) = \max_\pi T^\pi Q(s, a). \tag{7}$$

The soft policy improvement step involves updating the composite policy π to a new policy π' by choosing actions that maximize the current soft Q-values plus the entropy term across all skills:

$$\pi' = \arg\max_\pi \mathbb{E}_{a\sim\pi}\left[\sum_i Q^{\pi_i}(s, a) - \alpha \log \pi(a|s)\right]. \tag{8}$$

By the policy improvement theorem, the new policy π' achieves a Q-value that is greater than or equal to that of the composite policy π, i.e., $Q^{\pi'}(s,a) \geq Q^{\pi}(s,a)$ for all (s,a).

Since the action space $\mathcal{A}$ is finite, the sequence of composite policies $\{\pi_k\}$ obtained by alternating soft policy evaluation and improvement must eventually converge to a policy π^* that cannot be improved further, thus being the optimal policy with respect to the soft Bellman optimality equation. Therefore:

$$Q^{\pi^*}(s,a) = T^* Q^{\pi^*}(s,a), \tag{9}$$

for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. This concludes the proof that the sequence of composite policies converges to an optimal policy π^*. □

A.3 Proof of Theorem 2: Entropy Maximization Efficiency of SDSRA

Theorem A.3 (Entropy Maximization Efficiency of SDSRA). *Let π^{SAC} and π^{SDSRA} be the policies obtained from the SAC and SDSRA algorithms, respectively, when trained under identical conditions. Assume that both algorithms achieve convergence. Then, for any state $s \in \mathcal{S}$, the expected entropy of π^{SDSRA} is greater than or equal to that of π^{SAC}:*

$$\mathbb{E}_{a\sim\pi^{SDSRA}}[-\log \pi^{SDSRA}(a|s)] \geq \mathbb{E}_{a\sim\pi^{SAC}}[-\log \pi^{SAC}(a|s)], \tag{10}$$

or the time to reach an ϵ-optimal policy entropy for SDSRA is less than that for SAC:

$$t_{SDSRA}(\epsilon) \leq t_{SAC}(\epsilon), \tag{11}$$

where $t_{SDSRA}(\epsilon)$ and $t_{SAC}(\epsilon)$ denote the time to reach a policy entropy within ϵ of the maximum entropy for SDSRA and SAC, respectively.

Proof. Assume that both π^{SAC} and π^{SDSRA} have converged to their respective policy distributions for all states $s \in \mathcal{S}$. By the definition of convergence, we have that the policies are stationary and hence the expected entropy under each policy is constant over time.

The SDSRA algorithm incorporates a diverse set of skills, each represented by an individual policy. This diversity enables SDSRA to explore a broader action space compared to SAC, which is constrained by a single policy approach. The dynamic nature of skill selection in SDSRA, governed by a softmax function over the skills' relevance scores, allows for more adaptive and varied action selection.

Formally, let S be the set of all skills in SDSRA, and let r_i be the relevance score of skill i. Then the probability of selecting an action a given state s under policy π^{SDSRA} is given by a mixture of policies corresponding to each skill:

$$\pi^{\text{SDSRA}}(a|s) = \sum_{i=1}^{N} P(i|s)\pi_{\text{skill}_i}(a|s), \tag{12}$$

where $P(i|s)$ is the softmax probability of selecting skill i.

The mixture model in SDSRA leads to higher entropy due to the increased unpredictability and diversity of actions available from multiple skill policies, enhancing the system's randomness compared to a single-policy approach like SAC. Therefore, the expected entropy of π^{SDSRA} is greater than the expected entropy of π^{SAC}:

$$\mathbb{E}_{a\sim\pi^{\text{SDSRA}}}[-\log \pi^{\text{SDSRA}}(a|s)] \geq \mathbb{E}_{a\sim\pi^{\text{SAC}}}[-\log \pi^{\text{SAC}}(a|s)]. \tag{13}$$

Moreover, SDSRA's rapid exploration of diverse actions facilitates faster convergence to high-entropy policies compared to SAC:

$$t_{\text{SDSRA}}(\epsilon) \leq t_{\text{SAC}}(\epsilon), \tag{14}$$

thus completing the proof. □

B SDSRA ALGORITHM

Algorithm 1 Soft Actor-Critic with Skill-Driven Skill Recombination Algorithm (SDSRA)

Initialize action-value functions $Q_{\theta_1}, Q_{\theta_2}$ with parameters θ_1, θ_2
Initialize the policy π_ϕ with parameters ϕ
Initialize target value parameters $\theta'_1 \leftarrow \theta_1, \theta'_2 \leftarrow \theta_2$
Initialize skill set $S = \{\pi_{\theta_{\text{skill}_i}}\}_{i=1}^N$ with parameters θ_{skill_i}
Initialize relevance scores $r_i \leftarrow c, \forall i \in \{1, \ldots, N\}$
Initialize replay buffer D
for each iteration **do**
  for each environment step **do**
    Sample skill index i using probabilities $P(i|s) = \frac{e^{r_i}}{\sum_{j=1}^N e^{r_j}}$
    Select action $a_t \sim \pi_{\theta_{\text{skill}_i}}(s_t)$
    Execute a_t and observe reward r_t and new state s_{t+1}
    Store transition tuple $(s_t, a_t, r_t, s_{t+1}, i)$ in buffer D
  end for
  for each gradient step **do**
    Randomly sample a batch of transitions from D
    Compute target values using the Bellman equation
    Update $Q_{\theta_1}, Q_{\theta_2}$ by minimizing the loss:

$$L(\theta_i) = \mathbb{E}_{(s,a,r,s')\sim D}\left[\left(Q_{\theta_i}(s,a) - (r + \gamma(\min_{j=1,2} Q_{\theta'_j}(s', \pi_\phi(s')) - \alpha \log \pi_\phi(a|s'))\right)^2\right]$$

    Update policy π_ϕ using the policy gradient:

$$\nabla_\phi J(\pi_\phi) = \mathbb{E}_{s\sim D, a\sim\pi_\phi}\left[\nabla_\phi \log \pi_\phi(a|s) Q_{\theta_1}(s,a)\right]$$

    Update target networks: $\theta'_i \leftarrow \tau\theta_i + (1-\tau)\theta'_i$
  end for
  for each skill update interval **do**
    Evaluate and update the performance of each skill $\pi_{\theta_{\text{skill}_i}}$
    Update relevance scores r_i based on the performance
  end for
end for

B.1 SKILL ALGORITHM OVERVIEW

In the SDSRA framework, skills are conceptualized as distinct GaussianPolicy objects, each embodying a unique strategy set. Unlike static hyperparameters, these skills are dynamic, evolving through interaction with the environment. This evolution is guided by a learning process that adjusts each skill's policy parameters, optimizing for both predictive accuracy and policy entropy.

Skills are selected based on their relevance to the current state, with relevance scores r_i dynamically updated based on effectiveness in recent scenarios. The agent employs a probabilistic approach to select skills during each action phase:

$$P(i|s) = \frac{\exp(r_i(s))}{\sum_j \exp(r_j(s))}, \tag{15}$$

where $r_i(s)$ denotes the relevance score of skill i in state s, and the denominator normalizes the probabilities across all skills.

The action a taken by the agent in state s under this skill-based framework is determined by:

$$a = \sum_{i=1}^N P(i|s) \cdot a_{\text{skill}_i}(s), \tag{16}$$

where $a_{\text{skill}_i}(s)$ is the action proposed by skill i in state s.

This adaptive, skill-driven approach not only enhances the agent's performance but also ensures a more versatile and efficient response to diverse challenges encountered in the environment.