Sparse Reward Exploration via Novelty Search and Emitters
Giuseppe Paolo, Alexandre Coninx, Stéphane Doncieux, Alban Laflaquière

To cite this version:
Giuseppe Paolo, Alexandre Coninx, Stéphane Doncieux, Alban Laflaquière. Sparse Reward Exploration via Novelty Search and Emitters. The Genetic and Evolutionary Computation Conference 2021 (GECCO 2021), Jul 2021, Lille, France. 10.1145/3449639.3459314. hal-03200022

HAL Id: hal-03200022
https://hal.science/hal-03200022
Submitted on 16 Apr 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Sparse Reward Exploration via Novelty Search and Emitters
Giuseppe Paolo, Alexandre Coninx, Stéphane Doncieux, Alban Laflaquière

To cite this version:
Giuseppe Paolo, Alexandre Coninx, Stéphane Doncieux, Alban Laflaquière. Sparse Reward Exploration via Novelty Search and Emitters. The Genetic and Evolutionary Computation Conference 2021 (GECCO 2021), Jul 2021, Lille, France. 10.1145/3449639.3459314 . hal-03200022

HAL Id: hal-03200022
https://hal.archives-ouvertes.fr/hal-03200022
Submitted on 16 Apr 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
ABSTRACT

Reward-based optimization algorithms require both exploration, to find rewards, and exploitation, to maximize performance. The need for efficient exploration is even more significant in sparse reward settings, in which performance feedback is given sparingly, thus rendering it unsuitable for guiding the search process. In this work, we introduce the SparsE Reward Exploration via Novelty and Emitters (SERENE) algorithm, capable of efficiently exploring a search space, as well as optimizing rewards found in potentially disparate areas. Contrary to existing emitters-based approaches, SERENE separates the search space exploration and reward exploitation into two alternating processes. The first process performs exploration through Novelty Search, a divergent search algorithm. The second one exploits discovered reward areas through emitters, i.e., local instances of population-based optimization algorithms. A meta-scheduler allocates a global computational budget by alternating between the two processes, ensuring the discovery and efficient exploitation of disjoint reward areas. SERENE returns both a collection of diverse solutions covering the search space and a collection of high-performing solutions for each distinct reward area. We evaluate SERENE on various sparse reward environments and show it compares favorably to existing baselines.

KEYWORDS

Novelty search, sparse rewards, emitters, evolutionary algorithm, quality diversity

1 INTRODUCTION

Embodied agents solve tasks by learning a policy dictates how to act in different situations. This is done by evaluating the agent’s performance on the task through a reward function. Learning strategies for such agents can be divided in two groups: step-based and episode-based [33]. The former expects a reward after each step. On the contrary, episode-based ones need rewards only at the end of each training episode. So much reliance on the reward forces some constraints: the reward function must be well designed and provide feedback as frequently as possible. In many complex scenarios where the reward is given only if specific conditions are met, such constraints are impossible to respect. These are known as sparse reward situations and can prove very difficult to tackle. In this work, we consider sparse reward settings in which the reward is obtained only in small disjoint areas of the whole search space. One example would be a robotic arm trying to push an object to one of a few given positions. The search space consists of all the positions the object can achieve, while the reward is given only if the object reaches one...
of the goals. In such situations, a standard Reinforcement Learning (RL) [35] agent typically explores by trying random actions. The probability of finding a reward this way tends to zero, rendering learning impractical. Therefore, the way exploration is performed is fundamental when dealing with sparse rewards settings.

In recent years, many algorithms have been proposed to solve this problem [5, 7, 15, 24, 36]. Among them, Novelty Search (NS) is an evolutionary algorithm that focuses only on exploration, while ignoring any possible reward [24]. By doing so, NS tends towards a uniform exploration of the search space [14], avoiding the need for a well-defined reward function. At the same time, its strength is also its limitation: considering all the non-rewarding areas as valuable as the rewarding ones prevents the algorithm from finding the best possible solutions. Augmenting NS with the ability to shift its focus from pure exploration to reward exploitation could help address this issue. One possible way of doing so is by using multi-objective optimization methods like NSGA-II [12]. However, merging exploration and exploitation through a Pareto front can degrade the exploring power of the algorithm. A different approach is taken by Quality-Diversity (QD) algorithms, a family of methods that build a set of both diverse and high-quality solutions [32].

In this work, we introduce Sparse Reward Exploration via Novelty search and Emitters (SERENE), a QD algorithm addressing sparse reward problems. SERENE augments NS with emitters [17] to perform rewards maximization while keeping its exploration ability, thanks to a clear separation between the exploration and exploitation. Introduced as a way to improve the efficiency of MAP-Elites (ME) [27] in the CMA-ME method [17], emitters are instances of reward-based evolutionary algorithms scheduled to perform a local search in the search space. In the original formulation, ME acts as a scheduler by initializing emitters in different areas of the search space. The emitters then perform both exploration and exploitation of the reward, leading to degraded performances in settings with very sparse rewards, where not all policies can obtain a reward. Conversely, SERENE decouples exploration from exploitation to better deal with such situations. The former is performed through NS, completely ignoring the reward. Once a reward area is found, SERENE spawns emitters focusing solely on its maximization, with no dependency on the exploration process. This allows our algorithm to shift its focus between exploration and exploitation at any moment. Persisting in exploring even after some reward areas have been found is essential, since other reward areas could be present in the search space.

In the following, we will discuss other works tackling the sparse rewards problem in Section 2. In Section 3 we will analyze the methods SERENE draws from and explain in detail the concept of emitter. The method itself will be introduced in Section 4, tested in Section 5, and the results discussed in Section 6. We will conclude with Section 7 by pointing at possible extensions and improvements.

2 RELATED WORK

2.1 Sparse rewards

Step-based algorithms expect a reward at every step, making dealing with sparse reward particularly difficult; this is the case for many RL algorithms. Following the recently increased interest in the problem, many new approaches have been proposed to deal with this sparsity. Some methods work on improving the data efficiency of the search [1, 28]. Others introduce some artificial curiosity by counting the number of times a state is visited, and push exploration by making less-visited states more rewarding [2, 36]. Another strategy uses additional shaped rewards to aid in approaching the task [37]. A population of RL agents can also be used to increase exploration while learning a policy [13, 22, 30]. However, none of these methods explicitly separates exploration and exploitation.

Episode-based methods, and more specifically evolutionary algorithms [38], are better suited for dealing with sparse reward settings, given the more relaxed dependency on the reward. For this reason, many works combined evolutionary algorithms with RL. Some works use Evolutionary Strategy (ES) to collect the data over which a RL agent is then trained [23, 31]. These approaches take advantage of the exploration of evolution-based methods and the higher data efficiency of RL.

Separating exploration from exploitation has proven useful for overcoming deceptive gradients in sparse reward settings [6, 7, 15]. In the work from Colas et al. [7], a reward-agnostic exploration phase is first performed through Goal Exploration Processes [18]; then a RL-based policy is learned on the collected data. A similar two-step process is used in GO-Explore [15] to solve ATARI games. Conversely, QD-RL [6] separates exploration and exploitation by taking advantage of a QD population trained through an actor-critic approach. Half of the population is optimized for quality, while the other half is optimized for diversity.

2.2 Divergent search algorithms

Divergent search methods, as the one used by Cideron et al. [6], generate solutions by looking for a set of diverse policies. This prevents getting stuck in local optima that could limit the performance of the solutions. One of the first algorithms developed in this direction is NS [24]. Since, many divergent search algorithms have been developed, using different mechanisms to drive the search: curiosity [34], empowerment [5], surprise [20], diversity [10, 11, 16, 32], and novelty [25].

QD [11, 32] is a family of divergent search algorithms that searches for a set of diverse solutions while also improving on their quality. A well-known QD algorithm is ME [27], a method that drives the search for novel policies by discretizing the search space into a grid and filling its cells with high-performing solutions.

QD algorithms have been extended by combining them with ES [3] to increase their efficiency and speed of convergence [8, 9, 17]. Conti et al. [8] augment an ES with NS’s novelty objective to look for novel solutions while improving their performances. At the same time, the approach followed by Fontaine et al. [17], and then extended by Cully [9], uses ME as a scheduler for modified instances of CMA-ES [21], named emitters. Exploration of the search space and reward exploitation are both performed through emitters. However, fusing the two aspects can limit performances in sparse reward settings where reward-based algorithms struggle to explore.

In this work, we take inspiration from CMA-ME [17] by combining emitters with NS to keep the two aspects, i.e. exploration and exploitation, separated. This allows our method to avoid the shortcomings of exploring through emitters. In the next section, we
describe in detail how both NS and emitters work before detailing the functioning of SERENE.

3 BACKGROUND

The notation used in this work is based on the one introduced by Doncieux et al. [14] and is directly inspired by the RL literature.

3.1 Novelty Search

NS is an evolutionary algorithm that replaces the usual fitness metrics used by evolutionary algorithms with a novelty metric. This metric pushes the search towards novel areas of the search space. The novelty is calculated in a hand-defined behavior space $\mathcal{B}$ in which the behavior of each policy $\theta_i \in \Theta$ is represented. When a policy is evaluated, it traverses a sequence of states $\tau = [s_0, \ldots, s_T]$, where the initial state $s_0$ is constant for every policy. Traversed states are observed through some sensors generating a sequence of observations $\tau_o = [o_0, \ldots, o_T]$, with $o \in \mathcal{O}$. From the sequence of observations it is possible to extract a representation $b_i \in \mathcal{B}$ of the policy’s behavior by using an observer function $O_B : \mathcal{O} \to \mathcal{B}$. This whole process can be summarized by introducing a behavior function directly mapping a policy $\theta_i$ to its behavior descriptor $b_i$:

$$\phi(\theta_i) = b_i.$$  

(1)

Once computed, the behavior descriptors are used to calculate the policies’ novelty as:

$$\eta(\theta_i) = \frac{1}{J} \sum_{j \in J} \text{dist}(b_i, b_j) = \frac{1}{J} \sum_{j \in J} \text{dist}(\phi(\theta_i), \phi(\theta_j)), \tag{2}$$

where $J$ is the set of indexes of the $k$ policies closest to $\theta_i$ in the behavior space.

The novelty is calculated at each generation and used to choose the policies for the next generation. Moreover, $N_Q$ policies are sampled to be stored into an archive, returned as outcome of the algorithm. This archive is also used to keep track of the already explored areas of the space $\mathcal{B}$. This is done by choosing the $|J|$ closest neighbors used in equation (2) not only from the current population and offspring but also from the archive. By choosing the most novel policies from the previous generation to compose the population, the search is always pushed towards less explored areas of $\mathcal{B}$. Notwithstanding its capacity for exploration, NS cannot exploit the rewards potentially found during the search. This can lead to slow learning solutions.

3.2 Emitters

An emitter [9, 17] is an instance of a reward-based Evolutionary algorithm (EA), such as CMA-ES [21]. Its objective is to rapidly examine a small area of the search space while optimizing on the reward. The CMA-ME algorithm [9, 17] combines emitters with ME [27], by using the latter as a scheduler for the emitters evaluation. It works by initializing a population of policies $\theta$ by sampling their parameters from a distribution $\mathcal{N}(\mu, \Sigma)$ and adding them to the ME archive. The algorithm then samples one of these policies and uses it to initialize the population of the emitter $E_i$. At this point, $E_i$ is evaluated until a termination criterion is met; e.g., a lack of increase of the reward found. Moreover, the policies found during the evaluation of the emitter are added to the ME archive according to ME addition strategy. After the termination of $E_i$, a new emitter is initialized by sampling another policy from the archive. This is repeated until the whole evaluation budget is depleted.

Different types of algorithms can be used as emitters, changing how the search is performed and how the policies are selected. This shows the flexibility of the approach. At the same time, previous works [9, 17] perform exploration through reward-following emitters. This reduces performances in situations where the reward is very sparse and many of the policies do not get any reward. Decoupling the exploitation of the reward from the exploration allows to more efficiently deal with sparse rewards settings [7].

4 METHOD

SERENE disentangles the exploration of the behavior space $\mathcal{B}$ from the exploitation of the reward through a two-steps process. In the first phase, called exploration phase, $\mathcal{B}$ is explored by performing NS. As per equation (1), the policies $\theta_i$ found during exploration are assigned a behavior descriptor $\phi(\theta_i)$. A policy obtaining a reward means that its $\phi(\theta_i)$ belongs to the subspace of rewarding behaviors $\mathcal{B}_{\text{Rew}} \subseteq \mathcal{B}$. It is in this subspace that the exploitation of the reward happens. This is done in the second phase, called exploitation phase, in which emitters are initialized using the rewarding policies found in $\mathcal{B}_{\text{Rew}}$ during exploration. During the exploitation phase the most rewarding policies are stored to be returned as result of the algorithm. Moreover, particularly novel policies found by the emitters are also stored. By launching emitters only in the neighborhoods of the reward areas, SERENE keeps the exploitation of the reward separated from the exploration of the search space. This results in taking the best of both worlds: the exploration power of NS and the focused exploitation of reward-based algorithms.

The exploitation and exploration phases are alternated repeatedly through a meta-scheduler. This scheduler divides a total evaluation budget $B_u$ in smaller chunks of size $K_u$ and assigns them to either one of the two phases. The whole process is illustrated in Figure 1 and described in Algorithm 1.

Algorithm 1: SERENE

| INPUT: | evaluation budget $B_u$, budget chunk size $K_u$, population size $M$, emitter population size $M_e$, offspring per policy $m$, mutation parameter $\sigma$, number of policies added to novelty archive $N_Q$ |
| RESULT: | Novelty archive $A_{\text{Nov}}$, rewarding archive $A_{\text{Rew}}$; $A_{\text{Nov}} = \emptyset$; $A_{\text{Rew}} = \emptyset$; |
| $Q_{\text{Em}} = 0$; $Q_{\text{Cand,Nov}} = 0$; $Q_{\text{Cand,Em}} = 0$; |
| Sample population $I_0$; |
| Split $B_u$ in chunks of size $K_u$; |
| while $B_u$ not depleted do |
| if $I_0$ then |
| Evaluate $\theta_i \forall \theta_i \in I_0$; |
| Calculate $b_i = \phi(\theta_i) \in \mathcal{B}, \forall \theta_i \in I_0$; |
| ExplorationPhase ($K_u$, $m$, $\sigma$, $A_{\text{Nov}}$, $Q_{\text{Cand,Em}}$, $I_0$, $N_Q$); |
| if not $Q_{\text{Cand,Em}} = 0$ or not $Q_{\text{Em}} = 0$ then |
| ExploitationPhase ($K_u$, $m$, $Q_{\text{Em}}$); |
| $A_{\text{Nov}}$, $A_{\text{Rew}}$, $M_e$; |
To keep track of policies generated during the different phases, SERENE uses the following buffers and containers:

- **novelty archive \( \mathcal{A}_{\text{Nov}} \):** a repertoire of the novel policies found during the exploration phase, and returned as first output of SERENE;
- **reward archive \( \mathcal{A}_{\text{Rew}} \):** a repertoire of rewarding policies found during the exploitation phase, returned as second output of SERENE;
- **candidates emitter buffer \( Q_{\text{Cand\_Em}} \):** a buffer containing the rewarding policies \( \phi(\theta_j) \in \mathcal{B}_{\text{Rew}} \) found during the exploration phase and used in the exploitation phase to initialize emitters;
- **emitter buffer \( Q_{\text{Em}} \):** a buffer containing all the initialized emitters to be evaluated during the exploitation phase;
- **novelty candidates buffer \( Q_{\text{Cand\_Nov}} \):** a buffer containing the most novel policies found by the emitter. Each emitter has its own instance of this buffer and the policies in it are sampled for addition to the novelty archive \( \mathcal{A}_{\text{Nov}} \) once the emitter is terminated.

A high-level overview of how these sets interact during the two phases is given in Figure 2, and a more detailed description is proposed in the two following subsections.

### Exploration phase

SERENE starts by generating an initial population \( \Gamma_0 \) of size \( M \). This is done by sampling the parameters of the population’s policies \( \theta_j \) from a normal distribution \( \mathcal{N}(0, I) \). The population is used to explore the behavior space \( \mathcal{B} \) through NS. At each generation \( g \), a mutation operator generates \( m \) new policies \( \theta_j^g \) (offspring) from each of the policies \( \theta_j \in \Gamma_g \):

\[
\forall j, i \in \{1, \ldots, M\} \times \{1, \ldots, m\}, \theta_j^g = \theta_j + \epsilon, \text{ with } \epsilon \sim \mathcal{N}(0, \sigma I).
\] (3)

The resulting offspring population \( \Gamma_g^m \), of size \( m \times M \), is then evaluated to obtain the behavior descriptors \( \phi(\theta_j^g) = b_j^g \in \mathcal{B} \). The novelty of \( \Gamma_g \) and \( \Gamma_g^m \) is then calculated using equation (2) and is used to generate the next generation population \( \Gamma_{g+1} \) by taking the most novel policies from the current population and the offsprings. At the same time, \( N_Q \) policies among the offsprings are uniformly sampled to be added to the novelty archive \( \mathcal{A}_{\text{Nov}} \). Finally, all the rewarding policies found are stored in the candidates emitters buffer \( Q_{\text{Cand\_Em}} \).

The process just described is detailed in Algorithm 2.

The exploration phase is executed for the \( K_{\text{Bud}} \) evaluation steps in the given budget chunk, where each evaluation step corresponds to one policy evaluation. Once the chunk is depleted, the scheduler assigns the next chunk to the exploitation phase only if \( Q_{\text{Cand\_Em}} \neq \emptyset \). On the contrary, another exploration phase is performed. This means that in the worst case scenario where no reward can be discovered, i.e. \( \mathcal{B}_{\text{Rew}} = \emptyset \), SERENE performs exactly like NS.

### exploitation phase

The exploitation phase consists of two sub-steps: the bootstrapping step, in which the policies in the candidates emitter buffer \( Q_{\text{Cand\_Em}} \) are used to initialize and bootstrap emitters, and the emitter step, in which the initialized emitters are evaluated.

#### Bootstrap step

During this step, emitters are initialized from the rewarding policies \( \theta_j \) in the candidates emitter buffer, and their potential for reward improvement evaluated. This insures that only emitters capable of improving the rewards are considered for full evaluation, reducing wasted evaluation budget. The policies used to initialize the emitters are selected according to their novelty with respect to the reward archive \( \mathcal{A}_{\text{Rew}} \). This enables SERENE to focus on less explored areas of the rewarding behavior space \( \mathcal{B}_{\text{Rew}} \). The whole bootstrapping phase lasts \( K_{\text{Bud}}/3 \) evaluations.

As discussed in Section 3.2, an emitter is an instance of a reward-based EA. In contrast to previous work [9, 17], in this work we do not use estimation-of-distribution algorithms like CMA-ES [21] because the estimation of the covariance matrix \( \Sigma \) is unreliable when the population size is smaller than the dimension of the parameter space \( \Theta \). CMA-ES circumvents the issue by using information from previous generations to calculate \( \Sigma \). While stabilizing \( \Sigma \), this also leads to a less efficient use of the evaluation budget. Hence, in this work we use as emitter an elitist evolutionary algorithm that does not require any estimation of distribution. Conversely, it composes its population with the most rewarding policies from the previous generation’s population and offspring, while the offspring are generated according to equation 3.

An emitter \( \hat{E}_i \) based on this algorithm consists of: a population \( P \) containing \( M_E \) policies \( \theta_i \); a population of offspring \( P^m \) of size \( m \times M_E \); a generation counter \( \gamma \); a tracker for the maximum reward found so far \( R_F \); an improvement measure \( I(\cdot) \); a novelty measure \( \eta_j \) equal to the novelty of the policy used to initialize the emitter; and a novel candidate buffer \( \Gamma_{\text{Cand\_Nov}} \). The emitter \( \hat{E}_i \) is initialized from a policy \( \theta_i \) in the candidates emitter buffer by sampling its initial population \( P_0 \) from the distribution \( \mathcal{N}(\theta_i, \sigma I) \). To keep the emitter’s exploration local and prevent overlapping with the search space of possible nearby emitters, we initialize \( \sigma \) as:

\[
\sigma_l = \min_{j} \left( \frac{|\text{dist}(\theta_i, \theta_j)|}{3}, \forall \theta_j \in \Gamma_{g}^m \cup \Gamma_g \right).
\] (4)

This shapes \( \mathcal{N}(\theta_i, \sigma I) \) such that all other \( \theta_j \) are at least 3 standard deviation away from its center. Once \( \hat{E}_i \) has been initialized, its potential is evaluated by running it for \( \lambda \) generations and calculating its emitter improvement \( I(\hat{E}_i) \). This improvement is defined as the difference between the average rewards obtained during the most

---

**Algorithm 2: Exploration Phase**

**INPUT:** budget chunk \( K_{\text{Bud}} \), number of offspring per parent \( m \), mutation parameter \( \sigma \), novelty archive \( \mathcal{A}_{\text{Nov}} \), candidate emitters buffer \( Q_{\text{Cand\_Em}} \), population \( \Gamma_g \), number of policies \( N_Q \);

**while** \( K_{\text{Bud}} \) **not depleted** **do**

- Generate offspring \( \Gamma_g^m \) from population \( \Gamma_g \);
- Evaluate \( \theta_i, \forall \theta_i \in \Gamma_g^m \);
- Calculate \( b_1 = \phi(\theta_i) \in \mathcal{B} \), \( \forall \theta_i \in \Gamma_g^m \);
- Calculate \( \eta(\theta_i) = \frac{1}{|\mathcal{B}|} \sum_{j \in J} \text{dist}(b_j, b_j), \forall \theta_i \in \Gamma_g^m \cup \Gamma_g \);
- \( \mathcal{A}_{\text{Nov}} \leftarrow N_Q \) samples from \( \Gamma_g^m \);
- **if** \( \phi(\theta_i) \in \mathcal{B}_{\text{Rew}} \) **then**
  - \( Q_{\text{Cand\_Em}} \leftarrow \theta_i \)
- **Generate** \( \Gamma_{g+1} \) from most novel \( \theta_i \in \Gamma_g^m \cup \Gamma_g \);

---

---
Our approach combines exploration and exploitation in a novel manner. We use novelty search to guide exploration towards areas that are both less explored and more promising. This is achieved by maintaining a buffer of novelty candidates that are periodically evaluated for their potential contribution to the reward. The selection of the next evaluative step is based on the novelty of the candidates, with a focus on those that have not been explored recently.

The novelty calculation is performed using the formula:

\[
I(E_i) = \frac{1}{\lambda_M} \left( \sum_{y=1}^{T} \sum_{j=0}^{M_y} \sum_{j=0}^{M_{ij}} r(y,j) - \sum_{y=0}^{\gamma_y} \sum_{j=0}^{M_{ij}} r(y,j) \right). \tag{5}
\]

Here, \(T\) is the last evaluated generation, \(r(y,j)\) is the reward of policy \(\hat{\theta}_j \in P_y\), and \(\gamma_y\) is the generation at which the emitter is at the beginning of the exploitation phase; it is always \(\gamma_0 = 0\) for an emitter in the bootstrap step. If \(I(E_i) \leq 0\), the chances for the emitter to find better solutions than the initial ones are low, so it is not worth allotting more budget to its evaluation. On the contrary, \(I(E_i) > 0\) means that the emitter has high potential for improvement. Thus all the initialized emitters for which \(I(E_i) > 0\) are added to the emitter buffer \(Q_{EIm}\) for further evaluation.

**Emitter step.** The initialized emitters in the emitter buffer \(Q_{EIm}\) are run during this step. It starts by calculating the pareto front between the improvement \(I(E_i)\) and the novelty \(\eta(E_i)\) of each of the emitters \(E_i\) in the emitter buffer. The emitter to run is then sampled from the front of the non-dominated emitters. Using both the novelty and the fitness to select which emitter to run allows SERENE to focus both on the less explored and most promising areas of \(S_{Rew}\).

The policies \(\hat{\theta}_j\) generated by an emitter can be stored either for the reward they achieve or for their novelty. At every generation \(y\) all the policies \(\hat{\theta}_j\) in the current population with a reward \(r(\hat{\theta}_j) > R_{y-1}\) are added to the reward archive \(A_{Rew}\). Additionally, the policies \(\hat{\theta}_j\) with a novelty higher than the emitter novelty \(\eta_i\) are stored into the emitter’s novelty candidates buffer \(Q_{Cand_Nov}\). The emitter \(E_i\) is run until either the given budget chunk is depleted or a termination condition is met. In the first case, SERENE recalculates \(I(E_i)\) from the beginning of the emitter phase and assigns the next budget chunk to the exploration phase. On the contrary, if a termination condition is met, \(E_i\) is discarded and another emitter to evaluate is sampled from the Pareto front. There can be multiple termination conditions. The one used in this work is inspired from the stagnation criterion [21], stopping the emitter when there is no more improvement on the reward. A detailed definition of the termination condition is presented in Appendix C. Before starting the new emitter evaluation, \(N_{Q}\) policies from the terminated emitter’s novelty candidates buffer are uniformly sampled to be added to \(A_{Nov}\). In addition to saving particularly novel solutions as part of the final result, this prevents the exploration phase from re-exploring areas covered by emitters during the exploitation phase.

The whole exploitation phase is detailed in Algorithm 3. The code repository is available at: github.com/GPaolo/SERENE.

## 5 EXPERIMENTS

In this section we want to verify if SERENE can efficiently deal with sparse reward settings, find all disjoint reward areas, and optimize the reward in each of them. For the evaluation, we consider the four sparse rewards environment illustrated in Figure 3:

- **Curling:** A two Degrees of Freedom (DoF) robotic arm controlled by a 3 layers Neural Network (NN) with each layer of size 5. The arm has to push the blue ball into one of the two goal areas shown in orange and green. A reward is provided only if the ball stops in one of the two areas. The controller takes as input a 6-dimensional vector containing the ball pose \((x, y)\), and the two joints angles and velocities. The output of the controller is the speed of each joint at the next timestep. The size of the parameter space \(\Theta\) is 94, and each policy is run in the environment for 500 timesteps.

- **Hardmaze:** Introduced in the original NS paper [24], it consists of a two-wheeled robot, in blue, whose task is to navigate the maze and reach either one of the green and orange areas. Contrary to the original formulation, in which only a single binary-reward area was present, here the reward areas are two and provide continuous rewards. At the same time, the reward is only given if the robot stops in one of the two areas. The robot is controlled by a 2-layers NN with each layer of size 5. The controller takes as input the reading of the 5 distance sensors mounted on the robot, shown in red in Figure 3. Its output is the 2-dimensional vector containing the speed of the 2 wheels at the next timestep. The size of the parameter space \(\Theta\) is 63, and each policy is run in the environment for 2000 timesteps.

- **Redundant arm:** A 20-DoF robotic arm [26] in which the arm’s end-effector has to reach one of the 3 colored goal areas. The arm is controlled by a NN with 2 layers of size 5. The controller takes as input the 20-dimensional vector of each joint’s position, and outputs the 20-dimensional joint’s torque vector. The size of the parameter space \(\Theta\) is 228, and each policy is run in the environment for 600 timesteps.

- **Robotic ant maze:** Introduced by Cideron et al. [6], it consists in a 4-legged robotic ant in a maze. There are two goal areas and the task is for the ant to navigate the maze and reach the center of one of the two areas. The controller takes as input a 6-dimensional vector containing the ant’s position \((x, y)\), and the two joints angles and velocities. The output of the controller is the speed of each joint at the next timestep. The size of the parameter space \(\Theta\) is 125, and each policy is run in the environment for 2000 timesteps.
Robotic ant maze

6 RESULTS

This section discusses the results obtained during the experiments.

6.1 Budgeting

Balancing the exploration of the search space and the exploitation of the reward is an aspect of paramount importance for reward-based algorithms. Even more so in sparse reward environments. This balance can be studied by analyzing the amount of evaluation budget dedicated to either one of the two aspects. The exploration budget consists of all the evaluated policies that did not get any reward. On the contrary, the exploitation budget is obtained by counting all the evaluated policies that collected some reward from one of the reward areas.

Figure 3: Testing environments: Curling, HardMaze, Redundant arm, Robotic ant maze.

Baselines

We compare SERENE against 5 different baselines:

- NS[24]: vanilla NS, that performs pure exploration and does not attempt to improve on the reward;
- NSGA-II[12]: a multi-objective evolutionary algorithm optimizing both the novelty and the reward;
- CMA-ME[17]: the original algorithm introducing emitters that combines ME with emitters over a $50 \times 50$ grid covering the behavior space of all environments. Among the various emitters proposed by the authors we selected the “optimizing” emitter;
- ME[27]: vanilla MAP-Elites that uses a $50 \times 50$ grid to cover the behavior space of every environment;
- RND: pure random search in which no selection happens, and every policy is sampled from a normal distribution $N(0, I)$.

The parameters used during the experiments are listed in Appendix B. The statistical results are computed over 15 runs for each experiment.

Algorithm 3: Exploitation Phase

<table>
<thead>
<tr>
<th>INPUT:</th>
<th>budget chunk $K_{bud}$, candidate emitters buffer $Q_{Cand, Em}$, number of bootstrap generations $\lambda$, emitter population size $M_C$, number of offspring per policy $m$, emitters buffer $Q_{Em}$, rewarding archive $A_{Res}$, novelty archive $A_{Nov}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>while</td>
<td>$K_{bud}$/not depleted do</td>
</tr>
<tr>
<td>Select</td>
<td>most novel policy $\theta_i$ from $Q_{Cand, Em}$;</td>
</tr>
<tr>
<td>Calculate</td>
<td>$\sigma_i$;</td>
</tr>
<tr>
<td>Initialize:</td>
<td>$E_i, Q_{Cand, Nov}^i = \emptyset$, and $P_0$;</td>
</tr>
<tr>
<td>for $\gamma \in {0, \ldots, \lambda}$ do</td>
<td></td>
</tr>
<tr>
<td>if $P_0$ then</td>
<td></td>
</tr>
<tr>
<td>Evaluate $\tilde{\theta}_j, \forall \tilde{\theta}_j \in P_0$;</td>
<td></td>
</tr>
<tr>
<td>Generate offspring population $P^m_f$ from $P_f$;</td>
<td></td>
</tr>
<tr>
<td>Evaluate $\tilde{\theta}_j, \forall \tilde{\theta}_j \in P^m_f$;</td>
<td></td>
</tr>
<tr>
<td>Generate $P_{f+1}$ from best $\tilde{\theta}_j \in P^m_f \cup P_f$;</td>
<td></td>
</tr>
<tr>
<td>Calculate $I(E_i)$;</td>
<td></td>
</tr>
<tr>
<td>if $I(E_i) &gt; 0$ then</td>
<td></td>
</tr>
<tr>
<td>$Q_{Em} \leftarrow E_i$;</td>
<td></td>
</tr>
<tr>
<td>Calculate pareto fronts in $Q_{Em}$;</td>
<td></td>
</tr>
<tr>
<td>while $2/3K_{bud}$ not depleted do</td>
<td></td>
</tr>
<tr>
<td>Sample $E_i$ from non-dominated emitters in $Q_{Em}$;</td>
<td></td>
</tr>
<tr>
<td>while not terminate($E_i$) do</td>
<td></td>
</tr>
<tr>
<td>Generate offspring population $P^m_f$ from $P_f$;</td>
<td></td>
</tr>
<tr>
<td>Evaluate $\tilde{\theta}_j, \forall \tilde{\theta}_j \in P^m_f$;</td>
<td></td>
</tr>
<tr>
<td>$A_{Res} \leftarrow \tilde{\theta}_j, \forall \tilde{\theta}_j \in P^m_f \mid r(\tilde{\theta}_j) &gt; y_f$;</td>
<td></td>
</tr>
<tr>
<td>$Q_{Cand, Nov}^i \leftarrow \tilde{\theta}_j, \forall \tilde{\theta}_j \in P^m_f \mid \eta(\tilde{\theta}_j) &gt; \eta_i$;</td>
<td></td>
</tr>
<tr>
<td>Generate $P_{f+1}$ from best $\tilde{\theta}_j \in P^m_f \cup P_f$;</td>
<td></td>
</tr>
<tr>
<td>Update $I(E_i)$ and $R_f$;</td>
<td></td>
</tr>
<tr>
<td>if terminate($E_i$) then</td>
<td></td>
</tr>
<tr>
<td>$A_{Nov} \leftarrow N(Q)$ samples from $Q_{Cand, Nov}^i$;</td>
<td></td>
</tr>
<tr>
<td>Discard emitter $E_i$;</td>
<td></td>
</tr>
</tbody>
</table>

them. The robot is controlled by a 3-layers NN, with each layer of size 10. The input of the controller is the 29-dimensional observation returned by the environment at each step, while its output is the 8-dimensional joint’s torque control. The size of the parameter space $\Theta$ is 574, and each policy is run in the environment for 3000 timesteps.

For all environments, the reward is given only if inside the reward area, and as a continual value in the $[0, 1]$ range. The reward varies with the distance to the center of the area and is highest directly at the center.

It can be expressed as:

$$r(\theta) = \begin{cases} 0, & \text{if } d_r > \text{radius} \\ \frac{\text{radius} - d_r}{\text{radius}}, & \text{if } d_r \leq \text{radius} \end{cases}$$

where $d_r$ is the distance from the center and $\text{radius}$ is the radius of the reward area.
As Figure 4 shows, SERENE has a more balanced budget split between exploration (in blue) and exploitation (other colors) compared to the other baselines. In situations in which exploration is harder, a bigger part of the budget is assigned to exploration rather than exploitation of the reward. This is the case for the robotic ant maze environment. Additionally, due to the way emitters are selected, the algorithm can shift its exploitation focus among the different reward areas. Figure 4 shows that most of SERENE’s exploitation budget is assigned to the green reward area in the Curling, Hard maze and Robotic ant maze environments. As it can be seen in Figure 3, this area is more difficult to discover and to reach with respect to the orange area. This makes the exploitation of the orange reward area faster, having both the novelty and the improvement go to zero rapidly. On the contrary, being the green area harder to reach, its novelty will remain higher for longer, making SERENE select more emitters focused on it. The effect can also be seen in Figure 6, where the reward for area 1 quickly reaches higher values compared to the one of reward area 2. At the same time, in the Redundant arm environment where the 3 reward areas are equally easy to discover and to reach, this effect is less present and the exploitation budget is more evenly split between them. The ability to switch its focus is similar to intrinsic motivation based methods [4, 19] and allows SERENE to reach high rewards in all environments, while other baselines exhibit a less balanced distribution of the evaluation budget, as they do not explicitly separate exploration from exploitation.

### 6.3 Exploitation

Figure 6 shows the average maximum reward achieved by the algorithms in the reward areas of all environments. Emitters solely focusing on exploiting the reward allow SERENE to reach almost the maximum reward on the easiest to reach reward areas in less than $10^5$ evaluations. High rewards are also achieved on the harder to reach areas, even if the required time is higher. On the contrary, ME improves on the reward at a much slower pace. This is likely due to the random selection of policies from the archive to generate new policies. In a sparse reward environment in fact, the probability of selecting a rewarding policy is proportional to the ratio between the rewarding and non-rewarding areas. The sparser the reward is, i.e. the smaller the reward area is, the lower the probability of selecting a rewarding policy from the archive and the slower the exploitation gets. A similar trend is exhibited by CMA-ME: even if able to reach high rewards on the discovered reward areas, it is slow in its optimization. At the same time, even NS reached high rewards on almost all environments, but without any explicit reward optimization it did not exploit the reward areas to the maximum. The multi-objective approach NSGA-II can always find at least one of the multiple reward areas, but then tends to extensively focus on it, instead of also exploring other areas. For this reason only the easiest reward area is exploited to high values in all environments, while the harder reward area is seldom exploited.

### 7 CONCLUSION AND FUTURE WORK

In this work we introduced SERENE, a method that efficiently deals with sparse reward environments by augmenting NS with emitters. Contrary to similar methods using emitters, SERENE keeps exploration and exploitation of the reward as two distinct processes. Exploration is carried out by taking advantage of NS to discover all the reachable reward areas. These areas are then exploited by using...
local instances of population-based optimization algorithms called emitters. By using a meta-scheduler, SERENE can automatically assign the evaluation budget to either exploration or exploitation. This is advantageous also in situations in which no reward is present: in the absence of reward to exploit, SERENE performs exactly like NS.

SERENE has been tested on four different sparse reward environments, reaching high performances on all of them. Notwithstanding these encouraging results, the method still suffers from the same limitations as other QD methods, and first and foremost from the prior hand-design of the behavior space $\mathcal{B}$. In the future we will work on addressing this limitation by learning a behavior descriptor that could foster exploration towards rewarding solutions.

At the same time, it has been highlighted by Cully [9] that many kind of emitters can be used to address different kind of problems. Evaluating and combining different types of emitters is also an exciting line of work to extend the current method.
REFERENCES


A FINAL ARCHIVE DISTRIBUTION

In figure 7 we show the distribution of the behaviors of the policies in the final archive. Each point represents different policy. In blue are the policies that do not get any reward, thus considered exploratory, while in orange are rewarding policies, considered exploitative. For SERENE the exploratory policies are the ones in the novelty archive $\mathcal{A}_N$, while the exploitative policies are the ones in the rewarding archive $\mathcal{A}_R$.

We can see that even if the coverage metric values for SERENE are lower with respect to ME, the search space is well covered. Moreover, the reward areas are densely explored.

B HYPERPARAMETERS

The values of the hyperparameters used during the experiments are listed here. For each experiment we used a budget of $Bud = 500000$ evaluations, with the chunk size set to $K_{Bud} = 1000$. The population size is $M = 100$, and for each policy we generate $m = 2$ offspring. As mutation parameter we used $\sigma = 0.5$, while the number of policies uniformly sampled to be added to the novelty archive is $N_Q = 5$. SERENE uses an emitter population size of $M_E = 6$, with a bootstrap phase for each emitter of $\lambda = 6$ generations. For CMA-ME we used the same parameters used by Fontaine et al. [17]: 15 emitters, each one with a population size of 37. In every experiment, the policies parameters are bounded in the $[-5,5]$ range.

C TERMINATION CRITERION

The termination condition used for our emitters is inspired by the stagnation criteria introduced in [21]. We track the history of the rewards obtained over the last $120 + 20 \times n/\lambda$ emitter’s generations. Where $n$ is the size of the parameter space $\Theta$ and $\lambda$ is the emitter’s population size. The emitter is terminated if either the maximum or the median of the last 20 rewards is not better than the maximum or the median of the first 20 rewards.
Figure 7: Distribution of the behavior descriptors of the archived policies. On each column are shown the results for an environment, while on each row is shown the distribution for each experiment. The archive plotted are from the runs achieving highest coverage. In blue are the policies with no reward, in orange the policies with a reward. For SERENE in blue are the policies in the novelty archive and in orange the policies in the reward archive.