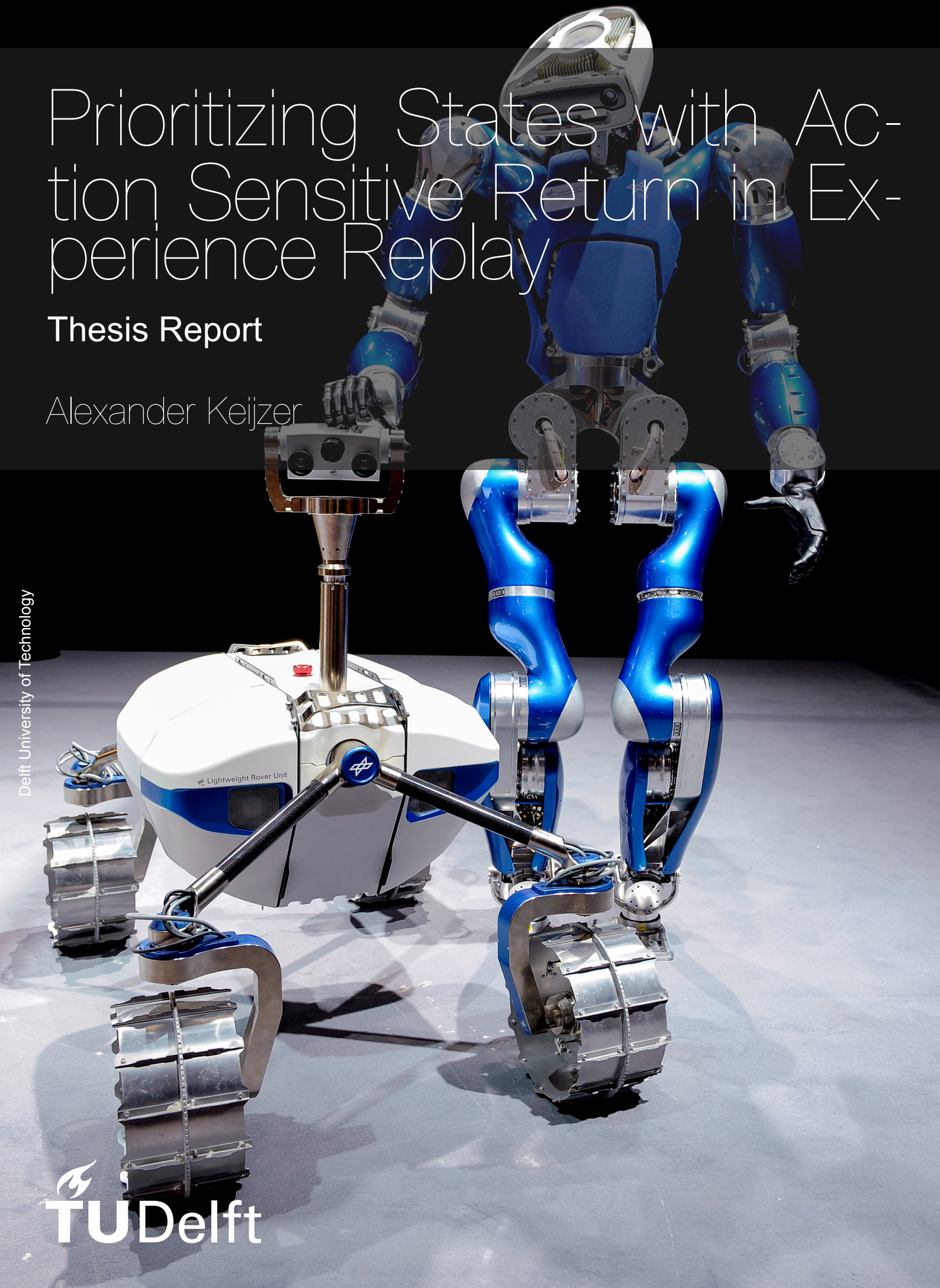
Prioritizing States with Action Sensitive Return in Experience Replay

Thesis Report

Alexander Keijzer



Delft University of Technology

TU Delft

Prioritizing States with Action Sensitive Return in Experience Replay

Thesis Report

by

Alexander Keijzer

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Friday June 16, 2023 at 16:00.

Student number:	4372697	
Period:	March 1, 2022 – June 16, 2023	
Thesis committee:	Dr.ing. J. Kober,	TU Delft, supervisor
	Prof.dr. R. Babuska,	TU Delft
	Dr. W. Böhmer,	TU Delft
	Ir. B. van der Heijden,	TU Delft

An electronic version of this thesis is available at http://repository.tudelft.nl/.

TU Delft

Contents

Prioritizing States with Action Sensitive Return in Experience Replay

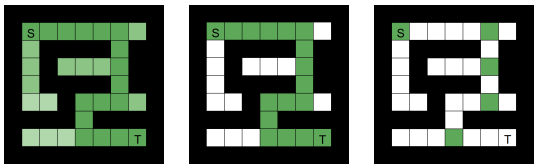
Alexander Keijzer
Cognitive Robotics
Delft University of Technology
Delft, The Netherlands

Abstract

Experience replay for off-policy reinforcement learning has been shown to improve sample efficiency and stabilize training. However, typical uniformly sampled replay includes many irrelevant samples for the agent to reach good performance. We introduce Action Sensitive Experience Replay (ASER), a method to prioritize samples in the replay buffer and selectively model parts of the state-space more accurately where choosing sub-optimal actions has a larger effect on the return. We experimentally show that this can make training more sample efficient and that this allows smaller function approximators – like neural networks with few neurons – to achieve good performance in environments where they would otherwise struggle.

1 Introduction

Reinforcement learning aims to find a policy that maximizes cumulative reward. This is often done through learning a compact representation of the value of states or state-action pairs in an environment by interacting with it. However, environment interactions can be costly and therefore necessitate efficient data use. Lin's introduction of experience replay in reinforcement learning [12] promotes efficient use and reuse of collected experience. A replay buffer stores transition samples from environment interactions, aiding sample efficiency and stabilizing training [13, 14]. Typically, learning of the value function in off-policy reinforcement learning is approached like a supervised learning problem where stochastic gradient descent is used to train function approximations such as neural networks. In supervised learning, data is usually sampled uniformly, however, for imbalanced datasets there are techniques to change the sampling distribution [6]. With reinforcement learning, we have a similar issue as not every sample is equally relevant for the performance of the agent. Lambert et al. [11] showed that in model-based reinforcement learning, a model that better explains the transition data (with higher likelihood) does not necessarily allow a better policy to be found on it. There is an objective mismatch when training. This is also the case for a value function



(a) Replay buffer data

(b) Only on-policy buffer data

(c) Only decision point buffer data

Figure 1: Toy maze environment where an agent moves from start S to termination T. The agent can choose which adjacent square to move to but can only move back to the square where it came from if there is no other valid action. The states sampled from a buffer of transitions are shown in green. After convergence, with normal experience replay the buffer will likely be filled with states from the whole maze. Previous work focused function approximator expressiveness on relevant states by prioritizing on-policy data. We introduce a method to prioritize "decision points" for the agent to further focus this expressiveness.

Preprint. Under review.

in off-policy reinforcement learning with function approximators. A lower value function error on the contents of a uniformly sampled buffer does not necessarily mean a better policy will result from it.

When using localized representations for the value network, such as in tabular reinforcement learning, replaying these irrelevant samples leads to computational overhead, but it does not impact the accuracy in the vital parts of the state-action space. However, many popular off-policy algorithms use global parametric function approximators like neural networks instead. Updates in irrelevant parts of the state-action space then yield a global effect, as these are often trained with stochastic gradient descent and minimize the mean square error across a batch of experiences [24]. Particularly in the case of smaller function approximators, such as neural networks with few neurons, the limited expressiveness that could have been allocated to model performance-critical areas is spent on inconsequential regions of the state-action space.

Off-policy algorithms may therefore, despite their ability to learn from experiences under different policies, still be influenced by the sampling distribution and buffer content when using global function approximators. To illustrate this, consider an extreme case: an agent would not see any performance improvement if it is trained using a replay buffer that contains states that the agent would never come across during rollouts. Other works have limited replay of states and/or actions unlikely under the current policy, the "off-policyness", which will reduce the replay of irrelevant samples and can improve performance [16, 22, 23]. However, just increased sampling of on-policy data does not prioritize what is really relevant: the importance of modeling these experiences accurately for the agent's performance during evaluation. A toy example that shows how we would ideally focus the approximators' expressiveness is shown in Fig. 1. In this example, an optimal policy could be learned with data from only 5 states while it is likely the replay buffer will contain data from all 30 states.

In this paper, we define a modeling importance criterion that measures sensitivity on return of taking a suboptimal action and introduce Action Sensitive Experience Replay (ASER), a method to change the replay distribution to match this criterion for off-policy reinforcement learning algorithms that use a state-action value function. This allows us to focus the expressiveness of function approximators on important parts of the state-space. We argue that while fitting the observation distribution (i.e. minimizing the total modeling error of the data collected by the agent along its rollouts) might seem logical, it is likely not the best strategy to achieve the maximum return efficiently. Since this leaves deprioritized parts of the state-space with lower accuracy, we use n-step returns to bootstrap only to states that are adequately modeled. We experimentally show that we find sample efficiency and/or final performance gains: (i) in a simple case, like the maze in Fig. 1 where importance is binary, this effective reduction of the state-space allows for significant improvements, (ii) when we transfer a learned importance criterion from a previously trained policy, (iii) in some cases, when learning the importance criterion during training.

2 Related Work

There have been several different proposals for non-uniform sampling or reweighting of replay buffers. Some, like CER [26], add extra samples to a uniform sampled batch while others change the sampling distribution altogether. Most techniques are aimed to satisfy one of the following four objectives.

Firstly, a common aim is to reduce the off-policyness of the selected samples for replay. This may improve learning speed and stability. Previous works estimate the off-policyness using importance weights [16] or multiple buffers [22] and then clip gradients or reweight experiences. In [23], off-policyness is instead reduced by sampling multiple batches and taking the most similar state distribution. Possible drawbacks of selecting mostly on-policy data are reduced robustness to policy or environment changes. Our approach does not prioritize on-policyness specifically, however, these approaches may work well together by further reducing the amount of modeled states.

Another objective may be to prioritize samples that are modeled poorly by the value function. Previous works [3, 21] propose prioritizing the replay of experiences based on their TD-error, which is analogous to how unexpected the transition is to the value function. This may be beneficial to integrating new experience faster and, like our approach, will prioritize states where the value function is volatile but can potentially focus the limited expressiveness on irrelevant parts of the state-space (reducing the final performance and convergence rate).

Heuristics can also be used to enforce good coverage of transition dynamics. For example, [19] introduce prioritized sampling based on Shannon's entropy of the state space vector. Other variants include selecting on reward [9] and state coverage [8]. Some of these works may do the opposite of our approach and the works described before, as, instead of focusing replay, they often enforce a broader coverage of the state space. In some cases, this may be preferred to increase robustness to environment changes, avoid catastrophic collapse, or for transfer learning.

Lastly, instead of using rules-based strategies to do prioritization, a learning-based approach to select experience from the buffer can be used. In [17, 25], training a replay policy to maximize the increase in cumulative reward by selecting transitions to train the agents' policy on is proposed. An interesting observation in these works is that these learning methods prefer replaying recent, likely on-policy, data. These methods, however, cannot learn a prioritization similar to our approach as the learning methods do not have the information needed as input.

Next to what the goal of prioritization is previous works differ in how they implement prioritization. The approach used by some works discussed here [16, 22] is reweighting errors to, for example, take larger update steps for more prioritized data. Instead, the sampling distribution can also be changed so more prioritized samples are seen more often [3, 19, 21]. With very large batch sizes and many update steps per new sample, these approaches will converge to effectively the same. Since this is often not the case we use the latter approach, however, reweighting may work too.

Most works discussed here prioritize samples from a first-in-first-out buffer. Instead, works such as [8] change in what order data is removed from the replay buffer as they often have a limited capacity. While this may also benefit our approach, we only change how a fixed-size buffer is sampled, not how data is removed.

Finally, our method makes use of n-step returns for off-policy reinforcement learning algorithms. A common way of accounting for the bias towards off-policy data introduced by n-step returns is by using importance sampling or tree backup [24]. We do not use these approaches as we argue our method naturally limits the impact of this bias (see Sect. 3.3).

3 Action Sensitive Experience Replay (ASER)

ASER aims to focus the expressiveness of function approximators used in off-policy reinforcement learning by changing the sampling distribution of the replay buffer. To do this, our method consists of three additions to standard algorithms. Firstly, a formal definition of the modeling importance criterion based on the Q-function of the algorithm is given in Sect. 3.1. Then, this criterion is used to change the sampling distribution as described in Sect. 3.2. Lastly, due to the changed sampling distribution, we need to skip over some states when bootstrapping as described in Sect. 3.3. In Sect. 3.4 we show the implementation of ASER for SAC [7].

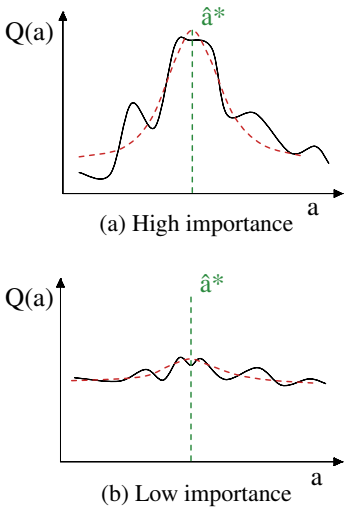


Figure 2: Visualization of Q-functions over actions where $\hat{a}^*$ is the estimate of the optimal action according to the policy. A good importance metric differentiates these, even though metrics like a second derivative may locally be similar. The dotted red line shows an approximation of this function, where a scaled entropy or variance could be a good metric, for example.

3.1 Modeling Importance Criterion

The modeling importance could be formalized in many ways, depending on reward distribution, action authority, etc. What is explored here is a measure to find areas of the state-action space where choosing a wrong action greatly affects the expected return for continuous state and action spaces. We define a function $p\colon \mathbb{R}^n \rightarrow \mathbb{R}^+$ that maps the state space $\mathbf{s}$ to the modeling importance. A possible modeling importance could be the norm of the second derivative of the Q-function to the action a at the optimal action. A highly negative concavity at the peak of the Q-function gives a local proxy for the rate of decrease in expected return when choosing a sub-optimal action. For

multi-dimensional actions, a matrix norm of the Hessian could be used instead. However, there are multiple issues with using this definition as the modeling importance. It is often not trivial to find the peak of the Q function in continuous domains. Taking the estimated optimal action in actor-critic methods instead, for example, may not exactly coincide with the peak of the Q-function and therefore give inaccurate values. Even if the peak of the Q-function could be found, the concavity is a local metric. The second-order derivatives may be large while the total value decrease is small.

Therefore, a less local metric for this value loss is needed. A solution to this is to approximate the Q-function over actions for every state with a tractable function where we could use metrics like entropy or variance. Fig. 2 shows examples of Q-functions where we would like a high difference in importance. The assumption here, however, is that the reinforcement learning algorithm used will model the reduction in expected return accurately. From experiments, we found that this is not generally the case. For example, we see that with ϵ-greedy exploration, DQN may find what the best action is but will likely not accurately model the value of selecting a sub-optimal action. On the other hand, maximum entropy reinforcement learning algorithms have a built-in incentive to find the sensitivity of their actions and perform more randomly in parts of the state space where the effect on the expected return is smaller.

In Sec. 3.4, we show that the policy of a SAC agent already suits these needs well. However, a similar tractable approximation and exploration incentive could be implemented for other algorithms.

3.2 Sampling Prioritization

The modeling importance criterion is used to select what priority samples should be selected for replay. Similarly to PER [21], the selection is stochastic depending on the criterion and we introduce a hyperparameter α that scales the prioritization of samples. The chance of selecting a sample i is:

$$P(i) = \frac{p_i^\alpha}{\sum_k p_k^\alpha}. \tag{1}$$

Sweeping the entire buffer to select data is too computationally expensive, therefore their prioritization will be updated with the current criterion where they are sampled. New experience samples are introduced into the buffer with the maximum prioritization that exists inside the buffer. An additional benefit is that this prioritizes new, on-policy data to integrate new experiences quickly.

Since PER's prioritization criterion is based on the TD-error, which depends on what next state the transition ends up in, a bias will be introduced that needs to be annealed during training. PER does this using importance sampling. The modeling importance criterion we introduce only depends on the current state, which is independent of any stochasticity in the environment, which means we do not need this correction. We change the distribution of states that is replayed compared to the distribution of states that is seen by the agent during rollouts, this will introduce a bias in modeling errors but not in value convergence.

3.3 Bootstrapping

Typically, off-policy reinforcement learning algorithms use TD(0) updates to learn the Q-function. Here, the Q-function is bootstrapped to the value of the next state-action pair. The sampling prioritization may result in inaccurate modeling of this next state and these inaccuracies will then be pulled into the value estimation of states that do need accurate modeling. This is mainly an issue when there is a significant difference in modeling importance like in the maze environment from Fig. 1.

To account for this, transitions can be unrolled further along the trajectory that they came from to areas where the Q-function is modeled accurately using n-step returns. Rewards along this trajectory are added up and discounted. The bootstrap target of the initially replayed transition to a transition n steps away is as follows:

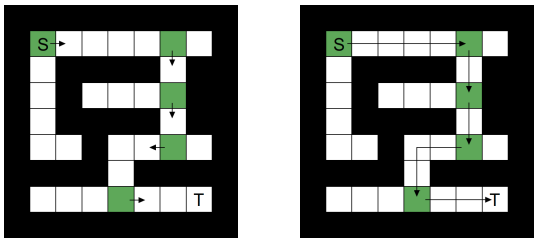
$$Q_{\text{targ}}(\mathbf{s}_0, \mathbf{a}_0) = r_0 + \gamma r_1 + \gamma^2 r_2 + ... + \gamma^{n-1} r_{n-1} + \gamma^n Q(\mathbf{s}_n, \pi(\mathbf{s}_n)) \tag{2}$$

where the subscripts indicate the steps away from the initial transition along the trajectory of the transition. A visualization of this bootstrapping can be found in Fig. 3.

The modeling importance criterion is used to decide whether the Q-function models a state well enough. We determine the criterion for an unprioritized batch and take a percentile of this batch as a threshold value b for bootstrapping. Since at the start of training the accuracy of the Q-values is low we increase this value during training. The k-th percentile target is determined as follows:

$$k = \min(k_m, tk_s) \tag{3}$$

where k_m is the maximum target percentile, t is the amount of training timesteps and k_s is the target percentile slope. This results in a simple bounded linearly increasing threshold. If the criterion is lower than this threshold, the Q-function is assumed to be inaccurate and the next transition in the trajectory is taken until a transition is equal to or higher than the threshold is found. A maximum bootstrap length h is used to bound the bootstrap distance when the criterion is low for many transitions in a trajectory.



(a) Standard bootstrapping

(b) Importance n-step bootstrapping

Figure 3: Bootstrapping visualized with arrows in the toy maze environment. White states, which are not modeled accurately due to the prioritized sampling, are skipped when boostrapping to avoid pulling this badly modeled data into the well-modeled green states.

It is important to note this n-step bootstrapping will cause the bootstrap target to be dependent on the policy that collected it. A different policy may have taken a different trajectory by taking different actions and have collected different rewards. This will result in a bias towards older policies that are still in the buffer with off-policy reinforcement learning algorithms. There are ways to solve this, for example, with importance sampling or tree backup [24]. We argue, however, that this effect is limited since the policy dependency is only in areas where actions have a limited effect on the expected return. As soon as a state is encountered where a different policy would have a significant effect, it is used to bootstrap instead. A different policy would therefore only have a limited effect on the value of the bootstrap target.

3.4 Implementation

As explained in Sect. 3.1, we would like a tractable function to approximate the Q-function in order to find a more global metric for sensitivity. In SAC, the policy is modeled by a Gaussian distribution where the mean and standard deviation depend on the state. This mean and standard deviation are learned to minimize the KL divergence with a normalized exponential of the Q-function. This results in a global approximation of the sensitivity of the Q-function described by a Gaussian. Using the log probability of the mode of this Gaussian works well to estimate the total sensitivity as it gives a combined metric when actions are multi-dimensional, but other metrics, like the determinant of the covariance matrix, may also work. A high probability of the mode will mean that the distribution is narrow, resulting in a fast reduction in value next to the optimal action.

This results in the following modeling importance:

$$p(\mathbf{s}) = \log \pi(\mathbf{a}|\mathbf{s}), \quad \mathbf{a} = \bar{\pi}(\mathbf{s}) \tag{4}$$

where $\bar{\pi}$ is the "greedy" version of the policy, in this case, the mode of the Gaussian. Since this can result in negative log probabilities when $\pi(\mathbf{a}|\mathbf{s}) < 1$ which cannot be used in Equation (1), we move the underlying probabilities by adding 1 to it. This will have a squashing effect on the prioritization. We, therefore, implement our method for SAC to evaluate its effectiveness. One thing to note is that in SAC the Q-value target includes an entropy term. When using n-step returns these entropy terms will need to be added in a similar way to the rewards in Equation (2) to get consistent results.

The complete algorithm for ASER is described in Algorithm 1. ASER can be used as a replacement for uniform sampling used in other reinforcement learning algorithms if a good importance criterion can be found. We use a replay buffer of transition tuples state $\mathbf{s}$, actions $\mathbf{a}$, reward r, next state $\mathbf{s}'$ and end of episode (or done flag) d with prioritization p. For every gradient step, ASER will modify the batch of transitions before they are used in the normal gradient step of the algorithm. The only change that needs to be done to the algorithm it is implemented on is that any bootstrapping will need to be discounted according to the bootstrap length used by ASER, so with γ^n instead of simply γ. Other than updating the prioritization the buffer contents are not modified.

Algorithm 1: ASER

Input: Replay buffer $\mathcal{D}$ of samples $T = (\mathbf{s}, \mathbf{a}, r, \mathbf{s}', d)$ with prioritization p, training timesteps t
Parameters: Maximum horizon h, maximum target percentile k_m, target percentile slope k_s, discount factor γ

for *each gradient step* **do**
  Sample with $P(i) = p_i^\alpha / \sum_j p_j^\alpha$ a batch of transitions $B = \{(\mathbf{s}, \mathbf{a}, r, \mathbf{s}', d)_i\}$ from $\mathcal{D}$;
  $b \leftarrow$ top k-th percentile of $p(\mathbf{s}')$ for every $T \in B$, where $k = \min(k_m, tk_s)$;
  for *each* $T = (\mathbf{s}, \mathbf{a}, r, \mathbf{s}', d)_i \in B$ **do**
    $n \leftarrow 1$;
    while $p(\mathbf{s}') < b$ ***and*** *not* d ***and*** $n < h$ **do**
      $(\hat{r}, \hat{\mathbf{s}}', \hat{d}) \leftarrow$ from next transition in trajectory from buffer $\mathcal{D}$;
      Modify T with: $r \leftarrow r + \gamma^n \hat{r}$; $\mathbf{s}' \leftarrow \hat{\mathbf{s}}'$; $d \leftarrow \hat{d}$;
      $n \leftarrow n + 1$;
    end
  end
  Normal gradient step with batch B of modified transitions T with bootstrap discount γ^n;
  Update in $\mathcal{D}$ prioritization p for every $T \in B$ with $p(\mathbf{s})$;
end

4 Results

In the following section, we introduce the setup of our experiments and the maze environment in Sect. 4.1 and then show the effect of ASER when it has access to prior information in Sect. 4.2. In Sect. 4.3 we show that ASER also works without prior information. Results that show the dependency on reward function or environment definition and an ablation study of the sampling prioritization and n-step bootstrapping are given in Sect. 4.4.

4.1 Maze environment and experiment setup

We first show the effect of ASER in a four times larger version of the maze shown in Fig. 1 and 3. In this environment, the agent can choose which adjacent square to move to but can only move back to the square where it came from if there is no other valid action. Therefore, in many states the chosen action will have no effect, as every action leads to the same next state but in some choosing the wrong action will result in the agent reaching a dead end. This means that the neural network only needs to model these decision-sensitive areas and can ignore the other states. We initially show the maze with a reward function where a penalty is applied every time the agent reaches a dead end. Since the reward function has a significant effect on ASER, in Sect. 4.4 we also show the effect when the reward is given when reaching the end of the maze linearly decreasing with the timesteps spent to get to the end.

We build upon the algorithm implementations of Stable Baselines 3 [18]. To create a fair comparison we run hyperparameter optimization for every algorithm on the shown environments. For this, we use Optuna [2] and maximize the sum of the average total reward of multiple rollouts in the evaluation environment during training to optimize for speed of convergence and final performance. Tables with all hyperparameters, both optimized values and SB3 defaults, can be found in Appendix A. We present all the sample efficiency graphs according to the recommendations of Agarwal et al. [1], using the interquartile mean with a 95% stratified bootstrap confidence interval from 16 training runs.

4.2 ASER with prior information

To show the effectiveness of only learning action-sensitive parts of the state space, we first compare the effect of our method with prior information about the sensitivity of actions. To do this, we create an "oracle" that has perfect information about the sensitivity of actions. This oracle is used instead of the modeling importance metric $p(\mathbf{s})$ from Equation (4), where a decision point has an importance of 1 and all other states have an importance of 0. Another way to use prior information is to use the importance criterion from a previously trained policy instead of learning it while training.

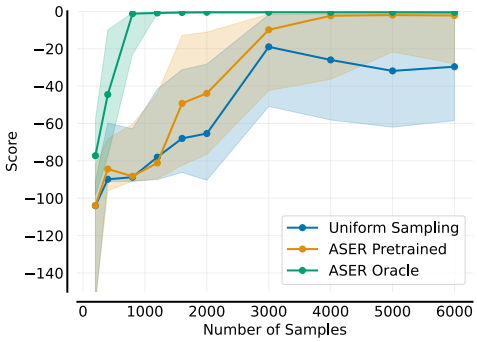


Figure 4: Sample efficiency comparison between SAC with uniform sampling, ASER with a pretrained importance criterion, and ASER with an importance criterion from an "oracle" on the Maze environment with a small, 24-neuron actor and critic network.

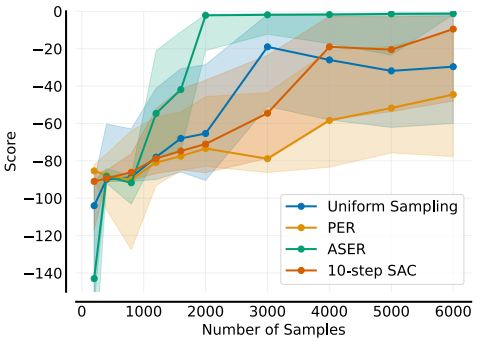


Figure 5: Sample efficiency comparison between SAC with standard replay, PER, ASER with an online learned importance criterion, and 10-step SAC on the Maze environment with a small, 24-neuron actor and critic network.

In Fig. 4 we show the sample efficiency curves for SAC with standard experience replay, ASER with the selection criterion based on a pretrained policy, and ASER with the oracle. All of these agents use a small, 24-neuron policy and critic network. SAC with standard replay struggles to reliably converge in the Maze environment with these small networks, even after hyperparameter optimization. It appears that at some point additional training samples no longer improve the agent's performance. A possible reason for this is that too much of the limited expressiveness of the neural network is used to model parts of the state-space with no impact on performance, as actions do not affect return there. Filling the replay buffer with more data will not be beneficial, as the neural network needs to filter out portions of these data to effectively incorporate it. ASER with a pretrained importance criterion helps with this. We see that the agent's performance keeps rising with more samples and we achieve a significant performance improvement over uniform sampling.

In this environment, however, the importance criterion is binary. Actions either have no effect at all or will lead to a negative reward. The pretrained mode probability that sets the importance criterion did not fit this perfectly. Therefore, if we use the oracle we see a large increase in sample efficiency. Both the spread of results between training runs and the average number of samples needed to converge is smaller. At 1.2k samples all of the runs have converged to the best policy.

4.3 ASER with learned importance

When the modeling importance criterion is not given and learned while training, we still see an improvement in performance against standard replay in Fig. 5. ASER consistently converges to the best policy. PER struggles in this environment, with performance slightly below standard replay. In this figure, we also show SAC with just n-step returns to rule out that this could be the only contributing factor in ASERs performance gain. 10-step returns were chosen using parameter optimization as explained in Sect. 4.1. Somewhat surprisingly 10-step SAC performs comparable, if not slightly better, than standard SAC even though this will introduce bias in the value estimation to older policies in the buffer without importance sampling or tree backup. Fig. 9 shows that when training a larger network of two layers of 256 neurons there is still a significant sample efficiency, stability, and final performance gain for ASER compared to standard experience replay and PER, but the effect is smaller than with a small network.

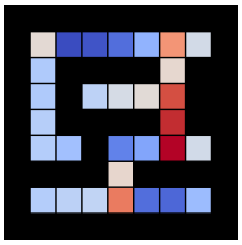


Figure 6: Sampling count relative to uniform with ASER in the Maze environment where blue is less sampling and red is more sampling than uniform.

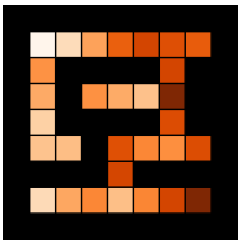


Figure 7: Bootstrap distance with ASER for the Maze environment where darker is a shorter distance. E.g. in the top row, bootstrapping distance decreases with the distance to the decision point.

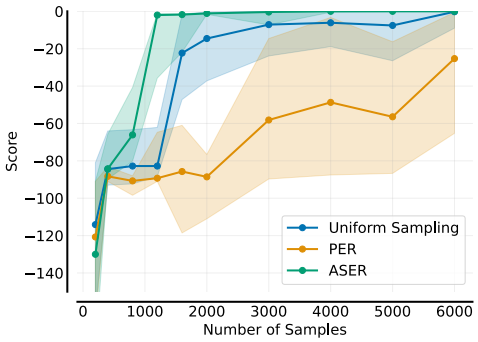


Figure 9: Sample efficiency comparison between SAC with uniform sampling, PER, and ASER with an online learned importance criterion on the Maze environment with a two-layer 256-neuron actor and critic network.

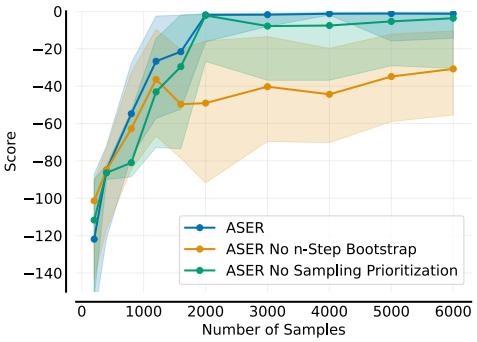
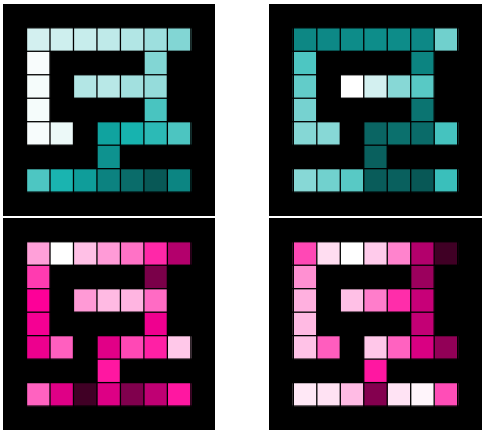


Figure 10: Sample efficiency comparison between ablations of SAC ASER on the Maze environment.

ASER without pretraining actually seems to perform better than ASER with pretraining. There could be multiple explanations for this. Firstly, the newly trained policy and the pretrained policy may not align and take different trajectories through the state-space. This is possible in this environment as the agent state includes the last movement direction, therefore approaching a decision point from a different direction is a different state. However, the options are limited, so this possible misalignment seems unlikely to have a significant impact. Secondly, the pretrained policy may not capture the sensitivity well. The pretrained policy does indeed not exactly track the oracle and still changes significantly after the policy converges to the best actions. However, the same will be true for an online learned sensitivity. Lastly, online learning of prioritization may help in finding policy improvement. Prioritization may temporarily be higher than with a converged policy in areas of the state-space where the agent is struggling and therefore receiving high variance in its returns. This may aid in faster learning.

In Fig. 6 and 7 we show, respectively, the sampling prioritization and the variable step bootstrapping of ASER with learned importance. In the best case, these should reflect Fig. 1c and 3b. We see very similar behavior in the sampling, where decision points have been learned and are prioritized, although with some effect on nearby states. The bootstrapping figure is less clear but gives a similar picture, especially in the bottom and top rows it is clear that it is bootstrapping to decision points and termination as the distance lowers the closer we get to these points.



(a) Sparse reward (b) Denser reward

Figure 8: Comparison of the value function (top) and log probability of the mode (bottom) between a sparse discounted reward at termination and a denser reward with a penalty each time the agent encounters a dead end. In the denser reward scenario figures (right), there is more immediate feedback resulting in a clearer difference in next state value at decision points and therefore a more accurate importance metric.

4.4 Ablation and limitations

Sample efficiency curves for ablation of both prioritization and n-step bootstrapping with a predetermined importance criterion are shown in Fig. 10. We see that the most significant efficiency improvement comes from variable step bootstrapping, with a smaller improvement from the prioritization. Both features are needed to properly "compress" the environment and achieve stable convergence.

In order for ASER to learn a good importance, there must be a significant change in value between actions. This can depend heavily on how the reward function is defined. In Fig. 8 we show two of the same mazes, where one receives only a reward at the end and one receives a penalty each time it

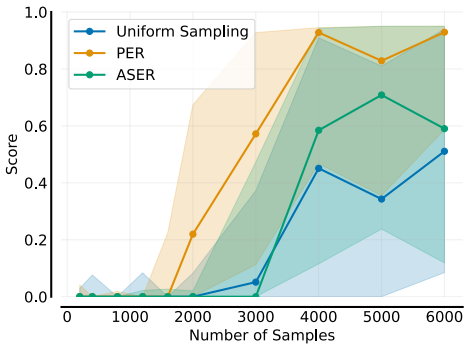


Figure 11: Sample efficiency comparison between SAC with uniform sampling, PER, and ASER with an online learned importance criterion on the Maze environment with a sparse reward function.

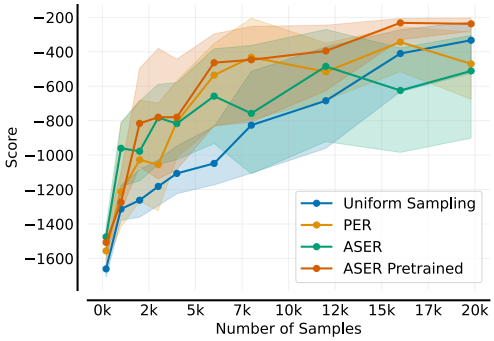


Figure 12: Sample efficiency comparison between SAC with uniform sampling, PER, ASER with an online learned importance criterion, and ASER with a pretrained importance criterion on the Pendulum environment with a small, 24-neuron actor and critic network.

needs to turn around. In the former, ASER does not improve performance as shown in Fig. 11 due to the smaller difference in value, resulting in less clear differences in the importance criterion.

In the Pendulum environment from OpenAI Gym [4], performance differences are smaller, however, increased sample efficiency can still be achieved with pretraining, as shown in Fig. 12. Online learned ASER is unable to achieve performance gains, possibly due to being unable to stably learn the importance criterion.

5 Conclusion

In this paper, we have presented Action Sensitive Experience Replay (ASER) which changes the experience replay distribution to prioritize states where the return is especially sensitive to non-optimal actions. This strategic reallocation of modeling resources enables parametric function approximators – such as neural networks – to focus their limited expressiveness on regions of the state-space that are most relevant for the agent's performance.

Our findings show improvements in sample efficiency, stability, and final performance, particularly when the action's sensitivity is either known in advance. When learning the sensitivity during training or carrying the sensitivity over from a previously trained policy we still observed some improvements. The improvements are more pronounced when the number of parameters in the function approximator is constrained.

By making reinforcement learning agents model just the important parts of the state-space it could be possible for them to learn good policies in larger state-spaces without increasing the size of the function approximators used in the learning algorithm. Allowing smaller neural networks to achieve good performance may help with multi-task reinforcement learning. For example, in policy distillation [20] the policy of a large network is extracted to train a smaller, more efficient network, a step that may be omitted with a method similar to ours. Another possible application of our method is in transfer learning and specifically sim-to-real transfer. It is possible that transferring a pretrained importance criterion is more robust to environmental changes than the learned policy which can be prone to overfitting to the simulator dynamics [5, 10, 15].

In order to achieve performance improvements, our method requires environments to have rewards shaped in a way that results in clear changes in the value in some states. We realize that the evaluation of our method is limited to a small number of environments in this paper. A larger study may give a better overview of the implications and possible limitations of our method. Usage in larger and more complex state- and action-spaces has the potential to show an even larger performance improvement to existing methods but may also run into transfer or stability issues.

References

[1] Rishabh Agarwal et al. “Deep Reinforcement Learning at the Edge of the Statistical Precipice”. In: *Advances in Neural Information Processing Systems* (2021).

[2] Takuya Akiba et al. “Optuna: A Next-generation Hyperparameter Optimization Framework”. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2019.

[3] Marc Brittain et al. “Prioritized Sequence Experience Replay”. In: *CoRR* abs/1905.12726 (2019).

[4] Greg Brockman et al. *OpenAI Gym*. 2016. eprint: `arXiv:1606.01540`.

[5] Rodney A Brooks. “Artificial life and real robots”. In: *Proceedings of the First European Conference on artificial life*. 1992, pp. 3–10.

[6] Mikel Galar et al. “A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches”. In: *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42.4 (2012), pp. 463–484.

[7] Tuomas Haarnoja et al. “Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor”. In: *International conference on machine learning*. PMLR. 2018, pp. 1861–1870.

[8] David Isele and Akansel Cosgun. “Selective experience replay for lifelong learning”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. 1. 2018.

[9] Max Jaderberg et al. “Reinforcement Learning with Unsupervised Auxiliary Tasks”. In: *International Conference on Learning Representations*. 2017.

[10] Nick Jakobi, Phil Husbands, and Inman Harvey. “Noise and the reality gap: The use of simulation in evolutionary robotics”. In: *Advances in Artificial Life: Third European Conference on Artificial Life Granada, Spain, June 4–6, 1995 Proceedings 3*. Springer. 1995, pp. 704–720.

[11] Nathan Lambert et al. “Objective Mismatch in Model-based Reinforcement Learning”. In: *Learning for Dynamics and Control*. PMLR. 2020, pp. 761–770.

[12] Long-Ji Lin. “Self-Improving Reactive Agents Based On Reinforcement Learning, Planning and Teaching”. In: *Machine learning* 8 (1992), pp. 293–321.

[13] Volodymyr Mnih et al. “Human-level control through deep reinforcement learning”. In: *Nature* 518 (2015).

[14] Volodymyr Mnih et al. “Playing Atari with Deep Reinforcement Learning”. In: *arXiv preprint arXiv:1312.5602* (2013).

[15] Stefano Nolfi and Dario Floreano. *Evolutionary robotics: The biology, intelligence, and technology of self-organizing machines*. MIT press, 2000.

[16] Guido Novati and Petros Koumoutsakos. “Remember and forget for experience replay”. In: *International Conference on Machine Learning*. PMLR. 2019, pp. 4851–4860.

[17] Youngmin Oh et al. “Learning to sample with local and global contexts in experience replay buffer”. In: *arXiv preprint arXiv:2007.07358* (2020).

[18] Antonin Raffin et al. “Stable-Baselines3: Reliable Reinforcement Learning Implementations”. In: *Journal of Machine Learning Research* 22.268 (2021), pp. 1–8.

[19] Mirza Ramicic and Andrea Bonarini. “Entropy-based prioritized sampling in deep Q-learning”. In: *2017 2nd international conference on image, vision and computing (ICIVC)*. IEEE. 2017, pp. 1068–1072.

[20] Andrei A Rusu et al. “Policy distillation”. In: *arXiv preprint arXiv:1511.06295* (2015).

[21] Tom Schaul et al. “Prioritized Experience Replay”. In: *International Conference on Learning Representations (ICLR)*. 2016.

[22] Samarth Sinha et al. “Experience replay with likelihood-free importance weights”. In: *Learning for Dynamics and Control Conference*. PMLR. 2022, pp. 110–123.

[23] Peiquan Sun, Wengang Zhou, and Houqiang Li. “Attentive experience replay”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 04. 2020, pp. 5900–5907.

[24] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[25] Daochen Zha et al. “Experience replay optimization”. In: *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. 2019, pp. 4243–4249.

[26] Shangtong Zhang and Richard S Sutton. “A Deeper Look at Experience Replay”. In: *arXiv preprint arXiv:1712.01275* (2017).

A Hyperparameters

Table 1: Optimized Hyperparameters Values Maze

		Standard	PER	ASER	Pretrained ASER	Oracle ASER
λ	Learning Rate	$2.480\text{e}{-2}$	$5.192\text{e}{-2}$	$2.356\text{e}{-2}$	$2.890\text{e}{-2}$	$3.215\text{e}{-2}$
γ	Discount Factor	$9.770\text{e}{-1}$	$9.549\text{e}{-1}$	$9.673\text{e}{-1}$	$9.851\text{e}{-1}$	$9.5267\text{e}{-1}$
τ	Target Smoothing Coefficient	$1.914\text{e}{-3}$	$9.203\text{e}{-2}$	$2.025\text{e}{-2}$	$9.270\text{e}{-2}$	$7.399\text{e}{-2}$
α	Sample Prioritization	-	$3.414\text{e}{-1}$	$5.541\text{e}{-1}$	$6.103\text{e}{-1}$	1.032
h	Max Bootstrap Length	-	-	7	13	8
k_m	Bootstrap Target Percentile Maximum	-	-	85.31%	57.97%	85.06%
k_s	Bootstrap Target Percentile Slope	-	-	$1.383\text{e}{-3}$	$3.250\text{e}{-2}$	$4.457\text{e}{-1}$

Table 2: Optimized Hyperparameters Values Pendulum

		Standard	PER	ASER [1]	Pretrained ASER
λ	Learning Rate	$3.281\text{e}{-2}$	$1.200\text{e}{-2}$	$6.402\text{e}{-2}$	$6.402\text{e}{-2}$
γ	Discount Factor	$9.357\text{e}{-1}$	$9.447\text{e}{-1}$	$9.508\text{e}{-1}$	$9.508\text{e}{-1}$
τ	Target Smoothing Coefficient	$7.518\text{e}{-2}$	$2.538\text{e}{-3}$	$3.880\text{e}{-2}$	$3.880\text{e}{-2}$
α	Sample Prioritization	-	$2.538\text{e}{-1}$	$6.425\text{e}{-1}$	$6.425\text{e}{-1}$
h	Max Bootstrap Length	-	-	3	3
k_m	Bootstrap Target Percentile Maximum	-	-	73.34%	73.34%
k_s	Bootstrap Target Percentile Slope	-	-	$2.495\text{e}{-3}$	$2.495\text{e}{-3}$

Table 3: Stable Baselines 3 Defaults

Replay Buffer Size	1e6
Env. Steps Before Learning Starts	100
Batch Size	256
Gradient Steps Per Env. Step	1
Entropy Coefficient	Learned
Target Entropy	-1

[1] Parameter optimization for ASER without pretaining converged to standard sampling, using pretrained values instead to show the effect of ASER.

Additional Appendices

B. Corridor-1D Environment and Results

Initial experiments were done on a simpler 1D environment. This has similar properties to the maze with only select states where actions matter but the environment would terminate immediately at a wrong action. Results are very similar to the Maze environment.

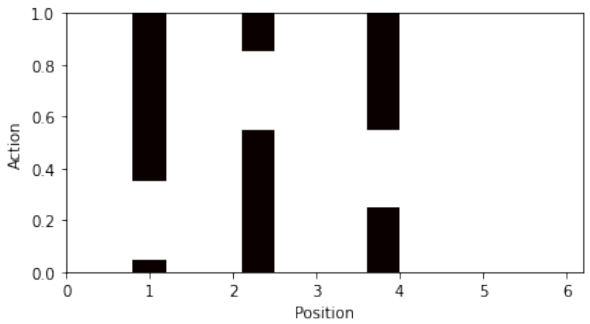


Figure B.1: Corridor-1D environment, the agent position increases with 0.1 each timestep but will terminate and get a negative reward if it performs an action within the black bars.

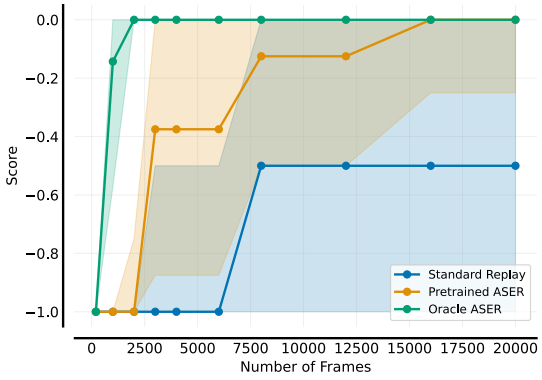


Figure B.2: Sample efficiency comparison between SAC with standard replay, SER with a pretrained importance criterion, and ASER with an importance criterion from an “oracle” on the 1D-Corridor environment with a small, 24-neuron actor and critic network.

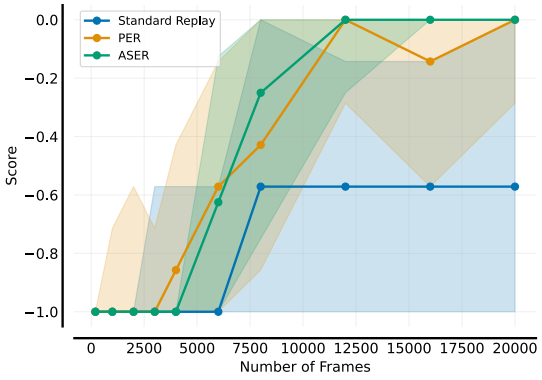


Figure B.3: Sample efficiency comparison between SAC with standard replay, PER, and ASER with an online learned importance criterion on the 1D-Corridor environment with a small, 24-neuron actor and critic network.

C. DQN Implementation Issues

ASER was also implemented for DQN. However, this turned out not to work well. As explained in the paper, ϵ-greedy exploration seems to not be enough to learn dropoff in value well enough for ASER to work. A short experiment was done in the Corridor-1D environment from Appendix B where for every 0.1 position step the Q-values for each action were sorted and plotted. First a set of Q-values from states where the action does not matter as the agent in a full white state:

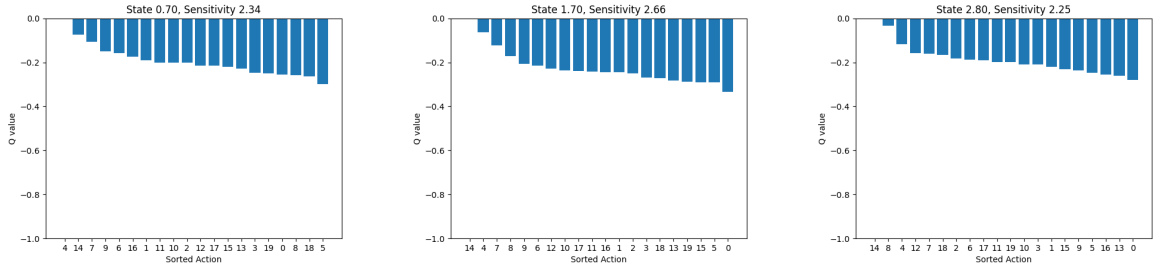


Figure C.1: Q-functions value loss compared to the best action for states where actions do not matter.

Now a set of Q-values where actions do matter as the agent is at a barrier:

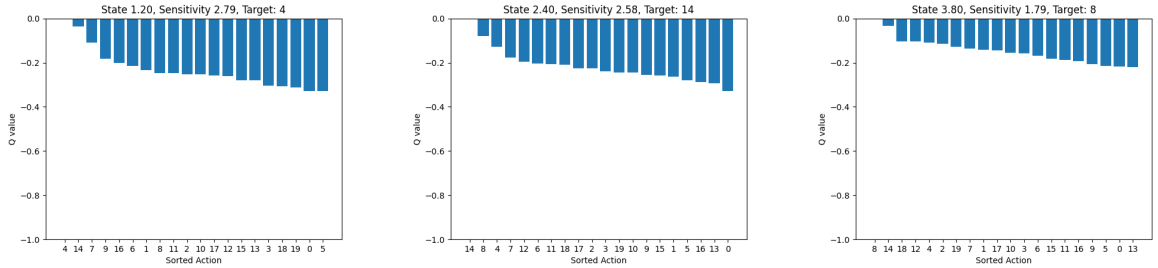


Figure C.2: Q-functions value loss compared to the best action for states where actions do matter.

There is no clear distinction to make between these two sets of figures. The agent does learn what the best action is but the values of the other states are inaccurate. They should all be 0 for the top set (as every state has the same results) and all be -1 except the few around the target action in the bottom set. An exploration incentive, therefore, seems to be needed.

D. Second Derivative and Mode Log Probability

While the second derivative to actions of the Q-function generally seems too unstable to function as an effective modeling importance criterion, especially when it is learned while training, it is interesting to see how it compares to the log probability of the mode of the policy in SAC. In the OpenAI Gym Pendulum environment we expect clear lines where a wrong action would cause the pendulum to irrecoverably overshoot or undershoot. This is indeed what we see in Fig. D.1 and D.2. The log probability shown shows a similar profile but softer and more patchy.

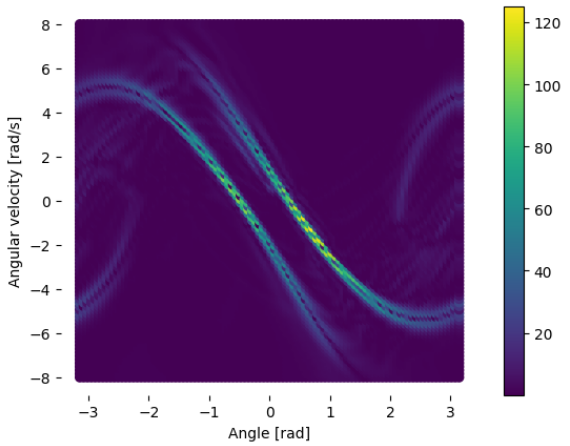


Figure D.1: Second derivative of the Q-function to the action at the optimal action in the OpenAI Gym Pendulum environment.

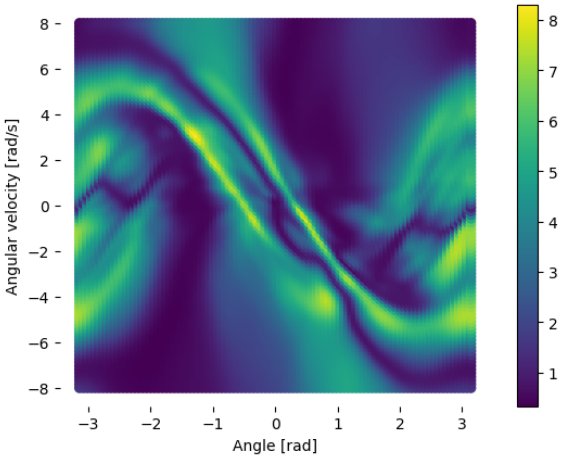


Figure D.2: Log probability of the mode of the policy distribution of SAC in the OpenAI Gym Pendulum environment.

From these plots, it seems like using the second derivative would work, possibly even better. This may be true as we haven't specifically tested this in Pendulum. For other environments, however, it is clear that the second derivative does not give a good importance criterion. In Fig. D.3 and D.4, for example, we show the second derivative and mode probability for the Maze environment where the second derivative does not appear to show anything but noise. We continued our research with the mode probability due to its more robust nature. However, in some environments, the second derivative may indeed perform better.

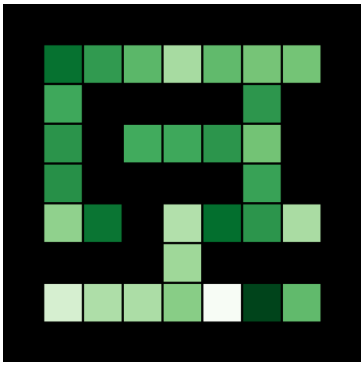


Figure D.3: Second derivative of the Q-function to the action at the optimal action in the Maze environment.

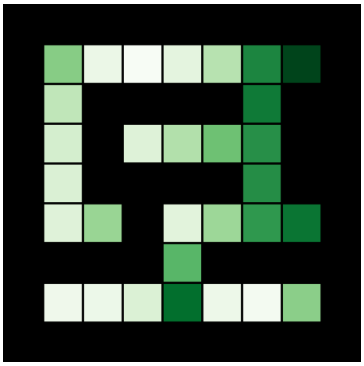


Figure D.4: Log probability of the mode of the policy distribution of SAC in the Maze environment.

E. Mode Probability and Log Probability

SAC aims to (among other objectives) maximize the entropy of the policy. The expectation of the negative log probability of random samples drawn from its policy distribution during training will be an estimate of the entropy. While we would like to use the entropy too, doing a similar estimation by drawing random samples will necessitate updating the buffer sample prioritization for all samples from that state. This is too computationally expensive for us to do. Instead, we use the log probability of just the mode, as the distribution is a Guassian and we expect this value to move with the entropy (but in a complex way).

As shown in the paper, using this log probability works as a metric. However, aside from this experimental evidence, the reasoning behind using it is weak, especially since prioritization does not work with negative values. We could, for example, also use the mode probability directly instead. Experiments with the mode probability resulted in generally weaker performance than the log probability (as shown in the paper), for example:

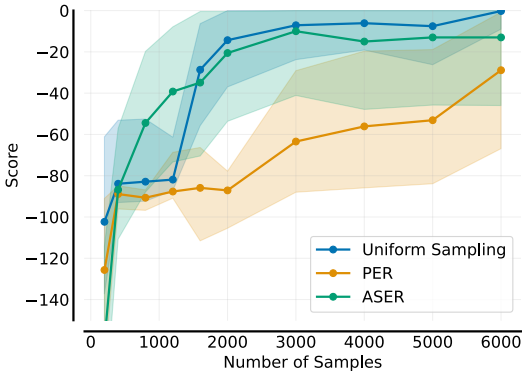


Figure E.1: Sample efficiency comparison between SAC with uniform sampling, PER, and the mode probability version ASER with an online learned importance criterion on the Maze environment with a two-layer 256-neuron actor and critic network (as paper Fig. 9).

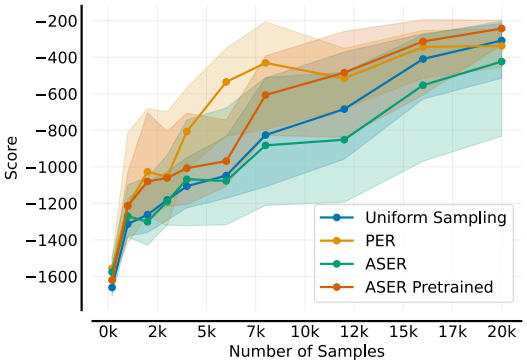


Figure E.2: Sample efficiency comparison between SAC with uniform sampling, PER, the mode probability version ASER with an online learned importance criterion, and the mode probability version ASER with a pretrained importance criterion on the Pendulum environment with a small, 24-neuron actor and critic network (as paper Fig. 12).

In future work, a developing better theoretical framework and understanding of how this criterion affects performance would be beneficial.