SACn: Soft Actor-Critic with n-step Returns Supplementary Materials

Jakub Łyskawa[1][a], Jakub Lewandowski[2], Paweł Wawrzyński[3][b]
[1]*Warsaw University of Technology*
[2]*Warsaw University*
[3]*IDEAS Research Institute*
jakub.lyskawa@pw.edu.pl

Keywords:

Abstract:

1 DENSITY MEASUREMENT

Figures 1 present which part of the probability density ratios are equal to or greater than given threshold values, measured for batches used for training the base SAC agent during the 1000-step periods at the time steps $\{10001, \ldots, 11000\}$, $\{19001, \ldots, 20000\}$, $\{49001, \ldots, 50000\}$, $\{99001, \ldots, 100000\}$, $\{199001, \ldots, 200000\}$, $\{499001, \ldots, 500000\}$, $\{999001, \ldots, 1000000\}$, to capture the distributions of the probability ratios during different parts of the training. The aggregated values were averaged over 5 runs.

The measured action probability density ratios allow to formulate the following observations:

1. Significantly below half of the samples in the buffer have a probability density ratio of 1 or greater, except during the very first steps of the training. It means that for the most of the samples in the buffer, the probability of the selected action decreased. This is likely caused by most of the explored actions resulting in suboptimal returns and the policy being fitted to match only the small of best actions.
2. During the whole training process samples with large importance sampling values keep occurring, and some of them, due to the precision of number representation, result in infinite probability density values. As the policy is likely fitted to replicate a small part of the actions selected by the exploration process, this underlines the need to properly handle such samples.

[a] https://orcid.org/0000-0003-0576-6235
[b] https://orcid.org/0000-0002-1154-0470

2 ABLATION RESULTS

Figures 2, 3, and 4 contain the learning curves for the experiments whose aim is to measure the influence of specific hyperparameters and components of SAC*n* algorithm, presented in Subsection 5.4. Each result is averaged over last $3 \cdot 10^4$ time steps to account for large amplitudes of the learning curves for some environments and over 12 runs, and the results are presented together with their standard errors. Specifically, figure 2 contains the learning curves for different values of the q_b and n hyperparameters. Figure 3 contains the learning curves for SAC*n* without *tau*-sampled entropy estimation. Figure 4 contains the learning curves for SAC with *tau*-sampled entropy estimation for different values of τ.

3 HYPERPARAMETER SETTINGS

In this section, we provide values of all the used hyperparameters for the experiments reported in Section 5. For SAC, we used the values provided by Raffin (2020), listed in the Table 1. For SAC*n*, we used the same values as for SAC where applicable. For SAC*n*, the default value of q_b was 0.75 and the values of n were provided for each result.

REFERENCES

Raffin, A. (2020). Rl baselines3 zoo. https://github.com/DLR-RM/rl-baselines3-zoo.

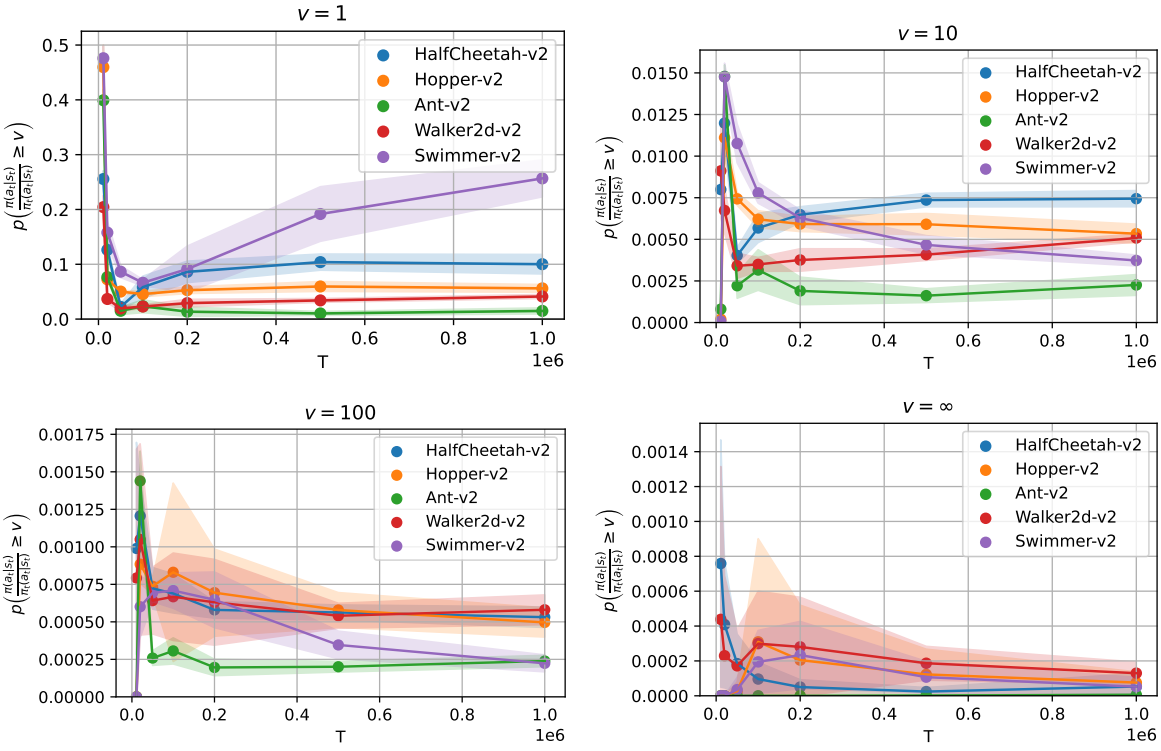

Figure 1: Part of samples that result in action probability densities equal to or exceeding given threshold value. Infinity values are the result of numerical precision of 32-bit floating point numbers typically used in machine learning.

Table 1: SAC and SACn hyperparameter values as provided for SAC by Raffin (2020). S - denotes the state space.

Hyperparameter	Value
Time steps	10^6
Batch size	256
Discount (Swimmer)	0.999
Discount (other)	0.99
Learning rate	$3 \cdot 10^{-4}$
Critic hidden layers	$\langle 256, 256 \rangle$
Actor hidden layers	$\langle 256, 256 \rangle$
Critic activation function	ReLU
Actor activation function	ReLU
Learning start	10^4
Training frequency	1
Num. of training steps	1
Target network update freq.	1
Target network update coeff.	0.005
Entropy target	$-dim(S)$

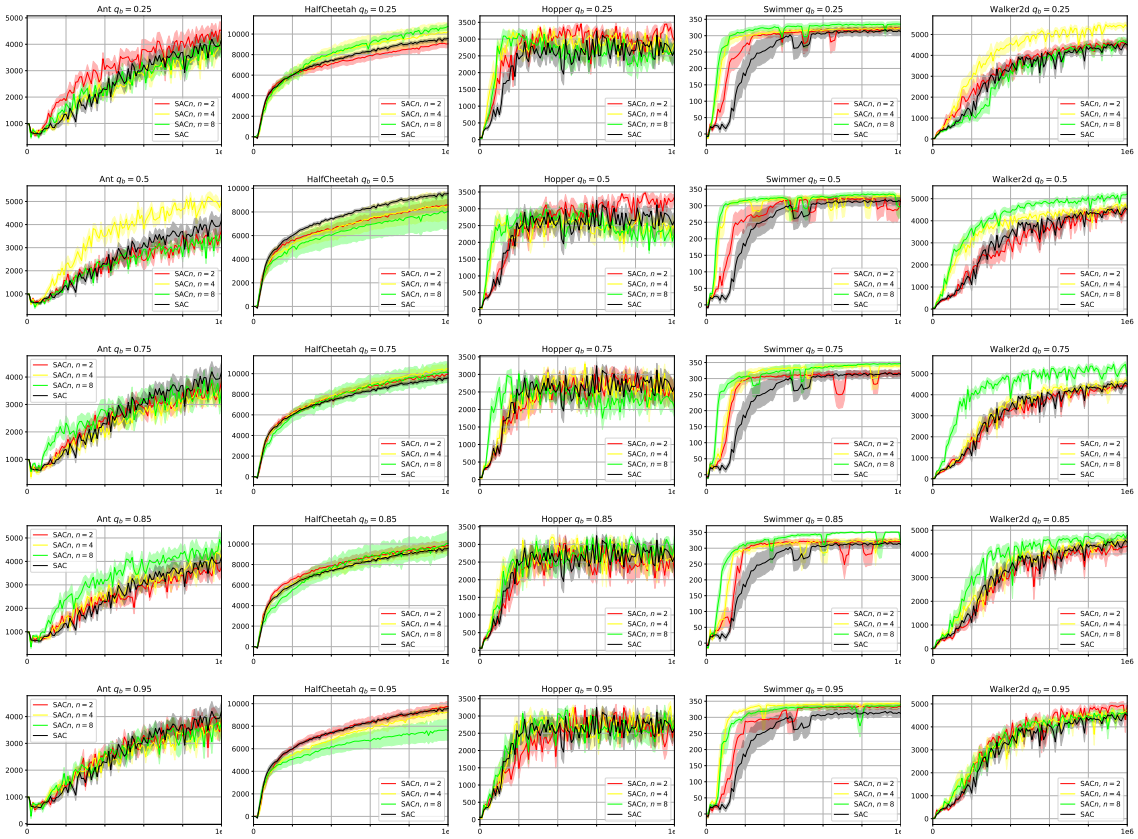

Figure 2: Learning curves for SACn with different values of q_b hyperparameter.

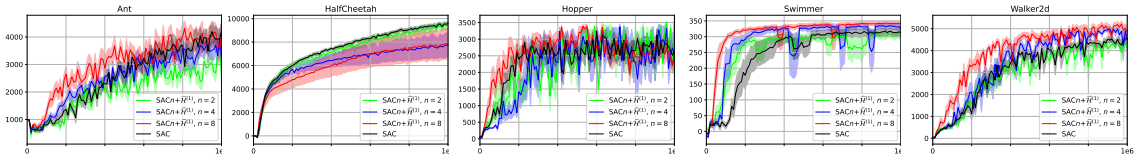

Figure 3: Learning curves for SACn without τ-sampled entropy estimation.

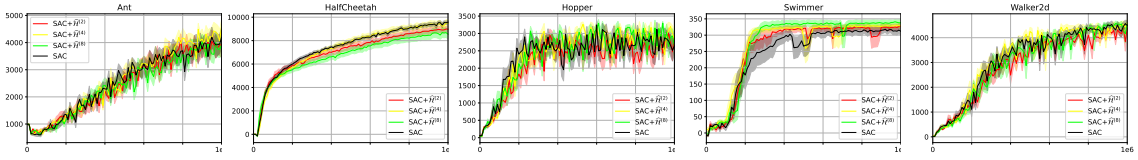

Figure 4: Learning curves for SAC with τ-sampled entropy estimation.