

A Off-Policy DRL algorithms

In this section, we briefly discuss the two off-policy learning algorithms, Soft Actor-Critic (SAC) [3] and Twin Delayed Deep Deterministic Policy Gradients (TD3) [2].

Twin Delayed Deep Deterministic Policy Gradients TD3 is a model-free, off-policy deep reinforcement learning algorithm. TD3 expands upon Deep Deterministic Policy Gradients (DDPG) [7] by employing a clipped double-Q trick to stabilize the actor’s learning [2]. The actor itself is parameterized as a neural mean function approximator $\mu_{\theta}(s)$ that predicts optimal mean actions given a state s . We build on the TD3 implementation provided by RLlib [6].

Soft Actor-Critic SAC is also a model-free, off-policy deep reinforcement learning algorithm. It aims to maximize expected improvement on an entropy-regularized objective and trades off exploration with exploitation. Here, we optimize the coefficient α and leverage the clipped double-Q trick that TD3 makes use of to stabilize learning. We build on the SAC implementation provided by RLlib [6].

B Nearest neighbor computation in NMER

In Neighborhood Mixup Experience Replay (NMER), nearest neighbors are computed using a concatenated state-action vector norm in tandem with Euclidean distance search over the Z-score standardized states and actions. Figure 1 illustrates this.

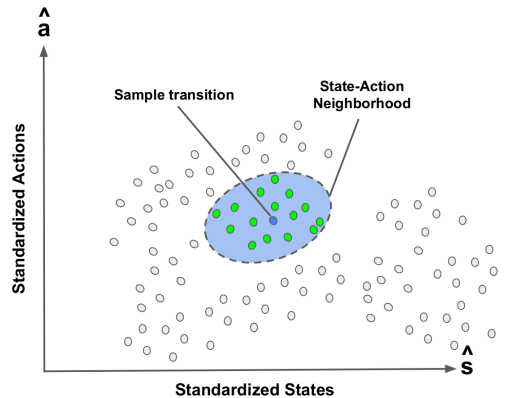


Figure 1: To compute nearest neighbors, NMER finds the closest transitions in the replay buffer to a given transition also sampled from the replay buffer in standardized state and action coordinates.

FAISS [5] is used to compute nearest neighbors using a C++ backend in tandem with GPU-based acceleration with CUDA [10]. Relative runtimes of NMER are given as averaged over four seeds in the Humanoid-v2 environment using our 200K environment interaction benchmark. As can be seen, NMER introduces minimal computational overhead, even for high-dimensional state and action spaces such as Humanoid-v2.

C Baselines

For our comparative evaluation studies, we implemented and evaluated several baseline replay buffers used in state-of-the-art off-policy experience replay.

Uniform Replay (U) Uniform Replay [9] is a circular replay buffer that selects stored transitions uniformly at random, in an independent and identically-distributed fashion. If the sample limit of this replay buffer is reached, samples are replaced in a first-in, first-out (FIFO) fashion.

Prioritized Experience Replay (PER) PER [13] extends Uniform Replay by selecting samples for training the DRL agent in a prioritized fashion, rather than uniformly at random. PER prioritizes these samples according to their estimated TD-error, which, intuitively, is a heuristic estimate for the “surprise”, and therefore learnability, a given transition induces on the DRL agent. For these comparative evaluation experiments, we make use of the stochastic prioritization variant of PER, in which samples are selected stochastically with sampling weights proportional to their priority. We make use of the PER implementation provided by [6].

Continuous Transition (CT) CT [8] is a replay buffer interpolation and data augmentation module designed for continuous state and action spaces. Specifically, it interpolates temporally-adjacent transitions, on the condition that they fall in the same episode. CT utilizes this temporal proximity as a heuristic, in tandem with a Mixup temperature (α) discriminator to ensure that interpolated transitions remain approximately on the transition manifold [8]. We implement our own variant of CT, as described in the experimental section of the manuscript.

D Additional Baselines

In addition to the baselines considered above, we also considered baselines to intuitively and rigorously demonstrate the strong empirical performance of NMER. These baselines include:

- **Noisy Replay** ($\mathcal{N}(0, \sigma^2)$): Perturbing the original transitions used to train the reinforcement learning agents with zero-mean, i.i.d. Gaussian noise. Below, is the standard deviation of the zero-mean i.i.d. Gaussian noise.
- **Naive Mixup** (Mixup): Standard Mixup - this is the limiting case where neighborhood size.
- **S4RL** (S4RL): The S4RL observation/next observation Mixup technique [14].
- **1-Nearest Neighbor Mixup** (1NN-Mixup): The other extreme neighborhood case. This baseline is paired with baseline 3 to better empirically quantify the effect of neighborhood size on agent performance.

The plots below compare the performance of the newly-added baselines to the replay buffers and baselines introduced in our manuscript found in Table 1.

E Other Ablative Factors Considered

In addition to running ablations over the replay ratio variable, we also consider the following ablation factors for future work in NMER:

1. **Neighborhood Size k** : Mixup and neighborhood sampling stand in somewhat stark contrast to one another - while Mixup encourages generalization by interpolating any pair of samples in the dataset [8], invoking a particularly strong inductive bias, neighborhood sampling reduces the degree of this inductive bias by restricting this bias to proximal neighborhoods of the inputs on the transition manifold. Systematically evaluating this trade-off can help to yield optimal k values for continuous control tasks and baseline RL algorithms (e.g. SAC, TD3).
2. **Fraction of Samples Interpolated**: Another ablative factor that can be used to measure the efficacy of NMER relative to the underlying baseline RL algorithms is the fraction of samples to interpolate in a mini-batch. An ablation study over this interpolation fraction can indicate whether a combination of observed and interpolated transitions results in better performance compared to entirely interpolated or entirely observed transitions.

F Reinforcement Learning Agent Hyperparameters

The hyperparameters in this section detail the hyperparameterization for the deep reinforcement learning agents applied for NMER. Namely, these hyperparameterizations are for Soft Actor-Critic (SAC) [3] and Twin Delayed Deep Deterministic Policy Gradients (TD3) [2]. Environment-specific parameters are provided in the following section.

F.1 Soft Actor-Critic (SAC) Parameters

Hyperparameters used for running comparative and ablation evaluation experiments on SAC [3] are provided in Table 1.

Table 1: Table of hyperparameters for Soft Actor-Critic (SAC) across the different environments on which these replay buffers were evaluated. These hyperparameters are adapted from [6].

Hyperparameter	Value
Actor Learning Rate	0.003
Critic Learning Rate	0.003
Alpha Learning Rate	0.003
Polyak (target network update coefficient) τ	0.005
Target Network Update Interval (Gradient Steps)	1
Entropy Target	‘auto’ ($\dim(\mathcal{A})$)
Initial Entropy Parameter (α)	1.0
Twin-Q	True
Normalize Actions	True
N-step	1
Gamma (γ)	0.99
Policy Parameterization	Squashed Gaussian
Clip Actions	False
Critic Hidden Units	[256, 256]
Critic activation function	ReLU
Actor Hidden Units	[256, 256]
Actor activation function	ReLU
Train batch size	256
Replay Buffer Size	1000000

F.1.1 SAC L2 regularization

In addition to the SAC hyperparameters given in Table 1, we empirically observe that all replay buffer variants, including Vanilla Replay, result in actor network divergence when combined with SAC and high replay ratios in many of the continuous control environments considered. To mitigate this divergence, we apply a small L2 regularization penalty to the actor network of the SAC agent. The ablation studies above detail this regularization value as ‘L’ for each replay buffer variant in each continuous control environment.

F.2 Twin Delayed Deep Deterministic Policy Gradients (TD3)

Hyperparameters used for running comparative and ablation evaluation experiments on TD3 [2] are provided in Table 2.

F.3 Prioritized Experience Replay Hyperparameters

Table 3 details the hyperparameter configurations for the stochastic prioritization variant of Prioritized Experience Replay (PER) [13] used for these replay buffers and variants. PER hyperparameters were kept constant for all environments and algorithms. We build on the PER implementation provided by RLlib [6].

G Environment details

For our comparative evaluation and ablation studies, we use the OpenAI MuJoCo continuous control task suite, as well as the OpenAI Classic Control suite. Dimensions of the continuous state and action spaces for each of these environments, as well as whether termination signals are automatically applied at the end of an episode, can be found in Table 4.

Table 2: Table of hyperparameters for Twin Delayed Deep Deterministic Policy Gradients (TD3) across the different environments on which these replay buffers were evaluated. These hyperparameters are adapted from [6].

Hyperparameter	Value
Actor Learning Rate	0.0005
Critic Learning Rate	0.0005
Polyak (target network update coefficient) τ	0.005
Target Network Update Interval (Gradient Steps)	1
Twin-Q	True
Normalize Actions	True
Policy Parameterization	Squashed Gaussian
Policy Delay	2
Smooth Target Policy	True
Target Noise	0.2
Target Noise Clip	0.5
Exploration Noise Type	$\mathcal{N}(0, 0.1)$
Random Steps	10000
N-step	1
Gamma (γ)	0.99
Clip Actions	False
Critic Hidden Units	[400, 300]
Actor Hidden Units	[400, 300]
Train batch size	100
L2 Regularization	0
Replay Buffer Size	1000000

Hyperparameter	Value
α	0.6
Initial β	0.4
Final β	0.4
β Annealing Time Steps	20000
ϵ	0.000001

Table 3: Table of Prioritized Experience Replay (PER) [13] hyperparameters used for all comparative and evaluation studies with the PER baseline. These hyperparameters are adapted from [6].

G.1 Determining appropriate episode termination signals

For these continuous control tasks, it is important to determine appropriate episode termination signals provided to the agent in order for them to properly learn which steps mark the end of an episode. Since many of these continuous control tasks are infinite-horizon in nature, simply setting the final step of an episode to have a termination signal may inadvertently train the agent to avoid this state-action pair in the future, since it is treated as the last step regardless of whether early episode termination conditions (e.g. the `Walker2d-v2` agent falling over) actually apply. On the contrary, it is also important that agents are provided with information related to termination signals when true termination conditions actually apply.

Therefore, in general, we turn termination signals off for these environments. For environments in which early termination signals can apply (e.g. the `Walker2d-v2` agent falling over), we allow for termination signals, but only apply them if the termination signal occurs before the horizon. If the termination signal occurs at the horizon, we do not use a termination signal, since this implies this step is the last in the horizon. Table 4 shows whether a termination signal is applied automatically on the final step of the environment. This additional conditional termination logic is applied on top of environments in which ‘No Done At End?’ is set to ‘False’.

Crucially, this logic is applied to all variants tested for our comparative evaluation and ablation studies in order to ensure fairness to the different replay buffer variants being evaluated.

Table 4: State and action space dimensions for the continuous control environments we ran comparative evaluation and ablation studies on with replay buffers.

Environment	Number of States	Number of Actions	No Done At End?
Ant-v2	111	8	False
HalfCheetah-v2	17	6	True
Hopper-v2	8	3	False
Swimmer-v2	8	2	True
Walker2d-v2	17	6	False
Humanoid-v2	376	17	False
InvertedPendulum-v2	4	1	False
InvertedDoublePendulum-v2	11	1	False
Reacher	2	11	True
Pendulum-v0	3	1	True
MountainCarContinuous-v0	2	1	True

H Computing environment

To run the aforementioned comparative evaluation and ablation studies, we run experiments in either one of two configurations based off of compute availability: (i) 5 Intel Xeon-p8 CPUs, (ii) 10 Intel Xeon-p6 CPUs in tandem with 1 Nvidia Volta V100 GPU. Our GPU jobs are typically allocated to NMER to accelerate nearest neighbor searches; however, when comparing run times of NMER to our other tested baselines, we compare these replay variants when they are all run on the second (CPU + GPU) hardware configuration.

H.1 Libraries

NMER is built using the following packages:

1. **RLlib** [6]: Integrating our NMER code with a scalable and efficient framework for running RL experiments. This framework uses PyTorch and OpenAI Gym and MuJoCo on the backend.
2. **OpenAI Gym** [1]: Running simulations in reinforcement learning libraries to evaluate the effectiveness and performance of our BIER module. The environments we will be using for this project use MuJoCo [15] on the backend.
3. **MuJoCo** [15]: Physics-based simulator for our reinforcement learning libraries - this will be used primarily for running simulations in tandem with OpenAI Gym.
4. **NumPy** [4]: Used for pre-processing, post-processing, and matrix computation.
5. **Scikit-Learn** [12]: Used for performing standardization of states and actions of the stored transition in order to compute nearest neighbors in standardized state-action space.
6. **FAISS** [5]: Used for improving runtime efficiency of nearest neighbor classes, particularly for large replay buffers. These classes are implemented using C++ and are GPU-compatible.
7. **CUDA** [10]: Used for GPU-based hardware acceleration. This is especially applicable for our NMER nearest neighbor search because this batched routine is highly amenable for efficient and parallelized vectorization.

H.2 RNG Seed Setting

To ensure reproducibility and to compare replay variants to one another directly, seeding the random number generators for Python packages that leverage stochastic routines or subroutines was performed. These packages included:

1. PyTorch (CPU) [11]
2. PyTorch (GPU) [11]
3. NumPy [4]

4. Random (Python)
5. Scikit-Learn [12]
6. RLlib [6]

I Codebase and trained agents

NMER source code can be found at <https://github.com/rmsander/interreplay>.

J Detailed Ablation Studies, OpenAI Gym

Ablation studies were to determine the optimal replay ratio for each replay buffer variant. Since this optimal replay ratio differs for each replay variant as well as for each environment, it was important to run ablations over several replay ratios, namely 1, 5, and 20. The graphs below provide insight as to how the optimal replay buffer was chosen. These graphs are divided according to the underlying deep reinforcement learning algorithm run.

Additionally, for Soft Actor-Critic experiments, ablation studies were performed to determine approximately minimal actor network L2 regularization needed in order to ensure policy stability. We empirically observe actor network divergence, particularly for higher replay ratios, in the absence of actor network regularization for SAC. In nearly all environments, the L2 regularization coefficient is set to be smaller for the baseline replay buffers than for NMER, in order to ensure that NMER was not gaining an unfair advantage through having less constraints on its neural policy.

Summarized, these ablation factors are:

1. **Replay ratio (RR)**: The number of gradient training steps for every environment interaction the agent has with its environment. The replay ratios tested in these experiments were 1, 5, and 20. These replay ratio values are denoted by ‘RR’.
2. **(SAC-only) L2 Actor Network Regularization (L)**: The L2 regularization coefficient used to ensure actor network stability for the Soft Actor-Critic studies. These values are denoted by ‘L’. Additionally, for the Hopper-v2 experiments, we consider a gradient clipping value of 40 (SAC-only).

These values are given in each of the graphs and tables below.

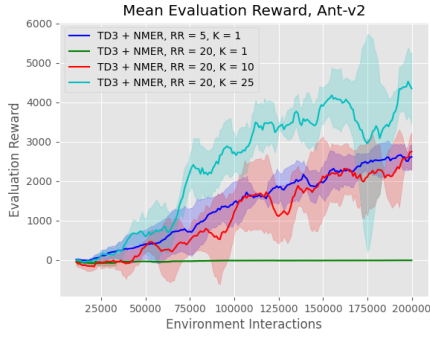
J.1 TD3 Ablation Studies

For each environment and replay buffer variant (U, PER, CT, and NMER), we run ablation studies in which we vary the replay ratio to determine the optimal replay ratio for each variant. These ablation studies are conducted to ensure that we compare NMER against competitive baselines. Empirically, these ablation studies are motivated by observing that for one subset of the following continuous control tasks, our baseline replay buffers perform better with lower replay ratios, and for another subset of these continuous control tasks, these same baseline replay buffers perform better with higher replay ratios.

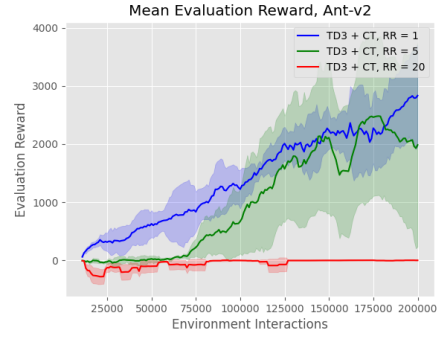
Results from each (environment, replay buffer) variant are given in the following tables and figures. The best ablation result for each replay buffer variant is shown in bold.

Table 5: Evaluation reward for TD3 and Ant-v2 outfitted with replay buffers with varying replay ratios.

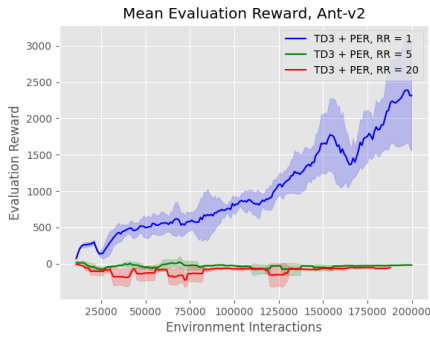
Replay Buffer	Reward (200K steps)
U, RR=1	2005 \pm 399
U, RR=5	-12 \pm 3
U, RR=20	-50 \pm 5
PER, RR=1 [13]	2317 \pm 756
PER, RR=5 [13]	-19 \pm 4
PER, RR=20 [13]	-50 \pm 9
CT, RR=1 [8]	2834 \pm 875
CT, RR=5 [8]	1986 \pm 1750
CT, RR=20 [8]	3 \pm 6
NMER, K=25, RR=20	4347 \pm 908



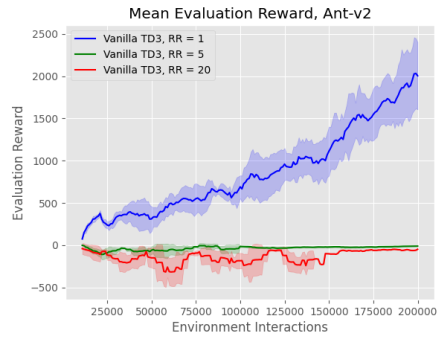
(a) Ablation study results with TD3 + NMER on the Ant-v2 environment. This ablation study allows for selecting the optimal NMER replay buffer to compare the other replay buffers to.



(b) Ablation study results with TD3 + CT on the Ant-v2 environment. This ablation study allows for selecting the optimal CT replay buffer to compare the other replay buffers to.



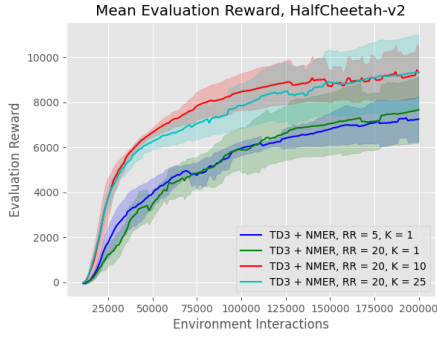
(a) Ablation study results with TD3 + PER on the Ant-v2 environment. This ablation study allows for selecting the optimal PER replay buffer to compare the other replay buffers to.



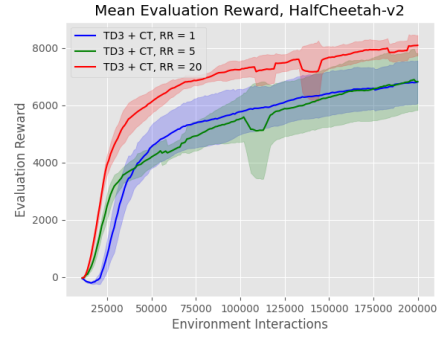
(b) Ablation study results with vanilla TD3 on the Ant-v2 environment. This ablation study allows for selecting the optimal vanilla replay buffer to compare the other replay buffers to.

Table 6: Evaluation reward for TD3 and HalfCheetah-v2 outfitted with replay buffers with varying replay ratios.

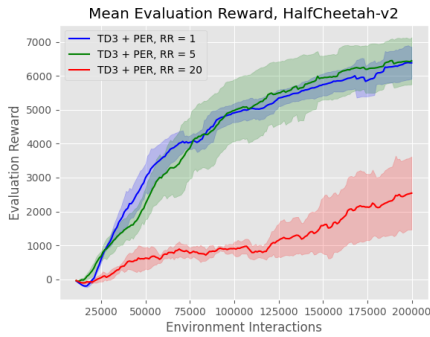
Replay Buffer	Reward (200K steps)
U, RR=1	6467 \pm 658
U, RR=5	6295 \pm 560
U, RR=20	1059 \pm 574
PER, RR=1 [13]	6388 \pm 474
PER, RR=5 [13]	6447 \pm 693
PER, RR=20 [13]	2542 \pm 1078
CT, RR=1 [8]	6821 \pm 745
CT, RR=5 [8]	6831 \pm 997
CT, RR=20 [8]	8097 \pm 358
NMER, K=10, RR=20	9327 \pm 1099



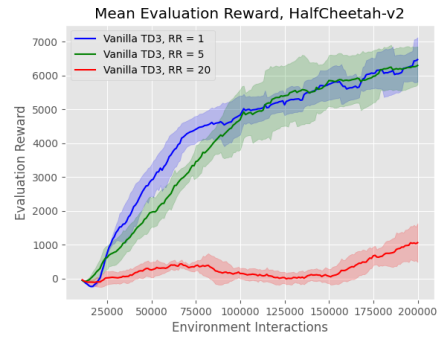
(a) Ablation study results with TD3 + NMER on the HalfCheetah-v2 environment. This ablation study allows for selecting the optimal NMER replay buffer to compare the other replay buffers to.



(b) Ablation study results with TD3 + CT on the HalfCheetah-v2 environment. This ablation study allows for selecting the optimal CT replay buffer to compare the other replay buffers to.



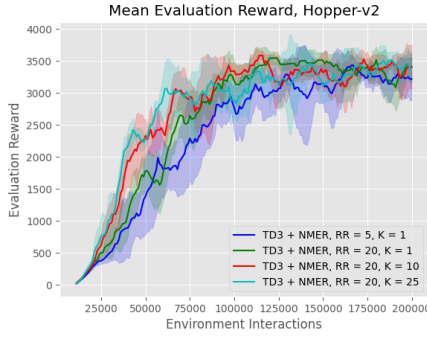
(a) Ablation study results with TD3 + PER on the HalfCheetah-v2 environment. This ablation study allows for selecting the optimal PER replay buffer to compare the other replay buffers to.



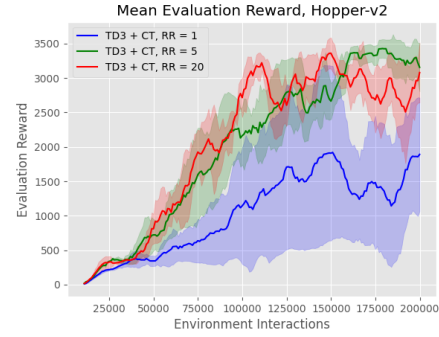
(b) Ablation study results with vanilla TD3 on the HalfCheetah-v2 environment. This ablation study allows for selecting the optimal vanilla replay buffer to compare the other replay buffers to.

Table 7: Evaluation reward for TD3 and Hopper-v2 outfitted with replay buffers with varying replay ratios.

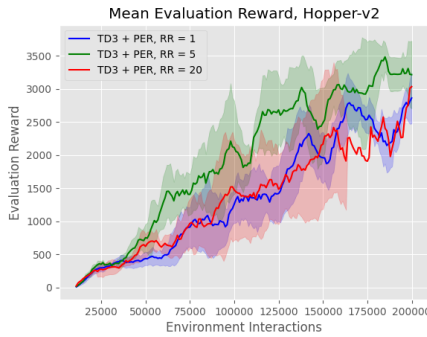
Replay Buffer	Reward (200K steps)
U, RR=1	2371 \pm 593
U, RR=5	2758 \pm 607
U, RR=20	3252 \pm 157
PER, RR=1 [13]	2860 \pm 387
PER, RR=5 [13]	3213 \pm 511
PER, RR=20 [13]	2190 \pm 530
CT, RR=1 [8]	1891 \pm 825
CT, RR=5 [8]	3156 \pm 351
CT, RR=20 [8]	3080 \pm 437
NMER, K=10, RR=20	3411 \pm 340



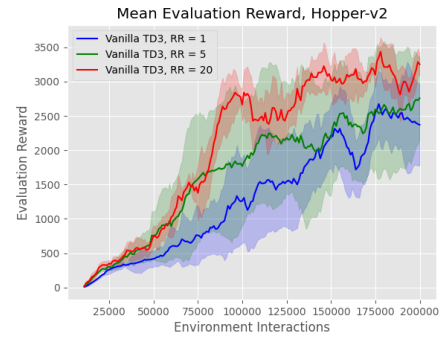
(a) Ablation study results with TD3 + NMER on the Hopper-v2 environment. This ablation study allows for selecting the optimal NMER replay buffer to compare the other replay buffers to.



(b) Ablation study results with TD3 + CT on the Hopper-v2 environment. This ablation study allows for selecting the optimal CT replay buffer to compare the other replay buffers to.



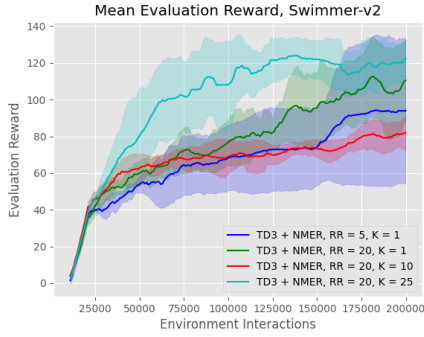
(a) Ablation study results with TD3 + PER on the Hopper-v2 environment. This ablation study allows for selecting the optimal PER replay buffer to compare the other replay buffers to.



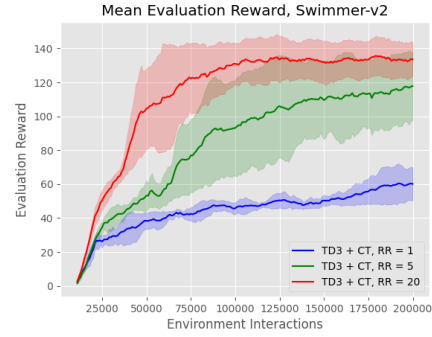
(b) Ablation study results with vanilla TD3 on the Hopper-v2 environment. This ablation study allows for selecting the optimal vanilla replay buffer to compare the other replay buffers to.

Table 8: Evaluation reward for TD3 and Swimmer-v2 outfitted with replay buffers with varying replay ratios.

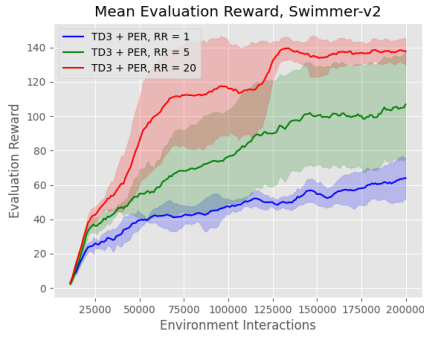
Replay Buffer	Reward (200K steps)
U, RR=1	56 ± 8
U, RR=5	92 ± 25
U, RR=20	131 ± 20
PER, RR=1 [13]	64 ± 12
PER, RR=5 [13]	107 ± 31
PER, RR=20 [13]	138 ± 8
CT, RR=1 [8]	60 ± 10
CT, RR=5 [8]	118 ± 20
CT, RR=20 [8]	134 ± 10
NMER, K=10, RR=20	131 ± 20



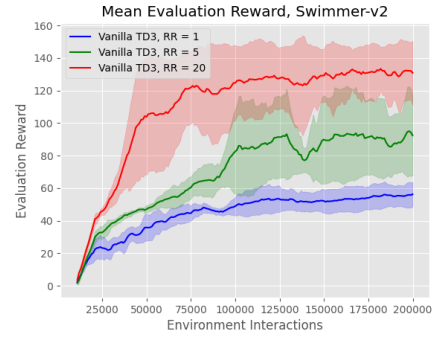
(a) Ablation study results with TD3 + NMER on the Swimmer-v2 environment. This ablation study allows for selecting the optimal NMER replay buffer to compare the other replay buffers to.



(b) Ablation study results with TD3 + CT on the Swimmer-v2 environment. This ablation study allows for selecting the optimal CT replay buffer to compare the other replay buffers to.



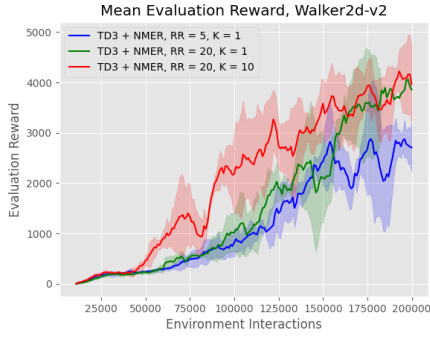
(a) Ablation study results with TD3 + PER on the Swimmer-v2 environment. This ablation study allows for selecting the optimal PER replay buffer to compare the other replay buffers to.



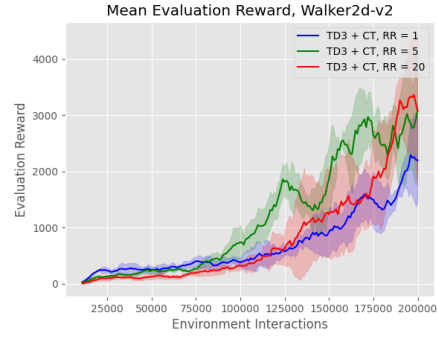
(b) Ablation study results with vanilla TD3 on the Swimmer-v2 environment. This ablation study allows for selecting the optimal vanilla replay buffer to compare the other replay buffers to.

Table 9: Evaluation reward for TD3 and Walker2d-v2 outfitted with replay buffers with varying replay ratios.

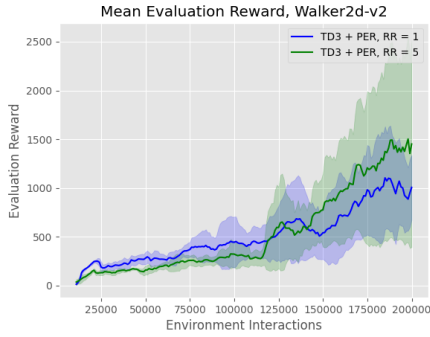
Replay Buffer	Reward (200K steps)
U, RR=1	2236 \pm 686
U, RR=5	1833 \pm 934
U, RR=20	682 \pm 806
PER, RR=1 [13]	1006 \pm 332
PER, RR=5 [13]	1452 \pm 1057
PER, RR=20 [13]	476 \pm 646
CT, RR=1 [8]	2198 \pm 770
CT, RR=5 [8]	3087 \pm 1058
CT, RR=20 [8]	3079 \pm 1331
NMER, K=10, RR=20	3960 \pm 777



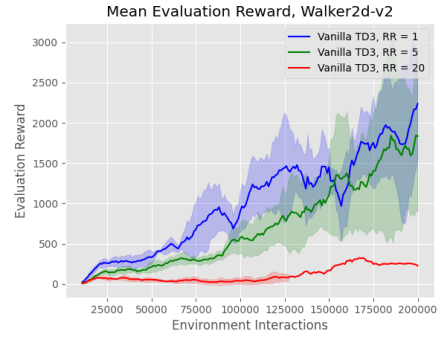
(a) Ablation study results with TD3 + NMER on the Walker2d-v2 environment. This ablation study allows for selecting the optimal NMER replay buffer to compare the other replay buffers to.



(b) Ablation study results with TD3 + CT on the Walker2d-v2 environment. This ablation study allows for selecting the optimal CT replay buffer to compare the other replay buffers to.



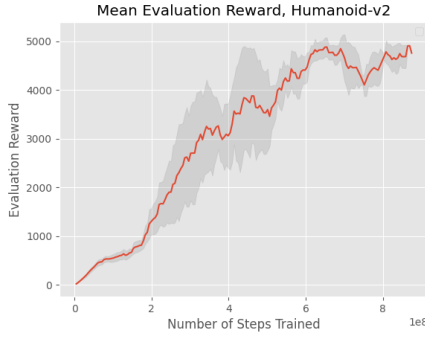
(a) Ablation study results with TD3 + PER on the Walker2d-v2 environment. This ablation study allows for selecting the optimal PER replay buffer to compare the other replay buffers to.



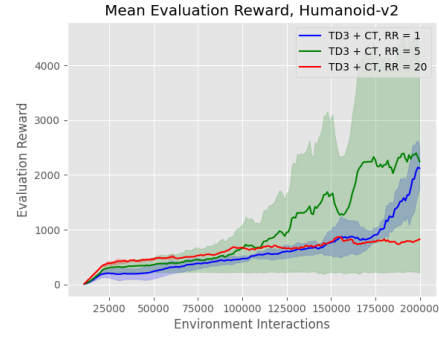
(b) Ablation study results with vanilla TD3 on the Walker2d-v2 environment. This ablation study allows for selecting the optimal vanilla replay buffer to compare the other replay buffers to.

Table 10: Evaluation reward for TD3 and Humanoid-v2 outfitted with replay buffers with varying replay ratios.

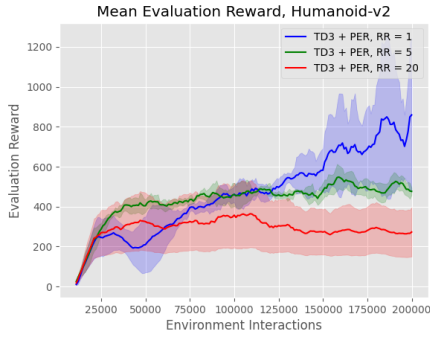
Replay Buffer	Reward (200K steps)
U, RR=1	293 ± 0
U, RR=5	388 ± 66
U, RR=20	385 ± 0
PER, RR=1 [13]	860 ± 385
PER, RR=5 [13]	476 ± 18
PER, RR=20 [13]	273 ± 123
CT, RR=1 [8]	2122 ± 372
CT, RR=5 [8]	2242 ± 2027
CT, RR=20 [8]	823 ± 0
NMER, K=10, RR=20	4791 ± 271



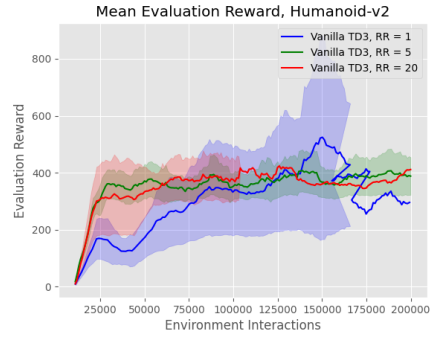
(a) Ablation study results with TD3 + NMER on the Humanoid-v2 environment. This ablation study allows for selecting the optimal NMER replay buffer to compare the other replay buffers to.



(b) Ablation study results with TD3 + CT on the Humanoid-v2 environment. This ablation study allows for selecting the optimal CT replay buffer to compare the other replay buffers to.



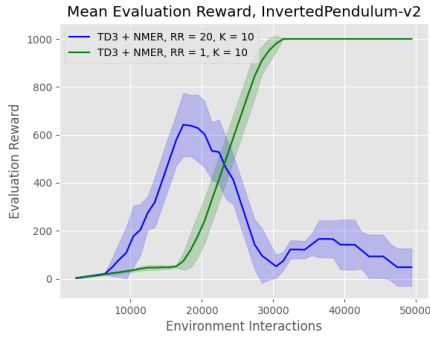
(a) Ablation study results with TD3 + PER on the Humanoid-v2 environment. This ablation study allows for selecting the optimal PER replay buffer to compare the other replay buffers to.



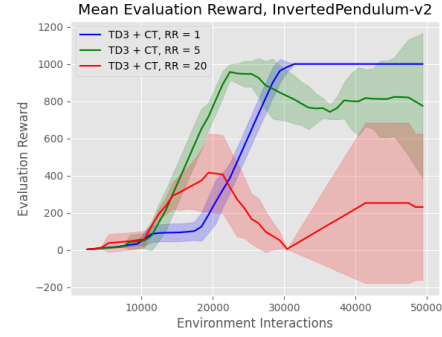
(b) Ablation study results with vanilla TD3 on the Humanoid-v2 environment. This ablation study allows for selecting the optimal vanilla replay buffer to compare the other replay buffers to.

Table 11: Evaluation reward for TD3 and InvertedPendulum-v2 outfitted with replay buffers with varying replay ratios.

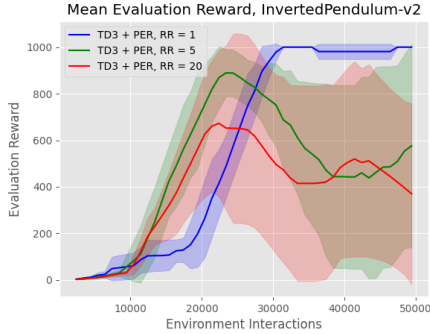
Replay Buffer	Reward (50K steps)
U, RR=1	1000 \pm 0
U, RR=5	982 \pm 32
U, RR=20	49 \pm 80
PER, RR=1 [13]	1000 \pm 0
PER, RR=5 [13]	576 \pm 437
PER, RR=20 [13]	370 \pm 388
CT, RR=1 [8]	1000 \pm 0
CT, RR=5 [8]	775 \pm 390
CT, RR=20 [8]	230 \pm 394
NMER, K=10, RR=1	1000 \pm 0
NMER, K=10, RR=20	48 \pm 79



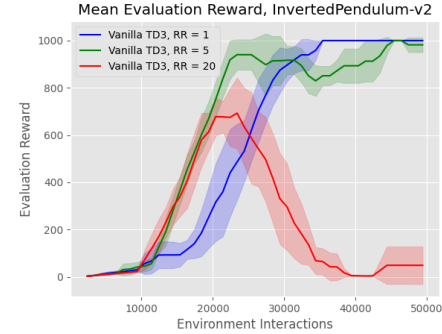
(a) Ablation study results with TD3 + NMER on the InvertedPendulum-v2 environment. This ablation study allows for selecting the optimal NMER replay buffer to compare the other replay buffers to.



(b) Ablation study results with TD3 + CT on the InvertedPendulum-v2 environment. This ablation study allows for selecting the optimal CT replay buffer to compare the other replay buffers to.



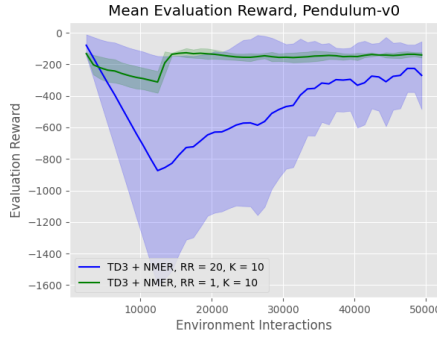
(a) Ablation study results with TD3 + PER on the InvertedPendulum-v2 environment. This ablation study allows for selecting the optimal PER replay buffer to compare the other replay buffers to.



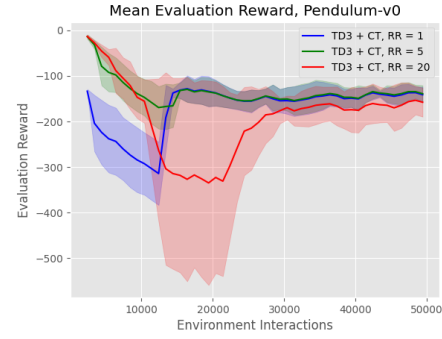
(b) Ablation study results with vanilla TD3 on the InvertedPendulum-v2 environment. This ablation study allows for selecting the optimal vanilla replay buffer to compare the other replay buffers to.

Table 12: Evaluation reward for TD3 and Pendulum-v0 outfitted with replay buffers with varying replay ratios.

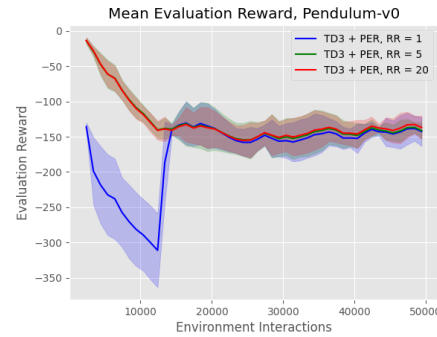
Replay Buffer	Reward (50K steps)
U, RR=1	-138 \pm 18
U, RR=5	-138 \pm 20
U, RR=20	-263 \pm 194
PER, RR=1 [13]	-143 \pm 21
PER, RR=5 [13]	-141 \pm 11
PER, RR=20 [13]	-137 \pm 16
CT, RR=1 [8]	-141 \pm 17
CT, RR=5 [8]	-139 \pm 19
CT, RR=20 [8]	-158 \pm 32
NMER, K=10, RR=1	-141 \pm 17
NMER, K=10, RR=20	-270 \pm 215



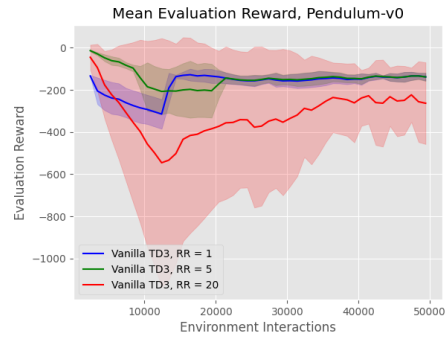
(a) Ablation study results with TD3 + NMER on the Pendulum-v0 environment. This ablation study allows for selecting the optimal NMER replay buffer to compare the other replay buffers to.



(b) Ablation study results with TD3 + CT on the Pendulum-v0 environment. This ablation study allows for selecting the optimal CT replay buffer to compare the other replay buffers to.



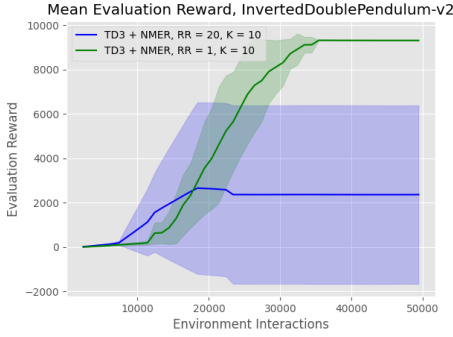
(a) Ablation study results with TD3 + PER on the Pendulum-v0 environment. This ablation study allows for selecting the optimal PER replay buffer to compare the other replay buffers to.



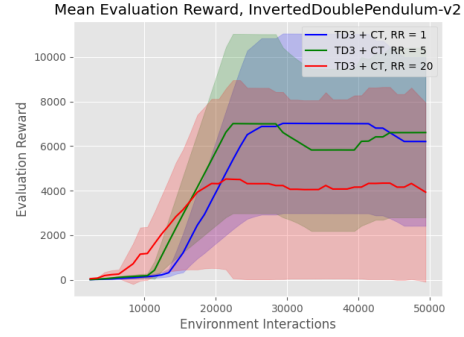
(b) Ablation study results with vanilla TD3 on the Pendulum-v0 environment. This ablation study allows for selecting the optimal vanilla replay buffer to compare the other replay buffers to.

Table 13: Evaluation reward for TD3 and InvertedDoublePendulum-v2 outfitted with replay buffers with varying replay ratios.

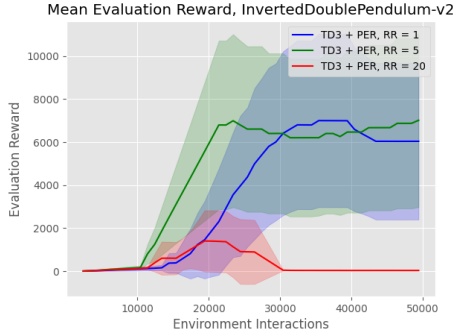
Replay Buffer	Reward (50K steps)
U, RR=1	6181 \pm 3790
U, RR=5	6451 \pm 3810
U, RR=20	41 \pm 12
PER, RR=1 [13]	6037 \pm 3640
PER, RR=5 [13]	7015 \pm 4033
PER, RR=20 [13]	36 \pm 4
CT, RR=1 [8]	6207 \pm 3791
CT, RR=5 [8]	6609 \pm 3812
CT, RR=20 [8]	3931 \pm 4026
NMER, K=10, RR=1	9312 \pm 1
NMER, K=10, RR=20	2366 \pm 4025



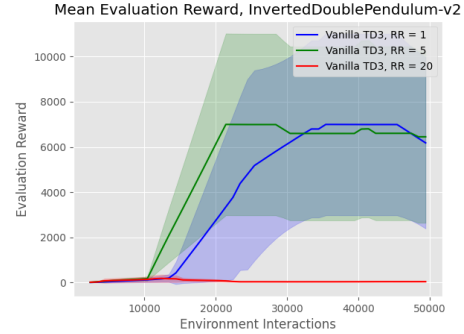
(a) Ablation study results with TD3 + NMER on the InvertedDoublePendulum-v2 environment. This ablation study allows for selecting the optimal NMER replay buffer to compare the other replay buffers to.



(b) Ablation study results with TD3 + CT on the InvertedDoublePendulum-v2 environment. This ablation study allows for selecting the optimal CT replay buffer to compare the other replay buffers to.



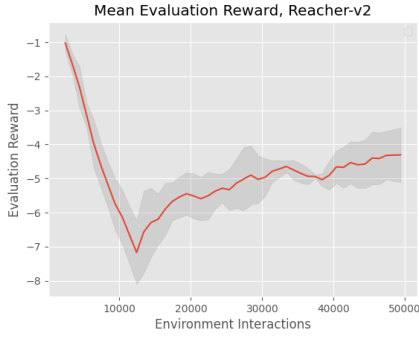
(a) Ablation study results with TD3 + PER on the InvertedDoublePendulum-v2 environment. This ablation study allows for selecting the optimal PER replay buffer to compare the other replay buffers to.



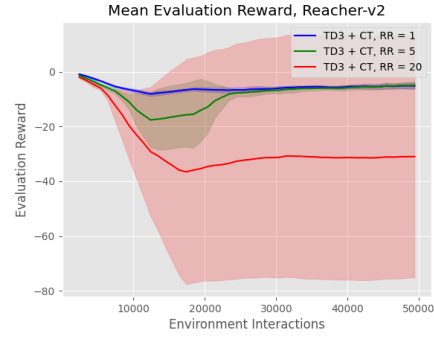
(b) Ablation study results with vanilla TD3 on the InvertedDoublePendulum-v2 environment. This ablation study allows for selecting the optimal vanilla replay buffer to compare the other replay buffers to.

Table 14: Evaluation reward for TD3 and Reacher-v2 outfitted with replay buffers with varying replay ratios.

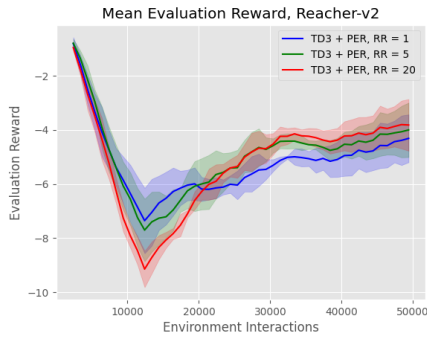
Replay Buffer	Reward (50K steps)
U, RR=1	-4 ± 1
U, RR=5	-4 ± 1
U, RR=20	-4 ± 1
PER, RR=1 [13]	-4 ± 1
PER, RR=5 [13]	-4 ± 1
PER, RR=20 [13]	-4 ± 1
CT, RR=1 [8]	-5 ± 1
CT, RR=5 [8]	-5 ± 1
CT, RR=20 [8]	-31 ± 44
NMER, K=10, RR=1	-4 ± 1



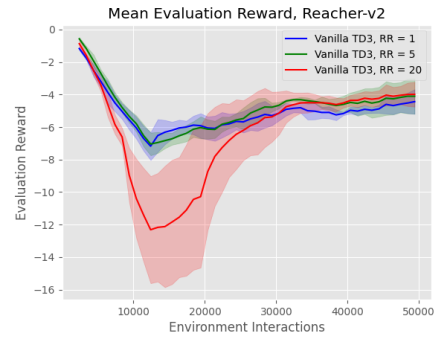
(a) Ablation study results with TD3 + NMER on the Reacher-v2 environment. This ablation study allows for selecting the optimal NMER replay buffer to compare the other replay buffers to.



(b) Ablation study results with TD3 + CT on the Reacher-v2 environment. This ablation study allows for selecting the optimal CT replay buffer to compare the other replay buffers to.



(a) Ablation study results with TD3 + PER on the Reacher-v2 environment. This ablation study allows for selecting the optimal PER replay buffer to compare the other replay buffers to.

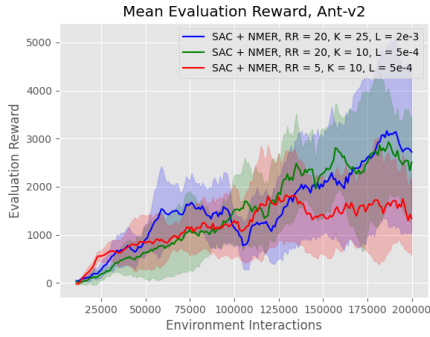


(b) Ablation study results with vanilla TD3 on the Reacher-v2 environment. This ablation study allows for selecting the optimal vanilla replay buffer to compare the other replay buffers to.

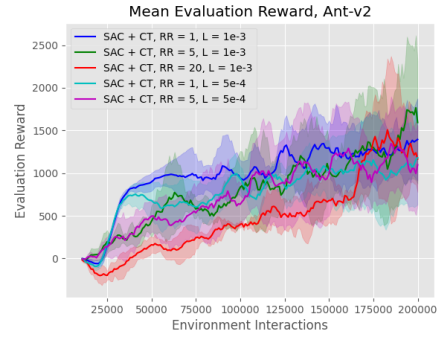
Table 15: Evaluation reward for SAC and Ant-v2 outfitted with replay buffers with varying replay ratios.

Replay Buffer	Reward (200K steps)
U, RR=1, L=1e-3	975 \pm 63
U, RR=1, L=5e-4	1111 \pm 235
U, RR=5, L=1e-3	1188 \pm 692
U, RR=5, L=5e-4	822 \pm 120
U, RR=20, L=1e-3	610 \pm 141
PER, RR=1, L=1e-3 [13]	878 \pm 239
PER, RR=5, L=1e-3 [13]	800 \pm 160
PER, RR=1, L=5e-4 [13]	655 \pm 257
PER, RR=5, L=5e-4 [13]	581 \pm 184
CT, RR=1, L=1e-3 [8]	1395 \pm 481
CT, RR=5, L=1e-3 [8]	1594 \pm 717
CT, RR=20, L=1e-3 [8]	1166 \pm 388
CT, RR=1, L=5e-4 [8]	1181 \pm 547
CT, RR=5, L=5e-4 [8]	1098 \pm 472
NMER, K=25, RR=20, L=2e-3	2723 \pm 1685
NMER, K=10, RR=20, L=5e-4	2507 \pm 962
NMER, K=10, RR=5, L=5e-4	1342 \pm 773

J.2 SAC Ablation Studies



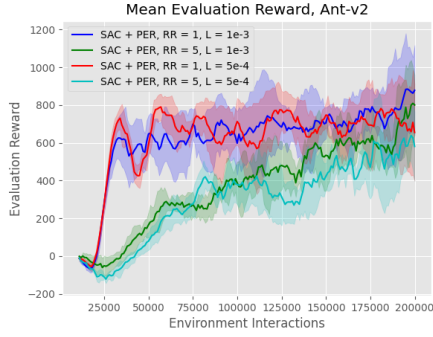
(a) Ablation study results with SAC + NMER on the Ant-v2 environment. This ablation study allows for selecting the optimal NMER replay buffer to compare the other replay buffers to.



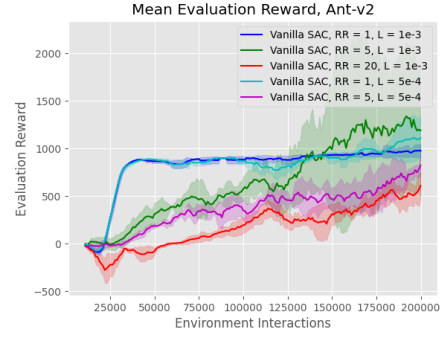
(b) Ablation study results with SAC + CT on the Ant-v2 environment. This ablation study allows for selecting the optimal CT replay buffer to compare the other replay buffers to.

Table 16: Evaluation reward for SAC and HalfCheetah-v2 outfitted with replay buffers with varying replay ratios.

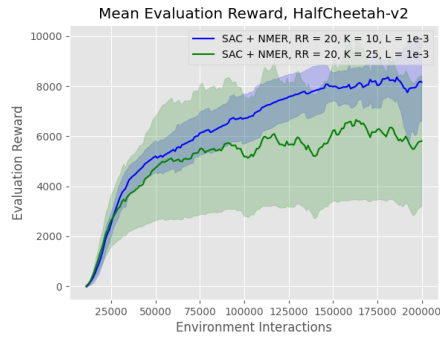
Replay Buffer	Reward (200K steps)
U, RR=1, L=1e-3	3424 \pm 2093
U, RR=5, L=1e-3	4918 \pm 1928
U, RR=20, L=5e-4	4297 \pm 1700
PER, RR=1, L=1e-3 [13]	4665 \pm 1926
PER, RR=5, L=1e-3 [13]	6518 \pm 1117
PER, RR=20, L=1e-3 [13]	6444 \pm 364
PER, RR=1, L=5e-4 [13]	5002 \pm 2093
PER, RR=5, L=5e-4 [13]	6880 \pm 886
PER, RR=20, L=5e-4 [13]	5358 \pm 415
PER, RR=20, L=1e-4 [13]	6134 \pm 891
CT, RR=1, L=1e-3 [8]	5102 \pm 592
CT, RR=5, L=1e-3 [8]	4271 \pm 1934
CT, RR=20, L=1e-3 [8]	5263 \pm 2146
CT, RR=1, L=5e-4 [8]	4108 \pm 1523
CT, RR=5, L=5e-4 [8]	5121 \pm 738
CT, RR=20, L=5e-4 [8]	4894 \pm 2002
NMER, K=10, RR=20, L=1e-3	8168 \pm 1585
NMER, K=25, RR=20, L=1e-3	5812 \pm 2619



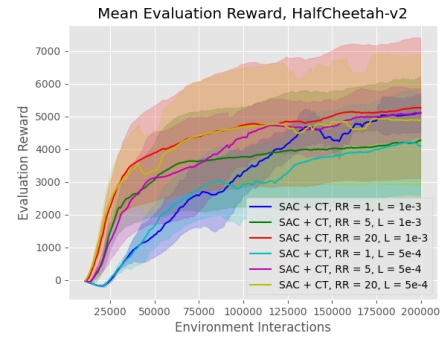
(a) Ablation study results with SAC + PER on the Ant-v2 environment. This ablation study allows for selecting the optimal PER replay buffer to compare the other replay buffers to.



(b) Ablation study results with vanilla SAC on the Ant-v2 environment. This ablation study allows for selecting the optimal vanilla replay buffer to compare the other replay buffers to.



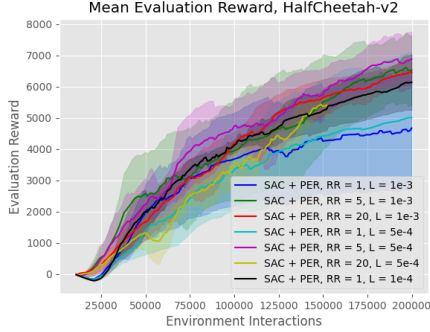
(a) Ablation study results with SAC + NMER on the HalfCheetah-v2 environment. This ablation study allows for selecting the optimal NMER replay buffer to compare the other replay buffers to.



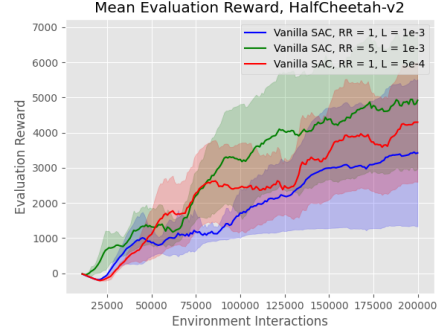
(b) Ablation study results with SAC + CT on the HalfCheetah-v2 environment. This ablation study allows for selecting the optimal CT replay buffer to compare the other replay buffers to.

Table 17: Evaluation reward for SAC and Hopper-v2 outfitted with replay buffers with varying replay ratios.

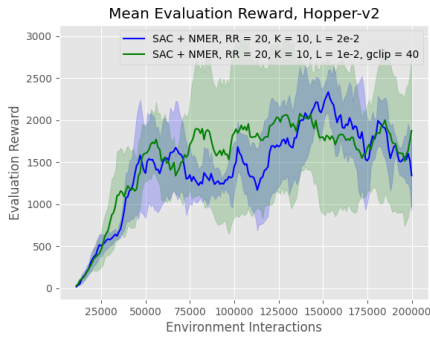
Replay Buffer	Reward (200K steps)
U, RR=1, L=1e-2	427 \pm 303
U, RR=5, L=1e-2	1333 \pm 1006
U, RR=20, L=1e-2	763 \pm 795
U, RR=5, L=1e-3, G=40	1692 \pm 1160
PER, RR=1, L=1e-2 [13]	2472 \pm 444
PER, RR=5, L=1e-2 [13]	2630 \pm 843
PER, RR=5, L=1e-3 [13]	2296 \pm 946
PER, RR=5, L=5e-3, G=40 [13]	2449 \pm 452
CT, RR=1, L=1e-2 [8]	412 \pm 164
CT, RR=5, L=1e-2 [8]	937 \pm 592
CT, RR=20, L=5e-3, G=40 [8]	1115 \pm 897
NMER, K=10, RR=20, L=2e-2	1343 \pm 424
NMER, K=10, RR=20, L=1e-2, G=40	1875 \pm 900



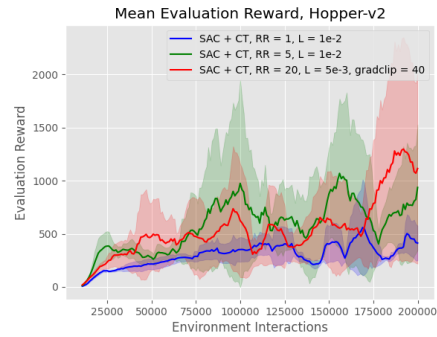
(a) Ablation study results with SAC + PER on the HalfCheetah-v2 environment. This ablation study allows for selecting the optimal PER replay buffer to compare the other replay buffers to.



(b) Ablation study results with vanilla SAC on the HalfCheetah-v2 environment. This ablation study allows for selecting the optimal vanilla replay buffer to compare the other replay buffers to.



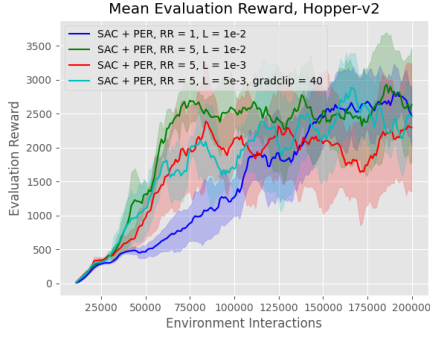
(a) Ablation study results with SAC + NMER on the Hopper-v2 environment. This ablation study allows for selecting the optimal NMER replay buffer to compare the other replay buffers to.



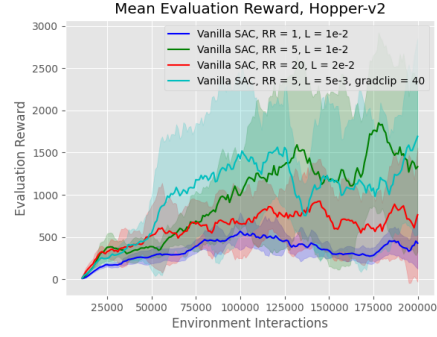
(b) Ablation study results with SAC + CT on the Hopper-v2 environment. This ablation study allows for selecting the optimal CT replay buffer to compare the other replay buffers to.

Table 18: Evaluation reward for SAC and Swimmer-v2 outfitted with replay buffers with varying replay ratios.

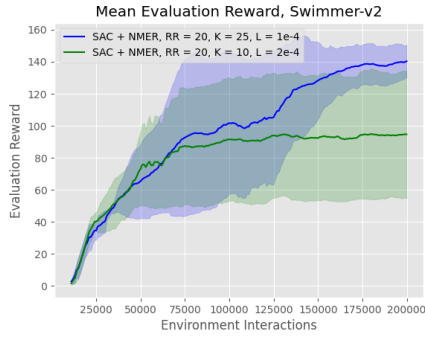
Replay Buffer	Reward (200K steps)
U, RR=1, L=1e-4	51 \pm 7
U, RR=5, L=1e-4	79 \pm 25
U, RR=20, L=1e-4	111 \pm 29
PER, RR=1, L=1e-4 [13]	48 \pm 1
PER, RR=5, L=1e-4 [13]	114 \pm 31
PER, RR=20, L=1e-4 [13]	121 \pm 42
CT, RR=1, L=1e-4 [8]	51 \pm 9
CT, RR=5, L=1e-4 [8]	74 \pm 20
CT, RR=20, L=5e-4 [8]	106 \pm 46
NMER, K=25, RR=20, L=1e-4	140 \pm 10
NMER, K=10, RR=20, L=2e-4	95 \pm 40



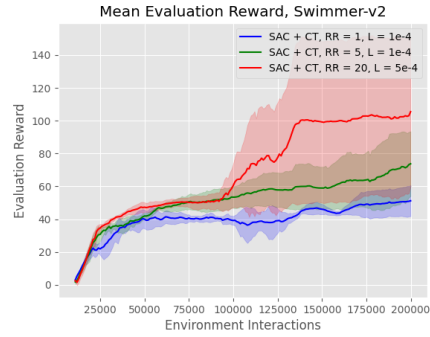
(a) Ablation study results with SAC + PER on the Hopper-v2 environment. This ablation study allows for selecting the optimal PER replay buffer to compare the other replay buffers to.



(b) Ablation study results with vanilla SAC on the Hopper-v2 environment. This ablation study allows for selecting the optimal vanilla replay buffer to compare the other replay buffers to.



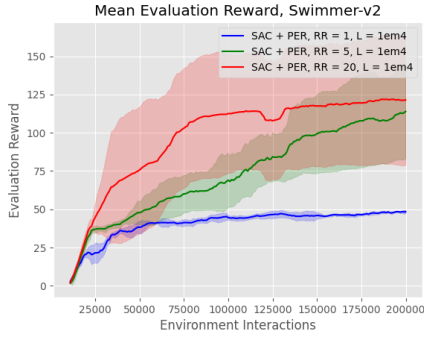
(a) Ablation study results with SAC + NMER on the Swimmer-v2 environment. This ablation study allows for selecting the optimal NMER replay buffer to compare the other replay buffers to.



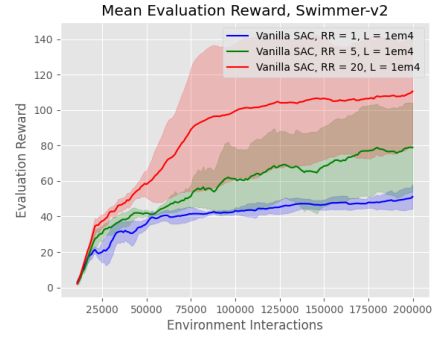
(b) Ablation study results with SAC + CT on the Swimmer-v2 environment. This ablation study allows for selecting the optimal CT replay buffer to compare the other replay buffers to.

Table 19: Evaluation reward for SAC and Walker2d-v2 outfitted with replay buffers with varying replay ratios.

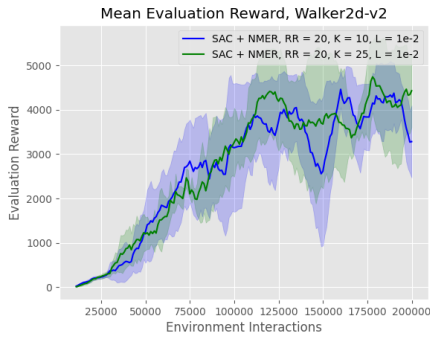
Replay Buffer	Reward (200K steps)
U, RR=1, L=1e-2	1481 \pm 546
U, RR=5, L=1e-2	4303 \pm 636
U, RR=20, L=1e-2	3295 \pm 901
PER, RR=1, L=1e-2 [13]	3466 \pm 784
PER, RR=5, L=1e-2 [13]	3376 \pm 611
PER, RR=20, L=1e-2 [13]	2513 \pm 728
CT, RR=1, L=1e-2 [8]	1982 \pm 920
CT, RR=5, L=1e-2 [8]	3319 \pm 898
CT, RR=20, L=1e-2 [8]	4696 \pm 1194
NMER, K=10, RR=20	3282 \pm 813
NMER, K=25, RR=20	4429 \pm 819



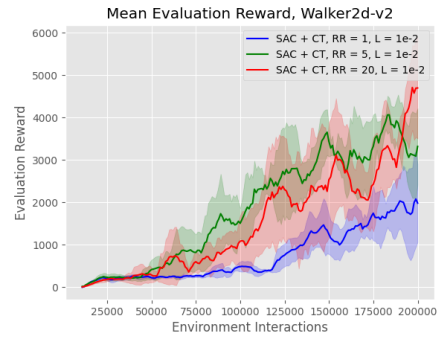
(a) Ablation study results with SAC + PER on the Swimmer-v2 environment. This ablation study allows for selecting the optimal PER replay buffer to compare the other replay buffers to.



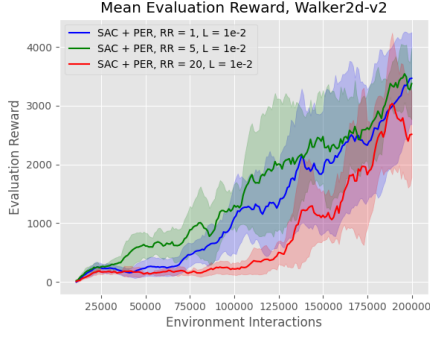
(b) Ablation study results with vanilla SAC on the Swimmer-v2 environment. This ablation study allows for selecting the optimal vanilla replay buffer to compare the other replay buffers to.



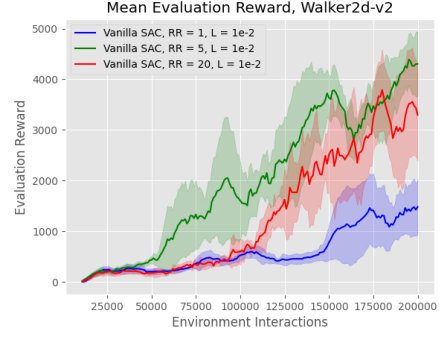
(a) Ablation study results with SAC + NMER on the Walker2d-v2 environment. This ablation study allows for selecting the optimal NMER replay buffer to compare the other replay buffers to.



(b) Ablation study results with SAC + CT on the Walker2d-v2 environment. This ablation study allows for selecting the optimal CT replay buffer to compare the other replay buffers to.



(a) Ablation study results with SAC + PER on the Walker2d-v2 environment. This ablation study allows for selecting the optimal PER replay buffer to compare the other replay buffers to.



(b) Ablation study results with vanilla SAC on the Walker2d-v2 environment. This ablation study allows for selecting the optimal vanilla replay buffer to compare the other replay buffers to.

K Optimal Replay Buffer Comparison, Additional Baselines

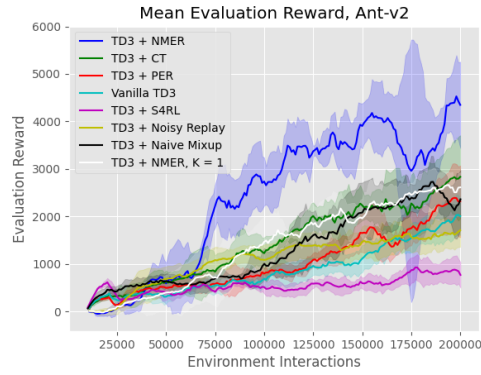


Figure 32: Best-performing replay buffers over Ant-v2 Gym environment. Note that the baselines that were not implemented in the manuscript were not included in these plots to avoid comparing too many curves simultaneously.

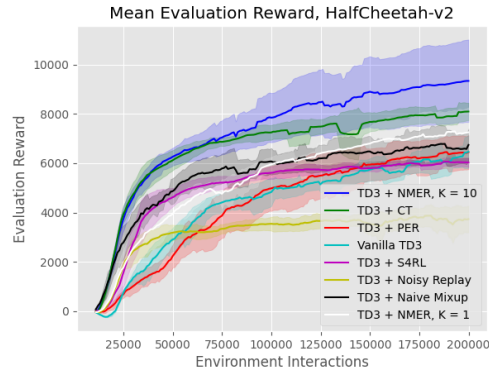


Figure 33: Best-performing replay buffers over HalfCheetah-v2 Gym environment. Note that the baselines that were not implemented in the manuscript were not included in these plots to avoid comparing too many curves simultaneously.

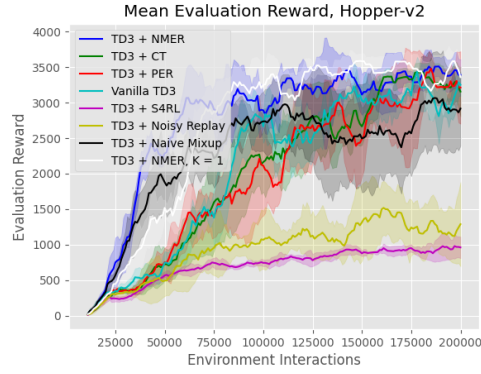


Figure 34: Best-performing replay buffers over Hopper-v2 Gym environment. Note that the baselines that were not implemented in the manuscript were not included in these plots to avoid comparing too many curves simultaneously.

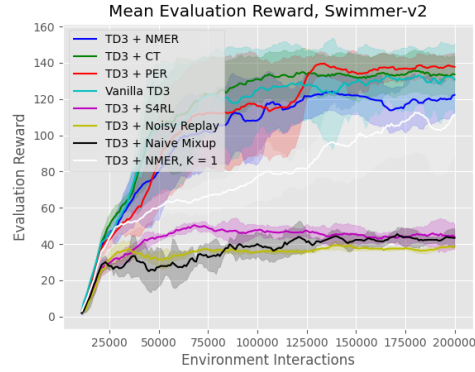


Figure 35: Best-performing replay buffers over Swimmer-v2 Gym environment. Note that the baselines that were not implemented in the manuscript were not included in these plots to avoid comparing too many curves simultaneously.

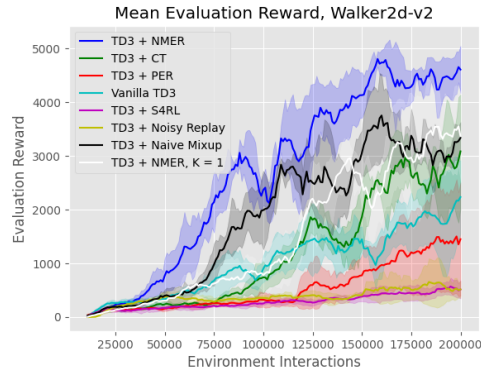


Figure 36: Best-performing replay buffers over Walker2d-v2 Gym environment. Note that the baselines that were not implemented in the manuscript were not included in these plots to avoid comparing too many curves simultaneously.

L DeepMind Control (DMC) Suite Experiments

In addition to evaluating these replay buffers on the OpenAI Gym environments, we also consider the state-based DeepMind Control (DMC) Suite environments.

L.1 Overview

The plots below visually depict ablation study results from running the evaluated replay buffers on the DeepMind Control Suite (DMC) environments. Please recall the following abbreviations for each replay buffer, used in the legends and captions of the plots below:

- **NMER**: Neighborhood Mixup Experience Replay
- **CT**: Continuous Transition
- **PER**: Prioritized Experience Replay
- **Vanilla TD3**: Uniform Experience Replay

Please find the best replay ratio variants for each type of replay buffer in the section immediately proceeding, as well as all the replay ratio ablation study results for each environment in the subsequent sections after.

Corresponding tabular results for these DMC experiments, with 200K environment interactions, can be found below.

DMC Environment	Vanilla (U)	PER	CT	NMER, K = 10
Walker-Walk	957 \pm 14	948 \pm 4	967 \pm 1	960 \pm 5
Walker-Run	680 \pm 63	689 \pm 71	748 \pm 27	802 \pm 18
Cheetah-Run	755 \pm 51	785 \pm 27	799 \pm 40	808 \pm 34
Quadruped-Run	477 \pm 118	571 \pm 68	818 \pm 75	689 \pm 151
Finger-Spin	933 \pm 43	962 \pm 33	965 \pm 11	917 \pm 30
Hopper-Hop	47 \pm 37	32 \pm 32	83 \pm 79	150 \pm 101
Hopper-Stand	724 \pm 233	349 \pm 192	724 \pm 282	340 \pm 281
Humanoid-Walk	58 \pm 98	29 \pm 46	71 \pm 76	91 \pm 110
Acrobot-Swingup	10 \pm 13	6 \pm 4	6 \pm 6	3 \pm 3
Pendulum-Swingup	492 \pm 329	399 \pm 297	506 \pm 200	439 \pm 326
Cartpole-Swingup	863 \pm 10	860 \pm 3	865 \pm 5	870 \pm 9

M Trained Agent Videos

Videos of trained agents can be accessed using the links in Table 20, as well as through [this provided link](#).

Table 20: Videos of agents trained TD3/SAC + NMER on OpenAI MuJoCo environments.

Environment	RL Agent	Link
Ant-v2	TD3	View
HalfCheetah-v2	TD3	View
Hopper-v2	TD3	View
Swimmer-v2	TD3	View
Walker2d-v2	TD3	View
Humanoid-v2	TD3	View
Reacher-v2	TD3	View
Pendulum-v0	TD3	View
InvertedPendulum-v2	TD3	View
InvertedDoublePendulum-v2	TD3	View
Ant-v2	SAC	View
HalfCheetah-v2	SAC	View
Hopper-v2	SAC	View
Swimmer-v2	SAC	View
Walker2d-v2	SAC	View

References

- [1] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba. Openai gym, 2016. [5](#)
- [2] S. Fujimoto, H. Hoof, and D. Meger. Addressing function approximation error in actor-critic methods. In *International Conference on Machine Learning*, pages 1587–1596. PMLR, 2018. [1](#), [2](#), [3](#)
- [3] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1861–1870. PMLR, 7 2018. [1](#), [2](#), [3](#)
- [4] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del R’io, M. Wiebe, P. Peterson, P. G’erard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, Sept. 2020. doi: 10.1038/s41586-020-2649-2. [5](#)
- [5] J. Johnson, M. Douze, and H. Jégou. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*, 2017. [1](#), [5](#)
- [6] E. Liang, R. Liaw, R. Nishihara, P. Moritz, R. Fox, K. Goldberg, J. Gonzalez, M. Jordan, and I. Stoica. Rllib: Abstractions for distributed reinforcement learning. In *International Conference on Machine Learning*, pages 3053–3062, 2018. [1](#), [2](#), [3](#), [4](#), [5](#), [6](#)
- [7] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra. Continuous control with deep reinforcement learning. In Y. Bengio and Y. LeCun, editors, *ICLR*, 2016. [1](#)
- [8] J. Lin, Z. Huang, K. Wang, X. Liang, W. Chen, and L. Lin. Continuous transition: Improving sample efficiency for continuous control problems via mixup. *arXiv preprint arXiv:2011.14487*, 2020. [2](#), [7](#), [8](#), [9](#), [10](#), [11](#), [12](#), [13](#), [14](#), [15](#), [16](#), [17](#), [18](#), [19](#), [20](#), [21](#)
- [9] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013. [1](#)
- [10] NVIDIA, P. Vingelmann, and F. H. Fitzek. Cuda, release: 10.2.89, 2020. [1](#), [5](#)
- [11] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani,

- S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems* 32. Curran Associates, Inc., 2019. 5
- [12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 5, 6
- [13] T. Schaul, J. Quan, I. Antonoglou, and D. Silver. Prioritized experience replay. In *ICLR (Poster)*, 2016. 2, 3, 4, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21
- [14] S. Sinha, A. Mandlekar, and A. Garg. S4RL: Surprisingly simple self-supervision for offline reinforcement learning in robotics. In *5th Annual Conference on Robot Learning*, 2021. 2
- [15] E. Todorov, T. Erez, and Y. Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033. IEEE, 2012. 5