A APPENDIX

Supplementary material to the paper: (Haider et al. 2023, Out-of-Distribution Detection for Reinforcement Learning Agents with Probabilistic Dynamics Models)

A.1 Explanation of Baselines

A.1.1 GAUSSIAN. A multivariate Gaussian distribution is fitted to the entire training data, i.e. the mean and variance/co-variance are calculated along each dimension. For inference, the negative log-likelihood under the fitted distribution serves as an anomaly score.

A.1.2 GMM. A mixture multivariate Gaussian distribution is fitted to the entire training data via Expectation–maximization. For inference, the negative log-likelihood under the fitted distributions serves as an anomaly score. For our experiments we 5 mixture components and 10 initialization performed best.

A.1.3 IFOREST. We directly followed the Isolation Forest algorithm from [**?**] with 100 random estimators. The anomaly score is the negative mean number of splittings required to isolate this point over all trees in the forest.

A.1.4 KNN. An index is built using the entire training data. The anomaly score is the distance to the k-th nearest neighbor in this index. In our experiments we set k=1, wich consistently performed best. The default version of KNN is only trained on single states. KNN+ calculates the distances between (s_t, a_t, s_{t+1})-pairs.

A.1.5 LSTM. A recurrent neural network with long short term memory units [**?**] is adopted to model the nonlinear temporal correlations between data instances. We fit this model on all training episodes. The mean-squared prediction-error of this model is used as anomaly score. In our experiments we used an LSTM with 8 recurrent layers, 32 hidden components, a dropout of 0.1, an additional fully connected layer with 128 neurons before the output layer and a learning rate of of 0.005. The default version is trained to predict s_{t+1} given s_t. LSTM+ receives the policy action as an additional input.

A.1.6 RIQN. We follow the Recurrent Implict Quantile Network approach introduced in [**?**] and use the suggested hyper parameters (64 neurons in the first fully connected layer, 64 memory in the GRU model, and 64 neurons in both the second and the third fully connected layers, skip connections and a learning rate of 0.001). The anomaly score is the average L1 distance between the predicted samples and the actual observation.

A.2 Further techniques to compute the anomaly score

1) Prediction-error-based, analytical

The likelihood of the observed state s_{t+1} (after applying a_t in the environment) under the predicted distribution gives a straightforward anomaly score:

$$p_\theta(s_{t+1}) = \mathcal{N}\left((s_{t+1}|\mu_\theta(s_t, a_t), \sigma_\theta(s_t, a_t)\right). \tag{7}$$

Alternatively, the observed state can be applied to a statistical hypothesis test, given the predictive distributions. In other words, we can calculate the probability of observing a sample at least as extreme as s_{t+1} as:

$$\begin{aligned} P_\theta(s_0 > |s_{t+1}|) &= \int_{-\infty}^{-|s_{t+1}|} (s_{t+1}|\mu_\theta, \sigma_\theta) \\ &\quad + \int_{-|s_{t+1}|}^{\infty} (s_{t+1}|\mu_\theta, \sigma_\theta) \\ &= 2\mathrm{cdf}_{\mathcal{N}}\Big(- |s_{t+1}|\Big|\mu_\theta, \sigma_\theta\Big). \end{aligned} \tag{8}$$

The final anomaly score $h(\cdot)$ is obtained by aggregating over the scores from all ensemble member via simple statistics (e.g. mean, max, min):

$$h(s_t, a_t, s_{t+1}) = \text{aggr.}\left(\{p_\theta(s_{t+1})\}^B\right). \tag{9}$$

2) Variance-only based A different approach to establishing OOD scores leverages the epistemic uncertainty of dynamics models. That is, we quantify model uncertainty as the variance over all individual predictions and use this as the final anomaly score:

$$\begin{aligned} h(s_t, a_t) &= \mathrm{Var}\left(\{s'_{t+1}\}^{K,B}\right) \\ &= \frac{1}{(K \cdot B) - 1} \sum_{i=1}^{K \cdot B} \left(s'_{t+1,i} - \mu'_{t+1,i}\right)^2, \end{aligned} \tag{10}$$

where μ' is the mean of all predicted particles.

A.3 Why sampling works better

The sampling-based technique consistently outperforms the two analytical approaches. This can be explained by the mismatch between likelihood and probability mass in high dimensional multivariate distributions. Samples from high-dimensional distributions on average have a non-zero distance to the distribution's mode due to the exponential growth in differential volume. The further away a sample lies from the distribution's mode however, the lower is it's likelihood (by definition). These two counteracting effects form the typical set, whose distance to the mean is higher as dimensionality is increased. Samples within this set are thus perfectly ID, even though they have a much lower density than the mode. For a visualization of this see the bottom graphs in figure 5. Samples with a norm ≈ 5.5 are well within the typical set (i.e. they are ID) but have a very low density (close to zero). Now consider two samples s_a and s_b. Let s_a be relatively close to the typical set and s_b be far away from the mean and far away from the typical set s.t. $d(s_a, \mu) < d(s_b, \mu)$. Therefore $p(s_a) > p(s_b)$. However, both $p(s_a)$ and $p(s_b)$ are very low, since they are not close the the mean, and thus $p(s_a) \approx p(s_b)$. Since only the order matters, this itself is not a problem. However, the distribution is only an estimate and it is subject to prediction errors, even on ID samples. These prediction errors directly amplify the density estimates. Therefore the likelihood estimate of an ID sample under one prediction can have a lower density estimate than OOD sample under another prediction, simply due to a tiny prediction error. This makes it difficult to draw a strict line between ID and OOD. The sampling based approach on the other hand is nothing else than an empirical solution of calculating the distance to the typical set. The distance to the typical set of an ID sample is by definition much smaller than of an OOD sample. This order doesn't change, even when smaller prediction errors are present.

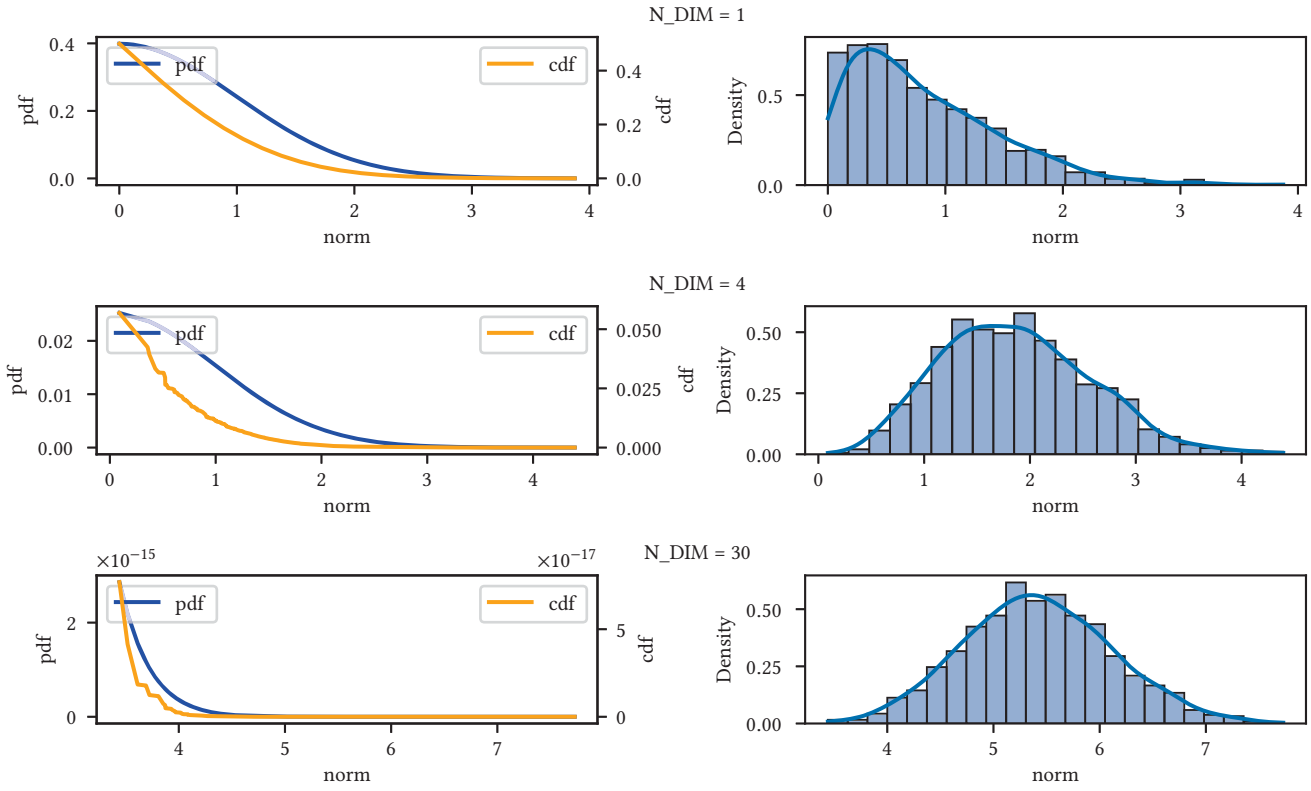

Figure 5: Probability-density-function(pdf) and cumulative distribution function (cdf) over samples from a multivariate Gaussian distribution sorted after their norm (left) and frequency of samples with the respective norm (right). The pdf and cdf concentrate around the mode and decrease with the distance from the mean. The frequency of samples drawn from this distribution however concentrates in a neighborhood called the typical set with a non-zero norm. As dimensionality increases, this typical set moves further and further away from norm 0.

Therefore, the sampling based approach is more robust to small prediction errors and scales better to higher dimensions.

A.4 Details on model training

PEDM In all our experiments we used $B = 5$ ensemble members for our approach. Each ensemble member has three fully connected layers, 500 neurons per layer (except four fully connected layers with 200 neurons for HalfCheetah), and swish activation functions. We train all models with a learning rate of 0.001 for 2000 epochs on a dataset from 45 episodes.

TD3 We use an implementation from [**?**] for the TD3 algorithm [13] with default hyper-parameters buffer_size: 1e6, learning_starts: 100, batch_size: 100, tau: 0.005, gamma: 0.99, noise_type: 'normal' noise_std: 0.1, target_policy_noise: 0.2, target_noise_clip: 0.5)

A.5 Detailed Environment Parameters

We consider continuous robotic control tasks in the mujoco physics engine [38] as basis for our evaluation environments. Specifically, we modify *CartPole*, *HalfCheetah*, *Reacher* and *Pusher* from [9] with the perturbation parameters provided in table 6.

Table 6: Detailed environment parameters for minor and severely disturbed environments form the mujoco physics engine.

	Cartpole	HalfCheetah	Pusher	Reacher
bm_factor_minor	1.01	1.01	1.01	1.01
act_factor_minor	1.01	1.01	1.01	1.01
act_noise_minor	0.01	0.01	0.01	0.01
act_offset_minor	0.01	0.01	0.01	0.01
force_vector_minor	[-0.1,0,0,0,0,0]	[-0.1,0,0,0,0,0]	[-0.5,0,0,0,0,0]	[0,0,-0.5,0,0,0]
bm_factor_severe	1.5	1.5	2	1.5
act_factor_severe	1.5	1.5	2	1.5
act_offset_severe	0.2	0.2	0.3	0.5
act_noise_severe	0.2	0.2	0.3	1
force_vector_severe	[-10, 0, 0, 0, 0, 0]	[-15, 0, 0, 0, 0, 0]	[-1.5, 0, 0, 0, 0, 0]	[0, 0, -5, 0, 0, 0]

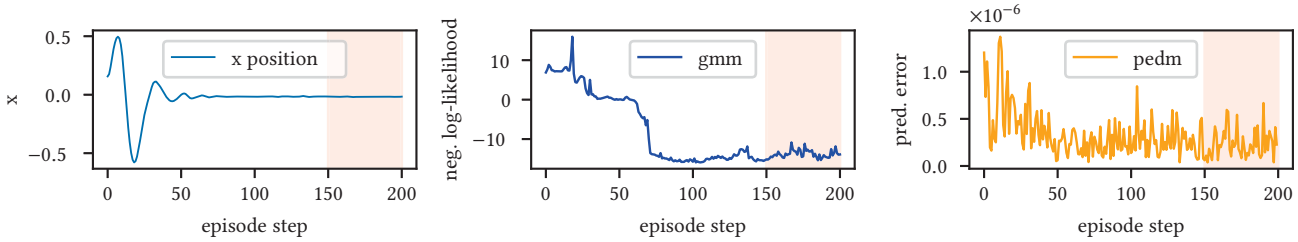

Figure 6: Comparison of anomaly scores from GMM and PEDM on an episode of Carpole. Anomaly occurs after 150 steps (shaded in red). The position of the cart oscillates heavily in the beginning of the episode but stays roughly constant after ≈ 50 time-steps (left). The inital phase has low likelihood under the GMM compared to the stable phase, as visibly by the drop in negative log-likelihood (center). The PEDM also has a higher prediction error in the beginning of the episode but the peak in the beginning is comparably smaller. As a result the PEDM has more similar anomaly scores throughout the episode, resulting in an AUC of ≈ 0.5

mod	reward	env_id	GAUSSIAN			GMM			ISOFOREST			KNN			KNN+			LSTM			LSTM+			RIQN			PEDM		
			auroc	ap	fpr95	auroc	ap	fpr95	auroc	ap	fpr95	auroc	ap	fpr95	auroc	ap	fpr95	auroc	ap	fpr95	auroc	ap	fpr95	auroc	ap	fpr95	auroc	ap	fpr95
bm_factor_minor	182.38	MJCartpole-v0	0,25	0,38	0,96	0,23	0,38	0,98	0,23	0,37	0,98	0,27	0,36	0,98	0,24	0,36	0,98	0,34	0,38	0,98	0,30	0,33	0,98	0,46	0,43	0,88	0,43	0,45	0,96
act_factor_minor	182.38	MJCartpole-v0	0,27	0,40	0,96	0,21	0,34	0,98	0,24	0,34	0,98	0,26	0,37	0,98	0,25	0,34	0,98	0,35	0,33	0,97	0,29	0,38	0,98	0,47	0,41	0,88	0,42	0,47	0,96
act_offset_minor	182.38	MJCartpole-v0	0,27	0,33	0,96	0,21	0,34	0,98	0,22	0,34	0,98	0,23	0,33	0,98	0,31	0,38	0,84	0,33	0,41	0,98	0,31	0,34	0,98	0,46	0,42	0,89	0,80	0,75	0,77
act_noise_minor	182.38	MJCartpole-v0	0,29	0,32	0,96	0,23	0,33	0,98	0,22	0,35	0,98	0,22	0,37	0,98	0,31	0,34	0,93	0,35	0,41	0,98	0,32	0,37	0,98	0,45	0,46	0,88	0,67	0,69	0,91
force_vector_minor	182.37	MJCartpole-v0	0,28	0,36	0,98	0,37	0,40	0,78	0,25	0,39	0,97	0,38	0,39	0,79	0,35	0,37	0,87	0,36	0,37	0,98	0,29	0,37	0,98	0,48	0,43	0,88	0,42	0,43	0,96
bm_factor_severe	182.16	MJCartpole-v0	0,26	0,37	0,93	0,22	0,34	1,00	0,18	0,34	1,00	0,25	0,36	0,98	0,36	0,42	0,90	0,34	0,39	0,98	0,32	0,38	0,98	0,48	0,41	0,87	0,89	0,92	0,75
act_factor_severe	182.18	MJCartpole-v0	0,26	0,37	0,96	0,28	0,41	0,98	0,22	0,38	0,98	0,27	0,36	0,97	0,46	0,46	0,81	0,56	0,56	0,95	0,57	0,59	0,94	0,48	0,41	0,87	0,94	0,96	0,51
act_offset_severe	182.32	MJCartpole-v0	0,26	0,35	0,98	0,43	0,42	0,60	0,29	0,42	0,92	0,48	0,39	0,58	0,71	0,59	0,33	0,37	0,36	0,97	0,55	0,46	0,85	0,50	0,40	0,84	1,00	1,00	0,00
act_noise_severe	182.23	MJCartpole-v0	0,41	0,41	0,99	0,48	0,44	0,63	0,47	0,43	0,71	0,48	0,48	0,68	0,73	0,59	0,51	0,74	0,70	0,79	0,84	0,83	0,56	0,47	0,38	0,89	0,98	0,99	0,05
force_vector_severe	165.91	MJCartpole-v0	0,77	0,58	0,29	0,98	0,98	0,22	0,94	0,89	0,27	0,99	0,99	0,08	1,00	1,00	0,01	0,76	0,60	0,45	0,94	0,86	0,12	0,80	0,65	0,51	1,00	1,00	0,00
avgs_minor	182.38	MJCartpole-v0	0,27	0,36	0,96	0,25	0,36	0,94	0,23	0,36	0,98	0,27	0,36	0,94	0,29	0,36	0,92	0,35	0,38	0,98	0,30	0,36	0,98	0,46	0,43	0,88	0,55	0,56	0,91
avgs_severe	178.96	MJCartpole-v0	0,39	0,42	0,83	0,48	0,52	0,69	0,42	0,49	0,78	0,49	0,52	0,66	0,65	0,61	0,51	0,55	0,52	0,83	0,64	0,62	0,69	0,55	0,45	0,80	0,96	0,97	0,26
bm_factor_minor	7342.09	MJHalfCheetah-v0	0,47	0,50	0,95	0,47	0,51	0,95	0,46	0,47	0,96	0,50	0,46	0,91	0,50	0,49	0,91	0,49	0,45	0,95	0,50	0,49	0,94	0,47	0,43	0,95	0,51	0,50	0,95
act_factor_minor	7387.1	MJHalfCheetah-v0	0,48	0,47	0,95	0,47	0,48	0,95	0,46	0,48	0,95	0,49	0,49	0,91	0,50	0,48	0,92	0,49	0,52	0,94	0,50	0,49	0,94	0,47	0,45	0,95	0,51	0,49	0,93
act_offset_minor	7354.2	MJHalfCheetah-v0	0,48	0,45	0,95	0,48	0,46	0,95	0,47	0,45	0,95	0,52	0,46	0,89	0,52	0,49	0,88	0,49	0,50	0,94	0,50	0,47	0,94	0,47	0,51	0,95	0,60	0,52	0,69
act_noise_minor	7339.28	MJHalfCheetah-v0	0,47	0,49	0,95	0,47	0,50	0,96	0,47	0,40	0,95	0,51	0,49	0,90	0,50	0,46	0,90	0,48	0,48	0,94	0,49	0,52	0,94	0,47	0,42	0,95	0,60	0,49	0,77
force_vector_minor	7358.66	MJHalfCheetah-v0	0,48	0,47	0,95	0,47	0,45	0,95	0,47	0,42	0,95	0,50	0,49	0,91	0,50	0,53	0,91	0,48	0,53	0,95	0,49	0,46	0,94	0,47	0,44	0,95	0,49	0,48	0,95
bm_factor_severe	5842.86	MJHalfCheetah-v0	0,75	0,71	0,67	0,83	0,76	0,52	0,85	0,79	0,53	0,97	0,95	0,08	0,97	0,95	0,07	0,92	0,89	0,25	0,93	0,89	0,24	0,75	0,74	0,64	0,97	0,97	0,06
act_factor_severe	6632.72	MJHalfCheetah-v0	0,65	0,63	0,88	0,65	0,64	0,85	0,66	0,63	0,88	0,84	0,82	0,51	0,86	0,84	0,46	0,77	0,76	0,69	0,77	0,77	0,71	0,55	0,56	0,91	0,84	0,85	0,68
act_offset_severe	5961.68	MJHalfCheetah-v0	0,77	0,76	0,71	0,85	0,85	0,52	0,84	0,83	0,63	0,97	0,96	0,09	0,97	0,96	0,09	0,90	0,87	0,35	0,91	0,90	0,37	0,74	0,76	0,76	0,98	0,99	0,01
act_noise_severe	6307.66	MJHalfCheetah-v0	0,71	0,68	0,77	0,80	0,78	0,61	0,81	0,80	0,63	0,96	0,94	0,12	0,96	0,94	0,12	0,88	0,85	0,36	0,88	0,82	0,39	0,66	0,56	0,78	0,98	0,97	0,08
force_vector_severe	6465.93	MJHalfCheetah-v0	0,58	0,58	0,95	0,60	0,60	0,90	0,64	0,66	0,86	0,88	0,82	0,37	0,88	0,84	0,33	0,74	0,69	0,73	0,73	0,72	0,74	0,64	0,67	0,84	0,90	0,84	0,18
avgs_minor	7356.27	MJHalfCheetah-v0	0,48	0,48	0,95	0,47	0,48	0,95	0,47	0,44	0,95	0,50	0,48	0,90	0,50	0,49	0,90	0,49	0,50	0,94	0,50	0,49	0,94	0,47	0,45	0,95	0,54	0,50	0,86
avgs_severe	6242.17	MJHalfCheetah-v0	0,69	0,67	0,80	0,75	0,73	0,68	0,76	0,74	0,71	0,92	0,90	0,23	0,93	0,91	0,21	0,84	0,81	0,48	0,84	0,82	0,49	0,67	0,66	0,79	0,93	0,92	0,20
bm_factor_minor	-58.66	MJPusher-v0	0,19	0,30	0,99	0,25	0,38	0,98	0,22	0,37	0,99	0,36	0,43	0,98	0,36	0,43	0,99	0,24	0,36	1,00	0,26	0,41	0,99	0,22	0,36	0,99	0,31	0,41	0,99
act_factor_minor	-58.69	MJPusher-v0	0,20	0,33	0,99	0,24	0,37	0,98	0,22	0,34	0,99	0,34	0,41	0,99	0,29	0,33	0,99	0,26	0,37	0,99	0,24	0,39	0,99	0,25	0,37	0,99	0,31	0,40	0,99
act_offset_minor	-58.5	MJPusher-v0	0,19	0,36	1,00	0,22	0,37	0,96	0,22	0,32	0,99	0,39	0,44	0,90	0,42	0,46	0,90	0,23	0,34	0,99	0,22	0,39	1,00	0,21	0,37	0,99	0,74	0,58	0,38
act_noise_minor	-58.66	MJPusher-v0	0,23	0,28	0,99	0,25	0,36	0,97	0,23	0,34	0,99	0,34	0,44	0,98	0,33	0,37	0,98	0,31	0,35	0,96	0,28	0,41	0,97	0,25	0,32	0,99	0,68	0,57	0,66
force_vector_minor	-57.07	MJPusher-v0	0,32	0,36	0,87	0,63	0,61	0,80	0,34	0,43	0,88	0,75	0,62	0,58	0,78	0,68	0,44	0,43	0,40	0,93	0,33	0,39	0,99	0,31	0,34	0,93	0,87	0,74	0,20
bm_factor_severe	-58.53	MJPusher-v0	0,24	0,38	0,98	0,35	0,39	0,95	0,30	0,39	0,98	0,56	0,55	0,91	0,49	0,50	0,95	0,28	0,44	1,00	0,33	0,42	0,98	0,28	0,36	0,98	0,56	0,50	0,92
act_factor_severe	-69.16	MJPusher-v0	0,32	0,40	1,00	0,40	0,51	0,96	0,34	0,44	0,97	0,58	0,57	0,92	0,61	0,67	0,92	0,54	0,62	0,99	0,53	0,64	0,99	0,31	0,36	0,99	0,77	0,78	0,92
act_offset_severe	-60.82	MJPusher-v0	0,94	0,93	0,22	0,98	0,99	0,03	0,89	0,88	0,44	0,99	0,99	0,03	1,00	0,99	0,01	0,56	0,54	0,91	0,85	0,78	0,48	0,56	0,40	0,72	0,99	0,96	0,03
act_noise_severe	-65.46	MJPusher-v0	0,55	0,58	0,82	0,89	0,91	0,52	0,61	0,62	0,81	0,90	0,88	0,35	0,97	0,97	0,11	0,97	0,96	0,08	0,97	0,94	0,09	0,41	0,38	0,90	0,98	0,97	0,04
force_vector_severe	-63.16	MJPusher-v0	0,85	0,87	0,62	0,95	0,96	0,35	0,73	0,64	0,72	0,96	0,97	0,22	0,98	0,98	0,12	0,58	0,53	0,85	0,74	0,61	0,54	0,54	0,47	0,77	0,96	0,91	0,07
avgs_minor	-58.32	MJPusher-v0	0,23	0,33	0,97	0,32	0,42	0,94	0,25	0,36	0,97	0,44	0,47	0,89	0,44	0,45	0,86	0,29	0,36	0,97	0,27	0,40	0,99	0,25	0,35	0,98	0,58	0,54	0,64
avgs_severe	-63.43	MJPusher-v0	0,58	0,63	0,73	0,71	0,75	0,56	0,57	0,59	0,78	0,80	0,79	0,49	0,81	0,82	0,42	0,59	0,62	0,77	0,68	0,68	0,62	0,42	0,39	0,87	0,85	0,82	0,40
bm_factor_minor	-44.74	MJReacher-v0	0,42	0,47	0,95	0,37	0,41	0,99	0,37	0,43	0,95	0,62	0,57	0,85	0,59	0,56	0,94	0,29	0,34	0,98	0,32	0,38	0,99	0,38	0,35	0,97	0,35	0,39	0,99
act_factor_minor	-44.79	MJReacher-v0	0,45	0,48	0,96	0,40	0,50	0,98	0,37	0,48	0,99	0,56	0,51	0,92	0,55	0,57	0,91	0,35	0,43	0,98	0,31	0,40	0,99	0,38	0,42	0,94	0,38	0,39	0,97
act_offset_minor	-44.68	MJReacher-v0	0,46	0,47	0,96	0,46	0,50	0,99	0,41	0,45	0,95	0,61	0,61	0,97	0,55	0,53	0,92	0,29	0,37	0,98	0,29	0,33	0,99	0,42	0,41	0,93	0,41	0,44	0,92
act_noise_minor	-44.74	MJReacher-v0	0,44	0,47	0,98	0,42	0,50	0,98	0,40	0,49	0,93	0,58	0,54	0,95	0,50	0,50	0,96	0,33	0,42	0,97	0,30	0,46	0,98	0,40	0,42	0,96	0,44	0,48	0,94
force_vector_minor	-45.03	MJReacher-v0	0,39	0,44	0,97	0,44	0,55	0,99	0,40	0,44	0,96	0,63	0,66	0,90	0,57	0,60	0,93	0,30	0,41	0,98	0,33	0,43	0,97	0,44	0,42	0,92	0,77	0,62	0,39
bm_factor_severe	-44.81	MJReacher-v0	0,40	0,44	0,94	0,50	0,52	0,98	0,38	0,44	0,93	0,63	0,60	0,88	0,62	0,56	0,89	0,32	0,41	0,99	0,31	0,40	0,99	0,40	0,37	0,96	0,59	0,53	0,93
act_factor_severe	-47.57	MJReacher-v0	0,49	0,52	0,92	0,63	0,65	0,91	0,47	0,50	0,93	0,74	0,73	0,75	0,69	0,68	0,78	0,50	0,51	0,95	0,48	0,53	0,96	0,42	0,47	0,95	0,95	0,88	0,13
act_offset_severe	-48.66	MJReacher-v0	0,58	0,57	0,86	0,77	0,80	0,77	0,51	0,56	0,83	0,87	0,88	0,49	0,86	0,82	0,40	0,48	0,49	0,94	0,62	0,58	0,83	0,44	0,46	0,91	0,98	0,93	0,02
act_noise_severe	-50.98	MJReacher-v0	0,43	0,49	0,98	0,61	0,67	0,91	0,44	0,53	0,91	0,78	0,76	0,63	0,76	0,69	0,64	0,95	0,91	0,15	0,95	0,91	0,15	0,44	0,40	0,92	0,99	0,97	0,02
force_vector_severe	-47.24	MJReacher-v0	0,58	0,55	0,88	0,78	0,80	0,79	0,64	0,64	0,82	0,89	0,90	0,52	0,85	0,86	0,50	0,52	0,47	0,83	0,49	0,48	0,84	0,58	0,54	0,87	0,98	0,93	0,03
avgs_minor	-44.8	MJReacher-v0	0,43	0,47	0,96	0,42	0,49	0,99	0,39	0,46	0,96	0,60	0,58	0,92	0,55	0,55	0,93	0,31	0,39	0,98	0,31	0,40	0,98	0,40	0,40	0,94	0,47	0,46	0,84
avgs_severe	-47.85	MJReacher-v0	0,50	0,51	0,92	0,66	0,69	0,87	0,49	0,53	0,88	0,78	0,77	0,65	0,76	0,72	0,64	0,55	0,56	0,77	0,57	0,58	0,75	0,46	0,45	0,92	0,90	0,85	0,23

Table 7: Results for all detectors, environments and disturbances