Experiments with REINFORCE algorithm

Rajesh Siraskar

14-Jun-2023

Agenda

Objectives

Research goal*: An optimal predictive maintenance policy for replacement of milling tool

1. Experiment with the very *basic, naïve*, REINFORCE algorithm

2. Implemented from "scratch"

3. Compare against industry gold-standard Stable-Baselines-3 implementation of DQN, A2C and PPO

4. REINFORCE trained for 800 episodes, with an inner loop of 300. SB-3 trained with 10,000 episodes

5. This is a purely experimental (empirical) exercise.

* Note: My "proposal" goal is "Deep RL for predicting RUL", this is a sub/related goal

RL environment

Two different sources – simulated and real milling data

$$VB = a \cdot t^{b_1} \cdot e^{b_2 \cdot t} \Big|_{t=0}^{t=t_k}$$

◄ Simulated. Dašić, Predrag (2006): Analysis of wear cutting tools by complex power exponential function for finishing turning of the hardened steel 20CrMo5 by mixed ceramic tools.

▼ Real data. IEEE – PHM Society

RL environment - Variants

Three environments and their variants:

1. Simulated. Based on Dašić (2006). Simple single-variate state (tool wear)

   - Variants: (1) No noise (2) Low noise and low break-down chance and (3) High noise and high break-down chance

2. PHM 2010 real data – Simple single-variate state (tool wear)

   - Variants: C-01, C-04 and C-06 data-sets

   - Variants: (1) No noise (2) Low noise and low break-down chance and (3) High noise and high break-down chance

3. PHM 2010 real data – Complex multivariate state (tool wear, 3-axis forces, 3-axis vibration and acoustic data)

   - Variants: C-01, C-04 and C-06 data-sets
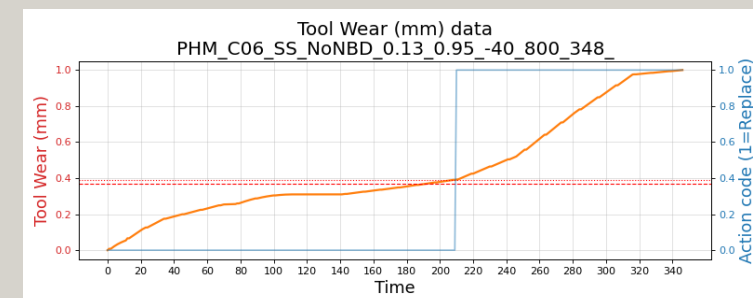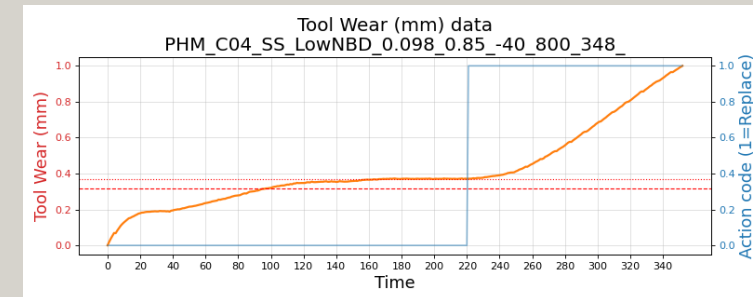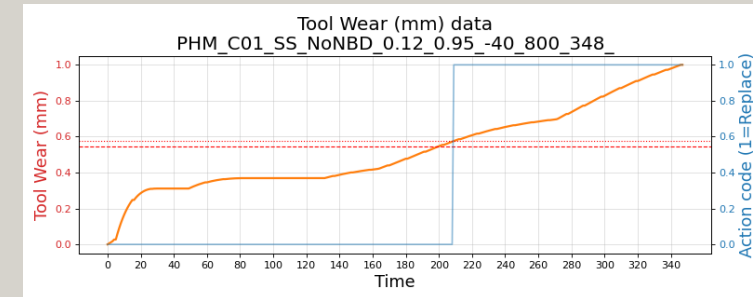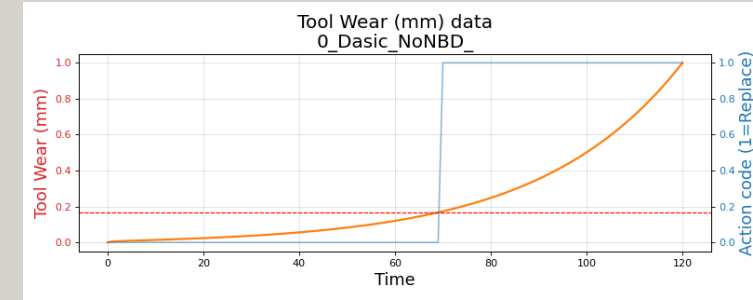
Wear plot – real data and its variants
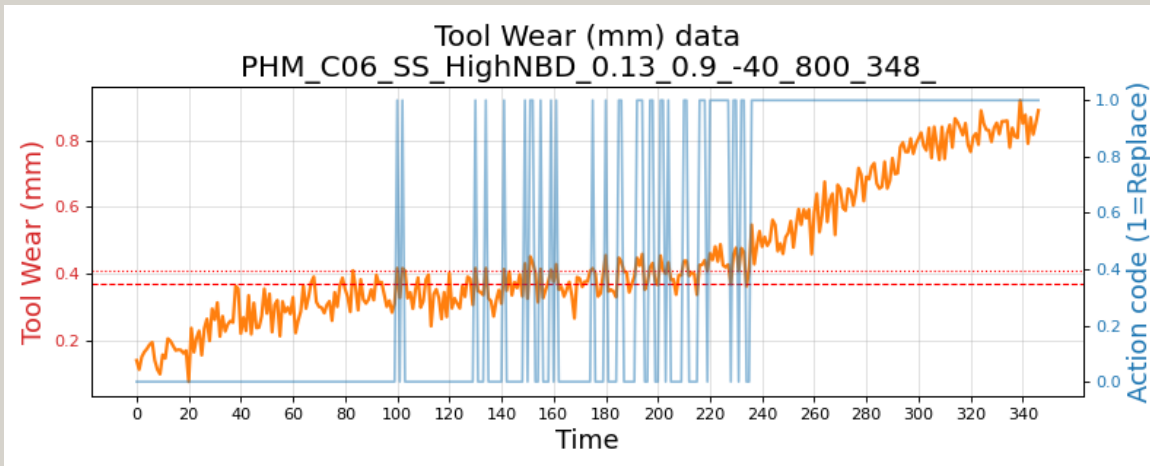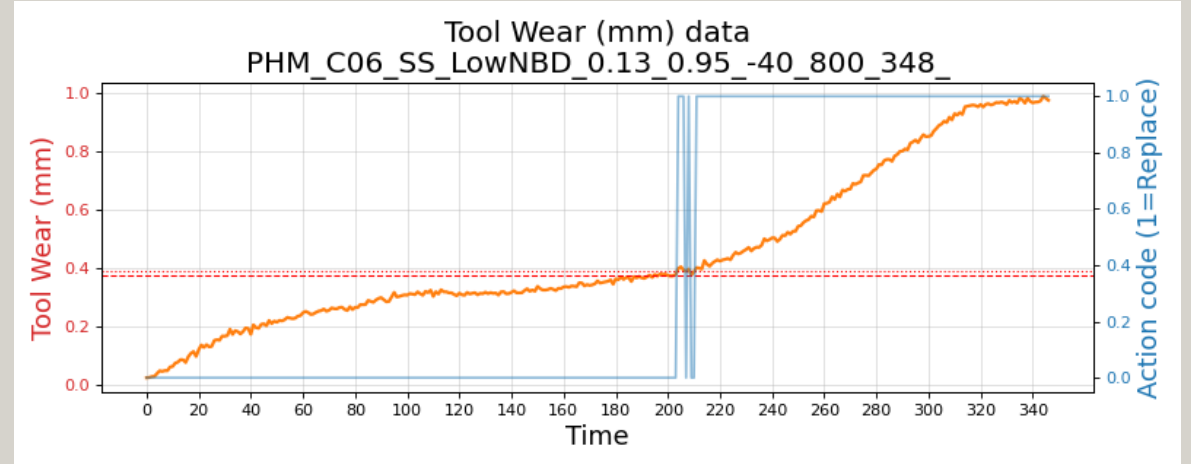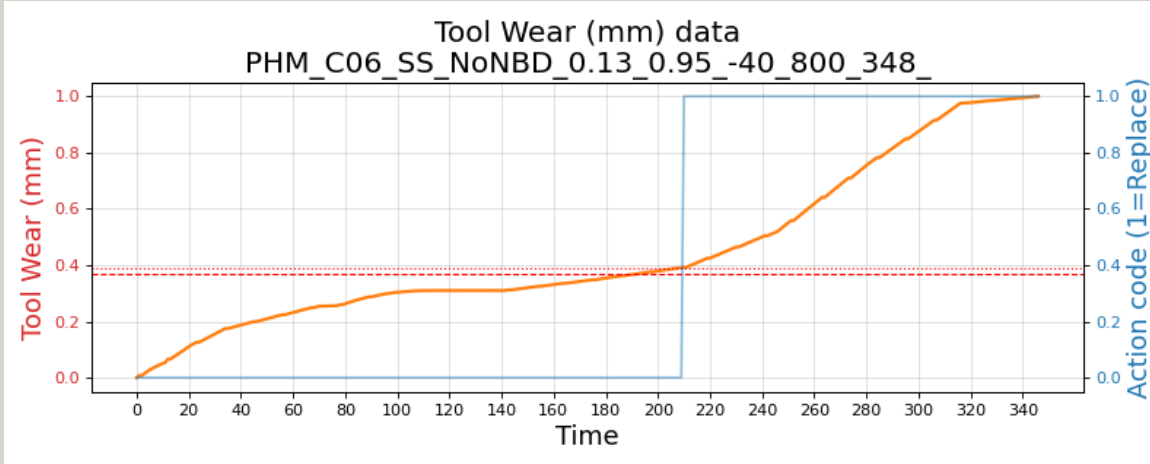
PHM 2010 C-06 data-set

- No noise, no break-down chance

- Low noise: 1e-3 and break-down chance 5%

- High noise: 1e-2 and break-down chance 10%

Evaluation strategy

- RL policy decides when to replace tool

- Compare against "human" replacement action

- Test from randomly sampled 40 points, 5 rounds

- Compute Precision, Recall and F1

- Compare mean and std. deviations

Findings

1. The naïve implementation of REINFORCE algorithm was implemented in PyTorch with an extremely simple architecture: Three layers and ReLU activation and an Adam optimizer.

2. In training, the SB-3 algorithms were extremely stable, but the F1 performance was always between 0.38-0.67. The naïve REINFORCE, on the other-hand, is unstable in training.

3. We saved models that performed better than 0.7 and these cases, the REINFORCE performs surprisingly better than DQN, A2C and PPO in all cases

4. The SB-3 algorithms almost always perform at about F1 0.47-0.50

5. REINFORCE performed between 0.72 (for complex real data environment) to 0.89 (for simulated environment)

6. Across precision, recall and F1; REINFORCE was better than the *best* performing SB-3 algorithm by 0.252 basis-points. The difference[1] in variance (σ^2) was also *lower*, though near negligible, for REINFORCE at -0.0003

7. Across precision on tool replacement, REINFORCE was better by 0.354, a lower variance of -0.004

1. w.r.t. lowest of SB-3 algorithms
2. Reference data-sheet: "1. Analysis_model_performance_summary_V3 (PPT ref.).xls"

Results: Overall – all environments and their variants

Algorithm	Precision		Recall		F1-score	
	Mean	Std.Dev	Mean	Std.Dev	Mean	Std.Dev
A2C	0.431	0.090	0.397	0.065	0.383	0.059
DQN	0.475	0.175	0.657	0.030	0.471	0.029
PPO	0.504	0.121	0.391	0.066	0.402	0.070
REINFORCE	**0.861**	0.039	**0.843**	0.051	**0.836**	0.041

Results: Simple single variate state. Including noise variants

Simulated – Dašić, Predrag (2006).

Algorithm	Precision		Recall		F1-score	
	Mean	Std.Dev	Mean	Std.Dev	Mean	Std.Dev
A2C	0.292	0.041	0.083	0.024	0.128	0.030
DQN	0.517	0.012	0.963	0.032	0.673	0.015
PPO	0.478	0.244	0.093	0.048	0.150	0.074
REINFORCE	**0.928**	0.032	**0.883**	0.041	**0.895**	0.039

PHM 2010: Single-variate environment, across three data sets C-01, C-04 and C-06

Algorithm	Precision		Recall		F1-score	
	Mean	Std.Dev	Mean	Std.Dev	Mean	Std.Dev
A2C	0.459	0.115	0.422	0.077	0.417	0.067
DQN	0.486	0.236	0.559	0.029	0.413	0.035
PPO	0.502	0.094	0.479	0.071	0.467	0.073
REINFORCE	**0.845**	0.040	**0.892**	0.051	**0.855**	0.038

Results: Complex, multi-variate environment

PHM 2010: Complex, multi-variate environment, across three data sets C-01, C-04 and C-06
No noise or break-down

Algorithm	Precision		Recall		F1-score	
	Mean	Std.Dev	Mean	Std.Dev	Mean	Std.Dev
A2C	0.485	0.067	0.637	0.072	0.534	0.065
DQN	0.398	0.156	0.647	0.031	0.444	0.025
PPO	0.537	0.080	0.423	0.068	0.458	0.055
REINFORCE	**0.842**	0.040	**0.657**	0.062	**0.721**	0.052

Results Summary (13-Jun-2023 run)

Model	REINFORCE			SB-3 A2C			SB-3 DQN			SB-3 PPO		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Simulated Dasic 2006 - No noise	1.00	0.74	0.85	0.51	0.62	0.56	0.30	0.03	0.05	0.57	0.43	0.49
Simulated Dasic 2006 - Low NBD	0.93	0.99	0.96	0.00	0.00	0.00	0.51	0.99	0.67	0.72	0.46	0.56
Simulated Dasic 2006 - High NBD	0.90	0.99	0.94	0.70	0.82	0.76	0.30	0.03	0.05	0.59	0.31	0.40
PHM C01 simple - No noise	0.87	0.98	0.92	0.47	0.26	0.33	0.40	0.03	0.06	0.27	0.13	0.17
PHM C01 simple - Low NBD	0.91	0.85	0.88	0.00	0.00	0.00	0.40	0.03	0.06	0.50	0.17	0.25
PHM C01 simple - High NBD	0.80	0.99	0.88	0.50	1.00	0.67	0.52	0.99	0.68	0.52	0.53	0.52
PHM C04 simple - No noise	0.90	0.93	0.91	0.53	0.62	0.57	0.05	0.01	0.02	0.37	0.16	0.22
PHM C04 simple - Low NBD	0.67	0.99	0.80	0.59	0.68	0.63	0.10	0.01	0.02	0.27	0.02	0.04
PHM C04 simple - High NBD	0.69	0.82	0.75	0.00	0.00	0.00	0.51	1.00	0.68	0.46	0.26	0.33
PHM C06 simple - No noise	1.00	0.61	0.76	0.50	1.00	0.67	0.51	0.99	0.67	0.30	0.05	0.08
PHM C06 simple - Low NBD	0.99	0.85	0.91	0.42	0.38	0.40	0.50	0.99	0.67	0.38	0.23	0.28
PHM C06 simple - High NBD	0.78	0.90	0.83	0.40	0.29	0.33	0.50	0.96	0.66	0.27	0.05	0.08
PHM C01 multi-variate state	0.83	0.94	0.88	0.49	0.64	0.55	0.50	0.98	0.66	0.58	0.37	0.44
PHM C04 multi-variate state	0.78	0.68	0.72	0.48	0.64	0.55	0.60	0.03	0.06	0.61	0.29	0.39
PHM C06 multi-variate state	1.00	0.58	0.73	0.50	1.00	0.67	0.50	1.00	0.67	0.42	0.38	0.40

Training time: Avg. over different variants

Average training time in secs.

| Variant	A2C	DQN	PPO	REINFORCE
Dasic 2006 Simulated - Single variate state	33.53	3.04	34.23	182.76
PHM-2010 Real data - Single variate state	21.53	2.37	24.27	313.99
PHM-2010 Real data - Multi-variate state	37.23	6.06	42.46	632.05
Overall average	**27.07**	**3.24**	**29.90**	**351.36**

Average training time per 1,000 time-steps, in secs. Stable-baseline algorithms are trained for 10,000 time-steps, while REINFORCE uses 96,800 for simulated and 278,400 for PHM based environments

Variant	A2C	DQN	PPO	REINFORCE
Dasic 2006 Simulated - Single variate state	3.35	0.30	3.42	1.89
PHM-2010 Real data - Single variate state	2.15	0.24	2.43	1.13
PHM-2010 Real data - Multi-variate state	3.72	0.61	4.25	2.27
Overall average	**9.23**	**1.15**	**10.10**	**1.73**

Some plots

Training plots for REINFORCE algorithm

PHM 2010 C-04 data-set: Variant: Single-state tool-wear with low noise (1e-3) and low break-down chance (5%)
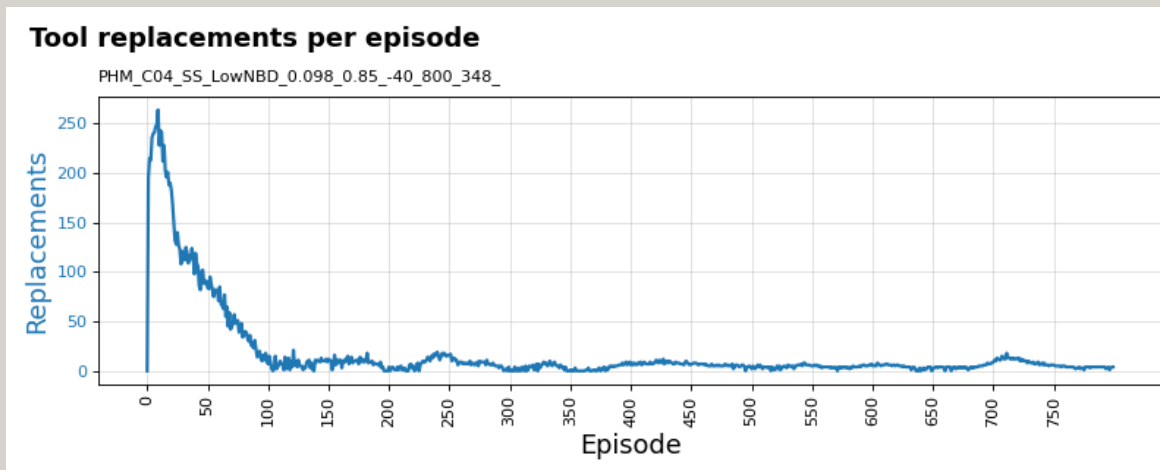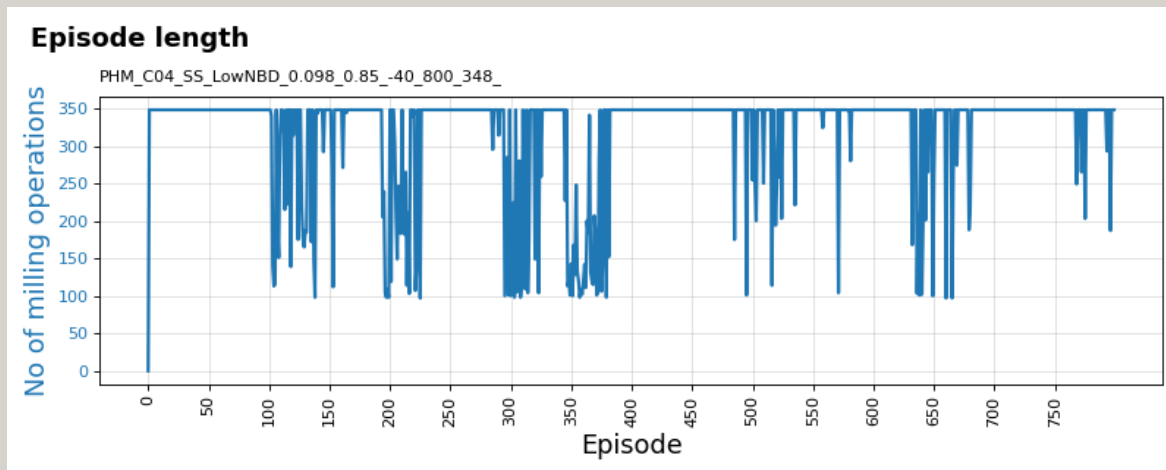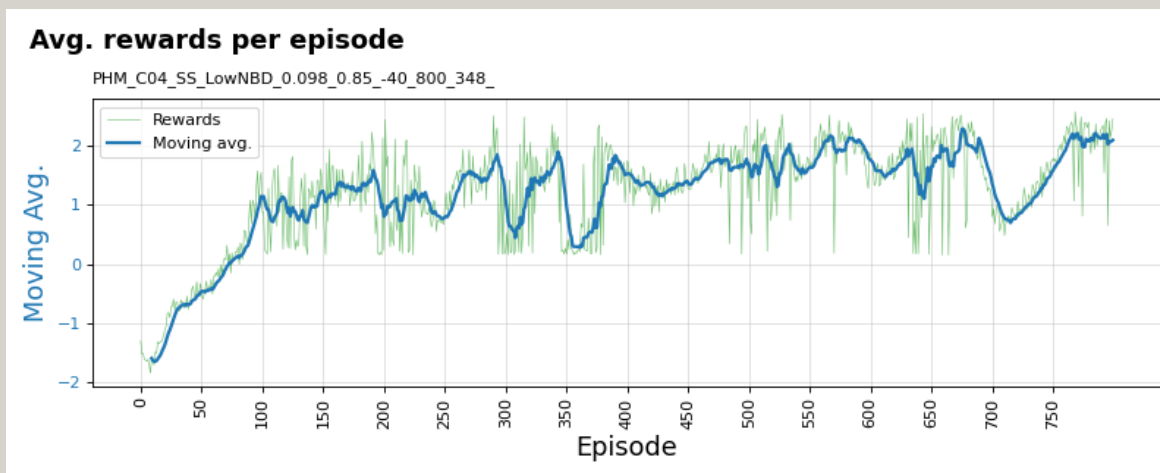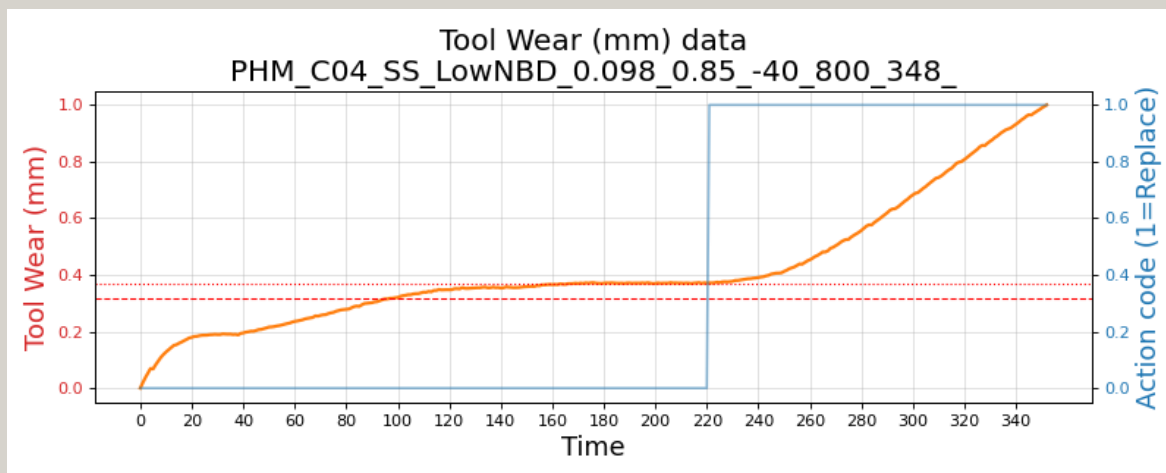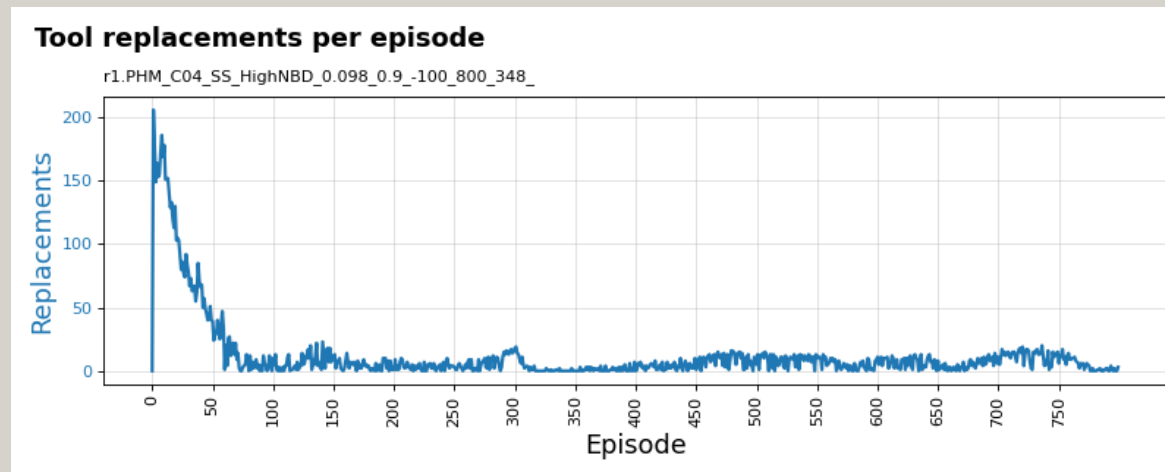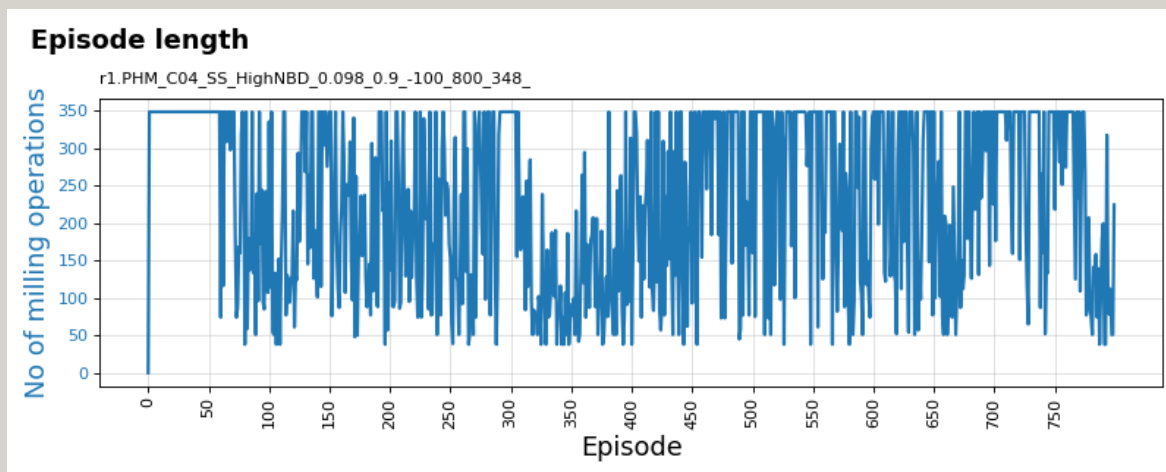
Training plots for REINFORCE algorithm

PHM 2010 C-04 data-set: Variant: Single-state tool-wear with high noise (1e-2) and low break-down chance (10%)
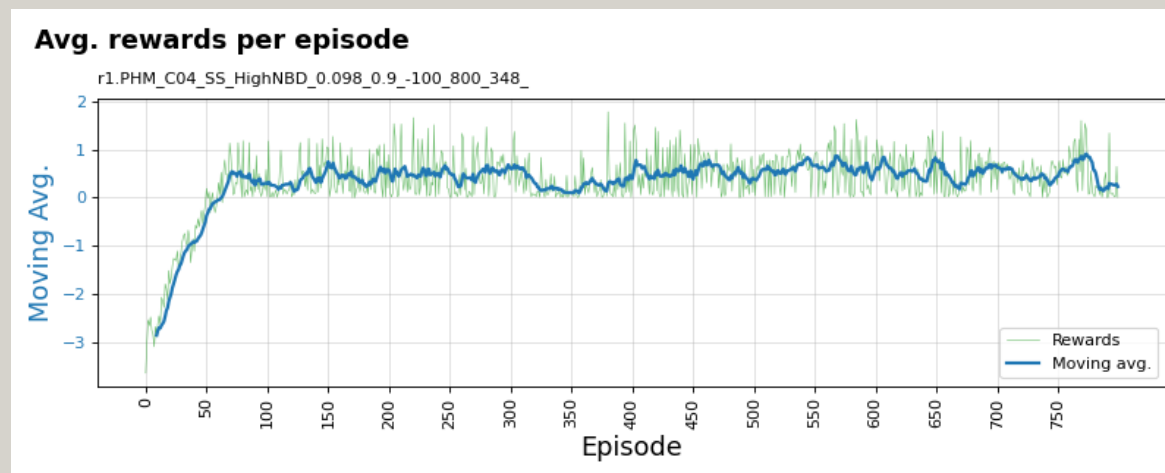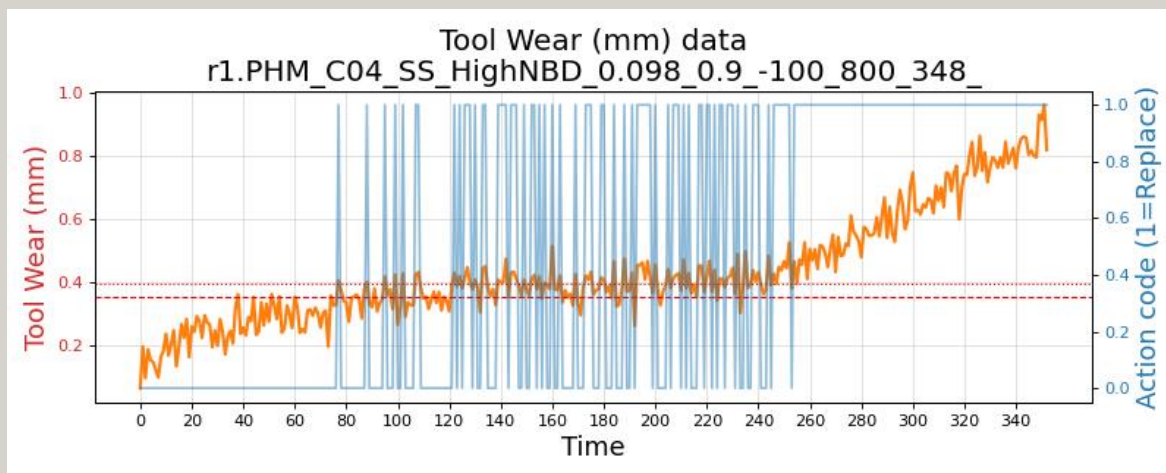
Tensorboard plots for SB-3 algorithms - PPO
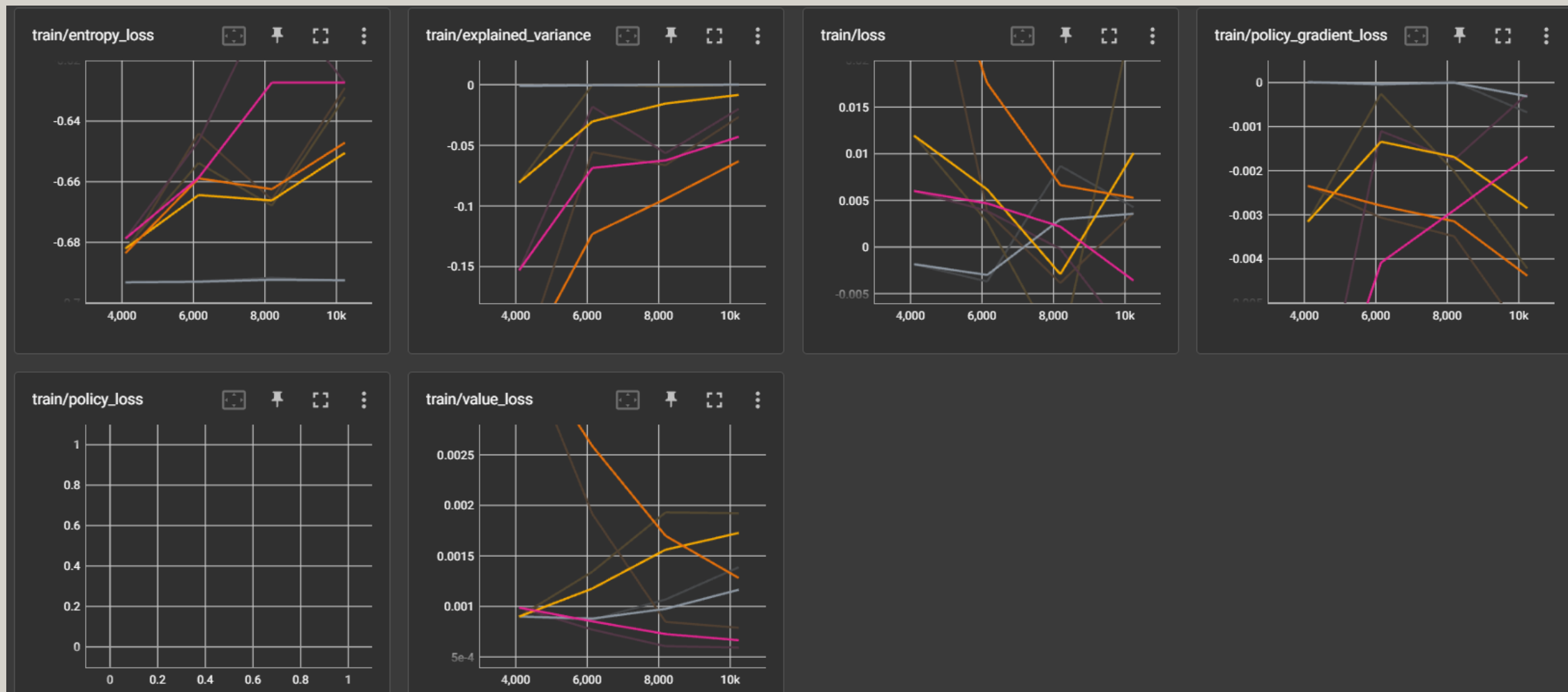
SB-3 stability doubts

SB-3 stability issue?

1. Earlier runs with same episodes = 800

2. But REINFORCE has an internal loop of 348

3. Increased SB-3 algo runs to 10,000, then 800 x 348 then 20,000

4. Re-installed SB-3 and upgraded the version

5. Tested on cart-pole and mountain-car environment

6. Re-trained SB-3 models and re-ran experiments for 20,000 – see next slide

Results: SB-3 updated. Episodes = 20 K

Model	REINFORCE			SB-3 A2C			SB-3 DQN			SB-3 PPO		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Simulated Dasic 2006 - No noise	1.00	0.71	0.83	0.00	0.00	0.00	0.50	0.96	0.66	0.07	0.01	0.02
Simulated Dasic 2006 - Low NBD	0.94	0.94	0.94	0.54	0.57	0.55	0.88	0.11	0.19	0.00	0.00	0.00
Simulated Dasic 2006 - High NBD	0.88	0.98	0.93	0.44	0.22	0.29	0.07	0.04	0.05	0.33	0.04	0.07
PHM C01 simple - No noise	0.90	0.97	0.93	0.00	0.00	0.00	0.09	0.03	0.04	0.53	0.73	0.61
PHM C01 simple - Low NBD	0.95	0.92	0.93	0.50	0.45	0.47	0.98	0.47	0.63	0.40	0.02	0.04
PHM C01 simple - High NBD	0.83	0.92	0.87	0.51	0.44	0.47	0.88	0.24	0.37	0.20	0.01	0.02
PHM C04 simple - No noise	0.85	0.98	0.91	0.50	1.00	0.67	0.80	0.05	0.09	0.51	0.25	0.33
PHM C04 simple - Low NBD	0.71	1.00	0.83	0.52	0.60	0.55	0.50	0.97	0.66	0.10	0.01	0.02
PHM C04 simple - High NBD	0.73	0.90	0.81	0.50	1.00	0.67	0.40	0.02	0.04	0.41	0.11	0.17
PHM C06 simple - No noise	1.00	0.58	0.73	0.59	0.53	0.56	0.20	0.01	0.02	0.56	0.66	0.60
PHM C06 simple - Low NBD	0.94	0.92	0.93	0.00	0.00	0.00	0.33	0.03	0.05	0.20	0.01	0.02
PHM C06 simple - High NBD	0.69	0.88	0.78	0.00	0.00	0.00	0.40	0.03	0.06	0.37	0.20	0.26
PHM C01 multi-variate state	0.86	0.90	0.88	0.31	0.08	0.13	0.36	0.54	0.43	0.58	0.12	0.19
PHM C04 multi-variate state	0.72	0.61	0.66	0.49	0.94	0.64	0.50	0.98	0.66	0.44	0.15	0.22
PHM C06 multi-variate state	1.00	0.58	0.73	0.50	1.00	0.67	0.51	0.99	0.67	0.53	0.36	0.43

Thank you