Experiments with REINFORCE algorithm

A purely experimental and empirical study

Rajesh Siraskar

14-Jun-2023

Agenda

1. Study objective
2. The RL environment
3. Evaluation strategy
4. Implementation architectures
5. Results
6. Findings
7. Discussion - *What could I possibly be doing wrong?*

Study Objective

Stable Baselines 3 implementations	Custom REINFORCE
Advanced algorithms – A2C and PPO Earlier algorithm – DQN	Very early algorithm – REINFORCE
Industry grade implementation	Custom, extremely naïve
Stable implementation	Un-stable (no effort made to improve stability)
Architecture: 2 layers x 64 | Tanh	Single layer 64 | ReLU

Important: This forms justification for how we choose the model file to be tested.

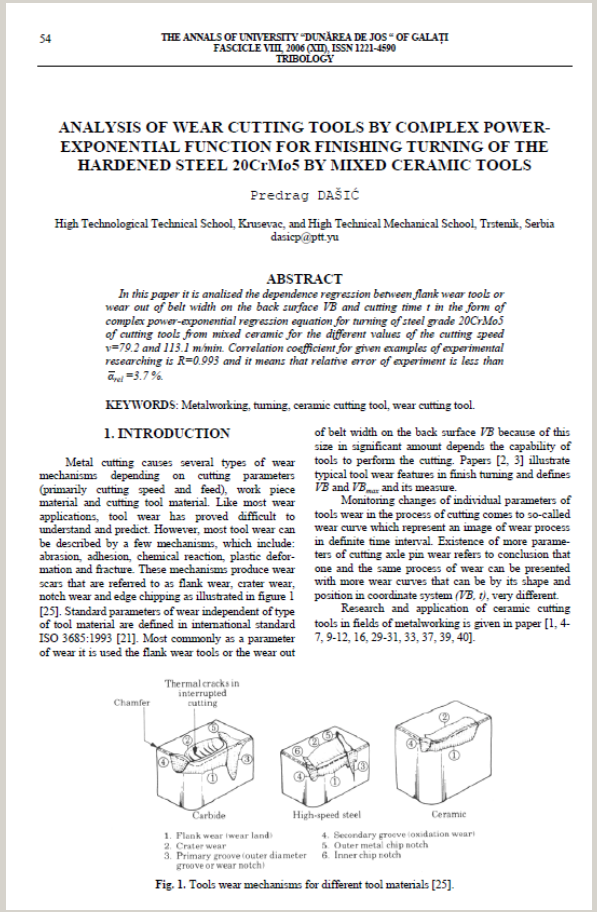
Study Objective *(contd.)*

1. My “proposal” goal is “Deep RL for predicting RUL”
2. Current sub/related goal: *An optimal predictive maintenance policy for replacement of milling tool*
3. REINFORCE - implemented from “scratch” – with hope of refining over time
4. This is a purely experimental (empirical) exercise.
5. Audience:
    1. RL/CS Researchers
    2. Practitioners, who do not understand algorithms deeply and deep learning or deep RL hyper-parameter tuning. Will want to use “default” algorithms

RL environment

Two different sources – simulated and real milling data

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



54 THE ANNALS OF UNIVERSITY "DUNĂREA DE JOS" OF GALAȚI
FASCICLE VIII, 2006 (XII), ISSN 1221-4590
TRIBOLOGY

ANALYSIS OF WEAR CUTTING TOOLS BY COMPLEX POWER-EXPONENTIAL FUNCTION FOR FINISHING TURNING OF THE HARDENED STEEL 20CrMo5 BY MIXED CERAMIC TOOLS

Predrag DAŠIĆ

High Technological Technical School, Kruševac, and High Technical Mechanical School, Trstenik, Serbia
dasicp@ptt.yu

ABSTRACT

In this paper it is analized the dependence regression between flank wear tools or wear out of belt width on the back surface VB and cutting time t in the form of complex power-exponential regression equation for turning of steel grade 20CrMo5 of cutting tools from mixed ceramic for the different values of the cutting speed v=79.2 and 113.1 m/min. Correlation coefficient for given examples of experimental researching is R=0.993 and it means that relative error of experiment is less than $\bar{a}_{rel}$ *=3.7 %.*

KEYWORDS: Metalworking, turning, ceramic cutting tool, wear cutting tool.

1. INTRODUCTION

Metal cutting causes several types of wear mechanisms depending on cutting parameters (primarily cutting speed and feed), work piece material and cutting tool material. Like most wear applications, tool wear has proved difficult to understand and predict. However, most tool wear can be described by a few mechanisms, which include: abrasion, adhesion, chemical reaction, plastic deformation and fracture. These mechanisms produce wear scars that are referred to as flank wear, crater wear, notch wear and edge chipping as illustrated in figure 1 [25]. Standard parameters of wear independent of type of tool material are defined in international standard ISO 3685:1993 [21]. Most commonly as a parameter of wear it is used the flank wear tools or the wear out of belt width on the back surface *VB* because of this size in significant amount depends the capability of tools to perform the cutting. Papers [2, 3] illustrate typical tool wear features in finish turning and defines *VB* and VB_{max} and its measure.

Monitoring changes of individual parameters of tools wear in the process of cutting comes to so-called wear curve which represent an image of wear process in definite time interval. Existence of more parameters of cutting axle pin wear refers to conclusion that one and the same process of wear can be presented with more wear curves that can be by its shape and position in coordinate system (*VB, t*), very different.

Research and application of ceramic cutting tools in fields of metalworking is given in paper [1, 4-7, 9-12, 16, 29-31, 33, 37, 39, 40].

Fig. 1. Tools wear mechanisms for different tool materials [25].

◄ 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

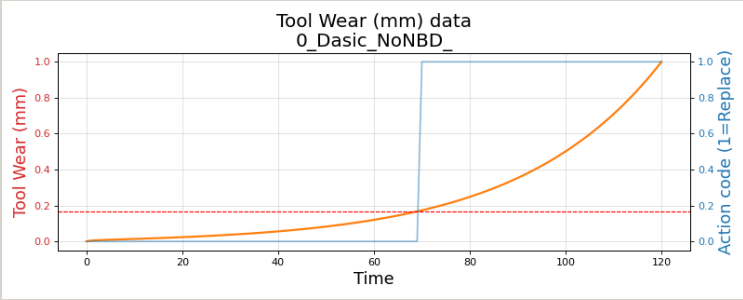
RL environment - Variants

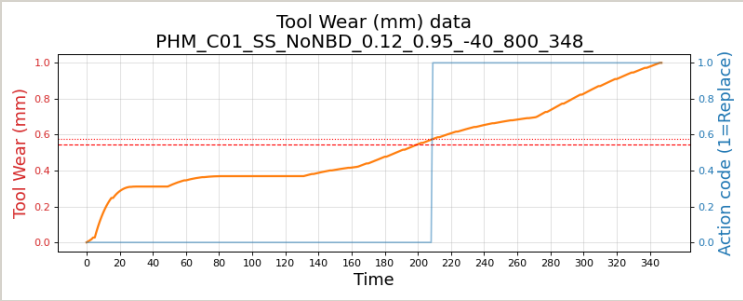
Three environments and their variants:

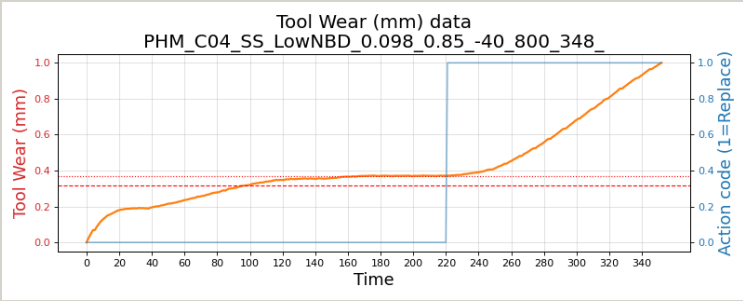
1. Simulated. Based on Dašić (2006). Simple single-variate state (tool wear)
    - Variants: (1) No noise (2) Low noise and low break-down chance and (3) High noise and high break-down chance

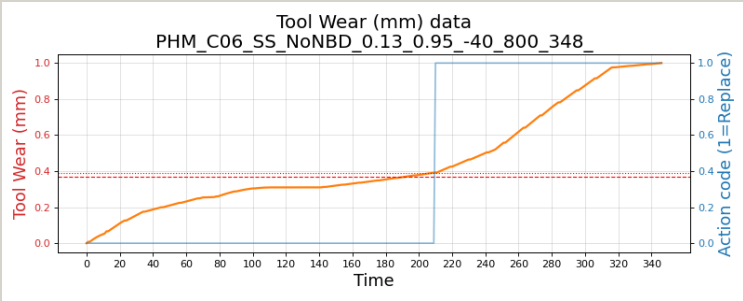
2. PHM 2010 real data – Simple single-variate state (tool wear)
    - Variants: C-01, C-04 and C-06 data-sets
    - Variants: (1) No noise (2) Low noise and low break-down chance and (3) High noise and high break-down chance

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

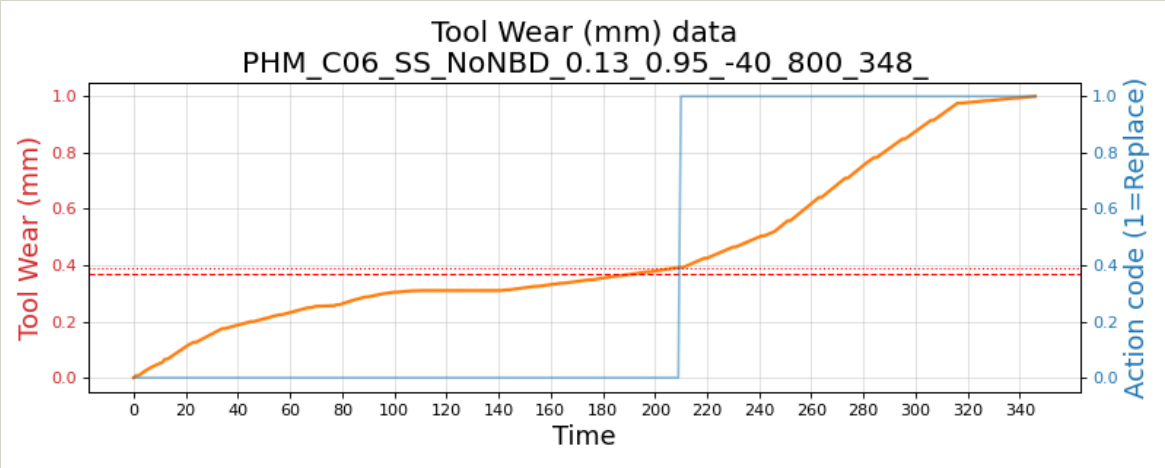


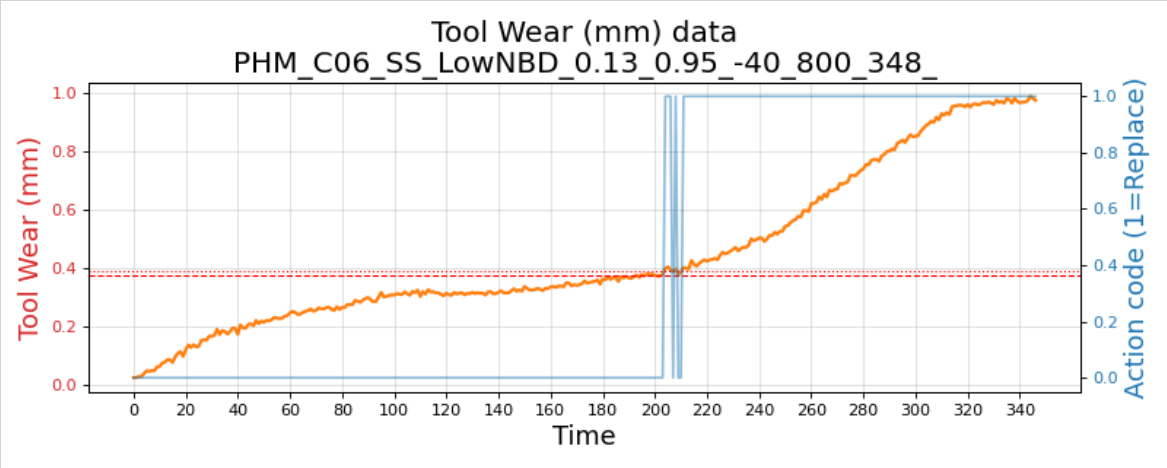


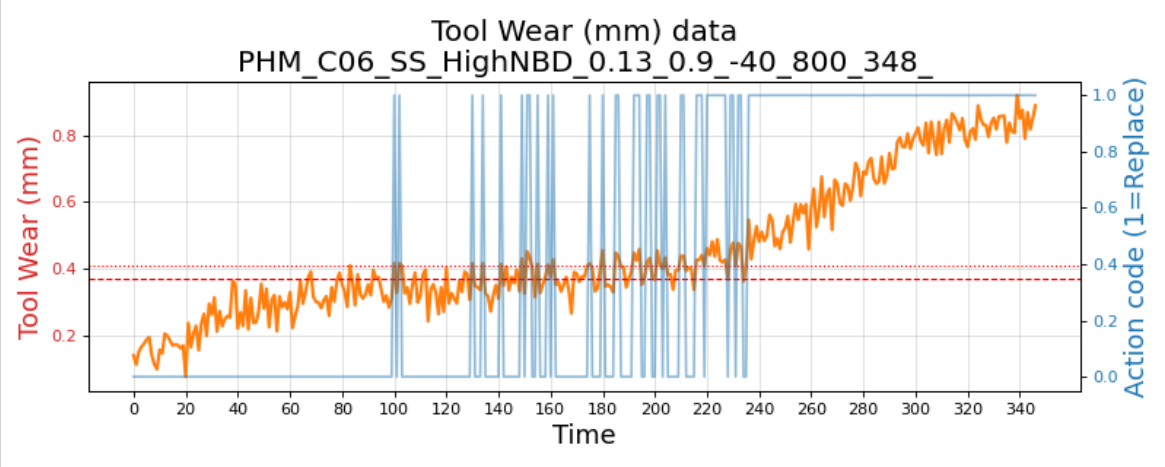




Wear plot - real data and its variants







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
- Train for 5 rounds. Each model tested over 10 rounds of randomly sampled 40 points
- Selecting the models to test (*“justification”*)
  - REINFORCE – model performance sometimes dropped to 0.5 or even 0.0. Select ones that performed > 0.7
  - SB-3 – almost always 0.5-0.6 on average, sometimes showing high performance (0.8)
- Compute Precision, Recall and F1
- Compare mean and std. deviations

Architectures

	A2C	DQN	PPO	REINFORCE
Network architecture	input dim x [64|Tanh x 64|Tanh] x output dim	input dim x [64|Tanh x 64|Tanh] x output dim	input dim x [64|Tanh x 64|Tanh] x output dim	input dim x [64|ReLU] x output dim
Layers	2	2	2	1
Units	64 x 64	64 x 64	64 x 64	64
Activation	Tanh, Tanh	Tanh, Tanh	Tanh, Tanh	ReLU
Learning rate	0.0007	0.0001	0.0003	0.01
Gamma	0.99	0.99	0.99	0.99
Optimizer	RMSprop	Adam	Adam	Adam

Results

	REINFORCE				A2C				DQN				PPO			
Environment	**Precision**	**Recall**	**F1**	**Fβ 0.5**	**Precision**	**Recall**	**F1**	**Fβ 0.5**	**Precision**	**Recall**	**F1**	**Fβ 0.5**	**Precision**	**Recall**	**F1**	**Fβ 0.5**
Simulated - No noise	0.999	0.645	0.782	0.898	0.335	0.359	0.344	0.338	0.348	0.597	0.410	0.352	0.392	0.211	0.252	0.303
Simulated - Low noise	0.943	0.954	0.948	0.945	0.409	0.318	0.349	0.379	0.273	0.064	0.076	0.108	0.359	0.173	0.205	0.255
Simulated - High noise	0.889	0.974	0.929	0.904	0.471	0.439	0.443	0.455	0.423	0.408	0.295	0.284	0.402	0.205	0.248	0.307
PHM C01 SS - No noise	0.886	0.978	0.928	0.902	0.294	0.337	0.305	0.296	0.350	0.405	0.291	0.269	0.517	0.494	0.471	0.476
PHM C01 SS - Low noise	0.916	0.893	0.903	0.911	0.526	0.645	0.568	0.540	0.321	0.591	0.404	0.343	0.490	0.415	0.443	0.468
PHM C01 SS - High noise	0.757	0.926	0.831	0.784	0.499	0.632	0.542	0.513	0.399	0.402	0.308	0.292	0.403	0.223	0.270	0.325
PHM C04 SS - No noise	0.865	0.959	0.908	0.881	0.515	0.676	0.575	0.535	0.365	0.497	0.383	0.348	0.431	0.239	0.265	0.311
PHM C04 SS - Low noise	0.722	0.980	0.831	0.762	0.399	0.393	0.391	0.393	0.409	0.589	0.410	0.361	0.438	0.299	0.334	0.377
PHM C04 SS - High noise	0.770	0.809	0.787	0.776	0.375	0.456	0.397	0.381	0.408	0.411	0.296	0.282	0.491	0.324	0.362	0.409
PHM C06 SS - No noise	0.996	0.609	0.751	0.879	0.463	0.454	0.455	0.459	0.538	0.780	0.585	0.523	0.402	0.410	0.374	0.370
PHM C06 SS - Low noise	0.968	0.854	0.905	0.941	0.508	0.615	0.548	0.522	0.395	0.593	0.411	0.362	0.454	0.342	0.367	0.404
PHM C06 SS - High noise	0.699	0.912	0.790	0.732	0.480	0.512	0.466	0.467	0.581	0.499	0.417	0.433	0.424	0.199	0.252	0.314
PHM C01 MS - No noise	0.824	0.895	0.856	0.836	0.444	0.284	0.315	0.358	0.313	0.215	0.165	0.175	0.513	0.347	0.395	0.448
PHM C04 MS - No noise	0.752	0.678	0.709	0.733	0.506	0.326	0.368	0.425	0.588	0.642	0.492	0.486	0.472	0.455	0.444	0.452
PHM C06 MS - No noise	1.000	0.643	0.779	0.896	0.499	0.731	0.575	0.523	0.520	0.239	0.209	0.256	0.509	0.260	0.330	0.409

Results: Overall – all environments and their variants

	Precision		Recall		F1 score		F beta (0.5)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
A2C	0.448	0.074	0.478	0.084	0.443	0.071	0.439	0.069
DQN	0.415	0.196	0.462	0.033	0.343	0.038	0.325	0.063
PPO	0.447	0.147	0.306	0.090	0.334	0.093	0.375	0.107
REINFORCE	0.866	0.042	0.847	0.054	0.842	0.043	0.852	0.042

Results: Simple single variate state. Including noise variants

Simulated – Dašić, Predrag (2006).

	Precision		Recall		F1 score		F beta (0.5)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
A2C	0.405	0.079	0.372	0.086	0.379	0.076	0.391	0.076
DQN	0.348	0.217	0.356	0.033	0.260	0.041	0.248	0.068
PPO	0.385	0.175	0.196	0.064	0.235	0.080	0.289	0.110
REINFORCE	0.944	0.029	0.858	0.041	0.886	0.032	0.916	0.030

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

	Precision		Recall		F1 score		F beta (0.5)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
A2C	0.451	0.064	0.524	0.085	0.472	0.067	0.456	0.063
DQN	0.418	0.172	0.530	0.032	0.389	0.034	0.357	0.055
PPO	0.450	0.146	0.327	0.095	0.349	0.095	0.384	0.106
REINFORCE	0.842	0.043	0.880	0.053	0.848	0.043	0.841	0.042

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

	Precision		Recall		F1 score		F beta (0.5)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
A2C	0.483	0.101	0.447	0.081	0.419	0.075	0.435	0.079
DQN	0.474	0.248	0.365	0.038	0.289	0.049	0.306	0.082
PPO	0.498	0.121	0.354	0.103	0.390	0.101	0.436	0.104
REINFORCE	0.859	0.053	0.739	0.069	0.781	0.055	0.822	0.052

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**

Findings

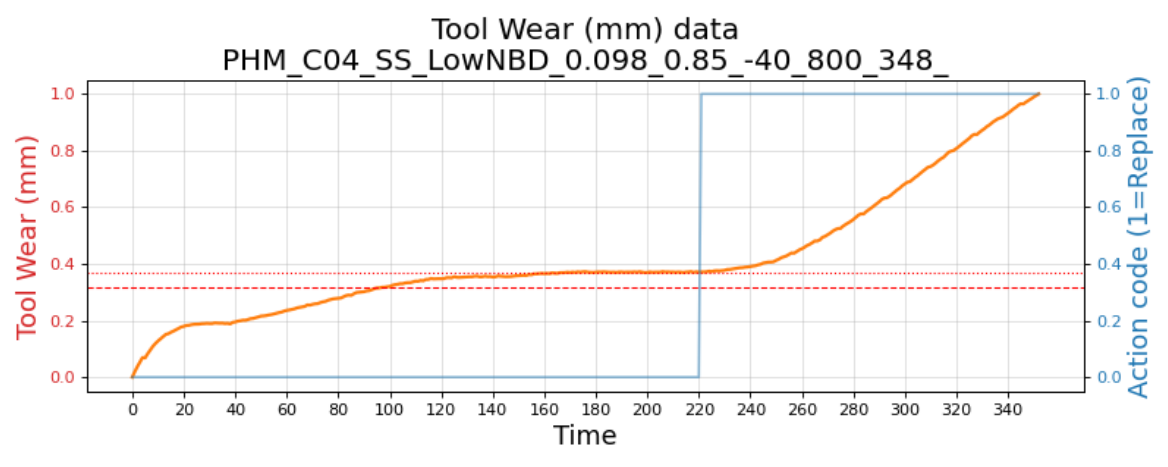
1. The naïve implementation of REINFORCE algorithm was implemented in PyTorch with an extremely simple architecture: Three layers and ReLU activation and an Adam optimizer.
2. In training, the SB-3 algorithms were extremely stable, but the F1 performance was always between 0.38-0.67. The naïve REINFORCE, 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 0.78 (for complex real data environment) to 0.88 (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

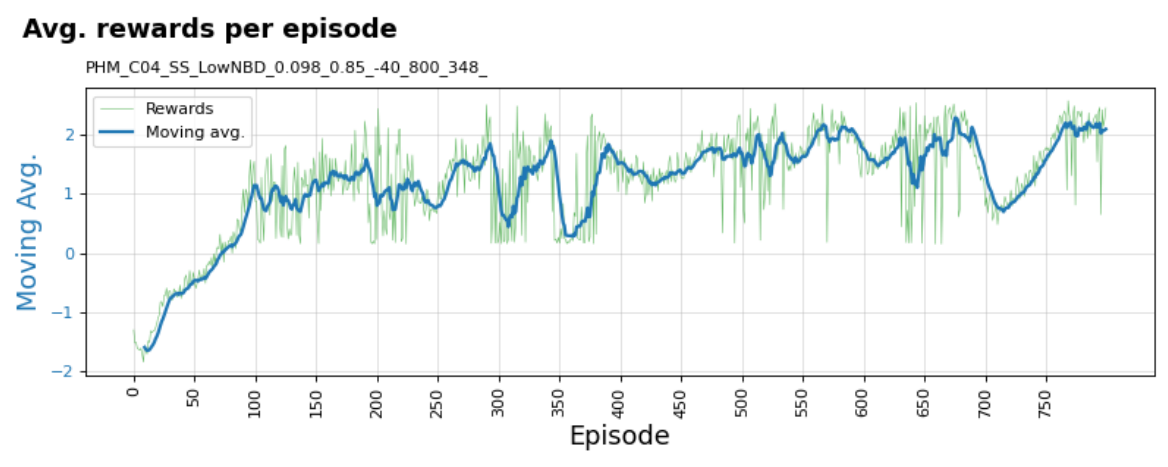
1. w.r.t. lowest of SB-3 algorithms
2. Reference data-sheet: "1. Analysis_model_performance_summary_V3 (PPT ref.).xls"

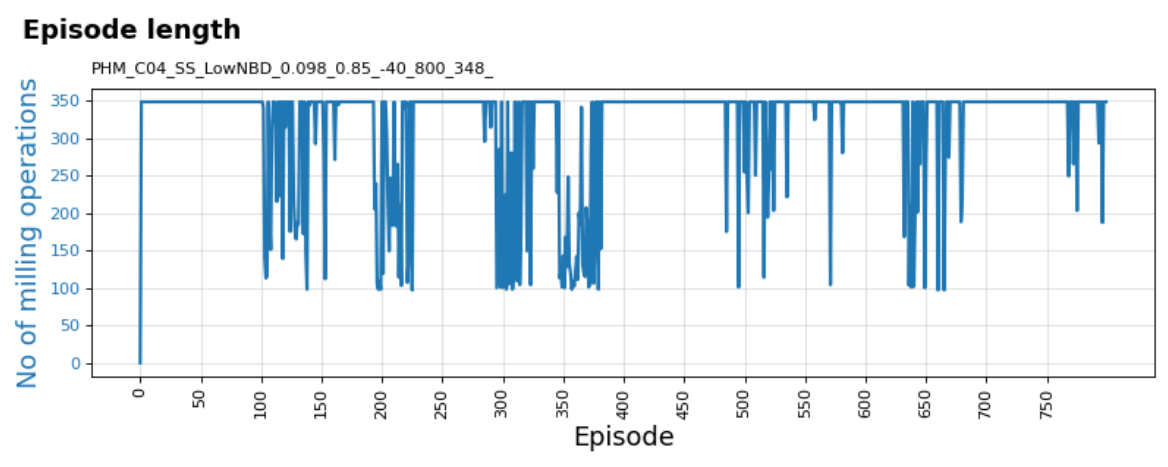
Some plots

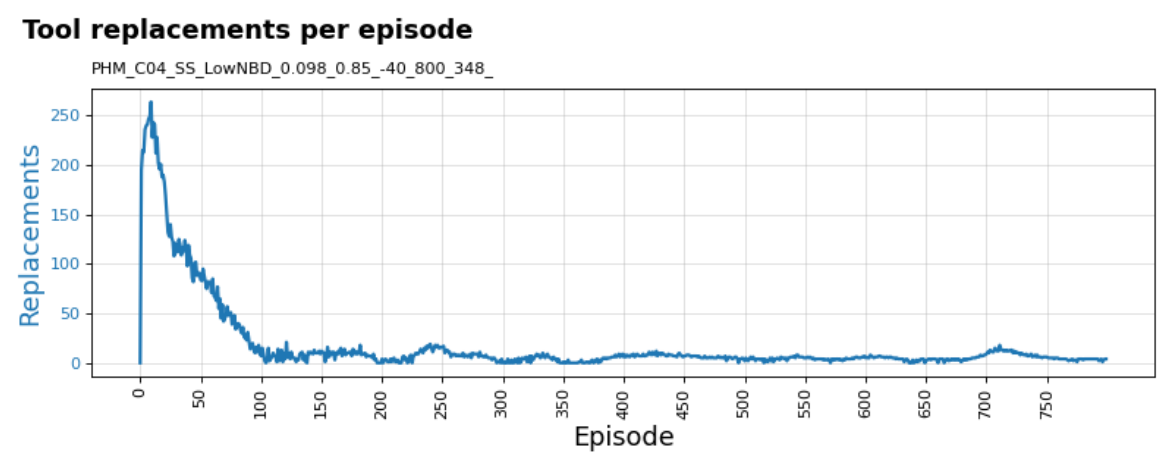
Training plots for REINFORCE algorithm

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



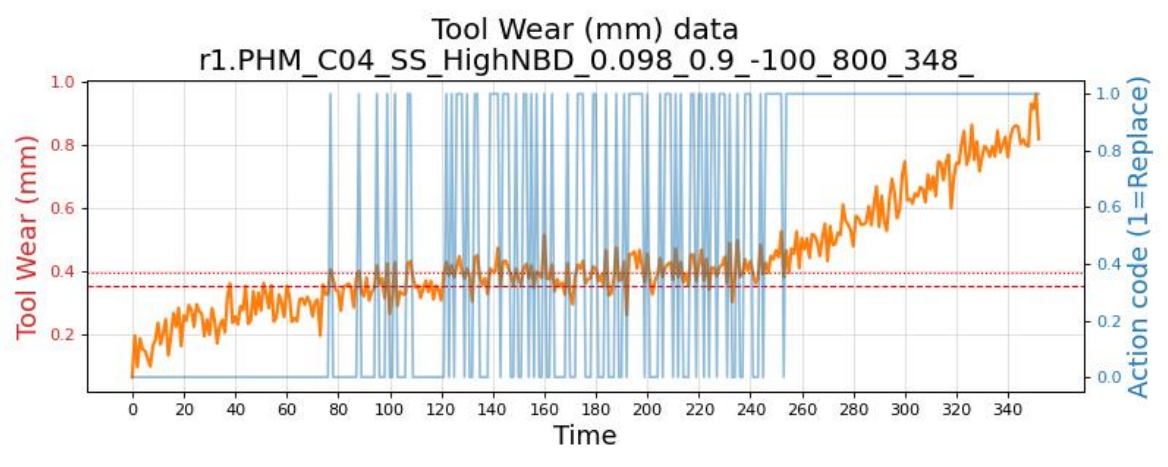


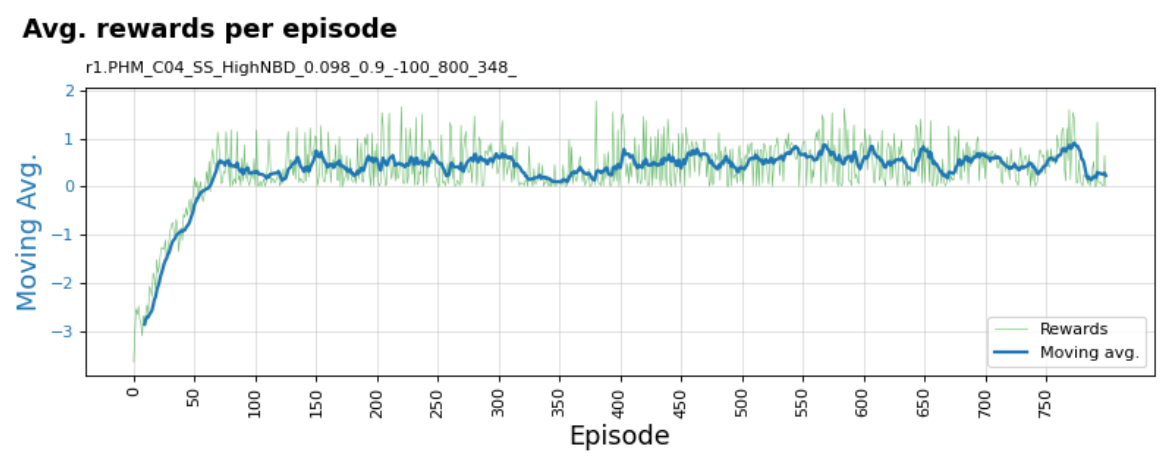


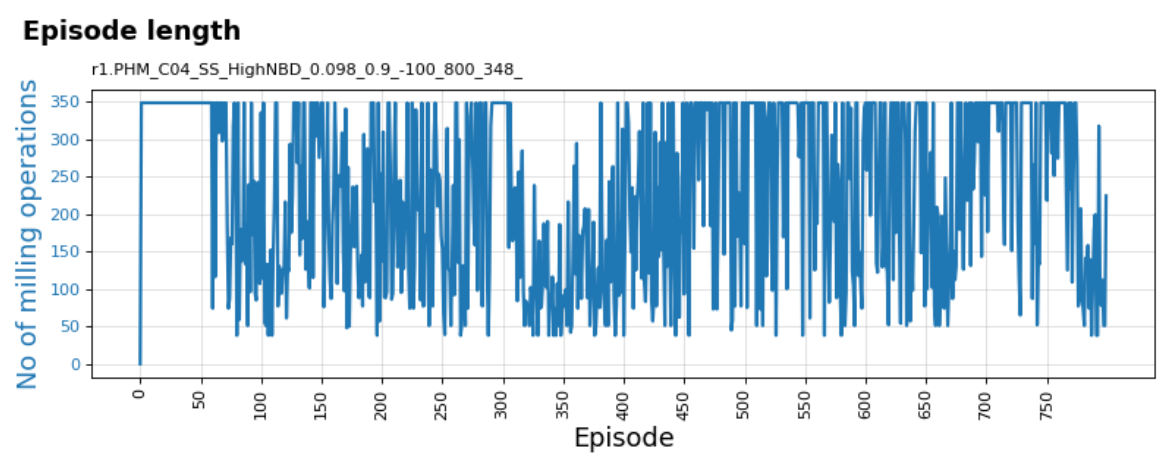


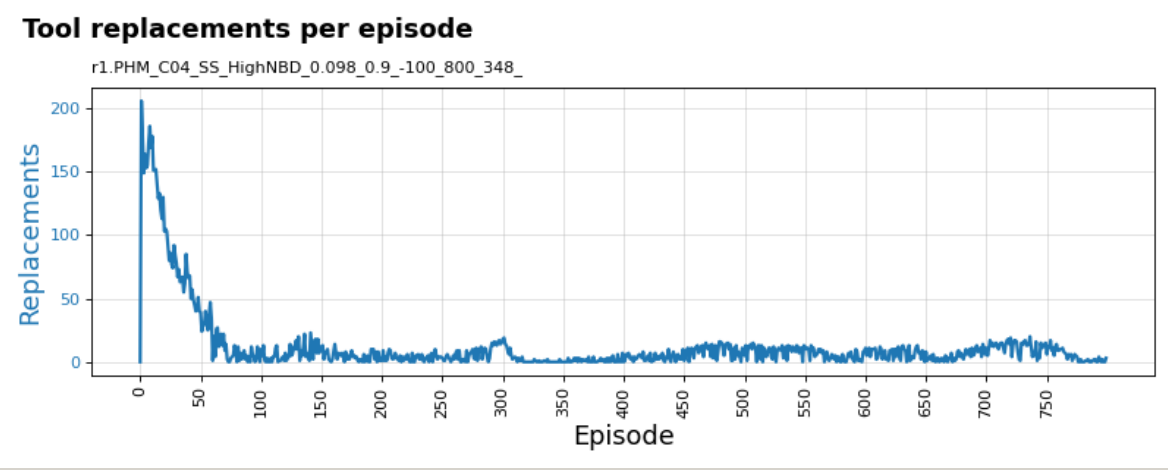
Training plots for REINFORCE algorithm

PHM 2010 C-04 data-set: Variant: Single-state tool-wear with high noise (1e-2) and low break-down chance (10%)



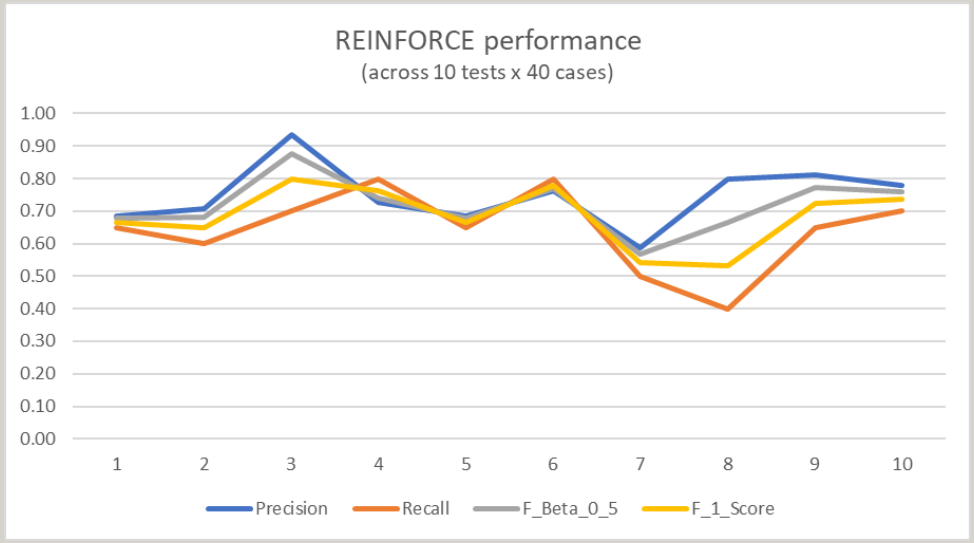


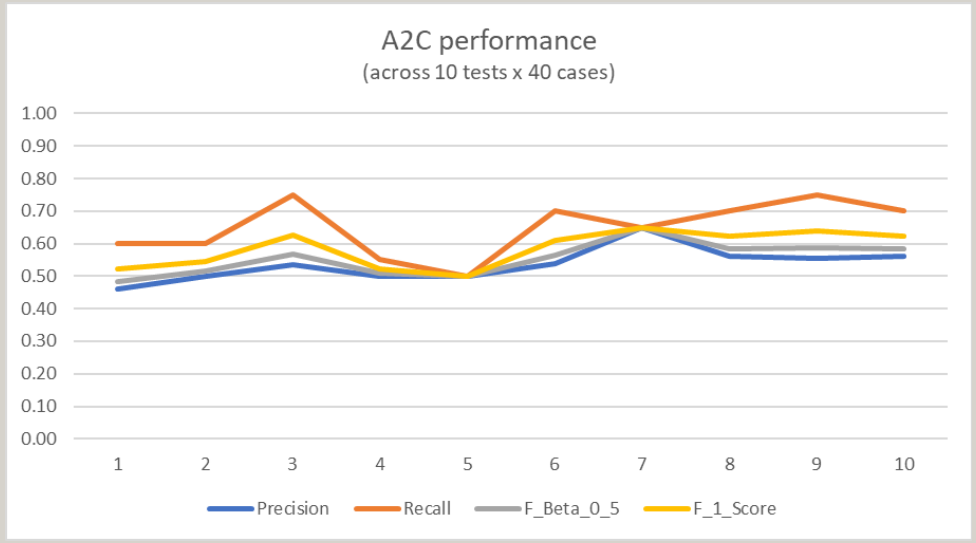


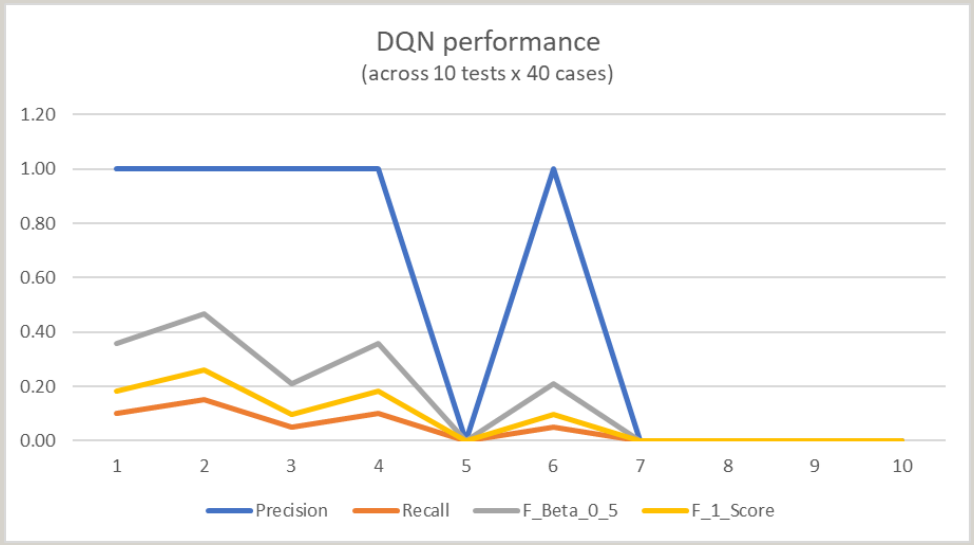


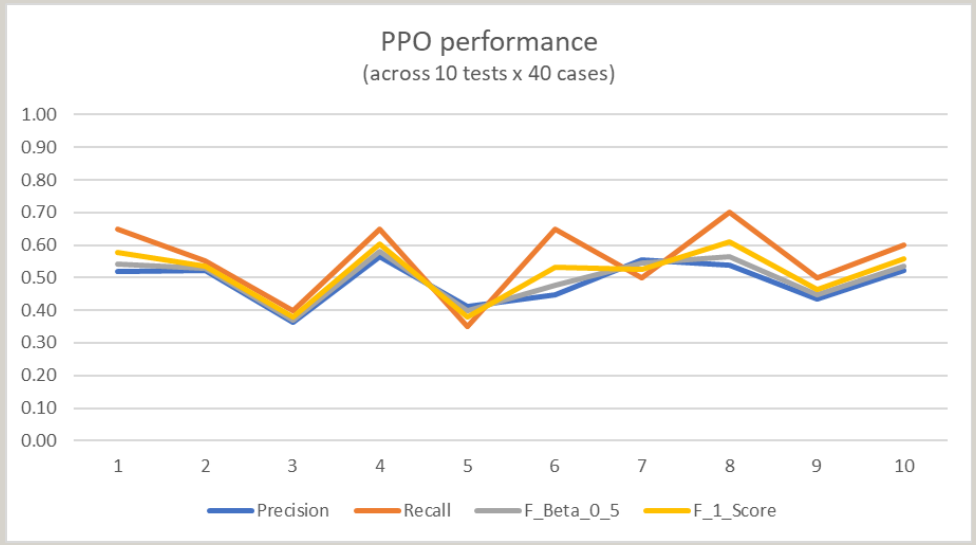
Sample test plots for algorithms

PHM C-04 data-set: Variant: Multi-variant state with no noise or break-down chance

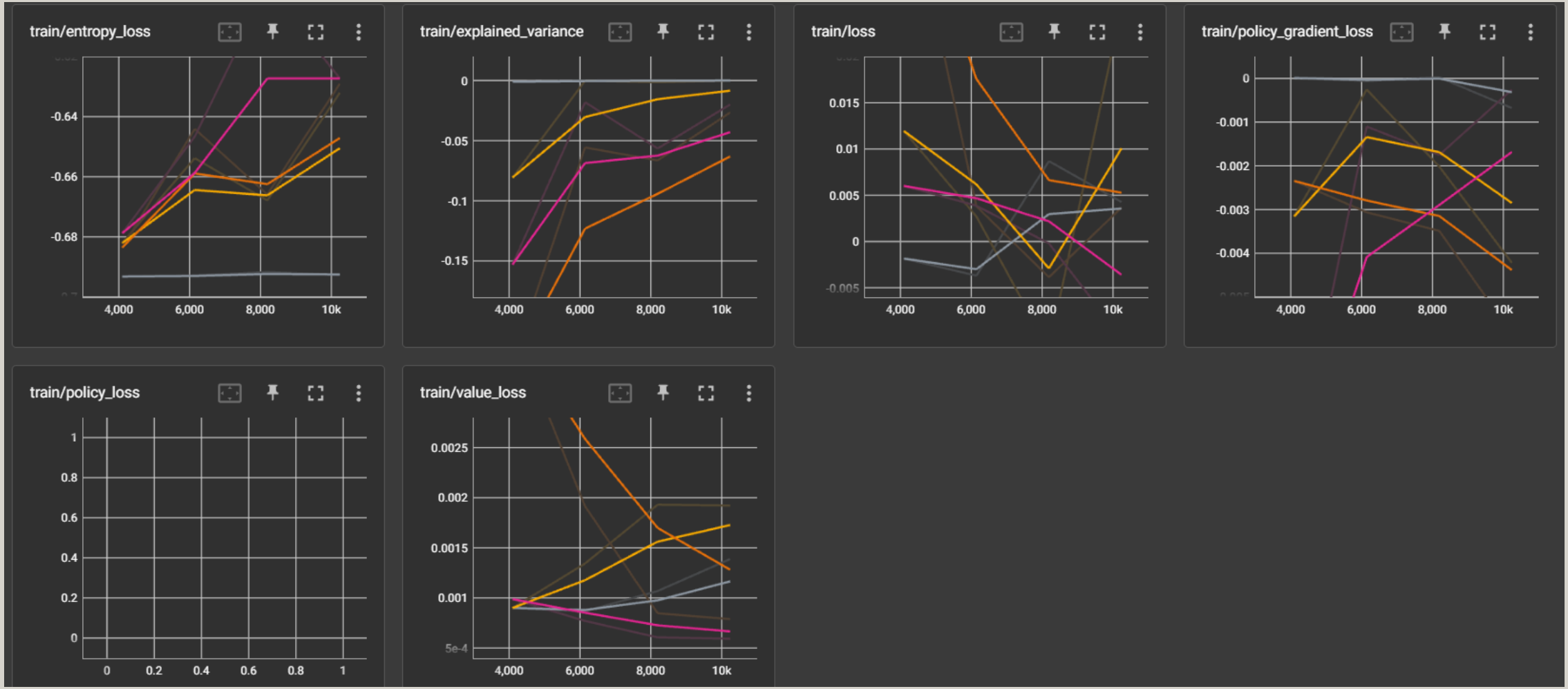








Tensorboard plots for SB-3 algorithms - PPO



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 10,000 and 20,000 episodes– see next slide

Results: SB-3 (10 K episodes) – single training round

	REINFORCE			SB-3 A2C			SB-3 DQN			SB-3 PPO		
Model	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Simulated Dasic 2006 - No noise	1.00	0.70	0.82	0.49	0.44	0.46	0.07	0.01	0.02	0.46	0.28	0.34
Simulated Dasic 2006 - Low NBD	0.94	0.90	0.92	0.00	0.00	0.00	0.45	0.03	0.05	0.37	0.06	0.10
Simulated Dasic 2006 - High NBD	0.90	1.00	0.94	0.50	0.53	0.51	0.50	0.96	0.66	0.58	0.15	0.24
PHM C01 simple - No noise	0.91	0.96	0.93	0.52	0.56	0.54	0.37	0.03	0.05	0.50	0.25	0.33
PHM C01 simple - Low NBD	0.89	0.80	0.84	0.38	0.13	0.19	0.40	0.02	0.04	0.46	0.14	0.21
PHM C01 simple - High NBD	0.78	0.93	0.85	0.00	0.00	0.00	0.51	0.96	0.66	0.52	0.21	0.29
PHM C04 simple - No noise	0.82	0.96	0.88	0.51	0.09	0.15	0.50	0.97	0.66	0.49	0.47	0.48
PHM C04 simple - Low NBD	0.74	0.99	0.85	0.47	0.51	0.49	0.71	0.72	0.71	0.53	0.28	0.37
PHM C04 simple - High NBD	0.67	0.78	0.72	0.52	0.55	0.53	0.50	0.98	0.66	0.47	0.25	0.32
PHM C06 simple - No noise	1.00	0.65	0.78	0.47	0.46	0.46	0.51	0.98	0.67	0.38	0.14	0.20
PHM C06 simple - Low NBD	0.98	0.84	0.90	0.49	0.53	0.51	0.95	0.58	0.72	0.50	0.38	0.43
PHM C06 simple - High NBD	0.72	0.88	0.79	0.50	0.47	0.48	0.51	0.98	0.67	0.37	0.11	0.17
PHM C01 multi-variate state	0.80	0.92	0.85	0.51	0.59	0.55	0.58	0.97	0.73	0.49	0.24	0.31
PHM C04 multi-variate state	0.77	0.69	0.73	0.48	0.53	0.50	0.51	0.97	0.66	0.49	0.34	0.40
PHM C06 multi-variate state	1.00	0.57	0.73	0.49	0.60	0.54	0.49	0.96	0.65	0.45	0.48	0.46

SB-3 algorithms do perform well sometimes *(green)*. On an average their performance is poor *(red-orange-yellow)*.

Results: SB-3 (20 K episodes) – single training round

	REINFORCE			SB-3 A2C			SB-3 DQN			SB-3 PPO		
Model	**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

SB-3 algorithms do perform well sometimes *(green)*. On an average their performance is poor *(red-orange-yellow)*.

Thank you