Model based and model free reinforcement learning at FERMI FEL

Simon Hirlaender
Univertsity of Salzburg, Salzburg, Austria

Niky Bruchon
University of Trieste, Trieste, IT
(Dated: November 25, 2020)

In this paper we discuss a model based reinforcement learning approach in comparison to a model free reinforcement learning approach applied at the FERMI FEL system. Both algorithms and approaches are new in this context and the main purpose of this paper is to show how reinforcement learning can be applied on an operational level in a feasible training time on real accelerator physics problems. In terms of sample-complexity the model-based approach is faster, while the final performance of the model free method is superior - the so called asymptotic performance. The model-based algorithm is done in a Dyna-style using an uncertainty aware model and the model-free algorithm is based on tailored deep Q-learning using some tricks to increase the sample efficiency.

- Noise benchmarks with pendulum - done
- Benchmark NAF2 with pendulum - done
- Compare NAF2 and AE-DYNA
- The first time model based on accelerator problem compared to a novel state of the art MFRL: NAF2
- Tailored to accelerators - short horizons are standard?
- Long waiting time in MBRL
- Noise sensivity as mentioned in Benchmark paper
- Improvements of the algorithms:
  - Adding the noise feature
  - Jumping feature using SAC
  - Noise in the policy
  - discuss different approaches to e.g. dynamic waiting time
  - usually prior not discussed in papers

I. INTRODUCTION AND MOTIVATION

In particle accelerators one main goal is to provide stable and reproducible performance. In order to achieve this, a number of control problems have to be considered simultaneously. Especially if there is no way to model the physics, one might use optimisation techniques as e.g. derivative free optimisers (DFOs) or model-based optimisations as Gaussian processes to restore or maintain the performance. Another promising way is to apply reinforcement learning which shows several advantages over optimisation methods:

- It covers a larger class of problems, RL optimises a sequence of decisions.
- It memorizes the problem and does not start always from scratch as an DFO.
- Existing data can be used.
- The underlying structure of the problem might be deduced.

One critical aspect using RL is the number of iterations needed to train a controller and the a second critical aspect is the robustness of the training itself.
In this paper, we present the study carried out to solve the maximisation-problem of the radiation intensity generated by a seeded Free-Electron Laser (FEL) on the Free Electron laser Radiation for Multidisciplinary Investigations (FERMI) at Elettra Sincrotrone Trieste. Three different algorithms were applied successfully to the FERMI FEL problem and reveal different advantages. The critical aspects are addressed and tailored solutions, generally applicable to accelerator control problems are presented.
The paper is organised as follows:

- Description of the problem set-up at FERMI
- Overview of RL
- Details of the implementations used in these studies and theoretical concerns
- Results
- Summary and outlook

II. THE PHYSICAL SET-UP OF THE STUDIED PROBLEM

In a seeded free-electron laser one of the most critical parameters is the temporal and transverse overlap of the electron and laser beam in the magnetic section called modulator undulator.
At FERMI several beam-based feedback systems are deployed to control the beams trajectories shot to shot with the main purpose of guaranteeing a steady intensity of

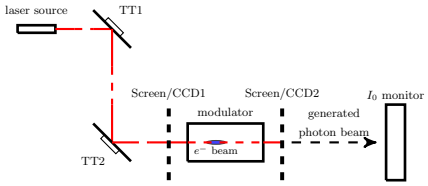

FIG. 1: A schematic view on the set-up of the FERMI FEL.

the light radiation. Nevertheless, in the last years various approaches have been studied to investigate their applicability in tuning operations.
A free-electron laser is a fourth-generation light source where the lasing medium consists of very-high-speed electron moving freely through a magnetic structure. The FERMI peculiarity is given by the usage of an external seeding source that provides several advantages, as the increased stability in pulse and photon energy, reduced size of the device, improved longitudinal coherence, more control on the pulse structure, and improved temporal synchronization with external timing systems.
The external optical laser signal provide is contribution to the FEL process in the modulator where it interacts with the relativistic electron beam modulating it in energy. The modulation in energy is the converted into a charge modulation in the dispersive section, and finally the density modulated beams radiation is amplified in the radiators section. The importance of ensuring the best possible overlapping between the seed laser and the electron beam in the modulator is therefore evident.
For this reason the proposed study focuses on the control of the seed laser trajectory in the modulator, looking at FERMI performance as a reference.

A. our system: the modulator and the seed laser

The most critical parameters in a seeded free-electron laser are the temporal and transverse overlap of the electron and laser beams in the modulator magnetic section. The problem is simplified by keeping constant the trajectory of the electron beam and the mechanical delay line that controls the temporal alignment. A schematic overview of the set-up is provided in fig. 1. Two mirrors, TT1 and TT2, are used to control the trajectory of the laser by tilting and inclining, which gives a total of four degrees of freedom (DOF). In turn the laser overlaps with the electron beam between the two screens, CCD1 and CCD2. Lastly the monitor measures the intensity I_0. The final problem faced consists in optimising the seed laser trajectory to match the electron beam and consequently increase the intensity of the FEL radiation.

III. DEEP REINFORCEMENT LEARNING

Assume states $\mathbf{s}$ the set of all states $\mathcal{S}$ and actions $\mathbf{a}$ all actions $\mathcal{A}$ and rewards r and the set of all rewards $\mathcal{R}$ an initial state distribution d_0. Goal of reinforcement learning is to find a $\pi^* : \mathbf{s} \mapsto \mathbf{a}$, which is the solution of:

$$\pi^* \operatorname{argmax} J(\pi) = \mathbb{E}_{\tau\sim p_\pi}\left[\sum_{t=0}^{H}\gamma^t r(\mathbf{s}_t,\mathbf{a}_t)\right], \qquad (1)$$

where $p_\pi = d_0(\mathbf{s}_0)\prod_{t_0}^{H}\pi(\mathbf{a}_t,\mathbf{s}_t)T(\mathbf{s}_{t+1}|\mathbf{s}_t,\mathbf{a}_t)$ is the distribution of all trajectories $\tau := (\mathbf{s}_0,\mathbf{a}_0,\mathbf{s}_1,\mathbf{a}_1,\dots\mathbf{s}_H,\mathbf{a}_H)$ drawn by π with a horizon H. $T(\mathbf{s}_{t+1}|\mathbf{s}_t,\mathbf{a}_t)$ characterises the probability to end in a state $\mathbf{s}_{t+1}$ if an action $\mathbf{a}_t$ is taken at state $\mathbf{s}_t$. In the modern field of RL one distinguishes if the policy $\pi(\mathbf{a}) \approx \pi_\phi(\mathbf{s})$ or the state-action value function $Q(\mathbf{s},\mathbf{a})$ is approximated using a high capacity function approximator, as e.g. a deep neural network. In the first case π is optimized directly and one speaks about policy gradient methods [1–6]. In the latter about approximate dynamic programming, which we now discuss.

IV. MODEL FREE REINFORCEMENT LEARNING - MFRL

We do not provide an exhaustive treatment of state of the art MFRL algorithms. Only the main principles are covered for the later statements and an overview is given e.g. in [1, 7]. We discuss in detail the normalised advantage function algorithm, which has good characteristics for accelerator control problems as shown in **??**.

A. Approximate dynamic programming

The state-value function $Q(\mathbf{s},\mathbf{a})$:

$$Q^\pi(\mathbf{s},\mathbf{a}) := \mathbb{E}_{\tau\sim p_\pi(\tau|\mathbf{s}_t,\mathbf{a}_t)}\left[\sum_{t=t'}^{H}\gamma^{(t'-t)} r(\mathbf{s}_t,\mathbf{a}_t)\right], \qquad (2)$$

is expressed as $Q^\pi_\theta(\mathbf{s},\mathbf{a})$, where θ denotes the parameters of the function approximator. By satisfying the Bellmann-optimality equation Q^π_θ can be trained towards the optimal $Q^*(\mathbf{s},\mathbf{a})$:

$$\min_\theta \left(\vec{Q}_\theta - \mathcal{B}^*\vec{Q}_\theta\right)^2. \qquad (3)$$

And π can be calculated via:

$$\pi_\theta(\mathbf{s}) = (\delta(\mathbf{a}) - \operatorname*{argmax}_{\mathbf{a}} Q_\theta(\mathbf{s},\mathbf{a})). \qquad (4)$$

The Bellman operator $\mathcal{B}^*$ has a unique fixed point but is non-linear, due to the max - operator:

$$\mathcal{B}^*\vec{Q}_\theta(\mathbf{s}_t,\mathbf{a}_t) := r(\mathbf{s}_t,\mathbf{a}_t) + \gamma \max_{\mathbf{a}}\left(Q_\theta(\mathbf{s}_{t+1},\mathbf{a}) - Q_\theta(\mathbf{s}_t,\mathbf{a}_t)\right). \qquad (5)$$

The form of this equation can cause overestimation and other complications, when using a function approximator. Several methods exist which try to mitigate the problems [8–12]. One way to reduce the effect is to take a simple analytical form of the Q-function.

B. Normalized advantage function

If a specific quadratic form of the Q function is assumed [11]:

$$Q_\theta(\mathbf{s},\mathbf{a}) = -\frac{1}{2}(\mathbf{a}-\mu_\theta(\mathbf{s}))P_\theta(\mathbf{a}-\mu_\theta(\mathbf{s}))^T + V_\theta(\mathbf{s}). \quad (6)$$

One modification, which is used in these tests is a twin network (weights for network i denoted by θ^i), where only one is used to obtain the policy, while the other is used for the update rule to avoid over-estimation. It is motivated by double Q-learning [8, 13]. The maximum is given analytically as $\max_{\mathbf{a}} Q(\mathbf{s},\mathbf{a}) = V(\mathbf{s})$, hence from eq. (3):

$$\min_\theta \left((r(\mathbf{s}_t,\mathbf{a}_t) + \gamma \min_{1,2} V_{\theta^i_{\text{targ}}}(\mathbf{s}_{t+1}) - (1+\gamma)Q_\theta(\mathbf{s}_t,\mathbf{a}_t) \right)^2 \quad (7)$$

θ_{targ} are the weights of a target network, which is softly updated. To stabilize the network training a small artificial noise is added to the actions in the update. This algorithm has an extremely good sample-efficiency for suitable problems as can be found often in accelerators and is used as the baseline for the considered control problem, as it shows very good results. An illustration and additional changes to the original proposal [11] and a previous implementation used for experiments using a prioritized replay buffer [14] are discussed in appendix A 1.

V. UNCERTAINTY AWARE DYNA-STYLE REINFORCEMENT LEARNING

The original Dyna algorithm [15] is modified here in several aspects. Generally, Dyna style algorithms [16] denote algorithms, where a MFRL algorithm is trained on purely synthetic data from an approximate dynamics model or on a mixture of synthetic and real data. We use only synthetic data to reduce the interaction with the real environment to a minimum. An overview of the used method is shown in fig. 2. At the beginning the data is collected or read in from a pre-collection. An uncertainty aware model is trained, using anchored ensembles [17] on the data, which allows to take the allegorical (measurement errors) as well as the epistemic uncertainty (lack of data) into account. Subsequently a MFRL algorithm is deployed to learn the controller on the learned model by only using the synthetic data, by taking the uncertainty into account. After a defined number of training steps the controller is testes on each model individually. If there is no improvement in a specific ratio< 1 of models, the controller is tested on the real environment for a number of episodes. The training is stopped if the controller solves the problem on the real environment.

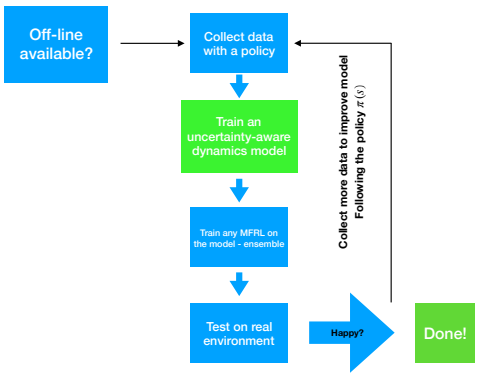

FIG. 2: A schematic overview of the AE-Dyna approach used in this paper.

A. Critical design decisions

In the following the most important aspects of a successful application of a Dyna-AE algorithm are listed:

- The ANN - including the prior - number of models - noise level - early stopping
- The number of data points and the policy.
- The uncertainty - how is this included?
- The episodic design - avoid long trajectories.
- The MFRL agent as e.g. the TRPO and PPO or the SAC and its training
- Tuning the algorithm on a model.

We try to discuss all items in some detail in the following. The anchoring ensemble methods usually yields good results already with a small number of different networks. The idea behind it, is to mimic the posterior probability of the dynamics model to capture first of all its own uncertainty due to the lack of data. Empirical results show that three to five models were sufficient to see clear advantages over a single network approach and the main goal is not to calculate the exact posterior of the model. A small two-layer network with around 15-25 nodes and the *tanh* activation were used. The last layer was linear and the inputs were normalized to the interval [-1,1]. The prior is controlled by the initialization of the weights, which were normalized by the number of nodes in each layer. The noise level is also respected

during the initialization. For the moment only homeoscetastic errors are used. Several methods to set the training method were implemented. Early stopping, a standard techniques showed good results, mainly to the fact, that the number of training steps is increasing with more data. The advantage is that the uncertainty at regions without data shrink in this way leading to a better training performance. A fixed loss value threshold was also tested, as also used in the original anchore-ensemble technique. In the experiments a combination of both was taken.
One of the most crucial points is how this uncertainties are taken into account by the RL algorithms. Several different approaches were tried. It is possible to take the average of the models and add Gaussian noise leveled by the standard deviation of the models. Another straight forward way is to randomly select a model to provide at each training step, which is the original implementation of the ME-TRPO algorithm and was used in the experiments labeled *ME-TRPO*. A pessimistic setting only would select the model resulting in the lowest predicted reward. Good results were obtained by following a randomly selected single model each full episode and were used in the experiments labeled *AE-Dyna*.
The number of data-points taken every time the model as improved was firstly determined by the number of initial data points, collect using a random policy. The initialization phase has to be chosen not to small to minimize the chance of getting trapped in a local minimum for too long. Afterwards at least on full episode should be taken. We decided to use a short total episode length to reduce the impact of the compound error, the accumulation error following a wrong model, as well known for Dyna-style approaches. The maximal number of steps was ten. During the data collected on the real system, the latest policy was taken with some small noise added to improve the exploration.
To decide which MFRL algorithm to use, two main algorithm classes have to be considered: the on and the off-policy algorithms. Online algorithm show a more stable convergence in general, while off-policy algorithms have the advantage that the data buffer can be filled with real data as well. A very attractive off-policy algorithm is the Soft-actor-critic. This algorithm tries to maximize the entropy of the actions to find a good trade-off between exploration and exploitation. On the other hand the on-policy algorithm TRPO provides some theoretical improvement guarantees. In the latter case the policy was improved over the whole training, while for the SAC the policy was reset, to take advantage of the exploration features.

B. Experiment results from FERMI RL tests

As discussed in the previous sections, several tests were performed on the FERMI XFEL. The main purpose was to test the newly implemented algorithms on a real system to evaluate their operational feasibility. The *NAF2* algorithm, as a representative for highly sample efficient MFRL algorithms, was tested first.

C. MFRL tests

In total four tests were carried out, two using a single network and two using the double network architecture. In both cases *smoothing* was applied, as described in appendix A 1.
Figure 3 displays the results, averaged over the two test. A training of 100 episodes was done. In the upper figure the number of iterations per episode is plotted including the cumulative number of steps. In the verification fig. 4 both algorithms show a similar performance, while the double network needs less training steps, and reveals a more stable overall performance.
Additionally the convergence metrics of the two algorithms is plotted in fig. 5 against the number of training steps. The blue curves shows the Bellmann error, which is comparable. The state-value function, which is a direct output of the neural net (eq. (6)), converges to a reasonable value for the double network within the shown 700 steps, whereas the single network tends to overestimate the value. In this case convergence is reached 1400 steps.

D. MBRL tests

The second test campaign was employing the *AE-DYNA* algorithm, as representative for pure a MBRL algorithms. Two variants were implemented: the *ME-TRPO* variant and the *AE-DYNA SAC* variant. In both cases an ensemble of three network was used and the epistemic and aleatoric uncertainty is measured via the anchor-ensemble technique as discussed in section V A. On top early stopping was used.
The first uses the trust region policy optimization - *TRPO*- [5] to train the controller. The TRPO monotonically converges to better policy and this property is exploited in the training. The convergence property can be seen in fig. 8. In the upper figure the total reward per batch is shown, as well as the number of data points used in the training of the dynamics model. In the lower plot the average cumulative reward as achieved by the TRPO on the individual models in the ensemble on a number of 10 episodes is drawn. The shade area shows the corresponding standard deviation to indicate the uncertainty of the dynamics model. To measure the convergence of the TRPO the logarithm of the standard deviation of the distribution drawn trajectories p_τ is also visualized. Here the training was stopped after 450 steps acquiring 25 steps each dynamics training. In the verification, as can be seen in fig. 9, all of the 50 episodes where successfully finished after a few steps.
Secondly, the *AE-DYNA SAC* was tested, which uses

the soft actor critic -SAC- algorithm [18]. The SAC not only tries to maximise J eq. (1), but also simultaneously but weights on the maximisation of the entropy in the action space. In this way exploration is encouraged to avoid getting stuck in a local optimum. In this test the controller is reset each time, when new data for the dynamics model training is acquired. Consequently, the performance drops each time, the dynamics model was retrained, as shown in fig. 6. Chances to get trapped in a local optimum are smaller by using this strategy of training. In difference to the first MBRL test, the batches of when acquiring new data are 50 with an initial random walk of 100 steps. The number of initial steps was chosen carefully large enough, because the convergence is slowed down and there is the risk that the training becomes unfeasibly. The training was stopped after the acquisition of 500 data points and a verification was done as in the first test. Again the success is 100%, but the number of iterations per step is smaller. It might be a result of the bigger number of data points, but in general, this method showed a better asymptotic performance than the $ME-TRPO$ variant.

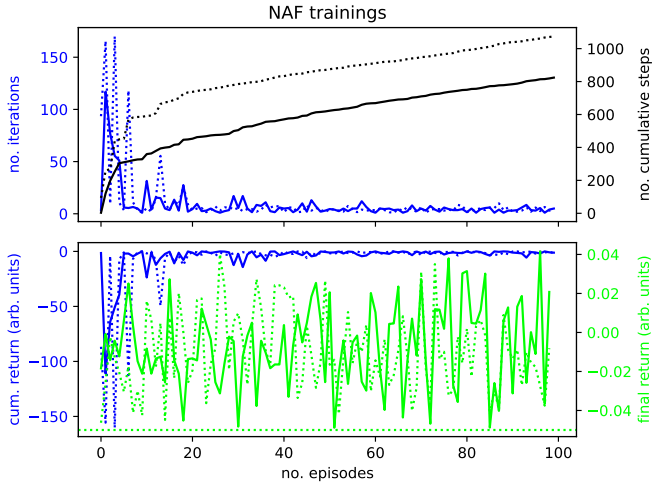

FIG. 3: The training of the *NAF2* on the FERMI FEL.

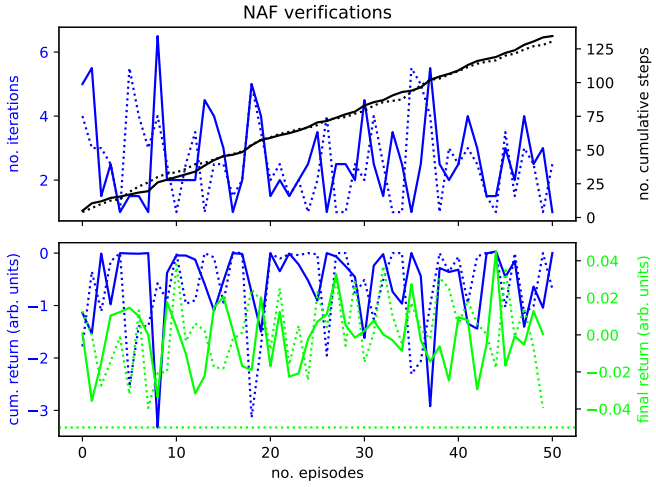

FIG. 4: The verification of the *NAF2* on the FERMI FEL.

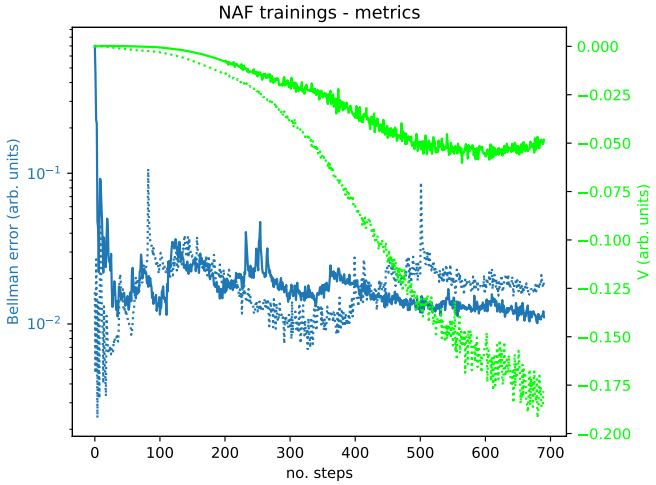

FIG. 5: The verification of the AE-DYNA-SAC on the FERMI FEL.

VI. DISCUSSION AND OUTLOOK

VII. CONCLUSIONS

Appendix A: A non-linear standard control problem

To provide some transparency of these studies for other labs we provide results on a famous classical standard control problem [19], the *inverted pendulum*. It is a non-linear low dimensional unsolved continuous control problem. Unsolved means there is no threshold for the reward to terminate an episode. The episode length is set to 200 steps. In the following several tests were carried out on the *inverted pendulum* to demonstrate the improvements of the selected algorithms, mainly in terms of noise handling. This is of importance when dealing with measurements on real systems.

1. NAF2 details

We compare the different *NAF* variants: *smoothing-double*, *smoothing single*, *standart*. The smoothing adds a small clipped noise on the actions to stabilize the network training as:

$$a'(s') = \text{clip}\left(\mu_{\theta_{\text{targ}}}(s') + \text{clip}(\epsilon, -c, c), a_{Low}, a_{High}\right), \tag{A1}$$

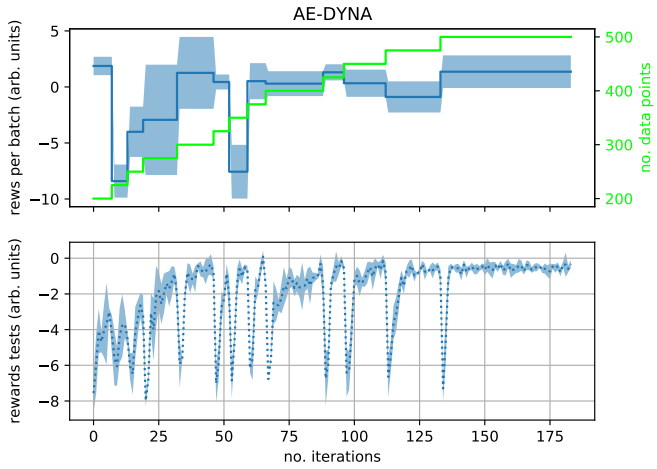

FIG. 6: The training of the AE-DYNA-SAC on the FERMI FEL.

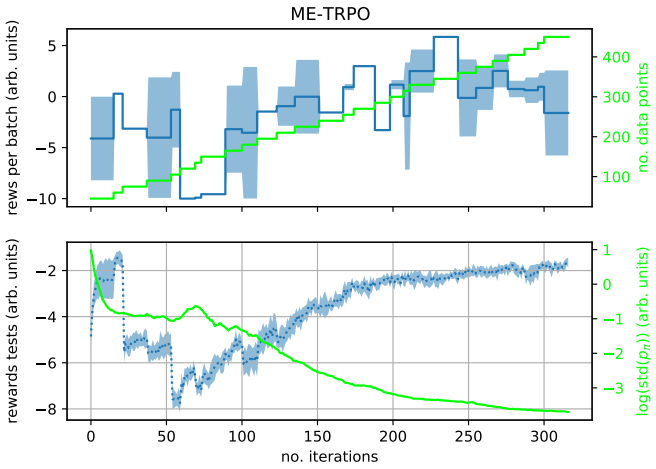

FIG. 8: The training of the ME-TRPO on the FERMI FEL.

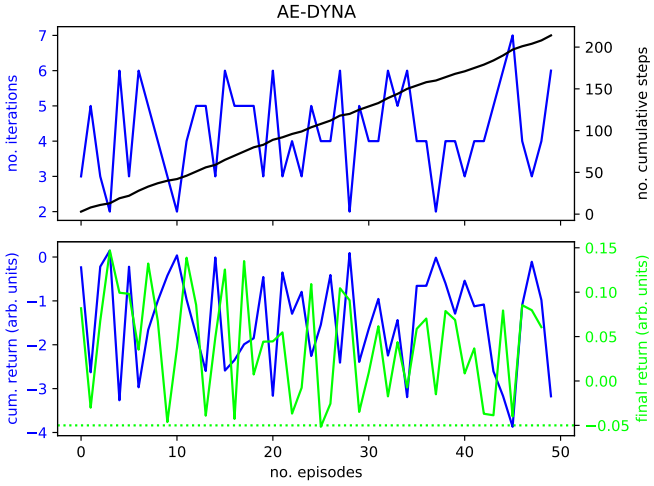

FIG. 7: The verification of the AE-DYNA-SAC on the FERMI FEL.

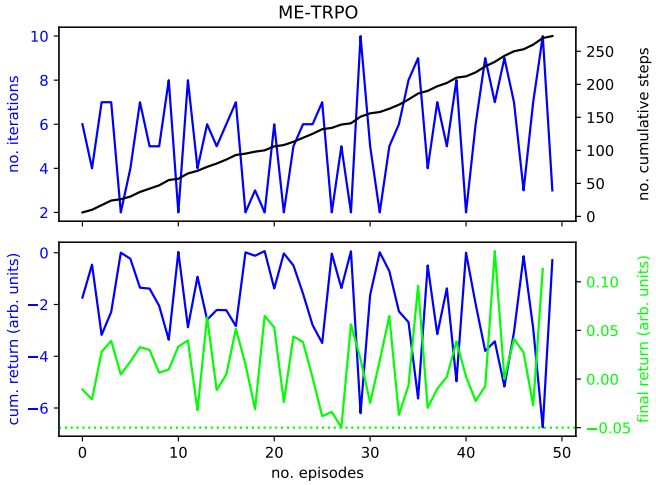

FIG. 9: The verification of the ME-TRPO on the FERMI FEL.

where $\epsilon \sim \mathcal{N}(0, \sigma)$. This method was used already in [20] to improve the deterministic policy gradient [21]. The double network was used in [18, 20] and is done in the following way:

$$y(r, s', d) = r + \gamma(1 - d) \min_{i=1,2} Q_{\phi_{i,\text{targ}}}(s', a'(s')) \quad \text{(A2)}$$

and then both are learned by regressing to this target:

$$L(\phi_1, \mathcal{D}) = \mathbb{E}_{(s,a,r,s',d)\sim\mathcal{D}} \left(Q_{\phi_1}(s, a) - y(r, s', d) \right)^2 \quad \text{(A3)}$$

and the policy is obtained via:

$$\max_{\theta} \mathop{\mathrm{E}}_{s\sim\mathcal{D}} \left[Q_{\phi_1}(s, \mu_\theta(s)) \right]. \quad \text{(A4)}$$

The results are shown in fig. 10. One sees the cumulative reward per episode for a training of a total of 50 episodes. The curve labelled *clipping* corresponds to the double network including the smoothing method and shows the best stability during the training yielding quickly a high reward. Also the single network shows good and comparable performance, except for the stability. The worst performance is achieved without smoothing and a single network, nevertheless the result is competing with state of the art model free methods as [22] as the benchmark in the *leaderboard* of openai gym [23].

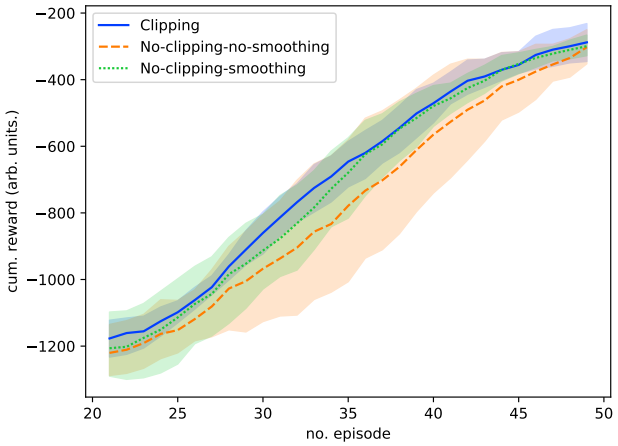

FIG. 10: Cumulative reward of different NAF implementations as discussed in the text.

2. The impact of noise

A test adding artificial Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma)$ with $\sigma = 0.05$ in the normalized observation space on the states is presented in fig. 11. There the difference of the three methods becomes even more evident.The results are shown in fig. 11. After around 65 episodes the single network without smoothing decreases before reaching the final performance at around 95 episodes, while the smoothing prevents this performance drop. Using the anchors in the dynamics model stabilizes the training of the *AE-DYNA* as illustrated in fig. 13. It the upper plot the mean cumulative reward on 10 test episodes on the real environment is shown during the training. It should indicate the result if the training is stopped at this training iteration. One cannot observe this during a real training, unless one does costly performance measurements during the training. The blue curve reaches the target much quicker exhibiting less variation.
The lower plot of fig. 13 shows the batch rewards a measured during the data collection, which is observable during the training.

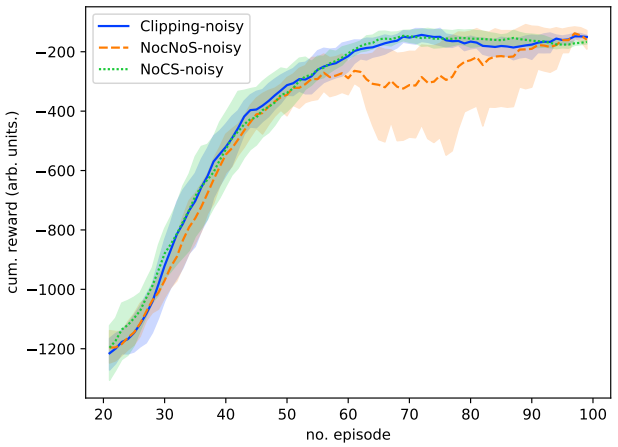

FIG. 11: Cumulative reward of different NAF implementations with artificial noise as discussed in the text.

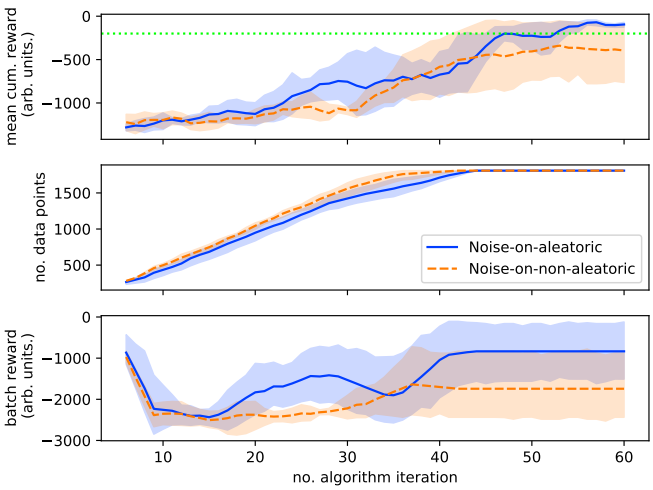

FIG. 12: Cumulative reward of AE-DYNA with artificial noise using the anchor.

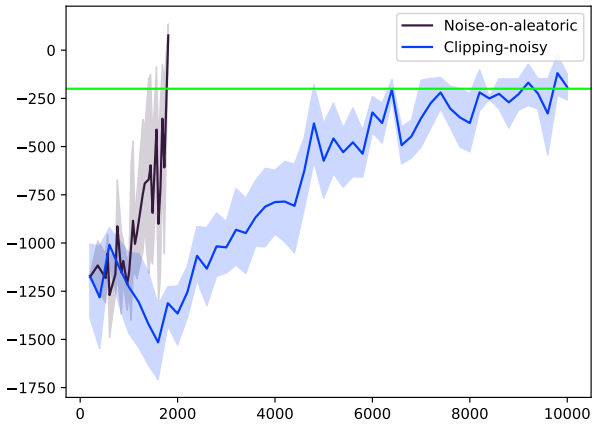

FIG. 13: The comparison of the NAF2 and the AE-DYNA on the noisy pendulum.

Appendix B: Theoretical aspects on the AE-DYNA on FERMI FEL

In this section we want to discuss some theoretical aspects concerning the AE-DYNA approach. A simulation

was used to obtain the presented analysis.
One issue of MBRL methods is the asymptotic performance. Altough good results were obtained using the AE-DYNA, there is a difference between the MFRL and MBRL. We call it the *reality gap*. There are ideas to attack the problem e.g. by learning a meta-policy [24] followed by a fine tuning on the residual physics, which minimizes the *reality gap* ot training on a simulator and closing than with a small amount of training iterations on the real system the gap [25].

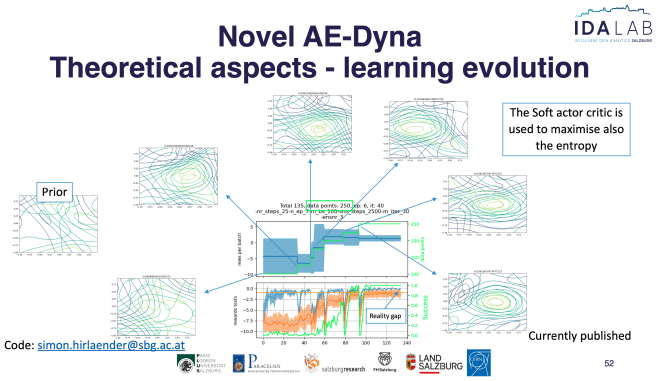

FIG. 14: The training .

[1] A. G. C.-D. A. L. L. B. Richard S. (University of Alberta) Sutton, *Reinforcement Learning* (MIT Press Ltd, 2018).

[2] R. J. Williams, Simple statistical gradient-following algorithms for connectionist reinforcement learning, Machine Learning **8**, 229 (1992).

[3] J. Baxter and P. L. Bartlett, Infinite-horizon policy-gradient estimation 10.1613/jair.806, 1106.0665.

[4] S. Levine and V. Koltun, Guided Policy Search, in *Proceedings of the 30th International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 28, edited by S. Dasgupta and D. McAllester (PMLR, Atlanta, Georgia, USA, 2013) pp. 1–9.

[5] J. Schulman, S. Levine, P. Moritz, M. I. Jordan, and P. Abbeel, Trust region policy optimization, 1502.05477.

[6] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, Proximal policy optimization algorithms, CoRR **abs/1707.06347** (2017), arXiv:1707.06347.

[7] S. Levine, A. Kumar, G. Tucker, and J. Fu, Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems, 2005.01643.

[8] H. van Hasselt, A. Guez, and D. Silver, Deep reinforcement learning with double q-learning, 1509.06461.

[9] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, Playing atari with deep reinforcement learning, 1312.5602.

[10] 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, 1509.02971.

[11] S. Gu, T. Lillicrap, I. Sutskever, and S. Levine, Continuous deep q-learning with model-based acceleration, 1603.00748.

[12] Z. Wang, T. Schaul, M. Hessel, H. van Hasselt, M. Lanctot, and N. de Freitas, Dueling network architectures for deep reinforcement learning, 1511.06581.

[13] H. Hasselt, Double q-learning, in *Advances in Neural Information Processing Systems*, Vol. 23, edited by J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta (Curran Associates, Inc., 2010) pp. 2613–2621.

[14] S. Hirlaender, Mathphyssim/per-naf: Initial release (2020).

[15] T. Kurutach, I. Clavera, Y. Duan, A. Tamar, and P. Abbeel, Model-ensemble trust-region policy optimization, 1802.10592.

[16] R. S. Sutton, Dyna, an integrated architecture for learning, planning, and reacting, ACM SIGART Bulletin **2**, 160 (1991).

[17] T. Pearce, F. Leibfried, A. Brintrup, M. Zaki, and A. Neely, Uncertainty in neural networks: Approximately bayesian ensembling, 1810.05546.

[18] T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. Ha, J. Tan, V. Kumar, H. Zhu, A. Gupta, P. Abbeel, and S. Levine, Soft actor-critic algorithms and applications, 1812.05905.

[19] K. Furuta, M. Yamakita, and S. Kobayashi, Swing up control of inverted pendulum, in *Proceedings IECON '91: 1991 International Conference on Industrial Electronics, Control and Instrumentation* (IEEE).

[20] S. Fujimoto, H. van Hoof, and D. Meger, Addressing function approximation error in actor-critic methods (2018), arXiv:1802.09477 [cs.AI].

[21] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, Deterministic policy gradient algorithms, in *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14 (JMLR.org, 2014) p. I–387–I–395.

[22] G. Barth-Maron, M. W. Hoffman, D. Budden, W. Dabney, D. Horgan, D. TB, A. Muldal, N. Heess, and T. Lillicrap, Distributed distributional deterministic policy gradients, 1804.08617.

[23] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, Openai gym, 1606.01540.

[24] I. Clavera, J. Rothfuss, J. Schulman, Y. Fujita, T. Asfour, and P. Abbeel, Model-based reinforcement learning via meta-policy optimization, 1809.05214.

[25] A. Zeng, S. Song, J. Lee, A. Rodriguez, and T. Funkhouser, Tossingbot: Learning to throw arbitrary objects with residual physics, 1903.11239.