Informed Sampling the Plasma Hyperspace for Surrogate Modelling

Mayur Bakrania[1] and Vignesh Gopakumar[2]

[1]University College London
[2]UK Atomic Energy Authority
[2]*vignesh.gopakumar@ukaea.uk*

September 15, 2021

Abstract

Digital twins capable of predicting plasma evolution ahead of plasma progression within a Tokamak is a crucial tool required for real-time plasma intervention and control. Considering speed and scale required, quite often these have to be purely data conditioned models as opposed to being physics conditioned, making data selection a vital component of model efficacy. However, as we move to the exascale regime, the amount of data generated tends to choke the data pipelines, introducing latency to the model. It might also be the case that some of the data available might be redundant and creating imbalances within the training dataset. In this work we demonstrate a machine learning pipeline that maps out in hyperspace the distributions of the plasma behaviours within a specific campaign. The embedding created through dimensionality reduction within the pipeline is then used as the sampling space for the training dataset for a Convolutional LSTM that maps the control signals to diagnostic signals in a sequential manner. We primarily experiment with MAST data with the control signals being plasma current, toroidal magnetic field, plasma shape, gas fuelling and auxiliary heating. The diagnostics of interest are the core density and temperature as measured by the Thomson scattering diagnostic. With initial focus on a single experimental campaign (M7), we demonstrate that our predictive model trained on all available data is capable of achieving a mean squared error of 0.0285. However, our pipeline demonstrates that by using a distance based informed sampling method to gather only 10 percent of the dataset we can achieve a comparable mean squared error of 0.0293.

We further demonstrate the robustness of the pipeline by extending the model to operate within the space of the M9 campaign in addition to the M7 campaign. Our work shows that a predictive model trained on all of the available data across both campaigns achieves a mean squared error of 0.0279, while the one sampled using the knowledge garnered from the cluster representations (mapped individually across each campaign) achieves an L2 error of 0.0282, while only relying on 10 percent of the dataset. We also engage with standard continual learning practices such as elastic weight consolidation and deep generative replay to move the model across various campaign data and cross reference against our pipeline.

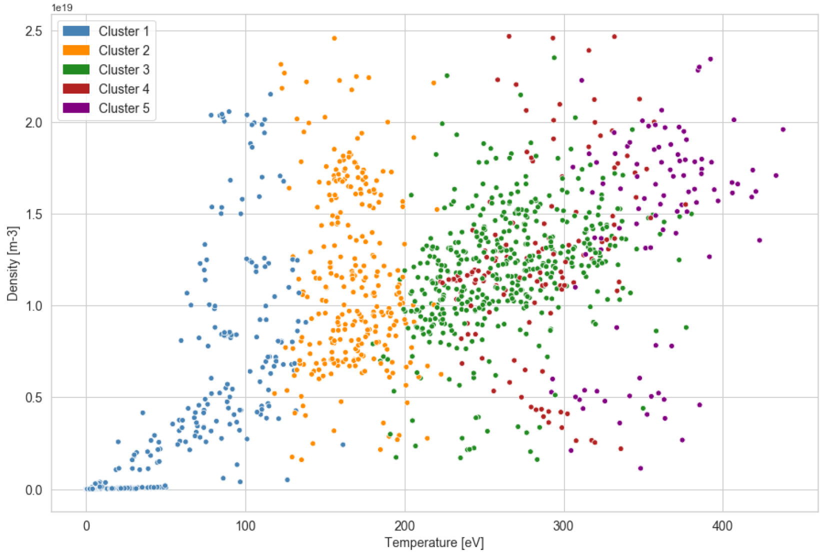

Figure 1: Average density vs average temperature across the M7 Campaign. Our dimensionality reduction approaches clusters together similar plasma behaviours across the campaign. This embedding in the hyperspace provides us with relevant information to curate an effective training dataset.