Neural network approximation of Bayesian models for the inference of ion and electron temperature profiles at W7-X

A Pavone1, J Svensson1, A Langenberg1, U Höfel1, S Kwak1, N Pablant2, R C Wolf1 and the Wendelstein 7-X Team1

1 Max-Planck-Institut für Plasmaphysik, Teilinstitut Greifswald, D-17491 Greifswald, DE, Germany
2 Princeton Plasma Physics Laboratory, 08540 Princeton, NJ, United States of America

E-mail: andrea.pavone@ipp.mpg.de

Received 22 January 2019, revised 8 March 2019
Accepted for publication 26 April 2019
Published 30 May 2019

Abstract

In this paper, we describe a method for training a neural network (NN) to approximate the full model Bayesian inference of plasma profiles from x-ray imaging diagnostic measurements. The modeling is carried out within the Minerva Bayesian modeling framework where models are defined as a set of assumptions, prior beliefs on parameter values and physics knowledge. The goal is to use NNs for fast ion and electron temperature profile inversion from measured image data. The NN is trained solely on artificial data generated by sampling from the joint distribution of the free parameters and model predictions. The training is carried out in such a way that the mapping learned by the network constitutes an approximation of the full model Bayesian inference. The analysis is carried out on images constituted of 20 × 195 pixels corresponding to binned lines of sight and spectral channels, respectively. Through the full model inference, it is possible to infer electron and ion temperature profiles as well as impurity density profiles. When the network is used for the inference of the temperature profiles, the analysis time can be reduced down to a few tens of microseconds for a single time point, which is a drastic improvement if compared to the ≈4 h long Bayesian inference. The procedure developed for the generation of the training set does not rely on diagnostic-specific features, and therefore it is in principle applicable to any other model developed within the Minerva framework. The trained NN has been tested on data collected during the first operational campaign at W7-X, and compared to the full model Bayesian inference results.

Keywords: stellarator, x-ray imaging, neural network, Bayesian inference, modeling, Minerva framework, real time

(Some figures may appear in colour only in the online journal)

1. Introduction

Neural networks (NNs) are a powerful tool when it comes to speed and approximation of complex functions. universal approximation theorems have been shown to be valid for NNs under different assumptions, as in [1, 2] and [3]. The real time capabilities of NNs have also been shown in different fusion experiments, e.g. ion temperature profile inference and disruption prediction at JET as in [4–6] and at ASDEX Upgrade [7]. NNs have also been used for the reconstruction of plasma parameters from diagnostic data as in the case of charge exchange spectra automatic analysis at JET for reconstruction of ion temperature, rotation velocity and impurity density [8, 9], and in the case of electron temperature from a soft-x-ray system at NSTX [10]. In this paper, we focus on an
application of NN algorithms on x-ray imaging crystal spectrometer (XICS) measurements and we will describe an approach based on a different paradigm of NN training and reconstruction suitable when the physics model of the diagnostic is available. In particular, we will make use of the Bayesian implementation of the model within the Minerva modeling framework [11].

The Minerva framework provides a language to develop models of complex systems in a Bayesian way: models implemented within it share the same structure and are highly modular, so that different operations can be performed on them independently of the specific model under consideration. These operations can be the inference of the free parameter posterior distributions through different optimization and sampling algorithms, sampling schemes—as the one we make use of in this work, and storage/sharing. The elements/modules of a Minerva model are called nodes: these can be pieces of code that perform computation of physical quantities and they are such that they can be easily replaced. This makes building new models out of existing nodes particularly easy, as well as testing different models simply by switching their nodes. In this work, we have taken advantage of the modularity and shared structure of the models so that the method described here is general and can be easily applied to different problems modeled within the same framework.

NNs applied to diagnostic data are typically trained on real measurements and the corresponding quantities of interest in situations where a model of the problem is missing. Such an approach has the advantage of providing the NN with actually measured data, but it also has the limitation of depending on a fixed and restricted amount of training samples, the feature of which depends on the performed experiments, and on a limited parameter space. A different case is illustrated in [9]: here the authors have first reconstructed the distribution of a large set of measured physics parameters; then new parameters have been sampled from it, and used as input in a forward model in order to create simulated measurements. A NN was then trained on both measured and simulated data.

The way we try to overcome these limitations is by training the network solely on data synthesized through the Bayesian model specified within Minerva, the same that is used for the standard inference [12]. The parameters, corresponding to physics quantities, used to produce the data are sampled from the corresponding prior distributions and the synthesized observations are sampled taking into account the error model. This conferes control over the features we put in the training set and, consequently, the features the NN will be learning and will be sensitive to when evaluated on measured data. In addition, an advantage of this approach concerns its generality and the possibility of performing automatic data analysis based on physics models, which comes as a consequence of the sampling procedure described in the following chapters. This becomes of greater relevance as the scale of fusion experiments grows larger and the duration of plasma shots becomes longer. During the first operation campaign (OP 1.1, see [13]) of the W7-X stellarator several diagnostics [14, 15] were involved in the measurements, and the number increased during the second one (OP 1.2a). Together with the number of diagnostics, a large proportion of which is currently implemented in the Minerva framework, e.g. [12, 16, 17], also the duration of the plasma shots increased. All of this makes fast and automatic data analysis very desirable. The technical implementation of our approach only makes use of features shared between all models in Minerva, and thus it is easily transferable and applicable to other diagnostics modeled in the framework.

The paper is organized as follows: in section 2 we give a general outline of the core idea behind the training method, that is the sampling scheme from the Bayesian model; this constitutes the main piece of novelty of this work. The specific details necessary to understand each part of the entire scheme are then delivered in the following sections. In section 3, we give a brief overview of the hardware setup and the physics involved in the measurements performed with the x-rays imaging diagnostic. It is followed by an explanation of Bayesian modeling in section 5, which starts generally from Bayes theorem and goes into more specific details of the Minerva implementation of the diagnostic model. In section 6 we illustrate the sampling scheme used to generate the training data from the XICS Bayesian model, whereas details about the NN architecture and training algorithm are given in section 7. In section 8, we will show how training sets generated with different models can be compared to each other in order to be able to choose the one that better describes the measurements; this is particularly relevant in the context of this work since the training data are generated synthetically. In section 9 we compare the NN reconstructed profiles to those traditionally inferred with the Minerva Bayesian model in the case of measurements performed during the first experimental campaign at the W7-X experiment. In section 10 we draw some conclusive considerations.

2. The method

The core of the method that we want to illustrate in this section is the sampling scheme used to generate the NN training data from the Minerva Bayesian model of the XICS diagnostic measurements. In order to understand how it is practically carried out, it is necessary to have some basic notions about the XICS diagnostic and the Minerva modeling framework. Details about the former are given in section 3, while the latter is described in section 5.

An outline of the idea behind the method is the following: the XICS Bayesian model is a model whose key elements are the free parameters—the ion and electron temperature profiles—and the observations, the XICS measured images. Before any measurement is taken into account, a prior distribution is assigned to the free parameters, and a likelihood function is assigned to the observations. The former is usually a broad distribution which may include features we believe are found in measured plasma profiles, whereas the latter reflects the uncertainties of the model in predicting the data, which typically include measurement errors. The product of the two probability distributions defines the joint probability
distribution of the model. The joint distribution completely describes the model, both from the point of view of its statistical properties—residing in the fact that it is a probabilistic distribution, and from the point of view of the physics implemented within it—found in the relation between free parameters and observations. Since we aim at training the NN to be an approximate replica of the full Bayesian model, we can create the training data set by sampling the free parameters and synthetic observations from the joint distribution. Afterwards, we train the NN on the inverse problem, which is the mapping of the observed XICS images to the ion and electron temperature profiles. We expect that, under the assumption that the training data properly resemble the measurements, the network can be used for fast reconstruction of temperature profiles from new experimental data. The goodness of this assumption for a given training data set can be tested quantitatively, so that it is possible to compare different models in their ability to produce data set that resemble the experimental measurements. In section 8 we describe a possible way to test this assumption.

3. XICS diagnostic

The XICS collects x-rays emitted in atomic processes involving ion impurities and plasma electrons occurring in the bean shaped cross section of the W7-X stellarator. The concept behind the diagnostic is described in [18]. A sketch of the view is shown in figure 1(a), where the plasma cross section and the line of sight (LOS) span are shown. For each LOS an integrated spectrum is collected. The set of all spectra forms one measured image. From an image it is possible to reconstruct plasma ion and electron temperature profiles, and impurity density profiles. In section 3.1 we will describe the setup of the diagnostic and in section 3.2 we will describe the physics involved in the measurements.

3.1. Setup

The system is equipped with a spherical bent crystal and the light is collected onto a CCD detector producing 2D images similar to the one shown in figure 1(b). The two dimensions in the image represent energy and spatial resolution respectively. The diagnostic is sensitive to the energy region of He-like Argon emission lines. The main emission lines constituting the spectrum are shown in figure 1(c) and they are the w, x, y and z for the \( n = 2 \) to \( n = 1 \) transitions in addition to numerous \( n \geq 2 \) dielectronic satellites, e.g. the \( k \) lines for \( n = 2 \). A study of the spectrum and the atomic processes involved can be found in [19–21]. The detector covers the wavelength range from \( \approx 3.94 \) to \( \approx 4.00 \) Å along the central LOS.

3.2. Physics

The core component of the physics processes involved in the measurements is given by the atomic processes giving rise to the spectral emission. A detailed description of the calculation of the emission intensity for the different lines can be found in

Figure 1. (a) Sketch of the XICS system view on the bean shaped cross section at W7-X. The viewing span of the lines of sight is represented by the ocher area. (b) A measured raw image detected on the CCD detector. The typical curved feature is due to the spherical bending of the crystal. (c) He-like Argon spectrum measured along one of the central lines of sight of the image in figure (b). The main emission lines are marked with their names.

| Atomic process | Plasma parameter dependency |
|----------------|-----------------------------|
| (1) Excitation from ground state of He-like ions | \( n_e, n_{\text{Atm}}, T_e, T_i \) |
| (2) Di-electronic recombination of He-like ions | \( n_e, n_{\text{Atm}}, T_e, T_i \) |
| (3) Recombination of H-like ions | \( n_e, n_{\text{Atm}}, T_e, T_i \) |
| (4) Inner shell excitation of Li-like ions | \( n_e, n_{\text{Atm}}, T_e, T_i \) |
reproduce the measurements in their absolute values. Therefore, both simulated and measured images are normalized dividing the intensity of each pixel by that of the brightest one.

### 4. Bayesian inference

When developing a Bayesian modeling and inference scheme, the first step is the definition of the model free parameters, \( \lambda \), and observed data, \( d \). Probability distributions are assigned to both quantities, and they take the name of prior distribution, \( P(\lambda) \), and likelihood function, \( P(d|\lambda) \). The prior distribution represents the \textit{a priori} knowledge that we have about the free parameters before taking the observations into account. The likelihood distribution represents instead the model uncertainties in the prediction of the data. The inference is the process of knowledge acquisition when new data are observed. According to Bayesian probability theory, it can be described as an update process of the \textit{a priori} distribution. This process is formally expressed through Bayes formula:

\[
P(\lambda|d) = \frac{P(d|\lambda)P(\lambda)}{P(d)}.
\]  

The numerator in equation (4.1) corresponds to the joint distribution \( P(d, \lambda) \) of the observed data and free parameters, \( P(d, \lambda) = P(d|\lambda)P(\lambda) \). The term \( P(d|\lambda) \) is called posterior distribution of the free parameter \( \lambda \) given the observed data \( d \), and it represents the new state of knowledge on the model free parameters as new data are collected. The quantity \( P(\lambda) \) is called evidence or prior predictive and, as first interpretation, it plays the role of a normalization factor:

\[
P(\lambda|d) = \int P(d|\lambda)P(\lambda)\,d\lambda.
\]  

The hidden relevance of this term becomes evident when we switch from Bayesian inference for parameter estimation to Bayesian model selection, see for example [22]. Since the integral in equation (4.2) is a marginalization over the free parameters of a given model, we see that \( P(\lambda|d) \) is the distribution of all the data that a model can describe, quantifying how likely each data point is to be generated under the assumptions of the model. Once it is evaluated on a data point \( d^* \), the evidence \( P(d^*) \) defines how likely our model is as an explanation of such data point. The value of this probability depends on the prior distributions of the parameters and the uncertainties attributed to the model prediction in the likelihood term. The dependency is such that broader prior distributions, or prior distributions defined over higher dimensional parameter spaces, i.e. more complex models, will be penalized, and the overall probability will always constitute a trade-off between model complexity and good fit to the data. This is indeed an application of the \textit{Occam’s razor} principle. An illuminating explanation of how Bayesian model selection naturally embodies Occam’s razor is given in [23].

As data are measured, Bayesian inference can be carried out to extract information about the free parameters. The
target of Bayesian inference is the posterior distribution of the parameters, \( P(w|d) \), the spread of which expresses the uncertainties on the inferred quantities.

5. The Minerva framework

In order to interpret the measured data, the physics describing the processes giving rise to the measurements is implemented in a forward model of the diagnostic within the Minerva framework [12, 24]. The Minerva modeling framework is a framework for the development of Bayesian models of complex systems. Within the framework it is possible to carry out both modeling and inference. The result of the modeling is an object that is called a graph. It contains all information about the physics and the probabilistic relations between the modeled quantities. It describes all the hypotheses used in the creation of the Bayesian model. A given model can be used to produce simulated observations and compare them to the measured observations. In this way, the model free parameters that better describe the data can be found. The advantage of doing Bayesian inference is that it formally defines a procedure to calculate the uncertainties of a solution, which simply involves the application of Bayesian probability rules.

The Minerva framework relies on graphical models [25] in order to express the conditional dependency between random variables in the model. For each model implemented in the framework, a graph object is created that describes the joint distribution of the free parameter and the measurements according to the forward model. A simplified version of the XICS model graph is shown in figure 3. In the graph the colored nodes are probabilistic nodes, where orange denotes the free parameters and blue denotes the observed quantities.

In the case of the XICS forward model, the free parameters can be the plasma parameters summarized in table 1, right column, and the observed data are the images constructed of spectra like the one shown in figure 2, calculated accounting for the atomic processes described in section 2. All the distributions in the graph are chosen to be normal distributions. Note that the likelihood function is then a normal distribution centered on the model prediction obtained with a given set of free parameter values:

\[
P(d|w) = \frac{1}{\sqrt{2\pi}\sigma^2} \exp\left(-\frac{(d - y(w))^2}{2\sigma^2}\right).
\]

where we used \( y(w) \) to denote the forward model function \( y \) dependent on the free parameter values \( w \), and \( d \) to denote a measured data point. The white squared nodes in the graph of figure 3 represent deterministic calculation nodes, and the cloud node is used to denote a data source, i.e. a node that communicates with a database, here the W7-X ArchiveDB, where information about the diagnostic, e.g. geometry setup etc. are stored together with measured and analyzed data. The arrows represent dependencies between nodes rather than a computational flow. For example, all arrows from the free parameters reach, directly or indirectly, the observed node, i.e. the probability distributions of the observed quantities (\( d \)) should be conditioned on the value of the free parameters, \( P(d|w) \). The joint probability distribution represented by the whole graph can be factorized and written as: \( P(w, d) = P(w)P(d|w) \).

The \( T_e \), \( T_i \) and \( n_e \) profiles, where \( n_e \) stands for Argon ion or electron density, are functions of the normalized effective radius \( \rho_{\text{eff}} = \sqrt{\psi/\psi_{\text{LCS}}} \). Thus, an equilibrium code, in this case VMEC [26], called from the corresponding node, is required to carry out the mapping to the 3D Cartesian coordinates. The impurity emission can then be calculated locally and integrated along the lines of sight. A background emission is also added to the spectrum. In order to calculate the detector pixel response, the instrumental function and the wavelength dispersion on the chip are also required. The data source node is used to provide the observed data for the inference and the i.o.s. geometry. The smoothness of the profiles is controlled by a zero mean Gaussian process (GP) prior [27] with a squared exponential covariance function, as defined in:

\[
\text{cov}_f = \sigma_f^2 \exp\left(-\frac{(\rho_i - \rho_j)^2}{2\sigma^2}\right) + \delta_{ij}\sigma_y^2.
\]

The quantity \( \text{cov}_f \) denotes the elements of the covariance matrix, \( \rho_i \) and \( \rho_j \) denote the location of any two profile points labeled with \( i \) and \( j \), and the quantities \( \sigma_f, \sigma_e \) and \( \sigma_y \) denote the function variance, the length scale and the noise variance of the profiles, respectively. Note that the quantity \( \text{cov}_f \) corresponds to the covariance between any two points in the profile, as function of the location \( \rho_{\text{eff}} \). The length scale \( \sigma_e \) describes how smooth a function is. Small length scale values mean that function values can change quickly in their domain, whereas large values describe functions that change slowly. In our case, this refers to plasma profiles characterized by either quick or slow spatial changes. The function variance \( \sigma_f \) is a scaling factor and determines function value variations around the mean. Large values will allow again for bigger variations, smaller values will describe less varying functions. It also determines, together with \( \sigma_y \), the value of the covariance matrix elements along the diagonal, where \( i = j \). The noise variance \( \sigma_y \) is used to allow for noise present in the data and it specifies the amount of expected noise.

When measured data are available, inference can be performed on a Minerva model. In a practical implementation, when a single value solution is desirable, a Maximum A Posteriori optimizer can be used to find the maximum of the posterior distribution. The full Bayesian answer to the inference problem is nevertheless provided by the full posterior distribution, the spread of which expresses the uncertainties on the inferred quantities. In order to provide this information, a Markov Chain Monte Carlo (MCMC) sampler is usually adopted to generate the posterior samples. The samples can then be stored and used in further independent calculations providing full non-linear uncertainty propagation. Profiles inferred using XICS data can then be used, for example, in impurity transport studies, see [24, 28].
The model depicted in figure 3 represents a sophisticated inference problem: first of all, the images fed to the inversion routines go through a preprocessing stage, occurring in the data source node, where they are (1) straightened, and (2) binned along the LOS direction in order to reduce the computational effort. At this point, the inversion problem consists of simultaneously fitting an image of 20 × 195 pixels along the LOS direction and the wavelength dispersion direction respectively, and doing a tomographic inversion of different plasma profiles. The full Bayesian inference takes from 1 to 4 h for each measured image, whereas a NN can process data at a time scale of tens of microseconds in good implementation conditions (e.g. on a GPU).

6. Creation of the training set

In order to generate the NN training set, we will only make use of the XICS Minerva model. All the features of the model are expressed in its joint distribution \( P(d, w) = P(d|w)P(w) \); the distribution of the variables \( (d, w) \) depends on the functional form of the forward model, appearing in the likelihood term \( P(d|w) \) as \( y(w) \) (equation (5.1)) and which expresses the dependence of \( d \) on \( w \), the uncertainties on the model prediction and the prior distribution \( P(w) \). As our goal is to create a (approximated) copy of the original full Bayesian model, we must provide the NN with training data having the same properties described by the model: this is achieved by generating the training set data from samples of the full model joint distribution.

Practically, such training set can be obtained by iterating over the following three steps:

i. Draw and store a sample from the joint prior distribution of the free parameters: \( P(w) = P(T_e, T_i, n_k, bg) = P(T_e)P(T_i)P(n_k)P(bg) \), where \( bg \) denotes the background l.o.s. integrated emission profile

ii. Run the forward model in order to calculate a synthetic observation with the given free parameters

iii. Store a number of samples drawn from the likelihood function of the synthetic observations, \( P(d|w) \), which is fully specified by the given set of sampled free parameters and the model uncertainties.

The sampling procedure taking place at step (iii) will introduce noisy samples in the training set since the likelihood distribution expresses the uncertainties of the model prediction. This will help making the NN stable against small perturbations in the input data when evaluated on measured images. This is equivalent to the technique known as data augmentation, [29, 30]. The modifications we inject into the samples are based on the noise model that has been assumed for the problem, in this case a Gaussian noise model. The NN
Bayesian model, the distribution of the target data $t$ given an input vector $x$, $p(t|x)$, in the limit of large training data set, corresponds to the posterior distribution of the ion or electron temperature profile of the full Bayesian model. Therefore, we can state that the NN mapping, in ideal circumstances, is given by the mean of the full model posterior distribution. In real world circumstances, the data set size is finite and the optimization is never perfect, so that we can say that the inversion provided by a NN trained in such a way constitutes an approximation of the full model Bayesian inference.

7. NN input, output and architecture

It is worth to summarize here what the NN input and output are at the different stages. At training time:

- Input: synthetic images, generated with the XICS forward model and the sampling procedure described in section 6. These images supposedly closely resemble the actual XICS measurements (after few pre-processing operation, i.e. row binning and straightening, see next bullet point and figure 5). They are made of $20 \times 195$ pixels/values.
- Target: the ion or electron temperature profiles used to generate the corresponding images. These are made of 15 points along the effective radius $r_{\text{eff}} = \sqrt{\psi/\psi_{\text{LCFS}}}$, where $\psi$ is the magnetic flux and $\psi_{\text{LCFS}}$ is the flux at the LCFS. The points are obtained by sampling from the GP prior distributions of the temperature profiles.

At evaluation time:

- Input: the pre-processed, actual XICS measurements. The measured images, originally showing the curved feature shown in figure 1(b), are straightened according to the detector and crystal geometry. Afterwards, the original $1475 \times 195$ pixels of the image are binned along the largest dimension, which corresponds to the spatial resolution, where neighboring lines of sight overlap significantly. The binned images are made of $20 \times 195$ pixels. An example of a binned image is shown on the leftmost side of figure 5.
- Output: the estimated ion or electron temperature profiles. In case of successful training, it will match, within uncertainties, with the profile inferred through the full Bayesian model.

Essentially, two NNs, identical for all the features except for the target profiles, have been trained and tested independently: one for the inversion of ion temperature profiles and the other for the inversion of the electron temperature profiles.

Since the network’s input are 2D images, the network architecture has been inspired by the LeNet-5 convolutional neural network (CNN) [32] and is shown in figure 5. This kind of architecture has been shown to be particularly effective when the input present a 2D structure. It has been successfully used in many image recognition problems, achieving state-of-the-art results [33, 34]. Here we expect to recurrently find across the image the features induced in the spectrum by the ion or electron temperature profiles, which...
affect line width and intensities, respectively. Two convolutional layers \( C_1 \) and \( C_2 \), each one followed by one sub-sampling layer, are used in a hierarchical feature extraction structure. A convolutional layer applies a convolution filter or kernel to the input image, extracting information that are recurrent across different locations in the image (for more details see [32]). The kernel dimension sizes used in the convolutional layers are respectively: \((3, 16)\) and \((2, 5)\), where the first and second dimensions refer to the LOS and wavelength dispersion dimension of the input images. The number of feature maps, i.e. sets of units whose weights are constrained to be identical [35, 36], is set to 30 for both convolutional layers. The sub-sampling layers use \( \text{max pooling} \) with a resolution of 2 by 2. Two fully connected layers \( M_1 \) and \( M_2 \) made of 20 and 18 units respectively, constitute the final layers, which will produce the desired 15 points ion or electron temperature profile output. The activation function in the convolutional layers is the rectified linear unit:

\[
f(x) = \max(0, x), \tag{7.1}
\]

whereas in the fully connected layers it is the hyperbolic tangent function.

CNNs are also especially suitable for parallelization on GPUs, which applied to our case made the training 30 times faster than on CPUs. The NN was implemented within Theano, a Python framework for fast symbolic computation [37].

The training was stopped either according to an early stopping criterion, i.e. when the network performance on the test starts degrading, or when the decay rate of the loss function is small enough. The loss function that the NN is trained to minimize is defined as:

\[
S(w) = \frac{\beta}{2} \sum_{n=1}^{N} (y^n - t^n)^2 + \sum_{k,i=1}^{L_N} \alpha_k \omega_k^2, \tag{7.2}
\]

where \( w \) denotes the network weight vector, the first term on the right-hand side is the sum-of-square error between the network output \( y_n \) and the target \( t_n \) and \( n \) is an index that goes through the \( N \) samples in the training set. The second term on the right-hand side is a regularizing term, where \( \omega_k \) denotes the network weight of unit \( i \) at layer \( k \). The parameters \( \beta \) and \( \alpha_k \) are scale parameters which control the relative importance of the two terms. An insightful interpretation, based on a Bayesian view of the NN training [23, 38], can be provided to such expression and can be useful in the choice of the values of \( \beta \) and \( \alpha_k \). The values of \( \beta = 10 \) and \( \alpha_k = (\alpha_{C1} = 68.00, \alpha_{C2} = 58.33, \alpha_{M1} = 576.67, \alpha_{M2} = 5.83) \) were used. More details about Bayesian NN training in the context of this study can be found in [39].

8. Training set comparison

When dealing with NN reconstructions, the following situation is likely to occur: the network might be able to reconstruct the target profiles from training data, but it might fail when applied to experimental data. One reason why this might happen is the following: the features of the training data do not resemble closely enough those found in the measurements. In this section, we illustrate how training data sets generated with different models can be compared to each other from the point of view of their adequacy in describing the experimental measurements. In particular, we will consider two cases: the case of a data set

![Figure 5. Architecture of the NN used. The input layer at the leftmost side is followed by a convolutional layer and a sub-sampling layer, which are followed again by a couple of convolutional and sub-sampling layers. Two fully connected feed forward layers follow up to the output layer. Each blue plane represents a feature map, where all the units share the same weights.](image)
generated by sampling unconstrained profiles that leads to a poorly performing NN, and the case of a data set where a realistic constraint is applied to the prior distributions, which in turn help improving the performance of the network. In general, one would try to introduce as little bias as possible to a given training set. As it is shown here, this does not always lead to a successful result. The reason of the failure can therefore be assessed with the method described here. It can also be used as a novelty detection system: it can be used to identify a novel incoming measured image as an outlier, i.e. a case that is relatively unknown to the network and on which it might perform poorly. In this way, one can be informed in advance about how reliable the NN output is going to be. In general, since several different prior distributions can be used to generate training data, this method provides a way to quantitatively compare them and choose the one that better describes the measurements before any training is performed. In section 8.1 we will describe the algorithm used for the comparison and in section 8.2 we will show its application to training sets generated with two different prior distributions.

8.1. The k-nearest neighbors algorithm

The features of the set of measurable images $D_m$ are determined by the properties that the plasma profiles have during the experiments. An absolutely free, unconstrained sampling of the $15 \times 7$ points in the plasma profiles introduced in table 1, will produce a set of synthetic data $D$ which likely will have little in common with $D_m$. Most of the samples in such a training set will have little use to our purposes, since they would not belong to the domain of the mapping that we want the NN to learn. In order the NN to be able to accurately predict temperature profiles from measured images, the $D_m$ space has then to be covered densely enough by the training data set: we are not interested in generating all possible 15-dimensional output vectors, but only those which represent realistic ion or electron temperature plasma profiles. This is accomplished by refining the prior distributions in such a way that sampling from them will generate a data set of synthetic images which densely encloses $D_m$.

In principle, the full distribution of the data described by a given model is determined by the prior predictive distribution, equation (4.2). As we have seen in section 6, the training images constitute samples from such distribution, therefore they can be used to estimate how closely a training set resemble the actual measurements.

Several methods can be used to study the adequacy of the training set to the coverage of the $D_m$ space of the measured images. Here we will describe an approach which relies on the k-nearest neighbors algorithm (k-NN). A k-NN algorithm is used to find a number $k$ of data points in the training set that are the closest to a given observed data point, according to a metric measure. In this case the Euclidean distance has been used. We expect that if we compare the distance from the training samples to a measured image and to the test set samples, the former will be much larger of the latter in the case where the measurement is not properly described by the training data. This expectation is justified by the fact that we know that the network performs well on the test data and therefore they can be used as reference. Distance based methods are often used in the framework of outlier and novelty detection and similar methods are presented for example in [40, 41].

8.2. Refining the priors

An application to our study is shown in figure 7, where we have compared two training sets obtained with two different models. The difference in the models is in the prior distribution of the plasma profiles: in one case, labeled as W/O, the temperature profiles were left unconstrained in the region of the last closed flux surface (LCFS), being allowed to assume any value between 0 and 10 keV; in the other case, labeled as W, the profiles were constrained to assume low values in the LCFS region (0.1 keV $\pm$ 0.5 keV) at $p_{\text{LCS}}$ = 0.99, a feature that is typically expected in such plasma profiles. Such a constraint enters the Minerva model as a so called virtual observation: at the level of the Minerva graph, this corresponds to a standard observed node connected to the profiles which states that the value of the profiles at the given position $x_p$ has been ‘virtually measured’ to have value $v_p$ with error $\epsilon_p$, as shown in figure 6. The only difference with the other observed nodes in the graph is that it does not correspond to an actual measurement. Its role is to constrain the profile shape, and this is what observations in Bayesian models do: they constrain the solution found for the free parameters. In figure 6, the dashed arrow and box represent the connection to the rest of the model of figure 3. The computation node on the left of the dashed one represents the evaluation of the $T_e$ profile at the $p_{\text{LCS}}$ = 1.0 position, which corresponds to the LCFS. This node is finally connected to the virtual observation, the blue circle, whose normal distribution is specified in term of its mean and standard deviation. If we exclude the graph represented in this figure...
from the rest of the model, and we then calculate the covariance matrix of the posterior distribution of the plasma profile, in this case $T_e$, given the virtual observation, we will find a covariance matrix which can be used later on to sample profiles which will feature the desired constraint. This has been done in order to obtain more realistic plasma profiles, and the effect of this on the training set is described in the next paragraph. Specifically, the virtual observation was implemented as a normal distribution with mean value of 0.1 keV and standard deviation of 0.05 keV at $\rho_{\text{eff}} = 0.99$ for both ion and electron temperature profiles.

The importance of sampling a realistic set of plasma profiles, which corresponds to a realistic set of synthetic measurements, is depicted in figure 7. The three plots on the top show a spectrum measured along three different lines of sight (blue line), and the 10 nearest neighbors (10-NN, gray lines) found among the samples in a training set where the electron temperature profiles were sampled without constraint (W/O) on the value they assume at the edge of the machine (three top plots), (2) the $T_e$ profiles are sampled with (W.) constraint to low values at the edge, as described in the text (three bottom plots). The constraint is applied to the prior probability distributions. Each column shows spectra integrated along a different line of sight. From left to right: a line-of-sight through the edge plasma, one crossing the mid-plane half-way to the core, and one core line of sight.

Figure 7. The $k$-NN algorithm is applied to find the 10 nearest neighbors of a measured image (blue line) among the training samples in two different training sets: (1) the $T_e$ profiles are let vary without (W/O) constraint on the values they assume at the edge of the machine (three top plots), (2) the $T_e$ profiles are sampled with (W.) constraint to low values at the edge, as described in the text (three bottom plots). The constraint is applied to the prior probability distributions. Each column shows spectra integrated along a different line of sight. From left to right: a line-of-sight through the edge plasma, one crossing the mid-plane half-way to the core, and one core line of sight.

Figure 8. The $T_e$ profile samples from the two training sets. Top, the profiles are sampled with (W.) constraints. Bottom, the profiles are sampled without constraints (W/O).

The constraint was a smoothness criterion induced by the GP prior. From left to right, the lines of sight of the plots are traversing the following regions of the machine: edge, half-way to the core and core. The intensity on the y-axis is normalized to the brightest pixel in the image, in this case a w spectral line along one of the central lines of sight. The three plots on the bottom, instead, show the same measured image and the 10-NN neighbors found among the samples in the training set where the electron temperature profiles were sampled without constraint (W/O) on the value assumed on any of the flux surfaces. The only constraint was a smoothness criterion induced by the GP prior. From left to right, the lines of sight of the plots are traversing the following regions of the machine: edge, half-way to the core and core. The intensity on the y-axis is normalized to the brightest pixel in the image, in this case a w spectral line along one of the central lines of sight. The three plots on the bottom, instead, show the same measured image and the 10-NN neighbors found among the samples in the training set with constraint (W.). The effect of such constraint on the sampled profiles is shown in figure 8, where 100 samples from each of the two training sets are drawn. It is evident that when the constraint is applied (bottom plot), the average electron temperature through the machine is lower. This brings the training samples closer to the measured data in two ways: (1) the intensities measured along the edge lines of sight (see leftmost bottom plot in figure 7) show a smaller signal to noise ratio, (2) the ratio between different emission lines is smaller, see middle and rightmost plots in figure 7. The distance of a measured data point from the 100 000 nearest neighbors in the training set can be compared to the distance of 10 test set samples from 100 000 training set samples to get an estimation of the proximity of the measured data with respect to the training samples. These values are shown in table 2 in the two cases of training set created with and without constraint. The distance for the measured data point is substantially reduced when the constraint is applied to the electron temperature profile prior distribution, getting closer to the value of the test set samples. It is worth to note
that, even with the constraint, the sampled profiles are not necessarily monotonically decreasing, as shown in figure 8.

The improvement on the prediction capabilities of the NN when applied to the measured image is remarkable and it is shown in figure 9. The blue line in the plot denotes the mean value of 800 000 samples of ion temperature profiles drawn from the posterior distribution inferred within Minerva, and the corresponding standard deviation. The agreement between the NN prediction and the full Bayesian inference result is visibly better when the constraint is applied.

9. Results

A NN with architecture described in section 7 and illustrated in figure 5 has been evaluated on data from plasma shots of the first operational campaign at W7-X.

The prior distributions of the free parameters of the model used for the creation of the training set were all normal distribution functions with lower truncation at 0.0 keV. The $T_e$ profile prior distribution had also upper truncation at 0.0 keV, whereas the other profiles had none. The values of the parameters of the GP squared exponential function defined in equation (5.2) were set to $\sigma_f = 2.0$, $\sigma_e = 0.3$, and $\sigma_i = 10^{-3} \sigma_f$ for the $T_i$ profile and to $\sigma_f = 5.0$, $\sigma_e = 0.3$, and $\sigma_i = 10^{-3} \sigma_f$ for the $T_e$ profile. The constraint discussed in section 8 was applied to the temperature profiles but not to the density profiles. The magnetic configuration was kept fixed during the sampling procedure and the NN was tested on data from shots having such configuration. A comparison between the standard Bayesian inference carried out within the Minerva framework and the NN inversion is shown in figure 10, for both ion and electron temperature profiles. The different plots in the figure refer to data from different shots and time points within a shot. In general, NN central prediction and full model central prediction, shown with the solid lines, are reasonably close to each other. In two cases the mismatch is especially pronounced: these are the cases of $T_e$ for measurement 160223.007@1.0-1.2s and 160310.029@0.6-0.65s in the core region; when the uncertainties are taken into account, however, the agreement is still good. Providing uncertainties together with the NN central prediction is, indeed, an important point, especially when a comparison is made. We will now give an overview of how the uncertainties have been calculated; the readers that are interested in understanding the details can look at the work presented in [39]. The predicted profiles have been obtained from a committee of networks: a set of 10 NNs with same architecture but different weight initialization has been trained on the same training set. The error bars are calculated in a Bayesian fashion: the training procedure is seen as an inference problem on the network’s weights, which gives as result an approximated Gaussian posterior distribution of the weights, centered on the network’s weight vector found with the training process and whose standard deviation depends on the Hessian matrix of the loss function. The spread in the posterior produces then a spread in the network’s prediction, and this is the source of the network’s error bars. This procedure has been applied to the 10 networks in the committee. The committee prediction is then obtained by sampling a random member network, sampling a set of weights from the corresponding weight’s posterior and then feeding the network with a sample of the input vector drawn from the XICS noise model. This corresponds to approximating the overall weight’s posterior with a Gaussian mixture, where each Gaussian of the mixture is obtained with the single Gaussian approximation of each committee member, and it is centered on the weight vectors found with the different starting values [31]. This is necessary because, in principle, the single Gaussian approximation is valid only around the solution found with the training, and different starting values will lead to different solutions, i.e. different local minima of the optimization problem. In this way, it is then possible to take into account this fact and put together each committee member predictive distribution. The error bars shown in figure 10 also include uncertainties in the XICS measured data.

The full Bayesian model prediction is obtained as the average of 800 000 samples drawn from the posterior distribution of the corresponding free parameter, obtained with a MCMC sampler, and the error bars are obtained as the standard deviation of the samples.

It is important to note that the training set has been generated sampling with the LCFS constraint described in
section 8: as a consequence, the profiles predicted by the networks show small variance towards $\rho = 1$. This is the reason why the uncertainties in the NN output of figure 10 are systematically lower for larger values of $\rho_{\text{eff}}$.

One of the main advantages in using NNs for data analysis is the speed-up that they provide. This is, indeed, substantial: the evaluation of one single data point takes $\sim 10 \mu s$ on a single CPU. The inference with MCMC sampling takes a few hours in similar conditions, thus the speed-up is of $10^9$ orders of magnitude.

10. Conclusions

We have shown a Bayesian model oriented approach to NN training for the inference of ion and electron temperature profiles from data measured with an x-ray imaging diagnostic at W7-X. The model implemented within the Minerva framework is used to generate the training set, sampling from the prior distribution of the free parameters and from the likelihood function of the simulated data, i.e. from the joint distribution of the model. Since the joint distribution summarizes all the relevant properties of the Bayesian model, the trained network can be thought of as an approximation of the full Bayesian model. What makes this approach uncommon is the fact that the network is trained on a problem for which a model, and therefore a solution, already exists. The fact that we make use of a model to generate the training data gives us control over the features that we introduce in the training set, both from a point of view of the statistics and the physics. These features are completely determined by the joint distribution of the Bayesian model.

Since we can manipulate the prior probability distribution functions, we can also knowingly choose the model that better describes the measurements we expect to perform. This gives us the possibility to have control on the training of the network from the point of view of the physics parameters that, through the forward model, generate the expected observations. This is precisely what allows us to find a better training set for more accurate NN predictions.

The NN has been tested on measurements from different plasma shots from the first operational campaign at W7-X and compared with the results of the standard Bayesian inference. The first major advantage of this approach is the speed-up of the analysis, which can be carried out in tens of micro-seconds. The second advantage is that the sampling procedure for the creation of the training data only requires generic, not diagnostic-specific features of the Minerva model and thus is
in principle automatically applicable to any other diagnostic developed within the same framework.

Acknowledgments

This work has been carried out within the framework of the EUROTHERM and has received funding from the Euratom research and training programme 2014–2018 and 2019–2020 under grant agreement No. 633053. The views and opinions expressed herein do not necessarily reflect those of the European Commission.

ORCID iDs

A Pavone @ https://orcid.org/0000-0003-2398-966X
A Langenberg @ https://orcid.org/0000-0002-2107-5488
R C Wolf @ https://orcid.org/0000-0002-2606-5289

References

[1] Cybenko G 1989 Degree of approximation by superpositions of a sigmoidal function Math. Control, Signals, Syst. 2 303–14
[2] Hornik K 1991 Approximation capabilities of multilayer feedforward networks Neural Netw. 4 251–7
[3] Sonoda S and Murata N 2017 Neural Network with Unboundedactivation Functions is Universal Approximator https://arxiv.org/abs/1505.03654
[4] Svensson J et al 1998 Real-time ion temperature profiles in the JET nuclear fusion experiment ICANN 98, Perspectives in Neural Computing (London: Springer London) (https://doi.org/10.1007/978-1-4471-1599-1_30)
[5] Cannas B, Fanni A, Sonato P and Zedda M K 2007 A prediction tool for real-time application in the disruption protection system at JET Nucl. Fusion 47 1559–69
[6] Wang Y Y et al 2016 Prediction of density limit disruptions on the J-TEXT tokamak Plasma Phys. Control. Fusion 58 055014
[7] Pautasso G and Tichmann C 2002 On-line prediction and mitigation of disruptions in ASDEX Upgrade Nucl. Fusion 42 100–8
[8] Bishop C M, Roach C M and von Hellermann M G 1993 Automatic analysis of JET charge exchange spectra using neural networks Plasma Phys. Control. Fusion 35 765–73
[9] Svensson J, von Hellermann M and König R W T 1999 Analysis of JET charge exchange spectra using neural networks Plasma Phys. Control. Fusion 41 315–38
[10] Clayton D J et al 2013 Electron temperature profile reconstructions from multi-energy SXR measurements using neural networks Plasma Phys. Control. Fusion 55 095015
[11] Svensson J and Werner A 2007 Large scale bayesian data analysis for nuclear fusion experiments IEEE Int. Symp. on Intelligent Signal Processing pp 1–6
[12] Langenberg A, Svensson J, Thomsen H, Marchuk O, Pablant N A, Burren R and Wolf R C 2016 Forward modeling of x-ray imaging crystal spectrometers within the Minerva Bayesian analysis framework Fusion Sci. Technol. 69 560–7
[13] Wolf R C et al 2017 Major results from the first plasma campaign of the Wendelstein 7-X stellarator Nucl. Fusion 57 102020
[14] König R et al 2015 The set of diagnostics for the first operation campaign of the Wendelstein 7-X stellarator J. Instrum. 10 P10002
[15] Kuychoviak M et al 2016 Overview of diagnostic performance and results for the first operation phase in Wendelstein 7-X (invited) Rev. Sci. Instrum. 87 11D304
[16] Bozhenkov S A et al 2017 The Thomson scattering diagnostic at Wendelstein 7-X and its performance in the first operation phase J. Instrum. 12 P11004
[17] Hoefel U et al 2019 Bayesian modelling of microwave radiometer calibration on the example of the Wendelstein 7-X electron cyclotron emission diagnostic Rev. Sci. Instrum. 90 043502
[18] Bitter M et al 2010 Objectives and layout of a high-resolution x-ray imaging crystal spectrometer for the large helical device Rev. Sci. Instrum. 81 1–5
[19] Marchuk O, Wolf R and Kunze H J 2004 Modeling of He-like spectra measured at the tokamaks TEXTOR and TORE SUPRA PhD Thesis Ruhr-Universität Bochum
[20] Group T F R et al 1985 Dielectronic satellite spectrum of Helium-like Argon: a contribution to the physics of highly charged ions and plasma impurity transport Phys. Rev. A 32 2374–83
[21] Vainshtein L A and Safronova U I 1978 Wavelet analysis and transition probabilities of satellites to resonance lines of H- and He-like ions At. Data Nucl. Data Tables 21 49–68
[22] Sivia D and Skilling J 2006 Data Analysis: A Bayesian Tutorial (Oxford: Oxford University Press)
[23] MacKay D J C 1991 Bayesian Methods for Adaptive Models PhD Thesis California Institute of Technology
[24] Langenberg A et al 2018 Inference of temperature and density profiles via forward modeling of an x-ray imaging crystal spectrometer within the minerva bayesian analysis framework Rev. Sci. Instrum. Submitted
[25] Pearl J 1986 Fusion, propagation, and structuring in belief networks Artif. Intell. 29 241–88
[26] Hirshman S P and Whitson J C 1983 Steepest-descent moment method for three-dimensional magnetohydrodynamic equilibria Phys. Fluids 26 3553–68
[27] Rasmussen C E and Williams C K I 2006 Gaussian Processes for Machine Learning (Cambridge, MA: MIT Press)
[28] Langenberg A et al 2017 Argon Impurity Transport Studies at Wendelstein 7-X using x-ray Imaging Spectrometer Measurements Nucl. Fusion 57 086013
[29] Kauderer-Abrams E 2017 Quantifying Translation-Invariance in Convolutional Neural Networks https://arxiv.org/abs/1801.01450
[30] Wong S C, Gatt A, Stamatescu V and McDonnell M D 2016 Understanding data augmentation for classification: when to warp? CoRR arXiv:abs/1609.08764
[31] Bishop C M 1995 Neural Networks for Pattern Recognition (Oxford: Oxford University Press)
[32] LeCun Y, Bottou L, Bengio Y and Haffner P 1998 Gradient-based learning applied to document recognition Proc. IEEE 86 2278–324
[33] Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R and Fei-Fei L 2014 Large-scale video classification with convolutional neural networks 2014 IEEE Conf. on Computer Vision and Pattern Recognition pp 1725–32
[34] Krizhevsky A, Sutskever I and Hinton G E 2012 Imagenet classification with deep convolutional neural networks Advances in Neural Information Processing Systems 25 ed F Pereira et al (USA: Curran Associates, Inc.) pp 1097–105
[35] LeCun Y 1989 Generalization and Network Design Strategies (Amsterdam: Elsevier)
| Number | Reference                                                                                           |
|--------|---------------------------------------------------------------------------------------------------|
| 36     | LeCun Y, Boser B E, Denker J S, Henderson D, Howard R E, Hubbard W E and Jackel L D 1990 Handwritten digit recognition with a back-propagation network *Advances in Neural Information Processing Systems* 2 ed D S Touretzky (San Mateo, CA: Morgan Kaufmann Publishers) pp 396–404 |
| 37     | Theano Development Team 2016 *Theano: A Python Framework for Fast Computation of Mathematical Expressions* https://arxiv.org/abs/1605.02688 |
| 38     | Neal R M 1995 Bayesian learning for neural networks *PhD Thesis* University of Toronto |
| 39     | Pavone A et al 2018 Bayesian uncertainty calculation in neural network inference of ion and electron temperature profiles at w7-x *Rev. Sci. Instrum.* **89** 10K102 |
| 40     | Knorr E M and Ng R T 1998 Algorithms for mining distance-based outliers in large datasets *Proc. 24rd Int. Conf. on Very Large Data Bases* pp 392–403 http://www.vldb.org/conf/1998/p392.pdf |
| 41     | Dang T T et al 2015 Distance-based k-nearest neighbors outlier detection method in large-scale traffic data *2015 IEEE Int. Conf. on Digital Signal Processing (DSP)* pp 507–10 |