Supplementary material: Implementation of Gaussian Process Regression to reconstruct 2-D data planes from discrete measurements and propagate error bars

Sayak Bose,[1, *] William Fox,[1, 2] Hantao Ji,[1, 2] Jongsoo Yoo,[1] Aaron Goodman,[3] Andrew Alt,[2] and Masaaki Yamada[1]

[1] *Princeton Plasma Physics Laboratory, Princeton, New Jersey 08540, USA*
[2] *Department of Astrophysical Sciences, Princeton University, New Jersey 08540, USA*
[3] *Department of Mechanical and Aerospace Engineering, Princeton University, New Jersey 08540, USA*

I. INTRODUCTION

Gaussian process regression (GPR) is a class of supervised machine learning algorithms that can be used to construct data profiles from discrete observations, predict uncertainty in the reconstructed profiles in a statistically rigorous manner and provide the framework for error propagation in the calculation of quantities derived from the data profiles [1–3]. Compared to traditional parametric fitting methods, GPR is particularly useful in data reconstruction from discrete observations where the plasma is non-uniform, and the function describing the non-uniform profile is not known beforehand. To elaborate, traditional fitting methods assume a parametric form of a function before fitting that function to a discrete set of observations. A limitation of the traditional method is that more than one function can fit a discrete set of observations, and often plasma physics theories don't predict which functional form is preferred. GPR is a Bayesian non-parametric regression technique that predicts a probability distribution over possible functions that fit a set of discrete data points. The mean of the probability distribution gives the most probable characterization of the data, i.e., the mean function (mean profile), while the standard deviation obtained from variance indicates the uncertainty in the prediction. Thus, GPR recognizes that there is an uncertainty in the function describing the data profiles, which is not acknowledged in traditional fitting methods where a functional form is assumed for fitting [1, 4].

GPR models assume that the measured discrete data points are part of a multivariate normal distribution, i.e., the data points measured at a particular spatial location satisfy a univariate normal distribution, while the data points at neighboring locations are correlated. The correlation between neighboring data points enables GPR to make predictions at locations where measurement does not exist. A covariance kernel like the Radial Basis Function (RBF) is used to incorporate the correlation between data points in the GPR model.

GPR models are classified as homoscedastic and heteroscedastic models depending on how the variability in the observation is modeled [5]. In the context of pulsed experiments like MRX[6], the shot-to-shot variation cause observation variability. In a homoscedastic GPR model, the variance in the observation is assumed to be constant throughout the input space. An example will be a set of observations where the spread of the data points about the mean is the same at each measurement location. In the heteroscedastic model, the variance in the observation can change depending on the location in the input space. Heteroscedastic models are more relevant for experiments where the error bars can differ at different measurement locations. Heteroscedastic GPR uses two $\mathcal{GP}$s (Gaussian Processes), one for modeling the mean function and the other for modeling the input space-dependent error bars. A combination of the two $\mathcal{GP}$s give a posterior distribution over the mean function and input dependent error bars [5]. The mathematical basis for GPR and its application are well described in Refs [1, 2, 4, 5].

Recently, GPR has been used to reconstruct 1-D density and electron temperature profiles of a Tokamak (fusion device) [7]. Typically for analysis of Tokamak data, discrete measurements made at multiple spatial locations simultaneously in a single shot are used to reconstruct the 1-D profiles [7]. However, in MRX, data points from a large number of shots are assembled for the reconstruction of 2-D profiles. The data variability due to shot-to-shot variation in MRX needs to be accounted for during profile reconstruction, gradient calculation, and uncertainty prediction. We describe the application of GPR for data analysis that considers the nuances of MRX data due to shot-to-shot variation.

II. APPLICATION OF GPR FOR ANALYZING MRX DATA

GPR was employed to construct 2-D planes from the discete datsets of electric probes like Langmuir and Mach probes. We demonstrate the implementation of GPR by using the density and potential data of MRX as examples. The density and potential were measured using Langmuir probes [8]. The Langmuir probes were used to make single point measurements in each shot. The probes were moved between shots to obtain discrete data points covering the r-z plane of MRX. Discrete data points from a large number of shots were assimilated to form a dataset for application of GPR. An inspection of the raw data showed that the spread of the discrete data points about the mean differs at different spatial locations. Furthermore, simultaneous measurements at multiple locations showed that the change in the measured values due to

[*] sbose@princeton.edu

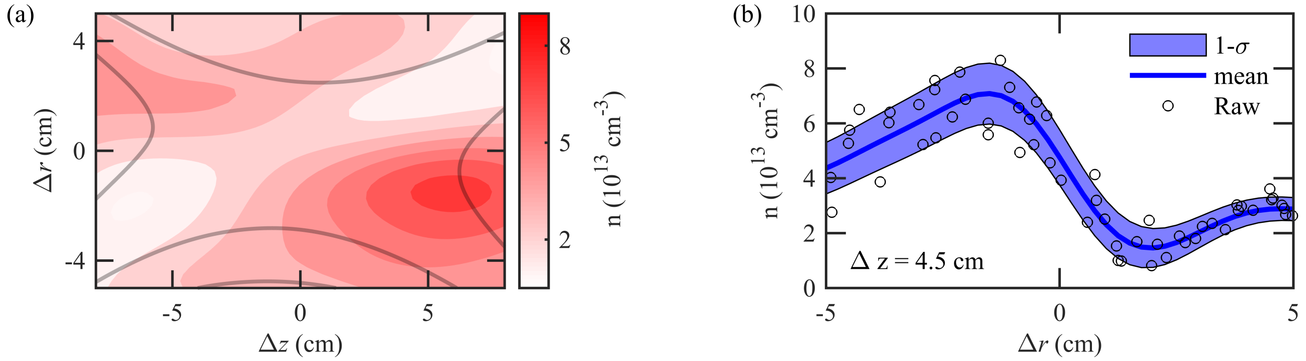

FIG. 1. (a) Two-dimensional mean profile of the electron density extracted from the posterior of a GPR model trained using the discrete raw data. The grey lines are representative magnetic field lines in the r-z plane for reference. (b) A 1-D plot showing the mean and standard deviation extracted from the posterior of the trained GPR model along with the raw data. The dark blue line is the mean density at $\Delta z = 4.5$ cm. The blue band gives the 1-σ uncertainty in the predicted mean profile from the trained GPR model, where σ is the standard deviation. The black circular dots are raw data points lying between $\Delta z = 4$ to 4.5 cm. Note that the 1-σ error bar follows the local spread in the data points. The error bar is thicker at $\Delta r \sim -2.5$ cm where the spread in the raw data is more, while the error bar is smaller at $\Delta r \sim 4.5$ cm where the spread in the raw data is less.

shot-to-shot variation is spatially correlated. To preserve these raw data features in our analysis, we adopted the heteroscedastic GPR model of Zhang and Ni [5]. The model implemented for data reconstruction can be written as

$$y(\mathbf{x}) = f(\mathbf{x}) + \mathcal{GP}\left(0, g(\mathbf{x})\, \rho(\mathbf{x}, \mathbf{x}')\, g(\mathbf{x}')\right), \qquad (1)$$

where $y(\mathbf{x})$ is a sample from the posterior of the trained model, $f(\mathbf{x})$ is the mean function, g^2 is a measure of the variance, and ρ is the correlation function that incorporates the spatial correlation between the change in the measured values at neighboring locations due to shot-to-shot variation. The formulation of Zhang and Ni [5] is used to estimate $f(\mathbf{x})$ and g from two $\mathcal{GP}$s employing RBF Kernels. An RBF kernel with a length scale l of 1.5 cm is used to specify ρ. Refer to subsection II B for the role of length scale of ρ in the GPR model. For the functional form of the RBF kernel, also known as the squared exponential kernel, refer to Rasmussen [1].

Henceforth for brevity, the heteroscedastic GPR model used for reconstructing MRX data profiles will be referred to as the GPR model. The codes for the model were written using the freely available Python packages scikit-learn [9] and NumPy. For introductory examples on how $\mathcal{GP}$s are trained refer to the help files of scikit-learn [10].

A. Reconstruction of 2-D data profiles and calculation of uncertainty in the predicted profile

We have used the density data to demonstrate the construction of 2-D data profile using GPR. As mentioned before, a Langmuir probe was used to measure density. The Langmuir probe was used to make a single-point measurement in each shot. A large number of shots were taken where the probe was moved in-between shots to sample the r-z plane of MRX. A total of 1,700 shots are scrutinized to assemble a refined dataset. We checked for consistency in magnetic-field structure and reference Langmuir probe data in the scrutiny. The X-point location of the reconnection layer had a minor shot-to-shot variation with respect to the system coordinates of MRX. We corrected for the slight change in the X-point location by using the relative positions of the probes measured from the X-point in our GPR analysis.

Fig. 1 demonstrates the effectiveness of GPR in the reconstruction of the density profile and in the prediction of the associated uncertainty. Fig. 1a shows the mean profile of the density extracted from the posterior of the trained GPR model. The 2-D density profile shows a pair of high density and low density regions in the vicinity of the separatrices. To validate these features, we compared the mean profile and error bars with the raw data. A visual comparison of a 1-D cut of the mean profile at $\Delta z = 4.5$ cm in Fig. 1b shows that the mean profile follows the trend in the raw data points. The 1-σ confidence interval predicted by GPR is shown by the blue band in Fig. 1b. A visual inspections shows that the width of the 1-σ confidence interval follows the local spread of the raw data points. For example, the error bar is larger at $\Delta r \sim -2.5$ cm where the spread in the raw data is more, while the error bar is smaller at $\Delta r \sim 4.5$ cm where the spread in the raw data is small. Furthermore, we quantitatively checked the 1-σ error bar by comparing the number of raw data points lying within $n_{\text{mean}} \pm 1\sigma$ to those lying outside the error bars. We found 67% of the raw data points to lie within the error bars, which is very close to the theoretical expectation where 68% of the data points are expected to lie within $\pm 1\sigma$ confidence interval. Thus the comparison between raw data and GPR predictions show that GPR is effective at predicting mean profiles and quantifying the uncertainty

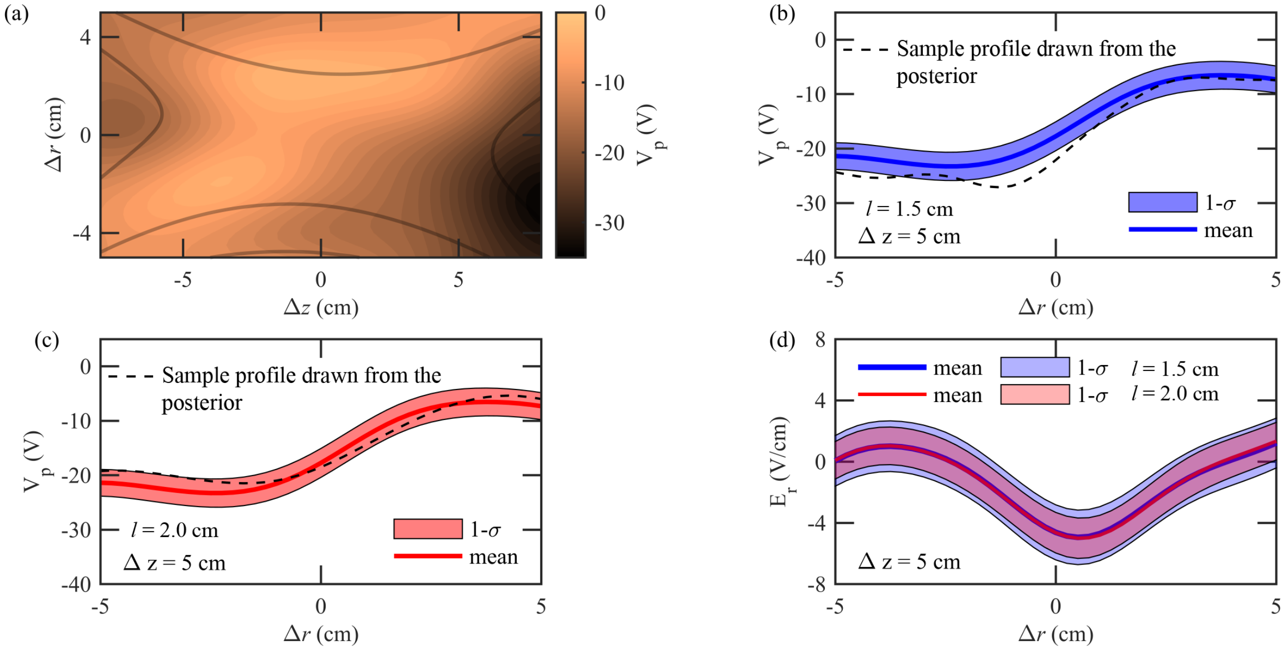

FIG. 2. Demonstration of gradient calculation and determination of the associated error bars using the GPR framework. We also show the effect of the length scale l of the correlation function ρ on the uncertainty in the calculated gradient. The two-dimensional mean profile of plasma potential V_p extracted from the posterior of a GPR model trained using the discrete raw data points is shown in (a). The grey lines in (a) are representative magnetic field lines in the r-z plane for reference. A 1-D plot showing the radial variation of the mean and standard deviation of V_p for $l = 1.5$ and 2 cm are shown in (b) and (c), respectively. Note that the mean and standard deviation in (b) and (c) does not depend on l, which is expected as l does not influence the variance. Rather l affects the smoothness of the sample drawn from the posterior of the trained model. This is because l affects the correlation between spatially separated data points extracted from the posterior. The radial electric field $E_\mathrm{r} = -\partial V_\mathrm{p}/\partial r$ calculated for the two cases $l = 1.5$ and 2 cm is shown in (d). Note that the error bar in the E_r calculation is smaller for the $l = 2$ cm case as the profiles drawn from the posterior are relatively smoother.

associated with those predictions.

B. Calculation of gradient using GPR framework

GPR is useful for computing gradients from data profiles and estimating the associated error bars. For calculating gradients, many sample data profiles were randomly drawn from the posterior of a trained GPR model. Spatial gradients were calculated for each data profile to obtain an ensemble of spatial gradient profiles. The multiple profiles of spatial gradients were averaged to get the mean profile of the data gradient, and the standard deviation was computed to estimate the uncertainty in the gradient calculation.

The magnitude of the error bar of the gradient calculated by the above method depends on the correlation length scale l used for ρ in Eq. 1. We demonstrate the effect of l on uncertainty estimates for gradients using the electric field calculation as an example.

The electric field was calculated from the negative gradient of the plasma potential profiles. The GPR model was trained using diescrete plasma potential data points to extract the plasma potential profiles from the posterior. The discrete data points of the plasma potential were calculated using $V_\mathrm{p} = V_\mathrm{f} + 3.7T_e$ [11]. Here V_p is the plasma potential, V_f is the floating potential, and T_e is the electron temperature. The V_f and T_e were measured using a Langmuir probe.

Fig. 2 shows the calculation of radial electric field, E_r, from V_p using $E_\mathrm{r} = -\partial V_\mathrm{p}/\partial r$. Fig. 2b and c shows the radial variation of V_p used for calculating E_r shown in Fig. 2d. The value of l was held at 1.5 cm in Fig. 2b, while l was 2 cm in Fig. 2c. Note that the mean and the error bar of V_p in Fig. 2b and c are the same as l does not effect the mean and the standard deviation. However, the samples drawn randomly from the posterior of GPR are smoother for the $l = 1.5$ cm compared to $l = 2$ cm as l affects spatial correlation of the posterior samples. The E_r profile for the two cases are shown in Fig. 2c. The mean E_r for the cases are nearly identical, however, the error bars are greater for the $l = 1.5$ cm case compared to $l = 2$ cm.

To estimate a value of l for reconstructing data profiles we compared the gradient of the floating potential directly measured using a radial floating probe array with predictions of GPR for different values of l. The results of direct measurement and GPR prediction were found to best agree for $l = 1.5$ cm. Therefore, we used $l = 1.5$ cm for defining ρ in Eq. 1.

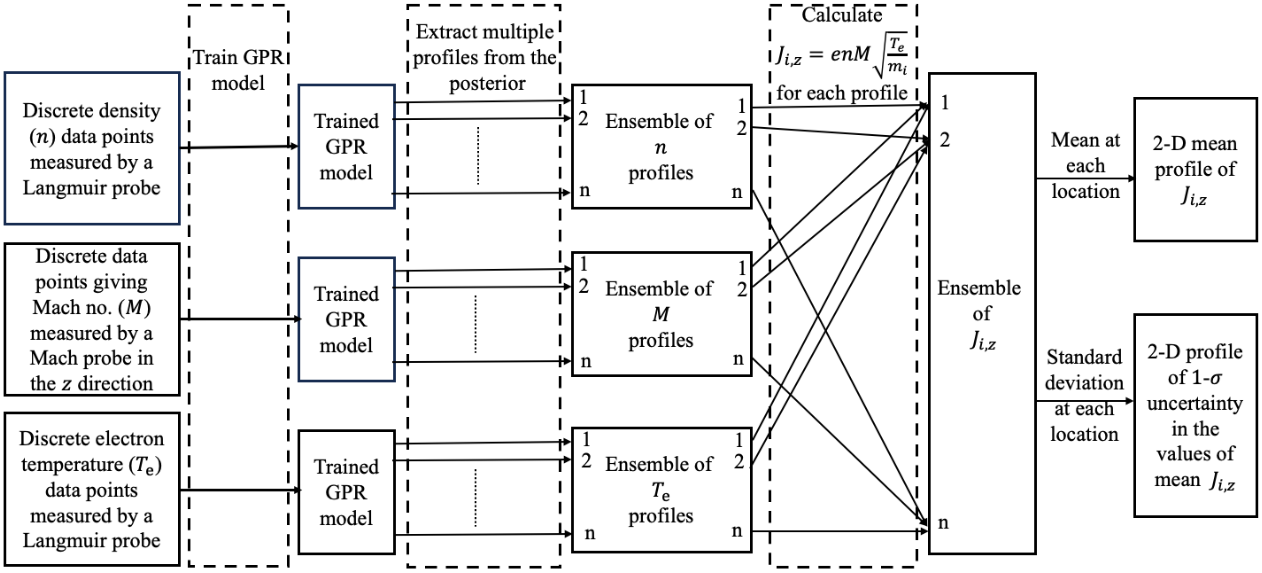
Discrete density (n) data points measured by a Langmuir probe
Discrete data points giving Mach no. (M) measured by a Mach probe in the z direction
Discrete electron temperature (T_e) data points measured by a Langmuir probe
Train GPR model
Trained GPR model
Trained GPR model
Trained GPR model
Extract multiple profiles from the posterior
Ensemble of n profiles
Ensemble of M profiles
Ensemble of T_e profiles
1
2
n
Calculate $J_{i,z} = enM\sqrt{\frac{T_e}{m_i}}$ for each profile
Ensemble of $J_{i,z}$
Mean at each location
2-D mean profile of $J_{i,z}$
Standard deviation at each location
2-D profile of 1-σ uncertainty in the values of mean $J_{i,z}$

FIG. 3. Schematic of uncertainty propagation in calculation of composite quantities.

C. Error propagation in calculation of composite quantities using a Monte-Carlo method and the GPR framework

GPR provides an effective framework for error propagation in the calculation of composite quantities derived from data profiles of multiple directly measured variables using a Monte-Carlo technique. We use the calculation of the z-component of the ion current density, $J_{i,z} = enM\sqrt{T_e/m_i}$, as an example to demonstrate error propagation. Here, e is the quantum of electric charge, M is the Mach number, T_e is the electron temperature, and m_i is the ion mass. The n and T_e are measured using a Langmuir probe, and M is measured using a Mach probe. The mach probe data was calibrated using ion velocity measured by an ion Doppler spectroscopy probe. The steps of the error propagation calculation are shown in the schematic given in Fig. 3. First, the discrete data points of n, M and T_e are used to train GPR models for each quantity. From the posterior of trained GPR models, many profiles of n, M, and T_e are drawn randomly to calculate $J_{i,z}$ and thus generate an ensemble of $J_{i,z}$ profiles. The multiple profiles of $J_{i,z}$ are averaged to obtain the mean profile of $J_{i,z}$, and the standard deviation gives the uncertainty.

III. DISCUSSION

In this supplementary material, we have described the use of GPR to reconstruct 2-D data planes and compute gradients of physical quantities (n, T_e, V_p, Mach number, etc.) from datasets of discrete data points obtained by electric probes. The electric probes were used to make single-point measurements in each shot, and the probes were moved in between shots to cover the r-z plane of MRX. Single point measurements from a large number of shots were assimilated to form a dataset of discrete data points for the application of GPR. The 2-D reconstructed data planes of the electric probe data, along with the magnetic field data of the 2-D B-dot probe array, are used for physics analysis of the reconnection layer. The error propagation in the calculation of composite quantities was done using a Monte-Carlo method.

[1] C. E. Rasmussen, *Gaussian processes for machine learning*, Adaptive Computation and Machine Learning series (MIT Press,, Cambridge, Mass. :, 2005-01-01).

[2] M. Chilenski, M. Greenwald, Y. Marzouk, N. Howard, A. White, J. Rice, and J. Walk, Nuclear Fusion **55**, 023012 (2015).

[3] A. Mathews and J. W. Hughes, IEEE Transactions on Plasma Science **49**, 3841 (2021).

[4] J. Wang, arXiv preprint arXiv:2009.10862 (2020).

[5] Q.-H. Zhang and Y.-Q. Ni, IEEE Transactions on Signal Processing **68**, 3450 (2020).

[6] M. Yamada, H. Ji, S. Hsu, T. Carter, R. Kulsrud, N. Bretz, F. Jobes, Y. Ono, and F. Perkins, Physics of

Plasmas **4**, 1936 (1997).

[7] M. Chilenski, M. Greenwald, A. Hubbard, J. Hughes, J. Lee, Y. Marzouk, J. Rice, and A. White, Nuclear Fusion **57**, 126013 (2017).

[8] J. Yoo, *Experimental studies of particle acceleration and heating during magnetic reconnection*, Ph.D. thesis, Princeton University (2013).

[9] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, Journal of Machine Learning Research **12**, 2825 (2011).

[10] Gaussian processes, https://scikit-learn.org/stable/modules/gaussian_process.html, [Online; accessed 5-July-2023].

[11] H. Ji, H. Toyama, K. Yamagishi, S. Shinohara, A. Fujisawa, and K. Miyamoto, Review of Scientific Instruments **62**, 2326 (1991).