Gaussian Process Regression for Maximum Entropy Distribution

Preprint · July 2019

DOI: 10.13140/RG.2.2.30319.59042

3 authors:

Mohsen Sadr
RWTH Aachen University
4 PUBLICATIONS **7** CITATIONS
SEE PROFILE

Manuel Torrilhon
RWTH Aachen University
134 PUBLICATIONS **2,271** CITATIONS
SEE PROFILE

Hossein Gorji
École Polytechnique Fédérale de Lausanne
27 PUBLICATIONS **214** CITATIONS
SEE PROFILE

Gaussian Process Regression for Maximum Entropy Distribution

Mohsen Sadr[a,*], Manuel Torrilhon[a], M. Hossein Gorji[b]

[a] *MATHCCES, Department of Mathematics, RWTH Aachen University, Schinkestrasse 2, D-52062 Aachen, Germany*
[b] *MCSS, Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland.*

Abstract

Maximum-Entropy Distributions offer an attractive family of probability densities suitable for moment closure problems. Yet finding the Lagrange multipliers which parametrize these distributions, turns out to be a computational bottleneck for practical closure settings. Motivated by recent success of Gaussian processes, we investigate the suitability of Gaussian priors to approximate the Lagrange multipliers as a map of a given set of moments. Examining various kernel functions, the hyperparameters are optimized by maximizing the log-likelihood. The performance of the devised data-driven Maximum-Entropy closure is studied for couple of test cases including relaxation of non-equilibrium distributions governed by the Bhatnagar-Gross-Krook kinetic model.

Keywords: Gaussian process regression, Maximum entropy distribution, Moment problem

1. Introduction

Estimating a probability density from a given set of moments known as the closure problem, naturally arises by representing a high-dimensional system with only a few moments. This inverse problem is ill-posed in general, and thus regularization/regression has to be pursued. In practice two frameworks have been developed: regression on the probability density versus regression on the logarithm of the probability density. The former includes orthogonal expansion techniques such as Hermite/Grad type expansions [1, 2, 3] besides

*Corresponding author
Email address: sadr@mathcces.rwth-aachen.de (Mohsen Sadr)

quadrature methods [4]. The latter leads to the family of Maximum Entropy Distributions (MEDs) [5, 6]. The MED is defined by maximizing an entropy functional of the distribution, subject to the given moment constraints. Regularizing the closure problem by maximizing the Shannon entropy is motivated by both physical and information theoretic considerations. The physical motivation relies on the Boltzmann H-theorem, whereas the latter is linked to the least-bias estimators. MEDs have been employed in various settings as diverse as natural language processing [7], image/signal processing [8, 9], geoscience [10], rarefied gas dynamics [11], solid state physics [12, 13] and climate forecast [14]. However besides theoretical difficulties [15], the use of Maximum-Entropy distributions has been restricted due to numerical challenges.

Following standard steps of the method of Lagrange multipliers, finding the MED reduces to computing the Lagrange multipliers arising from moment constraints [16]. Although the well-posedeness of such an optimization problem has been shown for bounded domains and realizable moments [17, 18, 19], in practice expensive iterations have to be employed for finding Lagrange multipliers. Commonly used iterative approaches are based on the gradient descent, Newton's method and the adaptive basis method. For invertable and Lipschitz continuous Hessians, Newton's method provides the fastest convergence. However since those conditions are not guaranteed in the considered setting, the adaptive basis method is suggested [20, 21].

As a numerically efficient alternative, here we reset the problem of finding the Lagrange multipliers to a Bayesian inference framework. The idea is to express the mapping from moments to Lagrange multipliers by a Gaussian Process (GP). Since computing moments for a given set of Lagrange multipliers is simple and cheap, the training data set can be obtained in a straight-forward way. Therefore, the hyperparameters of the considered GP prior are found by maximizing the log-likelihood over the training data set. Once the hyperparameters are found, the Lagrange multipliers for a new set of moments can be inferred by conditioning the constructed multivariate Gaussian distribution [22].

The motivation behind our approach is purely computational. Observe that all heavy computations including generating training data, finding an appropriate kernel, the Cholesky factorization of the covariance matrix and fitting the hyperparameters are done up-front (offline). For simulations, evaluation of the GP regression is done via a simple backward substitution.

Following the objective of constructing accurate GP estimators for the Lagrange multipliers of MED, the remainder of this manuscript is structured as the following. First in § 2, a short review of MED besides an iterative approach for computing the Lagrange multipliers are presented. Furthermore, a short description of the GP regression is provided. Then in § 3, several kernels such as radial basis and Matèrn family are evaluated for our problem setting. As the first test case, the accuracy of the fitted GP in predicting bi-modal distributions is studied in § 3.1. Then, in § 3.3 transition of a non-equilibrium distribution to the normal one governed by Bhatnagar-Gross-Krook (BGK) equation [23] is studied. At the end, a conclusion and an outlook for future studies are given in § 4.

2. Methods

In the following, first the MED framework is reviewed and the problem statement is refined. Next, a short description of the GP regression is presented.

2.1. Review of Maximum Entropy Problem

Consider the set of admissible probability densities defined over measurable functions as

$$\mathcal{P} \;=\; \left\{ f : \mathbb{R}^l \to [0,\infty) \middle| \int_\Omega f(x)dx = 1 \right\}, \tag{1}$$

where $\Omega \subseteq \mathbb{R}^l$. Suppose we are given a finite vector of moments $p \in \mathbb{R}^N$ of an unknown $f(v) \in \mathcal{P}$ such that

$$p_j = \int_\Omega f(v)\phi_j(v)dv; \qquad 1 \le j \le N \;, \tag{2}$$

where $\phi(v) : \mathbb{R}^l \to \mathbb{R}^N$ is a vector of polynomials. Here and hence forth the subscript indices denote a component of the quantity. The goal is to approximate f by some $f^{(s)} \in \mathcal{P}$ such that the (mathematical) entropy

$$S[f] := \int_\Omega f \ln(f) dv, \tag{3}$$

is minimized while the constraints

$$\int_\Omega \phi(v) f^{(s)} dv = p \tag{4}$$

are satisfied. Since $S[f]$ is convex and the constraints are linear, the solution of the above minimization problem is unique, once it exists. To leave out pathological cases [15], we focus on a bounded domain Ω, for which the minimization problem is well-posed for realizable moments. Using the method of Lagrange multipliers we get

$$C_N^\lambda[f^{(s)}] := \int_\Omega f^{(s)} \ln(f^{(s)}) dv - \sum_{j=1}^N \lambda_j \left(\int_\Omega f^{(s)} \phi_j dv - p_j \right), \tag{5}$$

which has its extremum at

$$f_N^\lambda(v) = Z_\lambda^{-1} \exp\left(- \sum_{j=1}^N \lambda_j \phi_j \right), \tag{6}$$

where Z_λ is the normalization factor [16]. By inserting f_N^λ into the constraints, the Lagrange multipliers $\lambda(p)$ can be computed. However it is more convenient to consider the dual formulation which provides an unconstrained convex minimization for Lagrange multipliers as

$$\lambda(p) \quad = \quad \underset{\lambda^* \in \mathbb{R}^N}{\arg\min} \left\{ Z_{\lambda^*} - \sum_j \lambda_j^* p_j \right\}. \tag{7}$$

Hence the maximum entropy regularization, reduces the closure problem to computing $\lambda(p)$ from Eq. (7). As a direct solution of the dual problem, the standard Newton's method for finding the Lagrange multipliers are reviewed in the following. Let $H(\lambda)$ and $g(\lambda)$ be the Hessian and the gradient of the objective function in Eq. (7), respectively. Following

Newton's method [24], the estimated Lagrange multipliers λ^n at step n, are updated by solving the linear system

$$\sum_{j=1}^{N} H_{ij}(\lambda^n)\Delta\lambda_j^n = g_i(\lambda^n) \tag{8}$$

for $\Delta\lambda^n$. Accordingly, the Lagrange multipliers get updated

$$\lambda_i^{n+1} \;=\; \lambda_i^n + \beta^n \Delta\lambda_i^n, \tag{9}$$

where β is a damping factor [24]. Although H is symmetric-positive-definite, it can become ill-conditioned which can be coped with by using an adaptive basis [25]. For example in [24], the Hermit polynomials are employed as the basis in order to keep the Hessian matrix close to a diagonal one. A more general approach which generates a diagonal Hessian for an arbitrary probability density is followed in [20, 21]. Yet high computational costs can become a limiting factor for this fully adaptive basis methodology.

2.2. Gaussian Process Regression

The high computational intensity of the direct iterative approach for solving the dual problem (7), motivates alternative methods. Here we focus on a data-driven approach based on GP. Let us first review the main idea behind GP based regressions. Suppose $\Psi(x) : \mathbb{R}^N \to \mathbb{R}^N$ is an unknown map, yet we have access to evaluations $\{\Psi(x^{(j)})\}_{j=1}^M$ at some data points $\mathcal{D} = \{x^{(1)}, x^{(2)}, ..., x^{(M)}\}$. Note that the superscript index denotes the corresponding data batch. Therefore the regression problem addresses estimating $\Psi(x)$ from the given $\{x^{(j)}, \Psi(x^{(j)})\}_{j=1}^M$. Consider a positive semi-definite (PSD) kernel function $\mathcal{K}(x, x') : \mathbb{R}^N \times \mathbb{R}^N \to [0, \infty)$, then the GP regression sets forth

$$\tilde{\Psi} \sim \mathcal{GP}(0, \mathcal{K}) \tag{10}$$

as an approximation of Ψ. Here $\mathcal{GP}$ denotes a random process whose distribution for a set of points is a joint normal with the covariance being the Gram matrix associated with $\mathcal{K}$. The merit of a regression of the type (10) can be addressed from different perspectives. More relevant to our setting, it can be shown that the conditional expectation

$\mathbb{E}[\tilde{\Psi}|\tilde{\Psi}(x^{(j)}) = \Psi(x^{(j)}), \forall x^{(j)} \in \mathcal{D}]$ provides an optimal recovery of Ψ in the sense of the relative error induced by the corresponding Reproducing-Kernel-Hilbert-Space [26]. In practice, we work with parametrized kernels $\mathcal{K}_\Theta$, where the hyperparameters embedded in Θ are found by maximizing the log-likelihood [27]. Furthermore, we construct the GP regressions component-wise. Hence we evaluate the hyperparameters for every $\tilde{\Psi}_i$ $(i = 1, ..., N)$, separately.

Several PSD kernel functions $\mathcal{K}$ have been introduced in the literature, see e.g. [27]. Here we consider the radial basis function (RBF) along with Matérn's family for each component $i, j \in \{1, ..., N\}$ we have

$$\mathcal{K}_{\Theta_i}^{\text{RBF}}(x, x') = \sigma_i \exp\left(-r_i^2/2\right) \ , \tag{11}$$

$$\mathcal{K}_{\Theta_i}^{\text{Matérn(12)}}(x, x') = \sigma_i \exp(-r_i) \ , \tag{12}$$

$$\mathcal{K}_{\Theta_i}^{\text{Matérn(32)}}(x, x') = \sigma_i(1 + \sqrt{3}r_i)\exp(-\sqrt{3}r_i) \quad \text{and} \tag{13}$$

$$\mathcal{K}_{\Theta_i}^{\text{Matérn(52)}}(x, x') = \sigma_i(1 + \sqrt{5}r_i + \frac{5}{3}\sqrt{r_i}\exp(-\sqrt{5}r_i) \ . \tag{14}$$

Note that $r_i^2 = \sum_j L_{ij}^{-1}(x_j - x'_j)^2$, where the positive-definite-matrix $L_{N\times N}$ encodes a characteristic length-scale. For each component $i \in \{1, ..., N\}$, the hyperparameters $\Theta_i = \{\sigma_i, L_{i1}^{-1}, ..., L_{iN}^{-1}\}$ can be found by maximizing the log-likelihood

$$\begin{aligned}\ln\left(\tilde{f}\left(\tilde{\Psi}_i(x) \,|\, x \in \mathcal{D}\right)\right) = -\frac{1}{2}\ln\left(\mathcal{K}_{\Theta_i}(x, x')\right) \\ + {\Psi_i}^T(x)\mathcal{K}_{\Theta_i}(x, x')^{-1}\Psi_i(x) - \frac{M}{2}\ln(2\pi),\end{aligned} \tag{15}$$

where $\tilde{f}$ denotes the probability density of $\tilde{\Psi}_i$ conditioned on the training points. The Broyden-Fletcher-Goldfarb-Shanno algorithm (BFGS) is used in this study to find the local minimum with respect to the hyperparameters [28]. It can be shown that the global minimum is attained as more data points are deployed [29]. Once the kernel function $\mathcal{K}$ and its hyperparameters are set, one can evaluate the distribution of $\tilde{\Psi}$ at an arbitrary point. Let x be composed of the training points, therefore

$$\left(\tilde{\Psi}_i(x^*)\middle|\tilde{\Psi}_i(x) = \Psi_i(x)\right) \sim \mathcal{N}(\bar{m}_i, \bar{\Sigma}_i), \tag{16}$$

where

$$\bar{m}_i = \mathcal{K}_{\Theta_i}(x^*, x')\mathcal{K}_{\Theta_i}(x, x')^{-1}\Psi_i(x) \tag{17}$$

$$\text{and} \qquad \bar{\Sigma}_i = \mathcal{K}_{\Theta_i}(x^*, x^*) - \mathcal{K}_{\Theta_i}(x^*, x)\mathcal{K}_{\Theta_i}(x, x')^{-1}\mathcal{K}_{\Theta_i}(x^*, x)\ . \tag{18}$$

Since the inversions appearing in Eqs. (17) and (18) only include the training points, the corresponding computations can be done up-front. Although more efficient GP models such as sparse GP [30] could be pursued, in this study we adopt the straight-forward GP regression model available on GPflow [31].

3. Results

In this section, first constructing GP maps for Lagrange multipliers are pursued and the performance of several covariance functions is assessed. Afterwards, the trained GP is employed for predicting different scenarios, relevant for kinetic systems. To further refine our setting, without loss of generality we restrict ourselves to a one-dimensional domain $\Omega = [v_{\min}, v_{\max}]$. Moreover the moments are computed for the polynomials $\phi_i = v^i$, for $i \in \{1, ..., N\}$. We shift and scale the coordinate such that zero mean and unity variance are obtained. After normalization, the sample space is set by adopting $v_{\max} = -v_{\min} = 10$.

3.1. Training Gaussian Process

3.1.1. Initializing data set

In order to create the data set, an algorithm is outlined in § 1. First a set of Lagrange multipliers $\{\tilde{\lambda}_i\}_{i=1}^N$ are sampled uniformly from $\Lambda \subset \mathbb{R}^N$, conditioned on $Z_{\tilde{\lambda}} < 1/\epsilon$. Here $\epsilon = 10^{-15}$ is the error tolerance. Hence a trial density $f_N^{\tilde{\lambda}}$ is generated. The mean μ and the variance σ^2 are computed from $f_N^{\tilde{\lambda}}$ using the Gaussian-quadrature. In order to find the corresponding Lagrange multipliers that guarantee zero mean and unity variance, we make use of the coordinate transformation $v' = (v - \mu)/\sigma$. Observe that by equality of measures we get

$$f_N^{\lambda}(v') \;=\; \sigma f_N^{\tilde{\lambda}}\left(\sigma v' + \mu\right). \tag{19}$$

Using the binomial expansion, it is straight-forward to find that

$$\lambda_i = \sigma^i \tilde{\lambda}_i + \sum_{j=i+1}^{N} \tilde{\lambda}_j \binom{j}{i} \sigma^i \mu^{j-i} \; ; \quad i \in \{1, ..., N\} \tag{20}$$

ensures Eq. (19).

Algorithm 1 Generating data set (λ, p) with $\int_\Omega v f_N^\lambda dv = 0$ and $\int_\Omega v^2 f_N^\lambda dv = 1$

Sample $\tilde{\lambda} \in \Lambda$ from a uniform distribution subject to $Z_{\tilde{\lambda}} < \frac{1}{\epsilon}$.

Compute $\mu = \int_\Omega v f_N^{\tilde{\lambda}} dv$ and $\sigma^2 = \int_\Omega v^2 f_N^{\tilde{\lambda}} dv$.

Compute λ according to Eq.(20).

Compute moments p .

return (λ, p).

For the training, we consider $N \in \{4, 6, 8\}$ and accordingly $\Lambda^{(4)} = [-1, 1]^3 \times [5\times10^{-2}, 5\times10^{-2}]$, $\Lambda^{(6)} = [-10, 10]^5 \times [10^{-4}, 10^{-4}]$ and $\Lambda^{(8)} = [-10, 10]^4 \times [-10^{-2}, 10^{-2}]^2 \times [-10^{-3}, 10^{-3}] \times [10^{-7}, 10^{-7}]$ are selected. Numerical integrations are carried out using Gaussian-quadrature with roughly 20 points. The total number of 1000 training pairs $\{(\lambda^{(k)}, p^{(k)})\}_{k=1}^{1000}$ were generated according to algorithm(1).

3.1.2. Pre-treatment of data set

Every $(\lambda_i^{(k)}, p_j^{(k)})$ component of the data set can have significant variations passing through different batches of $k \in \{1, ..., M\}$ (with $M = 1000$ for our data set). We follow the common recipe in data-driven methodologies which includes scaling and shifting of every data point $(\lambda_i^{(k)}, p_j^{(k)})$ by the standard-deviation and the average computed over N batches of the particular (i, j) component, respectively. Note that this does not have to be carried out for p_1 and p_2, since they have fixed values already.

3.1.3. Kernel comparison

First, let us consider the radial basis function (RBF) for the kernel choice. Once the hyper-parameters of Eq. (11) are found via maximizing the log-likelihood given by Eq. (15), the

accuracy of predictions over unseen data is investigated. As shown in Fig. 1, by increasing the number of data points M in the training set, the expectation and the variance of the relative error decay using the GP regression.

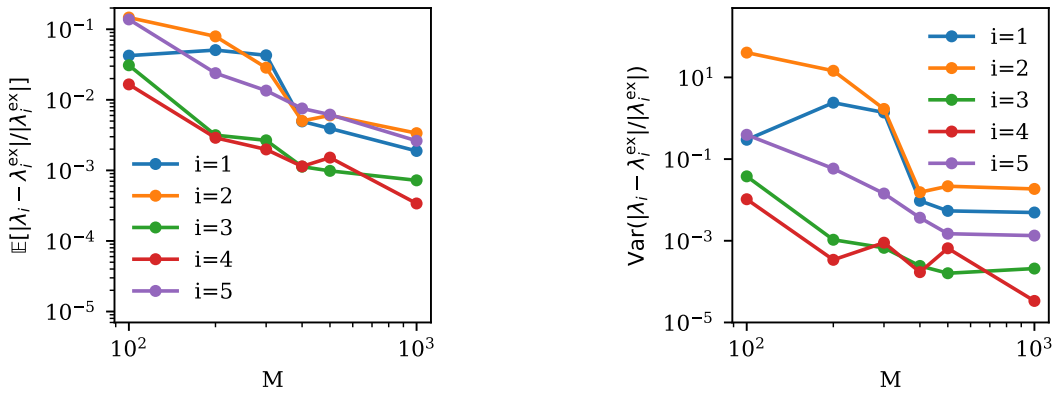

Figure 1: Expectation and variance of the relative error in predicting λs using RBF.

For comparison, several kernels from the Matérn family of functions, i.e. Matérn(12), Matérn(32) and Matérn(52), have been tested here for the training step. Based on our computational experiments as shown in Fig. 2, RBF provides a better estimation for this data set.

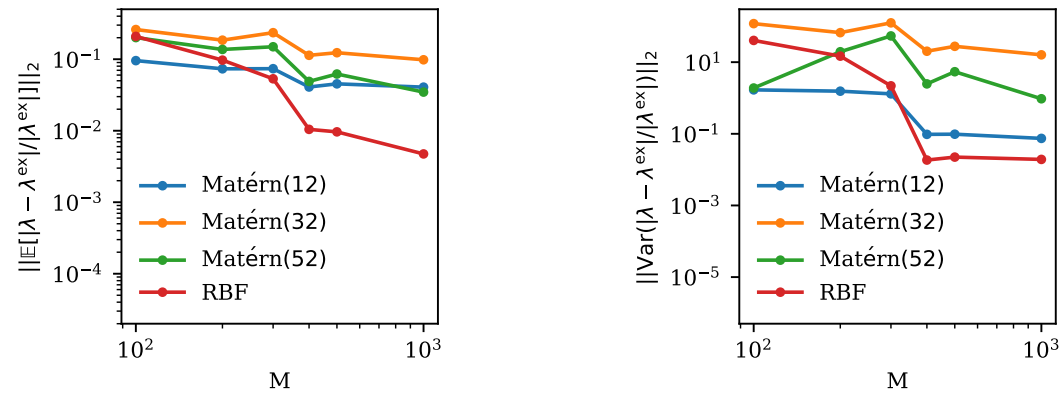

Figure 2: Convergence comparison using different kernels. The L^2-norm of the expectation and the variance of the relative error are shown.

3.2. Test case #1: recovering bi-modal density

We employ the trained GP with the RBF kernel to predict the bi-modal density of the form

$$f^{\text{ex}}(v|\mu_1,\sigma_1,\mu_2,\sigma_2) = \frac{1}{2}\left[f^{\mathcal{N}}(v|\mu_1,\sigma_1) + f^{\mathcal{N}}(v|\mu_2,\sigma_2)\right], \tag{21}$$

where $\mu_2 = -\mu_1$ and $\sigma_2 = \sqrt{2-(\sigma_1^2+2\mu_1^2)}$. Note that $f^{\mathcal{N}}(v|\mu,\sigma)$ is the normal density with the mean μ and the variance σ^2. To quantify the deviation of the estimated density from the exact one, the Kullback–Leibler divergence

$$D_{KL}(f^{\text{ex}}||f_N^{\lambda}) = \int_{\Omega} f^{\text{ex}}(v)\ln\left(f^{\text{ex}}(v)/f_N^{\lambda}(v)\right)dv \tag{22}$$

is used here. Three different scenarios of $(\mu_1,\sigma_1) \in \{(0.8,0.3),(0.9,0.2),(0.95,0.15)\}$ are considered, where predictions are provided based on the GP regression with $N = 4, 6$ and 8 moments. The results depicted in Fig. 3 show that even with f_4^{λ} a good recovery is achieved. It is important to emphasize that not much accuracy gain can be obtained by passing from $N = 4$ to $N = 8$. It seems that moments higher than those considered here are responsible for the observed deviations. As expected by merging the two modes, better agreement is obtained between the GP-accelerated MED and the bi-model one.

3.3. Test case #2: recovering BGK relaxation

This test case investigates the accuracy of the trained GP with the RBF kernel in predicting the evolution of a density $f(v|t)$ governed by

$$\frac{\partial f(v|t)}{\partial t} = \nu(f^{\mathcal{N}}(v|0,1) - f(v|t)) \ . \tag{23}$$

The collision frequency ν controls how quick the solution reaches the equilibrium. Given an initial condition $f(v|t_0)$, the exact solution reads

$$f^{\text{ex}}(v|t) = [1-\exp(-\nu t)]\, f^{\mathcal{N}}(v|0,1) + \exp(-\nu t) f(v|t_0) \ . \tag{24}$$

Here, we use bi-modal normal distribution described in § 3.2 with $(\mu_1,\sigma_1) = (0.98, 0.2)$ as the initial density.

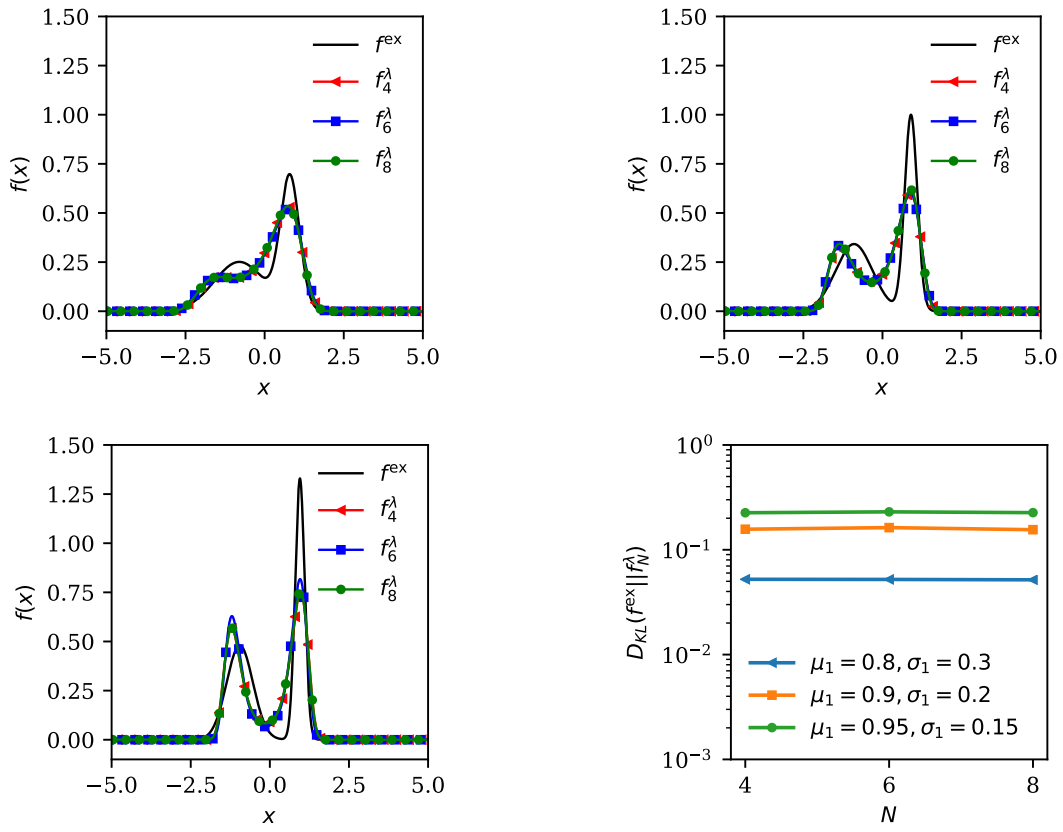

Figure 3: Recovering bi-modal probability densities using maximum entropy distributions accelerated by the Gaussian process regression with $N = 4, 6$ and 8 moments.

In order to solve Eq. (23) using MED, the Lagrange multipliers corresponding to the set of moments at time t need to be evaluated. Applying the devised GP regression, trained for $\lambda \in \mathbb{R}^N$ with $N = 4, 6$ and 8, the Lagrange multipliers are estimated. Observe that the moments $p(t)$ can be computed analytically from Eq. (23). Therefore, at each time instant, the moments and subsequently the trained GP map, are found. The estimated f_N^λ together with its moments are compared with respect to the corresponding exact solution as shown in Fig. 4. Here $\nu = 0.25$ and time intervals are $(t_0 = 0, t_1 = 3, t_2 = 8, t_3 = 20)$ are chosen. Improvements of the estimator are clearly visible by increasing the number of moments as shown in Fig. 5. It is encouraging to see that even with as few moments as $N = 4$, one can recover the bi-model density using the GP-estimated MED.

4. Conclusions

The moment closure problem arising from high dimensional systems continues to be a challenge for scientific computing. While MEDs offer an interesting solution framework for estimating the underlying probability density from a given set of moments, the computational cost associated with computing the Lagrange multipliers hindered their use for practical settings. In this study, we accelerate finding the MED by employing GPs as a regression map from moments to Lagrange multipliers. By taking advantage of the fact that computing the moments from Lagrange multipliers can be performed by simple numerical integrations, around 1000 training data points were generated. Appropriate preparation of the training set by ensuring zero mean and unity variance of MED, besides careful choice of the kernel function have been carried out for a one-dimensional bounded sample space. The results of capturing bi-modal distributions and a BGK type relaxation show encouraging performance of the GP-accelerated MED. For future studies, higher dimensional sample spaces besides sparse GPs will be pursued to further generalize the devised scheme.

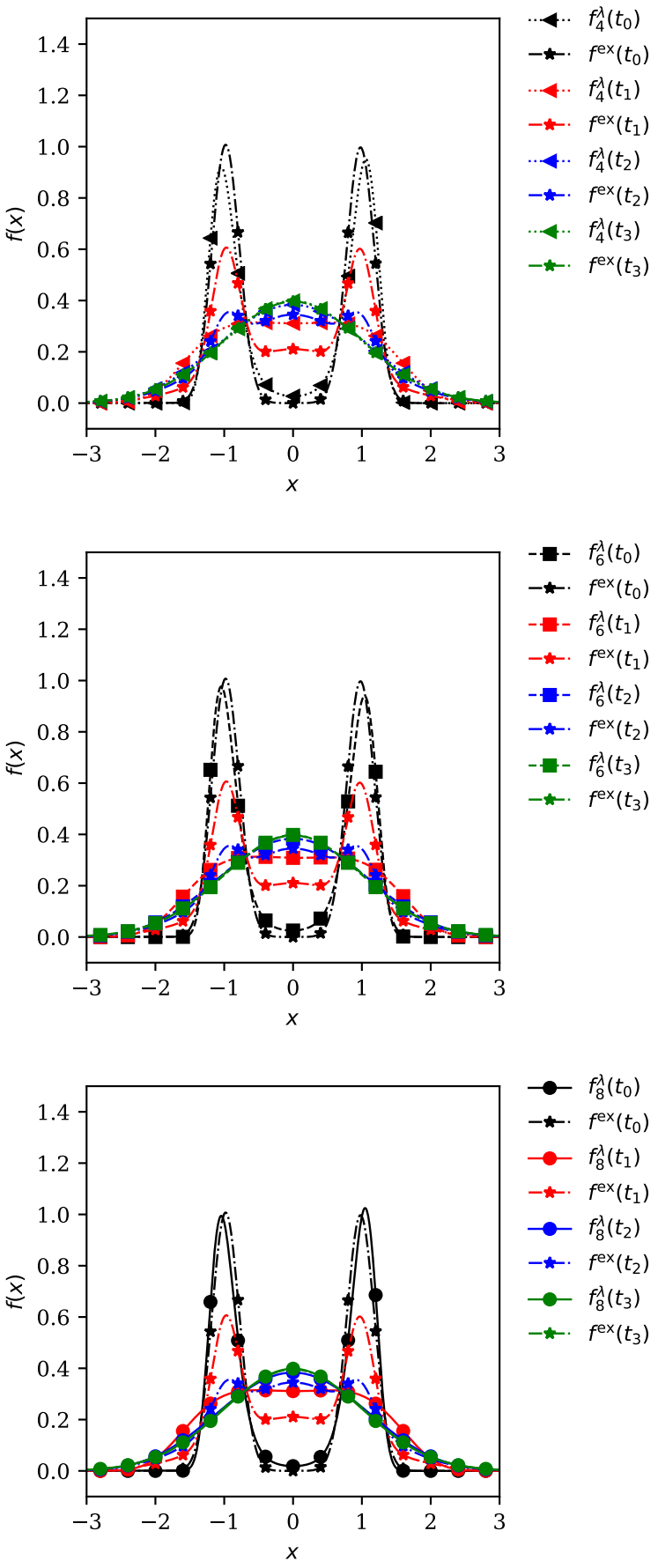

Figure 4: Capturing BGK relaxation using GP-accelerated MED for $N = 4, 6$ and 8 moments at time $(t_0 = 0, t_1 = 3, t_2 = 8, t_3 = 20)$.

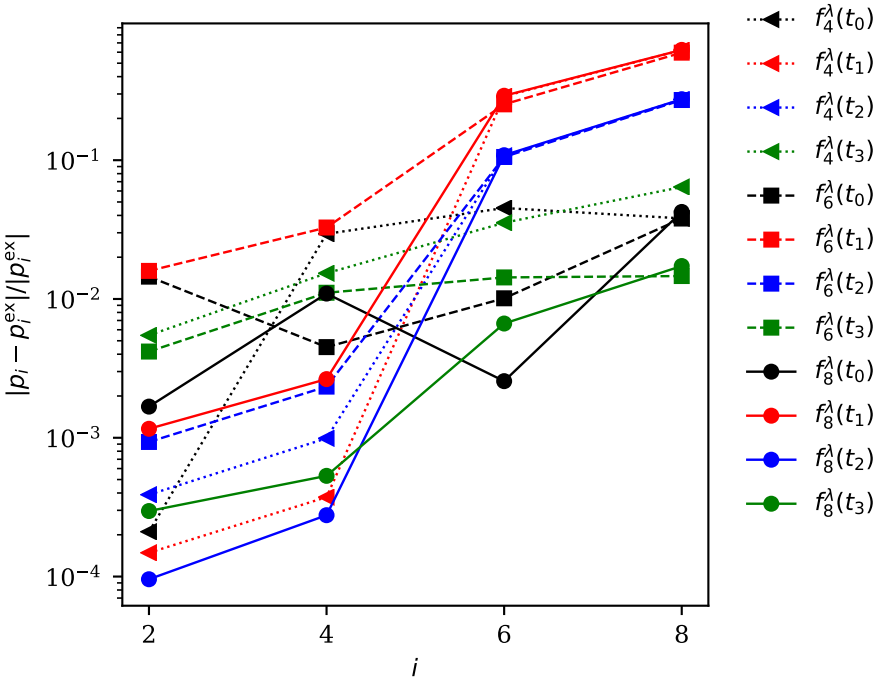

Figure 5: The error in predicting moments in the BGK relaxation problem using GP-accelerated MED for $N = 4, 6$ and 8 moments at time $(t_0 = 0, t_1 = 3, t_2 = 8, t_3 = 20)$. Note that p_i stands for the moment computed from GP regression of MED, while p_i^{ex} is the corresponding exact moment.

5. Acknowledgment

Hossein Gorji acknowledges the funding provided by Swiss National Science Foundation under the grant number 174060. Manuel Torrilhon and Mohsen Sadr acknowledge the funding provided by German Research Foundation (DFG) with the number IRTG-2379 .

References

[1] S. C. Schwartz, Estimation of probability density by an orthogonal series, The Annals of Mathematical Statistics (1967) 1261–1265.

[2] H. Grad, On the kinetic theory of rarefied gases, Communications on pure and applied mathematics 2 (4) (1949) 331–407.

[3] H. Struchtrup, M. Torrilhon, Regularization of grads 13 moment equations: derivation and linear analysis, Physics of Fluids 15 (9) (2003) 2668–2680.

[4] R. O. Fox, Higher-order quadrature-based moment methods for kinetic equations, Journal of Computational Physics 228 (20) (2009) 7771–7791.

[5] W. Dreyer, Maximisation of the entropy in non-equilibrium, Journal of Physics A: Mathematical and General 20 (18) (1987) 6505.

[6] C. D. Levermore, Moment closure hierarchies for kinetic theories, Journal of statistical Physics 83 (5-6) (1996) 1021–1065.

[7] F. J. Och, H. Ney, Discriminative training and maximum entropy models for statistical machine translation, in: Proceedings of the 40th annual meeting on association for computational linguistics, Association for Computational Linguistics, 2002, pp. 295–302.

[8] S. F. Gull, J. Skilling, Maximum entropy method in image processing, in: IEE Proceedings F (Communications, Radar and Signal Processing), Vol. 131, IET, 1984, pp. 646–659.

[9] M. Basseville, Distance measures for signal processing and pattern recognition, Signal processing 18 (4) (1989) 349–369.

[10] T. J. Ulrych, T. N. Bishop, Maximum entropy spectral analysis and autoregressive decomposition, Reviews of Geophysics 13 (1) (1975) 183–200.

[11] R. P. Schaerer, P. Bansal, M. Torrilhon, Efficient algorithms and implementations of entropy-based moment closures for rarefied gases, Journal of Computational Physics 340 (2017) 138–159.

[12] D. A. Drabold, O. F. Sankey, Maximum entropy approach for linear scaling in the electronic structure problem, Physical review letters 70 (23) (1993) 3631.

[13] I. Turek, A maximum-entropy approach to the density of states within the recursion method, Journal of Physics C: Solid State Physics 21 (17) (1988) 3251.

[14] T. Schneider, S. M. Griffies, A conceptual framework for predictability studies, Journal of climate 12 (10) (1999) 3133–3155.

[15] C. D. Hauck, C. D. Levermore, A. L. Tits, Convex duality and entropy-based moment closures: Characterizing degenerate densities, SIAM Journal on Control and Optimization 47 (4) (2008) 1977–2015.

[16] J. N. Kapur, Maximum-entropy models in science and engineering, John Wiley & Sons, 1989.

[17] A. Tagliani, Hausdorff moment problem and maximum entropy: a unified approach, Applied Mathematics and Computation 105 (2-3) (1999) 291–305.

[18] L. R. Mead, N. Papanicolaou, Maximum entropy in the problem of moments, Journal of Mathematical Physics 25 (8) (1984) 2404–2417.

[19] A. Y. Khinchin, Mathematical foundations of information theory, Courier Corporation, 2013.

[20] R. V. Abramov, An improved algorithm for the multidimensional moment-constrained maximum entropy problem, Journal of Computational Physics 226 (1) (2007) 621–644.

[21] R. V. Abramov, The multidimensional moment-constrained maximum entropy problem: A bfgs algorithm with constraint scaling, Journal of Computational Physics 228 (1) (2009) 96–108.

[22] K. P. Murphy, Machine learning: a probabilistic perspective, MIT press, 2012.

[23] P. L. Bhatnagar, E. P. Gross, M. Krook, A model for collision processes in gases. i. small amplitude processes in charged and neutral one-component systems, Physical review 94 (3) (1954) 511.

[24] R. P. Schaerer, M. Torrilhon, The 35-moment system with the maximum-entropy closure for rarefied gas flows, European Journal of Mechanics-B/Fluids 64 (2017) 30–40.

[25] G. W. Alldredge, C. D. Hauck, D. P. OLeary, A. L. Tits, Adaptive change of basis in entropy-based moment closures for linear kinetic equations, Journal of Computational Physics 258 (2014) 489–508.

[26] H. Owhadi, G. R. Yoo, Kernel flows: from learning kernels from data into the abyss, Journal of Computational Physics.

[27] C. E. Rasmussen, C. K. I. Williams, Gaussian Processes for Machine Learning, The MIT Press, 2006.

[28] J. Nocedal, S. Wright, Numerical optimization, Springer Science & Business Media, 2006.

[29] L. Györfi, M. Kohler, A. Krzyzak, H. Walk, A distribution-free theory of nonparametric regression, Springer Science & Business Media, 2006.

[30] E. Snelson, Z. Ghahramani, Sparse gaussian processes using pseudo-inputs, in: Advances in neural information processing systems, 2006, pp. 1257–1264.

[31] A. G. d. G. Matthews, M. van der Wilk, T. Nickson, K. Fujii, A. Boukouvalas, P. León-Villagrá, Z. Ghahramani, J. Hensman, GPflow: A Gaussian process library using TensorFlow, Journal of Machine Learning Research 18 (40) (2017) 1–6.
URL http://jmlr.org/papers/v18/16-537.html