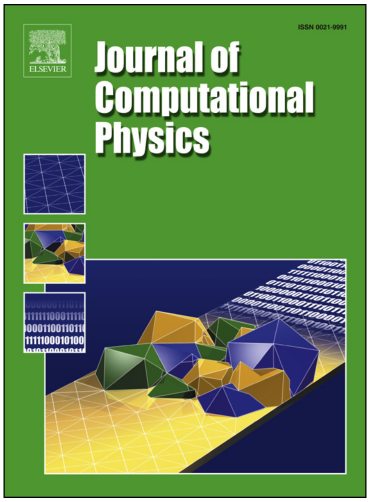


Gaussian process regression for maximum entropy distribution

Mohsen Sadr, Manuel Torrilhon, M. Hossein Gorji

PII: S0021-9991(20)30418-6
DOI: https://doi.org/10.1016/j.jcp.2020.109644
Reference: YJCPH 109644

To appear in: *Journal of Computational Physics*

Received date: 16 July 2019
Revised date: 11 April 2020
Accepted date: 5 June 2020

Highlights

- Devising Bayesian inference for fast evaluation of maximum entropy distribution.
- Adopting Radial basis function for covariance kernel.
- Excellent recovery of bi-modal densities among others.

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 Bhatnagar-Gross-Krook and Boltzmann kinetic equations.

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, training of the GP regression is pursued, where several kernels such as radial basis and Matèrn family are evaluated for our problem setting. Section 4 deals with the assessment of the devised GP-accelerated MED. As the first test case, the accuracy of the fitted GP in predicting bi-modal distributions is studied in § 4.1. In § 4.2, robustness of the GP regression is tested by predicting MED for moments obtained from noisy bi-modal distributions. Then § 4.3 and § 4.4 focus on relaxation of non-equilibrium distributions, governed by Bhatnagar-Gross-Krook (BGK) [23] and Boltzmann equations, respectively. At the end, a conclusion and an outlook for future studies are given in § 5.

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

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

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

$$\lambda(p) = \underset{\lambda^* \in \mathbb{R}^N}{\arg\min} \left\{ C(\lambda^*; p) \right\}, \tag{7}$$

$$\text{where} \qquad C(\lambda^*; p) := Z_{\lambda^*} - \sum_j \lambda_j^* p_j \ . \tag{8}$$

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 $\lambda^{(n)}$ at step n, are updated by solving the linear system

$$\sum_{j=1}^{N} H_{ij}(\lambda^{(n)})\Delta\lambda_j^{(n)} = g_i(\lambda^{(n)}) \tag{9}$$

for $\Delta\lambda^{(n)}$. After random initialization of the Lagrange multipliers, they get updated according to

$$\lambda_i^{(n+1)} \;\; = \;\; \lambda_i^{(n)} + \beta^{(n)}\Delta\lambda_i^{(n)}. \tag{10}$$

Here β is a damping factor and is chosen such that the cost function decreases efficiently. The damping $\beta^{(n)}$ is set to the largest value of the power series $\{s^k\}_{k=0}^{N_s}$ with $s \in (0,1)$ that guarantees the Armijo's rule

$$C(\lambda^{(n)} + \beta^{(n)}\Delta\lambda^{(n)}; p) < C(\lambda^{(n)}; p) + c\beta^{(n)}(\Delta\lambda^{(n)}, g(\lambda^{(n)})), \tag{11}$$

where $(.,.)$ indicates the dot product of vectors. Note that the values of c, s, and N_s need to be tuned appropriately. For our study, we set $c = 10^{-4}$, $s = 1/2$ and $N_s = 30$. A pseudocode describing this direct approach with corresponding values of the free parameters are provided in algorithm (1) [24, 25, 26]. Although H is symmetric-positive-definite, it can become ill-conditioned which can be coped with by using an adaptive basis [27]. For example in [24], Hermite 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

Algorithm 1 Direct approach to find Lagrange multipliers given the moments $p \in \mathbb{R}^N$

```
Set $n = 0$ and sample $\lambda^{(n)}$ uniformly from $[-0.1, 0.1]^N$
Set the tolerance $\epsilon = 10^{-10}$
while $C(\lambda^{(n)}; p) > \epsilon$ do
    Compute Hessian and gradient of the cost function $C(\lambda^{(n)}; p)$
    Solve the linear system in Eq. (9) for $\Delta\lambda^{(n)}$
    Find the largest $\beta^{(n)}$ that satisfies Armijo's rule (11)
    Compute the new guess $\lambda^{(n+1)}$ from Eq. (10)
    Increment n
end while
return $\lambda^{(n)}$
```

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

$$\tilde{\Psi} \sim \mathcal{GP}(0, \mathcal{K}) \tag{12}$$

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 (12) 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 [28]. In practice, we work with parametrized kernels $\mathcal{K}_\Theta$, where the hyperparameters embedded in Θ are found by maximizing the log-likelihood [29]. 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. [29]. Here we consider the radial basis function (RBF) along with Matérn's family for each component $i, j \in \{1, ..., N\}$ we have

$$\mathcal{K}_{\Theta_i}^{\mathrm{RBF}}(x,x') = \sigma_i \exp\left(-r_i^2/2\right) \ , \tag{13}$$

$$\mathcal{K}_{\Theta_i}^{\mathrm{Matérn(12)}}(x,x') = \sigma_i \exp(-r_i) \ , \tag{14}$$

$$\mathcal{K}_{\Theta_i}^{\mathrm{Matérn(32)}}(x,x') = \sigma_i(1+\sqrt{3}r_i)\exp(-\sqrt{3}r_i) \quad \text{and} \tag{15}$$

$$\mathcal{K}_{\Theta_i}^{\mathrm{Matérn(52)}}(x,x') = \sigma_i(1+\sqrt{5}r_i+\frac{5}{3}\sqrt{r_i})\exp(-\sqrt{5}r_i) \ . \tag{16}$$

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

$$\begin{aligned}\ln\left(\tilde{f}\left(\tilde{\Psi}_i(x) \,|\, x \in \mathcal{D}\right)\right) = -\frac{1}{2}\ln\left(\det(\mathcal{K}_{\Theta_i}(x,x'))\right)\\ -\frac{1}{2}{\Psi_i}^T(x)\mathcal{K}_{\Theta_i}(x,x')^{-1}\Psi_i(x') - \frac{M}{2}\ln(2\pi),\end{aligned} \tag{17}$$

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 [30]. It can be shown that the global minimum is attained as more data points are deployed [31]. Once the kernel function $\mathcal{K}$ and its hyperparameters are set, one can evaluate the distribution of $\tilde{\Psi}$ at an arbitrary input point $x^* \in \mathbb{R}^N$. Let x be composed of the training points, therefore

$$\left(\tilde{\Psi}_i(x^*)\middle|\tilde{\Psi}_i(x) = \Psi_i(x)\right) \sim \mathcal{N}(\bar{m}_i, \bar{\Sigma}_i), \tag{18}$$

where $\mathcal{N}(\bar{m}_i, \bar{\Sigma}_i)$ is a normal distribution with mean

$$\bar{m}_i = \mathcal{K}_{\Theta_i}(x^*,x')\mathcal{K}_{\Theta_i}(x,x')^{-1}\Psi_i(x) \tag{19}$$

and variance

$$\bar{\Sigma}_i = \mathcal{K}_{\Theta_i}(x^*,x^*) - \mathcal{K}_{\Theta_i}(x^*,x)\mathcal{K}_{\Theta_i}(x,x')^{-1}\mathcal{K}_{\Theta_i}(x^*,x) \ . \tag{20}$$

Note that $\tilde{m}$ and $\tilde{\Sigma}$ indicate posterior estimates of the mean and the variance, respectively. Since the inversions appearing in Eqs. (19) and (20) only include the training points, the corresponding computations can be done up-front. Therefore computational advantage is gained, as only matrix-vector multiplication is needed for predictions. Although more efficient GP models such as sparse GP [32] could be pursued, in this study we adopt the straight-forward GP regression model available on GPflow [33].

3. Training Gaussian Process

In this section, constructing GP maps for Lagrange multipliers are pursued. The performance of several covariance functions besides accuracy of the GP regression close to realizability limit are assessed.

3.1. Initializing data set

To construct a regression on the Lagrange multipliers as a map from moments, we need to construct a data set. Since the inverse map is cheaper to evaluate, the main idea here is to compute the moments based on a set of Lagrange multipliers. In order to do that we need to introduce a domain for Lagrange multipliers of the form $\Lambda = [\lambda_{\min}, \lambda_{\max}]^N$ to sample from. Furthermore, we need to have a boundary for values of the moments i.e. $\Omega_p = \prod_i [p_{i,\min}, p_{i,\max}]$. Since all scenarios can be shifted and scaled to a reference with zero mean and unity variance, finally we only include the data points corresponding to zero mean and unity variance MEDs. First, $\{\tilde{\lambda}_i\}_{i=1}^N$ are uniformly sampled from Λ resulting in a trial density $f_N^{\tilde{\lambda}}$. The mean μ and the variance σ^2 are computed from $f_N^{\tilde{\lambda}}$ using Gaussian-quadrature. In order to find the corresponding Lagrange multipliers that guarantee zero mean and unity variance, we make use of the coordinate transformation $v' = (v - \mu)/\sigma$. Let f_N^{λ} be the density with zero mean and unity variance. Observe that by equality of measures and assuming $v \in \mathbb{R}$ we get

$$f_N^{\lambda}(v') \quad = \quad \sigma f_N^{\tilde{\lambda}}\left(\sigma v' + \mu\right). \tag{21}$$

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

$$\lambda_i = \sigma^i \tilde{\lambda}_i + \sum_{j=i+1}^{N} \tilde{\lambda}_j \binom{j}{i} \sigma^i \mu^{j-i} \; ; \quad i \in \{1, ..., N\} \tag{22}$$

ensures Eq. (21). Yet since we deal with bounded $v \in \Omega$, an iterative scheme with the initial guess given by Eq. (22) is employed to ensure zero mean and unity variance. Algorithm (2) is introduced in order to create the data set.

Algorithm 2 Generating (λ, p) with $\int_\Omega v f_N^\lambda dv = 0$ and $\int_\Omega v^2 f_N^\lambda dv = 1$ given the moment space Ω_p, and sample space Λ for Lagrange multipliers

```
Set $p' = (0, ..., 0)^T \in \mathbb{R}^N$ with $N = \dim(\Omega_p)$
Set the tolerance $\epsilon = 10^{-10}$
while $p' \notin \Omega_p$ do
    Sample $\tilde{\lambda}$ uniformly from $\Lambda$
    Compute $\mu = \int_\Omega v f_N^{\tilde{\lambda}} dv$ and $\sigma^2 = \int_\Omega v^2 f_N^{\tilde{\lambda}} dv$
    while $\mu > \epsilon$ or $|\sigma - 1| > \epsilon$ do
        Compute $\lambda$ according to Eq. (22)
        Update $\mu = \int_\Omega v f_N^\lambda dv$ and $\sigma^2 = \int_\Omega v^2 f_N^\lambda dv$
        $\tilde{\lambda} \leftarrow \lambda$
    end while
    Update $p'_i = \int_\Omega v^i f_N^\lambda dv$ for $1 \le i \le N$
end while
$p \leftarrow p'$
return $\lambda$ and $p$
```

For the training, we consider $N \in \{4, 6, 8\}$ and numerical integrations are carried out using Gaussian-quadrature with roughly 20 points. The sample space for the Lagrange multipliers has been chosen carefully after a trial and error on the outcome moments obtained from the algorithm (2). As shown in Fig. 1, the generated data points using the uniform sample space of $\Lambda = [-b, b]^N$ with $b \in \{1, 10\}$, suggests that the tail of the moment distribution becomes longer as b increases. This implies that MED with a larger b is better equipped to capture rare events. The training points are generated by setting $b = 10$ and only keeping data points whose moments lie in the space $\Omega_p = [-\epsilon, \epsilon] \times [1-\epsilon, 1+\epsilon] \times [-1, 1] \times [1, 4] \times [-4, 4] \times [1, 15] \times [-25, 1] \times [1, 110]$.

3.2. 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.

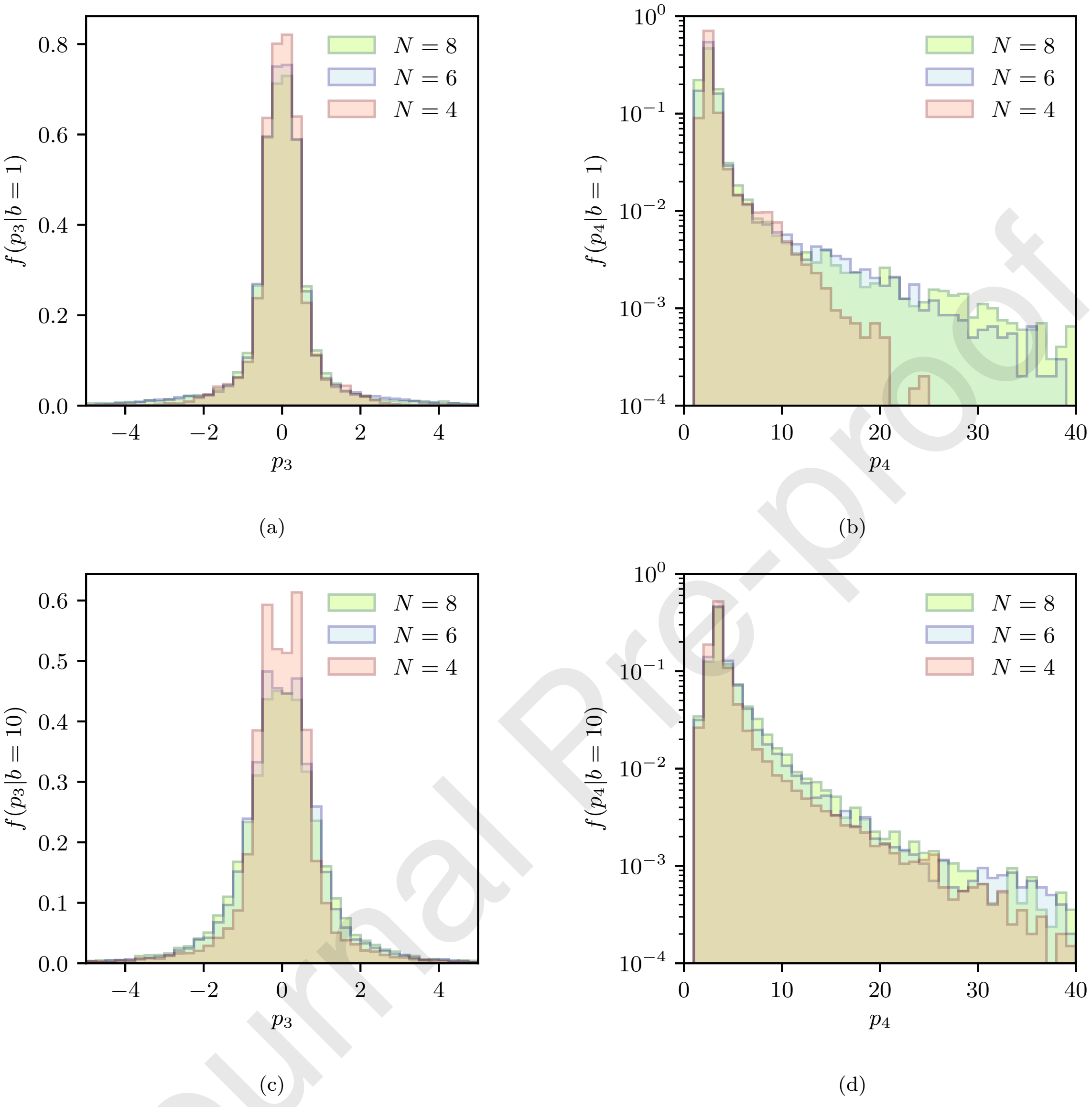
3.3. Kernel comparison

We consider MED with $N = 6$ moments as the target distribution, where the appropriate kernel and number of training data points M should be found. First, let us consider the radial basis function (RBF) for the kernel choice. Once the hyperparameters of Eq. (13) are found via maximizing the log-likelihood given by Eq. (17), the accuracy of predictions over unseen data is investigated. As shown in Fig. 2, 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.

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. 3, RBF provides a better estimation for this data set.

3.4. Cost of data generation and training

Although the described algorithm(2) is straight-forward, the iterations on $\tilde{\lambda}$ required to ensure $p \in \Omega_p$ besides zero mean and unity variance, can become costly. As it can be seen in Fig. 4, the computational time for generating the data set τ^{gen} scales almost linearly with the number of data points. However, the cost of generating relevant data points increases



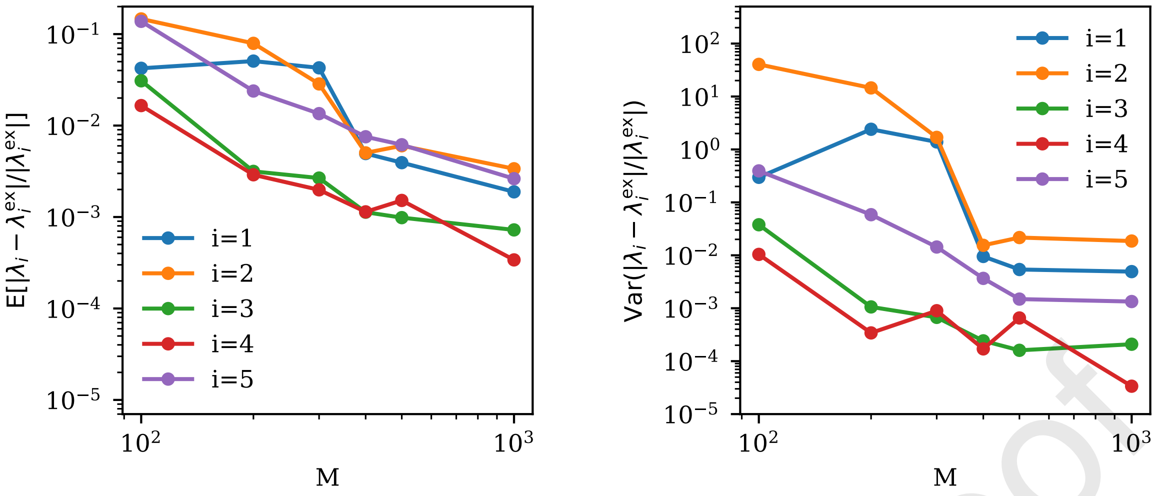
(a)

(b)

(c)

(d)

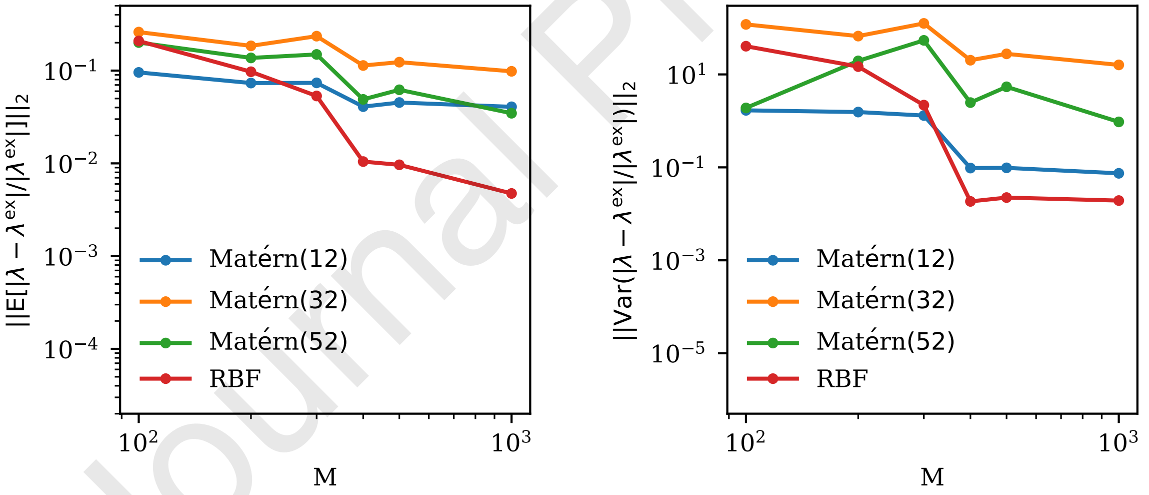
Figure 1: Probability density function of the moments p_3 and p_4 obtained from 10^4 data points using algorithm (2). The sample space $\Lambda = [-b, b]^N$ for the Lagrange multipliers varies with $N \in \{4, 6, 8\}$ and $b \in \{1, 10\}$.



(a)

(b)

Figure 2: Expectation and variance of the relative error in predicting λs using RBF. Here, $\lambda^{\mathrm{ex}} \in \mathbb{R}^N$ with $N = 6$ indicates the exact solution of MED taken from untrained subset of the data set. The statistics are performed over 2000 testing points.



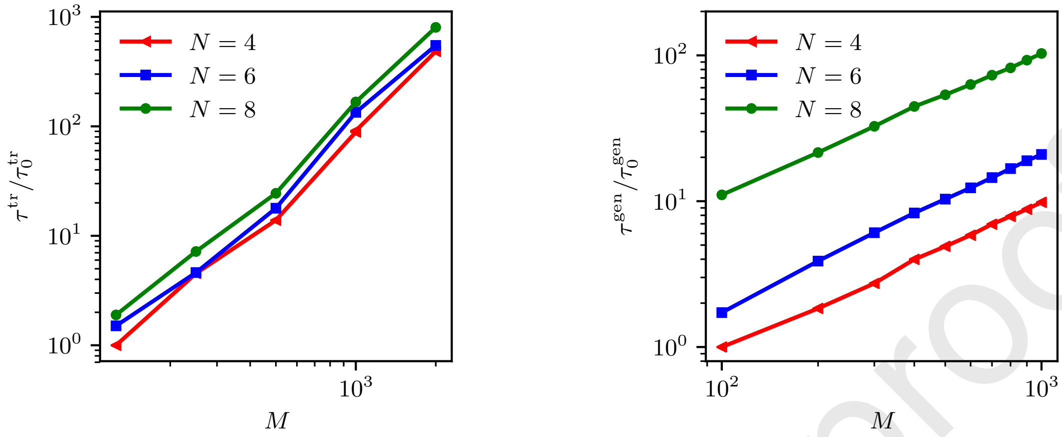
(a)

(b)

Figure 3: Convergence comparison using different kernels. The L^2-norm of the expectation and the variance of the relative error are shown. Here, $\lambda^{\mathrm{ex}} \in \mathbb{R}^N$ with $N = 6$ indicates the exact solution of MED taken from untrained subset of the data set. The statistics are performed over 2000 testing points.

as more moments are considered.

The GP hyperparameters are trained by maximizing the log-likelihood, as explained in § 2.2. The execution time for training τ^{tr} shown in Fig. 4 includes cost of the Cholesky factorization besides optimizing the hyperparameters with a tolerance 10^{-6}.



(a)

(b)

Figure 4: Averaged execution time τ^{gen} for tuning hyperparameters and τ^{tr} for generating the data set, are depicted at left and right, respectively. Here N denotes the number of moments and M the number of employed data points. Execution times are normalized by the computational time corresponding to the case of $N = 4$ and $M = 100$.

3.5. Accuracy in the limit of realizability

Here, we investigate the accuracy of the trained GP at the limit of moment realizability. Let us consider MED with $N = 4$ moments along with RBF as the kernel function. Following [34, 35], the moment problem is physically realizable for $N = 4$, if the sufficient condition

$$p_4 \geq p_3^2 + 1 \tag{23}$$

holds for the standardized moments $p \in \mathbb{R}^4$. In order to show accuracy of the trained GP at the points near the limit $p_4 = p_3^2 + 1$, we investigate the GP predictions for a set of standardized input moments

$$D_p^{\text{test}} = \left\{ \left(\begin{smallmatrix} 0 \\ 1 \\ \alpha \\ \alpha^2+1+d \end{smallmatrix} \right) \middle| \alpha = \alpha_{\min} + ih,\ i = 1, ..., N_{\text{test}},\ h = (\alpha_{\max} - \alpha_{\min})/N_{\text{test}} \right\} \tag{24}$$

where $d \in \{0.04, 0.02, 0.01\}$, $\alpha_{\max} = -\alpha_{\min} = 0.5$, and $N_{\text{test}} = 100$. As shown in Fig. 5, by decreasing d, the relative error and the variance of predictions increase. On the other hand, as expected, the accuracy improves by increasing the number of training points M. Therefore this investigation suggests that the GP regression for MED becomes less reliable once moments close to the realizability border are encountered. Moreover, the upper limit

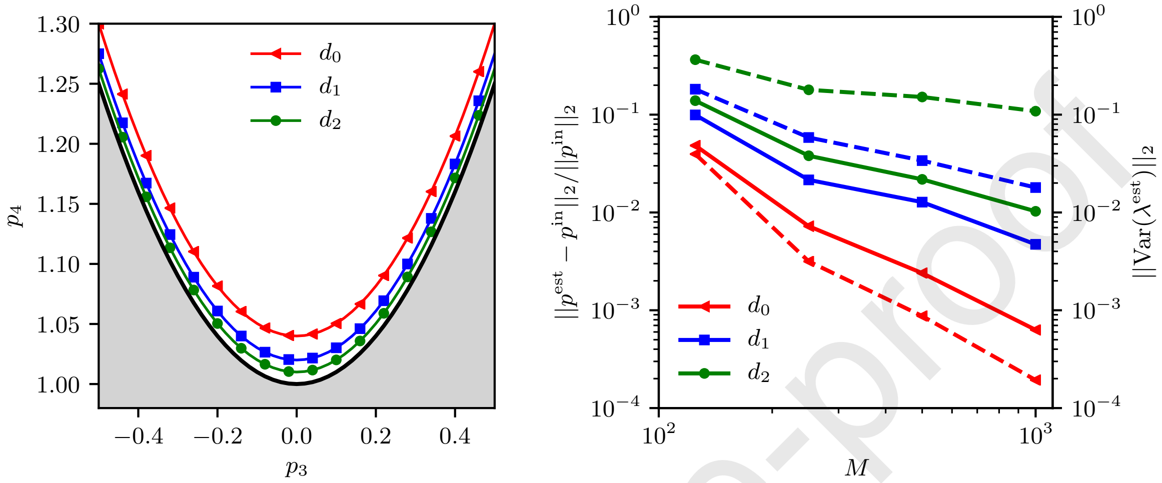


Figure 5: Left: Input points depicted in (p_3, p_4) plane with solid black curve on $p_4 = p_3^2+1$, Right: Average of moments relative error using data-driven MED for input moments $p^{\text{in}} \in D_p^{\text{test}}$, and variance of the predicted Lagrange multipliers shown in solid and dashed lines, respectively

of realizable moments is investigated here. First, since data points with 4th order moment $p_4 \in [1, 4]$ are considered here, let us evaluate the accuracy of GP in predicting MED as points of interest approaches the upper limit. Let us define a set of points as

$$U_p^{\text{test}} = \left\{ \begin{pmatrix} 0 \\ 1 \\ \beta \\ 4-d \end{pmatrix} \middle| \beta = \beta_{\min} + ih,\ i = 1, ..., N_{\text{test}},\ h = (\beta_{\max} - \beta_{\min})/N_{\text{test}} \right\} \qquad (25)$$

where $d \in \{0.1, 0.05, 0\}$ and $\beta_{\max} = -\beta_{\min} = 0.1$, and $N_{\text{test}} = 200$ illustrate the number of testing points. As shown in the Fig. 6, similar to the lower limit of physical realizability, the accuracy in prediction decreases as the upper limit of moments is approached.

Finally, the trained GP is tested in estimating the MED near the line $p_3 = 0$ with $p_4 > 3$ which are not realizable with MED. In order to evaluate the accuracy of MED with $N = 4$

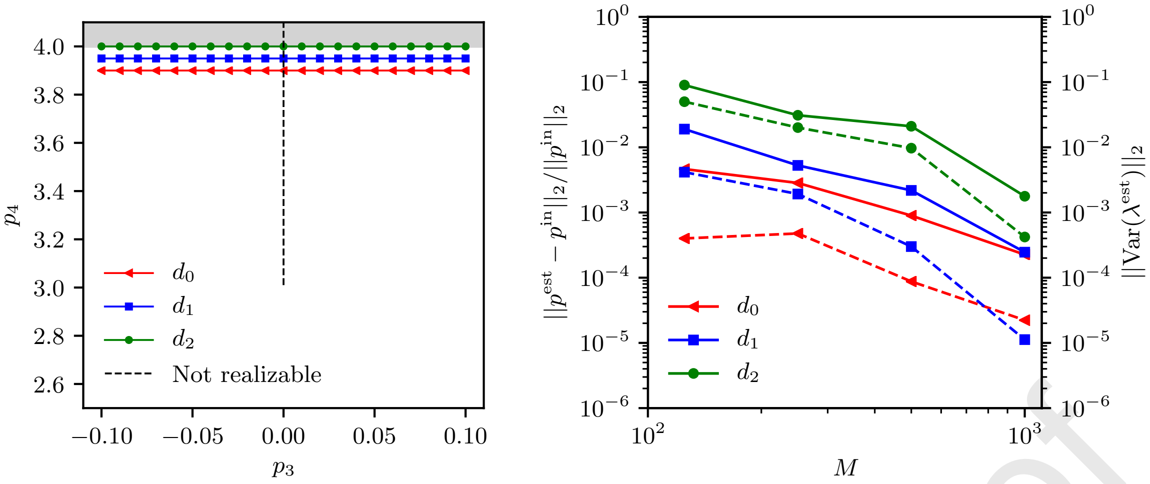


Figure 6: Left: Input points depicted in (p_3, p_4) plane with black dashed line depicting $p_3 = 0$ with $p_4 > 3$, Right: Average of moments relative error using data-driven MED for input moments $p^{\mathrm{in}} \in U_p^{\mathrm{test}}$, and variance of the predicted Lagrange multipliers shown in solid and dashed lines, respectively

around this limit of realizability, let us take moments from the set

$$S_p^{\mathrm{test}} = \left\{ \begin{pmatrix} 0 \\ 1 \\ \beta/d \\ (10\beta d)^2+3 \end{pmatrix} \middle| \beta = \beta_{\min} + ih, \ i = 1, ..., N_{\mathrm{test}}, \ h = (\beta_{\max} - \beta_{\min})/N_{\mathrm{test}} \right\} \qquad (26)$$

where the parameters are $d \in \{1, 8, 64\}$ and $N_{\mathrm{test}} = 100$ indicates the number of testing points. Similar to the lower bound of realizability, it can be observed from Fig. 7 that the relative error, as well as the variance of the predictions, decrease by deploying more training points. However, the error in predictions as the point of interest approaches the upper limit of realizability, i.e., by increasing d, is negligible, which can be explained by having an excess of testing points near the equilibrium point, i.e., $(p_3, p_4) = (0, 3)$.

4. Results

In this section, the trained GP is employed for predicting different scenarios relevant in kinetic problems. 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

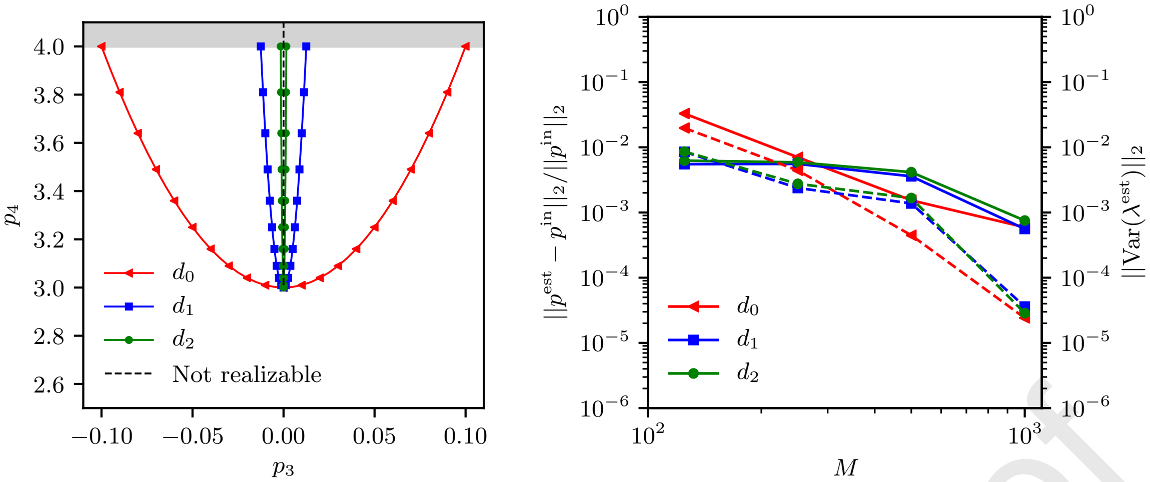


Figure 7: Left: Input points depicted in (p_3, p_4) plane with black dashed line depicting $p_3 = 0$ with $p_4 > 3$, Right: Average of moments relative error using data-driven MED for input moments $p^{\text{in}} \in S_p^{\text{test}}$, and variance of the predicted Lagrange multipliers shown in solid and dashed lines, respectively

mean and unity variance are obtained. After normalization, the velocity sample space is set by adopting $v_{\max} = -v_{\min} = 10$.

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

Bi-modal distributions are prototype of non-equilibrium phenomena in kinetic problems. For example they show up as simplified solutions of shock waves in rarefied gas kinetics [36]. We employ the trained GP with the RBF kernel to predict the bi-modal density of the form

$$f^{\text{bi}}(v|\mu_1, \sigma_1, \mu_2, \sigma_2) = \frac{1}{2}\left[f^{\mathcal{N}}(v|\mu_1, \sigma_1) + f^{\mathcal{N}}(v|\mu_2, \sigma_2)\right], \tag{27}$$

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

$$D_{KL}(f^{\text{bi}}||f_N^{\lambda}) = \int_{\Omega} f^{\text{bi}}(v) \ln\left(f^{\text{bi}}(v)/f_N^{\lambda}(v)\right) dv \tag{28}$$

is used here. Three different scenarios {(a),(b),(c)} corresponding to $(\mu_1, \sigma_1) \in \{(0.8, 0.3), (0.9, 0.2), (0.95, 0.15)\}$ are considered, where predictions are provided based on the GP regression with $N = 4, 6$ and 8 moments. The results depicted in Fig. 8 show that even with f_4^λ a good recovery is achieved. Although predictions of MED suggest that by increasing the number of moments better estimation of the bi-modal distribution can be obtained, such improvement were not observed for the predictions in the test case (b) and (c) from $N = 6$ to $N = 8$. This discrepancy can be explained by noticing high values for the variance of posterior for mentioned points, i.e. lack of training data near prediction points. As expected by merging the two modes, better agreement is obtained between the GP-accelerated MED and the bi-model one.

To further evaluate accuracy and performance of the data-driven MED, the bi-modal test case was also studied using the standard algorithm (1). While reasonable accuracy in predicting the exact Lagrange multipliers and their outcome moments are obtained via GP estimates as shown in Fig. 9, a speedup of at least two orders of magnitude compared to the direct approach is observed. The predictions are improved overall as the number of moments is increased.

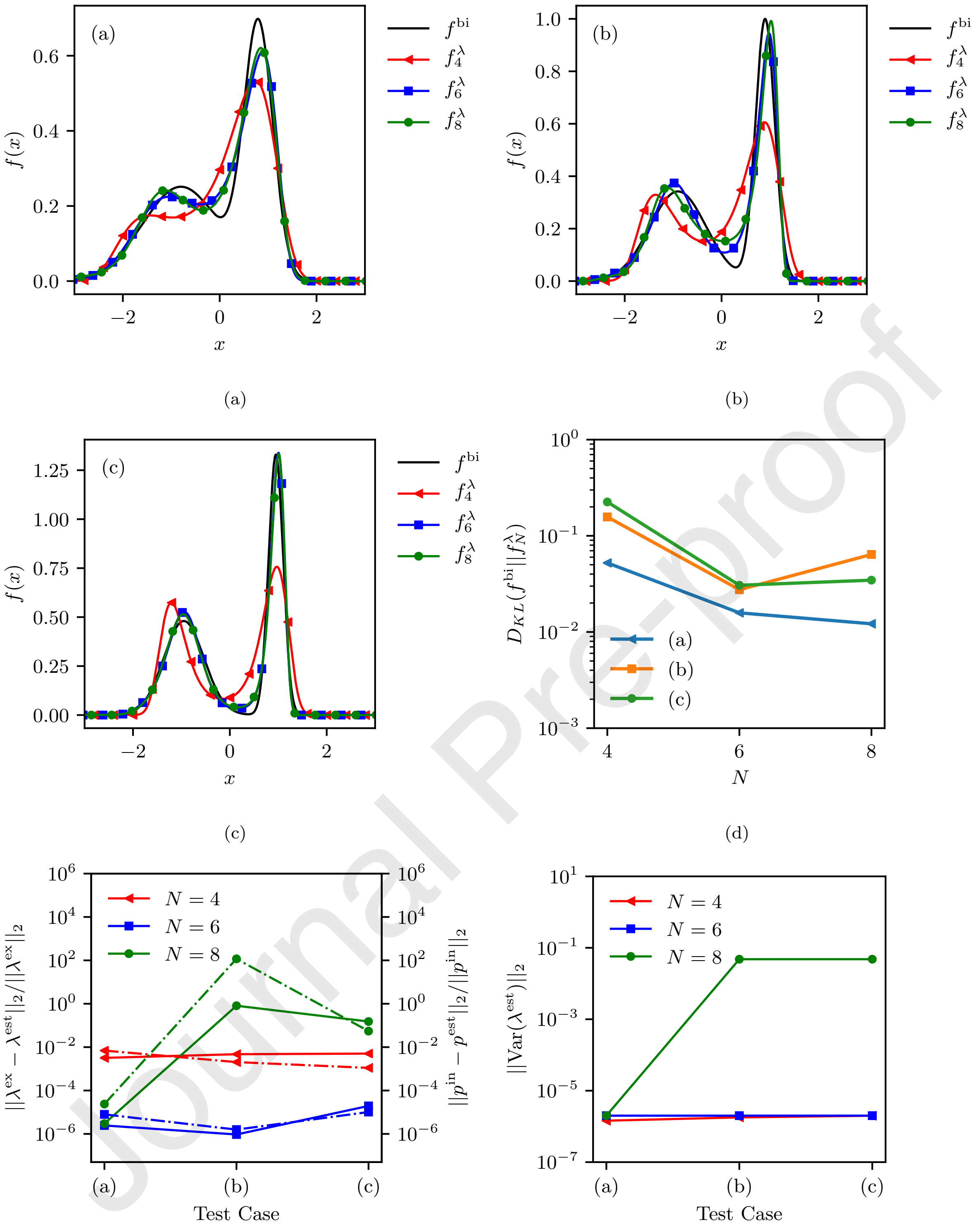
4.2. Test case #2: noisy bi-modal distribution

To test the robustness of our data-driven MED estimator, here we consider a perturbed bi-modal distribution

$$f_\epsilon^{\mathrm{bi}}(v|\mu_1, \sigma_1, \mu_2, \sigma_2) = f^{\mathrm{bi}}(v|\mu_1, \sigma_1, \mu_2, \sigma_2)(1 + \epsilon), \tag{29}$$

where ϵ is a random variable with the normal density $f^{\mathcal{N}}(0, 0.1)$. The values of (μ_1, σ_1) are taken from § 4.1.

As depicted in Fig. 10, devised GP estimators provide reasonable performance for perturbed scenarios. Here, it can be observed that although MED with higher moments has the potential of describing more complicated distributions, the sensitivity of higher-order Lagrange multipliers to the input moments reduces the robustness.



(a) (b)

(c) (d)

(e) (f)

Figure 8: Recovering bi-modal probability densities using MEDs accelerated by the GP regression with $N = 4, 6$ and 8 moments. The estimated densities are shown in sub-figures (a)-(c) for test cases (a)-(c), respectively. The KL-divergence between estimated MEDs and the bi-modal distribution, relative error of the GP estimator with respect to exact values of Lagrange multipliers and outcome moments, and variance of predictions are presented in (d)-(f), respectively.

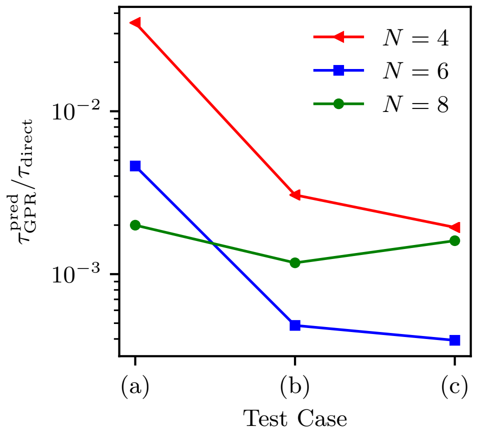


Figure 9: The ratio between computational times of the GP regression and the direct approach are shown for computing the Lagrange multipliers.

4.3. Test case #3: 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

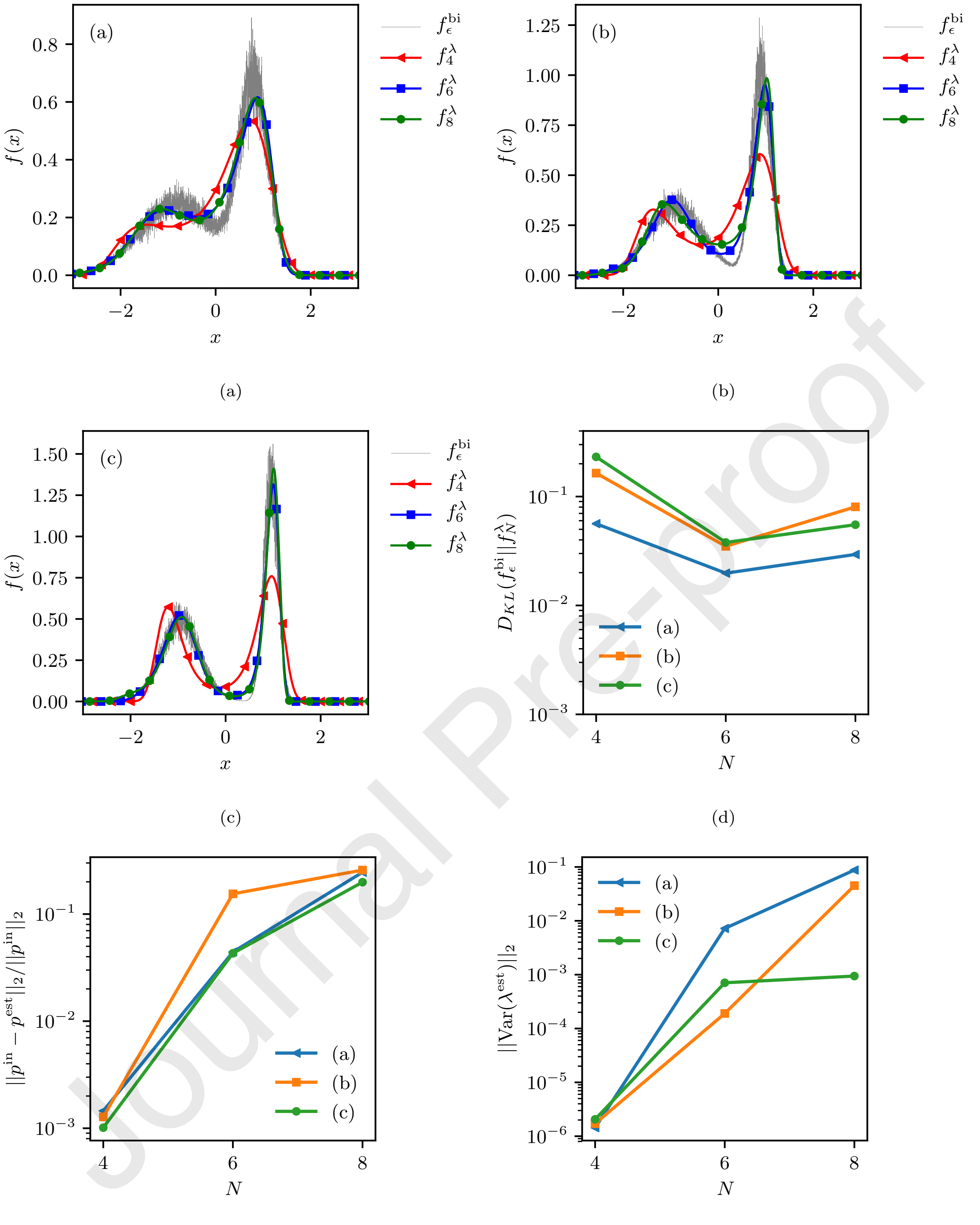
$$\frac{\partial f(v|t)}{\partial t} = \nu(f^{\mathcal{N}}(v|0,1) - f(v|t)) \ . \tag{30}$$

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

$$f^{\text{ex}}(v|t) = [1 - \exp(-\nu t)]\, f^{\mathcal{N}}(v|0,1) + \exp(-\nu t) f(v|t_0) \ . \tag{31}$$

Here, we use bi-modal normal distribution described in § 4.1 with $(\mu_1, \sigma_1) = (0.98, 0.2)$ as the initial density.

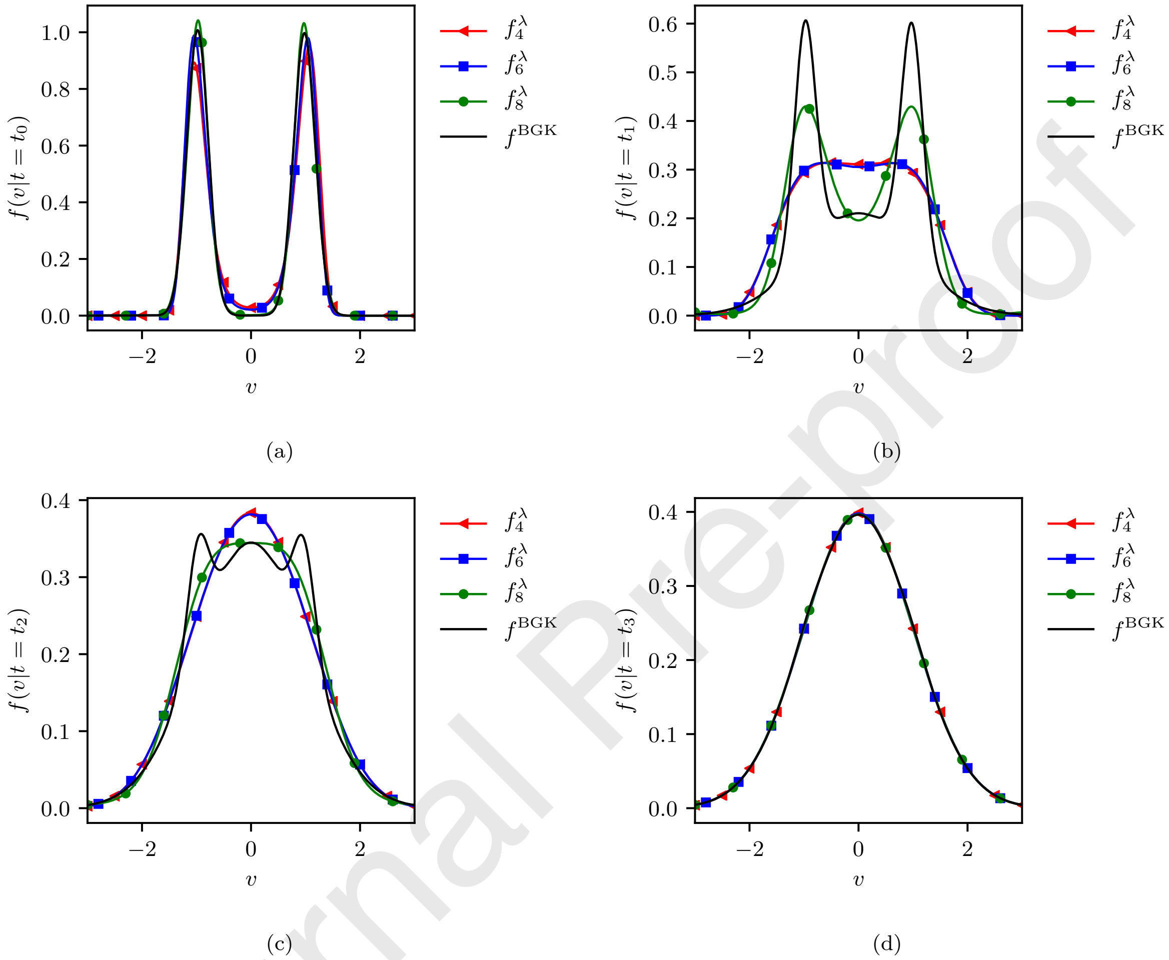
In order to solve Eq. (30) using MED, the Lagrange multipliers corresponding to the set of moments at time t need to be evaluated. Applying the devised GP regression, trained for $\lambda \in \mathbb{R}^N$ with $N = 4, 6$ and 8, the Lagrange multipliers are estimated. Observe that the moments $p(t)$ can be computed analytically from Eq. (30). Therefore, at each time instant, the moments and subsequently the trained GP map, are found. The estimated f^{λ}_N together with its moments are compared with respect to the corresponding exact solution as shown



(a) (b) (c) (d) (e) (f)

Figure 10: Estimating the noisy bi-modal distribution f_ϵ^{bi} with GP-accelerated MED f_N^λ, where $N \in \{4, 6, 8\}$ for test cases $\{(a), (b), (c)\}$. Probability density functions are shown in (a)-(c). Also the KL-divergence between distributions, the relative error in outcome moments and the variance of the Lagrange multipliers are shown in (d)-(f), respectively.

in Fig. 11. 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. 12. 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.



(a)

(b)

(c)

(d)

Figure 11: Capturing BGK relaxation using GP-accelerated MED for $N = 4, 6$ and 8 moments at time $t \in \{0, 3, 8, 20\}$.

4.4. Test case #4: recovering Boltzmann relaxation

In this section, we investigate accuracy of the devised data-driven MED in estimating an exact solution of the Boltzmann equation. Consider the homogeneous and dimension-less

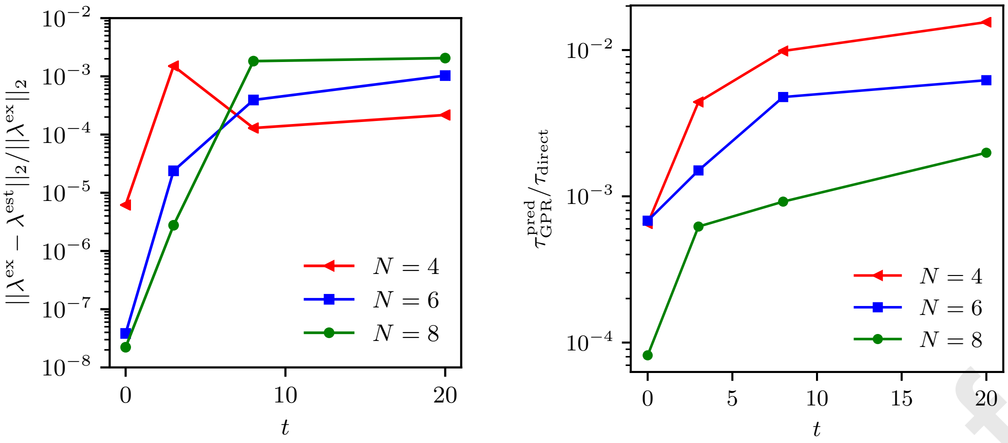


Figure 12: Left: Relative error of the GP regression in predicting Lagrange multipliers of MED for $N = 4, 6$ and 8, Right: The ratio between computational times required by direct and GP approaches for $N = 4, 6$ and 8 at time $t \in \{0, 3, 8, 20\}$.

Boltzmann equation in the velocity space v

$$\frac{\partial f(v, \hat{t})}{\partial \hat{t}} = \frac{1}{4\pi} \int \int [f(v', \hat{t}) f(w', \hat{t}) - f(v, \hat{t}) f(w, \hat{t})] \phi(\mathcal{X}) d\bar{\Omega} dw, \tag{32}$$

where $\hat{t}$ is the normalized time, superscript $(.)'$ denotes pre-collision velocities of the collision pair, w is the velocity of the collision partner and $d\bar{\Omega} = \sin(\mathcal{X}) d\mathcal{X} d\epsilon_0$ with scattering angle $\mathcal{X}$ and $\epsilon_0 \in [0, 2\pi]$. Here $g = |v - w|$ is the magnitude of relative velocity. Note that in the case of the isotropic scattering we have $\phi(\mathcal{X}) = 1$. As shown in [37], an exact solution

$$f^{\mathrm{Bolt}}(v, \hat{t}) = \frac{\exp(-v^2/2K(\hat{t}))}{2K(\hat{t})[2\pi K(\hat{t})]^{3/2}} \left[(5K(\hat{t}) - 3) + \frac{1 - K(\hat{t})}{K(\hat{t})} v^2 \right] / \mathcal{I} \tag{33}$$

can be obtained, where $\mathcal{I}$ is the normalizing factor and

$$K(\hat{t}) = 1 - \exp(\hat{t}/6)\ . \tag{34}$$

Note that Eq. (33) provides a valid solution of the Boltzmann equation once $\hat{t} \geq 6 \log(5/2)$. As derived in [37], the even moments for this isotropic setting evolve according to

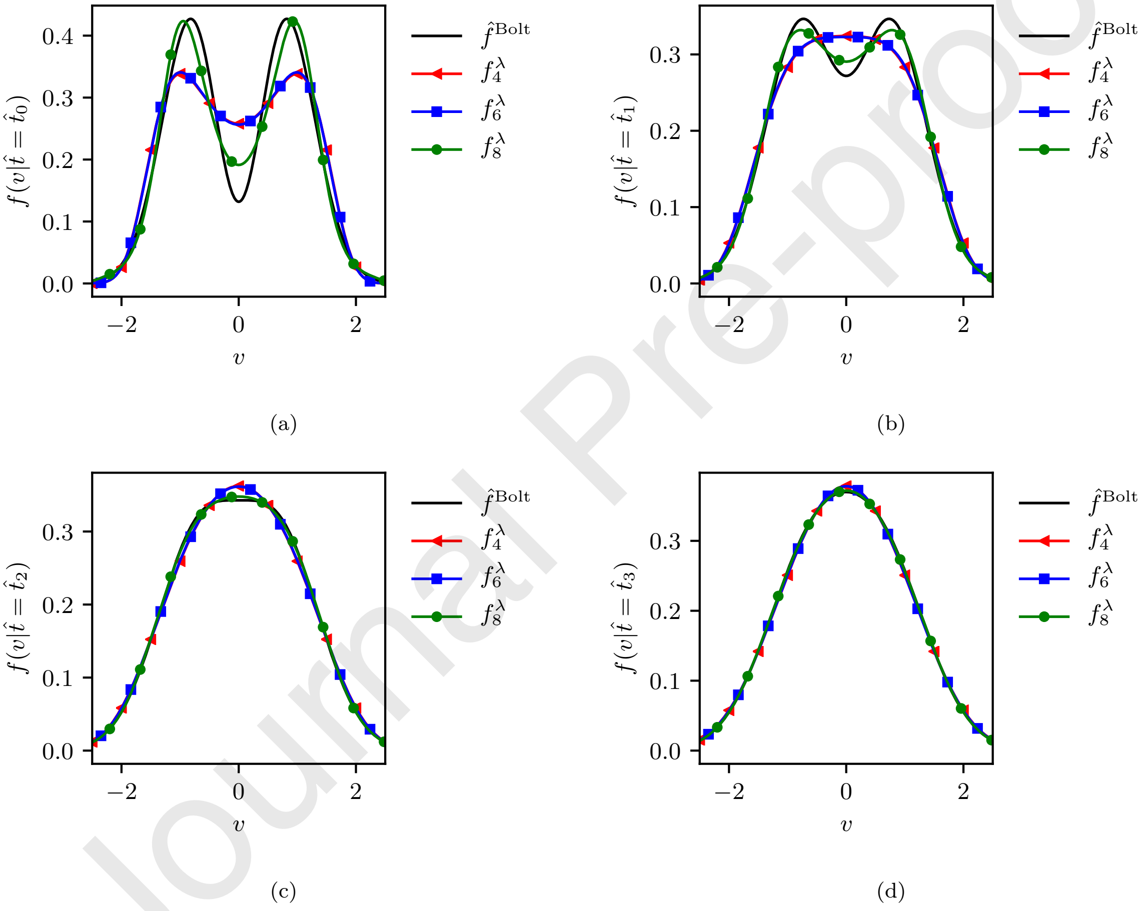
$$p_{2n}(\hat{t}) = \frac{(4n+1)!}{2^{2n}(2n)!} M_{2n} \quad \text{and} \tag{35}$$

$$M_{2n}(\hat{t}) = K^{2n-1}(\hat{t}) \left[2n - (2n-1) K(\hat{t}) \right], \tag{36}$$

for $n \in \{0, 1, ...\}$. In order to deploy our GP estimator of MED, the input moments need to be standardized via

$$\hat{p}_k(\hat{t}) = \frac{p_k(\hat{t})}{p_2(\hat{t})^{k/2}}, \quad \text{for } k = 1, ..., N \ . \tag{37}$$

By plugging standardized moments at any time $\hat{t}$ as the input in the trained GP, the outcome Lagrange multipliers are predicted. As shown in Figs. 13-14, the trained GP-accelerated MED estimator provides an accurate solution of the Boltzmann equation. As 205 expected, the accuracy in prediction improves once more moments are considered.



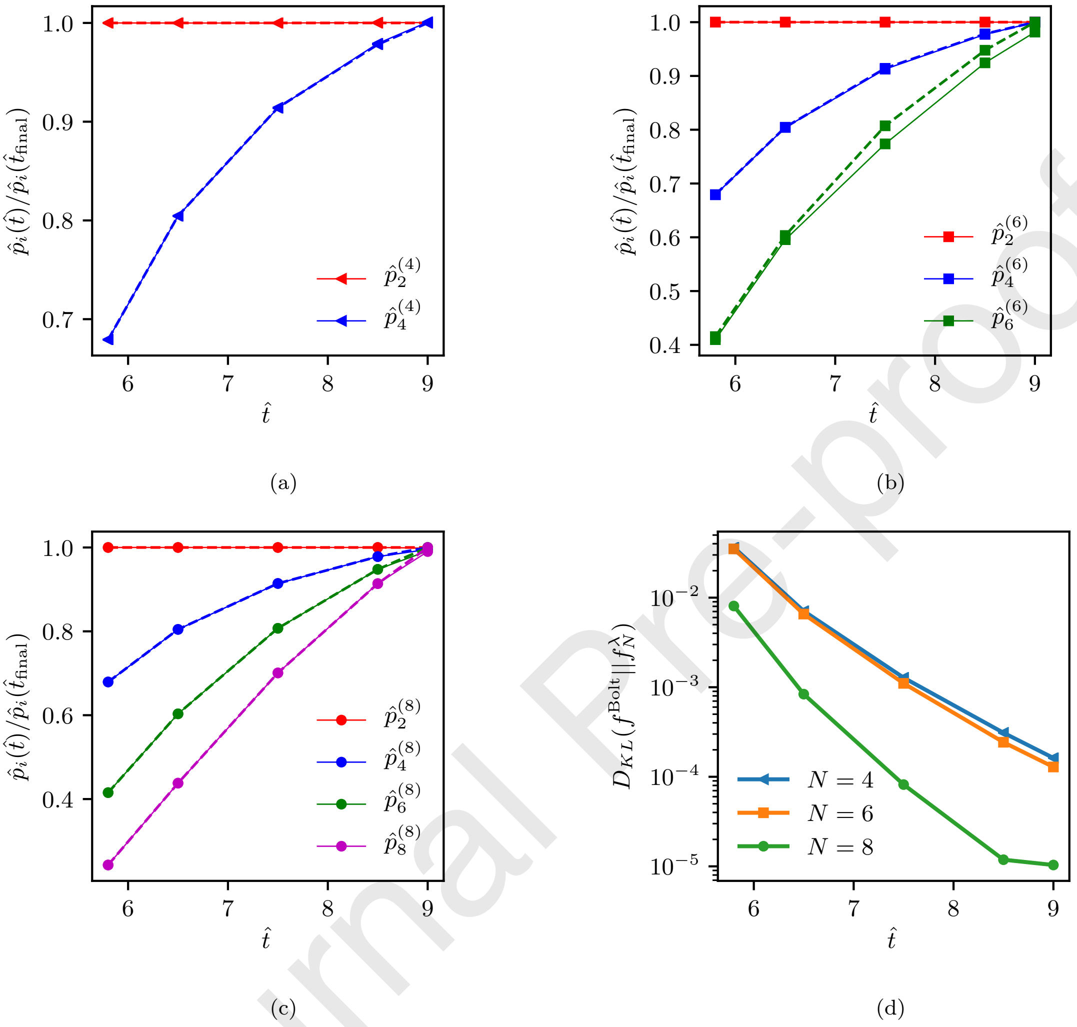
(a)

(b)

(c)

(d)

Figure 13: Estimating solution of the Boltzmann equation at time $\hat{t} \in \{5.8, 6.5, 7.5, 8.5\}$ by devised GP-accelerate MED for $N = 4, 6$ and 8 moments.



(a)

(b)

(c)

(d)

Figure 14: Relaxation of standardized moments $\hat{p}^{(N)} \in \mathbb{R}^N$ obtained from exact solution of the Boltzmann equation and devised MED estimation f_N^λ with $N \in \{4, 6, 8\}$ depicted by dashed and solid lines, respectively, together with the KL divergence between solutions.

5. 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, noisy distributions, and BGK/Boltzmann type relaxations show encouraging performance of the GP-accelerated MED. However, GP prediction of Lagrange-multipliers becomes less accurate once moments near the realizability limit are encountered. This issue can be tackled by enriching the training points near those limits. For future studies, higher dimensional sample spaces besides sparse GPs will be pursued to further generalize the devised scheme.

6. 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 grad's 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] P. Wolfe, Convergence conditions for ascent methods, SIAM review 11 (2) (1969) 226–235.

[26] P. Wolfe, Convergence conditions for ascent methods. ii: Some corrections, SIAM review 13 (2) (1971) 185–188.

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

[28] H. Owhadi, G. R. Yoo, Kernel flows: from learning kernels from data into the abyss, Journal of Computational Physics (2019).

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

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

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

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

[33] 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`

[34] J. McDonald, M. Torrilhon, Affordable robust moment closures for cfd based on the maximum-entropy hierarchy, Journal of Computational Physics 251 (2013) 500–523.

[35] N. I. Akhiezer, N. Kemmer, The classical moment problem: and some related questions in analysis, Vol. 5, Oliver & Boyd Edinburgh, 1965.

[36] H. M. Mott-Smith, The solution of the Boltzmann equation for a shock wave, Physical Review 82 (6) (1951) 885.

[37] M. Krook, T. T. Wu, Exact solutions of the Boltzmann equation, The Physics of Fluids 20 (10) (1977)

1589–1595.

Declaration of interests

■ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: