A Realizability Preserving Discretization and Entropy Modification for Intrusive Polynomial Moment Methods

1 Introduction

Today a lot of simulation applications make use of deterministic models in order to predict the behavior of physical systems. However, the question arises whether these models allow a good investigation of the problem in the case of non-deterministic inputs. Deterministic inputs are often not available, as for example the choice of model parameters as well as measurements of boundary or initial conditions contain non-predictable errors. As a result, one might ask the question whether we can rely on the results of a solution computed without taking into account these known and present uncertainties.

An approach to treat these uncertainties has been proposed in [10], where the so called Stochastic Galerkin (SG) method has been derived. This popular method is based on polynomial chaos [20] and promises pseudo-spectral convergence for smooth data. Simple applications, such as the steady diffusion equation [21] or the advection equation [11] show this expected spectral convergence. However, in the case of non-smooth data, which can for example arise when making use of non-smooth initial conditions or non-linear transport equations, the SG method can lead to oscillatory solutions. Furthermore, the problem can loose important characteristics, such as for systems hyperbolicity.

In [17] the Intrusive Polynomial Moment (IPM) method was introduced and has been applied to several problems [18] and [7]. It aims at preserving hyperbolicity of conservation equations and preventing oscillations in the space of uncertainties. The IPM approach is a minimum entropy (M_N) method applied to the moment system arising in Uncertainty Quantification. M_N methods are frequently used in the field of transport theory, see for example [14, 12, 8, 3, 4, 13, 16, 9]. The key idea of minimal entropy methods is to close the moment system by minimizing an entropy while preserving a set of given moments. This leads to the problem of finding a set of so called dual variables that fulfill the moment constraints. One property of these methods is the dissipation of the chosen entropy over time. To limit oscillations, this property is used to prohibit a solution that falls below or exceeds certain bounds u_- and u_+ by defining an entropy that is infinite at these two bounding values. In the context of oscillations, this means that the method will damp under- and overshoots. Aside from limiting oscillations, the resulting system will be hyperbolic. The main disadvantage of M_N and IPM methods is that they are time consuming as an optimization problem needs to be solved in every spatial cell for each time step. In order to reduce costs, high order schemes are desirable, which allow a coarser discretization. Note that in addition to using high order schemes, M_N methods can compete with classical methods, due to the growth of computing power and their ease of parallelization, see [13].

As already pointed out, one advantage of the IPM solution is that the entropy dissipation ensures a solution bounded by the two values u_- and u_+. To gain solutions with minimal over- and undershoots, two appropriate bounds u_- and u_+ must be chosen. In the context of scalar hyperbolic problems, the maximum principle holds, meaning that the exact solution will be bounded by the minimal and maximal value of the initial condition, which we denote by u_{min} and u_{max}. Therefore, one makes the choice $u_- = u_{min} - \Delta u$ and $u_+ = u_{max} + \Delta u$, where $\Delta u > 0$ is the distance between the bounds of the true solution and the IPM solution. Note that in order to guarantee a finite entropy at time $t = 0$, the IPM bounds are not allowed to touch the minimal and maximal value of the initial condition, meaning that $\Delta u = 0$ is not allowed. Consequently, the IPM method will violate the maximum principle. In order to approach the maximum principle and to obtain minimal over- and undershoots, which will be in the order of Δu, one wishes to set the bounds u_- and u_+ close to the minimal and maximal value of the initial condition. In addition to a solution with only small over- and undershoots, this will allow the choice of a sharp CFL condition as maximal velocities of the system are controlled by the bounds. Unfortunately, it can be seen that the standard M_N discretization will fail as soon as the distance between the exact and the IPM bounds Δu is chosen too small. This is due to the fact that the method encounters moments which do not lie in the so called realizable set, meaning that no set of corresponding dual variables exists. For M_N methods, this is an often occuring problem, which has been studied in [6, 2, 1]. Additionally, despite of preventing oscillations at the bounds, the IPM method yields solutions with oscillations at intermediate values.

In order to provide a better understanding of M_N closures, our aim is to further investigate the source of the realizability problem for a general hyperbolic equation containing uncertainties. The standard approach to ensure realizability is to use a weaker CFL condition, which has been derived in [2]. However, we will show that this CFL condition will not ensure realizability in the IPM framework as it only guarantees positivity. To overcome this problem, we introduce a small error onto the moments of the previous time step, yielding realizable moments at the next time step for a certain CFL condition. In addition to remaining realizable, the modified scheme can easily be extended to higher order. Having introduced a modified scheme, we are able to choose the bounds u_- and u_+ close to the exact solution. As a result, over- and undershoots at the bounds can heavily be damped, however we will see oscillations at values between u_- and u_+. We show that an entropy remaining finite at u_- and u_+ will suffice to introduce these bounds onto the solution. This allows picking a new entropy which is motivated by kinetic theory. In addition to fulfilling the maximum principle, the new entropy will yield less oscillatory results. Furthermore, one can use the CFL condition of the deterministic problem, allowing stability for big time steps without adding unwanted numerical diffusion. The approximation error of the new kinetic entropy will be studied numerically. Furthermore, we show that the kinetic entropy will dissipate oscillations in time.

This paper is structured as follows: In Section 2, we will briefly re-derive the Stochastic Galerkin and the Intrusive Polynomial Moment method for a scalar hyperbolic partial differential equation. Section 3 will discuss the numerical discretization of the derived moment system and will point out realizablility problems as well as approximation properties. The problem of realizability is further investigated in Section 4 and a CFL condition and a modification of the standard M_N algorithm is proposed to prevent this problem. In Section 5, we will discuss properties of the approximation with M_N, which will lead to the introduction of a new entropy. Section 6 will show numerical results for the uncertain Burger's and advection equations. In Section 7, we will summarize our findings and give an outlook on future work.

2 Stochastic Galerkin and IPM

In this Section, we derive the Stochastic Galerkin and IPM system for a scalar and hyperbolic equation

$$\partial_t u(t,x,\xi) + \partial_x f(u(t,x,\xi)) = 0, \tag{1a}$$

$$u(0,x,\xi) = u_0(x,\xi). \tag{1b}$$

From the theory of scalar, hyperbolic problems, we know that the solution will be bounded by the minimal and maximal value of the initial condition. Furthermore, any convex function $s(u) : \mathbb{R} \to \mathbb{R}$ is an entropy to this problem, meaning that there exists an entropy flux h with $h'(u) = s'(u)f'(u)$ such that

$$\partial_t s(u) + \partial_x h(u) = 0$$

for strong solutions. The uncertainty ξ with the probability distribution function f_Ξ enters by the initial condition and can be interpreted as measurement error of some physical field u at time $t = 0$. We wish to determine how this measurement error will propagate through the mathematical model over time. In order to derive methods such as Stochastic Garlerkin, we take moments of this equation with respect to orthogonal polynomials, meaning that we multiply with the basis function φ_i as well as the pdf f_Ξ and integrate over ξ. To simplify the notation, we introduce the bracket operator

$$\langle \cdot \rangle = \int (\cdot) f_\Xi(\xi) d\xi.$$

The resulting system now reads

$$\partial_t \langle u(t,x,\xi)\varphi_i(\xi)\rangle + \partial_x \langle f(u(t,x,\xi))\varphi_i(\xi)\rangle = 0, \tag{2a}$$

$$\langle u(0,x,\xi)\varphi_i(\xi)\rangle = \langle u_0(x,\xi)\varphi_i(\xi)\rangle. \tag{2b}$$

Note that if the basis functions are orthogonal polynomials, the terms $\langle u\varphi_i\rangle$ can be interpreted as Fourier coefficients. If the solution is sufficiently smooth, these coefficients will fall to zero rapidly, meaning that using only the first moments will yield a good approximation. Furthermore, the first moments can be used to determine properties such as expectation values and variances of the solution. This motivates calculating only the first $N+1$ moments

$$u_i(t,x) := \frac{\langle u(t,x,\xi)\varphi_i(\xi)\rangle}{\gamma_i} \quad \text{for} \quad i = 0, \cdots, N,$$

where $\gamma_i := \langle \varphi_i^2 \rangle$. The main problem is to find a good closure $u \approx \mathcal{U}(u_0, \cdots, u_N)$, which allows us to write (2) as a system that only depends on the moments u_i for $i = 0, \cdots, N$. In the case of Stochastic Galerkin, we choose this closure to be

$$u \approx u_{SG} = \sum_{i=0}^{N} u_i \varphi_i.$$

Note that in kinetic theory, this closure is known as the P_N closure, which has widely been applied, see for example [5, 15, 19]. Using the vector notation $\boldsymbol{u} = (u_0, \cdots, u_N)^T$, $\tilde{\boldsymbol{\varphi}} = (\varphi_0/\gamma_0, \cdots, \varphi_N/\gamma_N)^T$ and $\boldsymbol{\varphi} = (\varphi_0, \cdots, \varphi_N)^T$, the resulting Stochastic Galerkin system becomes

$$\partial_t \boldsymbol{u} + \partial_x \langle f(\boldsymbol{u}^T\boldsymbol{\varphi})\tilde{\boldsymbol{\varphi}}\rangle = 0,$$

$$\boldsymbol{u}(0,x) = \langle u_0(x,\xi)\tilde{\boldsymbol{\varphi}}(\xi)\rangle.$$

The low costs to solve this system and the pseudo-spectral convergence of u_{SG} to the correct solution u if this solution is sufficiently smooth make Stochastic Galerkin an attractive method to solve equations containing uncertainties. However, the main drawback in the case of scalar problems is that the solution will be prone to oscillations, leading to non-satisfactory approximation results if the correct solution is non-smooth. In the case of systems, the resulting moment system might no longer be hyperbolic, making it impossible to solve in time and space with standard finite volume methods. In addition to that, the maximal velocities of the system will grow as the solution u_{SG} is not bounded by a certain value. As a result, the choice of the CFL condition must be adapted after each time step to ensure as little numerical diffusion as possible. The IPM method, which has been introduced in [17] tackles these problems.

The idea of the IPM method is to choose a closure u_{ME} such that it minimizes a convex entropy $\langle s(u)\rangle$ under the constraint $\boldsymbol{u} = \langle u\tilde{\boldsymbol{\varphi}}\rangle$. Hence, we can write $u_{ME} = \mathcal{U}(\boldsymbol{u})$ as

$$\mathcal{U}(\boldsymbol{u}) = \arg\min_u \langle s(u)\rangle \quad \text{s.t. } \boldsymbol{u} = \langle u\tilde{\boldsymbol{\varphi}}\rangle.$$

This means, that solving a constraint optimization problem is necessary to evaluate the closure, making this method computationally expensive. To solve the constraint optimization problem, the saddle point of the Lagrangian L has to be determined, i.e.

$$\begin{aligned}
\max_{\boldsymbol{\lambda}} \min_u L(u,\boldsymbol{\lambda}) &= \max_{\boldsymbol{\lambda}} \min_u \left(\langle s(u)\rangle + \boldsymbol{\lambda}^T\left(\boldsymbol{u} - \langle u\tilde{\boldsymbol{\varphi}}\rangle\right)\right) \\
&= \max_{\boldsymbol{\lambda}} \left(\min_u \langle s(u) - u\boldsymbol{\lambda}^T\tilde{\boldsymbol{\varphi}}\rangle + \boldsymbol{\lambda}^T\boldsymbol{u}\right) \\
&= \max_{\boldsymbol{\lambda}} \left(-\langle s^*(\boldsymbol{\lambda}^T\tilde{\boldsymbol{\varphi}})\rangle + \boldsymbol{\lambda}^T\boldsymbol{u}\right) = \min_{\boldsymbol{\lambda}} \left(\langle s^*(\boldsymbol{\lambda}^T\tilde{\boldsymbol{\varphi}})\rangle - \boldsymbol{\lambda}^T\boldsymbol{u}\right).
\end{aligned} \tag{4}$$

Here, the Legendre transformation of the entropy $s^*(\boldsymbol{\lambda}^T\tilde{\boldsymbol{\varphi}})$ is used. The Lagrange multipliers $\boldsymbol{\lambda} \in \mathbb{R}^{N+1}$ will be referred to as dual variables. Defining the dual state $\Lambda := \boldsymbol{\lambda}^T\tilde{\boldsymbol{\varphi}}$, the Legendre transform of the entropy is given by

$$\langle s^*(\Lambda)\rangle := \langle -s(u_{ME}(\Lambda)) + u_{ME}(\Lambda)\Lambda\rangle \tag{5}$$

with $u_{ME}(\Lambda) = (s')^{-1}(\Lambda)$. Calculating the closure by solving a constraint optimization problem has now been reduced to finding the dual variables $\boldsymbol{\lambda}$ by solving the so called dual problem (4). Plugging the resulting dual variables $\boldsymbol{\lambda}$ into $u_{ME}(\Lambda) = (s')^{-1}(\Lambda)$ yields the closure

$$\mathcal{U}(\boldsymbol{u}) = u_{ME}\left(\left(\arg\min_{\boldsymbol{\lambda}}\left(\langle s^*(\boldsymbol{\lambda}^T\tilde{\boldsymbol{\varphi}})\rangle - \boldsymbol{\lambda}^T\boldsymbol{u}\right)\right)^T\tilde{\boldsymbol{\varphi}}\right). \tag{6}$$

Inserting this derivation into (2) leads to the closed moment system

$$\partial_t\boldsymbol{u} + \partial_x\langle f(\mathcal{U}(\boldsymbol{u}))\tilde{\boldsymbol{\varphi}}(\xi)\rangle = 0, \tag{7a}$$

$$\boldsymbol{u}(0,x) = \langle u_0(x,\xi)\tilde{\boldsymbol{\varphi}}(\xi)\rangle. \tag{7b}$$

This system is called IPM-, or when speaking in terms of kinetic theory M_N system. Note that the IPM closure (6) can only be calculated if the moment vector $\boldsymbol{u}$ lies in the so called realizable set

$$\mathcal{R} = \left\{\boldsymbol{u} \in \mathbb{R}^{N+1} \middle| \, \exists u(\xi) \in [u_-, u_+] \text{ s.t. } \boldsymbol{u} = \langle u\tilde{\boldsymbol{\varphi}}\rangle\right\}. \tag{8}$$

As already pointed out, the IPM moment system will dissipate the overall entropy

$$S(t) = \int \langle s(\mathcal{U}(\boldsymbol{u}(t,x)))\rangle dx$$

over time. In [17], the entropy

$$s(u) = -\ln(u - u_-) - \ln(u_+ - u)$$

is used. Assume, that the initial condition has a discontinuity in the random variable, meaning for $t = 0$ and fixed x, the solution will for example be a jump from u_L to $u_R < u_L$. If $s(u)$ is chosen s.t. a value of infinity will be reached at $u_+ := u_L + \Delta u$ and $u_- := u_R - \Delta u$ with a positive constant Δu, the initial condition's entropy $S(0)$ will be finite. Since a solution at a later time t with overshoots that exceed u_+ will have an infinite entropy, these values will be prohibited as $S(t) > S(0)$. The same holds for undershoots. Furthermore, the IPM system is hyperbolic, which would be important if the original problem (1) was a system. For a scalar problem, the Stochastic Galerkin system will also be hyperbolic, due to the fact that it can be interpreted as an IPM system with entropy $s(u) = \frac{1}{2}u^2$.

3 Numerical treatment of the IPM system

The IPM system is continuous in space and time. It can be rewritten as

$$\partial_t\boldsymbol{u} + \partial_x\boldsymbol{F}(\Lambda) = \boldsymbol{0}$$

with the physical system flux $\boldsymbol{F}(\Lambda) = \langle f(u_{ME}(\Lambda))\tilde{\boldsymbol{\varphi}}(\xi)\rangle$ depending on the dual state

$$\Lambda = \left(\arg\min_{\boldsymbol{\lambda}}\left(\langle s^*(\boldsymbol{\lambda}^T\tilde{\boldsymbol{\varphi}})\rangle - \boldsymbol{\lambda}^T\boldsymbol{u}\right)\right)^T\tilde{\boldsymbol{\varphi}}.$$

As this system is hyperbolic, it can be solved by a finite volume method. For this, the time and spatial domain are discretized into cells. The discrete unknowns are now chosen to be the spatial averages over each cell at time t_n, given by

$$u_{ij}^n = \frac{1}{\Delta x}\int_{x_{j-1/2}}^{x_{j+1/2}} u_i(t_n,x)dx.$$

If a moment vector in cell j at time t_n is denoted as $\boldsymbol{u}_j^n = (u_{0j}^n, \cdots, u_{Nj}^n)^T$, the finite volume scheme can then be written in conservative form with the numerical flux $\boldsymbol{G}$, which is consistent with the physical system flux $\boldsymbol{F}$ as

$$\boldsymbol{u}_j^{n+1} = \boldsymbol{u}_j^n - \frac{\Delta t}{\Delta x}\left[\boldsymbol{G}(\Lambda_j^n, \Lambda_{j+1}^n) - \boldsymbol{G}(\Lambda_{j-1}^n, \Lambda_j^n)\right].$$

The discretization of the dual state Λ at spatial cell j and time step n is given by

$$\Lambda_j^n = \left(\arg\min_{\boldsymbol{\lambda}} \left(\langle s^*(\boldsymbol{\lambda}^T\tilde{\boldsymbol{\varphi}})\rangle - \boldsymbol{\lambda}^T\boldsymbol{u}_j^n\right)\right)^T \tilde{\boldsymbol{\varphi}}.$$

Consistency of the numerical flux means that $\boldsymbol{G}(\Lambda, \Lambda) = \boldsymbol{F}(\Lambda)$. To ensure stability, a CFL condition has to be derived by investigating the Eigenvalues of $\nabla \boldsymbol{F}(\Lambda)$. In [17], a Roe scheme is used, which simplifies to an upwinding scheme for the Burgers' test case, meaning that $\boldsymbol{G}(\Lambda_1, \Lambda_2) = \boldsymbol{F}(\Lambda_1)$. For more complicated problems, a Roe matrix $\boldsymbol{A}(\Lambda_1, \Lambda_2)$ needs to be constructed. If ν_k are the Eigenvalues of $\boldsymbol{A}$, the CFL condition is given by $\max_k |\nu_k| \frac{\Delta t}{\Delta x} < CFL$. The standard M_N algorithm in pseudo code is

Algorithm 1 M_N Method for Uncertainty Quantification

```
for $n = 1$ to $NTimeSteps$ do
    for $j = 1$ to $NCells$ do
        $\boldsymbol{u}_j^{n+1} = \boldsymbol{u}_j^n - \frac{\Delta t}{\Delta x}\left[\boldsymbol{G}(\Lambda_j^n, \Lambda_{j+1}^n) - \boldsymbol{G}(\Lambda_{j-1}^n, \Lambda_j^n)\right]$
        $\Lambda_j^{n+1} = \left(\arg\min_{\boldsymbol{\lambda}} \left(\langle s^*(\boldsymbol{\lambda}^T\tilde{\boldsymbol{\varphi}})\rangle - \boldsymbol{\lambda}^T\boldsymbol{u}_j^{n+1}\right)\right)^T \tilde{\boldsymbol{\varphi}}$
    end for
end for
```

It is important to note that this scheme includes several evaluations of integrals, meaning that one needs to choose an appropriate quadrature rule

$$\langle h(\xi)\rangle \approx \sum_{k=1}^{N_q} h(\xi_k) f_\Xi(\xi_k) w_k,$$

where ξ_k are the quadrature points and w_k are the quadrature weights. A remaining task is to choose the solution bounds u_- and u_+ imposed by the entropy. To ensure a finite entropy at $t = 0$, the initial condition is not allowed to be larger or equal to u_+ as well as smaller or equal to u_-. If we define $u_{max} := \max_{x,\xi} u_0(x, \xi)$ and $u_{min} := \min_{x,\xi} u_0(x, \xi)$ we choose the bounds to be $u_+ = u_{max} + \Delta u$ and $u_- = u_{min} - \Delta u$. In [17], the chosen values are $\Delta u = 0.5$. Consequently, the resulting over- and undershoots will be in the order of 0.5. Furthermore, the maximal velocities will be $f'(u_{max} + 0.5)$. This means that the bigger the value of Δu, the smaller the time step Δt in order to fulfill the CFL condition, which increases the number of time steps and adds numerical viscosity. Therefore, choosing Δu closer to zero is desirable. Note that choosing $\Delta u = 0$ will not be allowed, since in this case the entropy $S(0)$ will be infinite and the entropy dissipation no longer prohibits over- and undershoots. Consequently, the IPM solution will not fulfill the maximum principle. When choosing small values of Δu one will observe that the algorithm to solve the dual problem no longer converges after a few time steps. It can be seen that in this case, the moment vector no longer lies in the realizable set $\mathcal{R}$, which was introduced in (8). To understand this behavior, we need to look at the numerical discretization of the problem.

4 Investigation of Realizability

To understand the source of the failure of the optimization method to solve the dual problem if Δu is chosen too small, we need to investigate the numerical discretization in space and time. We start by taking a closer look at the discretization of the IPM moment system, which was

$$\partial_t \boldsymbol{u} + \partial_x \langle f(u_{ME}(\Lambda))\tilde{\boldsymbol{\varphi}}(\xi)\rangle = 0, \tag{9a}$$

$$\Lambda = \left(\arg\min_{\boldsymbol{\lambda}} \left(\langle s^*(\boldsymbol{\lambda}^T \tilde{\boldsymbol{\varphi}})\rangle - \boldsymbol{\lambda}^T \boldsymbol{u}\right) \right)^T \tilde{\boldsymbol{\varphi}}, \tag{9b}$$

$$\boldsymbol{u}(0,x) = \langle u_0(x,\xi)\tilde{\boldsymbol{\varphi}}(\xi)\rangle. \tag{9c}$$

We wish to choose a numerical flux based on the flux of the original problem (1), which will allow finding a simple expression for the CFL condition, an easy implementation of limiters as well as ensure realizability. To construct a consistent numerical flux, we make the choice

$$\boldsymbol{G}(\Lambda_j^n, \Lambda_{j+1}^n) = \langle g(u_{ME}(\Lambda_j^n), u_{ME}(\Lambda_{j+1}^n))\tilde{\boldsymbol{\varphi}}\rangle,$$

where $g(u,v)$ is a consistent numerical flux of the deterministic problem, meaning that $g(u,u) = f(u)$. In terms of kinetic theory, the choice of such a numerical flux is called kinetic flux. The time update of the moment vector now becomes

$$\boldsymbol{u}_j^{n+1} = \boldsymbol{u}_j^n - \frac{\Delta t}{\Delta x}\left[\langle g(u_{ME}(\Lambda_j^n), u_{ME}(\Lambda_{j+1}^n))\tilde{\boldsymbol{\varphi}}\rangle - \langle g(u_{ME}(\Lambda_{j-1}^n), u_{ME}(\Lambda_j^n))\tilde{\boldsymbol{\varphi}}\rangle\right].$$

The time update for the dual state is then given by

$$\Lambda_j^{n+1} = \left(\arg\min_{\boldsymbol{\lambda}} \left(\langle s^*(\boldsymbol{\lambda}^T \tilde{\boldsymbol{\varphi}})\rangle - \boldsymbol{\lambda}^T \boldsymbol{u}_j^{n+1}\right) \right)^T \tilde{\boldsymbol{\varphi}}.$$

A standard choice for updating the dual states is the use of Newton's method making use of the stopping criterion

$$\left\| \boldsymbol{u}_j^{n+1} - \langle u_{ME}(\boldsymbol{\lambda}^T \tilde{\boldsymbol{\varphi}})\tilde{\boldsymbol{\varphi}}\rangle \right\| < \tau.$$

Note that obviously the dual state will not fulfill the moment constraint exactly. This is denoted by writing the inexact dual state as $\bar{\Lambda}$ hence

$$\boldsymbol{u}_j^n = \left\langle u_{ME}(\bar{\Lambda}_j^n)\tilde{\boldsymbol{\varphi}}\right\rangle + \tilde{\boldsymbol{\tau}}$$

whereas the exact dual state fulfills

$$\boldsymbol{u}_j^n = \left\langle u_{ME}(\Lambda_j^n)\tilde{\boldsymbol{\varphi}}\right\rangle. \tag{10}$$

The difference between the exact dual state Λ and the inexact dual state $\bar{\Lambda}$ is denoted by $\Delta\Lambda$ such that $\Lambda = \bar{\Lambda} + \Delta\Lambda$. Using this notation as well as (10) for the numerical moment update yields

$$\boldsymbol{u}_j^{n+1} = \left\langle u_{ME}(\Lambda_j^n)\tilde{\boldsymbol{\varphi}}\right\rangle - \frac{\Delta t}{\Delta x}\left[\langle g(u_{ME}(\bar{\Lambda}_j^n), u_{ME}(\bar{\Lambda}_{j+1}^n))\tilde{\boldsymbol{\varphi}}\rangle - \langle g(u_{ME}(\bar{\Lambda}_{j-1}^n), u_{ME}(\bar{\Lambda}_j^n))\tilde{\boldsymbol{\varphi}}\rangle\right].$$

This can be rewritten as $\boldsymbol{u}_j^{n+1} = \left\langle H(\bar{\Lambda}_{j-1}^n, \bar{\Lambda}_j^n, \bar{\Lambda}_{j+1}^n)\tilde{\boldsymbol{\varphi}}\right\rangle$ with

$$\begin{aligned} H(\bar{\Lambda}_{j-1}^n, \bar{\Lambda}_j^n, \bar{\Lambda}_{j+1}^n) =& u_{ME}(\bar{\Lambda}_j^n + \Delta\Lambda_j^n) \\ & - \frac{\Delta t}{\Delta x}\left[g(u_{ME}(\bar{\Lambda}_j^n), u_{ME}(\bar{\Lambda}_{j+1}^n)) - g(u_{ME}(\bar{\Lambda}_{j-1}^n), u_{ME}(\bar{\Lambda}_j^n))\right]. \end{aligned} \tag{11}$$

This means that the time updated moments are the moments of the function H, which depends on the inexact dual states of the previous time step and is continuous in ξ. Obviously, the time updated moment vector will lie in the realizable set (8) if $H(\bar{\Lambda}_{j-1}^n, \bar{\Lambda}_j^n, \bar{\Lambda}_{j+1}^n) \in [u_-, u_+]$. We will use this condition to construct a numerical scheme which remains realizable.

Note that the approach of investigating the underlying function from which one calculates the time updated moments has been investigated in [2] in the case of a linear transport equation for ensuring positivity. Here, the positivity of the underlying function is ensured by the choice of a weaker CFL condition. Note that the chosen CFL condition will only ensure positivity, while allowing the maximal value of the solution to grow. In our case, we need to show that H will be bounded by u_- and u_+ for a general scalar hyperbolic equation, which is why we need to choose a different strategy. It can be shown that the term $\Delta\Lambda_j^n$ can violate the bounds imposed by the entropy. In order to show this property for a non-linear problem with an arbitrary monotone flux g, we use standard techniques from Finite Volume methods. A standard result from Finite Volume schemes applied to the M_N context is that monotonicity will ensure a bounded solution:

Theorem 4.1 *If a scheme of the form* (11) *is monotone, meaning that it fulfills*

$$\frac{\partial H(\bar{\Lambda}_{j-1}^n, \bar{\Lambda}_j^n, \bar{\Lambda}_{j+1}^n)}{\partial \bar{\Lambda}_{j-1,j,j+1}^n} \geq 0, \tag{12}$$

the solution will be bounded from below by u_- and from above by u_+ for all $\xi \in [-1,1]$ in the next time step if the entropy fulfills

$$\lim_{u \to u_+} s'(u) \to \infty, \quad \lim_{u \to u_-} s'(u) \to -\infty.$$

Proof Since $s'(u) : [u_-, u_+] \to \mathbb{R}$ is bijective, we have that $(s')^{-1}(x) \in [u_-, u_+]$ for all $x \in \mathbb{R}$. We choose a fixed $\xi^* \in [-1,1]$ and assume that (12) holds, hence our scheme is monotone. Now, we look at a sequence $v = (\bar{\Lambda}_i^n)_{i\in\mathbb{Z}}$ and the constant sequence $w = (w_i)_{i\in\mathbb{Z}}$ with $w_i = \max_k v_k$ for all $i, k \in \mathbb{Z}$. Let us define $H_\Delta(v) := (H(v_{i-1}, v_i, v_{i+1}))_{i\in\mathbb{Z}}$. We now have

$$(H_\Delta(v))_j \overset{\text{H monotone}}{\leq} (H_\Delta(w))_j = u_{ME}(w_j + \Delta\Lambda_j^n) = (s')^{-1}(w_j + \Delta\Lambda_j^n) \leq u_+.$$

This essentially tells us that the numerical time update of the underlying function $H(\Lambda_{j-1}^n, \Lambda_j^n, \Lambda_{j+1}^n)$ fulfills $H \leq u_+$. The other side $H \geq u^-$ can be shown analogously with the sequence $w = (w_i)_{i\in\mathbb{Z}}$ with $w_i = \min_k v_k$. □

It is important to point out that this allows the entropy to be finite at the bounds, which will be used later to impose the maximum principle onto the solution. This seems to be striking as the entropy dissipation will no longer ensure boundedness, assuming that we can construct a convex entropy $\tilde{s}(u) \in C^2(\mathbb{R})$ such that

$$\tilde{s}(u) = \begin{cases} s_-(u) & \text{if } u < u_-, \\ s(u) & \text{if } u \in [u_-, u_+], \\ s_+(u) & \text{if } u > u_+, \end{cases}$$

and $s_{-,+}$ remains finite in a certain interval. However note that if $s'(u)$ is infinite at the bounds, it cannot be continued in a differentiable way at u_- and u_+, meaning that $s_{-,+}$ do not exist.
Also note that opposed to classical finite volume schemes, monotonicity will not ensure a min-max condition, meaning that the solution will not be bounded by the minimal and maximal value of the initial condition. Therefore, if $u_L < u_+$, the CFL condition of the deterministic problem

$$\frac{\Delta t}{\Delta x} f'(u_L) = 1$$

will lead to solutions violating the imposed bounds. This stems from the fact that the scheme (11) is not in classical conservation form.
Also note that ensuring monotonicity in the dual state is crucial for boundedness. A scheme which is only monotone in the solution values will violate the bounds u_- and u_+. To illustrate this, we will look at the following setting:

Example Assume that we wish to solve the uncertain linear advection equation

$$\partial_t u + a(\xi)\partial_x u = 0,$$
$$u(t=0,x,\xi) = u_0(x)$$

where ξ is uniformly distributed between -1 and 1 and $a(\xi) = 0.5(1+\xi)$. For the underlying kinetic scheme, we use an upwind flux $g(u,v) = a(\xi)u$. The underlying scheme is now given by

$$u_j^{n+1} = u_{ME}(\Lambda_j^n) - \frac{a(\xi)\Delta t}{\Delta x}\left(u_{ME}(\bar{\Lambda}_j^n) - u_{ME}(\bar{\Lambda}_{j-1}^n)\right),$$

where we choose u_{ME} s.t. the solution is bounded by $u_+ > u_- \geq 0$. Defining $u_j^n := u_{ME}(\Lambda_j^n)$ and $\bar{u}_j^n := u_{ME}(\bar{\Lambda}_j^n)$, as well as $\eta_j^n := \bar{u}_j^n / u_j^n$, we can rewrite this scheme in difference form as

$$u_j^{n+1} = \left(1 - \eta_j^n \frac{a(\xi)\Delta t}{\Delta x}\right) u_j^n + \frac{a(\xi)\Delta t}{\Delta x}\eta_{j-1}^n u_{j-1}^n.$$

Choosing the CFL condition

$$1 - \eta\frac{a(\xi)\Delta t}{\Delta x} > 0$$

with $\eta = \max_j \eta_j^n$ will ensure that the difference scheme has the form

$$u_j^{n+1} = c_j u_j^n + c_{j-1}u_{j-1}^n$$

with positive coefficients c_j and c_{j-1}. Obviously $u_j^{n+1} > 0$. However, an example, where this scheme will violate the upper bound u_+ can easily be constructed. Assume that $u_j^n = \bar{u}_{j-1}^n = u_{j-1}^n = 0.99 \cdot u_+$ and $\bar{u}_j^n = 0.9 \cdot u_j^n$. In this case

$$u_j^{n+1} = 0.99 \cdot u_+ \left(1 - 0.9\frac{a(\xi)\Delta t}{\Delta x} + \frac{a(\xi)\Delta t}{\Delta x}\right) = 0.99 \cdot u_+ \left(1 + 0.1\frac{a(\xi)\Delta t}{\Delta x}\right)$$

Note that since $\eta = 1$, we have that $\Delta t = \frac{\Delta x}{a(\xi)}$. In this case

$$1 + 0.1\frac{a(\xi)\Delta t}{\Delta x} = 1.1 > \frac{1}{0.99}$$

hence $u_j^{n+1} > u_+$.

According to 4.1, we need monotonicity w.r.t. the dual state in order to obtain boundedness. However, this monotonicity will not be assured by the scheme (11) under a desirable non-restricting CFL condition:

Remark Assume that g is monotone and consistent with the physical flux $f(u)$. If the CFL condition

$$\frac{\Delta t}{\Delta x}f'(u^+) = 1 \tag{13}$$

is chosen, the time update of the underlying function of the updated inexact moments, which is given by (11) will not necessarily be bounded by u^- and u^+.

Proof We again investigate the scheme

$$H(\bar{\Lambda}_{j-1}^n, \bar{\Lambda}_j^n, \bar{\Lambda}_{j+1}^n) = u_{ME}(\Lambda_j^n) - \frac{\Delta t}{\Delta x}\left[g(u_{ME}(\bar{\Lambda}_j^n), u_{ME}(\bar{\Lambda}_{j+1}^n)) - g(u_{ME}(\bar{\Lambda}_{j-1}^n), u_{ME}(\bar{\Lambda}_j^n))\right],$$

where the exact dual state is given by $\Lambda_j^n = \bar{\Lambda}_j^n + \Delta\Lambda_j^n$ and we choose a fixed $\xi^* \in [-1, 1]$. Note that since the numerical flux g belongs to a monotone scheme, we have that the scheme

$$G(u, v, w) = v - \frac{\Delta t}{\Delta x}\left[g(v, w) - g(u, v)\right]$$

is monotonically increasing in each argument. If we write $g = g(u_1, u_2)$, meaning that we name the first input u_1 and the second input u_2 and define $g_{j-1/2} := g(u, v)$ as well as $g_{j+1/2} := g(v, w)$, this implies

$$\begin{aligned} \frac{\partial}{\partial u_1} g_{j-1/2} &\geq 0, \\ \frac{\partial}{\partial u_2} g_{j+1/2} &\leq 0, \\ 1 - \frac{\Delta t}{\Delta x}\left(\frac{\partial}{\partial u_1} g_{j+1/2} - \frac{\partial}{\partial u_2} g_{j-1/2}\right) &\geq 0. \end{aligned}$$

From this, as well as our entropy ansatz, we can show that our original scheme H is monotone in the first and third input, since

$$\frac{\partial H}{\partial \bar{\Lambda}_{j-1}^n} = \frac{\Delta t}{\Delta x}\frac{\partial g_{j-1/2}}{\partial u_1} u'_{ME}(\bar{\Lambda}_{j-1}^n) = \frac{\Delta t}{\Delta x} \underbrace{\frac{\partial}{\partial u_1} g_{j-1/2}}_{\geq 0,\, G \text{ monotone}} \underbrace{\frac{1}{s''(u_{ME}(\bar{\Lambda}_{j-1}^n))}}_{\geq 0, \text{ s convex}} \geq 0,$$

$$\frac{\partial H}{\partial \bar{\Lambda}_{j+1}^n} = -\frac{\Delta t}{\Delta x} \underbrace{\frac{\partial}{\partial u_2} g_{j+1/2}}_{\leq 0,\, G \text{ monotone}} \underbrace{\frac{1}{s''(u_{ME}(\bar{\Lambda}_{j+1}^n))}}_{\geq 0, \text{ s convex}} \geq 0.$$

The error when solving the dual problem will destroy monotonicity for the second input:

$$\frac{\partial H}{\partial \bar{\Lambda}_j^n} = u'_{ME}(\bar{\Lambda}_j^n + \Delta\Lambda_j^n) - \frac{\Delta t}{\Delta x}\left(\frac{\partial}{\partial u_1} g_{j+1/2} u'_{ME}(\bar{\Lambda}_j^n) - \frac{\partial}{\partial u_2} g_{j-1/2} u'_{ME}(\bar{\Lambda}_j^n)\right). \tag{14}$$

We now add $0 = u'_{ME}(\bar{\Lambda}_j^n) - u'_{ME}(\bar{\Lambda}_j^n)$ to arrive at

$$\frac{\partial H}{\partial \bar{\Lambda}_j^n} = \frac{1}{s''(u_{ME}(\Lambda_j^n))} - \frac{1}{s''(u_{ME}(\bar{\Lambda}_j^n))} + \underbrace{\frac{1}{s''(u_{ME}(\bar{\Lambda}_j^n))}}_{\geq 0, \text{ s convex}} \underbrace{\left(1 - \frac{\Delta t}{\Delta x}\left(\frac{\partial}{\partial u_1} g_{j+1/2} - \frac{\partial}{\partial u_2} g_{j-1/2}\right)\right)}_{\geq 0,\, G \text{ monotone}}. \tag{15}$$

Note that if the last term is zero, we are left with

$$\frac{\partial H}{\partial \bar{\Lambda}_j^n} = \frac{1}{s''(u_{ME}(\Lambda_j^n))} - \frac{1}{s''(u_{ME}(\bar{\Lambda}_j^n))}.$$

This term can easily become negative, namely if the exact dual states are bigger than the inexact dual states, which is for example the case if the correct underlying function has a value of u_- or u_+. Therefore, with theorem 4, the min-max property is violated. □

We now need to think of ways to resolve this problem by for example imposing a weaker CFL condition. Looking at equation (14), one can derive such a CFL condition by

$$\frac{\partial H}{\partial \bar{\Lambda}_j^n} = u'_{ME}(\bar{\Lambda}_j^n + \Delta\Lambda_j^n) - \frac{\Delta t}{\Delta x}\left(\frac{\partial}{\partial u_1} g_{j+1/2} u'_{ME}(\bar{\Lambda}_j^n) - \frac{\partial}{\partial u_2} g_{j-1/2} u'_{ME}(\bar{\Lambda}_j^n)\right)\frac{u'_{ME}(\Lambda_j^n)}{u'_{ME}(\Lambda_j^n)}.$$

Defining

$$\gamma := \frac{u'_{ME}(\bar{\Lambda}_j^n)}{u'_{ME}(\bar{\Lambda}_j^n + \Delta\Lambda_j^n)},$$

we arrive at

$$\frac{\partial H}{\partial \bar{\Lambda}_j^n} = u'_{ME}(\Lambda_j^n)\left(1 - \gamma\frac{\Delta t}{\Delta x}\left(\frac{\partial}{\partial u_1}g_{j+1/2} - \frac{\partial}{\partial u_2}g_{j-1/2}\right)\right). \tag{16}$$

Monotonicity can thus be achieved by choosing the CFL condition

$$\gamma\frac{\Delta t}{\Delta x}f'(u_+) = 1.$$

However, a weaker CFL condition will smear out the solution forcing us to use a finer grid resolution, leading to higher numerical costs. The goal of using the sharp CFL condition (13) forces us to come up with a different strategy. This results in more spacial cells as well as a larger number of time steps, meaning that more expensive dual problems need to be solved. Let us discuss another ansatz: We keep the CFL condition (13) and make sure that

$$\frac{1}{s''(u_{ME}(\Lambda_j^n))} - \frac{1}{s''(u_{ME}(\bar{\Lambda}_j^n))} \overset{!}{=} 0. \tag{17}$$

This means that our scheme keeps its spatial accuracy, but will stay monotone, as (15) becomes

$$\frac{\partial H}{\partial \bar{\Lambda}_j^n} = \underbrace{\frac{1}{s''(u_{ME}(\bar{\Lambda}_j^n))}}_{\geq 0,\text{ s convex}}\underbrace{\left(1 - \frac{\Delta t}{\Delta x}\left(\frac{\partial}{\partial u_1}g_{j+1/2} - \frac{\partial}{\partial u_2}g_{j-1/2}\right)\right)}_{\geq 0,\ G\text{ monotone}} \geq 0. \tag{18}$$

We could achieve this goal in two ways: Either, we can solve the dual problem exactly, or we can modify the scheme by setting Λ_j^n to $\bar{\Lambda}_j^n$. Finding the exact solution of the dual problem is not possible, which is why one needs to make sure that the exact dual states are modified. Let us look back at equation (11), which was

$$\begin{aligned} H(\bar{\Lambda}_{j-1}^n, \bar{\Lambda}_j^n, \bar{\Lambda}_{j+1}^n) =& \langle u_{ME}(\bar{\Lambda}_j^n + \Delta\Lambda_j^n)\tilde{\varphi}_i\rangle \\ &- \frac{\Delta t}{\Delta x}\left[\langle g(u_{ME}(\bar{\Lambda}_j^n), u_{ME}(\bar{\Lambda}_{j+1}^n))\tilde{\varphi}_i\rangle - \langle g(u_{ME}(\bar{\Lambda}_{j-1}^n), u_{ME}(\bar{\Lambda}_j^n))\tilde{\varphi}_i\rangle\right]. \end{aligned}$$

Setting the exact dual states onto the inexact dual states leads to the modified scheme

$$\begin{aligned} \langle u_{ME}(\Lambda_j^{n+1})\tilde{\varphi}_i\rangle =& \langle u_{ME}(\bar{\Lambda}_j^n)\tilde{\varphi}_i\rangle \\ &- \frac{\Delta t}{\Delta x}\left[\langle g(u_{ME}(\bar{\Lambda}_j^n), u_{ME}(\bar{\Lambda}_{j+1}^n))\tilde{\varphi}_i\rangle - \langle g(u_{ME}(\bar{\Lambda}_{j-1}^n), u_{ME}(\bar{\Lambda}_j^n))\tilde{\varphi}_i\rangle\right]. \end{aligned}$$

Now one needs to discuss how to solve this scheme. Obviously, one only needs to update the moments with the inexact dual states meaning that we use $\bar{u}_{ij} = \langle u_{ME}(\bar{\Lambda}_j^n)\tilde{\varphi}_i\rangle$ for the time update instead of $u_{ij} = \langle u_{ME}(\Lambda_j^n)\tilde{\varphi}_i\rangle$. This gives us the slightly modified algorithm

Algorithm 2 IPM Method

```
for $n = 1$ to $NTimeSteps$ do
    for $j = 1$ to $NCells$ do
        $\bar{\boldsymbol{u}}_j^n = \langle u_{ME}(\tilde{\boldsymbol{\varphi}}^T\boldsymbol{\lambda}_j^n)\tilde{\boldsymbol{\varphi}}\rangle$
        $\boldsymbol{u}_j^{n+1} = \bar{\boldsymbol{u}}_j^n - \frac{\Delta t}{\Delta x}\left[\boldsymbol{G}(u_{ME}(\tilde{\boldsymbol{\varphi}}^T\boldsymbol{\lambda}_j^n), u_{ME}(\tilde{\boldsymbol{\varphi}}^T\boldsymbol{\lambda}_{j+1}^n)) - \boldsymbol{G}(u_{ME}(\tilde{\boldsymbol{\varphi}}^T\boldsymbol{\lambda}_{j-1}^n), u_{ME}(\tilde{\boldsymbol{\varphi}}^T\boldsymbol{\lambda}_j^n))\right]$
        $\boldsymbol{\lambda}_j^{n+1} = \arg\min_{\boldsymbol{\lambda}}\left(\langle s^*(\boldsymbol{\lambda}^T\tilde{\boldsymbol{\varphi}})\rangle - \boldsymbol{\lambda}^T\boldsymbol{u}_j^{n+1}\right)$
    end for
end for
```

Let us conclude, that the choice of the CFL condition $f'(u_+)\Delta t/\Delta x = 1$ as well as imposing the error of the dual problem onto the exact moments and our choice of the numerical flux will lead to a monotone scheme, which ensures realizability. This means that the dual problem can be solved for the moments of the next time step. To ensure boundedness, the monotonicity with respect to the dual state is crucial. A next step to improve this scheme could be the use of spatial limiters to increase the accuracy. This task can be easily done by choosing bound preserving limiters.
Note that after having resolved the issue of realizablity, choosing a small value of Δu is possible. As overshoots at u_L and undershoots at u_R will be in the area of Δu, we expect a non-oscillatory solution. However, it can be seen that for small values of Δu, significant oscillations will arise at intermediate values $u \in (u_-, u_+)$. In the following, we investigate this behavior and propose a new entropy, which is based on a less restricting condition. This less restrictive condition allows the choice $\Delta u = 0$, meaning that the resulting solution fulfills the maximum principle.

5 The Entropy Choice

In [17], the chosen entropy is

$$s(u) = -\ln(u - u_-) - \ln(u_+ - u). \tag{19}$$

The dissipation of

$$S(t) = \int \langle s(u(t, x, \xi))\rangle dx$$

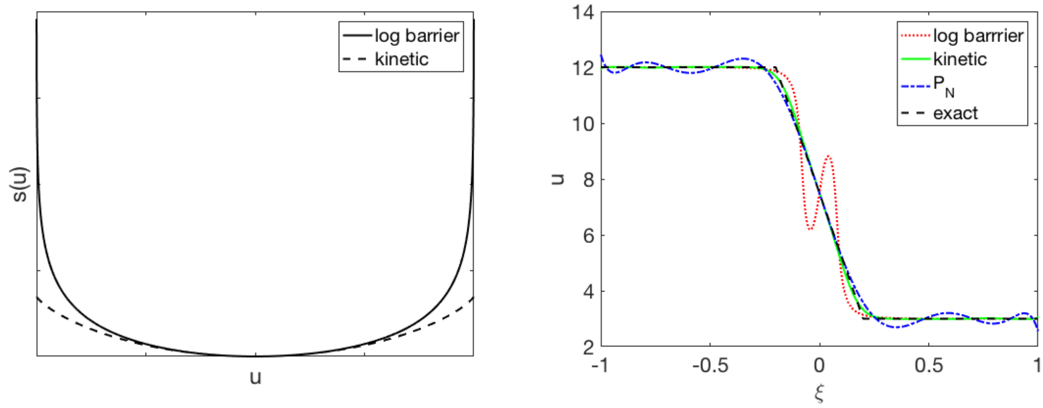
over time is used to limit oscillations, since over- and undershoots that reach values of u_- and u_+ lead to an infinite entropy. If the initial entropy $S(0)$ is finite, these over- and undershoots will therefore be prohibited, leading to a limitation of oscillations. Note that u_+ can not be chosen as the maximal value of the initial condition u_{max}, since in this case, the initial entropy is infinite. As a result, the IPM solution will have oscillations in the order of $\Delta u = u_+ - u_{max}$. To allow only small oscillations one can choose u_+ very close to the maximal value of the initial condition. We will show, that this will lead to an unreasonable manipulation of the solution which is not close to the maximal values. The same holds for the lower bound u_-.
To further investigate the approximation properties of the different entropy candidates, especially for entropy (19), it is useful to interpret the constraint entropy minimization in the following way: Solving the minimization under the moment constraint is equivalent to rewriting the closure as

$$\mathcal{U}(\boldsymbol{u}) = \boldsymbol{u}^T \boldsymbol{\varphi} + \sum_{i=N+1}^{\infty} \bar{u}_i \varphi_i, \tag{20}$$

where the additional Fourier coefficients $\bar{u}_i$ are determined by minimizing the entropy

$$\left\langle s\left(\boldsymbol{u}^T \boldsymbol{\varphi} + \sum_{i=N+1}^{\infty} \bar{u}_i \varphi_i\right)\right\rangle. \tag{21}$$

Hence, from this point of view the entropy determines the difference between the P_N closure $\boldsymbol{u}^T\boldsymbol{\varphi}$ and the M_N closure (20). It must be noted that the approximation of the exact solution is done by the first $N+1$ moments. The preceding moments $\bar{u}_i$ are chosen by the entropy, meaning that they will only contribute to the approximation if the entropy imposes properties of the exact solution.
In our case, we wish to obtain a solution, which is bounded by u_- and u_+. Note that the log barrier entropy will certainly ensure this property, as it will have infinite values at these bounds. However, the log barrier-entropy will add more characteristics to the solution: As an example, let us look at two solution values $u_M = 0.5(u_- + u_+)$ and $u_C = 0.99u_+$. Looking at the log barrier



(a)

(b)

Figure 1: (a) Comparison of entropies and (b) resulting approximation properties for a forming shock with $\Delta u = 0.01$ and $N = 10$.

entropy in Figure 1a, it gets clear that u_C leads to a big entropy and the entropy minimization will therefore focus on decreasing the solution values at u_C by choosing the coefficients $\bar{u}_i$ in (20). Certainly, these coefficients will modify the solution at u_M, which is allowed due to the fact that the entropy will only be slightly increased in this region. As a result, the log barrier entropy will heavily modify the solution at values close to u_M in order to slightly decrease the solution in regions close to u_C. Therefore, the entropy (19) will lead to bad approximation results as in Figure 1b. Let us now discuss how one should modify the entropy in order to obtain better approximation results in non-critical regions. It is import to state two remarks: The entropy should ensure that $u \in [u_-, u_+]$. A point where the solution is slightly smaller than u_+ should not be punished, or -in the context of entropy minimization- a decrease of its value should not result in a heavy decease of the entropy. In addition to this, a non-critical point should be as important as a critical point. Hence the entropy should ensure that modifying a critical and a non-critical point should result in a comparable decrease or increase of the entropy. As a second remark, setting the bounds onto the maximal and minimal value of the solution, meaning that $\Delta u = 0$ will lead to a non-oscillatory solution, which is certainly desirable. However, this can only be achieved by an entropy, which stays finite at u_+ and u_-. In this case, the entropy dissipation will no longer ensure that the solution will stay inside the prescribed bounds. However, it has been shown in theorem 4.1, that the entropy itself does not need to go to infinity as the IPM scheme fulfills a min-max property. This goal can already be achieved by choosing an entropy with a slope of plus or minus infinity, leading to a parametrization $u_{ME}(\Lambda) = (s')^{-1}(\Lambda) \in [u_-, u_+]$. The chosen entropy has the form

$$s(u) = -(u - u_-)\ln(u - u_-) - (u_+ - u)\ln(u_+ - u). \tag{22}$$

As it is similar to the entropy $\eta(f) = f \ln f - f$ often used in kinetic theory, we will call this entropy 'kinetic entropy'. A comparison to the original entropy, which we will call 'log barrier' entropy can be found in Figure 1a. Looking at this Figure we conclude: It can be seen that the kinetic entropy will not try to push the solution away from these bounds, as this will only slightly decrease the entropy. As soon as such a modification influences a point in a non-critical region, the entropy will again increase, meaning that the entropy (21) will not be minimized, hence such a modification is prohibited. Opposed to this behavior, the log barrier entropy will focus on pushing the solution away from the bounds as this will heavily minimize the entropy. The effect on the solution in non-critical is not included by the minimization, as this will only slightly influence the entropy, leading to unreasonable modifications of these regions. Let us sum up our findings: We have seen that the entropy should fulfill

$$\lim_{u\to u_+} s'(u) \to \infty, \quad \lim_{u\to u_-} s'(u) \to -\infty.$$

in order to ensure the min-max property of the solution if $\Delta u = 0$. Furthermore, the absolute value of the entropy's slope should be close to constant for $u \in (u_-, u_+)$ to prevent oscillations in non-critical regions. In order to give a better understanding of how the slope of the entropy in the interval (u_-, u_+) affects the reconstruction, we investigate the reconstruction of two shocks connecting the states u_L, u_M and u_R by a family of entropies $s_k(u)$. The starting point of constructing such a sequence will be the kinetic entropy (22). The bounds imposed by the kinetic entropy are chosen to be close to the left state u_L and the right state u_R, namely $u_+ = u_R + 0.1$ and $u_- = u_L - 0.1$. The family of entropies is now defined to be

$$s_k(u) = \left(\frac{s(u) - s\left(\frac{1}{2}(u_- + u_+)\right)}{s(u_L) - s\left(\frac{1}{2}(u_- + u_+)\right)} \right)^k.$$

The main modification compared to the kinetic entropy (22) is that it is exponentiated by k. Furthermore, we impose that $s_k(0.5(u_- + u_+)) = 0$ and $s_k(u_L) = s_k(u_R) = 1$ by a shift and a scaling term. For values of k from 1.0 to 3.0, the entropies are depicted in Figure 2a. It can be seen that as k grows, the entropy's slope at u_L and u_R will increase, whereas the entropy will develop a nearly constant region at the intermediate state

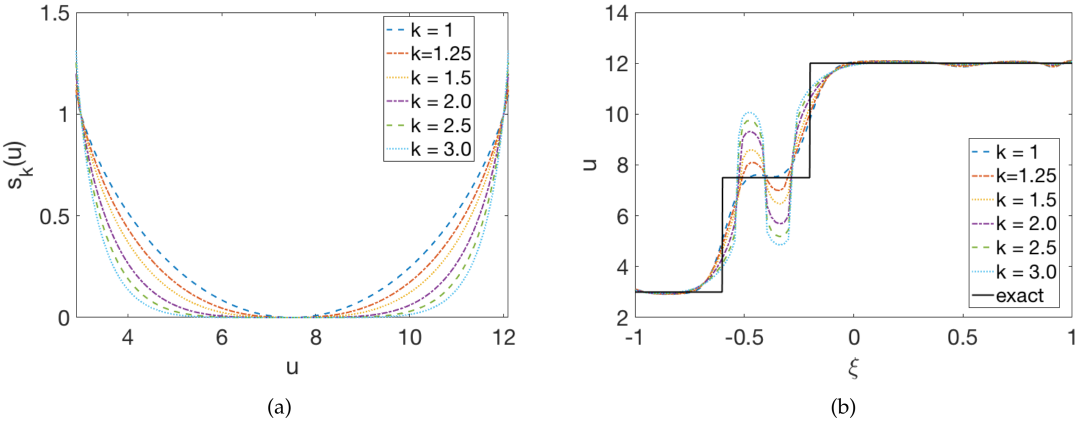


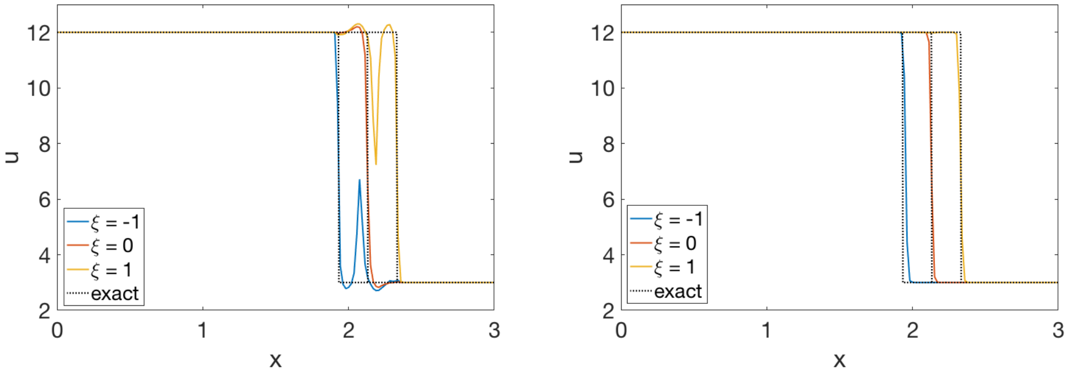
Figure 2: (a) Family of entropies and (b) corresponding reconstruction.

Let us assume, that we have an entropy which heavily decreases if the solution u is pushed away from the bounds u_- and u_+. Furthermore, this entropy is not affected at solutions away from these bounds. As the reconstruction minimizes this entropy under the moment constraint, all degrees of freedom will heavily try to push the solution away from u_- and u_+. Resulting modifications of intermediate values are not measured by the entropy and are therefore likely to occur. Looking at Figure 2b, one sees how the reconstructions done by s_k change for increasing k.

6 Numerical Results

6.1 Uncertain Burger's equation

In order to show the analytically derived behavior, we will look at two different hyperbolic equations with uncertain inputs. The first one is the uncertain Burger's equation, which has been



(a) Log Barrier Entropy.

(b) Kinetic entropy.

Figure 3: Solution for different entropies evaluated at $\xi \in \{-1, 0, 1\}$.

studied in [17]. It is given by

$$\partial_t u(t, x, \xi) + \partial_x \frac{u(t, x, \xi)^2}{2} = 0,$$
$$u(t = 0, x, \xi) = u_0(x, \xi).$$

This equation is interesting, as it will lead to shocks in the spatial and the random domain even for smooth initial conditions. As in [17], we choose the random initial condition

$$u_0(x, \xi) = \begin{cases} u_L & \text{for } x < x_0 \\ u_L + \frac{u_R - u_L}{x_0 - x_1}(x_0 + \sigma\xi - x) & \text{for } x \in [x_0, x_1] \\ u_R & \text{else} \end{cases},$$

which is a forming shock with a linear connection from x_0 to x_1. In our case, ξ is uniformly distributed on the interval $[-1, 1]$. Note that this initial condition cannot be exactly described by a uniformly random variable, meaning that already in the beginning, there will be an approximation error. When solving this problem with the standard M_N algorithm 1, the dual problem will not be solvable after a few time steps when using a small value for Δu. This behavior is expected, as the inexact solution of the dual problem will yield a non-monotone scheme which will not be bounded by u_- and u_+. Using a weaker CFL condition $f'(u_+)\Delta t/\Delta x = 0.8$ will resolve this problem, but will add unwanted diffusion. We now turn to the comparison of the different entropies. In Figure 6, the solution $u(T, x, \xi)$ for $\xi \in \{-1, 0, 1\}$ is plotted for both entropies. Since the log barrier entropy will be infinite for u_+ and u_-, we need to choose $\Delta u > 0$. As expected, a small choice of Δu will lead to a bad approximation, which already shows in the initial condition. In the following, we use $\Delta u = 0.5$ as in [17]. Note that the maximal velocity of the equation can be $u_- = u_L + 0.5$, meaning that the CFL condition of the deterministic problem, where velocities are bounded by u_L cannot be used. The kinetic entropy also shows good approximation results for small values of Δu, which is why we set this value to zero, allowing the use of the deterministic CFL condition. It can be seen that the log barrier entropy will have over- and undershoots, whereas the kinetic entropy nicely approximates the solution. Looking at the dependency on ξ for a fixed spatial cell in Figure 4, it can be seen that the log barrier entropy will have oscillations whereas the kinetic entropy will lead to a non-oscillatory solution. Furthermore, the solution obtained with the help of the kinetic entropy is L^∞ stable, due to the choice of $\Delta u = 0$.
Let us now turn to a new initial condition for the uncertain Burger's equation in order to investi-

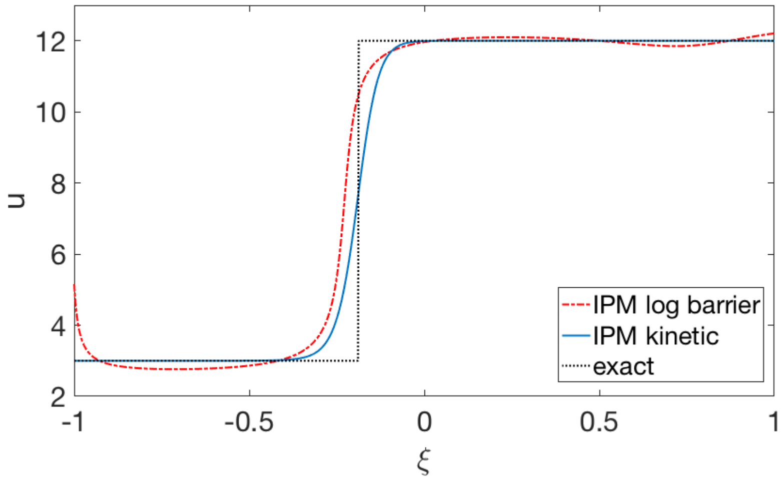


Figure 4: Solutions for log barrier and kinetic entropy at fixed spatial position x.

gate the oscillations arising at a non-critical state u_M:

$$u(t=0,x,\xi) = \begin{cases} u_L, & \text{if } x \leq x_0 + \sigma\xi \\ u_L + (u_M - u_L) \cdot \frac{x_0+\sigma\xi-x}{x_0-x_1}, & \text{if } x \in (x_0+\sigma\xi, x_1+\sigma\xi] \\ u_M, & \text{if } x \in (x_1+\sigma\xi, x_2+\sigma\xi] \\ u_M + (u_R - u_M) \cdot \frac{x_3+\sigma\xi-x}{x_3-x_2}, & \text{if } x \in (x_2+\sigma\xi, x_3+\sigma\xi] \\ u_R, & \text{if } x > x_3+\sigma\xi. \end{cases}$$

This initial condition describes two forming shocks that connect the three states u_L, u_M and u_R. At a later time t^*, two shock will form, similar to the solution in Figure 2b. As the slope of the kinetic entropy is of comparable magnitude for $u \in (u_-, u_+)$ compared to the double log entropy, we will expect to see little oscillations at the intermediate state u_M when making use of the kinetic entropy. As before, the states are chosen to be $u_L = 12.0$, $u_R = 3.0$ and $u_M = 0.5(u_L + u_R)$. We choose the positions of the forming shock such that the shocks will show for $t^* = 0.04$. For this time we investigate the forming shock at a fixed position $x = 1.5$. Since the solution has a complicated structure, compared to the shock solution from before, we will make the choice $N = 15$, meaning that 26 moments will be used to describe the solution. Again the choice for Δu will be zero for the kinetic entropy as well as 0.5 for the log barrier entropy.
The results for this problem can be seen in Figure 5. It can be seen that, as expected the solution of the log barrier entropy will be oscillatory, whereas the kinetic entropy shows only small oscillations. While the kinetic entropy is L^∞ stable, the log barrier entropy will show over- and undershoots in the order of Δu.

6.2 Uncertain advection equation

In a next step, we look at the uncertain advection equation, which is similar to kinetic equations like the Boltzmann equation without collision terms. The advection equation is given by

$$\begin{aligned} \partial_t u(t,x,\xi) + a(\xi)\partial_x u(t,x,\xi) = 0, \\ u(t=0,x) = u_0(x). \end{aligned}$$

The random variable ξ now enters from the time evolution of the deterministic initial condition. We choose $a(\xi)$ to be uniformly distributed in the interval $[10, 12]$. What is interesting about this equation is that the velocity of the system is not known, which means that our CFL condition

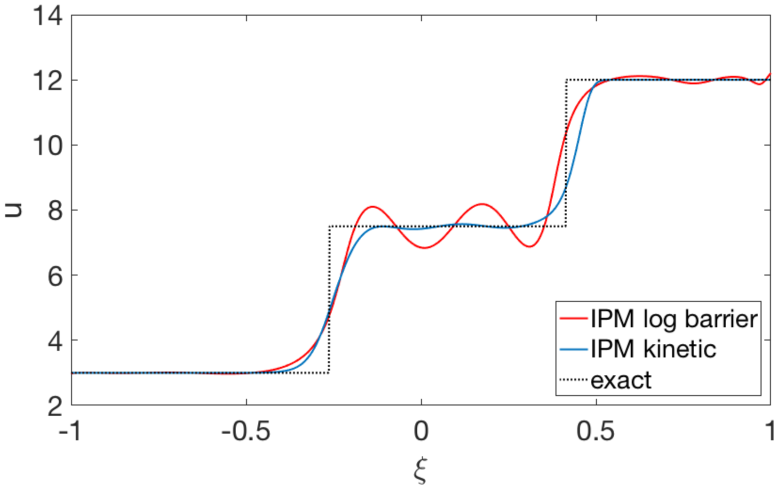
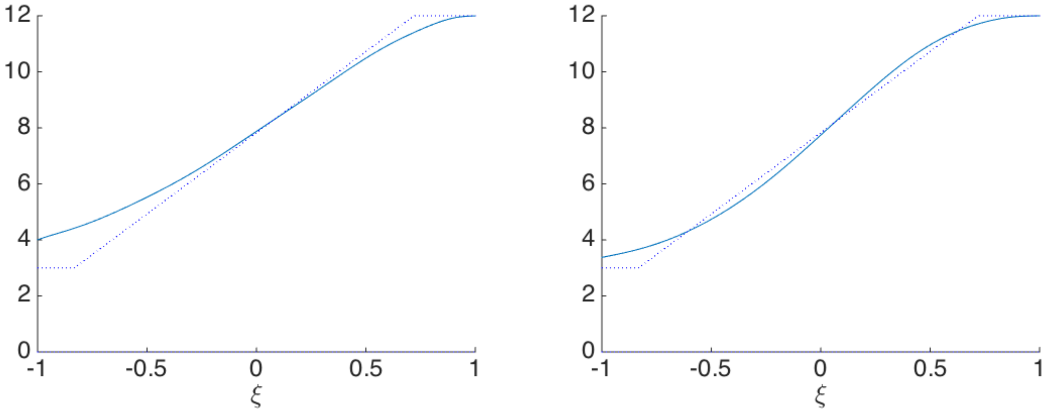


Figure 5: Solutions for log barrier and kinetic entropy at fixed spatial position x.



(a) No Limiter.

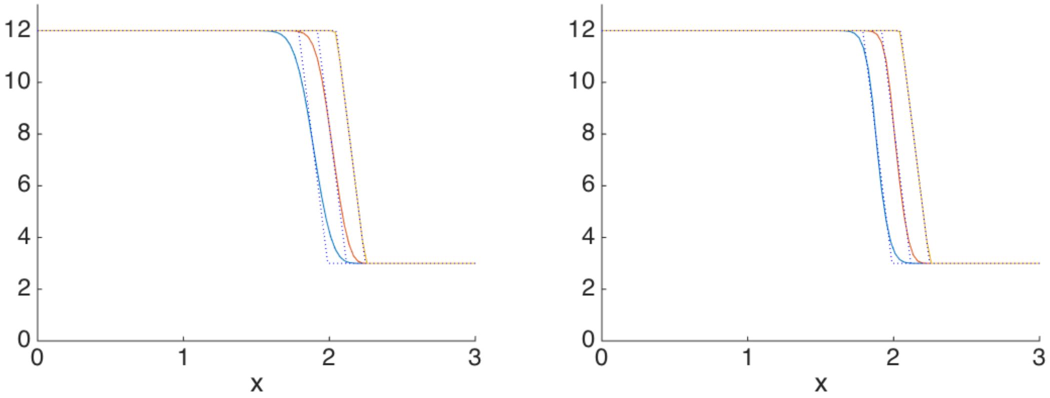
(b) Minmod Limiter.

Figure 6: Solution with and without a limiter.

will add diffusion to smaller velocities, while high velocities are well resolved. Furthermore, due to the equation's linear nature, the minmod limiter will preserve monotonicity of the scheme, allowing the use of this limiter. This means that we can obtain a high spatial resolution while preserving realizability, which will not be the case for general hyperbolic equations.

7 Conclusion and Outlook

In this paper, we have investigated minimal entropy closures in the context of uncertainty quantification. To resolve problems with realizability, we have introduced a new construction of the numerical flux, which together with the CFL condition $f'(u_+)\frac{\Delta t}{\Delta x} = 1$ leads to a monotone scheme if the residual of the dual problem is imposed on the moments. A min-max condition shows that monotonicity ensures realizable moments. In the case of linear equations, we were able to use a spatial limiter, which can easily be constructed by our choice of the numerical flux. This limiter will not destroy realizability. Furthermore, we were able to investigate the approximation error done by the minimal entropy reconstruction leading to a new choice of the entropy, which will ensure that non-critical regions of the solution will not suffer from the reconstruction, while



(a) No Limiter.

(b) Minmod Limiter.

Figure 7: Solution evaluated at $\xi \in \{-1, 0, 1\}$ with and without limiters.

the solution will lie in between the prescribed bounds. The resulting reconstructions show nice approximation behavior opposed to the log barrier entropy, which will be oscillatory in these regions. In order to demonstrate the analytically derived behavior, we looked at two test cases. It can be seen that the proposed modifications will lead to nice approximation results, while not leading to problems regarding realizability. In the case of the linear advection equation, it can be seen, that the limiter will lead to sharp approximations in space, but also in the random space, which heavily motivates the use of limiters.

We have seen that the use of spatial limiters will not necessarily lead to a realizable solution. However, these limiters will heavily improve the quality of the solution as well as allow a coarser grid spacing, which will reduce computational costs. It still needs to be investigated how these limiters can be applied for non-linear problems, as standard limiters will cause non-realizable moments.
Furthermore, the numerical results from the advection equation demonstrate that the derived modifications of the M_N method together with limiters will become interesting in the context of kinetic equations, which will be investigated in future work.

References

[1] Graham Alldredge and Florian Schneider. A realizability-preserving discontinuous galerkin scheme for entropy-based moment closures for linear kinetic equations in one space dimension. *Journal of Computational Physics*, 295:665–684, 2015.

[2] Graham W Alldredge, Cory D Hauck, and Andre L Tits. High-order entropy-based closures for linear transport in slab geometry ii: A computational study of the optimization problem. *SIAM Journal on Scientific Computing*, 34(4):B361–B391, 2012.

[3] Thomas A Brunner and James Paul Holloway. One-dimensional riemann solvers and the maximum entropy closure. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 69(5):543–566, 2001.

[4] Thomas A Brunner and James Paul Holloway. Two-dimensional time dependent riemann solvers for neutron transport. *Journal of Computational Physics*, 210(1):386–399, 2005.

[5] Kenneth M Case and Paul Frederick Zweifel. *Linear transport theory*. Addison-Wesley Pub. Co., 1967.

[6] Raúl E Curto and Lawrence A Fialkow. Recursiveness, positivity, and truncated moment problems. *Houston Journal of Mathematics*, 17(4):603–635, 1991.

[7] Bruno Després, Gaël Poëtte, and Didier Lucor. Robust uncertainty propagation in systems of conservation laws with the entropy closure method. In *Uncertainty quantification in computational fluid dynamics*, pages 105–149. Springer, 2013.

[8] B Dubroca and A Klar. Half-moment closure for radiative transfer equations. *Journal of Computational Physics*, 180(2):584–596, 2002.

[9] Martin Frank, Bruno Dubroca, and Axel Klar. Partial moment entropy approximation to radiative heat transfer. *Journal of Computational Physics*, 218(1):1–18, 2006.

[10] Roger G Ghanem and Pol D Spanos. *Stochastic finite elements: a spectral approach*. Courier Corporation, 2003.

[11] David Gottlieb and Dongbin Xiu. Galerkin method for wave equations with uncertain coefficients. *Commun. Comput. Phys*, 3(2):505–518, 2008.

[12] Cory Hauck and Ryan McClarren. Positive p_n closures. *SIAM Journal on Scientific Computing*, 32(5):2603–2626, 2010.

[13] Cory D Hauck. High-order entropy-based closures for linear transport in slab geometry. *Commun. Math. Sci*, 9(1):187–205, 2011.

[14] C David Levermore. Moment closure hierarchies for kinetic theories. *Journal of Statistical Physics*, 83(5-6):1021–1065, 1996.

[15] Elmer Eugene Lewis and Warren F Miller. Computational methods of neutron transport. 1984.

[16] Gerald N Minerbo. Maximum entropy eddington factors. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 20(6):541–545, 1978.

[17] Gaël Poëtte, Bruno Després, and Didier Lucor. Uncertainty quantification for systems of conservation laws. *Journal of Computational Physics*, 228(7):2443–2467, 2009.

[18] Gaël Poëtte, Bruno Després, and Didier Lucor. Treatment of uncertain material interfaces in compressible flows. *Computer Methods in Applied Mechanics and Engineering*, 200(1):284–308, 2011.

[19] Gerald C Pomraning. *The equations of radiation hydrodynamics*. Courier Corporation, 1973.

[20] Norbert Wiener. The homogeneous chaos. *American Journal of Mathematics*, 60(4):897–936, 1938.

[21] Dongbin Xiu and George Em Karniadakis. Modeling uncertainty in steady state diffusion problems via generalized polynomial chaos. *Computer Methods in Applied Mechanics and Engineering*, 191(43):4927–4948, 2002.