SIAM J. NUMER. ANAL.
Vol. 60, No. 4, pp. 2345–2371

AN ENTROPIC METHOD FOR DISCRETE SYSTEMS WITH GIBBS ENTROPY*

ZHENNING CAI†, JINGWEI HU‡, YANG KUANG§, AND BO LIN†

Abstract. We consider general systems of ordinary differential equations with monotonic Gibbs entropy and introduce an entropic scheme that simply imposes an entropy fix after every time step of any existing time integrator. It is proved that in the general case, our entropy fix has only infinitesimal influence on the numerical order of the original scheme, and in many circumstances, it can be shown that the scheme does not affect the numerical order. Numerical experiments on the linear Fokker–Planck equation and nonlinear Boltzmann equation are carried out to support our numerical analysis.

Key words. Gibbs entropy, entropic schemes, numerical accuracy

MSC code. 65L05

DOI. 10.1137/21M1429023

1. Introduction.

The second law of thermodynamics, discovered more than 170 years ago, states that the direction of the thermodynamic processes is driven by a physical quantity called entropy. The importance of this law cannot be overstated, and nearly every thermodynamic model has to respect such a property. Mathematically, there are a number of formulas to represent the entropy, among which the Gibbs entropy, formulated as the integral of $f \log f$ with f being the distribution function of the states, is widely used in a variety of models such as the heat equation, the Boltzmann equation, and the Fokker–Planck equation. In our discussion, we assume a finite number of states, so that the Gibbs entropy is defined by

$$\eta(\boldsymbol{f}) = \sum_{i=1}^{N} f_i \log f_i \Delta v_i,$$

where $\boldsymbol{f} = (f_1, \dots, f_N)^T \in \mathbb{R}_+^N$ describes the distribution of the N states and $\Delta v_i \in \mathbb{R}_+$ represents the weight of the ith state which might come from the size of the ith grid or the quadrature rule. The vector $\boldsymbol{f}$ is a vector function of time t, and we assume that it satisfies the initial value problem

$$\begin{aligned} &\frac{\mathrm{d}f_i(t)}{\mathrm{d}t} = Q_i(\boldsymbol{f}(t)), \quad i = 1, \dots, N, \\ &f_i(0) = f_i^0, \end{aligned} \tag{1.1}$$

*Received by the editors June 23, 2021; accepted for publication (in revised form) June 27, 2022; published electronically August 31, 2022.
https://doi.org/10.1137/21M1429023
Funding: The work of the first author was supported by the Academic Research Fund of the Ministry of Education of Singapore grants R-146-000-305-114 and R-146-000-326-112. The work of the second author was partially supported by the National Science Foundation CAREER grant DMS-2153208.

†Department of Mathematics, National University of Singapore, Singapore 119076 (matcz@nus.edu.sg, matbl@nus.edu.sg).
‡Department of Applied Mathematics, University of Washington, Seattle, WA 98195 USA (hujw@uw.edu).
§School of Mathematics and Statistics, Guangdong University of Technology, Guangzhou 510006, Guangdong, China (ykuang@gdut.edu.cn).

with the following properties:

(P1) conservation of mass: $\frac{\mathrm{d}}{\mathrm{d}t}\sum_{i=1}^{N} f_i(t)\Delta v_i = 0$;

(P2) nonnegativity: $f_i(t) \geq 0$ for all $1 \leq i \leq N, t \geq 0$;

(P3) monotonicity of entropy: $\frac{\mathrm{d}}{\mathrm{d}t}\sum_{i=1}^{N} f_i(t)\log f_i(t)\Delta v_i \leq 0$.

The ODE system of the form (1.1) appears frequently after discretizing the thermodynamic equations in space. For example, it may arise from the finite difference discretization of the heat equation and the Fokker–Planck-type equations [10, 6, 17, 8], or the discrete velocity method and the entropic Fourier method for the Boltzmann equation [11, 7].

Although the semidiscrete scheme (1.1) decays entropy, there is no guarantee that this property will carry over when time is discretized. In some special cases, the entropy decay can be proved for the fully discrete scheme [3, 14, 15], yet it often comes at a price of using implicit schemes and is highly problem and scheme dependent. Given the importance of entropy in thermodynamic processes, it would be desirable to have a fully discrete entropic scheme that is generic (e.g., does not require a specific type of time discretization) as well as easily implementable (e.g., does not require expensive nonlinear iterations).

Researchers have been trying to develop entropic schemes for explicit time integrators [4, 5, 20]. One recent progress is the relaxation Runge–Kutta method [13, 18], which enforces the entropy dissipation by adding a relaxation parameter to the final stage of each Runge–Kutta step. However, such a modification can cause slower convergence, which is typically one order less than the original Runge–Kutta method. Another recent work [1, 2] applies the deferred correction method under the residual distribution framework, which needs calculating additional entropy residual and solving a series of correction substeps to reach desired order of accuracy.

To bridge the above gap, we introduce an entropic scheme in this paper to achieve the following: one can apply any time discretization to the system (1.1) as long as it maintains the mass conservation and nonnegativity of the solution. After each time step, if the entropy goes in the wrong direction, we provide a simple fix to make it decay monotonically. Such a fix is done by a weighted average of the current solution and the solution with maximum entropy. This fix only at the final stage of each time step is similar to the idea of the relaxation Runge–Kutta method [18]. Via numerical analysis, we show that such a fix has only a tiny effect on the order of accuracy, and in various cases, it can be proven that the order of accuracy is not affected at all. Numerical experiments on the linear Fokker–Planck equation and nonlinear Boltzmann equation will also be carried out to support our findings.

The paper is organized as follows. In section 2, we first outline the procedure of our entropic method and summarize the main theorems of the method. The detailed proof of the theorems with some deeper understandings is illustrated in section 3. Section 4 provides the numerical experiments, and the conclusion follows in section 5.

2. Main results.

This section outlines the overall procedure of our entropic method and lists the main results of our numerical analysis. Before stating our theorems, we introduce the notations and review some basic properties of the Gibbs entropy.

2.1. Brief review of Gibbs entropy.

Due to the conservation hypothesis (P1), below we focus on the entropy functional defined by

$$H(\boldsymbol{f}) = \sum_{i=1}^{N} (f_i \log f_i - f_i)\Delta v_i := \sum_{i=1}^{N} h(f_i)\Delta v_i$$

with $h(x) = x \log x - x$. Note that $H(\boldsymbol{f})$ differs from $\eta(\boldsymbol{f})$ only by a constant.

Let $\boldsymbol{C} = (C, \dots, C)^T \in \mathbb{R}_+^N$ with

$$C = \frac{\sum_{i=1}^{N} f_i \Delta v_i}{\sum_{i=1}^{N} \Delta v_i}. \tag{2.1}$$

We denote by $\tilde{\boldsymbol{f}} = \boldsymbol{f}/C = (\tilde{f}_1, \dots, \tilde{f}_N)^T$ the normalized $\boldsymbol{f}$; then it can be checked that

$$C\eta(\tilde{\boldsymbol{f}}) = H(\boldsymbol{f}) - H(\boldsymbol{C}). \tag{2.2}$$

Furthermore, we define the L^p $(p = 1, 2)$ norm and L^∞ norm of any $\boldsymbol{f}$ as

$$\|\boldsymbol{f}\|_p = \left(\sum_{i=1}^{N} f_i^p \Delta v_i \right)^{1/p}, \qquad \|\boldsymbol{f}\|_\infty = \max_i |f_i|.$$

LEMMA 2.1. *$\boldsymbol{C}$ is the unique global minimum point of $H(\boldsymbol{f})$ for all $\boldsymbol{f} \in \mathbb{R}_+^N$ satisfying* (2.1) *with fixed C.*

The proof of Lemma 2.1 can be done by the concavity of $\log(x)$ and Jensen's inequality. Furthermore, a straightforward corollary of Lemma 2.1 is that $\mathbf{1} = (1, \dots, 1)^T \in \mathbb{R}_+^N$ is the unique global minimum point of $\eta(\tilde{\boldsymbol{f}})$ for all $\tilde{\boldsymbol{f}} \in \mathbb{R}_+^N$ satisfying $\|\tilde{\boldsymbol{f}}\|_1 = \|\mathbf{1}\|_1$. To ease the notation, we use $\|\mathbf{1}\|_1 = \sum_{i=1}^{N} \Delta v_i = V$ to denote the volume.

The notations hereafter will be focused on the relative entropy $\eta(\tilde{\boldsymbol{f}})$ and the normalized $\tilde{\boldsymbol{f}}$ for fixed C. One could find its relationship to the entropy function $H(\cdot)$ from (2.2). For simplicity, we would like to omit the tilde symbol in $\tilde{\boldsymbol{f}}$, and thus the average of the components of $\boldsymbol{f}$ will be 1 hereafter.

2.2. Main results. We assume after temporal discretization of (1.1), the properties (P1) and (P2) can be preserved. Specifically, if we let $\boldsymbol{f}^n \geq 0$ be the numerical solution at the nth time step, then we have

(H1) conservation: $\sum_{i=1}^{N} f_i^{n+1} \Delta v_i = \sum_{i=1}^{N} f_i^n \Delta v_i$;

(H2) nonnegativity: $f_i^{n+1} \geq 0$ for all $1 \leq i \leq N$.

We would like to design an entropic method such that it can fulfill a discrete version of (P3) while keeping (H1) and (H2).

Our numerical scheme is based on imposing a simple entropy fix after computing the numerical solution at every time step. Suppose that $\boldsymbol{f}^{n+1}$ is computed through evolving $\boldsymbol{f}^n$ by one time step. If $\eta(\boldsymbol{f}^{n+1}) \leq \eta(\boldsymbol{f}^n)$, nothing needs to be done. Otherwise, we revise the solution at the $(n+1)$th time step by a convex combination with the state of minimum entropy. According to Lemma 2.1, such a state is $\mathbf{1}$ under our conservation hypothesis, and thus the revised numerical solution is

$$\hat{\boldsymbol{f}}^{n+1} = \boldsymbol{f}^{n+1} + \beta_p(\mathbf{1} - \boldsymbol{f}^{n+1}), \tag{2.3}$$

where $\beta_p \in (0, 1]$ is chosen to satisfy

$$\eta(\boldsymbol{f}^{n+1} + \beta_p(\mathbf{1} - \boldsymbol{f}^{n+1})) = \eta(\boldsymbol{f}^n). \tag{2.4}$$

This guarantees that the entropy is always nonincreasing. In the case where the equilibrium is not a constant, the same technique can be applied, and we refer the readers to Remark 2.7 for some comments on such cases. It can be shown that the solution of (2.4) exists since the function $\phi(\beta) := \eta(\boldsymbol{f}^{n+1}+\beta(\mathbf{1}-\boldsymbol{f}^{n+1}))$ is convex and monotonic with respect to $\beta \in [0,1]$, and we have $\phi(0) > \eta(\boldsymbol{f}^n)$ and $\phi(1) \le \eta(\boldsymbol{f}^n)$. Therefore, we can use the bisection method to solve the equation efficiently.

In most cases, such a method stabilizes the solution since it reduces both the Gibbs entropy and the 2-norm of vectors, where the reduction of 2-norm can be shown by the convexity of the 2-norm and Jensen's inequality. Therefore we are mainly concerned about the magnitude of the fixing term $\beta_p(\mathbf{1} - \boldsymbol{f}^{n+1})$, and we hope that this term does not affect the numerical convergence order of the original scheme. Generally, the error estimation of this scheme can be analyzed in the following manner:

(2.5)
$$\begin{aligned}\|\hat{\boldsymbol{f}}^{n+1} - \boldsymbol{f}(t_{n+1})\| &\le \|\hat{\boldsymbol{f}}^{n+1} - \boldsymbol{f}^{n+1}\| + \|\boldsymbol{f}^{n+1} - \boldsymbol{f}(t_{n+1})\| \\ &\le \|\hat{\boldsymbol{f}}^{n+1} - \boldsymbol{f}^{n+1}\| + \|\boldsymbol{f}^{n+1} - \tilde{\boldsymbol{f}}(t_{n+1})\| + \|\tilde{\boldsymbol{f}}(t_{n+1}) - \boldsymbol{f}(t_{n+1})\|,\end{aligned}$$

where $\tilde{\boldsymbol{f}}(t)$ is the solution of the problem

$$\begin{aligned}&\frac{\mathrm{d}\tilde{f}_i(t)}{\mathrm{d}t} = Q_i(\tilde{\boldsymbol{f}}(t)), \quad i = 1, \dots, N, \\ &\tilde{f}_i(t_n) = f_i^n, \qquad\qquad i = 1, \dots, N,\end{aligned} \tag{2.6}$$

and hence $\|\boldsymbol{f}^{n+1} - \tilde{\boldsymbol{f}}(t_{n+1})\|$ is the "one-step error" of the scheme. The last term in (2.5) is usually controlled by the stability of the ODE problem with respect to the initial condition. If we assume that the scheme satisfies the consistency condition

$$\|\boldsymbol{f}(t_{n+1}) - \boldsymbol{f}^{n+1}\| \le O(\Delta t^{s+1}),$$

then the original scheme (before our entropy fix) is a scheme of order s. Here our purpose is to demonstrate that the first term in the second line of (2.5), i.e., $\|\beta_p(\mathbf{1} - \boldsymbol{f}^{n+1})\|$, can be controlled by the second term $\|\tilde{\boldsymbol{f}}(t_{n+1}) - \boldsymbol{f}^{n+1}\|$. In the ideal case, we may find a constant C such that

$$\|\beta_p(\mathbf{1} - \boldsymbol{f}^{n+1})]\| \le C\|\tilde{\boldsymbol{f}}(t_{n+1}) - \boldsymbol{f}^{n+1}\|;$$

then the numerical convergence order is not affected. Hereafter, for simplicity, we would like to omit the tilde and use $\boldsymbol{f}(t_{n+1})$ to denote the solution of (2.6) at time t_{n+1}. In other words, we assume that the solution at the nth time step $\boldsymbol{f}^n$ is exact ($\boldsymbol{f}(t_n) = \boldsymbol{f}^n$), so that $\boldsymbol{f}(t_{n+1})$ becomes identical to $\tilde{\boldsymbol{f}}(t_{n+1})$.

In the following theorems, we will study a stronger result,

$$\eta\left(\boldsymbol{f}^{n+1} + \beta(\mathbf{1} - \boldsymbol{f}^{n+1})\right) = \eta(\boldsymbol{f}(t_{n+1})), \tag{2.7}$$

where β_p in (2.4) is replaced by β and the solution at $(n+1)$th time step is revised to possess the same entropy as $\boldsymbol{f}(t_{n+1})$. Due to $\eta(\boldsymbol{f}(t_{n+1})) \le \eta(\boldsymbol{f}^n)$ and the monotonicity of $\eta(\boldsymbol{f}^{n+1} + \omega(\mathbf{1} - \boldsymbol{f}^{n+1}))$ with respect to ω, we see that $\beta_p \le \beta$. Therefore, it suffices to show that $\|\beta(\mathbf{1} - \boldsymbol{f}^{n+1})\|$ can be controlled by the difference between $\boldsymbol{f}(t_{n+1})$ and $\boldsymbol{f}^{n+1}$. Based on the commonly used 2-norm of vectors, we are going to prove this type of results in four different scenarios, which will be stated in the four theorems listed below.

In the first case, we have no assumptions on the structure of the solution, which may lead to a slight reduction of the numerical convergence order.

THEOREM 2.2. *Given a positive and conservative numerical scheme, i.e.,* $\boldsymbol{f}^{n+1} \in \mathbb{R}_+^N$ *and* $\|\boldsymbol{f}^{n+1}\|_1 = \|\boldsymbol{f}(t_{n+1})\|_1$*, when* $\eta(\boldsymbol{f}^{n+1}) > \eta(\boldsymbol{f}^n)$ *and* (2.7) *are satisfied, if* $\|\boldsymbol{f}(t_{n+1}) - \boldsymbol{f}^{n+1}\|_2 \le 1$*, then*

$$\|\beta(\mathbf{1} - \boldsymbol{f}^{n+1})\|_2 \le M\|\boldsymbol{f}(t_{n+1}) - \boldsymbol{f}^{n+1}\|_2 \left(1 + \left|\log\left(\|\boldsymbol{f}(t_{n+1}) - \boldsymbol{f}^{n+1}\|_2\right)\right|\right),$$

where $M > 0$ *is a constant which depends on* V*,* $\|\boldsymbol{f}^{n+1}\|_\infty$*, and* $\|\boldsymbol{f}(t_{n+1})\|_\infty$*.*

In this case, the right-hand side of the inequality contains a logarithmic term, which tends to infinity when $\|\boldsymbol{f}(t_{n+1}) - \boldsymbol{f}^{n+1}\|_2$ approaches zero. However, for any $\epsilon > 0$, we have

$$1 + \left|\log\left(\|\boldsymbol{f}(t_{n+1}) - \boldsymbol{f}^{n+1}\|_2\right)\right| < \|\boldsymbol{f}(t_{n+1}) - \boldsymbol{f}^{n+1}\|_2^{-\epsilon}$$

when $\|\boldsymbol{f}(t_{n+1}) - \boldsymbol{f}^{n+1}\|_2$ is sufficiently small, meaning that the numerical convergence order is reduced only by an arbitrarily small positive number. Nevertheless, we would still like to explore the conditions under which such a logarithmic term does not exist. The remaining three cases are related to this type of results.

Intuitively, the reason of the logarithmic term in Theorem 2.2 is the unboundedness of the function $h'(x)$ when x is close to zero. In the following result, we assume that the components of the numerical solution $\boldsymbol{f}^{n+1}$ have a lower bound C_0 such that $h'(x)$ becomes bounded.

THEOREM 2.3. *Given a positive and conservative numerical scheme, i.e.,* $\boldsymbol{f}^{n+1} \in \mathbb{R}_+^N$ *and* $\|\boldsymbol{f}^{n+1}\|_1 = \|\boldsymbol{f}(t_{n+1})\|_1$*, when* $\eta(\boldsymbol{f}^{n+1}) > \eta(\boldsymbol{f}^n)$ *and* (2.7) *are satisfied, if* $f_i^{n+1} \ge C_0 > 0$ *holds for all* $1 \le i \le N$*, then*

$$\|\beta(\mathbf{1} - \boldsymbol{f}^{n+1})\|_2 \le M\|\boldsymbol{f}(t_{n+1}) - \boldsymbol{f}^{n+1}\|_2,$$

where $M > 0$ *is a constant which depends on* C_0*,* $\|\boldsymbol{f}^{n+1}\|_\infty$*, and* $\|\boldsymbol{f}(t_{n+1})\|_\infty$*.*

The condition in this theorem disallows the numerical solution to be zero anywhere in the domain. When some components of the numerical solution equal zero, we can still show that the L^2 error after the entropy fix is small if the scheme can guarantee the numerical convergence order for the L^∞ error. This corresponds to our third case.

THEOREM 2.4. *Given a positive and conservative numerical scheme, i.e.,* $\boldsymbol{f}^{n+1} \in \mathbb{R}_+^N$ *and* $\|\boldsymbol{f}^{n+1}\|_1 = \|\boldsymbol{f}(t_{n+1})\|_1$*, when* $\eta(\boldsymbol{f}^{n+1}) > \eta(\boldsymbol{f}^n)$ *and* (2.7) *are satisfied, if* $\|\boldsymbol{f}(t_{n+1}) - \boldsymbol{f}^{n+1}\|_\infty \le 1/3$*, it holds that*

$$\|\beta(\mathbf{1} - \boldsymbol{f}^{n+1})\|_2 \le M\|\boldsymbol{f}(t_{n+1}) - \boldsymbol{f}^{n+1}\|_\infty,$$

where $M > 0$ *is a constant which depends on* V*,* $\|\boldsymbol{f}^{n+1}\|_\infty$*, and* $\|\boldsymbol{f}(t_{n+1})\|_\infty$*.*

The last case we consider can be regarded as a generalization of Theorem 2.3. We allow the numerical solution to be small on some part of the domain but require that the solution increases slowly. This will lead to a result similar to the conclusion of Theorem 2.3, where the L^2-magnitude of the entropy fix can be directly bounded by the L^2 error.

THEOREM 2.5. *Given a positive and conservative numerical scheme, i.e.,* $\boldsymbol{f}^{n+1} \in \mathbb{R}_+^N$ *and* $\|\boldsymbol{f}^{n+1}\|_1 = \|\boldsymbol{f}(t_{n+1})\|_1$*, we denote the components of* $\boldsymbol{f}^{n+1}$ *as* $f_1^{n+1} \le f_2^{n+1} \le \cdots \le f_N^{n+1}$*. For any* $C_1, C_f \in (0, 1]$*, there exist two positive constants* δ *and* M *such that*

$$\|\beta(\mathbf{1} - \boldsymbol{f}^{n+1})\|_2 \le M\|\boldsymbol{f}^{n+1} - \boldsymbol{f}(t_{n+1})\|_2$$

if all the following conditions hold:

- $\eta(\boldsymbol{f}^{n+1}) > \eta(\boldsymbol{f}^n)$ *and* $\eta(\boldsymbol{f}^{n+1} + \beta(\mathbf{1} - \boldsymbol{f}^{n+1})) = \eta(\boldsymbol{f}(t_{n+1}))$*;*
- $\|\boldsymbol{f}^{n+1} - \boldsymbol{f}(t_{n+1})\|_2 < \delta$*;*
- *The index* $I_1 = \min\{I \mid \sum_{i=1}^{I} \Delta v_i \geq C_1 V\}$ *satisfies*

$$\frac{1}{|\log\left(f_1^{n+1}\right)|} \geq \frac{C_f}{|\log\left(f_{I_1}^{n+1}\right)|}. \tag{2.8}$$

Here δ *depends on* C_1*,* C_f*, and* V*, and* M *depends on* C_1*,* C_f*,* V*,* $\|\boldsymbol{f}^{n+1}\|_\infty$*, and* $\|\boldsymbol{f}(t_{n+1})\|_\infty$*.*

In (2.8), the function $1/|\log x|$ is regarded as zero when x takes the value zero. The condition (2.8) allows the existence of small components in the solution. To better demonstrate the nature of this condition, two examples are presented below.

Example 1. This example assumes that $\boldsymbol{f}^{n+1}$ is the uniform discretization of a one-dimensional Gaussian, i.e.,

$$\Delta v_i = \Delta v,\ f_i^{n+1} = \frac{1}{C\sqrt{\pi}} \exp(-v_i^2), \qquad i = 1, \ldots, N+1,$$

where v_i are uniformly distributed in $[-L, L]$, $\Delta v = 2L/(N+1)$, and $L > 0$ is set to be sufficiently large such that $\exp(-L^2)$ is sufficiently small. The constant C is chosen such that $\|\boldsymbol{f}^{n+1}\|_1 = \|\mathbf{1}\|_1$. Furthermore, fox fixed L, we assume that N is an even number and large enough such that $C \geq 1/(4L)$. According to the assumption of Theorem 2.5, we set v_i to be

$$v_i = (-1)^i \lceil (N+1-i)/2 \rceil \frac{2L}{N}, \qquad i = 1, \ldots, N+1,$$

such that f_i^{n+1} increases with respect to i. For illustration, we plot the normalized Gaussian and its sorted version in Figure 1, where parameters are set as $L = 6$ and $N = 20$. In this example, we take $I_1 = \lceil (N+1)/2 \rceil = N/2 + 1$; then

$$\frac{\log(f_{I_1}^{n+1})}{\log(f_1^{n+1})} = \frac{\log \frac{1}{C\sqrt{\pi}} - v_{I_1}^2}{\log \frac{1}{C\sqrt{\pi}} - L^2} \geq \frac{v_{I_1}^2 - \log(\frac{4L}{\sqrt{\pi}})}{L^2} \geq \frac{v_{I_1}^2}{2L^2} \geq \frac{1}{8},$$

which satisfies (2.8) with $C_1 = 1/2$ and $C_f = 1/8$. This example shows a case where the values of f_i^{n+1} are nonzero but can be arbitrarily small.

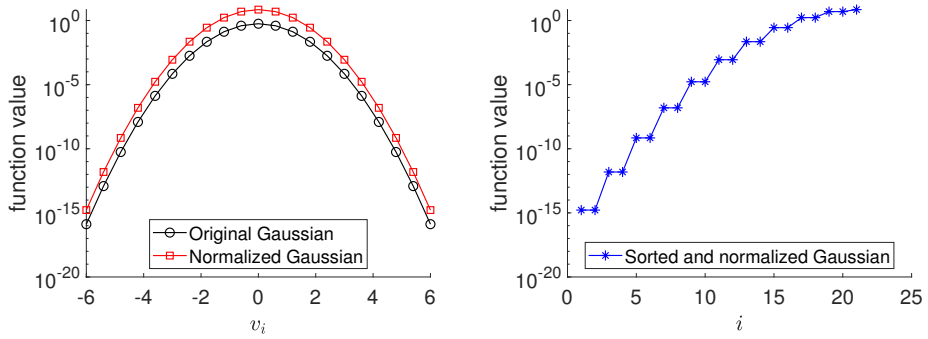


FIG. 1. *Discretized Gaussian, its normalization and sorted notation in Example* 1.

Example 2. The second example is for the case where some components of $\boldsymbol{f}^{n+1}$ are zero. We assume a uniform discretization on $[0,1]$ with $\Delta v_i = 1/N$ for $i = 1, \ldots, N$ and choose $\boldsymbol{f}^{n+1}$ to be

$$f_i^{n+1} = \begin{cases} 0, & i = 1, \ldots, I_1, \\ 1, & i = I_1 + 1, \ldots, N - I_1, \\ 2, & i = N - I_1 + 1, \ldots, N. \end{cases}$$

If I_1/N is a constant, the vector $\boldsymbol{f}^{n+1}$ approximates a piecewise constant function. In this case, Theorem 2.5 holds by choosing $C_1 = I_1/N$ and C_f to be any positive number in $(0,1]$. The blue lines in Figure 2 show the situation where $C_1 = 1/3$. However, if I_1/N decreases to zero as N increases, e.g., $I_1 \equiv 1$ for all N, such a constant C_1 cannot be found. This situation violates the condition of Theorem 2.5, which is illustrated as the red lines in Figure 2.

In general, the above theorems suggest that such an entropy fix can be safely used without sacrificing the numerical accuracy. Moreover, for a numerical scheme with sufficient accuracy, the violation of the entropy inequality will not always happen, meaning that the entropy fix may be needed only at a few time steps, resulting in even less significant impact on the numerical accuracy.

Remark 2.6. The value of β in the above theorems is determined by (2.7), which is not practical in real implementation as $\boldsymbol{f}(t_{n+1})$ is unknown. Nevertheless, as stated after (2.7), the computable β_p defined in (2.4) satisfies $\beta_p \le \beta$. Therefore, all the theorems still hold if we replace β by β_p in the theorems.

Remark 2.7. The above results can be easily generalized to the cases where the equilibrium is not a constant. Assume that $\boldsymbol{\mathcal{M}} = (\mathcal{M}_1, \ldots, \mathcal{M}_N)^T \in \mathbb{R}_N^+$ is the equilibrium state of (1.1), and the entropy functional (in this case, it is the relative entropy) is defined by

$$\eta[\boldsymbol{f}] = \sum_{i=1}^{N} f_i \log \frac{f_i}{\mathcal{M}_i} \Delta v_i.$$

We can let $g_i = f_i/\mathcal{M}_i$ and $\Delta w_i = \mathcal{M}_i \Delta v_i$ so that $\eta[\boldsymbol{f}]$ can be rewritten as

$$\eta[\boldsymbol{f}] = \sum_{i=1}^{N} g_i \log g_i \Delta w_i,$$

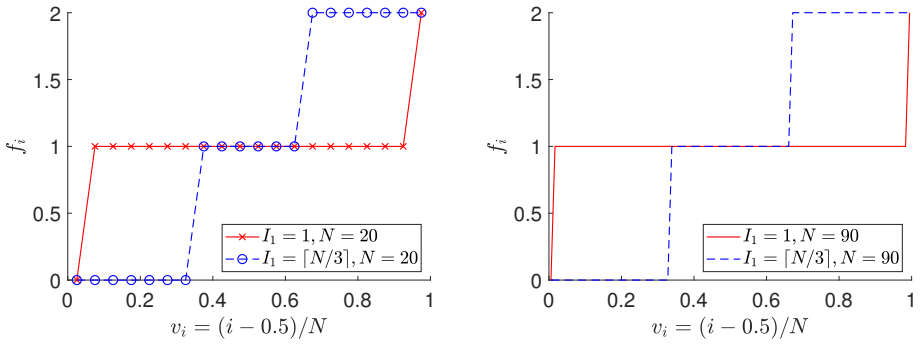


FIG. 2. *Illustration of Example* 2 *where* f_i^{n+1} *can only be chosen as* 0, 1, *or* 2.

which fits the entropy formulas in the theorems again. In this case, the entropy fix (2.3) applied to $\boldsymbol{g}^{n+1}$ is equivalent to the following fix applied to $\boldsymbol{f}^{n+1}$:

$$\hat{\boldsymbol{f}}^{n+1} = \boldsymbol{f}^{n+1} + \beta_p(\boldsymbol{\mathcal{M}} - \boldsymbol{f}^{n+1}). \tag{2.9}$$

By this transformation, our approach can also be applied to the linear Fokker–Planck equation. Please see the numerical section for more details.

3. Theoretical proofs of the error estimates. This section provides all the details of the proofs of the four theorems. Instead of proving these theorems in the order they are presented, below we will first provide the proof of Theorem 2.3, which can provide necessary tools needed in the proof of Theorem 2.2.

3.1. Proof of Theorem 2.3. Before proving the theorem, the relationship between the entropy function and the L^2 norm will be demonstrated by several lemmas. Among them, we will first estimate the entropy function $\eta(\boldsymbol{f})$ and its L^2 norm $\|\boldsymbol{f}\|_2$ in the following lemma.

LEMMA 3.1. *For $\boldsymbol{f} \in \mathbb{R}_+^N$ and $\|\boldsymbol{f}\|_1 = V$,*

$$\frac{1}{2\|\boldsymbol{f}\|_\infty}\|\boldsymbol{f} - \mathbf{1}\|_2^2 \le \eta(\boldsymbol{f}) \le \|\boldsymbol{f} - \mathbf{1}\|_2^2.$$

Proof. On one hand, for $x \ge 0$,

$$x\log x - (x-1) \le x(x-1) - (x-1) = (x-1)^2,$$

where the inequality above uses $\log x \le x-1$. On the other hand, by Taylor's theorem,

$$x\log x = (x-1) + \int_1^x \frac{1}{t}(x-t)\mathrm{d}t.$$

For $0 \le x \le \|\boldsymbol{f}\|_\infty$, the integral satisfies

$$\int_1^x \frac{1}{t}(x-t)\mathrm{d}t \ge \int_1^x \frac{1}{\max(x,1)}(x-t)\mathrm{d}t \ge \int_1^x \frac{1}{\|\boldsymbol{f}\|_\infty}(x-t)\mathrm{d}t = \frac{(x-1)^2}{2\|\boldsymbol{f}\|_\infty}.$$

Therefore,

$$(x-1) + \frac{(x-1)^2}{2\|\boldsymbol{f}\|_\infty} \le x\log x \le (x-1) + (x-1)^2.$$

The lemma can be proved by taking $x = f_i$ in the above inequality and summing up all $1 \le i \le N$. ∎

A straightforward corollary of the above lemma is given as follows.

LEMMA 3.2. *For $\boldsymbol{f}^{(1)} \in \mathbb{R}_+^N$ and $\boldsymbol{f}^{(2)} \in \mathbb{R}_+^N$ with $\|\boldsymbol{f}^{(1)}\|_1 = \|\boldsymbol{f}^{(2)}\|_1 = V$, if $\eta(\boldsymbol{f}^{(1)}) \le \eta(\boldsymbol{f}^{(2)})$, then it holds that*

$$\|\boldsymbol{f}^{(1)} - \mathbf{1}\|_2^2 \le 2\|\boldsymbol{f}^{(1)}\|_\infty \|\boldsymbol{f}^{(2)} - \mathbf{1}\|_2^2.$$

After showing the equivalence between the entropy function and the 2-norm, we will proceed to discuss the relationship between $\eta(\boldsymbol{f}^{(1)}) - \eta(\boldsymbol{f}^{(2)})$ and $\|\boldsymbol{f}^{(1)} - \boldsymbol{f}^{(2)}\|_2$ for any two vectors $\boldsymbol{f}^{(1)}$ and $\boldsymbol{f}^{(2)}$. By the definition of $\eta(\cdot)$, we are inspired to study the estimation of $h(x) - h(y)$. The result is presented in the following lemma.

LEMMA 3.3. *Given* $0 < C_0 \leq 1$, $y \geq 0$, *and* $x \geq C_0$, *if* $y \geq C_0$ *or* $h(x) > h(y)$, *then*

$$|h(x) - h(y)| \leq \max\left(2, 2|\log(C_0)|\right) |x - y| \left(|x - 1| + |y - 1|\right). \tag{3.1}$$

Proof. If $x = y$, it is obvious that the lemma is correct. It remains to prove the lemma when $x \neq y$.

By the mean value theorem,

$$h(x) - h(y) = \log(\xi)(x - y), \tag{3.2}$$

where ξ is between x and y. If $\log(\xi) \geq 0$, it holds that $\xi \geq 1$ and

$$|\log(\xi)| = \log(\xi) \leq \xi - 1 \leq \max(x - 1, y - 1) \leq |x - 1| + |y - 1|.$$

Therefore, if $\log(\xi) \geq 0$, (3.2) becomes

$$|h(x) - h(y)| \leq |x - y| \left(|x - 1| + |y - 1|\right). \tag{3.3}$$

Next we assume $h(x) > h(y)$. If $h(x) > h(y)$ and $x > y$, (3.2) implies $\log(\xi) > 0$, which gives (3.3). If $h(x) > h(y)$ and $x < y$, (3.2) implies $\log(\xi) < 0$ and $\xi \leq 1$. In this case, $y > x \geq C_0$, which implies $\xi \geq C_0$ and $\log(\xi) \geq \log(C_0)$. Therefore, (3.2) becomes

$$|h(x) - h(y)| = -\log(\xi)|x - y| \leq -\log(C_0)|x - y| = |\log(C_0)||x - y|. \tag{3.4}$$

On the other hand, by the mean value theorem,

$$-\log(\xi) = \log(1) - \log(\xi) = \frac{1}{\xi_2}(1 - \xi) \leq \frac{1}{C_0}\left(|x - 1| + |y - 1|\right),$$

where $\xi_2 \in [\xi, 1] \subset [C_0, 1]$. The above results can be summarized into the following estimation:

$$|h(x) - h(y)| \leq |x - y| \min\left(|\log(C_0)|, \frac{1}{C_0}\left(|x - 1| + |y - 1|\right)\right). \tag{3.5}$$

If we further assume $x \geq 1/2$ and $y \geq 1/2$, then (3.5) is satisfied with $C_0 = 1/2$, which becomes

$$|h(x) - h(y)| \leq |x - y| \min\left(\log 2, 2\left(|x - 1| + |y - 1|\right)\right) \leq 2|x - y| \left(|x - 1| + |y - 1|\right).$$

Otherwise, if $x < 1/2$ or $y < 1/2$, we have $2(|x - 1| + |y - 1|) \geq 1$. Therefore,

$$\min\left(|\log(C_0)|, \frac{1}{C_0}\left(|x - 1| + |y - 1|\right)\right) \leq |\log(C_0)| \leq 2|\log(C_0)| \left(|x - 1| + |y - 1|\right).$$

Combining the two results above yields the inequality (3.1) when $h(x) > h(y)$.

It remains only to consider the case $h(x) \leq h(y)$ and $y \geq C_0$. If $x < y$, (3.2) implies $\log(\xi) \geq 0$, which gives the result of (3.3). Otherwise, $x > y$ implies $\log(\xi) \leq 0$. Since $x \geq C_0$ and $y \geq C_0$, it holds that $\xi \geq C_0$, and therefore $0 \geq \log(\xi) \geq \log(C_0)$, which also yields (3.5). The rest of the proof is the same as the previous case. ∎

With the help of the above lemma, we could give an upper bound of the difference of entropy functions $\eta(\boldsymbol{f}^{(1)}) - \eta(\boldsymbol{f}^{(2)})$ in the following lemma.

LEMMA 3.4. *For* $\boldsymbol{f}^{(1)} = (f_1^{(1)}, \dots, f_N^{(1)}) \in \mathbb{R}_+^N$ *and* $\boldsymbol{f}^{(2)} = (f_1^{(2)}, \dots, f_N^{(2)}) \in \mathbb{R}_+^N$ *with* $\|\boldsymbol{f}^{(1)}\|_1 = \|\boldsymbol{f}^{(2)}\|_1 = V$, *given* $0 < C_0 \le 1$, *if* $f_i^{(1)} \ge C_0$ *for all* $1 \le i \le N$ *and* $\boldsymbol{f}^{(2)}$ *satisfies either of the conditions*

1. $f_i^{(2)} \ge C_0$ *for all* $1 \le i \le N$ *or*
2. $\eta(\boldsymbol{f}^{(2)}) < \eta(\boldsymbol{f}^{(1)})$,

then it holds that

$$\left|\eta(\boldsymbol{f}^{(1)}) - \eta(\boldsymbol{f}^{(2)})\right| \le \max\left(2, 2|\log(C_0)|\right) \|\boldsymbol{f}^{(1)} - \boldsymbol{f}^{(2)}\|_2 \left(\|\boldsymbol{f}^{(1)} - \mathbf{1}\|_2 + \|\boldsymbol{f}^{(2)} - \mathbf{1}\|_2\right).$$

Proof. For simplicity, we use M to denote the constant $\max\left(2, 2|\log(C_0)|\right)$ in this proof. In the first case, $f_i^{(2)} \ge C_0 > 0$ for all $1 \le i \le N$, we can plug $x = f_i^{(1)}$ and $y = f_i^{(2)}$ in Lemma 3.3 and sum over all $1 \le i \le N$. By using $\|\boldsymbol{f}^{(1)}\|_1 = \|\boldsymbol{f}^{(2)}\|_1$, we can obtain that

$$\left|\eta(\boldsymbol{f}^{(1)}) - \eta(\boldsymbol{f}^{(2)})\right| \le M \sum_{i=1}^{N} \left(\left|(f_i^{(1)} - 1)(f_i^{(1)} - f_i^{(2)})\right| + \left|(f_i^{(2)} - 1)(f_i^{(1)} - f_i^{(2)})\right|\right) \Delta v_i.$$

The lemma can be proven by the Cauchy–Schwarz inequality.

In the second case $\eta(\boldsymbol{f}^{(2)}) < \eta(\boldsymbol{f}^{(1)})$, we have

$$\begin{aligned}
&\eta(\boldsymbol{f}^{(1)}) - \eta(\boldsymbol{f}^{(2)}) \\
&= \sum_{h(f_i^{(1)}) \le h(f_i^{(2)})} \left(h(f_i^{(1)}) - h(f_i^{(2)})\right) \Delta v_i + \sum_{h(f_i^{(1)}) > h(f_i^{(2)})} \left(h(f_i^{(1)}) - h(f_i^{(2)})\right) \Delta v_i \\
&\le \sum_{h(f_i^{(1)}) > h(f_i^{(2)})} \left(h(f_i^{(1)}) - h(f_i^{(2)})\right) \Delta v_i \\
&\le M \sum_{h(f_i^{(1)}) > h(f_i^{(2)})} \left(\left|(f_i^{(1)} - 1)(f_i^{(1)} - f_i^{(2)})\right| + \left|(f_i^{(2)} - 1)(f_i^{(1)} - f_i^{(2)})\right|\right) \Delta v_i,
\end{aligned}$$

where the last inequality is again the result of Lemma 3.3. The lemma naturally follows by extending the range of summation of i to $1, \dots, N$ and applying the Cauchy–Schwarz inequality. ∎

In the proof of case 2, we applied Lemma 3.3 only to $f_i^{(1)}$ and $f_i^{(2)}$ with $h(f_i^{(1)}) > h(f_i^{(2)})$. This allows us to relax the condition "$f_i^{(1)} \ge C_0$ for all $1 \le i \le N$" in the case $\eta(\boldsymbol{f}^{(1)}) > \eta(\boldsymbol{f}^{(2)})$. In fact, we need $f_i^{(1)} > C_0$ only for the components that require Lemma 3.3. We write this result in the following corollary.

COROLLARY 3.5. *For* $\boldsymbol{f}^{(1)} = (f_1^{(1)}, \dots, f_N^{(1)}) \in \mathbb{R}_+^N$ *and* $\boldsymbol{f}^{(2)} = (f_1^{(2)}, \dots, f_N^{(2)}) \in \mathbb{R}_+^N$ *with* $\|\boldsymbol{f}^{(1)}\|_1 = \|\boldsymbol{f}^{(2)}\|_1 = V$, *we assume* $\eta(\boldsymbol{f}^{(1)}) > \eta(\boldsymbol{f}^{(2)})$. *If there exists* $0 < C_0 < 1$ *such that for any* $i = 1, \dots, N$, *either* $f_i^{(1)} \ge C_0$ *or* $h(f_i^{(1)}) \le h(f_i^{(2)})$ *is satisfied, then it holds that*

$$\left|\eta(\boldsymbol{f}^{(1)}) - \eta(\boldsymbol{f}^{(2)})\right| \le \max\left(2, 2|\log(C_0)|\right) \|\boldsymbol{f}^{(1)} - \boldsymbol{f}^{(2)}\|_2 \left(\|\boldsymbol{f}^{(1)} - \mathbf{1}\|_2 + \|\boldsymbol{f}^{(2)} - \mathbf{1}\|_2\right).$$

We are now ready to prove Theorem 2.3.

Proof of Theorem 2.3. The convexity of $\eta(\cdot)$ implies

$$\eta(\boldsymbol{f}^{n+1} + \beta(\mathbf{1} - \boldsymbol{f}^{n+1})) \le \beta\eta(\mathbf{1}) + (1 - \beta)\eta(\boldsymbol{f}^{n+1}).$$

By (2.7) with $\eta(\mathbf{1}) = 0$, the above inequality is equivalent to

$$\beta \le \frac{\eta(\boldsymbol{f}^{n+1}) - \eta(\boldsymbol{f}(t_{n+1}))}{\eta(\boldsymbol{f}^{n+1})}. \tag{3.6}$$

The numerator in (3.6) can be estimated by

$$\begin{aligned}
&\eta(\boldsymbol{f}^{n+1}) - \eta(\boldsymbol{f}(t_{n+1})) \\
&\quad \le M\|\boldsymbol{f}^{n+1} - \boldsymbol{f}(t_{n+1})\|_2 \left(\|\boldsymbol{f}^{n+1} - \mathbf{1}\|_2 + \|\boldsymbol{f}(t_{n+1}) - \mathbf{1}\|_2\right) \qquad \text{(Lemma 3.4)} \\
&\quad \le M\|\boldsymbol{f}^{n+1} - \boldsymbol{f}(t_{n+1})\|_2 \left(1 + \sqrt{\|\boldsymbol{f}(t_{n+1})\|_\infty}\right) \|\boldsymbol{f}^{n+1} - \mathbf{1}\|_2, \qquad \text{(Lemma 3.2)}
\end{aligned}$$

where $M = \max(2, 2|\log(C_0)|)$. On the other hand, according to Lemma 3.1, the denominator in (3.6) satisfies

$$\eta(\boldsymbol{f}^{n+1}) \ge \frac{1}{2\|\boldsymbol{f}^{n+1}\|_\infty} \|\boldsymbol{f}^{n+1} - \mathbf{1}\|_2^2.$$

Therefore,

$$\|\beta(\mathbf{1} - \boldsymbol{f}^{n+1})\|_2 \le 2M\|\boldsymbol{f}^{n+1}\|_\infty (1 + \sqrt{\|\boldsymbol{f}(t_{n+1})\|_\infty}) \|\boldsymbol{f}^{n+1} - \boldsymbol{f}(t_{n+1})\|_2. \qquad \square$$

In this case, we would like to give a remark on the practical choice of β_p in (2.4). Instead of solving $\eta(\boldsymbol{f}^{n+1} + \beta_p(\mathbf{1} - \boldsymbol{f}^{n+1})) = \eta(\boldsymbol{f}^n)$, we can simply take $\hat{\beta}_p = (\eta(\boldsymbol{f}^{n+1}) - \eta(\boldsymbol{f}^n))/\eta(\boldsymbol{f}^{n+1})$, which equals the upper bound in (3.6). Note that the convexity of function $\eta(\cdot)$ implies $\eta(\boldsymbol{f}^n) = (1-\hat{\beta}_p)\eta(\boldsymbol{f}^{n+1}) + \hat{\beta}_p\eta(\mathbf{1}) \ge \eta(\boldsymbol{f}^{n+1} + \hat{\beta}_p(\mathbf{1} - \boldsymbol{f}^{n+1}))$. Therefore, under the condition of Theorem 2.3, if we change the numerical solution at $(n+1)$th step to $\boldsymbol{f}^{n+1} + \hat{\beta}_p(\mathbf{1} - \boldsymbol{f}^{n+1})$, it still holds that $\|\hat{\beta}_p(\mathbf{1} - \boldsymbol{f}^{n+1})\|_2 \le M\|\boldsymbol{f}(t_{n+1}) - \boldsymbol{f}^{n+1}\|_2$.

3.2. Proof of Theorem 2.2. Different from the previous proof, in Theorem 2.2, we allow the solution to have components arbitrarily close to zero, so that Lemma 3.4 cannot be directly applied. To overcome this difficulty, we introduce a regularization term before using Lemma 3.4. The details are given as follows.

Proof of Theorem 2.2. For simplicity, we let $\varepsilon = \|\boldsymbol{f}^{n+1} - \boldsymbol{f}(t_{n+1})\|_2$. To avoid dealing with zero components, we first regularize the numerical solution $\boldsymbol{f}^{n+1}$ by

$$\boldsymbol{f}^{n+1,1} = \boldsymbol{f}^{n+1} + \varepsilon(\mathbf{1} - \boldsymbol{f}^{n+1}), \tag{3.7}$$

after which $f_i^{n+1,1} \ge \varepsilon$ for all $i = 1, \ldots, N$. On the other hand, since $\|\mathbf{1} - \boldsymbol{f}^{n+1}\|_\infty \le \max(1, \|\boldsymbol{f}\|_\infty - 1) \le \|\boldsymbol{f}\|_\infty$, the L^2 norm of the perturbation introduced by the regularization satisfies

$$\|\varepsilon(\mathbf{1} - \boldsymbol{f}^{n+1})\|_2 \le \sqrt{V}\|\boldsymbol{f}^{n+1}\|_\infty \varepsilon.$$

After perturbation, if $\eta(\boldsymbol{f}^{n+1,1}) < \eta(\boldsymbol{f}(t_{n+1}))$, then we have $\beta < \varepsilon$ so that the conclusion of the theorem is drawn. If $\eta(\boldsymbol{f}^{n+1,1}) > \eta(\boldsymbol{f}(t_{n+1}))$, we can find $\beta_2 \in (0, 1]$ such that

$$\eta\left(\boldsymbol{f}^{n+1,1} + \beta_2(\mathbf{1} - \boldsymbol{f}^{n+1,1})\right) = \eta(\boldsymbol{f}(t_{n+1})),$$

which is identical to (2.7) by replacing $\boldsymbol{f}^{n+1}$ to $\boldsymbol{f}^{n+1,1}$. Therefore, we can set $C_0 = \varepsilon$ in Theorem 2.3 to obtain

$$\|\beta_2(\mathbf{1} - \boldsymbol{f}^{n+1,1})\|_2 \le M_1\|\boldsymbol{f}^{n+1,1} - \boldsymbol{f}(t_{n+1})\|_2,$$

and by the proof of Theorem 2.3, we know that

$$\begin{aligned} M_1 &= 4\max(1, |\log\varepsilon|)\,\|\boldsymbol{f}^{n+1,1}\|_\infty(1+\sqrt{\|\boldsymbol{f}(t_{n+1})\|_\infty}) \\ &\le 4(1+|\log\varepsilon|)\|\boldsymbol{f}^{n+1}\|_\infty(1+\sqrt{\|\boldsymbol{f}(t_{n+1})\|_\infty}), \end{aligned}$$

since $\|\boldsymbol{f}^{n+1,1}\|_\infty \le \|\boldsymbol{f}^{n+1}\|_\infty$.

If we define

$$\hat{\boldsymbol{f}}^{n+1} = \boldsymbol{f}^{n+1,1} + \beta_2(\mathbf{1}-\boldsymbol{f}^{n+1,1}) = \boldsymbol{f}^{n+1} + (\varepsilon+\beta_2-\varepsilon\beta_2)(\mathbf{1}-\boldsymbol{f}^{n+1}), \tag{3.8}$$

then by $\eta(\hat{\boldsymbol{f}}^{n+1}) = \eta(\boldsymbol{f}(t_{n+1}))$ we know that $\beta = \varepsilon+\beta_2-\varepsilon\beta_2$. Thus it holds that

$$\begin{aligned} \|\beta(\mathbf{1}-\boldsymbol{f}^{n+1})\|_2 &= \|\hat{\boldsymbol{f}}^{n+1}-\boldsymbol{f}^{n+1}\|_2 \le \|\hat{\boldsymbol{f}}^{n+1}-\boldsymbol{f}^{n+1,1}\|_2 + \|\boldsymbol{f}^{n+1,1}-\boldsymbol{f}^{n+1}\|_2 \\ &\le M_1\|\boldsymbol{f}^{n+1,1}-\boldsymbol{f}(t_{n+1})\|_2 + \|\boldsymbol{f}^{n+1,1}-\boldsymbol{f}^{n+1}\|_2 \\ &\le M_1(\|\boldsymbol{f}^{n+1,1}-\boldsymbol{f}^{n+1}\|_2+\varepsilon) + \|\boldsymbol{f}^{n+1,1}-\boldsymbol{f}^{n+1}\|_2 \\ &\le M_1(\sqrt{V}\|\boldsymbol{f}^{n+1}\|_\infty\varepsilon+\varepsilon) + \sqrt{V}\|\boldsymbol{f}^{n+1}\|_\infty\varepsilon \le M_2\varepsilon(|\log\varepsilon|+1), \end{aligned}$$

where $M_2 = 8(\sqrt{V}\|\boldsymbol{f}^{n+1}\|_\infty+1)\|\boldsymbol{f}^{n+1}\|_\infty(1+\sqrt{\|\boldsymbol{f}(t_{n+1})\|_\infty})$. ∎

The proof of this theorem follows the two-step procedure, which will also be applied in the proof of Theorem 2.4.

3.3. Proof of Theorem 2.4. To prove Theorem 2.4, we deal with the components with $f_i^{n+1} < \frac{2}{3}$ and $f_i^{n+1} > \frac{2}{3}$ separately. The difference between these two cases can be seen from the following lemma:

LEMMA 3.6. *For $\boldsymbol{f}^{(1)} \in \mathbb{R}_+^N$ and $\boldsymbol{f}^{(2)} \in \mathbb{R}_+^N$ with $\|\boldsymbol{f}^{(1)}-\boldsymbol{f}^{(2)}\|_\infty \le \frac{1}{3}$, define*

$$\boldsymbol{f}^{(3)} = \boldsymbol{f}^{(1)} + \beta_1(\mathbf{1}-\boldsymbol{f}^{(1)}),$$

where $\beta_1 = 3\|\boldsymbol{f}^{(1)}-\boldsymbol{f}^{(2)}\|_\infty$. If $\|\boldsymbol{f}^{(1)}\|_1 = V$, then $\boldsymbol{f}^{(3)}$ satisfies following properties:

1. *For all k such that $f_k^{(1)} < \frac{2}{3}$, it holds that $h(f_k^{(3)}) \le h(f_k^{(2)})$.*
2. *For all k such that $f_k^{(1)} \ge \frac{2}{3}$, it holds that $f_k^{(3)} \ge \frac{2}{3}$.*
3. *$\|\boldsymbol{f}^{(3)}-\boldsymbol{f}^{(1)}\|_\infty \le 3\|\boldsymbol{f}^{(1)}\|_\infty\|\boldsymbol{f}^{(1)}-\boldsymbol{f}^{(2)}\|_\infty$.*

Proof. For those k such that $f_k^{(1)} < \frac{2}{3}$, we have $1-f_k^{(1)} \ge \frac{1}{3}$. Thus

$$f_k^{(3)} - f_k^{(1)} = \beta_1(1-f_k^{(1)}) \ge 3\left|f_k^{(1)}-f_k^{(2)}\right|\cdot(1-f_k^{(1)}) \ge \left|f_k^{(1)}-f_k^{(2)}\right| \ge f_k^{(2)}-f_k^{(1)},$$

which yields $f_k^{(3)} \ge f_k^{(2)}$. Since $f_k^{(3)}$ is the convex combination of 1 and $f_k^{(1)}$, we have $0 \le f_k^{(3)} \le 1$. Since $h(\cdot)$ is monotonically decreasing on $[0,1]$, we conclude that $h(f_k^{(3)}) \le h(f_k^{(2)})$.

The second property is obvious since $f_k^{(3)}$ lies between $f_k^{(1)}$ and 1.

As for the third property, it should be noted that $\|\boldsymbol{f}^{(1)}\|_1 = V$ implies $\|\boldsymbol{f}^{(1)}\|_\infty \ge 1$. Therefore,

$$\begin{aligned} \|\boldsymbol{f}^{(3)}-\boldsymbol{f}^{(1)}\|_\infty &= \beta_1\|\mathbf{1}-\boldsymbol{f}^{(1)}\|_\infty \\ &\le \max(1, \|\boldsymbol{f}^{(1)}\|_\infty-1)\beta_1 \le 3\|\boldsymbol{f}^{(1)}\|_\infty\|\boldsymbol{f}^{(1)}-\boldsymbol{f}^{(2)}\|_\infty. \end{aligned}$$ ∎

The first property in Lemma 3.6 shows how we deal with the small components, and this only holds when β_1 is proportional to the difference between $\boldsymbol{f}^{(1)}$ and $\boldsymbol{f}^{(2)}$ measured by the infinity norm, leading to the form of the right-hand side in the conclusion of Theorem 2.4. For the remaining terms, an $O(1)$ lower bound exists, so that the same technique as Theorem 2.3 can be applied. The details of the proof are given below.

Proof of Theorem 2.4. By Lemma 3.6, we could pick $\beta_1 = 3\|\boldsymbol{f}^{n+1} - \boldsymbol{f}(t_{n+1})\|_\infty$ and construct

$$\boldsymbol{f}^{n+1,1} = \boldsymbol{f}^{n+1} + \beta_1(\mathbf{1} - \boldsymbol{f}^{n+1}). \tag{3.9}$$

If $\eta(\boldsymbol{f}^{n+1,1}) \le \eta(\boldsymbol{f}(t_{n+1}))$, the proof is already completed. If $\eta(\boldsymbol{f}^{n+1,1}) > \eta(\boldsymbol{f}(t_{n+1}))$, we construct $\hat{\boldsymbol{f}}^{n+1}$ as (3.8) such that $\eta(\hat{\boldsymbol{f}}^{n+1}) = \eta(\boldsymbol{f}(t_{n+1}))$, and thus $\beta = \beta_1 + \beta_2 - \beta_1\beta_2$. According to Lemma 3.6, those components i where $h(f_i^{n+1,1}) > h(f_i(t_{n+1}))$ satisfy $f_i^{n+1,1} \ge \frac{2}{3}$. Therefore, Corollary 3.5 could be applied with $C_0 = \frac{2}{3}$, and we could mimic the proof of Theorem 2.3, replacing only Lemma 3.4 with Corollary 3.5 in the proof. As a result, by the conclusion of Theorem 2.3, it holds that

$$\|\beta_2(\mathbf{1} - \boldsymbol{f}^{n+1,1})\|_2 \le M_1\|\boldsymbol{f}(t_{n+1}) - \boldsymbol{f}^{n+1,1}\|_2,$$

where $M_1 = 4\max(1, |\log(C_0)|)\,\|\boldsymbol{f}^{n+1,1}\|_\infty(1 + \sqrt{\|\boldsymbol{f}(t_{n+1})\|_\infty})$ taken from the proof of Theorem 2.3. Moreover, $\boldsymbol{f}^{n+1,1}$ in M_1 could be replaced by $\boldsymbol{f}^{n+1}$ since $\|\boldsymbol{f}^{n+1,1}\|_\infty \le \|\boldsymbol{f}^{n+1}\|_\infty$. Then, similar to the second step in the proof of Theorem 2.2, it holds that

$$\begin{aligned}
\|\hat{\boldsymbol{f}}^{n+1} - \boldsymbol{f}^{n+1}\|_2 &\le \|\hat{\boldsymbol{f}}^{n+1} - \boldsymbol{f}^{n+1,1}\|_2 + \|\boldsymbol{f}^{n+1,1} - \boldsymbol{f}^{n+1}\|_2 \\
&\le M_1\|\boldsymbol{f}(t_{n+1}) - \boldsymbol{f}^{n+1,1}\|_2 + \|\boldsymbol{f}^{n+1,1} - \boldsymbol{f}^{n+1}\|_2 \\
&\le M_1\sqrt{V}\|\boldsymbol{f}(t_{n+1}) - \boldsymbol{f}^{n+1,1}\|_\infty + \sqrt{V}\|\boldsymbol{f}^{n+1,1} - \boldsymbol{f}^{n+1}\|_\infty \\
&\le M_1\sqrt{V}\|\boldsymbol{f}(t_{n+1}) - \boldsymbol{f}^{n+1}\|_\infty + \sqrt{V}(M_1+1)\|\boldsymbol{f}^{n+1,1} - \boldsymbol{f}^{n+1}\|_\infty \\
&\le \left(M_1\sqrt{V} + 3\sqrt{V}(M_1+1)\|\boldsymbol{f}^{n+1}\|_\infty\right)\|\boldsymbol{f}(t_{n+1}) - \boldsymbol{f}^{n+1}\|_\infty,
\end{aligned}$$

where the last "$\le$" is the result of Lemma 3.6. This completes the proof since $\|\beta(\mathbf{1} - \boldsymbol{f}^{n+1})\|_2 = \|\hat{\boldsymbol{f}}^{n+1} - \boldsymbol{f}^{n+1}\|_2$. ∎

3.4. Proof of Theorem 2.5. In this subsection, we will prove Theorem 2.5. Before that, we would like to introduce two lemmas. Lemma 3.7 comes from optimization, which illustrates the infinity norm of optimal solution could be bounded by the L^2 norm of it. Based on Lemma 3.7, we make a decomposition of the (relative) entropy function in (3.21) and then introduce Lemma 3.11 to estimate the difference of decomposed entropy functions.

As assumed in the theorem, we suppose all the components of $\boldsymbol{f}$ are sorted in the ascending order:

$$f_1 \le f_2 \le \cdots \le f_N = \|\boldsymbol{f}\|_\infty.$$

Note that this does not affect the definition of entropy and the numerical scheme for the entropy fix.

LEMMA 3.7. *For any $C_1, C_f \in (0,1]$ and positive integer N, let $I_1 = \min\{I \mid \sum_{i=1}^{I}\Delta v_i \ge C_1 V\}$. If $\boldsymbol{f} \in \mathbb{R}_+^N$ satisfies*

$$f_i \le 1/2 \quad \forall i = 1, \ldots, I_1 \qquad \text{and} \qquad \frac{1}{|\log f_1|} \ge \frac{C_f}{|\log f_{I_1}|},$$

then when $\varepsilon < \frac{1}{2}\sqrt{C_1 V}$*, the solution* $\boldsymbol{g}^* = (g_1^*, \dots, g_{I_1}^*)^T \in \mathbb{R}^{I_1}$ *of the optimization problem*

$$\operatorname*{argmin}_{g_1,\dots,g_{I_1}} \sum_{i=1}^{I_1} h(f_i + g_i)\Delta v_i \qquad s.t. \ \sum_{i=1}^{I_1} g_i^2 \Delta v_i \le \varepsilon^2 \tag{3.10}$$

satisfies $0 \le g_{I_1}^* \le \cdots \le g_1^* \le (\sqrt{C_1 V} C_f)^{-1}\varepsilon$ *and* $C_f \le g_{I_1}^*/g_1^* \le 1$.

Proof. The proof utilizes the Karush–Kuhn–Tucker sufficient conditions for optimization problems [19, Chapter 3.5]. It is easy to verify that both the objective function and the constraint are continuously differentiable convex functions with respect to $(g_1, \dots, g_{I_1})^T$. Therefore, if the following conditions hold for $\lambda^* \in \mathbb{R}$ and $\boldsymbol{g}^* = (g_1^*, \dots, g_{I_1}^*)^T$,

$$\begin{cases} h'(f_i + g_i^*) + 2\lambda^* g_i^* = 0 \qquad \forall\ 1 \le i \le I_1, \\ \displaystyle\sum_{i=1}^{I_1} (g_i^*)^2 \Delta v_i \le \varepsilon^2, \\ \lambda^* \ge 0, \\ \displaystyle\lambda^* \left(\sum_{i=1}^{I_1} (g_i^*)^2 \Delta v_i - \varepsilon^2 \right) = 0, \end{cases} \tag{3.11}$$

then $\boldsymbol{g}^*$ is the global minimum of the optimization problem.

First, we claim that $\lambda^* \neq 0$, so that

$$\sum_{i=1}^{I_1} (g_i^*)^2 \Delta v_i = \varepsilon^2 \tag{3.12}$$

due to the last equation in (3.11). If λ^* equals 0, then $h'(f_i + g_i^*) = 0$, which yields $g_i^* = 1 - f_i \ge \frac{1}{2}$. Therefore,

$$\sum_{i=1}^{I_1} (g_i^*)^2 \Delta v_i \ge \frac{\sum_{i=1}^{I_1} \Delta v_i}{4} \ge \frac{C_1 V}{4} > \varepsilon^2,$$

which contradicts with the second inequality in (3.11).

Now we would like to establish the existence and uniqueness of the solution. We first focus on the first equation in (3.11). For any $1 \le i \le I_1$ and fixed $\lambda^* > 0$, there exists one unique $g_i^* \in (0, 1)$ satisfying $h'(f_i + g_i^*) + 2\lambda^* g_i^* = 0$. This is because the function $\zeta_i(x) := h'(f_i + x) + 2\lambda^* x$ is monotonically increasing, and

$$\zeta(0) = \log f_i \le \log\left(\frac{1}{2}\right) < 0, \qquad \zeta(1) \ge 2\lambda^* > 0.$$

Thus it remains to demonstrate that λ^* is unique. Inspired by the first equation in (3.11), we define

$$\sigma(x) = -\frac{h'(f_i + x)}{2x} = -\frac{\log(f_i + x)}{2x}, \qquad x \in (0, 1 - f_i].$$

Then its inverse function $\sigma_i^{-1}(y)$ satisfies

$$y = -\frac{\log(f_i + \sigma_i^{-1}(y))}{2\sigma_i^{-1}(y)} \quad \text{and} \quad \sigma_i^{-1}(y) = \frac{W_0(2ye^{2yf_i}) - 2yf_i}{2y}, \tag{3.13}$$

where $W_0(\cdot)$ is the Lambert W function [9] satisfying $W_0(x)e^{W_0(x)} = x$. For $\sigma_i(\cdot)$ and $\sigma_i^{-1}(\cdot)$, we have the following properties:

1. $\sigma_i(x)$ is monotonically decreasing, and so is $\sigma_i^{-1}(x)$ (this requires $f_i \le \frac{1}{2}$);
2. $\sigma_i(g_i^*) = \lambda^*$ and $g_i^* = \sigma_i^{-1}(\lambda^*)$;
3. $\sigma_i^{-1}(0) = 1 - f_i \ge \frac{1}{2}$ and $\sigma_i^{-1}(y) \to 0$ as $y \to +\infty$.

Here the limit of $\sigma_i^{-1}(y)$ at $+\infty$ can be obtained by the inequality (see [12])

$$W_0(x) \le \log(x) - \log(\log(x)) + \frac{e}{e-1}\frac{\log(\log(x))}{\log(x)} \qquad \forall x \ge e.$$

Furthermore, if we define

$$\Xi(y) = \sum_{i=1}^{I_1} [\sigma_i^{-1}(y)]^2 \Delta v_i, \qquad y \in [0, +\infty),$$

then by the three properties of σ_i, we have

1. $\Xi(y)$ is a decreasing function since each $\sigma_i^{-1}(y)$ is monotonically decreasing;
2. $\Xi(\lambda^*) = \varepsilon^2$ according to (3.12);
3. $\Xi(0) \ge \frac{1}{4}\sum_{i=1}^{I_1} \Delta v_i > \varepsilon^2$, and $\Xi(y) \to 0$ as $y \to +\infty$.

These properties show the existence and uniqueness of λ^*.

Next, we will show $g_{I_1}^* \le \cdots \le g_1^*$. For any $1 \le i \le j \le I_1$, $f_i \le f_j$ implies

$$\sigma_j(g_j^*) = \lambda^* = \sigma_i(g_i^*) \ge \sigma_j(g_i^*).$$

Using the fact that $\sigma_j(\cdot)$ is decreasing, we see that $g_j^* \le g_i^*$. To get the bound of g_1^*, we need the following two results:

- By (3.13), we have

$$\lim_{y\to+\infty} \frac{\sigma_i^{-1}(y)}{\sigma_1^{-1}(y)} = \lim_{y\to+\infty} \frac{\log(f_i + \sigma_i^{-1}(y))}{\log(f_1 + \sigma_1^{-1}(y))} = \frac{\log(f_i)}{\log(f_1)} \ge C_f.$$

- By straightforward calculation, we have

$$\frac{\mathrm{d}}{\mathrm{d}y}\left(\frac{\sigma_i^{-1}(y)}{\sigma_1^{-1}(y)}\right) = \frac{W_0(2ye^{2yf_1}) - W_0(2ye^{2yf_i})}{y(1+W_0(2ye^{2yf_1}))(1+W_0(2ye^{2yf_i}))}\frac{\sigma_i^{-1}(y)}{\sigma_1^{-1}(y)} \le 0.$$

These results indicate that

$$\frac{g_i^*}{g_1^*} = \frac{\sigma_i^{-1}(\lambda^*)}{\sigma_1^{-1}(\lambda^*)} \ge C_f,$$

and thus

$$g_1^* = \varepsilon\left(\sum_{i=1}^{I_1} \frac{(g_i^*)^2}{(g_1^*)^2}\Delta v_i\right)^{-1/2} \le \varepsilon\left(\sum_{i=1}^{I_1} C_f^2 \Delta v_i\right)^{-1/2} \le \frac{\varepsilon}{\sqrt{C_1 V C_f}}.$$

This completes the proof. ∎

One corollary of the above lemma is the extension to a continuous version, with identical optimal solution $\boldsymbol{g}^*$ in the sense of a piecewise constant function. For the ease of this extension, we would like to introduce the (partial) sum of first i parameters Δv_i as

$$S_0 = 0, \qquad S_i = \sum_{j=1}^{i} \Delta v_j, \qquad i = 1, \ldots, N. \tag{3.14}$$

Then we have the following corollary.

COROLLARY 3.8. *Under the condition of Lemma* 3.7*, if a piecewise constant function defined on* $(0, S_{I_1}]$ *is introduced as*

$$f(v) = f_i, \qquad v \in (S_{i-1}, S_i], \quad i = 1, \dots, I_1,$$

then the solution $g^*(v) \in L^2((0, S_{I_1}])$ *of the following problem*

$$\operatorname*{argmin}_{g \in L^2((0,S_{I_1}])} \int_0^{S_{I_1}} h(f(v) + g(v)) \mathrm{d}v \qquad s.t.\ \|g\|_2^2 := \int_0^{S_{I_1}} (g(v))^2 \mathrm{d}v \le \varepsilon^2 \tag{3.15}$$

is equal to a piecewise constant function a.e. as

$$g^*(v) = g_i^*, \qquad v \in (S_{i-1}, S_i], \quad i = 1, \dots, I_1,$$

where g_i^* *is the component of the optimal solution* $\boldsymbol{g}^*$ *in Lemma* 3.7*.*

Proof. To prove the corollary, it suffices to show that for every $i = 1, \dots, I_1$, the function $g^*(v)$ is a constant on $(S_{i-1}, S_i]$ except for a set with measure zero, so that the optimization problem (3.15) is essentially equivalent to (3.10). Suppose that $g^*(v)$ is essentially not a constant on $(S_{i-1}, S_i]$ for some i. We define the function $\hat{g}(v)$ by

$$\hat{g}(v) = \begin{cases} \dfrac{1}{\Delta v_i} \displaystyle\int_{S_{i-1}}^{S_i} g^*(v)\, \mathrm{d}v & \text{if } v \in (S_{i-1}, S_i], \\ g^*(v) & \text{otherwise.} \end{cases}$$

By Hölder's inequality (on $(S_{i-1}, S_i]$), it is easy to find $\|\hat{g}\|_2^2 \le \|g^*\|_2^2 \le \varepsilon^2$. Moreover, using Jensen's inequality on the convex function $h(f_i + \cdot)$, we obtain

$$\int_{S_{i-1}}^{S_i} h(f(v) + \hat{g}(v))\, \mathrm{d}v = \Delta v_i h(f_i + \hat{g}(v)) \le \int_{S_{i-1}}^{S_i} h(f_i + g^*(v))\, \mathrm{d}v. \tag{3.16}$$

Note that $g^*(\cdot)$ is the optimal solution, implying that the equality must hold for (3.16). However, since $h(f_i + \cdot)$ is strictly convex, the equality holds only when $g^*(v)$ is a constant on $(S_{i-1}, S_i]$, which contradicts our assumption. This completes the proof of the corollary. ∎

Another important corollary of Lemma 3.7 is to pick $\beta_1 = O(\|\boldsymbol{f}^{n+1} - \boldsymbol{f}(t_{n+1})\|_2)$ and construct $\boldsymbol{f}^{n+1,1}$ following (3.9) such that the entropy of $\boldsymbol{f}^{n+1,1}$ is less than the entropy of $\boldsymbol{f}(t_{n+1})$ in the range of $i \le I_1$.

COROLLARY 3.9. *Let* $\varepsilon := \|\boldsymbol{f}^{n+1} - \boldsymbol{f}(t_{n+1})\|_2$*. Suppose* $\boldsymbol{f}^{n+1}$ *satisfies the condition of Lemma* 3.7 *and* $\varepsilon < \frac{\sqrt{C_1 V} C_f}{2}$*. Let* $\beta_1 = \frac{2\varepsilon}{\sqrt{C_1 V} C_f}$ *and*

$$\boldsymbol{f}^{n+1,1} = \boldsymbol{f}^{n+1} + \beta_1(\mathbf{1} - \boldsymbol{f}^{n+1}). \tag{3.17}$$

Then $\boldsymbol{f}^{n+1,1}$ *satisfies*

$$\sum_{i=1}^{I_1} h(f_i^{n+1,1}) \Delta v_i \le \sum_{i=1}^{I_1} h(f_i(t_{n+1})) \Delta v_i. \tag{3.18}$$

Proof. Let $g_1^*, \dots, g_{I_1}^*$ be the solution of the optimization problem (3.10). Since

$$\sum_{i=1}^{I_1} (f_i^{n+1} - f_i(t_{n+1}))^2 \Delta v_i \le \|\boldsymbol{f}^{n+1} - \boldsymbol{f}(t_{n+1})\|_2^2 = \varepsilon^2,$$

it holds that

$$\sum_{i=1}^{I_1} h(f_i^{n+1} + g_i^*)\Delta v_i \le \sum_{i=1}^{I_1} h(f_i(t_{n+1}))\Delta v_i.$$

To prove (3.18), it suffices to show

$$\sum_{i=1}^{I_1} h(f_i^{n+1,1})\Delta v_i \le \sum_{i=1}^{I_1} h(f_i^{n+1} + g_i^*)\Delta v_i. \tag{3.19}$$

By the conclusion of Lemma 3.7,

$$f_i^{n+1,1} = f_i^{n+1} + \beta_1(1 - f_i^{n+1}) \ge f_i^{n+1} + \frac{\beta_1}{2} = f_i^{n+1} + \frac{\varepsilon}{\sqrt{C_1 V C_f}} \ge f_i^{n+1} + g_i^*$$

for all $1 \le i \le I_1$. Noticing that $\beta_1 < 1$ by the constraint $\varepsilon < \frac{\sqrt{C_1 V C_f}}{2}$, we obtain $f_i^{n+1,1} < 1$. Hence, the monotonicity of $h(\cdot)$ yields

$$h(f_i^{n+1,1}) \le h(f_i^{n+1} + g_i^*) \qquad \forall i = 1, \dots, I_1.$$

Multiplying Δv_i and summing up the above inequalities for i yield (3.19). ∎

By Corollary 3.9, we have performed our first step that reduces the entropy of the smallest part of $\boldsymbol{f}^{n+1}$ (from f_1^{n+1} to $f_{I_1}^{n+1}$) below the entropy of the exact solution in the same section. If the smallest component beyond this section $f_{I_1+1}^{n+1}$ already has the magnitude $O(1)$, for instance, $f_{I_1+1}^{n+1} \ge \frac{1}{2}$, then the remaining part can be processed using the same technique as in Theorem 2.2 and Theorem 2.4. Therefore, below we will only focus on the case where $f_{I_1+1}^{n+1} < 1/2$, and this inspires us to further decompose the remaining components into two parts by introducing I_2 such that

$$f_{I_2}^{n+1} \le \frac{1}{2}, \qquad f_{I_2+1}^{n+1} > \frac{1}{2}. \tag{3.20}$$

Then we will have $\eta(\boldsymbol{f}) - V = H_1(\boldsymbol{f}) + H_2(\boldsymbol{f}) + H_3(\boldsymbol{f})$ for any $\boldsymbol{f} \in \mathbb{R}_+^N$, where

(3.21)

$$H_1(\boldsymbol{f}) = \sum_{i=1}^{I_1} h(f_i)\Delta v_i, \quad H_2(\boldsymbol{f}) = \sum_{i=I_1+1}^{I_2} h(f_i)\Delta v_i, \quad H_3(\boldsymbol{f}) = \sum_{i=I_2+1}^{N} h(f_i)\Delta v_i.$$

Note that this decomposition also includes the case $f_{I_1+1}^{n+1} \ge \frac{1}{2}$, for which we can choose $I_2 = I_1$, so that $H_2(\boldsymbol{f}^{n+1}) = 0$.

Lemma 3.11 will show some properties of above decomposition. Before that, a quotient $F(x, y, C)$, which will be used in the proof of Lemma 3.11, is introduced as

$$F(x, y, C) = \frac{h(x + y) - h(x + Cy)}{h(x) - h(x + y)}, \tag{3.22}$$

where $0 \le x \le 1/2$, $C > 1$ and $0 \le y \le 1/(2C)$. It is easy to find $F(x, y, C) \ge 0$ in its domain of definition. Furthermore, the following lemma gives the positive lower bound of $F(x, y, C)$ for fixed C, where the proof utilizes the (partial) derivatives of $F(x, y, C)$, and its detail is left in supplementary material SM1.

LEMMA 3.10. *For any $C_1 \in (0,1]$, there exists $C_2 > 1$ depending on C_1 such that $F(x,y,C_2)$ given in* (3.22) *satisfies*

$$F(x,y,C_2) \geq \frac{1}{C_1} \qquad \forall 0 \leq x \leq 1/2,\ 0 \leq y \leq \frac{1}{2C_2}.$$

LEMMA 3.11. *Under the condition of Corollary* 3.9 *and the decomposition of* (3.21)*, the following properties are satisfied:*

1. $H_2(\boldsymbol{f}^{n+1,1}) - H_2(\boldsymbol{f}(t_{n+1})) \leq \frac{1}{C_1}(H_1(\boldsymbol{f}^{n+1}) - H_1(\boldsymbol{f}^{n+1,1}))$.
2. *There exists a constant $M_1 > 1$ depending on C_1 such that when $\varepsilon \leq \frac{\sqrt{C_1 V} C_f}{2M_1}$, the vector*

$$\boldsymbol{f}^{n+1,2} = \boldsymbol{f}^{n+1,1} + M_1\beta_1(\mathbf{1} - \boldsymbol{f}^{n+1,1}) \tag{3.23}$$

satisfies $H_1(\boldsymbol{f}^{n+1,1}) - H_1(\boldsymbol{f}^{n+1,2}) \geq \frac{1}{C_1}(H_1(\boldsymbol{f}^{n+1}) - H_1(\boldsymbol{f}^{n+1,1}))$.

Proof. To prove the first statement, we use the convexity of $H_2(\cdot)$ to obtain

$$H_2(\boldsymbol{f}^{n+1,1}) = H_2(\boldsymbol{f}^{n+1} + \beta_1(\mathbf{1} - \boldsymbol{f}^{n+1})) \leq \max(H_2(\boldsymbol{f}^{n+1}), H_2(\mathbf{1})) = H_2(\boldsymbol{f}^{n+1}). \tag{3.24}$$

Therefore,

$$\begin{aligned} &H_2(\boldsymbol{f}^{n+1,1}) - H_2(\boldsymbol{f}(t_{n+1})) \\ &\quad = H_2(\boldsymbol{f}^{n+1,1}) - H_2(\boldsymbol{f}^{n+1}) + H_2(\boldsymbol{f}^{n+1}) - H_2(\boldsymbol{f}(t_{n+1})) \\ &\quad \leq H_2(\boldsymbol{f}^{n+1}) - H_2(\boldsymbol{f}(t_{n+1})) \leq H_2(\boldsymbol{f}^{n+1}) - H_2(\boldsymbol{f}^{n+1} + \boldsymbol{g}^{**}), \end{aligned} \tag{3.25}$$

where $\boldsymbol{g}^{**} = (g_1^{**}, \ldots, g_N^{**})^T \in \mathbb{R}^N$ is the solution of following minimization problem:

$$\operatorname*{argmin}_{\|\boldsymbol{g}\|_2 \leq \varepsilon} H_2(\boldsymbol{f}^{n+1} + \boldsymbol{g}).$$

The existence of $\boldsymbol{g}^{**}$ is because $H_2(\boldsymbol{f}^{n+1} + \boldsymbol{g})$ is a continuous function (with respect to $\boldsymbol{g}$) defined on a closed set, and the constraint $\|\boldsymbol{g}\|_2 \leq \varepsilon$ also gives a closed set for $\boldsymbol{g}$. The solution $\boldsymbol{g}^{**}$ satisfies that $g_i^{**} \geq 0$ for all $I_1 < i \leq I_2$, since replacing any negative component of $\boldsymbol{g}$ by zero will lead to a smaller value for the objective function.

For any $i = I_1 + 1, \ldots, I_2$ and $j = 1, \ldots, I_1$, the convexity of $h(\cdot)$ implies

$$h(f_i^{n+1}) - h(f_i^{n+1} + g_i^{**}) \leq h(f_j^{n+1}) - h(f_j^{n+1} + g_i^{**}). \tag{3.26}$$

To extend the above inequality to functions defined on $\mathbb{R}_+$ with support in $[0, V]$, which is convenient for our proof in the following step, we would like to follow the notation in (3.14) and represent $\boldsymbol{f}^{n+1}$ and $\boldsymbol{g}^{**}$ by piecewise constant functions $f^{n+1}(v)$ and $g^{**}(v)$, respectively, as

$$f^{n+1}(v) = f_i^{n+1}, \qquad g^{**}(v) = g_i^{**}, \qquad v \in (S_{i-1}, S_i],\ i = 1, \ldots, I_2,$$

and where both $f^{n+1}(v)$ and $g^{**}(v)$ equal zero if $v > S_{I_2}$. Using the functions $f^{n+1}(v)$ and $g^{**}(v)$, the inequality (3.26) is equivalent to the following: for any $w \in (S_{I_1}, S_{I_2})$ and $v \in (S_0, S_{I_1})$,

$$h(f^{n+1}(w)) - h(f^{n+1}(w) + g^{**}(w)) \leq h(f^{n+1}(v)) - h(f^{n+1}(v) + g^{**}(w)).$$

Since $g^{**}(w) = 0$ for $w \geq S_{I_2}$, the above inequality actually holds for any $w \in (S_{I_1}, +\infty)$. Therefore, we choose $w = v + kS_{I_1}$ with $k \geq 1$ to obtain

(3.27)
$$\begin{aligned}
&H_2(\boldsymbol{f}^{n+1}) - H_2(\boldsymbol{f}^{n+1} + \boldsymbol{g}^{**}) \\
&\quad = \sum_{i=I_1+1}^{I_2} \left(h(f_i^{n+1}) - h(f_i^{n+1} + g_i^{**})\right) \Delta v_i \\
&\quad = \int_{S_{I_1}}^{S_{I_2}} \left(h(f^{n+1}(v)) - h(f^{n+1}(v) + g^{**}(v))\right) \mathrm{d}v \\
&\quad = \sum_{k=1}^{\left\lceil \frac{S_{I_2} - S_{I_1}}{S_{I_1}} \right\rceil} \int_0^{S_{I_1}} \left(h(f^{n+1}(v + kS_{I_1})) - h(f^{n+1}(v + kS_{I_1}) + g^{**}(v + kS_{I_1}))\right) \mathrm{d}v \\
&\quad \leq \sum_{k=1}^{\left\lceil \frac{S_{I_2} - S_{I_1}}{S_{I_1}} \right\rceil} \int_0^{S_{I_1}} \left(h(f^{n+1}(v)) - h(f^{n+1}(v) + g^{**}(v + kS_{I_1}))\right) \mathrm{d}v.
\end{aligned}$$

Since $\|g^{**}\|_2^2 \leq \|\boldsymbol{g}^{**}\|_2^2 \leq \varepsilon^2$, for any $1 \leq k \leq \lceil \frac{S_{I_2} - S_{I_1}}{S_{I_1}} \rceil$, we have

(3.28)
$$\begin{aligned}
&\int_0^{S_{I_1}} h(f^{n+1}(v) + g^{**}(v + kS_{I_1}))\mathrm{d}v \\
&\quad \geq \int_0^{S_{I_1}} h(f^{n+1}(v) + g^{*}(v))\mathrm{d}v \\
&\quad = \sum_{j=1}^{I_1} h(f_j^{n+1} + g_j^*)\Delta v_j \geq \sum_{j=1}^{I_1} h(f_j^{n+1,1})\Delta v_j,
\end{aligned}$$

where $g^*(v)$ and g_i^* stand for the solutions of the optimization problem (3.15) and (3.10), respectively; the equality is the conclusion of Corollary 3.8, and the last "$\geq$" comes from the inequality (3.19). Inserting (3.28) into (3.27) yields

(3.29)
$$\begin{aligned}
H_2(\boldsymbol{f}^{n+1}) - H_2(\boldsymbol{f}^{n+1} + \boldsymbol{g}^{**}) &\leq \sum_{k=1}^{\lceil \frac{S_{I_2} - S_{I_1}}{S_{I_1}} \rceil} \sum_{j=1}^{I_1} \left(h(f_j^{n+1}) - h(f_j^{n+1,1})\right) \Delta v_j \\
&\leq \frac{V}{S_{I_1}} (H_1(\boldsymbol{f}^{n+1}) - H_1(\boldsymbol{f}^{n+1,1})).
\end{aligned}$$

Since the definition of I_1 implies $S_{I_1} \geq C_1 V$, concatenating (3.25) and (3.29) proves the first statement.

The second statement will be proved componentwise. We set $M_1 = 2C_2$, where C_2 is determined by Lemma 3.10 with C_1 being chosen as the constant C_1 appearing in the first statement. Then, for any $1 \leq i \leq I_1$, it holds that

$$f_i^{n+1,1} = f_i^{n+1} + \beta_1(1 - f_i^{n+1}) \leq f_i^{n+1} + \beta_1.$$

Moreover, when $\varepsilon \leq \frac{\sqrt{C_1 V C_f}}{2M_1}$, it could be found that $\beta_1 \leq 1/M_1$ and

$$\begin{aligned}
f_i^{n+1,2} &= f_i^{n+1} + (\beta_1 + M_1\beta_1 - M_1\beta_1^2)(1 - f_i^{n+1}) \\
&\geq f_i^{n+1} + M_1\beta_1(1 - f_i^{n+1}) \geq f_i^{n+1} + C_2\beta_1,
\end{aligned}$$

where we have used $f_i^{n+1} \le \frac{1}{2}$ and $M_1 = 2C_2$. Therefore, the monotonicity of $h(\cdot)$ in the interval of $[0,1]$ implies

$$\frac{h(f_i^{n+1,1}) - h(f_i^{n+1,2})}{h(f_i^{n+1}) - h(f_i^{n+1,1})} \ge \frac{h(f_i^{n+1} + \beta_1) - h(f_i^{n+1} + C_2\beta_1)}{h(f_i^{n+1}) - h(f_i^{n+1} + \beta_1)} = F(f_i^{n+1}, \beta_1, C_2) \ge \frac{1}{C_1},$$

where the function $F(\cdot,\cdot,\cdot)$ is defined in (3.22) and the last inequality is due to Lemma 3.10. By noticing $h(f_i^{n+1}) - h(f_i^{n+1,1}) \ge 0$, the second statement can then be easily derived. ∎

With the preparation of Lemma 3.11, we can start to prove Theorem 2.5.

Proof of Theorem 2.5. If $f_{I_1}^{n+1} \ge \frac{1}{2}$, (2.8) implies $\log(f_1^{n+1}) \ge -\frac{1}{C_f}\log(2)$, which means $f_1^{n+1} \ge 2^{-1/C_f}$. Then, from Theorem 2.3, we get $\|\beta(\mathbf{1} - \boldsymbol{f}^{n+1})\|_2 \le M\|\boldsymbol{f}^{n+1} - \boldsymbol{f}(t_{n+1})\|_2$, where $M > 0$ depends on $2^{-1/C_f}$, $\|\boldsymbol{f}^{n+1}\|_\infty$ and $\|\boldsymbol{f}(t_{n+1})\|_\infty$. This completes the proof.

Otherwise, if $f_{I_1}^{n+1} < \frac{1}{2}$, we would like to introduce I_2 and decompose $\eta(\boldsymbol{f})$ following (3.20) and (3.21). After that, we construct $\boldsymbol{f}^{n+1,1}$ and $\boldsymbol{f}^{n+1,2}$ from (3.9) with $\beta_1 = \|\boldsymbol{f}(t_{n+1}) - \boldsymbol{f}^{n+1}\|_2/(\sqrt{C_1 V} C_f)$ and (3.23) with $\beta_2 = M_1\beta_1$, respectively, where the M_1 is the constant in Lemma 3.11. Then we set $\delta = \sqrt{C_1 V} C_f/2$, and if $\|\boldsymbol{f}^{n+1} - \boldsymbol{f}(t_{n+1})\| < \delta$, it holds that

$$\begin{aligned}
&H_1(\boldsymbol{f}^{n+1,2}) + H_2(\boldsymbol{f}^{n+1,2}) - H_1(\boldsymbol{f}(t_{n+1})) - H_2(\boldsymbol{f}(t_{n+1})) \\
&\quad = \big(H_1(\boldsymbol{f}^{n+1,2}) - H_1(\boldsymbol{f}^{n+1,1})\big) + \big(H_1(\boldsymbol{f}^{n+1,1}) - H_1(\boldsymbol{f}(t_{n+1}))\big) \\
&\qquad + \big(H_2(\boldsymbol{f}^{n+1,2}) - H_2(\boldsymbol{f}^{n+1,1})\big) + \big(H_2(\boldsymbol{f}^{n+1,1}) - H_2(\boldsymbol{f}(t_{n+1}))\big) \\
&\quad \le \big(H_1(\boldsymbol{f}^{n+1,2}) - H_1(\boldsymbol{f}^{n+1,1})\big) + 0 \qquad \text{(Corollary 3.9)} \\
&\qquad + \left(-\frac{H_1(\boldsymbol{f}^{n+1}) - H_1(\boldsymbol{f}^{n+1,1})}{C_1}\right) + \left(\frac{H_1(\boldsymbol{f}^{n+1}) - H_1(\boldsymbol{f}^{n+1,1})}{C_1}\right) \qquad \text{(Lemma 3.11)} \\
&\quad = H_1(\boldsymbol{f}^{n+1,2}) - H_1(\boldsymbol{f}^{n+1,1}) \le 0,
\end{aligned}$$

where the last inequality is similar to (3.24) which utilizes the convexity of $H_1(\cdot)$.

Therefore, by the decomposition in (3.21),

$$\begin{aligned}
\eta(\boldsymbol{f}^{n+1,2}) - \eta(\boldsymbol{f}(t_{n+1})) &\le H_3(\boldsymbol{f}^{n+1,2}) - H_3(\boldsymbol{f}(t_{n+1})) \\
&= \sum_{f_i^{n+1} > \frac{1}{2}} \Big(h(f_i^{n+1,2}) - h(f_i(t_{n+1}))\Big)\,\Delta v_i.
\end{aligned} \tag{3.30}$$

From the construction of $\boldsymbol{f}^{n+1,2}$, we know $f_i^{n+1,2}$ is a convex combination of 1 and f_i^{n+1}, so $f_i^{n+1} > \frac{1}{2}$ implies $f_i^{n+1,2} > \frac{1}{2}$. Therefore (3.30) can be further extended as

$$\eta(\boldsymbol{f}^{n+1,2}) - \eta(\boldsymbol{f}(t_{n+1})) \le \sum_{f_i^{n+1,2} > \frac{1}{2}} \Big(h(f_i^{n+1,2}) - h(f_i(t_{n+1}))\Big)\,\Delta v_i. \tag{3.31}$$

The remaining part of the proof is similar to the proof of Theorem 2.4. If $\eta(\boldsymbol{f}^{n+1,2}) \le \eta(\boldsymbol{f}(t_{n+1}))$, the proof is done. Otherwise, we have $\eta(\boldsymbol{f}^{n+1,2}) > \eta(\boldsymbol{f}(t_{n+1}))$, and we can continue to find $\hat{\boldsymbol{f}}^{n+1}$ and β_3 such that

$$\hat{\boldsymbol{f}}^{n+1} = \boldsymbol{f}^{n+1,2} + \beta_3(\mathbf{1} - \boldsymbol{f}^{n+1,2})$$

and $\eta(\hat{\boldsymbol{f}}^{n+1}) = \eta(\boldsymbol{f}(t_{n+1}))$. Due to the inequality (3.31), we can follow the proof of Lemma 3.4 (case 2) and Theorem 2.3 to show

$$\|\beta_3(\mathbf{1} - \boldsymbol{f}^{n+1,2})\|_2 \le M_2 \|\boldsymbol{f}(t_{n+1}) - \boldsymbol{f}^{n+1,2}\|_2,$$

where $M_2 > 0$ is a constant depending on $\|\boldsymbol{f}^{n+1}\|_\infty$ (because $\|\boldsymbol{f}^{n+1,2}\|_\infty \le \|\boldsymbol{f}^{n+1}\|_\infty$) and $\|\boldsymbol{f}(t_{n+1})\|_\infty$. Therefore, $\eta(\hat{\boldsymbol{f}}^{n+1}) \le \eta(\boldsymbol{f}(t_{n+1}))$, and

$$\begin{aligned}
&\|\hat{\boldsymbol{f}}^{n+1} - \boldsymbol{f}^{n+1}\|_2 \\
&\quad \le \|\hat{\boldsymbol{f}}^{n+1} - \boldsymbol{f}^{n+1,2}\|_2 + \|\boldsymbol{f}^{n+1,2} - \boldsymbol{f}^{n+1}\|_2 \\
&\quad \le M_2 \|\boldsymbol{f}(t_{n+1}) - \boldsymbol{f}^{n+1,2}\|_2 + \|\boldsymbol{f}^{n+1,2} - \boldsymbol{f}^{n+1}\|_2 \\
&\quad \le M_2 \|\boldsymbol{f}(t_{n+1}) - \boldsymbol{f}^{n+1}\|_2 + (1 + M_2)\|\boldsymbol{f}^{n+1,2} - \boldsymbol{f}^{n+1}\|_2 \\
&\quad = M_2 \|\boldsymbol{f}(t_{n+1}) - \boldsymbol{f}^{n+1}\|_2 + (1 + M_2)(\beta_1 + \beta_2 - \beta_1\beta_2)\|\mathbf{1} - \boldsymbol{f}^{n+1}\|_2 \\
&\quad \le \left(M_2 + \frac{(1 + M_1)(1 + M_2)\|\boldsymbol{f}^{n+1}\|_\infty}{\sqrt{C_1} C_f} \right) \|\boldsymbol{f}(t_{n+1}) - \boldsymbol{f}^{n+1}\|_2,
\end{aligned}$$

where the last inequality utilizes $(\beta_1 + \beta_2 - \beta_1\beta_2) \le \beta_1 + \beta_2$ and $\|\mathbf{1} - \boldsymbol{f}^{n+1}\|_2 \le \sqrt{V}\|\boldsymbol{f}^{n+1}\|_\infty$. If we denote the constant in front of $\|\boldsymbol{f}(t_{n+1}) - \boldsymbol{f}^{n+1}\|_2$ as M, we have proved that

$$\begin{aligned}
\hat{\boldsymbol{f}}^{n+1} &= \boldsymbol{f}^{n+1,2} + \beta_3(\mathbf{1} - \boldsymbol{f}^{n+1,2}) \\
&= \boldsymbol{f}^{n+1} + (\beta_1 + \beta_2 + \beta_3 - \beta_1\beta_2 - \beta_2\beta_3 - \beta_1\beta_3 + \beta_1\beta_2\beta_3)(\mathbf{1} - \boldsymbol{f}^{n+1})
\end{aligned}$$

satisfies $\eta(\hat{\boldsymbol{f}}^{n+1}) \le \eta(\boldsymbol{f}(t_{n+1}))$ and $\|\hat{\boldsymbol{f}}^{n+1} - \boldsymbol{f}^{n+1}\|_2 \le M\|\boldsymbol{f}(t_{n+1}) - \boldsymbol{f}^{n+1}\|_2$. Due to the monotonicity of $H(\boldsymbol{f}^{n+1} + \beta(\mathbf{1} - \boldsymbol{f}^{n+1}))$ with respect to β, if we construct β from (2.7),

$$\|\beta(\mathbf{1} - \boldsymbol{f}^{n+1})\|_2 \le \|\hat{\boldsymbol{f}}^{n+1} - \boldsymbol{f}^{n+1}\|_2 \le M\|\boldsymbol{f}^{n+1} - \boldsymbol{f}(t_{n+1})\|_2. \qquad \square$$

4. Numerical examples. We now present three numerical examples to show the effect of our entropy fix. In order to construct cases where the numerical scheme frequently violates the entropy inequality, we deliberately select highly oscillatory initial data. We would like to remark that such an entropy fix may only need to be applied occasionally in many applications.

4.1. Convection equation. In this example, we consider the finite volume method for the scalar convection equation

$$\frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} = 0, \qquad x \in (0, 1), \quad t > 0,$$

with periodic boundary conditions. We define the semidiscrete solution on a uniform grid with $N = 256$ cells:

$$u_j(t) = \frac{1}{\Delta x} \int_{x_j}^{x_{j+1}} u(t, x)\, \mathrm{d}x, \qquad j = 0, \ldots, N - 1,$$

where $\Delta x = 1/256$ and $x_j = j\Delta x$. To preserve the conservation of entropy, we adopt the numerical method introduced in [20]:

$$\frac{\mathrm{d}u_j}{\mathrm{d}t} + \frac{1}{\Delta x}\left(F_{j+1/2} - F_{j-1/2}\right) = 0, \tag{4.1}$$

where

$$F_{j+1/2} = \frac{u_{j+1} - u_j}{\log(u_{j+1}) - \log(u_j)}. \tag{4.2}$$

The conservation of the Gibbs entropy can be shown by

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\sum_j (u_j \log u_j - u_j) &= \sum_j \frac{\mathrm{d}u_j}{\mathrm{d}t}\log u_j = -\sum_j \frac{F_{j+1/2} - F_{j-1/2}}{\Delta x}\log u_j \\ &= -\sum_j F_{j+1/2}\frac{\log u_{j+1} - \log u_j}{\Delta x} = -\sum_j \frac{u_{j+1} - u_j}{\Delta x} = 0.\end{aligned}$$

Following Example 1, we demonstrate Theorem 2.5 numerically by choosing the initial condition of u_j to be a Gaussian:

$$u_j(0) = \exp\left(-40\left[\left(j + \frac{1}{2}\right)\Delta x - \frac{1}{2}\right]^2\right).$$

Two different temporal discretizations are adopted to check the numerical order, which are the first-order forward Euler method and the second-order Heun's method. The numerical errors are computed by the comparison to the numerical solution using Heun's method and a smaller time step $\Delta t = 10^{-7}$ without entropic fix. Results are summarized in Tables 1 and 2 for the forward Euler method and Heun's method, respectively.

It can been seen from Tables 1 and 2 that both temporal schemes retain their orders of convergence with the entropic fix. In the forward Euler method, an entropy fix is required at every time step to maintain the monotonicity of entropy, and the entropy fix actually reduces the L^2 error for all sizes of the time steps. However, in Heun's method, the entropy fix is applied only at a few time steps, and its effect on the final L^2 error at $T = 1$ is negligible.

To numerically verify the general case described in Theorem 2.2, we design another test case with the initial condition:

$$u_j(0) = \left|j\Delta x - \frac{1}{2}\right| + \delta, \quad j = 1, \ldots, N, \tag{4.3}$$

where δ is a small positive number to avoid the appearance of $\log(0)$ in (4.2), and we choose $\delta = 10^{-12}$ in our test. With this initial condition, there exists only one u_j very close to zero, which implies a large constant M if we apply Theorem 2.3 or

TABLE 1
Errors of the convection equation (4.1) *with Gaussian initial values at* $T = 1$.

Forward Euler method				
Time step Δt	$1/2^{15}$	$1/2^{16}$	$1/2^{17}$	$1/2^{18}$
L^2 error of $\boldsymbol{u}$	4.71×10^{-4}	2.35×10^{-4}	1.18×10^{-4}	5.88×10^{-5}
Order		1.00	0.99	1.00
Forward Euler method with entropy fix (2.4)				
Time step Δt	$1/2^{15}$	$1/2^{16}$	$1/2^{17}$	$1/2^{18}$
L^2 error of $\boldsymbol{u}$	3.72×10^{-4}	1.86×10^{-4}	9.32×10^{-5}	4.66×10^{-5}
Order		1.00	1.00	1.00
No. of entropy fixes	2^{15}	2^{16}	2^{17}	2^{18}

TABLE 2
Errors of the convection equation (4.1) *with Gaussian initial values at* $T = 1$.

Heun's method				
Time step Δt	$1/2^{11}$	$1/2^{12}$	$1/2^{13}$	$1/2^{14}$
L^2 error of $\boldsymbol{u}$	1.74×10^{-5}	4.20×10^{-6}	9.06×10^{-7}	1.93×10^{-7}
Order		2.05	2.21	2.23
Heun's method with entropy fix (2.4)				
Time step Δt	$1/2^{11}$	$1/2^{12}$	$1/2^{13}$	$1/2^{14}$
L^2 error of $\boldsymbol{u}$	1.74×10^{-5}	4.20×10^{-6}	9.06×10^{-7}	1.93×10^{-7}
Order		2.05	2.21	2.23
No. of entropy fixes	50	65	88	96

TABLE 3
Errors of the convection equation (4.1) *with piecewise linear initial values at* $T = 1$.

Forward Euler method				
Time step Δt	$1/2^{15}$	$1/2^{16}$	$1/2^{17}$	$1/2^{18}$
L^2 error of $\boldsymbol{u}$	1.81×10^{-3}	7.47×10^{-4}	3.42×10^{-4}	1.64×10^{-4}
Order		1.28	1.13	1.06
Forward Euler method with entropy fix (2.4)				
Time step Δt	$1/2^{15}$	$1/2^{16}$	$1/2^{17}$	$1/2^{18}$
L^2 error of $\boldsymbol{u}$	1.80×10^{-3}	7.42×10^{-4}	3.39×10^{-4}	1.62×10^{-4}
Order		1.28	1.13	1.07
No. of entropy fixes	2^{15}	2^{16}	2^{17}	2^{18}

TABLE 4
Errors of the convection equation (4.1) *with piecewise linear initial values at* $T = 1$.

Heun's method				
Time step Δt	$1/2^{11}$	$1/2^{12}$	$1/2^{13}$	$1/2^{14}$
L^2 error of $\boldsymbol{u}$	7.04×10^{-4}	1.76×10^{-4}	4.46×10^{-5}	1.14×10^{-5}
Order		2.00	1.98	1.97
Heun's method with entropy fix (2.4)				
Time step Δt	$1/2^{11}$	$1/2^{12}$	$1/2^{13}$	$1/2^{14}$
L^2 error of $\boldsymbol{u}$	7.03×10^{-4}	1.75×10^{-4}	4.45×10^{-5}	1.14×10^{-5}
Order		2.01	1.98	1.96
No. of entropy fixes	2^{11}	2^{12}	2^{13}	2^{14}

Theorem 2.5. We therefore expect that the theoretical result in Theorem 2.2 better fits this test case.

Results of the two temporal discretizations are summarized in Tables 3 and 4. The general behavior of the entropy fix for the initial condition (4.3) is similar to the Gaussian case. According to Theorem 2.2, the entropy fix has only infinitesimal effect on the numerical order. As shown in Tables 3 and 4, the orders of convergence are well retained for both temporal schemes even if the entropy fix is applied on every time step.

4.2. Linear Fokker–Planck equation. In this example, we consider the one-dimensional linear Fokker–Planck equation (also known as the drift-diffusion equation):

$$f_t = f_{xx} + (V'(x)f)_x, \qquad t > 0, x \in (0,1), \tag{4.4}$$

with periodic boundary condition $f(t,0) = f(t,1)$ and potential function

$$V(x) = \frac{1}{2\pi}\cos(20\pi x).$$

Let $M(x) = \exp(-V(x))$; then (4.4) can be written equivalently as

$$f_t = \left(M\left(\frac{f}{M}\right)_x\right)_x. \tag{4.5}$$

If we further define $g(t,x) = f(t,x)/M(x)$, then (4.5) becomes

$$g_t = \frac{1}{M}(Mg_x)_x, \qquad t > 0, x \in (0,1). \tag{4.6}$$

We will focus on the discretization of (4.6). Initial condition is taken as

$$g(0,x) = 1.2 + \sum_{j=1}^{20} \frac{j}{210}\sin(2j\pi x),$$

Note that $\sum_{j=1}^{20} j = 210$, which yields $0.2 \le g(0,x) \le 2.2$ indicating that the condition of Theorem 2.3 is fulfilled after the spatial discretization. We partition $[0,1]$ into $N = 64$ grids uniformly with mesh size $\Delta x = 1/N$ and take central difference for spatial discretization. Denoting $g_j = g(t, j\Delta x)$, $M_j = M(j\Delta x)$, and $M_{j+1/2} = M((j+1/2)\Delta x)$ for $j = 0, \ldots, N-1$, (4.6) can be approximated by

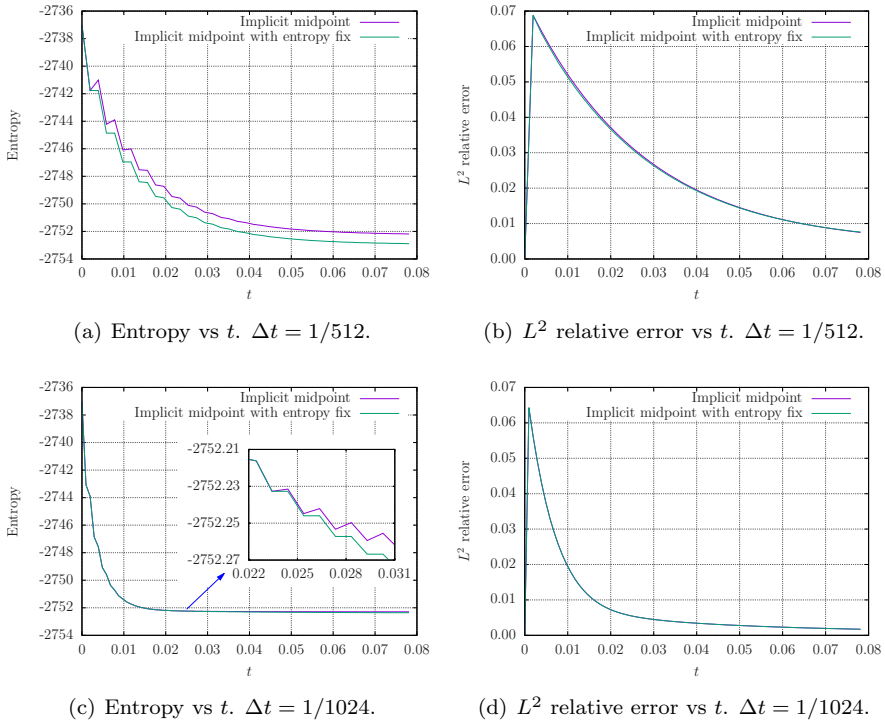
$$\frac{\mathrm{d}g_j}{\mathrm{d}t} = \frac{1}{M_j}\frac{M_{j+1/2}(g_{j+1} - g_j) - M_{j-1/2}(g_j - g_{j-1})}{(\Delta x)^2}. \tag{4.7}$$

The exact solution of (4.7) can be calculated by evaluating the eigenvalues and eigenvectors of the right-hand side of (4.7).

The semidiscrete scheme (4.7) (time is kept continuous) satisfies the conservation of mass and the monotonicity of entropy with weight M_j. In fact, it is easy to verify $\sum_{j=0}^{N-1} M_j g_j$ remains as constant. For the entropy, we have

$$\begin{aligned}\frac{\mathrm{d}\left(\sum_{j=0}^{N-1} M_j g_j \log g_j\right)}{\mathrm{d}t} &= \frac{1}{(\Delta x)^2}\sum_{j=0}^{N-1}\left(M_{j+1/2}(g_{j+1} - g_j) - M_{j-1/2}(g_j - g_{j-1})\right)\log g_j \\ &= -\frac{1}{(\Delta x)^2}\sum_{j=0}^{N-1} M_{j-1/2}(g_j - g_{j-1})(\log g_j - \log g_{j-1}) \le 0.\end{aligned} \tag{4.8}$$

We now discretize (4.7) by the implicit midpoint (i.e., Crank–Nicolson) method. This time discretization still conserves the mass. However, there is no guarantee that the entropy will decay monotonically in time (in fact, it does not). In Figure 3, we report the time evolution of the entropy with and without the entropy fix. Two different time steps $\Delta t = 1/512$ and $\Delta t = 1/1024$ are considered. In both cases, it is clear that the entropy decreases monotonically with the help of the entropy fix. Meanwhile, the L^2 error of the solution remains almost the same with and without the entropy fix. It is interesting to note that when $\Delta t = 1/512$, the entropy fix is only needed at the first few time steps. On the other hand, when $\Delta t = 1/1024$, the entropy fix is required only after $t = 0.02$.



(a) Entropy vs t. $\Delta t = 1/512$.

(b) L^2 relative error vs t. $\Delta t = 1/512$.

(c) Entropy vs t. $\Delta t = 1/1024$.

(d) L^2 relative error vs t. $\Delta t = 1/1024$.

FIG. 3. *Example of the linear Fokker–Planck equation. Time evolution of the entropy* $H(\boldsymbol{g}) = \sum_{j=0}^{N-1}(g_j \log g_j - g_j)M_j\Delta x$ *and the* L^2 *relative error* $\|\boldsymbol{g}-\boldsymbol{g}_{\text{exact}}\|_2/\|\boldsymbol{g}_{\text{exact}}\|_2 = (\sum_{j=0}^{N-1}(g_j - g_{\text{exact},j})^2 M_j\Delta x)^{1/2}/(\sum_{j=0}^{N-1}(g_{\text{exact},j})^2 M_j\Delta x)^{1/2}$, *where* $\Delta x = 1/64$, $\Delta t = 1/512$ *in the top two panels and* $\Delta t = 1/1024$ *in the bottom two panels.*

4.3. Nonlinear Boltzmann equation. In this example, we consider a nonlinear model introduced in [7], which results from a Fourier method for the spatially homogeneous Boltzmann equation. The governing equation reads

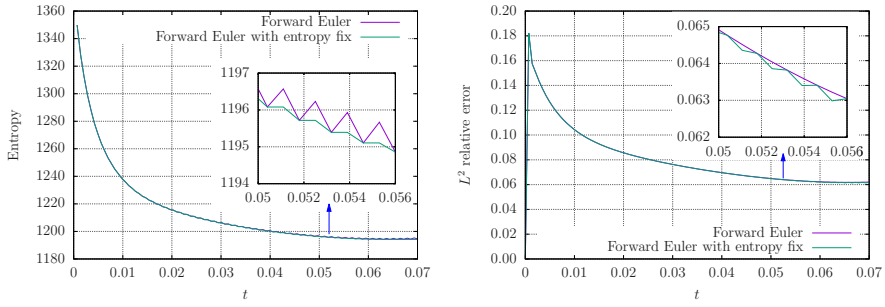
$$\frac{\mathrm{d}f_r(t)}{\mathrm{d}t} = \sum_{p,q,s\in\mathcal{X}} A_{pq}^{rs}\left(f_p(t)f_q(t) - f_r(t)f_s(t)\right), \quad r \in \mathcal{X}, \tag{4.9}$$

where f_r represents the approximation of the distribution function on a uniform three-dimensional lattice index set $\mathcal{X} = \{(r_1, r_2, r_3) \mid r_i = 0, \ldots, M-1 \text{ for } i = 1, 2, 3\}$. In [7], the coefficients A_{pq}^{rs} are determined in such a way that the semidiscrete scheme (4.9) decays the entropy. However, this property may not hold when the time is discretized.

In our experiment, we choose $M = 17$, and the values of A_{pq}^{rs} are given in Appendix A. The initial condition is taken as the one with lower bound so that condition of Theorem 2.3 is satisfied:

$$f_r(0) = 3.2 + \sum_{j=1}^{10} \frac{j}{55}\left[\sin\left(j\pi\left(\frac{r_1}{M} - \frac{1}{2}\right)\right) + \sin\left(j\pi\left(\frac{r_2}{M} - \frac{1}{2}\right)\right) + \sin\left(j\pi\left(\frac{r_3}{M} - \frac{1}{2}\right)\right)\right].$$

We solve (4.9) by the forward Euler method with time step $\Delta t = 0.0007$. The results are displayed in Figure 4, from which we can see that the entropy fix method guarantees the monotonicity of the entropy. The numerical error is computed by comparison



(a) Entropy vs t. (b) L^2 relative error vs t.

FIG. 4. *Example of the nonlinear Boltzmann equation. Time evolution of the entropy* $H(\boldsymbol{f}) = \sum_{r\in\mathcal{X}}(f_r \log f_r - f_r)\Delta v$ *and the* L^2 *relative error* $\|\boldsymbol{f} - \boldsymbol{f}_{\text{exact}}\|_2/\|\boldsymbol{f}_{\text{exact}}\|_2 = (\sum_{r\in\mathcal{X}}(f_r - f_{\text{exact},r})^2\Delta v)^{1/2}/(\sum_{r\in\mathcal{X}}(f_{\text{exact},r})^2\Delta v)^{1/2}$*, where* $\Delta v = (3(3+\sqrt{2})/17)^3$ *and* $\Delta t = 0.0007$*.* $\boldsymbol{f}_{\text{exact}}$ *is the numerical solution evaluated with time step* $\Delta t = 0.000175$*.*

with the numerical solution computed with a smaller time step $\Delta t = 0.000175$, with and without the entropy fix. It can be seen that the two error curves almost coincide with each other, meaning that the entropy fix does not ruin the numerical accuracy.

5. Conclusions. This paper focuses on the entropic method for a conservative and positive system of ODEs. When the numerical solution at the next time step violates the monotonicity of entropy, our entropic method revises it by a linear interpolation to the constant state. The resulting scheme decays the entropy monotonically, while the order of local truncation error has a slight reduction in general. However, in some special cases, the numerical order is proved to be retained after entropic revision. Numerical experiments validate our results. Future work includes the extension of the entropic method to spatially inhomogeneous kinetic equations such as the Boltzmann equation and the radiative transfer equations.

Appendix A. Coefficients in (4.9). The values of A_{pq}^{rs} are given by

$$A_{pq}^{rs} = \frac{1}{M^9}\sum_{l,h,k\in K}\hat{B}_M^{\sigma}(h-k,l-k)E_{-l}(p-s)E_{-h}(q-s)E_k(r-s), \tag{A.1}$$

where K is defined as $K = \{k \mid k = (k_1,k_2,k_3), -m \le k_1,k_2,k_3 \le m\}$ with $M = 2m+1$, and $E_k(v) = \exp(\frac{\mathbf{i}\pi}{T}k\cdot v)$ is the Fourier basis on the period $[-T,T]^3$. The kernel functions $\hat{B}_M^{\sigma}(\cdot,\cdot)$ are defined by

$$\hat{B}_M^{\sigma}(i,j) := \hat{B}(i \bmod M, j \bmod M)\sigma_M(i \bmod M)\sigma_M(j \bmod M),$$

where mod is the symmetric modulo function such that each component of $i \bmod M$ ranges from $-m$ to m, and $\sigma_M(i) = \tilde{\sigma}_M(i_1)\tilde{\sigma}_M(i_2)\tilde{\sigma}_M(i_3)$, where $\tilde{\sigma}_M(\beta)$ is the one-dimensional modified Jackson filter [21] given by

$$\tilde{\sigma}_M(\beta) = \frac{(m+1-|\beta|)\cos\left(\frac{\pi|\beta|}{m+1}\right) + \sin\left(\frac{\pi|\beta|}{m+1}\right)\cot\left(\frac{\pi}{m+1}\right)}{m+1}.$$

In the example in subsection 4.3, we adopt the kernel modes for the case of the Maxwell molecules presented in [16] with

$$\hat{B}(k,l) := \int_0^1 r^2\,\mathrm{Sinc}(\xi r)\,\mathrm{Sinc}(\eta r)\,\mathrm{d}r = \frac{(\xi+\eta)\sin(\xi-\eta) - (\xi-\eta)\sin(\xi+\eta)}{2\xi\eta(\xi^2-\eta^2)},$$

where $\xi = |k+l|\lambda\pi, \eta = |k-l|\lambda\pi$, and $\lambda = 2/(3+\sqrt{2})$. In the numerical simulation, we take $M = 17$ and $T = 3/\lambda$.

REFERENCES

[1] R. Abgrall, *High order schemes for hyperbolic problems using globally continuous approximation and avoiding mass matrices*, J. Sci. Comput., 73 (2017), pp. 461–494, https://doi.org/10.1007/s10915-017-0498-4.

[2] R. Abgrall, P. Öffner, and H. Ranocha, *Reinterpretation and extension of entropy correction terms for residual distribution and discontinuous Galerkin schemes: Application to structure preserving discretization*, J. Comput. Phys., 453 (2022), 110955, https://doi.org/10.1016/j.jcp.2022.110955.

[3] R. Bailo, J. A. Carrillo, and J. Hu, *Fully discrete positivity-preserving and energy-dissipating schemes for aggregation-diffusion equations with a gradient flow structure*, Commun. Math. Sci., 18 (2020), pp. 1259–1303.

[4] F. Bouchut, C. Bourdarias, and B. Perthame, *A MUSCL method satisfying all the numerical entropy inequalities*, Math. Comp., 65 (1996), pp. 1439–1461, https://doi.org/10.1090/S0025-5718-96-00752-1.

[5] C. Buet and S. Cordier, *Conservative and entropy decaying numerical scheme for the isotropic Fokker–Planck–Landau equation*, J. Comput. Phys., 145 (1998), pp. 228–245.

[6] C. Buet and S. Cordier, *Numerical analysis of conservative and entropy schemes for the Fokker–Planck–Landau equation*, SIAM J. Numer. Anal., 36 (2006), pp. 953–973.

[7] Z. Cai, Y. Fan, and L. Ying, *An entropic fourier method for the Boltzmann equation*, SIAM J. Sci. Comput., 40 (2018), pp. A2858–A2882, https://doi.org/10.1137/17M1127041.

[8] S. Chow, L. Dieci, and W. Li, *Entropy dissipation semi-discretization schemes for Fokker–Planck equations*, J. Dynam. Differential Equations, 31 (2019), pp. 765–792.

[9] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth, *On the LambertW function*, Adv. Comput. Math., 5 (1996), pp. 329–359.

[10] P. Degond and B. Lucquin-Desreux, *An entropy scheme for the Fokker-Planck collision operator of plasma kinetic theory*, Numer. Math., 68 (1994), pp. 239–262.

[11] D. Goldstein, B. Strutevant, and J. E. Broadwell, *Investigations of the motion of discrete-velocity gases*, in Rarefied Gas Dynamics: Theoretical and Computational Techniques, D. P. Weaver, D. H. Campbell, and E. P. Muntz, eds., American Institute of Aeronautics and Astronautics, Reston, VA, 1989, pp. 100–117.

[12] A. Hoorfar and M. Hassani, *Inequalities on the Lambert W function and hyperpower function*, J. Inequal. Pure Appl. Math., 9 (2008), pp. 1–5.

[13] D. I. Ketcheson, *Relaxation Runge–Kutta methods: Conservation and stability for inner-product norms*, SIAM J. Numer. Anal., 57 (2019), pp. 2850–2870, https://doi.org/10.1137/19M1263662.

[14] P. G. LeFloch, J. M. Mercier, and C. Rohde, *Fully discrete, entropy conservative schemes of arbitrary order*, SIAM J. Numer. Anal., 40 (2002), pp. 1968–1992.

[15] H. Liu and H. Yu, *An entropy satisfying conservative method for the Fokker–Planck equation of the finitely extensible nonlinear elastic dumbbell model*, SIAM J. Numer. Anal., 50 (2012), pp. 1207–1239.

[16] L. Pareschi and G. Russo, *Numerical solution of the Boltzmann equation* I*: Spectrally accurate approximation of the collision operator*, SIAM J. Numer. Anal., 37 (2000), pp. 1217–1245.

[17] L. Pareschi and M. Zanella, *Structure preserving schemes for nonlinear Fokker–Planck equations and applications*, J. Sci. Comput., 74 (2018), pp. 1575–1600.

[18] H. Ranocha, M. Sayyari, L. Dalcin, M. Parsani, and D. I. Ketcheson, *Relaxation Runge–Kutta methods: Fully discrete explicit entropy-stable schemes for the compressible Euler and Navier–Stokes equations*, SIAM J. Sci. Comput., 42 (2020), pp. A612–A638, https://doi.org/10.1137/19M1263480.

[19] A. Ruszczyński, *Nonlinear Optimization*, Princeton University Press, Princeton, NJ, 2006.

[20] E. Tadmor, *Entropy stability theory for difference approximations of nonlinear conservation laws and related time-dependent problems*, Acta Numer., 12 (2003), pp. 451–512, https://doi.org/10.1017/S0962492902000156.

[21] A. Weiße, G. Wellein, A. Alvermann, and H. Fehske, *The kernel polynomial method*, Rev. Modern Phys., 78 (2006), pp. 275–306, https://doi.org/10.1103/RevModPhys.78.275.