

SUPPLEMENTARY MATERIALS: AN ENERGY STABLE AND POSITIVITY-ENERGY STABLE AND POSITIVITY-PRESERVING SCHEME FOR THE MAXWELL-STEFAN DIFFUSION SYSTEM*

XIAOKAI HUO[†], HAILIANG LIU[‡], ATHANASIOS E. TZAVARAS[†], AND SHUAIKUN WANG[§]

SM1. Proof Theorem 4.2. We give the details of the proof of Theorem 4.2 here.

THEOREM SM1.1. *Assume $b_{ij} > 0$ and $b_{ij} = b_{ji}$ for $i \neq j$ and $i, j = 1, \dots, n$. Assume $\rho^k \in (\mathcal{C}_{\text{per}}^d)^n$ be positive. Then there exists a constant $\delta_0 > 0$, such that $\rho^{k+1} > 0$ is a solution of the numerical scheme (4.2)–(4.4) if and only if it is a minimizer of the optimization problem:*

$$(SM1.1) \quad \rho^{k+1} = \arg \min_{(\rho, w) \in K_\delta} \left\{ J = \frac{1}{4\Delta t} \left[\sum_{i,j=1}^n b_{ij} \hat{\rho}_i^k \hat{\rho}_j^k (w_i - w_j)^2 \right] + F_h(\rho) \right\},$$

where

$$K_\delta = \left\{ (\rho, w) : \rho \in (\mathcal{C}_{\text{per}}^d)^n, w \in (\mathcal{E}_{\text{per}}^d)^n; \rho_{i,\ell} \geq \delta, \rho_{i,\ell} - \rho_{i,\ell}^k + d_h(\hat{\rho}_i^k w_i)_\ell = 0, \right. \\ \left. \sum_{i=1}^n \hat{\rho}_{i,\ell_1, \dots, \ell_s + \frac{1}{2}, \dots, \ell_d}^k w_{i,\ell_1, \dots, \ell_s + \frac{1}{2}, \dots, \ell_d} = 0 \text{ and } \sum_{i=1}^n \rho_{i,\ell} = 1, \right. \\ \left. \forall i = 1, \dots, n, \forall \ell = (\ell_1, \dots, \ell_d) \in \{1, \dots, N\}^d, s = 1, \dots, d \right\},$$

for any $0 < \delta \leq \delta_0$.

The proof follows a similar strategy as the proof of Theorem 3.1 for the one dimensional case. We establish a multidimensional version of Lemma 3.2.

LEMMA SM1.2. *Suppose Φ is a $(n-1) \times (n-1)$ symmetric positive definite matrix, with $\Phi_{ij} \in \mathcal{E}_{\text{per}}^d$. Suppose $\phi \in (\mathcal{C}_{\text{per}}^d)^{n-1}$ satisfies $\|\phi\|_{L^\infty} \leq M$,*

$$\|\phi\|_{L^\infty} := \max_{\substack{i=1, \dots, n-1 \\ \ell_s=1, \dots, N \\ s=1, \dots, d}} |\phi_{i,\ell_1, \dots, \ell_d}|.$$

The following estimate holds

$$\|\mathcal{L}_\Phi^{-1} \phi\|_{L^\infty} \leq \frac{CM}{\lambda_{\min}} h^{-\frac{1}{2}} (n-1)^{\frac{1}{2}},$$

*Supplementary material for SINUM MS#M133866.

<https://doi.org/10.1137/20M1338666>

[†]Computer, Electrical and Mathematical Science and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955, Saudi Arabia (xiaokai.huo@kaust.edu.sa, athanasios.tzavaras@kaust.edu.sa).

[‡]Mathematics Department, Iowa State University, Ames, IA 50011 (hliu@iastate.edu).

[§]School of Mathematics, Shandong University, Jinan 250100, China (skwang@email.sdu.edu.cn).

where $C > 0$ depends only on the domain, λ_{\min} is the minimum of the eigenvalues of Φ over all grid points:

$$\lambda_{\min} = \min_{\substack{\ell_s=1,\dots,N \\ s=1,\dots,d}} \left\{ \lambda_{\ell_1,\dots,\ell_s+\frac{1}{2},\dots,\ell_d} \text{ the eigenvalue of } (\Phi_{ij,\ell_1,\dots,\ell_s+\frac{1}{2},\dots,\ell_d})_{(n-1)\times(n-1)} \right\}$$

Proof.

$$\begin{aligned} \|\phi\|_{L^2}^2 &:= h^d \sum_{\substack{i=1,\dots,n-1 \\ \ell^s=1,\dots,N \\ s=1,\dots,d}} |\phi_{i,\ell_1,\dots,\ell_d}|^2 \\ &\leq h^d \sum_{\substack{i=1,\dots,n-1 \\ \ell^s=1,\dots,N \\ s=1,\dots,d}} |M|^2 \leq (n-1)h^d N^d |M|^2 = (n-1)L^d |M|^2. \end{aligned}$$

Let $g = \phi$ and $f = \mathcal{L}_{\Phi}^{-1}g$ in (4.1), the norm satisfies

$$\begin{aligned} \lambda_{\min} \|D_h f\|_{L^2}^2 &\leq [D_h f, \Phi D_h f] \\ &= -\langle f, d_h(\Phi D_h f) \rangle = -\langle f, \phi \rangle \leq \|f\|_{L^2} \|\phi\|_{L^2} \leq C_P \|f\|_{L^2} \|\phi\|_{L^2}, \end{aligned}$$

according to the discrete Poincaré inequality. Therefore, we get

$$\|D_h f\|_{L^2} \leq \frac{C_P}{\lambda_{\min}} \|\phi\|_{L^2}.$$

Using an inverse inequality in $(\mathring{C}^d)_{\text{per}}^{n-1}$ leads to

$$\|f\|_{L^\infty} \leq C_1 h^{-\frac{1}{2}} \|D_h f\|_{L^2} \leq \frac{C_1 C_P}{\lambda_{\min}} h^{-\frac{1}{2}} L^{\frac{d}{2}} M (n-1)^{\frac{1}{2}} \leq \frac{CM}{\lambda_{\min}} h^{-\frac{1}{2}} (n-1)^{\frac{1}{2}}. \quad \square$$

Now we prove Theorem SM1.1.

Proof. In a fashion similar to the proof of the one dimensional case, there exists a unique solution to the optimization problem (SM1.1) for any $\delta > 0$. This follows from the same argument with notations replaced by the multidimensional version. To prove that the minimizer of (SM1.1) does not touch the boundary of K_δ , we use the equivalent optimization problem

$$(SM1.2) \quad \min_{\tilde{\rho} \in \mathring{K}_\delta} \left\{ J = \frac{1}{2\Delta t} \|\tilde{\rho} - \tilde{\rho}^k\|_{\mathcal{L}_{\tilde{D}^k}^{-1}}^2 + F_h(\tilde{\rho}) \right\},$$

over the set

$$\begin{aligned} \mathring{K}_\delta = \left\{ \tilde{\rho} : \tilde{\rho} - \tilde{\rho}^k \in (\mathring{C}_{\text{per}}^d)^{n-1}; \rho_{i,\ell} \geq \delta, \sum_{i=1}^{n-1} \rho_{i,\ell} \leq 1 - \delta, \right. \\ \left. \forall i = 1, \dots, n-1, \ell \in \{1, \dots, N\}^d \right\}. \end{aligned}$$

Assume the minimizer touches the boundary of \mathring{K}_δ at the grid point $\ell^0 = (\ell_1^0, \dots, \ell_d^0)$ for the i_0 -th component, i.e.

$$(SM1.3) \quad \rho_{i_0, \ell_1^0, \dots, \ell_d^0}^* = \delta.$$

Next we consider the following two cases:

(a)

$$\sum_{i=1}^{n-1} \rho_{i,\ell^0}^* \geq \frac{1}{2},$$

(b)

$$\sum_{i=1}^{n-1} \rho_{i,\ell^0}^* < \frac{1}{2}.$$

First consider the case (a). We also suppose $\{\rho_{i,\ell_1^0,\dots,\ell_d^0}^*\}_{i=1}^{n-1}$ achieves its maximum at the i_1 -th component, and $\{\rho_{i_0,\ell}^*\}_{\ell \in \{1,\dots,N\}^d}$ achieves its maximum at $\ell = \ell^1 = (\ell_1^1, \dots, \ell_d^1)$. We calculate the directional derivative of the objective function (SM1.2) along the direction

$$\nu_{i,\ell_1,\dots,\ell_d} = \begin{cases} 1, & \text{for } i = i_0, \ell_s = \ell_s^0, \forall s = 1, \dots, d, \\ -1, & \text{for } i = i_1, \ell_s = \ell_s^0, \forall s = 1, \dots, d, \\ -1, & \text{for } i = i_0, \ell_s = \ell_s^1, \forall s = 1, \dots, d, \\ 1, & \text{for } i = i_1, \ell_s = \ell_s^1, \forall s = 1, \dots, d, \\ 0, & \text{otherwise,} \end{cases}$$

and we get

$$\begin{aligned} & \left. \frac{1}{h^d} \frac{d}{ds} J(\tilde{\rho}^* + s\nu) \right|_{s=0} \\ &= \frac{1}{\Delta t} (\mathcal{L}_{\hat{D}^k}^{-1}(\tilde{\rho}^* - \tilde{\rho}^k))_{i_0,\ell^0} - \frac{1}{\Delta t} (\mathcal{L}_{\hat{D}^k}^{-1}(\tilde{\rho}^* - \tilde{\rho}^k))_{i_1,\ell^0} - \frac{1}{\Delta t} (\mathcal{L}_{\hat{D}^k}^{-1}(\tilde{\rho}^* - \tilde{\rho}^k))_{i_0,\ell^1} \\ & \text{(SM1.4)} \\ &+ \frac{1}{\Delta t} (\mathcal{L}_{\hat{D}^k}^{-1}(\tilde{\rho}^* - \tilde{\rho}^k))_{i_1,\ell^1} + \log \rho_{i_0,\ell^0}^* - \log \rho_{i_1,\ell^0}^* - \log \rho_{i_0,\ell^1}^* + \log \rho_{i_1,\ell^1}^*. \end{aligned}$$

Since ρ_{i_1,ℓ^0}^* is the maximum point and the assumption (a) that $\sum_{i=1}^{n-1} \rho_{i,\ell^0}^* \geq \frac{1}{2}$,

$$\text{(SM1.5)} \quad \rho_{i_1,\ell^0}^* \geq \frac{1}{2(n-1)}.$$

Since ρ_{i_0,ℓ^1}^* is the maximum point and

$$\sum_{\ell \in \{1,\dots,d\}^N} \rho_{i_0,\ell}^* = \sum_{\ell \in \{1,\dots,d\}^N} \rho_{i_1,\ell}^k,$$

we have

$$\text{(SM1.6)} \quad \rho_{i_0,\ell^1}^* \geq \frac{m}{h^d N^d},$$

where m is set to be

$$m = \min_{\{i=1,\dots,n-1\}} \left\{ h^d \sum_{\ell \in \{1,\dots,N\}^d} \rho_{i,\ell}^k \right\}.$$

In order to guarantee $\tilde{\rho}^* + s\nu \in \overset{\circ}{K}_\delta$, we assume

$$\delta \leq \frac{m}{2h^d N^d}$$

so that $\rho_{i_0, \ell^1}^* - s \geq \frac{G}{h^d N^d} - s \geq \delta$ for small s . One can check for other components and get $\tilde{\rho}^* + s\nu \in \overset{\circ}{K}_\delta$ for $\delta \leq \frac{1}{4(n-1)}$. We also have

$$\rho_{i_1, \ell^1}^* \leq 1 - \delta < 1.$$

Taking the above inequality and (SM1.3), (SM1.5)-(SM1.6) into (SM1.4) and applying Lemma SM1.2 leads to

$$\begin{aligned} & \left. \frac{1}{h^d} \frac{d}{ds} J(\tilde{\rho}^* + s\nu) \right|_{s=0} \\ & \leq \frac{8C}{\lambda_{\min}^k \Delta t} h^{-\frac{1}{2}} (n-1)^{\frac{1}{2}} + \log \delta - \log \frac{1}{2(n-1)} - \log \frac{m}{h^d N^d} + \log 1, \end{aligned}$$

where λ_{\min}^k is the minimum eigenvalue of \hat{D}^k . Taking

$$(SM1.7) \quad \delta_0 \leq \min \left\{ \frac{m}{4(n-1)h^d N^d} e^{-\frac{8C}{\lambda_{\min}^k \Delta t} h^{-\frac{1}{2}} (n-1)^{\frac{1}{2}}}, \frac{m}{2h^d N^d}, \frac{1}{4(n-1)} \right\}$$

leads to

$$\left. \frac{1}{h^d} \frac{d}{ds} J(\tilde{\rho}^* + s\nu) \right|_{s=0} \leq -\log 2 < 0,$$

which contradicts to the assumption that $\tilde{\rho}^*$ is a minimizer.

Next we consider the case (b). We also suppose $\{\rho_{i_0, \ell}^*\}_{\ell \in \{1, \dots, N\}^d}$ achieves its maximum at $\ell = \ell^1 = (\ell_1^1, \dots, \ell_d^1)$. We take

$$\nu_{i, \ell_1, \dots, \ell_d} = \begin{cases} 1, & \text{for } i = i_0, \ell_s = \ell_s^0, \forall s = 1, \dots, d, \\ -1, & \text{for } i = i_1, \ell_s = \ell_s^0, \forall s = 1, \dots, d, \\ 0, & \text{otherwise,} \end{cases}$$

and use (SM1.3), (b), (SM1.6) to get

$$\begin{aligned} & \left. \frac{1}{h^d} \frac{d}{ds} J(\tilde{\rho}^* + s\nu) \right|_{s=0} \\ & = \frac{1}{\Delta t} (\mathcal{L}_{\hat{D}^k}^{-1}(\tilde{\rho}^* - \tilde{\rho}^k))_{i_0, \ell^0} - \frac{1}{\Delta t} (\mathcal{L}_{\hat{D}^k}^{-1}(\tilde{\rho}^* - \tilde{\rho}^k))_{i_0, \ell^1} + \log \rho_{i_0, \ell^0}^* \\ & \quad - \log \left(1 - \sum_{i=1}^{n-1} \rho_{i_0, \ell_1}^* \right) - \log \rho_{i_0, \ell_1}^* + \log \left(1 - \sum_{i=1}^{n-1} \rho_{i_0, \ell_1}^* \right) \\ & \leq \frac{4C}{\lambda_{\min}^k \Delta t} h^{-\frac{1}{2}} (n-1)^{\frac{1}{2}} + \log \delta - \log \frac{1}{2} - \log \frac{m}{h^d N^d} + \log 1, \end{aligned}$$

Taking

$$(SM1.8) \quad \delta_0 \leq \min \left\{ \frac{m}{4h^d N^d} e^{-\frac{4C}{\lambda_{\min}^k h \Delta t} h^{-\frac{1}{2}} (n-1)^{\frac{1}{2}}}, \frac{m}{2h^d N^d} \right\}$$

leads to

$$\left. \frac{1}{h} \frac{d}{ds} J(\tilde{\rho}^* + s\nu) \right|_{s=0} = -\log 2 < 0,$$

which contradicts to the assumption that $\tilde{\rho}^*$ is a minimizer, and so the situation (b) cannot occur.

On the other hand, we suppose $\tilde{\rho}^*$ touches the other boundary with

$$(SM1.9) \quad \sum_{i=1}^{n-1} \rho_{i,\ell^0}^* = 1 - \delta.$$

Suppose ρ_{i_0,ℓ^0}^* achieves its maximum at i_0 , then

$$(SM1.10) \quad \rho_{i_0,\ell^0}^* \geq \frac{1 - \delta}{n - 1} \geq \frac{1}{2(n - 1)},$$

for $\delta \leq \frac{1}{2}$.

Since $\tilde{\rho}^* - \tilde{\rho}^k \in (\tilde{\mathcal{C}}_{\text{per}}^d)^{n-1}$, we have

$$\sum_{\ell \in \{1, \dots, N\}^d} \sum_{i=1}^{n-1} \rho_{i,\ell}^* = \sum_{\ell \in \{1, \dots, N\}^d} \sum_{i=1}^{n-1} \rho_{i,\ell}^k \leq N^d (1 - \rho_{\min}^k)$$

with

$$\rho_{\min}^k = \min_{\substack{i=1, \dots, n, \\ \ell \in \{1, \dots, N\}^d}} \rho_{i,\ell}^k.$$

Suppose $\sum_{i=1}^{n-1} \rho_{i,\ell^1}^*$ achieves its minimum at ℓ^1 , then we have

$$\begin{aligned} \sum_{i=1}^{n-1} \rho_{i,\ell^1}^* &\leq \frac{1}{N^d - 1} (N^d (1 - \rho_{\min}^k) - (1 - \delta)) \\ &\leq 1 - \frac{N^d \rho_{\min}^k - \delta}{N^d - 1} \leq 1 - \frac{2N^d - 1}{2(N^d - 1)} \rho_{\min}^k. \end{aligned}$$

if $\delta \leq \frac{1}{2} \rho_{\min}^k$.

We take

$$\nu_{i,\ell} = \begin{cases} -1, & \text{for } i = i_0, \ell = \ell^0, \\ 1, & \text{for } i = i_0, \ell = \ell^1, \\ 0, & \text{otherwise,} \end{cases}$$

and use the above inequality together with (SM1.9),(SM1.10) to obtain

$$\begin{aligned} &\left. \frac{1}{h} \frac{d}{ds} J(\tilde{\rho}^* + s\nu) \right|_{s=0} \\ &= -\frac{1}{\Delta t} (\mathcal{L}_{\tilde{D}^k}^{-1}(\tilde{\rho}^* - \tilde{\rho}^k))_{i_0,\ell^0} - \log \rho_{i_0,\ell^0}^* + \log \left(1 - \sum_{i=1}^{n-1} \rho_{i,\ell^0}^* \right) \\ &\quad + \frac{1}{\Delta t} (\mathcal{L}_{\tilde{D}^k}^{-1}(\tilde{\rho}^* - \tilde{\rho}^k))_{i_0,\ell^1} + \log \rho_{i_0,\ell^1}^* - \log \left(1 - \sum_{i=1}^{n-1} \rho_{i,\ell^1}^* \right) \\ &\leq \frac{4C}{\lambda_{\min}^k \Delta t} h^{-\frac{1}{2}} (n-1)^{\frac{1}{2}} - \log \frac{1}{2(n-1)} + \log \delta + \log 1 - \log \frac{2N^d - 1}{2(N^d - 1)} \rho_{\min}^k. \end{aligned}$$

Taking

$$(SM1.11) \quad \delta_0 \leq \min \left\{ \frac{(2N^d - 1)\rho_{\min}^k}{8(N^d - 1)(n - 1)} e^{-\frac{4C}{\lambda_{\min}^k \Delta t} h^{-\frac{1}{2}}(n-1)^{\frac{1}{2}}}, \frac{1}{2}\rho_{\min}^k, \frac{1}{4(n-1)} \right\}$$

leads to

$$\left. \frac{1}{h^d} \frac{d}{ds} J(\tilde{\rho}^* + s\nu) \right|_{s=0} \leq -\log 2 < 0,$$

which contradicts to the assumption that $\tilde{\rho}^*$ is a minimizer.

We conclude that there exists a δ_0 , which can be chosen to be the smaller value of (SM1.7), (SM1.8) and (SM1.11) that only depends on $h, \Delta t, \rho^k$ and the domain, such that the minimizer of (SM1.1) cannot touch the boundary.

To prove the equivalence of the numerical scheme with the minimizer of the optimization problem (SM1.1), we follow Step 3 of the proof of Theorem 4.1 for the one dimensional case. We omit the details here. \square

Theorem 4.1 is then proved in a fashion similar to the proof of Theorem 3.5.

SM2. Proof of consistency. Here we present detailed calculations of the truncation error defined by

$$\begin{aligned} \tau_i^1 &= \frac{P_i^{k+1} - P_i^k}{\Delta t} + d_h(\hat{P}_i^k V_i^{k+1}), \\ \tau_i^2 &= D_h \log P_i^{k+1} - \frac{1}{\sum_{j=1}^n \hat{P}_j^k} \sum_{i=1}^n \hat{P}_i^k D_h \log P_i^{k+1} + \sum_{j=1}^n b_{ij} \hat{P}_j^k (V_i^{k+1} - V_j^{k+1}), \\ \tau_i^3 &= \sum_{i=1}^n \hat{P}_i^k V_i^{k+1}. \end{aligned}$$

We first calculate τ_i^1 .

$$\begin{aligned} \tau_{i,\ell}^1 &= \frac{P_{i,\ell}^{k+1} - P_{i,\ell}^k}{\Delta t} + d_h \left(\hat{P}_i^k V_i^{k+1} \right)_\ell \\ &= \frac{P_{i,\ell}^{k+1} - P_{i,\ell}^k}{\Delta t} + \frac{1}{2h} \left((P_{i,\ell}^k + P_{i,\ell+1}^k) V_{i,\ell+\frac{1}{2}}^{k+1} - (P_{i,\ell}^k + P_{i,\ell-1}^k) V_{i,\ell-\frac{1}{2}}^{k+1} \right). \end{aligned}$$

The terms in the above equation can be calculated using Taylor's expansion as

$$\begin{aligned} P_{i,\ell}^{k+1} &= P_{i,\ell}^k + \partial_t P_{i,\ell}^k \Delta t + O(\Delta t^2), \\ P_{i,\ell \pm 1}^k &= P_{i,\ell}^k \pm h \partial_x P_{i,\ell}^k + \frac{1}{2} h^2 \partial_{xx} P_{i,\ell}^k + O(h^3), \\ V_{i,\ell \pm \frac{1}{2}}^{k+1} &= V_{i,\ell}^k \pm \frac{1}{2} h \partial_x V_{i,\ell}^k + \Delta t \partial_t V_{i,\ell}^k + \frac{1}{4} h^2 \partial_{xx} V_{i,\ell}^k + \frac{1}{2} \Delta t^2 V_{i,\ell}^k \pm \frac{1}{2} h \Delta t \partial_{xt} V_{i,\ell}^k \\ &\quad + O(h^3 + \Delta t h^2 + \Delta t^2 h + \Delta t^3). \end{aligned}$$

Taking these expressions into the previous equation leads to

$$\begin{aligned} \tau_{i,\ell}^1 &= \partial_t P_{i,\ell}^k - \frac{1}{2h} \left(2h \partial_x P_{i,\ell}^k \left(V_{i,\ell}^k + \Delta t \partial_t V_{i,\ell}^k + \frac{1}{4} h^2 \partial_{xx} V_{i,\ell}^k + \frac{1}{2} \Delta t^2 V_{i,\ell}^k \right) \right) \\ &\quad - \frac{1}{2h} \left(2P_{i,\ell}^k + \frac{1}{2} h^2 \partial_{xx} P_{i,\ell}^k \right) (h \partial_x V_{i,\ell}^k + h \Delta t \partial_{xt} V_{i,\ell}^k) \\ &\quad + O(\Delta t + h^2 + \Delta t h + \Delta t^2 + \Delta t^3) \\ &= (\partial_t P - \partial_x(PV))_{i,\ell}^k + O(\Delta t + h^2 + \Delta t^2 + \Delta t h + \Delta t^3). \end{aligned}$$

The terms τ^2 and τ^3 can be also approximated again using the Taylor expansion. The results are

$$\begin{aligned}\tau_{i,\ell+\frac{1}{2}}^2 &= 0 + \frac{h}{2} \partial_x \left(\frac{\partial_x P_{i,\ell}^k}{P_{i,\ell}^k} - \sum_{j=1}^n b_{ij} P_{j,\ell}^k (V_{i,\ell}^k - V_{j,\ell}^k) \right) + O(\Delta t + h^2) \\ &= O(\Delta t + h^2).\end{aligned}$$

$$\begin{aligned}\tau_{i,\ell+\frac{1}{2}}^3 &= \sum_{i=1}^n P_{i,\ell}^k V_{i,\ell}^k + \frac{1}{2} h \sum_{i=1}^n \partial_x (P_{i,\ell}^k V_{i,\ell}^k) + \Delta t \sum_{i=1}^n P_{i,\ell}^k \partial_t V_{i,\ell}^k + O(\Delta t^2 + h^2) \\ &= O(\Delta t + h^2).\end{aligned}$$

In summary, we conclude the result stated in Lemma 2.5, i.e., there exists $C > 0$ depending on (P, V) so that

$$|\tau_{i,\ell}^1|, |\tau_{i,\ell+\frac{1}{2}}^2|, |\tau_{i,\ell+\frac{1}{2}}^3| \leq C(\Delta t + h^2).$$