ESAIM: COCV 27 (2021) 27
https://doi.org/10.1051/cocv/2021020

ESAIM: Control, Optimisation and Calculus of Variations
www.esaim-cocv.org

WHEN BINGHAM MEETS BRATU: MATHEMATICAL AND COMPUTATIONAL INVESTIGATIONS*

Frederick J. Foss II[1,**] and Roland Glowinski[2,3]

Abstract. In this article, we discuss the numerical solution of the Bingham-Bratu-Gelfand (BBG) problem, a non-smooth nonlinear eigenvalue problem associated with the total variation integral and an exponential nonlinearity. Using the fact that one can view the nonlinear eigenvalue as a possible Lagrange multiplier associated with a constrained minimization problem from Calculus of Variations, we associate with the BBG problem an initial value problem (dynamical flow), well suited to time-discretization by operator-splitting. Various mathematical results are proved, including the convergence of a finite element approximation of the BBG problem. The operator-splitting/finite element methodology discussed in this article is robust and easy to implement. We validate the implementation by first solving the classical Bratu-Gelfand problem, obtaining and reporting results consistent with those found in the literature. We then explore the full capability of the implementation by solving the viscoplastic BBG problem, obtaining and reporting results for several values of the plasticity yield. We conclude by exhibiting and discussing the bifurcation diagrams corresponding to these same values of the plasticity yield, and by reporting and examining some finer details of the solver discovered during the course of our investigation.

Mathematics Subject Classification. 35P30, 49M15, 65K15, 74S05.

Received January 18, 2020. Accepted February 12, 2021.

1. Introduction: motivation and preliminary results

Some years ago, the second author (RG) was spending a summer month at Oak Ridge National Laboratory (ORNL) working there on super-conductivity related problems. During his stay at ORNL, RG had the opportunity to attend a lecture given by our colleague Qiang Du (then at Penn State, now at Columbia) on *saddle-point* computations (by *mountain-pass type algorithms*, as far as we remember). As a possible challenging test problem, RG suggested that Q. Du have a look at

$$\left\{\begin{array}{l}\text{Find } \{u,\lambda\}\in H_0^1(\Omega)\times\mathbf{R}_+ \text{ so that}\\ \mu\int_\Omega \nabla u\cdot\nabla(v-u)\,\mathrm{d}x+\tau_y\left[\int_\Omega|\nabla v|\,\mathrm{d}x-\int_\Omega|\nabla u|\,\mathrm{d}x\right]\geq\lambda\int_\Omega e^u(v-u)\,\mathrm{d}x,\\ \forall v\in H_0^1(\Omega),\end{array}\right. \tag{1.1}$$

* This article is dedicated to H.B. Keller and J.L. Lions.

Keywords and phrases: Non-smooth nonlinear eigenvalue problem, bingham viscoplastic flow, exponential nonlinearity, multiple solutions, turning points, operator-splitting time-discretization schemes, finite element approximations, lagrange multipliers.

[1] Freestyle Analytical & Quantitative Services, LLC, 2210 Parkview Dr., Longmont, CO 80504, USA.
[2] Department of Mathematics, University of Houston, 4800 Calhoun Rd, Houston, TX 77204-3008, USA.
[3] Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong.

** Corresponding author: foss@airmail.net

where in (1.1): Ω is a bounded domain of $\mathbf{R}^2$ and μ and τ_y are two positive constants. The solutions of the variational inequality problem (1.1) can be considered as the steady state solutions of the nonlinear, non-smooth reaction-diffusion problem

$$\begin{cases} \dfrac{\partial u}{\partial t} - \mu\nabla^2 u - \tau_y \nabla \cdot \dfrac{\nabla u}{|\nabla u|} = \lambda e^u \text{ in } \Omega \times (0, +\infty), \\ u = 0 \text{ on } \partial\Omega \times (0, +\infty), \\ u(0) = u_0. \end{cases}$$

Indeed, if $\tau_y = 0$, (1.1) is nothing but a variational formulation of the celebrated *Bratu problem*

$$\begin{cases} -\mu\nabla^2 u = \lambda e^u \text{ in } \Omega, \\ u = 0 \text{ on } \partial\Omega, \end{cases} \tag{1.2}$$

a classical nonlinear reaction-diffusion problem from *solid combustion* (see [2] for details). The bifurcation diagram reported in Figure 1 (borrowed from [11]) corresponds to $\Omega = (0,\, 1)^2$ and $\mu = 1$ in (1.2); it provides the graph of the function $\lambda \to u_h(0.5, 0.5)$, u_h being a finite element-computed approximate solution.

Figure 1 shows the existence of a critical value λ_c of λ beyond which problem (1.2) has no (real) solution. If $\lambda \in (0, \lambda_c)$, problem (2) has two solutions. The lower branch corresponds to *local minimizers* over $H_0^1(\Omega)$ of the functional J defined by $J(v) = \frac{\mu}{2}\int_\Omega |\nabla v|^2 \,\mathrm{d}x - \lambda \int_\Omega e^v \mathrm{d}x$. The upper branch, on the other hand, corresponds to stationary solutions of the above functional. These solutions can be viewed as *saddle points* since the associated linearized operator, namely $v \to -\mu\nabla^2 v - \lambda e^u v$ has a finite number of negative eigenvalues (and of course a countable infinity of positive eigenvalues); the number of negative eigenvalues increases as λ decreases and converges to $+\infty$ as $\lambda \to 0_+$. Let us suppose now that in (1.1), one replaces $\lambda \int_\Omega e^u(v-u)\mathrm{d}x$ by $\lambda \int_\Omega (v-u)\mathrm{d}x$, then problem (1.1) reduces to

$$\begin{cases} \text{Find } u \in H_0^1(\Omega) \text{ such that} \\ \mu \int_\Omega \nabla u \cdot \nabla (v-u)\,\mathrm{d}x + \tau_y \left[\int_\Omega |\nabla v|\,\mathrm{d}x - \int_\Omega |\nabla u|\,\mathrm{d}x\right] \geq \lambda \int_\Omega (v-u)\,\mathrm{d}x, \\ \forall v \in H_0^1(\Omega). \end{cases} \tag{1.3}$$

Problem (1.3) is well posed and is a model for the steady flow of a viscoplastic material of the *Bingham* type in an infinitely long cylinder of cross-section Ω, μ and τ_y being the fluid viscosity and plasticity yield, respectively, λ the pressure drop per unit length, and u the flow axial velocity. Actually, (1.3) can be viewed as a saddle point problem since one can associate with its solution u a (possibly non-unique) vector-valued function $\mathbf{p}$ such that

$$\begin{cases} \{u, \mathbf{p}\} \in H_0^1(\Omega) \times \mathbf{P}, \\ \mu \int_\Omega \nabla u \cdot \nabla v \mathrm{d}x + \tau_y \int_\Omega \mathbf{p} \cdot \nabla v \mathrm{d}x = \lambda \int_\Omega v \mathrm{d}x, \forall v \in H_0^1(\Omega), \\ \mathbf{p} \cdot \nabla u = |\nabla u|, \end{cases} \tag{1.4}$$

with $\mathbf{P} = \{\mathbf{q} | \mathbf{q} \in (L^2(\Omega))^2, |\mathbf{q}(x)| \leq 1 \text{ a.e. in } \Omega\}$ and $|\mathbf{z}| = \sqrt{z_1^2 + z_2^2}, \forall \mathbf{z} = \{z_1, z_2\}$. It follows from, *e.g.*, Chapter 6 of [10], that (1.4) implies that the pair $\{u, \mathbf{p}\}$ is a *saddle point* over $H_0^1(\Omega) \times \mathbf{P}$ of the *Lagrangian* functional $\mathcal{L}\colon H_0^1(\Omega) \times (L^2(\Omega))^2 \to \mathbf{R}$ defined by

$$\mathcal{L}(v, \mathbf{q}) = \frac{\mu}{2}\int_\Omega |\nabla v|^2 \,\mathrm{d}x + \tau_y \int_\Omega \mathbf{q} \cdot \nabla v \mathrm{d}x - \lambda \int_\Omega v \mathrm{d}x.$$

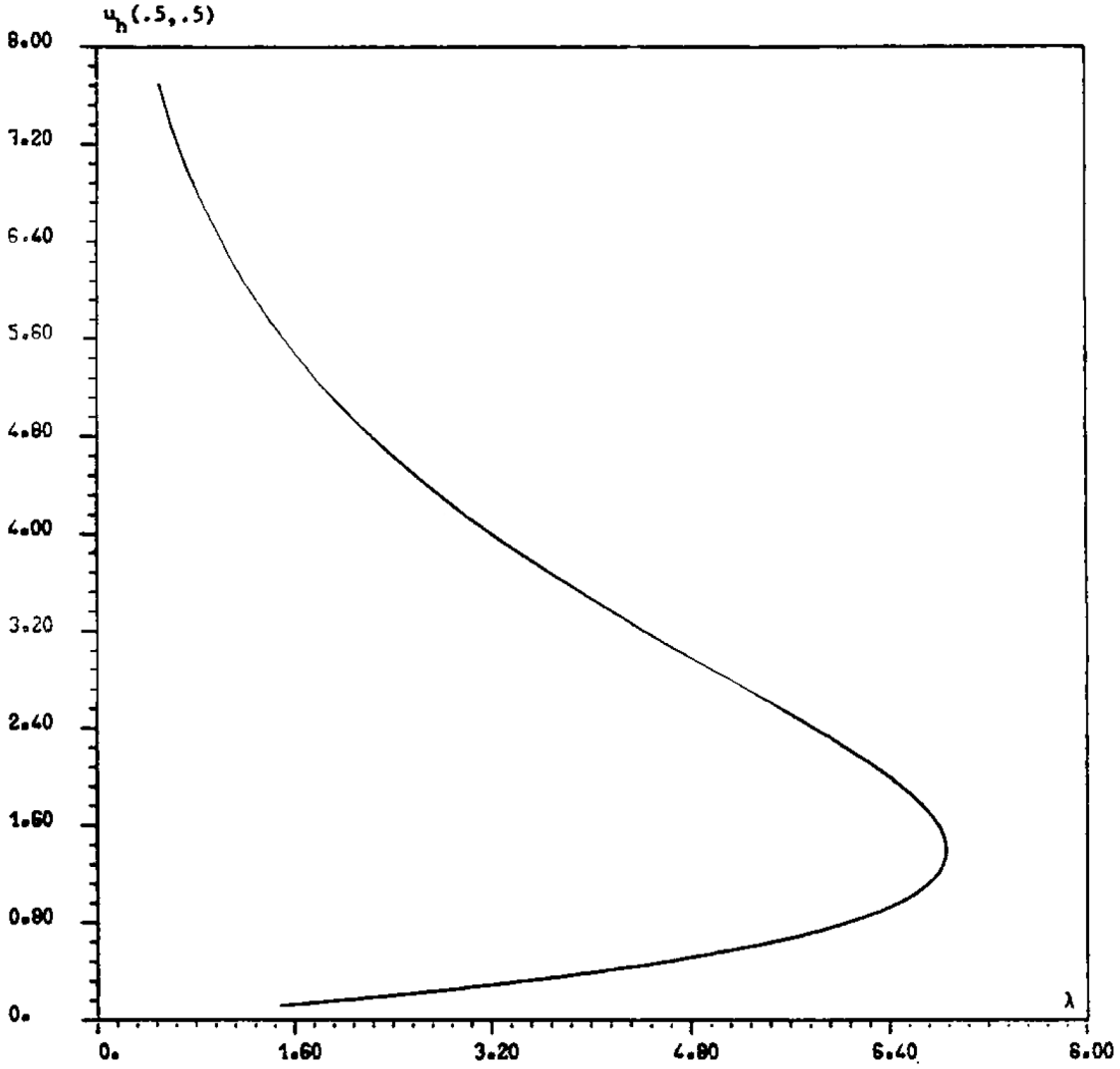

FIGURE 1. Bratu's problem bifurcation diagram ($\tau_y = 0, \mu = 1, \Omega = (0,1)^2$).

From the above considerations, one can also write problem (1.1) as

$$\begin{cases} \text{Find } \{u, \mathbf{p}, \lambda\} \in H_0^1(\Omega) \times \mathbf{P} \times \mathbf{R}_+ \text{ such that} \\ \mu \int_\Omega \nabla u \cdot \nabla v \mathrm{d}x + \tau_y \int_\Omega \mathbf{p} \cdot \nabla v \mathrm{d}x = \lambda \int_\Omega e^u v \mathrm{d}x, \forall v \in H_0^1(\Omega), \\ \mathbf{p} \cdot \nabla u = |\nabla u|, \end{cases} \tag{1.5}$$

which makes it a (kind of) *double saddle point problem*, explaining why we suggested it to our colleague Q. Du as a test problem worth investigating. If one has to classify problem (1.1), it makes sense to consider it as a nonlinear eigenvalue problem associated with a non-smooth elliptic operator. Indeed, formally, one can write

problem (1.1) as

$$\begin{cases} -\mu\nabla^2 u - \tau_y \nabla \cdot \dfrac{\nabla u}{|\nabla u|} = \lambda e^u \text{ in } \Omega, \\ u = 0 \text{ on } \partial\Omega. \end{cases} \tag{1.6}$$

Problem (1.6) clearly has the flavor of a *nonlinear non-smooth eigenvalue problem*. Formulation (1.6) of problem (1.1) is formal since it makes no sense on the subset of Ω where $\nabla u = \mathbf{0}$. An obvious way to overcome this difficulty is to approximate (1.6) by

$$\begin{cases} -\mu\nabla^2 u_\varepsilon - \tau_y \nabla \cdot \dfrac{\nabla u_\varepsilon}{\sqrt{\varepsilon^2 + |\nabla u_\varepsilon|^2}} = \lambda e^{u_\varepsilon} \text{ in } \Omega, \\ u_\varepsilon = 0 \text{ on } \partial\Omega, \end{cases} \tag{1.7}$$

where ε is a (small) positive parameter. The solutions of problem (1.7) can be computed by *continuation methods* (as done for example by D. Sorensen and the second author in some unpublished investigations); however this regularization approach has several drawbacks, among them: (i) The computational cost increases significantly with $1/\varepsilon$. (ii) It 'kills' the fact that the set $\{\{u, \lambda\} | u = 0, \lambda \in [0, \gamma\tau_y]\}$, with $\gamma = \inf\limits_{\phi \in H_0^1(\Omega)\setminus\{0\}} \dfrac{\int_\Omega |\nabla\phi| \, \mathrm{d}x}{\int_\Omega |\phi| \, \mathrm{d}x}$, forms a (trivial) branch of solutions to problem (1) (in [18] it is shown that $\gamma = 2/R$ for a disk of radius R, while $\gamma = (2 + \sqrt{\pi})/L$ for a square of side length L). This strongly suggests avoiding regularization à la (1.7) and instead addressing the solution of problem (1.1) directly. Actually, in [17] (see also [10], Chap. 7), the solution of a non-smooth eigenvalue problem reminiscent of problem (1.1) is discussed, namely

$$\begin{cases} \{u, \lambda\} \in H_0^1(\Omega) \times \mathbf{R}_+, \\ \mu \int_\Omega \nabla u \cdot \nabla(v - u) \mathrm{d}x + \tau_y \left[\int_\Omega |\nabla v| \mathrm{d}x - \int_\Omega |\nabla u| \mathrm{d}x \right] \geq \lambda \int_\Omega u(v - u) \mathrm{d}x, \\ \forall v \in H_0^1(\Omega), \\ \int_\Omega |u|^2 \mathrm{d}x = |\Omega| U^2, \end{cases} \tag{1.8}$$

with $|\Omega|$ = measure of Ω and U a constant with dimension of velocity. The method advocated in the two above references relies on *operator-splitting* and produces an algorithm closely related to the *inverse power method with shift* for computing the eigenvalues of real symmetric matrices. We will take a similar approach to solve problem (1.1), first showing then taking advantage of the fact that in problem (1.1), λ is possibly a Lagrange multiplier for the following constrained minimization problem

$$\begin{cases} u \in S_C, \\ J(u) \leq J(v), \forall v \in S_C, \end{cases} \tag{1.9}$$

with $C \in (0, +\infty)$ and

$$J(v) = \tfrac{\mu}{2} \int_\Omega |\nabla v|^2 \, \mathrm{d}x + \tau_y \int_\Omega |\nabla v| \, \mathrm{d}x, \tag{1.10}$$

$$S_C = \{v | v \in H_0^1(\Omega), \textstyle\int_\Omega (e^v - 1) \mathrm{d}x = C\}. \tag{1.11}$$

Proving that problem (1.9) has a solution is simple since: (i) The functional J is continuous, convex and coercive over $H_0^1(\Omega)$ (the last property meaning that $\lim_{\|v\|_{H_0^1(\Omega)} \to +\infty} J(v) = +\infty$). (ii) Next, the set S_C is clearly non-empty. (iii) Finally, the compactness of the nonlinear operator $v \to \mathrm{e}^v \colon H_0^1(\Omega) \to L^s(\Omega), \forall s \in [1, +\infty)$ (see [1]) implies that S_C is weakly closed in $H_0^1(\Omega)$. From these properties we can show, for $C > 0$, the existence of a minimizing sequence converging to a solution of (1.9).

The first result we are going to prove is that any solution of problem (1.9) is necessarily non-negative. This result comes from the following

Theorem 1.1. *Assume that u is solution of problem (1.9) with $C \in (0, +\infty)$. Then u is non-negative.*

Proof. Let u be solution of problem (1.9) and suppose that u does not verify $u \geq 0$; we have then

$$|\nabla |u|\,| = |\nabla u|\,, \text{ and } \int_\Omega (e^{|u|} - 1)\mathrm{d}x > \int_\Omega (e^u - 1)\mathrm{d}x,$$

implying

$$J(|u|) = J(u), \text{ and } \int_\Omega (e^{|u|} - 1)\mathrm{d}x > C. \tag{1.12}$$

Since the function $\alpha \to \int_\Omega (e^{\alpha|u|} - 1)\mathrm{d}x$ is continuous, strictly increasing and maps $[0, +\infty)$ onto $[0, +\infty)$, there exists a unique $\alpha_0 \in (0, 1)$ such that $\int_\Omega (e^{\alpha_0|u|} - 1)\mathrm{d}x = C$. Denote $\alpha_0|u|$ by w_0; we have then

$$J(w_0) = \alpha_0^2 \frac{\mu}{2} \int_\Omega |\nabla u|^2 \,\mathrm{d}x + \alpha_0 \tau_y \int_\Omega |\nabla u| \,\mathrm{d}x < J(u), \text{ and } \int_\Omega (e^{w_0} - 1)\mathrm{d}x = C,$$

contradicting the fact that u minimizes J over S_C, thus necessarily $u \geq 0$. □

Next, we are going to prove what we consider the most important theoretical result of this article, namely, the following

Theorem 1.2. *Suppose that for $C > 0$ given, u is a solution of the minimization problem (1.9). There exists then $\lambda > 0$ so that the pair $\{u, \lambda\}$ is a solution of problem (1.1).*

Proof. The following proof is conceptually simple and relies on a *penalization-regularization* approach: with $\varepsilon > 0$, we associate the pair $\{\varepsilon_1, \varepsilon_2\}$, where $\varepsilon_1 = \varepsilon$ and $\varepsilon_2 = \varepsilon^{1+\alpha}$, $\alpha > 0$. Next, we consider the following variant of problem (1.9):

$$\begin{cases} u_\varepsilon \in S_C, \\ J_\varepsilon(u_\varepsilon) \leq J_\varepsilon(v), \forall v \in S_C, \end{cases} \tag{1.13}$$

where

$$J_\varepsilon(v) = \frac{1}{2\varepsilon_1} \int_\Omega |\nabla(v - u)|^2 \,\mathrm{d}x + \tfrac{\mu}{2} \int_\Omega |\nabla v|^2 \,\mathrm{d}x + \tau_y \int_\Omega \sqrt{\varepsilon_2^2 + |\nabla v|^2}\mathrm{d}x. \tag{1.14}$$

The approach we used to prove the existence of a solution to problem (1.9) applies also to problem (1.13), implying the existence of a minimizer u_ε for J_ε over S_C. The following chain of inequalities holds (with $|\Omega|$ = measure of Ω):

$$\begin{aligned}
&\frac{\mu}{2} \int_\Omega |\nabla u|^2 \,\mathrm{d}x + \tau_y \int_\Omega |\nabla u| \,\mathrm{d}x + \tau_y |\Omega| \varepsilon_2 \\
&\geq \frac{\mu}{2} \int_\Omega |\nabla u|^2 \,\mathrm{d}x + \tau_y \int_\Omega \sqrt{\varepsilon_2^2 + |\nabla u|^2}\mathrm{d}x \\
&\geq \frac{1}{2\varepsilon_1} \int_\Omega |\nabla(u_\varepsilon - u)|^2 \,\mathrm{d}x + \tfrac{\mu}{2} \int_\Omega |\nabla u_\varepsilon|^2 \,\mathrm{d}x + \tau_y \int_\Omega \sqrt{\varepsilon_2^2 + |\nabla u_\varepsilon|^2}\mathrm{d}x \\
&> \frac{1}{2\varepsilon_1} \int_\Omega |\nabla(u_\varepsilon - u)|^2 \,\mathrm{d}x + \tfrac{\mu}{2} \int_\Omega |\nabla u_\varepsilon|^2 \,\mathrm{d}x + \tau_y \int_\Omega |\nabla u_\varepsilon| \,\mathrm{d}x \\
&\geq \frac{1}{2\varepsilon_1} \int_\Omega |\nabla(u_\varepsilon - u)|^2 \,\mathrm{d}x + \tfrac{\mu}{2} \int_\Omega |\nabla u|^2 \,\mathrm{d}x + \tau_y \int_\Omega |\nabla u| \,\mathrm{d}x,
\end{aligned}$$

implying

$$\int_\Omega |\nabla (u_\varepsilon - u)|^2 \, \mathrm{d}x \leq 2\tau_y |\Omega| \varepsilon_1 \varepsilon_2. \tag{1.15}$$

We have thus proved that

$$\lim_{\varepsilon \to 0} u_\varepsilon = u \text{ in } H_0^1(\Omega). \tag{1.16}$$

Let us denote the functional $v \to \int_\Omega (e^v - 1)\mathrm{d}x - C$: $H_0^1(\Omega) \to \mathbf{R}$ by G; functional G is differentiable, its differential DG at v being given by

$$\langle DG(v), w \rangle = \int_\Omega e^v w \mathrm{d}x, \forall v, w \in H_0^1(\Omega), \tag{1.17}$$

implying

$$DG(v) \neq 0, \forall v \in H_0^1(\Omega). \tag{1.18}$$

Since J_ε and G are both C^1 over $H_0^1(\Omega)$, and DG verifies (1.17), (1.18), one can associate with u_ε a *Lagrange multiplier* λ_ε such that the pair $\{u_\varepsilon, \lambda_\varepsilon\}$ verifies

$$\begin{cases} \{u_\varepsilon, \lambda_\varepsilon\} \in H_0^1(\Omega) \times \mathbf{R}, \\ \dfrac{1}{\varepsilon_1} \int_\Omega \nabla (u_\varepsilon - u) \cdot \nabla v \mathrm{d}x + \mu \int_\Omega \nabla u_\varepsilon \cdot \nabla v \mathrm{d}x + \tau_y \int_\Omega \dfrac{\nabla u_\varepsilon}{\sqrt{\varepsilon_2^2 + |\nabla u_\varepsilon|^2}} \cdot \nabla v \mathrm{d}x = \lambda_\varepsilon \int_\Omega e^{u_\varepsilon} v \mathrm{d}x, \\ \forall v \in H_0^1(\Omega), \\ \int_\Omega (e^{u_\varepsilon} - 1)\, \mathrm{d}x = C. \end{cases} \tag{1.19}$$

Take $v = u_\varepsilon$ in (1.19); we have then

$$\frac{1}{\varepsilon_1} \int_\Omega \nabla (u_\varepsilon - u) \cdot \nabla u_\varepsilon \mathrm{d}x + \mu \int_\Omega |\nabla u_\varepsilon|^2 \, \mathrm{d}x + \tau_y \int_\Omega \frac{|\nabla u_\varepsilon|^2}{\sqrt{\varepsilon_2^2 + |\nabla u_\varepsilon|^2}} \mathrm{d}x = \lambda_\varepsilon \int_\Omega e^{u_\varepsilon} u_\varepsilon \mathrm{d}x. \tag{1.20}$$

Several observations are in order concerning (1.20), namely:

(a) It follows from (1.16) that there exists $M > 0$ such that $\|u_\varepsilon\|_{H_0^1(\Omega)} \leq M, \forall \varepsilon > 0$. Combining the above inequality with (1.15), we obtain

$$\frac{1}{\varepsilon_1} \left| \int_\Omega \nabla (u_\varepsilon - u) \cdot \nabla u_\varepsilon \mathrm{d}x \right| \leq \frac{1}{\varepsilon_1} \|u_\varepsilon - u\|_{H_0^1(\Omega)} \|u_\varepsilon\|_{H_0^1(\Omega)} \leq \sqrt{2\tau_y |\Omega|} M \sqrt{\frac{\varepsilon_2}{\varepsilon_1}} = \sqrt{2\tau_y |\Omega|} M \varepsilon^{\alpha/2},$$

which implies in turn that

$$\lim_{\varepsilon \to 0} \frac{1}{\varepsilon_1} \int_\Omega \nabla (u_\varepsilon - u) \cdot \nabla u_\varepsilon \mathrm{d}x = 0. \tag{1.21}$$

(b) It follows from (1.16) and from Theorem 1.1 that

$$\lim_{\varepsilon \to 0} \int_\Omega e^{u_\varepsilon} u_\varepsilon \mathrm{d}x = \int_\Omega e^u u \mathrm{d}x > 0, \tag{1.22}$$

which implies

$$\int_\Omega e^{u_\varepsilon} u_\varepsilon \mathrm{d}x > 0 \,\text{for}\, \varepsilon \,\text{small enough.} \tag{1.23}$$

(c) Combining the relation

$$\int_\Omega \frac{|\nabla u_\varepsilon|^2}{\sqrt{\varepsilon_2^2 + |\nabla u_\varepsilon|^2}} \mathrm{d}x = \int_\Omega |\nabla u_\varepsilon| \, \mathrm{d}x + \int_\Omega |\nabla u_\varepsilon| \left[\frac{|\nabla u_\varepsilon|}{\sqrt{\varepsilon_2^2 + |\nabla u_\varepsilon|^2}} - 1 \right] \mathrm{d}x$$

with

$$-\varepsilon_2 < \xi \left[\frac{\xi}{\sqrt{\varepsilon_2^2 + \xi^2}} - 1 \right] \leq 0, \forall \xi \geq 0$$

and (1.16), we obtain

$$\lim_{\varepsilon \to 0} \int_\Omega \frac{|\nabla u_\varepsilon|^2}{\sqrt{\varepsilon_2^2 + |\nabla u_\varepsilon|^2}} \mathrm{d}x = \int_\Omega |\nabla u| \, \mathrm{d}x. \tag{1.24}$$

It follows from (1.20)–(1.24) that

$$\lim_{\varepsilon \to 0} \lambda_\varepsilon = \frac{\mu \int_\Omega |\nabla u|^2 \, \mathrm{d}x + \tau_y \int_\Omega |\nabla u| \, \mathrm{d}x}{\int_\Omega e^u u \mathrm{d}x} > 0. \tag{1.25}$$

We will denote by λ the ratio in (1.25) (a kind of *Rayleigh quotient*). Proving that the pair $\{u, \lambda\}$ verifies (1.1) is rather simple now: Indeed, it follows from (1.19) that

$$\begin{cases} \frac{1}{\varepsilon_1} \int_\Omega \nabla (u_\varepsilon - u) \cdot \nabla (v - u_\varepsilon) \, \mathrm{d}x + \mu \int_\Omega \nabla u_\varepsilon \cdot \nabla (v - u_\varepsilon) \, \mathrm{d}x + \\ \tau_y \int_\Omega \frac{\nabla u_\varepsilon}{\sqrt{\varepsilon_2^2 + |\nabla u_\varepsilon|^2}} \cdot \nabla (v - u_\varepsilon) \, \mathrm{d}x = \lambda_\varepsilon \int_\Omega e^{u_\varepsilon} (v - u_\varepsilon) \, \mathrm{d}x, \\ \forall v \in H_0^1(\Omega). \end{cases} \tag{1.26}$$

From the *convexity* of the functional $v \to \int_\Omega \sqrt{\varepsilon_2^2 + |\nabla v|^2} \mathrm{d}x$ we have

$$\varepsilon_2 |\Omega| + \int_\Omega |\nabla v| \, \mathrm{d}x - \int_\Omega |\nabla u_\varepsilon| \, \mathrm{d}x > \int_\Omega \sqrt{\varepsilon_2^2 + |\nabla v|^2} \mathrm{d}x - \int_\Omega \sqrt{\varepsilon_2^2 + |\nabla u_\varepsilon|^2} \mathrm{d}x \geq \int_\Omega \frac{\nabla u_\varepsilon}{\sqrt{\varepsilon_2^2 + |\nabla u_\varepsilon|^2}} \cdot \nabla (v - u_\varepsilon) \, \mathrm{d}x,$$

which combined with (1.26) implies

$$\begin{cases} \frac{1}{\varepsilon_1} \int_\Omega \nabla (u_\varepsilon - u) \cdot \nabla (v - u_\varepsilon) \, \mathrm{d}x + \mu \int_\Omega \nabla u_\varepsilon \cdot \nabla (v - u_\varepsilon) \, \mathrm{d}x + \\ \tau_y \varepsilon_2 |\Omega| + \tau_y \int_\Omega |\nabla v| \, \mathrm{d}x - \tau_y \int_\Omega |\nabla u_\varepsilon| \, \mathrm{d}x \geq \lambda_\varepsilon \int_\Omega e^{u_\varepsilon} (v - u_\varepsilon) \, \mathrm{d}x, \\ \forall v \in H_0^1(\Omega). \end{cases} \tag{1.27}$$

Taking the limit in (1.27), we obtain

$$\mu \int_\Omega \nabla u \cdot \nabla (v-u)\,\mathrm{d}x + \tau_y \int_\Omega |\nabla v|\,\mathrm{d}x - \tau_y \int_\Omega |\nabla u|\,\mathrm{d}x \geq \lambda \int_\Omega e^u (v-u)\,\mathrm{d}x,$$

which completes the proof of the theorem. □

Remark 1.3. Let u be a solution of problem (9). Since u verifies (1) with $e^u \in L^s(\Omega), \forall s \in [1,+\infty)$, it follows from [3] that $u \in H_0^1(\Omega) \cap H^2(\Omega)$, and that it belongs also to $C^0(\overline{\Omega})$. Actually, the numerical results reported in Section 4 suggest that $u \in H_0^1(\Omega) \cap W^{2,\infty}(\Omega)$.

If u is solution of problem (1.9), with $C > 0$, there exists $\lambda > 0$ so that the pair (u, λ) is solution of

$$\left\{ \begin{array}{l} \{u,\lambda\} \in H_0^1(\Omega) \times \mathbf{R}_+, \\ \mu \int_\Omega \nabla u \cdot \nabla (v-u)\,\mathrm{d}x + \tau_y \left[\int_\Omega |\nabla v|\mathrm{d}x - \int_\Omega |\nabla u|\mathrm{d}x\right] \geq \lambda \int_\Omega e^u (v-u)\,\mathrm{d}x, \\ \forall v \in H_0^1(\Omega), \\ \int_\Omega (e^u - 1)\,\mathrm{d}x = C, \end{array} \right. \tag{1.28}$$

or equivalently of

$$\left\{ \begin{array}{l} \{u, \mathbf{p}, \lambda\} \in H_0^1(\Omega) \times \mathbf{\Lambda} \times \mathbf{R}_+, \\ \mu \int_\Omega \nabla u \cdot \nabla v \mathrm{d}x + \tau_y \int_\Omega \mathbf{p} \cdot \nabla v \mathrm{d}x = \lambda \int_\Omega e^u v \mathrm{d}x, \forall v \in H_0^1(\Omega), \\ \mathbf{p} \cdot \nabla u = |\nabla u|, \\ \int_\Omega (e^u - 1)\,\mathrm{d}x = C. \end{array} \right. \tag{1.29}$$

Remark 1.4. If u is a solution of the non-convex minimization problem (1.9), one has necessarily (1.28), (1.29) as shown in Theorem 1.2, but of course the converse may not be true.

We will discuss in Section 2 the solution of problems (1.9), (1.28), (1.29) by an operator-splitting method. Next, we will address in Section 3 the finite element implementation of the above iterative method. Finally, we will report in Section 4 the results of numerical experiments.

Actually, before moving to the computational part of this article, we are going to prove a mathematical result of practical interest since it provides an estimate of the maximal value taken by the parameter λ in (1.28), (1.29) as C varies over $\mathbf{R}_+$, μ and τ_y being fixed. This result shows also that $\lambda \to 0$ when $C \to +\infty$, a well-documented result if $\tau_y = 0$ (see Fig. 1 for the visualization of a particular case). We have thus the following

Theorem 1.5. *Assume that μ and τ_y are fixed; then if the pair $\{u, \lambda\}$ is a solution of problem (1.9), (1.28) we have*

$$\lambda \leq \lambda_{MAX}(\mu, \tau_y) < +\infty, \forall C \geq 0, \tag{1.30}$$

and

$$\lim_{C \to +\infty} \lambda = 0. \tag{1.31}$$

Proof. For clarity, we have divided the proof into several steps. We assume also, without loss of generality, that $|\Omega| = 1$.

(i) Let us consider the following elliptic variational inequality:

$$\left\{ \begin{array}{l} \psi \in H_0^1(\Omega), \\ \mu \int_\Omega \nabla \psi \cdot \nabla (\phi - \psi)\,\mathrm{d}x + \tau_y \left[\int_\Omega |\nabla \phi|\mathrm{d}x - \int_\Omega |\nabla \psi|\mathrm{d}x\right] \geq \int_\Omega f(\phi - \psi)\,\mathrm{d}x, \forall \phi \in H_0^1(\Omega). \end{array} \right. \tag{1.32}$$

Suppose that $f \in L^2(\Omega)$ and is non-negative. It follows then from, *e.g.*, [8, 9, 12] that problem (1.32) has a unique solution, this solution being non-negative; actually ([3]), this solution has $H^2(\Omega)$-regularity (and therefore belongs to $C^0(\overline{\Omega})$) if $\partial\Omega$ is smooth and/or Ω is convex. Since $\lambda e^u \geq 0$ in (1.28) the above result implies the non-negativity of u (take $\psi = u$ and $f = \lambda e^u$ in (1.32)).(ii) Take now $v = 0$ and $v = 2u$ in (1.28). We obtain then

$$\mu \int_\Omega |\nabla u|^2 \mathrm{d}x + \tau_y \int_\Omega |\nabla u| \mathrm{d}x = \lambda \int_\Omega e^u u \mathrm{d}x. \tag{1.33}$$

It follows from (1.9)–(1.11) and (1.33) that

$$\begin{cases} 2J(v) \geq 2J(u) = \mu \int_\Omega |\nabla u|^2 \mathrm{d}x + 2\tau_y \int_\Omega |\nabla u| \mathrm{d}x > \\ \mu \int_\Omega |\nabla u|^2 \mathrm{d}x + \tau_y \int_\Omega |\nabla u| \mathrm{d}x = \lambda \int_\Omega e^u u dx, \\ \forall v \in S_C. \end{cases} \tag{1.34}$$

(iii) Observe that

$$\int_\Omega e^u u dx = \int_\Omega e^u \mathrm{d}x + \int_\Omega e^u \left(u-1\right) \mathrm{d}x = C + 1 + \int_{\Omega^+} e^u \left(u-1\right) \mathrm{d}x + \int_{\Omega^-} e^u \left(u-1\right) \mathrm{d}x, \tag{1.35}$$

where $\Omega^+ = \{x | x \in \Omega, u(x) > 1\}$ and $\Omega^- = \{x | x \in \Omega, u(x) < 1\}$. On Ω^+ (resp., Ω^-) one has $e^u(u-1) > 0$ (resp., $0 < e^u(1-u) \leq 1$)), which imply, combined with (1.34), (1.35), that

$$2J\left(v\right) > \lambda C, \forall v \in S_C. \tag{1.36}$$

(iv) Consider $w_0 \in H_0^1(\Omega) \cap C^0(\overline{\Omega})$ and verifying $w_0 \geq 0$ and $\int_\Omega w_0 \mathrm{d}x = 1$. Next, consider the function

$$\alpha \to \int_\Omega \left(e^{\alpha w_0} - 1\right) \mathrm{d}x.$$

This continuous function is strictly increasing and varies from 0 to $+\infty$ when α varies from 0 to $+\infty$. This property implies the existence (and the uniqueness) of a constant α_0 (≥ 0) such that $\int_\Omega (e^{\alpha_0 w_0} - 1) \mathrm{d}x = C$. Using the relations $|\Omega| = 1, \int_\Omega w_0 \mathrm{d}x = 1$ and the convexity of the exponential function we obtain (from the *Jensen inequality* for integrals) that

$$\exp\left(\alpha_0\right) = \exp\left(\alpha_0 \int_\Omega w_0 \mathrm{d}x\right) = \exp\left(\int_\Omega \alpha_0 w_0 \mathrm{d}x\right) \leq \int_\Omega e^{\alpha_0 w_0} \mathrm{d}x = 1 + C,$$

which implies in turn that

$$\alpha_0 \leq \ln\left(1 + C\right). \tag{1.37}$$

(v) Take $v = \alpha_0 w_0$ in (1.36). It follows then from (1.37) and from the Cauchy-Schwarz inequality that

$$\begin{aligned} \lambda &\leq \frac{2}{C} J\left(\alpha_0 w_0\right) = \frac{1}{C} \alpha_0 \left[\mu \alpha_0 \int_\Omega |\nabla w_0|^2 \mathrm{d}x + 2\tau_y \int_\Omega |\nabla w_0| \mathrm{d}x\right] \\ &\leq \frac{1}{C} \alpha_0 \|\nabla w_0\| \left[\mu \alpha_0 \|\nabla w_0\| + 2\tau_y\right] \\ &\leq \frac{1}{C} \ln\left(1 + C\right) \|\nabla w_0\| \left[\mu \ln\left(1 + C\right) \|\nabla w_0\| + 2\tau_y\right], \end{aligned} \tag{1.38}$$

with $\|\nabla w_0\| = \sqrt{\int_\Omega |\nabla w_0|^2 \mathrm{d}x}$ $(= \|w_0\|_{H_0^1(\Omega)})$.

Very clearly, (1.38) implies relations (1.30) and (1.31). To obtain an upper bound of $\lambda_{MAX}(\mu,\tau)$, we consider first the case $C \in [0, e-1]$: It follows from (1.38) and from the relation $\ln(1+C) \leq C, \forall C$, that

$$\lambda \leq \|\nabla w_0\| \left[\|\nabla w_0\| \mu + 2\tau_y\right], \forall C \in [0, e-1]. \tag{1.39}$$

Suppose now that $C > e-1$: Since $1 \leq \ln(1+C) \leq \ln(2C), \forall C \geq e-1$, it follows from (1.38) that

$$\lambda \leq \frac{8}{e^2} \|\nabla w_0\| \left[\|\nabla w_0\| \mu + 2\tau_y\right], \forall C > e-1. \tag{1.40}$$

Considering that $8e^{-2} = 1.082682....$, it follows from (1.39) and (1.40) that the upper bound for $\lambda_{MAX}(\mu,\tau)$ given by

$$\frac{8}{e^2} \|\nabla w_0\| \left[\|\nabla w_0\| \mu + 2\tau_y\right] \tag{1.41}$$

covers the full range of the values of C. □

Remark 1.6. Considering the way the upper bound given by (1.41) has been obtained, we doubt that it provides a sharp estimate of the maximal value $\lambda_{MAX}(\mu,\tau)$ of λ. Another difficulty is finding the right function w_0. A natural choice in that direction is to take for w_0 the function minimizing $\|\nabla\phi\|$ over the affine space $\{\phi | \phi \in H_0^1(\Omega), \int_\Omega \phi \mathrm{d}x = 1\}$; it is a simple exercise to show that this function is given by

$$w_0 = \frac{\tilde{w}_0}{\int_\Omega \tilde{w}_0 \mathrm{d}x}, \tag{1.42}$$

$\tilde{w}_0$ being the unique solution in $H_0^1(\Omega)$ of the Dirichlet problem

$$\begin{cases} -\nabla^2 \tilde{w}_0 = 1 \text{ in } \Omega, \\ \tilde{w}_0 = 0 \text{ on } \partial\Omega. \end{cases} \tag{1.43}$$

We have then

$$\|\nabla w_0\| = \frac{1}{\sqrt{\int_\Omega \tilde{w}_0 \mathrm{d}x}}. \tag{1.44}$$

In order to test the quality of estimate (1.41), we consider the case where $\Omega = (0,1)^2, \mu = 1$ and $\tau_y = 0$. Let us define w_0 by (1.42), (1.43). After space discretization, using a uniform mesh with $h = 1/64$, we obtain $\|\nabla w_0\|^2 \approx 28.4767$. Plugging the above w_0 into (1.41) we obtain the value 30.831. Since the actual value of $\lambda_{MAX}(1,0)$ is close to 6.81 (as shown in Fig. 1), this shows that in that particular case a big gap exists between the actual maximal value of λ and the upper bound given by (1.41).

Remark 1.7. The solution of nonlinear non-smooth elliptic variational problems with multiple solutions is discussed in [21, 22] (see also the references therein). The computational methods discussed in the above references should apply to the solution of problem (1.1). These methods rely on the *mountain pass theorem*, a powerful tool indeed, but rather complicated conceptually (see our comment in [10], Chap. 7), explaining why we decided, eventually, to use a simpler method, well suited to the specificities of the problem under consideration, this method being based on *operator-splitting*. In [6, 7], the authors applied operator-splitting methods to the numerical solution of nonlinear eigenvalue problems for elliptic operators.

Remark 1.8. It has been our experience that, very often, a well-chosen toy problem may help with predicting the qualitative behavior of the solutions of a much more complicated one we intend to solve. Problem (1.1) is no exception, an associated toy problem being

$$\begin{cases} \text{Find } \{u,\lambda\} \in \mathbf{R} \times \mathbf{R}_+ \text{ such that} \\ u\,(v-u) + \tau\,[|v| - |u|] \geq \lambda e^u\,(v-u)\,, \forall v \in \mathbf{R}, \end{cases} \tag{1.45}$$

with $\tau \geq 0$. Actually, problem (1.45) has solutions, these solutions verifying

$$u^2 + \tau|u| = \lambda u e^u. \tag{1.46}$$

Since $\lambda \geq 0$, relation (1.46) implies $u \geq 0$. On the other hand, the set $\{\{0,\lambda\}\}_{\lambda\in[0,\tau]} (= \{0\} \times [0,\,\tau])$ consists of (trivial) solutions to (1.45), the branch of non-trivial solutions being defined by

$$\lambda = e^{-u}\,(u+\tau)\,, u \geq 0. \tag{1.47}$$

Some non-trivial branches of solutions have been reported in Figure 2, which shows that for $\tau \geq 1$ the turning point is located on the positive part of the λ– axis. The numerical results of solving problem (1.1) reported in Section 4, while a bit more complicated and subtle, essentially exhibit similar qualitative behavior to this toy problem as τ_y varies.

2. An operator-splitting scheme for the solution of problem (1.9), (1.28)

2.1. An operator-splitting scheme

A warning is in order: Indeed, several parts of this second section are mathematically formal; their justification deserves further investigations. Problem (1.28) being a non-smooth nonlinear eigenvalue problem, various methods are natural candidates to solve it, among them *semi-smooth Newton's* and *mountain pass* algorithms. Actually, we are going to employ a method based on *operator-splitting*, which is nothing but a relatively simple generalization of the *inverse power method with shift* for the solution of linear eigenvalue problems. Our starting point is to associate with (1.28) the following initial value problem:

$$\begin{cases} (u(t),\lambda(t)) \in H_0^1(\Omega) \times \mathbf{R}_+, 0 < t < +\infty, \\ \begin{cases} \left\langle \dfrac{\partial u}{\partial t}, v - u(t) \right\rangle + \mu \int_\Omega \nabla u \cdot \nabla (v - u(t)) \mathrm{d}x \\ +\tau_y \left[\int_\Omega |\nabla v| \mathrm{d}x - \int_\Omega |\nabla u(t)| \mathrm{d}x \right] \geq \lambda(t) \int_\Omega e^{u(t)} (v - u(t)) \mathrm{d}x, \forall v \in H_0^1(\Omega), \\ \int_\Omega (e^{u(t)} - 1) \mathrm{d}x = C, \\ u(0) = u_0. \end{cases} \end{cases} \tag{2.1}$$

Above, $\langle \cdot, \cdot \rangle$ denotes the duality pairing between $H^{-1}(\Omega)$ (the dual space of $H_0^1(\Omega)$) and $H_0^1(\Omega)$, and $u(t)$ the function $x \to u(x,t)$. An obvious choice for u_0 is $u_0 = 0$, but more sophisticated alternatives do exist, as shown hereafter. From the non-negativity of the solutions of problem (1.9), (1.28), one can replace $H_0^1(\Omega)$ by $H_0^1(\Omega) \cap L^2_+(\Omega)$ in (2.1) (with $L^2_+(\Omega) = \{\phi | \phi \in L^2(\Omega), \phi \geq 0\}$).

Following [10] and [13], we advocate using the *Marchuk-Yanenko scheme* for its robustness and simplicity. Let Δt (> 0) be a time-discretization step; using a three-operator decomposition, we obtain (among other possible schemes) the following time-discretization of problem (2.1):

$$u^0 = u_0. \tag{2.2}$$

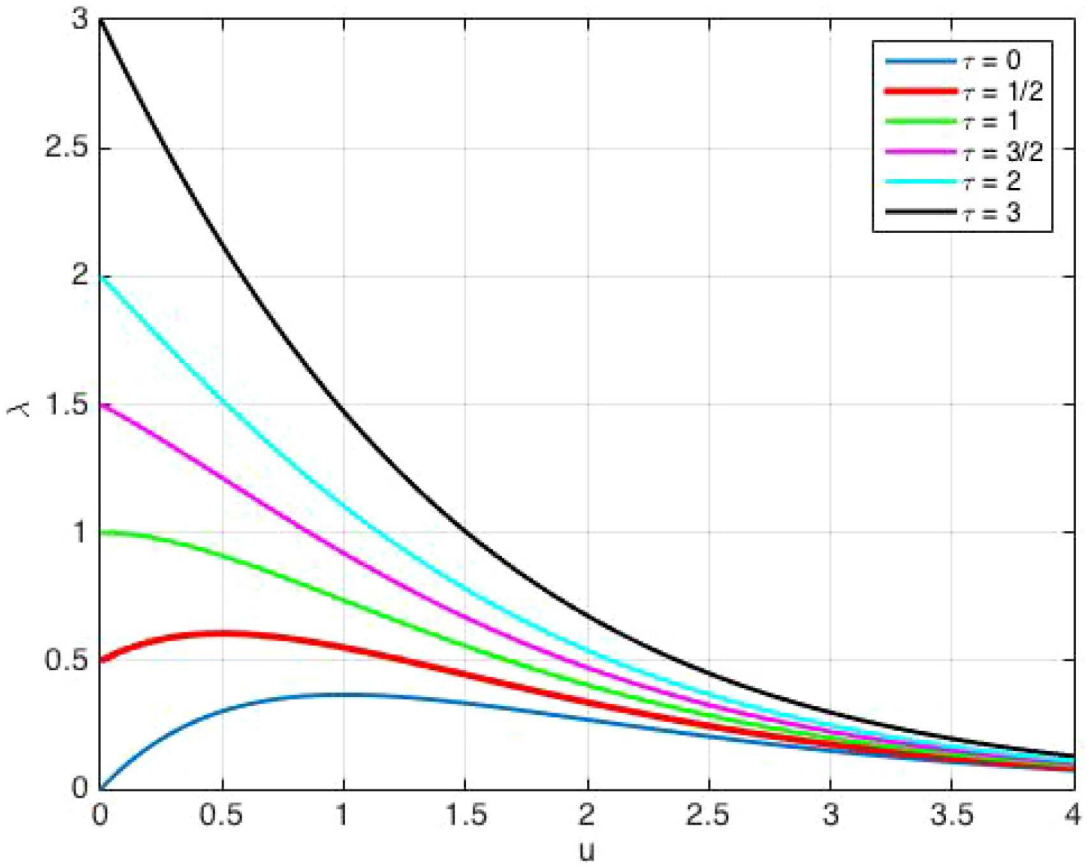

FIGURE 2. Visualization of the solutions of problem (1.45) for various values of τ.

For $n \geq 0$, $u^n \to u^{n+1/3} \to u^{n+2/3} \to u^{n+1}$ *via* the solution of

$$\begin{cases} u^{n+1/3} \in H_0^1(\Omega), \\ \int_\Omega \dfrac{u^{n+1/3}-u^n}{\Delta t}(v-u^{n+1/3})\mathrm{d}x + \mu \int_\Omega \nabla u^{n+1/3} \cdot \nabla (v-u^{n+1/3})\mathrm{d}x + \tau_y \left[\int_\Omega |\nabla v|\mathrm{d}x - \int_\Omega |\nabla u^{n+1/3}|\mathrm{d}x\right] \geq 0, \\ \forall v \in H_0^1(\Omega), \end{cases} \tag{2.3}$$

$$\begin{cases} \dfrac{u^{n+2/3}-u^{n+1/3}}{\Delta t} = \lambda^{n+2/3} e^{u^{n+2/3}}, \\ \int_\Omega (e^{u^{n+2/3}}-1)\mathrm{d}x = C, \end{cases} \tag{2.4}$$

$$\begin{cases} u^{n+1} \in L^2_+(\Omega), \\ \int_\Omega \dfrac{u^{n+1}-u^{n+2/3}}{\Delta t}(v-u^{n+1})\mathrm{d}x \geq 0, \forall v \in L^2_+(\Omega). \end{cases} \tag{2.5}$$

We will be discussing in Sections 2.2, 2.3 and 2.4 the solution of the sub-problems (2.3), (2.4) and (2.5), respectively, the finite element implementation of (2.2)–(2.5) being discussed in Section 3.

2.2. On the solution of problem (2.3)

One can rewrite problem (2.3) (an elliptic variational inequality problem) as

$$u^{n+1/3} = \arg \min_{v \in H_0^1(\Omega)} \left[\frac{1}{2} \int_\Omega |v|^2 \mathrm{d}x + \frac{\Delta t}{2} \mu \int_\Omega |\nabla v|^2 \mathrm{d}x + \Delta t \tau_y \int_\Omega |\nabla v| \mathrm{d}x - \int_\Omega u^n v \mathrm{d}x \right]. \tag{2.6}$$

Problem (2.6) is a particular case of the following family (parameterized by $u_* \in L^2(\Omega)$) of variational problems:

$$u = \arg \min_{v \in H_0^1(\Omega)} \left[\frac{1}{2} \int_\Omega |v|^2 \mathrm{d}x + \frac{\Delta t}{2} \mu \int_\Omega |\nabla v|^2 \mathrm{d}x + \Delta t \tau_y \int_\Omega |\nabla v| \mathrm{d}x - \int_\Omega u_* v \mathrm{d}x \right]. \tag{2.7}$$

Problem (2.7) has a unique solution (verifying $u_* \geq 0 \Rightarrow u \geq 0$). It follows from, *e.g.*, [8] and Chapter 6 of [10] that an equivalent formulation of problem (2.7) reads as:

$$\begin{cases} \{u, \mathbf{p}\} \in H_0^1(\Omega) \times \mathbf{P}, \\ \int_\Omega uv \mathrm{d}x + \Delta t \mu \int_\Omega \nabla u \cdot \nabla v \mathrm{d}x + \Delta t \tau_y \int_\Omega \mathbf{p} \cdot \nabla v \mathrm{d}x = \int_\Omega u_* v \mathrm{d}x, \\ \forall v \in H_0^1(\Omega), \\ \mathbf{p} \cdot \nabla u = |\nabla u| \Leftrightarrow \mathbf{p} = P_{\mathbf{P}}(\mathbf{p} + \tau_y \rho \nabla u), \forall \rho > 0, \end{cases} \tag{2.8}$$

where in (2.8) (i) $\mathbf{P} = \{\mathbf{q} | \mathbf{q} \in (L^2(\Omega))^2, |\mathbf{q}(x)| \leq 1 \text{ a.e. on } \Omega\}$, and (ii) operator $P_{\mathbf{P}}$ is the orthogonal projector from $(L^2(\Omega))^2$ onto $\mathbf{P}$, that is

$$P_{\mathbf{P}}(\mathbf{q})(x) = \frac{\mathbf{q}(x)}{\sup[1, |\mathbf{q}(x)|]}, \forall \mathbf{q} \in \left(L^2(\Omega)\right)^2, \text{ a.e. on } \Omega.$$

Many algorithms are available for the solution of elliptic variational inequalities such as problem (2.7), (2.8), ranging from those (including *Alternating Direction Methods of Multipliers* (*ADMM*)-based algorithms) discussed in, *e.g.* [4, 10, 14] to semi-smooth Newton's algorithms developed more recently and discussed in, *e.g.* [5, 15, 16, 19, 20]. Privileging simplicity over performance, and given the *fixed-point* flavor of (2.8), we selected the well-documented (see Chap. 6 of [10] and references therein) algorithm below:

$$\mathbf{p}^0 = \mathbf{p}_* \text{ is given in } \mathbf{P}. \tag{2.9}$$

For $k \geq 0, \mathbf{p}^k \to u^k \to \mathbf{p}^{k+1}$ as follows:
Solve

$$\begin{cases} u^k \in H_0^1(\Omega), \\ \int_\Omega u^k v \mathrm{d}x + \Delta t \mu \int_\Omega \nabla u^k \cdot \nabla v \mathrm{d}x = \int_\Omega u_* v \mathrm{d}x - \Delta t \tau_y \int_\Omega \mathbf{p}^k \cdot \nabla v \mathrm{d}x, \forall v \in H_0^1(\Omega), \end{cases} \tag{2.10}$$

and update $\mathbf{p}^k$ via

$$\mathbf{p}^{k+1} = P_{\mathbf{P}}\left(\mathbf{p}^k + \tau_y \rho \nabla u^k\right), \text{ or more generally, } \mathbf{p}^{k+1} = \frac{\gamma}{\gamma + \Delta t} \mathbf{p}_* + \frac{\Delta t}{\gamma + \Delta t} P_{\mathbf{P}}\left(\mathbf{p}^k + \tau_y \rho \nabla u^k\right), \gamma > 0. \tag{2.11}$$

Concerning the convergence of algorithm (2.9)–(2.11), it is well-known (see, *e.g.*, [4, 10] and the references therein) that if

$$0 < \rho < 2 \frac{\mu}{\tau_y^2}, \tag{2.12}$$

one then has, $\forall \mathbf{p}_*$,

$$\lim_{k\to+\infty} u^k = u \text{ in } H_0^1(\Omega), \tag{2.13}$$

where u is the solution of problem (2.7), (2.8), and it is further shown in [14] that we additionally obtain strong convergence of the sequence $\{\mathbf{p}^k\}_{k\geq 0}$ using the generalized form of the update $\mathbf{p}^{k+1}$ in (2.11).

Remark 2.1. In practice, we will employ a discrete variant of algorithm (2.9)–(2.11). If one uses (as we will do in Section 3) a finite element implementation of the above algorithm, the conditions on ρ implying convergence become

$$0 < \rho < \frac{2}{\tau_y^2}\left(\mu + \frac{1}{\Delta t \Lambda_{Mh}}\right), \tag{2.14}$$

where Λ_{Mh} denotes the largest eigenvalue of the discrete analogue of $-\nabla^2$ operating on $H_0^1(\Omega)$; we recall that $\Lambda_{Mh} = O(h^{-2})$, making (2.14) less restrictive than (2.12), considering in particular that in practice a small value of Δt may be required in order to keep small the splitting error associated with the Marchuk-Yanenko scheme.

2.3. On the solution of problem (2.4)

From the context of problem (1.1), we interpret (2.4) as the formal Euler-Lagrange equation of the following problem from Calculus of Variations

$$u^{n+2/3} = \arg\min_{v\in\tilde{S}_c}\left[\frac{1}{2}\int_\Omega |v|^2\,\mathrm{d}x - \int_\Omega u^{n+1/3} v \mathrm{d}x\right], \tag{2.15}$$

with $\tilde{S}_C = \{v | v \in L^2(\Omega), \int_\Omega (e^v - 1)\mathrm{d}x = C\}$. Relation (2.15) shows that $u^{n+2/3}$ is obtained from $u^{n+1/3}$ by $L^2(\Omega)$-orthogonal projection over (the non-convex set) $\tilde{S}_C$. The set $\tilde{S}_C$ not being weakly closed for the $L^2(\Omega)$-norm, problem (2.15) has no solution in general, unlike its discrete analogues (obtained, for example, by finite element approximation). Below, we will employ the (relatively simple) formalism of the continuous problem (2.15) to discuss its iterative solution, albeit knowing that it makes sense only for its discrete analogues. Problem (2.15) is a particular case of

$$u = \arg\min_{v\in\tilde{S}_c}\left[\frac{1}{2}\int_\Omega |v|^2\,\mathrm{d}x - \int_\Omega f v \mathrm{d}x\right], \tag{2.16}$$

where in (2.16), $f \in H_0^1(\Omega) \cap L_+^2(\Omega)$. Optimality conditions for problem (2.16) read as

$$\begin{cases} u - \lambda\Delta t e^u = f, \\ \int_\Omega (e^u - 1)\,\mathrm{d}x = C, \end{cases} \tag{2.17}$$

$\lambda\Delta t$ here playing the role of a Lagrange multiplier. Actually, to simplify this presentation, we will denote $\lambda\Delta t$ by Λ. In order to solve system (2.17) we will proceed as follows: For α given, we solve the equation

$$u_\alpha - \alpha e^{u_\alpha} = f \tag{2.18}$$

then define the real valued function H by

$$H(\alpha) = \int_\Omega (e^{u_\alpha} - 1)\,\mathrm{d}x - C. \tag{2.19}$$

We clearly have

$$H(\Lambda) = 0. \tag{2.20}$$

We intend to solve (2.20) by Newton's method. This requires computing $H'\left(=\dfrac{dH}{d\alpha}\right)$. It follows from (2.18), (2.19) (and from the *implicit function theorem*) that

$$H'(\alpha) = \int_\Omega \frac{e^{2u_\alpha}}{1-\alpha e^{u_\alpha}}\mathrm{d}x. \tag{2.21}$$

Applying Newton's method to the solution of problem (2.20) leads to the following algorithm:

$$\Lambda^0 = \Lambda_0. \tag{2.22}$$

For $k \geq 0$, $\Lambda^k \to \Lambda^{k+1}$ as follows:

$$\Lambda^{k+1} = \Lambda^k - \frac{H\left(\Lambda^k\right)}{H'\left(\Lambda^k\right)}. \tag{2.23}$$

The initialization and termination of algorithm (2.22), (2.23) will be discussed in Section 2.5. Computing $H(\Lambda^k)$ and $H'(\Lambda^k)$ requires solving the equation

$$u^k - \Lambda^k e^{u^k} = f. \tag{2.24}$$

This can be done pointwise using Newton's method. In Section 2.5 we will discuss some practicalities associated with the Newton solution of equation (2.24).

2.4. On the solution of problem (2.5)

Obviously, the solution of problem (2.5) is

$$u^{n+1} = \sup\left(0, u^{n+2/3}\right). \tag{2.25}$$

Remark 2.2. We anticipate that if we initialize properly scheme (2.2)–(2.5), we will have $u^{n+2/3} \geq 0$ making (2.5) and (2.25) superfluous. Actually, the numerical experiments reported in Section 4 confirm this prediction.

2.5. On initializations and terminations

2.5.1. General strategy

We will consider first the case $\tau_y = 0$ (the pure Bratu problem). We intend to solve it for C varying from 0 to a value C_{MAX} large enough so that the corresponding λ will be well below its maximal value. Then we will increase τ_y by a well-chosen increment and solve for C between 0 and a value large enough so that again the corresponding λ will be well below its maximal value. We will denote by $\{u^{n+1/3}, \mathbf{p}^{n+1}\}$ the solution of problem (2.8) associated with $u_* = u^n$.

2.5.2. Initialization and termination of scheme (2.2)–(2.5)

(a) Suppose that $\tau_y = 0$ and that the first value of C we consider is C_1 (> 0 and small; indeed, we will take $\Delta C = C_1$ to increment C). Suppose that $u_0 = 0$ in scheme (2.2)–(2.5); we have, then $u^{1/3} = 0$, and $u^{2/3} = u^1 = \ln\left(1 + \dfrac{C_1}{|\Omega|}\right)$, the corresponding value of the Lagrange multiplier $\Lambda = \lambda\Delta t$ in (2.17) being $\dfrac{|\Omega|}{|\Omega| + C_1} \ln\left(1 + \dfrac{C_1}{|\Omega|}\right)$. Of course, we could have chosen directly $u_0 = \ln\left(1 + \dfrac{C_1}{|\Omega|}\right)$. Still assuming that $\tau_y = 0$, suppose now that C describes a finite sequence $\{C_q\}_{q=0}^{Q_{MAX}}$ with $C_0 = 0$ and $C_{q+1} = C_q + \Delta C, \forall q = 0, 1, ..., Q_{MAX} - 1$. When applying scheme (2.2)–(2.5) to the solution of problem (1.28) for $C = C_{q+1}$, with $q \geq 1$, we suggest to take for u_0 the solution u of problem (1.28) associated with $C = C_q$.

(b) Suppose now that $\tau_y > 0$. It makes sense to have τ_y describing a finite sequence $\{\tau_y^s\}_{s=0}^{S_{MAX}}$ defined (with $\Delta\tau > 0$) by $\tau_y^0 = 0$ and $\tau_y^{s+1} = \tau_y^s + \Delta\tau, \forall s = 0, 1, ..., S_{MAX} - 1$. For simplicity, we assume that C describes the same sequence as in the case $\tau_y = 0$. Suppose one wants to compute the solutions of problem (1.28) associated with $\tau_y = \tau_y^s, s \in \{1, ..., S_{MAX}\}$, and $C \in \{C_q\}_{q=1}^{Q_{MAX}}$, using scheme (2.2)–(2.5). Among the possible initializations of the above scheme, let us mention the following three:

(i) When using algorithm (2.2)–(2.5) to compute the solution of (1.28) associated with τ_y^s and $C_1 (= \Delta C)$, take $u_0 = 0$ in (2.2), and then proceed as in the case $\tau_y = 0$ for computing the solution associated with C_q for $q \geq 2$.

(ii) Use the variant of (i) where, when computing the solution of (1.28) associated with τ_y^s and $C_1 (= \Delta C)$, one takes for u_0 in (2.2), the solution u of (1.28) associated with $\tau_y^{s-1}$and C_1.

(iii) Figure 2 suggests that when C increases, the influence of τ_y on the solution of (1.28) is of decreasing importance. One can take advantage of this property when computing the solutions of (1.28) for $\tau_y = \tau_y^s$ (with $s \geq 1$) and $C \in \{C_q\}_{q=1}^{Q_{MAX}}$. Indeed, instead of computing first the solution of (1.28) associated with τ_y^s and C_1, one computes first the solution of (1.28) associated with τ_y^s and $C_{Q_{MAX}}$, taking for u_0 in (2.2), the solution u of (1.28) associated with τ_y^{s-1} and $C_{Q_{MAX}}$. To compute the solution of (1.28) associated with τ_y^s and C_q, $q < Q_{MAX}$, we advocate taking for u_0 in (2.2), the solution u of (1.28) associated with τ_y^s and C_{q+1}, that is, proceed by decreasing values of q. The coarse grid numerical results reported in Section 4 have been obtained using this strategy.

Remark 2.3. Another approach, necessarily more memory demanding, is the following: Suppose that for $\tau_y = \tau_y^s$ we solved (1.28) for all $C \in \{C_q\}_{q=1}^{Q_{MAX}}$ and saved the associated pairs $\{u, \lambda\}$. We can take advantage of these savings when using scheme (2.2)–(2.5) to compute the solution of (1.28) associated with C_q and τ_y^{s+1}: indeed, we advocate taking for u_0 in (2.2) the solution u of (1.28) associated with C_q and τ_y^s.

Concerning the termination of scheme (2.2)–(2.5) we advocate stopping the computations when

$$||u^{n+1/3} - u^{(n-1)+1/3}||_{L^2(\Omega)} \leq \theta|\Delta t|^2 \max\left[1, ||u^{n+1/3}||_{L^2(\Omega)}\right], \tag{2.26}$$

with $0 < \theta \leq 1$, $\theta = 1/10$ being a reasonable choice. The rationale with (2.26) is that in general the Marchuk-Yanenko scheme produces sequences $\{u^{n+j/J}\}_{n\geq 0}$ converging to different limits (as shown in, *e.g.*, Chap. 6 of [9]), J being the number of operators resulting from the splitting ($J = 3$ here).

Once one has computed (approximately, in practice) the solution u of problem (1.28) associated with τ_y^s and C_q, the safest way to compute the associated multiplier λ is to use relation

$$\lambda = \frac{\mu \int_\Omega |\nabla u_c|^2 \, \mathrm{d}x + \tau_y \int_\Omega |\nabla u_c| \, \mathrm{d}x}{\int_\Omega e^{u_c} u_c \mathrm{d}x}, \tag{2.27}$$

u_c being, in (2.27), the computed solution of (1.28). Relation (2.27) is a direct consequence of (1.28).

2.5.3. On the initialization and termination of algorithm (2.9)–(2.11)

In this section, we adapt initialization and termination procedures for algorithms discussed at length in [4, 10, 14] to algorithm (2.9)–(2.11), using the notation of Section 2.5.2. Suppose that C_q and τ_y^s are given with $q \geq 1$ and $s \geq 1$. If $n \geq 1$ in scheme (2.2)–(2.5), we advocate $\mathbf{p}^n$ as initializer in (2.9), when computing $u^{n+1/3}$ and $\mathbf{p}^{n+1}$ *via* algorithm (2.9)–(2.11). If $n = 0$, how do we pick the initializer in (2.9) to compute $u^{1/3}$ and $\mathbf{p}^1$? The answer depends on the way one varies C_q for a given value of τ_y^s. If one proceeds by increasing values of q (with $q \in \{2, \ldots, Q_{MAX}\}$), take for initializer in (2.9) the (computed) limit of the sequence $\{\mathbf{p}^n\}_{n\geq 0}$ associated with $\{u^n\}_{n\geq 0}$ *via* (2.8) with $u^* = u^n$ when using scheme (2.2)–(2.5) to solve problem (1.28), (1.29) for $C = C_{q-1}$ and $\tau_y = \tau_y^s$. If one proceeds by decreasing values of q (with $q \in \{1, \ldots, Q_{MAX} - 1\}$), do as above with C_{q-1} replaced by C_{q+1}.

We still have to discuss the initialization of algorithm (2.9)–(2.11) when using scheme (2.2)–(2.5) to compute the solution of problem (1.28), (1.29) associated with $C = C_1$ $(= \Delta C)$ or $C_{Q_{MAX}}$ and $\tau_y = \tau_y^1 (= \Delta\tau)$. We intend to use the information provided by the solution of the pure Bratu problem (problem (1.2)). It is thus natural to proceed by decreasing values of q in order to compute the solutions of (1.28), (1.29) associated with $C_q, q = 1, \ldots, Q_{MAX}$ and $\tau_y = \tau_y^1 (= \Delta\tau)$, using scheme (2.2)–(2.5) combined with algorithm (2.9)–(2.11) (the main reason being that the size of the subdomain of Ω where ∇u vanishes is a decreasing function of C). In order to compute the pair $\{u^{1/3}, \mathbf{p}^1\}$ associated with $C_{Q_{MAX}}$ and τ_y^1, using algorithm (2.9)–(2.11), we suggest taking as initializer in(2.9) the vector-valued function $\mathbf{p}_0^0$ defined by

$$\mathbf{p}_0^0(x) = \begin{cases} \dfrac{\nabla u_{Q_{MAX}}^0(x)}{\left|\nabla u_{Q_{MAX}}^0(x)\right|} \text{ if } \nabla u_{Q_{MAX}}^0(x) \neq \mathbf{0}, \\ \mathbf{0} \text{ if } \nabla u_{Q_{MAX}}^0(x) = \mathbf{0}, \end{cases} \tag{2.28}$$

where $u_{Q_{MAX}}^0$ is the solution of the pure Bratu problem associated with $C = C_{Q_{MAX}}$ (actually, a similar strategy can be used for all the terms of the finite sequence $\{\tau_y^s\}_{s=1}^{S_{MAX}}$ (assuming that one always proceeds by decreasing values of C_q)).

Since algorithm (2.9)–(2.11) is nested in scheme (2.2)–(2.5), it makes sense to use a stopping criterion less demanding than the one advocated for the latter scheme, namely the one in (2.26). This leads us to use

$$||u^{k+1} - u^k||_{L^2(\Omega)} \leq |\Delta t|^2 \theta' \max\left[1, ||u^{k+1}||_{L^2(\Omega)}\right], \tag{2.29}$$

with $\theta' >> \theta$, θ being the coefficient used in (2.26); $\theta' = 10$ is a reasonable choice (validated, as with (2.26), by numerical experiments).

2.5.4. On the initialization and termination of the Newton algorithm (2.22), (2.23)

Let us denote by Λ^{n+1} the Lagrange multiplier associated with the solution $u^{n+2/3}$ of problem (2.4). Suppose that $n \geq 1$: In order to compute $\{\Lambda^{n+1}, u^{n+2/3}\}$ *via* algorithm (2.22), (2.23), we advocate initializing the above algorithm with Λ^n (the Lagrange multiplier associated with $u^{(n-1)+2/3}$). If $n = 0$, the initialization of algorithm (2.22), (2.23) depends upon the choice of u^0 in scheme (2.2)–(2.5): (i) If one takes $u^0 = 0$ in (2.2), then $u^{1/3} = 0$, $u^{2/3} = u^1 = \ln\left(1 + \dfrac{C}{|\Omega|}\right)$, and $\Lambda^1 = \dfrac{|\Omega|}{|\Omega| + C} \ln\left(1 + \dfrac{C}{|\Omega|}\right)$. (ii) For each of the more sophisticated choices of u^0 discussed in Section 2.5.2, obvious initializations do exist when applying algorithm (2.22), (2.23) to the computation of $\{u^{2/3}, \Lambda^1\}$.

For the termination of algorithm (2.22), (2.23) a natural stopping criterion is provided by

$$\left|\frac{\Lambda^{k+1} - \Lambda^k}{\Lambda^{k+1}}\right| \leq tol \tag{2.30}$$

with $10^{-8} \leq tol \leq 10^{-6}$. Another important issue we still have to discuss is the initialization of the Newton algorithm we use for the pointwise solution of problem (2.24). Fortunately, this is a simple issue since, if $k \geq 1$, we will use u^{k-1} as initializer. For $k = 0$, we will take $u^{(n-1)+2/3}$ as initializer (assuming that we are applying algorithm (2.22), (2.23) to the computation of $\{\Lambda^{n+1}, u^{n+2/3}\}$).

2.6. Further remarks

Remark 2.4. (*On a variant of scheme* (2.2)–(2.5)). In order to overcome the systematic *splitting error* associated with the scheme of Marchuk-Yanenko, using small values of Δt is a necessity. One can take advantage of this requirement by using a *linearized version* of problem (2.4), namely

$$u^{n+2/3} = \arg \min_{v \in D\tilde{S}_C(u^{n+1/3})} \left[\frac{1}{2} \int_\Omega |v|^2 - \int_\Omega u^{n+1/3} v \mathrm{d}x \right], \tag{2.31}$$

where the affine space $D\tilde{S}_C(u^{n+1/3})$ is defined by

$$D\tilde{S}_C(u^{n+1/3}) = \left\{ v | v \in L^2(\Omega), \int_\Omega \exp(u^{n+1/3})(v - u^{n+1/3} + 1)\mathrm{d}x = C + |\Omega| \right\}. \tag{2.32}$$

The closed form solution of problem (2.31), (2.32) is given by

$$u^{n+2/3} = u^{n+1/3} + \frac{C + |\Omega| - \int_\Omega \exp(u^{n+1/3})\mathrm{d}x}{\int_\Omega \exp(2u^{n+1/3})\mathrm{d}x} \exp(u^{n+1/3}). \tag{2.33}$$

Other linearization possibilities do exist; one can, for example, replace $u^{n+1/3}$ by $u^{(n-1)+2/3}$ in (2.33).

Remark 2.5. The Marchuk-Yanenko scheme (2.2)–(2.5) is the simplest operator-splitting method we can think about concerning the solution of the minimization problem (1.9). Actually, other solution methods based on operator-splitting do exist for problem (1.9), some of them of the ADMM type. Unlike the Marchuk-Yanenko scheme (2.2)–(2.5), ADMM algorithms are splitting error free but require proper adjustment of the so-called *Lagrangian augmentation parameters*, a delicate issue in general (as shown in, *e.g.*, [13]).

3. On the finite element approximation of problem (1.1), (1.9)

3.1. The basic discrete sets

From now on, we assume that Ω is a polygonal domain of $\mathbf{R}^2$. Next, we consider a family $\{\mathcal{T}_h\}_h$ of triangulations of Ω verifying the usual (see, *e.g.*, Appendix 1 of [8, 9] and Chap. 1 of [10]) assumptions below:

(i) If T in $\mathcal{T}_h$, T is closed and $\cup_{T \in \mathcal{T}_h} T = \overline{\Omega}$.
(ii) h is the length of the largest edge(s) of $\mathcal{T}_h$.
(iii) If T_1, $T_2 \in \mathcal{T}_h$ with $T_1 \neq T_2$, we have either $T_1 \cap T_2 = \varnothing$, or T_1, T_2 have only one common vertex, or T_1, T_2 have only one full edge in common.

To these classical assumptions, we will add the following one:

$$\text{all the angles of } \mathcal{T}_h \text{ are } \leq \frac{\pi}{2}. \tag{3.1}$$

We associate with $\mathcal{T}_h$ the two following finite dimensional spaces:

$$V_h = \{ v | v \in C^0(\overline{\Omega}), v|_T \in \mathcal{P}_1, \forall T \in \mathcal{T}_h \}, \tag{3.2}$$

and

$$V_{0h} = \{v | v \in V_h, v|_{\partial\Omega} = 0\} \left(= V_h \cap H_0^1(\Omega)\right), \tag{3.3}$$

with $\mathcal{P}_1$ the space of the two variable polynomials of degree ≤ 1. Let N_h (resp., N_{0h}) be the dimension of space V_h (resp., V_{0h}); we recall that

$$\begin{cases} v = \sum_{i=1}^{N_h} v(P_i) w_i, \forall v \in V_h, \\ v = \sum_{i=1}^{N_{0h}} v(P_i) w_i, \forall v \in V_{0h}, \end{cases} \tag{3.4}$$

where $\sum_h = \{P_i\}_{i=1}^{N_h}$(resp., $\sum_{0h} = \{P_i\}_{i=1}^{N_{0h}}$) is the set of the vertices of $\mathcal{T}_h$ (resp., is the set of the vertices of $\mathcal{T}_h$ which do not belong to $\partial\Omega$), and the function w_i is uniquely defined by

$$\begin{cases} w_i \in V_h, \forall i = 1, \ldots, N_h, \\ w_i(P_i) = 1, w_i(P_j) = 0, \forall j = 1, \ldots, N_h, j \neq i, \end{cases}$$

implying that the support of function w_i is the union of those triangles of $\mathcal{T}_h$ which have P_i as a common vertex. The sets $\mathcal{B}_h = \{w_i\}_{i=1}^{N_h}$ and $\mathcal{B}_{0h} = \{w_i\}_{i=1}^{N_{0h}}$are vector bases of the spaces V_h and V_{0h}, respectively. Concerning the approximation of the set S_C, one has two natural options, the first one being

$$S_{Ch}^1 = \left\{v | v \in V_{0h}, \int_\Omega (e^v - 1)\, \mathrm{d}x = C\right\} (= S_C \cap V_{0h}). \tag{3.5}$$

Note that while $\int_T e^v w dx$ can be computed exactly, $\forall T \in \mathcal{T}_h$, $\forall v, w \in \mathcal{P}_1$, as a function of the values taken by v and w at the vertices of T (see Chap. 3 of [9] for details) we strongly recommend (our second option) approximating S_C by

$$S_{Ch}^2 = \left\{v | v \in V_{0h}, \sum_{i=1}^{N_{0h}} |\omega_i| \left(e^{v(P_i)} - 1\right) = 3C\right\}, \tag{3.6}$$

where ω_i is the polygonal union of those triangles of $\mathcal{T}_h$ which have P_i as a common vertex, and $|\omega_i|$= measure of ω_i. Clearly, one obtains $S_{Ch}^2$from $S_{Ch}^1$by application of the *trapezoidal rule* on each triangle of $\mathcal{T}_h$, taking advantage of the relation $\int_\Omega (e^v - 1)\, \mathrm{d}x = \sum_{T \in \mathcal{T}_h} \int_T (e^v - 1)\, \mathrm{d}x$.

3.2. Formulation of the approximate problems. Basic properties of the approximate solutions

Section 3.1 suggests approximating problem (1.9) by either

$$\begin{cases} u_h^1 \in S_{Ch}^1, \\ J\left(u_h^1\right) \leq J(v), \forall v \in S_{Ch}^1, \end{cases} \tag{3.7}$$

or

$$\begin{cases} u_h^2 \in S_{Ch}^2, \\ J\left(u_h^2\right) \leq J(v), \forall v \in S_{Ch}^2, \end{cases} \tag{3.8}$$

with $J(v) = \dfrac{\mu}{2} \int_\Omega |\nabla v|^2\, \mathrm{d}x + \tau_y \int_\Omega |\nabla v|\, \mathrm{d}x$.

Problem (3.8) is more 'friendly' computationally than (3.7), however it is more complicated to discuss mathematically when convergence is concerned, as shown in Section 3.3. Both problems (3.7) and (3.8) have solutions as shown by the following.

Theorem 3.1. *Problems* (3.7) *and* (3.8) *have solutions. Moreover, all of the solutions of* (3.7) *and* (3.8) *are non-negative.*

Proof. The sets S^1_{Ch} and S^2_{Ch} being non-empty closed subsets of the finite dimensional space V_{0h}, and functional J being coercive (that is $\lim_{||v||_{H^1_0(\Omega)} \to +\infty} J(v) = +\infty$), the existence of solutions to both problems follows by very simple standard arguments. To prove that $u^l_h \geq 0, \forall l = 1, 2$, suppose that the above property is not true, that is u^l_h takes negative values in Ω. With u^l_h, we associate u^{l+}_h defined uniquely as follows:

$$\begin{cases} u^{l+}_h \in V_{0h}, \\ u^{l+}_h (P_i) = \max \left[0, u^l_h (P_i)\right], \forall i = 1, \ldots, N_{0h}. \end{cases}$$

We clearly have $u^{l+}_h \geq u^l_h$ and $u^{l+}_h - u^l_h \neq 0$, these two relations implying

$$\int_\Omega \left(e^{u^{1+}_h} - 1\right) \mathrm{d}x > \int_\Omega \left(e^{u^1_h} - 1,\right) \mathrm{d}x = C \left(\text{resp., } \sum_{i=1}^{N_{0h}} |\omega_i| \left(e^{u^{2+}_h (P_i)} - 1\right) > \sum_{i=1}^{N_{0h}} |\omega_i| \left(e^{u^2_h (P_i)} - 1\right) = 3C\right), \tag{3.9}$$

if one uses (3.7) (resp., (3.8)) as the approximate problem. Moreover, since triangulation $\mathcal{T}_h$ verifies (3.1), we clearly have

$$\left|\nabla u^{l+}_h\right| \leq \left|\nabla u^l_h\right| \text{ over each triangle } T \in \mathcal{T}_h. \tag{3.10}$$

Relation (3.10) implies

$$J\left(u^{l+}_h\right) \leq J\left(u^l_h\right). \tag{3.11}$$

It follows from (3.9) that the non-negativity of u^{l+}_h implies the existence of a constant $\alpha_l \in (0, 1)$ such that

$$\int_\Omega \left(e^{\alpha_1 u^{1+}_h} - 1\right) \mathrm{d}x = C \left(\text{resp., } \sum_{i=1}^{N_{0h}} |\omega_i| \left(e^{\alpha_2 u^{2+}_h (P_i)} - 1\right) = 3C\right). \tag{3.12}$$

Denote $\alpha_l u^{l+}_h$ by ψ_l; we clearly have

$$J(\psi_l) < \alpha_l J\left(u^{l+}_h\right) < J\left(u^{l+}_h\right) \leq J\left(u^l_h\right). \tag{3.13}$$

Relations (3.12) and (3.13) show that $u^l_h$is not a minimizer of J over S^l_{Ch}, implying that such a minimizer is necessarily non-negative. □

Proceeding as in Section 1, we can easily prove the following discrete analogue of Theorem 1.2:

Theorem 3.2. *Suppose that $C > 0$. Then if u^1_h (resp., u^2_h) is a solution to problem* (3.7) *(resp.,* (3.8)*) there exists a constant $\lambda^1_h > 0$ (resp., $\lambda^2_h > 0$) such that*

$$\begin{cases} u^1_h \in V_{0h}, \\ \mu \int_\Omega \nabla u^1_h \cdot \nabla \left(v - u^1_h\right) \mathrm{d}x + \tau_y \left[\int_\Omega |\nabla v| \, \mathrm{d}x - \int_\Omega \left|\nabla u^1_h\right| \mathrm{d}x\right] \\ \quad \geq \lambda^1_h \int_\Omega e^{u^1_h} \left(v - u^1_h\right) \mathrm{d}x, \forall v \in V_{0h}, \\ \quad \int_\Omega \left(e^{u^1_h} - 1\right) \mathrm{d}x = C \end{cases} \tag{3.14}$$

(resp.,

$$\begin{cases} u_h^2 \in V_{0h}, \\ \mu \int_\Omega \nabla u_h^2 \cdot \nabla \left(v - u_h^2 \right) \mathrm{d}x + \tau_y \left[\int_\Omega |\nabla v| \, \mathrm{d}x - \int_\Omega \left| \nabla u_h^2 \right| \mathrm{d}x \right] \\ \geq \dfrac{\lambda_h^2}{3} \sum_{i=1}^{N_{0h}} |\omega_i| \, e^{u_h^2(P_i)} \left[v\left(P_i\right) - u_h^2\left(P_i\right) \right], \forall v \in V_{0h}, \\ \sum_{i=1}^{N_{0h}} |\omega_i| \left(e^{u_h^2(P_i)} - 1 \right) = 3C.) \end{cases} \tag{3.15}$$

Relations (3.14) *and* (3.15) *imply*

$$\begin{cases} \lambda_h^1 = \dfrac{\mu \int_\Omega \left| \nabla u_h^1 \right|^2 \mathrm{d}x + \tau_y \int_\Omega \left| \nabla u_h^1 \right| \mathrm{d}x}{\int_\Omega e^{u_h^1} u_h^1 \mathrm{d}x}, \\ \lambda_h^2 = 3 \dfrac{\mu \int_\Omega \left| \nabla u_h^2 \right|^2 \mathrm{d}x + \tau_y \int_\Omega \left| \nabla u_h^2 \right| \mathrm{d}x}{\sum_{i=1}^{N_{0h}} |\omega_i| \, e^{u_h^2(P_i)} u_h^2\left(P_i\right)}. \end{cases} \tag{3.16}$$

3.3. Convergence of the approximate solutions

The convergence results presented below rely partly on the following

Lemma 3.3. *Let us denote by* S_C^+ *and* $\mathcal{D}(\Omega)$ *the sets defined by*

$$S_C^+ = \{ v | v \in S_C, v \geq 0 \}$$

and

$$\mathcal{D}\left(\Omega\right) = \left\{ v | v \in C^\infty\left(\overline{\Omega}\right), v \, \text{has a compact support in} \, \Omega \right\},$$

respectively. Then,$\forall C > 0$,

$$\overline{S_C^+ \cap \mathcal{D}\left(\Omega\right)}^{H_0^1(\Omega)} = S_C^+. \tag{3.17}$$

Proof. Let us consider $v \in S_C^+$. Since $C > 0$, we have $v \neq 0$. It follows from Chapter 1 of [8] that there exists a sequence $\{v_n\}_{n \geq 0}$ such that

$$v_n \in \mathcal{D}\left(\Omega\right), v_n \geq 0, \forall n \geq 0, \lim_{n \to +\infty} v_n = v, \text{in} \, H_0^1\left(\Omega\right). \tag{3.18}$$

Combined with (3.18), the property $v \neq 0$ implies that $v_n \neq 0$ for n sufficiently large. Let us consider $v_n \neq 0$. The function $\alpha \to \int_\Omega \left[e^{\alpha v_n} - 1 \right] \mathrm{d}x$ being a continuous and strictly increasing map from $[0, +\infty)$ onto $[0, +\infty)$, there exists a unique positive real number α_n such that

$$\int_\Omega \left[e^{\alpha_n v_n} - 1 \right] \mathrm{d}x = C. \tag{3.19}$$

It follows from (3.19) and the Jensen inequality for integrals that

$$\exp\left[\frac{\alpha_n}{|\Omega|} \int_\Omega v_n \mathrm{d}x \right] \leq \frac{1}{|\Omega|} \int_\Omega e^{\alpha_n v_n} \mathrm{d}x = 1 + \frac{C}{|\Omega|}. \tag{3.20}$$

Relation (3.20) implies in turn that for n sufficiently large

$$0 < \alpha_n \leq \frac{|\Omega|}{\int_\Omega v_n \mathrm{d}x} \ln\left(1 + \frac{C}{|\Omega|}\right). \tag{3.21}$$

It follows from (3.21) and from $\lim_{n\to+\infty} \int_\Omega v_n \mathrm{d}x = \int_\Omega v \mathrm{d}x > 0$, that there exists $A > 0$ such that

$$0 < \alpha_n \leq A, \forall n \geq 0. \tag{3.22}$$

Interval $[0,\ A]$ being compact, the sequence $\{\alpha_n\}_{n\geq 0}$ contains converging sub-sequences; one can easily show that all of these converging subsequences converge to 1, implying that the whole sequence $\{\alpha_n\}_{n\geq 0}$ converges to 1 also, and therefore that $\lim_{n\to+\infty} \alpha_n v_n = v \,\text{in}\, H_0^1(\Omega)$. Let us denote $\alpha_n v_n$ by w_n. Then the sequence $\{w_n\}_{n\geq 0}$ verifies all of the properties listed in (3.18); in addition, it verifies $w_n \in S_C^+, \forall n \geq 0$, which proves the lemma. □

Concerning the convergence of the families$\{u_h^1\}_{h>0}$ and $\{u_h^2\}_{h>0}$, the first result we are going to prove is the simplest one, namely, the convergence of $\{u_h^1\}_{h>0}$. This follows from

Theorem 3.4. *Assume that $C > 0$. Assume also that*

$$\text{the angles of } \mathcal{T}_h \text{ are bounded from below by } \theta_0 > 0, \theta_0 \text{ being independent of } h. \tag{3.23}$$

Then from the family $\{\{u_h^1, \lambda_h^1\}\}_{h>0}$ of the solutions of problem (3.7), (3.14), we can extract a sub-sequence (still denoted by$\{\{u_h^1, \lambda_h^1\}\}_h$) such that

$$\lim_{h\to 0} \{u_h^1, \lambda_h^1\} = \{u, \lambda\} \text{ in } H_0^1(\Omega) \times \mathbf{R}, \tag{3.24}$$

where $\{u, \lambda\}$ is a solution to (1.9), (1.28). All of the converging sub-sequences extracted from the family $\{\{u_h^1, \lambda_h^1\}\}_{h>0}$ have a similar property.

Proof. Albeit simple conceptually, the proof is a bit lengthy, justifying breaking it into several well-identified logical steps.

Step 1: *Weak convergence*

Define $r_h \colon S_C^+ \cap \mathcal{D}(\Omega) \to S_{Ch}^1$ as follows: With $v \in S_C^+ \cap \mathcal{D}(\Omega)$ we associate $\pi_h v$ uniquely defined by

$$\begin{cases} \pi_h v \in V_{0h}, \\ \pi_h v\,(P) = v\,(P)\,, \forall P \in \Sigma_h, \end{cases}$$

that is, $\pi_h v$ is the piecewise linear interpolant of v over $\overline{\Omega}$ associated with $\mathcal{T}_h$. We clearly have $\pi_h v \geq 0$, and $\pi_h v \neq 0$ for h small enough. It follows from (3.23) and, *e.g.*, Appendix 1 of [8] that

$$\|\nabla\,(\pi_h v - v)\|_{(L^\infty(\Omega))^2} \leq \frac{3h}{\sin\theta_0} \left\|\mathbf{D}^2 v\right\|_{\left(C^0(\overline{\Omega})\right)^{2\times 2}}, \tag{3.25}$$

where $\mathbf{D}^2 v$ is the Hessian matrix of function v. Relation (3.25) implies

$$\lim_{h\to 0} \pi_h v = v \,\text{in}\, H_0^1(\Omega) \cap L^\infty(\Omega). \tag{3.26}$$

From the properties $\pi_h v \geq 0$, and $\pi_h v \neq 0$ for h small enough, there exists a unique real number $\alpha_h > 0$ such that

$$\int_\Omega \left(e^{\alpha_h \pi_h v} - 1\right) \mathrm{d}x = C. \tag{3.27}$$

Proceeding as in the proof of Lemma 3.3, one can easily show that $\lim_{h\to 0}\alpha_h = 1$, which implies in turn that

$$\lim_{h\to 0}\alpha_h\pi_h v = v \text{ in } H_0^1(\Omega). \tag{3.28}$$

Denote $\alpha_h\pi_h v$ by $r_h v$. Operator r_h maps $S_C^+\cap\mathcal{D}(\Omega)$ into S_{Ch}^1, and, from (3.28),

$$\lim_{h\to 0} r_h v = v \text{ in } H_0^1(\Omega). \tag{3.29}$$

It follows from (3.7), and from the properties of operator r_h that

$$J\left(u_h^1\right) \le J\left(r_h v\right), \forall v \in S_C^+\cap\mathcal{D}\left(\Omega\right). \tag{3.30}$$

Let u^* be an arbitrary element of $S_C^+\cap\mathcal{D}(\Omega)$; since $C>0$, we have $u^*\neq 0$. From (3.29), we have

$$\lim_{h\to 0} r_h u^* = u^* \text{ in } H_0^1\left(\Omega\right),$$

a relation implying

$$\lim_{h\to 0} J(r_h u^*) = J(u^*). \tag{3.31}$$

It follows from (3.30) and (3.31) that

$$J\left(u_h^1\right) \le J\left(r_h u^*\right) \le 2J\left(u^*\right) \tag{3.32}$$

for h small enough. Relations (3.32) imply that the family $\{u_h^1\}_{h>0}$ is bounded in $H_0^1(\Omega)$, a property implying in turn the existence of a sub-sequence – still denoted by$\{u_h^1\}_h$– and of $u\in H_0^1(\Omega)$ such that

$$\lim_{h\to 0} u_h^1 = u, \text{ weakly in } H_0^1(\Omega). \tag{3.33}$$

It follows from (3.33) that $u\ge 0$ and $\int_\Omega\left(e^u-1\right)\mathrm{d}x = C$, the second relation resulting from the fact that the mapping $\phi\to e^\phi$ is compact (completely continuous) from $H_0^1(\Omega)$ into $L^s(\Omega)$,$\forall s\in[1,+\infty)$. Now, the functional J being convex continuous over $H_0^1(\Omega)$ is weakly lower semi-continuous on the above space. Combining this weak lower semi-continuity property with (3.29) and (3.30), we obtain

$$J\left(u\right) \le \liminf_{h\to 0} J\left(u_h^1\right) \le \limsup_{h\to 0} J\left(u_h^1\right) \le J\left(v\right), \forall v\in S_C^+\cap\mathcal{D}\left(\Omega\right). \tag{3.34}$$

Since, from Lemma 3.3, $\overline{S_C^+\cap\mathcal{D}(\Omega)}^{H_0^1(\Omega)} = S_C^+$, relation (3.34) implies

$$J\left(u\right) \le \liminf_{h\to 0} J\left(u_h^1\right) \le \limsup_{h\to 0} J\left(u_h^1\right) \le J\left(v\right), \forall v\in S_C^+. \tag{3.35}$$

From Theorem 1.1, any solution of problem (1.9) belongs to S_C^+. It follows then from (3.35) that u, a minimizer of J over S_C^+, is also a solution of problem (1.9).

Step 2: *Strong convergence*

We keep Step 1 notation and still consider the above sub-sequence $\{u_h^1\}_h$ converging weakly to u in $H_0^1(\Omega)$. Taking $v=u$ in (3.35) shows that

$$\lim_{h\to 0} J(u_h^1) = J(u). \tag{3.36}$$

Observe now that $J(v) = J_0(v) + J_1(v)$, with

$$\begin{cases} J_0(v) = \dfrac{\mu}{2}\int_\Omega |\nabla v|^2 \,\mathrm{d}x, \\ J_1(v) = \tau_y \int_\Omega |\nabla v| \,\mathrm{d}x, \end{cases} \quad \forall v \in H_0^1(\Omega).$$

Both functionals J_0 and J_1 being convex continuous are weakly lower semi-continuous. We have then from (3.33) and (3.36)

$$J(u) = \liminf_{h\to 0}\left[J_0(u_h^1) + J_1(u_h^1)\right] \geq \liminf_{h\to 0} J_0(u_h^1) + \liminf_{h\to 0} J_1(u_h^1) \geq J_0(u) + J_1(u) = J(u). \quad (3.37)$$

Relations (3.37) imply $\lim_{h\to 0}\left(J_0(u_h^1), J_1(u_h^1)\right) = (J_0(u), J_1(u))$, which implies in turn (from the definition of J_0) that

$$\lim_{h\to 0}\int_\Omega \left|\nabla u_h^1\right|^2 \mathrm{d}x = \int_\Omega |\nabla u|^2 \,\mathrm{d}x. \quad (3.38)$$

Combining the weak convergence property (113) with the convergence of the norm relation (3.38) shows that $\lim_{h\to 0} u_h^1 = u$ in the Hilbert space $H_0^1(\Omega)$, that is

$$\lim_{h\to 0}\int_\Omega \left|\nabla\left(u_h^1 - u\right)\right|^2 \mathrm{d}x = 0, \quad (3.39)$$

Step 3: *Convergence of* $\{\lambda_h^1\}_h$
It follows from Section 1 that

$$\lambda = \frac{\mu\int_\Omega |\nabla u|^2 \,\mathrm{d}x + \tau_y \int_\Omega |\nabla u| \,\mathrm{d}x}{\int_\Omega e^u u \mathrm{d}x}. \quad (3.40)$$

Combining (3.39), (3.40) with the first equation in (3.16), we obtain

$$\lim_{h\to 0}\lambda_h^1 = \lim_{h\to 0}\frac{\mu\int_\Omega \left|\nabla u_h^1\right|^2 \mathrm{d}x + \tau_y\int_\Omega \left|\nabla u_h^1\right| \mathrm{d}x}{\int_\Omega e^{u_h^1} u_h^1 \mathrm{d}x} = \frac{\mu\int_\Omega |\nabla u|^2 \,\mathrm{d}x + \tau_y\int_\Omega |\nabla u| \,\mathrm{d}x}{\int_\Omega e^u u \mathrm{d}x} = \lambda,$$

which completes the proof of the theorem. □

If the pair $\{u,\lambda\}$ is the unique solution of problem (1.9), (1.28), the whole family$\{\{u_h^1, \lambda_h^1\}\}_{h>0}$ converges to $\{u,\lambda\}$ in $H_0^1(\Omega) \times \mathbf{R}$.

Previously, we advocated using (3.8) to approximate problem (1.9), (1.28), problem (3.8) being easier to handle computationally than (3.7). On the other hand, proving the convergence of the family $\{\{u_h^2, \lambda_h^2\}\}_{h>0}$ is technically more complicated, compared to proving the convergence of $\{\{u_h^1, \lambda_h^1\}\}_{h>0}$. A fact that will appear clearly from the proof of the following

Theorem 3.5. *Assume that $C > 0$. Assume also that there exists σ (≥ 1) such that*

$$\frac{h}{h_{\min}} \leq \sigma, \forall T \in \mathcal{T}_h, \quad (3.41)$$

where $h_{\min}$ is the length of the smallest edge(s) of $\mathcal{T}_h$. Then, from the family $\{\{u_h^2, \lambda_h^2\}\}_{h>0}$ of solutions of problem (3.8), (3.15), we can extract a sub-sequence (still denoted by$\{\{u_h^2, \lambda_h^2\}\}_h$) such that

$$\lim_{h\to 0}\left\{u_h^2, \lambda_h^2\right\} = \{u, \lambda\} \text{ in } H_0^1(\Omega) \times \mathbf{R}, \quad (3.42)$$

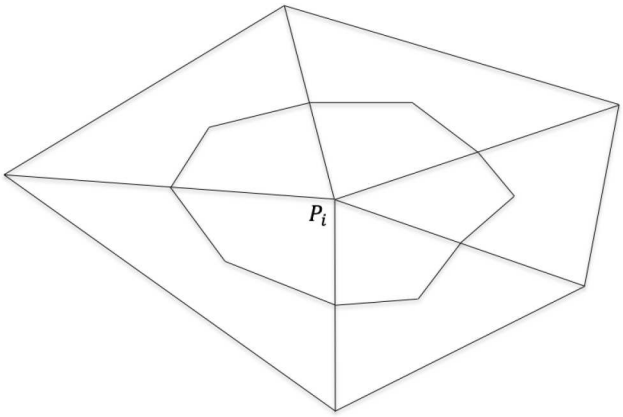

FIGURE 3. Visualization of the polygonal Q_i associated with the vertex P_i.

where $\{u, \lambda\}$ *is a solution to* (1.9), (1.28). *All of the converging sub-sequences extracted from the family* $\{\{u_h^2, \lambda_h^2\}\}_{h>0}$ *have a similar property.*

Proof. As for Theorem 3.4, we have divided the proof into several logical steps. Before we move into these steps we will make two preliminary observations, namely: (i) Relation (3.41) implies the angle condition (3.23) with

$$\theta_0 = \sin^{-1}\left(\frac{1}{\sigma}\frac{\sqrt{3}}{2}\right). \tag{3.43}$$

(ii) An alternative definition of the set S_{Ch}^2 is given by

$$S_{Ch}^2 = \left\{v | v \in V_{0h}, \int_\Omega \left(e^{s_h v} - 1\right) \mathrm{d}x = C\right\}, \tag{3.44}$$

where the operator $s_h \colon V_{0h} \to L^\infty(\Omega)$ is defined by

$$s_h v = \sum_{i=1}^{N_{0h}} v\left(P_i\right) \chi_i; \tag{3.45}$$

and in (3.45), χ_i is the characteristic function of the polygonal union of those quadrilaterals Q_i, having P_i as a vertex, obtained by joining the center of mass of any triangle T of $\mathcal{T}_h$, having P_i as a vertex, to the mid-points of the two edges of T which have P_i as a common extremity (see Fig. 3 for a visualization). We clearly have $|Q_i| = \dfrac{|\omega_i|}{3}$.

A fundamental relation, very easy to prove *via* a simple Taylor expansion, reads as

$$\int_\Omega |s_h v - v|^p \, \mathrm{d}x \leq \left(\frac{2}{3}\right)^p h^p \int_\Omega |\nabla v|^p \, \mathrm{d}x, \forall v \in V_{0h}, \forall p \in [1, +\infty). \tag{3.46}$$

Step 1: *Weak convergence*

Taking the proof of Theorem 3.4 as a guideline, with $v \in S_C^+ \cap \mathcal{D}(\Omega)$ we associate $\pi_h v$ uniquely defined by

$$\begin{cases} \pi_h v \in V_{0h}, \\ \pi_h v\left(P\right) = v\left(P\right), \forall P \in \Sigma_h. \end{cases}$$

We clearly have $\pi_h v \geq 0$, and $\pi_h v \neq 0$ for h small enough. Relation (3.25) still holds with θ_0 defined by (3.43), implying

$$\lim_{h\to 0} \pi_h v = v \text{ in } H_0^1(\Omega) \cap L^\infty(\Omega). \tag{3.47}$$

With v as above, we consider next the function $s_h \pi_h v$. Assume that $T \in \mathcal{T}_h$ and verifies $|Q_i \cap T| > 0$. We then have

$$\begin{cases} |s_h \pi_h v(x) - \pi_h v(x)| = |v(P_i) - \pi_h v(x)| = \\ \left|\nabla(\pi_h v|_T) \cdot \overrightarrow{P_i x}\right| \leq \dfrac{2}{3} h \|\nabla \pi_h v\|_{(L^\infty(\Omega))^2}, \\ \forall x \in T \cap Q_i. \end{cases} \tag{3.48}$$

It follows from relation (3.48) that

$$\|s_h \pi_h v(x) - \pi_h v(x)\|_{L^\infty(\Omega)} \leq \frac{2}{3} h \|\nabla \pi_h v\|_{(L^\infty(\Omega))^2}. \tag{3.49}$$

Combining (3.49) with (3.25) we obtain

$$\lim_{h\to 0} \|s_h \pi_h v(x) - \pi_h v(x)\|_{L^\infty(\Omega)} = 0. \tag{3.50}$$

Relations (3.26) and (3.50) imply in turn that

$$\begin{array}{l} \lim_{h\to 0} \int_\Omega [\exp(s_h \pi_h v) - 1]\, dx = \lim_{h\to 0} \int_\Omega [\exp(\pi_h v) - 1]\, dx \\ = \int_\Omega (e^v - 1)\, dx = C, \forall v \in S_C^+ \cap \mathcal{D}(\Omega). \end{array} \tag{3.51}$$

Relations (3.51) implies the existence of a positive number family $\{\alpha_h\}_h$ such that

$$\begin{cases} \lim_{h\to 0} \alpha_h = 1, \\ \lim_{h\to 0} \alpha_h \pi_h v = v \text{ in } H_0^1(\Omega) \cap L^\infty(\Omega), \\ \int_\Omega [\exp(s_h \alpha_h \pi_h v) - 1]\, dx = C. \end{cases} \tag{3.52}$$

Take $v \in S_C^+ \cap \mathcal{D}(\Omega)$; it follows then from (3.8), (3.44) and (3.52) that

$$\begin{cases} u_h^2 \in S_{Ch}^2, \\ J\left(u_h^2\right) \leq J(\alpha_h \pi_h v), \forall v \in S_C^+ \cap \mathcal{D}(\Omega). \end{cases} \tag{3.53}$$

It follows from (3.52) that $\lim_{h\to 0} J(\alpha_h \pi_h v) = J(v)$, implying that the family $\{J(\alpha_h \pi_h v)\}_h$ is bounded, implying in turn (from the definition of functional J) that the family $\left\{u_h^2\right\}_h$ is bounded in $H_0^1(\Omega)$. There exists thus a subsequence – still denoted by $\left\{u_h^2\right\}_h$ – and $u \in H_0^1(\Omega)$, such that

$$\lim_{h\to 0} u_h^2 = u \text{ weakly in } H_0^1(\Omega). \tag{3.54}$$

Relations (3.52), (3.53), (3.54), and the property $u_h^2 \geq 0$ (see Thm. 3.1), imply, since functional J is convex over $H_0^1(\Omega)$, that

$$\begin{cases} u \in H_0^1(\Omega), u \geq 0, \\ J(u) \leq \liminf_{h\to 0} J\left(u_h^2\right) \leq \limsup_{h\to 0} J\left(u_h^2\right) \leq J(v), \forall \in v \in S_C^+ \cap \mathcal{D}(\Omega). \end{cases} \tag{3.55}$$

Since (Lem. 3.3) $\overline{S_C^+ \cap \mathcal{D}(\Omega)}^{H_0^1(\Omega)} = S_C^+$, relation (3.55) implies

$$\begin{cases} u \in H_0^1(\Omega), u \geq 0, \\ J(u) \leq J(v), \forall \in v \in S_C^+. \end{cases} \tag{3.56}$$

From (3.56), it suffices to prove

$$\int_\Omega (e^u - 1)\,\mathrm{d}x = C, \tag{3.57}$$

to show that u is solution to problem (1.9), (1.28). Since $u_h^2 \in S_{Ch}^2$, we have

$$\int_\Omega \left(e^{s_h u_h^2} - 1\right)\mathrm{d}x = C,$$

that is

$$C = \int_\Omega \left(e^{s_h u_h^2} - e^{u_h^2}\right)\mathrm{d}x + \int_\Omega \left(e^{u_h^2} - 1\right)\mathrm{d}x. \tag{3.58}$$

It follows from (3.54), and from the complete continuity (compactness) of the map

$$v \to e^v \colon H_0^1(\Omega) \to L^p(\Omega), p \in [1, +\infty),$$

that

$$\lim_{h\to 0} \int_\Omega \left(e^{u_h^2} - 1\right)\mathrm{d}x = \int_\Omega (e^u - 1)\,\mathrm{d}x. \tag{3.59}$$

If we can prove that

$$\lim_{h\to 0} \int_\Omega \left(e^{s_h u_h^2} - e^{u_h^2}\right)\mathrm{d}x = 0, \tag{3.60}$$

relations (3.58)–(3.60) will imply (3.57); (3.57) combined with (3.56) will imply in turn that u is a solution to problem (1.9), (1.28). To prove (3.60), observe that the Cauchy-Schwarz inequality in $L^2(\Omega)$ implies

$$\begin{cases} \left|\int_\Omega \left(e^{s_h u_h^2} - e^{u_h^2}\right)\mathrm{d}x\right| = \left|\int_\Omega e^{u_h^2}\left(e^{s_h u_h^2 - u_h^2} - 1\right)\mathrm{d}x\right| \\ \leq \left(\int_\Omega e^{2u_h^2}\mathrm{d}x\right)^{1/2}\left(\int_\Omega \left(e^{s_h u_h^2 - u_h^2} - 1\right)^2 \mathrm{d}x\right)^{1/2}. \end{cases} \tag{3.61}$$

Since $\lim_{h\to 0} \int_\Omega e^{2u_h^2}\mathrm{d}x = \int_\Omega e^{2u}\mathrm{d}x$, proving

$$\lim_{h\to 0} \int_\Omega \left(e^{s_h u_h^2 - u_h^2} - 1\right)^2 \mathrm{d}x = 0 \tag{3.62}$$

is sufficient to prove (3.60). We have

$$\left|e^{s_h u_h^2 - u_h^2} - 1\right| \leq \sum_{n\geq 1} \frac{1}{n!}\left|s_h u_h^2 - u_h^2\right|^n. \tag{3.63}$$

Since $\sum_{n\geq 1}\frac{1}{n!}=e-1$, it follows from (3.63), and from the Hölder inequality for series, that

$$\left|e^{s_h u_h^2-u_h^2}-1\right|^2\leq (e-1)\sum_{n\geq 1}\frac{1}{n!}\left|s_h u_h^2-u_h^2\right|^{2n}. \tag{3.64}$$

Relation (3.64) implies

$$\int_\Omega\left|e^{s_h u_h^2-u_h^2}-1\right|^2 \mathrm{d}x\leq (e-1)\sum_{n\geq 1}\frac{1}{n!}\int_\Omega\left|s_h u_h^2-u_h^2\right|^{2n}\mathrm{d}x. \tag{3.65}$$

Combining (3.65) with (3.46), we obtain

$$\int_\Omega\left|e^{s_h u_h^2-u_h^2}-1\right|^2 \mathrm{d}x\leq (e-1)\sum_{n\geq 1}\frac{1}{n!}\left(\frac{2}{3}\right)^{2n}h^{2n}\int_\Omega\left|\nabla u_h^2\right|^{2n}\mathrm{d}x. \tag{3.66}$$

We have

$$\left\{\begin{array}{l}\int_\Omega\left|\nabla u_h^2\right|^{2n}\mathrm{d}x=\sum_{T\in\mathcal{T}_h}|T|\left|\nabla u_h^2|_T\right|^{2n}=\\ \sum_{T\in\mathcal{T}_h}\frac{|T|^n}{|T|^{n-1}}\left|\nabla u_h^2|_T\right|^{2n}\leq\frac{1}{|T_{\min}|^{n-1}}\sum_{T\in\mathcal{T}_h}|T|^n\left|\nabla u_h^2|_T\right|^{2n}\\ \leq\frac{1}{|T_{\min}|^{n-1}}\left(\sum_{T\in\mathcal{T}_h}|T|\left|\nabla u_h^2|_T\right|^2\right)^n=\frac{1}{|T_{\min}|^{n-1}}\left\|\nabla u_h^2\right\|^{2n}_{(L^2(\Omega))^2},\end{array}\right. \tag{3.67}$$

with $T_{\min}=\arg_{T\in\mathcal{T}_h}\min|T|$. It follows from (3.41) that

$$|T_{\min}|\geq\frac{\sqrt{3}}{4}\frac{h^2}{\sigma^2}. \tag{3.68}$$

Combining (3.66), (3.67) and (3.68), we obtain

$$\int_\Omega\left|e^{s_h u_h^2-u_h^2}-1\right|^2 \mathrm{d}x\leq (e-1)\frac{\sqrt{3}}{4}\frac{h^2}{\sigma^2}\left[\exp\left(\frac{16}{9\sqrt{3}}\sigma^2\left\|\nabla u_h^2\right\|^2_{(L^2(\Omega))^2}\right)-1\right]. \tag{3.69}$$

Since the weak convergence property (3.54) implies the boundedness of $\left\{\left\|\nabla u_h^2\right\|_{(L^2(\Omega))^2}\right\}_h$, it follows from (3.69) that relation (3.62) is verified, a property implying that u minimizes J over S_C.

Step 2: *Strong convergence of* $\left\{u_h^2\right\}_h$ *and convergence of* $\left\{\lambda_h^2\right\}_h$

Modifying the proof of Theorem 3.4 to prove the strong convergence of $\left\{u_h^2\right\}_h$ and the convergence of $\left\{\lambda_h^2\right\}_h$ is a (relatively) simple issue. We leave it to the reader as an exercise. □

4. Numerical results

In this section, we present selected results from numerical experiments performed using our solver on the initial value problem (2.1) associated with the variational form of the Euler-Lagrange equations (1.28) or (1.29) for the Bingham-Bratu-Gelfand (BBG) constrained minimization problem (1.9)–(1.11). We present and discuss results for the viscous non-rigid ($\tau_y=0$) Bratu-Gelfand (BG) and the viscoplastic $\tau_y=2.5$ BBG cases, comparing the qualitative and quantitative behaviors of the solver in these two cases, culminating in solution branch bifurcation diagrams for these and a few other values of the plasticity yield τ_y. Note that the viscosity coefficient μ was fixed at unity for all numerical experiments. We conclude with some remarks on the finer details of the solver discovered during our investigation.

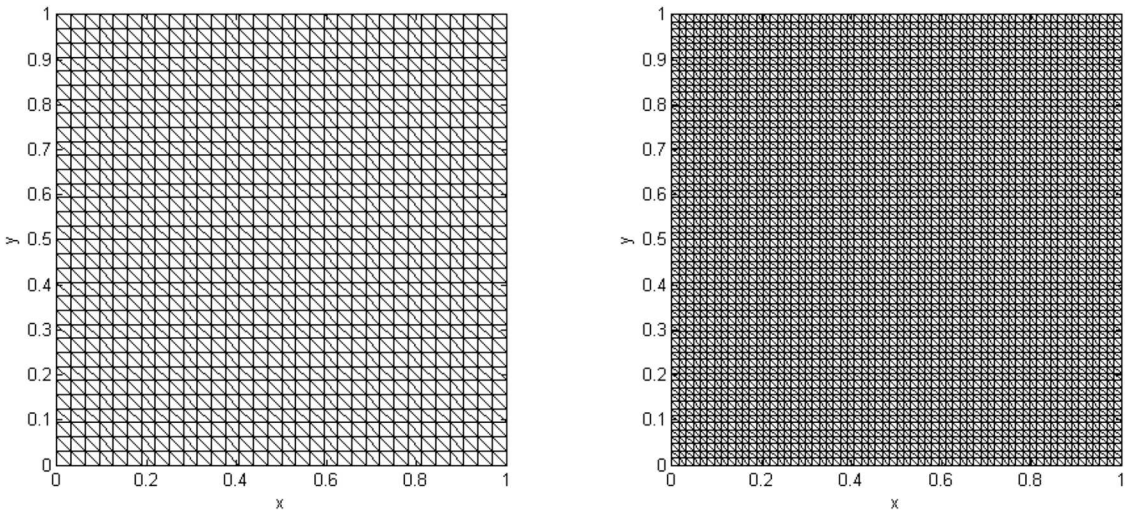

FIGURE 4. Domain $\Omega = (0,1)^2$ partitioned with the coarsest (*left*) and once refined (*right*) uniform right FE triangulations, with $h_x = h_y = h = \frac{1}{32}$ and $h_x = h_y = h = \frac{1}{64}$, respectively. Not shown is the finest triangulation used in this study with $h_x = h_y = h = \frac{1}{128}$.

4.1. Numerical framework

Time discretization was performed using the Marchuk-Yanenko (MY) operator-splitting scheme (2.2)–(2.5) of the initial value problem (2.1), while space discretization was performed using a particular case of the classical finite element (FE) method developed for the current problem in Section 3, specifically the discrete problem formulation (3.6), (3.8) on particular uniform right FE triangulations of the unit square (two of which are shown in Fig. 4). For all of the numerical experiments, we parametrized the various discrete problems with at least the 26 values of C within the set

$$\mathscr{C} \equiv \{C(q)\}_{q=1,\cdots,26} = \{0, 0.025, 0.05, 0.1, 0.2, \cdots, 1.6, 1.8, 2.0, 2.2, 2.5, 2.8, 3.15, 3.5\},$$

which was sufficient to capture the qualitative behavior of the solution branches corresponding to the various values of the parameter τ_y used in the experiments and the solver's quantitative performance in computing the various steady states along these branches. Of particular interest, as it turns out, is the behavior of the solver for vanishingly small values of C, namely $C = 0.025$ and smaller, discussed in Remark 4.3.

4.2. Numerical experiments

Among our first numerical experiments were solving the discrete BG flow problem (2.2)–(2.5) with the largest flow time step size ($\Delta t = \frac{1}{1000}$) on the coarsest ($h = \frac{1}{32}$) FE triangulation for all values of $C \in \mathscr{C}$ to verify recovery of these now classical and well-known numerical results, noting that the operator-splitting step (2.3) degenerates into a linear elliptic problem when τ_y vanishes (henceforth we refer to operator-splitting step (2.3) as the "Bingham step" and steps (2.4), (2.5) together as the "Bratu step"). Figure 5 shows five numerical bifurcation diagrams resulting from plotting the λ_h^{n+1} values computed *via* the Newton algorithm (2.22), (2.23) and *via* the Rayleigh quotient (RQ) (2.27) formed using the four possible combinations of the Bingham step $u_h^{n+\frac{1}{3}}$ and Bratu step u_h^{n+1} steady states in the numerator and/or denominator, for each value of $C \in \mathscr{C}$ (note the axes are transposed compared with the classical bifurcation diagram in Fig. 1). Plotting these bifurcation diagrams on the same axes gives some insight into the effect of operator-splitting on the numerical results. Careful examination and comparison of these diagrams and corresponding computed values reveals:

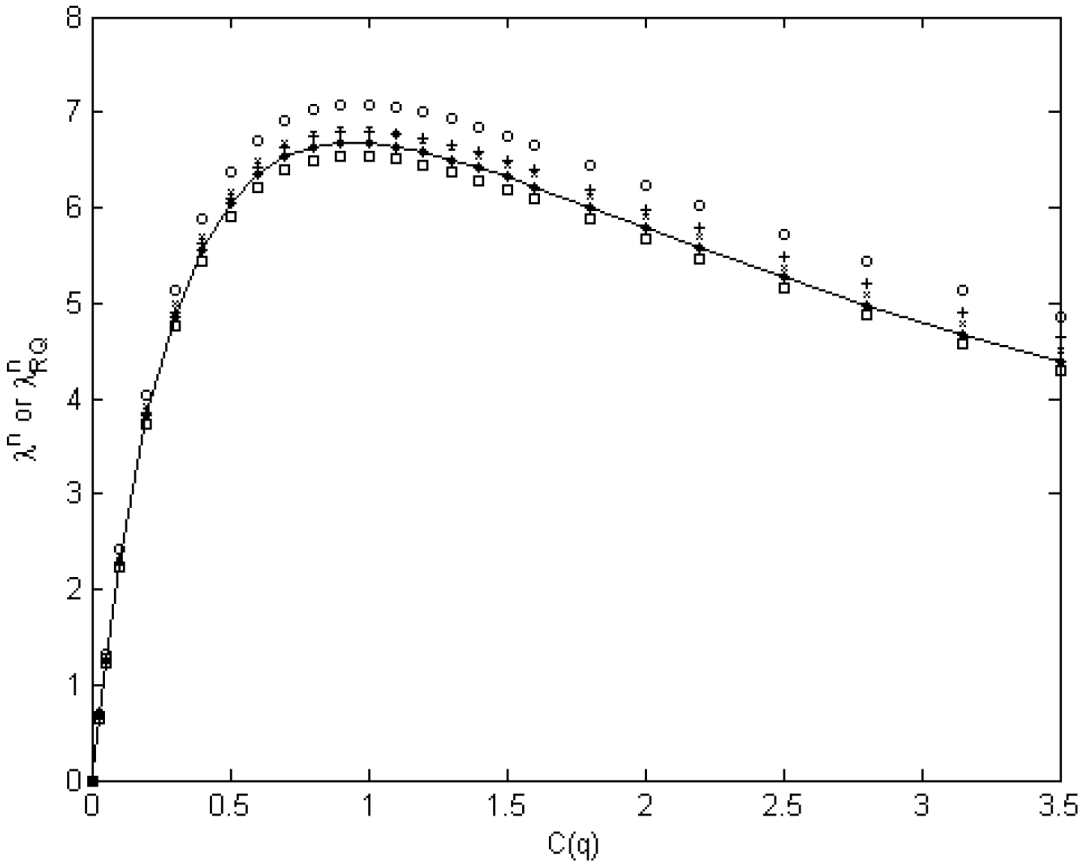

FIGURE 5. Bifurcation diagrams for the Bratu-Gelfand ($\tau_y = 0$) problem (2.2)–(2.5) computed with flow time step size $\Delta t = \frac{1}{1000}$ on the coarsest FE triangulation with $h_x = h_y = h = \frac{1}{32}$ plotted using the value of λ_h^{n+1} computed *via* the Newton algorithm (2.22), (2.23) (solid line connecting the ♦) and *via* the Ralyleigh quotient (2.27) using the numerator/denominator Bingham step/Bratu step (□), Bingham step/Bingham step (+), Bratu step/Bratu step (×), and Bratu step/Bingham step (o) steady state combinations.

– The heterogeneous RQ values (computed using the Bingham resp. Bratu step steady state in the numerator and Bratu resp. Bingham step steady state in the denominator of (2.27)) sandwich both homogeneous RQ values (computed using either the Bingham or Bratu step steady state in both numerator and denominator of (2.27)) as well as the Newton algorithm λ_h^{n+1} values, for every $C \in \mathscr{C}$. Furthermore, the Bingham/Bratu resp. Bratu/Bingham RQ values are uniformly smaller resp. larger than the corresponding former three values, for every $C \in \mathscr{C}$.
– The homogeneous Bingham step RQ values are smaller resp. larger than the corresponding Bratu step RQ values before resp. after the turning point, and these homogenous RQ values are uniformly larger than the Newton algorithm values, for every $C \in \mathscr{C}$.
– Since the product of the heterogeneous RQ values is identical to that of the corresponding homogeneous RQ values (by construction) for every $C \in \mathscr{C}$, the geometric means of the heterogeneous RQ values and corresponding homogeneous RQ values are also identical for every $C \in \mathscr{C}$, and these values are uniformly larger than the corresponding Newton algorithm λ_h^{n+1} values, for every $C \in \mathscr{C}$.

TABLE 1. Values of the multiplier λ_h^{n+1} for $C = 0.9$ resp. $C = 0.3$ near the turning point of the Bratu-Gelfand ($\tau_y = 0$) resp. Bingham-Bratu-Gelfand ($\tau_y = 2.5$) problem bifurcation diagram computed *via* the Newton algorithm (2.22)–(2.23) and by forming the Rayleigh quotient (2.27) using the indicated numerator/denominator pairs of operator-splitting step steady states for the indicated flow time step and FE mesh sizes.

		$\tau_y = 0, C = 0.9, h =$			$\tau_y = 2.5, C = 0.3, h =$		
Δt	Steady state λ_h^{n+1} calculated via:	$\frac{1}{32}$	$\frac{1}{64}$	$\frac{1}{128}$	$\frac{1}{32}$	$\frac{1}{64}$	$\frac{1}{128}$
$\frac{1}{1000}$	Newton algorithm	6.6742	6.6749	6.6751	13.4905	13.4864	13.4855
	Bingham/Bratu RQ	6.5392	6.5390	6.5386	12.7669	12.7518	12.7451
	Bingham/Bingham RQ	6.7938	6.7944	6.7945	13.7766	13.7720	13.7708
	Bratu/Bratu RQ	6.8151	6.8210	6.8301	13.8833	13.9340	14.0338
	Bratu/Bingham RQ	7.0803	7.0874	7.0974	14.9813	15.0488	15.1632
	Geometric mean of RQs	6.8044	6.8077	6.8123	13.8298	13.8528	13.9016
	Implied operator-splitting error	1.9135%	1.9507%	2.0140%	2.4534%	2.6450%	2.9932%
$\frac{1}{2000}$	Newton algorithm	6.7380	6.7388	6.7390	13.6238	13.6197	13.6188
	Bingham/Bratu RQ	6.6692	6.6695	6.6694	13.2550	13.2453	13.2415
	Bingham/Bingham RQ	6.7987	6.7995	6.7996	13.7694	13.7650	13.7640
	Bratu/Bratu RQ	6.8083	6.8106	6.8132	13.8115	13.8227	13.8479
	Bratu/Bingham RQ	6.9405	6.9434	6.9463	14.3474	14.3650	14.3944
	Geometric mean of RQs	6.8035	6.8050	6.8064	13.7904	13.7938	13.8059
	Implied operator-splitting error	0.9627%	0.9728%	0.9911%	1.2081%	1.2622%	1.3552%
$\frac{1}{4000}$	Newton algorithm	6.7704	6.7712	6.7714	13.6991	13.6917	13.6898
	Bingham/Bratu RQ	6.7356	6.7362	6.7362	13.4785	13.4673	13.4633
	Bingham/Bingham RQ	6.8010	6.8018	6.8019	13.7651	13.7576	13.7557
	Bratu/Bratu RQ	6.8055	6.8068	6.8077	13.7996	13.7987	13.8065
	Bratu/Bingham RQ	6.8715	6.8731	6.8741	14.0931	14.0962	14.1063
	Geometric mean of RQs	6.8032	6.8043	6.8048	13.7823	13.7781	13.7811
	Implied operator-splitting error	0.4821%	0.4865%	0.4908%	0.6037%	0.6271%	0.6625%

For $C = 0.9$ resp. $C = 0.3$ near their respective turning points along the corresponding bifurcation diagrams for the BG resp. $\tau_y = 2.5$ BBG case, we solved (2.2)–(2.5) using three flow time step sizes $\Delta t = \frac{1}{1000}$, $\frac{1}{2000}$ and $\frac{1}{4000}$ and three FE mesh sizes $h = \frac{1}{32}$, $\frac{1}{64}$ and $\frac{1}{128}$ for all $C \in \mathscr{C}$, reporting in Table 1 the resulting Newton algorithm (2.22), (2.23) along with the corresponding RQ (2.27) λ_h^{n+1} values, as well as the corresponding geometric means of the RQ values and the implied operator-splitting errors, each of the latter defined (we justify this definition below) as the relative difference between the geometric mean of the RQ and corresponding Newton algorithm λ_h^{n+1} values expressed as a percentage of the former. Examining the values in Table 1, we observe the following:

- For the BG case, the reported approximate turning point occurs near $C = 0.9$ with value consistent with the classical $\lambda_c \approx 6.81$ value reported in the literature (*e.g.* [10], p. 223). For the BBG case, the reported approximate turning point occurs near $C = 0.3$ closer to the λ-axis, with the Newton algorithm and various RQ λ_h^{n+1} values larger, than their BG counterparts. It is important to note that we made no attempt to find and use the value of C corresponding to the turning point in our numerical experiments, nor did we attempt to estimate the turning point value from the neighboring λ_h^{n+1} values computed for the corresponding C values used in the numerical experiments.
- In the BG resp. $\tau_y = 2.5$ BBG case, the Newton algorithm λ_h^{n+1} values increase with flow time step size in both cases, and with larger variation, compared with only slight increase resp. decrease with FE mesh size. Specifically, the maximum (taken across values for all reported FE mesh sizes) relative variation

(as a fraction of the value for the largest flow time step size) in the Newton algorithm λ_h^{n+1} values as a function of flow time step size is 1.4427% resp. 1.5463%, while the maximum (taken across values for all reported flow time step sizes) relative variation (as a fraction of the value on the coarsest FE mesh) in the Newton algorithm λ_h^{n+1} values as a function of FE mesh size is just 0.0148% resp. 0.0679%, a flow time step-to-FE mesh size sensitivity ratio of 97.48:1 resp. 22.77:1. This shows that, while the $\tau_y = 2.5$ BBG case is significantly more sensitive than the BG case to FE mesh size, both cases are far more sensitive to flow time step size.

- In both the BG and $\tau_y = 2.5$ BBG cases, the maximum relative variation (or "spread") in RQ λ_h^{n+1} values (expressed as a fraction of the minimum RQ value) with flow time step size refinement decreases by roughly the flow time step refinement factor for fixed FE mesh size, the widest such spreads occurring for the finest FE mesh and tighten from 8.5462% to 4.1518% to 2.0471% in the BG case, and even more so from 18.9728% to 8.7067% to 4.7759% in the BBG case, for the largest to medium to smallest flow time step size. On the other hand, the RQ spreads are almost invariant as a function of FE mesh size in both the BG and BBG cases, ever so slightly widening from coarsest to medium to finest FE mesh by 0.2507% to 0.0806% to 0.0289% in the former, and widening more from 1.6279% to 0.4653% to 0.2161% in the latter, for largest to medium to smallest flow time step size.
- In both the BG and $\tau_y = 2.5$ BBG cases, the geometric mean of the RQ λ_h^{n+1} values is relatively invariant with both flow time step and FE mesh sizes. Specifically, for the BG resp. BBG cases, the maximum (taken across values for all reported FE mesh sizes) relative variation (as a fraction of the value for the largest flow time step size) in the geometric means of the RQ λ_h^{n+1} values as a function of flow time step size is 0.1101% resp. 0.8668%, and the maximum (taken across values for all reported flow time step sizes) relative variation (as a fraction of the value on the coarsest FE mesh) in the geometric means of the RQ λ_h^{n+1} values as a function of FE mesh size is 0.1161% resp. 0.5192%, a nearly 1:1 flow time step size to FE mesh size sensitivity ratio of 0.95:1 in the BG case and just slightly larger at 1.67:1 in the BBG case.
- Finally, the implied operator-splitting error defined as the relative difference between the geometric mean of the RQ and Newton algorithm λ_h^{n+1} values, expressed as a fraction of the former, decreases with flow time step size refinement by roughly the flow time step refinement factor for fixed FE mesh size in both the BG and $\tau_y = 2.5$ BBG cases, with slightly larger magnitudes in the latter case.

Our primary conclusions based on these observations are (1) because it is essentially invariant as a function of both flow time step and FE mesh sizes, we posit that the geometric mean of the RQ λ_h^{n+1} values computed *via* (2.27) gives the best approximation of the "true" RQ value computed *via* (1.25) in the absence of operator-splitting, and (2) because the Newton algorithm λ_h^{n+1} values are much more sensitive to flow time step than to FE mesh size, we further posit that the relative difference between the geometric mean of the RQ and corresponding Newton algorithm λ_h^{n+1} values, expressed as a fraction of the former, provides an implicit measure of the operator-splitting error for every value of $C \in \mathscr{C}$.

Next we examine some steady state solutions of the discrete BG and $\tau_y = 2.5$ BBG flow problems (2.2)–(2.5) in detail. Figure 6 in the BG case resp. Figure 7 in the BBG case shows the Bratu step u_h^{n+1} steady states, respectively, for $C = 0.1$, 0.5, and 1.0 computed using a flow time step of $\Delta t = \frac{1}{4000}$ and FE mesh sizes $h = \frac{1}{32}, \frac{1}{64}$ and $\frac{1}{128}$. Compared with the BG case, the effect of rigidity in the BBG case is clearly visible as a plateau region in the steady state's extrema where the gradient vanishes, the area of such increasing with τ_y for fixed C and decreasing with C for fixed τ_y. Closer examination of these steady states and their Bingham counterparts, comparing (for example) their $L^\infty(\Omega)$ norms as exhibited in Table 2 (also for the larger time step sizes $\Delta t = \frac{1}{1000}$ and $\frac{1}{2000}$), we see that they are slightly different, with the norms of the Bingham step steady states slightly smaller than their Bratu step counterparts for a given value of C. Moreover, it can also be seen in the table that the effect of refining the flow time step size is to increase the norm of the computed solutions for a given value of C, with the effect more pronounced in the Bingham step than in the Bratu step steady state.

In Figure 8 we visualize the dependency of the discrete $\tau_y = 2.5$ BBG flow problem (2.2)–(2.5) on flow time step size by plotting the complete Bratu step steady state solution branch λ_h^{n+1} vs. $||u_h^{n+1}||_{L^\infty(\Omega)}$ over the entire range of $C \in \mathscr{C}$ for each flow time step size. Examining the figure, the dependency of the solution

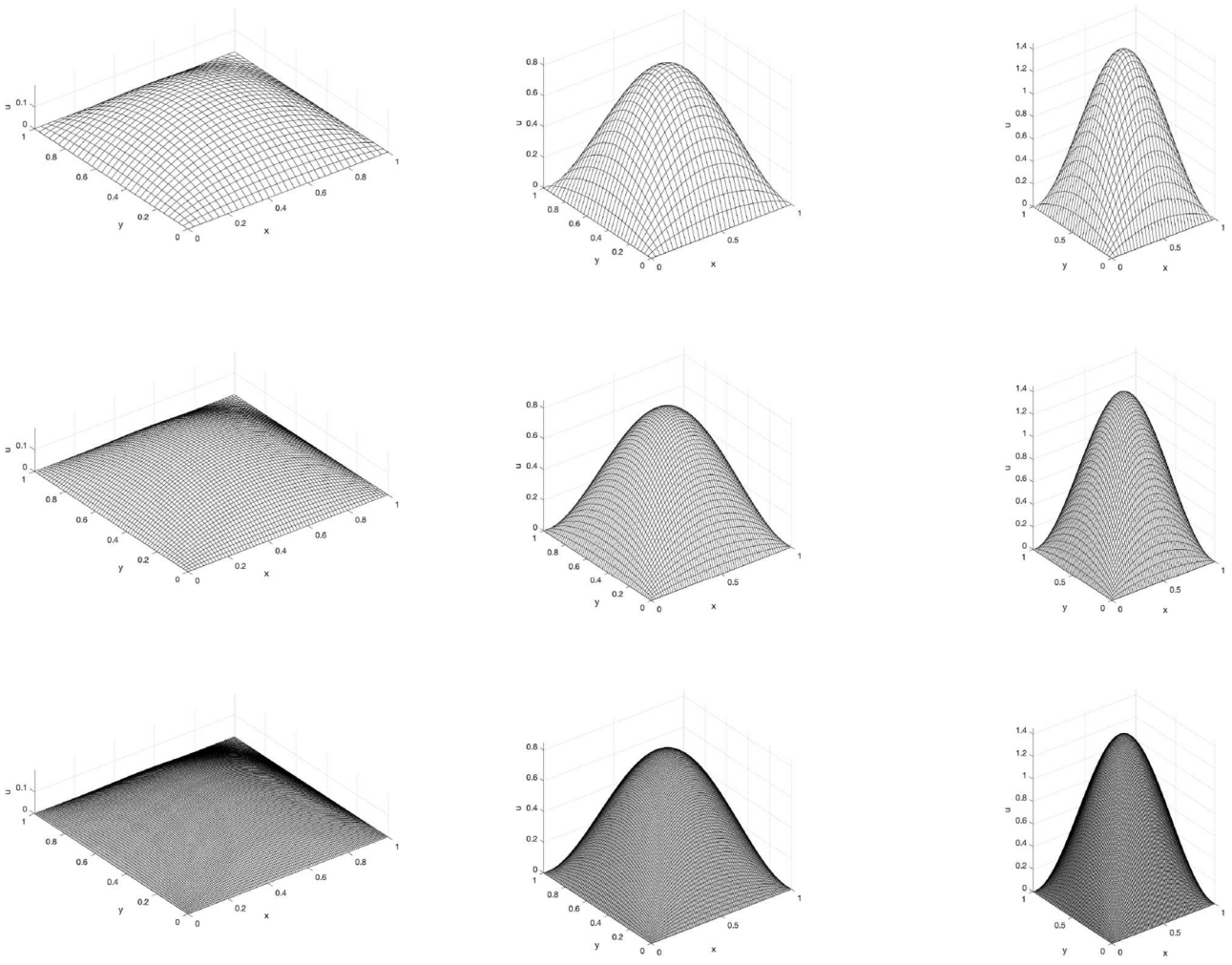

FIGURE 6. Selected Bratu-Gelfand flow steady states u_h^{n+1} solving the Bratu step (2.4), (2.5) computed *via* Euler-Lagrange equations (2.17) using algorithm (2.22)–(2.24) with flow time step size $\Delta t = \frac{1}{4000}$ and FE mesh sizes $h = \frac{1}{32}$ (*top*), $h = \frac{1}{64}$ (*middle*), $h = \frac{1}{128}$ (*bottom*) for $C = 0.1$ (*left*), $C = 0.5$ (*center*), and $C = 1.0$ (*right*).

pair $\left(u_h^{n+1}, \lambda_h^{n+1}\right)$ for each value of C on flow time step size is reflected in the drift of the largest to smallest marker symbol (dot) representing the solution pair computed using the largest to smallest flow time step size. Concerning the behavior of λ_h^{n+1} as a function of flow time step size, we can clearly see that successive flow time step size refinement leads to a corresponding successive, albeit diminishing, increase in λ_h^{n+1} for each value of C. The λ_h^{n+1} spread across all flow time step sizes generally becomes wider with *decreasing* C from a "stationary point" (*i.e.* a point where the norm is invariant, and λ-spread is a minimum, across flow time step size) along the branch in the vicinity of $C = 1.5$. Most significantly, the λ_h^{n+1} spread becomes widest for C closest to zero, indicating that the sensitivity of the solver to flow time step size increases as $C \searrow 0$ (this stands in contrast to the BG case where the flow time step size sensitivity decreases as $C \searrow 0$). The $L^\infty(\Omega)$ norms generally increase (decrease) with flow time step refinement for C before (after) the "stationary point". The $L^\infty(\Omega)$ norm spreads become larger the further C is from zero or the "stationary point", especially noticeable for relatively large C. Note that we omit a similar presentation and analysis of the steady state solution branch dependency on FE

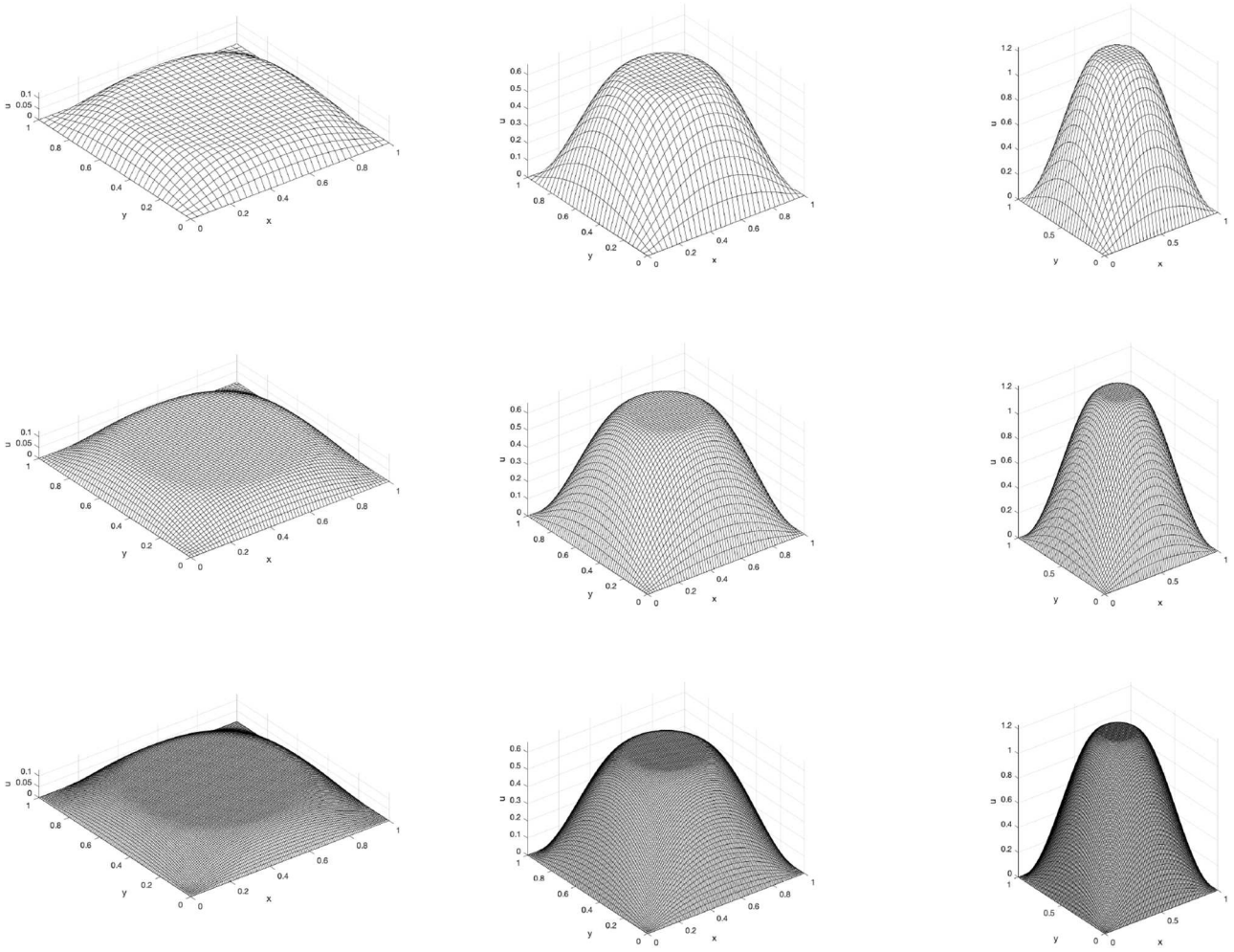

FIGURE 7. Selected Bingham-Bratu-Gelfand flow steady states u_h^{n+1} solving the Bratu step (2.4), (2.5) computed *via* Euler-Lagrange equations (2.17) using algorithm (2.22)–(2.24) with flow time step size $\Delta t = \frac{1}{4000}$ and FE mesh sizes $h = \frac{1}{32}$ (*top*), $h = \frac{1}{64}$ (*middle*), $h = \frac{1}{128}$ (*bottom*) for $\tau_y = 2.5$, $C = 0.1$ (*left*), $C = 0.5$ (*center*), and $C = 1.0$ (*right*).

mesh size, since there is not enough variation across all FE mesh sizes to visualize the dependency in such a presentation.

The performance of our solver in computing the BG resp. $\tau_y = 2.5$ BBG Bingham step and Bratu step steady states (the latter shown in Fig. 6 resp. Fig. 7) is reported in Table 3, where the solver employed the global operator-splitting convergence criterion (2.26) with Bingham step convergence criterion (2.29) and Bratu step convergence criterion (2.30). Examining the table, we see that in the BG case, refining the flow time step size Δt by a factor of two unsurprisingly roughly doubles the number of such steps required for convergence of the operator-splitting scheme (2.2)–(2.5). We also see that for a given flow time step size, the scheme's convergence rate is essentially invariant with FE mesh size, and the maximum and minimum numbers of Newton iterations required for convergence of the Bratu step (2.4), (2.5) either remains unchanged or decreases with FE mesh size refinement. Comparing these BG performance numbers with those in the BBG case, we see that while the Bratu step (2.4), (2.5) Newton algorithm performance is quite similar, the relationship between the flow time

TABLE 2. $L^\infty(\Omega)$ norm of Bingham step $u_h^{n+\frac{1}{3}}$ and Bratu step u_h^{n+1} steady states of the Bratu-Gelfand ($\tau_y = 0$) and Bingham-Bratu-Gelfand ($\tau_y = 2.5$) flows for the indicated values of C and flow time step and FE mesh sizes.

			C					
			0.1		0.5		1.0	
τ_y	Δt	h	$\|\|u_h^{n+\frac{1}{3}}\|\|_{L^\infty(\Omega)}$	$\|\|u_h^{n+1}\|\|_{L^\infty(\Omega)}$	$\|\|u_h^{n+\frac{1}{3}}\|\|_{L^\infty(\Omega)}$	$\|\|u_h^{n+1}\|\|_{L^\infty(\Omega)}$	$\|\|u_h^{n+\frac{1}{3}}\|\|_{L^\infty(\Omega)}$	$\|\|u_h^{n+1}\|\|_{L^\infty(\Omega)}$
0	$\frac{1}{1000}$	$\frac{1}{32}$	0.1952	0.1980	0.8319	0.8459	1.4229	1.4513
		$\frac{1}{64}$	0.1948	0.1976	0.8307	0.8448	1.4214	1.4499
		$\frac{1}{128}$	0.1946	0.1974	0.8304	0.8444	1.4210	1.4495
	$\frac{1}{2000}$	$\frac{1}{32}$	0.1977	0.1991	0.8412	0.8484	1.4374	1.4518
		$\frac{1}{64}$	0.1973	0.1987	0.8402	0.8474	1.4361	1.4505
		$\frac{1}{128}$	0.1972	0.1986	0.8400	0.8471	1.4357	1.4501
	$\frac{1}{4000}$	$\frac{1}{32}$	0.1989	0.1997	0.8460	0.8496	1.4448	1.4520
		$\frac{1}{64}$	0.1986	0.1993	0.8451	0.8486	1.4435	1.4508
		$\frac{1}{128}$	0.1985	0.1992	0.8448	0.8484	1.4432	1.4504
2.5	$\frac{1}{1000}$	$\frac{1}{32}$	0.1101	0.1241	0.6274	0.6531	1.1847	1.2247
		$\frac{1}{64}$	0.1093	0.1232	0.6258	0.6514	1.1828	1.2227
		$\frac{1}{128}$	0.1089	0.1228	0.6253	0.6508	1.1821	1.2220
	$\frac{1}{2000}$	$\frac{1}{32}$	0.1195	0.1266	0.6458	0.6588	1.2089	1.2292
		$\frac{1}{64}$	0.1189	0.1260	0.6444	0.6574	1.2072	1.2274
		$\frac{1}{128}$	0.1186	0.1257	0.6439	0.6569	1.2066	1.2269
	$\frac{1}{4000}$	$\frac{1}{32}$	0.1243	0.1279	0.6551	0.6617	1.2213	1.2315
		$\frac{1}{64}$	0.1238	0.1273	0.6538	0.6604	1.2196	1.2298
		$\frac{1}{128}$	0.1236	0.1272	0.6535	0.6600	1.2191	1.2293

step size and the operator-splitting scheme (2.2)–(2.5) convergence rate is far less linear in the BBG case where, in general, refining the flow time step size demands significantly more time steps than the refinement factor for convergence, and for a given flow time step size, refining the FE mesh size significantly improves the convergence rate, this latter behavior standing in contrast to the BG case. Finally, we note that, unlike the BG case, the convergence rate of the operator-splitting scheme becomes significantly slower as $C \searrow 0$, indicating that the solver has greater difficulty resolving the solutions for smaller values of C, in the BBG case (see later discussion on this point).

We finish our review of the BG and $\tau_y = 2.5$ BBG cases with a comparison of the BG (Fig. 9) and $\tau_y = 2.5$ BBG (Fig. 10) flow problems' Bingham step steady states $u_h^{n+\frac{1}{3}}$, corresponding gradient magnitudes $|\nabla u_h^{n+\frac{1}{3}}|$ and the theoretically equivalent steady state Bingham multipliers (computed *via* fixed point iteration in (2.11), with τ_y omitted in BG case) dotted with the gradients $\mathbf{p}_h^{n+1} \cdot \nabla u_h^{n+\frac{1}{3}}$, computed using the smallest flow time step size ($\Delta t = \frac{1}{4000}$) on the finest ($h = \frac{1}{128}$) FE triangulation for three values of $C = 0.05$, 1, and 3.5 (nearly spanning the entire range of $C \in \mathscr{C}$). The vanishing of the central areas in $|\nabla u_h^{n+\frac{1}{3}}|$ (or equivalently $\mathbf{p}_h^{n+1} \cdot \nabla u_h^{n+\frac{1}{3}}$) in the

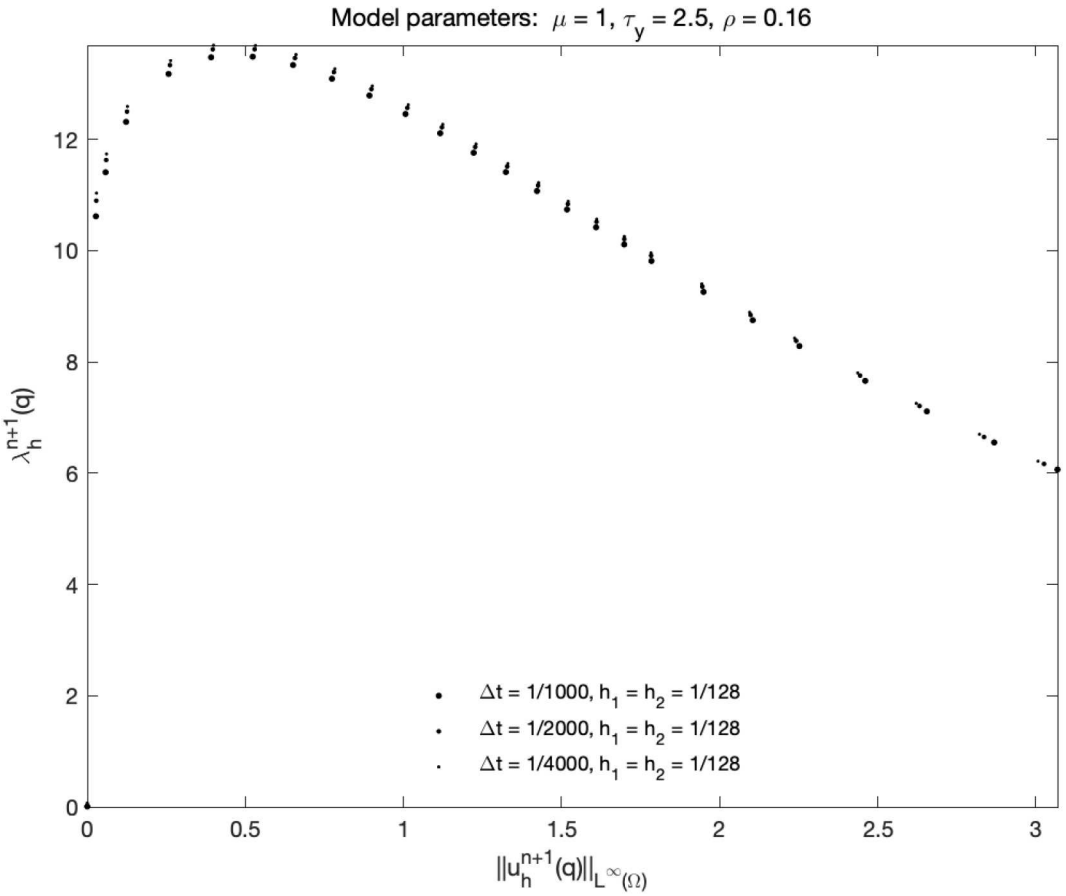

FIGURE 8. Bingham-Bratu-Gelfand λ_h^{n+1} vs. $||u_h^{n+1}||_{L^\infty(\Omega)}$ bifurcation diagrams for $\tau_y = 2.5$, all $C \in \mathscr{C}$ resulting from three flow time step sizes $\Delta t = \frac{1}{1000}$, $\frac{1}{2000}$ and $\frac{1}{4000}$ for fixed FE mesh size $h = \frac{1}{128}$.

BBG compared with the corresponding BG steady states highlights the former's "rigidity plateau". It is natural to wonder why we include both $|\nabla u_h^{n+\frac{1}{3}}|$ and $\mathbf{p}_h^{n+1} \cdot \nabla u_h^{n+\frac{1}{3}}$ for comparison, since the latter is theoretically equivalent to the former. The reason is that such a comparison is useful as an *a posteriori* monitor of whether or not the flow time step size Δt in the solver is sufficiently small to resolve the *correct* Bingham multiplier and steady state upon convergence (according to criterion (2.29)) of the flow. The match between $|\nabla u_h^{n+\frac{1}{3}}|$ and $\mathbf{p}_h^{n+1} \cdot \nabla u_h^{n+\frac{1}{3}}$ for each value of C in Figure 9 or 10 indicates that the chosen time step size is sufficiently small.

Beyond the BG and $\tau_y = 2.5$ BBG cases discussed in detail above, we performed many additional numerical experiments with our solver for other values of τ_y and summarize all results in each of the four panels of Figure 11 showing five computed bifurcation diagrams for $\tau_y = 0.0, 1.0, 2.5, 5.0$, and 10.0 and all $C \in \mathscr{C}$ parametrizing the corresponding numerical experiments, each $\bullet$ representing a computed steady state solution pair $\left(u_h^{n+\frac{j}{3}}(q), \lambda_h^{n+1}(q)\right)$ corresponding to $C(q)$ for $q = 1, \cdots, 26$, $j = 1$ (Bingham step) or 3 (Bratu step) on the corresponding diagram. Three of the panels show λ_h^{n+1} vs. $L^\infty(\Omega)$ (upper left), $L^2(\Omega)$ (lower left), and $H_0^1(\Omega)$ (lower right) norms of the Bratu step solution u_h^{n+1}, the remaining panel showing λ_h^{n+1} vs. the $H_0^1(\Omega)$ (upper

TABLE 3. Number of Bratu-Gelfand ($\tau_y = 0$) and Bingham-Bratu-Gelfand ($\tau_y = 2.5$) flow time steps $n^{\Delta t}$ required to obtain the steady states of operator-splitting scheme (2.2)–(2.5), and the maximum $n_{\max}^{\text{Newton}}$ and minimum $n_{\min}^{\text{Newton}}$ numbers of Newton iterations required for convergence of the Bratu step (2.4), (2.5), for the indicated values of C and flow time step and FE mesh sizes.

			C								
			0.1			0.5			1.0		
τ_y	Δt	h	$n^{\Delta t}$	$n_{\max}^{\text{Newton}}$	$n_{\min}^{\text{Newton}}$	$n^{\Delta t}$	$n_{\max}^{\text{Newton}}$	$n_{\min}^{\text{Newton}}$	$n^{\Delta t}$	$n_{\max}^{\text{Newton}}$	$n_{\min}^{\text{Newton}}$
0	$\frac{1}{1000}$	$\frac{1}{32}$	123	9	3	125	8	3	118	7	2
		$\frac{1}{64}$	123	8	3	125	7	3	118	6	2
		$\frac{1}{128}$	123	7	3	125	7	2	118	6	2
	$\frac{1}{2000}$	$\frac{1}{32}$	256	9	3	260	8	2	247	7	2
		$\frac{1}{64}$	256	8	3	259	7	2	246	7	2
		$\frac{1}{128}$	256	7	2	259	7	2	246	6	2
	$\frac{1}{4000}$	$\frac{1}{32}$	538	10	2	547	9	2	521	7	2
		$\frac{1}{64}$	537	8	2	545	8	2	519	7	2
		$\frac{1}{128}$	537	7	2	544	7	2	519	6	2
2.5	$\frac{1}{1000}$	$\frac{1}{32}$	287	10	2	157	8	2	130	7	2
		$\frac{1}{64}$	114	8	2	109	7	3	130	6	2
		$\frac{1}{128}$	96	7	2	117	7	2	129	6	2
	$\frac{1}{2000}$	$\frac{1}{32}$	1668	10	1	735	8	1	559	7	2
		$\frac{1}{64}$	997	9	1	693	7	2	592	7	2
		$\frac{1}{128}$	206	8	1	233	7	2	275	6	2
	$\frac{1}{4000}$	$\frac{1}{32}$	3939	10	1	1423	8	1	1083	7	1
		$\frac{1}{64}$	2756	9	1	1177	8	1	699	7	1
		$\frac{1}{128}$	1412	8	1	979	7	1	887	6	1

right) norm of the Bingham step solution $u_h^{n+\frac{1}{3}}$. To compute each solution pair, we utilized the initialization, branch transit strategies and termination criteria described in Section 2.5.

From the graphs in Figure 11, it is clear that the qualitative behavior exhibited in Figure 2 by the toy problem of Remark 1.8 largely carries over to the behavior observed in the results produced with our solver, specifically the turning points move closer to the λ-axis as τ_y increases. This behavior indicates that with continued increase in τ_y to some critical value, the turning point would eventually intersect the λ-axis. While we did not perform enough numerical experiments to find (or even bracket) this critical value of τ_y, in principle it should be possible using smaller values of C and larger values of τ_y, assuming our solver is robust enough for the task. One peculiarity with the operator-splitting scheme (2.2)–(2.5), as evidenced in the bifurcation diagrams of Figure 11, is that the flows initialized with $C_2 = 0.025$ steady states to obtain the $C_1 = 0$ steady states for the various values of τ_y never produced appreciably non-zero values of λ_h^{n+1} intersecting the λ-axis using the flow convergence criterion prescribed in (2.26). There is indication in Figure 8 that this may be, in part, a consequence of the increased sensitivity of the computed value of $\lambda_h^{n+1}(q)$ to the flow time step size as the

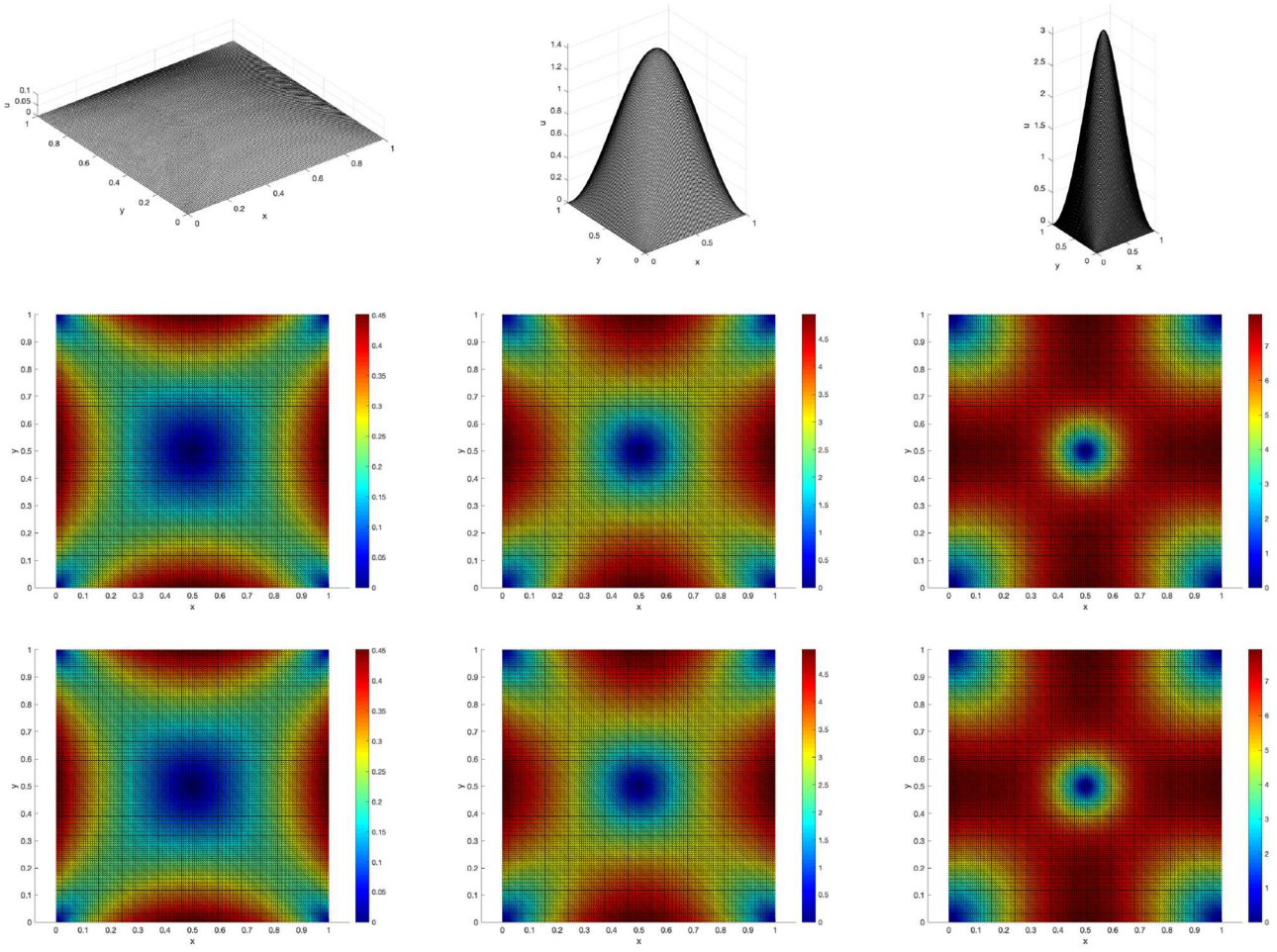

FIGURE 9. Selected Bratu-Gelfand flow steady states $u_h^{n+\frac{1}{3}}$ solving the Bingham step (2.3) computed using the generalized variant of algorithm (2.9)–(2.11) with flow time step size $\Delta t = \frac{1}{4000}$ and FE mesh size $h = \frac{1}{128}$ (*top*) together with the Bingham step steady state gradient magnitude $|\nabla u_h^{n+\frac{1}{3}}|$ (*middle*) and steady state Bingham multiplier dotted with the gradient $\mathbf{p}_h^{n+1} \cdot \nabla u_h^{n+\frac{1}{3}}$ (*bottom*) for $C = 0.05$ (*left*), $C = 1$ (*center*), and $C = 3.5$ (*right*).

value of C gets closer to zero (we explore this further in Remark 4.3). In comparing the $H_0^1(\Omega)$ norm Bingham and Bratu step bifurcation diagrams, we see that they are very similar, with the Bratu step diagrams slightly right-shifted (*i.e.* larger solution norms) compared with the corresponding Bingham step diagrams.

4.3. Remarks on some finer details of the solver

Remark 4.1. During the course of computing some 375 flows for $C \in \mathscr{C}$, $C > 0$ using operator-splitting scheme (2.2)–(2.5) with the largest flow time step size of $\Delta t = \frac{1}{1000}$, we encountered just two cases $(\tau_y, C) = (0, 0.025)$ and $(\tau_y, C) = (10, 0.025)$ where this flow time step size was too large for properly resolving the Bingham and/or Bratu step steady states, as evidenced by a mismatch between $\mathbf{p}_h^{n+1} \cdot \nabla u_h^{n+\frac{1}{3}}$ and $|\nabla u_h^{n+\frac{1}{3}}|$ and/or incorrect λ_h^{n+1} values upon convergence. We exploited these *a posteriori* clues to successively refine the flow time step size until the proper steady states emerged, thereby improving the numerical results produced by the solver.

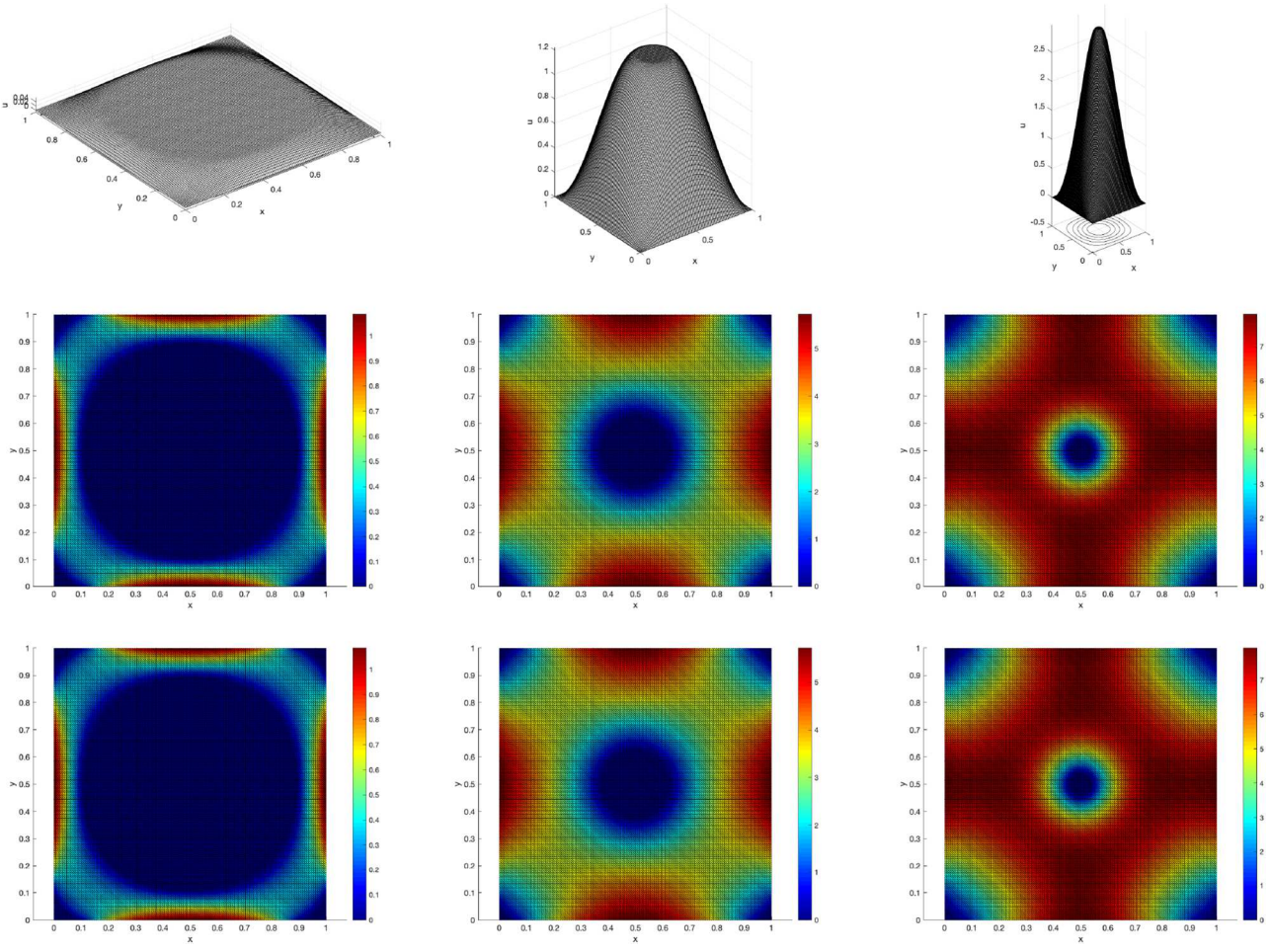

FIGURE 10. Selected $\tau_y = 2.5$ Bingham-Bratu-Gelfand flow steady states $u_h^{n+\frac{1}{3}}$ solving the Bingham step (2.3) computed using the generalized variant of algorithm (2.9)–(2.11) with $\Delta t = \frac{1}{4000}$, $h = \frac{1}{128}$ (*top*) together with the Bingham step steady state gradient magnitude $|\nabla u_h^{n+\frac{1}{3}}|$ (*middle*) and the steady state Bingham multiplier dotted with the gradient $\mathbf{p}_h^{n+1} \cdot \nabla u_h^{n+\frac{1}{3}}$ (*bottom*) for $C = 0.05$ (*left*), $C = 1$ (*center*), and $C = 3.5$ (*right*).

Remark 4.2. Solving (2.24) (with explicit dependence on h added to the notation) $u_h^k - \lambda_h^k \Delta t e^{u_h^k} = u_h^{n+\frac{1}{3}}$ pointwise on Ω for u_h^k only depends on h through the Bingham step steady state values $u_h^{n+\frac{1}{3}}$ (which decrease to nearly zero adjacent to the boundary corners) and the λ_h^k values (which are positive and bounded below away from zero) as h is refined. So, in the limit as $h \searrow 0$, (2.24) becomes $u^k - \lambda^k \Delta t e^{u^k} = 0$ on $\partial\Omega$, which very definitely has positive solutions u^k for Δt small enough. Thus, in the limit as $h \searrow 0$, there is an inconsistency of the Bratu operator-splitting step (2.4) as formulated in (2.22)–(2.24) with the homogeneous Dirichlet boundary condition in the (formal) continuous problem (1.6).

Remark 4.3. As implied by the essentially coincident $C = C_{\min} = 0$ steady state solution points $\left(||u_h^{n+1}||, \lambda_h^{n+1}\right)$ near the origin for the various values of $\tau_y > 0$ in any of the bifurcation diagrams of Figure 11, resolving an accurate numerical approximation λ_h^{n+1} of the continuous problem steady state λ-axis

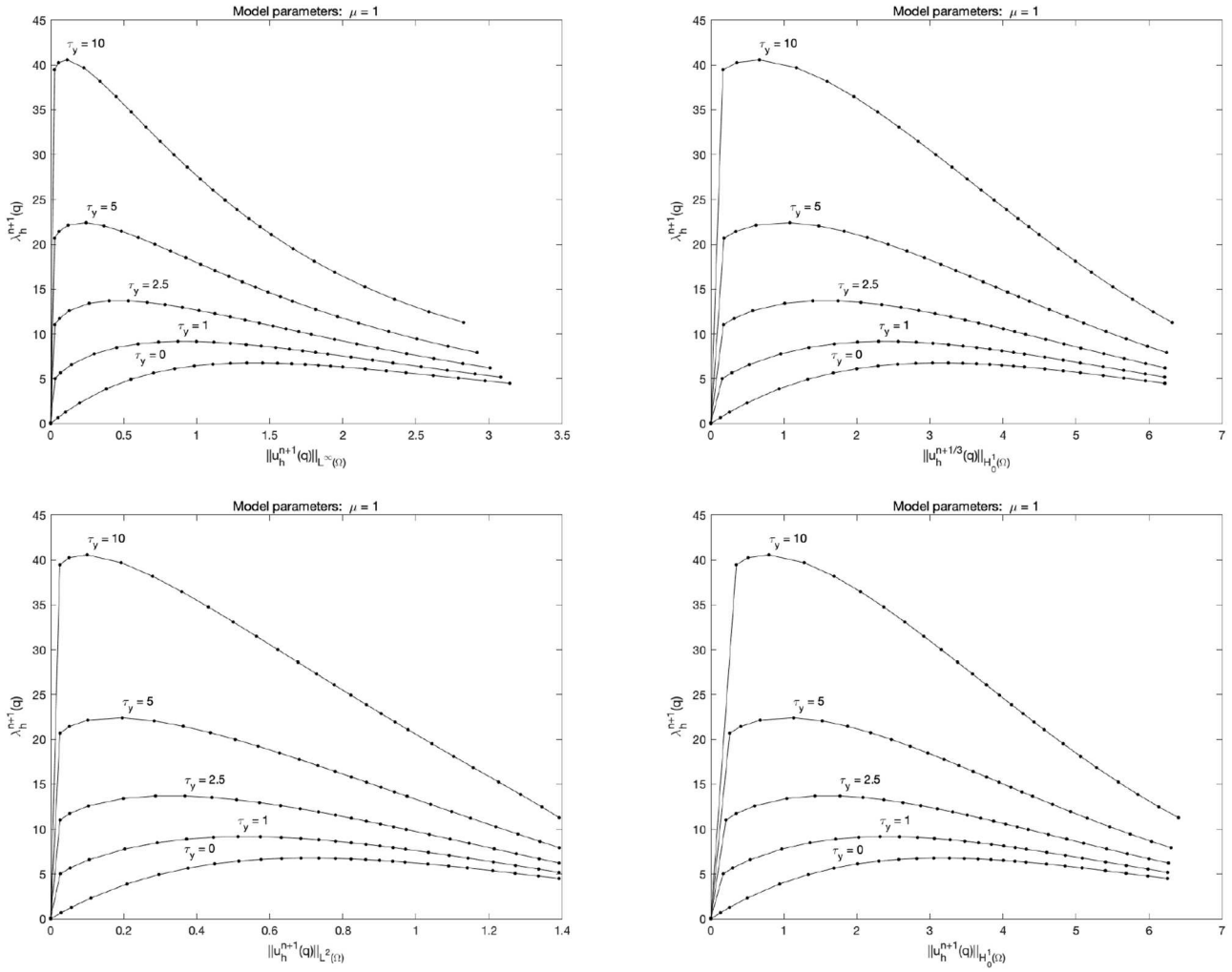

FIGURE 11. Bifurcation diagrams corresponding to parameter values $\tau_y = 0.0, 1.0, 2.5, 5.0, 10.0$, plotted with the $L^\infty(\Omega)$ (*upper left*) and $L^2(\Omega)$ (*lower left*) norms of the Bratu step steady states, and the $H_0^1(\Omega)$ norm of the Bingham step (*upper right*) and Bratu step steady states, each diagram corresponding to a sequence $\left\{(u_h^{n+\frac{j}{3}}(q), \lambda_h^{n+1}(q)\right\}_{q=1,\cdots,26;\, j=1 \text{ or } 3}$ of steady state pairs (denoted by $\bullet$ on the bifurcation diagrams) computed with parameter sequence $\mathscr{C}$, where $n = N(q)$ is the number of iterates required to achieve flow convergence, with flow time step size $\Delta t = \frac{1}{4000}$ and FE mesh size $h = \frac{1}{128}$.

intersection for each value of $\tau_y > 0$ proved to be the most difficult (in fact an unattainable) task for our solver as implemented for the work discussed herein. Recall that, based on the analysis of the toy problem discussed in Remark 1.8 of Section 1, we expect the solution branches for $\tau_y > 0$, $C = 0$ to intersect the λ-axis at values significantly greater than zero, in fact near the λ_h^{n+1} values of the $C = 0.025$ steady state solution points along these branches. However, while the resulting $||u_h^{n+1}||$ values were numerically close to (but never exactly) zero, as expected, for all values of $\tau_y > 0$ in which the $C = 0$ BBG steady states were computed using flow time step size $\Delta t = \frac{1}{4000}$ (and allowing the BBG flows to converge according to criterion (86)), it turns out that the computed

TABLE 4. Flow time step sizes Δt (4^{th} column) and corresponding numbers of time steps $n^{\Delta t}$ (5^{th} column) required for convergence (according to criterion (2.26)) of operator-splitting scheme (2.2)–(2.5) in obtaining the steady states, with the indicated $||u_h^{n+1}||$ and λ_h^{n+1} values (6^{th}–8^{th} columns), corresponding to the indicated C values (3^{rd} column) for the $\tau_y = 2.5$ Bingham-Bratu-Gelfand initial value problem (2.1) initialized with steady states corresponding to the indicated initializing steady state C values (1^{st} column) obtained with the indicated flow time step sizes Δt (2^{nd} column).

Initializing steady state		Computed steady state					
C	Δt	C	Δt	$n^{\Delta t}$	$\|\|u_h^{n+1}\|\|_{L^\infty(\Omega)}$	$\|\|u_h^{n+1}\|\|_{L^2(\Omega)}$	λ_h^{n+1}
0.05	$\frac{1}{4000}$	0.025	$\frac{1}{4000}$	1588	2.9150e-02	2.5933e-02	11.0326
0.025	$\frac{1}{2000}$	0.0025	$\frac{1}{8000}$	3035	2.6501e-03	2.5370e-03	9.8092
0.0025	$\frac{1}{8000}$	0.00025	$\frac{1}{64000}$	15783	2.6133e-04	2.5306e-04	9.5635
0.00025	$\frac{1}{64000}$	2.5e-05	$\frac{1}{512000}$	48512	2.5868e-05	2.5243e-05	9.5333
0.025	$\frac{1}{4000}$	0.0	$\frac{1}{4000}$	15206	1.7960e-05	7.9171e-06	0.0721

values of λ_h^{n+1} were also close to zero, a result that was unexpected and undesirable. Besides these erroneous values of λ_h^{n+1}, the tell-tale sign that these $\tau_y > 0$, $C = 0$ steady states were not the desired ones was, again, a mismatch between $\mathbf{p}_h^{n+1} \cdot \nabla u_h^{n+\frac{1}{3}}$ and $|\nabla u_h^{n+\frac{1}{3}}|$. Moreover, the flow convergence rates were all significantly slower than those for the corresponding $C = 0.025$ flows with the same flow time step size $\Delta t = \frac{1}{4000}$.

In order to gain some insight into the source of the difficulty in resolving the BBG solution branch λ-axis intersection points, we arbitrarily focused our attention on the $\tau_y = 2.5$ BBG dynamical flow problem for additional scrutiny of the solver's behavior in the case of very small values of C. Specifically, we added the values $C = 0.0025$, 0.00025 and $2.5e-05$ to the set $\mathscr{C}$ and computed the corresponding BBG flows on the $h = \frac{1}{128}$ FE mesh. The flows were initialized with steady states previously obtained from BBG flows whose C and Δt values are listed in Table 4, and each flow was restarted with time step sizes Δt successively refined by factors of 2 until apparent, from the behavior of the developing λ_h^{n+1} sequence, that the flow was converging to the *correct* steady state (as verified after convergence by a match between $\mathbf{p}_h^{n+1} \cdot \nabla u_h^{n+\frac{1}{3}}$ and $|\nabla u_h^{n+\frac{1}{3}}|$). Note that the flows were allowed to run for the numbers of flow time steps $n^{\Delta t}$ required for convergence according to criterion (2.26). From the results of these additional numerical experiments, we indeed gained some additional insight into the solver's behavior as $C \searrow 0$ as summarized below:

- Examining the values reported in rows 1–4 of Table 4, the correct steady states corresponding to the successively (order-of-magnitude) smaller values of C were attainable, but only by successively refining the flow time step size $\Delta t = \frac{1}{4000}$ used in the $C = 0.025$ case by factors of 2, 16, and 128 for the $C = 0.0025, 0.00025$ and $2.5e-05$ cases, respectively. This shows (specifically for $\tau_y = 2.5$, but likely the other $\tau_y > 0$ as well) that it is possible to resolve the correct steady state for the BBG problem with arbitrarily small C provided a sufficiently small flow time size is used. However, we see from these results that the flow time step size may become too small to practically compute the entire flow to obtain the steady state for vanishingly small C. So, it seems clear that the solver's source of difficulty in resolving the BBG solution branch λ-axis intersection points is its sensitivity to the flow time step size as $C \searrow 0$.
- The norms of the Bratu step steady states u_h^{n+1} reported in columns 6 & 7 of Table 4 are all consistent with the corresponding values of C (column 3) in the BBG flows used to compute the steady states, while the norms together with the λ_h^{n+1} value in the $C = 0$ case (last row) are for the *incorrect* steady state resulting from flow time step size $\Delta t = \frac{1}{4000}$ and are poor approximations to the expected $||u|| = 0$ and $\lambda \gg 0$ λ-axis intersection point.

– It would seem that, with our solver as implemented, we can only find small-C approximations to the $C = 0$ steady state λ-axis intersection points for the continuous problem solution branches. Examining the values of λ_h^{n+1} for the successively smaller non-zero values of C in Table 4, they appear to be converging. With these values, we were able to use *Richardson extrapolation* (we omit the details) to estimate the correct $\tau_y = 2.5$, $C = 0$ steady state λ-axis intersection point as 9.5291, which stands in stark contrast to the incorrect steady state λ_h^{n+1} value reported in the last row of the table.

Acknowledgements. The authors thank the two reviewers for their very careful reading of the article and their very helpful comments and suggestions.

References

[1] Th. Aubin, Nonlinear Analysis on Manifolds. Monge-Ampère Equations. Number 252 in *Grundlehren der mathematischen Wissenschaften.* Springer, Berlin (1982).
[2] J. Bebernes and D. Eberly, Mathematical Problems from Combustion Theory. Springer, New York, NY (1989).
[3] H. Brezis, Monotonicity methods in Hilbert spaces and some applications to nonlinear partial differential equations. In Contributions to Nonlinear Functional Analysis, edited by E.H. Zarantonello. Academic Press, New York, NY (1971) 101–156.
[4] E.J. Dean, R. Glowinski and G. Guidoboni, On the numerical simulation of Bingham visco-plastic flow: Old and new results. *J. Non-Newtonian Fl. Mech.* **142** (2007) 36–62.
[5] F. Facchinei, A. Fischer and C. Kanzow, Inexact Newton methods for semi-smooth equations with applications to variational inequality problems. In Nonlinear Optimization and Applications, edited by G. Di Pillo and F. Giannessi. Springer, Boston, MA (1996) 125–139.
[6] F.J. Foss, II, R. Glowinski and R.H.W. Hoppe, On the numerical solution of a semilinear elliptic eigenproblem of Lane-Emden type. I: Problem formulation and description of the algorithms. *J. Num. Math.* **15** (2007) 181–208.
[7] F.J. Foss, II, R. Glowinski and R.H.W. Hoppe, On the numerical solution of a semilinear elliptic eigenproblem of Lane-Emden type. II: Numerical experiments. *J. Num. Math.* **15** (2007) 277–298.
[8] R. Glowinski, Numerical Methods for Nonlinear Variational Problems. *Springer Series in Computational Physics.* Springer-Verlag, New York, NY (1984).
[9] R. Glowinski, Finite element methods for incompressible viscous flow. In Numerical Methods for Fluids (Part 3), edited by P.G. Ciarlet and J.L. Lions. In volume IX of *Handbook of Numerical Analysis.* North-Holland, Amsterdam (2003) 3–1176.
[10] R. Glowinski, Variational methods for the numerical solution of nonlinear elliptic problems. Number 86 in *CBMS-NSF Regional Conference Series in Applied Mathematics.* SIAM, Philadelphia, PA (2015).
[11] R. Glowinski, H.B. Keller and L. Reinhart, Continuation-conjugate gradient methods for the least squares solution of nonlinear boundary value problems. *SIAM J. Sci. Stat. Comp.* **6** (1985) 793–832.
[12] R. Glowinski, J.L. Lions and R. Trémolières, Numerical analysis of variational inequalities. Volume 8 of *Studies in Mathematics and its Applications.* North-Holland, Amsterdam (1981).
[13] R. Glowinski, S.J. Osher and W. Yin, Splitting Methods in Communication, Imaging, Science, and Engineering. Scientific Computation. Springer International Publishing, Switzerland (2016).
[14] J.W. He and R. Glowinski, Steady Bingham fluid flow in cylindrical pipes: a time dependent approach to the iterative solution. *Num. Lin. Alg. Appl.* **7** (2000) 381–428.
[15] K. Ito and K. Kunish, Semi-smooth Newton methods for variational inequalities of the 1st kind. *ESAIM: M2AN* **37** (2003) 41–62.
[16] K. Ito and K. Kunish, Semi-smooth Newton methods for the Signorini problem. *Appl. Math.* **53** (2008) 455–468.
[17] K. Majava, R. Glowinski and T. Kärkkäinen, Solving a non-smooth eigenvalue problem using operator-splitting. *Intl. J. Comp. Math.* **84** (2007) 825–846.
[18] P.P. Mosolov and V.P. Miasnikov, Variational methods in the theory of the fluidity of a visco-plastic medium. *J. Appl. Math. Mech.* **29** (1965) 468–492.
[19] L. Qi and D. Sun, Smoothing functions and smoothing Newton method for complementarity and variational inequality problems. *J. Optim. Th. Appl.* **113** (2002) 121–147.
[20] M. Ulbrich, editor. Semi-smooth Newton Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces. *MOS-SIAM Series on Optimization.* SIAM, Philadelphia, PA (2011).
[21] X. Yao, A minimax method for finding saddle critical points of upper semi-differentiable locally Lipschitz continuous functional in Hilbert space and its convergence. *Math. Comp.* **82** (2013) 2087–2136.
[22] X. Yao, Convergence analysis of a minimax method for finding multiple solutions of hemivariational inequality in Hilbert space. *Adv. Comp. Math.* **42** (2016) 1331–1362.