SIAM J. NUMER. ANAL.
Vol. 58, No. 1, pp. 153–188

HIGH-ORDER MASS-LUMPED SCHEMES FOR NONLINEAR DEGENERATE ELLIPTIC EQUATIONS*

JEROME DRONIOU† AND ROBERT EYMARD‡

Abstract. We present and analyze a numerical framework for the approximation of nonlinear degenerate elliptic equations of the Stefan or porous medium types. This framework is based on piecewise constant approximations for the functions, which we show are essentially necessary to obtain convergence and error estimates. Convergence is established without regularity assumption on the solution. A detailed analysis is then performed to understand the design properties that enable a scheme, despite these piecewise constant approximations and the degeneracy of the model, to satisfy high-order error estimates if the solution is piecewise smooth. Numerical tests, based on continuous and discontinuous approximation methods, are provided on a variety of one- and two-dimensional problems, showing the influence on the convergence rate of the nature of the degeneracy and of the design choices.

Key words. nonlinear degenerate elliptic equations, gradient discretization method, error estimate, mass-lumping, finite elements, discontinuous Galerkin, Stefan problem, porous medium equation, numerical scheme

AMS subject classifications. 65N12, 65N15, 65N30, 35J70

DOI. 10.1137/19M1244500

1. Introduction.

The goal of numerical methods for partial differential equations is to approximate, as accurately as possible, the continuous solution. For mesh-based methods, it is well-known that when the problem is linear and the solution has sufficient regularity properties, for a fixed number of degrees of freedom high-order methods provide more accurate solutions than low-order methods. This result must, however, be questioned in the case of nonlinear problems for which, even if the solution is smooth enough, stability and high-order estimates might not be achievable without the proper structure of the chosen discretization. We propose in this work to explore this question, considering the following nonlinear degenerate elliptic equation as the basis of our discussion:

$$\begin{aligned} \beta(\overline{u}) - \operatorname{div}(\Lambda \nabla \zeta(\overline{u})) &= f + \operatorname{div}(F) \qquad \text{in } \Omega, \\ \zeta(\overline{u}) &= 0 \qquad\qquad\quad \text{on } \partial\Omega. \end{aligned} \tag{1.1}$$

The corresponding weak formulation for this problem is

$$\begin{aligned} &\text{Find } \overline{u} \in L^2(\Omega) \text{ such that } \zeta(\overline{u}) \in H_0^1(\Omega) \text{ and} \\ &\int_\Omega \beta(\overline{u})\overline{v} + \int_\Omega \Lambda \nabla \zeta(\overline{u}) \cdot \nabla \overline{v} = \int_\Omega f\overline{v} - \int_\Omega F \cdot \nabla \overline{v} \qquad \forall \overline{v} \in H_0^1(\Omega). \end{aligned} \tag{1.2}$$

*Received by the editors February 13, 2019; accepted for publication (in revised form) October 28, 2019; published electronically January 8, 2020.
https://doi.org/10.1137/19M1244500
Funding: The work of the authors was supported by the Australian government through the Australian Research Council's Discovery Projects funding scheme grant DP170100605 and by the French government through the Agence Nationale de la Recherche (ANR) project CHARMS, grant ANR-16-CE06-0009.
†School of Mathematics, Monash University, Melbourne, Australia (jerome.droniou@monash.edu).
‡Laboratoire d'Analyse et de Mathématiques Appliquées, Université Paris-Est Marne-la-Vallée Champs-sur-Marne, France (Robert.Eymard@u-pem.fr).

Throughout the paper, we denote by $\|\cdot\|_{L^2}$ the norms in $L^2(\Omega)$ or $L^2(\Omega)^d$, and we make the following assumptions:

(1.3a) • Ω is an open bounded connected subset of $\mathbb{R}^d$ ($d \in \mathbb{N}^\star$),

• $\zeta : \mathbb{R} \to \mathbb{R}$ is continuous and nondecreasing, $\zeta(0) = 0$ and,

(1.3b) for some $M_0, M_1 > 0$, $|\zeta(s)| \geq M_0|s| - M_1 \ \forall\ s \in \mathbb{R}$,

• $\beta : \mathbb{R} \to \mathbb{R}$ is continuous and nondecreasing, $\beta(0) = 0$ and,

(1.3c) for some $K_0, K_1 > 0$, $|\beta(s)| \leq K_0|s| + K_1 \ \forall\ s \in \mathbb{R}$,

(1.3d) • $\beta + \zeta : \mathbb{R} \to \mathbb{R}$ is strictly increasing,

• $\Lambda : \Omega \to \mathcal{M}_d(\mathbb{R})$ is measurable and there exists $\overline{\lambda} \geq \underline{\lambda} > 0$ such that,

(1.3e) for a.e. $\boldsymbol{x} \in \Omega$, $\Lambda(\boldsymbol{x})$ is symmetric with eigenvalues in $[\underline{\lambda}, \overline{\lambda}]$.

(1.3f) • $f \in L^2(\Omega)$ and $F \in L^2(\Omega)^d$.

Remark 1.1 (growth assumptions). The superlinearity of ζ assumed in (1.3b) ensures that, even though the model is degenerate, it allows for proper a priori estimates on the solution: since $\zeta(\overline{u}) \in H^1_0(\Omega)$, the superlinearity ensures that $\overline{u}$ belongs to $L^2(\Omega)$ (at least). The sublinearity of β is assumed in (1.3c) to make sure that $\beta(\overline{u})$ belongs to the same Lebesgue space as $\overline{u}$, as this nonlinearity is continuous in this space; this is essential to pass to the limit in the numerical approximations.

The theoretical study of (1.1) is covered by the pioneering paper [7], extended in [3], on problems also including a nonlinear convection term; using techniques that necessitates to multiply the equation by various functions of the unknown, existence, and uniqueness of an entropy solution are obtained. An existence and uniqueness result for the simpler problem considered here is given by Theorem A.1 in Appendix A, without referring to entropy solutions.

The case $\zeta = \mathrm{Id}$ fits into quasilinear second-order elliptic problems, the approximation of which is covered in a rather large literature; see, e.g., [23, 8, 26, 4]. The case $\zeta \neq \mathrm{Id}$, on which we focus in this paper, raises major issues and is less often considered in the literature, especially when considering the question of high-order schemes. First, for such a problem, the solution can display discontinuities when ζ has plateaux. Moreover, the nonlinearities challenge the design of numerical methods that simultaneously (i) only require computing integrals of polynomials (integrals that can be exactly computed in general), (ii) are amenable to error estimates (or, at the very least, proven to be convergent), and (iii) are of order higher than 1.

Extending the entropy method used in [7] to the notion of entropy process solutions, the convergence of a two-point flux approximation (TPFA) finite volume method is proved in [19] for a time-dependent version of (1.1) with $\Lambda = \mathrm{Id}$. The entropy method requires us to consider $\phi(u)$ as test functions for various nonlinear functions ϕ, a process that can only be reproduced at the discrete level for the TPFA scheme; see [13, section 7] and [17]. Unfortunately, the TPFA scheme is only applicable on very specific grids, which usually forces $\Lambda = \mathrm{Id}$, and only low-order error estimates can be expected from the application of the doubling variable technique as in [18].

In the general case of an anisotropic heterogeneous field Λ, we need to consider more versatile schemes than the TPFA scheme, which will necessarily reduce the range of admissible test functions. Nevertheless, an important feature to preserve,

if one wants to ensure the stability of the discretization, is the capacity to choose appropriate test functions to simultaneously get diffusion estimates from the gradient terms, and a positive sign from the reaction term.

Let us first consider the case of conforming Galerkin methods. Given a subspace V_h of $H_0^1(\Omega)$, a conforming scheme for (1.2) is written

(1.4)
$$\text{Find } u \in V_h \text{ such that } \int_\Omega \beta(u)v + \int_\Omega \Lambda\nabla\zeta(u)\cdot\nabla v = \int_\Omega fv - \int_\Omega F\cdot\nabla v \quad \forall v \in V_h.$$

If $u \in V_h$ and ζ is globally Lipschitz continuous, we have $\zeta(u) \in H_0^1(\Omega)$ and the key of the convergence analysis is that the chain rule $\nabla\zeta(u) = \zeta'(u)\nabla u$ enables us to take $v = u \in V_h$ as a test function in the scheme. This choice creates from the diffusion term the quantity of interest $\zeta'(u)|\nabla u|^2$, while the reaction term is nonnegative since $\beta(u)u \geq 0$. However, to deduce any sort of estimate from this choice of test function, we are forced to set $F = 0$, since the term $\int_\Omega F\cdot\nabla u$ cannot in general be estimated using $\int_\Omega \zeta'(u)|\nabla u|^2$. A better choice of test function to estimate the term resulting from the presence of F would be $v = \zeta(u)$ since the diffusion term would provide the quantity $\int_\Omega |\nabla\zeta(u)|^2$, which can be used to estimate $\int_\Omega F\cdot\nabla\zeta(u)$. However, $v = \zeta(u)$ is not a valid test function in the scheme since it does not belong to V_h in general. Fixing $F = 0$, the convergence of (1.4) can nonetheless be proved, but no error estimate can be derived—the reason for this being, again, the lack of freedom in choosing suitable test functions in the scheme. The analysis of conforming approximations is sketched in Appendix B (in which (1.2) is first recast before applying the Galerkin method).

Coming back to the general case of (1.2) with possibly $F \neq 0$, we consider numerical methods for which the chain rule does not hold at the discrete level (as is the case for the majority of nonconforming methods). Unless the model problem is recast with a different form of nonlinearity as in [5, 6], the only reasonable test function to consider in order to get estimates is $v = \zeta(u)$, which formally provides $|\nabla\zeta(u)|^2$ from the diffusion term. More precisely, let us consider a scheme where the discrete unknowns $z = (z_i)_{i\in I}$ represent pointwise values of the solutions at certain nodes, and functions z_h are reconstructed from these values and used in the weak formulation (this is the choice made in [16, 1] in the case of the transient problem, through the use of "Lagrange interpolation operators"). Then, for $u = (u_i)_{i\in I}$, one can easily define $v = \zeta(u)$ pointwise, setting $v_i = \zeta(u_i)$ for all $i \in I$. The weak formulation then involves $\nabla(\zeta(u))_h \cdot \nabla v_h$ and, taking $v = \zeta(u)$, this diffusion term generates the quantity $|\nabla(\zeta(u))_h|^2$.

With this choice of v, the reaction term creates the quantity $\beta(u_h)(\zeta(u))_h$. This function is, at the considered nodes, equal to $\beta(u_i)\zeta(u_i) \geq 0$ (see (1.3b)–(1.3c)). However, outside the nodes, no particular sign can be ensured for $\beta(u_h)(\zeta(u))_h$ and it is not clear that this reaction term will indeed lead to proper estimates on the solution to the scheme.

The way to solve this conundrum is, in the reaction term, to use a different reconstruction of functions than the natural reconstruction $(\cdot)_h$ used for the diffusion term. Utilizing, for example, a piecewise constant reconstruction, in which the only values taken by the reconstruction are nodal values, ensures that the positivity of the reaction term—valid at these nodal values—extends to the entire domain (this is again done for low-order methods in [16, 1] to handle the accumulation terms issued from the time derivative, and in [15] on a variational inequality equivalent to

(1.1) with $\zeta = \mathrm{Id}$ and β multivalued). For linear models, using piecewise constant reconstructions for reaction/accumulation terms leads to what is called mass-lumped schemes. There is a large literature on the mass-lumping of finite element methods for second-order problems; see, e.g., [10, 20, 24, 22] and references therein. In most of these references, though, the construction of mass-lumped versions of high-order methods is justified by a need to reduce computational costs: for explicit discretizations of time-dependent linear problems, a mass-lumped scheme ensures a diagonal mass matrix which, unlike the standard mass matrix, is trivial to invert at each time step. This property of diagonal mass matrix has also been heavily used in schemes for eigenvalues problems related to linear elliptic operators (see, for example, [2] and references therein). On the contrary, for a nonlinear degenerate model as (1.1), as explained above the mass-lumping is not just a way to improve the method's efficiency, but appears as an imperative to establish convergence and error estimates—and thus rigorously ensure that the scheme has high-order approximation properties. Additionally, the usual interpretation of mass-lumping as a specific choice of quadrature rules for the mass matrix is meaningful mostly in the linear setting. For nonlinear models, the less standard interpretation based on piecewise constant reconstructions is more appropriate (even though, as we will see, there is still some link to exploit with local quadrature rules). Finally, let us notice that, to our best knowledge, mass-lumping techniques seem to only be considered in the literature on finite element methods, not in the literature covering other high-order polynomial-based methods such as discontinuous Galerkin (DG). This is understandable when the goal is to simplify the inversion of the mass matrix; mass-lumping is then not much useful to methods such as DG schemes, for which the standard mass matrix is easy to invert due to its block diagonal structure (which can also easily be made fully diagonal by a simple choice of orthogonal local polynomial basis). However, when the primary objective of mass-lumping is to enable convergence and error estimates for nonlinear models, the question of designing mass-lumped DG (or other methods based on local polynomials) is fully relevant.

Our goal in this paper is to design high-order mass-lumped schemes for the nonlinear degenerate model (1.1). Our main contributions can be summarized as follows:

- design of a general analysis framework that treats in a unified way many different methods, including finite element and DG methods (and others);
- proof of error estimates in this general framework;
- identification of conditions on the mass-lumping to ensure high-order convergences (when the exact solution is piecewise smooth), despite the nonlinearities and degeneracy in the model;
- extensive numerical tests, using both $\mathbb{P}^k$ finite element and DG schemes, on realistic test cases (porous medium, Stefan) to validate the theoretical analysis.

Let us describe the organization of this paper. We first provide in section 2 a general formulation of numerical schemes, based on schemes written in fully discrete form: approximate functions and gradients are reconstructed without direct relation, and the approximate functions are piecewise constant. This construction is performed in the gradient discretization method (GDM) [14], a framework that provides efficient notation and notions for the design and analysis of such schemes. After proving a first convergence result (Theorem 2.9) in section 2.1, we establish in section 2.2 error estimates on the approximation of $\zeta(\overline{u})$ when using mass-lumped schemes (Theorem 2.12 and Corollary 2.15). As demonstrated in section 2.3, this general error estimate yields a high-order convergence rate (Theorem 2.24) for piecewise smooth solutions to

(1.2), provided the mass-lumping is performed in a way that corresponds to sufficiently high-order local quadrature rules. These conditions on the local quadrature rules are similar to those highlighted for finite elements in [9, 10] but, interestingly, they appear here from the need of estimating quite different error terms than in the case of linear models as in these references. Extensive numerical tests are presented in section 3, both on porous medium equations and on Stefan problems, using mass-lumped finite element and DG schemes; the results confirm that high-order approximations are obtained only if the aforementioned local quadrature rules hold, even if the theoretical assumptions are not fully satisfied (e.g., the solution is not piecewise smooth). The paper is completed with a short conclusion (section 4) and two appendices. In Appendix A, the properties of the continuous problem are analyzed, and Appendix B sketches the study of the conforming scheme (1.4) with $F = 0$ and highlights its limitations compared to the method in section 2: strong convergence of the gradients only under some regularity assumption on the continuous solution, no error estimate, no uniqueness of the discrete solution.

2. Schemes with piecewise constant approximation. To present the discretization of (1.1), we use the GDM [14], a generic numerical analysis framework for diffusion equations that encompasses many different discretizations: finite element, finite volumes, etc. Using this framework enables a unified treatment of all these different schemes and also gives an efficient setting and tools to deal with them, including the notion of mass-lumping that will be essential to design a scheme for which an error estimate can be established.

The principle of the GDM is to introduce discrete elements—a finite dimensional space, an operator that reconstructs functions, and an operator that reconstructs gradients—together called a gradient discretization (GD), and to replace the continuous counterparts in the weak formulation (1.2) with these discrete elements, leading to a gradient scheme (GS) for (1.1).

DEFINITION 2.1 (gradient discretization). *A GD is* $\mathcal{D} = (X_{\mathcal{D},0}, \Pi_{\mathcal{D}}, \nabla_{\mathcal{D}}, Q_{\mathcal{D}})$ *such that*

- $X_{\mathcal{D},0}$ *a finite dimensional space;*
- $\Pi_{\mathcal{D}} : X_{\mathcal{D},0} \to L^2(\Omega)$ *and* $\nabla_{\mathcal{D}} : X_{\mathcal{D},0} \to L^2(\Omega)^d$ *are linear operators reconstructing, respectively, a function and a gradient;* $\nabla_{\mathcal{D}}$ *must be chosen such that* $\|\cdot\|_{\mathcal{D}} := \|\nabla_{\mathcal{D}}\cdot\|_{L^2}$ *is a norm on* $X_{\mathcal{D},0}$*;*
- $Q_{\mathcal{D}} : L^2(\Omega) \to L^2(\Omega)$ *is a quadrature operator.*

Remark 2.2 (quadrature operator). Quadrature rules for source terms are usually not accounted for in the definition and analysis of GS. In the context of mass-lumped schemes, however, accounting for quadrature rules is essential to establishing optimal high-order error estimates.

Note that $Q_{\mathcal{D}}$ is not assumed to be bounded. This enables different choices of quadrature rules depending on the regularity of the considered functions: $Q_{\mathcal{D}} f$ could be computed using pointwise values of f if f is continuous, or using averaged values of f otherwise.

To deal with the nonlinearity in the derivatives in (1.1) we need the following notion.

DEFINITION 2.3 (piecewise constant reconstruction). *Let* $\mathcal{D}$ *be a GD such that, for some finite sets* I *and* $I_{\partial\Omega} \subset I$*, it holds that*

$$X_{\mathcal{D},0} = \{v = (v_i)_{i\in I} \,:\, v_i \in \mathbb{R} \quad \forall i \in I\,,\ v_i = 0 \quad \forall i \in I_{\partial\Omega}\}.$$

We say that the reconstruction $\Pi_{\mathcal{D}}$ is piecewise constant if there exists a partition $U = (U_i)_{i\in I}$ of Ω (some of the U_i can be empty) such that

$$\forall v = (v_i)_{i\in I} \in X_{\mathcal{D},0}\,, \quad \Pi_{\mathcal{D}} v = \sum_{i\in I} v_i \mathbf{1}_{U_i}, \tag{2.1}$$

where $\mathbf{1}_{U_i}$ is the characteristic function of U_i. In other words, $(\Pi_{\mathcal{D}} v)_{|U_i} = v_i$ for all $i \in I$.

Remark 2.4 (reconstruction operator). Note that if some U_i are empty, then $\Pi_{\mathcal{D}}$ is not injective. This is a classical situation in the GDM; see, for example, the mass-lumped $\mathbb{P}^2$ scheme in Remark 3.1 or the HMM method in Remark 2.8 and [14, Chapter 13].

In the setting of this definition, if $g : \mathbb{R} \to \mathbb{R}$ is a function satisfying $g(0) = 0$, we define (with an abuse of notation) $g : X_{\mathcal{D},0} \to X_{\mathcal{D},0}$ by applying g coefficient by coefficient:

$$\forall v = (v_i)_{i\in I}\,,\ g(v) = (g(v_i))_{i\in I}. \tag{2.2}$$

We note that this definition actually depends on the choice of the basis of $X_{\mathcal{D},0}$. In practice, this basis being canonical and chosen once and for all, we do not make explicit the dependency of $g(v)$ with respect to it. If $\Pi_{\mathcal{D}}$ is a piecewise constant reconstruction, then (2.2) leads to

$$\forall v \in X_{\mathcal{D},0}\,,\ \Pi_{\mathcal{D}} g(v) = g(\Pi_{\mathcal{D}} v). \tag{2.3}$$

The accuracy properties of a GD are assessed through the following quantities. The first one measures a discrete Poincaré constant of $\mathcal{D}$, the second one is an interpolation error, and the last one measures the conformity defect of the method (how well a discrete divergence formula holds).

$$C_{\mathcal{D}} := \max_{v\in X_{\mathcal{D},0}\setminus\{0\}} \frac{\|\Pi_{\mathcal{D}} v\|_{L^2}}{\|v\|_{\mathcal{D}}}, \tag{2.4}$$

$$\forall \varphi \in H^1_0(\Omega)\,, \quad S_{\mathcal{D}}(\varphi) = \min_{v\in X_{\mathcal{D},0}} \left(\|\nabla_{\mathcal{D}} v - \nabla\varphi\|_{L^2} + \|\Pi_{\mathcal{D}} v - \varphi\|_{L^2}\right), \tag{2.5}$$

$$\forall \boldsymbol{\psi} \in H_{\mathrm{div}}(\Omega)\,,\ W_{\mathcal{D}}(\boldsymbol{\psi}) := \max_{v\in X_{\mathcal{D},0}\setminus\{0\}} \frac{1}{\|v\|_{\mathcal{D}}} \left| \int_\Omega \nabla_{\mathcal{D}} v \cdot \boldsymbol{\psi} + \Pi_{\mathcal{D}} v \mathrm{div}\boldsymbol{\psi} \right|. \tag{2.6}$$

In the following, unless otherwise specified, the notation $a \lesssim b$ means that $a \le Cb$ with $C > 0$ depending only on the data in assumption (1.3) and on an upper bound of $C_{\mathcal{D}}$.

Given a GD $\mathcal{D} = (X_{\mathcal{D},0}, \Pi_{\mathcal{D}}, \nabla_{\mathcal{D}}, Q_{\mathcal{D}})$ with piecewise constant reconstruction as in Definition 2.3, the GS for (1.1) is (compare with the weak formulation (1.2))

(2.7)

Find $u \in X_{\mathcal{D},0}$ such that

$$\int_\Omega \beta(\Pi_{\mathcal{D}} u)\Pi_{\mathcal{D}} v + \int_\Omega \Lambda \nabla_{\mathcal{D}} \zeta(u) \cdot \nabla_{\mathcal{D}} v = \int_\Omega Q_{\mathcal{D}} f\, \Pi_{\mathcal{D}} v - \int_\Omega F \cdot \nabla_{\mathcal{D}} v \quad \forall v \in X_{\mathcal{D},0}.$$

Remark 2.5 (quadrature for $\mathrm{div}(F)$). A quadrature operator could also be introduced for F (for example, considering $Q_{\mathcal{D}}$ componentwise, or selecting a different quadrature operator more appropriate to the structure of the gradient reconstruction). For simplicity of the presentation we decided not to include it in the analysis.

2.1. Convergence analysis. We first prove an a priori estimate on the solution to the GS. This estimate is used to prove the existence of this solution, its convergence, and the error estimate (2.12).

LEMMA 2.6 (bounds on the solution to the GS). *Let $\mathcal{D}$ be a GD with piecewise constant reconstruction as in Definition* 2.3*, and let $u \in X_{\mathcal{D},0}$ be a solution to the GS* (2.7)*. Then*

$$\|\Pi_{\mathcal{D}} u\|_{L^2} + \|\Pi_{\mathcal{D}} \beta(u)\|_{L^2} + \|\zeta(u)\|_{\mathcal{D}} \lesssim \|Q_{\mathcal{D}} f\|_{L^2} + \|F\|_{L^2} + 1. \tag{2.8}$$

Proof. Letting $v = \zeta(u)$ in (2.7) we get

$$\int_\Omega \beta(\Pi_{\mathcal{D}} u)\zeta(\Pi_{\mathcal{D}} u) + \int_\Omega \Lambda \nabla_{\mathcal{D}} \zeta(u) \cdot \nabla_{\mathcal{D}} \zeta(u) = \int_\Omega Q_{\mathcal{D}} f \, \Pi_{\mathcal{D}} \zeta(u) - \int_\Omega F \cdot \nabla_{\mathcal{D}} \zeta(u),$$

where we have used (2.3) to write $\Pi_{\mathcal{D}} \zeta(u) = \zeta(\Pi_{\mathcal{D}} u)$ in the first integral term. By monotonicity of β, ζ and $\beta(0) = \zeta(0) = 0$, we have $\beta(s)\zeta(s) \geq 0$ and the equation above thus gives, by definition of $C_{\mathcal{D}}$ and assumption (1.3e),

$$\begin{aligned} \underline{\lambda} \|\nabla_{\mathcal{D}} \zeta(u)\|_{L^2}^2 &\leq \|Q_{\mathcal{D}} f\|_{L^2} \|\Pi_{\mathcal{D}} \zeta(u)\|_{L^2} + \|F\|_{L^2} \|\nabla_{\mathcal{D}} \zeta(u)\|_{L^2} \\ &\leq \left(C_{\mathcal{D}} \|Q_{\mathcal{D}} f\|_{L^2} + \|F\|_{L^2} \right) \|\nabla_{\mathcal{D}} \zeta(u)\|_{L^2}. \end{aligned}$$

Recalling that $\|\zeta(u)\|_{\mathcal{D}} = \|\nabla_{\mathcal{D}} \zeta(u)\|_{L^2}$, this estimate yields the bound on $\zeta(u)$ in (2.8). Using again the definition of $C_{\mathcal{D}}$, we infer that $\|\Pi_{\mathcal{D}} \zeta(u)\|_{L^2} \lesssim \|Q_{\mathcal{D}} f\|_{L^2} + \|F\|_{L^2}$. By (2.3) this gives an $L^2(\Omega)$-estimate on $\zeta(\Pi_{\mathcal{D}} u)$ and, using assumption (1.3b), translates into the bound on $\Pi_{\mathcal{D}} u$ in (2.8). The estimate on $\beta(\Pi_{\mathcal{D}} u)$ follows from the sublinearity of β stated in assumption (1.3c). ∎

LEMMA 2.7 (existence and uniqueness for the GS). *Assume* (1.3) *and let $\mathcal{D}$ be a GD with a piecewise constant reconstruction as in Definition* 2.3*. Then there exists a solution to the GS* (2.7) *and, if (u_1, u_2) are two solutions to this scheme, then $\zeta(u_1) = \zeta(u_2)$ and $\Pi_{\mathcal{D}} u_1 = \Pi_{\mathcal{D}} u_2$.*

Remark 2.8 (counterexample to $u_1 = u_2$). In general, we cannot claim that $u_1 = u_2$, as the following counterexample shows. Consider $\beta(s) = s$ and $\zeta : \mathbb{R} \to \mathbb{R}$ such that $\zeta(s) = 0$ for all $s \in [0, 1]$. Take $F = 0$ and $f \in L^2(\Omega)$ such that $0 \leq f \leq 1$ almost everywhere, and consider an HMM GS [14, Chapter 13] on a polytopal mesh of Ω (which assumes that Ω is polytopal). Denoting by $\mathcal{M}$ and $\mathcal{F}$, respectively, the sets of cells and faces of this mesh, the corresponding GD satisfies

$$X_{\mathcal{D},0} = \{v = ((v_K)_{K\in\mathcal{M}}, (v_\sigma)_{\sigma\in\mathcal{F}}) \,:\, v_K \in \mathbb{R}\,,\ v_\sigma \in \mathbb{R}\,,\ v_\sigma = 0 \text{ if } \sigma \subset \partial\Omega\}$$

and $(\Pi_{\mathcal{D}} v)_{|K} = v_K$ for all $K \in \mathcal{M}$. We select $Q_{\mathcal{D}} = \mathrm{Id}$, and the precise expression of $\nabla_{\mathcal{D}} v$ is irrelevant to our counterexample. Then any $u = ((u_K)_{K\in\mathcal{M}}, (u_\sigma)_{\sigma\in\mathcal{F}}) \in X_{\mathcal{D},0}$ that satisfies

$$u_K = \frac{1}{|K|} \int_K f \quad \forall K \in \mathcal{M}\,, \quad u_\sigma \in [0,1] \quad \forall \sigma \in \mathcal{F}\,, \quad u_\sigma = 0 \text{ if } \sigma \subset \partial\Omega \tag{2.9}$$

is a solution to the GS (2.7). Indeed, all the components of such a vector belong to $[0, 1]$ and thus $\zeta(u) = 0$. The scheme equation on u thus reduces to

$$\int_\Omega \Pi_{\mathcal{D}} u \Pi_{\mathcal{D}} v = \int_\Omega f \Pi_{\mathcal{D}} v \quad \forall v \in X_{\mathcal{D},0},$$

which holds given the choice of the cell values $(u_K)_{K\in\mathcal{M}}$. Since there is an infinite number of u satisfying (2.9) (as the values on internal faces are free in $[0,1]$), this establishes that, when considering the HMM scheme, uniqueness fails for (2.7) with these f, β, ζ.

Proof of Lemma 2.7. The existence is obtained via a topological degree argument; we refer the reader to [11] for the definition and properties of this degree. Fix an arbitrary Euclidean structure, with inner product $\langle\cdot,\cdot\rangle$, on the finite dimensional space $X_{\mathcal{D},0}$. For $a\in[0,1]$ let $\zeta_a(s)=a\zeta(u)+(1-a)u$. Define $\mathfrak{F}:[0,1]\times X_{\mathcal{D},0}\to X_{\mathcal{D},0}$ the following way: for $a\in[0,1]$ and $u\in X_{\mathcal{D},0}$, $\mathfrak{F}(a,u)$ is the unique element of $X_{\mathcal{D},0}$ such that, for all $v\in X_{\mathcal{D},0}$,

$$\langle\mathfrak{F}(a,u),v\rangle=\int_\Omega a\beta(\Pi_{\mathcal{D}}u)\Pi_{\mathcal{D}}v+\int_\Omega\Lambda\nabla_{\mathcal{D}}\zeta_a(u)\cdot\nabla_{\mathcal{D}}v-\int_\Omega aQ_{\mathcal{D}}f\,\Pi_{\mathcal{D}}v-\int_\Omega aF\cdot\nabla_{\mathcal{D}}v.$$

We note that u is a solution to the GS (2.7) if and only if $\mathfrak{F}(1,u)=0$.

By continuity of β and ζ, and the finite dimension of $X_{\mathcal{D},0}$, the mapping $\mathfrak{F}$ is clearly continuous. Assume that $\mathfrak{F}(a,u)=0$ for some $a\in[0,1]$. The arguments in the proof of Lemma 2.6, using $v=\zeta_a(u)$ as a test function, show that $\|\zeta_a(u)\|_{\mathcal{D}}\le C_1$ with C_1 not depending on a; by equivalence of norms on the finite dimensional space $X_{\mathcal{D},0}$, this shows that $\|\zeta_a(u)\|_\infty\le C_2$ with C_2 still independent on a and $\|\cdot\|_\infty$ the supremum norm in $X_{\mathcal{D},0}$ on an arbitrary basis. The mapping ζ_a satisfies (1.3b) with M_0,M_1 independent of a. As a consequence, the bound on $\|\zeta_a(u)\|_\infty$ shows that $\|u\|_\infty<R$ with R independent of a.

Hence, any solution to $\mathfrak{F}(a,u)=0$ lies in the open ball B_R of $X_{\mathcal{D},0}$, centered at 0 and of radius R in the norm $\|\cdot\|_\infty$. This ball being independent of a, the topological degree theory ensures that $\deg(\mathfrak{F}(1,\cdot),B_R,0)=\deg(\mathfrak{F}(0,\cdot),B_R,0)$. The mapping $\mathfrak{F}(0,\cdot):X_{\mathcal{D},0}\to X_{\mathcal{D},0}$ is linear and the estimate obtained on the solutions to $\mathfrak{F}(0,u)=0$ shows that $\mathfrak{F}(0,\cdot)$ has a trivial kernel and is therefore invertible. This implies $\deg(\mathfrak{F}(0,\cdot),B_R,0)\neq0$ and thus $\deg(\mathfrak{F}(1,\cdot),B_R,0)\neq0$, which proves that the equation $\mathfrak{F}(1,u)=0$ has a solution $u\in B_R$.

We now consider the uniqueness of the solution to the scheme. Subtracting the equations satisfied by u_1 and u_2 and taking $v=\zeta(u_1)-\zeta(u_2)\in X_{\mathcal{D},0}$ as a test function, we have

$$\begin{aligned}&\int_\Omega(\beta(\Pi_{\mathcal{D}}u_1)-\beta(\Pi_{\mathcal{D}}u_2))\Pi_{\mathcal{D}}(\zeta(u_1)-\zeta(u_2))\\&\qquad+\int_\Omega\Lambda\nabla_{\mathcal{D}}(\zeta(u_1)-\zeta(u_2))\cdot\nabla_{\mathcal{D}}(\zeta(u_1)-\zeta(u_2))=0.\end{aligned}$$

Property (2.3) and the monotonicity of β and ζ show that

$$\begin{aligned}&(\beta(\Pi_{\mathcal{D}}u_1)-\beta(\Pi_{\mathcal{D}}u_2))\Pi_{\mathcal{D}}(\zeta(u_1)-\zeta(u_2))\\&\qquad=(\beta(\Pi_{\mathcal{D}}u_1)-\beta(\Pi_{\mathcal{D}}u_2))(\zeta(\Pi_{\mathcal{D}}u_1)-\zeta(\Pi_{\mathcal{D}}u_2))\ge0.\end{aligned}$$

Hence, $\|\nabla_{\mathcal{D}}(\zeta(u_1)-\zeta(u_2))\|_{L^2}=0$ which, by property of $\nabla_{\mathcal{D}}$, ensures that $\zeta(u_1)=\zeta(u_2)$.

We now come back to the equations satisfied by u_1 and u_2, subtract them, and take $v=\beta(u_1)-\beta(u_2)\in X_{\mathcal{D},0}$ as a test function to get

$$\int_\Omega(\beta(\Pi_{\mathcal{D}}u_1)-\beta(\Pi_{\mathcal{D}}u_2))^2+\int_\Omega\nabla_{\mathcal{D}}(\zeta(u_1)-\zeta(u_2))\cdot\nabla_{\mathcal{D}}(\beta(u_1)-\beta(u_2))=0.$$

Since $\zeta(u_1) = \zeta(u_2)$, we infer that $\beta(\Pi_{\mathcal{D}} u_1) - \beta(\Pi_{\mathcal{D}} u_2) = 0$. Owing to hypothesis (1.3d), we conclude that $\Pi_{\mathcal{D}} u_1 = \Pi_{\mathcal{D}} u_2$ from $\beta(\Pi_{\mathcal{D}} u_1) + \zeta(\Pi_{\mathcal{D}} u_1) = \beta(\Pi_{\mathcal{D}} u_2) + \zeta(\Pi_{\mathcal{D}} u_2)$. ∎

The next theorem is our first main convergence result. It states the strong convergence of the solution to the GS without assuming any regularity property on the continuous solution.

THEOREM 2.9 (convergence of the scheme). *Assume* (1.3) *and let* $(\mathcal{D}_m)_{m\in\mathbb{N}}$ *be a sequence of GDs with piecewise constant reconstructions as in Definition* 2.3. *Assume moreover that the following properties hold:*

- (Coercivity) *The sequence* $(C_{\mathcal{D}_m})_{m\in\mathbb{N}}$ *is bounded, where* $C_{\mathcal{D}_m}$ *is defined by* (2.4) *for* $\mathcal{D} = \mathcal{D}_m$.
- (Consistency) *Recalling the definition* (2.5), *there holds*

$$\forall \varphi \in H_0^1(\Omega),\ \lim_{m\to\infty} S_{\mathcal{D}_m}(\varphi) = 0 \quad \text{and} \quad \lim_{m\to\infty} \|Q_{\mathcal{D}_m} f - f\|_{L^2} = 0. \tag{2.10}$$

- (Limit-conformity) *Recalling the definition* (2.6), *there holds*

$$\forall \boldsymbol{\psi} \in H_{\mathrm{div}}(\Omega)\,,\ \lim_{m\to\infty} W_{\mathcal{D}_m}(\boldsymbol{\psi}) = 0. \tag{2.11}$$

- (Compactness) *For any* $(v_m)_{m\in\mathbb{N}}$ *such that* $v_m \in X_{\mathcal{D}_m,0}$ *for all* $m \in \mathbb{N}$ *and* $(\nabla_{\mathcal{D}_m} v_m)_{m\in\mathbb{N}}$ *is bounded in* $L^2(\Omega)^d$, *the set* $\{\Pi_{\mathcal{D}_m} v_m : m \in \mathbb{N}\}$ *is relatively compact in* $L^2(\Omega)$.

For any $m \in \mathbb{N}$, *let* u_m *be a solution of scheme* (2.7). *Then there exists* $\overline{u}$ *solution to* (1.2) *such that, as* $m \to \infty$, $\Pi_{\mathcal{D}_m}\zeta(u_m) \to \zeta(\overline{u})$ *strongly in* $L^2(\Omega)$, $\nabla_{\mathcal{D}_m}\zeta(u_m) \to \nabla\zeta(\overline{u})$ *strongly in* $L^2(\Omega)^d$, *and* $\Pi_{\mathcal{D}_m}\beta(u_m) \to \beta(\overline{u})$ *weakly in* $L^2(\Omega)$.

Proof. Using estimate (2.8) and (2.10) (which shows that $\|Q_{\mathcal{D}_m} f\|_{L^2}$ is bounded), the compactness, and the limit-conformity of $(\mathcal{D}_m)_{m\in\mathbb{N}}$, [14, Lemma 2.15] gives $Z \in H_0^1(\Omega)$ and $B \in L^2(\Omega)$ such that, up to a subsequence (not made explicit in the following), $\Pi_{\mathcal{D}_m}\zeta(u_m) \to Z$ strongly in $L^2(\Omega)$, $\nabla_{\mathcal{D}_m}\zeta(u_m) \to \nabla Z$ weakly in $L^2(\Omega)^d$, and $\beta(\Pi_{\mathcal{D}_m} u_m) \to B$ weakly in $L^2(\Omega)$. By weak/strong convergence we infer that

$$\lim_{m\to\infty} \int_\Omega \Pi_{\mathcal{D}_m}\beta(u_m)\Pi_{\mathcal{D}_m}\zeta(u_m) = \int_\Omega BZ.$$

The monotonicity properties of β and ζ then enable us to apply [14, Lemma D.10] (a Minty's trick) to get $\overline{u} \in L^2(\Omega)$ such that $Z = \zeta(\overline{u})$ and $B = \beta(\overline{u})$.

We now show that $\overline{u}$ solves (1.2). Let $\varphi \in H_0^1(\Omega)$ and let $v_m \in X_{\mathcal{D}_m,0}$ be an element that realizes the minimum defining $S_{\mathcal{D}_m}(\varphi)$. By (2.10) we have $\Pi_{\mathcal{D}_m} v_m \to \varphi$ in $L^2(\Omega)$ and $\nabla_{\mathcal{D}_m} v_m \to \nabla\varphi$ in $L^2(\Omega)^d$. Use v_m as a test function in GS (2.7) satisfied by u_m. The convergence properties of $\beta(\Pi_{\mathcal{D}_m} u_m)$ and $\nabla_{\mathcal{D}_m}\zeta(u_m)$ toward $B = \beta(\overline{u})$ and $\nabla Z = \nabla\zeta(\overline{u})$, together with the convergence $Q_{\mathcal{D}_m} f \to f$ in $L^2(\Omega)$ stated in (2.10), enable us to take the limit $m \to \infty$ of the scheme to see that $\overline{u}$ is a solution to (1.2). The uniqueness of $\overline{u}$ (see Theorem A.1) shows that the convergence properties hold for the whole sequence $(u_m)_{m\in\mathbb{N}}$ instead of just along the subsequence previously extracted.

It remains to establish the strong convergence of $\nabla_{\mathcal{D}_m}\zeta(u_m)$. We let $m \to +\infty$ in the GS (2.7) with $\mathcal{D} = \mathcal{D}_m$ and $v = \zeta(u_m)$, that is,

$$\begin{aligned}\int_\Omega \beta(\Pi_{\mathcal{D}_m} u_m)\Pi_{\mathcal{D}_m}\zeta(u_m) + \int_\Omega \Lambda\nabla_{\mathcal{D}_m}\zeta(u_m)\cdot\nabla_{\mathcal{D}_m}\zeta(u_m)\\ = \int_\Omega Q_{\mathcal{D}_m} f\,\Pi_{\mathcal{D}_m}\zeta(u_m) - \int_\Omega F\cdot\nabla_{\mathcal{D}}\zeta(u_m).\end{aligned}$$

This yields

$$\lim_{m\to\infty}\int_\Omega \Lambda\nabla_{\mathcal{D}_m}\zeta(u_m)\cdot\nabla_{\mathcal{D}_m}\zeta(u_m) = \int_\Omega (f\zeta(\overline{u}) - F\cdot\nabla\zeta(\overline{u}) - \beta(\overline{u})\zeta(\overline{u})) = \int_\Omega \Lambda\nabla\zeta(\overline{u})\cdot\nabla\zeta(\overline{u}),$$

where the conclusion follows using $\overline{v}=\overline{u}$ in (1.2). Since $(\xi,\eta)\mapsto\int_\Omega \Lambda\xi\cdot\eta$ is an inner product on $L^2(\Omega)$, this relation and the weak convergence of $(\nabla_{\mathcal{D}_m}\zeta(u_m))_{m\in\mathbb{N}}$ imply the strong convergence of $\nabla_{\mathcal{D}_m}\zeta(u_m)$ to $\nabla\zeta(\overline{u})$ in L^2. ∎

2.2. Error estimate. The analysis above pinpoints the required structure on a numerical scheme to ensure proper bounds and the convergence of the solution—namely, the piecewise constant reconstruction property. We now want to establish error estimates to better assess this convergence. In practice, one usually starts from a given numerical method and would like to apply it to the model under consideration. Following our discussion above, if the given method does not have a piecewise constant function reconstruction, it has to be modified into a method that has such a reconstruction. This process is called the mass-lumping of the original scheme. In the context of the GDM, this notion is translated in the following definition.

DEFINITION 2.10 (mass-lumped GD). *Let $\mathcal{D}_* = (X_{\mathcal{D},0}, \Pi_{\mathcal{D}_*}, \nabla_{\mathcal{D}}, Q_{\mathcal{D}})$ be a GD. A GD $\mathcal{D}$ is a mass-lumped version of $\mathcal{D}_*$ if it only differs from $\mathcal{D}_*$ through the function reconstruction (that is, $\mathcal{D} = (X_{\mathcal{D},0}, \Pi_{\mathcal{D}}, \nabla_{\mathcal{D}}, Q_{\mathcal{D}})$) and if $\Pi_{\mathcal{D}}$ is a piecewise constant reconstruction in the sense of Definition* 2.3.

Remark 2.11. Of course, if $\mathcal{D}_*$ already has a piecewise constant reconstruction as in Definition 2.3, one can take $\mathcal{D}=\mathcal{D}_*$.

The following theorem states a general error estimate on the GS (2.7).

THEOREM 2.12 (error estimate for the GS). *Assume* (1.3) *and let $\mathcal{D}$ be a mass-lumped version of a GD $\mathcal{D}_*$, in the sense of Definition* 2.10. *Let u be a solution to the GS* (2.7), *and let $\overline{u}$ be the solution to* (1.2) *(see Theorem* A.1*). Then, for any $\mathcal{I}_{\mathcal{D}}\zeta(\overline{u})\in X_{\mathcal{D},0}$, there holds*

$$\begin{aligned}(2.12)\quad &\|\nabla_{\mathcal{D}}[\mathcal{I}_{\mathcal{D}}\zeta(\overline{u}) - \zeta(u)]\|_{L^2}\\ &\lesssim W_{\mathcal{D}_*}(\Lambda\nabla\zeta(\overline{u}) + F) + \|\nabla_{\mathcal{D}}\mathcal{I}_{\mathcal{D}}\zeta(\overline{u}) - \nabla\zeta(\overline{u})\|_{L^2} + R_{\mathcal{D},\mathcal{D}_*}(\overline{u},f) + \mathfrak{T}_{\mathcal{D}}(\overline{u},u),\end{aligned}$$

where

(2.13)
$$R_{\mathcal{D},\mathcal{D}_*}(\overline{u},f) = \max_{v\in X_{\mathcal{D},0}\setminus\{0\}} \frac{1}{\|\nabla_{\mathcal{D}}v\|_{L^2}}\left|\int_\Omega \Pi_{\mathcal{D}}v[\beta(Q_{\mathcal{D}}\overline{u}) - Q_{\mathcal{D}}f] - \Pi_{\mathcal{D}_*}v[\beta(\overline{u}) - f]\right|,$$

and

(2.14)
$$\mathfrak{T}_{\mathcal{D}}(\overline{u},u) = \left(\max\left\{\int_\Omega \left[\beta(Q_{\mathcal{D}}\overline{u}) - \beta(\Pi_{\mathcal{D}}u)\right]\left[\zeta(Q_{\mathcal{D}}\overline{u}) - \Pi_{\mathcal{D}}\mathcal{I}_{\mathcal{D}}\zeta(\overline{u})\right]; 0\right\}\right)^{1/2}.$$

Remark 2.13 (choice of $\mathcal{I}_{\mathcal{D}}\zeta(\overline{u})$). The element $\mathcal{I}_{\mathcal{D}}\zeta(\overline{u})$ can be any vector in $X_{\mathcal{D},0}$. However, the estimate (2.12) is obviously useful only if $\nabla_{\mathcal{D}}\mathcal{I}_{\mathcal{D}}\zeta(\overline{u})$ is close to $\nabla\zeta(\overline{u})$. This is usually achieved selecting for $\mathcal{I}_{\mathcal{D}}\zeta(\overline{u})$ a suitable interpolate of $\zeta(\overline{u})$, which is why we used this notation.

Remark 2.14 (approximation of $\zeta(\overline{u})$). Introducing $\pm\nabla_{\mathcal{D}}(\mathcal{I}_{\mathcal{D}}\zeta(\overline{u}))$ and using a triangle inequality, we have

$$\|\nabla_{\mathcal{D}}\zeta(u)-\nabla\zeta(\overline{u})\|_{L^2}\le\|\nabla_{\mathcal{D}}[\zeta(u)-\mathcal{I}_{\mathcal{D}}\zeta(\overline{u})]\|_{L^2}+\|\nabla_{\mathcal{D}}\mathcal{I}_{\mathcal{D}}\zeta(\overline{u})-\nabla\zeta(\overline{u})\|_{L^2}.$$

Similarly, introducing $\pm\Pi_{\mathcal{D}_*}\mathcal{I}_{\mathcal{D}}\zeta(\overline{u})$ and using the triangle inequality and the definition of $C_{\mathcal{D}_*}$, we have

$$\begin{aligned}\|\Pi_{\mathcal{D}_*}\zeta(u)-\zeta(\overline{u})\|_{L^2}&\le\|\Pi_{\mathcal{D}_*}[\zeta(u)-\mathcal{I}_{\mathcal{D}}\zeta(\overline{u})]\|_{L^2}+\|\Pi_{\mathcal{D}_*}\mathcal{I}_{\mathcal{D}}\zeta(\overline{u})-\zeta(\overline{u})\|_{L^2}\\&\le C_{\mathcal{D}_*}\|\nabla_{\mathcal{D}}[\zeta(u)-\mathcal{I}_{\mathcal{D}}\zeta(\overline{u})]\|_{L^2}+\|\Pi_{\mathcal{D}_*}\mathcal{I}_{\mathcal{D}}\zeta(\overline{u})-\zeta(\overline{u})\|_{L^2}.\end{aligned}$$

An estimate on $\nabla_{\mathcal{D}}[\zeta(u)-\mathcal{I}_{\mathcal{D}}\zeta(\overline{u})]$ as in Theorem 2.12 therefore also yields an estimate on $\nabla_{\mathcal{D}}\zeta(u)-\nabla\zeta(\overline{u})$ and $\Pi_{\mathcal{D}_*}\zeta(u)-\zeta(\overline{u})$, modulo the additional interpolation errors $\nabla_{\mathcal{D}}\mathcal{I}_{\mathcal{D}}\zeta(\overline{u})-\nabla\zeta(\overline{u})$ and $\Pi_{\mathcal{D}_*}\mathcal{I}_{\mathcal{D}}\zeta(\overline{u})-\zeta(\overline{u})$. If $\mathcal{D}_*$ has function and gradient reconstructions that are piecewise polynomial of high-order, these interpolation errors can be expected to have a high rate of convergence with respect to the mesh size.

The same argument also gives an error estimate on $\Pi_{\mathcal{D}}\zeta(u)-\zeta(\overline{u})$, but the corresponding interpolation error $\Pi_{\mathcal{D}}\mathcal{I}_{\mathcal{D}}\zeta(\overline{u})-\zeta(\overline{u})$ is limited to a first-order convergence since $\Pi_{\mathcal{D}}$ is a piecewise constant reconstruction.

Proof of Theorem 2.12. Since $\overline{u}$ is the solution to (1.2), we have $\mathrm{div}(\Lambda\nabla\zeta(\overline{u})+F)=\beta(\overline{u})-f\in L^2(\Omega)$. Hence, by definition (2.6) of $W_{\mathcal{D}}$ applied to $\mathcal{D}_*$, for any $v\in X_{\mathcal{D},0}$,

$$\begin{aligned}\|v\|_{\mathcal{D}}W_{\mathcal{D}_*}(\Lambda\nabla\zeta(\overline{u})+F)&\ge\int_\Omega\nabla_{\mathcal{D}}v\cdot(\Lambda\nabla\zeta(\overline{u})+F)+\Pi_{\mathcal{D}_*}v\mathrm{div}(\Lambda\nabla\zeta(\overline{u})+F)\\&=\int_\Omega\nabla_{\mathcal{D}}v\cdot\Lambda\nabla\zeta(\overline{u})+F\cdot\nabla_{\mathcal{D}}v+\Pi_{\mathcal{D}_*}v[\beta(\overline{u})-f].\end{aligned}$$

Substituting the term involving F using (2.7), we get

$$\begin{aligned}(2.15)\quad&\int_\Omega\Lambda\nabla_{\mathcal{D}}v\cdot(\nabla\zeta(\overline{u})-\nabla_{\mathcal{D}}\zeta(u))+(Q_{\mathcal{D}}f-\beta(\Pi_{\mathcal{D}}u))\Pi_{\mathcal{D}}v+\Pi_{\mathcal{D}_*}v[\beta(\overline{u})-f]\\&\qquad\le\|v\|_{\mathcal{D}}W_{\mathcal{D}_*}(\Lambda\nabla\zeta(\overline{u})+F).\end{aligned}$$

Introducing $\pm\Lambda\nabla_{\mathcal{D}}v\cdot\nabla_{\mathcal{D}}\mathcal{I}_{\mathcal{D}}\zeta(\overline{u})$ and $\pm\beta(Q_{\mathcal{D}}\overline{u})\Pi_{\mathcal{D}}v$ in the left-hand side, using the Cauchy–Schwarz inequality, and recalling that $\|v\|_{\mathcal{D}}=\|\nabla_{\mathcal{D}}v\|_{L^2}$, we infer

$$\begin{aligned}(2.16)\quad&\int_\Omega\Lambda\nabla_{\mathcal{D}}v\cdot(\nabla_{\mathcal{D}}\mathcal{I}_{\mathcal{D}}\zeta(\overline{u})-\nabla_{\mathcal{D}}\zeta(u))+(\beta(Q_{\mathcal{D}}\overline{u})-\beta(\Pi_{\mathcal{D}}u))\Pi_{\mathcal{D}}v\\&\qquad+\int_\Omega\Pi_{\mathcal{D}_*}v[\beta(\overline{u})-f]-\Pi_{\mathcal{D}}v[\beta(Q_{\mathcal{D}}\overline{u})-Q_{\mathcal{D}}f]\\&\qquad\lesssim\|\nabla_{\mathcal{D}}v\|_{L^2}\left[W_{\mathcal{D}_*}(\Lambda\nabla\zeta(\overline{u})+F)+\|\nabla_{\mathcal{D}}\mathcal{I}_{\mathcal{D}}\zeta(\overline{u})-\nabla\zeta(\overline{u})\|_{L^2}\right].\end{aligned}$$

Choose $v=\mathcal{I}_{\mathcal{D}}\zeta(\overline{u})-\zeta(u)$. Introducing $\pm\zeta(Q_{\mathcal{D}}\overline{u})$ and using the monotonicity of ζ and β (which yields $[\beta(b)-\beta(a)][\zeta(b)-\zeta(a)]\ge0$ for all $a,b\in\mathbb{R}$) together with (2.3), we have

$$\begin{aligned}(\beta(Q_{\mathcal{D}}\overline{u})-\beta(\Pi_{\mathcal{D}}u))\Pi_{\mathcal{D}}v&=[\beta(Q_{\mathcal{D}}\overline{u})-\beta(\Pi_{\mathcal{D}}u)]\,[\Pi_{\mathcal{D}}\mathcal{I}_{\mathcal{D}}\zeta(\overline{u})-\zeta(Q_{\mathcal{D}}\overline{u})]\\&\quad+[\beta(Q_{\mathcal{D}}\overline{u})-\beta(\Pi_{\mathcal{D}}u)]\,[\zeta(Q_{\mathcal{D}}\overline{u})-\zeta(\Pi_{\mathcal{D}}u)]\\&\ge[\beta(Q_{\mathcal{D}}\overline{u})-\beta(\Pi_{\mathcal{D}}u)]\,[\Pi_{\mathcal{D}}\mathcal{I}_{\mathcal{D}}\zeta(\overline{u})-\zeta(Q_{\mathcal{D}}\overline{u})].\end{aligned}$$

Plugging this into (2.16) and using (1.3e) leads to

$$\begin{aligned}\|\nabla_{\mathcal{D}}[\mathcal{I}_{\mathcal{D}}\zeta(\overline{u})-\zeta(u)]\|_{L^2}^2 &\lesssim \|\nabla_{\mathcal{D}}v\|_{L^2}\left[W_{\mathcal{D}_*}(\Lambda\nabla\zeta(\overline{u})+F)+\|\nabla_{\mathcal{D}}\mathcal{I}_{\mathcal{D}}\zeta(\overline{u})-\nabla\zeta(\overline{u})\|_{L^2}\right]\\ &\quad+\int_\Omega \Pi_{\mathcal{D}}v[\beta(Q_{\mathcal{D}}\overline{u})-Q_{\mathcal{D}}f]-\Pi_{\mathcal{D}_*}v[\beta(\overline{u})-f]\\ &\quad+\int_\Omega [\beta(Q_{\mathcal{D}}\overline{u})-\beta(\Pi_{\mathcal{D}}u)]\,[\zeta(Q_{\mathcal{D}}\overline{u})-\Pi_{\mathcal{D}}\mathcal{I}_{\mathcal{D}}\zeta(\overline{u})]\\ &\lesssim \|\nabla_{\mathcal{D}}v\|_{L^2}\left[W_{\mathcal{D}_*}(\Lambda\nabla\zeta(\overline{u})+F)+\|\nabla_{\mathcal{D}}\mathcal{I}_{\mathcal{D}}\zeta(\overline{u})-\nabla\zeta(\overline{u})\|_{L^2}+R_{\mathcal{D},\mathcal{D}_*}(\overline{u},f)\right]\\ &\quad+\mathfrak{T}_{\mathcal{D}}(\overline{u},u)^2.\end{aligned}$$

Using the Young inequality on the first term in the right-hand side and recalling that $v=\mathcal{I}_{\mathcal{D}}\zeta(\overline{u})-\zeta(u)$ leads to

$$\begin{aligned}&\|\nabla_{\mathcal{D}}[\mathcal{I}_{\mathcal{D}}\zeta(\overline{u})-\zeta(u)]\|_{L^2}^2\\ &\lesssim \left[W_{\mathcal{D}_*}(\Lambda\nabla\zeta(\overline{u})+F)+\|\nabla_{\mathcal{D}}\mathcal{I}_{\mathcal{D}}\zeta(\overline{u})-\nabla\zeta(\overline{u})\|_{L^2}+R_{\mathcal{D},\mathcal{D}_*}(\overline{u},f)\right]^2+\mathfrak{T}_{\mathcal{D}}(\overline{u},u)^2.\end{aligned}$$

The proof of (2.12) is complete taking the square root of this estimate and using $\sqrt{a^2+b^2}\le a+b$ for all $a,b\ge 0$. ∎

From the general estimate (2.12) we deduce the following bound on the error, which often leads to (low-order) rates of convergence as noted in Remark 2.16. This estimate will be improved, for situations corresponding to classical mass-lumping versions of schemes with nodal interpolates, in section 2.3.

COROLLARY 2.15. *Under the assumptions of Theorem* 2.12*, define*

$$\alpha_{\mathcal{D},\mathcal{D}_*}=\max_{v\in X_{\mathcal{D},0}\setminus\{0\}}\frac{\|\Pi_{\mathcal{D}}v-\Pi_{\mathcal{D}_*}v\|_{L^2}}{\|\nabla_{\mathcal{D}}v\|_{L^2}}$$

and let $\mathcal{I}_{\mathcal{D}}\zeta(\overline{u})$ *be given by* $\mathcal{I}_{\mathcal{D}}\zeta(\overline{u})=\operatorname{argmin}_{v\in X_{\mathcal{D},0}}(\|\nabla_{\mathcal{D}}v-\nabla\zeta(\overline{u})\|_{L^2}+\|\Pi_{\mathcal{D}}v-\zeta(\overline{u})\|_{L^2})$*. Then,*

$$\begin{aligned}\|\nabla_{\mathcal{D}}[\mathcal{I}_{\mathcal{D}}\zeta(\overline{u})-\zeta(u)]\|_{L^2} &\lesssim W_{\mathcal{D}_*}(\Lambda\nabla\zeta(\overline{u})+F)+S_{\mathcal{D}}(\zeta(\overline{u}))\\ &\quad+\alpha_{\mathcal{D},\mathcal{D}_*}+\|\beta(Q_{\mathcal{D}}\overline{u})-\beta(\overline{u})\|_{L^2}+\|Q_{\mathcal{D}}f-f\|_{L^2}\\ &\quad+\left(S_{\mathcal{D}}(\zeta(\overline{u}))+\|\zeta(\overline{u})-\zeta(Q_{\mathcal{D}}\overline{u})\|_{L^2}\right)^{\frac12},\end{aligned}\tag{2.17}$$

where the hidden multiplicative constant in $\lesssim$ *additionally depends on* $\|\beta(Q_{\mathcal{D}}\overline{u})\|_{L^2}$ *and* $\|Q_{\mathcal{D}}f\|_{L^2}$*.*

Remark 2.16 (rate of convergence). For all classical mass-lumping of schemes based on a mesh of size h, we have $\alpha_{\mathcal{D},\mathcal{D}_*}=\mathcal{O}(h)$ (see, e.g., [14, equations (8.18) and (9.46)]). Likewise, any reasonable quadrature rule is locally exact on piecewise constant functions and thus, if β,ζ are globally Lipschitz continuous and $\overline{u},f$ are locally H^1, we expect $\mathcal{O}(h)$ estimates on $\|\beta(Q_{\mathcal{D}}\overline{u})-\beta(\overline{u})\|_{L^2}$, $\|Q_{\mathcal{D}}f-f\|_{L^2}$, and $\|\zeta(\overline{u})-\zeta(Q_{\mathcal{D}}\overline{u})\|_{L^2}$. The estimate (2.17) can thus be expected, most of the time, to provide an $\mathcal{O}(h^{\frac12})$ rate of convergence, the limiting factor in the right-hand side of (2.17) being the last one, coming from $\mathfrak{T}_{\mathcal{D}}(\overline{u},u)$. We will see in section 2.3 that this estimate is, however, very pessimistic and, in many cases, can be improved to higher powers of h (see Remark 2.27).

Proof. We estimate each term, except the first one, in the right-hand side of (2.12). By choice of $\mathcal{I}_{\mathcal{D}}\zeta(\overline{u})$ and definition (2.5) of $S_{\mathcal{D}}$,

$$\|\nabla_{\mathcal{D}}\mathcal{I}_{\mathcal{D}}\zeta(\overline{u}) - \nabla\zeta(\overline{u})\|_{L^2} + \|\Pi_{\mathcal{D}}\mathcal{I}_{\mathcal{D}}\zeta(\overline{u}) - \zeta(\overline{u})\|_{L^2} = S_{\mathcal{D}}(\zeta(\overline{u})).$$

Hence the term $\|\nabla_{\mathcal{D}}\mathcal{I}_{\mathcal{D}}\zeta(\overline{u}) - \nabla\zeta(\overline{u})\|_{L^2}$ in (2.12) is bounded above by $S_{\mathcal{D}}(\zeta(\overline{u}))$.

Using the definition of $\alpha_{\mathcal{D},\mathcal{D}_*}$ and of $C_{\mathcal{D}}$, we have

$$\begin{aligned}
R_{\mathcal{D},\mathcal{D}_*}(\overline{u}, f) &\le \alpha_{\mathcal{D},\mathcal{D}_*}\|\beta(\overline{u}) - f\|_{L^2} \\
&\quad + \max_{v\in X_{\mathcal{D},0}\setminus\{0\}} \frac{1}{\|\nabla_{\mathcal{D}} v\|_{L^2}} \left| \int_\Omega \Pi_{\mathcal{D}} v[\beta(Q_{\mathcal{D}}\overline{u}) - Q_{\mathcal{D}} f] - \Pi_{\mathcal{D}} v[\beta(\overline{u}) - f] \right| \\
&\lesssim \alpha_{\mathcal{D},\mathcal{D}_*} + C_{\mathcal{D}}\|[\beta(Q_{\mathcal{D}}\overline{u}) - Q_{\mathcal{D}} f] - [\beta(\overline{u}) - f]\|_{L^2} \\
&\lesssim \alpha_{\mathcal{D},\mathcal{D}_*} + \|\beta(Q_{\mathcal{D}}\overline{u}) - \beta(\overline{u})\|_{L^2} + \|Q_{\mathcal{D}} f - f\|_{L^2}.
\end{aligned}$$

This gives the third, fourth, and fifth terms in the right-hand side of (2.17).

For the last term in this estimate, we write, using the Cauchy–Schwarz inequality and the a priori bound (2.8) on $\beta(\Pi_{\mathcal{D}} u)$,

$$\begin{aligned}
\mathfrak{T}_{\mathcal{D}}(\overline{u}, u)^2 &\le \|\beta(Q_{\mathcal{D}}\overline{u}) - \beta(\Pi_{\mathcal{D}} u)\|_{L^2}\|\Pi_{\mathcal{D}}\mathcal{I}_{\mathcal{D}}\zeta(\overline{u}) - \zeta(Q_{\mathcal{D}}\overline{u})\|_{L^2} \\
&\lesssim (\|\beta(Q_{\mathcal{D}}\overline{u})\|_{L^2} + \|Q_{\mathcal{D}} f\|_{L^2} + \|F\|_{L^2}) \\
&\qquad \times \left(\|\Pi_{\mathcal{D}}\mathcal{I}_{\mathcal{D}}\zeta(\overline{u}) - \zeta(\overline{u})\|_{L^2} + \|\zeta(\overline{u}) - \zeta(Q_{\mathcal{D}}\overline{u})\|_{L^2}\right).
\end{aligned}$$

The proof is complete taking the square root and recalling that $\|\Pi_{\mathcal{D}}\mathcal{I}_{\mathcal{D}}\zeta(\overline{u}) - \zeta(\overline{u})\|_{L^2} \le S_{\mathcal{D}}(\zeta(\overline{u}))$. ∎

2.3. Suitable quadrature rules lead to high-order estimates. Let us first make the following broken regularity assumption on the data and solution.

ASSUMPTION 2.17 (data and exact solution). *$F = 0$ and, $\overline{u}$ being the solution to (1.2) and $s \ge 1$ being an integer, f and $\beta(\overline{u})$ belong to the broken Sobolev space*

$$W^{s,\infty}(\mathcal{M}) := \left\{ g \in L^\infty(\Omega) \,:\, g_{|K} \in W^{s,\infty}(K) \quad \forall K \in \mathcal{M} \right\}.$$

This space is endowed with the norm $\|g\|_{W^{s,\infty}(\mathcal{M})} := \max_{K\in\mathcal{M}} \|g\|_{W^{s,\infty}(K)}$.

Remark 2.18 (piecewise continuity and local smoothness). $W^{s,\infty}(\mathcal{M})$ is a subspace of

$$C(\mathcal{M}) := \{ g \in L^\infty(\Omega) \,:\, g_{|K} \in C(\overline{K}) \quad \forall K \in \mathcal{M} \}. \tag{2.18}$$

Assumption 2.17 only imposes a local smoothness of f and $\beta(\overline{u})$, which can in particular be discontinuous across cell interfaces. It is also worthwhile noticing that, since $F = 0$, $\zeta(\overline{u})$ is continuous (see Theorem A.1). Hence, the values of $\overline{u}$ at one of its discontinuities must belong to a plateau of ζ; in particular, if ζ does not have any plateau, then $\overline{u}$ is globally continuous.

Note that if $F \neq 0$ (and F is not smooth), as in Test Case S3 in section 3, $\zeta(\overline{u})$ is not expected to have any additional regularity beyond H^1 and therefore high-order estimates, even if they can theoretically be established, are of little use. Actually, numerical tests show that high-order schemes can deliver estimates that are no better than low-order schemes; see, e.g., Table 10.

Remark 2.19 ($\overline{u}$ is in $C(\mathcal{M})$). Theorem A.1 and Assumption 2.17 show that $\zeta(\overline{u}) \in C(\overline{\Omega})$ and $\beta(\overline{u}) \in C(\mathcal{M})$. By (1.3b)–(1.3d), $\beta + \zeta$ is a homeomorphism of $\mathbb{R}$. Hence, $\overline{u} = (\beta + \zeta)^{-1}(\beta(\overline{u}) + \zeta(\overline{u}))$ and thus $\overline{u} \in C(\mathcal{M})$.

In the rest of this section, we consider a slightly more precise setting than in section 2.2. We assume that $\mathcal{D}_*$ has a piecewise polynomial function reconstruction

(possibly of high-order) and unknowns associated to nodes in the domain, and that specific local quadrature rules can be chosen. Typically, $\mathbb{P}^k$ or $\mathbb{Q}^k$ finite elements and symmetric interior penalty discontinuous Galerkin (SIPG) schemes, with mass-lumping constructed using dual meshes around the nodes, fit into this setting. In what follows, h_X denotes the diameter of a set $X \subset \mathbb{R}^d$.

ASSUMPTION 2.20 (structure of $\mathcal{D}_*$, $\mathcal{D}$ and $\mathcal{I}_{\mathcal{D}}\zeta(\overline{u})$).

(1) (Mesh) $\Omega \subset \mathbb{R}^d$ *(with $d \leq 3$) is a polytopal open set and $\mathcal{M}$ is a polytopal mesh of Ω, in the sense of* [14, Definition 7.2] *(this definition actually represents the mesh as a quadruple of sets of cells, faces, points, and vertices that will not be useful to our purpose; we therefore confuse the mesh with the set of cells). The mesh size is $h = \max_{K\in\mathcal{M}} h_K$.*

(2) (Space) *There is a finite set I, partitioned into I_Ω and $I_{\partial\Omega}$, such that*

$$X_{\mathcal{D},0} = \{v = (v_i)_{i\in I} \,:\, v_i \in \mathbb{R} \quad \forall i \in I\,,\ v_i = 0 \quad \forall i \in I_{\partial\Omega}\}\,.$$

(3) (Local polynomial reconstructions) *There is a polynomial degree $k \geq 1$ such that, for all $K \in \mathcal{M}$ and all $v \in X_{\mathcal{D},0}$, $(\Pi_{\mathcal{D}_*} v)_{|K} \in \mathbb{P}^k$.*

(4) (Broken gradient bound) *There is $C_\nabla \geq 0$ such that, for all $v \in X_{\mathcal{D},0}$, $\|\nabla_h(\Pi_{\mathcal{D}_*} v)\|_{L^2} \leq C_\nabla \|\nabla_{\mathcal{D}} v\|_{L^2}$, where ∇_h is the usual broken gradient on $\mathcal{M}$.*

(5) (Nodes) *There is a family $(\boldsymbol{x}_i)_{i\in I}$ of points in $\overline{\Omega}$, and subsets $(I_K)_{K\in\mathcal{M}}$ of I, such that $I = (\cup_{K\in\mathcal{M}} I_K) \cup I_{\partial\Omega}$ and, for all $v = (v_i)_{i\in I} \in X_{\mathcal{D},0}$, all $K \in \mathcal{M}$, and all $i \in I_K$, we have $\boldsymbol{x}_i \in \overline{K}$ and $v_i = (\Pi_{\mathcal{D}_*} v)_{|K}(\boldsymbol{x}_i)$. Additionally, $\boldsymbol{x}_i \in \partial\Omega$ whenever $i \in I_{\partial\Omega}$.*

(6) (Mass-lumping) $\mathcal{D} = (X_{\mathcal{D},0}, \Pi_{\mathcal{D}}, \nabla_{\mathcal{D}}, Q_{\mathcal{D}})$ *is a mass-lumped version of the GD $\mathcal{D}_* = (X_{\mathcal{D},0}, \Pi_{\mathcal{D}_*}, \nabla_{\mathcal{D}}, Q_{\mathcal{D}})$ in the sense of Definition* 2.10*, which means that $\Pi_{\mathcal{D}}$ is piecewise constant on a partition $U = (U_i)_{i\in I}$ in the sense of Definition* 2.3*. We further assume that $U_i \cap K \neq \emptyset$ only if $i \in I_K$.*

(7) (Interpolate) $\mathcal{I}_{\mathcal{D}}\zeta(\overline{u}) \in X_{\mathcal{D},0}$ *is given by the nodal values of $\zeta(\overline{u})$, that is, $(\mathcal{I}_{\mathcal{D}}\zeta(\overline{u}))_i = \zeta(\overline{u})(\boldsymbol{x}_i)$ for all $i \in I$. This is well-defined since $\zeta(\overline{u}) \in C(\overline{\Omega})$ (see Remark* 2.18*).*

(8) (Quadrature rule) *The quadrature $Q_{\mathcal{D}}$ is defined on $C(\mathcal{M})$ (see* (2.18)*) by*

$$\forall g \in C(\mathcal{M})\,,\ \forall K \in \mathcal{M}\,, \quad (Q_{\mathcal{D}} g)_{|K} = \sum_{i\in I_K} g_{|K}(\boldsymbol{x}_i)\mathbf{1}_{U_i\cap K}. \tag{2.19}$$

(9) (Mesh regularity) *There exists $\rho > 0$ such that*
- *any $K \in \mathcal{M}$ is star-shaped with respect to all points in a ball of radius ρh_K,*
- *for all $i \in I$, $\rho h_{U_i} \leq h$.*

A few remarks are in order.

Remark 2.21 (local polynomial space). The space $\mathbb{P}^k$ in item (3) could be replaced by any of its subspace P_K that contains $\mathbb{P}^1$; the analysis would not be hindered, and some assumptions could even be weakened (see Remark 2.26). We chose to use $\mathbb{P}^k$ to simplify the presentation.

Remark 2.22 (nodes). The same i can belong to several I_K, as is the case for conforming finite elements. Conversely, in the case of DG schemes, for example, the following may occur (see the numerical example in section 3.3):
- one can have $\boldsymbol{x}_i = \boldsymbol{x}_j$ for $i \neq j$,
- I_K does not necessarily contain all the indices $i \in I$ such that $\boldsymbol{x}_i \in \overline{K}$,
- there can exist $i \in I_{\partial\Omega} \setminus (\cup_{K\in\mathcal{M}} I_K)$—but, in that case, $U_i = \emptyset$.

Remark 2.23 (quadrature rule). The GS (2.7) is usually implemented by assembling cell contributions. When the source term f is continuous on each cell, and since $\Pi_{\mathcal{D}} v$ is constant on each U_i, it is customary to use the simple—apparently low-order—quadrature rule defined by (2.19).

We also note that, since f and $\overline{u}$ belong to $C(\mathcal{M})$ by Remarks 2.18 and 2.19, the formula (2.19) can be used to compute $Q_{\mathcal{D}} f$ and $Q_{\mathcal{D}} \overline{u}$. These are the only values of $Q_{\mathcal{D}}$ of interest in the following analysis.

In the rest of this section, we write $a \lesssim b$ as shorthand for "$a \le Cb$ with C not depending on $\mathcal{M}$ or U, but possibly depending on ρ, k, and C_∇."

THEOREM 2.24 (high-order error estimate). *Under Assumption* 2.20*, let $\ell \ge 0$ be an integer and suppose that the local quadrature rules defined by $Q_{\mathcal{D}}$ are exact at degree $k+\ell$ (where k is the degree in item* (3) *of Assumption* 2.20*), that is,*

$$\forall K \in \mathcal{M}, \forall q \in \mathbb{P}^{k+\ell}, \quad \int_K q = \int_K Q_{\mathcal{D}} q = \sum_{i \in I_K} |U_i \cap K| q(\boldsymbol{x}_i). \tag{2.20}$$

Let $s \in \{1, \dots, \ell+2\}$ be such that Assumption 2.17 *holds. Then, the solution u to* (2.7) *satisfies*

$$\begin{aligned} &\|\nabla_{\mathcal{D}}[\mathcal{I}_{\mathcal{D}} \zeta(\overline{u}) - \zeta(u)]\|_{L^2} \\ &\lesssim W_{\mathcal{D}_*}(\Lambda \nabla \zeta(\overline{u})) + \|\nabla_{\mathcal{D}} \mathcal{I}_{\mathcal{D}} \zeta(\overline{u}) - \nabla \zeta(\overline{u})\|_{L^2} + h^s (1 + C_{\mathcal{D}_*}) \|\beta(\overline{u}) - f\|_{W^{s,\infty}(\mathcal{M})}. \end{aligned} \tag{2.21}$$

Let us make a few remarks.

Remark 2.25 (quadrature rule). The quadrature rule (2.20) bears similarities with the conditions on quadrature rules highlighted for finite elements in [9, 10]. However, in the proof below, the exactness condition (2.20) responds to a different need than the ones encountered in the analysis of mass-lumped finite elements for linear equations.

We also note that the precise geometry of the sets U_i is not important as long as (2.20) holds. This is due to the fact that, in the scheme (2.7), given the definitions (2.1) and (2.19) of $\Pi_{\mathcal{D}}$ and $Q_{\mathcal{D}}$, these sets U_i only appear through the quantities $|U_i \cap K|$.

Remark 2.26 (local polynomial space). Following Remark 2.21, if $\mathbb{P}^k$ is replaced by P_K in item (3) of Assumption 2.20, then an inspection of the proof below (see in particular the polynomial (2.26)) shows that (2.20) only has to be assumed for q belonging to the smaller space $P_K \mathbb{P}^l$. This is similar to what has been noticed in [20], in the context of mass-lumped $\mathbb{P}^k$ finite elements for linear equations.

Remark 2.27 (rates of convergence). If $\mathcal{D}_*$ is the GD corresponding to conforming $\mathbb{P}^k$ finite elements, we have $W_{\mathcal{D}_*} \equiv 0$ and, if $\zeta(\overline{u}) \in H^{k+1}(\Omega)$, $\|\nabla_{\mathcal{D}} \mathcal{I}_{\mathcal{D}} \zeta(\overline{u}) - \nabla \zeta(\overline{u})\|_{L^2} \lesssim h^k$; see [14, Proposition 8.11 and Remark 8.12]. In this case, (2.21) yields an $\mathcal{O}(h^{\min(s,k)})$ estimate on $\|\nabla_{\mathcal{D}}[\mathcal{I}_{\mathcal{D}} \zeta(\overline{u}) - \zeta(u)]\|_{L^2}$, which is a drastic improvement over (2.17) (see Remark 2.16).

The same $\mathcal{O}(h^{\min(s,k)})$ bound on $\|\nabla_{\mathcal{D}}[\mathcal{I}_{\mathcal{D}} \zeta(\overline{u}) - \zeta(u)]\|_{L^2}$ holds for the GD corresponding to DG schemes of degree k, provided that $\Lambda \nabla \zeta(\overline{u}) \in H^{\min(s,k)}(\Omega)^d$ (see [14, Lemmas 11.14 and 11.15]).

Before proving Theorem 2.24, we describe in Tables 1 and 2 a few examples of choices of $(\boldsymbol{x}_i, U_i \cap K)_{i \in I_K}$ that satisfy (2.20) in dimensions one and two. These rules will be used in the numerical tests in section 3, and they assume that the cell K is a simplex (interval if $d = 1$, triangle if $d = 2$). Note that some of these rules are

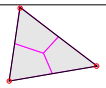
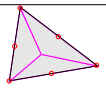
TABLE 1

Examples of quadrature rules satisfying (2.20) *in dimension* $d = 1$, *with* $K = (a, b)$. *DOE stands for degree of exactness and corresponds to* $k + \ell$ *in* (2.20). *In the illustrations, the circles represent the nodes* $\boldsymbol{x}_i$ *and the sets* $U_i \cap K$ *are the intervals delimited by vertical bars.*

Name	$(\boldsymbol{x}_i)_{i\in I_K}$	$(\lvert U_i \cap K\rvert)_{i\in I_K}$	DOE	Illustration
Trapezoidal	(a, b)	$(\frac{1}{2}\lvert K\rvert, \frac{1}{2}\lvert K\rvert)$	1	
Simpson	$(a, \frac{a+b}{2}, b)$	$(\frac{1}{6}\lvert K\rvert, \frac{2}{3}\lvert K\rvert, \frac{1}{6}\lvert K\rvert)$	3	
Equi6	$(a, \frac{2a+b}{3}, \frac{a+2b}{3}, b)$	$(\frac{1}{6}\lvert K\rvert, \frac{1}{3}\lvert K\rvert, \frac{1}{3}\lvert K\rvert, \frac{1}{6}\lvert K\rvert)$	1	
Equi8	$(a, \frac{2a+b}{3}, \frac{a+2b}{3}, b)$	$(\frac{1}{8}\lvert K\rvert, \frac{3}{8}\lvert K\rvert, \frac{3}{8}\lvert K\rvert, \frac{1}{8}\lvert K\rvert)$	3	
Gauss–Lobatto	$(a, \frac{5+\sqrt{5}}{10}a + \frac{5-\sqrt{5}}{10}b, \frac{5-\sqrt{5}}{10}a + \frac{5+\sqrt{5}}{10}b, b)$	$(\frac{1}{12}\lvert K\rvert, \frac{5}{12}\lvert K\rvert, \frac{5}{12}\lvert K\rvert, \frac{1}{12}\lvert K\rvert)$	5	

TABLE 2

Examples of quadrature rules satisfying (2.20) *in dimension* $d = 2$, *with* K *triangle* $(\mathbf{a}, \mathbf{b}, \mathbf{c})$. *DOE stands for degree of exactness and corresponds to* $k + \ell$ *in* (2.20). *In the illustrations, the nodes* $\boldsymbol{x}_i$ *are the circles and the sets* $U_i \cap K$ *are the regions delimited by straight lines.*

Name	$(\boldsymbol{x}_i)_{i\in I_K}$	$(\lvert U_i \cap K\rvert)_{i\in I_K}$	DOE	Illustration
Vertex	$(\mathbf{a}, \mathbf{b}, \mathbf{c})$	$(\frac{1}{3}\lvert K\rvert, \frac{1}{3}\lvert K\rvert, \frac{1}{3}\lvert K\rvert)$	1	
Vertex+Edge Midpoint	$(\mathbf{a}, \mathbf{b}, \mathbf{c}, \frac{\mathbf{a}+\mathbf{b}}{2}, \frac{\mathbf{a}+\mathbf{c}}{2}, \frac{\mathbf{b}+\mathbf{c}}{2})$	$(0, 0, 0, \frac{1}{3}\lvert K\rvert, \frac{1}{3}\lvert K\rvert, \frac{1}{3}\lvert K\rvert)$	2	

suboptimal in terms of degree of exactness versus number of quadrature points; they will serve to illustrate the optimality of Theorem 2.24.

Proof of Theorem 2.24. The inequality (2.21) follows from Theorem 2.12, estimating in the present context the terms $\mathfrak{T}_{\mathcal{D}}(\overline{u}, u)$ and $R_{\mathcal{D},\mathcal{D}_*}(\overline{u}, f)$.

(i) *Term* $\mathfrak{T}_{\mathcal{D}}(\overline{u}, u)$. For all $K \in \mathcal{M}$, all $i \in I_K$, and all $\boldsymbol{x} \in U_i \cap K$, Definition 2.3 of $\Pi_{\mathcal{D}}$ and items (7) and (8) in Assumption 2.20 imply that

$$\Pi_{\mathcal{D}}\mathcal{I}_{\mathcal{D}}\zeta(\overline{u})(\boldsymbol{x}) = (\mathcal{I}_{\mathcal{D}}\zeta(\overline{u}))_i = \zeta(\overline{u})(\boldsymbol{x}_i) = (\zeta(\overline{u}))_{|K}(\boldsymbol{x}_i) = \zeta(\overline{u}_{|K}(\boldsymbol{x}_i)) = \zeta(Q_{\mathcal{D}}\overline{u}(\boldsymbol{x})).$$

Hence, $\Pi_{\mathcal{D}}\mathcal{I}_{\mathcal{D}}\zeta(\overline{u}) = \zeta(Q_{\mathcal{D}}\overline{u})$ and $\mathfrak{T}_{\mathcal{D}}(\overline{u}, u) = 0$.

(ii) *Term* $R_{\mathcal{D},\mathcal{D}_*}(\overline{u}, f)$. For the sake of brevity, set $g = \beta(\overline{u}) - f$. By definition (2.19) of $Q_{\mathcal{D}}$, we have $Q_{\mathcal{D}}g = \beta(Q_{\mathcal{D}}\overline{u}) - Q_{\mathcal{D}}f$ and thus, to bound $R_{\mathcal{D},\mathcal{D}_*}(\overline{u}, f)$ above by the last term in (2.21), we have to establish that, for all $v \in X_{\mathcal{D},0}$,

$$\left|\int_\Omega (Q_{\mathcal{D}}g\,\Pi_{\mathcal{D}}v - g\,\Pi_{\mathcal{D}_*}v)\right| \lesssim h^s(1 + C_{\mathcal{D}_*})\|g\|_{W^{s,\infty}(\mathcal{M})}\|\nabla_{\mathcal{D}}v\|_{L^2}. \tag{2.22}$$

Let $A_{\mathcal{D},\mathcal{D}_*}(g, v)$ be the integral in the left-hand side of (2.22). We have

$$\begin{aligned} A_{\mathcal{D},\mathcal{D}_*}(g, v) &:= \sum_{K\in\mathcal{M}} \left(\sum_{i\in I_K} |U_i \cap K| g_{|K}(\boldsymbol{x}_i)v_i - \int_K g\,\Pi_{\mathcal{D}_*}v\right) \\ &= \sum_{K\in\mathcal{M}} \left(\sum_{i\in I_K} |U_i \cap K| g_{|K}(\boldsymbol{x}_i)(\Pi_{\mathcal{D}_*}v)_{|K}(\boldsymbol{x}_i) - \int_K g\,\Pi_{\mathcal{D}_*}v\right) = \sum_{K\in\mathcal{M}} \mathfrak{E}_K(g\Pi_{\mathcal{D}_*}v), \end{aligned} \tag{2.23}$$

where, in the second line, we have used $(\Pi_{\mathcal{D}_*}v)_{|K}(\boldsymbol{x}_i) = v_i$ (see item (5) in Assumption 2.20), and we have defined the error in the local quadrature rule on K by

$$\forall w \in C(\overline{K}), \quad \mathfrak{E}_K(w) := \sum_{i \in I_K} |U_i \cap K| w_{|K}(\boldsymbol{x}_i) - \int_K w.$$

By (2.20) and a straightforward estimate,

$$\forall q \in \mathbb{P}^{k+\ell}, \quad \mathfrak{E}_K(q) = 0, \text{ and} \tag{2.24}$$

$$\forall w \in C(\overline{K}), \quad |\mathfrak{E}_K(w)| \le 2|K| \|w\|_{L^\infty(K)}. \tag{2.25}$$

For a polynomial degree $r \ge 0$, let $\mathrm{Pr}_K^r : L^2(K) \to \mathbb{P}^r$ denote the $L^2(K)$-orthogonal projector on $\mathbb{P}^r$ and notice that, since $(\Pi_{\mathcal{D}_*} v)_{|K} \in \mathbb{P}^k$ (item (3) in Assumption 2.20) and $k \ge 1$, the function

$$q := (\mathrm{Pr}_K^\ell g)(\Pi_{\mathcal{D}_*} v)_{|K} + (\mathrm{Pr}_K^0(\Pi_{\mathcal{D}_*} v)_{|K})(\mathrm{Pr}_K^{\ell+1} g - \mathrm{Pr}_K^\ell g) \tag{2.26}$$

belongs to $\mathbb{P}^{\ell+k} + \mathbb{P}^{0+\ell+1} \subset \mathbb{P}^{k+\ell}$. Using (2.24) with this q yields

$$\begin{aligned}
\mathfrak{E}_K(g\Pi_{\mathcal{D}_*} v) &= \mathfrak{E}_K\left(g\Pi_{\mathcal{D}_*} v - (\mathrm{Pr}_K^\ell g)(\Pi_{\mathcal{D}_*} v)_{|K} - (\mathrm{Pr}_K^0(\Pi_{\mathcal{D}_*} v)_{|K})(\mathrm{Pr}_K^{\ell+1} g - \mathrm{Pr}_K^\ell g)\right) \\
&= \mathfrak{E}_K\Big([g - \mathrm{Pr}_K^\ell g][(\Pi_{\mathcal{D}_*} v)_{|K} - \mathrm{Pr}_K^0(\Pi_{\mathcal{D}_*} v)_{|K}] \\
&\qquad + (\mathrm{Pr}_K^0(\Pi_{\mathcal{D}_*} v)_{|K})[g - \mathrm{Pr}_K^{\ell+1} g]\Big).
\end{aligned}$$

Invoking then the bound (2.25) and the straightforward estimate $\|\mathrm{Pr}_K^0(\Pi_{\mathcal{D}_*} v)_{|K}\|_{L^\infty(K)} \le \|\Pi_{\mathcal{D}_*} v\|_{L^\infty(K)}$, we infer

$$\begin{aligned}
|\mathfrak{E}_K(g\Pi_{\mathcal{D}_*} v)| \le{}& 2\|g - \mathrm{Pr}_K^\ell g\|_{L^\infty(K)} |K| \|(\Pi_{\mathcal{D}_*} v)_{|K} - \mathrm{Pr}_K^0(\Pi_{\mathcal{D}_*} v)_{|K}\|_{L^\infty(K)} \\
&+ 2|K| \|\Pi_{\mathcal{D}_*} v\|_{L^\infty(K)} \|g - \mathrm{Pr}_K^{\ell+1} g\|_{L^\infty(K)}.
\end{aligned} \tag{2.27}$$

Under item (9) of Assumption 2.20, [12, Lemma 3.4] shows that, for any natural numbers $a \ge 0$ and $b \in \{0, \dots, a+1\}$, and any $w \in W^{b,\infty}(K)$,

$$\|w - \mathrm{Pr}_K^a w\|_{L^\infty(K)} \lesssim h_K^b \|w\|_{W^{b,\infty}(K)}.$$

Applying this estimate with $(a, b, w) = (\ell, \min(s, \ell+1), g)$, $(a, b, w) = (0, 1, (\Pi_{\mathcal{D}_*} v)_{|K})$ and $(a, b, w) = (\ell+1, s, g)$, (2.27) leads to

$$\begin{aligned}
|\mathfrak{E}_K(g\Pi_{\mathcal{D}_*} v)| \lesssim{}& h_K^{\min(s,\ell+1)} \|g\|_{W^{\min(s,\ell+1),\infty}(K)} |K| h_K \|\nabla(\Pi_{\mathcal{D}_*} v)_{|K}\|_{L^\infty(K)} \\
&+ |K| \|\Pi_{\mathcal{D}_*} v\|_{L^\infty(K)} h_K^s \|g\|_{W^{s,\infty}(K)}.
\end{aligned}$$

The discrete inverse Lebesgue embedding of [12, Lemma 5.1] gives, if $q \in \mathbb{P}^k(K)$, $|K| \|q\|_{L^\infty(K)} \lesssim |K|^{\frac{1}{2}} \|q\|_{L^2(K)}$. Applied to $q = (\Pi_{\mathcal{D}_*} v)_{|K}$ and $q =$ components of $\nabla(\Pi_{\mathcal{D}_*} v)_{|K}$, and since $\min(s, \ell+1) + 1 = \min(s+1, \ell+2) \ge s$, we obtain

$$|\mathfrak{E}_K(g\Pi_{\mathcal{D}_*} v)| \lesssim h_K^s \|g\|_{W^{s,\infty}(K)} |K|^{\frac{1}{2}} \left(\|\nabla(\Pi_{\mathcal{D}_*} v)_{|K}\|_{L^2(K)} + \|\Pi_{\mathcal{D}_*} v\|_{L^2(K)}\right).$$

Plugging this estimate into (2.23), using a discrete Cauchy–Schwarz inequality on the sums, and recalling item (4) in Assumption 2.20, we obtain

$$|A_{\mathcal{D}}(g, v)| \lesssim h^s \|g\|_{W^{s,\infty}(\mathcal{M})} \left(\|\nabla_{\mathcal{D}} v\|_{L^2} + \|\Pi_{\mathcal{D}_*} v\|_{L^2}\right).$$

The estimate (2.22) follows recalling the definition (2.4) of $C_{\mathcal{D}_*}$. ∎

3. Numerical illustrations. In this section, we present numerical tests to explore the optimality of the estimate in Theorem 2.24 and the necessity of the condition (2.20) on the chosen quadrature rules. This exploration will be conducted using mass-lumped finite elements and mass-lumped SIPG DG schemes. As seen in Remark 2.27, when $\zeta(\overline{u})$ is smooth enough, the expected rate of convergence of these schemes is $h^{\min(\ell+2,k)}$, where k is the degree of the underlying finite element or DG scheme. We will illustrate through examples that this rate can be optimal and that if (2.20) is not even satisfied for $\ell = 0$, then the rate of convergence falls to h (basic order one convergence for mass-lumped schemes; see [14, sections 8.4 and 9.6]). This illustration will be performed on both the porous medium equation and the Stefan model, in a variety of situations: with or without forcing term f (the latter being closer to genuine physical models), and also in the case where the right-hand side contains a term $\mathrm{div}F$ (in which case the convergence is hindered by the lack of regularity of $\zeta(\overline{u})$; see Remark 2.18).

In the following, each considered mass-lumped GD $\mathcal{D}$ shares the same elements $(X_{\mathcal{D},0}, \nabla_{\mathcal{D}}, \mathcal{I}_{\mathcal{D}}, Q_{\mathcal{D}})$ as the corresponding $\mathcal{D}_*$. We therefore start by describing the non-mass-lumped $\mathcal{D}_*$, after which, in the context of Assumption 2.20, $\mathcal{D}$ is completely determined by specifying the particular choices of nodes $(\boldsymbol{x}_i)_{i\in I_K}$ and weights $(|U_i \cap K|)_{i\in I_K}$ for each cell K, that is, of the local quadrature rules (2.20). The rules described in Tables 1 and 2 will serve as examples to construct the mass-lumped GDs $\mathcal{D}$.

3.1. Setting for the tests. The convergences are assessed through the following quantities:

$$E^{\Pi}_{\beta,\mathcal{I}_{\mathcal{D}}} = \|\beta(Q_{\mathcal{D}}\overline{u}) - \Pi_{\mathcal{D}}\beta(u)\|_{L^2(\Omega)}, \quad E^{\Pi}_{\zeta,\mathcal{I}_{\mathcal{D}}} = \|\Pi_{\mathcal{D}}(\mathcal{I}_{\mathcal{D}}\zeta(\overline{u}) - \zeta(u))\|_{L^2(\Omega)},$$
$$E^{\nabla}_{\zeta,\mathcal{I}_{\mathcal{D}}} = \|\nabla_{\mathcal{D}}(\mathcal{I}_{\mathcal{D}}\zeta(\overline{u}) - \zeta(u))\|_{L^2(\Omega)}, \quad E^{\nabla}_{\zeta} = \|\nabla\zeta(\overline{u}) - \nabla_{\mathcal{D}}\zeta(u)\|_{L^2(\Omega)},$$

measuring approximation errors on $\beta(Q_{\mathcal{D}}\overline{u})$, the interpolation of $\zeta(\overline{u})$ (for both function and gradient reconstruction), and on $\nabla\zeta(\overline{u})$ using high-order quadrature rules. A first-order polynomial fit is done on the logarithms of these errors with respect to $-\frac{1}{d}\log(\mathrm{Card}(I))$, which yields an approximation under the form

$$E \simeq C\mathrm{Card}(I)^{-\alpha/d}.$$

Our outputs give the numerical values of C and α, the latter providing a numerical convergence order with respect to an evaluation of the mesh size (the number of unknowns, $\mathrm{Card}(I)$, growing linearly with the number of cells).

All the one-dimensional (1D) and 2D tests refer to the following situations: $\Lambda = \mathrm{Id}$, $\beta = \mathrm{Id}$, and $\zeta \in \{\mathrm{Id}, \zeta_p, \zeta_s\}$, where the "porous medium" function ζ_p is defined by

$$\forall s \in \mathbb{R},\ \zeta_p(s) = \max(s,0)^2,$$

and the "Stefan" function ζ_s is defined by

$$\forall s \in \mathbb{R},\ \zeta_s(s) = \begin{cases} s & \text{if } s < 0, \\ 0 & \text{if } 0 \le s \le 1, \\ s-1 & \text{if } 1 < s. \end{cases}$$

In all the numerical tests, the approximate solution remains numerically bounded. There is therefore no need to redefine ζ_p on the negative axis in order to explicitly

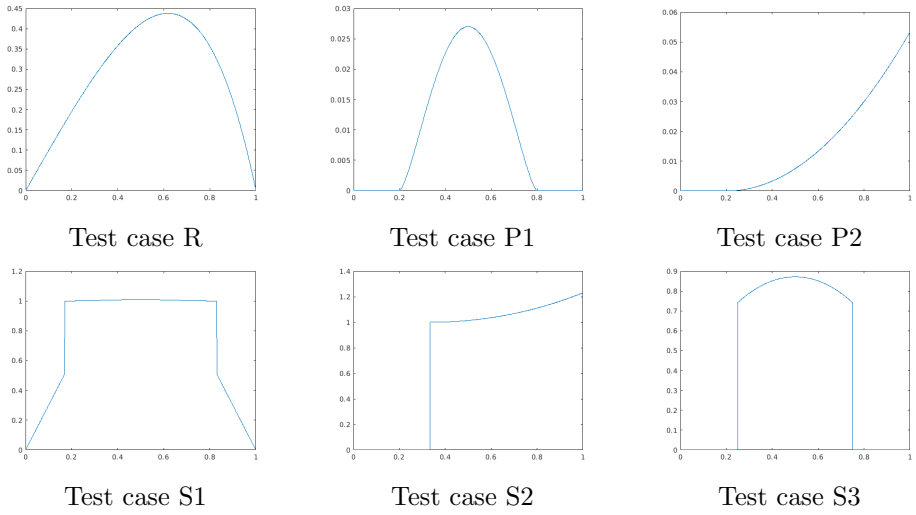


FIG. 1. *Profiles of the various exact solutions for the* 1*D tests.*

satisfy the superlinear bound in assumption (1.3b). Let us now give the complete continuous cases which are approximated below in one or two dimensions. The profiles of the corresponding exact solutions are presented in Figure 1.

Test case R: Regular problem, $f \neq 0$, $F = 0$. This problem corresponds to $\zeta = \mathrm{Id}$ (the model is therefore linear, but we still apply the mass-lumping process) and, for $x \in (0,1)$, the source term and solution are given by $f(x) = 4xe^x$ and $\overline{u}(x) = x(1-x)e^x$.

Test case P1: Porous medium problem, homogeneous Dirichlet BC, $f \neq 0$, $F = 0$. This test is on the porous medium equation, with $\zeta = \zeta_p$. The source term and exact solutions are defined as follows: for $x \in (0,1)$, setting $y_x = \max(x - 0.2, 0)$ and $z_x = \max(0.8 - x, 0)$, we take

$$f(x) = (y_x z_x)^{3/2} - 6y_x z_x(z_x^2 - 3y_x z_x + y_x^2) \quad \text{and} \quad \overline{u}(x) = (y_x z_x)^{3/2}.$$

Test case P2: Porous medium problem, nonhomogeneous Dirichlet BC, $f = 0$, $F = 0$. Still taking for ζ the porous medium function $\zeta = \zeta_p$, this test takes $\overline{u}(x) = \max(x - \frac{1}{5}, 0)^2/12$ for $x \in (0,1)$, which corresponds to the source term $f = 0$, and nonhomogeneous Dirichlet boundary conditions are imposed on $\zeta(\overline{u})$.

Test case S1: Stefan problem, homogeneous Dirichlet BC, $f \neq 0$, $F = 0$. In this test, the nonlinearity is given by the Stefan-like function $\zeta = \zeta_s$. The source term is given by $f(x) = 3(\frac{1}{2} - g(x))$ for $x \in (0,1)$, where $g(x) = |\frac{1}{2} - x|$. To describe $\overline{u}$, we first let $\gamma \in (0, \frac{1}{2})$ such that $\overline{u}(x) = f(x)$ for $g(x) \in (\gamma, \frac{1}{2})$ and $\overline{u}(x) \geq 1$ for $g(x) \in (0, \gamma)$. The ODE in (1.1) can then be solved on each subinterval and gives $\zeta(\overline{u}(x)) = 0$ for $g(x) \in (\gamma, \frac{1}{2})$ and, for some $a, b \in \mathbb{R}$, $\zeta(\overline{u}(x)) = ae^{g(x)} + be^{-g(x)} + 3(\frac{1}{2} - g(x)) - 1$ for $g(x) \in (0, \gamma)$. Hence, $\overline{u}(x) = ae^{g(x)} + be^{-g(x)} + 3(\frac{1}{2} - g(x))$ for $g(x) \in (0, \gamma)$. These values a, b, and γ are found by expressing the matching conditions ensuring that $\zeta(\overline{u}) \in H^2(0,1)$ (since $(\zeta(\overline{u}))'' = \overline{u} - f \in L^2(0,1)$), namely $\zeta(\frac{1}{2} \pm \gamma) = 0$ and $\zeta'(\frac{1}{2} \pm \gamma) = 0$; the symmetry of the problem also imposes $\zeta'(\frac{1}{2}) = 0$. This leads to the following equations:

$$3\left(\frac{1}{2}-\gamma\right)-1+ae^{\gamma}+be^{-\gamma}=0,\quad -3+ae^{\gamma}-be^{-\gamma}=0,\quad \text{and}\quad -3+a-b=0.$$

Numerically solving this nonlinear system of equations gives $\gamma\simeq 0.33036$, $a\simeq 1.2545$ and $b\simeq -1.7455$.

Test case S2: Stefan problem, nonhomogeneous Dirichlet BC, $f=0$, $F=0$.

As in the previous test, $\zeta=\zeta_s$. The source term is fixed at $f=0$ and, for any $\gamma\in[0,1]$, a solution is given by

$$\overline{u}(x)=\begin{cases}\frac{1}{2}(e^{x-\gamma}+e^{-(x-\gamma)})=\cosh(x-\gamma) & \forall x\in(\gamma,1),\\ 0 & \forall x\in(0,\gamma).\end{cases}$$

Nonhomogeneous Dirichlet conditions are imposed at $x=1$ to match the value of $\overline{u}$ there, and the tests are run with $\gamma=\frac{1}{3}$.

Test case S3: Stefan problem, homogeneous Dirichlet BC, $f\neq 0$, and $F\neq 0$. We let $\zeta=\zeta_s$ and

$$\begin{array}{llll} f(x)=0\,, & F(x)=4\frac{\sinh(1/4)}{\cosh(1/4)}\,, & \overline{u}(x)=0 & \forall x\in(0,\frac{1}{4});\\ f(x)=5\,, & F(x)=0\,, & \overline{u}(x)=5-4\frac{\cosh(x-1/2)}{\cosh(1/4)} & \forall x\in(\frac{1}{4},\frac{3}{4});\\ f(x)=0\,, & F(x)=-4\frac{\sinh(1/4)}{\cosh(1/4)}\,, & \overline{u}(x)=0 & \forall x\in(\frac{3}{4},1).\end{array}\tag{3.1}$$

3.2. Mass-lumped finite elements. For a conforming simplicial mesh $\mathcal{M}$ and using the notation in Assumption 2.20, the GD $\mathcal{D}_*=\mathcal{D}_*^{k,\mathrm{FE}}$, for $k\in\{1,2,3\}$, corresponding to conforming $\mathbb{P}^k$ finite elements, are defined by the following elements:

- The points $(\boldsymbol{x}_i)_{i\in I}$ are
  - if $k=1$, the mesh vertices (in dimension 1 or 2);
  - if $k=2$, the mesh vertices and one point in each cell (in dimension 1), or the mesh vertices and the edge midpoints (in dimension 2);
  - if $k=3$, the mesh vertices and two points in each cell (in dimension 1), or the mesh vertices, two additional points on each edge, and one point in each cell (in dimension 2).

  $I_{\partial\Omega}$ is the set of indices $i\in I$ such that $\boldsymbol{x}_i\in\partial\Omega$ and, for $K\in\mathcal{M}$, $I_K:=\{i\in I:\boldsymbol{x}_i\in\overline{K}\}$.
- For each simplex $K\in\mathcal{M}$ and $v=(v_i)_{i\in I}\in X_{\mathcal{D},0}$, $(\Pi_{\mathcal{D}_*}v)_{|K}$ is the unique polynomial in $\mathbb{P}^k$ that takes the values v_i at $\boldsymbol{x}_i$ for all $i\in I_K$.
- The gradient reconstruction is given by $(\nabla_{\mathcal{D}}v)_{|K}=\nabla(\Pi_{\mathcal{D}_*}v)_{|K}$ for all $v\in X_{\mathcal{D},0}$ and $K\in\mathcal{M}$.

3.2.1. Numerical tests for mass-lumped finite elements in dimension 1. We consider two families of meshes of $\Omega=(0,1)$ with N cells each, for $N\in\{16,32,64,512,1024,2048\}$. The first one is the uniform mesh $\mathcal{M}_N^{\mathrm{u}}$ with mesh step $h=1/N$. The second one is a random mesh $\mathcal{M}_N^{\mathrm{r}}$ such that each cell has size $h_i=H_i/\sum_j H_j$, where $H_i=(3+\rho_i)$ with ρ_i following a random uniform law on $(0,1)$. As mentioned above, all the GDs are mass-lumped versions of the corresponding $\mathbb{P}^k$ GD. We describe in Table 3 the remaining elements to fully define each GD, that is, the degree k and the chosen quadrature rule, in each cell, using the nomenclature introduced in Table 1. The nodes of each FE method are the union of the quadrature nodes in each cell; it is easily checked that the chosen quadrature rules always have the correct number of nodes to uniquely define the corresponding $\mathbb{P}^k$ functions in each cell, and the global continuity of the FE space is ensured since all considered quadrature rules include the cell endpoints as nodes.

TABLE 3
Descriptions of the mass-lumped finite element GDs in dimension $d = 1$. The degree k is that of the local polynomial space, and ℓ is the degree in (2.20) for the chosen quadrature rule ("–" means that (2.20) is not even satisfied with $\ell = 0$). The subscript g *can take the values* u*, when the considered mesh is uniform, or* r*, for random meshes.*

Name of GD	Degree k	Quadrature rule	ℓ
$\mathcal{D}_{\mathrm{g}}^{1,\mathrm{FE}}$	1	Trapezoidal	0
$\mathcal{D}_{\mathrm{g}}^{2,\mathrm{FE}}$	2	Simpson	1
$\mathcal{D}_{\mathrm{g},a}^{3,\mathrm{FE}}$	3	Equi6	–
$\mathcal{D}_{\mathrm{g},b}^{3,\mathrm{FE}}$	3	Equi8	0
$\mathcal{D}_{\mathrm{g},c}^{3,\mathrm{FE}}$	3	Gauss–Lobatto	2

TABLE 4
Constants and rates for Test Case R.

	$E^{\Pi}_{\beta,\mathcal{I}_\mathcal{D}}$		$E^{\Pi}_{\zeta,\mathcal{I}_\mathcal{D}}$		$E^{\nabla}_{\zeta,\mathcal{I}_\mathcal{D}}$		E^{∇}_{ζ}	
GD	C	α	C	α	C	α	C	α
$\mathcal{D}_{\mathrm{u}}^{1,\mathrm{FE}}$	4.6e-01	2.00	4.6e-01	2.00	4.4e-01	2.00	1.3e+00	1.00
$\mathcal{D}_{\mathrm{r}}^{1,\mathrm{FE}}$	2.5e-01	1.89	2.5e-01	1.89	3.1e-01	1.90	1.2e+00	0.99
$\mathcal{D}_{\mathrm{u}}^{2,\mathrm{FE}}$	8.8e-02	3.83	8.8e-02	3.83	1.4e-01	3.00	4.4e-01	2.00
$\mathcal{D}_{\mathrm{r}}^{2,\mathrm{FE}}$	7.4e-02	3.77	7.4e-02	3.77	1.3e-01	2.98	4.2e-01	1.98
$\mathcal{D}_{\mathrm{u},a}^{3,\mathrm{FE}}$	1.8e-01	2.00	1.8e-01	2.00	1.5e-01	1.00	1.5e-01	1.00
$\mathcal{D}_{\mathrm{r},a}^{3,\mathrm{FE}}$	2.0e-01	2.01	2.0e-01	2.01	1.5e-01	1.00	1.5e-01	1.00
$\mathcal{D}_{\mathrm{u},b}^{3,\mathrm{FE}}$	9.4e-02	3.00	9.4e-02	3.00	2.0e-01	2.00	2.0e-01	2.00
$\mathcal{D}_{\mathrm{r},b}^{3,\mathrm{FE}}$	9.6e-02	2.99	9.6e-02	2.99	2.0e-01	1.99	2.0e-01	1.99
$\mathcal{D}_{\mathrm{u},c}^{3,\mathrm{FE}}$	6.4e-08	1.73	6.4e-08	1.73	2.0e-04	2.95	7.2e-02	3.00
$\mathcal{D}_{\mathrm{r},c}^{3,\mathrm{FE}}$	9.0e-08	1.80	9.0e-08	1.80	2.4e-04	2.97	7.6e-02	3.00

TEST CASE R: REGULAR PROBLEM, $f \neq 0$. The results provided in Table 4 show a superconvergence, for $k = 1, 2$, of the function and gradient reconstruction when quadrature or interpolation are accounted for in the measure of the error: the errors $E^{\Pi}_{\beta,\mathcal{I}_\mathcal{D}}$, $E^{\Pi}_{\zeta,\mathcal{I}_\mathcal{D}}$, and $E^{\nabla}_{\zeta,\mathcal{I}_\mathcal{D}}$ appear to be at least $\mathcal{O}(h^{k+1})$ (even almost $\mathcal{O}(h^{k+2})$ for $k = 2$ and the function reconstruction errors). The rate for E^{∇}_{ζ} falls to h^k as expected since this is the optimal rate, when using piecewise $\mathbb{P}^k$ polynomials, to approximate a smooth nonpolynomial function. These rates for the approximation of the gradient are actually above those predicted by our analysis: as seen in Table 3, for all these methods we can take $\ell = 0$ in Theorem 2.24 and thus, following Remark 2.27, the expected decay of $E^{\nabla}_{\zeta,\mathcal{I}_\mathcal{D}}$ is $h^{\min(2,k)} = h^k$.

Regarding $k = 3$, the schemes based on the $\mathcal{D}_{\mathrm{g},c}^{3,\mathrm{FE}}$ variant appear to have a worse rate for the errors $E^{\Pi}_{\beta,\mathcal{I}_\mathcal{D}}$ and $E^{\Pi}_{\zeta,\mathcal{I}_\mathcal{D}}$ than $\mathcal{D}_{\mathrm{g},a}^{3,\mathrm{FE}}$ or $\mathcal{D}_{\mathrm{g},b}^{3,\mathrm{FE}}$, but focusing on the constant C we notice that these errors are actually much better. The scheme $\mathcal{D}_{\mathrm{g},c}^{3,\mathrm{FE}}$ also clearly outperforms the other two variants when considering the gradient reconstruction. Focusing on the latter, the convergence rate $\mathcal{O}(h)$ for $\mathcal{D}_{\mathrm{g},a}^{3,\mathrm{FE}}$ can be explained recalling

TABLE 5
Constants and rates for Test Case P1.

	$E^{\Pi}_{\beta,\mathcal{I}_\mathcal{D}}$		$E^{\Pi}_{\zeta,\mathcal{I}_\mathcal{D}}$		$E^{\nabla}_{\zeta,\mathcal{I}_\mathcal{D}}$		E^{∇}_{ζ}	
GD	C	α	C	α	C	α	C	α
$\mathcal{D}^{1,\mathrm{FE}}_{\mathrm{u}}$	2.3e+02	1.68	5.6e+00	2.01	1.2e+01	2.00	3.2e+00	1.00
$\mathcal{D}^{1,\mathrm{FE}}_{\mathrm{r}}$	3.4e+02	1.71	5.3e+00	2.00	1.2e+01	1.98	3.4e+00	1.01
$\mathcal{D}^{2,\mathrm{FE}}_{\mathrm{u}}$	1.9e+02	1.71	1.3e+00	2.69	4.3e+00	2.45	6.9e+00	2.01
$\mathcal{D}^{2,\mathrm{FE}}_{\mathrm{r}}$	9.5e+01	1.50	1.3e+01	3.16	1.6e+01	2.69	6.4e+00	1.99
$\mathcal{D}^{3,\mathrm{FE}}_{\mathrm{u},a}$	8.0e+01	1.82	4.4e-01	2.01	4.1e-01	1.03	4.0e-01	1.02
$\mathcal{D}^{3,\mathrm{FE}}_{\mathrm{r},a}$	9.2e+01	1.74	5.2e-01	2.03	4.2e-01	1.03	4.2e-01	1.03
$\mathcal{D}^{3,\mathrm{FE}}_{\mathrm{u},b}$	8.6e+01	1.74	2.8e+00	2.90	2.7e+00	1.99	2.7e+00	1.99
$\mathcal{D}^{3,\mathrm{FE}}_{\mathrm{r},b}$	3.0e+01	1.46	3.9e+00	3.01	2.3e+00	1.95	2.4e+00	1.96
$\mathcal{D}^{3,\mathrm{FE}}_{\mathrm{u},c}$	1.7e+01	1.41	1.0e+00	2.92	1.2e+00	2.42	2.7e+00	2.41
$\mathcal{D}^{3,\mathrm{FE}}_{\mathrm{r},c}$	3.5e+01	1.52	7.0e-02	2.17	2.2e-01	1.98	1.7e+00	2.28

that this variant does not even satisfy (2.20); even though $\mathfrak{T}_\mathcal{D}(\overline{u}, u) = 0$ for this method (see the proof of Theorem 2.24), we do not have any better estimate on $R_{\mathcal{D},\mathcal{D}_*}(\overline{u}, f)$ than in the proof of Corollary 2.15, which was precisely expected to be $\mathcal{O}(h)$ (see Remark 2.16).

On the contrary, referring again to Table 3, $\mathcal{D}^{3,\mathrm{FE}}_{\mathrm{g},b}$ enables us to take $\ell = 0$ in Theorem 2.24 and we recover the expected $\mathcal{O}(h^2)$ estimate on $E^{\nabla}_{\zeta,\mathcal{I}_\mathcal{D}}$ mentioned in Remark 2.27. For $\mathcal{D}^{3,\mathrm{FE}}_{\mathrm{g},c}$ we can even take $\ell = 2$ and Table 4 clearly shows that this leads to an improved and optimal $\mathcal{O}(h^3)$ estimate on the gradient (again, something predicted in Remark 2.27).

These results clearly demonstrate that the key factor in choosing a proper mass-lumped version for a high-order scheme is the exactness property (2.20)—not satisfying this property leads to decreased rates of convergence. They also indicate, at least for $k = 3$, the sharpness of the error estimate established in Theorem 2.24.

TEST CASE P1: POROUS MEDIUM PROBLEM, HOMOGENEOUS DIRICHLET BC, $f \neq 0$. The results are presented in Table 5. The functions f and $\overline{u}$ are only piecewise smooth, and the discontinuity of their derivatives is not necessarily aligned with the mesh. As a consequence, Assumption 2.17 does not hold. Compared to the smooth case studied in Test Case R, the convergence is overall degraded. However, the rates mostly remain not far from the linear case (especially for gradient approximations), and the main features discussed for the smooth case can also be found here: some errors display superconvergence behaviors (when using quadrature or interpolation of the exact solution), and the rates of convergence drop drastically if the local quadrature rule (2.20) does not hold with a high enough ℓ.

TEST CASE P2: POROUS MEDIUM PROBLEM, NONHOMOGENEOUS DIRICHLET BC, $f = 0$. Table 6 details the outcomes of this test. The source term is obviously smooth, but the solution is only piecewise smooth. Despite this, the results show that, except for the very small constants previously observed for $\mathcal{D}^{3,\mathrm{FE}}_{\mathrm{g},c}$, the schemes behave here in a very similar way as for the completely smooth situation of Test Case R. Here again we notice the importance of choosing proper local quadrature rules (2.20) when designing mass-lumped schemes from high-order methods.

TABLE 6
Constants and rates for Test Case P2.

	$E^{\Pi}_{\beta,\mathcal{I}_\mathcal{D}}$		$E^{\Pi}_{\zeta,\mathcal{I}_\mathcal{D}}$		$E^{\nabla}_{\zeta,\mathcal{I}_\mathcal{D}}$		E^{∇}_{ζ}	
GD	C	α	C	α	C	α	C	α
$\mathcal{D}^{1,\mathrm{FE}}_{\mathrm{u}}$	1.2e+01	1.99	2.2e-01	2.00	1.9e-01	2.00	1.3e+00	1.00
$\mathcal{D}^{1,\mathrm{FE}}_{\mathrm{r}}$	1.5e+01	2.02	3.9e-01	2.09	3.8e-01	1.97	1.4e+00	1.01
$\mathcal{D}^{2,\mathrm{FE}}_{\mathrm{u}}$	2.9e+00	2.50	2.1e-01	3.97	1.7e-01	2.99	5.3e-01	2.00
$\mathcal{D}^{2,\mathrm{FE}}_{\mathrm{r}}$	2.0e+00	2.41	2.2e-01	3.94	1.8e-01	2.98	5.2e-01	1.99
$\mathcal{D}^{3,\mathrm{FE}}_{\mathrm{u},a}$	3.9e+00	2.00	2.3e-01	2.00	1.4e-01	1.00	1.4e-01	1.00
$\mathcal{D}^{3,\mathrm{FE}}_{\mathrm{r},a}$	4.1e+00	2.00	2.4e-01	2.00	1.5e-01	1.00	1.5e-01	1.00
$\mathcal{D}^{3,\mathrm{FE}}_{\mathrm{u},b}$	3.9e+00	2.50	1.9e-01	3.00	2.4e-01	2.00	2.4e-01	2.00
$\mathcal{D}^{3,\mathrm{FE}}_{\mathrm{r},b}$	3.5e+00	2.47	2.0e-01	2.99	2.4e-01	1.99	2.4e-01	1.99
$\mathcal{D}^{3,\mathrm{FE}}_{\mathrm{u},c}$	2.7e-01	2.40	5.2e-07	2.33	2.7e-04	3.10	9.9e-02	3.00
$\mathcal{D}^{3,\mathrm{FE}}_{\mathrm{r},c}$	1.4e+00	2.76	2.8e-06	2.64	1.4e-03	3.46	1.1e-01	3.00

TABLE 7
Constants and rates for Test Case S1.

	$E^{\Pi}_{\beta,\mathcal{I}_\mathcal{D}}$		$E^{\Pi}_{\zeta,\mathcal{I}_\mathcal{D}}$		$E^{\nabla}_{\zeta,\mathcal{I}_\mathcal{D}}$		E^{∇}_{ζ}	
GD	C	α	C	α	C	α	C	α
$\mathcal{D}^{1,\mathrm{FE}}_{\mathrm{u}}$	1.8e+01	0.41	1.2e+01	1.97	1.2e+01	1.87	2.8e+00	1.00
$\mathcal{D}^{1,\mathrm{FE}}_{\mathrm{r}}$	3.7e+01	0.54	1.6e+01	2.15	3.2e+00	1.66	2.6e-01	-0.17
$\mathcal{D}^{2,\mathrm{FE}}_{\mathrm{u}}$	6.0e+01	0.76	1.1e+00	2.04	6.2e-01	1.54	2.5e+00	1.61
$\mathcal{D}^{2,\mathrm{FE}}_{\mathrm{r}}$	3.2e+01	0.50	1.2e+00	2.04	1.0e+00	1.65	3.3e-03	-0.16
$\mathcal{D}^{3,\mathrm{FE}}_{\mathrm{u},a}$	7.9e+01	0.84	1.2e+00	2.03	3.7e-01	1.03	4.4e-01	1.06
$\mathcal{D}^{3,\mathrm{FE}}_{\mathrm{r},a}$	1.0e+02	0.83	1.2e+00	2.15	3.8e-01	1.05	3.6e-02	0.28
$\mathcal{D}^{3,\mathrm{FE}}_{\mathrm{u},b}$	8.6e+01	0.84	3.8e-01	1.95	7.2e-01	1.61	8.9e-01	1.53
$\mathcal{D}^{3,\mathrm{FE}}_{\mathrm{r},b}$	5.1e+01	0.64	3.4e-01	1.84	2.8e-01	1.53	1.4e+03	2.77
$\mathcal{D}^{3,\mathrm{FE}}_{\mathrm{u},c}$	5.4e+01	0.67	4.6e-01	2.08	3.6e-01	1.58	8.5e-01	1.56
$\mathcal{D}^{3,\mathrm{FE}}_{\mathrm{r},c}$	5.1e+01	0.61	2.9e+00	2.41	3.9e-01	1.59	4.7e-01	0.37

TEST CASE S1: STEFAN PROBLEM, HOMOGENEOUS DIRICHLET BC, $f \neq 0$. The results for this test case are presented in Table 7. This test case is a much more severe one than the porous medium case, since the solution $\overline{u}$ is discontinuous. This explains the poor convergence of $E^{\Pi}_{\beta,\mathcal{I}_\mathcal{D}}$ for all considered methods. On the contrary, $\zeta(\overline{u})$ is continuous and $E^{\Pi}_{\zeta,\mathcal{I}_\mathcal{D}}$ thus behaves much better, with an order 2 decay for all schemes. The order of decay of $E^{\nabla}_{\zeta,\mathcal{I}_\mathcal{D}}$ is also similar for all methods (around 1.6), except for the GDs $\mathcal{D}^{3,\mathrm{FE}}_{\mathrm{g},a}$, for which it drops to 1; this reduction can be explained, as in the previous case, by recalling that these GDs do not satisfy the local quadrature rules (2.20) even for $\ell = 0$.

Based on our previous discussion, we could expect the schemes corresponding to $\mathcal{D}^{3,\mathrm{FE}}_{\mathrm{g},c}$ to have a higher rate of convergence than the other methods, but it should be

TABLE 8
Constants and rates for Test Case S1*, with errors computed excluding the discontinuities.*

	$E^{\Pi}_{\beta,\mathcal{I}_\mathcal{D}}$		$E^{\Pi}_{\zeta,\mathcal{I}_\mathcal{D}}$		$E^{\nabla}_{\zeta,\mathcal{I}_\mathcal{D}}$		E^{∇}_{ζ}	
GD	C	α	C	α	C	α	C	α
$\mathcal{D}^3_{\mathrm{u},a}$	1.2e+00	2.03	1.2e+00	2.03	3.0e-01	1.04	3.0e-01	1.04
$\mathcal{D}^3_{\mathrm{r},a}$	1.1e+00	2.21	1.1e+00	2.21	2.7e-01	1.03	1.9e-02	0.19
$\mathcal{D}^3_{\mathrm{u},b}$	3.3e-01	1.96	3.3e-01	1.96	1.5e+00	2.00	1.2e+00	1.94
$\mathcal{D}^3_{\mathrm{r},b}$	6.4e-01	2.08	6.4e-01	2.08	4.3e-01	1.75	7.1e-02	0.70
$\mathcal{D}^3_{\mathrm{u},c}$	3.9e-01	2.09	3.9e-01	2.09	1.1e-02	2.10	7.1e-05	0.49
$\mathcal{D}^3_{\mathrm{r},c}$	2.0e-01	1.88	2.0e-01	1.88	6.8e-01	2.12	3.2e-07	-2.04

noted that $\zeta(\overline{u})$ belongs only to H^2, not H^3, since $(\zeta(\overline{u}))'' = \overline{u} - f$ is discontinuous. This limits the application of Theorem 2.24 to $s = 2$ (despite $\ell = 2$ being a valid choice in this case).

We notice that E^{∇}_{ζ} has a quite poor convergence (or does not seem to converge) on random meshes. Given that the difference between this error and $E^{\nabla}_{\zeta,\mathcal{I}_\mathcal{D}}$ solely lies in the interpolation error $\|\nabla_\mathcal{D}\mathcal{I}_\mathcal{D}\zeta(\overline{u}) - \nabla\zeta(\overline{u})\|_{L^2}$, this apparently indicates that this interpolation error does not converge on random meshes. It is actually not the case, but for these meshes the regularity factor and maximum size oscillate a lot from one mesh to the other; combined with the low regularity of the solution (which implies an expected slow rate of convergence), this explains that the regression performed on the interpolation errors struggles to capture the correct convergence when considering a finite family of meshes.

For this test case where the singularities of $(\zeta(\overline{u}))''$ are located at specific points (namely, the discontinuities $\frac{1}{2} \pm \gamma$ of $\overline{u}$), it is interesting to consider the errors far from these points. Table 8 presents the regression data when the errors are computed excluding the two intervals of length 2/10 around $\frac{1}{2} \pm \gamma$. We observe that $\mathcal{D}^{3,\mathrm{FE}}_{\mathrm{g},b}$ and $\mathcal{D}^{3,\mathrm{FE}}_{\mathrm{g},c}$ then yield a second-order rate of convergence for $E^{\nabla}_{\zeta,\mathcal{I}_\mathcal{D}}$ (at least on uniform meshes), which is an improvement over the rate $h^{1.6}$ obtained when errors are computed over the entire domain (Table 7). We, however, do not recover a full h^3 rate of convergence for the methods $\mathcal{D}^{3,\mathrm{FE}}_{\mathrm{g},c}$, a sign that the localized lack of regularity of $\zeta(\overline{u})$ impacts the errors over the entire domain (which is expected for an elliptic equation with an infinite propagation speed). Noticeably, the variants $\mathcal{D}^{3,\mathrm{FE}}_{\mathrm{g},a}$ still only provide an order one rate of convergence, which is not surprising since the accuracy of these schemes is limited not by a lack of regularity of the solution but by an improper choice of quadrature rules, which impacts the error on the entire domain.

TEST CASE S2: STEFAN PROBLEM, NONHOMOGENEOUS DIRICHLET BC, $f = 0$. The results, presented in Table 9, are comparable to those obtained with a nonzero source term in Test Case S1 (reduced convergence for $\mathcal{D}^{3,\mathrm{FE}}_{\mathrm{g},a}$, limitation of the convergence for $\mathcal{D}^{3,\mathrm{FE}}_{\mathrm{g},c}$ due to the limited regularity of $\zeta(\overline{u})$). In this case, however, the gradient $\nabla_\mathcal{D}\mathcal{I}_\mathcal{D}\zeta(\overline{u})$ of the interpolate seems to enjoy better convergence property even on random meshes, which preserve a reasonable convergence of E^{∇}_{ζ}.

TEST CASE S3: STEFAN PROBLEM, HOMOGENEOUS DIRICHLET CONDITIONS, $f \neq 0$ AND $F \neq 0$. The term $-\int_\Omega F \cdot \nabla_\mathcal{D} v$ in the GS (2.7) is exactly computed, without numerical quadrature, using the relation

$$-\int_\Omega F\cdot\nabla_\mathcal{D} v = -\int_0^1 F(s)(\Pi_{\mathcal{D}_*} v)'(s)ds = -4\frac{\sinh(1/4)}{\cosh(1/4)}\left(\Pi_{\mathcal{D}_*} v\left(\frac{1}{4}\right) + \Pi_{\mathcal{D}_*} v\left(\frac{3}{4}\right)\right).$$

TABLE 9
Constants and rates for Test Case S2.

	$E^{\Pi}_{\beta,\mathcal{I}_\mathcal{D}}$		$E^{\Pi}_{\zeta,\mathcal{I}_\mathcal{D}}$		$E^{\nabla}_{\zeta,\mathcal{I}_\mathcal{D}}$		E^{∇}_{ζ}	
GD	C	α	C	α	C	α	C	α
$\mathcal{D}^{1,\mathrm{FE}}_{\mathrm{u}}$	2.0e+00	0.50	2.6e-01	1.98	1.5e-01	1.48	7.7e-01	1.00
$\mathcal{D}^{1,\mathrm{FE}}_{\mathrm{r}}$	4.7e+00	0.67	5.2e-02	1.78	3.8e-02	1.32	7.6e-01	1.00
$\mathcal{D}^{2,\mathrm{FE}}_{\mathrm{u}}$	2.3e+00	0.49	1.2e-01	2.02	8.6e-02	1.50	2.0e-01	1.50
$\mathcal{D}^{2,\mathrm{FE}}_{\mathrm{r}}$	2.4e+00	0.53	1.6e-01	2.13	1.0e-01	1.61	2.1e-01	1.51
$\mathcal{D}^{3,\mathrm{FE}}_{\mathrm{u},a}$	3.4e+00	0.50	9.3e-02	2.00	8.9e-02	1.01	9.2e-02	1.01
$\mathcal{D}^{3,\mathrm{FE}}_{\mathrm{r},a}$	2.9e-02	-0.21	6.8e-02	1.94	8.5e-02	1.00	8.8e-02	1.00
$\mathcal{D}^{3,\mathrm{FE}}_{\mathrm{u},b}$	4.1e+00	0.53	5.6e-02	2.03	8.0e-02	1.50	1.1e-01	1.50
$\mathcal{D}^{3,\mathrm{FE}}_{\mathrm{r},b}$	1.7e+01	1.10	1.5e-01	2.28	1.8e-01	1.75	9.3e-02	1.51
$\mathcal{D}^{3,\mathrm{FE}}_{\mathrm{u},c}$	3.1e+00	0.50	4.9e-02	2.01	5.3e-02	1.49	9.3e-02	1.50
$\mathcal{D}^{3,\mathrm{FE}}_{\mathrm{r},c}$	3.1e+00	0.71	6.0e-02	2.16	7.8e-02	1.78	6.1e-02	1.52

TABLE 10
Constants and rates for Test Case S3.

	$E^{\Pi}_{\beta,\mathcal{I}_\mathcal{D}}$		$E^{\Pi}_{\zeta,\mathcal{I}_\mathcal{D}}$		$E^{\nabla}_{\zeta,\mathcal{I}_\mathcal{D}}$		E^{∇}_{ζ}	
GD	C	α	C	α	C	α	C	α
$\mathcal{D}^{1,\mathrm{FE}}_{\mathrm{u}}$	3.8e+01	0.50	3.5e+01	2.01	7.7e+00	1.49	1.2e+00	0.71
$\mathcal{D}^{1,\mathrm{FE}}_{\mathrm{r}}$	2.3e+01	0.42	4.0e-01	0.81	5.8e-01	0.82	5.7e-01	0.34
$\mathcal{D}^{2,\mathrm{FE}}_{\mathrm{u}}$	2.2e+01	0.50	3.6e+00	2.00	1.6e+00	1.50	3.7e-01	0.51
$\mathcal{D}^{2,\mathrm{FE}}_{\mathrm{r}}$	1.2e+01	0.42	3.1e-03	-0.29	7.2e-02	0.26	6.6e-01	0.44
$\mathcal{D}^{3,\mathrm{FE}}_{\mathrm{u},a}$	2.2e+01	0.50	3.3e+00	2.01	6.5e-01	1.18	3.6e-01	0.51
$\mathcal{D}^{3,\mathrm{FE}}_{\mathrm{r},a}$	2.9e+00	0.17	4.7e-03	-0.03	6.3e-02	0.14	3.6e-01	0.35
$\mathcal{D}^{3,\mathrm{FE}}_{\mathrm{u},b}$	1.8e+01	0.50	2.3e+00	2.00	1.0e+00	1.50	3.6e-01	0.50
$\mathcal{D}^{3,\mathrm{FE}}_{\mathrm{r},b}$	5.4e+00	0.22	4.3e-01	0.95	2.2e-01	0.46	7.0e-01	0.50
$\mathcal{D}^{3,\mathrm{FE}}_{\mathrm{u},c}$	1.5e+01	0.50	8.8e-01	2.00	5.7e-01	1.50	3.5e-01	0.50
$\mathcal{D}^{3,\mathrm{FE}}_{\mathrm{r},c}$	5.5e+00	0.26	9.1e-03	-0.05	8.3e-02	0.33	6.4e-01	0.48

The outcome of the test can be seen in Table 10. We note that these data do not satisfy the assumptions of Theorem 2.24, and no high-order rate can therefore be expected. Actually, the solution displays a very low regularity since $\zeta(\overline{u})$ only belongs to H^1, not even H^2. This is represented in the results by the fact that, for each given error and type of mesh (random/uniform), the rates of convergence for all degrees k are in the same range. We notice also that, across the board, the schemes perform better on regular grids rather than random grids.

3.2.2. Numerical tests for mass-lumped finite elements in dimension 2. In the following 2D cases, we consider the domain $\Omega = (0,1) \times (0,1)$, the polynomial degrees $k = 1, 2$, and the following meshes (see Figure 2):

- Triangular meshes which are as equilateral as possible, with edge length $1/N$ for $N \in \{25, 50, 100\}$. The GDs on these meshes will have the subscript "e," e.g., $\mathcal{D}^k_{\mathrm{e}}$.

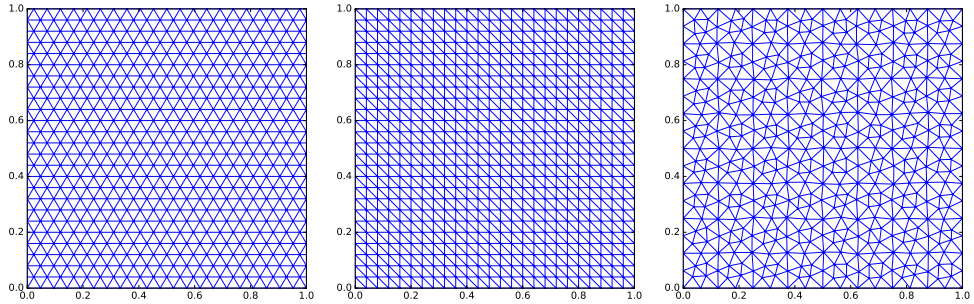


FIG. 2. *The three types of 2D meshes: "*e*" mesh (left), "*s*" mesh (middle), "*r*" mesh (right).*

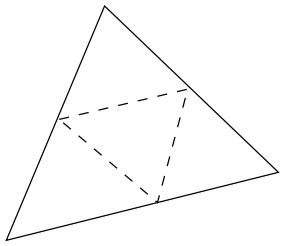


FIG. 3. *Division of a mesh triangle into four to construct* $\mathcal{D}^{1,\mathrm{FE}}_{\mathrm{g},1/4}$.

TABLE 11

Descriptions of the mass-lumped finite element GDs in dimension $d = 2$. *The degree* k *is that of the local polynomial space, and* ℓ *is the degree in* (2.20) *for the chosen quadrature rule. Here,* $\mathrm{g} \in \{\mathrm{e}, \mathrm{s}, \mathrm{r}\}$ *depending on whether the GD uses the triangular equilateral meshes, the triangular meshes coming from splitting rectangles in two, or the random triangular meshes.*

Name of GD	Degree k	Quadrature rule	ℓ
$\mathcal{D}^{1,\mathrm{FE}}_{\mathrm{g}}$	1	Vertex	0
$\mathcal{D}^{2,\mathrm{FE}}_{\mathrm{g}}$	2	Vertex+Edge Midpoint	0
$\mathcal{D}^{1,\mathrm{FE}}_{\mathrm{g},1/4} = \mathcal{D}^{1,\mathrm{FE}}_{\mathrm{g}}$ on the submesh described in Figure 3.			

- Rectangular triangular meshes obtained by splitting N^2 squares in 2, for $N \in \{25, 50, 100\}$. We use the subscript "s" for these GDs, e.g., $\mathcal{D}^k_{\mathrm{s}}$.
- Random meshes based on the three meshes `mesh1_3`, `mesh1_4`, and `mesh1_5` from the FVCA5 benchmark [21]. The randomness is obtained moving the internal nodes by a uniform random factor, and we use the subscript "r" for these GDs, such as in $\mathcal{D}^k_{\mathrm{r}}$.

Based on these meshes, the mass-lumped versions of $\mathcal{D}^k_*$ $(k = 1, 2)$ are described in Table 11. The quadrature rules refer to the rules described in Table 2, and the nodes of the finite element method are the union of the quadrature nodes in all the cells. Note that $\mathcal{D}^{1,\mathrm{FE}}_{\mathrm{g},1/4}$ has degree $k = 1$ but has the same unknowns as $\mathcal{D}^{2,\mathrm{FE}}_{\mathrm{g}}$, corresponding to $k = 2$, on the original mesh.

Remark 3.1 (implementation for $k = 2$). If $k = 2$, the function reconstruction obtained by mass-lumping does not see the vertex unknowns ($U_i = \emptyset$ if $\boldsymbol{x}_i$ is a

mesh vertex). The corresponding mass matrix is therefore singular, which is of course an issue when considering explicit discretisations of time-dependent (even linear) problems; solving this issue requires the usage of enriched $\mathbb{P}^2$ elements [10]. However, in the context of implicit time stepping, or equivalently of stationary problems, this is not an issue since the stiffness matrix is always nonsingular.

The case of stationary nonlinear degenerate equations such as (1.1) requires nonetheless an implementation trick. Since the diffusion term acts on $\zeta(u)$, in the nonlinear iterations the stiffness matrix is multiplied by $\zeta'(u)$, which can vanish, and the diffusion term does not yield in itself a control of all the unknowns u_i. It does, however, enable a control of the unknowns $\zeta(u)_i = \zeta(u_i)$. Even though these unknowns, especially for the Stefan problem, do not determine u entirely, this gives a way to implement the scheme in a nonsingular way. Instead of writing an equation on $(u_i)_{i\in I}$, we write an equation on $((u_i)_{i\in I_e}, (\zeta(u)_i)_{i\in I_v})$, where I_e is the set of indices corresponding to edge midpoints, and I_v is the set of indices corresponding to the vertices. The unknowns $(u_i)_{i\in I_e}$ are controlled by the mass-lumped reaction term and the unknowns $(\zeta(u)_i)_{i\in I_v}$ by the diffusion term. This implementation does not entirely determine a solution u to the scheme, only its values at the edge midpoints and the values of $\zeta(u)$ at the vertices, but this is expected given Lemma 2.7 and Remark 2.8.

Remark 3.2 (the case $k = 3$). If $k = 3$, it is possible to satisfy the local quadrature rules (2.20) with $\ell = 0$ (i.e., to have rules exact for third-degree polynomials). This is done fixing $\alpha \in (0, (\frac{3}{44})^{1/2})$ and choosing the nodes $\boldsymbol{x}_i$ and weight proportions $|U_i \cap K|/|K|$ as follows:

- the vertices of the mesh, each one associated with proportion $\frac{3-44\alpha^2}{60(1-4\alpha^2)}$;
- two points on each edge located at the barycentric coordinates $(\frac{1}{2} \pm \alpha, \frac{1}{2} \mp \alpha)$ on the edge, associated with proportion $\frac{1}{15(1-4\alpha^2)}$;
- the centers of mass of the triangles, associated with proportion $9/20$.

To have local quadratures of degree four (that is, (2.20) with $\ell = 1$), one must set $\alpha^2 = 1/12$, which leads to the negative weight proportion $-1/60$ at the vertices of the triangle, a situation which is incompatible with the mass-lumping setting. To properly mass-lump the $\mathbb{P}^3$ finite elements while preserving their high-order, an enriched version of these elements must be considered [10].

The data we consider in the following test case are the same as for the 1D case, using the diagonal as a 1D coordinate. For example, if g is a solution or source term for a 1D test case, the solution or source term for the corresponding 2D case is computed by setting $\widetilde{g}(x, y) = g((x + y)/\sqrt{2})$. All these 2D test cases therefore have nonhomogeneous Dirichlet boundary conditions.

Tests with $\mathcal{D}_{\mathrm{g}}^k$, $k = 1, 2$. Tables 12–15 present the results for the 2D versions of the Test Cases P1, P2, S1, and S2, that is, porous medium with $f \neq 0$, porous

Table 12
Constants and rates for the 2D version of Test Case P1.

GD	$\mathcal{D}_{\mathrm{e}}^{1,\mathrm{FE}}$		$\mathcal{D}_{\mathrm{e}}^{2,\mathrm{FE}}$		$\mathcal{D}_{\mathrm{s}}^{1,\mathrm{FE}}$		$\mathcal{D}_{\mathrm{s}}^{2,\mathrm{FE}}$		$\mathcal{D}_{\mathrm{r}}^{1,\mathrm{FE}}$		$\mathcal{D}_{\mathrm{r}}^{2,\mathrm{FE}}$	
	C	α	C	α	C	α	C	α	C	α	C	α
$E^{\Pi}_{\beta,\mathcal{I}_\mathcal{D}}$	9.1e-03	2.06	2.9e-02	2.95	5.0e-03	2.03	3.0e-03	2.59	2.0e-02	2.02	8.1e-03	2.50
$E^{\Pi}_{\zeta,\mathcal{I}_\mathcal{D}}$	3.3e-04	2.07	2.8e-03	3.53	1.5e-04	2.04	2.0e-03	4.04	1.5e-03	2.03	1.2e-03	3.01
$E^{\nabla}_{\zeta,\mathcal{I}_\mathcal{D}}$	1.4e-03	1.51	7.7e-03	2.52	7.5e-04	2.04	6.9e-03	3.02	2.0e-03	1.02	2.8e-03	2.02

TABLE 13
Constants and rates for the 2D version of Test Case P2.

GD	$\mathcal{D}_{\rm e}^{1,\rm FE}$		$\mathcal{D}_{\rm e}^{2,\rm FE}$		$\mathcal{D}_{\rm s}^{1,\rm FE}$		$\mathcal{D}_{\rm s}^{2,\rm FE}$		$\mathcal{D}_{\rm r}^{1,\rm FE}$		$\mathcal{D}_{\rm r}^{2,\rm FE}$	
	C	α	C	α	C	α	C	α	C	α	C	α
$E^{\Pi}_{\beta,\mathcal{I}_\mathcal{D}}$	1.4e-01	1.90	8.2e-02	1.69	4.8e-02	1.70	3.7e-03	1.02	3.1e-02	1.42	9.3e-02	1.63
$E^{\Pi}_{\zeta,\mathcal{I}_\mathcal{D}}$	1.8e-03	2.06	4.0e-02	3.48	9.5e-04	2.05	1.8e-02	3.22	1.7e-03	2.02	1.8e-03	2.56
$E^{\nabla}_{\zeta,\mathcal{I}_\mathcal{D}}$	2.5e-02	2.03	1.2e-01	2.52	1.3e-02	2.01	3.9e-02	2.38	2.6e-03	1.18	5.4e-02	2.29

TABLE 14
Constants and rates for the 2D version of Test Case S1.

GD	$\mathcal{D}_{\rm e}^{1,\rm FE}$		$\mathcal{D}_{\rm e}^{2,\rm FE}$		$\mathcal{D}_{\rm s}^{1,\rm FE}$		$\mathcal{D}_{\rm s}^{2,\rm FE}$		$\mathcal{D}_{\rm r}^{1,\rm FE}$		$\mathcal{D}_{\rm r}^{2,\rm FE}$	
	C	α	C	α	C	α	C	α	C	α	C	α
$E^{\Pi}_{\beta,\mathcal{I}_\mathcal{D}}$	1.3e-01	0.45	1.9e-01	0.52	1.7e-01	0.57	7.6e-02	0.32	8.3e-02	0.38	6.0e-02	0.29
$E^{\Pi}_{\zeta,\mathcal{I}_\mathcal{D}}$	1.7e-02	2.04	4.5e-02	2.64	4.1e-03	1.88	4.6e-03	1.95	1.2e-02	1.96	2.7e-02	2.22
$E^{\nabla}_{\zeta,\mathcal{I}_\mathcal{D}}$	6.9e-02	1.66	6.0e-02	1.53	1.1e-02	1.33	4.8e-02	1.50	1.5e-02	1.07	1.0e-01	1.64

TABLE 15
Constants and rates for the 2D version of Test Case S2.

GD	$\mathcal{D}_{\rm e}^{1,\rm FE}$		$\mathcal{D}_{\rm e}^{2,\rm FE}$		$\mathcal{D}_{\rm s}^{1,\rm FE}$		$\mathcal{D}_{\rm s}^{2,\rm FE}$		$\mathcal{D}_{\rm r}^{1,\rm FE}$		$\mathcal{D}_{\rm r}^{2,\rm FE}$	
	C	α	C	α	C	α	C	α	C	α	C	α
$E^{\Pi}_{\beta,\mathcal{I}_\mathcal{D}}$	2.5e-01	0.48	4.0e-01	0.58	8.1e-02	0.35	5.0e-01	0.68	1.4e-01	0.37	7.7e-01	0.72
$E^{\Pi}_{\zeta,\mathcal{I}_\mathcal{D}}$	2.4e-02	2.10	1.6e-02	2.10	2.4e-02	2.24	3.2e-02	2.23	3.8e-02	2.06	1.5e-02	1.86
$E^{\nabla}_{\zeta,\mathcal{I}_\mathcal{D}}$	7.5e-02	1.57	9.5e-02	1.51	7.2e-02	1.71	9.5e-02	1.52	5.0e-02	1.02	9.8e-02	1.49

medium with $f = 0$, Stefan problem with $f \neq 0$, and Stefan problem with $f = 0$. Plots of solutions for the Stefan problems are given in Figure 4 (2D version of Test Case S1, $f \neq 0$) and Figure 5 (2D version of Test Case S2, $f = 0$).

All the considered GDs satisfy the local quadrature rules (2.20) with $\ell = 0$. Accordingly, if the solution and source were smooth, rates of convergence for $E^{\nabla}_{\zeta,\mathcal{I}_\mathcal{D}}$ should be $\mathcal{O}(h^k)$ for $k = 1, 2$ (see Remark 2.27). The results show that, for the porous medium case, we are above these rates for all meshes, except for the random mesh—probably more representative of genuine situations—where we are at these rates (or slightly above). As in the 1D case, the Stefan problem is more challenging and, probably due to the loss of regularity of the solution, the rates are a little bit worse. They do, however, remain at or above $\mathcal{O}(h)$ for $k = 1$ and drop only to around $\mathcal{O}(h^{1.5})$ for $k = 2$.

TEST WITH $\mathcal{D}_{\rm g,1/4}^{1,\rm FE}$ AND $\mathcal{D}_{\rm g}^{2,\rm FE}$: COMPARISON BETWEEN DEGREE 1 AND DEGREE 2. To properly assess the interest of using a second-order scheme over a first-order scheme, we now look, on the same triangular mesh, at the outputs of $\mathcal{D}_{\rm r}^{2,\rm FE}$ and $\mathcal{D}_{\rm r,1/4}^{1,\rm FE}$. This makes for a fair comparison since these two schemes have the same number of unknowns. For each errors $E = E^{\Pi}_{\beta,\mathcal{I}_\mathcal{D}}$, $E = E^{\Pi}_{\zeta,\mathcal{I}_\mathcal{D}}$, and $E = E^{\nabla}_{\zeta,\mathcal{I}_\mathcal{D}}$, letting E_k be the error corresponding to $\mathcal{D}_{\rm r,1/4}^{1,\rm FE}$ if $k = 1$, or to $\mathcal{D}_{\rm r}^{2,\rm FE}$ if $k = 2$, we compute the ratios $r = E_2/E_1$ for all the tests on the three random meshes based on `mesh1_3`, `mesh1_4`, and `mesh1_5`. Assuming that each error E_k is of the form $C_k(\frac{h}{h_0})^{\alpha_k}$, where h_0 is the size of the reference mesh `mesh1_3`, the ratio between the two errors should be given

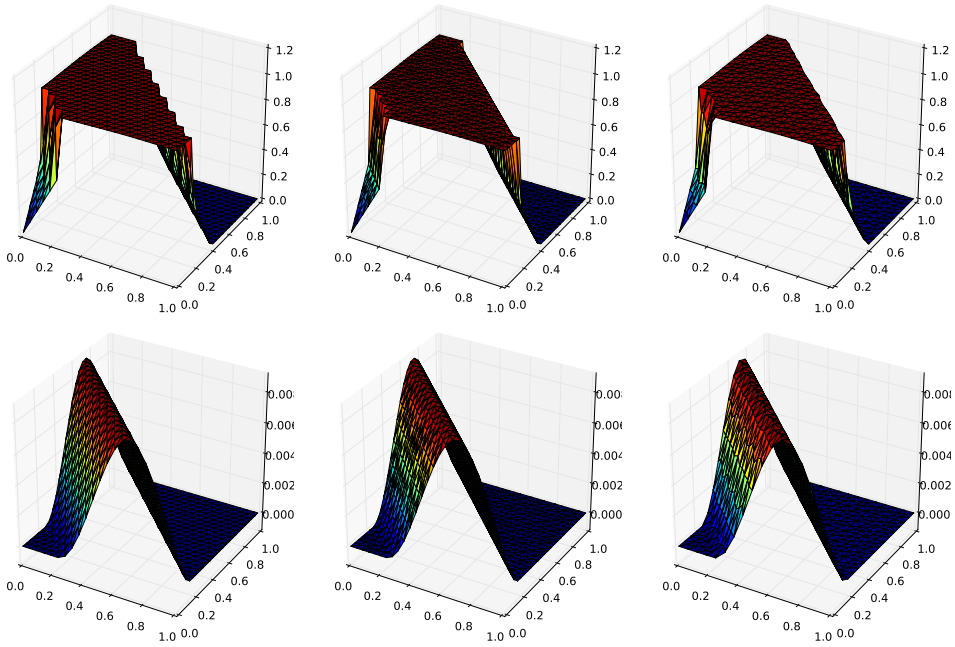


FIG. 4. *Approximate functions (top: u; bottom: $\zeta(u)$) for the 2D version of Test Case* S1. *From left to right:* $\mathcal{D}_{\mathrm{e}}^{1,\mathrm{FE}}$, $\mathcal{D}_{\mathrm{s}}^{1,\mathrm{FE}}$, $\mathcal{D}_{\mathrm{r}}^{1,\mathrm{FE}}$.

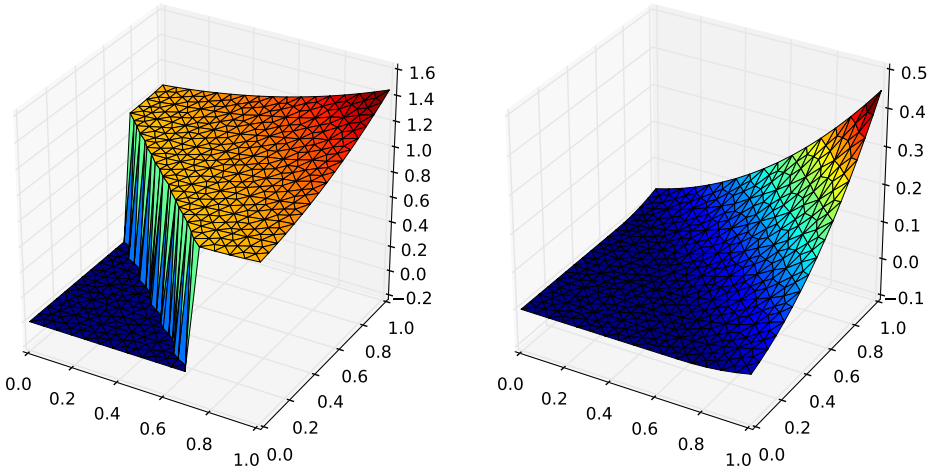


FIG. 5. *Approximate functions u (left) and $\zeta(u)$ (right) for the 2D version of Test Case* S2, *using* $\mathcal{D}_{\mathrm{r}}^{1,\mathrm{FE}}$.

by

$$r = \frac{E_2}{E_1} = \frac{C_2}{C_1}\left(\frac{h}{h_0}\right)^{\alpha_2-\alpha_1}.$$

Table 16 performs a $C(h/h_0)^{\alpha}$ regression of the ratio r. Hence, the C values in this table can be considered as approximations of $\frac{C_2}{C_1}$ and the α values as approximations

TABLE 16
Constants and rates for the comparison of first-/second-order with the same number of degrees of freedom.

Case	Test Case P1		Test Case P2		Test Case S1		Test Case S2	
	C	α	C	α	C	α	C	α
$E^{\Pi}_{\beta,\mathcal{I}_\mathcal{D}}$ 2/1	5.8e-02	1.63	1.3e+00	0.99	1.2e+00	0.37	1.5e+00	0.53
$E^{\Pi}_{\zeta,\mathcal{I}_\mathcal{D}}$ 2/1	1.6e-02	2.00	1.4e-01	1.75	6.9e-01	1.27	7.2e-01	0.97
$E^{\nabla}_{\zeta,\mathcal{I}_\mathcal{D}}$ 2/1	4.4e-02	1.83	4.2e-01	2.00	8.5e-01	1.26	4.6e-01	1.34

of $\alpha_2 - \alpha_1$. The results show a clear advantage (smaller C_k, larger α_k) of the second-order method over the first-order method and also that this advantage still holds, albeit reduced, for irregular (Stefan) test cases.

3.3. Numerical tests for mass-lumped DG schemes. The mesh $\mathcal{M}$ being a general polytopal mesh as in [14, Definition 7.2], still using the notation in Assumption 2.20, the GD $\mathcal{D}_* = \mathcal{D}_*^{k,\mathrm{DG}}$ for the SIPG method of order k is defined as follows:

- For each cell $K \in \mathcal{M}$, points $(\boldsymbol{x}_i)_{i\in I_K}$ are chosen such that for each choice of real numbers $(w_i)_{i\in I_K}$ there is a unique $q \in \mathbb{P}^k$ such that $q(\boldsymbol{x}_i) = w_i$ for all $i \in I_K$. Then $I = (\cup_{K\in\mathcal{M}} I_K) \cup I_{\partial\Omega}$ is the family that gathers the indices of all these points for all the cells, and of all the boundary points where a jump is accounted for in the expression of $\nabla_\mathcal{D}$.
- For each $K \in \mathcal{M}$ and $v = (v_i)_{i\in I} \in X_{\mathcal{D},0}$, $(\Pi_{\mathcal{D}_*} v)_{|K}$ is the unique polynomial in $\mathbb{P}^k$ that takes the values v_i at $\boldsymbol{x}_i$ for all $i \in I_K$.
- The gradient reconstruction is given by $(\nabla_\mathcal{D} v)_{|K} = \nabla(\Pi_{\mathcal{D}_*} v)_{|K} + S_K(v)$ for all $v \in X_{\mathcal{D},0}$ and $K \in \mathcal{M}$, where $S_K(v)$ is an appropriate stabilization term accounting for the jumps appearing in the DG scheme (see [14, Definition 11.1] for details).

Remark 3.3 (embedding the SIPG method into the GDM). The SIPG stabilization term is accounted for in the design of the gradient reconstruction $\nabla_\mathcal{D}$ through a penalization parameter (denoted by β in [14, Chapter 11]) which is fixed at 0.6 in all the tests.

We take $k = 3$ and consider the same families of uniform and random meshes of $\Omega = (0, 1)$ with N cells each as in section 3.2.1. Table 17 provides the remaining elements to fully define the GDs. These elements follow closely the choices made for the 1D finite element meshes in section 3.2.1, but there is a major difference in the choice of the global nodes. Since DG methods do not have any continuity conditions at the mesh vertices, each of these vertices corresponds to two different nodes (one for each cell the vertex belongs to, or one additional node to encode the boundary conditions for the domain endpoints), with different associated values of the unknowns/test functions. The nodes are therefore $\boldsymbol{x}_0 = \boldsymbol{x}_1 = 0 < \boldsymbol{x}_2 < \boldsymbol{x}_3 < \boldsymbol{x}_4 = \boldsymbol{x}_5 < \ldots \boldsymbol{x}_{4i} = \boldsymbol{x}_{4i+1} < \boldsymbol{x}_{4i+2} < \boldsymbol{x}_{4i+3} < \ldots \boldsymbol{x}_{4N} = \boldsymbol{x}_{4N+1} = 1$, with each cell corresponding to $(\boldsymbol{x}_{4i+1}, \boldsymbol{x}_{4i+4})$ for $i = 0, \ldots, N-1$.

The results in Table 18 are in line with what we already observed for finite element methods (numerical tests conducted for $k \in \{1, 2\}$, not presented here, lead to similar conclusions). The better local quadrature rules of $\mathcal{D}^{3,\mathrm{DG}}_{\mathrm{g},c}$ enable this variant to outperform $\mathcal{D}^{3,\mathrm{DG}}_{\mathrm{g},b}$ and $\mathcal{D}^{3,\mathrm{DG}}_{\mathrm{g},a}$ (this latter being badly hindered by its very low-order local quadrature rule) and preserve the expected order 3 convergence for smooth data and solutions. This optimal convergence is even noticed in the fully nonlinear Test Case P2. As before, the Stefan problem is much more challenging due to its reduced

TABLE 17

Descriptions of the mass-lumped DG GDs in dimension $d = 1$. The degree k is that of the local polynomial spaces, and ℓ is the degree in (2.20) *for the chosen quadrature rule ("–" means that* (2.20) *does not even hold for $\ell = 0$). Here,* g = u *or* r *depending if the GD uses the uniform or random meshes.*

Name of GD	Degree k	Quadrature rule	ℓ
$\mathcal{D}_{\mathrm{g},a}^{3,\mathrm{DG}}$	3	Equi6	–
$\mathcal{D}_{\mathrm{g},b}^{3,\mathrm{DG}}$	3	Equi8	0
$\mathcal{D}_{\mathrm{g},c}^{3,\mathrm{DG}}$	3	Gauss–Lobatto	2

TABLE 18

Constants and rates for $E_{\zeta,\mathcal{I}_\mathcal{D}}^{\nabla}$ with DG applied to some 1D test cases.

	Test Case R		Test Case P1		Test Case P2		Test Case S2	
GD	C	α	C	α	C	α	C	α
$\mathcal{D}_{\mathrm{u},a}^{3,\mathrm{DG}}$	1.5e-01	1.01	4.2e-01	1.03	1.5e-01	1.01	9.0e-02	1.01
$\mathcal{D}_{\mathrm{r},a}^{3,\mathrm{DG}}$	1.7e-01	1.02	4.2e-01	1.03	1.5e-01	1.01	8.5e-02	1.00
$\mathcal{D}_{\mathrm{u},b}^{3,\mathrm{DG}}$	2.2e-01	2.00	2.9e+00	1.98	2.7e-01	2.00	8.2e-02	1.50
$\mathcal{D}_{\mathrm{r},b}^{3,\mathrm{DG}}$	2.2e-01	1.99	2.9e+00	1.97	2.8e-01	2.00	7.4e-02	1.57
$\mathcal{D}_{\mathrm{u},c}^{3,\mathrm{DG}}$	2.3e-02	3.25	1.4e+00	2.39	3.9e-02	3.42	5.4e-02	1.49
$\mathcal{D}_{\mathrm{r},c}^{3,\mathrm{DG}}$	1.1e-02	3.01	1.0e+00	2.32	1.9e-02	3.08	5.8e-02	1.58

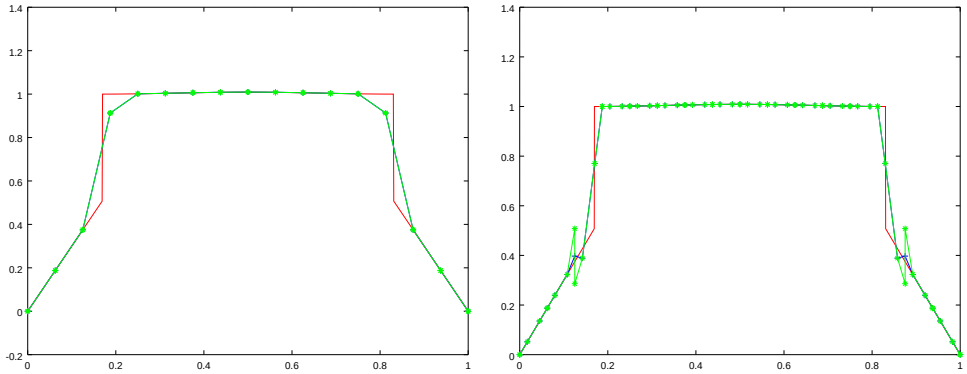


FIG. 6. *Comparison of approximate u on Test Case* S1*: $k = 1$ (left), $k = 3$ Gauss–Lobatto (right), finite element (blue, "+"), DG (green, "$*$"), exact (red), $N = 16$, uniform mesh.*

regularity, but even for this one we notice an interest in selecting a method with high enough local quadrature rules.

Figure 6 shows the solutions u obtained with finite element and DG schemes, for $k = 1$ and 3, on Test Case S1 and with a relatively coarse mesh ($N = 16$). As expected, the solutions obtained with $k = 3$ are much more accurate. They, however, present oscillations (more severe for DG than for finite element) in the vicinity of the discontinuity of u. The solutions (not shown here) for Test Case S2, corresponding to $f = 0$, do not display such oscillations.

4. Conclusion. We presented a generic analysis framework, covering a range of methods, for the numerical approximation of nonlinear degenerate elliptic equations, stationary version of the Stefan, or porous medium problems. We identified a particular structure of the method, the piecewise constant function reconstruction, which is sufficient and also appears to be necessary to establish the robustness of the schemes and to obtain error estimates. We showed how to design mass-lumping versions of high-order numerical methods in order to preserve, despite the usage of piecewise constant approximations in the scheme, high-order approximations of the solution to this severely nonlinear model. Our numerical tests on mass-lumped finite element and DG schemes corroborated the theoretical findings, showing that even for nonsmooth solutions an elevated rate of convergence is obtained only if the mass-lumping is designed to satisfy proper local quadrature rules.

Appendix A. Existence and uniqueness of the weak solution.

THEOREM A.1 (existence and uniqueness of the weak solution). *Under Assumption* (1.3)*, there is a unique solution* $\overline{u}$ *to* (1.2)*. This solution has the following regularity properties:*

- $\Lambda\nabla\zeta(\overline{u}) + F \in H_{\rm div}(\Omega)$*;*
- *if* $d \le 3$ *and* $F \in L^p(\Omega)^d$ *for some* $p > d$*, then* $\zeta(\overline{u}) \in C^\theta(\overline{\Omega})$ *for some* $\theta \in (0,1)$ *depending only on* Ω*,* Λ *and* p*;*
- *if* $F = 0$*,* Ω *is convex, and* Λ *is Lipschitz continuous, then* $\zeta(\overline{u}) \in H^2(\Omega)$*.*

Proof. The existence of a solution is a consequence of Theorem 2.9, together with Lemma A.2 that establishes the existence of a proper sequence of GDs. To prove the uniqueness of this solution, consider $\overline{u}_1$ and $\overline{u}_2$ two solutions to (1.2), subtract their respective equations, and take $v = \zeta(\overline{u}_1) - \zeta(\overline{u}_2) \in H_0^1(\Omega)$ as a test function to get

$$\int_\Omega (\beta(\overline{u}_1) - \beta(\overline{u}_2))(\zeta(\overline{u}_1) - \zeta(\overline{u}_2)) + \int_\Omega \Lambda\nabla(\zeta(\overline{u}_1) - \zeta(\overline{u}_2)) \cdot \nabla(\zeta(\overline{u}_1) - \zeta(\overline{u}_2)) = 0.$$

The first term is nonnegative since β and ζ are nondecreasing, and thus $\nabla(\zeta(\overline{u}_1) - \zeta(\overline{u}_2)) = 0$. This shows that $\zeta(\overline{u}_1) = \zeta(\overline{u}_2)$. The weak formulation (1.2) also shows that $\beta(\overline{u}_1) - \Delta\zeta(\overline{u}_1) = f + \mathrm{div}(F) = \beta(\overline{u}_2) - \Delta\zeta(\overline{u}_2)$ in the sense of distributions on Ω; since $\zeta(\overline{u}_1) = \zeta(\overline{u}_2)$, this yields $\beta(\overline{u}_1) = \beta(\overline{u}_2)$. Hence, $\beta(\overline{u}_1) + \zeta(\overline{u}_1) = \beta(\overline{u}_2) + \zeta(\overline{u}_2)$ and hypothesis (1.3d) shows that $\overline{u}_1 = \overline{u}_2$.

We finally consider the regularity properties of $\zeta(\overline{u})$. This function is a weak solution of

$$\zeta(\overline{u}) \in H_0^1(\Omega) \text{ and } -\mathrm{div}(\Lambda\nabla\zeta(\overline{u}) + F) = f - \beta(\overline{u}) \in L^2(\Omega).$$

This readily shows that $\Lambda\nabla\zeta(\overline{u}) + F \in H_{\rm div}(\Omega)$. If $d \le 3$, then $L^2 \subset W^{-1,q}(\Omega)$ for some $q > d$ and thus, assuming that $F \in L^p(\Omega)^d$ for $p > d$, $\zeta(\overline{u})$ is a solution in $H_0^1(\Omega)$ of $-\mathrm{div}(\Lambda\nabla\zeta(\overline{u})) = f + \mathrm{div}(F) - \beta(\overline{u}) \in W^{-1,\min(q,p)}(\Omega)$; the results of [25] then show that $\zeta(\overline{u})$ has the Hölder regularity stated in the theorem. Finally, the H^2 regularity property is a straightforward consequence of the optimal elliptic regularity on convex domains for Lipschitz-continuous diffusion tensors. ∎

LEMMA A.2 (existence of suitable sequences of GDs). *Under Assumption* (1.3a)*, there exists a sequence* $(\mathcal{D}_m)_{m\in\mathbb{N}} = (X_{\mathcal{D}_m,0}, \Pi_{\mathcal{D}_m}, \nabla_{\mathcal{D}_m}, Q_{\mathcal{D}_m})_{m\in\mathbb{N}}$ *of GDs, with piecewise constant reconstructions, that satisfy the coercivity, consistency, limit-conformity, and compactness properties stated in Theorem* 2.9*.*

Proof. Let $(\widetilde{\mathcal{M}}_m)_{m\in\mathbb{N}}$ be a sequence of conformal simplicial meshes of $\mathbb{R}^d$ (see, e.g., [14, Definition 7.4]), such that $\lim_{m\to\infty}\max_{T\in\widetilde{\mathcal{M}}_m}\operatorname{diam}(T)\to 0$ and $(\mathcal{M}_m)_{m\in\mathbb{N}}$ is regular in the sense that the ratio of the diameter of $T\in\widetilde{\mathcal{M}}_m$ over the largest ball inside T is bounded uniformly with respect to T and m. We let $\mathcal{M}_m=\{T\in\widetilde{\mathcal{M}}_m : T\subset\Omega\}$ and define the polyhedral set $\Omega_m\subset\Omega$ as the interior of $\cup_{T\in\mathcal{M}_m}\overline{T}_m$.

The GD $\mathcal{D}_m=(X_{\mathcal{D}_m,0},\Pi_{\mathcal{D}_m},\nabla_{\mathcal{D}_m},Q_{\mathcal{D}_m})$ is defined as the mass-lumped conforming $\mathbb{P}^1$ GD on the mesh $\mathcal{M}_m$ of Ω_m [14, section 8.4], with extensions to Ω by 0 outside Ω_m, and no quadrature rule. Letting $\mathcal{V}_m$ be the set of vertices of $\mathcal{M}_m$, we therefore set

- $X_{\mathcal{D}_m,0}=\{v=(v_i)_{i\in\mathcal{V}_m} : v_i\in\mathbb{R},\ v_i=0 \text{ if } i\in\partial\Omega_m\}$;
- for $v\in X_{\mathcal{D}_m,0}$, $(\Pi_{\mathcal{D}_m}v)_{|\Omega_i}=v_i$ for all $i\in\mathcal{V}_m$, where $(\Omega_i)_{i\in\mathcal{V}_m}$ is the dual (Donald) mesh of $\mathcal{M}_m$, and $\Pi_{\mathcal{D}_m}v=0$ on $\Omega\backslash\Omega_m$;
- for $v\in X_{\mathcal{D}_m,0}$, $\nabla_{\mathcal{D}_m}v$ is on Ω_m the gradient of the conforming $\mathbb{P}^1$ reconstruction from the vertex values $(v_i)_{i\in\mathcal{V}_m}$, and $\nabla_{\mathcal{D}_m}v=0$ on $\Omega\backslash\Omega_m$;
- $Q_{\mathcal{D}_m}=\mathrm{Id}: L^2(\Omega)\to L^2(\Omega)$.

Since the functions and gradient reconstructions are extended by 0 outside Ω_m, $C_{\mathcal{D}_m}$ and $W_{\mathcal{D}_m}$ can be computed using norms and integrals over Ω_m. The properties of mass-lumped $\mathbb{P}^1$ GDs on Ω_m (see [14, Theorem 8.17]) then show that $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is coercive, limit-conforming, and compact. It remains to analyze the consistency of $(\mathcal{D}_m)_{m\in\mathbb{N}}$.

As seen in [14, Lemma 2.16], the consistency follows if we prove that $S_{\mathcal{D}_m}(\varphi)\to 0$ when $\varphi\in C_c^\infty(\Omega)$. In that case, for m large enough, $\varphi\in C_c^\infty(\Omega_m)$ and the norms in $S_{\mathcal{D}_m}(\varphi)$ can be restricted to Ω_m. The estimate in [14, Remark 8.18] then shows that $S_{\mathcal{D}_m}(\varphi)\le C_\varphi\max_{T\in\mathcal{M}_m}\operatorname{diam}(T)$ with C_φ not depending on m. This shows that $S_{\mathcal{D}_m}(\varphi)\to 0$ as $m\to\infty$, as required. ∎

Appendix B. Conforming scheme. Throughout this section, we assume that $F=0$. Using assumptions (1.3b), (1.3c), and (1.3d), we see that $\beta+\zeta:\mathbb{R}\to\mathbb{R}$ is bijective and we can therefore set $\mu(t)=\zeta((\beta+\zeta)^{-1}(t))$ and $\rho(t):=t-\mu(t)=\beta((\beta+\zeta)^{-1}(t))$. These functions are nondecreasing and 1-Lipschitz continuous and, setting $\overline{w}=(\beta+\zeta)(\overline{u})$, we see that (1.2) is equivalent to: find $w\in L^2(\Omega)$ such that $\mu(\overline{w})\in H_0^1(\Omega)$ and

$$\int_\Omega\rho(\overline{w})\overline{v}+\int_\Omega\Lambda\nabla\mu(\overline{w})\cdot\nabla\overline{v}=\int_\Omega f\overline{v}\qquad\forall\overline{v}\in H_0^1(\Omega). \tag{B.1}$$

Given a family $(V_m)_{m\in\mathbb{N}}$ of finite dimensional subspaces of $H_0^1(\Omega)$, conforming schemes for (B.1) are written: find $w_m\in V_m$ such that

$$\int_\Omega\rho(w_m)v+\int_\Omega\Lambda\nabla\mu(w_m)\cdot\nabla v=\int_\Omega fv\qquad\forall v\in V_m. \tag{B.2}$$

Introducing the function $\nu:\mathbb{R}\to\mathbb{R}$ defined by $\nu(s)=\int_0^s\sqrt{\mu'(r)}dr$, we can then state the following convergence theorem.

THEOREM B.1 (convergence of the scheme). *Assume that* (1.3) *holds and that, for all* $\varphi\in H_0^1(\Omega)$, $\lim_{m\to\infty}\inf_{v\in V_m}\|\varphi-v\|_{H_0^1(\Omega)}=0$. *Then, for any* $m\in\mathbb{N}$, *there exists* w_m *a solution to* (B.2) *and, if* $\overline{w}$ *is the solution to* (B.1), *as* $m\to\infty$, *we have* $\mu(u_m)\to\mu(\overline{w})$ *weakly in* $H_0^1(\Omega)$ *and strongly in* $L^2(\Omega)$, $\nu(w_m)\to\nu(\overline{w})$ *weakly in* $H_0^1(\Omega)$ *and strongly in* $L^2(\Omega)$, *and* $\rho(w_m)\to\rho(\overline{w})$ *weakly in* $L^2(\Omega)$.

Moreover, if the energy equality

$$\int_\Omega \rho(\overline{w})\overline{w} + \int_\Omega \Lambda\nabla\nu(\overline{w})\cdot\nabla\nu(\overline{w}) = \int_\Omega f\overline{w} \tag{B.3}$$

holds, then $\nabla\nu(w_m)\to\nabla\nu(\overline{w})$ *and* $w_m\to\overline{w}$ *strongly in* $L^2(\Omega)$.

Remark B.2 (on condition (B.3)). We observe that (B.3) holds in the case where $\overline{w}\in H^1_0(\Omega)$ since it can then be taken as a test function in (B.1). But it may also hold in some less regular situations.

Proof. We only sketch the proof. Assuming the existence of a solution w_m to the scheme, we let $v=v_m$ in (B.2), use the monotonicity of μ and ρ, the relation $\mu'(w_m)|\nabla w|^2=|\nabla\nu(w_m)|^2$, the coercivity of Λ, and the Poincaré inequality, and we write (with $a\lesssim b$ meaning $a\le Cb$ with C independent of m):

$$\underline{\lambda}\|\nabla\nu(w_m)\|_{L^2}^2 \le \|f\|_{L^2}\|w_m\|_{L^2} \lesssim \|f\|_{L^2}(1+\|\mu(w_m)\|_{L^2}) \lesssim \|f\|_{L^2}(1+\|\nabla\mu(w_m)\|_{L^2}). \tag{B.4}$$

We have $|\nabla\mu(w_m)|^2=\mu'(w_m)|\nabla\nu(w_m)|^2\le|\nabla\nu(w_m)|^2$ and the estimate above therefore gives a bound on $\nu(w_m)$ in $H^1_0(\Omega)$, and thus also on $\mu(w_m)$. Using a coercivity property of μ similar to that of ζ we infer bounds in $L^2(\Omega)$ on w_m and $\rho(w_m)$. A topological degree argument, similar to the one developed in the proof of Lemma 2.7, then ensures the existence of at least one solution w_m to (B.2).

These bounds give $\overline{v}\in H^1_0(\Omega)$ and $\overline{w}\in L^2(\Omega)$ such that, up to a subsequence, $\mu(w_m)\to\overline{v}$ strongly in $L^2(\Omega)$, $\nabla\mu(w_m)\to\nabla\overline{v}$ weakly in $L^2(\Omega)^d$, and $w_m\to\overline{w}$ weakly in $L^2(\Omega)$. By weak/strong convergence we infer that

$$\lim_{m\to\infty}\int_\Omega w_m\mu(w_m) = \int_\Omega \overline{w}\,\overline{v}$$

and a Minty argument [14, Lemma D.10] yields $\overline{v}=\mu(\overline{w})$, and thus $\rho(w_m)\to\overline{w}-\mu(\overline{w})=\rho(\overline{w})$ weakly in $L^2(\Omega)$. We have $(\nu(a)-\nu(b))^2\le(b-a)(\mu(b)-\mu(a))$ and the strong convergence of $\mu(w_m)$ in L^2 therefore shows that $\nu(w_m)\to\nu(\overline{w})$ in $L^2(\Omega)$. Since $(\nu(w_m))_{m\in\mathbb{N}}$ is bounded in $H^1_0(\Omega)$, this convergence also holds weakly in this space.

Letting $\varphi\in H^1_0(\Omega)$ and taking $v_m:=\operatorname{argmin}_{v\in V_m}\|\varphi-v\|_{H^1_0(\Omega)}$ in (B.2), the above convergences enable us to take the limit as $m\to\infty$ to see that $\overline{w}$ is the solution to (B.1). The uniqueness of $\overline{w}$ shows that the convergence property holds for the whole sequence.

Assuming that (B.3) holds, we apply (B.2) with $v=w_m$ to get

$$\begin{aligned}\lim_{m\to\infty}\left(\int_\Omega\rho(w_m)w_m+\int_\Omega\Lambda\nabla\nu(w_m)\cdot\nabla\nu(w_m)\right) &= \int_\Omega f\overline{w}\\ &= \int_\Omega\rho(\overline{w})\overline{w}+\int_\Omega\Lambda\nabla\nu(\overline{w})\cdot\nabla\nu(\overline{w}).\end{aligned} \tag{B.5}$$

The weak convergence of $\nu(w_m)$ in $H^1_0(\Omega)$ ensures that

$$\liminf_{m\to\infty}\int_\Omega\Lambda\nabla\nu(w_m)\cdot\nabla\nu(w_m)\ge\int_\Omega\Lambda\nabla\nu(\overline{w})\cdot\nabla\nu(\overline{w}). \tag{B.6}$$

Developing the relation $\int_\Omega(\rho(w_m)-\rho(\overline{w}))(w_m-\overline{w})\ge0$ and using the weak convergences $w_m\to\overline{w}$ and $\rho(w_m)\to\rho(\overline{w})$ in $L^2(\Omega)$ we have

$$\liminf_{m\to\infty}\int_\Omega\rho(w_m)w_m\ge\int_\Omega\rho(\overline{w})\overline{w}. \tag{B.7}$$

Using (B.6) and (B.7) together with (B.5) yields

$$\int_\Omega \Lambda\nabla\nu(w_m)\cdot\nabla\nu(w_m) \to \int_\Omega \Lambda\nabla\nu(\overline{w})\cdot\nabla\nu(\overline{w}) \quad \text{and} \quad \int_\Omega \rho(w_m)w_m \to \int_\Omega \rho(\overline{w})\overline{w}.$$

The first relation classically shows that $\nabla\nu(w_m) \to \nabla\nu(\overline{w})$ strongly in $L^2(\Omega)$. Using the second relation and a weak/strong convergence argument on $\mu(w_m)w_m$ we infer that

$$\int_\Omega w_m^2 = \int_\Omega \rho(w_m)w_m + \mu(w_m)w_m \to \int_\Omega \rho(\overline{w})\overline{w} + \mu(\overline{w})\overline{w} = \int_\Omega \overline{w}^2,$$

which gives the strong convergence in $L^2(\Omega)$ of $\overline{w}$. ∎

Remark B.3 (about the assumption $F = 0$). If $F \neq 0$, then an additional term $\int_\Omega F\cdot\nabla w_m$ appears in the sequence of inequalities (B.4), and this term cannot be estimated since no a priori bound is expected on w_m in $H_0^1(\Omega)$.

REFERENCES

[1] G. Amiez and P.-A. Gremaud, *Error estimates for Euler forward scheme related to two-phase Stefan problems*, RAIRO Modél. Math. Anal. Numér., 26 (1992), pp. 365–383, https://doi.org/10.1051/m2an/1992260203651.

[2] A. B. Andreev, V. A. Kascieva, and M. Vanmaele, *Some results in lumped mass finite-element approximation of eigenvalue problems using numerical quadrature formulas*, J. Comput. Appl. Math., 43 (1992), pp. 291–311, https://doi.org/10.1016/0377-0427(92)90016-Q.

[3] B. Andreianov, M. Bendahmane, K. H. Karlsen, and S. Ouaro, *Well-posedness results for triply nonlinear degenerate parabolic equations*, J. Differential Equations, 247 (2009), pp. 277–302, https://doi.org/10.1016/j.jde.2009.03.001.

[4] C. Bi and V. Ginting, *Global superconvergence and a posteriori error estimates of the finite element method for second-order quasilinear elliptic problems*, J. Comput. Appl. Math., 260 (2014), pp. 78–90, https://doi.org/10.1016/j.cam.2013.09.042.

[5] C. Cancès and C. Guichard, *Numerical analysis of a robust free energy diminishing finite volume scheme for parabolic equations with gradient structure*, Found. Comput. Math., 17 (2017), pp. 1525–1584, https://doi.org/10.1007/s10208-016-9328-6.

[6] C. Cancés, F. Nabet, and M. Vohralik, *Convergence and A Posteriori Error Analysis for Energy-stable Finite Element Approximations of Degenerate Parabolic Equations*, https://hal.archives-ouvertes.fr/hal-01894884, 2018.

[7] J. Carrillo, *Entropy solutions for nonlinear degenerate problems*, Arch. Ration. Mech. Anal., 147 (1999), pp. 269–361.

[8] Z. Chen, *Expanded mixed finite element methods for quasilinear second order elliptic problems.* II, RAIRO Modél. Math. Anal. Numér., 32 (1998), pp. 501–520, https://doi.org/10.1051/m2an/1998320405011.

[9] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, Stud. Math. Appl. 4, North-Holland, Amsterdam, 1978.

[10] G. Cohen, P. Joly, J. E. Roberts, and N. Tordjman, *Higher order triangular finite elements with mass lumping for the wave equation*, SIAM J. Numer. Anal., 38 (2001), pp. 2047–2078, https://doi.org/10.1137/S0036142997329554.

[11] K. Deimling, *Nonlinear Functional Analysis*, Springer-Verlag, Berlin, 1985.

[12] D. A. Di Pietro and J. Droniou, *A hybrid high-order method for Leray–Lions elliptic equations on general meshes*, Math. Comp., 86 (2017), pp. 2159–2191, https://doi.org/10.1090/mcom/3180.

[13] J. Droniou, *Finite volume schemes for diffusion equations: Introduction to and review of modern methods*, Math. Models Methods Appl. Sci., 24 (2014), pp. 1575–1619, https://doi.org/10.1142/S0218202514400041.

[14] J. Droniou, R. Eymard, T. Gallouët, C. Guichard, and R. Herbin, *The Gradient Discretisation Method*, Math. Appl. 82, Springer, New York, 2018, https://doi.org/10.1007/978-3-319-79042-8.

[15] C. M. Elliott, *On the finite element approximation of an elliptic variational inequality arising from an implicit time discretization of the Stefan problem*, IMA J. Numer. Anal., 1 (1981), pp. 115–125, https://doi.org/10.1093/imanum/1.1.115.

[16] C. M. Elliott, *Error analysis of the enthalpy method for the Stefan problem*, IMA J. Numer. Anal., 7 (1987), pp. 61–71, https://doi.org/10.1093/imanum/7.1.61.
[17] R. Eymard, T. Gallouët, C. Guichard, R. Herbin, and R. Masson, *TP or not TP, that is the question*, Comput. Geosci., 18 (2014), pp. 285–296, https://doi.org/10.1007/s10596-013-9392-9.
[18] R. Eymard, T. Gallouët, and R. Herbin, *Error estimate for approximate solutions of a nonlinear convection-diffusion problem*, Adv. Differential Equations, 7 (2002), pp. 419–440, https://projecteuclid.org:443/euclid.ade/1356651802.
[19] R. Eymard, T. Gallouët, R. Herbin, and A. Michel, *Convergence of a finite volume scheme for nonlinear degenerate parabolic equations*, Numer. Math., 92 (2002), pp. 41–82.
[20] S. Geevers, W. A. Mulder, and J. J. W. van der Vegt, *New higher-order mass-lumped tetrahedral elements for wave propagation modelling*, SIAM J. Sci. Comput., 40 (2018), pp. A2830–A2857, https://doi.org/10.1137/18M1175549.
[21] R. Herbin and F. Hubert, *Benchmark on discretization schemes for anisotropic diffusion problems on general grids*, in Finite Volumes for Complex Applications V, ISTE, London, 2008, pp. 659–692.
[22] S. Jund and S. Salmon, *Arbitrary high-order finite element schemes and high-order mass lumping*, Int. J. Appl. Math. Comput. Sci., 17 (2007), pp. 375–393, https://doi.org/10.2478/v10006-007-0031-2.
[23] F. A. Milner, *Mixed finite element methods for quasilinear second-order elliptic problems*, Math. Comp., 44 (1985), pp. 303–320, https://doi.org/10.2307/2007954.
[24] A. Rösch and G. Wachsmuth, *Mass lumping for the optimal control of elliptic partial differential equations*, SIAM J. Numer. Anal., 55 (2017), pp. 1412–1436, https://doi.org/10.1137/16M1074473.
[25] G. Stampacchia, *Le problème de Dirichlet pour les équations elliptiques du second ordre à coefficients discontinus*, Ann. Inst. Fourier, 15 (1965), pp. 189–258.
[26] J.-P. Zeng and H.-X. Yu, *Error estimates of the lumped mass finite element method for semilinear elliptic problems*, J. Comput. Appl. Math., 236 (2012), pp. 1993–2004, https://doi.org/10.1016/j.cam.2011.11.009.