A CORRECTION TECHNIQUE FOR THE DISPERSIVE EFFECTS OF MASS LUMPING FOR TRANSPORT PROBLEMS*

JEAN-LUC GUERMOND†‡ R. PASQUETTI§

Abstract. This paper addresses the well-known dispersion effect that mass lumping induces when solving transport-like equations. A simple anti-dispersion technique based on the lumped mass matrix is proposed. The method does not require any non-trivial matrix inversion and has the same anti-dispersive effects as the consistent mass matrix. A novel quasi-lumping technique for $\mathbb{P}_2$ finite elements is introduced. Higher-order extensions of the method are also discussed.

Key words. Finite elements, Transport equation, Mass lumping, Dispersion.

AMS subject classifications. 65N35, 65D30

1. Introduction. Lumping the mass matrix is a routine procedure in the finite element community when solving the heat equation, the wave equation and the time-dependent transport equation. This technique consists of replacing the consistent mass matrix by a diagonal surrogate usually referred to as the lumped mass matrix. This process avoids having to invoke sophisticated linear algebra arguments to invert the consistent mass matrix at each time step. The mantra in the literature dedicated to mass lumping is that mass lumping produces explicit algorithms for the transport and the wave equations that are algebra-free.

The lumped mass matrix is generally obtained by using a quadrature formula instead of exact integration. It is usually believed that lumping is a benign operation since it does not affect the overall accuracy of the method provided the quadrature is accurate enough. For instance, it is known that using quadrature formulas that are exact for $\mathbb{P}_{2k-2}$ polynomials is sufficient to preserve the overall accuracy of the Galerkin method when solving the wave equation or some eigenvalue problems on simplex meshes, [1, 7, 12, 11, 20]. Although it is convenient numerically, it is well-known that lumping the mass matrix induces dispersion errors that have adverse effects when solving transport-like equations, see e.g., [5, 6, 14, 22]. The objectives of the present work are as follows:

i We propose a simple correction technique based on the lumped mass matrix that does not involve sophisticated linear algebra and that has the same anti-dispersive effects as the consistent mass matrix. Although this correction technique relies on a matrix series, we show theoretically and numerically that only considering the first term in this series is enough to correct the dominating dispersion error.

ii We introduce a novel quasi-lumping technique for $\mathbb{P}_2$ finite elements, where the new $\mathbb{P}_2$ quasi-lumped mass matrix is triangular. We show also that the proposed mass correction technique is efficient when using this $\mathbb{P}_2$ quasi-lumped mass matrix.

iii We investigate higher-order extensions of the correction method and demonstrate satisfactory results for the $\mathbb{P}_3$ approximation.

To the best of our knowledge, the correction technique and the quasi-lumping technique for $\mathbb{P}_2$ finite elements are original.

This paper is organized as follows. The anti-dispersive effects of the consistent $\mathbb{P}_1$ mass matrix on the transport equation are analyzed in §2. We focus in this section on the linear transport equation

*This material is based upon work supported in part by the National Science Foundation grants DMS–0811041 and DMS-1015984, by the Air Force Office of Scientific Research, USAF, under grant/contract number FA9550-09-1-0424, and by Award No. KUS-C1-016-04, made by King Abdullah University of Science and Technology (KAUST). Draft version, April 30, 2012

†Department of Mathematics, Texas A&M University 3368 TAMU, College Station, TX 77843, USA

‡On leave from CNRS, France.

§Lab. J.A. Dieudonné, UMR CNRS 6621, UNS, 06108 Nice, France.

in one space dimension. Most of the material therein is standard. A mass correction technique based on the lumped mass matrix is presented in §3. The method has the same algebraic complexity as when using the lumped mass matrix. It is also proved for $\mathbb{P}_1$ elements in one space dimension that using one correction term only is enough to obtain the same anti-dispersive effect as when using the consistent mass matrix. The mass correction method is further evaluated numerically in two space dimension on $\mathbb{P}_1$ finite elements in §4. A new $\mathbb{P}_2$ quasi-lumping technique is introduced in §5. To the best of our knowledge, the $\mathbb{P}_2$ quasi-lumping technique presented in §5.3 and §5.4 and the mass correction technique introduced in §3 are original. Higher-order suboptimal variants of the method are considered in §6. Conclusions are reported in §7.

2. One-dimensional heuristics. The objective of this section is to analyze in details the effects of mass lumping in one space dimension for the linear transport equation using piece-wise linear finite elements. The material herein is certainly not new, see e.g., [6, 14, 17, 22], but it is useful to comprehend the rest of the paper. Let us consider the following one-dimensional transport equation in the domain $\Omega = (a, b)$

$$\partial_t \mathsf{u} + \beta \partial_x \mathsf{u} = 0, \quad \mathsf{u}(x, 0) = \mathsf{u}_0(x), \quad (x, t) \in (a, b) \times \mathbb{R}_+, \tag{2.1}$$

equipped with periodic boundary conditions. The velocity field β is assumed to be constant to simplify the presentation.

2.1. Galerkin linear approximation. Let us partition $\Omega = (a, b)$ into N intervals $[x_i, x_{i+1}]$, $i = 0, \ldots, N-1$. Let $h_{i+\frac{1}{2}} := |x_{i+1} - x_i|$ be the diameter of the cell $[x_i, x_{i+1}]$. We introduce the family $\{\psi_0, \ldots, \psi_N\}$ composed of continuous and piecewise linear Lagrange functions associated with the nodes $\{x_0, \ldots, x_N\}$, and we define the $\mathbb{P}_1$ finite element space

$$X_h = \{v \in \mathcal{C}^0_{\#}(\overline{\Omega}; R),\ v|_{[x_i, x_{i+1}]} \in \mathbb{P}_1,\ i = 0, \ldots, N-1\} = \operatorname{span}(\psi_0, \ldots, \psi_N), \tag{2.2}$$

where $\mathcal{C}^0_{\#}(\overline{\Omega}; R)$ denotes the space of the real-valued functions that are periodic and continuous over $\overline{\Omega}$. Let u_0 be a reasonable approximation of u_0, say the Lagrange interpolate or L^2-projection thereof. An approximate solution to (2.1) is constructed by means of the Galerkin technique. We seek $u \in \mathcal{C}^1((0, T); X_h)$ so that $u(0) = u_0$ and

$$\int_\Omega (\partial_t u + \beta \partial_x u) v \, \mathrm{d}x = 0, \qquad \forall v \in X_h. \tag{2.3}$$

The approximate solution $u(x, t)$ is expanded with respect to the basis $\{\psi_0, \ldots, \psi_N\}$ as follows: $u(x, t) = \sum_{j=0}^N u_j(t) \psi_j(x)$. A system of ordinary differential equations is obtained by testing (2.3) with the members of the basis $\{\psi_0, \ldots, \psi_N\}$.

Upon testing (2.3) with ψ_i, $i = 0, \ldots, N$, the term involving the time derivative gives

$$\int_\Omega \partial_t u(x, t) \psi_i(x) \, \mathrm{d}x = \sum_{j=0}^N M_{ij} \partial_t u_j(t), \tag{2.4}$$

where the coefficients of the so-called mass matrix are

$$M_{ij} := \int_{x_{i-1}}^{x_{i+1}} \psi_i(x) \psi_j(x) \, \mathrm{d}x = \begin{cases} \frac{1}{6} h_{i \pm \frac{1}{2}} & \text{if } j = i \pm 1 \\ \frac{1}{3}(h_{i-\frac{1}{2}} + h_{i+\frac{1}{2}}) & \text{if } j = i \\ 0 & \text{otherwise} \end{cases} \tag{2.5}$$

with the convention that $h_{-\frac{1}{2}} = h_{N-\frac{1}{2}}$ and $h_{N+\frac{1}{2}} = h_{\frac{1}{2}}$. The transport term in (2.3) is handled as follows:

$$\begin{aligned}\int_\Omega \psi_i(x)\beta\partial_x u(x,t)\,\mathrm{d}x &= -\int_{x_i-1}^{x_{i+1}} \beta u(x,t)\partial_x\psi_i(x)\,\mathrm{d}x \\ &= \frac{\beta}{2}(u_{i+1}(t)+u_i(t)) - \frac{\beta}{2}(u_i(t)+u_{i-1}(t)),\end{aligned} \tag{2.6}$$

giving

$$\int_\Omega \psi_i(x)\beta\partial_x u(x,t)\,\mathrm{d}x = \beta\frac{1}{2}(u_{i+1}(t)-u_{i-1}(t)), \tag{2.7}$$

with the convention $u_{-1}(t) = u_{N-1}(t)$ and $u_{N+1}(t) = u_1(t)$.

Recalling that we are looking for a periodic solution, the above computation shows that the vector $(u_0(t),\dots,u_{N-1}(t))^T \in \mathbb{R}^N$ solves the following system of ordinary differential equations:

$$\sum_{j=i-1}^{i+1} M_{ij}\partial_t u_j(t) = -\beta\frac{1}{2}(u_{i+1}(t)-u_{i-1}(t)), \quad 0 \le i,j < N, \tag{2.8}$$

where $u_N(t) = u_0(t)$ and $u_{-1}(t) = u_{N-1}(t)$. The above system can be written in matrix form as follows:

$$M\partial_t U(t) = F(U(t)), \tag{2.9}$$

with $U(t) := (u_0(t),\dots,u_{N-1}(t))^T$, and the entries of F are defined by $F_i(U) := -\beta\frac{1}{2}(u_{i+1}-u_{i-1})$, $0 \le i < N$, and where M is the consistent mass matrix defined in (2.5) taking into account the periodicity in the first and last lines.

2.2. Dispersion and mass lumping. It is common in the literature to approximate (2.9) in time by means of explicit time stepping. To avoid having to solve linear systems involving the mass matrix at each time step, it also common to simplify (2.8) by lumping the mass matrix. Mass lumping can be shown in one space dimension to be equivalent to approximate the consistent mass matrix by using the following trapezoidal quadrature rule:

$$\int_r^s f(x)\,\mathrm{d}x \approx (s-r)\frac{1}{2}(f(r)+f(s)). \tag{2.10}$$

This quadrature is exact for linear polynomials. Using this quadrature, the mass matrix coefficients can be approximated as follows:

$$\int_{x_{i-1}}^{x_{i+1}} \psi_i(x)\psi_j(x)\,\mathrm{d}x \approx \frac{1}{2}(h_{i-\frac{1}{2}}+h_{i+\frac{1}{2}})\delta_{ij} =: \overline{M}_{ij}, \tag{2.11}$$

where δ_{ij} is the Kronecker symbol. The so-called lumped mass matrix $\overline{M}$ thus computed is diagonal. Upon denoting $\overline{h}_i := \frac{1}{2}(h_{i-\frac{1}{2}}+h_{i+\frac{1}{2}})$ and replacing the consistent mass matrix by the lumped mass matrix, we obtain a new approximate form of transport equation as follows:

$$\partial_t\tilde{u}_i(t) + \beta\frac{\tilde{u}_{i+1}-\tilde{u}_{i-1}}{2\overline{h}_i} = 0. \tag{2.12}$$

The approximation thus constructed is second-order accurate. More precisely, the consistency error of (2.12) is characterized by the following

PROPOSITION 2.1. *Provided the mesh is uniform, of mesh size h, the dominating term in the consistency error of* (2.12) *at the grid points $\{x_i\}_{0\le i\le N}$ is dispersive and is equal to $\beta\frac{h^2}{6}\partial_{xxx}\mathsf{u}(x_i,t)$.*

Proof. Using $x_{i\pm1}=x_i\pm h$ and upon using the Taylor expansion $\mathsf{u}(x_i\pm h,t)=\mathsf{u}(x_i)\pm h\partial_x\mathsf{u}(x,t)+\frac{1}{2}h^2\partial_{xx}\mathsf{u}(x_i,t)\pm\frac{1}{6}h^3\partial_{xxx}\mathsf{u}(x_i,t)+\frac{1}{24}h^4\partial_{xxxx}\mathsf{u}(x_i,t)+\mathcal{O}(h^5)$, we infer that

$$\partial_t\mathsf{u}(x_i,t)+\beta\frac{\mathsf{u}(x_{i+1},t)-\mathsf{u}(x_{i-1},t)}{2h}=(\partial_t\mathsf{u}+\beta\partial_x\mathsf{u})(x_i,t)+\beta\frac{h^2}{6}\partial_{xxx}\mathsf{u}(x_i,t)+\mathcal{O}(h^4).$$

This proves the statement of the proposition and proves also in passing that the equivalent limit equation is

$$\partial_t\tilde{u}+\beta\partial_x\tilde{u}+\beta\frac{h^2}{6}\partial_{xxx}\tilde{u}=0, \tag{2.13}$$

which is clearly dispersive. □

Remark 2.1. The key observation here is that the consistency error induced by mass lumping is second-order and dispersive.

Remark 2.2. The approximation (2.12) is exactly what a finite volume and a second-order finite difference approximation would give on a uniform mesh.

2.3. Anti-dispersive effect of the mass matrix. Let us now consider (2.8) where the mass matrix is not approximated, and let us redo the consistency analysis for this discrete system.

PROPOSITION 2.2. *Provided the mesh is uniform, of mesh size h, the dominating term in the consistency error of* (2.8) *at the grid points $\{x_i\}_{0\le i\le N}$ is equal to $\beta\frac{h^4}{180}\partial_{xxxxx}\mathsf{u}(x_i,t)$.*

Proof. Using the definition of the mass matrix (2.5), the discrete system (2.8) can be re-written as follows:

$$\frac{1}{h}\sum_{j=i-1}^{i+1}M_{ij}\partial_t u_j=\partial_t u_i+\frac{1}{6}(\partial_t u_{i-1}-2\partial_t u_i+\partial_t u_{i+1}).$$

Using Taylor expansions at x_i we obtain

$$\begin{aligned}\frac{1}{h}\sum_{j=i-1}^{i+1}M_{ij}\partial_t\mathsf{u}(x_j,t)&=\partial_t\mathsf{u}(x_i,t)+\frac{h^2}{6}\partial_{txx}\mathsf{u}(x_i,t)+\frac{h^4}{72}\partial_{txxxx}\mathsf{u}(x_i,t)+\mathcal{O}(h^6)\\&=\partial_t\mathsf{u}(x_i,t)-\beta\frac{h^2}{6}\partial_{xxx}\mathsf{u}(x_i,t)-\beta\frac{h^4}{72}\partial_{xxxxx}\mathsf{u}(x_i,t)+\mathcal{O}(h^6).\end{aligned}$$

By proceeding again as in the proof of Proposition 2.1 and using $\mathsf{u}(x_i\pm h,t)=\mathsf{u}(x_i)\pm h\partial_x\mathsf{u}(x,t)+\frac{1}{2}h^2\partial_{xx}\mathsf{u}(x_i,t)\pm\frac{1}{6}h^3\partial_{xxx}\mathsf{u}(x_i,t)+\frac{1}{24}h^4\partial_{xxxx}\mathsf{u}(x_i,t)\pm\frac{1}{120}h^5\partial_{xxxxx}\mathsf{u}(x_i,t)+\mathcal{O}(h^6)$, we infer that

$$\begin{aligned}&\frac{1}{h}\sum_{j=i-1}^{i+1}M_{ij}\partial_t\mathsf{u}(x_j,t)+\beta\frac{\mathsf{u}(x_{i+1},t)-\mathsf{u}(x_{i-1},t)}{2h}\\&\qquad=\partial_t\mathsf{u}(x_i,t)+\beta\partial_x\mathsf{u}(x_i,t)-\beta\frac{1}{180}h^4\partial_{xxxxx}\mathsf{u}(x_i,t)+\mathcal{O}(h^6),\end{aligned} \tag{2.14}$$

thereby proving the statement. This also prove in passing that the equivalent limit equation is

$$\partial_t\tilde{u}+\beta\partial_x\tilde{u}-\beta\frac{h^4}{180}\partial_{xxxxx}\tilde{u}=0, \tag{2.15}$$

which is again dispersive. Note however that the dispersion error is now forth-order whereas it is second-order in (2.13). □

When comparing Proposition 2.1 and Proposition 2.2 we now understand that accounting properly for the mass matrix limits the dispersion error of the centered approximation.

Remark 2.3. It is remarkable that the result of Proposition 2.2 holds in higher-space dimension. For instance, it is shown in the appendix A that the result holds on quadrangular grids with $\mathbb{Q}_1$ elements, independently of the transport direction.

Remark 2.4. The consistent mass matrix does not have anti-dispersive effect on the wave equation $\partial_{tt}u - c^2\partial_{xx}u = 0$, however, a simple computation as above shows that using $\frac{1}{2}(\overline{M} + M)$ is the right combination to do the job with $\mathbb{P}_1$ finite elements on uniform grids. See [5] and references therein for other details.

2.4. Fourier analysis. Fourier analysis is useful to evaluate numerical dispersion, and the purpose of this section is to revisit the statements of Proposition 2.1 and Proposition 2.2 from the Fourier analysis perspective. Let k be a real number and assume that $\mathsf{u}_0(x) = \alpha e^{\mathrm{i}kx}$, $\mathrm{i}^2 = -1$, then the exact solution to (2.1) is $\mathsf{u}(x,t) = \alpha e^{\mathrm{i}k(x-\beta t)}$. Let us now compare this solution to what (2.8) and (2.12) give, respectively.

PROPOSITION 2.3. *If the initial data to* (2.8) *and* (2.12) *is* $\{\alpha e^{\mathrm{i}kx_i}\}_{0\le i\le N}$*, the solution to* (2.8) *and* (2.12) *is* $\{\alpha e^{\mathrm{i}k(x_i-c_1(k)t)}\}_{0\le i\le N}$ *and* $\{\alpha e^{\mathrm{i}k(x_i-c_2(k)t)}\}_{0\le i\le N}$*, respectively, where*

$$c_1(k) = 3\beta\frac{\sin(kh)}{kh(2+\cos(kh))}, \qquad c_2(k) = \beta\frac{\sin(kh)}{kh}. \tag{2.16}$$

Proof. This result is not new (see e.g., [14, p. 136]), but we give the proof for the sake of completeness. Let us assume that the solution to (2.8) is given by $\{\alpha e^{\mathrm{i}k(x_i-c_1(k)t)}\}_{0\le i\le N}$, where $c_1(k)$ is yet to be determined. Then by inserting this expression into the following equivalent form of (2.8)

$$\partial_t u_i + \frac{1}{6}(\partial_t u_{i-1} - 2\partial_t u_i + \partial_t u_{i+1}) + \beta\frac{u_{i+1}-u_{i-1}}{2h} = 0,$$

we infer that the following must hold:

$$\begin{aligned} 0 &= -\mathrm{i}kc_1(k)(1 + \frac{1}{6}(e^{\mathrm{i}kh} - 2 + e^{-\mathrm{i}kh})) + \beta\frac{1}{2h}(e^{\mathrm{i}kh} - e^{-\mathrm{i}kh})) \\ &= -\mathrm{i}kc_1(k)\frac{1}{3}(2+\cos(kh)) + \mathrm{i}\beta\frac{1}{h}\sin(kh), \end{aligned}$$

which is equivalent to the expression of $c_1(k)$ in (2.16). The same argument gives $c_2(k)$. ∎

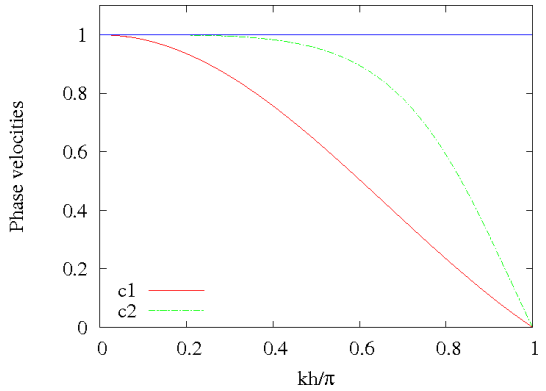


FIGURE 2.1. *Phase velocities*

The graph of the phase velocities $c_1(k)/\beta$ and $c_2(k)/\beta$ for $k \in [0, \pi/h]$ are shown in Figure 2.1. This figure shows that phase velocity $c_2(k)/\beta$ is closer to the perfect value 1 than $c_1(k)/\beta$, i.e., (2.8)

transports the high frequencies better than (2.12), thereby confirming again that the consistent mass matrix has anti-dispersive properties. The anti-dispersive effect of the consistent mass matrix is illustrated numerically on the one-dimensional linear transport equation in the Appendix B.1.

3. Mass matrix corrections. Since solving the mass matrix at each time step may be perceived as a drawback of the finite element method, we describe in this section a technique that has the same anti-dispersive effect as the consistent mass matrix but whose complexity is nearly the same as when using the lumped mass matrix.

3.1. An abstract result. The generic form of the system (2.9) can be re-written as follows:

$$\partial_t U + M^{-1} F U = 0, \tag{3.1}$$

and our goal in this section is to approximate M^{-1} efficiently. To this end we set $M = \overline{M} + M - \overline{M}$, where we assume that $\overline{M}$ is easy to invert, e.g., $\overline{M}$ can be the lumped mass matrix. At this point one may factorize $\overline{M}$ on the left, right, or symmetrically as follows:

$$M = \overline{M}(I + \overline{M}^{-1}(M - \overline{M})), \tag{3.2}$$

$$M = (I + (M - \overline{M})\overline{M}^{-1})\overline{M}, \tag{3.3}$$

$$M = \overline{M}^{1/2}(I + \overline{M}^{-1/2}(M - \overline{M})\overline{M}^{-1/2})\overline{M}^{1/2}. \tag{3.4}$$

Note that the symmetric factorization is legitimate provided $\overline{M}$ is symmetric and non-negative, and it of interest only if $\overline{M}$ is diagonal, since $\overline{M}^{1/2}$ is easy to compute in this case. Depending on factorization which is chosen we introduce the following matrices:

$$A_r = \overline{M}^{-1}(\overline{M} - M), \quad A_s = \overline{M}^{-1/2}(\overline{M} - M)\overline{M}^{-1/2}, \quad \text{or} \quad A_l = (\overline{M} - M)\overline{M}^{-1}. \tag{3.5}$$

We then obtain the following three possible representations for M^{-1}:

$$M^{-1} = (I + A_r + A_r^2 + \cdots)\overline{M}^{-1}, \tag{3.6}$$

$$M^{-1} = \overline{M}^{-1/2}(I + A_s + A_s^2 + \cdots)\overline{M}^{-1/2}, \tag{3.7}$$

$$M^{-1} = \overline{M}^{-1}(I + A_l + A_l^2 + \cdots). \tag{3.8}$$

Of course these representations are valid only if the series are convergent, which is the case if and only if the spectral radius of A is less than 1.

LEMMA 3.1. *The spectra of A_r, A_s (provided $\overline{M}$ is symmetric and non-negative) and A_l are identical.*

Proof. That the spectra of A_r and A_l are identical is the consequence of the standard result that the spectra of CD and DC are identical for all square matrices C, D. Let us now assume that $\overline{M}$ is symmetric and non-negative, then A_s is symmetric, thus diagonalizable. Let Λ_s and V_s be the matrices of the eigenvalues and eigenvectors of A_s, respectively. Then using the definition $A_s V_s = V_s \Lambda_s$ we infer that

$$\overline{M}^{-1/2} A_s \overline{M}^{1/2} \overline{M}^{-1/2} V_s = \overline{M}^{-1/2} V_s \Lambda_s$$

which in turn implies $A_r \overline{M}^{-1/2} V_s = \overline{M}^{-1/2} V_s \Lambda_s$, thereby proving that the spectra of A_r and A_s are identical. ∎

One of the key results of this paper is that the $\mathbb{P}_1$ lumped mass matrix in one space dimension and in higher dimensions is such that the above series are convergent, and that using only one term in the series, i.e., $1 + A$, is enough to compensate exactly the dominating dispersive effects of mass lumping.

3.2. One-dimensional argumentation. We show in this section that using $(1+A)\overline{M}^{-1}$ is enough to correct the dispersive effects of mass lumping in one space dimension with $\mathbb{P}_1$ elements. Note that in one space dimension and with $\mathbb{P}_1$ finite elements $A_r = A_s = A_l$ when the mesh is uniform.

PROPOSITION 3.2. *Provided the mesh is uniform, of meshsize h, the dominating term of the consistency error at the grid points $\{x_i\}_{0\le i\le N}$ is $\mathcal{O}(h^4)$ when using only one correction in* (3.6).

Proof. Observe that $\overline{M} = hI$ and $A_r = I - h^{-1}M$. This implies that $(I+A_r)\overline{M}^{-1} = h^{-1}(2I - h^{-1}M)$. The approximation equation is

$$\partial_t u_i + \frac{\beta}{2h}\sum_{j=i-1}^{i+1}\left[2\delta_{ij} - h^{-1}M_{ij}\right](u_{j+1}-u_{j-1}) = 0,$$

giving

$$\partial_t u_i + \frac{\beta}{2h}\left(\frac{1}{6}(u_{i-2}-u_{i+2}) + \frac{4}{3}(u_{i+1}-u_{i-1})\right) = 0.$$

Using Taylor expansions at x_i, we obtain that

$$\partial_t u(x_i,t) + \frac{\beta}{2h}\left(\frac{1}{6}(u(x_{i-2},t)-u(x_{i+2},t)) + \frac{4}{3}(u(x_{i+1},t)-u(x_{i-1},t))\right) \\ = \partial_t u(x_i,t) + \beta\partial_x u(x_i,t) + \mathcal{O}(h^4),$$

which completes the proof. Note that this result is similar to what has been obtained in (2.14) when using the consistent mass matrix. □

The above result is illustrated in the Appendix B.2 in one space dimension.

4. Application to $\mathbb{P}_1$ finite elements. We show in this section that the observations made in one space dimension generalize to two space dimensions. We restrict ourselves to two space dimensions for the sake of simplicity, but most of what is said hereafter generalizes to higher space dimensions.

4.1. The lumped $\mathbb{P}_1$ mass matrix. Let Ω be a two-dimensional polygonal domain and consider an affine finite element mesh $\mathcal{T}_h$ of Ω composed of simplices. Consider a cell in the mesh, $K \in \mathcal{T}_h$, and let $\boldsymbol{S}_1, \boldsymbol{S}_2, \boldsymbol{S}_3$ be the three vertices of K and ϕ_1, ϕ_2, ϕ_3 be the associated local nodal shape functions. The local mass matrix M^K associated to K is defined to be

$$M^K_{ij} := \int_K \phi_i(\boldsymbol{x})\phi_j(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = |K|\Phi_i^T W \Phi_j, \tag{4.1}$$

where $\{\Phi_1, \Phi_2, \Phi_3\}$ is the canonical basis of $\mathbb{R}^3$ and the matrix W is given by

$$W = \begin{bmatrix} \frac{1}{6} & \frac{1}{12} & \frac{1}{12} \\ \frac{1}{12} & \frac{1}{6} & \frac{1}{12} \\ \frac{1}{12} & \frac{1}{12} & \frac{1}{6} \end{bmatrix}. \tag{4.2}$$

Once M^K is computed for all $K \in \mathcal{T}_h$, the mass matrix M is obtained by the so-called assembling procedure.

The standard mass lumping process advocated in the literature consists of using the following approximate quadrature rule:

$$\int_K f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = |K|(\frac{1}{3}f(\boldsymbol{S}_1) + \frac{1}{3}f(\boldsymbol{S}_2) + \frac{1}{3}f(\boldsymbol{S}_3)), \qquad \forall f \in \mathbb{P}_1, \tag{4.3}$$

to approximate $\int_K \phi_i(\boldsymbol{x})\phi_j(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$. The local lumped matrix $\overline{M}^K$ obtained by this technique is

$$\overline{M}^K_{ij} := |K|\Phi_i^T\overline{W}\Phi_j, \tag{4.4}$$

where the matrix $\overline{W}$, computed by means of the above quadrature rule is

$$\overline{W} = \begin{bmatrix} \frac{1}{3} & 0 & 0 \\ 0 & \frac{1}{3} & 0 \\ 0 & 0 & \frac{1}{3} \end{bmatrix}. \tag{4.5}$$

Of course, since $\overline{W}$ is diagonal, $\overline{M}^K$ is diagonal and the assembled matrix $\overline{M}$ is also diagonal.

The popularity of the lumped $\mathbb{P}_1$ mass matrix, $\overline{M}$, comes from the fact that it can be shown to be a satisfactory alternative of the consistent mass matrix, M, in terms of approximation and convergence rate, at least for the heat and the wave equation, [1, 7, 20]. That the matrix $\overline{W}$ is indeed a good approximation of W is also expressed in the following

PROPOSITION 4.1. *The three eigenvalues of* $\overline{W}^{-1}(\overline{W} - W)$ *are* $(0, \frac{3}{4}, \frac{3}{4})$.

4.2. Numerical illustrations. We illustrate the efficiency of the correction algorithm in this section. We show in particular that using one term in the correction series is sufficient to remove the dominating dispersion error. Let us consider the scalar transport equation

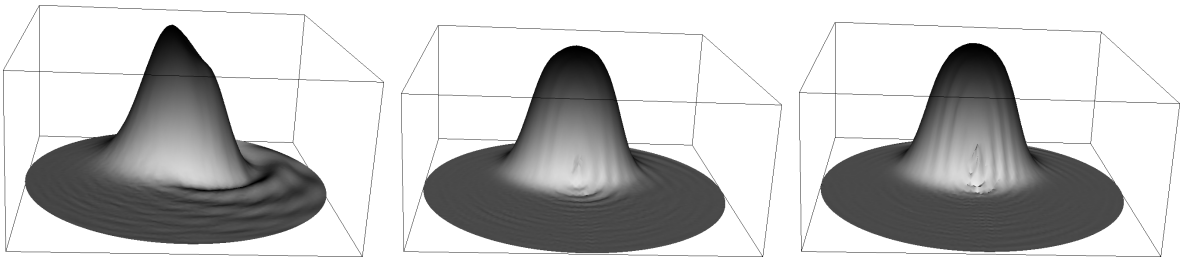
$$\partial_t \mathsf{u} + \boldsymbol{\beta}\cdot\nabla\mathsf{u} = 0, \quad \mathsf{u}(\boldsymbol{x}, 0) = \mathsf{u}_0(\boldsymbol{x}), \tag{4.6}$$

in the unit disk $\Omega = \{(x, y) \in \mathbb{R}^2, \sqrt{x^2 + y^2} < 1\}$. The velocity field is a solid rotation of angular velocity 2π, i.e., $\boldsymbol{\beta} = 2\pi(-y, x)$. The initial field u_0 is defined by

$$\mathsf{u}_0(\boldsymbol{x}) = \frac{1}{2}\left(1 - \tanh\left(\frac{(x - x_0)^2 + y^2}{a^2} - 1\right)\right), \quad x_0 = 0.4, a = 0.3. \tag{4.7}$$

We solve (4.6) with the Galerkin method with $\mathbb{P}_1$ finite elements on a mesh composed of 6293 $\mathbb{P}_1$ nodes. The time stepping is done with the standard RK4 method (RK3 and RK4 techniques are known to be stable under a CFL condition for the linear transport equation, see e.g., [16]); this ensures that the error induced by the time approximation is small compared to the spatial error. The solution is computed at $T = 2$, i.e., after two revolutions.

The results are shown in Figure 4.1. The solution obtained with mass lumping is shown in 4.1(a). The dispersive effect is clear and needs not be commented. We show in Figure 4.1(b) and



(a) 1 (b) $1 + A$ (c) $1 + A + A^2 + A^3 + A^4$

FIGURE 4.1. *Mass matrix corrections on a 2D Delaunay triangulation,* $\mathbb{P}_1$ *finite elements,* $h \approx 0.025$ *(6293* $\mathbb{P}_1$ *nodes),* $T = 2$.

4.1(c) the solutions obtained by replacing the inverse of the lumped mass matrix by $(1 + A)\overline{M}^{-1}$

and $(1 + A + A^2 + A^3 + A^4)\overline{M}^{-1}$, respectively, where $A := \overline{M}^{-1}(\overline{M} - M)$. The effect of applying only one correction to the lumped mass matrix is spectacular, the dispersive waves have completely disappeared.

h	Consist. Mass	4 corrections	1 correction	0 correction
0.1000	9.653E-2	1.003E-1	1.488E-1	4.444E-1
0.0500	1.990E-2	1.999E-2	3.191E-2	1.827E-1
0.0250	5.790E-3	5.706E-3	6.460E-3	6.369E-2
0.0125	2.120E-3	2.046E-3	1.186E-3	1.747E-2
0.0100	1.644E-3	1.576E-3	7.644E-4	1.124E-2

TABLE 4.1

L^2-norm of error, $\mathbb{P}_1$ finite elements, $T = 1$. Computations done with the consistent mass matrix, the lumped mass matrix corrected four times, and with the lumped mass matrix with no correction.

We have verified in tests not reported here that the solution obtained with four corrections is visually indistinguishable from that obtained by inverting exactly the consistent mass matrix. To make this statement more precise, we solve the above linear transport problem on various grids ($h = 0.1, 0.05, 0.025, 0.0125, 0.01$) and we compute the L^2-norm of the error at $T = 1$. The convergence results are reported in Table 4.1. For all practical purposes, the errors obtained by using the consistent mass matrix and by applying four corrections to the lumped mass matrix are identical. This series of tests clearly shows that correcting the lumped mass matrix four times is enough to obtain results that cannot be distinguished from those computed with the consistent mass matrix.

Let us finish this section by justifying the convergence of the Neumann expansion in (3.6). This is done by evaluating the spectral radius of the mass correction.

PROPOSITION 4.2. *The spectral radius of $A := \overline{M}^{-1}(\overline{M} - M))$ is less than $\frac{3}{4}$.*

Proof. Let (Y, λ) be an eigenpair of $\overline{M}^{-1}(\overline{M} - M))$, i.e., $Y^T(\overline{M} - M)Y = \lambda Y^T \overline{M} Y$. Then, using the fact that the mesh is affine, we infer

$$|Y^T(\overline{M} - M)Y| = |\sum_{K\in\mathcal{T}_h} Y_K^T(\overline{M}^K - M^K)Y_K| \le \sum_{K\in\mathcal{T}_h} |K|\|Y_K\|\|\overline{W} - W\|\|Y_K\|,$$

where Y_K is the vector of the three components of Y that are associated to the vertices of the triangle K and where $\|\cdot\|$ denotes the Euclidian norm. Owing to Proposition 4.1 we infer that $\|\overline{W} - W\| \le \frac{1}{4}$, which in turns implies

$$|Y^T(\overline{M} - M)Y| \le \frac{3}{4}\sum_{K\in\mathcal{T}_h} \frac{1}{3}|K|\|Y_K\|^2 = \frac{3}{4}\sum_{K\in\mathcal{T}_h} |K|Y_K^T\overline{M}^K Y_K = \frac{3}{4}Y^T\overline{M}Y.$$

In conclusion $|Y^T(\overline{M} - M)Y| = |\lambda|Y^T\overline{M}Y \le \frac{3}{4}Y^T\overline{M}Y$, which concludes the proof. ∎

Table 4.2 shows the largest eigenvalue of $A := \overline{M}^{-1}(\overline{M} - M))$ on the five Delaunay grids used in the convergence tests above. This table confirm that the spectral radius of A is indeed uniformly bounded by 0.75.

h	0.1	0.2	0.025	0.0125	0.01
$\rho(A)$	0.7428	0.7472	0.7488	0.7496	0.7497

TABLE 4.2

Spectral radius of $A := \overline{M}^{-1}(\overline{M} - M))$ vs. h

5. $\mathbb{P}_2$ finite elements. We now extend the above considerations to higher-order finite-elements. We particularly focus our attention in this section on the $\mathbb{P}_2$ mass matrix.

5.1. Terminology. The terminology "mass lumping" comes from the operation that consists of replacing the consistent mass matrix by a diagonal matrix whose entry in row i is the sum of all the entries of the consistent mass matrix in row i. When using Lagrange finite elements, this operation is equivalent to choosing an approximate quadrature based on the interpolation points to compute the diagonal surrogate. This statement is made more precise in the following

PROPOSITION 5.1. *Mass lumping and using the interpolation points as quadrature points to approximate the mass matrix give the same diagonal matrix.*

Proof. This result is standard, but we give the proof for completeness. Clearly, the proposition holds for the assembled matrices M and $\overline{M}$ if it holds for the local matrices M^K and $\overline{M}^K$. Let us then focus on the local matrices.

Let $\widehat{K}$ be the reference finite element and let $\widehat{\boldsymbol{A}}_1, \dots \widehat{\boldsymbol{A}}_L$ be the Lagrange nodes on $\widehat{K}$ and $\widehat{\phi}_1, \dots \widehat{\phi}_L$ be the corresponding nodal shape functions. The following quadrature rule holds

$$\int_{\widehat{K}} \widehat{f}(\widehat{\boldsymbol{x}})\,\mathrm{d}\widehat{\boldsymbol{x}} = |\widehat{K}| \sum_{i=1}^{L} \omega_i \widehat{f}(\widehat{\boldsymbol{A}}_i) := \mathcal{I}_{\widehat{K}}(\widehat{f}), \qquad \forall \widehat{f} \in \mathrm{span}(\widehat{\phi}_1, \dots, \widehat{\phi}_L), \tag{5.1}$$

provided the weights are defined as follows:

$$\omega_i = \frac{1}{|\widehat{K}|} \int_{\widehat{K}} \widehat{\phi}_i(\widehat{\boldsymbol{x}})\,\mathrm{d}\widehat{\boldsymbol{x}}, \qquad \forall i \in \{1, \dots, L\}.$$

Let M^K be the local mass matrix associated to element K; then, the sum of the entries of M^K in row i is computed as follows:

$$\sum_{l=1}^{L} M_{il}^K = \sum_{l=1}^{L} \int_K \phi_i(\boldsymbol{x})\phi_l(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int_K \phi_i(\boldsymbol{x}) \sum_{l=1}^{L} \phi_l(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \frac{|K|}{|\widehat{K}|} \int_{\widehat{K}} \widehat{\phi}_i(\widehat{\boldsymbol{x}})\,\mathrm{d}\widehat{\boldsymbol{x}} = |K|\omega_i.$$

where we used $\sum_{l=1}^{L} \phi_l(\boldsymbol{x}) = 1$. Now let us use the quadrature (5.1) defined above to approximate the entries of M^K; in other words, with obvious notations let us evaluate $\mathcal{I}_K(\phi_i\phi_j)$:

$$\mathcal{I}_K(\phi_i\phi_j) = \frac{|K|}{|\widehat{K}|} \mathcal{I}_{\widehat{K}}(\widehat{\phi}_i\widehat{\phi}_j) = \delta_{ij}|K|\omega_i. \tag{5.2}$$

In conclusion we have $\delta_{ij} \sum_{j=1}^{L} M_{ij}^K = \mathcal{I}_K(\phi_i\phi_j)$ for all element $K \in \mathcal{T}_h$, which in turns implies that the result holds also for assembled matrices M and $\overline{M}$. This concludes the proof. ∎

In the remainder of this paper we are going to use approximate quadratures to construct approximations of the consistent mass matrix. Some of these quadratures do not satisfy (5.1) and consequently the techniques that we are going to introduce are not mass lumping in the sense of Proposition 5.1. We are nevertheless going to make an abuse of language by referring to these alternative approaches as quasi-lumping.

5.2. The $\tilde{\mathbb{P}}_k$ construction. The above mass lumping technique is known to work properly only for the $\mathbb{P}_1$ finite element in the class of the simplicial finite elements with the Lagrange nodes equally distributed on a uniform lattice on the reference elements, see §4. For instance mass lumping fails for $\mathbb{P}_2$ finite elements in two space dimensions. Although the argumentation is standard, let us recall why mass lumping fails for the $\mathbb{P}_2$ finite elements. H^1-conformity and elementary symmetry considerations impose that there is a unique choice for the Lagrange nodes of the $\mathbb{P}_2$ finite element;

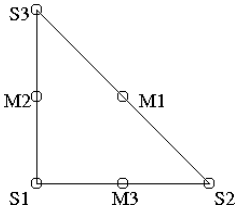


FIGURE 5.1. $\mathbb{P}_2$ *Lagrange finite element in two space dimensions*

this unique set of nodes is shown in Figure 5.1. The interpolation points are the vertices $\{\boldsymbol{S}_1, \boldsymbol{S}_2, \boldsymbol{S}_3\}$ and the mid-edges $\{\boldsymbol{M}_1, \boldsymbol{M}_2, \boldsymbol{M}_3\}$.

The quadrature based on this set of nodes is the following:

$$\int_K f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \frac{|K|}{3}(f(\boldsymbol{M}_1) + f(\boldsymbol{M}_2) + f(\boldsymbol{M}_3)), \qquad \forall f \in \mathbb{P}_2. \tag{5.3}$$

By virtue of Proposition 5.1 it immediately follows that the lumped mass matrix is singular since the weights at the vertices are zero. A similar result holds in three space dimensions.

Following the work of [8], it is now well understood that the mass lumping method can be salvaged by selecting the Lagrange nodes on a non-uniform lattice on the reference element and by augmenting the polynomial space $\mathbb{P}_k$ with extra degrees of freedoms so that the resulting augmented space $\tilde{\mathbb{P}}_k$ produces a quadrature with positive weights. For instance, it is shown in [4, 9, 10] that the following space $\tilde{\mathbb{P}}_2 := \mathbb{P}_2 \oplus \mathrm{span}(b)$ is suitable for this purpose, where $b(\boldsymbol{x}) := \lambda_1(\boldsymbol{x})\lambda_2(\boldsymbol{x})\lambda_3(\boldsymbol{x})$ is the bubble function and $\lambda_1(\boldsymbol{x})$, $\lambda_2(\boldsymbol{x})$, $\lambda_3(\boldsymbol{x})$ are the barycentric coordinates over K. The quadrature associated with this polynomial space is as follows:

$$\begin{aligned}\int_K f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = |K|(&\frac{1}{20}(f(\boldsymbol{S}_1) + f(\boldsymbol{S}_2) + f(\boldsymbol{S}_3)) \\ &+ \frac{2}{15}(f(\boldsymbol{M}_1) + f(\boldsymbol{M}_2) + f(\boldsymbol{M}_3)) + \frac{9}{20}f(\boldsymbol{G})), \quad \forall f \in \mathbb{P}_4,\end{aligned} \tag{5.4}$$

where $\boldsymbol{G}$ is the barycenter of K. Higher-order versions of these ideas are proposed in [4, 9, 13, 19].

We propose in the next two sections two quasi-lumping techniques for $\mathbb{P}_2$ finite elements that do not require the extra barycentric degree of freedom invoked by $\tilde{\mathbb{P}}_2$.

5.3. Construction of a diagonal $\mathbb{P}_2$ quasi-lumped mass matrix. We present in this section a first attempt at quasi-lumping the mass matrix based on the standard Lagrange $\mathbb{P}_2$ nodes, see Figure 5.1, and using a diagonal matrix.

Again, the local mass matrix M^K is given by the expression

$$M^K_{ij} := \int_K \phi_i(\boldsymbol{x})\phi_j(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = |K|\Phi_i^T W \Phi_j, \tag{5.5}$$

where $(\Phi_1, \ldots, \Phi_6)$ is the canonical basis of $\mathbb{R}^6$ and the matrix W is given by

$$W = \begin{bmatrix} \frac{1}{30} & -\frac{1}{180} & -\frac{1}{180} & -\frac{1}{45} & 0 & 0 \\ -\frac{1}{180} & \frac{1}{30} & -\frac{1}{180} & 0 & -\frac{1}{45} & 0 \\ -\frac{1}{180} & -\frac{1}{180} & \frac{1}{30} & 0 & 0 & -\frac{1}{45} \\ -\frac{1}{45} & 0 & 0 & \frac{8}{45} & \frac{4}{45} & \frac{4}{45} \\ 0 & -\frac{1}{45} & 0 & \frac{4}{45} & \frac{8}{45} & \frac{4}{45} \\ 0 & 0 & -\frac{1}{45} & \frac{4}{45} & \frac{4}{45} & \frac{8}{45} \end{bmatrix}. \tag{5.6}$$

The coefficients of W are equal to $|K|^{-1}\int_K \phi_i(\boldsymbol{x})\phi_j(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$, $1\le i,j\le 6$, where $\phi_1,\ldots,\phi_6$ are the local nodal shape functions.

Since we have seen above that (5.3) is the only possible quadrature that is exact for $\mathbb{P}_2$ polynomials, we propose to lower our expectations by constructing a convex combination between (4.3) and (5.3) as follows:

$$\int_K f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \gamma\frac{|K|}{3}(f(\boldsymbol{S}_1)+f(\boldsymbol{S}_2)+f(\boldsymbol{S}_3)) + (1-\gamma)\frac{|K|}{3}(f(\boldsymbol{M}_1)+f(\boldsymbol{M}_2)+f(\boldsymbol{M}_3)). \tag{5.7}$$

This gives a family of integration rules parameterized by γ that are exact only in $\mathbb{P}_1$ for all $\gamma\in(0,1)$. Since there are polynomials in $\mathbb{P}_2$ that are not integrated exactly with these rules, this choice certainly forbids any hope that the resulting method can be optimal in terms of approximation, but we nevertheless persists in this direction. The quasi-lumped local mass matrix that results from this strategy is the following:

$$\overline{M}_{ij}^K := |K|\Phi_i^T\overline{W}\Phi_j\,, \tag{5.8}$$

where

$$\overline{W} := \begin{bmatrix} \frac{1}{3}\gamma I_3 & 0 \\ 0 & \frac{1}{3}(1-\gamma)I_3 \end{bmatrix}, \tag{5.9}$$

where I_3 is the 3×3 identity matrix.

Our goal is to use $\overline{W}$ to approximate the matrix W defined in (5.6). The matrix $\overline{W}$ is a good approximation of W if the spectral radius of $\overline{W}^{-1}(\overline{W}-W)$ is smaller than 1. The spectral radius of $\overline{W}^{-1}(\overline{W}-W)$ can be computed exactly with the help of Maple. We show in Figure 5.2 the spectral radius of $\overline{W}^{-1}(\overline{W}-W)$ as a function of γ in the range $0.01\le\gamma\le0.79$. The minimum is reached for $\gamma\approx\frac{1}{5}$ and the largest eigenvalue has a modulus less than 0.875 in the range $\gamma\in[0.08,0.25]$. In conclusion, any value of γ in the range $[0.08,0.25]$ gives a quasi-lumped mass matrix for which all the Neumann series (3.6)–(3.8) converge. We have observed numerically that $\gamma=\frac{1}{12}$ gives the best performance.

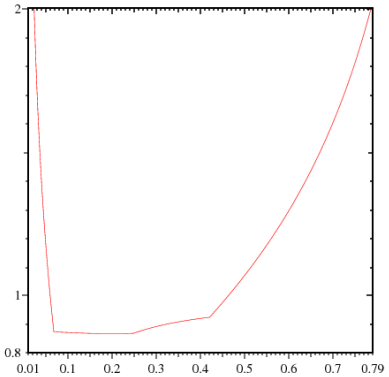


FIGURE 5.2. *Spectral radius of* $\overline{W}^{-1}(\overline{W}-W)$.

Since the quadrature rule (5.7) is not exact in $\mathbb{P}_2$, we cannot expect the mass correction method introduced in §3 to converge optimally with a fixed number of corrections. We illustrate this statement by performing the converge tests on the linear transport equation described in §4.2. We solve the linear transport problem on various grids ($h=0.2,0.1,0.05,0.025,0.0125$) using $\gamma=\frac{1}{12}$, and we compute the L^2-norm of the error at $T=1$. The results are reported in Table 5.1.

h	Consist. Mass	var. corrections		4 corrections
0.2	5.053E-2	5.536E-2	2	3.846E-2
0.1	1.522E-2	1.226E-2	4	1.226E-2
0.05	2.676E-3	2.773E-3	6	4.966E-3
0.025	5.589E-4	5.865E-4	8	3.577E-3
0.0125	1.446E-4	1.486E-4	10	3.448E-3

TABLE 5.1

L^2-norm of error at $T = 1$, $\mathbb{P}_2$ finite elements. Computations done with the consistent mass matrix, the quasi-lumped mass matrix corrected a variable number of times, and the quasi-lumped mass matrix corrected four times.

In conclusion, although the proposed quasi-lumped mass matrix does not give an optimally convergent method when corrected a fixed number of times, we claim that the matrix $\overline{M}$ is nevertheless a good preconditioner of M and can certainly be used as such within any Krylov-based iterative technique. This claim in confirmed by Table 5.2 where we report the condition number of $\overline{M}^{-1}M$ as a function of the mesh size h for the five grids used in the above convergence tests and for two values of γ.

	h	0.2	0.1	0.05	0.025	0.0125
Cond($\overline{M}^{-1}M$)	$\gamma = \frac{1}{12}$	5.910	5.952	5.995	6.015	6.022
	$\gamma = \frac{1}{5}$	5.342	5.384	5.420	5.435	5.440

TABLE 5.2

Condition number of $\overline{M}^{-1}M$ vs. h

5.4. Construction of a triangular $\mathbb{P}_2$ quasi-lumped mass matrix. Since it does not seem to be possible to construct a diagonal quasi-lumped $\mathbb{P}_2$ mass matrix with optimal convergence properties, see §5.3, we propose to consider the next best alternative which is to construct a triangular approximate mass matrix. Recall that triangular matrices are as easy to invert as diagonal matrices. This possibility has never been explored yet, to the best of our knowledge.

We propose to consider the following non-symmetric bilinear quadrature rule

$$\int_K u(\boldsymbol{x})v(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \approx |K| U^T \overline{W} V. \tag{5.10}$$

where the matrix $\overline{W}$ is defined by

$$\overline{W} := \begin{bmatrix} \alpha & 0 & 0 & \gamma & \delta & \delta \\ 0 & \alpha & 0 & \delta & \gamma & \delta \\ 0 & 0 & \alpha & \delta & \delta & \gamma \\ 0 & 0 & 0 & \beta & 0 & 0 \\ 0 & 0 & 0 & 0 & \beta & 0 \\ 0 & 0 & 0 & 0 & 0 & \beta \end{bmatrix}, \tag{5.11}$$

and with

$$U^T := (u(\boldsymbol{S}_1), u(\boldsymbol{S}_2), u(\boldsymbol{S}_3), u(\boldsymbol{M}_1), u(\boldsymbol{M}_2), u(\boldsymbol{M}_3), \tag{5.12}$$

$$V^T := (v(\boldsymbol{S}_1), v(\boldsymbol{S}_2), v(\boldsymbol{S}_3), v(\boldsymbol{M}_1), v(\boldsymbol{M}_2), v(\boldsymbol{M}_3). \tag{5.13}$$

We now try to make this formula as accurate as possible. Let ϕ_1, ϕ_2, ϕ_3 be the $\mathbb{P}_2$ nodal shape functions associated with the vertices $\boldsymbol{S}_1, \boldsymbol{S}_2, \boldsymbol{S}_3$, and ϕ_4, ϕ_5, ϕ_6 be the nodal shape functions associated with the mid-edges $\boldsymbol{M}_1, \boldsymbol{M}_2, \boldsymbol{M}_3$.

LEMMA 5.2. *The formula* (5.10) *is exact for all* $(u,v) \in \mathbb{P}_2\times\mathbb{P}_0 \cup \mathbb{P}_0\times\mathbb{P}_1$ *provided the following holds:*

$$\alpha+\gamma+2\delta=0, \quad \beta=\frac{1}{3}, \quad \forall\gamma,\delta\in\mathbb{R}. \tag{5.14}$$

(i) (5.10) *is also exact for all* $(u,v)\in \mathrm{span}(\phi_1,\phi_2,\phi_3)\times\mathbb{P}_1$ *if* (5.14) *holds and*

$$\delta+\gamma=-\frac{1}{30}, \quad \forall\gamma\in\mathbb{R}. \tag{5.15}$$

(ii) (5.10) *is exact for all* $(u,v)\in\mathbb{P}_1\times\mathbb{P}_1$ *if* (5.14) *holds and*

$$\delta+\gamma=0, \quad \forall\gamma\in\mathbb{R}. \tag{5.16}$$

Proof. (1) The condition $\int_K \phi_i(\boldsymbol{x})1\,\mathrm{d}\boldsymbol{x}=0$, $i=1,2,3$ implies $\alpha+\gamma+2\delta=0$. The condition $\int_K \phi_i(\boldsymbol{x})1\,\mathrm{d}\boldsymbol{x}=\frac{1}{3}|K|$, $i=4,5,6$ implies $\beta=\frac{1}{3}$. In conclusion (5.10) holds for all $(u,v)\in\mathbb{P}_2\times\mathbb{P}_0$ provided $\alpha+\gamma+2\delta=0$ and $\beta=\frac{1}{3}$. Moreover one easily verifies that $\int_K 1\lambda_i(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}=\frac{1}{3}|K|$ is computed exactly if $\alpha+\beta+\gamma+2\delta=\frac{1}{3}$, which with $\beta=\frac{1}{3}$ gives again $\alpha+\gamma+2\delta=0$. This implies that (5.10) holds also for all $(u,v)\in\mathbb{P}_0\times\mathbb{P}_1$. Note that these identities hold for the (singular) lumped mass matrix for which $\alpha=\gamma=\delta=0$ and $\beta=\frac{1}{3}$.

(2) The condition $\int_K \phi_i(\boldsymbol{x})\lambda_i(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}=\frac{1}{30}|K|$, $i=1,2,3$ implies $\alpha+\frac{1}{2}\delta+\frac{1}{2}\delta=\frac{1}{30}$. In conclusion we have

$$\alpha+\gamma+2\delta=0, \quad \alpha+\delta=\frac{1}{30}, \quad \beta=\frac{1}{3},$$

which is clearly equivalent to (5.14)-(5.15). Let $i\in\{1,2,3\}$ and let $\{j_1,j_2\}=\{1,2,3\}\backslash\{i\}$, then let us show that (5.10) evaluates exactly $\int_K \phi_i(\boldsymbol{x})\lambda_{j_1}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$ and $\int_K \phi_i(\boldsymbol{x})\lambda_{j_2}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$, which will conclude the proof of (i). Since the symmetries of the triangle K imply that $\int_K \phi_i(\boldsymbol{x})\lambda_{j_1}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$ equals $\int_K \phi_i(\boldsymbol{x})\lambda_{j_2}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$, we have

$$\begin{aligned}\int_K \phi_i(\boldsymbol{x})\lambda_{j_1}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} &= \frac{1}{2}\int_K \phi_i(\boldsymbol{x})(\lambda_{j_1}(\boldsymbol{x})+\lambda_{j_2}(\boldsymbol{x}))\,\mathrm{d}\boldsymbol{x}\\ &= \frac{1}{2}\int_K \phi_i(\boldsymbol{x})(\lambda_i(\boldsymbol{x})+\lambda_{j_1}(\boldsymbol{x})+\lambda_{j_2}(\boldsymbol{x}))\,\mathrm{d}\boldsymbol{x}-\frac{1}{2}\int_K \phi_i(\boldsymbol{x})\lambda_i(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\\ &= -\frac{1}{2}\int_K \phi_i(\boldsymbol{x})\lambda_i(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}.\end{aligned}$$

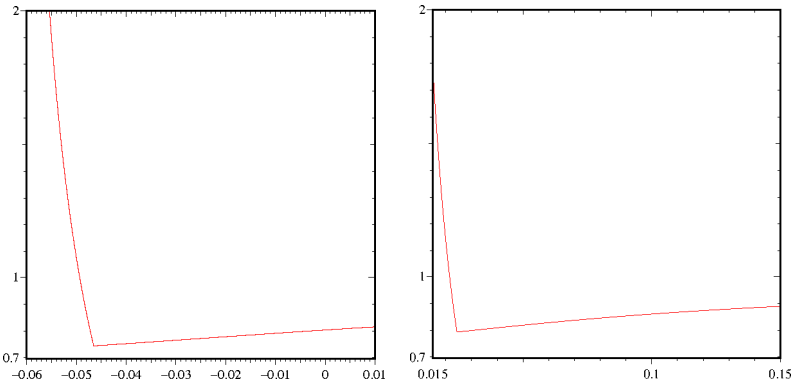
The conclusion follows readily owing to the fact that (5.10) satisfies all the symmetries used above and (5.10) evaluates exactly $\int_K \phi_i(\boldsymbol{x})1\,\mathrm{d}\boldsymbol{x}$ and $\int_K \phi_i(\boldsymbol{x})\lambda_j(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$, $i,j=1,2,3$.

(3) The proof of (ii) is similar. We observe first that (5.10) evaluates exactly $\int_K \lambda_i(\boldsymbol{x})\lambda_i(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$, $i\in\{1,2,3\}$ provided

$$\alpha+\delta+\frac{1}{2}\beta=\frac{1}{6},$$

which together with the results of step (1) imply (5.14)-(5.16). Proving then that (5.10) evaluates exactly $\int_K \lambda_i(\boldsymbol{x})\lambda_j(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$ for $j=\{1,2,3\}\backslash\{i\}$ can be done by using the symmetry properties of the quadrature as above. ∎

We now have two families of bilinear integration rules parameterized by γ. Our goal is to use $\overline{W}$ to approximate the matrix W defined in (5.6). We expect $\overline{W}$ to be a good approximation of W if



(a) Quadrature family (5.14)-(5.15) (b) Quadrature family (5.14)-(5.16)

FIGURE 5.3. *Spectral radius of* $\overline{W}^{-1}(\overline{W}-W)$ *as a function of* γ.

the spectral radius of $\overline{W}^{-1}(\overline{W}-W)$ is smaller than 1. We show in Figure 5.3(a) the spectral radius of $\overline{W}^{-1}(\overline{W}-W)$ as a function of γ in the range $-0.06 \leq \gamma \leq 0.01$ for the integration rule defined by (5.14)-(5.15). The minimum is reached for $\gamma \approx -\frac{1}{21}$ and the range $\gamma \in [-0.042, -0.03]$ is acceptable. The particular value $\gamma = -\frac{1}{30}$ has the advantage of simplifying the expression of $\overline{W}$ since $\delta = 0$ for this value. We have found numerically that indeed the choice $\gamma = -\frac{1}{30}$ works very well. We show in Figure 5.3(b) the spectral radius of $\overline{W}^{-1}(\overline{W}-W)$ as a function of γ in the range $0.015 \leq \gamma \leq 0.15$ for the integration rule defined in (5.14)-(5.16). The minimum is reached for $\gamma \approx \frac{1}{41}$ and the range $\gamma \in [0.03, 0.07]$ is acceptable. We have found numerically that the pair $\gamma = \frac{1}{30}$ works well for the integration rule (5.14)-(5.16).

Another important property to consider is to make sure that the quasi-lumped mass matrix is definite positive. This property holds as soon as the elementary matrix $\overline{W}$ is definite positive. We verify that $\overline{W}$ is definite positive by inspecting the smallest eigenvalue of $\frac{1}{2}(\overline{W}+\overline{W}^T)$.

PROPOSITION 5.3. *(i) With the choice of parameters* (5.14)-(5.15)*, the smallest eigenvalues of* $\frac{1}{2}(\overline{W}+\overline{W}^T)$ *is*

$$\min\left(\frac{1}{5}+\frac{1}{2}\gamma-\frac{1}{30}\sqrt{17-90\gamma+450\gamma^2}, \frac{1}{5}+\frac{1}{2}\gamma-\frac{1}{60}\sqrt{65-360\gamma+4500\gamma^2}\right) \tag{5.17}$$

(ii) With the choice of parameters (5.14)-(5.16)*, the smallest eigenvalues of* $\frac{1}{2}(\overline{W}+\overline{W}^T)$ *is*

$$\frac{1}{6}+\frac{1}{2}\gamma-\frac{1}{6}\sqrt{1-6\gamma+45\gamma^2}. \tag{5.18}$$

One can verify that matrix $\overline{W}$ is definite positive for the two choices (5.14)-(5.15) and (5.14)-(5.16) in the ranges considered above, $\gamma \in [-0.042, -0.03]$ and $\gamma \in [0.03, 0.07]$, respectively.

Remark 5.1. The mass matrix $\overline{M}$ preserves the block structure of the local mass matrices $\overline{M}^K$. For instance $\overline{M}$ is upper triangular if the vertices of the mesh are enumerated before the mid-edges.

Remark 5.2. The idea of using non-diagonal matrices to represent a quadrature rule can be traced back to [24, p.5] and [18, (A.2)]. The novelty of the technique presented here is that we are using a triangular matrix to represent a quadrature rule involving the product of two functions. The resulting bilinear form is obviously not a scalar product.

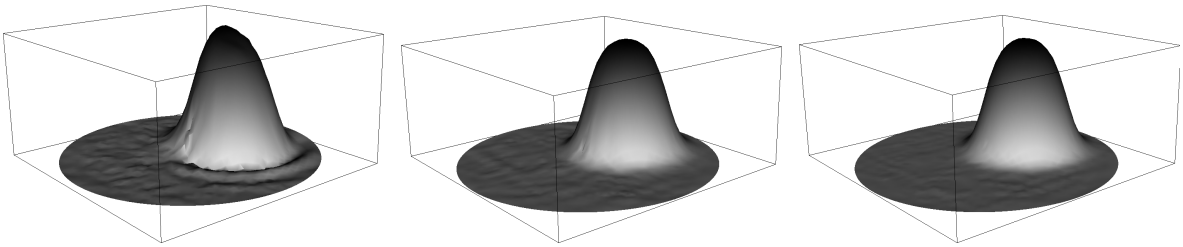
Remark 5.3. Instead of considering the bilinear quadrature rule defined by (5.11), one may think of using the transpose of the matrix $\overline{W}$ thus giving a lower triangular quadrature rule. The

counterpart of Lemma 5.2 follows immediately by permuting the polynomial spaces. One can then define another quasi-lumped mass matrix. This leads to two quasi-lumped matrices, say $\overline{M}_l$ and $\overline{M}_u$, where subscripts l and u are for lower or upper triangular. Let us then define the matrix $\overline{M} = (\frac{1}{2}\overline{M}_l^{-1} + \frac{1}{2}\overline{M}_u^{-1})^{-1}$. This new matrix $\overline{M}$ is clearly symmetric and can be used as a quasi-lumped mass matrix: the matrix vector product $\overline{M}^{-1}y$ is realized by solving $\overline{M}_u z_u = y$ and $\overline{M}_l z_l = y$ and by setting $\overline{M}^{-1}y = \frac{1}{2}(z_u + z_l)$. The resulting algorithm is of course a little more time consuming, but the quasi-lumped mass matrix is now symmetric. This route has been investigated, but the results are somewhat disappointing. It seems that $\overline{M}_l$ is not nearly as effective as $\overline{M}_u$ when applying the dispersion correction formula with one term only. The two matrices give similar results after four corrections though. This phenomenon is not yet well understood. We conjecture that it is important to associate the largest polynomial space with the test functions in the quadrature rule (5.11); the quadrature associated with $\overline{M}_l$ is exact in $\mathbb{P}_0 \times \mathbb{P}_2$, where $\mathbb{P}_0$ is the test space and $\mathbb{P}_2$ the trial space, whereas the quadrature associated with $\overline{M}_u$ is exact in $\mathbb{P}_2 \times \mathbb{P}_0$. Note finally that there is no local counterpart to the matrix $\overline{M} = (\frac{1}{2}\overline{M}_l^{-1} + \frac{1}{2}\overline{M}_u^{-1})^{-1}$ that defines a bilinear quadrature with properties similar to those mentioned in Lemma 5.2.

5.5. Numerical illustrations/Galerkin. We illustrate the efficiency of the construction proposed above by testing it on the linear transport equation (4.6)-(4.7) with the quadrature rule (5.14)-(5.15) using $\gamma = -\frac{1}{30}$, (i.e., $\alpha = \frac{1}{30}$, $\beta = \frac{1}{3}$, $\delta = 0$). Note in passing that the value $\gamma = -\frac{1}{30}$ is such that the three smallest eigenvalues of $\frac{1}{2}(\overline{W} + \overline{W}^T)$ are equal (see (5.17)). In this case we have

$$\overline{W} = \begin{bmatrix} \frac{1}{30} I_3 & -\frac{1}{30} I_3 \\ 0 & \frac{1}{3} I_3 \end{bmatrix} \tag{5.19}$$

The space approximation is done by using the Galerkin method on a mesh composed of 6293 $\mathbb{P}_2$ nodes. The time stepping is done with the standard RK4 method to ascertain that the error in time is negligible with respect to the spatial error. The solution is computed at $T = 2$, i.e., after two revolutions. The results are shown in Figure 5.4. The solution obtained with quasi-lumping is shown in 5.4(a). The dispersive effect associated with quasi-lumping is clear. We show in Figure 4.1(b) and



(a) 1 (b) $1 + A$ (c) $1 + A + A^2 + A^3 + A^4$

FIGURE 5.4. *Mass matrix corrections on 2D Delaunay triangulation, $\mathbb{P}_2$ finite elements, $h \approx 0.05$ (6293 $\mathbb{P}_2$ nodes), $T = 2$.*

4.1(c) the solutions obtained by replacing the inverse of the quasi-lumped mass matrix by $(1+A)\overline{M}^{-1}$ and $(1 + A + A^2 + A^3 + A^4)\overline{M}^{-1}$, respectively, where $A := \overline{M}^{-1}(\overline{M} - M)$. The conclusion is the same as for $\mathbb{P}_1$ finite elements: Applying one correction to the quasi-lumped mass matrix is enough to correct the dispersion effect.

We finish this section by performing convergence tests on the linear transport problem (4.6)-(4.7). The space approximation is done by using the Galerkin method on various grids ($h = 0.2, 0.1, 0.05, 0.025, 0.0125$). The time stepping is done with RK4 with CFL = 0.7. The L^2-norm

h	Consist. Mass	4 corrections	1 correction	0 correction
0.2	5.053E-2	3.726E-2	9.744E-2	3.045E-1
0.1	1.522E-2	1.159E-2	2.171E-2	1.467E-1
0.05	2.676E-3	2.231E-3	4.076E-3	4.610E-2
0.025	5.589E-4	4.658E-4	1.465E-3	1.233E-2
0.0125	1.446E-4	1.091E-4	2.756E-4	3.094E-3

TABLE 5.3

L^2-norm of error at $T = 1$, $\mathbb{P}_2$ finite elements. Computations done with the consistent mass matrix, the quasi-lumped mass matrix corrected four times, and the quasi-lumped mass matrix with no correction.

of the error is computed at $T = 1$. The results are reported in Table 5.3. It is remarkable that the technique using the uncorrected quasi-lumped mass matrix is second-order convergent. To the best of our knowledge, the technique presented here is the first convergent quasi-lumping technique for $\mathbb{P}_2$ finite elements using only the standard Lagrangian nodes. It is also remarkable that for all practical purposes, the results obtained by using the consistent mass matrix and by applying four mass corrections to the quasi-lumped mass matrix are identical. This test confirms the observations already made with $\mathbb{P}_1$ finite elements.

5.6. Numerical illustrations/Galerkin+Stabilization. Since it is known that the Galerkin method is suboptimal for linear first-order PDE's, we now investigate the performance of the mass correction when used jointly with stabilization techniques.

We consider first the so-called edge stabilization technique, [3]. Edge stabilization consists of augmenting the Galerkin formulation with a penalty term acting on the jump of the normal derivative of the unknown across all the internal faces of the mesh. Upon denoting X_h the finite element space, the edge stabilization technique consists of seeking $u \in \mathcal{C}^1((0,T); X_h)$ so that

$$\int_\Omega (\partial_t u + \boldsymbol{\beta}\cdot\nabla u)v \,\mathrm{d}\boldsymbol{x} + \chi \sum_{F\in\mathcal{F}_h^i} h_F^2 \|\boldsymbol{\beta}\|_{L^\infty(\Delta_F)} \int_F [\![\partial_n u]\!][\![\partial_n v]\!] \,\mathrm{d}\boldsymbol{x} = 0, \qquad \forall v \in X_h, \tag{5.20}$$

where $\mathcal{F}_h^i$ is the collection of the internal faces, h_F is the diameter of F, and Δ_F is the union of the two elements sharing the interface F. The coefficient χ is user-dependent; we have chosen $\chi = 0.01$ in the computations reported below. The time stepping is again explicit and done using RK4. The edge stabilization bilinear form is made explicit. The resulting scheme is known to be stable under the usual CFL condition in [2]. We used CFL $= 0.7$ in the computations reported in Table 5.4. By

h	Consist. Mass	4 corrections	1 correction	0 correction
0.2	2.904E-2	2.809E-2	8.269E-2	2.927E-1
0.1	5.633E-3	5.078E-3	1.523E-2	1.429E-1
0.05	5.707E-4	5.694E-4	2.417E-3	4.473E-2
0.025	8.421E-5	9.582E-5	6.911E-4	1.178E-2
0.0125	1.338E-5	1.764E-5	2.161E-4	2.918E-3

TABLE 5.4

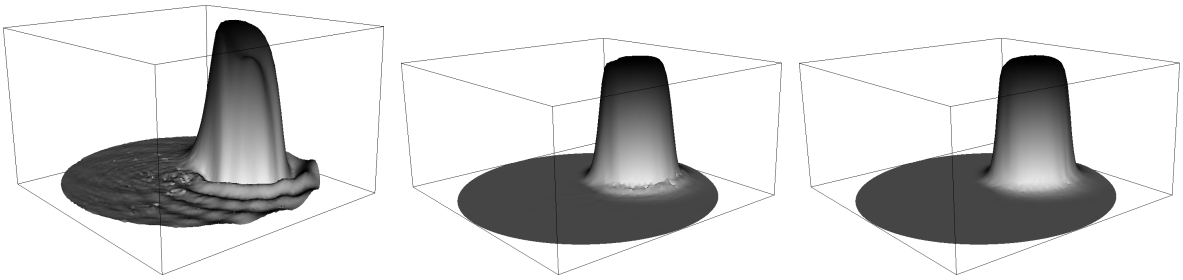
L^2-norm of error at $T = 1$, $\mathbb{P}_2$ finite elements with edge stabilization.

comparing Table 5.3 and Table 5.4, we observe that, as expected, the edge-stabilized technique is more accurate than the Galerkin technique. The results from Table 5.4 show that the technique with the quasi-lumped mass matrix corrected four times has roughly the same convergence rate as the technique using the consistent mass matrix.

We now consider the so-called entropy viscosity technique introduced in [15]. The method consists of adding a nonlinear dissipation to the Galerkin formulation to stabilize the method:

$$\int_\Omega (\partial_t u + \boldsymbol{\beta}\cdot\nabla u)v\,\mathrm{d}\boldsymbol{x} + \sum_{K\in\mathcal{T}_h}\int_K \nu_h(u)\nabla u\cdot\nabla v\,\mathrm{d}\boldsymbol{x} = 0, \qquad \forall v\in X_h, \tag{5.21}$$

The nonlinear viscosity is proportional to an entropy residual and is at most equal to $c_1\|\boldsymbol{\beta}\|_{L^\infty(K)}h_K/k$, where h_K is the diameter of K, k is the polynomial degree of approximation, and $c_1 = 1/4k$. We solve the linear transport equation (4.6)-(4.7) with the initial data $u_0(\boldsymbol{x}) = 1$ if $\|\boldsymbol{x}-\boldsymbol{x}_0\|\le a$ and $u_0(\boldsymbol{x}) = 0$ otherwise. We use the quadrature rule (5.14)-(5.15) with $\gamma = -\frac{1}{30}$ to evaluate the quasi-lumped mass matrix. Again, the mesh is composed of 6293 $\mathbb{P}_2$ nodes, the time stepping is done with the standard RK4, and the solution is computed at $T = 2$. The results are shown in Figure 5.5. We observe that using one mass matrix correction only is enough to remove most of the dispersion effect induced by the quasi-mass lumping.



(a) 1 (b) $1+A$ (c) $1+A+A^2+A^3+A^4$

FIGURE 5.5. *Mass matrix corrections on 2D Delaunay triangulation, $\mathbb{P}_2$ finite elements with entropy viscosity stabilization, $h\approx 0.05$ (6293 $\mathbb{P}_2$ nodes), $T = 2$.*

We have also performed tests with the quadrature (5.14)-(5.16) using $\gamma = \frac{1}{30}$, (i.e., $\alpha = \frac{1}{30}$, $\beta = \frac{1}{3}$, $\delta = -\frac{1}{30}$). The performance of the method is similar to what has been described above. We do not report these tests here for the sake of brevity.

The tests shown in this section confirm that the mass correction method is robust with respect to both the edge stabilization and the entropy viscosity technique.

6. $\mathbb{P}_N$ extensions. We finish this paper by exploring mass lumping for higher-order simplicial Lagrange finite elements in two space dimensions. We are going to restrict ourselves to $\mathbb{P}_N$ finite elements, $N\ge 3$, and investigate whether it is possible to find lattices on the reference simplex that give lumped mass matrices with positive weights and determine whether the mass matrix correction from §3 can be applied. Of course the quadrature associated with mass lumping in $\mathbb{P}_N$ is exact in $\mathbb{P}_N$ only, which is suboptimal since quadratures must be exact in $\mathbb{P}_{2N-2}$ to yield optimal error estimates in the energy norm, [1, 7, 20, 11].

6.1. $\mathbb{P}_3$ approximation. We begin with the $\mathbb{P}_3$ approximation. As generally advocated in the spectral element literature, the interpolation points for $\mathbb{P}_3$ Lagrange finite elements must be the images, by appropriate mappings, of the four one-dimensional Gauss-Lobatto Legendre points $\{-1,-1/\sqrt{5},1/\sqrt{5},1\}$ on the edges of the triangle and the center of gravity of the triangle. The quadrature associated with these points is exact in $\mathbb{P}_3$ only, but since all the weights are positive (suboptimal) mass lumping is possible for this finite element. Furthermore we have verified numerically that the spectral radius of the local matrix $\overline{W}^{-1}(\overline{W}-W)$ approximately equals $0.702<1$, thereby confirming that the mass matrix correction algorithm proposed in §3 is convergent.

Let us now illustrate the mass correction algorithm on the two-dimensional transport problem defined in §4.2. We have performed computations on four grids ($h = 0.157, 0.0753, 0.0575, 0.039$) with the consistent mass matrix, the lumped mass matrix, and the lumped mass matrix corrected up to 8 times. The computations have been done with the symmetric form of the mass correction matrix A, see (3.7), but this particular choice does not affect the spectral radius of A as shown in Lemma 3.1. The results are reported in Table 6.1.

h	Consist. Mass	8 correct.	4 correct.	2 correct.	1 correct.	no correct.
0.157	5.1078E-2	5.0866E-2	4.9149E-2	5.5802E-2	1.2791E-1	1.1185E-1
0.0753	5.7404E-3	5.7940E-3	7.4701E-3	1.7421E-2	4.8163E-2	4.2324E-2
0.0575	1.6734E-3	1.6168E-3	2.3904E-3	8.2706E-3	2.9825E-2	2.2364E-2
0.039	4.5458E-4	4.3058E-4	1.0650E-3	4.2248E-3	1.5986E-2	1.5216E-2

TABLE 6.1
L^2-norm of error at $T = 1$, $\mathbb{P}_3$ finite elements.

As can be observed in Table 6.1, the convergence rate obtained with the standard mass lumping is less than second-order (as expected), whereas it is close to fourth-order with the consistent mass matrix. One observes significant improvements with the mass correction algorithm. This is remarkable since the quadrature based on the interpolation points is exact in $\mathbb{P}_3$ and is not in $\mathbb{P}_{2N-2=4}$. Moreover, as already observed for the $\mathbb{P}_2$ approximation, the mass correction algorithm gives slightly better accuracy than when using the consistent mass matrix when the number of mass corrections is larger enough. This seems to indicate that the convergence of the Neumann series (3.7) always occurs from below.

6.2. Higher-order variants. Let us now consider higher-order polynomials, i.e., $N \in \{4, 5, 6\}$. For $N \in \{4, 5\}$ we consider the interpolation points given by the "warp & blend " technique from [23] and, for $N = 6$, we use the so-called Fekete points. The list of the Fekete points in the reference triangle for $N \in \{3, 6, \cdots, 18\}$ can be found in [21]. For both these families, the interpolation nodes coincide with the Gauss-Lobatto-Legendre points on the edges of the reference triangle. The Lebesgue constant for both these families is small; for instance, it is less than 10 for polynomial of degrees at most 12.

The first difficulty we encounter when computing the weights of the quadrature associated with the warp & blend points is that the weights at the vertices are negative for $N = 4$. The second problem is that the spectral radius of the local matrix A grows with N and is larger than 1 for $N \geq 4$ as shown in Table 6.2 for $N \in \{3, 4, 5, 6\}$.

N	3	4	5	6
$\rho(A)$	0.702	1.36	6.33	12.33

TABLE 6.2
Spectral radius of the local matrix A.

The above negative results show that standard mass-lumping fails for $N > 3$ in two space dimensions for standard Lagrange elements. This situation can be fixed by using the augmented spaces $\tilde{\mathbb{P}}_N$ mentioned in §5.2. This idea has been shown to work up to $N = 6$ in two space dimensions and up to $N = 4$ in three space dimensions in [4]. We think however that the quasi-lumping technique that we developed for $\mathbb{P}_2$ finite elements in §5.4 can be extended to higher-order polynomial degree. We think in particular that it should be possible in principle to construct triangular quasi-lumped mass matrices as alternatives to the $\tilde{\mathbb{P}}_N$ construction.

7. Conclusions. A new mass correction technique has been introduced to correct the dispersion error of mass lumping. The method has been shown to have the same anti-dispersive effect as when working with the consistent mass matrix. Two quasi-lumping techniques for $\mathbb{P}_2$ finite elements have been introduced. The $\mathbb{P}_2$ quasi-lumping technique based on the idea of using a triangular lumped mass matrix, as presented in §5.4, is new to the best of our knowledge. The mass correction technique introduced in §3 has been shown to perform very well with $\mathbb{P}_1$ lumping and $\mathbb{P}_2$ quasi-lumping. It seems that for these two elements using only one correction term only is enough to remove the dispersion error. We have verified that, although suboptimal, satisfactory results can also be obtained for the $\mathbb{P}_3$ approximation.

The idea of applying the mass matrix corrections and using triangular quasi-lumped mass matrices could be extended to higher-order finite elements. These venues will be explored in future works.

Appendix A. Anti-dispersive effect of the $\mathbb{Q}_1$-mass matrix on 2D Cartesian grids. Consider the two-dimensional transport equation:

$$\partial_t \mathsf{u} + \boldsymbol{\beta}\cdot\nabla \mathsf{u} = 0,$$

with constant velocity field $\boldsymbol{\beta} = (\beta_x, \beta_y)$. Consider a Cartesian grid with mesh sizes h_x and h_y in x- and y-directions, respectively.

PROPOSITION A.1. *The dominating term in the consistency error of the $\mathbb{Q}_1$ Galerkin approximation is $\mathcal{O}(h_x^4 + h_y^4)$ at the grid points (x_i, y_j).*

Proof. The test functions are the tensor products of one-dimensional functions, say $\psi_i^x(x)\psi_j^y(y)$, and the $\mathbb{Q}_1$ Galerkin approximation is represented as follows:

$$u(x,y,t) = \sum_i \sum_j u_{i,j}(t)\psi_i^x(x)\psi_j^y(y).$$

Applying twice the Simpson quadrature rule, the term involving the time derivative becomes

$$\int_{y_{j-1}}^{y_{j+1}} \psi_j^y \int_{x_{i-1}}^{x_{i+1}} \partial_t u\, \psi_i^x \, dxdy = h_x h_y \partial_t \Big(\frac{4}{9}u_{i,j} + \frac{1}{9}(u_{i\pm1,j} + u_{i,j\pm1}) + \frac{1}{36}u_{i\pm1,j\pm1}\Big),$$

where the notation $u_{i\pm1,j}$ stands for $u_{i-1,j} + u_{i+1,j}$ and $u_{i\pm1,j\pm1}$ stands for $u_{i-1,j-1} + u_{i+1,j-1} + u_{i-1,j+1} + u_{i+1,j+1}$, etc. Similarly, for the transport term we obtain:

$$\begin{aligned}\int_{y_{j-1}}^{y_{j+1}} \psi_j^y \int_{x_{i-1}}^{x_{i+1}} \boldsymbol{\beta}\cdot\nabla u\, \psi_i^x \,\mathrm{d}x\,\mathrm{d}y = {}& h_y\beta_x\left(\frac{1}{3}(u_{i+1,j} - u_{i-1,j}) + \frac{1}{12}(u_{i+1,j\pm1} + u_{i-1,j\pm1})\right)\\ &+ h_x\beta_y\left(\frac{1}{3}(u_{i,j+1} - u_{i,j-1}) + \frac{1}{12}(u_{i\pm1,j+1} + u_{i\pm1,j-1})\right).\end{aligned}$$

After inserting the exact solution, u, in the $\mathbb{Q}_1$ Galerkin approximation of the transport equation, using Taylor expansions, and dividing by $h_x h_y$, we obtain:

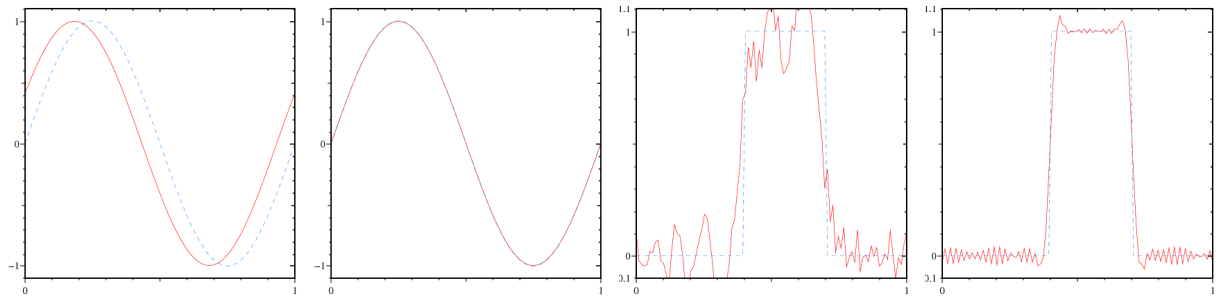
$$\begin{aligned}\frac{1}{h_x h_y}\int_{S_{ij}} (\partial_t \mathsf{u}_{ij} + \boldsymbol{\beta}\cdot\nabla \mathsf{u}_{ij})\psi_i^x(x)\psi_j^y(y)\,\mathrm{d}x\,\mathrm{d}y = {}& \partial_t(\mathsf{u}_{ij} + \frac{h_x^2}{6}\partial_{xx}\mathsf{u}_{ij} + \frac{h_y^2}{6}\partial_{yy}\mathsf{u}_{ij})\\ &+ \beta_x(\partial_x \mathsf{u}_{ij} + \frac{h_x^2}{6}\partial_{xxx}\mathsf{u}_{ij} + \frac{h_y^2}{6}\partial_{xyy}\mathsf{u}_{ij})\beta_x(\partial_y \mathsf{u}_{ij} + \frac{h_y^2}{6}\partial_{yyy}\mathsf{u}_{ij} + \frac{h_x^2}{6}\partial_{yxx}\mathsf{u}_{ij}) + \mathcal{O}(h_x^4 + h_y^4),\end{aligned}$$

where $S_{ij} = [x_{i-1}, x_{i+1}] \times [y_{i-1}, y_{i+1}]$ and $\mathsf{u}_{ij} := \mathsf{u}(x_i, y_j)$. Taking into account that $\partial_t \mathsf{u}(x_i, y_j, t) = -\beta_x \partial_x \mathsf{u}(x_i, y_j, t) - \beta_y \partial_y \mathsf{u}(x_i, y_j, t)$, one observes that the consistency error is of order 4. ∎

The above proposition shows that using the consistent matrix has an anti-dispersive effect for the 2D transport equation. One may conjecture that such a result holds in any dimension.

Appendix B. One-dimensional numerical illustrations.

B.1. Dispersive effects of mass lumping. We illustrate here the anti-dispersive effect of the consistent mass matrix in one space dimension. We show in Figure B.1(a) and B.1(b) the

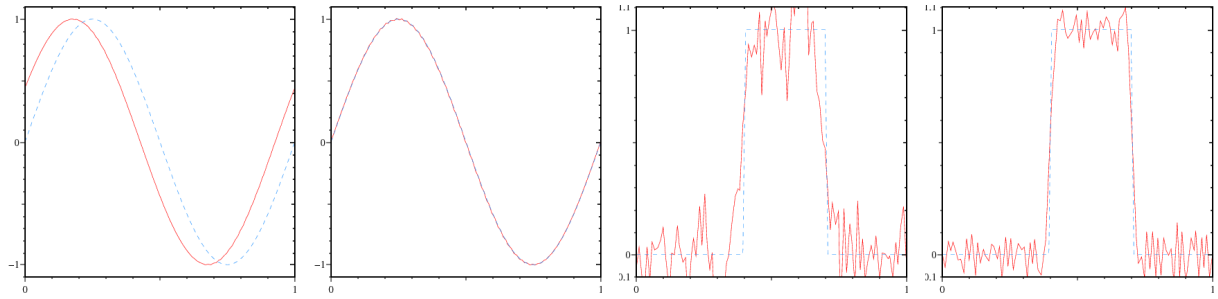


(a) Lumped mass matrix (b) Consist. mass matrix (c) Lumped mass matrix (d) Consist. mass matrix

FIGURE B.1. *Consistent vs. lumped mass matrix, uniform mesh, 100 cells, $T = 100$. Dashed line: exact solution; solid line: numerical approximation.*

Galerkin solution to the transport equation $\partial_t u + \partial_x u = 0$ over the interval $\Omega = (0, 1)$ with periodic boundary conditions and initial data $u(x, 0) = \sin(2\pi x)$. The solution is computed at $T = 100$, i.e., 100 periods, on a uniform mesh composed of 100 $\mathbb{P}_1$ cells. The time stepping is done using the standard explicit forth-order Runge Kutta (RK4) method so that the error induced by the time approximation is negligible with respect to the spatial error. The CFL number is 0.7. We show in Figure B.1(c) and B.1(d) the Galerkin solution with the initial data $u(x, 0) = 1$ if $0.4 < x < 0.7$ and $u(x, 0) = 0$ otherwise. The solution is computed at $T = 1$. The solutions shown in Figure B.1(a) and Figure B.1(c) are computed with the lumped mass matrix, and those shown in Figure B.1(b) and Figure B.1(d) are computed with the consistent mass matrix. The anti-dispersive effects of the consistent mass matrix are clearly visible on these two examples.

Since the dispersion analysis has been done assuming that the mesh is uniform, it is not clear a priori that the anti-dispersive effects of the mass matrix are robust with respect to mesh non-uniformity. This issue can be explored numerically by repeating the above numerical experiments on non-uniform meshes. The results are shown in Figure B.2. The mesh is composed of 100 cells with random size and the anisotropy factor is 3, that is to say the size ratio between two neighboring cells is at most 3. These experiments show that mesh non-uniformity does no have a notable influence on the anti-dispersive effects of the consistent mass matrix, and the conclusions of the dispersion analysis hold when the mesh is moderately non-uniform.

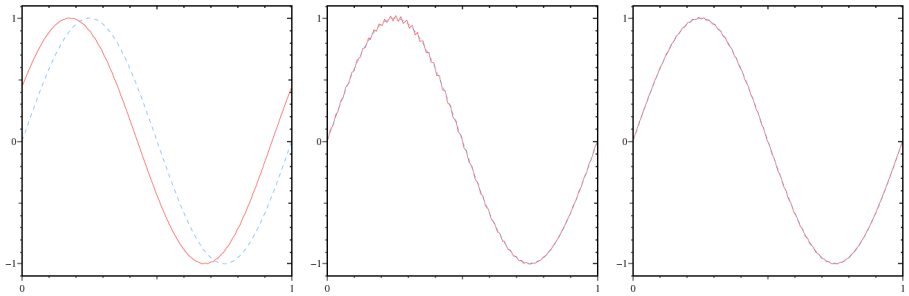


(a) Lumped mass matrix (b) Consist. mass matrix (c) Lumped mass matrix (d) Consist. mass matrix

FIGURE B.2. *Consistent vs. lumped mass matrix, random mesh, 100 cells, $T = 1$. Dashed line: exact solution; solid line: numerical approximation.*

B.2. Numerical illustrations of the correction technique. We illustrate numerically the correction technique introduced in §3 in one space dimension with $\mathbb{P}_1$ finite elements.

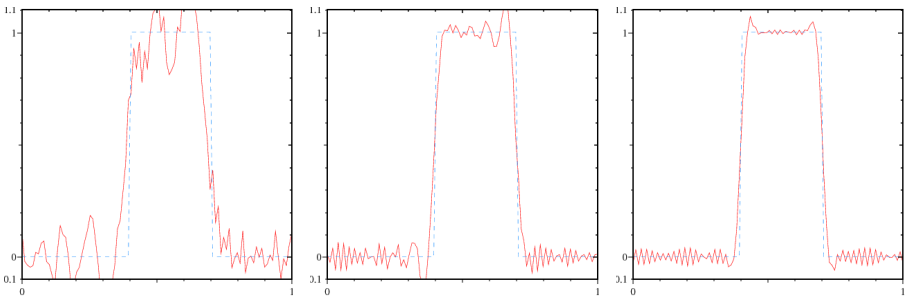
We show in Figure B.3 the effects of replacing the inverse of the lumped mass matrix by (3.6). The setting is the same as in Section B.1 and the initial data is the smooth sine function. We show in Figure B.3(a) the Galerkin solution using the lumped mass matrix on a random mesh composed of 100 cells at $T = 100$. The solutions shown in Figure B.3(b) and Figure B.3(c) have been obtained by replacing $\overline{M}^{-1}$ by $(1 + A)\overline{M}^{-1}$ and $(1 + A + A^2 + A^3 + A^4)\overline{M}^{-1}$, respectively. Figure B.3(a) clearly illustrates the dispersion effect of mass lumping; the phase error is $O(1)$ after 100 turnover times. Figure B.3(b) supports our claim that replacing M^{-1} by $(1+A)\overline{M}^{-1}$ corrects the dispersion error of the lumped mass matrix.



(a) 1 (b) $1 + A$ (c) $1 + A + A^2 + A^3 + A^4$

FIGURE B.3. *Mass matrix corrections on a random mesh, 100 cells, $T = 100$. Dashed line: exact solution; solid line: numerical approximation.*

We show in Figure B.4 the Galerkin solution of the one-dimensional transport problem with a step function as initial data. The solution shown in Figure B.4 has been computed at $T = 1$ using the lumped mass matrix on a uniform mesh composed of 100 cells. The solution shown in Figure B.3(b) and Figure B.3(c) have been obtained by replacing $\overline{M}^{-1}$ by $(1 + A)\overline{M}^{-1}$ and $(1+A+A^2+A^3+A^4)\overline{M}^{-1}$, respectively. Figure B.4 also confirms that replacing M^{-1} by $(1+A)\overline{M}^{-1}$ corrects the dispersion error of the lumped mass matrix even for non-smooth solutions.



(a) 1 (b) $1 + A$ (c) $1 + A + A^2 + A^3 + A^4$

FIGURE B.4. *Mass matrix corrections on a uniform mesh, 100 cells, $T = 1$. Dashed line: exact solution; solid line: (un-stabilized) Galerkin approximation.*

Of course, the Galerkin method must be stabilized to get rid of the spurious oscillations. As shown in §5.6, stabilization has a marginal effect on dispersion.

REFERENCES

[1] G. A. Baker and V. A. Dougalis. The effect of quadrature errors on finite element approximations for second order hyperbolic equations. *SIAM J. Numer. Anal.*, 13(4):577–598, 1976.

[2] E. Burman, A. Ern, and M. A. Fernández. Explicit Runge-Kutta schemes and finite elements with symmetric stabilization for first-order linear PDE systems. *SIAM J. Numer. Anal.*, 48(6):2019–2042, 2010.

[3] E. Burman and P. Hansbo. Edge stabilization for Galerkin approximations of convection-diffusion-reaction problems. *Comput. Methods Appl. Mech. Engrg.*, 193(15-16):1437–1453, 2004.

[4] M. J. S. Chin-Joe-Kong, W. A. Mulder, and M. Van Veldhuizen. Higher-order triangular and tetrahedral finite elements with mass lumping for solving the wave equation. *J. Engrg. Math.*, 35(4):405–426, 1999.

[5] M. A. Christon. The influence of the mass matrix on the dispersive nature of the semi-discrete, second-order wave equation. *Computer Methods in Applied Mechanics and Engineering*, 173(12):147 – 166, 1999.

[6] M. A. Christon, M. J. Martinez, and T. E. Voth. Generalized fourier analyses of the advection-diffusion equation-part i: one-dimensional domains. *International Journal for Numerical Methods in Fluids*, 45(8):839–887, 2004.

[7] P. Ciarlet. *The Finite Element Method for Elliptic Problems*. North Holland, Amsterdam, 1978.

[8] E. Cohen, Gary C. Eléments finis triangulaires $\mathbb{P}_2$ avec condensation de masse pour l'équation des ondes. Rapport de recherche 2418, INRIA, 1994.

[9] 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(6):2047–2078 (electronic), 2001.

[10] G. Cohen, P. Joly, and N. Tordjman. Higher order triangular finite elements with mass lumping for the wave equation. In *Mathematical and numerical aspects of wave propagation (Mandelieu-La Napoule, 1995)*, pages 270–279. SIAM, Philadelphia, PA, 1995.

[11] G. C. Cohen. *Higher-order numerical methods for transient wave equations*. Scientific Computation. Springer-Verlag, Berlin, 2002. With a foreword by R. Glowinski.

[12] I. Fried. Numerical integration in the finite element method. *Computers and Structures*, 4(5):921 – 932, 1974.

[13] F. X. Giraldo and M. A. Taylor. A diagonal-mass-matrix triangular-spectral-element method based on cubature points. *J. Engrg. Math.*, 56(3):307–322, 2006.

[14] P. Gresho, R. Sani, and M. Engelman. *Incompressible flow and the finite element method: advection-diffusion and isothermal laminar flow*. Incompressible Flow & the Finite Element Method. Wiley, 1998.

[15] J.-L. Guermond, R. Pasquetti, and B. Popov. Entropy viscosity method for nonlinear conservation laws. *Journal of Computational Physics*, 230:4248–4267, 2011.

[16] E. Hairer and G. Wanner. *Solving ordinary differential equations. II*, volume 14 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 1991. Stiff and differential-algebraic problems.

[17] F. Ham. Improved scalar transport for unstructured finite volume methods using simplex superposition. Annual research briefs, Center for Turbulence Research, 2008.

[18] J. S. Hesthaven and T. Warburton. Nodal high-order methods on unstructured grids: I. time-domain solution of maxwell's equations. *Journal of Computational Physics*, 181(1):186 – 221, 2002.

[19] S. Jund and S. Salmon. Arbitrary high-order finite element schemes and high-order mass lumping. *Int. J. Appl. Math. Comput. Sci.*, 17(3):375–393, 2007.

[20] P.-A. Raviart. The use of numerical integration in finite element methods for solving parabolic equations. In *Topics in numerical analysis (Proc. Roy. Irish Acad. Conf., University Coll., Dublin, 1972)*, pages 233–264. Academic Press, London, 1973.

[21] M. Taylor, B. Wingate, and R. Vincent. An algorithm for computing fekete points in the triangle. *SIAM J. Numer. Anal.*, 38(5):1707–1720, 2000.

[22] T. E. Voth, M. J. Martinez, and M. A. Christon. Generalized fourier analyses of the advectiondiffusion equation-part ii: two-dimensional domains. *International Journal for Numerical Methods in Fluids*, 45(8):889–920, 2004.

[23] T. Warburton. An explicit construction of interpolation nodes on the simplex. *J. Eng. Math.*, 56:247–262, 2006.

[24] T. Warburton, L. F. Pavarino, and J. S. Hesthaven. A pseudo-spectral scheme for the incompressible navier-stokes equations using unstructured nodal elements. *Journal of Computational Physics*, 164(1):1–21, 2000.