An Alternating-Direction Iteration Method for Helmholz Problems

Jim Douglas, Jr.* Jeffrey L. Hensley† Jean E. Roberts‡

Abstract. An alternating-direction iterative procedure is described for a class of Helmholz-like problems. An algorithm for the selection of the iteration parameters is derived; the parameters are complex with some having positive real part and some negative, reflecting the noncoercivity and nonsymmetry of the finite element or finite difference matrix. Examples are presented, with an application to wave propagation.

1. Introduction

The (complex–valued) Helmholz problem

$$\begin{aligned} -\Delta u - \omega^2 u &= f(x,y), \quad (x,y) \in \Omega, && \text{(1a)} \\ u_\nu + i\omega u &= 0, \quad (x,y) \in \partial\Omega && \text{(1b)} \end{aligned}$$

where $\Omega = [0,1]^2$ is the unit square, ν the outer unit normal to $\partial\Omega$, and $\omega > 0$, arises in the space–frequency treatment of the scalar wave problem

$$\begin{aligned} v_{tt} - \Delta v &= g, \quad x \in \Omega, \quad t > 0, && \text{(2a)} \\ v_t + v_\nu &= 0, \quad x \in \partial\Omega, \quad t > 0, && \text{(2b)} \\ v &\equiv 0, \quad x \in \Omega, \quad t \le 0. && \text{(2c)} \end{aligned}$$

The conditions (1b) and (2b) represent first–order absorbing boundary conditions that allow normally incident waves to pass out of Ω transparently. It is implicitly assumed that the support of f lies well inside the interior of Ω.

Consider a finite difference approximation to (1). Let $h = N^{-1}$ and set $(x_j, y_k) = (jh, kh)$. Let δ_x^2 denote the centered second difference with respect to x and set $\Delta_h = \delta_x^2 + \delta_y^2$. Let ∂_ν denote the centered first difference in the direction of the outer normal (here, an exterior bordering of the domain is assumed). Then, one proper approximation to (1) is given by seeking a grid function u_h such that

$$\begin{aligned} -\Delta_h u_h - \omega^2 u_h &= f_h, \quad (x_j, y_k) \in \Omega_h, && \text{(3a)} \\ \partial_\nu u_h + i \sin \xi h \, u_h &= 0, \quad (x_j, y_k) \in \Gamma_h, && \text{(3b)} \end{aligned}$$

*Department of Mathematics, Purdue University, West Lafayette, IN 47907-1395
†Center for Parallel and Scientific Computing, University of Tulsa, Tulsa, OK 74104-3189
‡INRIA, 78153 Le Chesnay, France

where $\Omega_h = \{(x_j, y_k)|j,k = 0,\ldots,N\}$ and Γ_h consists of the boundary grid points. The parameter ξ is given by

$$\xi = \frac{2}{h}\arcsin(\omega h/2) \tag{4}$$

and is found by asking that a normally incident grid wave be absorbed. It is a standard argument to show the convergence of the solution of (3) to that of (1) as $h \to 0$.

The question to be treated in this paper is that of finding the solution of the algebraic system defined by (3), after (3b) has been applied to eliminate from (3a) parameters corresponding to points outside Ω_h, in an effective and computationally efficient fashion. In addition to having a complex–valued solution, (3) is neither Hermitian symmetric nor coercive; as a consequence, most standard iterative methods either fail to converge or converge so slowly as to be impractical. The purposes here are to define an alternating–direction iteration procedure, indicate a method for choosing iteration parameters so as to assure convergence, and to present briefly an application of the method to the wave problem (2). See [1], [5], [6], [2], [4], [3], [8], and [9] for various discussions of alternating-direction iteration methods for finite difference or finite element procedures for elliptic problems with real solutions.

2. An Alternating–Direction Iteration Method

Alternating–direction iteration procedures are derived from time–stepping methods for parabolic analogues of the elliptic problem being solved. Here, we shall consider a direct extension of the classical method [1], [8] for the Dirichlet problem. Denote a cycle of iteration parameters (i.e., (reciprocals of) pseudo–time–steps) by

$$\rho_m \in \mathbf{C}, \quad m = 1,\ldots,M, \tag{5}$$

where M is the cycle length. Note that ρ_m can (and will) be complex; it will also be the case that $\mathrm{Re}(\rho_m)$ will be negative for some of the parameters. Let $u^{(0)}$ be an initial guess for the solution of (3) on Ω_h, and define iterates $u_h^{(m)}$, $m = 1,\ldots,M$, through first an x–sweep given by

$$\rho_m(u_h^{m-1/2)} - u_h^{(m-1)}) - \left(\delta_x^2 + \tfrac{1}{2}\omega^2\right)u_h^{(m-1/2)} \tag{6a}$$
$$-\left(\delta_y^2 + \tfrac{1}{2}\omega^2\right)u_h^{(m-1)} = f_h \qquad \text{on } \Omega_h,$$
$$\partial_\nu u_h^{(m-1/2)} + i\sin\xi h u_h^{(m-1/2)} = 0 \qquad \text{on } \Gamma_h^1, \tag{6b}$$
$$\partial_\nu u_h^{(m-1)} + i\sin\xi h\, u_h^{(m-1)} = 0 \qquad \text{on } \Gamma_h^2, \tag{6c}$$

where Γ_h^1 consists of the boundary points along $\{x = 0\}$ and $\{x = 1\}$ and Γ_h^2 those along $\{y = 0\}$ and $\{y = 1\}$, followed by a y–sweep:

$$\rho_m(u_h^{(m)} - u_h^{(m-1/2)}) - \left(\delta_x^2 + \tfrac{1}{2}\omega^2\right)u_h^{(m-1/2)} \tag{7a}$$
$$-\left(\delta_y^2 + \tfrac{1}{2}\omega^2\right)u_h^{(m)} = f_h \qquad \text{on } \Omega_h,$$

$$\partial_\nu u_h^{(m)} + i \sin \xi h \, u_h^{(m)} = 0 \qquad \text{on } \Gamma_h^2, \tag{7b}$$

$$\partial_\nu u_h^{(m-1/2)} + i \sin \xi h \, u_h^{(m-1/2)} = 0 \qquad \text{on } \Gamma_h^1. \tag{7c}$$

In the computational algorithm, the boundary conditions are used to eliminate algebraically values associated with the external grid points, so that only values associated with Ω_h are computed.

The choice of the iteration parameters is facilitated by considering the equations for the error

$$e_h^{(m)} = u_h - u_h^{(m)}, \quad m = 0, \frac{1}{2}, 1, \ldots, M. \tag{8}$$

First, $e_h^{(0)}$ is arbitrary on Ω_h, with the bordering values of $u_h^{(0)}$ (3b). Then, for $m = 1, \ldots, M$, $e_h^{(m)}$ satisfies the homogeneous equations associated with (6) and (7) obtained by setting $f_h = 0$. As in the case of the Dirichlet problem, a tensor–product eigenfunction expansion can be carried out for the error [3]. The argument below pertains to the case in which the x–and y–discretizations are identical; different discretizations could be handled analogously.

Let A be the tridiagonal matrix

$$A = \begin{bmatrix} b_0 & c_0 & & \\ a_1 & b_1 & c_1 & \\ \ddots & \ddots & & \\ & & a_N & b_N \end{bmatrix}, \tag{9}$$

where

$$a_n = \begin{cases} -1, & n = 1, \ldots, N-1, \\ -2, & n = N, \end{cases}$$

$$b_n = \begin{cases} 2, & n = 1, \ldots, N-1, \\ 2(1 + ih \sin \xi h), & n = 0 \text{ or } N, \end{cases}$$

$$c_n = \begin{cases} -2, & n = 0, \\ -1, & n = 1, \ldots, N-1. \end{cases}$$

A corresponds to the operator $-h^2\delta_x^2$, subject to the absorbing boundary condition. Now, solve the eigenvalue problem

$$A\varphi = \mu\varphi; \tag{10}$$

call the resulting eigenvalue–eigenfunction pairs (μ_n, φ_n) where the eigenvalues have been ordered so that

$$\mathrm{Re}\mu_0 \leq \mathrm{Re}\mu_1 \leq \ldots \leq \mathrm{Re}\mu_N. \tag{11}$$

For the Helmholz problem, in contrast to the Dirichlet problem [3], the actual eigenvalues must be found to carry out the calculations; the eigenfunctions are not required in the calculations, though their existence will enter the analysis. The eigenvalues can

be found with sufficient accuracy using public–domain codes, such as ones available in EISPACK or LAPACK. It should be noted that it is easy to see that both $\mathrm{Re}\mu_n$ and $\mathrm{Im}\mu_n$ are nonnegative. Moreover, $\mu_k \neq \mu_\ell$ for $k \neq \ell$, so that the eigenfunctions are complete.

Set

$$\psi_{pq} = \varphi_p(x)\varphi_q(y), \quad p, q = 0, \ldots, N. \tag{12}$$

Then, expand $e_h^{(m)}$ in the form

$$e_h^{(m)} = \sum_{p,q=0}^{N} \alpha_{pq}^{(m)} \psi_{pq}, \quad m = 0, \frac{1}{2}, 1, \ldots, M. \tag{13}$$

Then, if

$$\zeta_m = \rho_m h^2 \tag{14}$$

and

$$\lambda_p = \mu_p - \frac{1}{2}\omega^2 h^2, \tag{15}$$

it follows from (6) and (7) that

$$\alpha_{pq}^{(m)} = \alpha_{pq}^{(m-1)} \frac{\zeta_m - \lambda_p}{\zeta_m + \lambda_p} \cdot \frac{\zeta_m - \lambda_q}{\zeta_m + \lambda_q}. \tag{16}$$

Let

$$R(\zeta, \lambda) = \frac{\zeta - \lambda}{\zeta + \lambda}. \tag{17}$$

Then,

$$\alpha_{pq}^{(M)} = \alpha_{pq}^{(0)} \prod_{m=1}^{M} R(\zeta_m, \lambda_p) R(\zeta_m, \lambda_q), \tag{18}$$

a relation that is completely analogous to the error reduction relation arising in the treatment of the Dirichlet problem. However, the eigenvalues $\{\lambda_p\}$ no longer lie on a subinterval of $(0, \infty)$; in fact, for the more interesting values of ω, at least one λ_p has negative real part and all have positive imaginary part. Assume that

$$\mathrm{Re}\lambda_0 < \mathrm{Re}\lambda_1 < \ldots < \mathrm{Re}\lambda_r \leq 0 < \mathrm{Re}\lambda_{r+1} < \ldots < \mathrm{Re}\lambda_N. \tag{19}$$

Taking a pseudo–time–step ζ with positive real part gives growth for $p = 0, \ldots, r$; i.e.,

$$|R(\zeta, \lambda_p)| = \left| \frac{\zeta - \lambda_p}{\zeta + \lambda_p} \right| \geq 1, \qquad p = 0, \ldots, r, \tag{20}$$

while taking $\mathrm{Re}\zeta < 0$ gives growth for the remaining $|R(\zeta, \lambda_p)|$. In general, there is no choice of ζ that is stable for all modes, and it is necessary to use a cycle of different pseudo–time–steps in order to obtain convergence.

Assume for the moment that exact arithmetic is used in a calculation; of course, this is not actually feasible and it will be necessary to modify the parameter sequence in a

small way to avoid excessive rounding problems. These modifications will be discussed in the next section.

Let $M > r$, and choose

$$\zeta_m = \lambda_{m-1}, \quad m = 1, \ldots, r+1. \tag{21}$$

Thus, after $r+1$ double sweeps, the error associated with ψ_{pq} for $\min(p,q) \leq r$ will be totally eliminated, since at least one of the $R(\zeta_m, \lambda_p)$'s or $R(\zeta_m, \lambda_q)$'s vanishes, and the remaining pseudo–time–steps can be chosen so as to reduce the error associated with the collection of eigenfunctions ψ_{pq} with $\min(p,q) > r$. It should be noted that the number r is independent of h for sufficiently small h; thus, r/N tends to zero as $h \to 0$. When $\min(p,q) > r$, ψ_{pq} is a stable mode for pseudo–time–steps with positive real parts. For large p (i.e., p close to N), $\mathrm{Re}\lambda_p \approx 4$ and $\mathrm{Im}\lambda_p \ll \mathrm{Re}\lambda_p$. It is easy to see, using an argument similar to the one employed for the Dirichlet problem, that, given $\varepsilon > 0$, a set $\{\zeta_m : m = r+2, \ldots, M\}$, with

$$M - r = \mathcal{O}\Big(\log \frac{1}{h} \cdot \log \frac{1}{\varepsilon} \Big), \tag{22}$$

can be constructed [8], [3] so that

$$\max_{p>r} \prod_{m=r+2}^{M} \Big| \frac{\zeta_m - \lambda_p}{\zeta_m + \lambda_p} \Big| \leq \varepsilon. \tag{23}$$

If the cycle $\{\zeta_m : m = 1, \ldots, M\}$ is employed, then

$$\begin{aligned} &\max_{p,q>r} \Big| \prod_{m=1}^{M} R(\zeta_m, \lambda_p) R(\zeta_m, \lambda_q) \Big| \\ &\quad \leq \varepsilon^2 \max_{p,q>r} \Big| \prod_{m=1}^{r+1} R(\zeta_m, \lambda_p) R(\zeta_m, \lambda_q) \Big| \\ &\quad = K_1 \varepsilon^2; \end{aligned} \tag{24}$$

K_1 is computable, since the eigenvalues $\{\mu_p\}$ were found as the first step in the iterative procedure.

Theoretically, we can assure convergence by the argument that follows; the finite word length that necessarily arises in any actual computation will force a modification of the choice of iteration parameters.

Norm the vector $e_h^{(m)}$ by

$$\|e_h^{(m)}\| = \Big[\sum_{p,q=0}^{N} |\alpha_{pq}^{(m)}|^2 \Big]^{\frac{1}{2}}. \tag{25}$$

Then, if the cycle $\{\zeta_1, \ldots, \zeta_M\}$ above is used,

$$\begin{aligned} \|e_h^{(M)}\|^2 &= \sum_{p,q} |\alpha_{pq}^{(0)}|^2 \Big| \prod_{m=1}^{M} R(\zeta_m, \lambda_p) R(\zeta_m, \lambda_q) \Big|^2 \\ &\leq K_1^2 \varepsilon^4 \sum_{p,q>r} |\alpha_{pq}^{(0)}|^2. \end{aligned} \tag{26}$$

Take the iteration parameters for succeeding cycles to be $\{\zeta_{r+2}, \ldots, \zeta_M\}$; then

$$\|e_h^{(M+j(M-r-1))}\|^2 \le K_1^2 \varepsilon^{4(j+1)} \sum_{p,q>r} |\alpha_{pq}^{(0)}|^2, \tag{27}$$

so that convergence takes place.

3. Some Experimental Observations

The inclusion of the values λ_m, $m = 0, \ldots, r$, in the cycle of iteration parameters was found to be necessary in the trial calculations, but simply taking all of them first, as indicated in (21), was not found to be satisfactory. In fact, in some preliminary experiments simply taking the first $r+1$ parameters equal to the collection of λ_p having negative real parts led to strong divergence. Our first attempt to control this problem was to add the first $r+1$ λ's with least positive real parts to the iteration parameter sequence; if r is small, then this works reasonably well. But, if $r > 10$, say, this did not necessarily produce convergence when the word length was set at "complex*16", though it did in all our tests when we shifted to "complex*32". Since the extended arithmetic is very slow on any commonly available computer, this was also unsatisfactory, but it did show that the problem was due to rounding. Since the rounding problem was caused by the unstable growth of the coefficients of the modes that are stable for advancing pseudo–time, we tried the procedure that follows.

Assume $r+1$, the number of λ_p such that $\mathrm{Re}\lambda_p \le 0$, to be positive. Then, include in the cycle not only λ_p, $0 \le p \le r$, but also λ_p, $r+1 \le p \le 2r+2$. Alternate these parameters, one with negative real part and then one with positive real part. Then, add a modest number, say six to twelve, real ζ_m distributed geometrically between $\mathrm{Re}\lambda_{2r+3}$ and 4. In the experiments, the entire cycle, not just those pseudo–time–steps with positive real parts, was repeated. This somewhat arbitrary rule for selecting iteration parameters has proved experimentally to be practical for obtaining rapid and effective convergence.

4. Modifications to Include Attenuation

Attenuation leads to a modification in both the differential equation (1a) and the first-order absorbing boundary condition (1b), which change under the addition of a generalized friction to

$$-\Delta u - \Big(\omega^2 + ib(\omega)\omega\Big)u = \quad f \quad \text{on } \Omega, \tag{28a}$$

$$u_\nu + i\alpha(\omega) = \quad 0 \quad \text{on } \partial\Omega, \tag{28b}$$

where

$$\alpha(\omega) = \frac{\omega}{\sqrt{2}}\left(1 + \left(1 + \omega^{-2}b(\omega)^2\right)^{\frac{1}{2}}\right)^{\frac{1}{2}} - i\frac{b(\omega)}{\sqrt{2}}\left(1 + \left(1 + \omega^{-2}b(\omega)^2\right)^{\frac{1}{2}}\right)^{-\frac{1}{2}}. \tag{29}$$

If the function $b(\omega)$ is independent of ω, then by inverse Fourier transformation (28a) corresponds to the differential equation

$$v_{tt} + bv_t - \Delta v = g. \tag{30}$$

However, the boundary condition (28b) fails to reduce to the Fourier transform of a differential condition; thus, it is not a local-in-time condition, so that a pseudo-differential problem is generated even in this special case if an absorbing boundary condition is imposed to limit the computational domain. If $b(\omega)$ is not a polynomial in ω, then the problem is inherently pseudo-differential in nature.

The discrete problem is given by

$$-\Delta_h u_h - \Big(\omega^2 + ib(\omega)\omega\Big)u_h = \quad f_h, \quad (x_j, y_k) \in \Omega_h, \tag{31a}$$

$$\partial_\nu u_h + i\alpha_h u_h = \quad 0, \quad (x_j, y_k) \in \Gamma_h, \tag{31b}$$

where

$$\alpha_h = \alpha_h(\omega) = \sin \gamma h, \;\; \sin^2\left(\frac{\gamma h}{2}\right) = \frac{h^2}{4}(\omega^2 - i\omega b(\omega)). \tag{32}$$

Alternating-direction iteration takes the same form (6)-(7) for (29) as for the unattenuated equations (3), with the obvious substitutions of $\frac{1}{2}(\omega^2 + ib(\omega)\omega)$ for $\frac{1}{2}\omega^2$ and α_h for $\sin \xi h$. A choice of iteration parameters can be made in a fashion analogous to the procedure outlined in the simpler case. Let the matrix A, defined by (9), be changed only by replacing $\sin \xi h$ by α_h in the evaluation of two of its elements, and use the same notation $\{\mu_p\}$ for its eigenvalues, again ordered as in (11). Shift the definition of λ_p to be

$$\lambda_p = \mu_p - \frac{1}{2}h^2\Big(\omega^2 + i\alpha_h\omega\Big). \tag{33}$$

Then, the coefficients of the error remain propagated by (18). The same method for selecting iteration parameters for exact arithmetic discussed above leads to the same convergence estimate (27). Again, the rounding problem forces the choice of a cycle of parameters of the same type as was taken in the unattenuated case; note that the convergence proof remains applicable for the modified cycle. Experimental calculations confirmed the effectiveness of the practical algorithm; some results for approximating an attenuated wave will be presented in the next section.

5. An Application to Wave Approximation

The problem described by (2) and its generalization (28) to include attenuation were treated as examples of the applicability of the techniques introduced above. The domain Ω was the square $[-3.22, 3.22]^2$, and the source function was given by

$$g(x,t) = g_1(x)g_2(t),$$

where

$$g_1(x) \sim \delta_x + \delta_y,$$

and

$$g_2(t) = Cte^{-\alpha t} \sin \omega_0 t, \qquad t \geq 0.$$

The principal frequency ω_0 was taken to be 2π and α was taken equal to 1.5792. The spatial grid was rectangular, with 161 equal intervals on each side, so that the number of nodes per wave length for the principal frequency was 25. The Fourier transform of the source was truncated linearly between 15 and 20, and the Helmholz problems were solved at the midpoints of 120 equal intervals on $[0, 20]$; note that conjugate symmetry holds, so that this is equivalent to using 240 intervals on $[-20, 20]$. A maximum cycle length of 60 parameters was imposed; shorter cycles were used except at the high end of the spectrum, where the spectral density was quite small.

Figure 1 presents a snapshot of the unattenuated wave at time 3.175. The effect of an attenuation equivalent to the addition of v_t to (2a) can be seen by comparing Figure 2 to Figure 1. Figure 3 shows a wave at the same time, but with a source consisting of two dipole terms, one at the origin as in the previous two figures and the other along a diagonal. Interference between the two elementary waves can be observed.

Figures 4 and 5 present traces of displacements. Four receiver positions are indicated in Figure 4; note that the maximum amplitude of the trace at $(2.8, 2.8)$ is greater than those at the points $(1.8, 0)$ and $(2.7, 0)$, even though the latter points are closer to the source, reflecting the directed movement of the wave from a dipole source. Figure 5 shows the effect of attenuation on the trace at a given receiver.

More extensive experimental results are discussed for a single space variable problem treated by essentially the same approximation technique in [7].

REFERENCES

[1] Douglas, J., Jr. On the numerical integration of $u_{xx} + u_{yy} = u_t$ by implicit methods. *J. Soc. Indust. Appl. Math*, 3:42–65, 1955.

[2] Douglas, J., Jr. Alternating direction methods for three space variables. *Numerische Mathematik*, 4:41–63, 1962.

[3] Douglas, J., Jr., and Dupont, T. Alternating–direction Galerkin methods on rectangles. In *Numerical Solution of Partial Differential Equations II*, pages 133–214. Academic Press, New York, 1971. Burt Hubbard (ed.).

[4] Douglas, J., Jr., and Gunn, J. E. A general formulation of alternating direction methods, I. Parabolic and hyperbolic problems. *Numerische Mathematik*, 6:428–453, 1964.

[5] Douglas, J., Jr., and Peaceman, D. W. Numerical solution of two dimensional heat flow problems. *A.I.Ch.E. Jour*, 1:505–512, 1955.

[6] Douglas, J., Jr., and Rachford, H. H., Jr. On the numerical solution of heat conduction problems in two and three space variables. *Trans. Amer. Math Soc*, 82:421–439, 1956.

[7] Douglas, J., Jr., Santos, J. E., Sheen, D. and Bennethum, L. S. Frequency domain treatment of one–dimensional scalar waves. *Mathematical Models and Methods in Applied Sciences*, 1993. to appear.

[8] Peaceman, D. W. and Rachford, H. H., Jr. The numerical solution of parabolic and elliptic differential equations. *J. Soc. Ind. Appl. Math*, 3:28–41, 1955.

[9] Pearcy, C. M. On convergence of alternating direction procedures. *Numerische Mathematik*, 4:172–176, 1962.

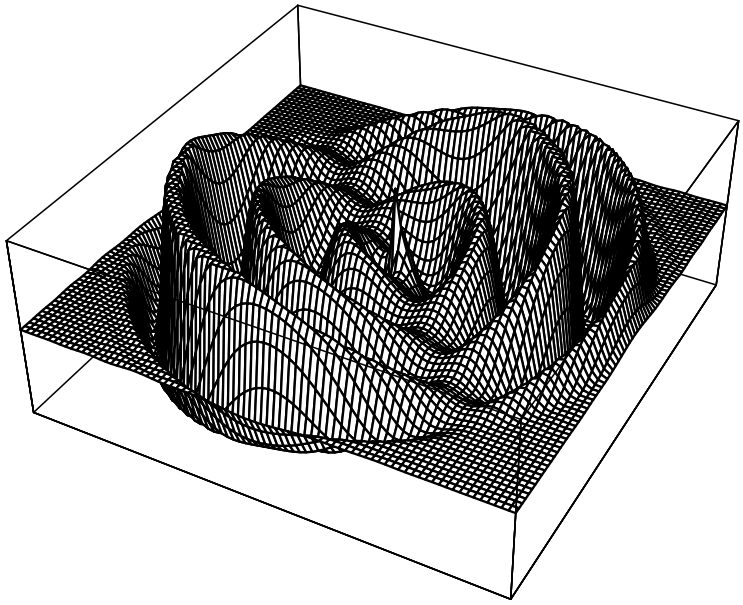

Figure 1: 3D plot of wave when $t = 3.175$. No attenuation.

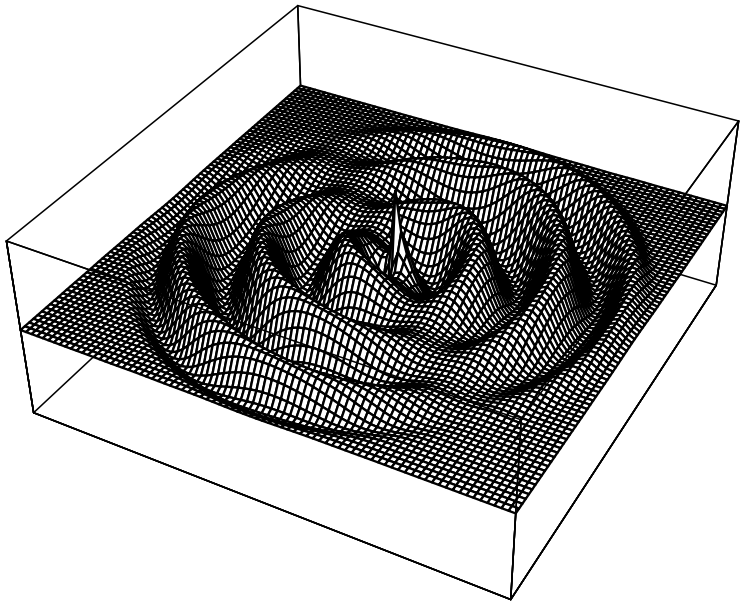

Figure 2: 3D plot of wave when $t = 3.175$. Attenuation is $b(\omega) = 1.0$.

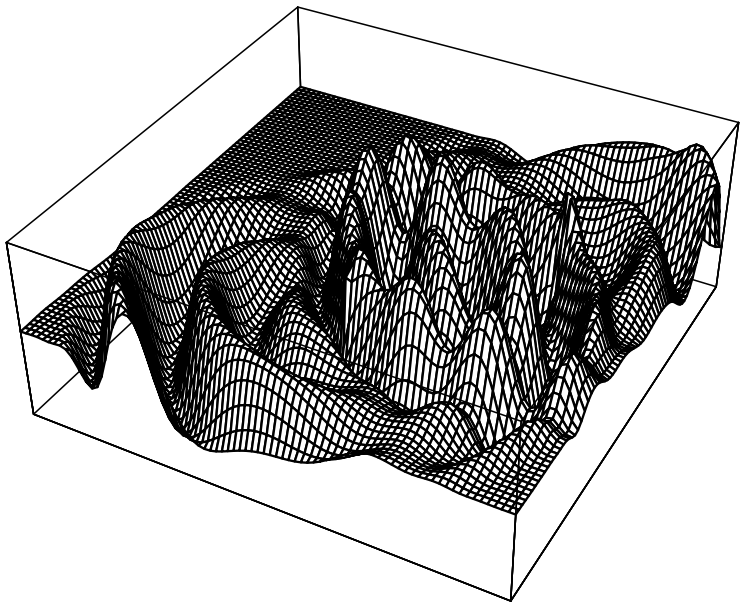

Figure 3: 3D plot at $t = 3.175$ with two sources and no attenuation.

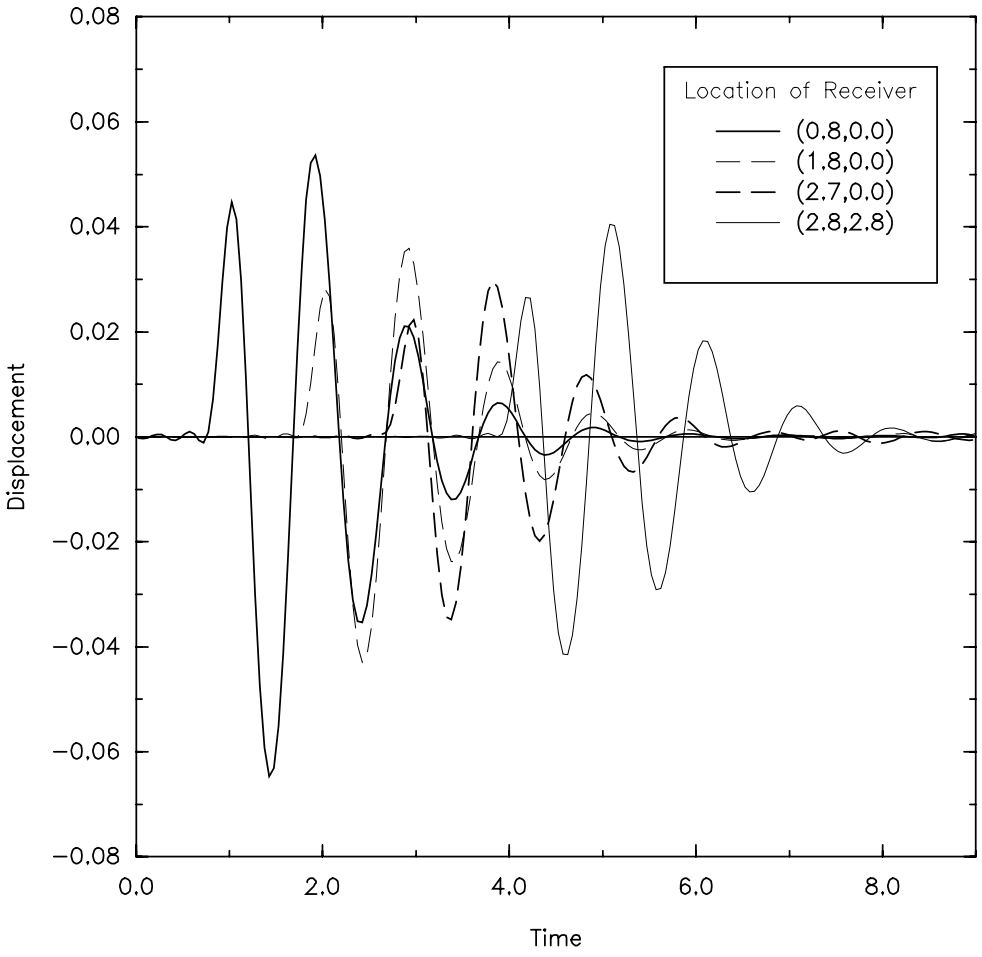

Figure 4: Traces at four different receivers in the domain. The source is at the center of the domain and there is no attenuation.

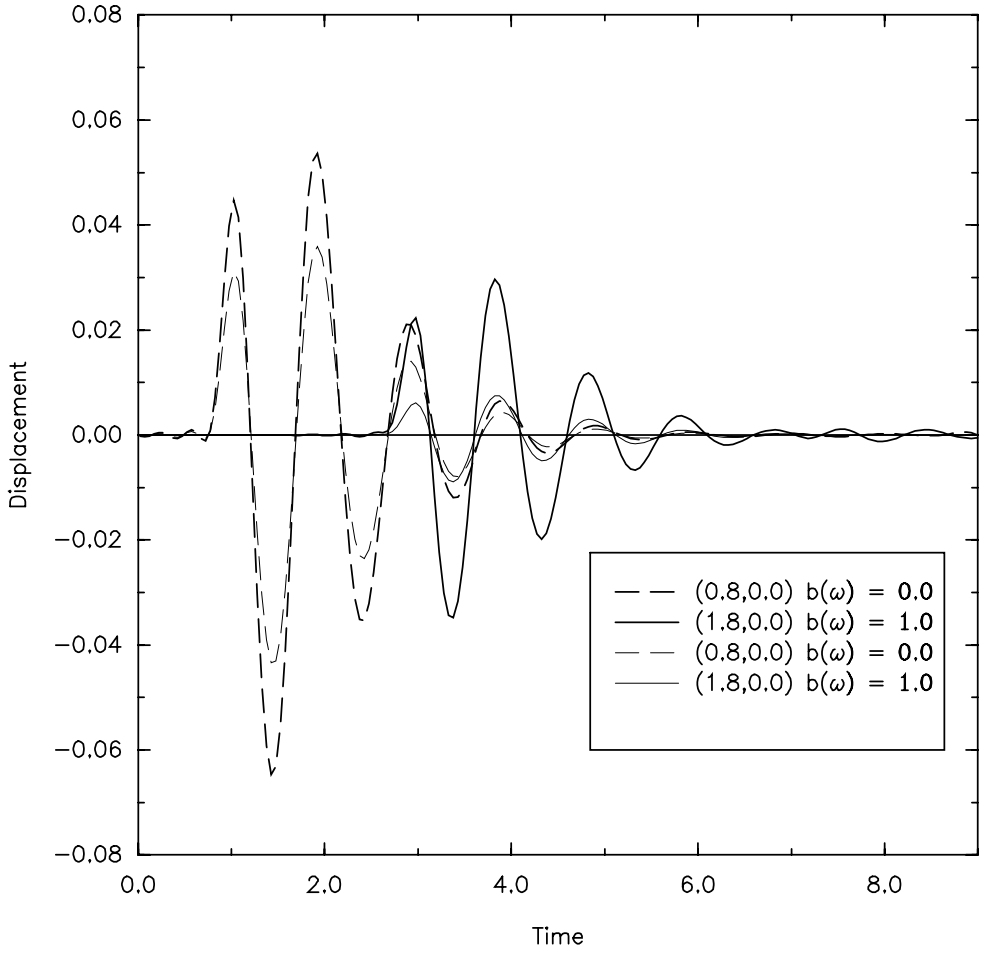

Figure 5: Traces at two different receivers for the two cases where $b(\omega) = 0$ and $b(\omega) = 1$.