Project 1 - TMA4212
Numerical Solution of Differential Equations by Difference Methods

Mari Norfolk
Lisa Marie F. Austad
Ola Rønnestad Abrahamsen

February 2024

1 Introduction

This paper develops and implements finite difference schemes for elliptic 2d-problems in complex domains. In particular we are going to analyse an equation that models the heat distribution in anisotropic and isotropic materials, that is materials where the flow of heat is uneven and even, respectively.

The Poisson equation is used to model the stationary temperature distribution T of a solid Ω. If the heat conductivity is κ, and the solid has internal heat sources f (an energy density), then conservation of energy, Fourier's law for the heat flux, and stationarity ($\partial_t T = 0$), give the following model

$$-\nabla \cdot (\kappa \nabla T) = f \quad \text{in } \Omega. \tag{1}$$

1.1 Heat distribution in anisotropic materials

In anisotropic materials, the heat flows faster in some directions than others. We will focus on two-dimensional models with two distinguished directions for the heat flow:

$$\vec{d_1} = (1,0) \quad \text{and} \quad \vec{d_2} = (1,r) \quad \text{where } r \in \mathbb{R}.$$

After normalization, this gives a heat conductivity of the form

$$\kappa = \begin{pmatrix} a+1 & r \\ r & r^2 \end{pmatrix} \quad \text{and} \quad R := \frac{a}{|\vec{d_2}|^2} = \frac{a}{1+r^2}$$

where $a > 0$ is a constant and R is the relative strength of the conductivity in the $\vec{d_1}$ versus $\vec{d_2}$ direction. In this case

$$\nabla \cdot (\kappa \nabla T) = (a+1)\partial_x^2 u + 2r\partial_x\partial_y u + r^2\partial_y^2 u = a\partial_x^2 u + (\vec{d_2} \cdot \nabla)^2 u = -f. \tag{2}$$

1.2 Heat distribution in isotropic materials

In isotropic materials, the heat flows evenly in all directions, such that we solve the same problem only with $\kappa = I$. Thus,

$$\nabla \cdot (\kappa \nabla T) = \Delta T = -f \quad \text{in } \Omega.$$

One can see that this equation simplifies to the well-known Poisson equation, $\Delta u = \partial_x^2 u + \partial_y^2 u$. We will investigate this problem on an irregular grid, and we will explore two different methods of solving as well as discussing how these methods compare to each other.

2 Numerical model

2.1 Unit grid

Firstly, we will consider a simple domain $\Omega = [0,1] \times [0,1]$ (unit square) with Dirichlet boundary conditions $u = g$ on $\partial\Omega$. Further, in the first couple tasks we let $r = 1$ and derive the scheme to solve the problem numerically using a finite difference scheme with second order central differences. Since we aim to avoid discretized mixed derivatives, we do the discretization in the $\vec{d_1}$- and $\vec{d_2}$-direction.

As the observant reader might have realised, the $\vec{d_1}$-direction is simply the $\vec{x}$-direction. The resulting discretization for the $\vec{d_1}$-derivative is therefore,

$$\frac{\partial^2 u}{\partial x^2} = \frac{U_{i+1,j} - 2U_{i,j} + U_{i-1,j}}{h^2},$$

and for the $\vec{d_2}$-derivative we obtain

$$(\vec{d_2} \cdot \nabla)^2 u = (\sqrt{2}\frac{\vec{d_2}}{|\vec{d_2}|} \cdot \nabla)^2 u = D^2_{\vec{d_2}} u = \frac{U_{i+1,j+1} - 2U_{i,j} + U_{i-1,j-1}}{k^2}.$$

Thus, we obtain the following scheme with $k = |r|h \overset{r=1}{=} h$ is

$$a\partial_x^2 u + (\vec{d_2}\nabla)^2 u = a\frac{U_{i+1,j} - 2U_{i,j} + U_{i-1,j}}{h^2} + \frac{U_{i+1,j+1} - 2U_{i,j} + U_{i-1,j-1}}{k^2} = -f. \tag{3}$$

2.2 Irregular grid

Now we let Ω be the domain in the first quadrant enclosed by the curve $y = h(x) := \frac{1}{2}(\cos(\pi x) + 1)$ and the two axes. The boundary, $\partial\Omega$, then consists of the curves

$$\gamma_1 = [0, 1] \times \{0\}, \gamma_2 = \{0\} \times [0, 1], \gamma_3 = \{(x, h(x)) : x \in [0, 1]\}.$$

Further, we assume we are observing the isotropic case discussed in section 1.2, where

$$\nabla \cdot (\kappa \nabla T) = \Delta T = (\partial_x^2 + \partial_y^2)T,$$

which yields the scheme

$$\partial_x^2 u + \partial_y^2 u = \frac{U_{i+1,j} - 2U_{i,j} + U_{i-1,j}}{h^2} + \frac{U_{i,j+1} - 2U_{i,j} + U_{i,j-1}}{k^2} = -f$$

2.2.1 Fattening the boundary

To deal with the inconsistent boundary, one possibility is to "fatten" the boundary, i.e. extending the boundary conditions. We solved this by projecting the outside point onto the boundary:

$$U_P = g(\pi_{\partial\Omega}(P)), P \in \partial\Omega, \pi_{\partial\omega}(P) \in \partial\Omega$$

However, by doing this we introduce an error that comes from the projection of the point onto the boundary.

$$e_P = u_P - U_P = g(\pi_{\partial\Omega}(P)) + (P - \pi_{\partial\Omega}(P))u_x(\xi) - g(\pi_{\partial\Omega}(P) = (P - \pi_{\partial\Omega}(P))u_x(\xi) = O(h)$$

Thus, we can expect linear convergence for a method using fattening the boundary, as the projection error will dominate.

2.2.2 Modifying the discretisation near the boundary

Another way to deal with the irregular grid is to modify the boundary. Here we introduce new points on the boundary to keep the original stencil, see stencil 2. Doing this we need to modify the scheme to adapt for the differences in stepsizes around the boundary. Denote h_E and h_N for the stepsize in east and north direction.

$$\partial_x^2 u_p - \tau_{x,p} = a_{E'} u_{E'} + a_p u_p + a_w u_w$$

Then we can use Taylor expansion and the method of undetermined coefficients and arrive at:

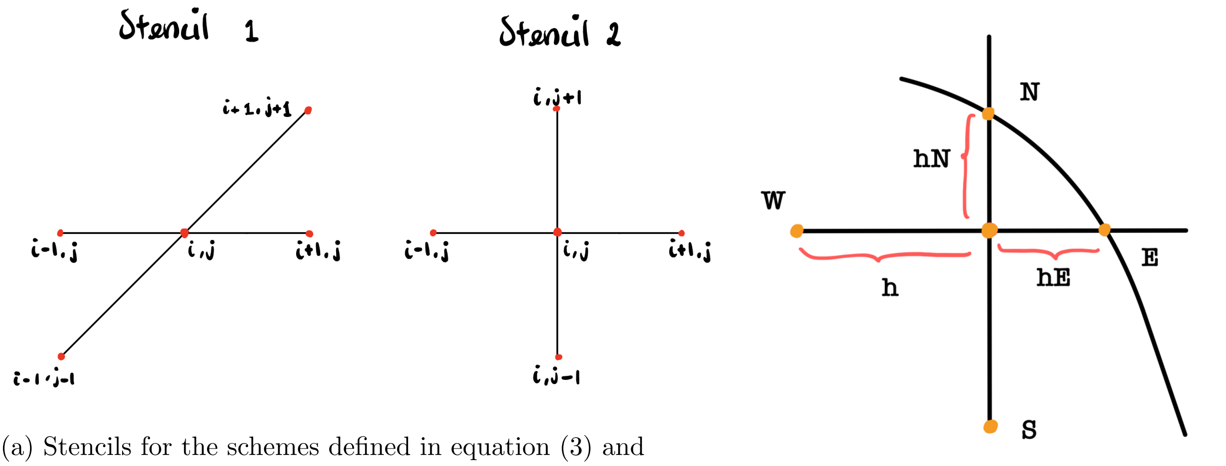
$$\begin{cases} a_{E'} + a_P + a_W = 0 \\ a_{E'}\psi h - h a_W = 0 \\ \frac{1}{2}a_{E'}(\psi h)^2 + \frac{1}{2}a_W h^2 = 1 \end{cases} \implies \begin{cases} a_{E'} = \frac{2}{h^2}\frac{1}{\psi(\psi+1)} \\ a_P = -\frac{2}{h^2}\frac{1}{\psi} \\ a_W = \frac{2}{h^2}\frac{1}{\psi+1} \end{cases}$$

Giving the new formula:

$$\frac{2}{h^2}\left(\frac{U_E}{\psi(\psi+1)} + \frac{U_W}{(\psi+1)} + \frac{U_N}{\phi(\phi+1)} + \frac{U_S}{(\phi+1)} - \frac{\psi+\phi}{\psi\phi}U_p\right), \psi = \frac{h_E}{h} \text{ and } \phi = \frac{h_N}{h} \tag{4}$$

2.3 Stencils and matrix form

The stencils for the various schemes discussed are displayed in the figure below, showing the points P at the respective indices.



(a) Stencils for the schemes defined in equation (3) and equation (4), respectively.

(b) Stencil for MTB on irregular domain

The schemes can also be expressed in matrix form with

$$A_h\mathbf{U} = h^2\mathbf{f} + \mathbf{g},$$

where $A_h = \text{blocktridiag}\{-L_{(M-1)}, B, -U_{(M-1)}\}$. Here $L_{(M-1)}$ and $U_{(M-1)}$ are the lower and upper shifting matrices, and $B = \text{tridiag}\{-a, 2a+2, -a\} \in \mathbb{M}_{(M-1)\times(M-1)}(\mathbb{R})$. Here $\mathbf{f}$ and $\mathbf{g}$ are initial and boundary conditions.

3 Analysis

Subsequently, we want to prove convergence of the scheme and discuss what changes when r is chosen arbitrarily. In order to prove convergence, we first need to prove stability and consistency.

3.1 Monotonicity

Firstly, to prove stability we start by proving monotonicity.

Definition: *A scheme is considered to be monotonic, thus have positive coefficients, if the scheme is on the form*

$$-\mathcal{L}_h U_P = \alpha_{PP}U_P - \sum_{Q\neq P}\alpha_{PQ}U_Q = f_P, \quad P \in \Omega,\ Q \in \Omega \cup \partial\Omega = \overline{\Omega}$$

satisfying

$$\alpha_{PP} > 0, \quad \alpha_{PQ} \geq 0, \quad \alpha_{PP} \geq \sum_{Q\neq P}\alpha_{PQ} \quad \textit{for all } P \in \Omega.$$

Claim: *The scheme defined in equation* (3) *is monotone.*

Proof: The scheme is on the form

$$-\mathcal{L}_h U_P = \frac{1}{h^2}((2a+2)U_{i,j} - aU_{i+1,j} - aU_{i-1,j} - U_{i+1,j+1} - U_{i-1,j-1}) = f_P,$$

in which the coefficients are

$$\alpha_{PP} = \frac{2a+2}{h^2}, \quad \alpha_{PQ} = \frac{a}{h^2} \quad \text{and} \quad \alpha_{PL} = \frac{1}{h^2}.$$

and since $a > 0$, the coefficients satisfy,

$$\alpha_{PP}, \alpha_{PQ}, \alpha_{PL} > 0$$

$$2a + 2 = \alpha_{PP} \geq \sum_{Q\neq P}\alpha_{PQ}a + a + 1 + 1 = 2a + 2.$$

Thus, the claim holds and the monotonicity of the scheme is proven. □

3.2 L^∞-stability

Further, we need to show that the scheme is L^∞-stable. To achieve this we first want to introduce the following theorem, as well as the supersolution ϕ.

Theorem 1 (DMP): The discrete maximum principle (DMP) for elliptic PDEs is formulated as Let $\mathcal{L}_h$ be the differential operator on the scheme. If

$$-\mathcal{L}_h U_P \leq 0, \quad \forall P \in \Omega \implies \max_{P\in\Omega} U_P \leq \max_{P\in\partial\Omega} U_P.$$

The supersolution $\phi(x) = \frac{1}{2}x(1-x)$ which is a solution to our equation with the following properties

$$\phi_P \geq 0 \text{ for all } P \in \Omega$$
$$-\mathcal{L}_h \phi_P = (a+1)(\geq 1), \forall P \in \Omega.$$

Now we are fully equipped to show the stability of this scheme.

Definition: *For any given any function f and a solution U to our discrete problem satisfying $U_P = 0\ \forall P$ in the boundary $\partial\Omega$, the scheme is stable with respect to the right-hand side if there exists a constant $C > 0$, independent of both h and the right-hand side, such that $\|U\|_\infty \leq C\|f\|_\infty$.*

Claim: *The scheme defined in equation (4) is L^∞-stable.*

Proof: Let V_P be a solution to the right-hand side such that,

$$V_p = \begin{cases} -\mathcal{L}_h V_p = f_p, & \text{in } \Omega \\ V_p = 0, & \text{in } \partial\Omega \end{cases},$$

and define $W_P = V_P - \|f\|_\infty \phi_P$. Then,

$$-\mathcal{L}_h(W_P) = -\mathcal{L}_h V_P - \|f\|_\infty(-\mathcal{L}_h\phi_P) \overset{(-\mathcal{L}_h\phi_P)\geq 1}{\leq} f_P - \|f\|_\infty \leq 0$$

Thus, by Theorem 1 (DMP) we have

$$\max_{P\in\Omega} W_P \leq \max_{P\in\partial\Omega}(V_P - \|f\|_\infty\phi_P) \overset{V_P=0\,\forall P\in\Omega}{\leq} 0 \implies V_P \leq \|f\|_\infty\phi_P.$$

In order to prove the same for $-V_P$ we introduce the bijection α:

$$\alpha : (v, w, f) \to (-v, -w, -f) \implies -V_P \leq \|-f\|_\infty\phi_P = \|f\|_\infty\phi_P.$$

This shows that we get the same result for $-V_P$, and thus we can take the norm to obtain the claimed result. Finally, along with taking the maximum norm and that $\max_{P\in\Omega}\phi_P = \phi(\frac{1}{2}) = \frac{1}{8}$ we obtain the following stability result

$$\max\|V_P\| \leq \max_{P\in\Omega}\|f\|_\infty\phi_P = \max_{P\in\Omega}\phi_P\|f\|_\infty = \frac{1}{8}\|f\|_\infty \implies \|V_P\|_\infty \leq \frac{1}{8}\|f\|_\infty.$$

Thus, we have shown that the scheme is stable with $C = \frac{1}{8}$. □

3.3 Consistency

Definition: *A consistent scheme is one in which the truncation error tends to zero as h, $k \to 0$.*

Claim: *The scheme is consistent.*

Proof: The truncation error is

$$\tau_p = -\mathcal{L}u_p - (-\mathcal{L}_h u_p) = -(a\partial_x^2 u + \partial_x^2 u + 2\partial_x\partial_y u + \partial_y^2 u) - (-a(\frac{U_{i+1,j} - 2U_{i,j} + U_{i-1,j}}{h^2}) + \frac{U_{i+1,j+1} - 2U_{i,j} + U_{i-1,j-1}}{k^2})$$

The taylor expansion of the scheme can be obtained by the following,

$$U(x \pm h, y \pm k) = U_{i\pm 1, j\pm 1} \approx \sum_{m=0}^{n} \frac{1}{m!}\left(\pm h \frac{\partial}{\partial x} + \pm k \frac{\partial}{\partial y}\right)^m u(x,y) + r_n$$

where r_n represents the remainder term containing higher-order terms.

Using the abbreviation denoted earlier $h = |r|k = k$, the truncation error is then

$$\tau_p = (a+1)\frac{1}{12}h^2 u_{xxxx} + \frac{1}{3}h^2 u_{xxxy} - \frac{1}{2}h^2 u_{xxyy} + \frac{1}{3}h^2 u_{xyyy} + \frac{1}{12}h^2 u_{yyyy} + \mathcal{O}(h^3)$$

Let K $= (a+1)\|u_{xxxx}\|_{L^\infty(\Omega)} + 4\|u_{xxxy}\|_{L^\infty(\Omega)} - 6\|u_{xxyy}\|_{L^\infty(\Omega)} + 4\|u_{xyyy}\|_{L^\infty(\Omega)} + \|u_{yyyy}\|_{L^\infty(\Omega)} < \infty$.
Then we obtain the following

$$|\tau_P| < \frac{1}{12}Kh^2 < \infty. \tag{5}$$

Thus, $\tau_p \overset{h\to 0}{\to} 0$ and we have proven consistency and boundedness of the truncation error for the scheme when we have a regular domain with $h = k$. □

3.4 Error bound and convergence

Since $-\mathcal{L}_h e_p = -\tau_p \implies \|e_p\|_\infty \leq C\|\tau_p\|_\infty$. Thus, we can use the bound on the truncation error found in the previous section in equation (5), K as defined in section 3.3 and that $C = \max_{P\in\Omega}\phi_p = \frac{1}{8}$ to obtain the following error bound,

$$\|e_p\|_\infty \leq C\|\tau_p\|_\infty = \frac{1}{8}\frac{1}{12}Kh^2 = \frac{1}{96}Kh^2.$$

For a smooth function one can easily see that the convergence rate is 2. However, recall that in section 3.3 on consistency the order of consistency was obtained due to the cancellation of linear and constant terms when $h = k$. When the grid is irregular, i.e. $h \neq k$, this is not the case and as a result we get a linear convergence instead.

3.5 Choosing r arbitrarily

Further we want to introduce a grid with uneven step sizes. To do this we will use the step size $h = \frac{1}{M}$ in x-direction and the step size $k = |r|h$ in y-direction, where we allow for r to be irrational.

Claim: *The grid will miss the upper boundary at $y = 1$.*

Proof:

Assume r irregular.

$$1 = Nk = N|r|h = \frac{N}{M}|r| \implies |r| = \frac{M}{N},$$

which is a contradiction if r is irrational. □

To account for this we use the method described over, namely fattening the boundary in the y-direction. That is we will go past the boundary to the closest grid point, and then project this point down to the boundary. Then we are sure to hit the boundary, even when r is irrational.

4 Results and discussion

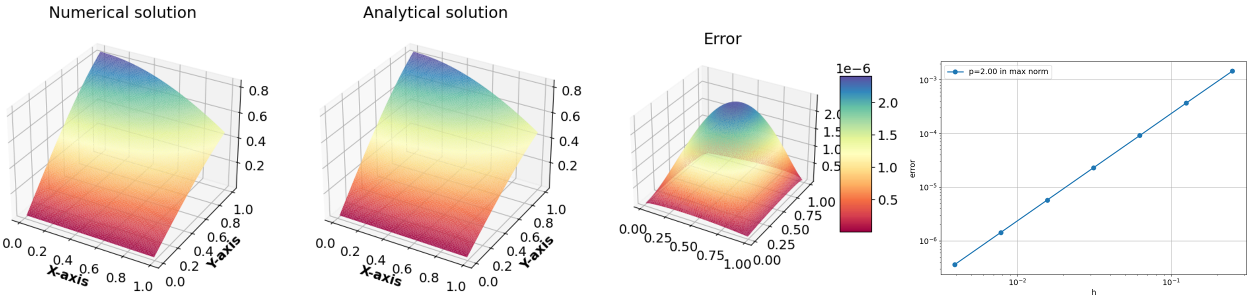
4.1 Anisotropic

Now we can implement the scheme defined in equation (3) and test it on suitable test functions.

Firstly, we choose the test function

$$u_1 = \cos(x)\sin(y).$$

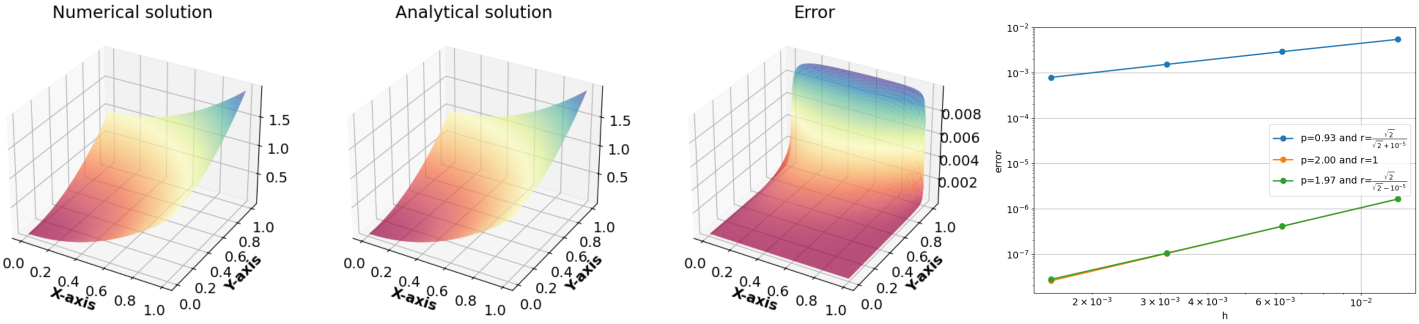
This gives the following plots for the heat distribution, error and convergence rate of the solver.



In the plots one can see that the analytical and numerical solution coincides with each other. One can also note that we get a very small maximum error, in the magnitude of 10^{-6}. From the convergence plot, our analysis in section 3.4 is verified, as we get quadratic convergence.

4.1.1 Error when fattening the boundary

In the error plot we use the test equation $u = x^2 + y^2$ with Dirichlet boundary conditions, since on a regular grid we can solve this exact. Whereas on the convergence plot we use $u = x^2 + x\sin(y)$.



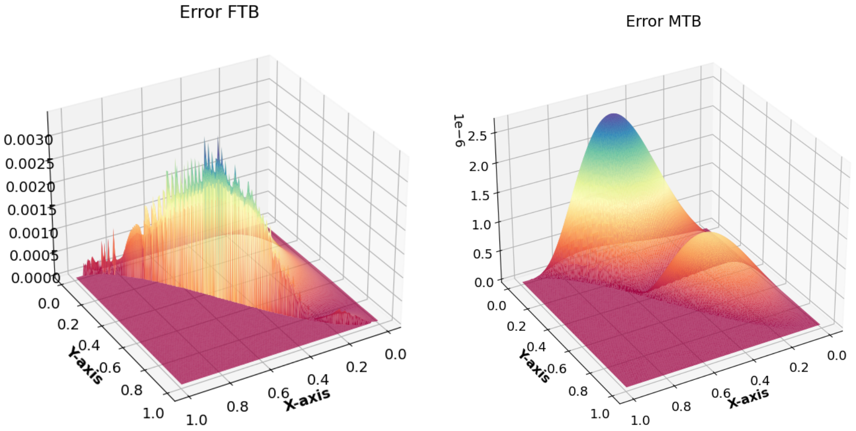
From the error plot one can see that the numerical and analytical solutions agree well, expect for on the fattened upper boundary, $y = 1$. Here we chose a problem that one could solve with very little error without irregular r, such that one can clearly see the error that the fattening imposes.

From the convergence plot, we observe that that our choice of r directly relates to the converge rate. Note that the orange line is hard to see as it is almost directly below the green.

For a r slightly above 1, we achieve a convergence rate slightly below 2. This is due to just slightly overshooting the boundary, and we are basically solving the problem like before with $r = 1$. However, once r is slightly below 1, we miss the boundary by a much higher margin. Thus we obtain a convergence rate slightly below 1, as the theory suggests.

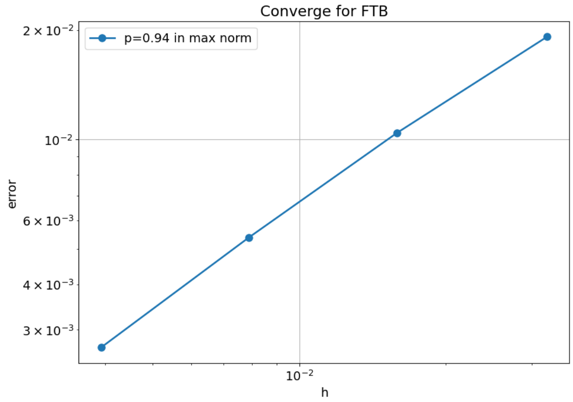
4.2 Isotropic case

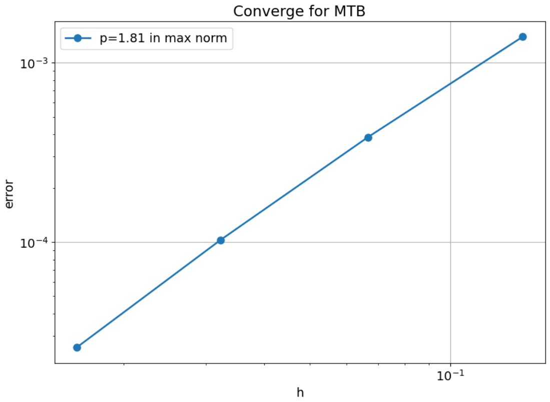
For the last part of the report we use the test equation $u = x^2\cos(5y)$.



Both of these plots are generated with $M = 200$. Here we denote fattening the boundary by FTB and modifying the discretisation on the boundary by MTB. The approach of modifying the discretisation on the boundary shows significantly higher accuracy compared to the approach of fattening the boundary, 10^{-6} vs. 10^{-3}. This outcome is rational as in MTB we solve exact by modifying the scheme, whereas in FTB one introduces an error related to the projection, similar to the anisotropic case.

The error using FTB is also a lot less smooth than in the MTB case. This can be reasoned by the fact that when using fattening the boundary we have an element of randomness in the error, consisting of how close to the original boundary our extended point is. The further away from the boundary the extended point is, the more error the method imposes in that point. Thus we have such a spike effect where the error can vary based on each point.





Using the MTB method we see that the error plot is a lot more smooth, more similar to the error plot of the scheme with the unit grid. This is rational as we now solve the boundary exact, but instead modify the scheme slightly. The convergence plot also showcases an convergence rate above one, but less than two.

So in this instance we have a second-order method, but with the modifications to the scheme we end up with an rate effectively between $O(h)$ and $O(h^2)$. Observations from the MTB error plot suggest that, for the grid resolutions achievable by our computational resources, the maximum error does not originate from the altered section of the scheme. Rather, it appears that the intrinsic error, also present in the solution on a regular domain, prevails. Given that the underlying scheme is characterized by quadratic convergence, we infer that for sufficiently large values of M, the boundary-related error would dominate due to the linear convergence rate in this region. However, within the computationally feasible range and for the problems we have studied, a definitive linear convergence has yet to be observed.

Our impression is that the FTB is easier to implement, as the code is exactly the same except for computing the projection. There is also no need to change the stencil or compute weights, which makes it easier.

For our implementation MTB was quicker than FTB. This can be reasoned by the fact that in FTB we also solve the optimisation problem of finding the closest point using the least squares method. Other than that the methods are quite similar. In our numerical solver we explicitly use a sparse matrix implementation and a sparse solver from scipy for the linear problem. This makes the most expensive part of the method a lot faster and reduces our run time in both cases a lot.

Metric	**MTB**	**FTB**
Mean time per loop	88 ms	140 ms
Standard deviation	±2.09 ms	±880 μs

Table 1: Performance Metrics from 10 runs with 30 loops each on an Apple M2Pro 10C16C-GPU.

5 Conclusion

In this project we have investigated and implemented various finite differences schemes in order to solve the Poisson equation on regular domains with rational and irrational directions, and irregular domains.