AN INTRODUCTION TO: DIFFERENTIAL EQUATIONS AND THEIR NUMERICAL APPROXIMATION

Todd Arbogast

Department of Mathematics
and
Center for Subsurface Modeling,
Institute for Computational Engineering and Sciences (ICES)

The University of Texas at Austin

Summer School in

Geophysical Porous Media:
Multidiscplinary Science from Nano-to-Global-Scale

July 17-28, 2006
Purdue University, West Lafayette, Indiana

Outline

	Topic	Main application
1.	Ordinary differential equations (ODE)	Reaction dynamics
2.	Elliptic Partial differential equations (PDE)	Steady state single phase flow or diffusive equilibrium
3.	Parabolic Partial differential equations	Tracer diffusion
4.	Hyperbolic Partial differential equations	Tracer transport

I. Ordinary Differential Equations (ODEs)

(for Reaction dynamics)

An Ordinary Differential Equation (ODE)

The problem is to find a function $\mathbf{u}(t)$ of a single independent variable t, which we will call "time," such that

$$\begin{cases} \mathbf{u}' = \mathbf{f}(\mathbf{u}, t), \quad t \geq 0, & \text{(the equation)} \\ \mathbf{u}(0) = \mathbf{u}_0. & \text{(the initial condition)} \end{cases}$$

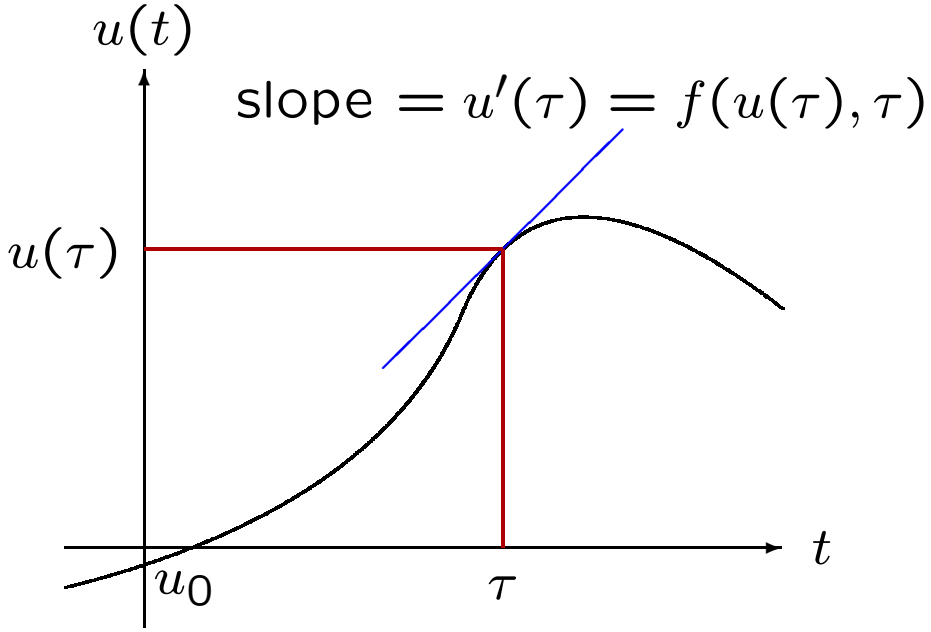
The unknown function $\mathbf{u}$ and the data $\mathbf{u}_0$, and $\mathbf{f}$ could be vectors in $\mathbb{R}^d$. Then we have a system of first order equations

$$\begin{cases} \begin{bmatrix} u_1' = f_1(u_1, \ldots u_d, t), \quad t \geq 0, \\ \vdots \\ u_d' = f_d(u_1, \ldots u_d, t), \end{bmatrix} \\ \begin{bmatrix} u_1(0) = u_{0,1}, \\ \vdots \\ u_d(0) = u_{0,d}. \end{bmatrix} \end{cases}$$

Meaning

The time rate of change of $\mathbf{u}$ depends on the time t and the present value of $\mathbf{u}$, i.e., $\mathbf{u}(t)$, but not on the past history, nor on the future.

Graphically: For $d = 1$,



As an integral: For $d = 1$,

$$u(t) = u_0 + \int_0^t f(u(\tau), \tau)\, d\tau.$$

Examples

Example: Radionuclide Decay

Let u be the mass of a radioactive substance.

Modeling assumption: The rate of change of the mass is proportional to the amount present. That is, for some $\lambda > 0$,

$$u' = -\lambda u.$$

Question: Why the negative sign?

Example: Radionuclide Decay Chains

Suppose we have the decay chain

$$\boxed{\text{Uranium-238} \longrightarrow \text{Thorium-234}}$$

Let u_1 be the amount of Uranium-238, decaying with constant λ_1,
u_2 be the amount of Thorium-234, decaying with constant λ_2.

Then

$$\begin{cases} u_1' = -\lambda_1 u_1, \\ u_2' = \lambda_1 u_1 - \lambda_2 u_2, \end{cases} \qquad \text{or} \qquad \mathbf{u}' = f(\mathbf{u}),$$

where

$$\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \qquad \text{and} \qquad f(\mathbf{u}) = \begin{pmatrix} -\lambda_1 u_1 \\ \lambda_1 u_1 - \lambda_2 u_2 \end{pmatrix} = - \begin{pmatrix} \lambda_1 & 0 \\ -\lambda_1 & \lambda_2 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}.$$

Question: Where does the Uranium-328 go?

Example: A Chemical Reaction

Suppose that

X is microbial biomass concentration,

C is a substrate concentration (food).

Single Monod kinetics:

$$\frac{dC}{dt} = -kX\frac{C}{K+C}.$$

Note that the slope is negative, so $C \to 0$ as $t \to \infty$.

1. If C is large, then
$$\frac{dC}{dt} \approx -kX.$$
and C decreases proportional to the number of microbes eating it.
2. If C is small, then
$$\frac{dC}{dt} \approx -(k/K)XC,$$
and the decrease of C is inhibited because microbes cannot find it.

Question: Is the system complete? What is missing?

Linear Equations

Solution of a Single Linear Equation

If $f(u) = -\lambda u$ is linear, then the equation is called linear. Now

$$\boxed{u' = -\lambda u,}$$

or

$$\frac{u'}{u} = -\lambda \implies \frac{d}{dt}\log|u| = -\lambda \implies \log|u| = \log|u_0| - \lambda t.$$

Thus, after exponentiating,

$$\boxed{u(t) = u_0 e^{-\lambda t}.}$$

Note that λ tells you how fast u decreases (or grows, if $\lambda < 0$).

Solution of Multiple Linear Equations

Now $\mathbf{f}(\mathbf{u}) = -A\mathbf{u}$, where A is a $d \times d$ matrix, so

$$\boxed{\mathbf{u}' = A\mathbf{u}.}$$

Suppose that A is diagonalizable. That is, there is a change of variables $\mathbf{v} = C\mathbf{u}$ so that

$$CAC^{-1} = D = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \lambda_d \end{pmatrix}.$$

Then the λ's are the eigenvalues of A, and

$$C\mathbf{u}' = CA\mathbf{u} \implies C\mathbf{u}' = CAC^{-1}C\mathbf{u} \implies \mathbf{v}' = D\mathbf{v},$$

which breaks into

$$v_i' = \lambda_i v_i \implies v_i = v_{0,i}\, e^{\lambda_i t},$$

where $\mathbf{v}_0 = C\mathbf{u}_0$. Thus

$$\boxed{\mathbf{u}(t) = C^{-1}\mathbf{v}.}$$

Application to Radionuclide Decay—1

Uranium-238 $\longrightarrow$ Thorium-234

$$f(\mathbf{u}) = A\mathbf{u} \quad \text{and} \quad A = \begin{pmatrix} -\lambda_1 & 0 \\ \lambda_1 & -\lambda_2 \end{pmatrix}.$$

The eigenvalues of A are $-\lambda_1$ and $-\lambda_2$. For our change of variables $\mathbf{v} = C\mathbf{u}$, the eigenvectors of A form the columns of C^{-1}:

$$C^{-1} = \begin{pmatrix} \lambda_2 - \lambda_1 & 0 \\ \lambda_1 & 1 \end{pmatrix} \quad \Longrightarrow \quad C = \frac{1}{\lambda_2 - \lambda_1} \begin{pmatrix} 1 & 0 \\ -\lambda_1 & \lambda_2 - \lambda_1 \end{pmatrix}.$$

Therefore

$$D = CAC^{-1} = \begin{pmatrix} -\lambda_1 & 0 \\ 0 & -\lambda_2 \end{pmatrix},$$

so

$$v_1 = v_{0,1} e^{-\lambda_1 t},$$
$$v_2 = v_{0,2} e^{-\lambda_2 t}.$$

Application to Radionuclide Decay—2

Now the initial $\mathbf{v}$ is

$$\mathbf{v}_0 = C\mathbf{u}_0 = \frac{1}{\lambda_2 - \lambda_1}\begin{pmatrix} u_{0,1} \\ -\lambda_1 u_{0,1} + (\lambda_2 - \lambda_1)u_{0,2} \end{pmatrix},$$

so

$$\begin{aligned} v_1 &= \frac{1}{\lambda_2 - \lambda_1} u_{0,1} e^{-\lambda_1 t}, \\ v_2 &= \frac{1}{\lambda_2 - \lambda_1}\Big(-\lambda_1 u_{0,1} + (\lambda_2 - \lambda_1)u_{0,2}\Big) e^{-\lambda_2 t}, \end{aligned}$$

and, since $\mathbf{u} = C^{-1}\mathbf{v}$,

$$\begin{aligned} u_1 &= u_{0,1} e^{-\lambda_1 t}, \\ u_2 &= u_{0,2} e^{-\lambda_2 t} + \frac{\lambda_1 u_{0,1}}{\lambda_2 - \lambda_1}\Big(e^{-\lambda_1 t} - e^{-\lambda_2 t}\Big) \\ &= \frac{\lambda_1 u_{0,1}}{\lambda_2 - \lambda_1} e^{-\lambda_1 t} + \left(u_{0,2} - \frac{\lambda_1 u_{0,1}}{\lambda_2 - \lambda_1}\right) e^{-\lambda_2 t}. \end{aligned}$$

Time scales. Note that there are two time scales $1/\lambda_1$ and $1/\lambda_2$, corresponding to the two eigenvalues of the matrix A.

Application to Radionuclide Decay—3

Half-life. The half-life is the time until half the substance decays away. Thus

$$1/2 = e^{-\lambda\tau} \qquad \Longrightarrow \qquad \tau = \log(2)/\lambda.$$

For us

	Uranium-238	Thorium-234
τ	4.5×10^9 years	24.5 days
λ	1.54×10^{-9}/ year	10.3/ year
$1/\lambda$	6.49×10^9 year	0.097 year

Thorium decay is almost instantaneous compared to Uranium! Thus

$$u_2 \approx \frac{\lambda_1 u_{0,1}}{\lambda_2 - \lambda_1} e^{-\lambda_1 t}.$$

A General Nonlinear System

$$\boxed{\mathbf{u}' = \mathbf{f}(\mathbf{u})}$$

Question: How do we deal with nonlinear functions?
Answer: We linearize using Taylor's theorem!

$$\mathbf{u}' \approx \mathbf{f}(\mathbf{u}_0) + J(\mathbf{u} - \mathbf{u}_0) = \Big(\mathbf{f}(\mathbf{u}_0) - J\mathbf{u}_0\Big) + J\mathbf{u},$$

where J is the Jacobian matrix (evaluated at $\mathbf{u}_0$),

$$J_{ij} = \frac{\partial f_i}{\partial u_j}.$$

Rewriting in terms of $\mathbf{v} = \mathbf{u} - \mathbf{u}_0 + J^{-1}\mathbf{f}(\mathbf{u}_0)$, this is

$$\mathbf{v}' \approx J\mathbf{v},$$

and the behavior is determined by the eigenvalues of J.

Remark. Now the eigenvalues change with time,

$$\lambda_i = \lambda_i(\mathbf{u}, t),$$

so we only talk about the behavior for "a little while into the future."

Some Theory

Lipschitz Continuity

Definition. The function $\mathbf{f}$ is said to be Lipschitz continuous if there is some constant $L \geq 0$ such that

$$\|\mathbf{f}(\mathbf{u}) - \mathbf{f}(\mathbf{v})\| \leq L\|\mathbf{u} - \mathbf{v}\|$$

for all $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$.

Lemma. If $\left|\dfrac{\partial f_i}{\partial u_j}\right| \leq C$ for all i and j, then $\mathbf{f}$ is Lipschitz.

For a single equation, we have

$$\begin{aligned} & f(u) = f(v) + f'(\xi)(u - v) \\ \Longrightarrow \quad & |f(u) - f(v)| = |f'(\xi)||u - v| \leq \max_{\xi} |f'(\xi)||u - v|, \end{aligned}$$

so

$$L = \max_{\xi} |f'(\xi)|.$$

Remark: Lipschitx continuous functions are continuous and have bounded derivatives. However, we can have a few points where there is no derivative, such as $f(u) = |u|$.

Existence and Uniqueness Theory

Theorem. If $\mathbf{f}$ is continuous in $\mathbf{u}$ and t, and Lipschitz in $\mathbf{u}$, then the system

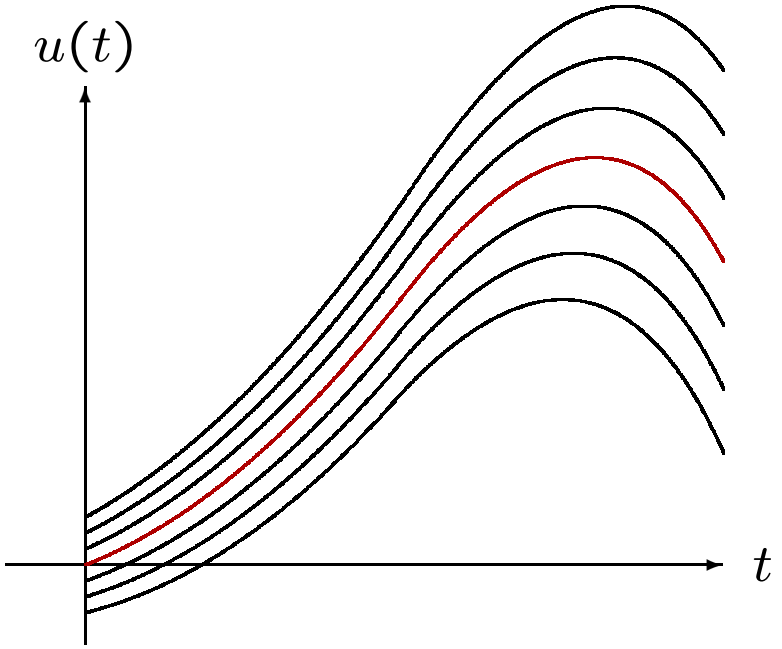
$$\begin{cases} \mathbf{u}' = \mathbf{f}(\mathbf{u}, t), & t \geq 0, \quad \text{(the equation)} \\ \mathbf{u}(0) = \mathbf{u}_0, & \quad \text{(the initial condition)} \end{cases}$$

has a unique solution.

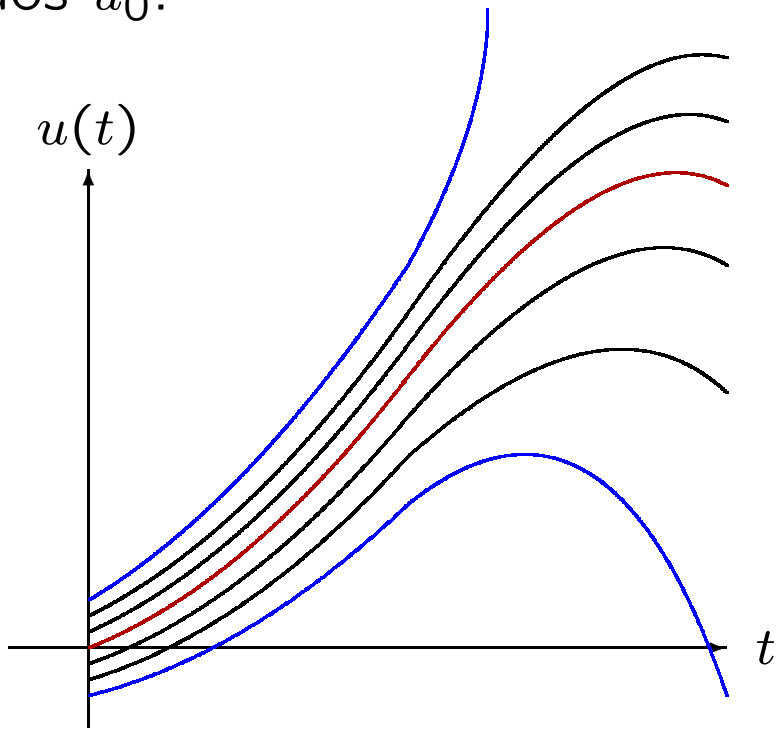
Question: Is this enough? Do you want to know more about the ODE?

Integral Curves for a Single Equation

Plot the solution of $u' = f(u,t)$ for variuos u_0:



A stable example. If u_0 varies, $u(t)$ does not change too much.



An unstable example. If u_0 varies, $u(t)$ may change a lot!

Question: Why might u_0 be in error?

Stability

Definition. The problem is stable, or has continuous dependence on the data $\mathbf{u}_0$ and $\mathbf{f}$, if there is some constant $C \geq 0$ such that for any small enough perturbations $\max_t \|\delta(t)\| \leq \epsilon$ and $\|\delta_0\| \leq \epsilon$, the solution to

$$\begin{cases} \mathbf{v}' = \mathbf{f}(\mathbf{v}, t) + \delta(t), & t \geq 0, \\ \mathbf{v}(0) = \mathbf{u}_0 + \delta_0, \end{cases}$$

satisfies

$$\max_t \|\mathbf{u}(t) - \mathbf{v}(t)\| \leq C\epsilon.$$

That is, the magnitude of the difference between the solutions is bounded by a (fixed) multiple of the magnitude of the differences of the initial conditions and the slope functions.

> More simply, if the data does not change much, then neither does the solution.

Remark. Note that this is a critical property for numerical approximation! This is also an important property for physical systems. If this fails, we have chaos.

Autonomous Systems

Stationary Points

Definition. If $\mathbf{f}(\mathbf{u}, t) = \mathbf{f}(\mathbf{u})$ depends on $\mathbf{u}$ only (not on t explicitly, only on t implicitly through $\mathbf{u}(t)$), then the system is autonomous. Moreover, if $\mathbf{f}(\mathbf{u}_0) = 0$, then $\mathbf{u}_0$ is a stationary point (or a critical point) for the autonomous system.

Lemma. If $\mathbf{u}_0$ is a stationary point, then $\mathbf{u}(t) = \mathbf{u}_0$.
Question: Can you show this fact?

Definition. A stationary point $\mathbf{v}_0$ is stable if whenever $\mathbf{u}_0 \approx \mathbf{v}_0$, then $\mathbf{u}(t) \approx \mathbf{v}_0$ for all $t > 0$. More precisely, given $\epsilon > 0$, there is $\delta > 0$ such that whenever $\|\mathbf{u}_0 - \mathbf{v}_0\| < \epsilon$, $\max_t \|\mathbf{u}(t) - \mathbf{v}_0\| < \delta$. A stationary point $\mathbf{v}_0$ is asymptotically stable if whenever $\mathbf{u}(t) - \mathbf{v}_0$ is sufficiently small, then

$$\lim_{t \to \infty} \mathbf{u}(t) = \mathbf{v}_0.$$

Remark. Unstable includes the case where some $\mathbf{u}_0$ stay close to $\mathbf{v}_0$ and some do not (a "saddle-point" instability).

Example: Preditor-Prey—1

R is the number of prey ("rabbits")
F is the number of preditors ("foxes")

a is the net reproduction rate of R
b is the death rate of R per encounter with F
c is the net reproduction of F per R eaten ($c < b$)
d is the death rate of F in the absence of food (R)

The equations are

$$R' = aR - bRF = R(a - bF),$$
$$F' = cRF - dF = F(cR - d).$$

The stationary points are

$$F = R = 0 \quad \text{and} \quad R = d/c,\ F = a/b.$$

Question: What does the case $F = R = 0$ mean?

Thus, in steady state, we would expect $R = d/c$ and $F = a/b$. But is this case stable? That is, how resilient is this natural system to perturbations?

Linear Stability Analysis—1

For the linear problem $\mathbf{u}' = A\mathbf{u}$, stationary points are those for which $A\mathbf{u} = 0$, such as $\mathbf{u} = 0$.

Suppose that we have an eigenvector $\mathbf{v}_0$ such that the eigenvalue $\lambda_0 = a + ib$ has positive real part $a > 0$. Then $A\mathbf{v}_0 = \lambda_0\mathbf{v}_0$, so

$$\mathbf{v}' = A\mathbf{v} \quad \text{and} \quad \mathbf{v}(0) = \epsilon\mathbf{v}_0$$

is solved by

$$\mathbf{v}(t) = \epsilon e^{\lambda_0 t}\mathbf{v}_0,$$

since

$$\mathbf{v}' = \epsilon e^{\lambda_0 t}\lambda_0\mathbf{v}_0 = \epsilon e^{\lambda_0 t}A\mathbf{v}_0 = A(\epsilon e^{\lambda_0 t}\mathbf{v}_0) = A\mathbf{v}.$$

No matter how small $\epsilon \neq 0$ is, the solution grows with time, since

$$e^{\lambda_0 t} = e^{(a+ib)t} = e^{at}(\cos bt + i \sin bt) \longrightarrow \infty.$$

Theorem. If A is invertible (so 0 is the only stationary point), then:

0 is a stable stationary point $\iff$ all eigenvalues have real part $a \leq 0$,

0 is asymptotically stable $\iff$ all eigenvalues have real part $a < 0$.

Linear Stability Analysis—2

For the nonlinear problem $\mathbf{u}' = \mathbf{f}(\mathbf{u})$ with stationary point $\mathbf{u}_0$, we use Taylor's theorem

$$\mathbf{u}' \approx \mathbf{f}(\mathbf{u}_0) + J(\mathbf{u} - \mathbf{u}_0) = J(\mathbf{u} - \mathbf{u}_0),$$

Rewriting in terms of $\mathbf{v} = \mathbf{u} - \mathbf{u}_0$, this is

$$\mathbf{v}' \approx J\mathbf{v},$$

and the linear stability behavior is determined by the eigenvalues of the Jacobian matrix J (evaluated at $\mathbf{u}_0$), where

$$J_{ij} = \frac{\partial f_i}{\partial u_j}.$$

Remark. We call this linear stability analysis, since it is not quite the same as analyzing the full stability of the nonlinear system. That is, the higher order terms from the Taylor approximation can sometimes turn a stationary point that looks stable into an unstable point.

Example: Preditor-Prey—2

$$R' = aR - bRF = R(a - bF),$$
$$F' = cRF - dF = F(cR - d),$$

with stationary point $R = d/c$, $F = a/b$.

Linear stability analysis requires the Jacobian

$$J = \frac{\partial f_i}{\partial u_j} = \begin{pmatrix} \dfrac{\partial R(a-bF)}{\partial R} & \dfrac{\partial R(a-bF)}{\partial F} \\ \dfrac{\partial F(cR-d)}{\partial R} & \dfrac{\partial F(cR-d)}{\partial F} \end{pmatrix} = \begin{pmatrix} a - bF & -bR \\ cF & cR - d \end{pmatrix} = \begin{pmatrix} 0 & -\frac{bd}{c} \\ \frac{ac}{b} & 0 \end{pmatrix}.$$

The eigenvalues satisfy

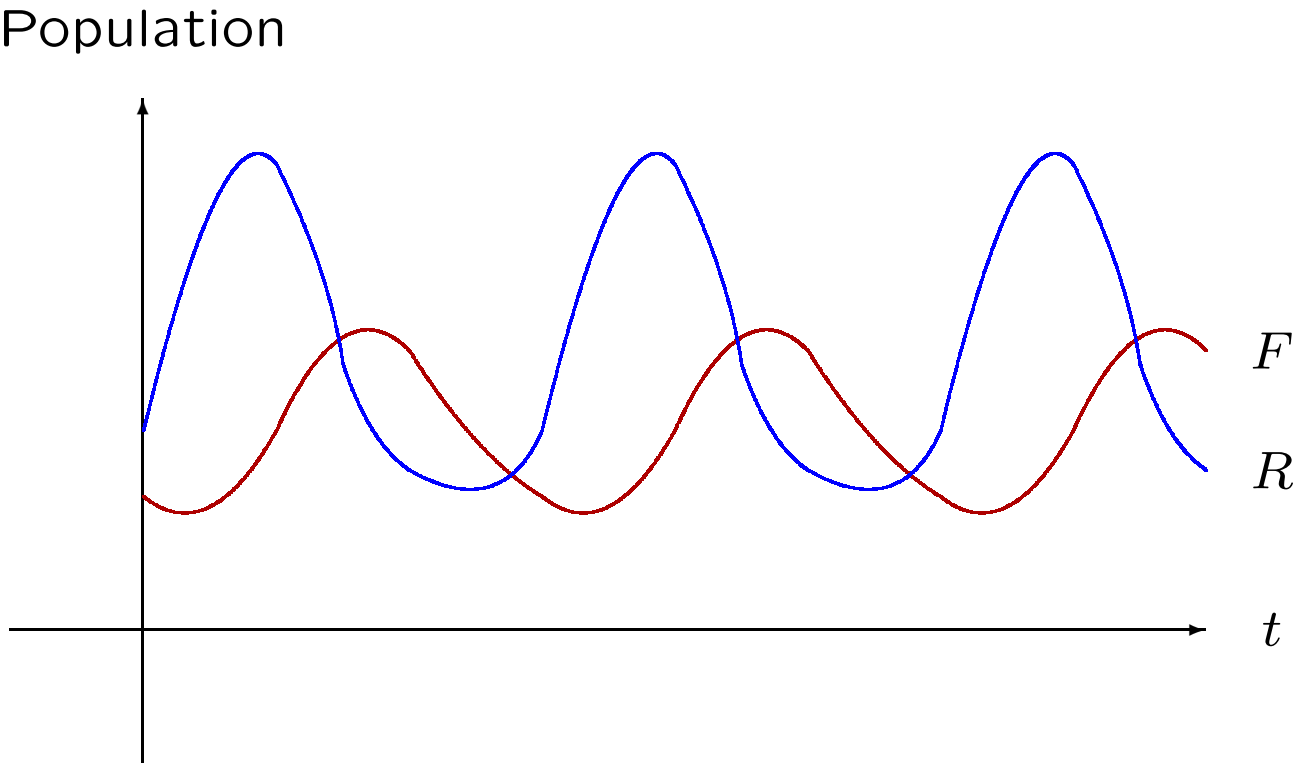
$$\lambda^2 + ad = 0 \qquad \Longrightarrow \qquad \lambda = \pm i\sqrt{ad}.$$

Hence the stationary point is stable (real part is 0), but *not* asymptotically so.

Conclusion: The R and F populations oscillate around the stationary configuration, never straying too far from it, but not converging to it.

Example: Preditor-Prey—3

The solution is periodic and looks something like



Example: Preditor-Prey—4

$$R' = aR - bRF = R(a - bF),$$
$$F' = cRF - dF = F(cR - d),$$

with stationary point $R = 0$, $F = 0$.

Now the Jacobian is

$$J = \begin{pmatrix} a - bF & -bR \\ cF & cR - d \end{pmatrix} = \begin{pmatrix} a & 0 \\ 0 & -d \end{pmatrix}.$$

The eigenvalues are $a > 0$ and $-d < 0$, so the stationary point is unstable.

Conclusion: If R and F populations are very small but positive, the populations may grow (i.e., they will not become extinct) in some cases.

Remark: In fact, if $F = 0$, $R \to \infty$, and if $R = 0$, $F \to 0$.
Question: Why is this obvious?

Numerical Approximation

The Euler Method

$$\begin{cases} \mathbf{u}' = \mathbf{f}(\mathbf{u}, t), & t \geq 0, \\ \mathbf{u}(0) = \mathbf{u}_0. \end{cases}$$

We discretize the time axis into small time steps

$$0 = t_0 < t_1 < t_2 < \cdots$$

and appromimate

$$\mathbf{u}(t_n) \approx \mathbf{u}_n.$$

Taylor's theorem implies

$$\begin{aligned} \mathbf{u}(t_{n+1}) &= \mathbf{u}(t_n) + \mathbf{u}'(t_n)(t_{n+1} - t_n) + R_n \\ &= \mathbf{u}(t_n) + \mathbf{f}(\mathbf{u}(t_n), t_n)\,(t_{n+1} - t_n) + R_n. \end{aligned}$$

Dropping the remainder, we have

$$\boxed{\mathbf{u}_{n+1} = \mathbf{u}_n + \mathbf{f}(\mathbf{u}_n, t_n)\,(t_{n+1} - t_n), \quad n = 0, 1, 2, ...,}$$

where $\mathbf{u}_0$ is known. This is called Euler's method.

Question: Why can we drop the remainder?

A Simple Variant of Euler's Method

Without using vectors, Euler's Method is

$$\begin{cases} u_{1,n+1} = u_{1,n} + f_1(u_{1,n}, u_{2,n}, ..., u_{d,n}, t_n)(t_{n+1} - t_n), \\ u_{2,n+1} = u_{2,n} + f_2(u_{1,n}, u_{2,n}, ..., u_{d,n}, t_n)(t_{n+1} - t_n), \\ \qquad \vdots \\ u_{d,n+1} = u_{d,n} + f_d(u_{1,n}, u_{2,n}, ..., u_{d,n}, t_n)(t_{n+1} - t_n). \end{cases}$$

Note that we compute all new $u_{i,n+1}$ using only the $u_{i,n}$.

A "Gauss-Seidel" version of Euler's Method is

$$\begin{cases} u_{1,n+1} = u_{1,n} + f_1(u_{1,n}, u_{2,n}, ..., u_{d,n}, t_n)(t_{n+1} - t_n), \\ u_{2,n+1} = u_{2,n} + f_2(u_{1,n+1}, u_{2,n}, ..., u_{d,n}, t_n)(t_{n+1} - t_n), \\ \qquad \vdots \\ u_{d,n+1} = u_{d,n} + f_d(u_{1,n+1}, u_{2,n+1}, ..., u_{d,n}, t_n)(t_{n+1} - t_n). \end{cases}$$

We use the new values as soon as we have computed them.

Convergence and Big Oh Notation

Theorem. For Euler's Method, if

$$\Delta t = \max_n (t_{n+1} - t_n),$$

then there is $C \geq 0$, depending on $\mathbf{f}$, such that

$$\max_n \|\mathbf{u}(t_n) - \mathbf{u}_n\| \leq C\Delta t.$$

Definition. We say that a function $f(x)$ is big oh of x as x tends to 0 and write

$$f(x) = \mathcal{O}(x)$$

if there is $C \geq 0$ such that

$$\lim_{x \to 0} \frac{|f(x)|}{|x|} \leq C.$$

Remark. Essentially, $f(x) \leq Cx$. Thus Eulers method has error

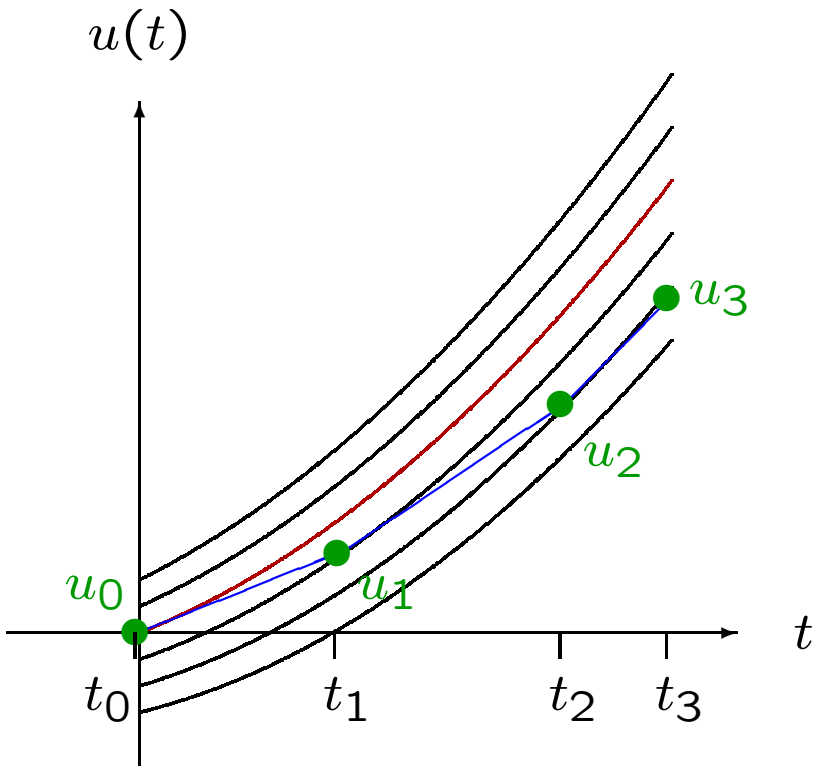
$$\max_n \|\mathbf{u}(t_n) - \mathbf{u}_n\| = \mathcal{O}(\Delta t),$$

which is first order accurate, meaning the power of Δt is 1.

Question: Why would it be better if the power had been 2 or more?

Graphical View of Euler's Method for a Single Equation

Consider a stable differential equation.



Euler's method is

$$u_{n+1} = u_n + f(u_n, t_n)(t_{n+1} - t_n),$$

where f is the slope of $u(t)$ (i.e., $u' = f$). So we have

$$\begin{aligned} u_1 &= u_0 + f_0\,(t_1 - t_0), \\ u_2 &= u_1 + f_1\,(t_2 - t_1), \\ u_3 &= u_2 + f_2\,(t_3 - t_2), \\ &\vdots \end{aligned}$$

Question: Do you see why we need a *stable* differential equation?

Euler moves along the known slope each time step. We move off the true trajectory (integral curve), but nearby trajectories are like the correct one, so we maintain accuracy.

Euler's Method for a Single Linear Equation

For the single linear equation

$$u' = f(u) = \lambda u,$$

we have the solution

$$u(t) = u_0 e^{\lambda t}.$$

This equation is stable only for $\lambda \leq 0$ (so we assume this).

Suppose we use equal time steps $\Delta t = (t_{n+1} - t_n)$. Then Euler becomes

$$\begin{aligned} u_{n+1} &= u_n + (\lambda u_n)\Delta t \\ &= (1 + \lambda \Delta t) u_n \\ &= (1 + \lambda \Delta t)^2 u_{n-1} \\ &\vdots \\ &= (1 + \lambda \Delta t)^{n+1} u_0. \end{aligned}$$

That is, shifting the index n,

$$\boxed{u_n = (1 + \lambda \Delta t)^n u_0.}$$

Sources of Error in Euler's Method

We have three sources of error:

1. discretization error (the error $\approx C\Delta t$ and $\Delta t \neq 0$);
2. inaccurate data;
3. numerical rounding error (finite number of digits).

We saw 1 was OK (using exact data and no rounding error). But 2 and 3 can be serious!

Example: Suppose we have error only in u_0. We get

$$v_n = (1 + \lambda\Delta t)^n(u_0 + \delta_0) = u_n + (1 + \lambda\Delta t)^n\delta_0.$$

The error

$$|(1 + \lambda\Delta t)^n\delta_0| \longrightarrow \infty$$

grows if, and only if,

$$|(1 + \lambda\Delta t)^n| > 1.$$

To be precise, let $u_0 = 0$, $\delta_0 = 10^{-12}$, $\lambda = -10$ and $\Delta t = 0.5$. Then

$$u = 0 \quad \text{and} \quad u_n = (1 - 10 \times 0.5)^n \times 10^{-12} = (-4)^n \times 10^{-12}.$$

Note that $4^{20} = 1.1 \times 10^{12}$, so $u_{20} = 1.1 \not\approx 0$, and $u_{40} = 1.2 \times 10^{12}$(!).

Thus the error completely hides the true solution $u = 0$ as $n \to \infty$.

Linear Stability of Euler's Method

Definition. A numerical method is stable if small errors in the data and small rounding errors do not lead to large changes in the numerical approximation.

Definition. We say that a numerical scheme for our single ODE is linearly stable if it is stable for the equation $u' = \lambda u$, with $\lambda < 0$.

Theorem. For a single equation, Euler's method is (linearly) stable if, and only if, $|(1 + \lambda \Delta t)| \leq 1$. That is,

$$-1 \leq 1 + \lambda \Delta t \leq 1 \quad \Longrightarrow \quad \boxed{0 < \Delta t \leq \frac{2}{|\lambda|}.}$$

For multiple equations, we consider the linearized equation. We require that all eigenvalues λ_i of the Jacobian matrix be negative, and require that Δt satisfy

$$\boxed{0 < \Delta t \leq \min_i \frac{2}{|\lambda_i|} = \frac{2}{\max_i |\lambda_i|}.}$$

So it is the maximal eigenvalue (i.e., shortest time scale) that determines the choice of Δt.

Example: Radionuclide Decay Chains Revisited

For the decay chain

$$\boxed{u_1 = \text{Uranium-238} \longrightarrow u_2 = \text{Thorium-234}}$$

we have

$$\begin{cases} u_1' = -\lambda_1 u_1, \\ u_2' = \lambda_1 u_1 - \lambda_2 u_2, \end{cases} \qquad \text{or} \qquad \mathbf{u}' = f(\mathbf{u}),$$

where $\lambda_1 = 1.54 \times 10^{-9}/\,\text{year}$ and $\lambda_2 = 10.3/\,\text{year}$.

Uranium only. To stably compute u_1, we need

$$\Delta t \leq \frac{2}{1.54 \times 10^{-9}} = 1.3 \times 10^9\,\text{year}.$$

The coupled system. To stably compute both, we need to find the eigenvalues, which are $-\lambda_1$ and $-\lambda_2$. Thus we need

$$\Delta t \leq \frac{2}{10.3} = 0.19\,\text{year!}$$

Question: Can we compute this for millions of years? How many steps is this?

The Backward Euler Method

$$\begin{cases} \mathbf{u}' = \mathbf{f}(\mathbf{u}, t), & t \geq 0, \\ \mathbf{u}(0) = \mathbf{u}_0. \end{cases}$$

Taylor's theorem implies

$$\begin{aligned} \mathbf{u}(t_n) &= \mathbf{u}(t_{n+1}) - \mathbf{u}'(t_{n+1})(t_{n+1} - t_n) + R_{n+1}) \\ &= \mathbf{u}(t_{n+1}) - \mathbf{f}(\mathbf{u}(t_{n+1}), t_{n+1})\,(t_{n+1} - t_n) + R_{n+1}. \end{aligned}$$

Dropping the remainder, we have the Backward Euler method.

(Forward) Euler:

$$\boxed{\mathbf{u}_{n+1} = \mathbf{u}_n + \mathbf{f}(\mathbf{u}_n, t_n)\,(t_{n+1} - t_n), \quad n = 0, 1, 2, \ldots.}$$

This method is explicit, since we simply compute u_{n+1}.

Backward Euler:

$$\boxed{\mathbf{u}_{n+1} = \mathbf{u}_n + \mathbf{f}(\mathbf{u}_{n+1}, t_{n+1})\,(t_{n+1} - t_n), \quad n = 0, 1, 2, \ldots.}$$

This method is implicit, since we do not have a simple formula for $\mathbf{u}_{n+1}$.

Question: How do we find $\mathbf{u}_{n+1}$?

Solution of Backward Euler

Given $\mathbf{u}_n$, t_n, and t_{n+1}, we need to find $\mathbf{v} = \mathbf{u}_{n+1}$ such that

$$\mathcal{F}(\mathbf{v}) = \mathbf{v} - \mathbf{u}_n - \mathbf{f}(\mathbf{v}, t_{n+1})\,(t_{n+1} - t_n) = 0.$$

This is a root finding problem (to be discussed later). Basically, you find the solution by iterating:

- You guess the solution $\mathbf{v}_0$ ($\mathbf{v}_0 = \mathbf{u}_n$, perhaps);
- You use $\mathbf{v}_0$ to define a better guess $\mathbf{v}_1$;
- You use $\mathbf{v}_1$ to define a better guess $\mathbf{v}_2$;
- etcetera.

After some time $\mathbf{v}_{N-1} \approx \mathbf{v}_N$, so you stop and set $\mathbf{v}_N \approx \mathbf{u}_{n+1}$.

Backward Euler is computationally more expensive than forward Euler!

Theorem. For backward Euler, if

$$\Delta t = \max_n (t_{n+1} - t_n),$$

then there is $C \geq 0$, depending on $\mathbf{f}$, such that

$$\max_n \|\mathbf{u}(t_n) - \mathbf{u}_n\| \leq C \Delta t.$$

Note that backward Euler is *not* more accurate than forward Euler.

Question: So, why would anyone use backward Euler?

Backward Euler for a Single Linear Equation

For the single linear equation (recall stability requires that $\lambda < 0$)

$$u' = f(u) = \lambda u,$$

and equal time steps $\Delta t = (t_{n+1} - t_n)$, backward Euler becomes

$$\begin{aligned} u_{n+1} = u_n + (\lambda u_{n+1})\Delta t \quad &\Longrightarrow \quad (1 - \lambda\Delta t)u_{n+1} = u_n \\ &\Longrightarrow \quad u_{n+1} = \frac{1}{1 - \lambda\Delta t}\, u_n, \end{aligned}$$

so

$$\begin{aligned} u_{n+1} &= \frac{1}{1 - \lambda\Delta t}\, u_n \\ &= \frac{1}{(1 - \lambda\Delta t)^2}\, u_{n-1} \\ &\vdots \\ &= \frac{1}{(1 - \lambda\Delta t)^{n+1}}\, u_0. \end{aligned}$$

That is,

$$\boxed{u_n = \frac{1}{(1 - \lambda\Delta t)^n}\, u_0.}$$

Linear Stability of Backward Euler

Example: Suppose we have error only in u_0. We get

$$v_n = \frac{1}{(1-\lambda\Delta t)^n}(u_0 + \delta_0) = u_n + \frac{1}{(1-\lambda\Delta t)^n}\,\delta_0.$$

Since $\lambda < 0$, the error

$$\left|\frac{1}{(1-\lambda\Delta t)^n}\,\delta_0\right| \longrightarrow 0$$

no matter what!

Theorem. Backward Euler is unconditionally (linearly) stable (i.e., it is stable for any choice of Δt).

Conclusions.

1. Forward Euler is easy to use and each time step is quick to compute, but the method can suffer from stability problems (roundoff and data errors can destroy the calculation) unless Δt is small enough, so we have to take lots of time steps.
2. Backward Euler is harder to use and more expensive per step, but it does not have stability problems.
3. Thus, we would probably use forward Euler if we can afford to take the very small Δt required, and backward Euler otherwise.

Improved Accuracy via Smaller Time Steps

Both forward and backward Euler are first order accurate:

$$\text{Error} \leq C\Delta t.$$

You can improve the accuracy of the computed solution by solving the problem using, say, time step $\Delta t/c$, where c is the "cut" factor.

First order accuracy. Assuming that $\text{Error} \approx C\Delta t$, we have

$$\text{Error}_{\Delta t} \approx C\Delta t \quad \text{and} \quad \text{Error}_{\Delta t/c} \approx C\Delta t/c = (\text{Error}_{\Delta t})/c$$

That is, if we cut Δt by c, the error is cut by c as well.

- If $c = 2$ (cut time step in half), then the error is also cut in half.
- To get one more decimal point (i.e., cut the error by $c = 10$), you use step size $\Delta t/10$, and take 10 times as many steps. So the computation is about 10 times slower!

Improved Accuracy via Higher Order Methods

Higher order accuracy. Suppose that instead Error $\leq C_p \Delta t^p$. Then

p	Δt	Error	Δt	Error
1	1/2	$C_1/2$	0.1	$C_1/10$
2	1/2	$C_2/4$	0.1	$C_2/100$
3	1/2	$C_3/8$	0.1	$C_3/1000$
4	1/2	$C_4/16$	0.1	$C_4/10000$

If $C_1 \approx C_2 \approx C_3 \approx C_4$, the errors improve as p increases (C_p may grow!).

Higher order accuracy and time step reduction. For a fixed p,

$$\text{Error}_{\Delta t} \approx C_p \Delta t^p \quad \text{and} \quad \text{Error}_{\Delta t/c} \approx C_p (\Delta t/c)^p = (\text{Error}_{\Delta t})/c^p$$

So whatever C_p is, if we cut Δt by c, the error is cut by c^p:

p	c	New Error	Extra Work	c	New Error	Extra Work
1	2	1/2	2	10	1/10	10
2	2	1/4	2	$\sqrt{10} = 3.2$	1/10	3.2
3	2	1/8	2	$10^{1/3} = 2.2$	1/10	2.2
4	2	1/16	2	$10^{1/4} = 1.8$	1/10	1.8

Conclusion. Higher order methods are more efficient.

Stiff ODE's

Caution. The constant C_p in the error estimate

$$\text{Error} \leq C_p \Delta t^p$$

depends on the size of the pth order derivatives of the solution $\mathbf{u}$:

$$\text{Error} \leq C \left\{ \int_0^T \|\mathbf{u}^{(p)}\|^2 \, dt \right\}^{1/2} \Delta t^p.$$

Thus, it is useless to use a higher order method if

- $\mathbf{u}$ does not have a pth order derivative;
- or, the pth order derivative of $\mathbf{u}$ is very large.

Definition. A differential equation whose solution has large derivatives is called stiff.

A stiff ODE requires a small Δt to solve accurately, so it may be better to use a lower order method for this type of ODE.

Runge-Kutta Methods

Euler's Method is

$$\mathbf{u}_{n+1} = \mathbf{u}_n + \mathbf{f}(u_n, t_n)\Delta t = (\text{current solution}) + (\text{slope}) \times (\text{time step})$$

We should try to improve the slope, by "sampling" more points. A class of such methods are called Runge-Kutta Methods.

A popular method is called RK4.

Theorem. For RK4, if

$$\Delta t = \max_n (t_{n+1} - t_n),$$

then there is $C \geq 0$, depending on $\mathbf{f}$, such that

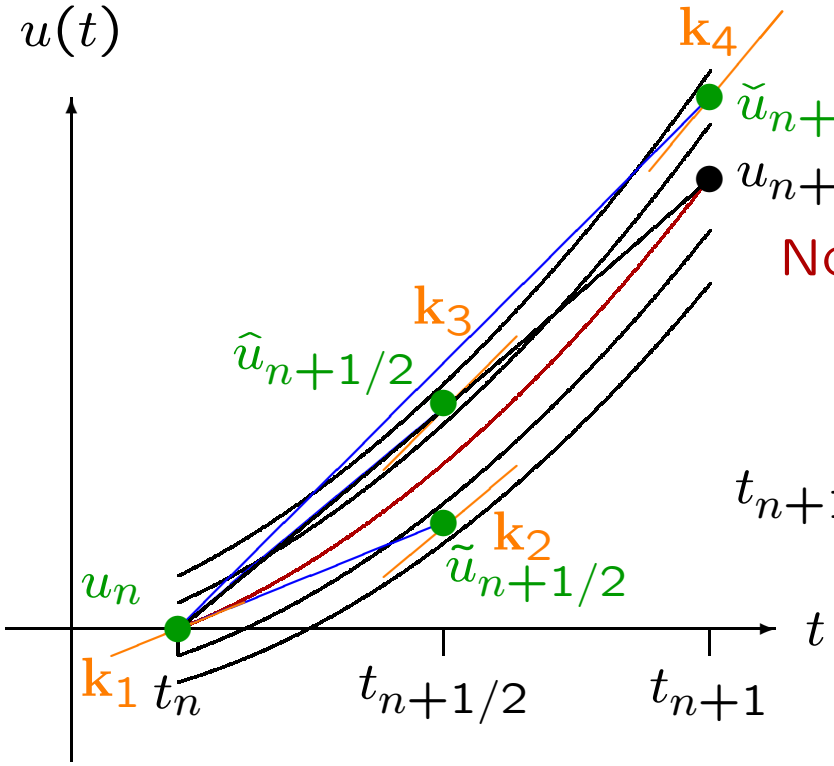
$$\max_n \|\mathbf{u}(t_n) - \mathbf{u}_n\| \leq C\Delta t^4.$$

Moreover, the method is only conditionally stable (i.e., for Δt sufficiently small).

Question: Why do I not bother to state the stability condition precisely?

Remark. There is an implicit version of RK4 which is stable.

Standard Runge-Kutta 4 (RK4)



Note how accurate this is!

$$t_{n+1/2} = (t_n + t_{n+1})/2$$

For RK4, we sample (slope)$\times\Delta t$ four times

$$\mathbf{k}_1 = \mathbf{f}(\mathbf{u}_n, t_n)\Delta t \qquad \tilde{\mathbf{u}}_{n+1/2} = \mathbf{u}_n + \tfrac{1}{2}\mathbf{k}_1 \quad \text{(Euler to } t_{n+1/2}\text{)}$$

$$\mathbf{k}_2 = \mathbf{f}(\tilde{\mathbf{u}}_{n+1/2}, t_{n+1/2})\Delta t \qquad \hat{\mathbf{u}}_{n+1/2} = \mathbf{u}_n + \tfrac{1}{2}\mathbf{k}_2 \quad \text{(Euler-like to } t_{n+1/2}\text{)}$$

$$\mathbf{k}_3 = \mathbf{f}(\hat{\mathbf{u}}_{n+1/2}, t_{n+1/2})\Delta t \qquad \breve{\mathbf{u}}_{n+1} = \mathbf{u}_n + \mathbf{k}_3 \quad \text{(Euler-like to } t_{n+1}\text{)}$$

$$\mathbf{k}_4 = \mathbf{f}(\breve{\mathbf{u}}_{n+1}, t_{n+1})\Delta t$$

and then set $$\mathbf{u}_{n+1} = \frac{1}{6}\Big[\mathbf{k}_1 + 2\mathbf{k}_2 + 2\mathbf{k}_3 + \mathbf{k}_4\Big].$$

Higher Order Ordinary Differential Equations

Conversion to a First Order System

Sometimes we have an rth order ODE like

$$\begin{cases} y^{(r)} = f\Big(y, y', y'', ..., y^{(r-1)}, t\Big), \\ y(0) = y_0,\ y'(0) = y_1,\ \dots\ , y^{(r-1)}(0) = y_{r-1}. \end{cases}$$

We can solve this by rewriting it as a system of first order equations (and using the previous methods).

Let

$$u_1 = y,$$

and define u_2, u_3, ... , u_r by

$$\begin{cases} u_1' = u_2 & (u_2 = y'), \\ u_2' = u_3 & (u_3 = y''), \\ \quad\vdots & \\ u_{r-1}' = u_r & (u_r = y^{r-1}), \\ u_r' = f\Big(u_1, u_2, u_3, ..., u_{r-1}, t\Big), & \end{cases}$$

where the last equation is the ODE itself, and set initial conditions

$$u_1(0) = y_0,\ u_2(0) = y_1,\ ..., u_{r-1}(0) = y_{r-1}.$$

Software

Matlab ODE Solvers

Some ODE solvers from the matlab manual.

Solver	Problem Type	Order of Accuracy	When to Use
ode45	Nonstiff	Medium	Most of the time. This should be the first solver you try.
ode23	Nonstiff	Low	For problems with crude error tolerances or for solving moderately stiff problems.
ode113	Nonstiff	Low to high	For problems with stringent error tolerances or for solving computationally intensive problems.
ode15s	Stiff	Low to medium	If ode45 is slow because the problem is stiff.
ode23t	Moderately Stiff	Low	For moderately stiff problems if you need a solution without numerical damping.

Remark. Solver ode45 uses RK4 and RK5 to estimate the error and adjust the time step to achieve a desired accuracy. Solver ode23 is similar but uses RK2/RK3.

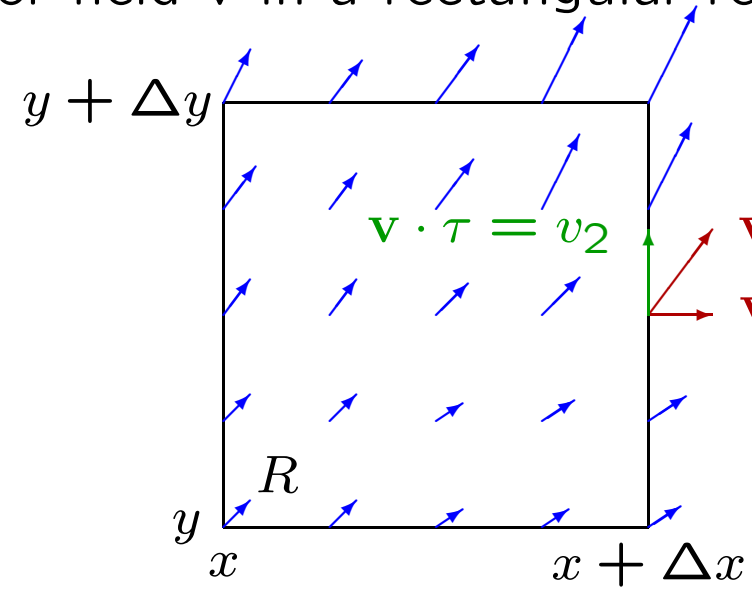
II. Elliptic Partial differential equations (PDE)

(for steady state single phase flow or diffusive equilibrium)

Conservation of Continuous Fluids

The Divergence Theorem—1

Consider a vector field $\mathbf{v}$ in a rectangular region R of space.



ν is the outer unit normal vector to ∂R

The total flow through the boundary ∂R is

$$\begin{aligned}
\int_{\partial R} \mathbf{v} \cdot \nu \, dS &= \int_y^{y+\Delta y} \Big(v_1(x + \Delta x, s) - v_1(x, s) \Big) ds \\
&\quad + \int_x^{x+\Delta x} \Big(v_2(r, y + \Delta y) - v_2(r, y) \Big) dr \\
&= \int_y^{y+\Delta y} \int_x^{x+\Delta x} \frac{\partial v_1(r,s)}{\partial x} \, dr \, ds + \int_x^{x+\Delta x} \int_y^{y+\Delta y} \frac{\partial v_2(r,s)}{\partial y} \, ds \, dr \\
&= \iint_R \left(\frac{\partial v_1(r,s)}{\partial x} + \frac{\partial v_2(r,s)}{\partial y} \right) dr \, ds.
\end{aligned}$$

The Divergence Theorem—2

Definition. In 3-D, let

$$\nabla \cdot \mathbf{v} = \frac{\partial v_1}{\partial x} + \frac{\partial v_2}{\partial y} + \frac{\partial v_3}{\partial z} = \sum_{i=1}^{d} \frac{\partial v_i}{\partial x}$$

be the divergence of $\mathbf{v}$ (in 2-D, omit the partial derivative in z).

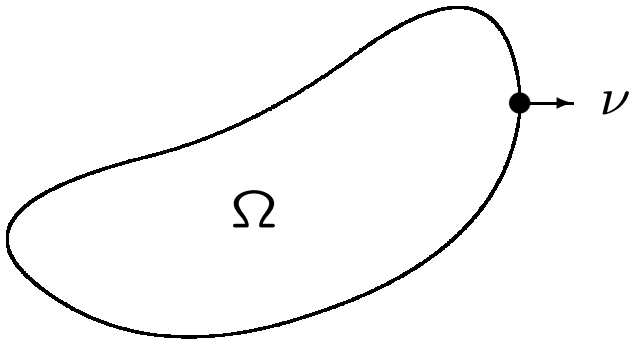
Theorem.

$$\iiint_R \nabla \cdot \mathbf{v}\, dx\, dy\, dz = \iint_{\partial R} \mathbf{v} \cdot \nu\, dS.$$

We can fill any region with cubes and add the results above.

Theorem. For any region Ω with unit outer normal vector ν,

$$\iiint_\Omega \nabla \cdot \mathbf{v}\, dx\, dy\, dz = \iint_{\partial\Omega} \mathbf{v} \cdot \nu\, dS.$$



Conservative Fluid Flow

Suppose

ξ is a conserved quantity ξ (mass/volume)

$\mathbf{v}$ is the fluid velocity (length/time)

$\xi\mathbf{v}$ is the *flux* of ξ (mass/area/time)

q is an external source or sink of fluid (mass/volume/time)

Within a region of space R, the total amount of ξ changes in time by

$$\underbrace{\frac{d}{dt}\iiint_R \xi\,dx\,dy\,dz}_{\text{Change in } R} = \underbrace{-\iint_{\partial R} \xi\mathbf{v}\cdot\nu\,dS}_{\text{Flow across } \partial R} + \underbrace{\iiint_R q\,dx\,dy\,dz}_{\text{Sources/sinks}}$$

$\Longrightarrow$ conservation locally on R

$$\iiint_R \xi_t\,dx\,dy\,dz = \underbrace{-\iiint_R \nabla\cdot(\xi\mathbf{v})\,dx\,dy\,dz}_{\text{Divergence Theorem}} + \iiint_R q\,dx\,dy\,dz$$

This is true for each region R, so in fact

$$\boxed{\xi_t + \nabla\cdot(\xi\mathbf{v}) = q}$$

Flow in Porous Media and Elliptic Partial Differential Equations (PDE's)

Darcy's Law

Darcy's law tells us that the fluid flux is

$$\boxed{\mathbf{u} = -\frac{K}{\mu}(\nabla p + \rho \mathbf{g}),}$$

where

$p(\mathbf{x}, t)$	is the fluid pressure
$\mathbf{u}(\mathbf{x}, t)$	is the Darcy velocity
$K(\mathbf{x})$	is the permeability of the medium
μ	is the fluid viscosity
$\rho(\mathbf{x}, t)$	is the fluid density
$\mathbf{g}$	is the gravitational constant vector

Remark. If we neglect gravity, Darcy's law tells us that fluid flows from high pressure to low pressure. We determine the direction of flow by taking the gradient of the pressure, and multiplying by K/μ.

Question: Why does K need to be positive?

Single Phase Darcy Flow

Conservation and Darcy's law requires that

$$\frac{\partial \phi \rho}{\partial t} + \nabla \cdot (\rho \mathbf{u}) = Q \quad \Longrightarrow \quad \frac{\partial \phi \rho}{\partial t} - \nabla \cdot \left(\rho \frac{K}{\mu} (\nabla p + \rho \mathbf{g}) \right) = Q.$$

where

$p(\mathbf{x}, t)$	is the fluid pressure
$\mathbf{u}(\mathbf{x}, t)$	is the Darcy velocity $(\mathbf{v} = \mathbf{u}/\phi)$
$\phi(\mathbf{x}, t)$	is the porosity of the medium
$K(\mathbf{x})$	is the permeability of the medium
μ	is the fluid viscosity
$\rho(\mathbf{x}, t)$	is the fluid density $(\xi = \phi\rho)$
$\mathbf{g}$	is the gravitational constant vector
$Q(\mathbf{x}, t)$	is the source/sink (i.e., wells)

Incompressible Single Phase Darcy Flow

If the fluid and medium are incompressible (ρ is constant and ϕ is constant in time), and we neglect gravity ($\mathbf{g} = 0$), then we have

$$\nabla \cdot \mathbf{u} = q \qquad \Longrightarrow \qquad -\nabla \cdot (k \nabla p) = q,$$

where

$q(\mathbf{x})$ is $Q(\mathbf{x})/\rho$ (assuming q does not change in time)

$k(\mathbf{x})$ is $K(\mathbf{x})/\mu$

In coordinate form, this is

$$-\frac{\partial}{\partial x}\left(k\frac{\partial p}{\partial x}\right) - \frac{\partial}{\partial y}\left(k\frac{\partial p}{\partial y}\right) - \frac{\partial}{\partial z}\left(k\frac{\partial p}{\partial z}\right) = q.$$

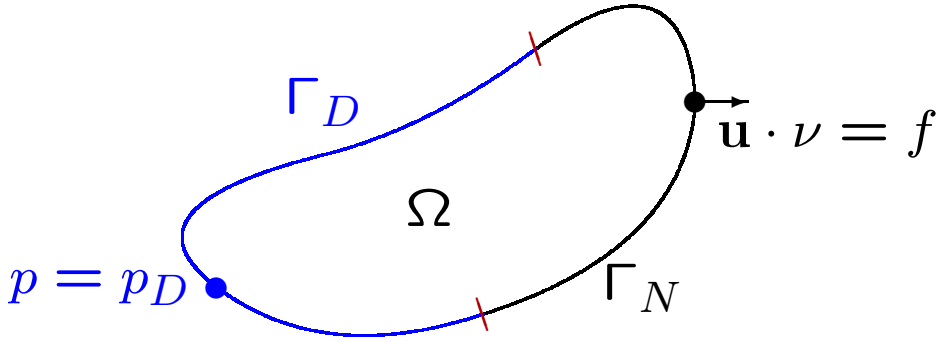
It is convenient to write a partial derivative using a subscript. Then we have more simply

$$-(k\,p_x)_x - (k\,p_y)_y - (k\,p_z)_z = q.$$

Boundary Conditions (BC's)

The equation holds in the interior of the porous formation, which we call $\Omega \subset \mathbb{R}^3$. We must also specify what happens on the boundary, $\partial\Omega$.

We consider two types of boundary conditions. Decompose $\partial\Omega$ into nonoverlapping regions Γ_D and Γ_N (so $\partial\Omega = \Gamma_D \cup \Gamma_N$ and $\Gamma_D \cap \Gamma_N = \emptyset$).



1. Dirichlet. Specify p_D, the pressure on Γ_D:

$$p = p_D.$$

2. Neumann. Specify f, the outward normal flux on Γ_N:

$$\mathbf{u} \cdot \nu = f.$$

   This is the fluid that enters or leaves the domain Ω. For example, $f = 0$ for a sealed boundary which cannot support flow.

Remark. If $p_D = 0$ and $f = 0$, the BC's are said to be homogeneous.

Compatibility Condition

If $\Gamma_N = \partial\Omega$ (i.e., $\Gamma_D = \emptyset$), then we may have a conservation problem. Recall the derivation. Within a region of space R, the total amount of $\xi = \phi\rho$ is now constant in time, so

$$0 = \underbrace{\frac{d}{dt}\iiint_R \phi\rho\, dx}_{\text{Change in } R} = \underbrace{-\iint_{\partial R} \rho\mathbf{u}\cdot\nu\, da(x)}_{\text{Flow across } \partial R} + \underbrace{\iiint_R \rho q\, dx}_{\text{Sources/sinks}}$$

If $R = \Omega$, we must have

$$\iiint_\Omega q\, dx = \iint_{\partial\Omega} \mathbf{u}\cdot\nu\, da(x)$$

That is, the data must satisfy the compatibility condition

$$\boxed{\iiint_\Omega q\, dx = \iint_{\partial\Omega} f\, da(x) \text{ when } \Gamma_N = \partial\Omega.}$$

Remark. If you inject some fluid through f on the boundary or through q in the interior, you must take it out somewhere else! This is basically what it means to have an incompressible fluid and medium.

We will always assume that the compatibility condition holds in this case.

Question: Does the Dirichlet BC require a compatibility condition?

A Second Order Elliptic Boundary Value Problem

In summary, incompressible single phase flow is a second order elliptic boundary value problem of the form

$$\begin{cases} -\nabla \cdot (k \nabla p) = q, & \text{in } \Omega, \\ p = p_D, & \text{on } \Gamma_D, \\ -(k \nabla p) \cdot \nu = f, & \text{on } \Gamma_N, \end{cases}$$

which we may sometimes write as

$$\begin{cases} \mathbf{u} = -k \nabla p, & \text{in } \Omega, \\ \nabla \cdot \mathbf{u} = q, & \text{in } \Omega, \\ p = p_D, & \text{on } \Gamma_D, \\ \mathbf{u} \cdot \nu = f, & \text{on } \Gamma_N. \end{cases}$$

Remark. If k is constant, we can divide it out of the equation. We then have Poisson's equation

$$\Delta p = \nabla^2 p = \nabla \cdot \nabla p = -q/k$$

wherein we call $\Delta = \nabla^2 = \nabla \cdot \nabla$ the Laplacian.

Existance and Uniqueness of a Solution

$$\begin{cases} -\nabla \cdot (k \nabla p) = q, & \text{in } \Omega, \\ p = p_D, & \text{on } \Gamma_D, \\ -(k \nabla p) \cdot \nu = f, & \text{on } \Gamma_N, \end{cases}$$

Theorem. If $k \geq k_* > 0$ for some k_*, then this boundary value problem has a unique solution p and $\mathbf{u}$ for any (reasonable) Ω, Γ_D, $\Gamma_N = \partial\Omega \setminus \Gamma_D$, k, q, p_D, and f. However, if $\Gamma_N = \partial\Omega$, we require the compatibility condition on q and f, and uniqueness of p holds only up to a constant.

Remark. The PDE has many solutions, but only one satisfies the BC's.

Remark. In the pure Neumann case, if p solves the full problem, so does $p + C$, for any constant C. In other words, if p_1 and p_2 both solve the problem, then $p_1 - p_2$ is a constant.

The Maximum Principle

Theorem (Maximum Principle). If $q \leq 0$ in a region R, then there are no local maxima inside R (maxima could exist on ∂R).

Proof. The case $q = 0$ is a bit tricky, but if $q < 0$, it is easy to see this.

Suppose that $R \subset \mathbb{R}^2$ for simplicity. If we had a local maximum, the derivatives are 0 and the curvature is negative at that point:

$$p_x = p_y = 0, \quad p_{xx} \leq 0, \quad \text{and} \quad p_{yy} \leq 0.$$

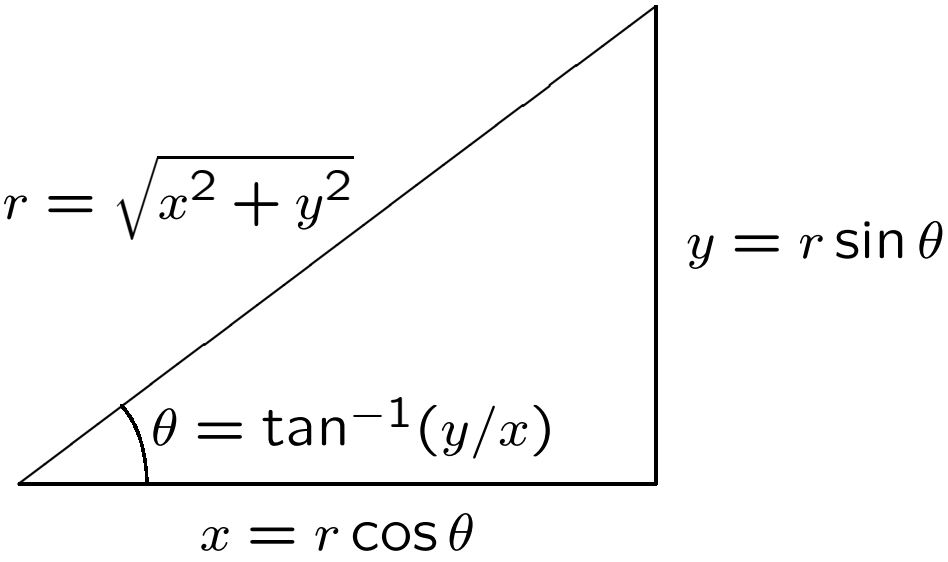
But then

$$0 < -q = \nabla \cdot (k \nabla p) = (k p_x)_x + (k p_y)_y = k_x p_x + k_y p_y + k(p_{xx} + p_{yy}) \leq 0,$$

a contradiction. □

Remark. Wells are local, so $q = 0$ most places. Thus we cannot have a build-up of pressure at any point away from the wells. (If we did, fluid would have to flow away from that point, reducing the pressure until it was not a local maximum.)

Polar and Cylindrical Coordinates



The polar coordinate transformation in $\mathbb{R}^2$ is

$$\begin{cases} r = \sqrt{x^2 + y^2}, \\ \tan\theta = y/x, \end{cases} \iff \begin{cases} x = r\cos\theta, \\ y = r\sin\theta, \end{cases}$$

with $z = z$ being unchanged for cylindrical coordinates in $\mathbb{R}^3$.

The Laplacian is

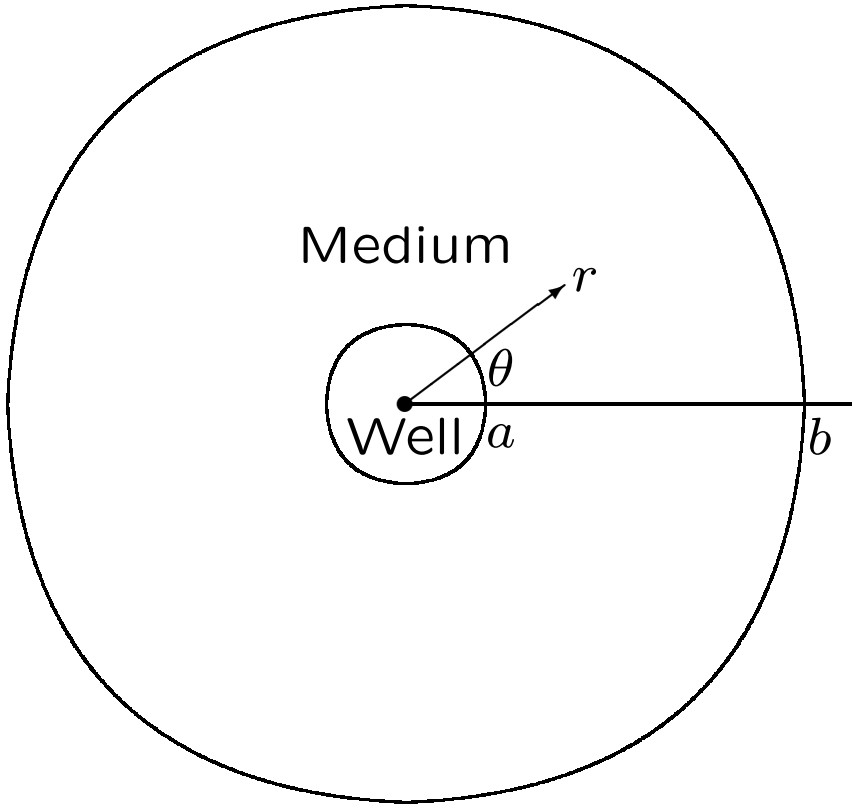
$$\Delta p = \nabla^2 p = \nabla \cdot \nabla p = \boxed{\frac{\partial^2 p}{\partial r^2} + \frac{1}{r}\frac{\partial p}{\partial r} + \frac{1}{r^2}\frac{\partial p}{\partial \theta} + \frac{\partial^2 p}{\partial z^2},}$$

wherein the z derivatives are missing in $\mathbb{R}^2$.

A Single Well

Assume that

- The permeability k is constant.
- There is a single vertical well of radius a.
- The domain is symmetric and annular of outer radius b.
- The well maintains a fixed pressure p_a.
- The outer boundary maintains a fixed pressure p_b.



By symmetry, $p = p(r)$ only. Thus, in cylindrical coordinates, we have

$$\begin{cases} \dfrac{\partial^2 p}{\partial r^2} + \dfrac{1}{r}\dfrac{\partial p}{\partial r} = 0, \quad a < r < b, \\ p(a) = p_a, \ p(b) = p_b. \end{cases}$$

The solution is

$$p(r) = p_a + (p_b - p_a)\frac{\log(r/a)}{\log(b/a)}.$$

We note that the pressure is logarithmically decaying.

Linearity

Our problem is linear, in the sense that for constants α and β,

$$\begin{cases} \nabla \cdot \Big(k\nabla(\alpha p_1 + \beta p_2)\Big) = \alpha \nabla \cdot (k\nabla p_1) + \beta \nabla \cdot (k\nabla p_2), & \text{in } \Omega, \\ (\alpha p_1 + \beta p_2) = \alpha p_1 + \beta p_2, & \text{on } \Gamma_D, \\ \Big(k\nabla(\alpha p_1 + \beta p_2)\Big) \cdot \nu = \alpha (k\nabla p_1) \cdot \nu + \beta (k\nabla p_2) \cdot \nu, & \text{on } \Gamma_N, \end{cases}$$

That is, if

$$\begin{cases} -\nabla \cdot (k\nabla p_1) = q_1 \quad \text{and} \quad -\nabla \cdot (k\nabla p_2) = q_2, & \text{in } \Omega, \\ p_1 = p_{D,1} \quad \text{and} \quad p_2 = p_{D,2}, & \text{on } \Gamma_D, \\ (k\nabla p_1) \cdot \nu = f_1 \quad \text{and} \quad (k\nabla p_2) \cdot \nu = f_2, & \text{on } \Gamma_N, \end{cases}$$

then if $p = p_1 + p_2$,

$$\begin{cases} -\nabla \cdot (k\nabla p) = q_1 + q_2 = q, & \text{in } \Omega, \\ p = p_{D,1} + p_{D,2} = p_D, & \text{on } \Gamma_D, \\ (k\nabla p) \cdot \nu = f_1 + f_2 = f, & \text{on } \Gamma_N, \end{cases}$$

Superposition. A hard problem can sometimes be broken into multiple simpler, solvable problems. If so, the full solution is the superposition (i.e., sum) of the simpler solutions (e.g., $p = p_1 + p_2$).

Simple Example of Linearity

$$\begin{cases} p_{xx} + p_{yy} = 2, \quad 0 < x < a,\ 0 < y < b, \\ p(0,y) = p(a,y) = p(x,0) = p(x,b) = 0. \end{cases}$$

Easily $p_1(x,y) = x(x-a)$ solves

$$\begin{cases} p_{1,xx} + p_{1,yy} = 2, \quad 0 < x < a,\ 0 < y < b, \\ p_1(0,y) = p_1(a,y) = 0,\ p_1(x,0) = p_1(x,b) = x(x-a). \end{cases}$$

So we just need to solve

$$\begin{cases} p_{2,xx} + p_{2,yy} = 0, \quad 0 < x < a,\ 0 < y < b, \\ p_2(0,y) = p_2(a,y) = 0,\ p_2(x,0) = -x(x-a),\ p_2(x,b) = 0. \end{cases}$$

and

$$\begin{cases} p_{3,xx} + p_{3,yy} = 0, \quad 0 < x < a,\ 0 < y < b, \\ p_3(0,y) = p_3(a,y) = p_3(x,0) = 0,\ p_3(x,b) = -x(x-a). \end{cases}$$

and then our solution is

$$p(x,y) = x(x-a) + p_2(x,y) + p_3(x,y).$$

We can use separation of variables for p_2 and p_3.

Solution by Separation of Variables

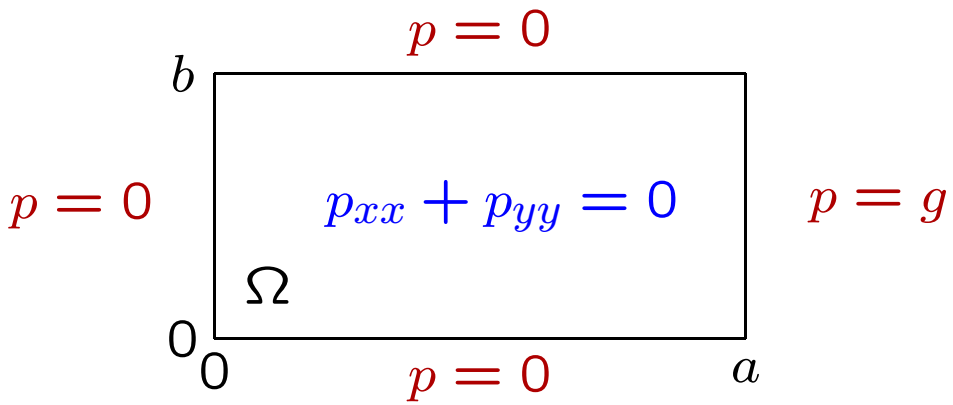
Limitations on the Problem to be Solved

- For simplicity only, assume that $\Omega \subset \mathbb{R}^2$ (i.e., everything is independent of the third coordinate, so we can reduce the problem to 2-D).
- Assume that the domain Ω is a rectangle.
- Assume that k is constant (take $k = 1$).

Then we have $\boxed{-p_{xx} - p_{yy} = q.}$

Simple example. We take $q = 0$ and the BC's:

$$\begin{cases} p_{xx} + p_{yy} = 0, \quad 0 < x < a, \ 0 < y < b, \\ p(a, y) = g(y), \\ p(0, y) = p(x, 0) = p(x, b) = 0. \end{cases}$$



Separation of Variables—1

Step 1, separate the variables. We look for solutions to

$$p_{xx} + p_{yy} = 0$$

of the form

$$\boxed{p(x, y) = X(x)Y(y).}$$

Thus

$$X''Y + XY'' = 0 \quad \Longrightarrow \quad \frac{X''}{X} + \frac{Y''}{Y} = 0 \Longrightarrow \quad \frac{X''}{X} = -\frac{Y''}{Y}.$$

Note that X''/X depends on x only, and Y''/Y depends on y only, so if these are equal, they must be constant! Let λ^2 be this constant:

$$\frac{X''}{X} = -\frac{Y''}{Y} = \lambda^2,$$

or

$$\boxed{X'' - \lambda^2 X = 0} \quad \text{and} \quad \boxed{Y'' + \lambda^2 Y = 0}.$$

Remark. We take the constant to be a square for convenience. The constant may be negative if λ is imaginary.

Step 2, solve the problem with two BC= 0 (homogeneous BC's). The general solution for Y is

$$Y(y) = \begin{cases} \alpha\cos(\lambda y) + \beta\sin(\lambda y), & \text{if } \lambda > 0, \\ \alpha\cosh(\lambda y) + \beta\sinh(\lambda y), & \text{if } \lambda < 0, \\ \alpha + \beta y, & \text{if } \lambda = 0, \end{cases}$$

where

$$\cosh x = \tfrac{1}{2}\left(e^x + e^{-x}\right) \quad \text{and} \quad \sinh x = \tfrac{1}{2}\left(e^x - e^{-x}\right).$$

We now need to recall the BC's for $y = 0$ and $y = b$, which are

$$p(x,0) = p(x,b) = 0 \quad \text{or} \quad X(x)Y(0) = X(x)Y(b) = 0.$$

The only way to have his true for all x is to require

$$Y(0) = Y(b) = 0,$$

Since $Y(0) = 0$, we must take $\alpha = 0$.

Separation of Variables—3

For the other BC,

$$0 = \begin{cases} \beta \sin(\lambda b), & \text{if } \lambda > 0, \\ \beta \sinh(\lambda b), & \text{if } \lambda < 0, \\ \beta b, & \text{if } \lambda = 0. \end{cases}$$

Now $\beta = 0$ is not interesting, since then $Y = 0$ and so $p = 0$. But

$$\beta \sinh(\lambda b) > 0 \quad \text{and} \quad \beta b > 0,$$

so $\lambda \leq 0$ is not possible. Now $\lambda > 0$ and we have

$$Y(y) = \beta \sin(\lambda y),$$

and the requirement that

$$\sin(\lambda b) = 0.$$

This fails in general, but it is true for certain λ, namely if

$$\boxed{\lambda = \lambda_n = \frac{n\pi}{b}, \quad n = 1, 2, 3, \ldots.}$$

Let

$$\boxed{Y_n = \sin(\lambda_n y) = \sin\left(\frac{n\pi y}{b}\right), \quad n = 1, 2, 3, \ldots.}$$

Step 3, solve the other problem. Now for $\lambda = \lambda_n$, we must have

$$X(x) = \alpha \cosh(\lambda_n x) + \beta \sinh(\lambda_n x).$$

The boundary condition

$$p(0, y) = X(0)Y(y) = 0 \qquad \Longrightarrow \qquad X(0) = 0,$$

so $\alpha = 0$. Let

$$\boxed{X_n(x) = \sinh(\lambda_n x).}$$

Step 4, use superposition. We have derived infinitely many solutions

$$p_n(x,y) = X_n(x)Y_n(y) = \sinh(\lambda_n x)\sin(\lambda_n y), \quad n = 1,2,3,...,$$

to the problem

$$\begin{cases} p_{n,xx} + p_{n,yy} = 0, \quad 0 < x < a,\ 0 < y < b, \\ p_n(0,y) = p_n(x,0) = p_n(x,b) = 0. \end{cases}$$

By superposition, for any constants c_n,

$$\boxed{p(x,y) = \sum_{n=1}^{\infty} c_n p_n(x,y) = \sum_{n=1}^{\infty} c_n \sinh(\lambda_n x)\sin(\lambda_n y),}$$

where $\lambda_n = n\pi/b$, $p(x,y)$ solves the same problem, which is the one we want to solve except that we need to require the final BC:

$$p(a,y) = \sum_{n=1}^{\infty} \Big(c_n \sinh(\lambda_n a)\Big)\sin(\lambda_n y) = g(y).$$

Question: That is, can we find c_n such that this holds?

Fourier Series

Definition. Given a function $f(x)$ for $0 < x < \ell$, let

$$a_n = \frac{2}{\ell}\int_0^\ell f(x)\,\cos(2n\pi x/\ell)\,dx \quad \text{and} \quad b_n = \frac{2}{\ell}\int_0^\ell f(x)\,\sin(2n\pi x/\ell)\,dx.$$

Then the Fourier series of f is

$$F(x) = \tfrac{1}{2}a_0 + \sum_{n=1}^{\infty}\Big(a_n\cos(2n\pi x/\ell) + b_n\sin(2n\pi x/\ell)\Big)$$

Definition. Given a function $f(x)$ for $0 < x < \ell$, let

$$a_n = \frac{2}{\ell}\int_0^\ell f(x)\,\cos(n\pi x/\ell)\,dx \quad \text{and} \quad b_n = \frac{2}{\ell}\int_0^\ell f(x)\,\sin(n\pi x/\ell)\,dx.$$

Then the Fourier cosine series of f is

$$C(x) = \tfrac{1}{2}a_0 + \sum_{n=1}^{\infty} a_n\cos(n\pi x/\ell),$$

and the Fourier sine series of f is

$$S(x) = \sum_{n=1}^{\infty} b_n\sin(n\pi x/\ell).$$

Theorem. If x is such that $0 < x < \ell$ and $f(x)$ is continuous, then $F(x) = C(x) = S(x) = f(x)$.

Step 5, set the remaining BC. It remains to show that

$$g(y) = \sum_{n=1}^{\infty} \Big(c_n \sinh(\lambda_n a)\Big) \sin(\lambda_n y).$$

But we now know that

$$g(y) = \sum_{n=1}^{\infty} b_n \sin(n\pi y/b) \quad \text{for} \quad b_n = \frac{2}{b}\int_0^b g(y)\, \sin(n\pi y/b)\, dy,$$

so

$$\boxed{c_n = b_n/\sinh(\lambda_n a) = \frac{2}{b\sinh(\lambda_n a)}\int_0^b g(y)\, \sin(n\pi y/b)\, dy,}$$

and

$$\boxed{p(x,y) = \sum_{n=1}^{\infty} c_n \sinh(n\pi x/b)\sin(n\pi y/b)}$$

solves the original problem.

Remark. The solution has modes that involve sine waves in y.

Solution by Fourier Transforms and Green's Functions

The Fourier Transform

Closely related to Fourier series is the Fourier transform.

Definition. Given a function $f(x)$, we define its Fourier transform by

$$\boxed{\widehat{f}(\omega) = \int_{-\infty}^{\infty} f(x)\, e^{-i\omega x}\, dx,}$$

and the Fourier inverse transform by

$$\boxed{\check{f}(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} f(x)\, e^{i\omega x}\, dx.}$$

Theorem. The transforms are inverses of each other:

$$\check{\widehat{f}} = f \quad \text{and} \quad \widehat{\check{f}} = f.$$

Remark. We transform from physical space x to Fourier space ω. Since

$$e^{i\theta} = \cos\theta + i \sin\theta$$

involves harmonic functions, we decompose f info harmonic waves $\widehat{f}$, which can be reconstructed to return f.

Some Properties of The Fourier Transform

Definition. The convolution of f and g is the function

$$(f * g)(x) = \int_{-\infty}^{\infty} f(y)\, g(x-y)\, dy.$$

Theorem. For functions f and g, $f * g = g * f$. Therefore

$$(f * g)(x) = \int_{-\infty}^{\infty} f(x-y)\, g(y)\, dy.$$

Theorem.

1. $(\alpha f_1 + \beta f_2)\widehat{\ } = \alpha \widehat{f}_1 + \beta \widehat{f}_2$ (F.T. is linear).
2. $\widehat{f * g} = (2\pi)^{-1} \widehat{f}\widehat{g}$ (F.T. converts convolution to multiplication).
3. $\widehat{f'}(\omega) = i\omega f(\omega)$ (F.T. converts differentiation to multiplication by $i\omega$).

Similar results hold for the Fourier inverse transformation.

Question: Can you show these results using only calculus?

Limitations on the Problem to be Solved

- Assume the domain $\Omega = \mathbb{R}^3$ is all of space! Then we have no BC's.
- Assume that k is constant (take $k = 1$).

Then we have

$$-p_{xx} - p_{yy} - p_{zz} = q,$$

which we convert to Fourier space by taking three Fourier transforms:

$$\begin{aligned}\widehat{f}(\omega_1, \omega_2, \omega_3) &= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} f(x, y, z)\, e^{-i\omega_1 x}\, dx \right) e^{-i\omega_2 y}\, dy \right] e^{-i\omega_3 z}\, dz \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y, z)\, e^{-i(\omega_1 x + \omega_2 y + \omega_3 z)}\, dx\, dy\, dz.\end{aligned}$$

We obtain (recall that $i^2 = -1$)

$$\begin{aligned}-\widehat{p_{xx}} - \widehat{p_{yy}} - \widehat{p_{zz}} &= -(i\omega_1)^2 \widehat{p} - (i\omega_2)^2 \widehat{p} - (i\omega_3)^2 \widehat{p} \\ &= (\omega_1^2 + \omega_2^2 + \omega_3^2) \widehat{p}.\end{aligned}$$

Thus, in Fourier space, the equation is

$$\boxed{(\omega_1^2 + \omega_2^2 + \omega_3^2)\, \widehat{p}(\omega_1, \omega_2, \omega_3) = \widehat{q}(\omega_1, \omega_2, \omega_3).}$$

Solution by Fourier Transform

$$(\omega_1^2 + \omega_2^2 + \omega_3^2)\widehat{p} = \widehat{q}$$

There are no derivatives, so we solve easily as

$$\widehat{p} = \frac{\widehat{q}}{\omega_1^2 + \omega_2^2 + \omega_3^2} \qquad \Longrightarrow \qquad p = \left(\frac{\widehat{q}}{\omega_1^2 + \omega_2^2 + \omega_3^2}\right)^{\vee},$$

if this makes sense! Suppose that

$$\widehat{\kappa}(\omega_1, \omega_2, \omega_3) = \frac{2\pi}{\omega_1^2 + \omega_2^2 + \omega_3^2}.$$

Then

$$p(x, y, z) = (2\pi)^{-1}(\widehat{q}\widehat{\kappa})^{\vee} = q * \kappa(x, y, z),$$

where the multivariable convolution is

$$q * \kappa(x, y, z) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} q(X, Y, Z)\, \kappa(x - X, y - Y, z - Z)\, dX\, dY\, dZ.$$

In fact, this does make sense, since we can find

$$\kappa(x, y, z) = \left(\frac{2\pi}{\omega_1^2 + \omega_2^2 + \omega_3^2}\right)^{\vee}.$$

Dirac Delta Function

The Dirac delta function (or Dirac mass) is a "generalized function" (or "distribution") that looks like

$$\delta_0(x) = \begin{cases} 0, & \text{if } x \neq 0, \\ +\infty, & \text{if } x = 0, \end{cases}$$

and has the property that for any continuous function f,

$$\boxed{\int_{-\infty}^{\infty} f(x)\,\delta_0(x)\,dx = f(0).}$$

Note that

$$f * \delta_0(x) = \int_{-\infty}^{\infty} f(x-y)\,\delta_0(y)\,dy = f(x),$$

and

$$(f*g)'(x) = \frac{d}{dx}\int_{-\infty}^{\infty} f(x-y)\,g(y)\,dy = \int_{-\infty}^{\infty} f'(x-y)\,g(y)\,dy = f'*g(x).$$

Theorem.

- $f * \delta_0(x) = f(x)$.
- $(f*g)' = f'*g = f*g'$.

A Fundamental Solution–1

In any dimension $\mathbb{R}^d$, let κ be

$$\kappa = \begin{cases} -\frac{1}{2}|x|, & \text{if } d = 1, \\ -\frac{1}{2\pi} \log \sqrt{x^2 + y^2}, & \text{if } d = 2, \\ \frac{1}{4\pi}(x^2 + y^2 + z^2)^{-1/2}, & \text{if } d = 3, \end{cases}$$

Theorem.

$$\boxed{-\Delta\kappa = \delta_0.}$$

Note that δ_0 represents an infinitely small well at the origin of unit strength, and κ solves our flow problem. We saw the logarithmic singularity earlier for a well in 2-D.

Definition. We call κ a fundamental solution to the differential equation $-\Delta p = q$ because $-\Delta\kappa = \delta_0$.

A Fundamental Solution–2

Since

$$-\Delta(q * \kappa) = -q * \Delta\kappa = q * \delta_0 = q,$$

we have the following theorem.

Theorem.

$$\boxed{p(x, y, z) = q * \kappa(x, y, z).}$$

Remark. By translation, $\kappa(x - X, y - Y, z - Z)$ is the response to a Dirac mass (i.e., source or well) at (X, Y, Z), since

$$-\Delta\kappa(x - X, y - Y, z - Z) = \delta_0(x - X, y - Y, z - Z) = \delta_{(X,Y,Z)}(x, y, z).$$

Therefore, our solution is an infinite superposition of fundamental solutions:

$$p(x, y, z) = \underbrace{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}}_{\text{"Infinite sum"}} \underbrace{q(X, Y, Z)}_{\substack{\text{Source at} \\ (X,Y,Z)}} \underbrace{\kappa(x - X, y - Y, z - Z)}_{\substack{\text{Response at } (x,y,z) \\ \text{to the source at } (X,Y,Z)}} dX\, dY\, dZ$$

In general, $q * \kappa$ does not satisfy the BC's. But we can fix them up!

The Green's Function

We return to the case where $k \neq 1$. We find the response to a Dirac mass at (X, Y, Z) with homogeneous BC's by solving

$$\begin{cases} -\nabla \cdot (k \nabla G) = \delta_{(X,Y,Z)}, & \text{in } \Omega, \\ G = 0, & \text{on } \Gamma_D, \\ -(k \nabla G) \cdot \nu = 0, & \text{on } \Gamma_N, \end{cases}$$

Note that $G = G(x, y, z; X, Y, Z)$ is a function of 6 variables: for each source point (X, Y, Z), we have a response at point (x, y, z). The derivatives above are taken in (x, y, z) with (X, Y, Z) fixed.

Definition. We call G the Green's function for the problem.

Theorem. The function

$$p(x, y, z) = \iiint_\Omega q(X, Y, Z)\, G(x, y, z; X, Y, Z)\, dX\, dY\, dZ$$

solves the homogeneous BC problem

$$\begin{cases} -\nabla \cdot (k \nabla p) = q, & \text{in } \Omega, \\ p = 0, & \text{on } \Gamma_D, \\ -(k \nabla p) \cdot \nu = 0, & \text{on } \Gamma_N. \end{cases}$$

Question: Can you verify this?

The Dirichlet Problem

For the pure Dirichlet problem,

$$\begin{cases} -\nabla \cdot (k \nabla p) = q, & \text{in } \Omega, \\ p = p_D, & \text{on } \partial\Omega, \end{cases}$$

the solution is given by the Poisson integral formula

$$p(x,y,z) = \iiint_\Omega q(X,Y,Z)\, G(X,Y,Z;x,y,z)\, dX\, dY\, dZ \\ - \iint_{\partial\Omega} p_D(X,Y,Z)\, \nabla G(X,Y,Z;x,y,z) \cdot \nu\, dS,$$

wherein the normal derivative of G is taken in the first set of variables.

Remark. Note that we have interchanged the two sets of variables in this formula from what we had before!

The Neumann Problem

For the pure Neumann problem,

$$\begin{cases} -\nabla \cdot (k\nabla p) = q, & \text{in } \Omega, \\ -(k\nabla p) \cdot \nu = f, & \text{on } \partial\Omega, \end{cases}$$

the solution is given by

$$p(x,y,z) = \iiint_\Omega q(X,Y,Z)\, G(X,Y,Z;x,y,z)\, dX\, dY\, dZ \\ - \iint_{\partial\Omega} f(X,Y,Z)\, G(X,Y,Z;x,y,z)\nu\, dS.$$

Again, we have interchanged the two sets of variables in this formula.

Remark. The Green's function data does not satisfy the compatibility condition, since

$$\iiint_\Omega \delta_{(X,Y,Z)}(x,y,z)\, dx\, dy\, dz = 1 \neq \iint_{\partial\Omega} 0\, dA.$$

In fact we solve

$$\begin{cases} -\nabla \cdot (k\nabla G) = \delta_{(X,Y,Z)} - 1/V, & \text{in } \Omega, \\ -(k\nabla G) \cdot \nu = 0, & \text{on } \partial\Omega, \end{cases}$$

where the volume of Ω is

$$V = \iiint_\Omega dx\, dy\, dz.$$

Numerical Solution by Vertex Centered Finite Differences

Vertex Centered Rectangular Grid

- For simplicity, we work in 2-D. Everything generalizes to 3-D.
- Assume that the porous medium domain Ω is the rectangle

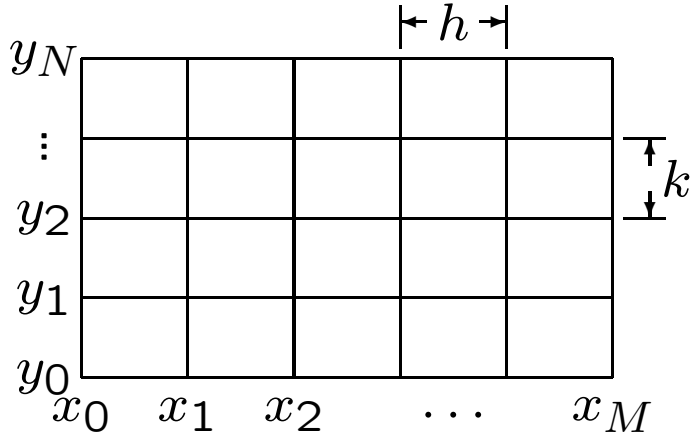
$$0 < x < L_1 \quad \text{and} \quad 0 < y < L_2.$$

We define a uniform rectangular grid on Ω by choosing integers $M \geq 1$ and $N \geq 1$, and setting the grid spacings to be

$$h = L_1/M \quad \text{and} \quad k = L_2/N,$$

and defining the grid vertex points

$$x_0 = 0,\ x_1 = h,\ \ldots,\ x_i = ih,\ \ldots,\ x_M = L_1,$$
$$y_0 = 0,\ y_1 = k,\ \ldots,\ y_j = jk,\ \ldots,\ y_N = L_2.$$



Remark. We have taken a uniform grid (i.e., constant x and y spacing) for simplicity. Most things generalize to nonuniform grids.

Approximation

We approximate

$$\boxed{p(x_i, y_j) \approx p_{ij} \quad \text{for } i = 0, 1, ..., M, \;\; j = 0, 1, ..., N.}$$

Strategy. Our strategy is to

- approximate derivatives of p using only these values p_{ij};
- replace the derivatives of p in the PDE by these approximations;
- solve the discretized PDE equations for the p_{ij};
- interpolate values of p_{ij} if we need p at a point other than a grid point.

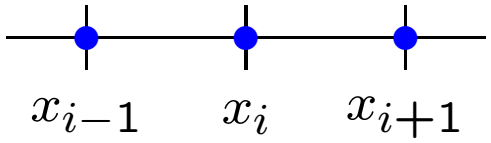
Remark. We will sometimes need the "half"-grid points, defined by

$$x_{i+1/2} = \frac{x_i + x_{i+1}}{2} = (i + 1/2)h$$

and

$$y_{j+1/2} = \frac{y_j + y_{j+1}}{2} = (j + 1/2)k.$$

Two Point Finite Difference



Given a function f(x), we have the two Taylor expansions

$$f(x_{i\pm1}) = f(x_i) \pm f'(x_i)h + \mathcal{O}(h^2).$$

Forward difference. We have the forward difference approximation

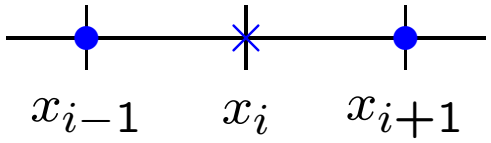
$$f'(x_i) = \frac{f(x_{i+1}) - f(x_i)}{h} + \mathcal{O}(h) \approx \frac{f_{i+1} - f_i}{h}.$$

Backward difference. We have the backward difference approximation

$$f'(x_i) = \frac{f(x_{i-1}) - f(x_i)}{-h} + \mathcal{O}(h) \approx \frac{f_i - f_{i-1}}{h}.$$

Theorem. Both approximations are first order accurate, meaning the error is $\mathcal{O}(h)$.

Two Point Centered Finite Differences



Recall the two Taylor approximations (with an additional term)

$$f(x_{i+1}) = f(x_i) + f'(x_i)h \quad + \tfrac{1}{2}f''(x_i)h^2 + \mathcal{O}(h^3),$$
$$f(x_{i-1}) = f(x_i) - f'(x_i)h \quad + \tfrac{1}{2}f''(x_i)h^2 + \mathcal{O}(h^3).$$

Subtracting, we get

$$f(x_{i+1}) - f(x_{i-1}) = 0 \qquad + 2f'(x_i)h + 0 \qquad\qquad + \mathcal{O}(h^3).$$

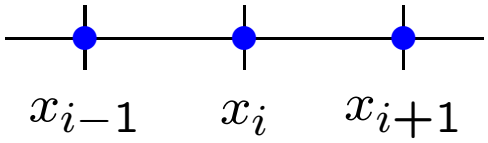
Centered difference. We have the centered difference approximation

$$f'(x_i) = \frac{f(x_{i+1}) - f(x_{i-1})}{2h} + \mathcal{O}(h^2) \approx \frac{f_{i+1} - f_{i-1}}{2h}.$$

Theorem. The centered approximation is second order accurate, meaning the error is $\mathcal{O}(h^2)$.

Remark. Note that $\mathcal{O}(h^3) - \mathcal{O}(h^3) = \mathcal{O}(h^3)$ (*not* 0), and the error is $\mathcal{O}(h^3)/2h = \mathcal{O}(h^2)$, since the big oh notation does not keep track of the constant multiple.

Summary of Two Point Finite Differences



Forward difference.

$$f'(x_i) \approx \frac{f_{i+1} - f_i}{h}, \quad \text{accurate to } \mathcal{O}(h).$$

Backward difference. approximation

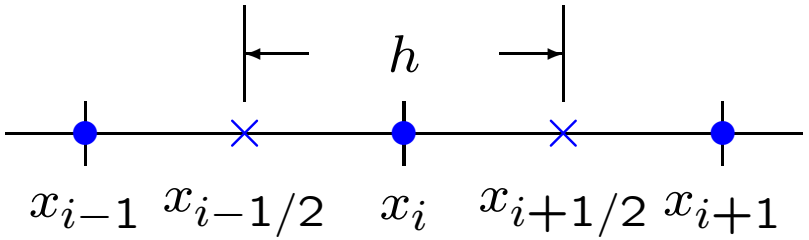
$$f'(x_i) \approx \frac{f_i - f_{i-1}}{h}, \quad \text{accurate to } \mathcal{O}(h).$$

Centered difference.

$$f'(x_i) \approx \frac{f_{i+1} - f_{i-1}}{2h}, \quad \text{accurate to } \mathcal{O}(h^2).$$

The centered difference is the most accurate!

Finite Differences for Second Order Derivatives



The standard formula is

$$\boxed{f''(x_i) \approx \frac{f_{i+1} - 2f_i + f_{i-1}}{h^2}.}$$

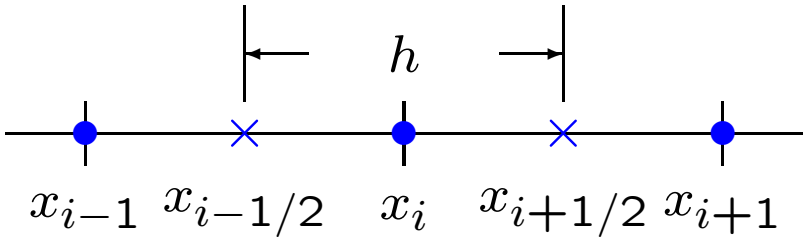
To see this, note that

$$\begin{aligned} f''(x_i) = (f')'(x_i) &\approx \frac{f'(x_{i+1/2}) - f'(x_{i-1/2})}{h} \\ &\approx \frac{1}{h}\left(\frac{f_{i+1} - f_i}{h} - \frac{f_i - f_{i-1}}{h}\right). \end{aligned}$$

Theorem. This approximation is accurate to second order, $\mathcal{O}(h^2)$.

Remark. It would appear that the error is $\mathcal{O}(h)$. To see that the error is in fact $\mathcal{O}(h^2)$, a more careful analysis is needed involving looking more carefully at the Taylor remainder terms.

Second Order Derivatives with a Coefficient



Suppose instead we have a coefficient

$$(Kf')'(x_i) \approx \frac{(Kf')(x_{i+1/2}) - (Kf')(x_{i-1/2})}{h}$$
$$\approx \frac{1}{h}\left(K_{i+1/2}\frac{f_{i+1}-f_i}{h} - K_{i-1/2}\frac{f_i - f_{i-1}}{h} \right).$$

The formula is

$$(Kf')'(x_i) \approx \frac{1}{h^2}\Big[K_{i+1/2}f_{i+1} - \Big(K_{i+1/2} + K_{i-1/2}\Big)f_i + K_{i-1/2}f_{i-1}\Big].$$

Theorem. This approximation is accurate to second order, $\mathcal{O}(h^2)$.

Approximation of the Dirichlet Problem

In the pure Dirichlet case (write K for the permeability),

$$\begin{cases} -\nabla \cdot (K \nabla p) = q, & \text{in } \Omega = (0, L_1) \times (0, L_2), \\ p = p_D, & \text{on } \partial\Omega, \end{cases}$$

we know how to set the boundary values

$$\begin{aligned} p_{i0} &= p_D(x_i, 0), \ p_{iN} = p_D(x_i, L_2), \quad i = 0, 1, ..., M, \\ p_{0j} &= p_D(0, y_j), \ p_{Mj} = p_D(L_1, y_j), \quad j = 0, 1, ..., N. \end{aligned}$$

It remains to solve the PDE at the interior grid points. We have

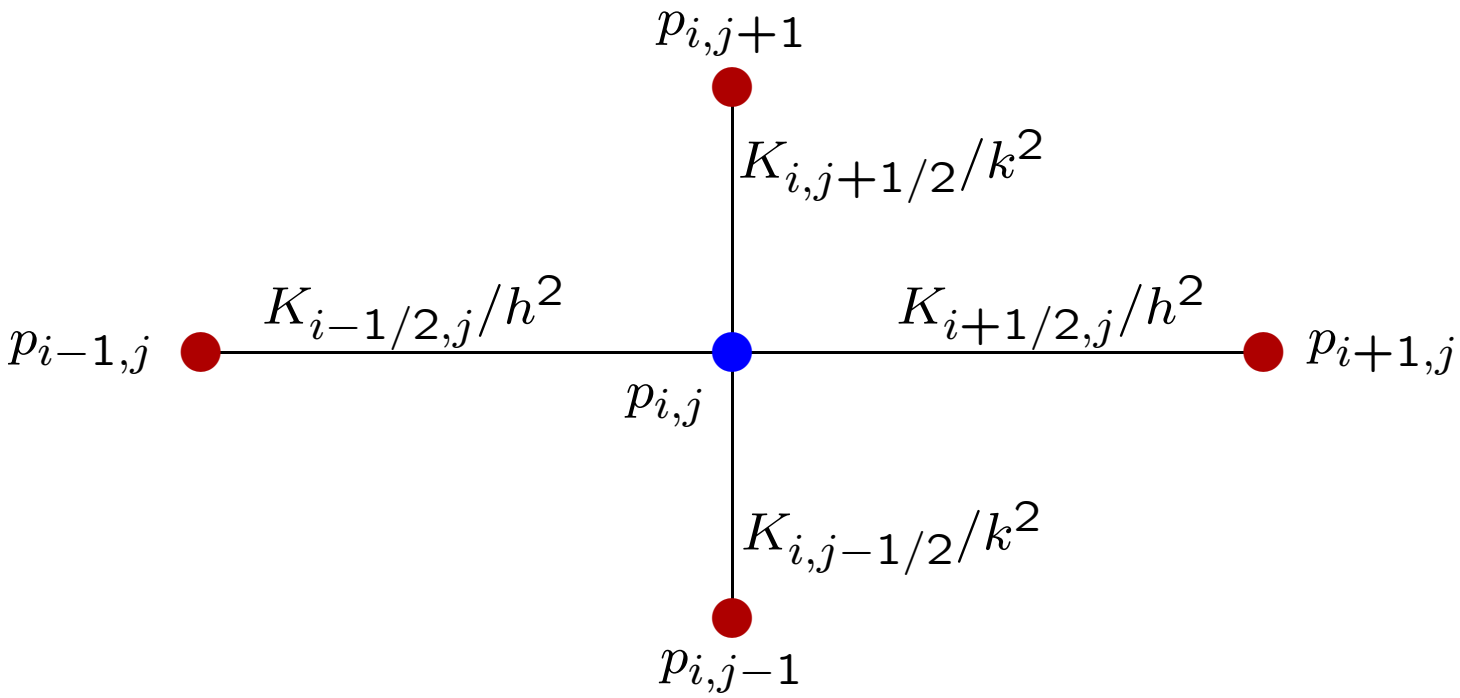
$$\begin{aligned} -\nabla \cdot (K \nabla p)(x_i, y_j) &= -(K p_x)_x(x_i, y_j) - (K p_y)_y(x_i, y_j) \\ &\approx -\frac{1}{h^2}\Big[K_{i+1/2,j} p_{i+1,j} - \Big(K_{i+1/2,j} + K_{i-1/2,j}\Big) p_{i,j} + K_{i-1/2,j} p_{i-1,j}\Big] \\ &\quad -\frac{1}{k^2}\Big[K_{i,j+1/2} p_{i,j+1} - \Big(K_{i,j+1/2} + K_{i,j-1/2}\Big) p_{i,j} + K_{i,j-1/2} p_{i,j-1}\Big]. \end{aligned}$$

The Finite Difference Stencil

Rearranging terms, the (i,j)th equation is

$$\begin{aligned}\frac{1}{h^2k^2}\Big[&\Big(k^2K_{i+1/2,j}+k^2K_{i-1/2,j}+h^2K_{i,j+1/2}+h^2K_{i,j-1/2}\Big)p_{i,j}\\&-k^2K_{i+1/2,j}p_{i+1,j}-k^2K_{i-1/2,j}p_{i-1,j}\\&-h^2K_{i,j+1/2}p_{i,j+1}-h^2K_{i,j-1/2}p_{i,j-1}\Big]\\=\;&q_{i,j},\quad i=1,2,...,M-1,\;\; j=1,2,...,N-1.\end{aligned}$$

This method has a five point stencil: p_{ij} is related to its four nearest neighbors.



Linear System for the Dirichlet Problem—1

Our approximation is a system of linear equations for the interior p_{ij}. If we assume that K is constant and $h = k$, then we have simply

$$\frac{K}{h^2}\Big[4p_{i,j} - p_{i+1,j} - p_{i-1,j} - p_{i,j+1} - p_{i,j-1}\Big] = q_{i,j} \quad i = 1, 2, ..., M-1, \;\; j = 1, 2, ..., N-1.$$

We linearly order the interior grid points:

$$\boxed{m = i + (M-1)(j-1),}$$

which gives a unique m to each (i, j) in the interior. Then

$$\mathbf{p} = \Big(p_{11},\; p_{21},\; \cdots,\; p_{M-1,1},\; p_{12},\; p_{22},\; \cdots,\; p_{M-1,2},\; \cdots,\; p_{M-1,N-1}\Big)^T$$

and the right-hand-side (RHS) vector is

$$\mathbf{b} = \Big(q_{11} + \frac{K}{h^2}(p_{01} + p_{10}),\; q_{21} + \frac{K}{h^2}p_{20},\; \cdots,$$

$$q_{M-1,1} + \frac{K}{h^2}(p_{M1} + p_{M-1,0}), ..., q_{M-1,N-1} + \frac{K}{h^2}(p_{M,N-1} + p_{M-1,N})\Big)^T,$$

wherein we have included any p_{ij} on the boundary. Note that the corners p_{00}, $p_{M-1,0}$, $p_{0,N-1}$, and $p_{M-1,N-1}$ are never used.

Linear System for the Dirichlet Problem—2

The matrix is

$$A = \frac{K}{h^2}\begin{pmatrix} 4 & -1 & 0 & 0 & \dots & -1 & 0 & 0 & \dots & 0 \\ -1 & 4 & -1 & 0 & \dots & 0 & -1 & 0 & \dots & 0 \\ 0 & -1 & 4 & -1 & \dots & 0 & 0 & -1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ -1 & 0 & 0 & 0 & \dots & 4 & -1 & 0 & \dots & 0 \\ 0 & -1 & 0 & 0 & \dots & -1 & 4 & -1 & \dots & 0 \\ 0 & 0 & -1 & 0 & \dots & 0 & -1 & 4 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 4 \end{pmatrix},$$

which is $(M-1)(N-1) \times (M-1)(N-1)$. It has a banded structure.

We are reduced to solving the linear system

$$\boxed{A\mathbf{p} = \mathbf{b}.}$$

Theorem. There exists a unique solution to the linear system, and the error satisfies

$$\max_{i,j} \left| p(x_i, y_j) - p_{ij} \right| = \mathcal{O}(h^2 + k^2).$$

Approximation of the Neumann Problem

In the pure Neumann case,

$$\begin{cases} -\nabla \cdot (k\nabla p) = q, & \text{in } \Omega = (0, L_1) \times (0, L_2), \\ -(k\nabla p) \cdot \nu = f, & \text{on } \partial\Omega, \end{cases}$$

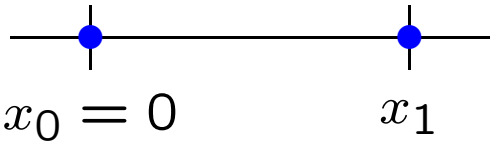
and we need to determine all p_{ij}, *not* just those in the interior.

We have the same discretized equations as in the Dirichlet case for the interior grid points

$$\begin{aligned} \frac{1}{h^2k^2}\Big[\Big(&k^2K_{i+1/2,j} + k^2K_{i-1/2,j} + h^2K_{i,j+1/2} + h^2K_{i,j-1/2}\Big)p_{i,j} \\ &- k^2K_{i+1/2,j}p_{i+1,j} - k^2K_{i-1/2,j}p_{i-1,j} \\ &- h^2K_{i,j+1/2}p_{i,j+1} - h^2K_{i,j-1/2}p_{i,j-1}\Big] \\ = q_{i,j}, \quad & i = 1, 2, ..., M-1, \ j = 1, 2, ..., N-1, \end{aligned}$$

and the same five point stencil.

Approximation of the Neumann BC—1



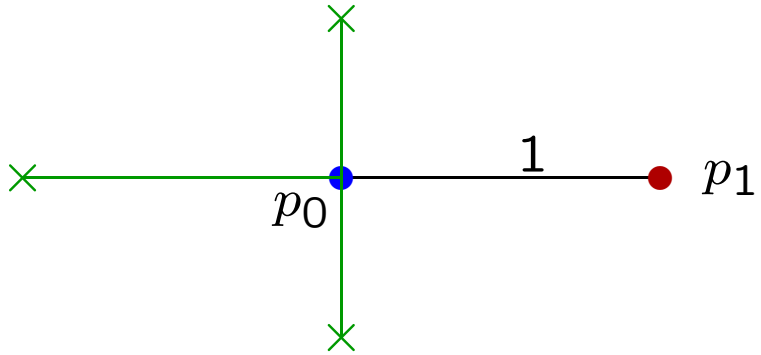
Consider the left edge of the domain where $x = 0$. Let $y = y_j$ be fixed (so we can ignore it for now). The BC says that

$$Kp'(0) = f(0),$$

If we make a standard finite difference approximation in terms of the two values p_0 and p_1, we obtain

$$K\frac{p_1 - p_0}{h} = f_0 \qquad \Longrightarrow \qquad p_0 = p_1 - \frac{h}{K}f_0.$$

Thus we add this to the equations we have. Note that we maintain the "five" point stencil (with three points missing), although the weights are different.



Approximation of the Neumann BC—2

Restoring the y variable, and treating the other three sides, gives us the following four sets of equations to add to the interior equations:

$$p_{0,j} = p_{1,j} - \frac{h}{K} f_{0,j}, \ \ j = 1, 2, ..., N,$$

$$p_{M+1,j} = p_{M,j} - \frac{h}{K} f_{M+1,j}, \ \ j = 1, 2, ..., N,$$

$$p_{i,0} = p_{i,1} - \frac{k}{K} f_{i,0}, \ \ i = 1, 2, ..., M,$$

$$p_{i,N+1} = p_{i,N} - \frac{k}{K} f_{i,N+1}, \ \ i = 1, 2, ..., M.$$

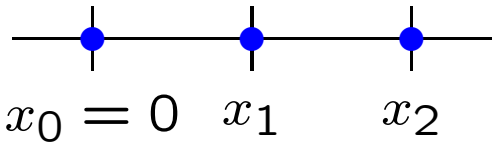
Problem. The BC approximation is only $\mathcal{O}(h + k)$ accurate, so we lose overall accuracy of the finite difference method.

Question: In fact, the convergence is $\mathcal{O}(h^{3/2} + k^{3/2})$. Can you explain why?

Approximation of the Neumann BC—3

Solutions.

- Approximate the boundary derivative by three points to obtain $\mathcal{O}(h^2)$ accuracy.

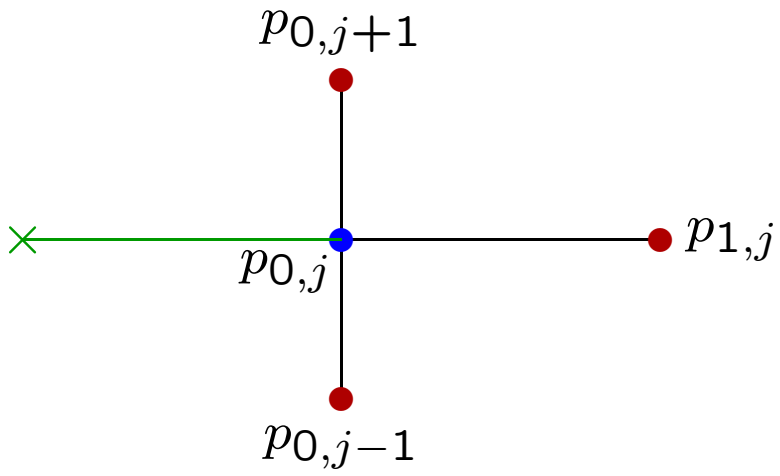


  Question: What happens to the stencil?

- Be more clever near the boundary, by using the fact that

$$p_{xx} + p_{yy} = -q.$$

  One can get an approximation within the five point stencil with one missing point that is $\mathcal{O}(h^2 + k^2)$ this way.



Linear System for the Neumann Problem–1

We linearly order the grid points, counting from left to right, and bottom to top. We have

$$(M+1)\times(N+1)-4$$

points, since the four corners are never used (or approximated). This gives us a linear system to solve for

$$\mathbf{p}=\left(p_{10},\ p_{20},\ p_{30},\ \cdots,\ p_{M,0},\ p_{01},\ p_{11},\ \cdots,\ p_{M+1,1},\ \cdots,\ p_{M,N+1}\right)^T$$

Again, we are reduced to solving a linear system

$$\boxed{A\mathbf{p}=\mathbf{b}.}$$

Linear System for the Neumann Problem–2

This system is singular, since the solution to the pure Neumann problem is not unique up to a constant. That is,

$$\mathbf{c} = \Big(1,\ 1,\ 1,\ \ldots\ , 1\Big)^T$$

is in the null space of A (i.e., $A\mathbf{c} = 0$). To solve this system, we need to replace one equation (the last?) by either

- $\sum_i \sum_j p_{ij} = 0$ (average pressure is zero),
- or $p_{10} = 0$ (or any other p_{ij}).

Theorem. There exists a unique solution to the linear system. Moreover, the error is $\mathcal{O}(h^{3/2} + k^{3/2})$ with the first order BC method, and $\mathcal{O}(h^2 + k^2)$ with the improved methods.

Remark. We can obtain higher order accuracy by taking a larger stencil. But this means that we need more computation, and so the code runs more slowly.

Numerical Solution by Cell Centered Finite Differences

Cell Centered Rectangular Grid

We define a uniform rectangular grid on $\Omega = \{0 < x < L_1,\ 0 < y < L_2\}$ by choosing integers $M \geq 1$ and $N \geq 1$, and setting the grid spacings to be

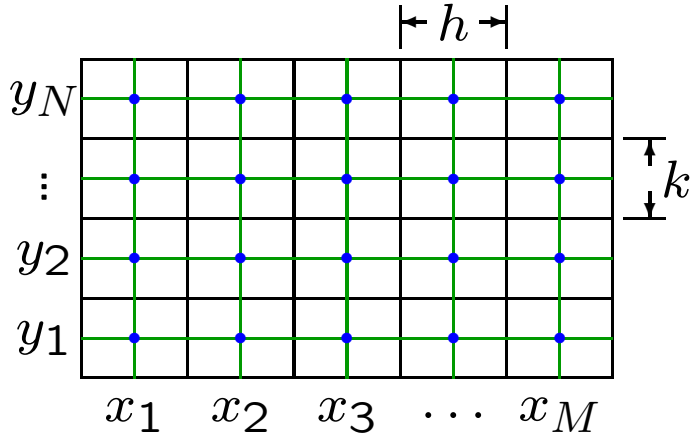
$$h = L_1/M \quad \text{and} \quad k = L_2/N,$$

and defining the grid lines at the half-grid points

$$x_{1/2} = 0,\ x_{3/2} = h,\ \ldots,\ x_{i-1/2} = (i-1/2)h,\ \ldots,\ x_{M+1/2} = L_1,$$
$$y_{1/2} = 0,\ y_{3/2} = h,\ \ldots,\ y_{j-1/2} = (j-1/2)h,\ \ldots,\ y_{N+1/2} = L_2.$$

What we care about now are the cell centers

$$x_1 = h/2,\ x_2 = 3h/2,\ \ldots,\ x_i = (i-1/2)h,\ \ldots,\ x_M = L_1 - h/2,$$
$$y_1 = k/2,\ y_2 = 3k/2,\ \ldots,\ y_j = (j-1/2)k,\ \ldots,\ y_N = L_2 - k/2.$$



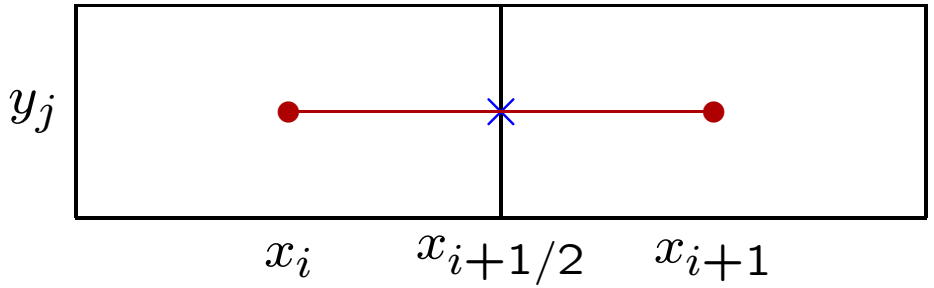
Remark. We have taken a uniform grid (i.e., constant x and y spacing) for simplicity. Most things gerneralize to nonuniform grids, and then the cell centered grid is fundamentally different from a vertex centered grid.

Approximation of the Velocity—1

We approximate the mixed system

$$\begin{cases} \mathbf{u} = -k\nabla p, & \text{in } \Omega, \\ \nabla \cdot \mathbf{u} = q, & \text{in } \Omega. \end{cases}$$

We begin with the velocity, and approximate it on grid element edges.



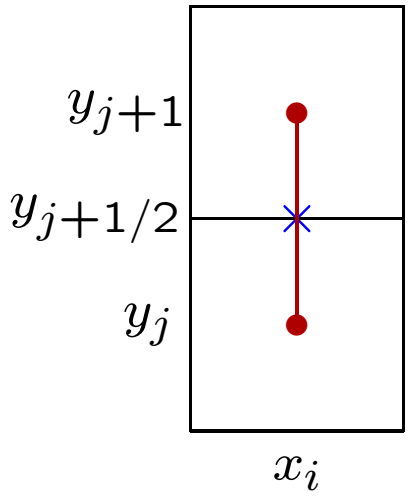
x-edges. For an x-edge, we approximate the normal velocity (i.e., u_1) as

$$u_1(x_{i+1/2}, y_j) = -Kp_x$$

$$\approx \boxed{u_{i+1/2,j} = -K_{i+1/2,j}\frac{p_{i+1,j} - p_{i,j}}{h}.}$$

This is $\mathcal{O}(h^2)$ accurate, and it works for nonuniform grids by simply replacing h by $x_{i+1} - x_i$.

Approximation of the Velocity—2

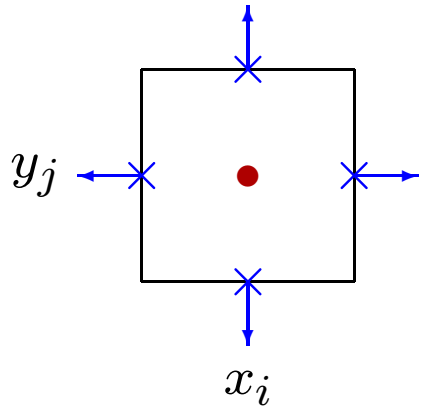


y-edges. For an y-edge, we approximate the normal velocity (i.e., u_2) as

$$u_2(x_{i+1/2}, y_j) = -Kp_y$$

$$\approx \boxed{u_{i,j+1/2} = -K_{i,j+1/2}\frac{p_{i,j+1} - p_{i,j}}{k}.}$$

Approximation of the Divergence



Recall

$$\nabla \cdot \mathbf{u} = q \qquad \Longrightarrow \qquad q = u_{1,x} + u_{2,y}.$$

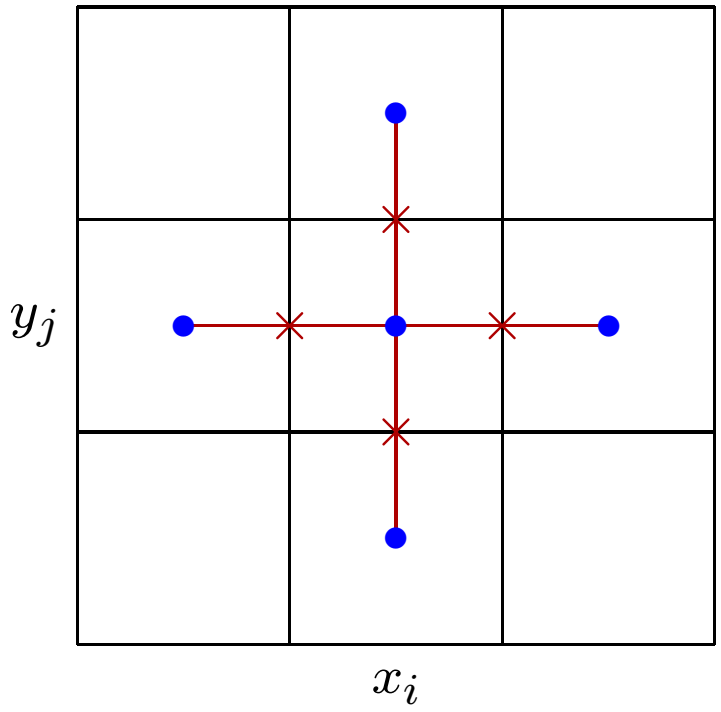
Thus we approximate at (x_i, y_j)

$$q_{i,j} = \frac{u_{i+1/2,j} - u_{i-1/2,j}}{h} + \frac{u_{i,j+1/2} - u_{i,j-1/2}}{k}.$$

Overall Approximation Stencil—1

Combining, we have a five point stencil

$$
\begin{aligned}
q_{i,j} &= \frac{u_{i+1/2,j} - u_{i-1/2,j}}{h} + \frac{u_{i,j+1/2} - u_{i,j-1/2}}{k} \\
&= -K_{i+1/2,j}\frac{p_{i+1,j} - p_{i,j}}{h} + K_{i-1/2,j}\frac{p_{i,j} - p_{i-1,j}}{h} \\
&\quad - K_{i,j+1/2}\frac{p_{i,j+1} - p_{i,j}}{k} + K_{i,j-1/2}\frac{p_{i,j} - p_{i,j-1}}{k}.
\end{aligned}
$$



Overall Approximation Stencil—2

Boundary Conditions. The Neumann BC is easily incorporated. For example, on an x-edge, above we would use

$$u_{1/2,j} = f(0, y_j).$$

For Dirichlet conditions, we simply approximate, for example,

$$u_{1/2,j} = -K_{1/2,j} \frac{p_{1,j} - p_D(0, y_j)}{h/2}.$$

Question: How accurate is this?

Theorem. The pressure is accurate to $\mathcal{O}(h^2 + k^2)$, and the velocity is $\mathcal{O}(h + k)$. This result holds for any BC's and nonuniform grids.

Question: Is this surprising?

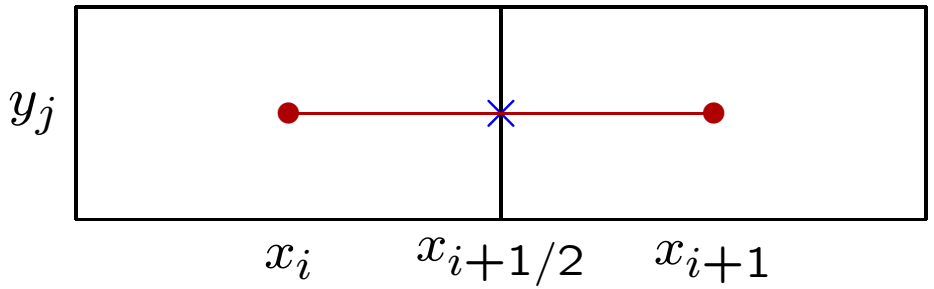
Evaluation of K on the Edges—1

For an x-edge,

$$u_{i+1/2,j} = -K_{i+1/2,j}\frac{p_{i+1,j} - p_{i,j}}{h}.$$

Often, one assumes that K is piecewise constant on grid elements, so where we need it, it is discontinuous.

Question: Should we just average the two values?



We have essentially 1-D flow from (x_i, y_j) to $(x_{i+1,j}, y_j)$.

Evaluation of K on the Edges—2

Since $q = 0$ most places, we almost have

$$\begin{cases} -(Kp')' = 0, \\ p(x_i) = p_i \text{ and } p(x_{i+1)} = p_{i+1}. \end{cases}$$

Thus

$$-Kp' = \text{Constant} = u,$$

so

$$p' = \begin{cases} u/K_i, & x_i < x < x_{i+1/2}, \\ u/K_{i+1}, & x_{i+1/2} < x < x_{i+1}, \end{cases}$$

which implies that

$$p = \begin{cases} p_i + (u/K_i)(x - x_i), & x_i < x < x_{i+1/2}, \\ p_{i+1} - (u/K_{i+1})(x_{i+1} - x), & x_{i+1/2} < x < x_{i+1}. \end{cases}$$

But p must be continuous at $x = x_{i+1/2}$.

Question: Why?

Evaluation of K on the Edges—3

So

$$p_i - \frac{u}{K_i}\frac{h}{2} = p_{i+1} - \frac{u}{K_{i+1}}\frac{h}{2},$$

which implies that

$$u = -2\left(\frac{1}{K_i} + \frac{1}{K_{i+1}}\right)^{-1}\frac{p_{i+1} - p_i}{h} = -K_{i+1/2}\frac{p_{i+1} - p_i}{h}.$$

Therefore, we should take

$$K_{i+1/2} = 2\left(\frac{1}{K_i} + \frac{1}{K_{i+1}}\right)^{-1} = \frac{2K_iK_{i+1}}{K_i + K_{i+1}},$$

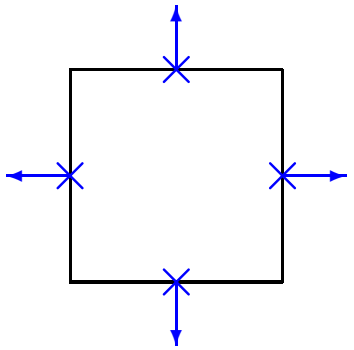
which is the harmonic average of K_i and K_{i+1}, which emphasizes small values. It is the reciprocal of the average of the reciprocals.

Local Mass Conservation

If E is the (i,j) grid element, then

$$\iint_E q\,dx\,dy = \iint_E \nabla\cdot\mathbf{u}\,dx\,dy = \int_{\partial E}\mathbf{u}\cdot\nu\,dS$$

is the integral of the normal velocities around ∂E.



If we define the source value as $q_{ij} = \dfrac{1}{hk}\iint_E q\,dx\,dy$, then our method is

$$\begin{aligned} q_{ij}\,h\,k &= [u_{i+1/2,j} - u_{i-1/2,j}]k + [u_{i,j+1/2} - u_{i,j-1/2}]h \\ &= \int_{y_{j-1/2}}^{y_{j+1/2}} [u_{i+1/2,j} - u_{i-1/2,j}]\,dy + \int_{x_{i-1/2}}^{x_{i+1/2}} [u_{i,j+1/2} - u_{i,j-1/2}]\,dx, \end{aligned}$$

which is also the integral of the normal velocities around ∂E.

That is, our method reproduces the local mass conservation principle exactly over each grid element. We say that it is locally conservative.

Vertex centered finite differences do *not* have this property!

Question: Why is local conservation important?

The Maximum Principle

Recall that local maxima cannot occur if $q \leq 0$ in the true solution.

Question: Can they occur in the numerical solution?

For the discrete solution, a local maximum would be a point such that

$$p_{ij} > \max\{p_{i-1,j},\ p_{i+1,j},\ p_{i,j-1},\ p_{i,j+1}\}.$$

Theorem. For any of our methods, this will imply flow away from the point (i.e., $q > 0$), so we indeed satisfy the maximum principle.

Question: If we write a code and find that the solution violates the maximum principle (or the local conservation principle for cell centered finite differences), what should we conclude?

Remark. Methods that approximate derivatives with more points (and so are more accurate, error $\mathcal{O}(h^r)$, $r > 2$) do not satisfy this property! Sometimes, we see small "nonphysical" flows, i.e., a small amount of fluid flowing in the wrong direction. We call such features numerical artifacts.

Question: Is this a paradox? More accurate but fails to satisfy the maximum principle.

Numerical Solution by Finite Elements

Integration by Parts

Theorem. The product rule for divergences is

$$\boxed{\nabla \cdot (\varphi \mathbf{v}) = \nabla \varphi \cdot \mathbf{v} + \varphi \nabla \cdot \mathbf{v}.}$$

Theorem. The integration by parts formula is

$$\boxed{\iiint_\Omega \nabla \varphi \cdot \mathbf{v} \, dx \, dy \, dz = \iint_\Omega \varphi \mathbf{v} \cdot \nu \, dS - \iiint_\Omega \varphi \nabla \cdot \mathbf{v} \, dx \, dy \, dz,}$$

or

$$\boxed{\iiint_\Omega \varphi \nabla \cdot \mathbf{v} \, dx \, dy \, dz = \iint_\Omega \varphi \mathbf{v} \cdot \nu \, dS - \iiint_\Omega \nabla \varphi \cdot \mathbf{v} \, dx \, dy \, dz.}$$

The proof is just an application of the divergence theorem and the product rule.

Question: Why do we call this "integration by parts?"

The Weak Form of the Problem—1

Our PDE is

$$-\nabla \cdot (k \nabla p) = q, \quad \text{in } \Omega.$$

Strategy.

- Multiply the PDE by a function $\varphi(x, y, z)$, called a test function.
- Integrate over Ω.
- Integrate by parts to even out the derivatives on the solution p and the test function φ.

Remark. We lose nothing by this "testing process," since if

$$\iiint f \varphi \, dx \, dy \, dz = \iiint g \varphi \, dx \, dy \, dz$$

for every test function φ, then

$$f = g.$$

That is, if two two functions agree in all their tests, they are the same!

The Weak Form of the Problem—2

The full problem is

$$\begin{cases} -\nabla \cdot (k\nabla p) = q, & \text{in } \Omega, \\ p = p_D, & \text{on } \Gamma_D, \\ -(k\nabla p) \cdot \nu = f, & \text{on } \Gamma_N. \end{cases}$$

Following our strategy, we get

$$\begin{aligned} \iiint_\Omega q\,\varphi\,dV &= -\iiint_\Omega \nabla \cdot (k\nabla p)\varphi\,dV \\ &= \iiint_\Omega k\nabla p \cdot \nabla\varphi\,dV - \iint_{\partial\Omega} k\nabla p \cdot \nu\varphi\,dS. \end{aligned}$$

Boundary Conditions. Neumann: We replace the normal derivative by f. Dirichlet: We impose the BC directly (i.e., not via the test function).

The weak form of the equations. Assume that

$$p = p_D \quad \text{and} \quad \varphi = 0 \quad \text{on } \Gamma_D.$$

Then we have the weak form of the equations

$$\boxed{\iiint_\Omega k\nabla p \cdot \nabla\varphi\,dV = \iiint_\Omega q\,\varphi\,dV - \iint_{\Gamma_N} f\,\varphi\,dS.}$$

Theorem. The PDE problem is equivalent to the weak problem.

The Galerkin Method

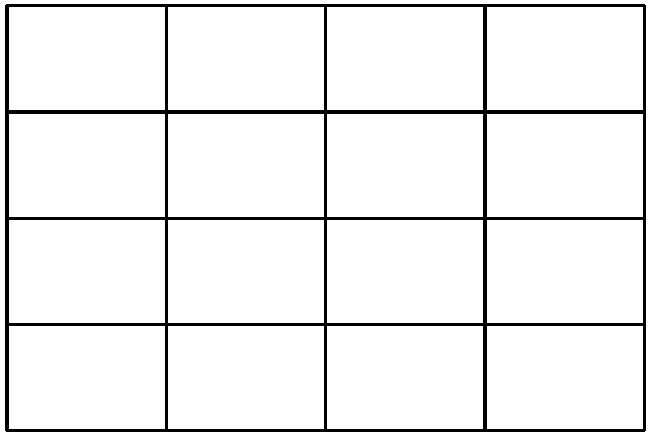
The basic ideas of Galerkin's method.

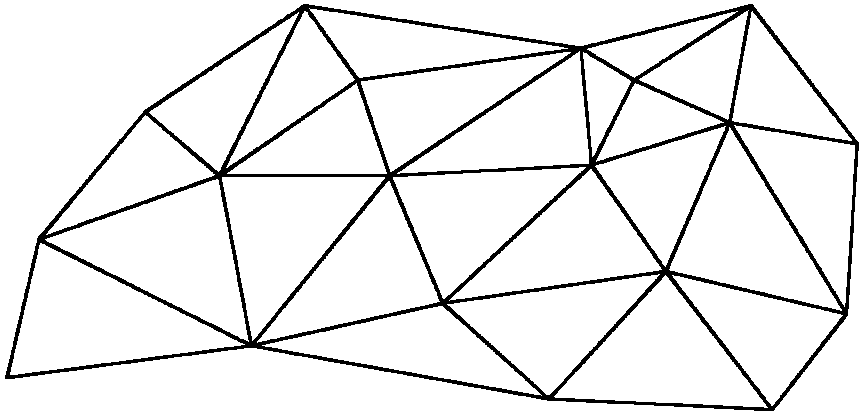
- Work on the weak problem, rather than the PDE problem.
- Approximate the weak problem by replacing p and φ by simple functions.
- Solve for the simplified p.

The finite element method (FEM) is a divide-and conquer strategy to find a simple representation for p (and φ).

The Computational Grid or Mesh

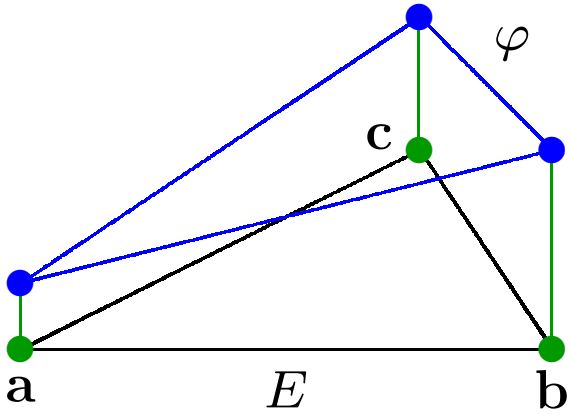
Let us restrict for simplicity to 2-D. We divide the domain into rectangles and/or triangles. If we use triangles, Ω can be irregularly shaped.





Note that all grid elements completely share their faces.

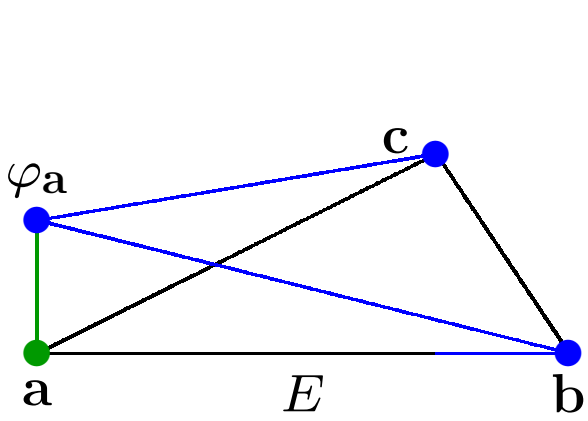
Piecewise Linear Polynomials on Triangles—1

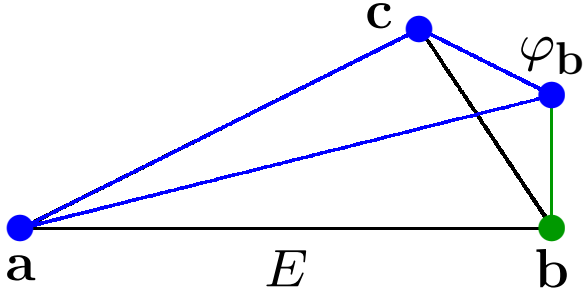


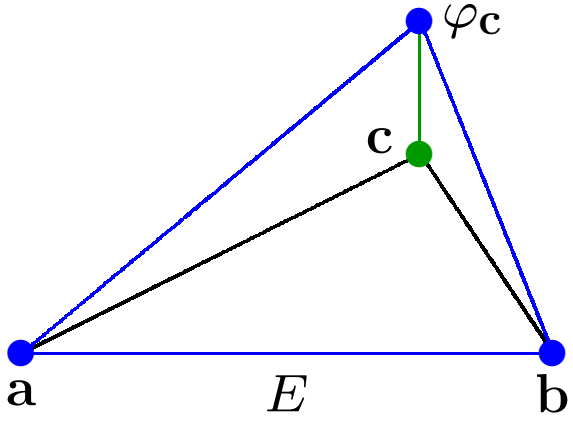
Over each triangle, let φ be a linear polynimial

$$\varphi(x, y) = \alpha + \beta x + \gamma y.$$

This is a three dimensional vector space, with the following basis.





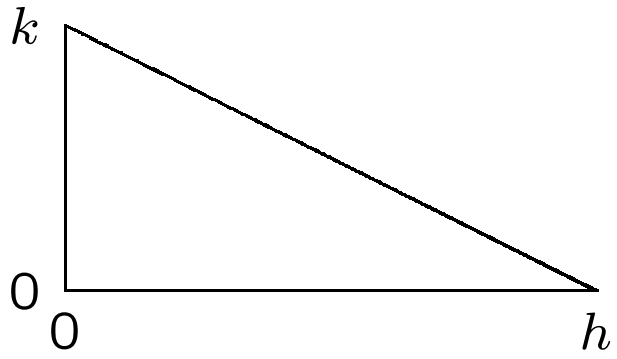


Then

$$\varphi(x, y) = \varphi(\mathbf{a})\varphi_{\mathbf{a}}(x, y) + \varphi(\mathbf{b})\varphi_{\mathbf{b}}(x, y) + \varphi(\mathbf{c})\varphi_{\mathbf{c}}(x, y).$$

Piecewise Linear Polynomials on Triangles—2

Example.



Then

$$\varphi_{0,0}(x,y) = 1 - \frac{x}{h} - \frac{y}{k}, \quad \varphi_{h,0}(x,y) = \frac{x}{h}, \quad \varphi_{0,k}(x,y) = \frac{y}{k},$$

since three points determine a plane and

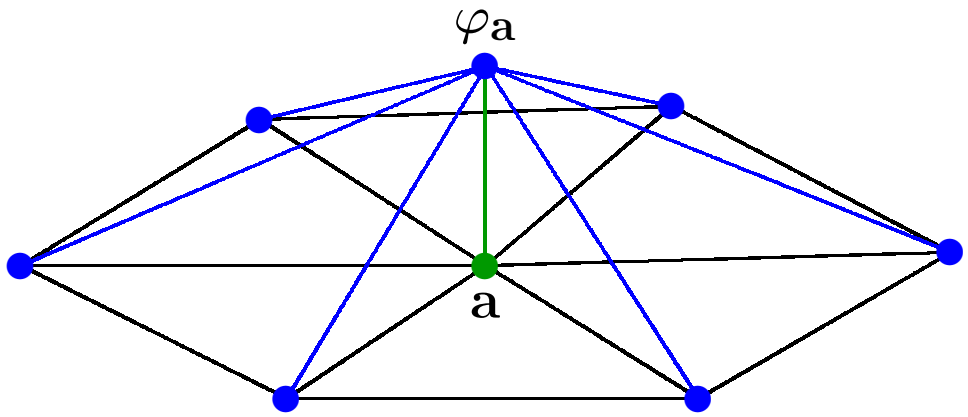
$$\begin{aligned}
&\varphi_{0,0}(0,0) = 1, && \varphi_{h,0}(0,0) = 0, && \varphi_{0,k}(0,0) = 0,\\
&\varphi_{0,0}(h,0) = 0, && \varphi_{h,0}(h,0) = 1, && \varphi_{0,k}(h,0) = 0,\\
&\varphi_{0,0}(0,k) = 0, && \varphi_{h,0}(0,k) = 0, && \varphi_{0,k}(0,k) = 1.
\end{aligned}$$

Finally, for example,

$$\begin{aligned}
\varphi(x,y) &= 3 + 2x - 4y\\
&= 3\,\varphi_{0,0}(x,y) + (3+2h)\,\varphi_{h,0}(x,y) + (3-4k)\,\varphi_{0,k}(x,y).
\end{aligned}$$

Piecewise Linear Polynomials on Triangles—3

We now piece these functions together to form a global basis function, as follows.



Finite Element Approximation

We approximate p by

$$P(x,y) = \sum_{\mathbf{a}\notin\Gamma_D} \alpha_{\mathbf{a}}\varphi_{\mathbf{a}}(x,y) + \sum_{\mathbf{a}\in\Gamma_D} p_D(\mathbf{a})\varphi_{\mathbf{a}}(x,y)$$

and find the coefficients $\alpha_{\mathbf{a}}$ satisfying the weak equations

$$\iint_\Omega k\nabla P\cdot\nabla\varphi_{\mathbf{b}}\,dV = \iint_\Omega q\,\varphi_{\mathbf{b}}\,dV - \int_{\partial\Omega} f\,\varphi_{\mathbf{b}}\,dS,$$

for all $\mathbf{b}\notin\Gamma_D$. This is,

$$\sum_{\mathbf{a}\notin\Gamma_D} \alpha_{\mathbf{a}} \iint_\Omega k\nabla\varphi_{\mathbf{a}}\cdot\nabla\varphi_{\mathbf{b}}\,dV = \iint_\Omega q\,\varphi_{\mathbf{b}}\,dV + \int_{\partial\Omega} f\,\varphi_{\mathbf{b}}\,dS - \sum_{\mathbf{a}\in\Gamma_D} p_D(\mathbf{a}) \iint_\Omega k\nabla\varphi_{\mathbf{a}}\cdot\nabla\varphi_{\mathbf{b}}\,dV,$$

Again, this is a matrix problem (since we have two indices, $\mathbf{a}$ and $\mathbf{b}$).

The Linear System

We solve

$$\boxed{A\mathbf{P} = \mathbf{r},}$$

where, for $\mathbf{a}$ and $\mathbf{b}$ not in Γ_D,

$$\begin{aligned} P_{\mathbf{a}} &= \alpha_{\mathbf{a}}, \\ A_{\mathbf{a},\mathbf{b}} &= \iint_{\Omega} k \nabla \varphi_{\mathbf{a}} \cdot \nabla \varphi_{\mathbf{b}} \, dV, \\ r_{\mathbf{b}} &= \iint_{\Omega} q \, \varphi_{\mathbf{b}} \, dV + \int_{\partial\Omega} f \, \varphi_{\mathbf{b}} \, dS \\ &\quad - \sum_{\mathbf{c} \in \Gamma_D} p_D(\mathbf{c}) \iint_{\Omega} k \nabla \varphi_{\mathbf{a}} \cdot \nabla \varphi_{\mathbf{b}} \, dV, \end{aligned}$$

Remark. Most matrix entries are 0. We call such a matrix sparse.

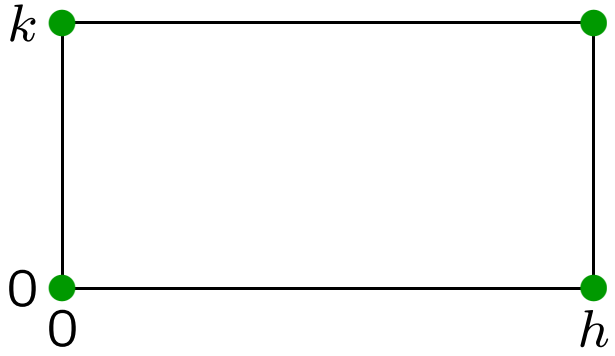
Question: How do we find these integrals?

Piecewise Bilinears on Rectangles—1

On a rectangle, we use a bilinear polynomial of the form

$$\varphi(x,y) = \alpha + \beta x + \gamma y + \delta xy.$$

This four dimensional vector space has the following basis.

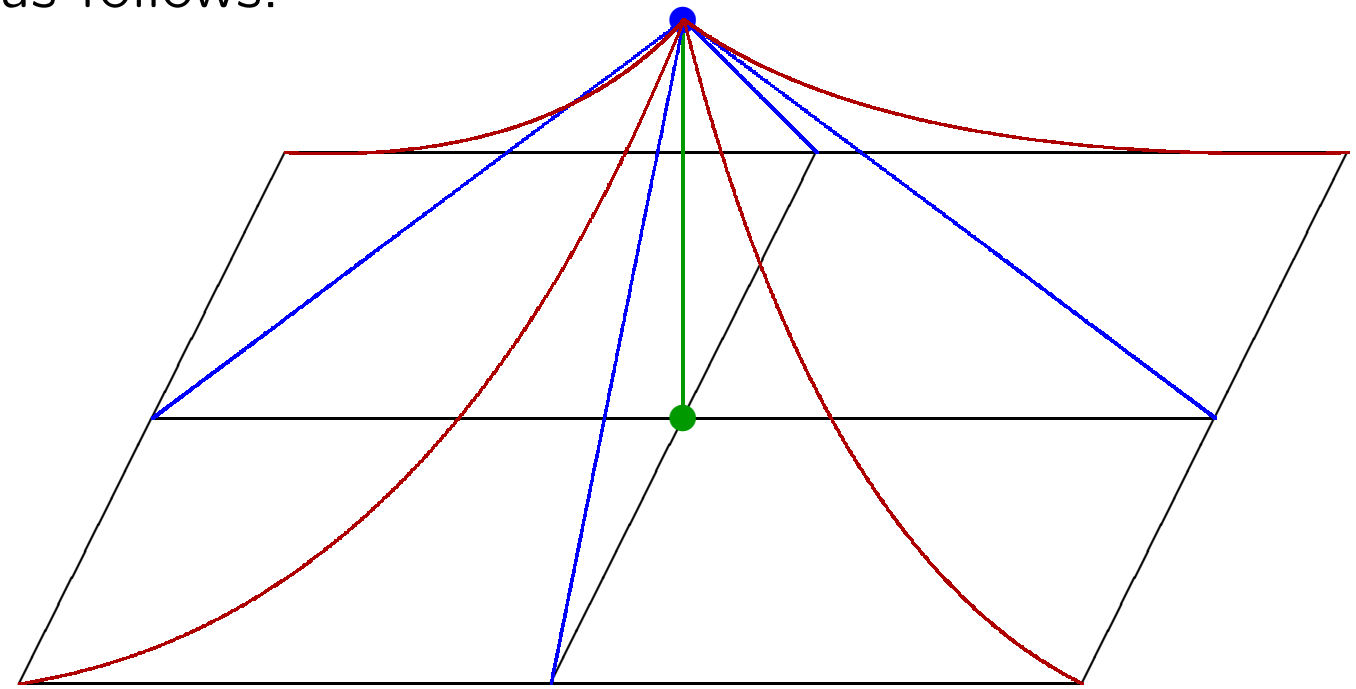


$$\varphi_{0,0} = \left(1 - \frac{x}{h}\right)\left(1 - \frac{y}{k}\right), \quad \varphi_{h,0} = \frac{x}{h}\left(1 - \frac{y}{k}\right),$$

$$\varphi_{0,k} = \left(1 - \frac{x}{h}\right)\frac{y}{k}, \quad \varphi_{h,k} = \frac{x}{h}\frac{y}{k}.$$

Piecewise Bilinears on Rectangles—2

These functions piece together to form a global basis function over four rectangles, as follows.



Convergence—1

Taylor's theorem says that if p has derivatives of order $r+1$, then it is a polynomial of degree r plus a remainder

$$p(x,y) = P_n(x,y) + R_n(x,y),$$

where the remainder R_n depends on the $(r+1)$st derivatives. In 1-D,

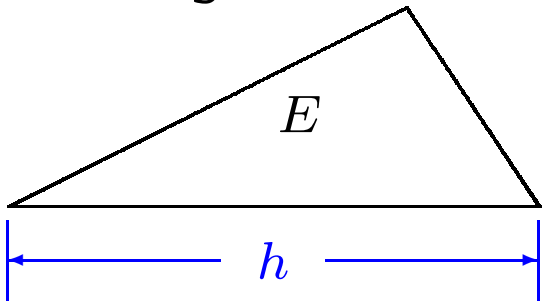
$$R_n(x) = \frac{1}{n!}\int_0^x f^{(n+1)}(t)\,(x-t)^n\,dt.$$

In the finite element method, we locally approximate p by a polynomial of degree r, where $r = 1$ for piecewise linears or bilinears, but we could also use higher order polynomials. Hopefully, then, the error in the method is like Taylor's theorem, depending on

- the size of the deviation from the base point;
- the size of the $(r+1)$st derivatives.

Convergence—2

Let h be the maximal size of the grid elements.



Theorem. If the finite elements are based on polynomials of degree r, then the error satisfies

$$\left\{ \iint_\Omega \Big(p(x,y) - P(x,y)\Big)^2 \, dx\, dy \right\}^{1/2} = \mathcal{O}(h^{r+1}),$$

$$\left\{ \iint_\Omega \|\nabla\Big(p(x,y) - P(x,y)\Big)\|^2 \, dx\, dy \right\}^{1/2} = \mathcal{O}(h^{r}).$$

Remarks.

- This bounds the error on the average rather than at points.
- It is unprofitable to use high degree polynomials if p has large derivatives! This is typically the case for porous media.
- The method fails to satisfy the local mass conservation principle.
- The method does not necessarily satisfy the maximum principle.