Understanding the Rational Approximation of the Exponential Integrator (REXI)

Martin Schreiber <M.Schreiber@exeter.ac.uk>
Pedro S. Peixoto <pedrosp@ime.usp.br>

May 10, 2018

This document serves as the basis for implementing the Rational approximation of the EXponential Integrator (REXI). Here, we purely focus on the linear part of the shallow-water equations (SWE) and show the different steps to approximate solving this linear part with an exponential integrator. This paper mainly summarises previous work on REXI.

1 Problem formulation

We use linearised shallow water equations (SWE) with respect to a rest state with mean water depth of η_0 and defined for perturbations of height η (see [1]). The linear operator (L) may be written as

$$L(U) := \begin{pmatrix} 0 & -\eta_0 \partial_x & -\eta_0 \partial_y \\ -g\partial_x & 0 & f \\ -g\partial_y & -f & 0 \end{pmatrix} U$$

where $U := (\eta, u, v)^T$. Here, we neglect all non-linear terms and consider f constant (f-plane approximation). The time evolution of the PDE, with the subscript t denoting the derivative in time, is given by

$$U_t = L(U).$$

It is further worth noting, that this system describes an oscillatory system (2D wave equation), hence the operator L is hyperbolic and has imaginary eigenvalues. Linear initial value differential problems are well known to be solvable with exponential integrators for arbitrary time step sizes via

$$U(t) = e^{Lt} U(0).$$

see e.g. [3]. However, this is typically quite expensive to compute and analytic solutions only exist for some simplified system of equations, see e.g. [1] for f-plane shallow-water equations. These exponential integrators can be approximated with rational functions and this paper is on giving insight into this approximation.

2 1D rational approximation

Terry et al. [2] developed a rational approximation of the exponential integrator. First, we like to get more insight into it with a one-dimensional formulation before applying REXI to a rational approximation of a linear operator. Our main target is to find an approximation of an operator with a *complex exponential shape*, in our case e^{ix}, which (in one-dimension) is given as a function $f(x)$. We will end up in an approximation given by the following rational approximation:

$$e^{ix} \approx \sum_{n=-N}^{N} Re\left(\frac{\beta_n}{ix+\alpha_n}\right)$$

with complex coefficients α_n and β_n. We point out that the coefficients α_n will always have non zero real part, so no singularity occurs with the rational function.

2.1 Step A: Approximation of solution space

First, we assume that we can use Gaussian curves as basis functions for our approximation. So first we find an approximation of one of our underlying Gaussian basis function

$$\psi_h(x) := (4\pi)^{-\frac{1}{2}} e^{-x^2/(4h^2)}$$

In this formulation, h can be interpreted as the horizontal "stretching" of the basis function. Note the similarities to the Gaussian distribution, but by dropping certain parts of the vertical scaling as it is required for probability distributions. We can now approximate our function $f(x)$ with a superposition of basis functions $\psi_h(x)$ by

$$f(x) \approx \sum_{m=-M}^{M} b_m \psi_h(x+mh)$$

with M controlling the interval of approximation (~size of "domain of interest") and h will be related to the accuracy of integration (~resolution in "domain of interest"). We choose h small enough so that the support of the Fourier transform of f is mainly localised within $[-1/(2h), 1/(2h)]$, i.e. almost zero outside this interval. M is chosen such that the approximation will be adequate in the interval $|x| < Mh$.

To compute the coefficients b_m, we rewrite the previous equation in Fourier space with

$$\frac{\hat{f}(\xi)}{\hat{\psi}_h(\xi)} = \sum_{m=\infty}^{\infty} b_m e^{2\pi i m h \xi},$$

where the $\hat{\cdot}$ symbols indicate the Fourier transforms of the respective functions. The b_m are now the Fourier coefficients of the series for the function $\frac{\hat{f}(\xi)}{\hat{\psi}_h(\xi)}$ and

can be calculated as (see [2], page 11),

$$b_m = h \int_{-\frac{1}{2h}}^{\frac{1}{2h}} e^{-2\pi i m h \xi} \frac{\hat{f}(\xi)}{\hat{\psi_h}(\xi)} d\xi,$$

for m integer and $1/h$ defines the periodicity of the trigonometric basis function. The reformulation from the discrete sum over this formulation can be understood by a Riemann integral over one of the intervals. Therefore, we also don't require the translated Gaussian since the integrated interval is already shifted to the origin.

Since we are interested in approximating $f(x) = e^{ix}$, we can simplify the equation by using the response in frequency space $\hat{f}(\xi) = \delta(\xi - \frac{1}{2\pi})$, where here δ is the Dirac distribution, and

$$b_m = h\, e^{-imh} \hat{\psi_h} \left(\frac{1}{2\pi} \right)^{-1}.$$

The Fourier transform of the Gaussian function is well known and given by

$$\begin{aligned}
\hat{\psi}_h(\xi) &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{4\pi}} e^{-\left(\frac{x}{2h}\right)^2} e^{-2\pi i x \xi} dx \\
&= \frac{1}{\sqrt{4\pi}} \int_{-\infty}^{\infty} e^{-\left(\left(\frac{x}{2h}\right)^2 + 2\pi i x \xi + (2h\pi i \xi)^2 - (2h\pi i \xi)^2\right)} dx \\
&= \frac{1}{\sqrt{4\pi}} e^{-(2h\pi\xi)^2} \int_{-\infty}^{\infty} e^{-\left(\frac{x}{2h} + 2h\pi i \xi\right)^2} dx \\
&= \frac{1}{\sqrt{4\pi}} e^{-(2h\pi\xi)^2} \int_{-\infty}^{\infty} e^{-\left(\frac{x}{2h}\right)^2} dx \\
&= h e^{-(2h\pi\xi)^2}
\end{aligned}$$

where we used that $\int_{-\infty}^{\infty} e^{-\left(\frac{x}{2h}\right)^2} dx = h\sqrt{4\pi}$ and completed squares in the exponential term. For the Gaussian function including a shift $\psi_h(x + hm)$, we get

$$\begin{aligned}
\hat{\psi}_h(\xi) &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{4\pi}} e^{-\left(\frac{x+hm}{2h}\right)^2} e^{-2\pi i x \xi} dx \\
&= h e^{-(2h\pi\xi)^2} e^{i 2\pi \xi (hm)}
\end{aligned}$$

For the case $\xi = \frac{1}{2\pi}$, we get

$$\hat{\psi}_h\left(\frac{1}{2\pi}\right) = h\, e^{-h^2}$$

or including the translation in the Gaussian we get

$$\hat{\psi}_h\left(\frac{1}{2\pi}\right) = h\, e^{-h^2} e^{ihm}.$$

Finally, one can obtain the equation

$$b_m = h\, e^{-imh} \frac{1}{h\, e^{-h^2}} = e^{-imh} e^{h^2}$$

to compute the coefficients b_m for $f(x) = e^{ix}$.

2.2 Step B: Approximation of basis function

The second step is the approximation of the basis function $\psi_h(x)$ itself with a rational approximation, see [4]. Our basis function is given by

$$\psi_h(x) := (4\pi)^{-\frac{1}{2}} e^{-x^2/(4h^2)}$$

and a close-to-optimal approximation of $\psi_1(x)$ with a sum of rational functions is given by

$$\psi_1(x) \approx Re\left(\sum_{l=-K}^{K} \frac{a_l}{ix + (\mu + i\, l)}\right)$$

with the μ and a_l given in [4], Table 1. We can generalise this approximation to arbitrary chosen h via

$$\psi_h(x) \approx Re\left(\sum_{l=-K}^{K} \frac{a_l}{i\frac{x}{h} + (\mu + i\, l)}\right).$$

The theory of how these coefficients are calculated are in [4] and will not be described here. Therefore, we assume that the coefficients a_l are given.

2.3 Step C: Approximation of the approximation

We then combine the approximation (B) into the approximation (A), yielding the approximation $\tilde{f}$ for f given by

$$\tilde{f}(x) = \sum_{m=-M}^{M} b_m \psi_h(x + mh) = \sum_{m=-M}^{M} b_m Re\left(\sum_{l=-K}^{K} \frac{a_l}{i\frac{x+mh}{h} + (\mu + i\, l)}\right)$$

$$= \sum_{m=-M}^{M} b_m \sum_{k=-K}^{K} Re\left(\frac{h a_l}{ix + h(\mu + i(m+k))}\right).$$

We further like to simplify this equation and we observe, that for $n := m+k$, the denominator is equal. We can hence express parts of the denomiator $ix + h(\mu + i(m+k))$ in terms of $n := m+k$ by

$$\alpha_n := h(\mu + i(m+k)) = h(\mu + in)). \tag{1}$$

Now, we merge the b_m and a_l coefficients and first have a look at the b_m which is complex valued. We observe the following property: Assuming that we want to compute the real value of $f(x)$, only the real value of b_m has to be merged with the sum, since the imaginary component would be dropped afterwards. This allows us to move the $Re(b_m)$ values inside the $\sum_K$:

$$Re(f(x)) := Re\left(\sum_{m=-M}^{M}\sum_{k=-K}^{K}\frac{Re(b_m)\,ha_l}{ix + h(\mu + i(m+k))}\right).$$

Now we can collect all nominators with equivalent denominator (if $n = m+k$ and by using δ as the Kronecker delta), yielding

$$\beta_n^{Re} := h\sum_{m=-M}^{M}\sum_{k=-K}^{K} Re(b_m)a_k\delta(n,\, m+k) \tag{2}$$

for real values $f(x)$ and

$$\beta_n^{Im} := h\sum_{m=-M}^{M}\sum_{k=-K}^{K} Im(b_m)a_k\delta(n,\, m+k) \tag{3}$$

for complex values of $f(x)$. Note, that the complexity of this operation is $O(N\,K)$, which is negligible for small N and K. This can be optimized by using $L_1 = max(-K, n-M)$ and $L_2 = min(K, n+M)$ (see [2]) and we can compute β_n^{Re} with

$$\beta_n^{Re} := h\sum_{k=L_1}^{L_2} Re(b_{n-k})a_k, \tag{4}$$

and β_n^{Im} correspondingly.

Finally, we get the REXI approximation

$$e^{ix} \approx \sum_{n=-N}^{N} Re\left(\frac{\beta_n^{Re}}{ix+\alpha_n}\right) + i\,Re\left(\frac{\beta_n^{Im}}{ix+\alpha_n}\right).$$

2.4 Merging real and imaginary exp-approximation

With the previous formulation, we would have to extract the real components from the complex values. However, this is not applicable if we use this for a linear operator. Therefore we require the system to cancel out the imaginary parts.

$$
\begin{aligned}
e^{ix} &\approx \sum_{n=-N}^{N} Re\left(\frac{\beta_n^{Re}}{ix+\alpha_n}\right) + i\,Re\left(\frac{\beta_n^{Im}}{ix+\alpha_n}\right) \\
&= \frac{1}{2}\sum_{n=-N}^{N}\left(\frac{\beta_n^{Re}}{ix+\alpha_n}\right) + \overline{\left(\frac{\beta_n^{Re}}{ix+\alpha_n}\right)} + i\left(\left(\frac{\beta_n^{Im}}{ix+\alpha_n}\right) + \overline{\left(\frac{\beta_n^{Im}}{ix+\alpha_n}\right)}\right) \\
&= \frac{1}{2}\sum_{n=-N}^{N}\left(\frac{\beta_n^{Re}+i\beta_n^{Im}}{ix+\alpha_n}\right) + \overline{\left(\frac{\beta_n^{Re}-i\beta_n^{Im}}{ix+\alpha_n}\right)} \\
&= \frac{1}{2}\sum_{n=-N}^{N}\left(\frac{\beta_n^{Re}+i\beta_n^{Im}}{ix+\alpha_n}\right) + \left(\frac{\overline{\beta_n^{Re}-i\beta_n^{Im}}}{\overline{ix+\alpha_n}}\right) \\
&= \frac{1}{2}\sum_{n=-N}^{N}\left(\frac{\beta_n^{Re}+i\beta_n^{Im}}{ix+\alpha_n}\right) + \left(\frac{\overline{\beta_n^{Re}-i\beta_n^{Im}}}{-ix+\overline{\alpha_n}}\right) \\
&= \frac{1}{2}\sum_{n=-N}^{N}\left(\frac{\beta_n^{Re}+i\beta_n^{Im}}{ix+\alpha_n}\right) + \left(\frac{\overline{-\beta_{-n}^{Re}+i\beta_{-n}^{Im}}}{ix-\alpha_n}\right) \\
&= \frac{1}{2}\sum_{n=-N}^{N}\left(\frac{\beta_n^{Re}+i\beta_n^{Im}}{ix+\alpha_n}\right) + \left(\frac{\overline{-\beta_{-n}^{Re}}+\overline{i\beta_{-n}^{Im}}}{ix-\alpha_n}\right) \\
&= \frac{1}{2}\sum_{n=-N}^{N}\left(\frac{\beta_n^{Re}+i\beta_n^{Im}}{ix+\alpha_n}\right) - \left(\frac{\overline{\beta_{-n}^{Re}}+i\overline{\beta_{-n}^{Im}}}{ix-\alpha_n}\right) \\
&= \frac{1}{2}\sum_{n=-N}^{N}\left(\frac{\beta_n^{Re}+i\beta_n^{Im}}{ix+\alpha_n}\right) - \left(\frac{\beta_n^{Re}-i\beta_n^{Im}}{ix-\alpha_n}\right)
\end{aligned}
$$

Where we used $\beta_n^{Re} = \overline{\beta_{-n}^{Re}}$ and $\beta_n^{Im} = -\overline{\beta_{-n}^{Im}}$. Finally, we gain new β and α coefficients which don't require application of the $Re()$ operator. However, one drawback is that the number of coefficients is enlarged by a factor of two.

2.5 Example coefficients

For a better understanding and discussion of the poles, we provide some explanatory coefficients for α_n and β_n^{Re} computed with $M := 2$ and $h := 0.2$:

n	α_n	β_n^{Re}
-13	(-0.863064302175, -2.6)	(-2.0794560075645e-08,5.312368394177e-09)
-12	(-0.863064302175, -2.4)	(-1.8562925598646e-08,-1.6892470811809e-07)
-11	(-0.863064302175, -2.2)	(6.8570271350932e-07,-4.4377515257134e-08)
-10	(-0.863064302175, -2)	(1.9470768200785e-07,2.1186231739561e-06)
-9	(-0.863064302175, -1.8)	(3.037169144916e-06,-3.8007524015554e-06)
-8	(-0.863064302175, -1.6)	(-0.00020292956274934,-9.4793805592883e-05)
-7	(-0.863064302175, -1.4)	(0.00051562027155282,0.0033198141762956)
-6	(-0.863064302175, -1.2)	(0.023802856324805,-0.020097812439831)
-5	(-0.863064302175, -1)	(-0.16210306892042,-0.057527918763957)
-4	(-0.863064302175, -0.8)	(0.083936569694558,0.55379453117192)
-3	(-0.863064302175, -0.6)	(0.87683903065806,-0.58136186212318)
-2	(-0.863064302175, -0.4)	(-0.87618099667542,-0.6444132979014)
-1	(-0.863064302175, -0.2)	(-0.2112750856805,0.51693268636776)
0	**(-0.863064302175, 0)**	**(0.21113064943379,1.1012434042446e-07)**
1	(-0.863064302175, 0.2)	(-0.2112752777559,-0.51693263772868)
2	(-0.863064302175, 0.4)	(-0.87618105783081,0.6444131761443)
3	(-0.863064302175, 0.6)	(0.87683907406497,0.58136183517238)
4	(-0.863064302175, 0.8)	(0.083936534477108,-0.55379454106338)
5	(-0.863064302175, 1)	(-0.16210304313401,0.057527824955638)
6	(-0.863064302175, 1.2)	(0.023802980792584,0.020097827969804)
7	(-0.863064302175, 1.4)	(0.00051562077173168,-0.0033196934926057)
8	(-0.863064302175, 1.6)	(-0.0002030221163996,9.4802957184526e-05)
9	(-0.863064302175, 1.8)	(3.0281700967037e-06,3.7434363774526e-06)
10	(-0.863064302175, 2)	(2.2311216999616e-07,-2.1234907990132e-06)
11	(-0.863064302175, 2.2)	(6.871098037128e-07,5.5123982746463e-08)
12	(-0.863064302175, 2.4)	(-2.1322288893395e-08,1.6899352278552e-07)
13	(-0.863064302175, 2.6)	(-2.0738399377275e-08,-5.6642992624128e-09)

Merges alpha and beta coefficients have similar properties.

3 REXI on linear operators

In this section, we investigate the linear operator L with the rational approximation.

3.1 Cancellation of imaginary values

For sake of simplicity, we only consider the first or second half of the merged α and β from the previous section! In particular, with $Im(\alpha_0) = Im(\beta_0) = 0$, there is a zero imaginary number for the central pole. Then, there is an anti-symmetry around the central pole with

$$\alpha_{-n} = \overline{\alpha_n} \tag{5}$$

and

$$\beta_{-n} = \overline{\beta_n}. \tag{6}$$

We start with REXI given by

$$\begin{aligned} e^{\tau L} &\approx \sum_{n=-N}^{N} \beta_n (\tau L + \alpha_n)^{-1} \\ &= \sum_{n=-N}^{-1} \beta_n (\tau L + \alpha_n)^{-1} + \beta_0 (\tau L + \alpha_0)^{-1} + \sum_{n=-N}^{-1} \beta_{-n} (\tau L + \alpha_{-n})^{-1} \end{aligned}$$

and using the properties (5) and (6), we can write this as

$$\begin{aligned} & \sum_{n=-N}^{-1} \beta_n \left(\tau L + \alpha_n\right)^{-1} + \beta_0 (\tau L + \alpha_0)^{-1} + \sum_{n=-N}^{-1} \overline{\beta_n} (\tau L + \overline{\alpha_n})^{-1} \\ = & \sum_{n=-N}^{-1} \beta_n \left(\tau L + \alpha_n\right)^{-1} + \beta_0 (\tau L + \alpha_0)^{-1} + \sum_{n=-N}^{-1} \overline{\beta_n (\tau L + \alpha_n)^{-1}}. \end{aligned}$$

which shows that the imaginary parts cancel out.

This is also shows (see Sec. 3.3 in [2])

$$\overline{(L+\alpha)^{-1} U(0)} = (L + \overline{\alpha})^{-1} U(0).$$

3.2 Reducing number of computations for L

For L and $U(0)$ real-valued only and with the overbar denoting the complex conjugate. We can then reformulate the approximation

$$\begin{aligned} e^{\tau L} &\approx \sum_{n=-N}^{-1} \beta_n \left(\tau L + \alpha_n\right)^{-1} + \beta_0 \left(\tau L + \alpha_0\right)^{-1} + \sum_{n=-N}^{-1} \beta_{-n} \left(\tau L + \alpha_{-n}\right)^{-1} \\ &= \sum_{n=-N}^{-1} \beta_n \left(\tau L + \alpha_n\right)^{-1} + \beta_0 \left(\tau L + \alpha_0\right)^{-1} + \sum_{n=-N}^{-1} \overline{\beta_{N-1-n}} \left(\tau L + \overline{\alpha_{N-1-n}}\right)^{-1} \end{aligned}$$

and using the properties (5) and (6), we can write this as

$$\begin{aligned} & \sum_{n=-N}^{-1} \beta_n \left(\tau L + \alpha_n\right)^{-1} + \beta_0 \left(\tau L + \alpha_0\right)^{-1} + \sum_{n=-N}^{-1} \overline{\beta_{N-1-n}} \left(\tau L + \overline{\alpha_{N-1-n}}\right)^{-1} \\ = & \sum_{n=-N}^{-1} \beta_n \left(\tau L + \alpha_n\right)^{-1} + \beta_0 \left(\tau L + \alpha_0\right)^{-1} + \sum_{n=-N}^{-1} \beta_n \left(\tau L + \alpha_n\right)^{-1} \end{aligned}$$

which shows that the imaginary parts cancel out.

Since the imaginary parts cancel out for $Re(a+\overline{a}) = a+\overline{a}$ and with α_0 and β_0 being only real-valued, we can simplify the equation to

$$e^{\tau L} \approx \sum_{n=-N}^{N} (\tau L + \alpha_n)^{-1} = \sum_{n=-N}^{-1} \left(\frac{2\beta_n}{\tau L + \alpha_n} \right) + \frac{\beta_0}{\tau L + \alpha_0}.$$

This allows us to reduce the computational amount almost by a factor of two for solving $(L+\alpha)^{-1}$ giving the real valued solution

$$U(\tau) = e^{\tau L} U(0) \approx \sum_{n=0}^{N} \gamma_n \left(\tau L + \alpha_n\right)^{-1} U(0)$$

with

$$\gamma_n := \begin{cases} \beta_0 & for\, n = 0 \\ 2\beta_n & else \end{cases}$$

3.3 Matrix exponential

We would like to apply REXI to a formulation such as

$$U(t) = e^{\tau L} U(0).$$

To see the relationship between the approximation of e^{ix} with $e^{\tau L}$ we assume that L is skew hermitian and therefore has only purely imaginary eigenvalues, and maybe decomposed as $\Sigma \Lambda \Sigma^H$, yielding

$$e^{\tau L} = \sum_{k=0}^{\infty} \frac{(\tau L)^k}{k!} = \sum_{k=0}^{\infty} \frac{\tau \Sigma \Lambda^k \Sigma^H}{k!} = \Sigma \left(\sum_{k=0}^{\infty} \frac{(\tau \Lambda)^k}{k!} \right) \Sigma^H = \Sigma e^{\tau \Lambda} \Sigma^H, \qquad (7)$$

where we used the orthonormality of Σ to cancel it out from the summation, and

$$e^{\tau \Lambda} = \begin{pmatrix} \cdots & & \\ & e^{i\lambda_j \tau} & \\ & & \cdots \end{pmatrix}$$

where we have explicitly detached the imaginary unit from the eigenvalues, therefore λ_n are assumed real. Since $e^{\tau \Lambda}$ is diagonal, it can be eigenvalue-wise approximated in the same way as in e^{ix} with REXI.

Although L has imaginary eigenvalues, we wish to evaluate the $e^{\tau L} U(0)$, which is real valued. Therefore, we will use the real approximation of e^{ix}

$$exp(\tau L) \approx Re \left(\sum_{n=-N}^{N} \beta_n (\tau L + \alpha_n)^{-1} \right), \qquad (8)$$

where β_n is given by equation (2) and α_n by equation (1).

Given the linear operator L to be applied on $U(0)$, we continue to show the relation of the linear operator to the REXI approximation of the real term

$$e^{\tau L} \approx \sum_{n=-N}^{N} Re\left(\frac{\beta_n^{Re}}{\tau L + \alpha_n}\right)$$

as we use it later on. Next, we rewrite L with our EV decomposition as

$$\begin{aligned}
e^{\tau L} &\approx \sum_{n=-N}^{N} Re\left(\beta_n^{Re}(\tau\Sigma\Lambda\Sigma^{-1} + \alpha_n I)^{-1}\right) \\
&= \sum_{n=-N}^{N} Re\left(\beta_n^{Re}(\Sigma(\tau\Lambda\Sigma^{-1} + \alpha_n\Sigma^{-1}))^{-1}\right) \\
&= \sum_{n=-N}^{N} Re\left(\beta_n^{Re}(\Sigma^{-1}(\tau\Lambda + \alpha_n\Sigma^{-1}\Sigma))^{-1}\Sigma^{-1}\right) \\
&= \sum_{n=-N}^{N} Re\left(\beta_n^{Re}\Sigma(\tau\Lambda + \alpha_n)^{-1}\Sigma^{-1}\right) \\
&= \approx \Sigma e^{\tau\Lambda}\Sigma^{H}
\end{aligned}$$

Note, that we use the same REXI approximation for different λ_j. Hence, the approximation has to be sufficiently accurate over the entire range of all Eigenvaluesλ_j which is what we use the approximation for.

3.4 Choosing h and M

Some important points about the choice of M and h have to be made at this point. We know that e^{ix} is accurately approximated with REXI for the interval $|x| < hM$, where h is chosen small enough to obtain a good approximation in step (A), and M will define the interval size and number of approximation points. In the matrix case, M has to be chosen so that $hM > t\bar{\lambda}$, where $\bar{\lambda} = \max_n |\lambda_n|$, in order to capture all wavelengths of L. In other other words, hM need to be set to capture the fastest wave. Note that if this is used as a time stepping method, with time step $t = \tau$, then, the larger the timestep, the larger M will be. Exact evaluations of the choices for h and M may be done based on equation (3.6) of [2].

3.5 Handling τ in REXI

We reformulate the REXI approximation scheme given by

$$(\tau L + \alpha)U(\tau) = U(0)$$

and by factoring τ out, yielding

$$(L + \frac{\alpha}{\tau})U(\tau) = \tau^{-1}U(0)$$

So instead of solving for $U(\tau)$, we are solving for $U^\tau(0) := U(0)\tau^{-1}$ as well as $\alpha^\tau := \frac{\alpha}{\tau}$.

To summarize, we have to solve the system of equations given by

$$(L + \alpha^\tau)U(\tau) = \tau^{-1}U(0) \tag{9}$$

with $U(0)$ the initial conditions. For sake of simplicity, we stick to the formulation without the τ notation.

3.6 Computing inverse of $(L + \alpha)^{-1}$

For computing the inverse, arbitrary solvers can be used. However we like to note, that α is a complex number. Hence, requiring solvers with support for solving in complex space.

We wish to solve the differential problem for each time step

$$(L + \alpha)U = U(0)$$

so that $U = (L + \alpha)^{-1}U_0$. Note, the change in sign before L for convenience which has to be accounted for afterwards. We will do this converting the problem into and elliptic equation.

First, lets expand the equations with the definition of L,

$$\begin{aligned} fv - g\eta_x + \alpha u &= u_0, & (10)\\ -fu - g\eta_y + \alpha v &= v_0, & (11)\\ -\bar{\eta}(u_x + v_y) + \alpha\eta &= \eta_0. & (12) \end{aligned}$$

Let f be constant (f-plane approximation), $\delta := u_x + v_y$ be the wind divergence, $\zeta := v_x - u_y$ be the wind (relative) vorticity and $\Delta\eta := \eta_{xx} + \eta_{yy}$ the Laplacian of the fluid depth. We will re-write the problem in a divergence-vorticity formulation by taking 2 steps. First, sum the ∂_x of equation (10) and the ∂_y of equation (11), yielding

$$f\zeta - g\Delta\eta + \alpha\delta = \delta_0. \tag{13}$$

Then subtract the ∂_y of equation (10) from the ∂_x of equation (11), yielding

$$-f\delta + \alpha\zeta = \zeta_0. \tag{14}$$

Using equation (13) in equation (14) gives us

$$\delta = \frac{1}{f^2 + \alpha^2}\left(\alpha g\Delta\eta + \alpha\delta_0 - f\zeta_0\right).$$

Finally, substituting δ in equation (12), that reads $\bar{\eta}\delta + \alpha\eta = \eta_0$, results in

$$-\frac{\bar{\eta}}{f^2+\alpha^2}\left(\alpha g \Delta\eta + \alpha\delta_0 - f\zeta_0\right) + \alpha\eta = \eta_0,$$

which can be written as an eliptic equation with

$$(\kappa - \overline{\eta} g\Delta)\eta = r_0 \tag{15}$$

where

$$\kappa = f^2 + \alpha^2$$

and

$$r_0 = \frac{\kappa}{\alpha}\eta_0 - \overline{\eta}\frac{f}{\alpha}\zeta_0 + \overline{\eta}\delta_0.$$

We can computeη with a solver for elliptic problems. Special attention has to be given to the LHS of the form $(\gamma - \Delta)\eta$ for a spectral solver and this is the reason for this brief excursion. We use $\tilde{}$ to annotate an operator or quantity to be given in spectral space. Then, we factor $\tilde{\eta}$ in and write form of the LHS in spectral space with the identity matrix I as

$$(\gamma I\tilde{\eta} - \tilde{\Delta}\tilde{\eta}) = (\gamma I - \tilde{\Delta})\tilde{\eta}$$

Here, the minus operator in spectral space has to be applied on all elements of $\tilde{\Delta}$ and not only to the 0^{th} mode as it is typically the case for adding a constant in spectral space.

We need to retrieve the velocities by solving the 2×2 system formed by equations (10) and (11), which reads

$$A_\alpha U = U_0 + g\nabla\eta$$

with

$$A_\alpha = \begin{pmatrix} \alpha & f \\ -f & \alpha \end{pmatrix}.$$

The solution is

$$U = A_\alpha^{-1}(U_0 + g\nabla\eta)$$

where

$$A_\alpha^{-1} = \frac{1}{f^2+\alpha^2}\begin{pmatrix} \alpha & -f \\ f & \alpha \end{pmatrix}$$

3.7 Interpretation of τ

We like to close this section with a brief discussion of τ by having a look on the REXI reformulation

$$\left(L + \frac{\alpha}{\tau}\right)^{-1} U\tau^{-1} = U_0$$

We see, that for an increasing τ, hence an integration in time over a larger time period, the poles given by α are getting closer. This can possibly lead to a loss

in accuracy for the data sampled by the outer poles α_{-N} and α_N. Therefore, the number N of poles is expected to scale linearly with the size of the coarse time step,

$$|N| \propto \tau.$$

Indeed, we saw in section 3.3 that for larger τ, M needs to be larger.

3.8 Errors and fault tolerance

TODO: CHECK THIS, IT MIGHT BE WRONG

One question arising is what the error is in case that one term of the REXI approximation is not computed. We assume that this will depend on α_i and continue investigating this further

Directly analysing the errors with the formulation $(L-\alpha)^{-1} U = U_0$ seems to be relatively challenging and we therefore decide to start direclty with the Helmholtz formulation

$$((\alpha^2 + f^2) - g\bar{\eta}\Delta)\eta = \frac{f^2+\alpha^2}{\alpha}\eta_0 - \bar{\eta}\delta_0 - \frac{f\bar{\eta}}{\alpha}\zeta_0$$

and use the spectral space with selected frequencies $\eta = exp(ikx)$ for all quantities.

Using the spectral representation leads to

$$((\alpha^2 + f^2) + g\bar{\eta}k^2)\hat{\eta} = \frac{f^2+\alpha^2}{\alpha}\hat{\eta}_0 - \eta\bar{i}k(\hat{u}+\hat{v}) - \frac{f\bar{\eta}}{\alpha}ik(\hat{v}-\hat{u})$$

Solving this for $\hat{\eta}$ yields

$$\hat{\eta} = \frac{\frac{f^2+\alpha^2}{\alpha}\hat{\eta}_0 - \eta\bar{i}k(\hat{u}+\hat{v}) - \frac{f\bar{\eta}}{\alpha}ik(\hat{v}-\hat{u})}{((\alpha^2 + f^2) + g\bar{\eta}k^2)}$$

Considering the tabulated alpha values, we observe a constant real alpha value and increasing imaginary parts for larger N. Hence, also the denominator increases and results in relatively small solutions $\hat{\eta}$. The same accounts for the explicit formulation to compute the velocities.

We tested this imperically with the code in helmholtz_problem_fault_tolerance.py.

Additionally, we can also observe significantly decreasing β_i for increasing i.

To conclude, the amplitude (the real part) decreases significantly with increasing α_i values. **TODO: CHECK THIS, IT MIGHT BE WRONG**

3.9 Numerical Dispersion

We continue to (try to) analyze possible dispersion effects of the REXI approximation. Here, we assume that we use an accurate solver to compute $(L-\alpha_i)^{-1}$. We are interested in answering the question which error is introduced when not using enough terms in the approximation.

According to 3.3, we can decompose the linear operator into EVals and EVects and write the approximation in the following way:

$$e^{\tau L} \approx \Sigma \left[\sum_{n=0}^{N} Re\left(\gamma_n (\tau\Lambda + \alpha_n)^{-1}\right) \right] \Sigma^H.$$

Alternatively without solving a linear operator, one can also think directly about approximating a set of exponents given by Λ with

$$e^{\tau\Lambda} \approx \sum_{n=0}^{N} \left(\gamma_n (\tau\Lambda + \alpha_n)^{-1}\right).$$

We can then use the spectral representation of the solution

$$e^{\tau \hat{L}} = \hat{\Sigma} e^{\tau \hat{\Lambda}} \hat{\Sigma}^H$$

with $\hat{L} = \hat{\Sigma}\hat{\Lambda}$the spectral EValue/EVector decomposition. We can then approximate the exponential $e^{\tau\Lambda}$ by

$$e^{\tau \hat{L}} \approx \hat{\Sigma} \sum_{n=0}^{N} \left(\gamma_n (\tau\hat{\Lambda} + \alpha_n)^{-1}\right) \hat{\Sigma}^H$$

For selected modes (k_1, k_2) at a specific point in time $\tau = \sigma$, the solution and operators in spectral space are then given by

$$\begin{aligned}
\hat{U}(k_1, k_2, \tau) =& e^{\tau \hat{L}_{k_1,k_2}} \hat{U}(k_1, k_2, 0) \\
\approx& \hat{\Sigma}_{k_1,k_2} \left[\sum_{n=-N}^{N} Re\left(\beta_n^{Re} (\tau\hat{\Lambda}_{k_1,k_2} + \alpha_n)^{-1}\right) \right] \hat{\Sigma}^H_{k_1,k_2} \hat{U}(k_1, k_2, 0).
\end{aligned}$$

Case A) For $k_1 \neq 0$ and $k_2 \neq 0$, we know that $\hat{\Lambda}$contains eigenvalues

Vortical mode:	$\omega_0 = 0$
Gravitational modes:	$\omega_{\pm 1} = \pm\sqrt{4\pi^2(\eta_0 g k_1^2 + \eta_0 g k_2^2) + f^2}$

Case B) For $k_1 = k_2 = 0$, we get the eigenvalues

Vortical mode:	$\omega_0 = 0$
Gravitational modes:	$\omega_{\pm 1} = \pm f$

These eigenvalues describe the frequency in $\hat{\Lambda}$ which we approximate. As soon as this frequency cannot be approximated anymore with REXI, this results in errors in the dispersion of this particular frequency.

With the computational requirements (M) of REXI being related to the fastest waves, we can identify the requirement $M \sim \sqrt{\eta_0 g}$ and $M \sim f$.

Vortical modes: Since the vortical modes are always zero, we do not expect any (analytical) errors in these modes. However, due to discretization, errors can accumulate and show up.

Gravitational modes: We can observe that errors in the frequency $\omega_{\pm 1}$ in case A area generated. The dominating frequency is only dependent on η_0 and g in case of

$$4\pi^2(\eta_0 g k_1^2 + \eta_0 g k_2^2) = 4\pi^2 \eta_0 g |k|^2 > f^2$$

and f otherwise. The factor $4\pi^2$ shows up since our domain is on $\Omega = [0;1]^2$. The Coriolis frequency f is also independent of the spatial frequency, hence also the resolution.

Furthermore, we can observe a relation to the Rossby radius given by

$$L_R := 4\pi^2 \frac{\sqrt{g\eta_0}}{f}$$

which describes at which scale at the Coriolis effect also strongly contributes to the simulation results compared to the gravitational and height values.

3.10 Complex solver

We are interested in solving a system of the form,

$$(-L + \alpha)U = U_0,$$

where α is a complex number, therefore we must allow complex solutions for U. The system can be transformed to have only real arithmetic in the following way.

First we decompose U into its real and imaginary parts, $U = U^r + iU^i$, and allow U_0 to be decomposed in the same way. Although α can be a general complex number, it real part is always constant in REXI, given by μ. We will therefore write $\alpha = \mu + i\alpha^i$, and we can absorb μ into L writing $D = -L + \mu I$, where I is the identity matrix. Now

$$(D + i\alpha^i I)(U^r + iU^i) = U_0^r + iU_0^i,$$
$$DU^r - \alpha^i U^i + i(\alpha^i U^r + DU^i) = U_0^r + iU_0^i,$$

therefore

$$DU^r - \alpha^i U^i = U_0^r \quad (16)$$
$$\alpha^i U^r + DU^i = U_0^i, \quad (17)$$

which in matrix notation gives

$$\begin{pmatrix} D & -\alpha^i I \\ \alpha^i I & D \end{pmatrix} \begin{pmatrix} U^r \\ U^i \end{pmatrix} = \begin{pmatrix} U_0^r \\ U_0^i \end{pmatrix},$$

or,

$$\begin{pmatrix} -L + \mu I & -\alpha^i I \\ \alpha^i I & -L + \mu I \end{pmatrix} \begin{pmatrix} U^r \\ U^i \end{pmatrix} = \begin{pmatrix} U_0^r \\ U_0^i \end{pmatrix}.$$

A similar approach may be taken with the elliptic equation

$$\Delta\eta + \theta\eta = r_0, \quad (18)$$

where η, θ and r_0 may be complex. The resulting system is given by

$$\Delta\eta^r + \theta^r\eta^r - \theta^i\eta^i = r_0^r, \quad (19)$$
$$\Delta\eta^i + \theta^i\eta^r + \theta^r\eta^i = r_0^i, \quad (20)$$

which can also be written in matrix notation and solved with arbitrary elliptic equation solvers.

4 Compatibility table (merging + halving)

Different reformulations of the REXI coefficients for better approximation and reduced workload were suggested in the previous sections. Those were (a) the merging of real and imaginary approximation of $\exp(ix)$ (see Sec. 2.4) as well as halving the number of poles by exploiting symmetries and real-only valued linear operator (see Sec. 3.2).

The test cases are described as follows:

- Approximation of real parts of complex-valued ODE:

$$Re(\exp(ix))u_0 = Re\left(\sum_n \beta_n \left(ix+\alpha_n\right)^{-1}\right) u_0$$

- Approximation of real and imaginary parts of complex-valued ODE

$$\exp(ix)u_0 = \sum_n \beta_n \left(ix+\alpha_n\right)^{-1} u_0$$

- Approximation of (real) linear operator

$$L = \begin{bmatrix} 0 & x \\ -x & 0 \end{bmatrix}$$

$$\exp(L)U_0 = \sum_n \beta_n \left(L+\alpha_n\right)^{-1} U_0$$

  where

$$U_0 = [Re(u_0), Im(u_0)].$$

We test with two different initial conditions:

- $u_0 = 1.0$:

Merging	-	merge	-	merge
Halve	-	-	halve	halve
Resulting REXI setup	β^{Re} only	**merge** β^{Re} and β^{Im}	**halve**	**merge+halve**
$Re\left(\exp(ix)\right)$	OK	OK	fail	fail
$\exp(ix)$	fail	OK	fail	fail
$Re(\exp(L))$	fail	OK	fail	OK

  - Neither merging, nor halving: Only the real values of the ODE are solved properly.
  - Merging only: Results in a successful execution for all test cases.
  - Only halving: Full failure, in particular because the real-valued property of the linear operator is failing
  - merge + halve: Only works for linear operator L which is the only real-valued one.

- $u0 = (1 + 2j)/\sqrt{5}$:

Merging	**-**	**merge**	**-**	**merge**
Halve	**-**	**-**	**halve**	**halve**
Resulting REXI setup	β^{Re} only	**merge** β^{Re} and β^{Im}	**halve**	**merge+halve**
$Re\,(\exp(ix))$	**fail**	OK	fail	fail
$\exp(ix)$	fail	OK	fail	fail
$Re(\exp(L))$	fail	OK	fail	OK

  - Neither merging, nor halving: Even the real approximation now failes. The reason for this is that a correct approximation of the imaginary value of $\exp(ix)$ is required.
  - All other benchmarks behave as expected.

5 Filtering

To ensure that the REXI is bounded by unit, a filtering process is proposed in [2]. REXI is prone to exceed unit in the neighborhood of $|t\lambda| \approx hM$, therefore in the highest frequencies. The idea is to construct a rational function $S(ix)$ that is approximately 1 in a smaller interval $|t\lambda| < hM_0$, with $M_0 < M$, and decays very fast to zero outside this interval. This filter can be applied as a separate "time step" to get rid of such fast modes.

6 Bringing everything together

Using the spectral methods (e.g. in SWEET), we can directly solve the Helmholtz problem for the height and then solver for the velocity. Note that the Helmholtz problem is in complex space, as α_n are complex. This is straightforward with spectral methods. For finite difference/element methods, the problem can be split into its real and imaginary parts.

Then, the problem is reduced to computing the REXI as given in Eq. (8).

7 Notes on HPC

- The terms in REXI to solve are all independent. Hence, for latency avoiding, the communication can be interleaved with computations.
- The iterative solvers are memory bound. Instead of computing $c := a * b$ for the stencil operations, we could compute $\vec{c} := a\vec{b}$ with a one coefficient in the stencil. This allows vectorization over c and b on accelerator cards with strided memory access.
- It is unknown which method is more efficient to solve the system of equations:
  - iterative solvers have low memory access,

  - inverting the system and storing it as a sparse matrix allows fast direct solving but can yield more memory access operations.

- Splitting the solver into real and complex number would store them consecutively in memory. This has a potential to avoid non-strided memory access and using the same SIMD operations (Just a rough idea, TODO: check if this is really the case).

8 Acknowledgements

Thanks to Terry for the feedback & discussions!

References

[1] Formulations of the shallow-water equations, M. Schreiber, P. Peixoto et al.

[2] High-order time-parallel approximation of evolution operators, T. Haut et al.

[3] Nineteen Dubious Ways to Compute the Exponential of a Matrix, Twenty-Five Years Later, Cleve Moler and Charles Van Loan, SIAM review

[4] Near optimal rational approximations of large data sets, Damle, A., Beylkin, G., Haut, T. S. & Monzon