

# Asymptotic Error Analysis with Noisy Function Evaluations

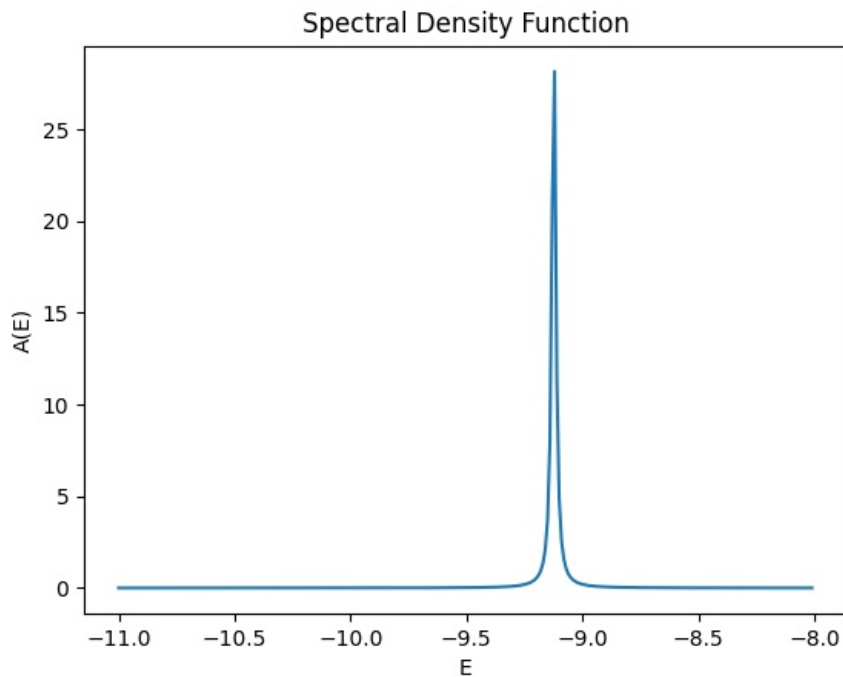
James Wu 92277235

## 1 Background

In one of my condensed matter physics research projects, I am looking to fit a vector of parameters  $\mathbf{v}$  to describe the interactions of a quantum mechanical system (the system is parametrized by  $\mathbf{v}$ ). There is another vector of real numbers  $\mathbf{k}$  (the system's momentum) that also parametrizes the system. Now, the system's ground state energy  $E_0(\mathbf{v}, \mathbf{k})$  is to be matched with target energies  $E_T(\mathbf{k})$ . This is done by interpolating for a finite number of  $\mathbf{k}$  points (in practice this seems to match  $E_T(\mathbf{k})$  for all  $\mathbf{k}$  quite well). One way to find an appropriate  $\mathbf{v}$  is to use gradient descent on the cost function

$$J(\mathbf{v}) = \|\mathbf{E}_0(\mathbf{v}) - \mathbf{E}_T\|_2$$

Here  $\mathbf{E}_0$  and  $\mathbf{E}_T$  are vectors with entries that are energy values at various  $\mathbf{k}$  values. In short, we use a least squares fit.  $E_0(\mathbf{v}, \mathbf{k})$  can be computed as the energy where the *spectral density function*  $A(E)$  spikes. For example, in the following figure, we see that  $E \approx -9.1$ :



Numerically, we restrict ourselves to an interval  $I$  with  $N$  evenly spaced grid points to evaluate  $A(E)$ . We thus have

$$E_0(\mathbf{v}, \mathbf{k}) = \arg \max_{E \in I} A(E, \mathbf{v}, \mathbf{k})$$

There is another parameter  $\eta$ , which controls the spikes width, that is set equal to the grid spacing  $dE$  to ensure that the spike will be detected. Locally, this spike is symmetric (a Lorentzian, in fact) in the limit  $\eta \rightarrow 0^+$ . So the error in this method is roughly bounded by  $dE/2$ . This error propagates to the  $J(\mathbf{v})$  function to also produce an error bound of  $dE/2$  if we use a dimensionality-agnostic 2-norm (i.e. divide by the square root of the dimension of the energy vectors).

In the following analyses, we consider single-variable functions for simplicity. Let  $f(x)$  be such a function that we are interested in and  $g(x) = f(x) + \delta(x)$  be a noisy function with  $|\delta(x)| \leq \Delta$  for some positive error bound  $\Delta$  (e.g.  $\Delta = dE/2$ ).

To motivate this project, I seek to answer the question of how  $\Delta$  should be scaled relative to the other numerical parameters to obtain the best accuracy for the least computation time. I will therefore be determining the asymptotic errors for various numerical methods for a given number of computations.

## 2 Derivatives

### 2.1 First Derivative: Central Difference

Suppose we want to compute the derivative of  $f(x)$  but only have access to evaluating the noisy function  $g(x)$ . A central difference yields:

$$f'(a) \approx \frac{g(a+k) - g(a-k)}{2k}$$

The error in this method is

$$\begin{aligned} f'(a) - \frac{g(a+k) - g(a-k)}{2k} &= f'(a) - \frac{f(a+k) - f(a-k)}{2k} - \frac{\delta(a+k) - \delta(a-k)}{2k} \\ &= -\frac{1}{12}f'''(\xi)k^2 + \frac{1}{12}f'''(\zeta)k^2 - \frac{\delta(a+k) - \delta(a-k)}{2k} \end{aligned}$$

for some  $\xi \in (a, a+k)$  and  $\zeta \in (a-k, a)$ . A bound for this error  $E(k, \Delta)$  is then

$$E(k, \Delta) = \frac{1}{6}K_3k^2 + \frac{\Delta}{k}$$

where  $K_3$  is a bound on the absolute value of the third derivative of  $f(x)$ . Note that Taylor analysis is not applied to  $\delta(x)$  because it is not continuous— $\delta(x)$  is discretized, for example, in the spectral density function case.

Now, the computation time  $C(k, \Delta)$  is independent of  $k$  because only a single finite difference is computed.  $C(k, \Delta)$  is inversely proportional to  $\Delta$  in the spectral density function scenario,

however, because the number of  $A(E)$  evaluations for a given interval size  $|I|$  is approximately  $|I|/dE$ :

$$C(k, \Delta) \sim \Delta^{-1}$$

The optimal  $k$  that produces a minimal error bound given a fixed  $C(k, \Delta)$  and therefore fixed  $\Delta$  is given by

$$\frac{dE}{dk} = \frac{\partial E}{\partial k} = \frac{1}{3}K_3k - \frac{\Delta}{k^2} = 0$$

This is indeed a minimum by the second derivative test:

$$\frac{d^2E}{dk^2} = \frac{\partial^2 E}{\partial k^2} = \frac{1}{3}K_3 + \frac{2\Delta}{k^3} > 0$$

So the optimal  $k$ - $\Delta$  relationship is

$$\Delta = \frac{1}{3}K_3k^3$$

The cubic relationship makes sense because the  $\frac{1}{6}K_3k^2$  and  $\frac{\Delta}{k}$  terms in  $E(k, \Delta)$  should scale the same—if they did not, the term that vanishes the slowest would become a bottleneck and could be made smaller for faster error convergence. The error bound is then

$$E(k, \Delta) = \frac{1}{2}K_3k^2 = \frac{3^{2/3}}{2}K_3^{2/3}\Delta^{2/3}$$

The error bound therefore scales with the computation time as:

$$E \sim C^{-2/3}$$

## 2.2 Second Derivative

We now consider the task of estimating  $f''(a)$ :

$$f''(a) \approx \frac{g(a+k) - 2g(a) + g(a-k)}{k^2}$$

The error in this method is

$$\begin{aligned} f''(a) - \frac{g(a+k) - 2g(a) + g(a-k)}{k^2} &= f''(a) - \frac{f(a+k) - 2f(a) + f(a-k)}{k^2} \\ &\quad - \frac{\delta(a+k) - 2\delta(a) + \delta(a-k)}{k^2} \\ &= -\frac{1}{24}f'''(\xi)k^2 + \frac{1}{24}f'''(\zeta)k^2 - \frac{\delta(a+k) - 2\delta(a) + \delta(a-k)}{k^2} \end{aligned}$$

for some  $\xi \in (a, a+k)$  and  $\zeta \in (a-k, a)$ . A bound for this error  $E(k, \Delta)$  is then

$$E(k, \Delta) = \frac{1}{12}K_4k^2 + \frac{4\Delta}{k^2}$$

where  $K_4$  is a bound on the absolute value of the fourth derivative of  $f(x)$ . Again, the computation time is scaled as such:

$$C(k, \Delta) \sim \Delta^{-1}$$

The optimal  $k$ - $\Delta$  relationship is given by:

$$\frac{dE}{dk} = \frac{\partial E}{\partial k} = \frac{1}{6}K_4k - \frac{8\Delta}{k^3} = 0$$

This is indeed a minimum by the second derivative test:

$$\frac{d^2E}{dk^2} = \frac{\partial^2 E}{\partial k^2} = \frac{1}{6}K_4 + \frac{24\Delta}{k^4} > 0$$

The ideal  $k$ - $\Delta$  relationship is thus

$$\Delta = \frac{1}{48}K_4k^4$$

Similar to the central difference case, the  $\frac{1}{12}K_4k^2$  and  $\frac{4\Delta}{k^2}$  terms scale the same under this relationship. The error bound is then

$$E(k, \Delta) = \frac{1}{6}K_4k^2 = \frac{2}{\sqrt{3}}K_4^{1/2}\Delta^{1/2}$$

The error bound therefore scales with the computation time as:

$$E \sim C^{-1/2}$$

### 3 Finit Difference Methods

#### 3.1 1-Periodic $-u'' + u = f$

Consider the following 1-periodic ODE:

$$-u''(x) + u(x) = f(x)$$

We may solve this problem using finite differences by discretizing the interval  $[0, 1)$  into  $N$  uniformly-spaced grid points, with  $k = 1/N$  being the grid spacing. Now let  $\mathbf{u}$  be  $u(x)$  evaluated at these points and  $\mathbf{U}$  be our numerical estimation of  $\mathbf{u}$ . Then

$$\begin{aligned} A\mathbf{U} &= \mathbf{G} = \mathbf{F} + \delta \\ A\mathbf{u} &= \mathbf{F} - \tau \end{aligned}$$

Here  $A = -D_2 + I$  (and  $A$  is nonsingular),  $\mathbf{F}$  is  $f(x)$  evaluated at the grid points,  $\mathbf{G}$  is our noisy evaluation of  $\mathbf{F}$  with error  $\delta$ , where  $\|\delta\|_2 \leq \Delta$ , and  $\tau$  is the truncation error in the (non-noisy) finite difference approximation. This numerical recipe takes the following steps to complete:

1.  $\mathbf{G}$  needs to be constructed. There are  $N = k^{-1}$  entries, each of which require  $O(\Delta^{-1})$  computations. This step has a runtime of  $O(k^{-1}\Delta^{-1})$ .

2.  $A$  needs to be constructed. Since  $A$  is sparse, this step takes  $O(N) = O(k^{-1})$  computations.
3. The system  $A\mathbf{U} = \mathbf{G}$  needs to be solved. Since  $A$  is sparse, this step also has a runtime of  $O(k^{-1})$ .

The bottleneck occurs in the first step, giving a total runtime of

$$\boxed{C \sim k^{-1} \Delta^{-1}}$$

Meanwhile, the error  $\mathbf{E} = \mathbf{u} - \mathbf{U}$  is given by

$$\begin{aligned} A\mathbf{E} &= A\mathbf{u} - A\mathbf{U} = -\tau - \delta \\ \mathbf{E} &= -A^{-1}\tau - A^{-1}\delta \end{aligned}$$

We know that  $\|A^{-1}\|_2 = 1$  from von Neumann analysis. The truncation error  $\tau = \mathbf{F} - A\mathbf{u}$  is bounded by the non-noisy finite difference error bound:

$$\|\tau\|_2 \leq \frac{1}{12} K_4 k^2$$

This bounds the error as:

$$\|\mathbf{E}\|_2 \leq \|\tau\|_2 + \|\delta\|_2 = \frac{1}{12} K_4 k^2 + \Delta$$

For a fixed  $C(k, \Delta) = k^{-1} \Delta^{-1}$ , we can treat  $\Delta$  as a function of  $k$ :  $\Delta = C^{-1} k^{-1}$ . Then optimizing the error bound gets us:

$$\begin{aligned} E(k, \Delta) &= \frac{1}{12} K_4 k^2 + \Delta \\ \frac{dE}{dk} &= \frac{\partial E}{\partial k} + \frac{\partial E}{\partial \Delta} \frac{d\Delta}{dk} = \frac{1}{6} K_4 k - C^{-1} k^{-2} = \frac{1}{6} K_4 k - k^{-1} \Delta = 0 \end{aligned}$$

$$\boxed{\Delta = \frac{1}{6} K_4 k^2}$$

Again, both error terms scale the same in this relationship. Finally, we verify that this is indeed a minimum:

$$\begin{aligned} \frac{d^2 E}{dk^2} &= \frac{\partial^2 E}{\partial k^2} + \frac{\partial^2 E}{\partial \Delta^2} \left( \frac{d\Delta}{dk} \right)^2 + \frac{\partial^2 E}{\partial \Delta \partial k} \frac{d\Delta}{dk} + \frac{\partial E}{\partial \Delta} \frac{d^2 \Delta}{dk^2} \\ &= \frac{\partial^2 E}{\partial k^2} + \frac{\partial E}{\partial \Delta} \frac{d^2 \Delta}{dk^2} = \frac{1}{6} K_4 + 2C^{-1} k^{-2} > 0 \end{aligned}$$

The error bound is now

$$\boxed{E(k, \Delta) = \frac{1}{4} K_4 k^2 = \frac{3}{2} \Delta}$$

The runtime under this  $k$ - $\Delta$  relationship scales as  $C \sim k^{-3} \sim \Delta^{-3/2}$ . The error therefore scales with runtime as

$$\boxed{E \sim C^{-2/3}}$$

## 4 Time Stepping

### 4.1 Forward Euler

#### 4.1.1 Error Analysis

Consider the initial value problem:

$$u'(t) = f(u, t), \quad u(0) = u_0, \quad t \in [0, \infty)$$

We can approximate  $u(t)$  numerically up to a cutoff time  $T$  using Euler's method. This is done by estimating  $u(t)$  at evenly-spaced time intervals of  $k$ . Let  $U_n$  denote the estimate of  $u(nk)$ . Then  $U^0 = u_0$  and

$$U_{n+1} = U_n + kg(U_n, nk)$$

where  $g(u, t) = f(u, t) + \delta(u, t)$  and  $|\delta| \leq \Delta$ . We find the local error by assuming  $U_n = u(nk)$ :

$$\begin{aligned} u((n+1)k) - U_{n+1} &= u((n+1)k) - u(nk) - ku'(nk) - k\delta(u(nk), nk) \\ &= -\frac{k^2}{2}f''(\xi) - k\delta(u(nk), nk) \end{aligned}$$

We can then bound the local error by:

$$|u((n+1)k) - U_{n+1}| \leq L(k, \Delta) = \frac{1}{2}K_2k^2 + k\Delta$$

where  $K_2$  is a bound on  $|u''(t)|$ . As a rough estimate, the global error scales as  $k^{-1}L(k, \Delta)$ . We shall therefore minimize the following global error estimate for a fixed number of computations:

$$E(k, \Delta) = \frac{1}{2}K_2k + \Delta$$

Regarding the computational complexity of this method, there are  $T/k$  time steps, and each time step requires  $O(\Delta^{-1})$  evaluations of  $g(u, t)$ . The computation time thus scales as

$$\boxed{C \sim k^{-1}\Delta^{-1}}$$

Next we optimize

$$\begin{aligned} E &= \frac{1}{2}K_2k + C^{-1}k^{-1} \\ \frac{dE}{dk} &= \frac{1}{2}K_2 - C^{-1}k^{-2} = \frac{1}{2}K_2 - k^{-1}\Delta = 0 \end{aligned}$$

We thus obtain the relationship

$$\boxed{\Delta = \frac{1}{2}K_2k}$$

We also verify the second derivative test:

$$\frac{d^2E}{dk^2} = 2C^{-1}k^{-3} > 0$$

The global error thus scales as:

$$\boxed{E \sim K_2 k \sim \Delta}$$

In terms of time complexity, we have  $C \sim k^{-2} \sim \Delta^{-2}$ . That means

$$\boxed{E \sim C^{-1/2}}$$

### 4.1.2 Stability

To investigate the stability of Forward Euler for a noisy function, we consider the test problem

$$u' = \lambda u \quad \therefore \quad u(t) = e^{\lambda t}$$

Then

$$U_{n+1} = U_n + kf(U_n, nk) + k\delta(U_n, nk) = \left(1 + k\lambda + \frac{k\delta}{U_n}\right) U_n$$

We attain stability when

$$\left|1 + k\lambda + \frac{k\delta}{U_n}\right| \leq 1$$

By the triangle inequality, this condition is satisfied when

$$\boxed{|1 + k\lambda| + \frac{k\Delta}{|U_n|} \leq 1}$$

In the case of minimal error, we have  $\Delta = \frac{1}{2}K_2k$ , so the stability condition becomes

$$|1 + k\lambda| + \frac{K_2k^2}{2|U_n|} \leq 1$$

As  $k$  vanishes (for a fixed  $T$  and therefore maximal  $|U_n|$ ), the  $\frac{K_2k^2}{2|U_n|}$  term becomes negligible compared to the  $k\lambda$  term. So the region of stability in the  $z = k\lambda$  plane remains roughly the same in the limit of small  $k$ .

## 4.2 Improved Euler

### 4.2.1 Error Analysis

The Improved Euler uses the following recursion:

$$U_{n+1} = U_n + \frac{k}{2} (g(U_n, nk) + g(U_n + kg(U_n, nk), (n+1)k))$$

Again, the computational complexity scales as

$$\boxed{C \sim k^{-1} \Delta^{-1}}$$

although with thrice as many evaluations of  $g(u, t)$  at each time step as the Forward Euler method. Elsewhere, the local error is

$$\begin{aligned}
u((n+1)k) - U_{n+1} &= u((n+1)k) - u(nk) \\
&\quad - \frac{k}{2} (f(u(nk), nk) + f(u(nk) + kg(u(nk), nk), (n+1)k)) \\
&\quad - \frac{k}{2} (\delta(u(nk), nk) + \delta(u(nk) + kg(u(nk), nk), (n+1)k)) \\
&= u((n+1)k) - u(nk) \\
&\quad - \frac{k}{2} (f(u(nk), nk) + f(u(nk) + kf(u(nk), nk), (n+1)k)) \\
&\quad - \frac{k}{2} (\delta(u(nk), nk) (1 + kf_u + O(k^2)) + \delta(u(nk) + kg(u(nk), nk), (n+1)k)) \\
&= u((n+1)k) - u(nk) \\
&\quad - \frac{k}{2} \left( 2f + k(f_u f + f_t) + \frac{k^2}{2}(f_{uu}f^2 + f_{tt} + 2f_{ut}f) + O(k^3) \right) \\
&\quad - \frac{k}{2} (\delta(u(nk), nk) (1 + kf_u + O(k^2)) + \delta(u(nk) + kg(u(nk), nk), (n+1)k)) \\
&= \frac{1}{6} \ddot{f} k^3 - \frac{1}{4} (f_{uu}f^2 + f_{tt} + 2f_{ut}f) k^3 \\
&\quad - \frac{k}{2} (\delta(u(nk), nk) (1 + kf_u + O(k^2)) + \delta(u(nk) + kg(u(nk), nk), (n+1)k))
\end{aligned}$$

Now let  $B$  be an upper bound on the second derivatives of  $f$  as such:

$$\left| \frac{1}{6} \ddot{f} - \frac{1}{4} (f_{uu}f^2 + f_{tt} + 2f_{ut}f) \right| \leq B$$

Then the local error is bounded by

$$|u((n+1)k) - U_{n+1}| \leq Bk^3 + k\Delta$$

Factoring out  $k^{-1}$  gives us an estimate for global error:

$$E(k, \Delta) = Bk^2 + \Delta$$

Differentiating,

$$\begin{aligned}
E &= Bk^2 + C^{-1}k^{-1} \\
\frac{dE}{dk} &= 2Bk - C^{-1}k^{-2} = 0 \\
\frac{d^2E}{dk^2} &= 2B + 2C^{-1}k^{-3} > 0
\end{aligned}$$

The first derivative test yields

$$\boxed{\Delta \sim k^2}$$

This gives an error that scales as

$$\boxed{E \sim k^2 \sim \Delta \sim C^{-2/3}}$$



### 4.2.2 Stability

Again considering the test problem  $u' = \lambda u$ , we have

$$\begin{aligned} U_* &= U_n + kg(U_n, nk) = \left(1 + k\lambda + \frac{k\delta(U_n, nk)}{U_n}\right) U_n \\ U_{n+1} &= U_n + \frac{k}{2} (g(U_n, nk) + g(U_*, (n+1)k)) \\ &= U_n + \frac{k\lambda}{2} \left(2 + k\lambda + \frac{k\delta(U_n, nk)}{U_n}\right) U_n + \frac{k}{2} (\delta(U_n, nk) + \delta(U_*, (n+1)k)) \end{aligned}$$

So the method is stable when

$$\left|1 + k\lambda + \frac{k^2\lambda^2}{2}\right| + \frac{k\Delta}{2|U_n|} (1 + |1 + k\lambda|) \leq 1$$

If we scale  $\Delta \sim k^2$ , we again see that this approaches the non-noisy stability condition  $\left|1 + k\lambda + \frac{k^2\lambda^2}{2}\right| \leq 1$  when we neglect terms of order  $k^3$  or higher.