TODO TITLE

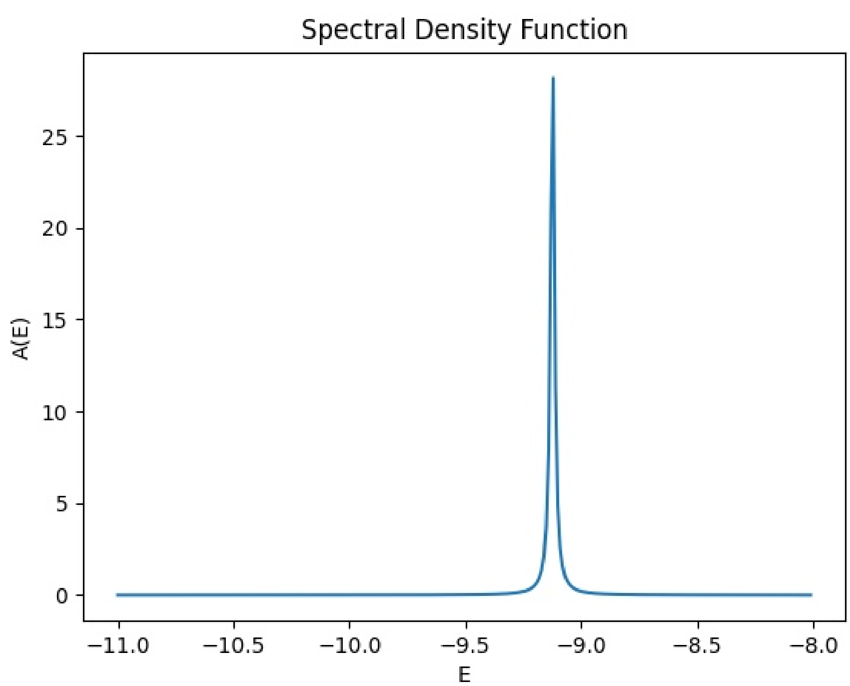
James Wu 92277235

1 Background

In one of my condensed matter physics research projects, I am looking to fit a vector of paramters $\mathbf{v}$ 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 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 η, 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 $\eta \to 0^+$. 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 Δ (e.g. $\Delta = dE/2$).

To motivate this project, I seek to answer the question of how Δ 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 Finite Difference

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

$$E(k, \Delta) = \frac{1}{6} K_3 k^2 + \frac{\Delta}{k}$$

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

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

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

$$\frac{dE}{dk} = \frac{\partial E}{\partial k} = \frac{1}{3}K_3 k - \frac{\Delta}{k^2} = 0$$

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-Δ relationship is

$$\boxed{\Delta = \frac{1}{3}K_3 k^3}$$

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

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

The error bound therefore scales with the comptuation time as:

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

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:

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

The optimal k-Δ relationship is given by:

$$\frac{dE}{dk} = \frac{\partial E}{\partial k} = \frac{1}{6}K_4 k - \frac{8\Delta}{k^3} = 0$$

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-Δ relationship is thus

$$\boxed{\Delta = \frac{1}{48}K_4 k^4}$$

Similar to the central difference case, the $\frac{1}{12}K_4 k^2$ and $\frac{4\Delta}{k^2}$ terms scale the same under this relationship. The error bound is then

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

The error bound therefore scales with the comptuation time as:

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