Problem Formulation

1 Problem Formulation

1.1 Overview of the Interferometric Signal

The primary objective of this analysis is to extract the phase shift ϕ from a set of timeseries. The measurement relies on the ratio of intensities between potentially a reference pulse and a signal pulse, which are temporally separated by a known edge time, t_{edge}.

The raw signal $S(t)$ is divided into two distinct regions of interest:

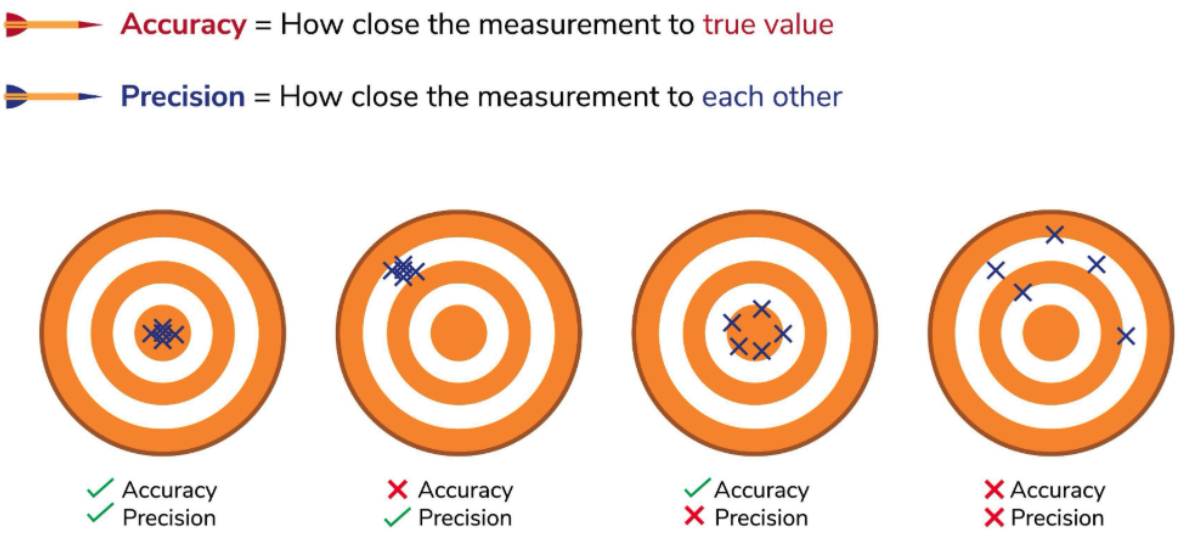
- **Region 1 (Pre-edge, $t < t_{edge}$):** Contains pulse 1 (reference).
- **Region 2 (Post-edge, $t > t_{edge}$):** Contains pulse 2 (signal).

The phase is encoded in the amplitude ratio R:

$$R = \frac{A_{pre}}{A_{post}}$$

Some notes on the unique challenge of detecting R

- **Robust Estimation:** The challenge is not one of peak detection here, but of parameter estimation and inference. Because the pulses do not necessarily peak within the recorded time window, the solution requires robust parameter estimation (non-linear least squares) capable of extrapolation, where the covariance of the fit serves as the primary metric for reliability.
- **Model-Dependent Solution:** The estimation of the peak amplitude ratio is fundamentally ill-conditioned and dependent on the structural prior applied mainly to Region 2. As later described, the pulse in Region 2 is assumed as a sum of three partially captured Gaussian pulses. My analysis showed that any simplification of this model introduces systematic bias (specification error) that propagates directly to the final phase shift estimation.
- **Precision Characterization:** In the absence of an external reference signal or ground truth, the reported uncertainty quantifies the statistical precision of the model fit (Type A uncertainty) rather than absolute accuracy.

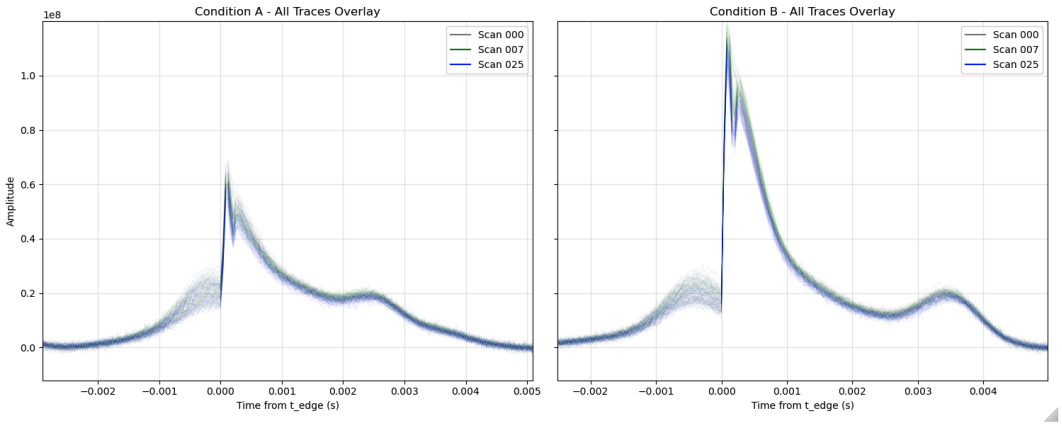


1.2 Region 2 Pulse Structure

Initial analysis of the Region 2 signal revealed that a standard single-Gaussian model was insufficient. The figure below displays all traces from a single file, overlaid one upon the other. This clearly shows that the signal in Region 2 contains a complex with a detectable signature. This signature is characterized by three distinct components:

1. **The Primary Sharp Peak:** A high-amplitude, narrow pulse occurring immediately after t_{edge}.
2. **The Secondary Sharp Peak:** A closely spaced "echo" or secondary mode, appearing < 1 ms after the primary peak.
3. **The Broad "Cloud" Component:** A significantly wider, lower-amplitude feature appearing later in the time window (> 5 ms post-edge).

Standard fitting algorithms often fail to separate these components, converging on a single broad Gaussian that underestimates the true peak amplitude. To accurately quantify A_{post}, a multi-component model with **strict shape constraints** is required.



1.3 Mathematical Framework

1.3.1 Region 1 Model

Pulse 1 is modeled as a single Gaussian with a DC offset:

$$S_1(t) = A_{pre} \exp\left(-\frac{(t-\mu_{pre})^2}{2\sigma_{pre}^2}\right) + C$$

Note: For this assignment, sensor bias (C) is not included in the calculation of Pulse 1's peak.

1.3.2 Region 2 Model

To capture the unique signature of the pulse in the second region, it is modeled as a superposition of three independent Gaussian functions:

$$S_2(t) = \sum_{i=1}^{3} A_i \exp\left(-\frac{(t-\mu_i)^2}{2\sigma_i^2}\right) + C$$

Where the components are constrained as follows:

- **Components 1 & 2 :** Constrained to be "sharp" ($\sigma < \sigma_{thresh}$) and located near t_{edge}.
- **Component 3 :** Allowed to be "broad" ($\sigma > \sigma_{thresh}$) and located later in the time window.

Also similar to Region 1, sensor bias C is not included in the final calculation of Pulse 2's peak.

1.3.3 Amplitude Extraction

Unlike the single-pulse case, the amplitude A_{post} is not simply one of the parameters A_i. Due to the significant overlap between the first two sharp peaks, the effective signal strength is the maximum value of the composite model above the baseline:

$$A_{post} = \max_t \left(S_2(t) - C\right)$$

This formulation ensures that constructive interference between the closely spaced modes is correctly accounted for in the final ratio calculation.

1.4 Estimation Strategy

The parameters are estimated using non-linear least squares optimization (`scipy.optimize.curve_fit`). To ensure physical validity, bounded constraints are applied:

- **Width Constraints:** $\sigma_{1,2} \in [0, 3.0\Delta t]$ to force sharpness (tunable).
- **Positional Constraints:** $\mu_{1,2}$ are bounded within the immediate post-edge window.

The uncertainty in the final ratio R is propagated from the covariance matrix of the fit, using Monte Carlo sampling to estimate the variance of the composite peak A_{post}.

1.5 Phase Extraction

Once the amplitude ratio R is determined for each trace in a scan, the interferometric phase shift ϕ is extracted by analyzing the variation of R as a function of the scan parameter θ (typically a rotation angle in degrees or radians).

The relationship between the measured ratio and the scan parameter is modeled as a sinusoidal fringe pattern:

$$R(\theta) = C_{fringe} + A_{fringe} \sin(\theta + \phi)$$

Where:

- A_{fringe} is the fringe amplitude (contrast).
- C_{fringe} is the fringe offset (DC component).
- ϕ is the phase shift of interest.

The phase ϕ is estimated by fitting this model to the dataset $\{(\theta_i, R_i)\}$ using a weighted non-linear least squares approach. The weights w_i are inversely proportional to the squared uncertainty of the individual ratio measurements ($\sigma_{R_i}^2$):

$$w_i = \frac{1}{\sigma_{R_i}^2}$$

This weighting scheme ensures that ratios derived from noisy traces or poor Gaussian fits contribute significantly less to the final phase determination.

2 End-to-end error propagation:

pulse fits $\rightarrow$ ratios $\rightarrow$ phase (closed form)

This is a closed-form uncertainty pipeline for the challenge: estimate a sinusoidal phase shift from ratio measurements derived from pulse fits.

2.1 Problem formulation

Estimate the **phase** ϕ of a sinusoid relating a derived scalar y_i to scan x_i, and report ϕ with uncertainty (e.g., standard error or CI):

$$\phi \pm \mathrm{SE}(\phi).$$

Derived scalar (ratio)

For each trace i , define two peak amplitudes:

- $A_{1,i}$: peak amplitude in **Region 1** (pre-t_{edge}).
- $A_{2,i}$: peak amplitude in **Region 2** (post-t_{edge}). In this challenge, Region 2 is treated as a **single composite pulse** formed by three underlying Gaussians.

Then:

$$y_i = \frac{A_{1,i}}{A_{2,i}}.$$

Uncertainty propagation flows as:

$$\text{trace noise} \Rightarrow \text{pulse-fit parameter uncertainty} \Rightarrow \text{peak uncertainty} \Rightarrow \text{ratio uncertainty} \Rightarrow \text{phase uncertainty}.$$

2.2 Pulse model + fitting (per trace, per region)

2.2.1 Define your pulse model

For each trace i , fit a parametric model separately in each region:

- Region 1 model: $m_1(t;\theta_{1,i})$
- Region 2 model: $m_2(t;\theta_{2,i})$

Examples:

- Gaussian + baseline: $m(t) = b + A\exp(-(t-\mu)^2/(2\sigma^2))$
- Template scaling + baseline: $m(t) = b + a\,g(t-\tau)$ where g is a fixed template.

The template approach will be discussed later.

Design choice: In this assignment, I considered a few different options for the Region 2 model m_2 : including a single Gaussian pulse, a summation of two Gaussian Pulses, or a summation of three Gaussian Pulses. **I recognize the importance of what the Region 2 pulse actually represents and the assumptions involved; however, I tackled this problem primarily to illustrate my thought process.**

2.2.2 Fit parameters by (nonlinear) least squares

Let $\mathbf{t}_r$ and $\mathbf{s}_{r,i}$ be the time grid and samples in region $r \in 1,2$ for trace i.

Fit:

$$\hat{\theta}_{r,i} \;=\; \arg\min_{\theta} \|\mathbf{s}_{r,i} - m_r(\mathbf{t}_r;\theta)\|^2.$$

2.2.3 Estimate residual variance (noise level in the window)

Compute residuals:

$$\mathbf{e}_{r,i} = \mathbf{s}_{r,i} - m_r(\mathbf{t}_r;\hat{\theta}_{r,i}).$$

Estimate residual variance:

$$\hat{\sigma}^2_{r,i} = \frac{\sum_k e_{r,i}(k)^2}{n_r - p_r},$$

Where n_r is the number of samples in region r and p_r is the number of fitted parameters.

Interpretation: this $\hat{\sigma}^2_{r,i}$ is the local noise/mismatch level after fitting.

2.2.4 Parameter covariance from the Jacobian

Let $J_{r,i}$be the Jacobian of the model w.r.t. parameters evaluated at$\hat{\theta}_{r,i}$:

$$J_{r,i}(k,:) = \left.\frac{\partial m_r(t_k;\theta)}{\partial \theta^T}\right|_{\theta=\hat{\theta}_{r,i}}.$$

Approximate parameter covariance (standard NLLS result):

$$\boxed{\mathrm{Cov}(\hat{\theta}_{r,i}) \;\approx\; \hat{\sigma}^2_{r,i}\,(J_{r,i}^T J_{r,i})^{-1}}$$

What this means:

- Bigger residual variance$\hat{\sigma}^2_{r,i} \rightarrow$ larger parameter uncertainty.
- Ill-conditioned$J^T J$ (non-identifiable fit) $\rightarrow$ large parameter uncertainty.

Caveat (model mismatch): this covariance assumes the model family is adequate. If the model is structurally wrong, covariance may understate bias. I will discuss this more later when looking at the goodness of fit for different assumptions of the Region 2 pulse.

2.3 Peak amplitude estimate + uncertainty (per trace)

2.3.1 Define peak as a function of parameters

Define the peak amplitude functional for each region:

$$A_{r,i} = f_r(\theta_{r,i}) \quad (r = 1, 2).$$

2.3.2 Propagate parameter covariance to peak variance (delta method)

Compute gradient at the estimate:

$$\mathbf{g}_{r,i} = \nabla_\theta f_r(\hat{\theta}_{r,i}).$$

Then:

$$\boxed{\mathrm{Var}(\hat{A}_{r,i}) \;\approx\; \mathbf{g}_{r,i}^T \,\mathrm{Cov}(\hat{\theta}_{r,i})\, \mathbf{g}_{r,i}}$$

and

$$\mathrm{SE}(\hat{A}_{r,i}) = \sqrt{\mathrm{Var}(\hat{A}_{r,i})}$$

Engineering note: If $A$is a direct fitted parameter (very common), this reduces to just reading the appropriate diagonal entry of $\mathrm{Cov}(\hat{\theta})$.

2.3.3 Ratio estimate + uncertainty (per trace)

Define ratio

$$\hat{y}_i = \frac{\hat{A}_{1,i}}{\hat{A}_{2,i}}.$$

Propagate peak uncertainties to ratio variance

Treat $y(A_1, A_2) = A_1/A_2$. The gradient is:

$$\frac{\partial y}{\partial A_1} = \frac{1}{A_2}, \qquad \frac{\partial y}{\partial A_2} = -\frac{A_1}{A_2^2}.$$

General 2-variable propagation:

$$\mathrm{Var}(\hat{y}_i) \approx \nabla y^T \begin{bmatrix} \mathrm{Var}(\hat{A}_{1,i}) & \mathrm{Cov}(\hat{A}_{1,i}, \hat{A}_{2,i}) \\ \mathrm{Cov}(\hat{A}_{1,i}, \hat{A}_{2,i}) & \mathrm{Var}(\hat{A}_{2,i}) \end{bmatrix} \nabla y.$$

Expanded:

$$\boxed{\mathrm{Var}(\hat{y}_i) \approx \left(\frac{1}{\hat{A}_{2,i}}\right)^2 \mathrm{Var}(\hat{A}_{1,i}) + \left(\frac{\hat{A}_{1,i}}{\hat{A}_{2,i}^2}\right)^2 \mathrm{Var}(\hat{A}_{2,i}) - 2\left(\frac{\hat{A}_{1,i}}{\hat{A}_{2,i}^3}\right)\mathrm{Cov}(\hat{A}_{1,i}, \hat{A}_{2,i})}$$

Typical simplification: because regions are disjoint in time, it is often reasonable to assume $\mathrm{Cov}(\hat{A}_{1,i}, \hat{A}_{2,i}) \approx 0$, giving:

$$\mathrm{Var}(\hat{y}_i) \approx \left(\frac{1}{\hat{A}_{2,i}}\right)^2 \mathrm{Var}(\hat{A}_{1,i}) + \left(\frac{\hat{A}_{1,i}}{\hat{A}_{2,i}^2}\right)^2 \mathrm{Var}(\hat{A}_{2,i}).$$

Define ratio standard deviation:

$$s_i \equiv \mathrm{SE}(\hat{y}_i) = \sqrt{\mathrm{Var}(\hat{y}_i)}.$$

Interpretation: s_i is the typical jitter you'd expect in the ratio if you repeated the same trace measurement and refit.

2.3.4 Sinusoid fit for phase (across scan settings)

2.3.4.1 Linear-in-parameters sinusoid model

Fit the ratio-vs-scan relation:

$$\hat{y}_i \approx c + b\sin(\omega x_i) + d\cos(\omega x_i) + \varepsilon_i, \quad \omega = 2\pi/360.$$

Define design row:

$$\mathbf{x}_i = [1,\ \sin(\omega x_i),\ \cos(\omega x_i)].$$

Stack rows into matrix X and observations into vector $\mathbf{y}$.

2.3.4.2 Use ratio uncertainties as weights (WLS)

Assume

$$\mathrm{Var}(\varepsilon_i) \approx \mathrm{Var}(\hat{y}_i) = s_i^2$$

. Then set weights:

$$w_i = \frac{1}{s_i^2}.$$

Let $W = \mathrm{diag}(w_i)$. Weighted least squares estimator:

$$\boxed{\hat{\beta} = (X^T W X)^{-1} X^T W \mathbf{y}, \quad \hat{\beta} = [\hat{c}, \hat{b}, \hat{d}]^T}$$

2.3.4.3 Covariance of sinusoid parameters

If the s_i are absolute standard deviations (not just relative), then:

$$\boxed{\mathrm{Cov}(\hat{\beta}) \approx (X^T W X)^{-1}}$$

If weights are only relative (common), include a scale estimated from weighted residuals:

$$\hat{\sigma}_w^2 = \frac{\sum_i w_i (y_i - \hat{y}_i)^2}{N - p}, \quad \mathrm{Cov}(\hat{\beta}) \approx \hat{\sigma}_w^2 (X^T W X)^{-1}.$$

2.3.5 Phase estimate + uncertainty from b, d

2.3.5.1 Phase definition

With $b\sin + d\cos$ form, define phase:

$$\boxed{\hat{\phi} = \operatorname{atan2}(\hat{d}, \hat{b})}$$

5.2 Delta-method propagation to phase variance

Let Σ_{bd} be the 2×2 submatrix of $\operatorname{Cov}(\hat{\beta})$ corresponding to (b, d):

$$\Sigma_{bd} = \begin{bmatrix} \operatorname{Var}(\hat{b}) & \operatorname{Cov}(\hat{b}, \hat{d}) \\ \operatorname{Cov}(\hat{b}, \hat{d}) & \operatorname{Var}(\hat{d}) \end{bmatrix}.$$

For

$\phi = \operatorname{atan2}(d, b)$, gradient is:

$$\frac{\partial \phi}{\partial b} = -\frac{d}{b^2 + d^2}, \qquad \frac{\partial \phi}{\partial d} = \frac{b}{b^2 + d^2}.$$

Closed form:

$$\boxed{\operatorname{Var}(\hat{\phi}) \approx \frac{d^2 \operatorname{Var}(b) + b^2 \operatorname{Var}(d) - 2bd \operatorname{Cov}(b, d)}{(b^2 + d^2)^2}}$$

Standard error:

$$\boxed{\operatorname{SE}(\hat{\phi}) = \sqrt{\operatorname{Var}(\hat{\phi})}}$$

Convert to milliradians:

$$\operatorname{SE}_{\text{mrad}} = 1000 \cdot \operatorname{SE}(\hat{\phi}).$$

2.3.5.3 Confidence interval (approx.)

A simple symmetric CI (assuming approximate normality):

$$\hat{\phi} \pm z \operatorname{SE}(\hat{\phi}), \quad z = 1.96 \text{ for } 95\%.$$

If the interval crosses $\pm\pi$, unwrap around $\hat{\phi}$ before reporting.

Method

The processing pipeline comprises three principal stages:

1. **Data Loading:** Initial step to import necessary data.
2. **Trace/Condition Processing:** Detailed analysis for each trace and condition within every file:
    - For each trace:
        - Region 1: Fit a single Gaussian plus a bias term, and determine the associated uncertainty.
        - Region 2: Fit a summation of three Gaussians, and determine the associated uncertainty.
        - Ratio Calculation: Calculate the ratio and its uncertainty based on the peaks of the two regions.
    - The output of this step is a dictionary containing the calculated ratio, uncertainty, and other diagnostic information.
3. **Phase Shift Estimation:** Fit a sine wave curve to the results using weighted least squares to estimate the phase shift and its associated uncertainty.

Accompanying this report is my GitHub repository, which details the code structure, and a Jupyter notebook documenting the various experiments conducted for this assignment.

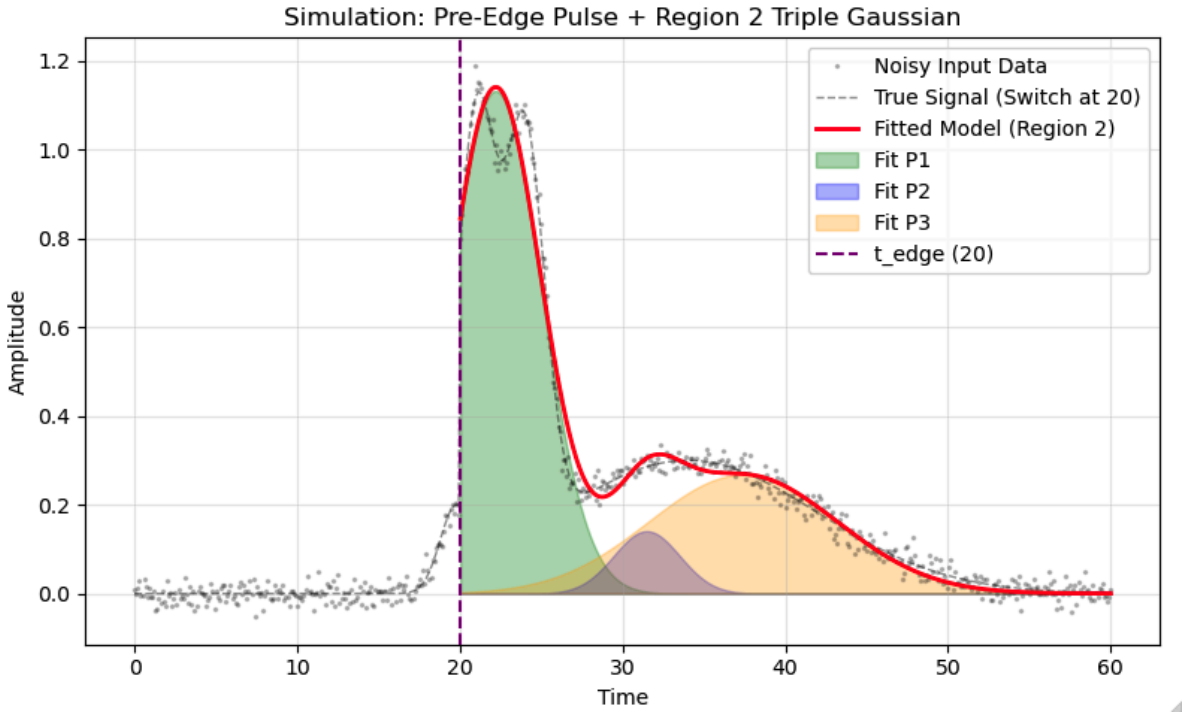
This section outlines my thought process and approach to the unique challenges presented by this assignment. Instead of providing a step-by-step account of the code writing, I focus on individual, related points where I had to address specific aspects of the problem and the associated dataset. I must reiterate that, due to the limited time allocated for this assignment, I was unable to fully explore every facet mentioned.

Synthetic Signal

Synthetic Signal

In the absence of a reference signal, I first created a synthetic signal that mimicked the data type found in the Nomad Atomics *.npz files. This synthetic data allowed me to manipulate the signal's parameters to assess the effectiveness of my curve fitting. Specifically, I tested its ability to extract the signal, even with (1) noise and (2) the challenge of partial data capture and (3) with the unique shape of Region 2 pulse. For instance, I could manually adjust the captured portion of each pulse and observe if the curve fitting accurately identified the peak locations and the ratio between them. The full code is available in the Jupyter Notebook.

Given the limited time for the assignment, this approach was primarily used to explore the curve fitting process, focusing on the impact of: (1) initialization, and (2) constraints applied to the Gaussian shapes. This investigation helped detect certain corner cases. For example, I realized, as somewhat expected, that if the curve fitting problem is poorly defined in the second region, the algorithm can converge to a local minimum ("get lazy"). This manifests as approximating the two narrow pulses (which I assumed were present) in that region with a single, broader "soft" pulse.



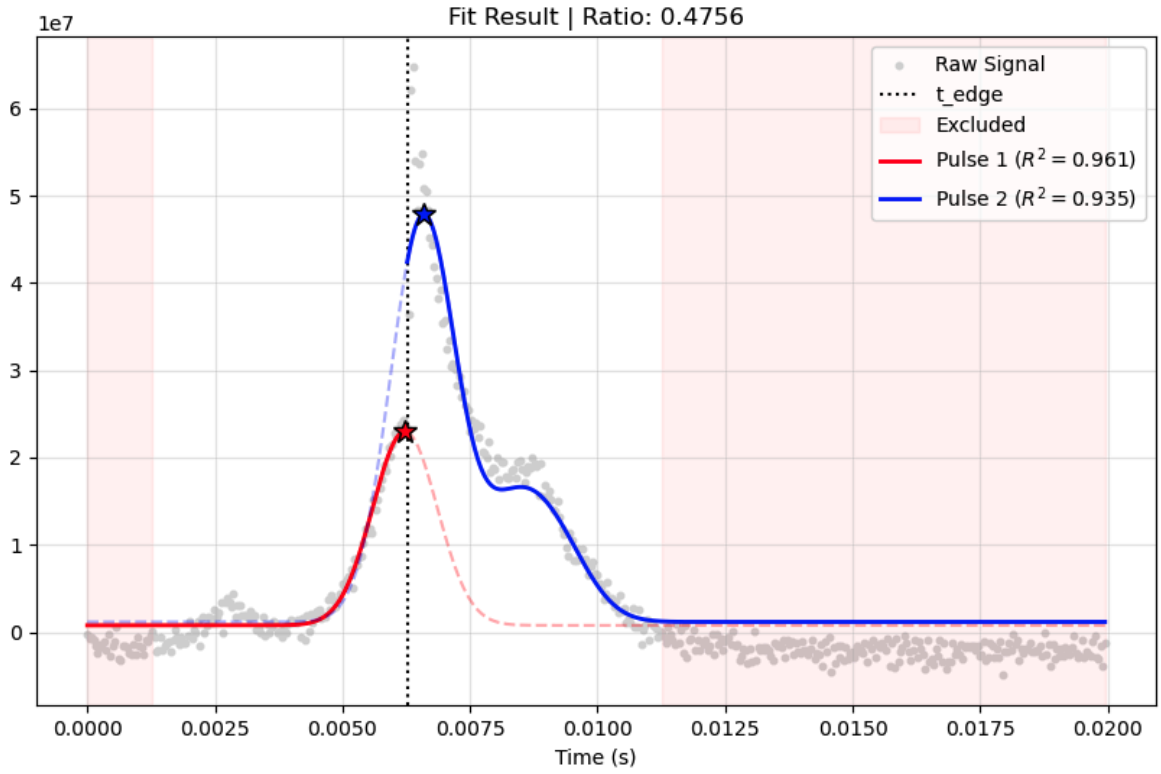
Signal Windowing

Signal Windowing

The success of phase identification is heavily dependent on the quality of the curve fitting to the noisy signal. To improve this process, I explored methods to assist the curve fit function. One approach involved limiting the signal window passed to the curve fit in each region. This decision aimed to reduce the number of points in the least squares problem (which is central to the curve fit), particularly because the Signal-to-Noise Ratio (SNR) is higher near the t_{edge}.

While this windowing option remains in the final submission, I ultimately did not use it. I found that bounding the estimated parameters had a more significant positive impact on the success of the curve fit.

The figure below simply illustrates the application of the windowing technique. The excluded portion of each section is indicated in the figure. For this example, all data points occurring after and before 0.005 s from the t_{edge} were omitted from the curve fit.



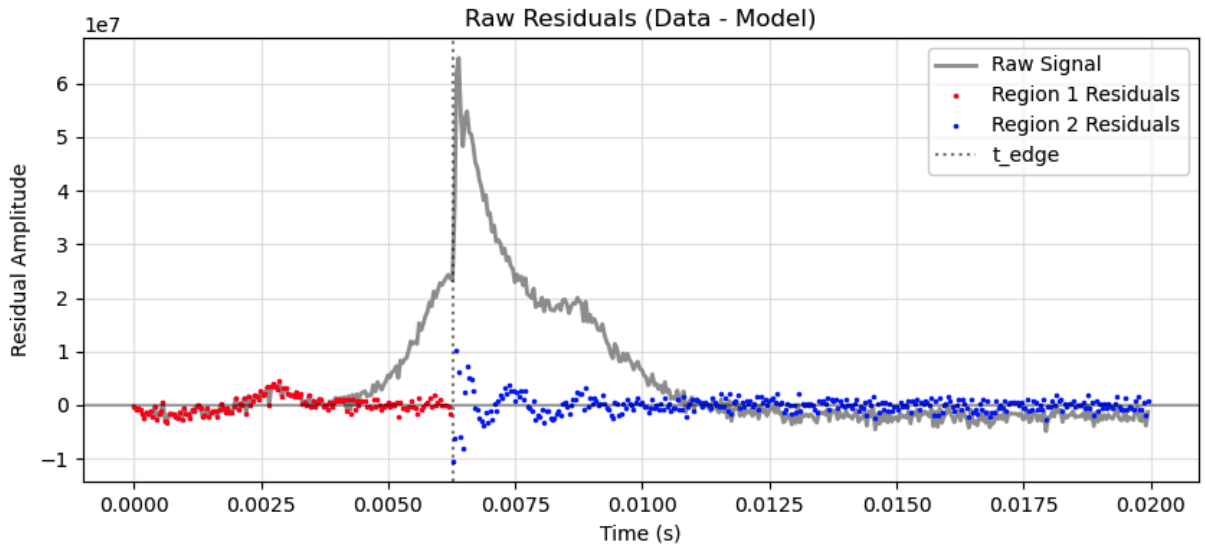
Soft Cropping

Soft Cropping

Related to signal windowing to focus the curve fitting on the regions where the signal to noise ratio is higher, another idea was to use weighted least square (instead of ordinary least square) with some form of weights that focuses the curve fit more around the

Goodness of fit

The following is an example of how I evaluated the quality of fit as I went through the assignment. I plotted the raw residual to visualize how well the model fit the data in each region. Below is one example:



Here is a breakdown of what the "shapes" in the plot are indicating:

1. Region 1 (Red): The "Sine Wave" Error

- **Observation:** The residuals do not scatter randomly. They dip negative, swing high positive (around 0.003s), and then return to zero. This forms a clear "S-curve" or sine-wave shape. However I am happy that closer to the t_edge the residual is closer to zero. The earlier sine-wave shape is more likely due to the unmodelled local maxima in the data.

2. Region 2 (Blue): The "Transient" Failure

- **Observation:** There is a massive explosion of error right at the t_edge boundary (the vertical dotted line). The residuals spike up to 10^7 and oscillate wildly before settling down.
- **Diagnosis:** This is most likely a Model Structural Failure.
    - The solver is completely failing to match the "Sum of 3 Gaussians" at the leading edge.
    - It looks like I might be fitting that my Width (σ) constraints are forcing the model to be too "stiff."
    - The oscillation suggests the model is "chasing" the data but cannot curve fast enough to match the complex shape of the 3 overlapping pulses.

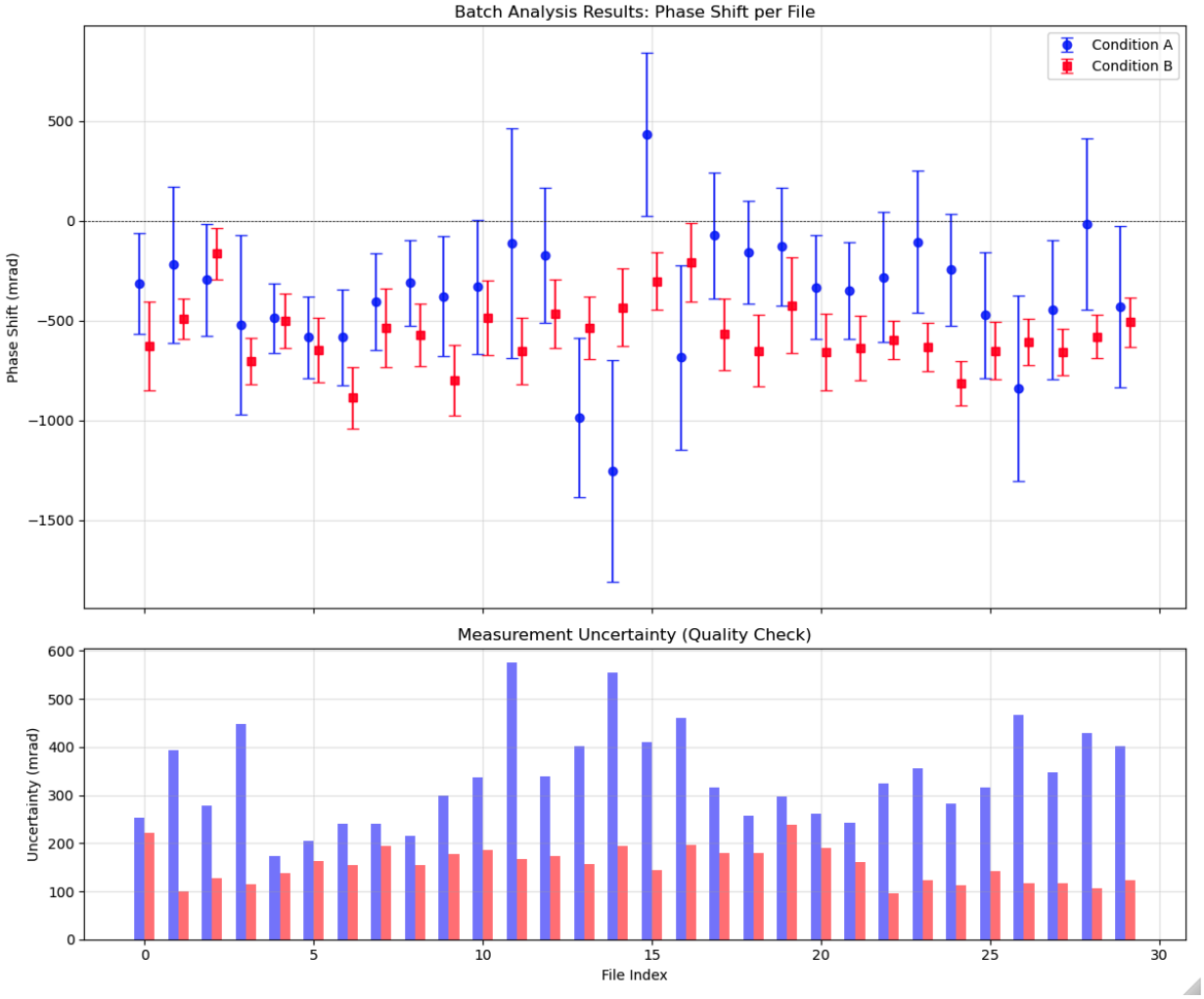
Trace Rejection

Trace Rejection

The curve fit function

Primary Task Results

The estimated phase shift, including its uncertainty, is displayed for each file and corresponding condition in the figure below.



Generally, the phase shift for Condition B is estimated with lower uncertainty. This is likely due to the clearer estimate of the peak in Region 1, as the peak appears to have been captured in the Condition B traces.

In contrast, the estimates for the Condition A traces were mostly poor, resulting in a less reliable phase shift. In some cases, the possibility of a negative phase shift, which could also be positive, suggests an unreliable result. I also note that the “reliability” depends significantly on the acceptability criteria for this application.

Secondary Task

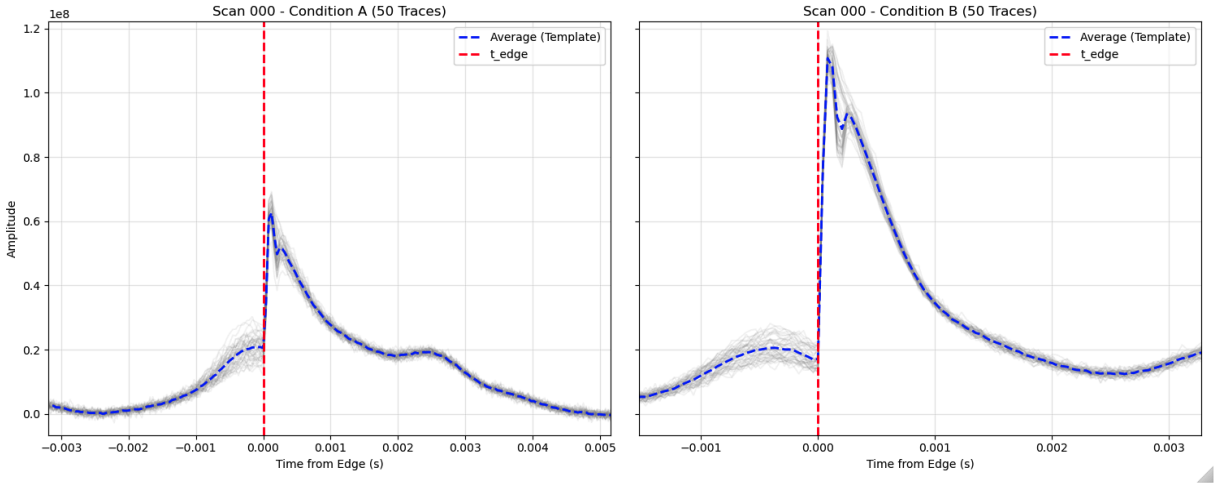
Secondary task

“If you needed to achieve similar uncertainty with significantly fewer data points, what would you do? Demonstrate.”

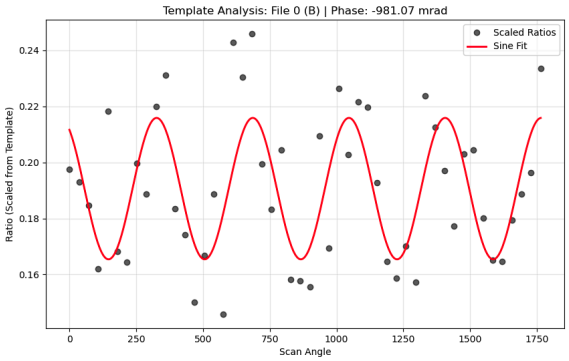
While the secondary task warrants further exploration, a general principle holds true: increased assumptions about any aspect of the estimation allow for a proportional reduction in data required to maintain the same level of uncertainty.

To me “fewer data points” can be interpreted as:
1- Fewer samples per scan
2- Fewer scans (shots)

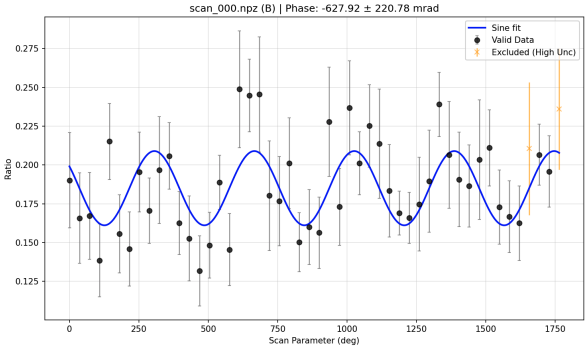
To achieve the goal of fewer samples per scan, a promising approach is to employ a signal template for each region. This strategy is based on the observation that all plotted signals share a consistent underlying structure. This critical prior knowledge regarding the pulse shape can be leveraged to fix certain parameters within the models for Region 1 and Region 2. Specifically, the center (μ_i) and spread (σ_i) could be predefined, simplifying the estimator function to focus solely on determining the scaling factor (A_i) for each region.



The initial examples demonstrated promising results. Nevertheless, additional work is required to more precisely define the uncertainty, which will enable a proper comparison against the baseline approach. A straightforward illustration, found in the final chapter of the notebook, presents a side-by-side view: the template method's result is on the left, and the baseline method's result is on the right.



Template method for Scan 000 Condition B



Baseline method for Scan 000 Condition B