

Positional Encoding

Sinusoidal and Learned Absolute Positional Encodings

Positional encoding injects position information into transformer inputs to break the permutation invariance of self-attention. Sinusoidal encodings use geometrically spaced frequencies to enable relative position learning.

Key properties demonstrated:

- $PE(pos+k) = M_k @ PE(pos)$ via rotation matrices
- Dot products depend only on relative distance (Toeplitz)
 - Self-dot = $d_{model}/2$ (from $\sin^2 + \cos^2 = 1$)
- Wavelengths form geometric progression: 2π to $\sim 10000 \cdot 2\pi$

Random seed: 42

Number of visualizations: 6

Examples: 6

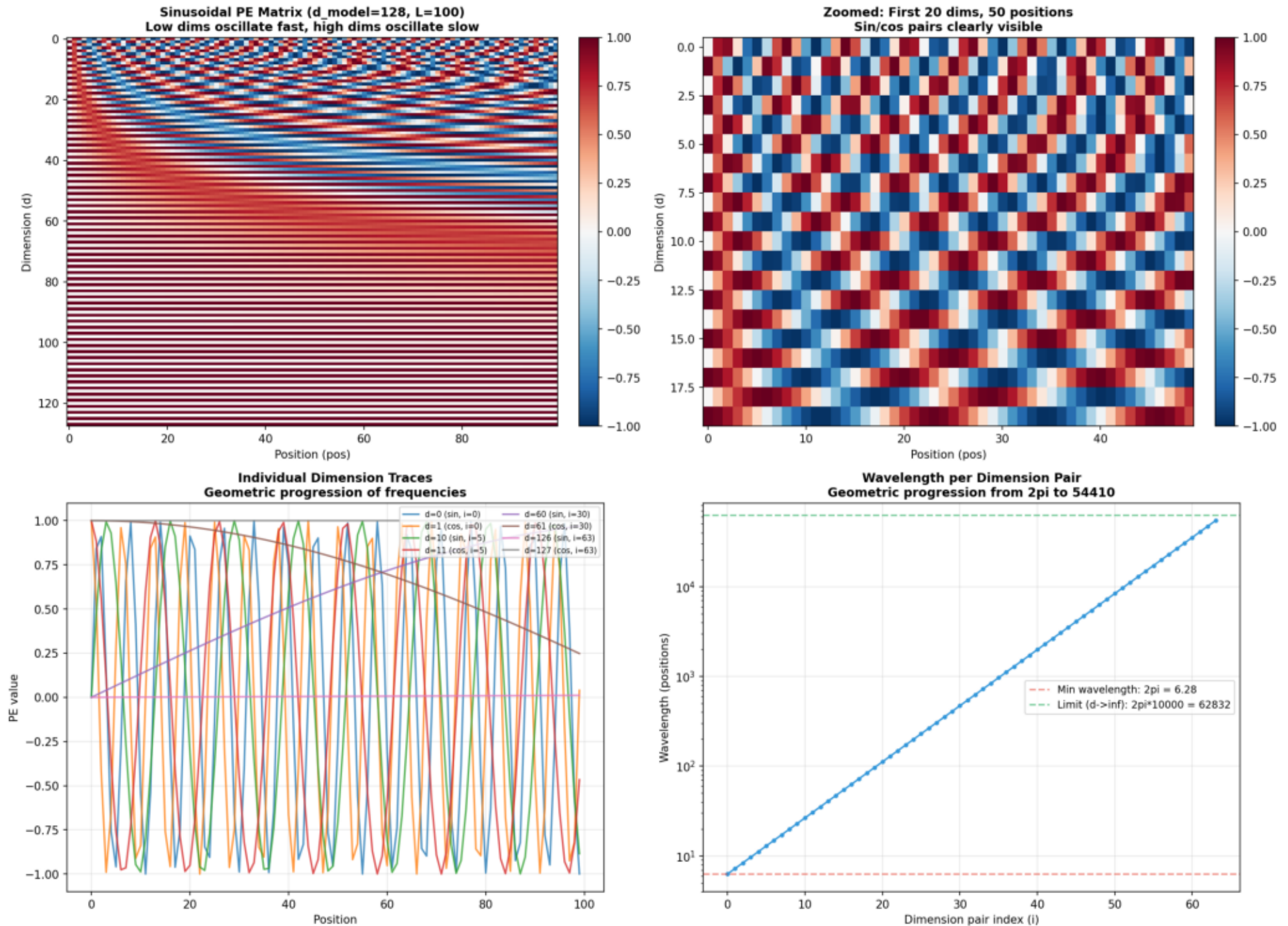
Generated by demo.py

Summary of Findings

1. PE Matrix Structure: Sinusoidal PE shows characteristic wave patterns.
Low dimensions ($i=0$) oscillate every ~ 6 positions ($\text{wavelength}=2\pi$).
High dimensions ($i=63$) have wavelength ~ 54000 positions ($d=128$).
Frequencies form a geometric progression with ratio $10000^{(2/d)}$.
2. Dot Product Distance: PE @ PE.T has perfect Toeplitz structure.
 $D[i,j] = \sum_k \cos(\omega_k * (i-j))$, depends ONLY on relative distance.
Self-dot product = $d_{\text{model}}/2$ exactly (from $\sin^2 + \cos^2 = 1$ per pair).
Variation at same distance: $\sim 1e-14$ (floating-point rounding only).
3. Relative Position Property: $\text{PE}(\text{pos}+k) = M_k @ \text{PE}(\text{pos})$ EXACTLY.
 M_k is block-diagonal with $d/2$ rotation blocks $R_i(k)$.
Reconstruction error: $\sim 1e-14$ (analytically exact, limited by float64).
This enables attention to learn relative positions via linear transforms.
4. Sinusoidal vs Learned: Sinusoidal has constant norm $\sqrt{d/2}$, perfect Toeplitz dot products, and unlimited extrapolation.
Learned (random init) has variable norms, no Toeplitz structure, and is capped at max_seq_len .
5. Attention Impact: Without PE, attention is permutation-equivariant (diff ~ 0). With PE, permuting input changes output significantly, making attention position-aware.
6. Frequency Analysis: Wavelengths span 2π to $\sim 10000 \cdot 2\pi$ (exact max depends on d).
Variance drops from ~ 0.5 (fast dims) to ~ 0 (slow dims).
Crossover where wavelength = seq_len determines which dims are informative for a given sequence length.

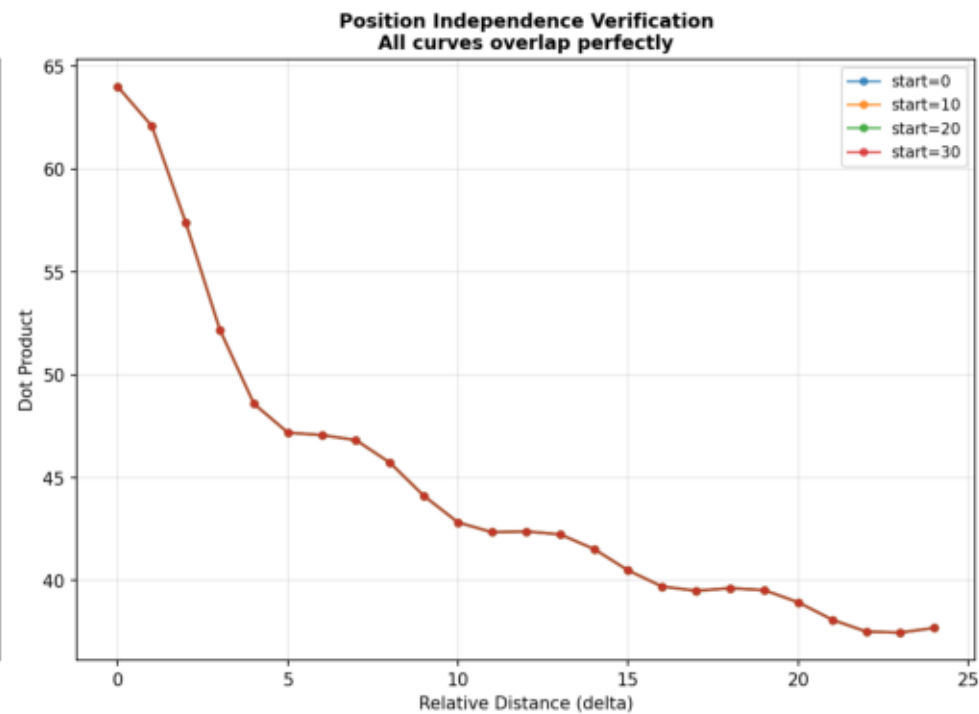
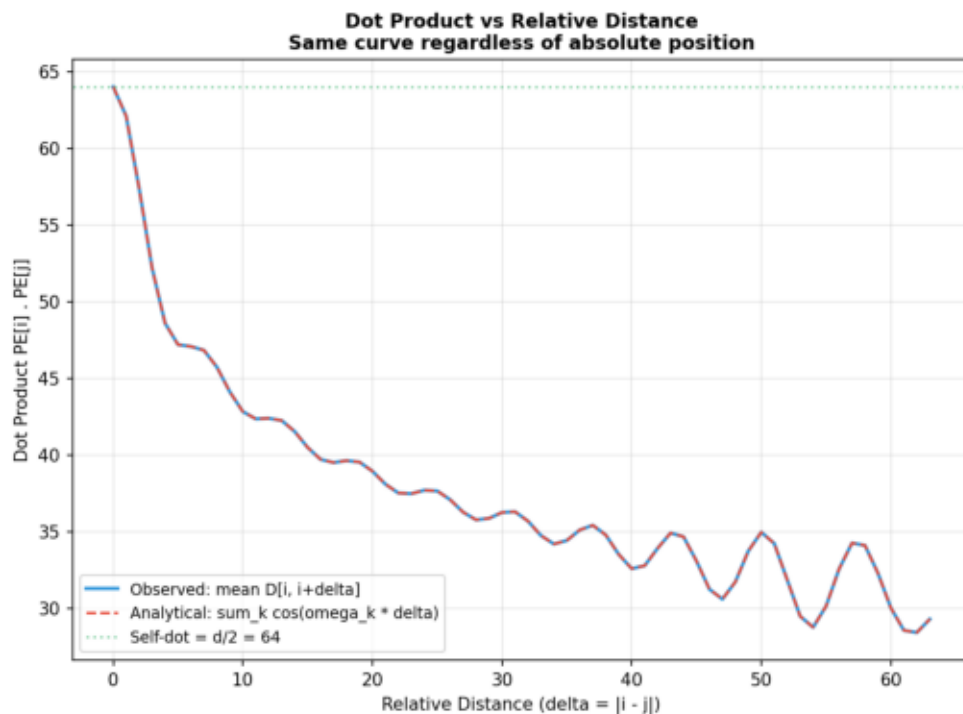
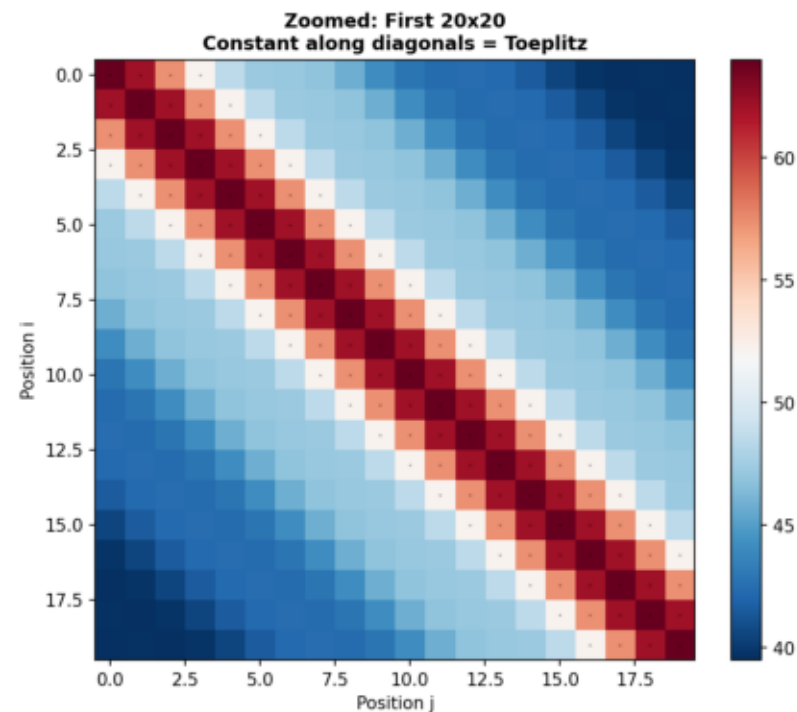
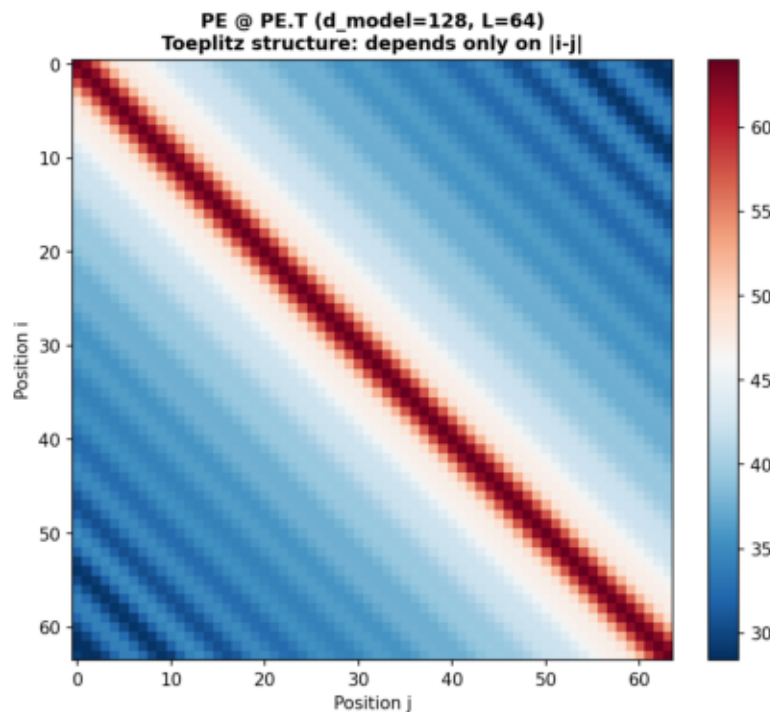
Example 1: PE Matrix Heatmap and Frequency Analysis

Sinusoidal Positional Encoding: Structure and Frequencies



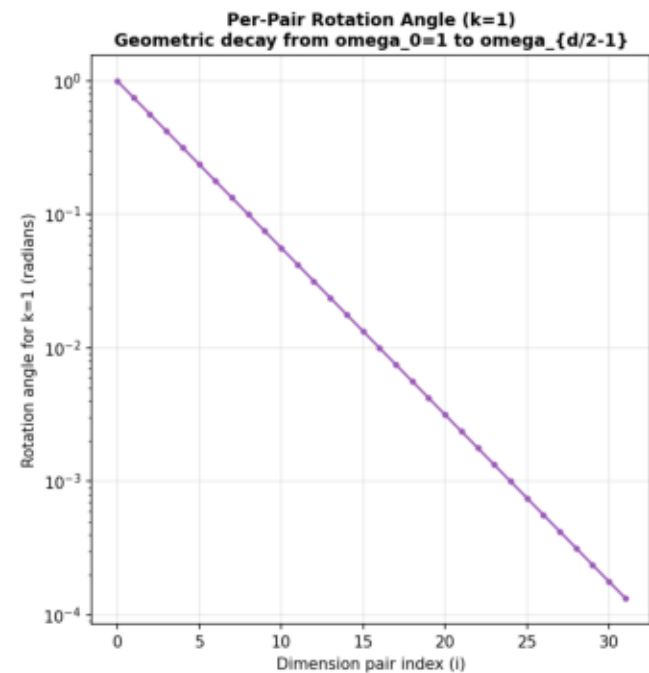
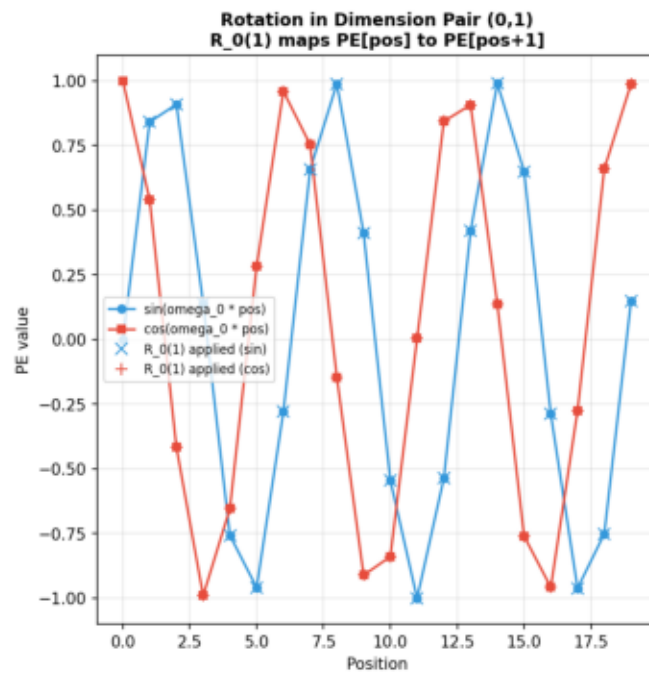
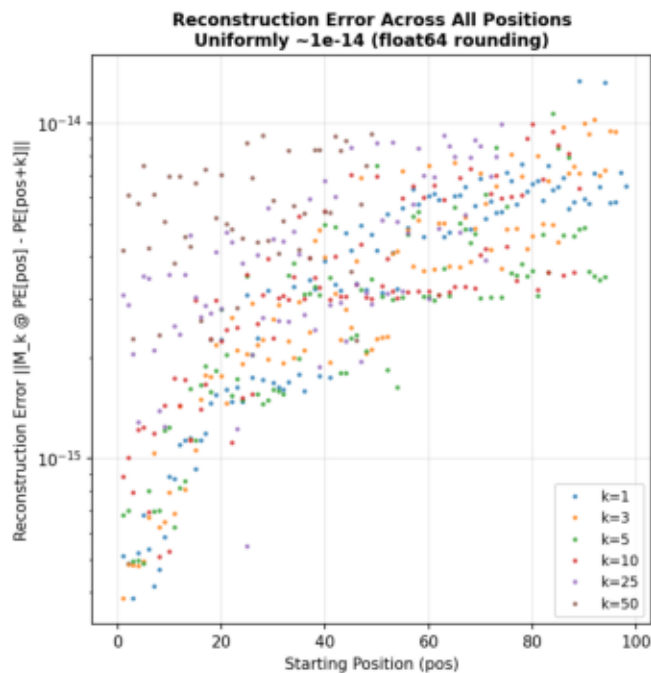
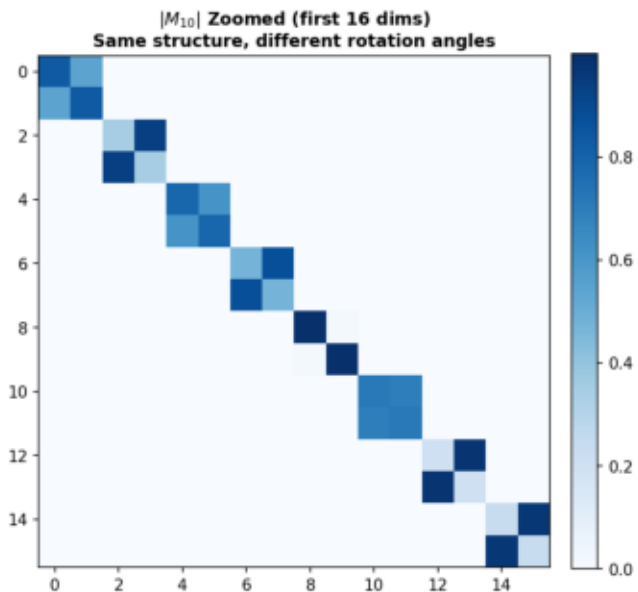
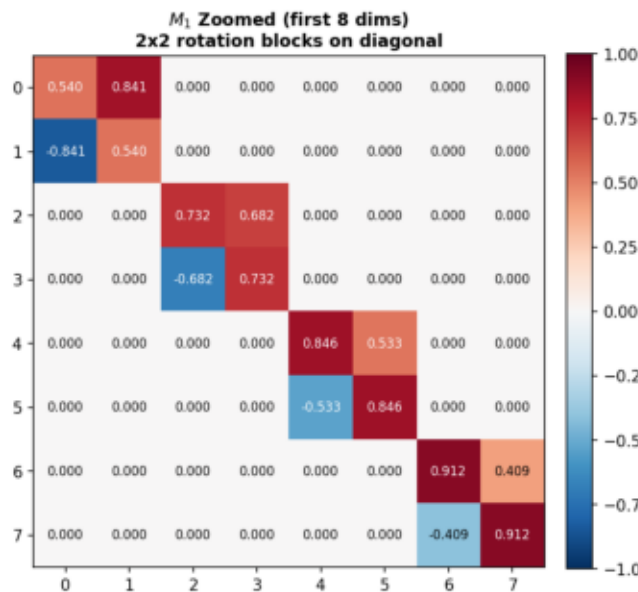
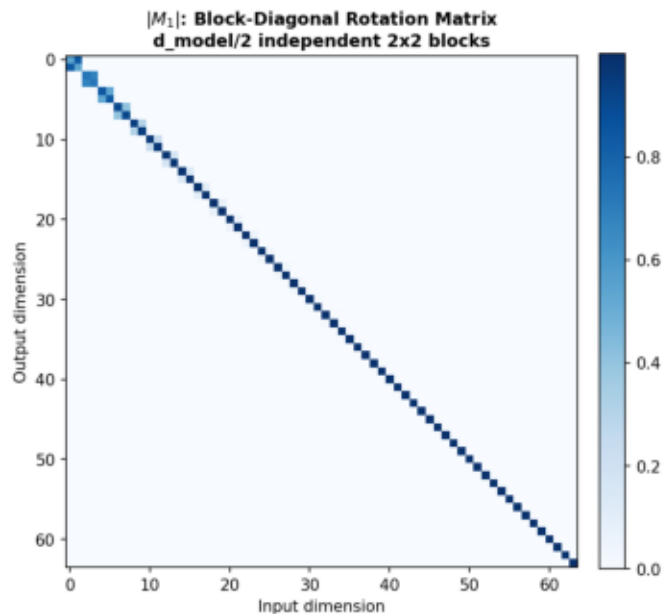
Example 2: Dot Product Distance Structure (Toeplitz Property)

Dot Product Distance Structure: Toeplitz Property of Sinusoidal PE



Example 3: Relative Position Property (Rotation Matrices)

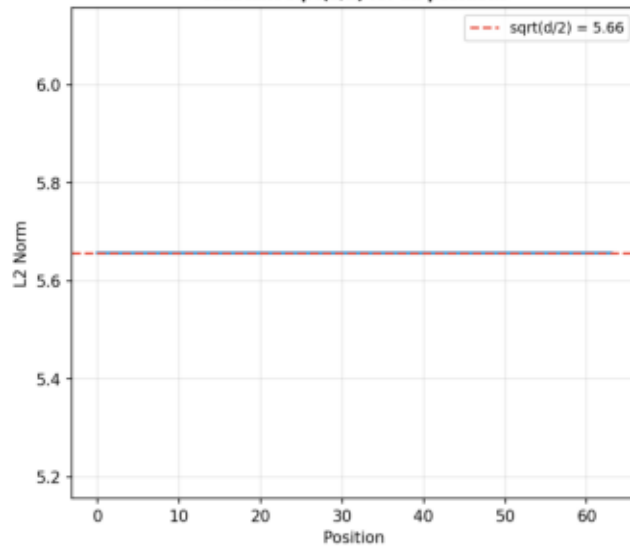
Relative Position Property: $PE(pos+k) = M_k @ PE(pos)$



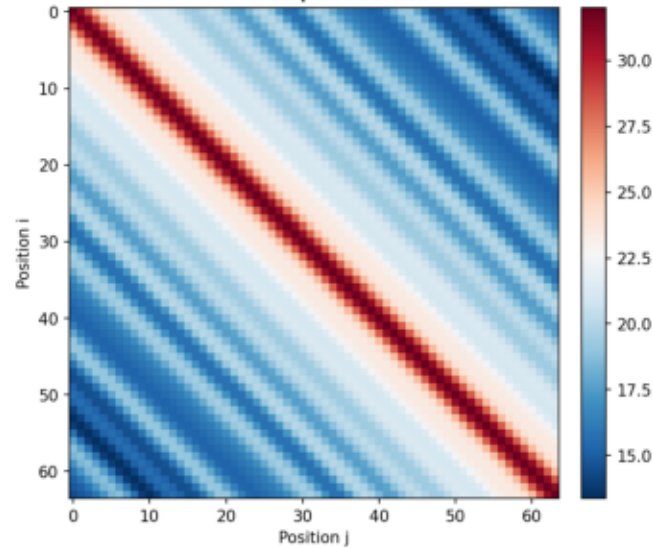
Example 4: Sinusoidal vs Learned Encoding Comparison

Sinusoidal vs Learned Positional Encoding Comparison

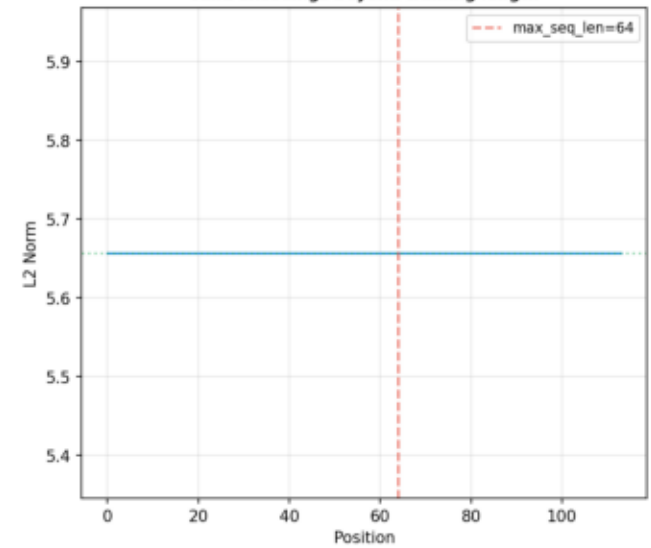
Sinusoidal: Position Norms
Constant $\sqrt{d/2}$ for all positions



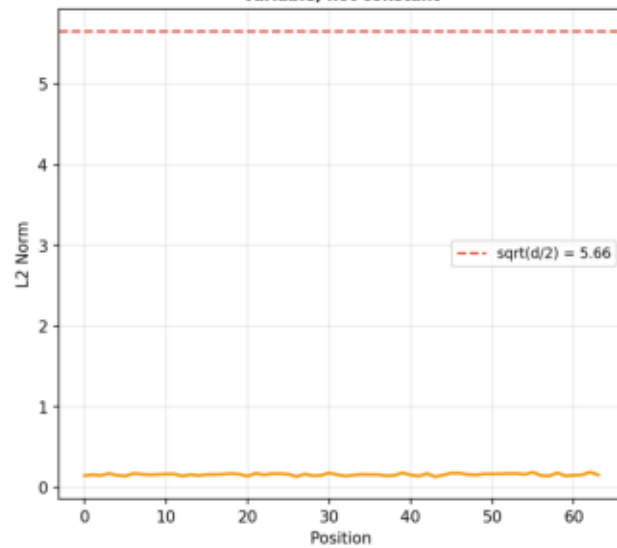
Sinusoidal: Dot Product Matrix
Perfect Toeplitz structure



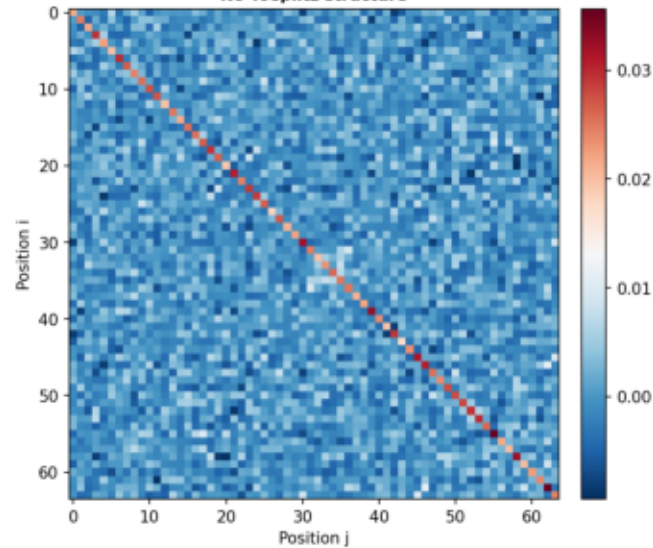
Sinusoidal: Extrapolation
Valid encodings beyond training length



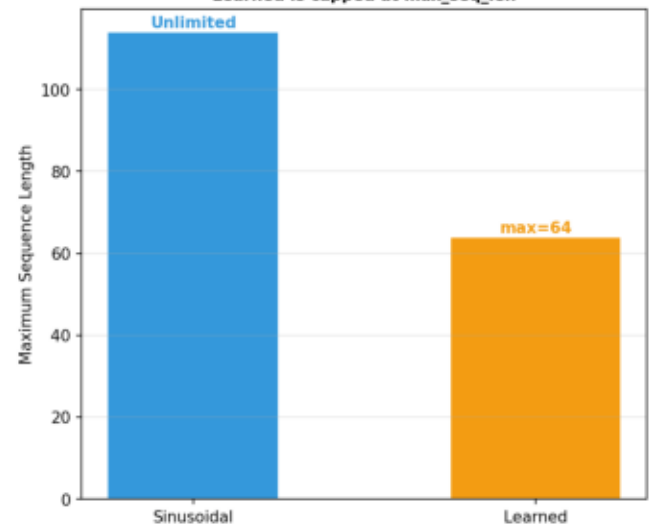
Learned (random init): Position Norms
Variable, not constant



Learned (random init): Dot Product Matrix
No Toeplitz structure

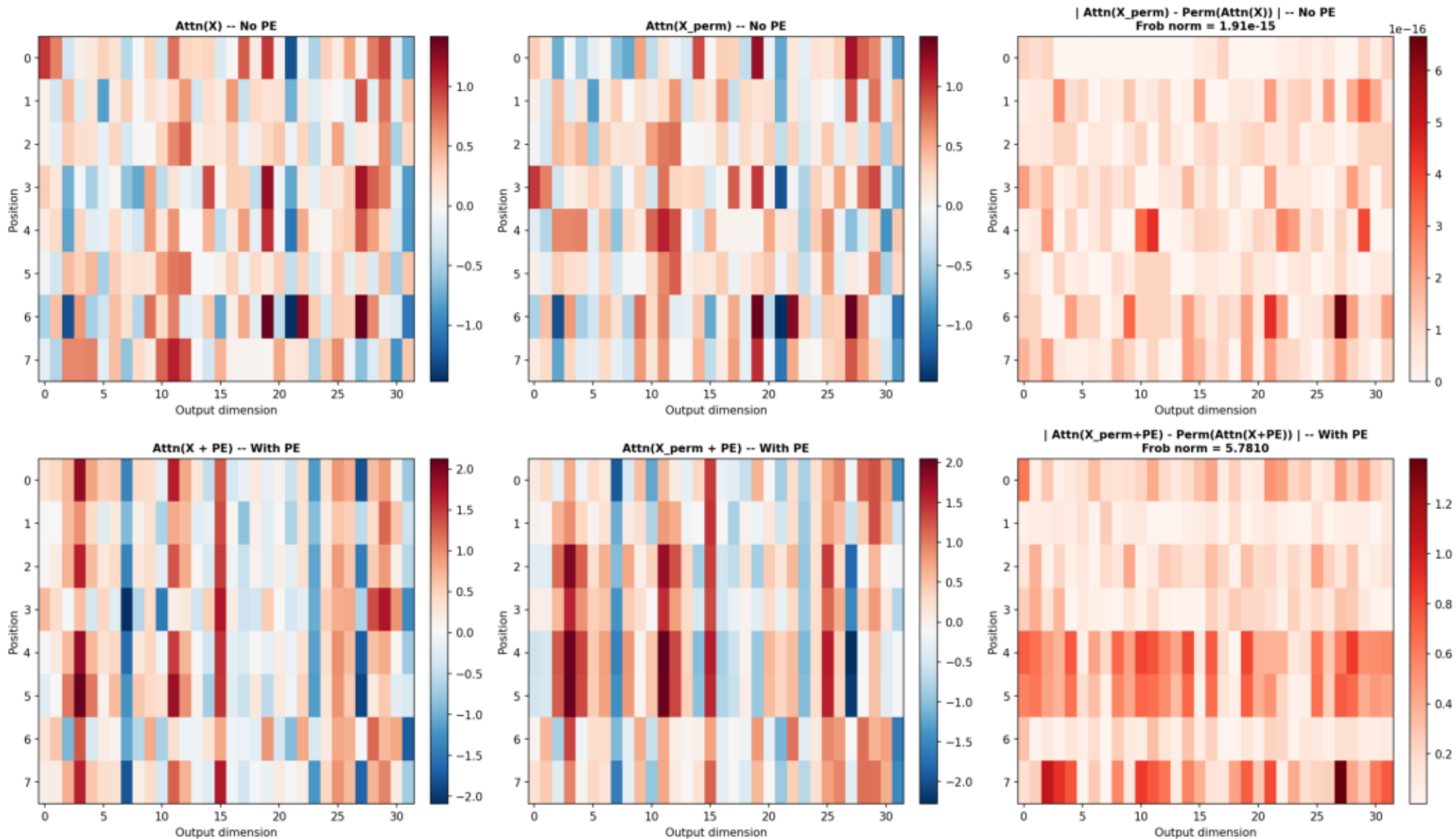


Extrapolation Capability
Learned is capped at max_seq_len



Example 5: Impact on Self-Attention (Permutation Invariance)

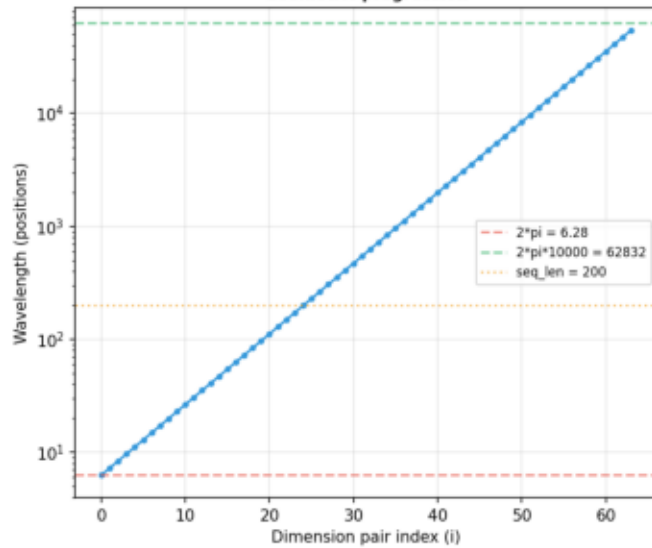
Impact of Positional Encoding on Self-Attention
Without PE: permutation-equivariant (top row, diff ~ 0). With PE: position-aware (bottom row, diff $\gg 0$).



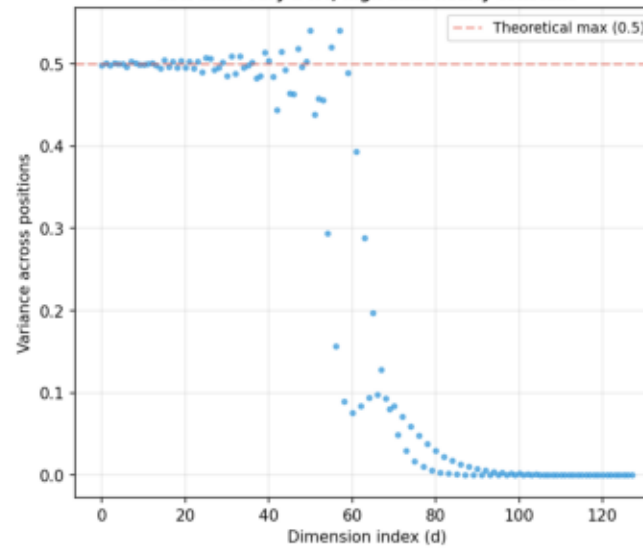
Example 6: Frequency Structure and Variance Analysis

Frequency Structure Analysis of Sinusoidal Positional Encoding

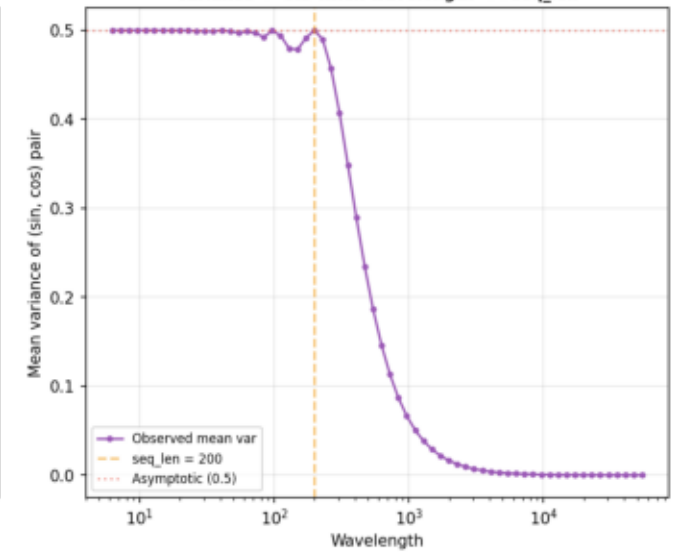
Wavelength per Dimension Pair
Geometric progression



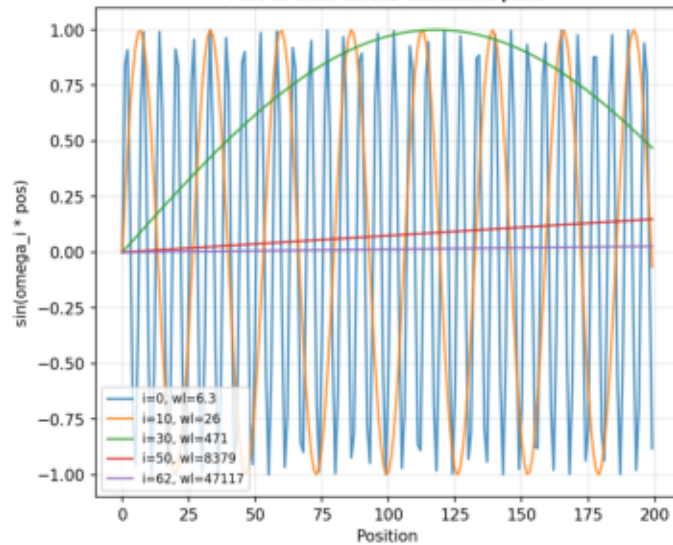
Variance per Dimension
Low dims vary a lot, high dims nearly constant



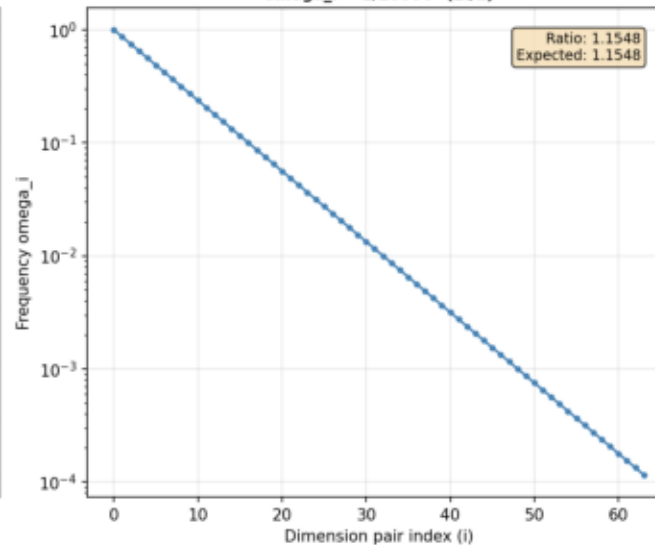
Variance vs Wavelength
Variance $\rightarrow 0.5$ when wavelength $\ll \text{seq_len}$



Sample Waveforms (sin component)
Fast to slow across dimension pairs



Frequency per Dimension Pair
 $\omega_i = 1/10000^{(2i/d)}$



Periods Completed in $\text{seq_len}=200$
Low dims: many periods; high dims: < 1 period

