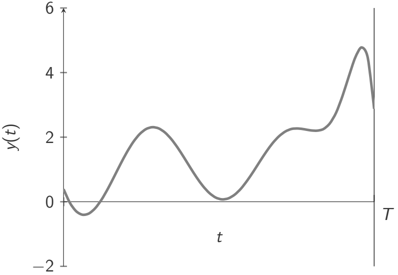
The Statistical Cost of Robust Kernel Hyperparameter Turning

Raphael A. Meyer with Christopher Musco

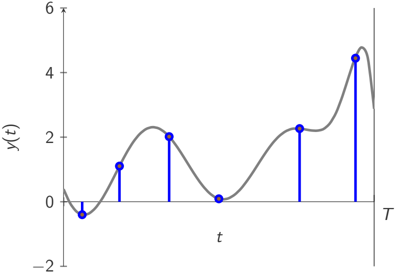
NYU, Tandon School of Engineering

Robust Active Interpolation



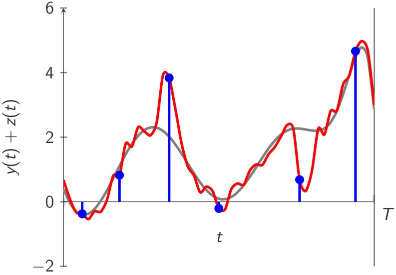
- We want to interpolate $y(t)$ for $0 \leq t \leq T$

Robust Active Interpolation



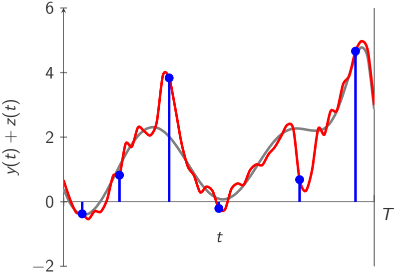
- We want to interpolate $y(t)$ for $0 \leq t \leq T$

Robust Active Interpolation

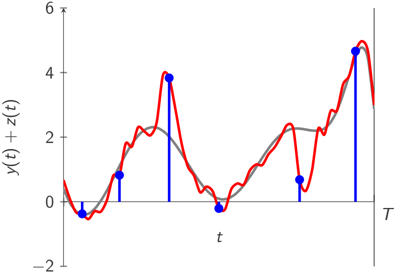


- We want to interpolate $y(t)$ for $0 \leq t \leq T$
- We observe $y(t) + z(t)$ for unknown $z(t)$

Robust Active Interpolation

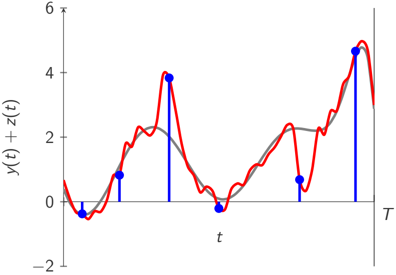


- We want to interpolate $y(t)$ for $0 \leq t \leq T$
- We observe $y(t) + z(t)$ for unknown $z(t)$
- How many observations are needed to interpolate $y(t)$?



- We want to interpolate $y(t)$ for $0 \leq t \leq T$
- We observe $y(t) + z(t)$ for unknown $z(t)$
- How many observations are needed to interpolate $y(t)$?
- Prior Work: For fixed kernel k, Kernel Ridge Regression with $\tilde{O}(\text{stat-dim}(k))$ observations suffice.

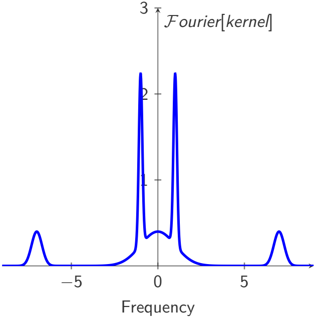
Robust Active Interpolation



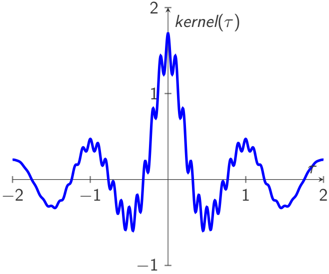
- We want to interpolate $y(t)$ for $0 \leq t \leq T$
- We observe $y(t) + z(t)$ for unknown $z(t)$
- How many observations are needed to interpolate $y(t)$?
- Prior Work: For fixed kernel k, Kernel Ridge Regression with $\tilde{O}(\text{stat-dim}(k))$ observations suffice.

NYU

Spectral Mixture Kernel

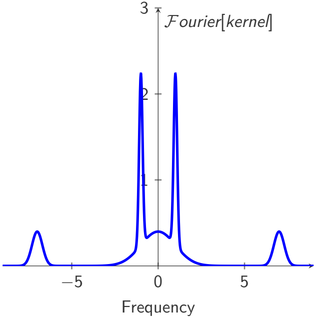


$\Longrightarrow$

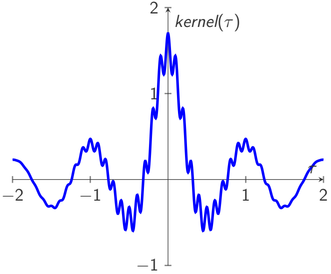


- Introduced in [WA13]; popular in Gaussian Process literature

Spectral Mixture Kernel

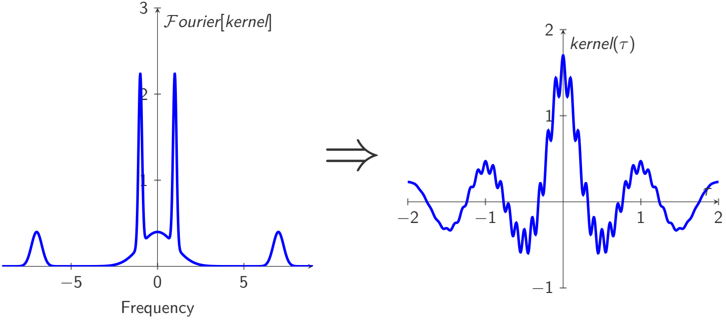


$\Rightarrow$



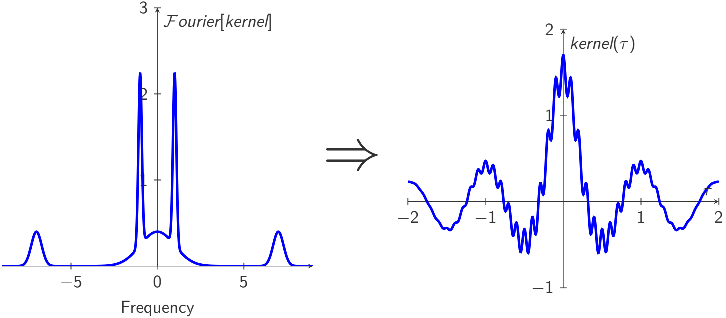
- Introduced in [WA13]; popular in Gaussian Process literature
- It is hard to find good hyperparameters in practice

Spectral Mixture Kernel



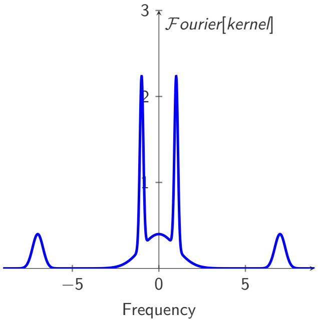
- Introduced in [WA13]; popular in Gaussian Process literature
- It is hard to find good hyperparameters in practice
  - Many Parameters: one mean, variance, & weight per Gaussian

Spectral Mixture Kernel

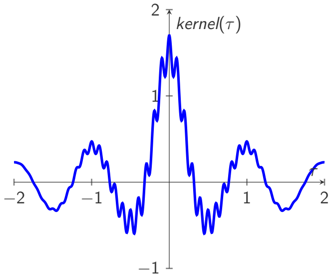


- Introduced in [WA13]; popular in Gaussian Process literature
- It is hard to find good hyperparameters in practice
  - Many Parameters: one mean, variance, & weight per Gaussian
  - Is this because we need more observations?

Spectral Mixture Kernel



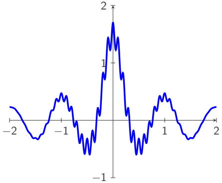
$\Rightarrow$



- Introduced in [WA13]; popular in Gaussian Process literature
- It is hard to find good hyperparameters in practice
  - Many Parameters: one mean, variance, & weight per Gaussian
  - Is this because we need more observations?
  - Is this because we need new algorithms?

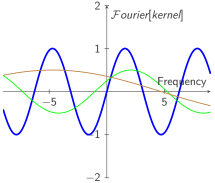
Our Core Contribution

Kernel Ridge Regression



⇒

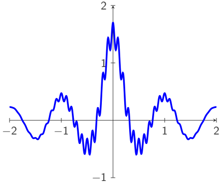
Sparse Fourier Fitting



- Reduce Kernel Learning to a Sparse Fourier Fitting problem

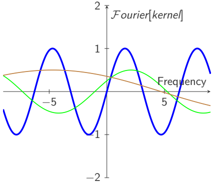
Our Core Contribution

Kernel Ridge Regression



$\Rightarrow$

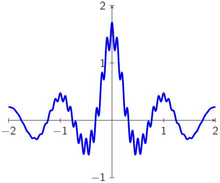
Sparse Fourier Fitting



- Reduce Kernel Learning to a Sparse Fourier Fitting problem
- Number of Observations needed for learning a Spectral Mixture Kernel with Q Gaussians is $\tilde{O}(Q^2)$

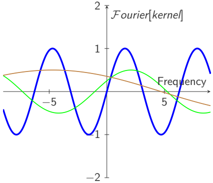
Our Core Contribution

Kernel Ridge Regression



$\Rightarrow$

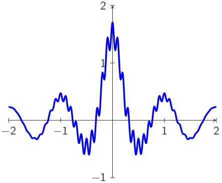
Sparse Fourier Fitting



- Reduce Kernel Learning to a Sparse Fourier Fitting problem
- Number of Observations needed for learning a Spectral Mixture Kernel with Q Gaussians is $\tilde{O}(Q^2)$
- Learning Spectral Mixture kernels is not statistically difficult

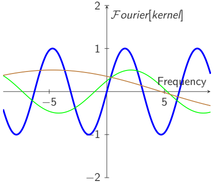
Our Core Contribution

Kernel Ridge Regression



$\Rightarrow$

Sparse Fourier Fitting



- Reduce Kernel Learning to a Sparse Fourier Fitting problem
- Number of Observations needed for learning a Spectral Mixture Kernel with Q Gaussians is $\tilde{O}(Q^2)$
- Learning Spectral Mixture kernels is not statistically difficult
- Techniques generalize to other Stationary Kernels

NYU

Thank You!

Andrew Wilson and Ryan Adams.
Gaussian process kernels for pattern discovery and extrapolation.
In International Conference on Machine Learning, pages 1067–1075, 2013.