UNIVERSITY OF CAMBRIDGE

Integrated Fourier Features for Fast Sparse Variational Gaussian Process Regression

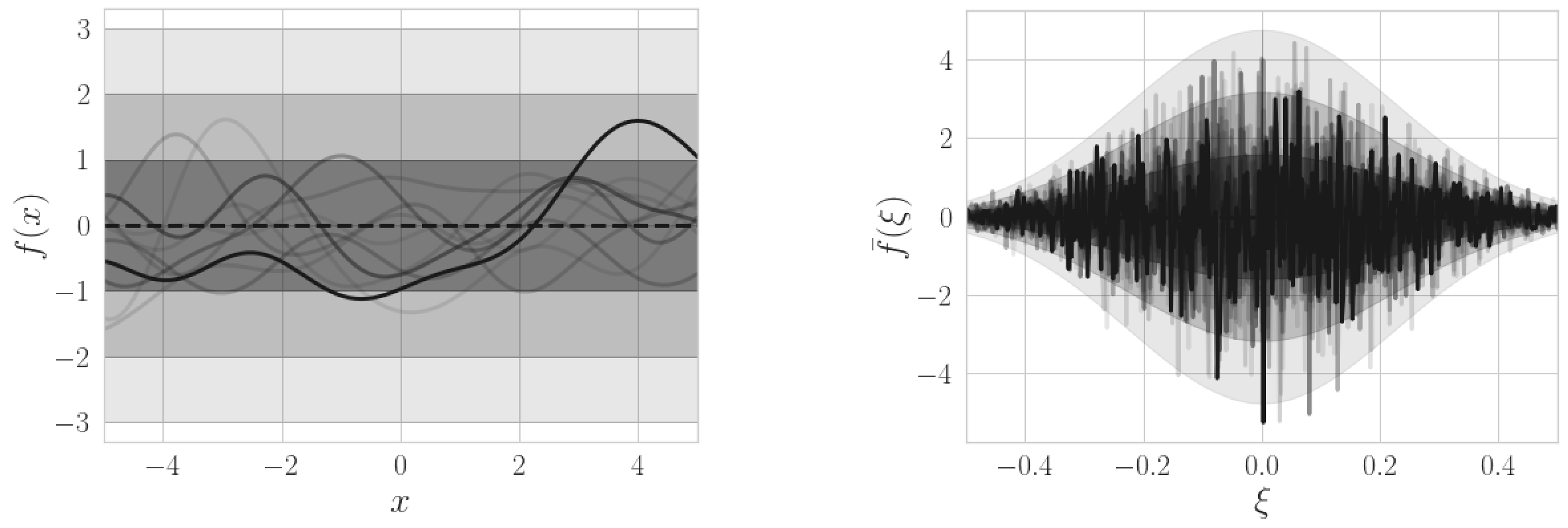
Talay M Cheema (tmc49@cam.ac.uk)

computational &biological learning laboratory

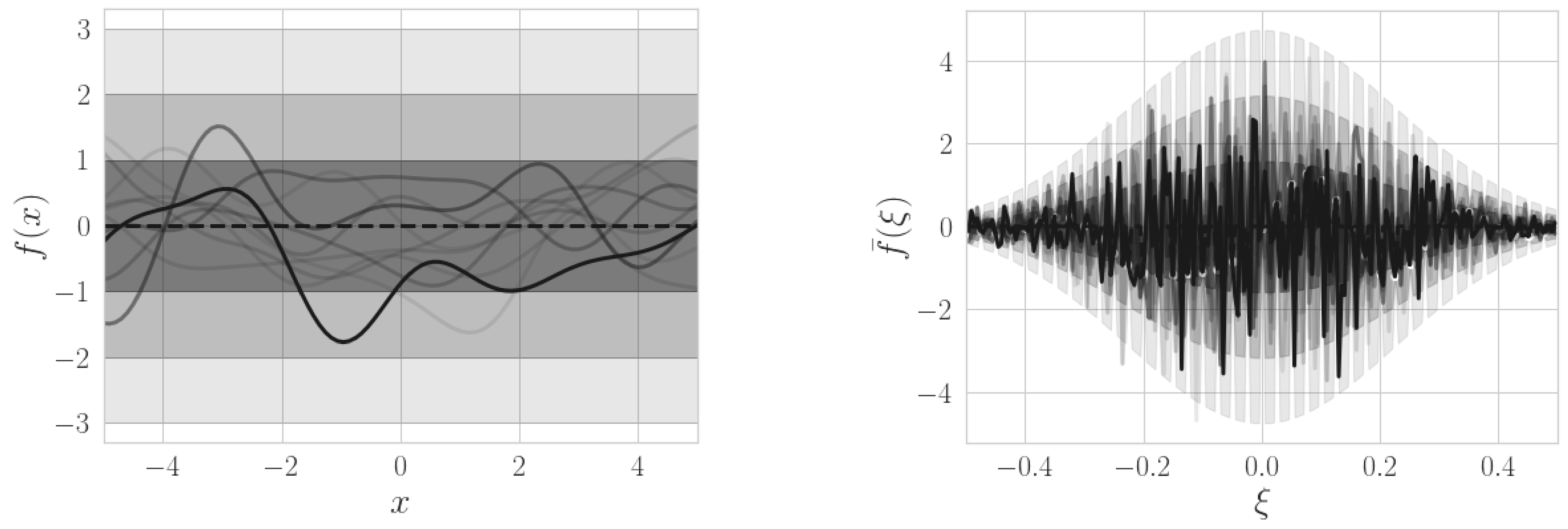
Background

$$y_n = f(x_n) + \rho_n \quad n \in \{1:N\}$$
$$f \sim \mathcal{GP}(0,k), \;\; \rho_n \sim \mathcal{N}(0,\sigma^2)$$

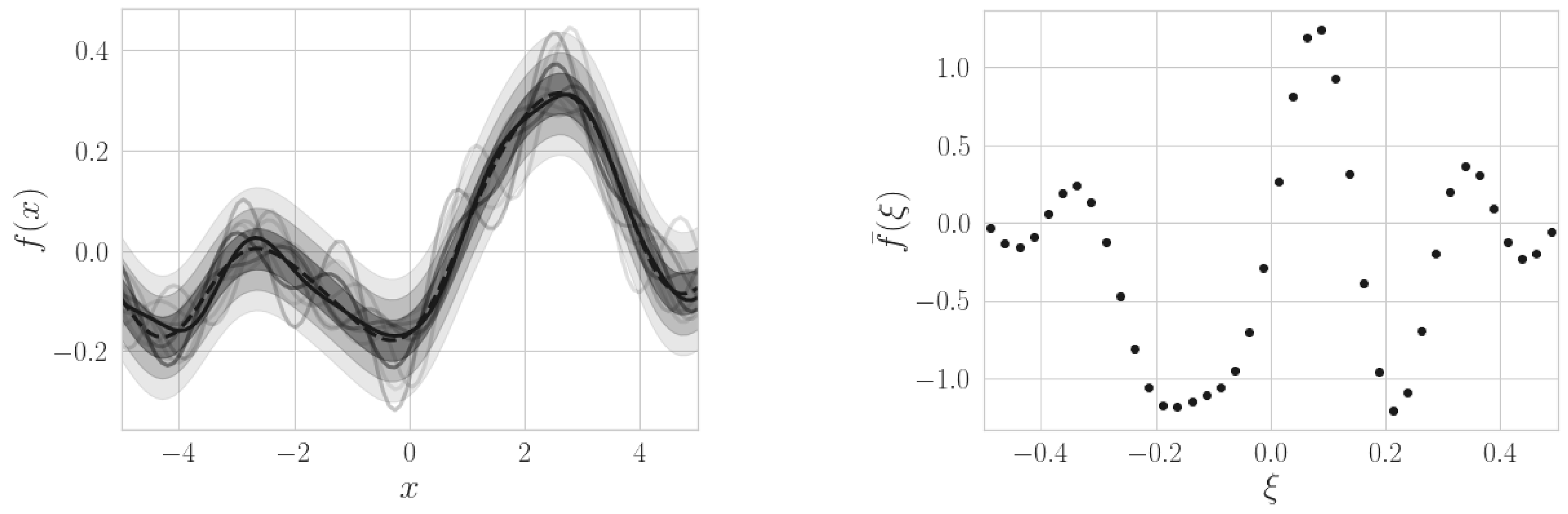
- Exact GP regression has $O(N^3)$ cost
- Variational GP regression introduces $u_m = \int f(x)\phi(x)dx, m \in \{1:M\}$ reducing the cost to $O(NM^2)$.
- Classic sparse GP regression (SGPR) uses $\phi(x) = \delta(z_m)$
- **Fourier features can have $O(M^3)$ cost per optimiser step – moving all the $O(N)$ cost out of the loop.**
- If f is stationary ($k(x,x')$ depends only on $x - x'$) then $\phi(x) = e^{-i2\pi z_m x}$ generates **independent features**, but they have **unbounded variance**.
- **Previously** proposed variational Fourier features work around this using various tricks – but are **limited to only a few choices of** k **and restrictive choices on the approximating frequencies** z_m.



(a) Prior



(b) Conditioning on Fourier features



(c) Conditioning on Integrated Fourier Features

Figure 1: Means in dashed, confidence intervals shaded, samples in solid lines. The Fourier transforms on the right correspond to the functions on the left.

Integrated Fourier Features

The underlying problem is that if the spectral density of k is s, **the Fourier transform of** f **is a white noise process**,

$$\bar{f}(\xi) \sim \mathcal{GP}(0, s(\xi)\delta(\xi - \xi'))$$

so conditioning on M points is ineffective. We propose to instead **sample by local averaging**.

$$u_m = \varepsilon^{-1} \int_{z_m-\varepsilon/2}^{z_m+\varepsilon/2} \frac{\bar{f}(\xi)}{\sqrt{s(\xi)}} d\xi$$

Then the correlation between u_m and f is hard to evaluate. Avoid this by **assuming ε is small**.

$$\mathbb{E}[u_m f(x)] = \varepsilon^{-1} \int_{z_m-\varepsilon/2}^{z_m+\varepsilon/2} \sqrt{s}(\xi) e^{-i2\pi\xi x} d\xi \approx \sqrt{s(z_m)} e^{-i2\pi z_m x}$$

Convergence

Theorem. *Convergence for large N with sub-Gaussian density.* Assume that s has bounded first and second derivatives everywhere, and that we have a tail bound $\int_\xi^\infty \bar{s}(\xi')d\xi' \in O(e^{-\xi})$. Select the inducing features ε apart centred on the origin, that is $z_m = (-(M+1)/2 + m)\varepsilon$, with M even. Let $\varepsilon \in O(M^{-1+a})$ for some $a \in (0,1)$. Then if y is sampled from the generative model. For any $\Delta, \delta > 0$, there exists $M_0, \alpha > 0$ such that for $M \geq M_0$

$$\Pr[D_{KL}(q(f) \| p(f|y))/N > \Delta/N] \leq \delta \iff M \leq \left(\frac{\alpha}{\Delta\delta} N\right)^{\frac{1}{2-3a}}$$

Since we can take any $a \in (0,1)$, we can optimise the rate by taking $a \to 0$, which leads to $M \in O(\sqrt{N})$.

- $D_{KL}(q(f)\|p(f|y))$ is the KL divergence from the approximate posterior to the true posterior.
- Convergence is dominated by the need to make ε small.
- Generalises to heavier tailed spectral densities and higher dimensions.
- However, M **goes up exponentially in dimension**.

Computational cost

Use $\bar{K}_{zz} = \mathbb{E}[uu^*] = \varepsilon^{-1}I$, $C_{zx} = \mathbb{E}[uf(x)^*]$, and S for a diagonal matrix of spectral densities.

$$\mathcal{F}(\mu_u, \Sigma_u) = \log \mathcal{N}(y|0,\; C_{zx}^*\bar{K}_{zz}^{-1}C_{zx} + \sigma^2 I) - \frac{1}{2}\sigma^{-2}\mathrm{tr}(K_{xx} - C_{zx}^*\bar{K}_{zz}^{-1}C_{zx})$$

Rearranging using matrix determinant lemma/matrix inversion lemma yields that the dominant cost relates to

$$\varepsilon^{-1}S^{-1} + \sigma^{-2}S^{-1/2}C_{zx}C_{zx}^*S^{-1/2}$$

- $\sigma^{-2}S^{-1/2}C_{zx}C_{zx}^*S^{-1/2}$ costs $O(NM^2)$ to form–but doesn't depend on the hyperparameters.
- The $O(N)$ cost is taken out of the loop if the frequencies are kept fixed.
- When N is large, this is much faster than SGPR.
- Theory suggests making ε small is the limiting factor in convergence – but in practice, ε around the inverse data diameter is sufficient.
- Flexible choice of z_m opens the way to better performance in higher dimensions.

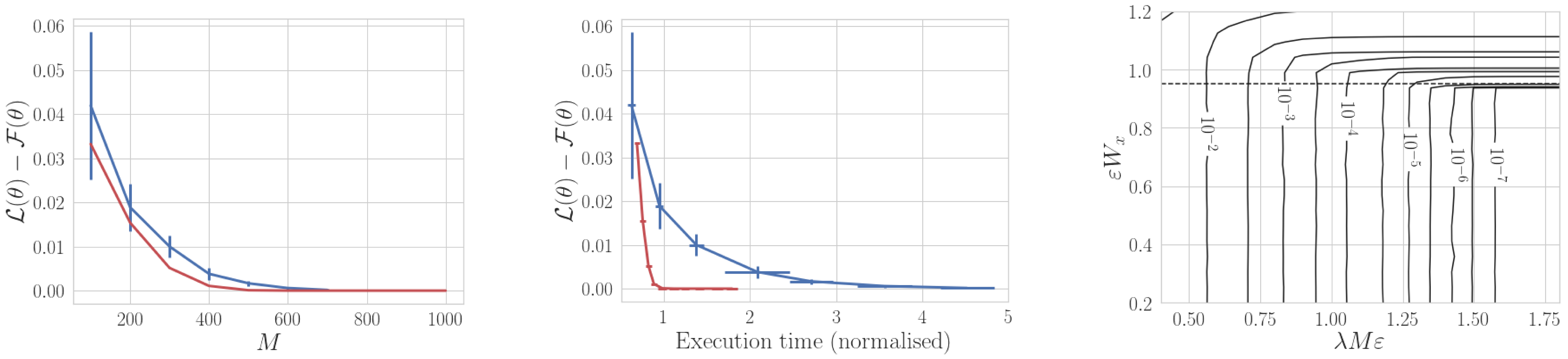


Figure 2: Synthetic plots. Orange is IFF and blue is SGPR initialised with K means. $\mathcal{L}$ is the log marginal likelihood, $\mathcal{F}(\theta)$ is the variational lower bound, each at learnt hyperparmaeters θ. The right most plot is $\mathcal{L} - \mathcal{F}$ for different settings of lengthscale λ and data diameter W_x.