

# Spectral Methods in Gaussian Modelling

## Topic 3: Variational Inference

James Requiema and Wessel Bruinsma

University of Cambridge and Invenia Labs

20 December 2019

- RFFs: approximate kernel and perform inference.

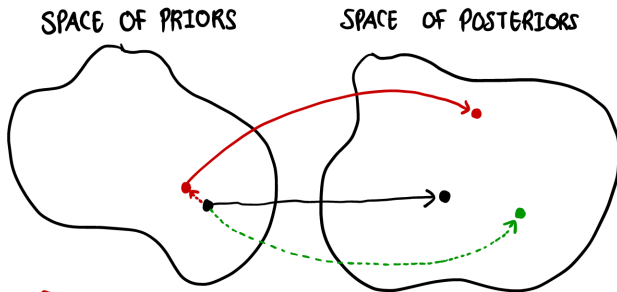
- RFFs: approximate kernel and perform inference.
- ⇒ Exact inference in approximate model.

- RFFs: approximate kernel and perform inference.
- ⇒ Exact inference in approximate model.
- Can affect posterior in unforeseen/undesired ways.

- RFFs: approximate kernel and perform inference.
- ⇒ Exact inference in approximate model.
- Can affect posterior in unforeseen/undesired ways.
  - Overfitting

- RFFs: approximate kernel and perform inference.
- ⇒ **Exact** inference in **approximate** model.
  - Can affect posterior in unforeseen/undesired ways.
  - Overfitting
- Should perform **approximate** inference in **exact** model.

- RFFs: approximate kernel and perform inference.
- ⇒ **Exact** inference in **approximate** model.
  - Can affect posterior in unforeseen/undesired ways.
  - Overfitting
- Should perform **approximate** inference in **exact** model.
  - No overfitting!



EXACT INFERENCE IN APPROXIMATE MODEL

APPROXIMATE INFERENCE IN EXACT MODEL



- Goal: compute  $p(f | \mathcal{D})$ .

- Goal: compute  $p(f | \mathcal{D})$ .
- Introduce **approximate posterior**  $q(f)$ :

- Goal: compute  $p(f | \mathcal{D})$ .
- Introduce **approximate posterior**  $q(f)$ :

$$\theta^* = \arg \min_{q \in \mathcal{Q}} D(q(f) \| p(f | \mathcal{D})).$$

- Goal: compute  $p(f | \mathcal{D})$ .
- Introduce **approximate posterior**  $q(f)$ :

$$\theta^* = \arg \min_{q \in \mathcal{Q}} D(q(f) \| p(f | \mathcal{D})).$$

- Often  $D =$  KL divergence: **variational inference**.

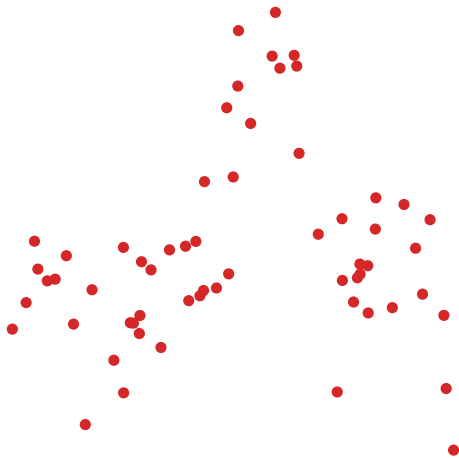
- Goal: compute  $p(f | \mathcal{D})$ .
- Introduce **approximate posterior**  $q(f)$ :

$$\theta^* = \arg \min_{q \in \mathcal{Q}} D(q(f) \| p(f | \mathcal{D})).$$

- Often  $D =$  KL divergence: **variational inference**.
- How to construct  $q(f)$ ?

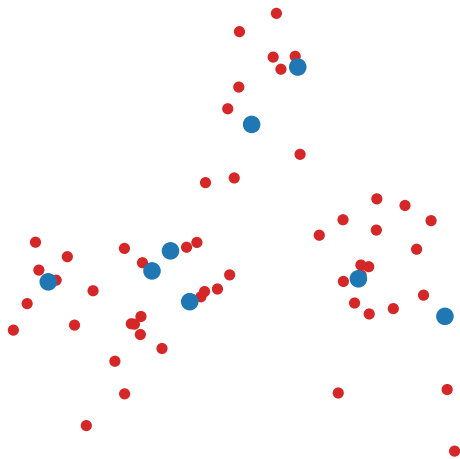
## Approximating the Posterior (2)

5/21



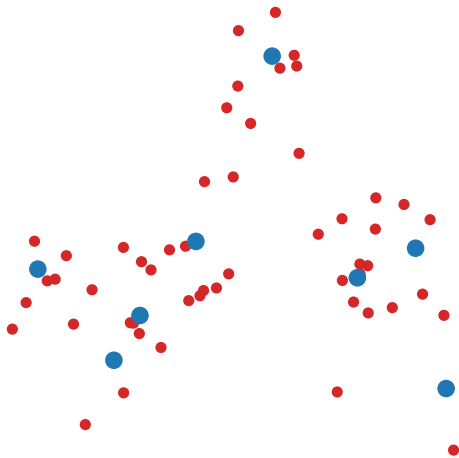
## Approximating the Posterior (2)

5/21



## Approximating the Posterior (2)

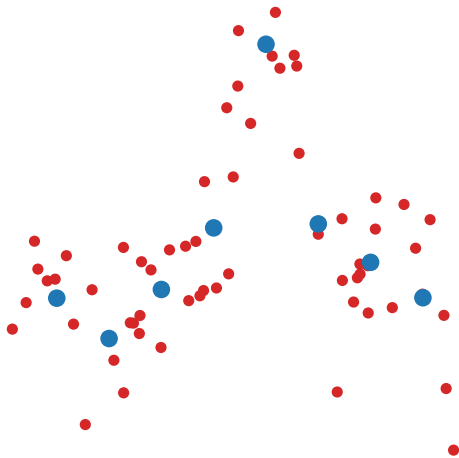
5/21





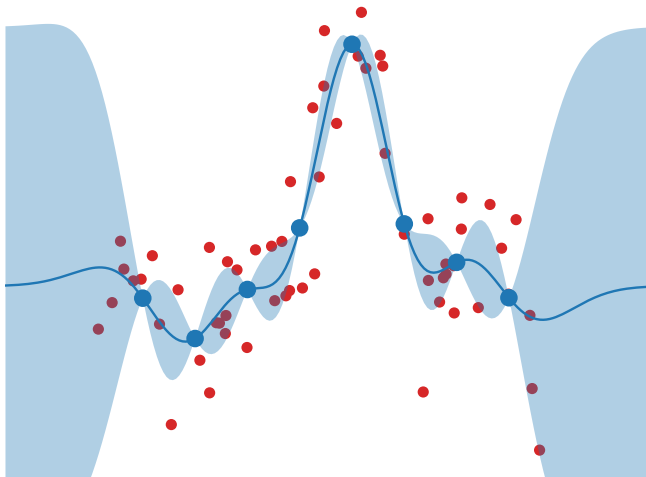
## Approximating the Posterior (2)

5/21



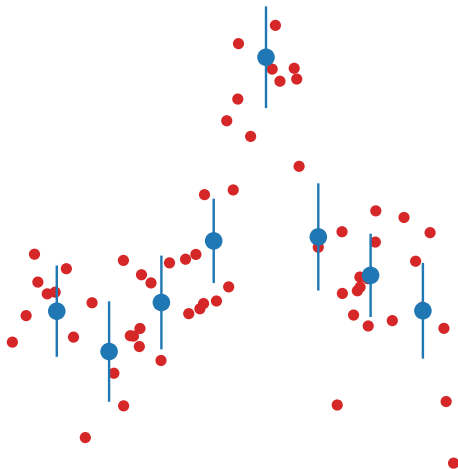
## Approximating the Posterior (2)

5/21



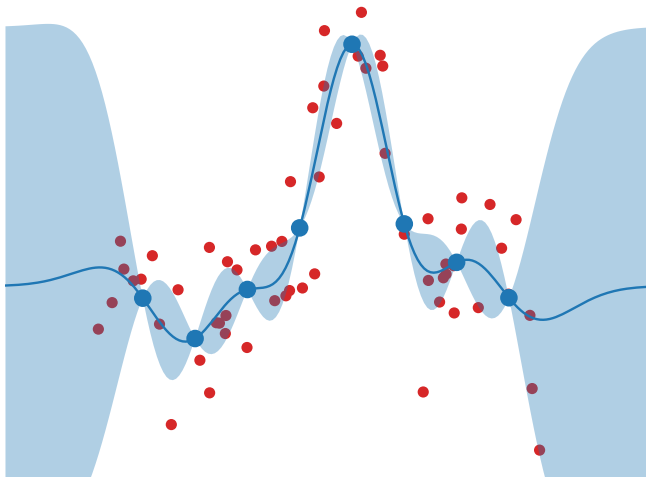
## Approximating the Posterior (2)

5/21



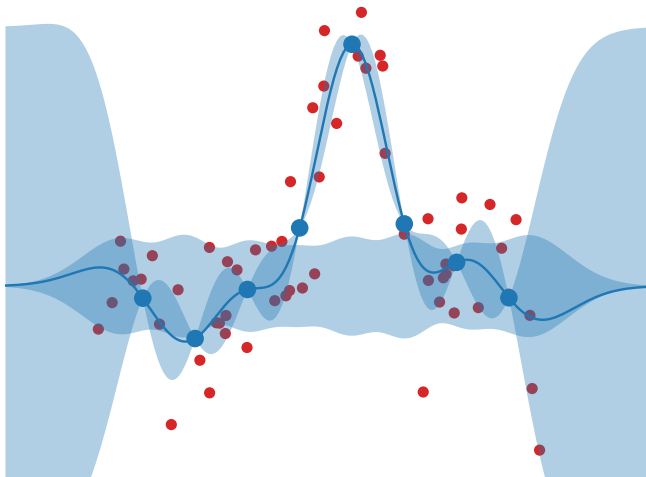
## Approximating the Posterior (2)

5/21



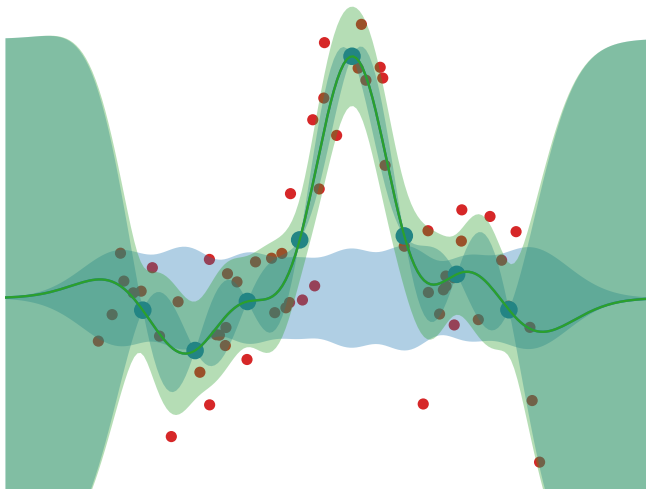
## Approximating the Posterior (2)

5/21



## Approximating the Posterior (2)

5/21



- Introduce **inducing points**:  $u = (f(t_{u,1}), \dots, f(t_{u,M}))$ .

- Introduce **inducing points**:  $u = (f(t_{u,1}), \dots, f(t_{u,M}))$ .
- Assume distribution  $q(u)$ .



- Introduce **inducing points**:  $u = (f(t_{u,1}), \dots, f(t_{u,M}))$ .
- Assume distribution  $q(u)$ .
- Inducing point approximation (Titsias, 2009):

$$p(f | \mathcal{D}) \approx q(f) \equiv \int p(f | u)q(u) du$$

- Introduce **inducing points**:  $u = (f(t_{u,1}), \dots, f(t_{u,M}))$ .
- Assume distribution  $q(u)$ .
- Inducing point approximation (Titsias, 2009):

$$p(f | \mathcal{D}) \approx q(f) \equiv \int p(f | u)q(u) du,$$
$$q^*(u) = \arg \min_{q(u)} D_{\text{KL}}(q(f) \| p(f | \mathcal{D}))$$

- Introduce **inducing points**:  $u = (f(t_{u,1}), \dots, f(t_{u,M}))$ .
- Assume distribution  $q(u)$ .
- Inducing point approximation (Titsias, 2009):

$$p(f | \mathcal{D}) \approx q(f) \equiv \int p(f | u)q(u) du,$$
$$q^*(u) = \arg \min_{q(u)} D_{\text{KL}}(q(f) \| p(f | \mathcal{D}))$$
$$\propto p(u) \exp \int p(f | u) \log p(\mathcal{D} | f) df.$$

- Introduce **inducing points**:  $u = (f(t_{u,1}), \dots, f(t_{u,M}))$ .
- Assume distribution  $q(u)$ .
- Inducing point approximation (Titsias, 2009):

$$p(f | \mathcal{D}) \approx q(f) \equiv \int p(f | u)q(u) du,$$

$$q^*(u) = \arg \min_{q(u)} D_{\text{KL}}(q(f) \| p(f | \mathcal{D}))$$

$$\propto p(u) \exp \int p(f | u) \log p(\mathcal{D} | f) df.$$

- Complexities:
 

	Time	Memory
Full posterior	$O(N^3)$	$O(N^2)$
Inducing points		

- Introduce **inducing points**:  $u = (f(t_{u,1}), \dots, f(t_{u,M}))$ .
- Assume distribution  $q(u)$ .
- Inducing point approximation (Titsias, 2009):

$$p(f | \mathcal{D}) \approx q(f) \equiv \int p(f | u)q(u) du,$$

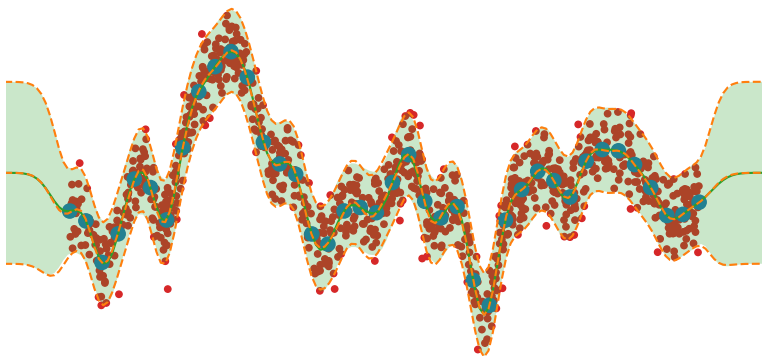
$$q^*(u) = \arg \min_{q(u)} D_{\text{KL}}(q(f) \| p(f | \mathcal{D}))$$

$$\propto p(u) \exp \int p(f | u) \log p(\mathcal{D} | f) df.$$

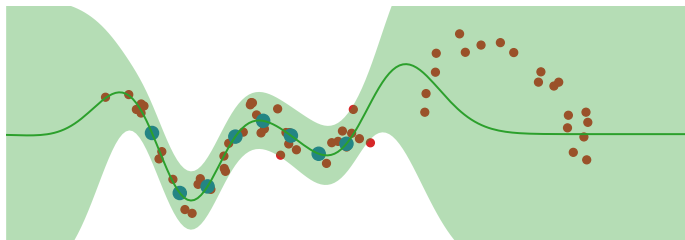
- Complexities:
 

	Time	Memory
Full posterior	$O(N^3)$	$O(N^2)$
Inducing points	$O(NM^2)$	$O(NM)$

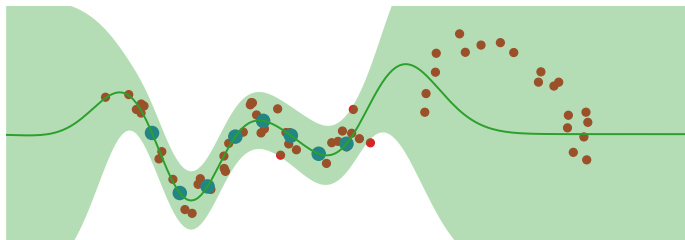
- $N = 1000$  and  $M = 40$ ;  $25\times$  compression!



- Inducing points: local in time/space



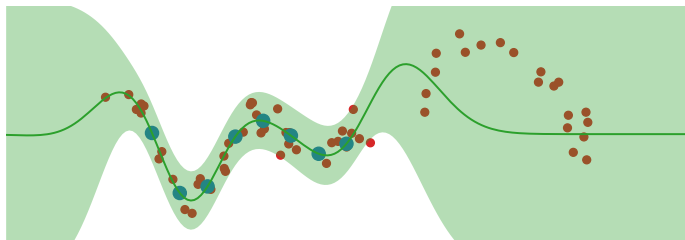
- Inducing points: local in time/space



- Need  $M \propto N$

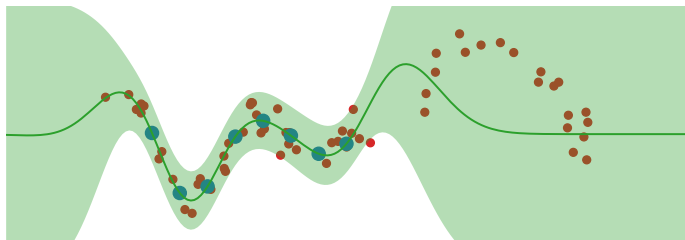


- Inducing points: local in time/space



- Need  $M \propto N$ : hidden  $O(N^3)$  scaling!

- Inducing points: local in time/space



- Need  $M \propto N$ : hidden  $O(N^3)$  scaling!
- SSA: local in spectrum

Inducing points:

- + Appealing approximative construction
- Hidden  $O(N^3)$  scaling

SSA:

- + Representative power
- Overfitting

Inducing points:

- + Appealing approximative construction
- Hidden  $O(N^3)$  scaling

SSA:

- + Representative power
- Overfitting

Best of both worlds?

Inducing points:

- + Appealing approximative construction
- Hidden  $O(N^3)$  scaling

SSA:

- + Representative power
- Overfitting

Best of both worlds?

Yes: Variational Fourier Features (VFFs)!  
(Hensman et al., 2016)

- Extension of inducing point method (Lázaro-Gredilla and Vidal, 2009)

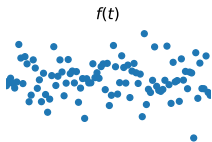
- Extension of inducing point method (Lázaro-Gredilla and Vidal, 2009)
- Introduce pseudo-observations for a linear transform of  $f$ :

$$g(\xi) | f = \int_{-\infty}^{\infty} h(\xi, t) f(t) dt, \quad u = (g(\xi_{u,1}), \dots, g(\xi_{u,M})).$$

- Extension of inducing point method (Lázaro-Gredilla and Vidal, 2009)
- Introduce pseudo-observations for a linear transform of  $f$ :

$$g(\xi) | f = \int_{-\infty}^{\infty} h(\xi, t) f(t) dt, \quad u = (g(\xi_{u,1}), \dots, g(\xi_{u,M})).$$

- In some cases, necessary:

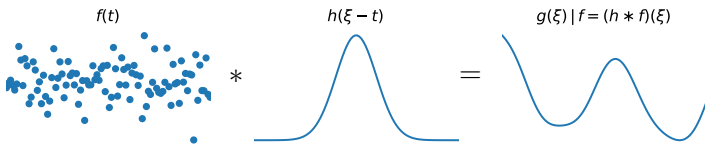




- Extension of inducing point method (Lázaro-Gredilla and Vidal, 2009)
- Introduce pseudo-observations for a **linear transform of  $f$** :

$$g(\xi) | f = \int_{-\infty}^{\infty} h(\xi, t) f(t) dt, \quad u = (g(\xi_{u,1}), \dots, g(\xi_{u,M})).$$

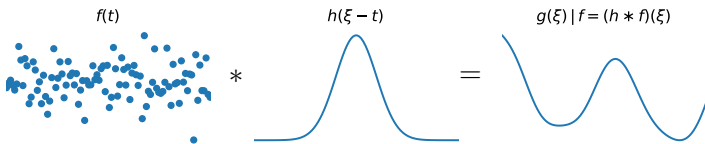
- In some cases, **necessary**:



- Extension of inducing point method (Lázaro-Gredilla and Vidal, 2009)
- Introduce pseudo-observations for a **linear transform of  $f$** :

$$g(\xi) | f = \int_{-\infty}^{\infty} h(\xi, t) f(t) dt, \quad u = (g(\xi_{u,1}), \dots, g(\xi_{u,M})).$$

- In some cases, **necessary**:



- In other cases, **improve**  $q(f)$ .

- Predictive mean in SSA:

$$\hat{f}^{(\text{SSA})}(t) = \sum_{i=1}^M \alpha_i \cos(2\pi\xi_i t) + \sum_{i=1}^M \beta_i \sin(2\pi\xi_i t).$$

- Predictive mean in SSA:

$$\hat{f}^{(\text{SSA})}(t) = \sum_{i=1}^M \alpha_i \cos(2\pi\xi_i t) + \sum_{i=1}^M \beta_i \sin(2\pi\xi_i t).$$

- Predictive mean for inter-domain inducing points:

$$\hat{f}^{(\text{IDIP})}(t) = \sum_{i=1}^M \alpha_i \int_{-\infty}^{\infty} k(t, \tau) h(\xi_{u,i}, \tau) d\tau.$$

- Predictive mean in SSA:

$$\hat{f}^{(\text{SSA})}(t) = \sum_{i=1}^M \alpha_i \cos(2\pi\xi_i t) + \sum_{i=1}^M \beta_i \sin(2\pi\xi_i t).$$

- Predictive mean for inter-domain inducing points:

$$\hat{f}^{(\text{IDIP})}(t) = \sum_{i=1}^M \alpha_i \int_{-\infty}^{\infty} k(t, \tau) h(\xi_{u,i}, \tau) d\tau.$$

- VFFs: engineer  $h$  such that  $\hat{f}^{(\text{VFF})}$  is also a Fourier expansion.

## Variational Fourier Features (2)

12/21

Attempt 1:  $g(\xi) | f = \int_{-\infty}^{\infty} e^{-2\pi i \xi t} f(t) dt.$

Attempt 1:  $g(\xi) | f = \int_{-\infty}^{\infty} e^{-2\pi i \xi t} f(t) dt.$

- Inducing points become **inducing frequencies**.

Attempt 1:  $g(\xi) | f = \int_{-\infty}^{\infty} e^{-2\pi i \xi t} f(t) dt.$

- Inducing points become **inducing frequencies**.

- $\hat{f}^{(\text{IDIP})}(t) = \sum_{i=1}^M \alpha_i \hat{k}(\xi_{u,i}) e^{-2\pi i \xi_{u,i} t}.$



Attempt 1:  $g(\xi) | f = \int_{-\infty}^{\infty} e^{-2\pi i \xi t} f(t) dt.$

- Inducing points become **inducing frequencies**.

- $\hat{f}^{(\text{IDIP})}(t) = \sum_{i=1}^M \alpha_i \hat{k}(\xi_{u,i}) e^{-2\pi i \xi_{u,i} t}.$

- But  $g$  is white noise...

Attempt 1:  $g(\xi) | f = \int_{-\infty}^{\infty} e^{-2\pi i \xi t} f(t) dt.$

- Inducing points become **inducing frequencies**.

- $\hat{f}^{(\text{IDIP})}(t) = \sum_{i=1}^M \alpha_i \hat{k}(\xi_{u,i}) e^{-2\pi i \xi_{u,i} t}.$

- But  $g$  is white noise...

Attempt 2:  $g(\xi) | f = \int_a^b e^{-2\pi i \xi (t-a)} f(t) dt.$

Attempt 1:  $g(\xi) | f = \int_{-\infty}^{\infty} e^{-2\pi i \xi t} f(t) dt.$

- Inducing points become **inducing frequencies**.

- $\hat{f}^{(\text{IDIP})}(t) = \sum_{i=1}^M \alpha_i \hat{k}(\xi_{u,i}) e^{-2\pi i \xi_{u,i} t}.$

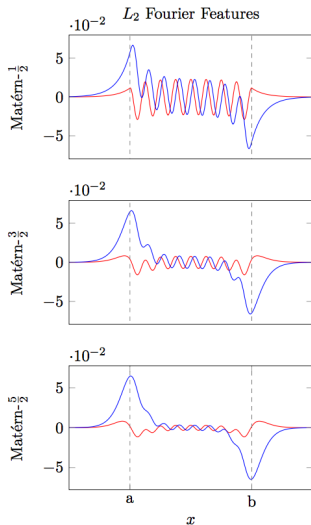
- But  $g$  is white noise...

Attempt 2:  $g(\xi) | f = \int_a^b e^{-2\pi i \xi(t-a)} f(t) dt.$  ( $L^2[a, b]$ -VFFs)

- Works, but edge effects near boundaries.

# Variational Fourier Features (3)

13/21



(Figure taken from Hensman et al. (2016).)

- Use tools from RKHS theory to eliminate edge effects.

- Use tools from RKHS theory to eliminate edge effects.
- Mercer's Theorem:

$$k(t, t') = \sum_{i=1}^{\infty} \lambda_i \phi_i(t) \phi_i^*(t')$$

where  $(\phi_i)_{i=1}^{\infty}$  and  $(\lambda_i)_{i=1}^{\infty}$  e.f.'s and e.v.'s of

$$T_k f = t \mapsto \langle k(t, \cdot), f \rangle, \quad \langle f, g \rangle = \int_a^b f(t) g^*(t) d\mu(t).$$

- Use tools from RKHS theory to eliminate edge effects.
- Mercer's Theorem:

$$k(t, t') = \sum_{i=1}^{\infty} \lambda_i \phi_i(t) \phi_i^*(t')$$

where  $(\phi_i)_{i=1}^{\infty}$  and  $(\lambda_i)_{i=1}^{\infty}$  e.f.'s and e.v.'s of

$$T_k f = t \mapsto \langle k(t, \cdot), f \rangle, \quad \langle f, g \rangle = \int_a^b f(t) g^*(t) d\mu(t).$$

- Why?

- Use tools from RKHS theory to eliminate edge effects.
- Mercer's Theorem:

$$k(t, t') = \sum_{i=1}^{\infty} \lambda_i \phi_i(t) \phi_i^*(t')$$

where  $(\phi_i)_{i=1}^{\infty}$  and  $(\lambda_i)_{i=1}^{\infty}$  e.f.'s and e.v.'s of

$$T_k f = t \mapsto \langle k(t, \cdot), f \rangle, \quad \langle f, g \rangle = \int_a^b f(t) g^*(t) d\mu(t).$$

- Why?

$$\hat{f}^{(\text{IDIP})}(t) = \sum_{i=1}^M \alpha_i T_k(h(\xi_{u,i}, \cdot))(t).$$



- Desire  $T_k(h(\xi, \cdot))(t) = \psi(\xi, t) = e^{-2\pi i \xi t}$ .

## Variational Fourier Features (5)

15/21

- Desire  $T_k(h(\xi, \cdot))(t) = \psi(\xi, t) = e^{-2\pi i \xi t}$ .
- If  $\psi(\xi, \cdot) \in \text{span}\{\phi_i\}_{i=1}^{\infty}$ , then

$$\psi(\xi, t) = \sum_{i=1}^{\infty} \phi_i(t) \langle \psi(\xi, \cdot), \phi_i \rangle.$$

- Desire  $T_k(h(\xi, \cdot))(t) = \psi(\xi, t) = e^{-2\pi i \xi t}$ .
- If  $\psi(\xi, \cdot) \in \text{span}\{\phi_i\}_{i=1}^{\infty}$ , then

$$\psi(\xi, t) = \sum_{i=1}^{\infty} \phi_i(t) \langle \psi(\xi, \cdot), \phi_i \rangle.$$

- Solution:

$$h(\xi, t) = \sum_{i=1}^{\infty} \frac{1}{\lambda_i} \phi_i(t) \langle \psi(\xi, \cdot), \phi_i \rangle$$

- Desire  $T_k(h(\xi, \cdot))(t) = \psi(\xi, t) = e^{-2\pi i \xi t}$ .
- If  $\psi(\xi, \cdot) \in \text{span}\{\phi_i\}_{i=1}^{\infty}$ , then

$$\psi(\xi, t) = \sum_{i=1}^{\infty} \phi_i(t) \langle \psi(\xi, \cdot), \phi_i \rangle.$$

- Solution:

$$h(\xi, t) = \sum_{i=1}^{\infty} \frac{1}{\lambda_i} \phi_i(t) \langle \psi(\xi, \cdot), \phi_i \rangle$$
$$\implies T_k(h(\xi, \cdot))(t) = \sum_{i=1}^{\infty} \frac{1}{\lambda_i} T_k(\phi_i)(t) \langle \psi(\xi, \cdot), \phi_i \rangle$$

- Desire  $T_k(h(\xi, \cdot))(t) = \psi(\xi, t) = e^{-2\pi i \xi t}$ .
- If  $\psi(\xi, \cdot) \in \text{span}\{\phi_i\}_{i=1}^{\infty}$ , then

$$\psi(\xi, t) = \sum_{i=1}^{\infty} \phi_i(t) \langle \psi(\xi, \cdot), \phi_i \rangle.$$

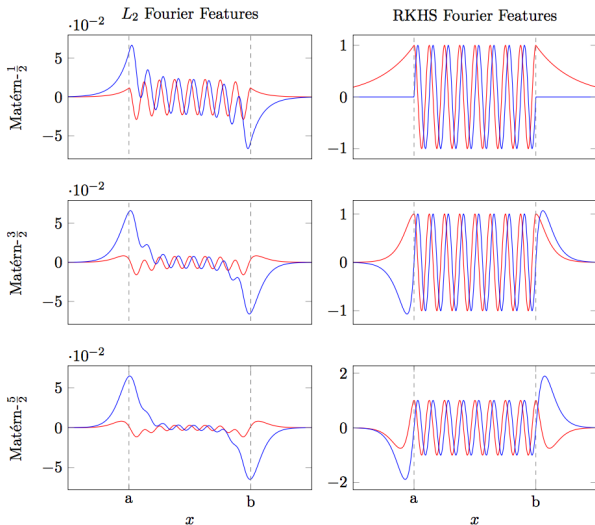
- Solution:

$$\begin{aligned} h(\xi, t) &= \sum_{i=1}^{\infty} \frac{1}{\lambda_i} \phi_i(t) \langle \psi(\xi, \cdot), \phi_i \rangle \\ \implies T_k(h(\xi, \cdot))(t) &= \sum_{i=1}^{\infty} \frac{1}{\lambda_i} T_k(\phi_i)(t) \langle \psi(\xi, \cdot), \phi_i \rangle \\ &= \sum_{i=1}^{\infty} \phi_i(t) \langle \psi(\xi, \cdot), \phi_i \rangle = \psi(\xi, t). \end{aligned}$$

- Works for Matérn kernels of half-integer order!
- ⇒ VFFs (RKHS-VFFs)

# Variational Fourier Features (7)

17/21



(Figure taken from Hensman et al. (2016).)

- RFFs **simple**: amenable to convergence analysis.



- RFFs **simple**: amenable to convergence analysis.
- Inducing points **more complex**: convergence analysis hard.

- RFFs **simple**: amenable to convergence analysis.
- Inducing points **more complex**: convergence analysis hard.
- Recent result by Burt et al. (2018):

## Theorem (Burt et al. (2018))

Fix  $\varepsilon > 0$  and  $\delta > 0$ . Let  $(t_i)_{i=1}^{\infty}$  be sampled i.i.d. from  $\mathcal{N}(0, \alpha)$ , let  $k$  be an exponentiated-quadratic kernel, and let  $\mu$  have density  $\mathcal{N}(0, \beta)$  with  $\beta > 2\alpha$ . Then there are  $\tilde{N}$  and  $\tilde{C}$  such that, for all  $N > \tilde{N}$ , the inter-domain point method with  $M = \tilde{C} \log N$  eigenfunction inducing features achieves  $D_{\text{KL}}(q(f) \parallel p(f | \mathcal{D})) \leq \varepsilon$  with probability at least  $1 - \delta$ .

# Rates of Convergence: Eigenfunction Inducing Features

19/21

- Eigenfunction inducing features:

$$u_i | f = \frac{1}{\sqrt{\lambda_i}} \int f(t) \phi_i(t) d\mu(t).$$

# Rates of Convergence: Eigenfunction Inducing Features

19/21

- Eigenfunction inducing features:

$$u_i | f = \frac{1}{\sqrt{\lambda_i}} \int f(t) \phi_i(t) d\mu(t).$$

- Nice behaviour:

$$\mathbb{E}[u_i u_j] = 1 \text{ if } i = j \text{ else } 0, \quad \mathbb{E}[f(t_i) u_j] = \sqrt{\lambda_j} \phi_j(t_i).$$

# Rates of Convergence: Eigenfunction Inducing Features

19/21

- Eigenfunction inducing features:

$$u_i | f = \frac{1}{\sqrt{\lambda_i}} \int f(t) \phi_i(t) d\mu(t).$$

- Nice behaviour:

$$\mathbb{E}[u_i u_j] = 1 \text{ if } i = j \text{ else } 0, \quad \mathbb{E}[f(t_i) u_j] = \sqrt{\lambda_j} \phi_j(t_i).$$

- Key quantity:

$$c = \text{tr}(K_{ff} - K_{fu} K_{uu}^{-1} K_{fu}),$$

$$(K_{ff})_{ij} = \mathbb{E}[f(t_i) f(t_j)], \quad (K_{fu})_{ij} = \mathbb{E}[f(t_i) u_j],$$

$$(K_{uu})_{ij} = \mathbb{E}[u_i u_j].$$

- Can compute e.f.'s and e.v.'s for EQ kernel.

- Can compute e.f.'s and e.v.'s for EQ kernel.
- Key inequality:

$$D_{\text{KL}}(q(f) \parallel p(f \mid \mathcal{D})) \leq \frac{c}{2\sigma^2} \left( 1 + \frac{\|y\|^2}{\sigma^2 + c} \right).$$

- Can compute e.f.'s and e.v.'s for EQ kernel.
- Key inequality:

$$D_{\text{KL}}(q(f) \parallel p(f | \mathcal{D})) \leq \frac{c}{2\sigma^2} \left( 1 + \frac{\|y\|^2}{\sigma^2 + c} \right).$$

- Bound follows from application of Chebyshev's to

$$\frac{1}{N}c = \sum_{m=M+1}^{\infty} \lambda_m \left[ \frac{1}{N} \sum_{i=1}^N \phi_m^2(t_i) \right]$$

combined with  $\sum_{m=M+1}^{\infty} \lambda_m = O(A^M)$  for some  $A \in (0, 1)$ .



- RFFs: **exact** inference in **approximate** model.

- RFFs: **exact** inference in **approximate** model.
  - + Good representative power

- RFFs: **exact** inference in **approximate** model.
  - + Good representative power
- Inducing points: **approximate** inference in **exact** model.

- RFFs: **exact** inference in **approximate** model.
  - + Good representative power
- Inducing points: **approximate** inference in **exact** model.
  - + No overfitting

- RFFs: **exact** inference in **approximate** model.
  - + Good representative power
- Inducing points: **approximate** inference in **exact** model.
  - + No overfitting
- VFFs = inducing points + representative power of RFFs.

- RFFs: **exact** inference in **approximate** model.
  - + Good representative power
- Inducing points: **approximate** inference in **exact** model.
  - + No overfitting
- VFFs = inducing points + representative power of RFFs.
- Can design inducing point methods amenable to convergence analysis.

# Appendix

## References

- Burt, D., Rasmussen, C. E., & van der Wilk, M. (2018). Explicit rates of convergence for sparse variational inference in Gaussian process regression. In *Symposium on advances in approximate Bayesian inference 1*.
- Hensman, J., Durrande, N., & Solin, A. (2016). Variational fourier features for Gaussian processes. *arXiv preprint arXiv:1611.06740*. eprint: <https://arxiv.org/abs/1611.06740>
- Lázaro-Gredilla, M., & Vidal, A. F. (2009). Inter-domain Gaussian processes for sparse inference using inducing features. *Advances in Neural Information Processing Systems*, 22, 1087–1095.
- Titsias, M. K. (2009). Variational learning of inducing variables in sparse Gaussian processes. *Artificial Intelligence and Statistics*, 12, 567–574.