Spectral Methods in Gaussian Modelling

Topic 3: Variational Inference

James Requiema and Wessel Bruinsma

University of Cambridge and Invenia Labs

18 January 2019

Approximate Inference

- RFFs: approximate kernel and perform inference.

- RFFs: approximate kernel and perform inference.

$\Rightarrow$ **Exact** inference in **approximate** model.

Approximate Inference

- RFFs: approximate kernel and perform inference.

$\Rightarrow$ **Exact** inference in **approximate** model.

  - Can affect posterior in unforeseen/undesired ways.

Approximate Inference

- RFFs: approximate kernel and perform inference.

⇒ **Exact** inference in **approximate** model.

  - Can affect posterior in unforeseen/undesired ways.
  - Overfitting

Approximate Inference

- RFFs: approximate kernel and perform inference.

$\Rightarrow$ **Exact** inference in **approximate** model.

  - Can affect posterior in unforeseen/undesired ways.
  - Overfitting

- Should perform **approximate** inference in **exact** model.

Approximate Inference

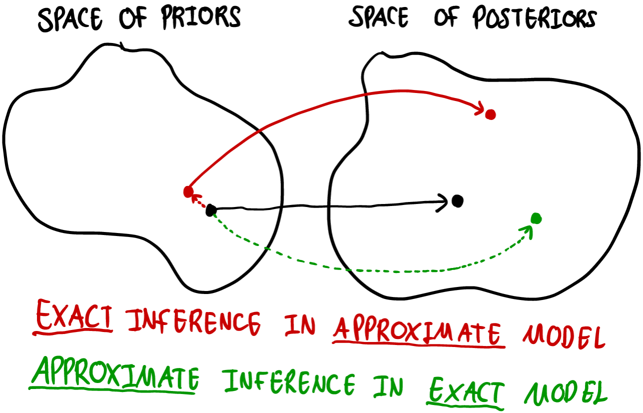
- RFFs: approximate kernel and perform inference.

$\Rightarrow$ **Exact** inference in **approximate** model.

  - Can affect posterior in unforeseen/undesired ways.
  - Overfitting

- Should perform **approximate** inference in **exact** model.
  - No overfitting!

Approximate Inference (2)



Approximating the Posterior

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

Approximating the Posterior

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

Approximating the Posterior

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

$$\theta^* = \operatorname*{arg\,min}_{q \in \mathcal{Q}} \mathrm{D}(q(f) \,\|\, p(f \mid \mathcal{D})).$$

Approximating the Posterior

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

$$\theta^* = \operatorname*{arg\,min}_{q \in \mathcal{Q}} \mathrm{D}(q(f) \,\|\, p(f \mid \mathcal{D})).$$

- Often $\mathrm{D} = \mathrm{KL}$ divergence: **variational inference**.

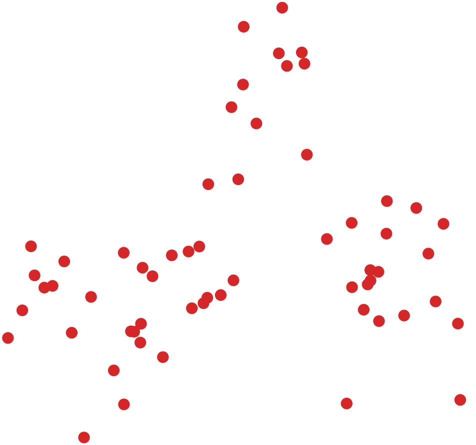
Approximating the Posterior

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

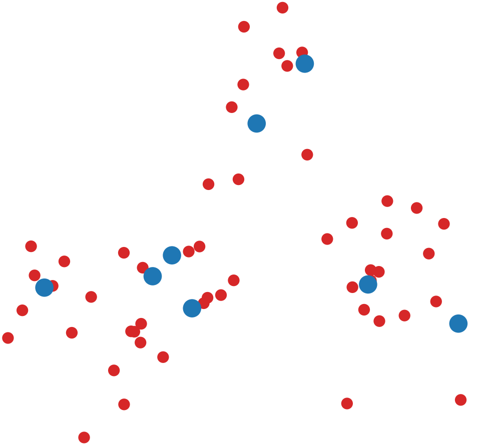
$$\theta^* = \operatorname*{arg\,min}_{q \in \mathcal{Q}} \mathrm{D}(q(f) \,\|\, p(f \mid \mathcal{D})).$$

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

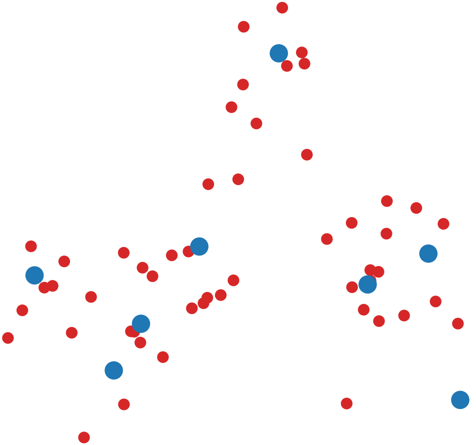
Approximating the Posterior (2)



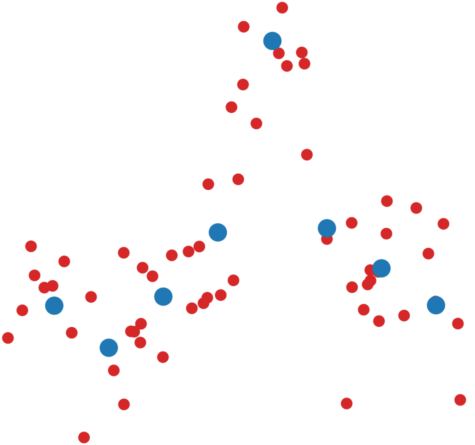
Approximating the Posterior (2)



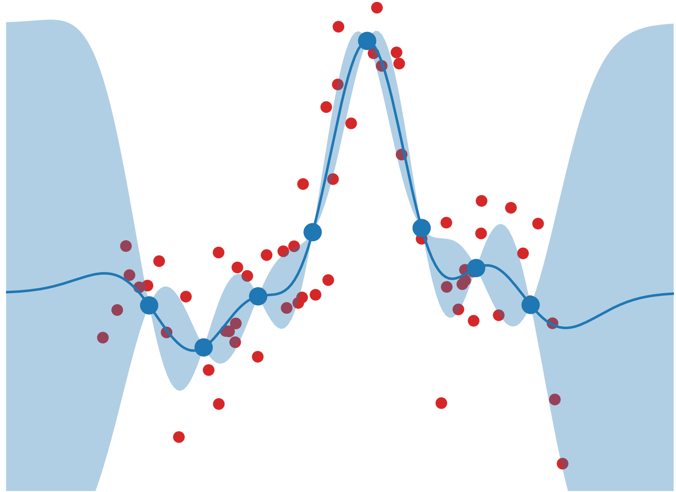
Approximating the Posterior (2)



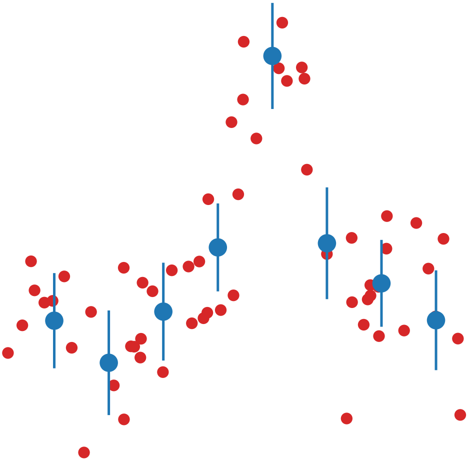
Approximating the Posterior (2)



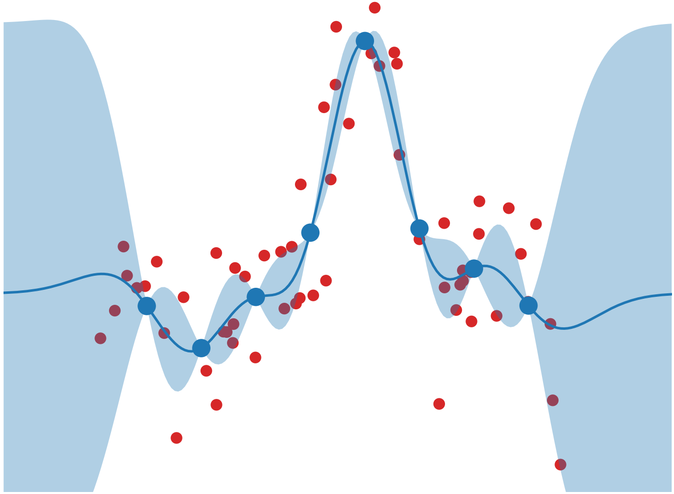
Approximating the Posterior (2)



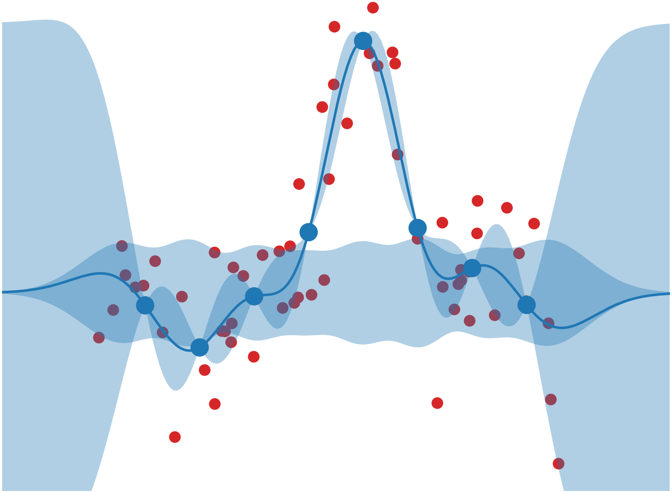
Approximating the Posterior (2)



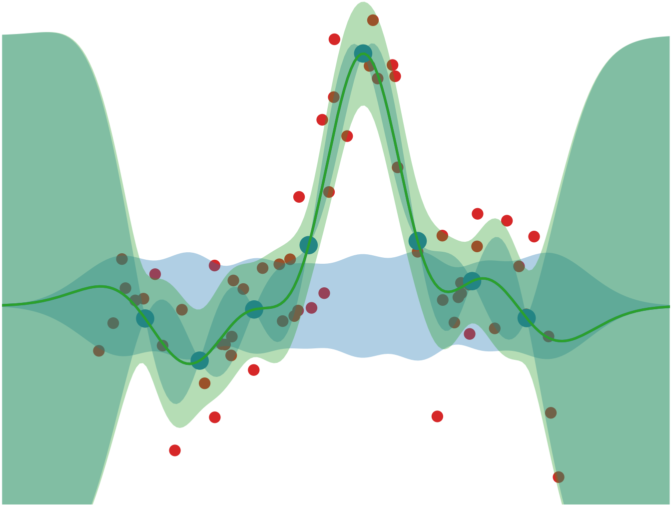
Approximating the Posterior (2)



Approximating the Posterior (2)



Approximating the Posterior (2)



Approximating the Posterior (3)

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

Approximating the Posterior (3)

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

Approximating the Posterior (3)

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

$$p(f \mid \mathcal{D}) \approx q(f) \equiv \int p(f \mid u) q(u) \, \mathrm{d}u$$

Approximating the Posterior (3)

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

$$p(f \mid \mathcal{D}) \approx q(f) \equiv \int p(f \mid u) q(u) \, \mathrm{d}u,$$

$$q^*(u) = \operatorname*{arg\,min}_{q(u)} \mathrm{D}_{\mathrm{KL}}(q(f) \,\|\, p(f \mid \mathcal{D}))$$

Approximating the Posterior (3)

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

$$p(f \mid \mathcal{D}) \approx q(f) \equiv \int p(f \mid u) q(u) \, \mathrm{d}u,$$

$$\begin{aligned} q^*(u) &= \operatorname*{arg\,min}_{q(u)} \mathrm{D}_{\mathrm{KL}}(q(f) \,\|\, p(f \mid \mathcal{D})) \\ &\propto p(u) \exp \int p(f \mid u) \log p(\mathcal{D} \mid f) \, \mathrm{d}f. \end{aligned}$$

Approximating the Posterior (3)

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

$$p(f \mid \mathcal{D}) \approx q(f) \equiv \int p(f \mid u) q(u) \,\mathrm{d}u,$$

$$\begin{aligned} q^*(u) &= \underset{q(u)}{\arg\min}\ \mathrm{D}_{\mathrm{KL}}(q(f) \,\|\, p(f \mid \mathcal{D})) \\ &\propto p(u) \exp \int p(f \mid u) \log p(\mathcal{D} \mid f) \,\mathrm{d}f. \end{aligned}$$

- Complexities:

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

Approximating the Posterior (3)

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

$$p(f \mid \mathcal{D}) \approx q(f) \equiv \int p(f \mid u) q(u) \, \mathrm{d}u,$$

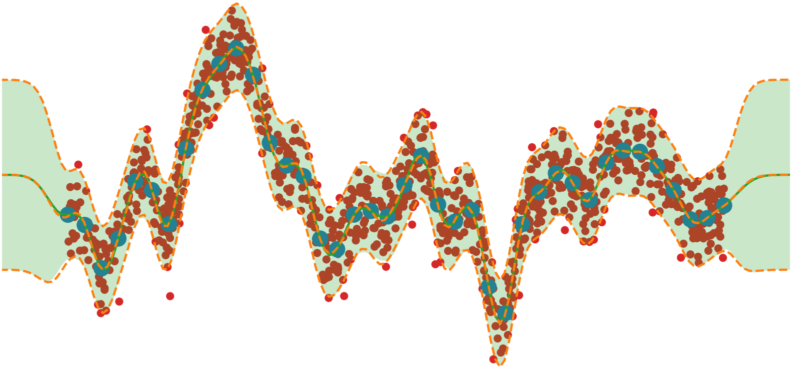
$$\begin{aligned} q^*(u) &= \operatorname*{arg\,min}_{q(u)} \mathrm{D}_{\mathrm{KL}}(q(f) \,\|\, p(f \mid \mathcal{D})) \\ &\propto p(u) \exp \int p(f \mid u) \log p(\mathcal{D} \mid f) \, \mathrm{d}f. \end{aligned}$$

- Complexities:

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

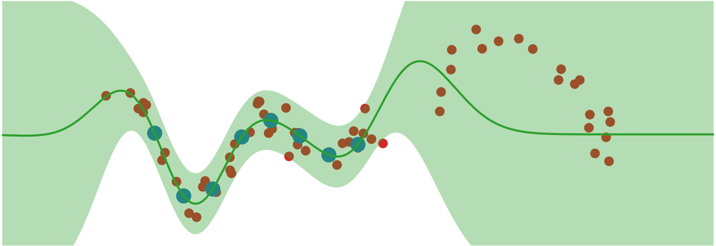
Example

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



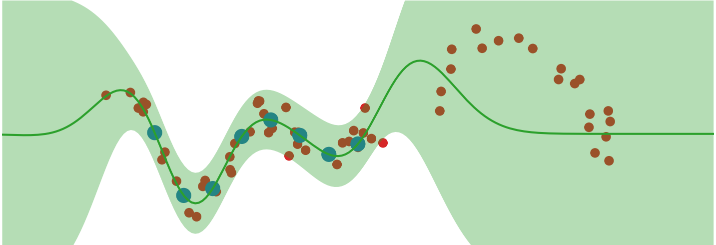
Inducing Points Versus SSA

- Inducing points: **local in time/space**



Inducing Points Versus SSA

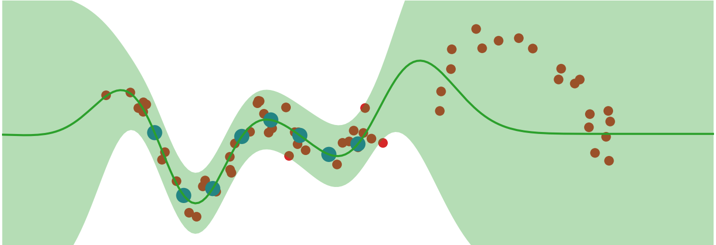
- Inducing points: **local in time/space**



- Need $M \propto N$

Inducing Points Versus SSA

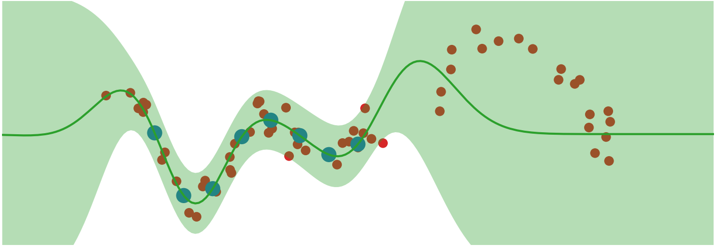
- Inducing points: **local in time/space**



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

Inducing Points Versus SSA

- Inducing points: **local in time/space**



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

Inducing Points Versus SSA (2)

Inducing points:

+ Appealing approximative construction

– Hidden $O(N^3)$ scaling

SSA:

+ Representative power

– Overfitting

Inducing Points Versus SSA (2)

Inducing points:

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

SSA:

+ Representative power
– Overfitting

Best of both worlds?

Inducing Points Versus SSA (2)

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)

Inter-Domain Inducing Points

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

Inter-Domain Inducing Points

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

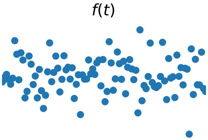
$$g(\xi) \mid f = \int_{-\infty}^{\infty} h(\xi, t) f(t) \, \mathrm{d}t, \quad u = (g(\xi_{u,1}), \ldots, g(\xi_{u,M})).$$

Inter-Domain Inducing Points

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

$$g(\xi) \mid f = \int_{-\infty}^{\infty} h(\xi, t) f(t) \, \mathrm{d}t, \quad u = (g(\xi_{u,1}), \ldots, g(\xi_{u,M})).$$

- In some cases, **necessary**:

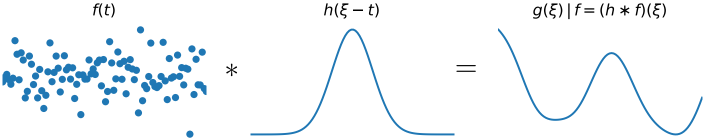


Inter-Domain Inducing Points

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

$$g(\xi) \mid f = \int_{-\infty}^{\infty} h(\xi, t) f(t) \,\mathrm{d}t, \quad u = (g(\xi_{u,1}), \ldots, g(\xi_{u,M})).$$

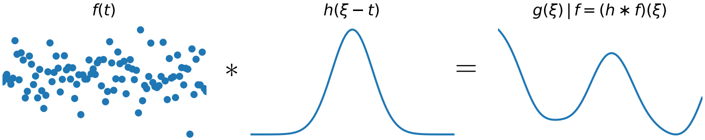
- In some cases, **necessary**:



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

$$g(\xi) \mid f = \int_{-\infty}^{\infty} h(\xi, t) f(t) \, \mathrm{d}t, \quad u = (g(\xi_{u,1}), \ldots, g(\xi_{u,M})).$$

- In some cases, **necessary**:



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

Variational Fourier Features

- Predictive mean in SSA:

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

- Predictive mean in SSA:

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

- Predictive mean for inter-domain inducing points:

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

- Predictive mean in SSA:

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

- Predictive mean for inter-domain inducing points:

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

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

Variational Fourier Features (2)

Attempt 1: $g(\xi) \mid f = \int_{-\infty}^{\infty} e^{-2\pi\iota\xi t} f(t) \, \mathrm{d}t.$

Variational Fourier Features (2)

Attempt 1: $g(\xi) \mid f = \displaystyle\int_{-\infty}^{\infty} e^{-2\pi\iota\xi t} f(t)\,\mathrm{d}t.$

- Inducing points become **inducing frequencies**.

Variational Fourier Features (2)

Attempt 1: $g(\xi) \mid f = \int_{-\infty}^{\infty} e^{-2\pi\iota\xi t} f(t) \, \mathrm{d}t.$

- Inducing points become **inducing frequencies**.
- $\hat{f}^{(\mathrm{IDIP})}(t) = \sum_{i=1}^{M} \alpha_i \hat{k}(\xi_{u,i}) e^{-2\pi\iota\xi_{u,i}t}.$

Variational Fourier Features (2)

Attempt 1: $g(\xi) \mid f = \int_{-\infty}^{\infty} e^{-2\pi\iota\xi t} f(t)\,\mathrm{d}t$.

- Inducing points become **inducing frequencies**.
- $\hat{f}^{(\mathrm{IDIP})}(t) = \sum_{i=1}^{M} \alpha_i \hat{k}(\xi_{u,i}) e^{-2\pi\iota\xi_{u,i} t}$.
- But g is white noise...

Variational Fourier Features (2)

Attempt 1: $g(\xi) \mid f = \int_{-\infty}^{\infty} e^{-2\pi\iota\xi t} f(t)\,\mathrm{d}t.$

- Inducing points become **inducing frequencies**.
- $\hat{f}^{(\mathrm{IDIP})}(t) = \sum_{i=1}^{M} \alpha_i \hat{k}(\xi_{u,i}) e^{-2\pi\iota\xi_{u,i}t}.$
- But g is white noise...

Attempt 2: $g(\xi) \mid f = \int_{a}^{b} e^{-2\pi\iota\xi(t-a)} f(t)\,\mathrm{d}t.$

Variational Fourier Features (2)

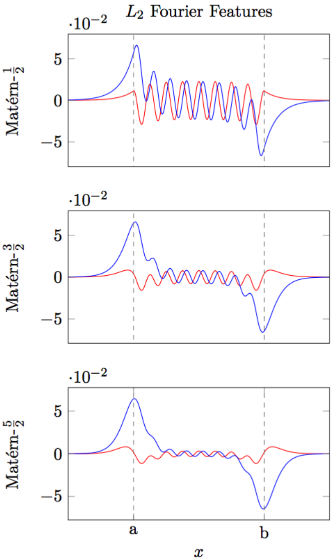
Attempt 1: $g(\xi) \mid f = \int_{-\infty}^{\infty} e^{-2\pi\iota\xi t} f(t)\,\mathrm{d}t.$

- Inducing points become **inducing frequencies**.
- $\hat{f}^{(\mathrm{IDIP})}(t) = \sum_{i=1}^{M} \alpha_i \hat{k}(\xi_{u,i}) e^{-2\pi\iota\xi_{u,i}t}.$
- But g is white noise...

Attempt 2: $g(\xi) \mid f = \int_{a}^{b} e^{-2\pi\iota\xi(t-a)} f(t)\,\mathrm{d}t.$ ($L^2[a,b]$-VFFs)

- Works, but edge effects near boundaries.

Variational Fourier Features (3)



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

Variational Fourier Features (4)

- Use tools from RKHS theory to eliminate edge effects.

Variational Fourier Features (4)

- 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

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

Variational Fourier Features (4)

- 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

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

- Why?

Variational Fourier Features (4)

- 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

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

- Why?

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

Variational Fourier Features (5)

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

Variational Fourier Features (5)

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

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

Variational Fourier Features (5)

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

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

- Solution:

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

Variational Fourier Features (5)

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

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

- Solution:

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

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

Variational Fourier Features (5)

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

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

- Solution:

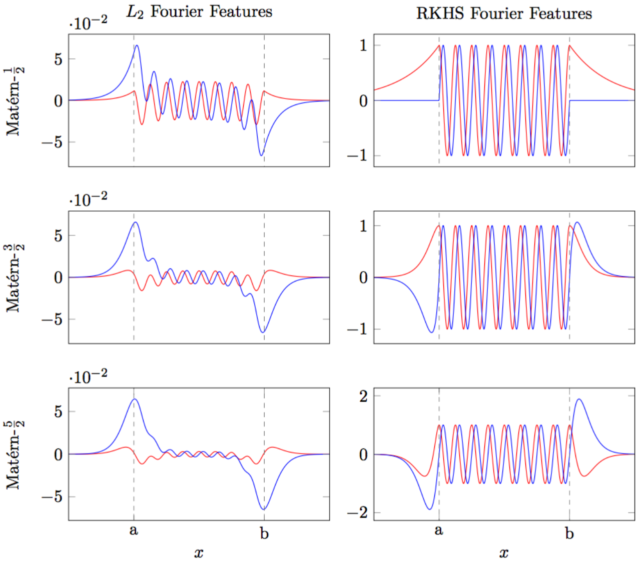
$$\begin{aligned} h(\xi, t) &= \sum_{i=1}^{\infty} \frac{1}{\lambda_i} \phi_i(t) \langle \psi(\xi, \cdot), \phi_i \rangle \\ \implies T_k(h(\xi, \cdot))(t) &= \sum_{i=1}^{\infty} \frac{1}{\lambda_i} T_k(\phi_i)(t) \langle \psi(\xi, \cdot), \phi_i \rangle \\ &= \sum_{i=1}^{\infty} \phi_i(t) \langle \psi(\xi,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

Variational Fourier Features (6)

- Works for Matérn kernels of half-integer order!

$\Rightarrow$ VFFs (RKHS-VFFs)

Variational Fourier Features (7)



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

Rates of Convergence

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

Rates of Convergence

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

Rates of Convergence

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

Theorem

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 μ 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 $\mathrm{D_{KL}}(q(f) \,\|\, p(f \mid \mathcal{D})) \leq \varepsilon$ with probability at least $1 - \delta$.

Rates of Convergence (2)

Eigenfunction Inducing Features

- Eigenfunction inducing features:

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

Eigenfunction Inducing Features

- Eigenfunction inducing features:

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

- Nice behaviour:

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

Eigenfunction Inducing Features

- Eigenfunction inducing features:

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

- Nice behaviour:

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

- Key quantity:

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

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

Sketch of Proof

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

Sketch of Proof

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

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

Rates of Convergence (3)

Sketch of Proof

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

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

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

Conclusion

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

Conclusion

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

Conclusion

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

Conclusion

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

Conclusion

- 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.

Conclusion

- 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.