A Tutorial on Sparse Gaussian Processes

Shengyang Sun

Feb 26th, 2021

Gaussian Processes

GP Inferences
using inducing points

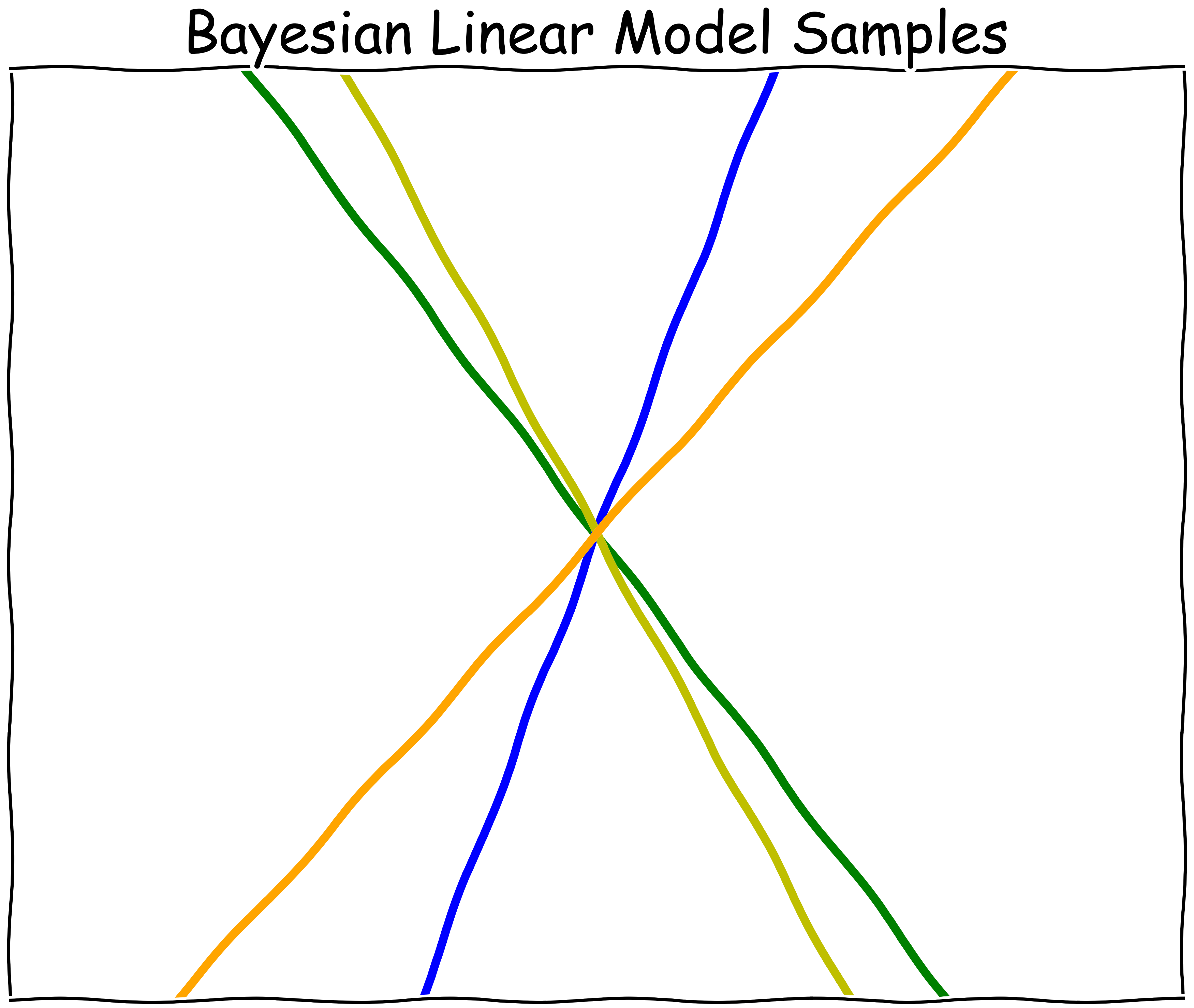
Composite GPs

Inducing Points
Beyond GPs

Bayesian Linear Models

- We are interested at the underlying function f of a problem.
- To characterize the function, linear models are the simplest, $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$
- Bayesian Linear Regression further characterizes the uncertainty with a prior on $\mathbf{w}$

$$f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}, \ \mathbf{w} \sim \mathcal{N}(0, \nu^2 \mathbf{I})$$



Bayesian Linear Models

- The prior in Bayesian linear regression enables various plausible explanations,

$$f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}, \;\; \mathbf{w} \sim \mathcal{N}(0, \nu^2 \mathbf{I})$$

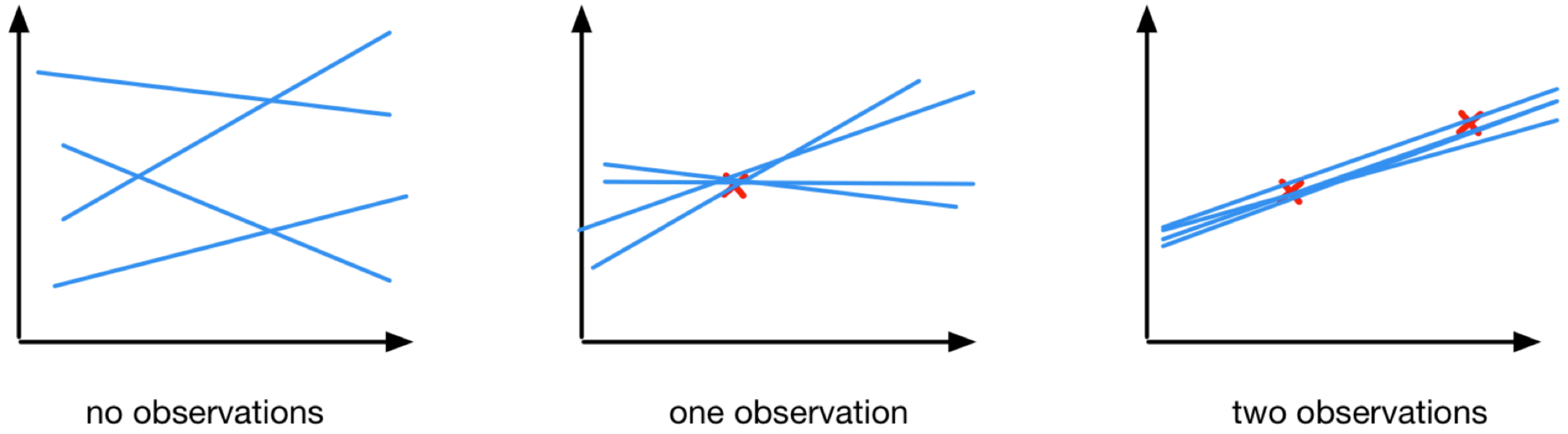
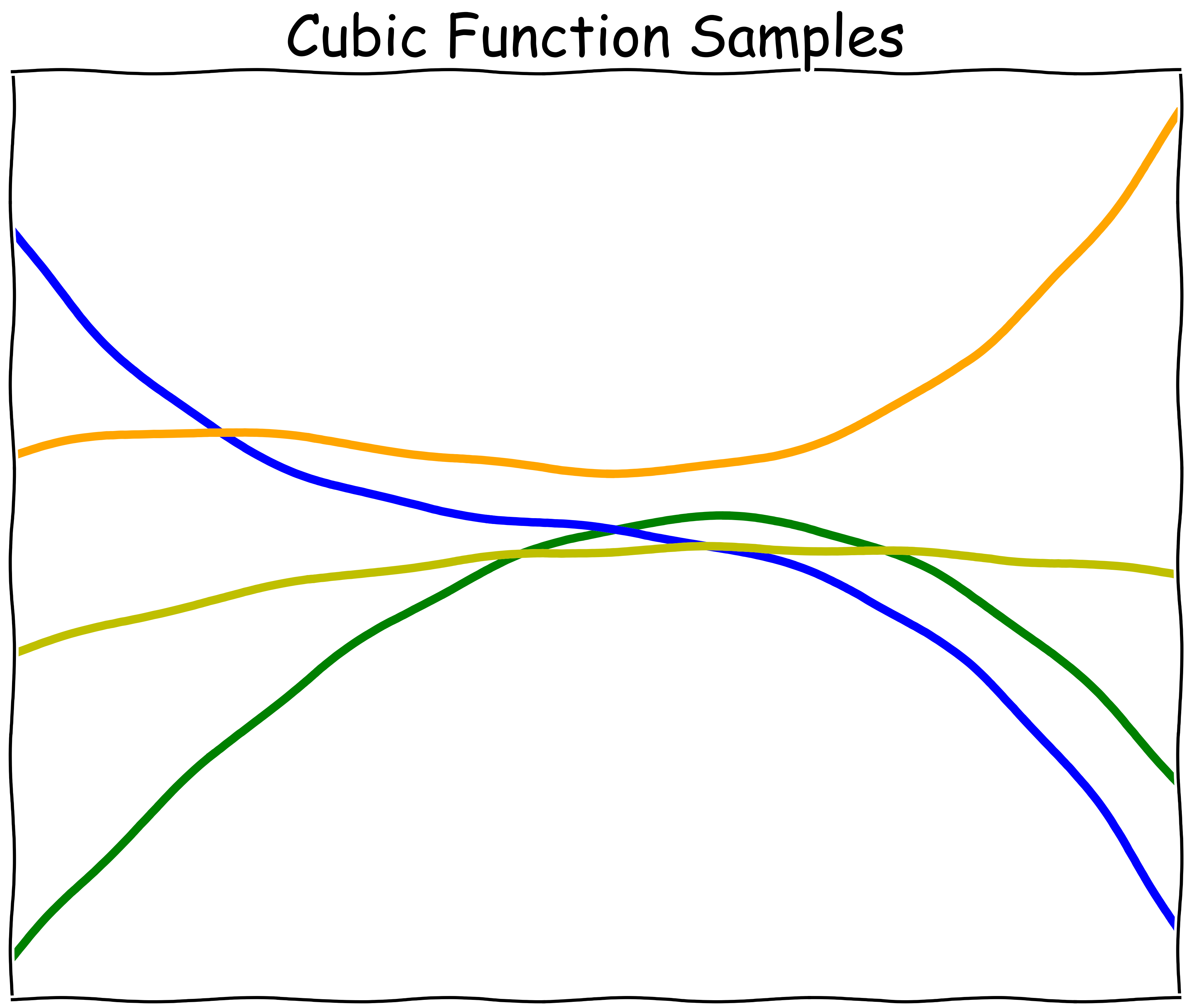


Image Courtesy of Roger Grosse, csc2541-f17

From Linear Models to Gaussian Processes

- What if the underlying function cannot be well approximated by a linear model ?
- Resort to the linear regression on non-linear features of the inputs.

$$f(\mathbf{x}) = \mathbf{w}^\top \varphi(\mathbf{x}), \ \mathbf{w} \sim \mathcal{N}(0, \nu^2 \mathbf{I})$$



From Linear Models to Gaussian Processes

- Bayesian linear regression,

$$f(\mathbf{x}) = \mathbf{w}^\top \varphi(\mathbf{x}),\ \mathbf{w} \sim \mathcal{N}(0, \nu^2 \mathbf{I})$$

- The weight-space prior defines a prior on the function values,
- Consider inputs $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$, whose function values $\mathbf{f} = [f(\mathbf{x}_1), f(\mathbf{x}_2), ..., f(\mathbf{x}_n)]^\top$

$$\mathbf{f} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}),$$

- Each element of the kernel matrix depends only on the corresponding pair of inputs.

$$\mathbf{K}_{ij} = \nu^2 \phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_j)$$

From Linear Models to Gaussian Processes

- The prior on finite sets of function values fully characterizes the distribution.

- *Given a kernel function* $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ *, a Gaussian process* $\mathcal{GP}(0, k)$ *is a distribution of functions. For any finite set of inputs* $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$ *, their function values satisfy a multivariate Gaussian distribution,*

$$\mathbf{f} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}),$$

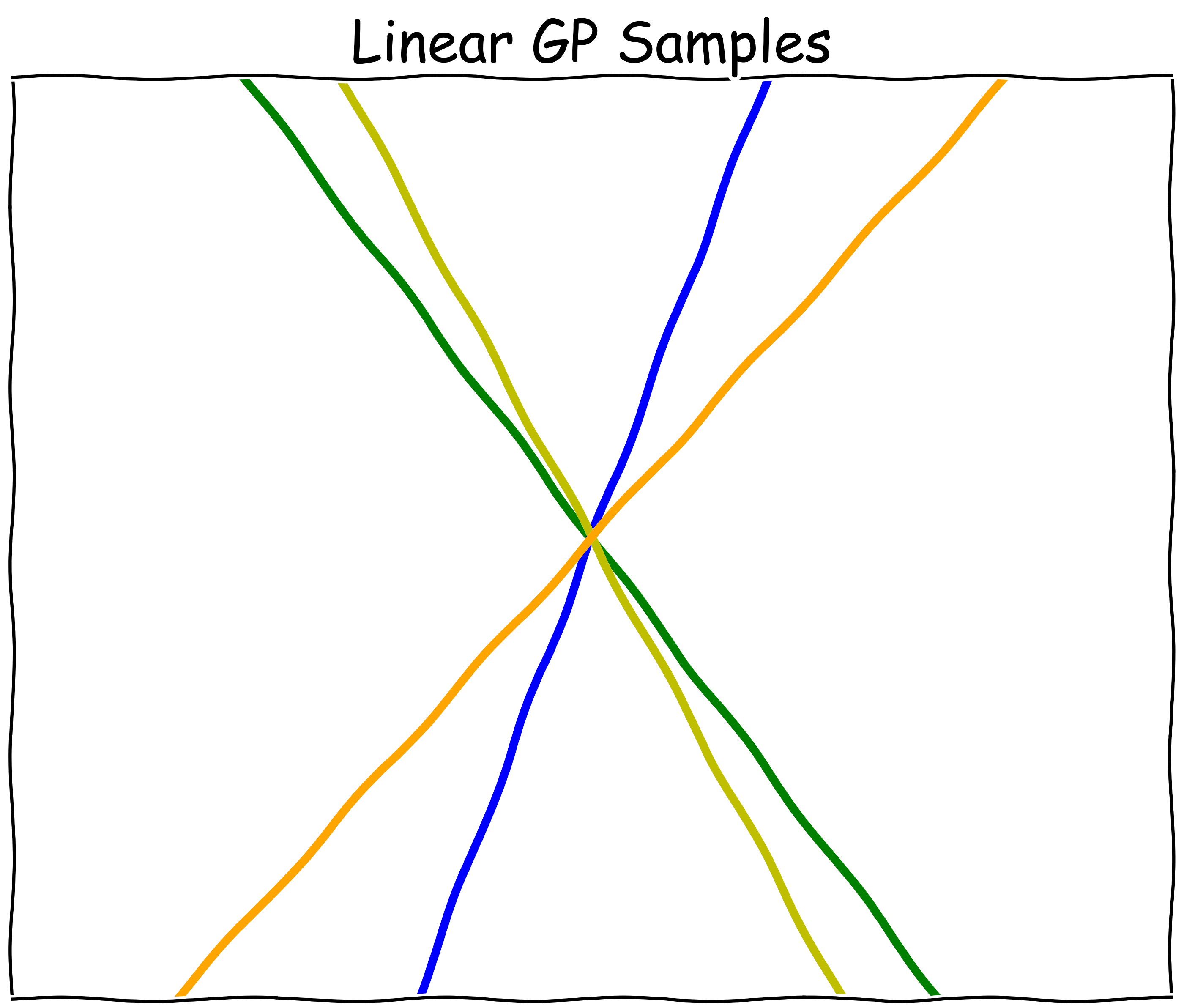
  Where $\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$

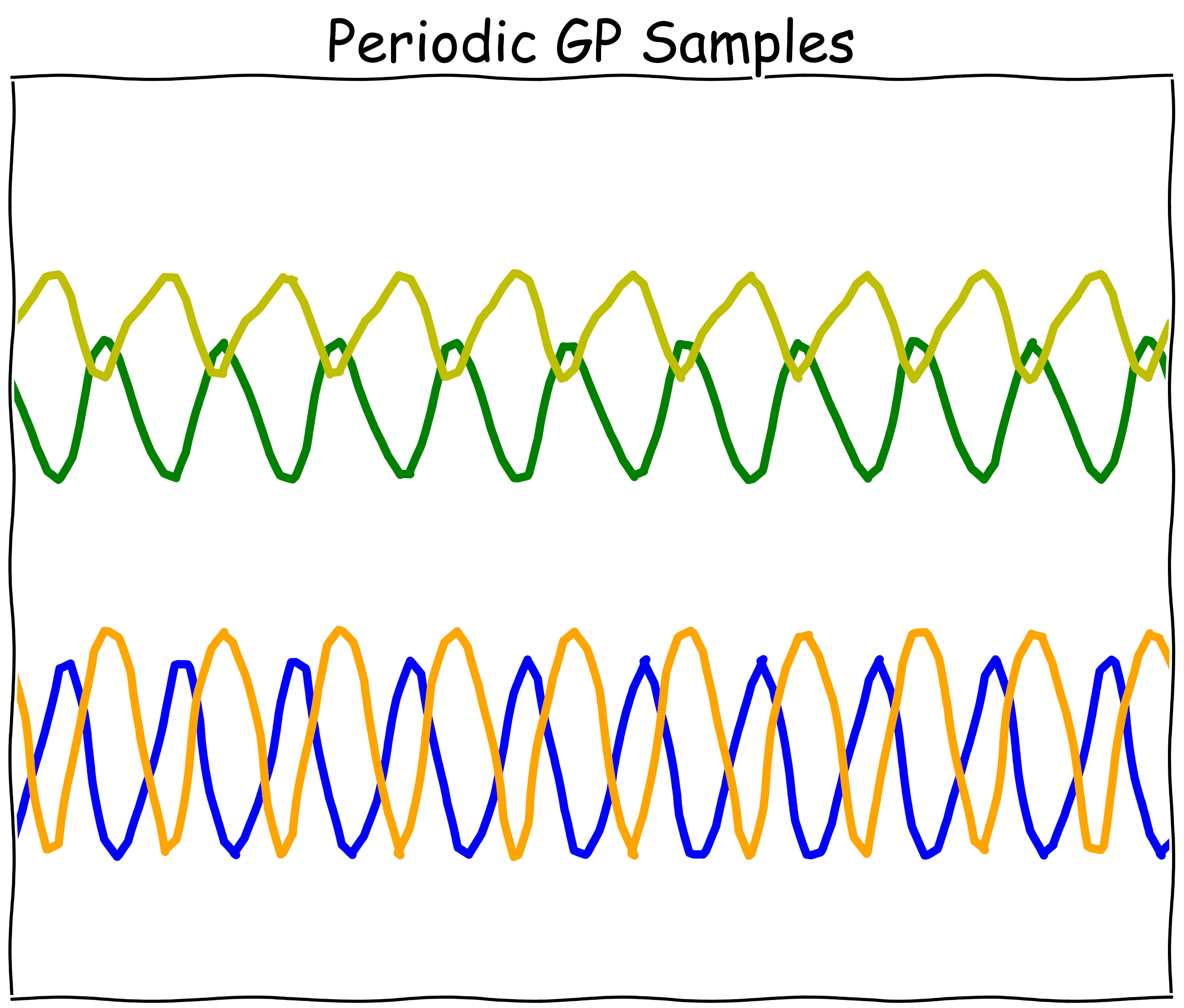
- Gaussian Processes are Bayesian linear regressions on nonlinear feature maps.

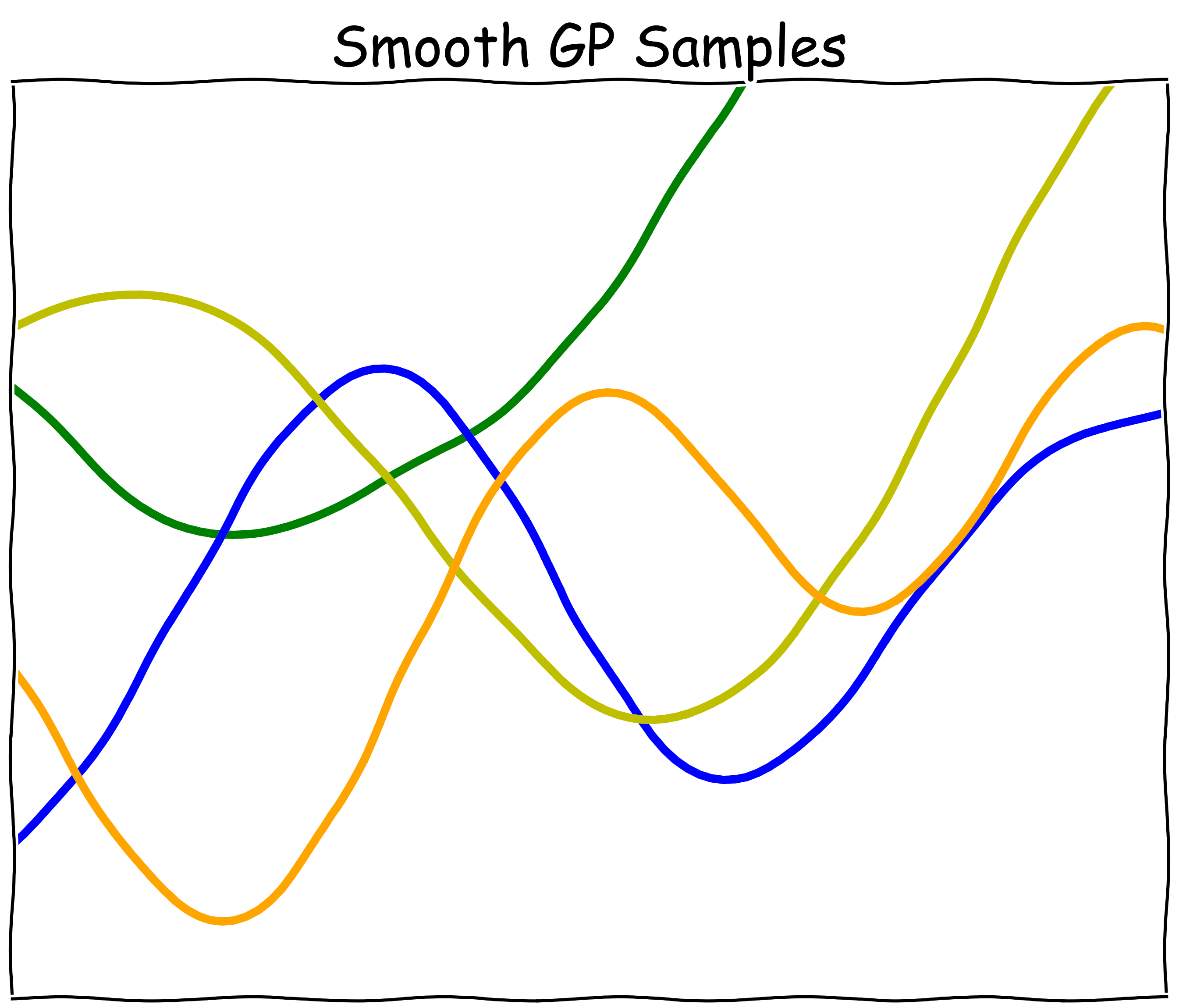
Kernels enable flexible Model Design

- Different kernels specify widely varying structures,

.



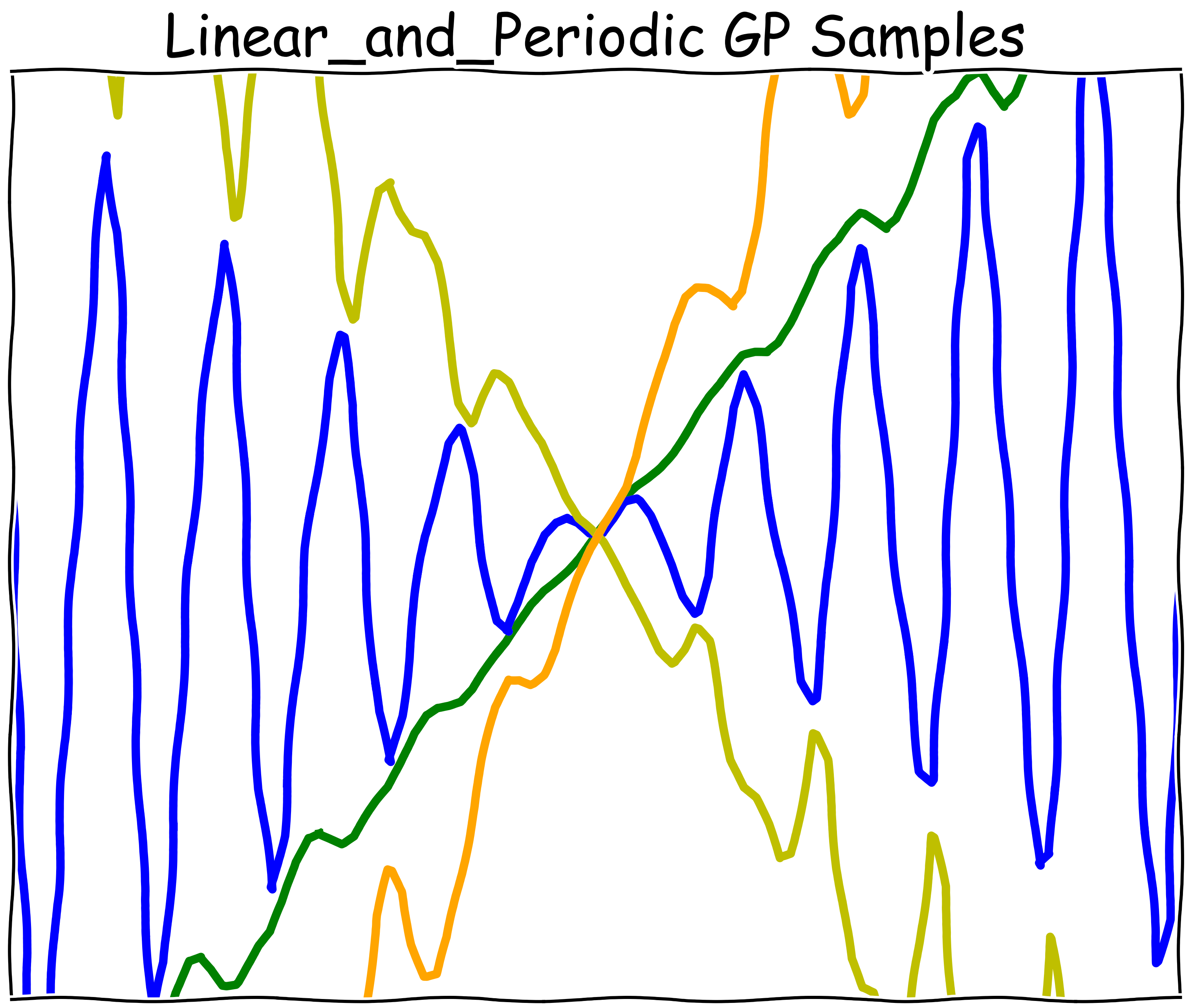


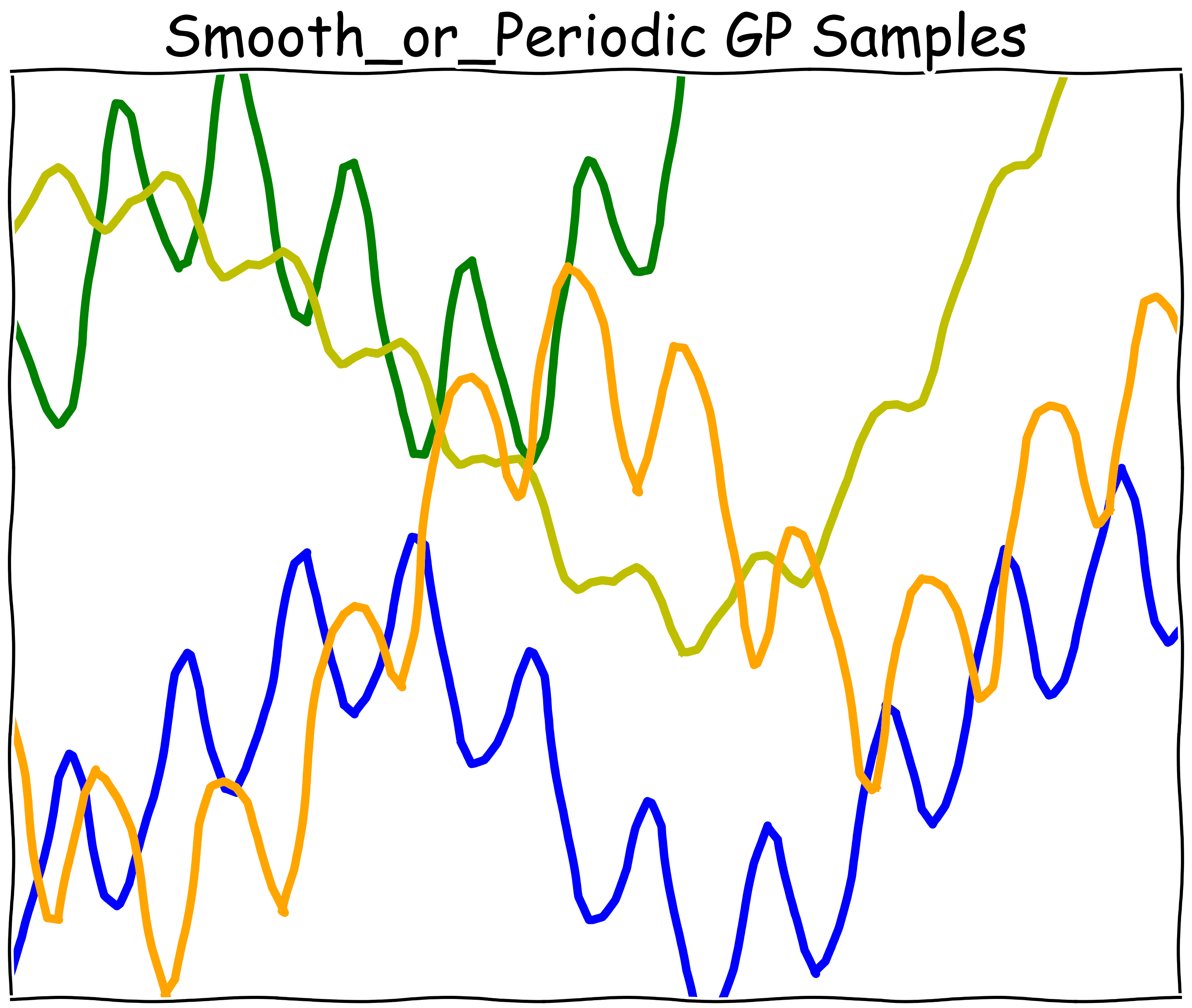


Kernels enable flexible Model Design

- Kernels can be combined to specify a composite of structures,

.





Kernels enable flexible Model Design

- ECG signals monitor the heart beat, which are generally periodic with variations.
- For a pregnant patient, the ECG is the composite of the mother's and the baby's.

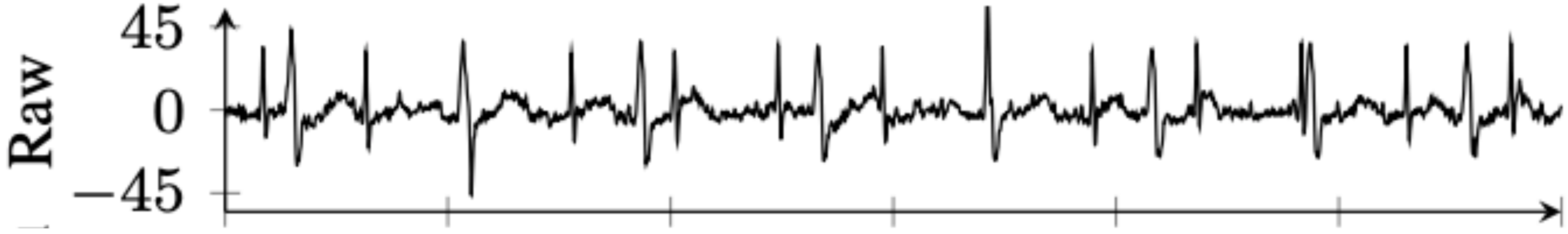


Image Courtesy of Graßhoff et. al., (2020)

Kernels enable flexible Model Design

- Gaussian processes specify the composite structure easily,

$$k(t, t') = k_{baby}(t, t) + k_{mother}(t, t')$$

- Inferences for the GP decomposes the composite signals,

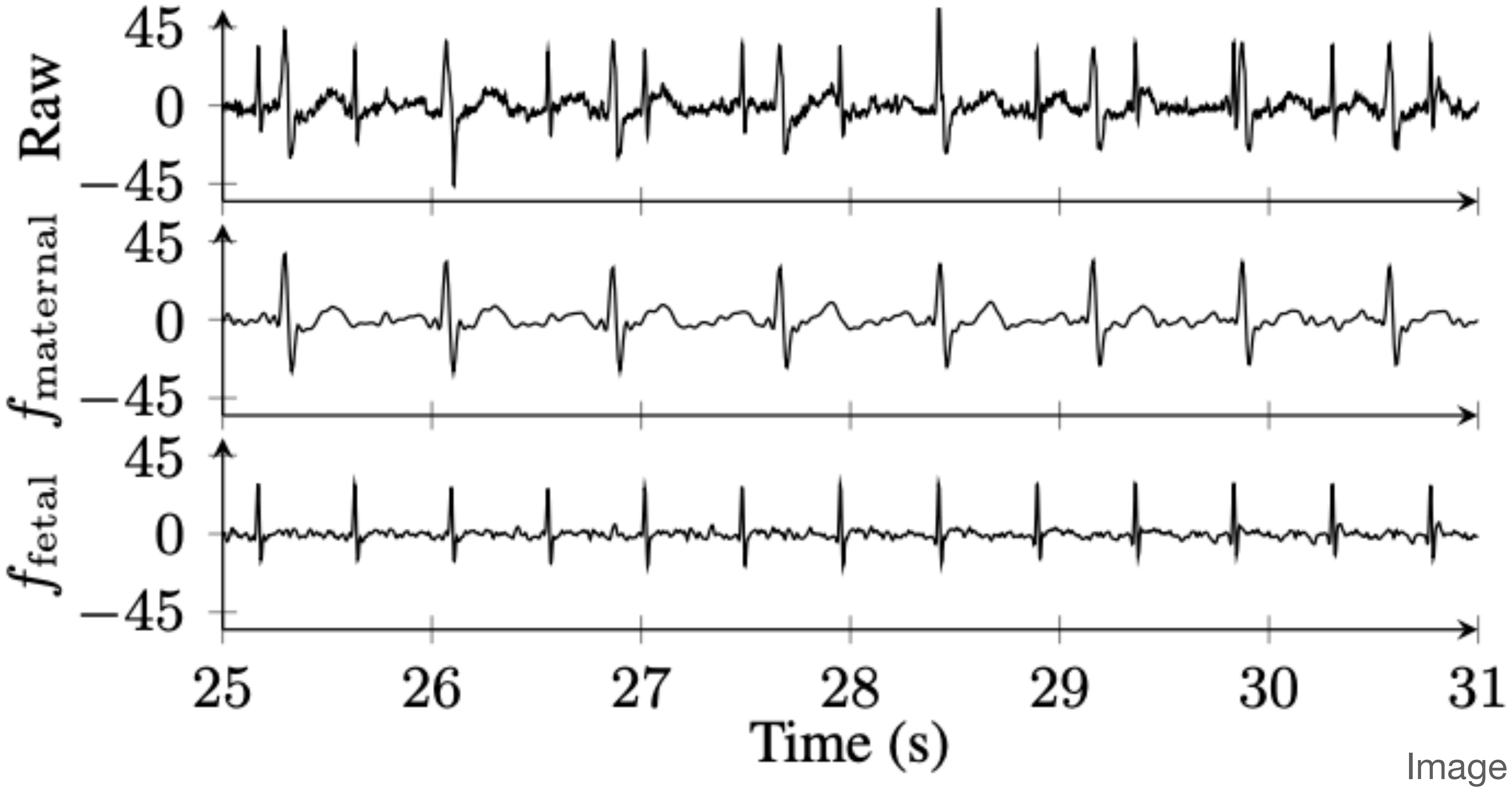


Image Courtesy of Graßhoff et. al., (2020)

What are ongoing research directions?

- Designing Flexible Kernels
  - Deep Kernel Learning, Spectral Mixture Kernels
- Automatic Kernel Selection
  - Automatic Statistician, Neural Kernel Network
- The function-space and weight-space contradistinctions
  - Neural Tangent Kernel, Neural network Gaussian process
- Gaussian processes for structured spaces
  - Convolutional Gaussian processes, graph convolutional Gaussian processes

Gaussian Processes

GP Inferences using inducing points

Composite GPs

Inducing Points Beyond GPs

GP Predictions from the Posterior

- Given a GP prior $\mathcal{GP}(0, k)$, and a dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ from $p\big(y \big| f(\mathbf{x})\big)$
- We are interested at inferring the posterior

$$p(f|\mathcal{D}) = \frac{p(\mathcal{D}|f)p(f)}{p(\mathcal{D})}$$

- The GP posterior can be used for making predictions on testing locations,

$$p(y_\star|\mathbf{x}_\star, \mathcal{D}) = \int p(y|f(\mathbf{x}_\star))p(f|\mathcal{D})df$$

GP Predictions from the Posterior

- Given a GP prior $\mathcal{GP}(0, k)$, and a dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ from $p(y|f(\mathbf{x}))$
- We are interested at inferring the posterior

$$p(f|\mathcal{D}) = \frac{p(\mathcal{D}|f)p(f)}{p(\mathcal{D})} \approx q(f)$$

- The GP posterior can be used for making predictions on testing locations,

$$p(y_\star|\mathbf{x}_\star, \mathcal{D}) \approx \int p(y|f(\mathbf{x}_\star))q(f)df$$

Full Data "Exact"

Inducing Points Approxi

Gaussian Likelihoods

MCMC

Variational Inference

MCMC[1]

Variational Inference

[1] Appendix

Conditionals of Multivariate Gaussians

- Consider a multivariate Gaussian,

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}_\star \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \mathbf{K}_{ff} & \mathbf{K}_{f\star} \\ \mathbf{K}_{\star f} & \mathbf{K}_{\star\star} \end{bmatrix} \right)$$

- The conditional distribution is a multivariate Gaussian,

$$\mathbf{f}_\star | \mathbf{f} \sim \mathcal{N}(\mathbf{K}_{\star f}\mathbf{K}_{ff}^{-1}\mathbf{f}, \mathbf{K}_{\star\star} - \mathbf{K}_{\star f}\mathbf{K}_{ff}^{-1}\mathbf{K}_{f\star})$$

GP Posteriors under Gaussian Likelihoods

- Under a Gaussian likelihood,

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_\star \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \mathbf{K}_{ff} + \sigma^2\mathbf{I} & \mathbf{K}_{f\star} \\ \mathbf{K}_{\star f} & \mathbf{K}_{\star\star} \end{bmatrix}\right)$$

- The posterior is a multivariate Gaussian,

$$\mathbf{f}_\star | \mathbf{y} \sim \mathcal{N}(\mathbf{K}_{\star f}(\mathbf{K}_{ff} + \sigma^2\mathbf{I})^{-1}\mathbf{y}, \mathbf{K}_{\star\star} - \mathbf{K}_{\star f}(\mathbf{K}_{ff} + \sigma^2\mathbf{I})^{-1}\mathbf{K}_{f\star})$$

- The "function" posterior $p(f|\mathcal{D})$ can be seen as a "vector" posterior $p(\mathbf{f}_\star|\mathcal{D})$

MCMC for Gaussian Processes

- Markov Chain Monte Carlo evolves particles according to the unnormalized density, whose distribution is the stationary distribution of the Markov Chain.
- How can we update an infinite-dimensional function ?

MCMC for Gaussian Processes

- Markov Chain Monte Carlo evolves particles according to the unnormalized density, whose distribution is the stationary distribution of the Markov Chain.

- How can we update an infinite-dimensional function ?

- Consider a augmented posterior,

$$p(f, \mathbf{f}|\mathcal{D}) \propto p(\mathcal{D}|f, \mathbf{f})p(f, \mathbf{f}) = p(\mathcal{D}|\mathbf{f})p(\mathbf{f})p(f|\mathbf{f}) \propto p(\mathbf{f}|\mathcal{D})p(f|\mathbf{f})$$

MCMC for Gaussian Processes

- Markov Chain Monte Carlo evolves particles according to the unnormalized density, whose distribution is the stationary distribution of the Markov Chain.
- How can we update an infinite-dimensional function ?
- Consider a augmented posterior,

$$p(f, \mathbf{f}|\mathcal{D}) \propto p(\mathcal{D}|f, \mathbf{f})p(f, \mathbf{f}) = p(\mathcal{D}|\mathbf{f})p(\mathbf{f})p(f|\mathbf{f}) \propto p(\mathbf{f}|\mathcal{D})p(f|\mathbf{f})$$

- $\mathbf{f}$ is finite-dimensional! MCMC can obtain samples from $p(\mathbf{f}|\mathcal{D})$

- MCMC is applicable to general likelihoods.

MCMC for Gaussian Processes

- Evolving MCMC particles requires evaluating the unnormalized log probability,

$$\log p(\mathcal{D}|\mathbf{f}) + \log p(\mathbf{f}) = \sum_{i=1}^{n} \log p(y_i|\mathbf{f}_i) - \frac{1}{2}\mathbf{f}^\top \mathbf{K}_{ff}^{-1}\mathbf{f} + const$$

- The exact posterior under Gaussian likelihoods,

$$\mathbf{f}_\star|\mathbf{y} \sim \mathcal{N}(\mathbf{K}_{\star f}(\mathbf{K}_{ff} + \sigma^2\mathbf{I})^{-1}\mathbf{y}, \mathbf{K}_{\star\star} - \mathbf{K}_{\star f}(\mathbf{K}_{ff} + \sigma^2\mathbf{I})^{-1}\mathbf{K}_{f\star})$$

- Is it possible to circumvent the cubic computations from matrix inversions ?

Variational Inference of GPs

- Variational Inference is another class of techniques for approximate posteriors, which optimizes a variational posterior by maximizing the Evidence Lower Bound (ELBO),

$$\log p(\mathcal{D}) \geq \mathbb{E}_{q(f)}[\log p(\mathcal{D}|f)] - \mathrm{KL}[q(f)\|p(f)]$$

Variational Inference of GPs

- Variational Inference is another class of techniques for approximate posteriors, which optimizes a variational posterior by maximizing the Evidence Lower Bound (ELBO),

$$\log p(\mathcal{D}) \geq \mathbb{E}_{q(f)}[\log p(\mathcal{D}|f)] - \mathrm{KL}[q(f)\|p(f)]$$

- To specify the variational posterior for f, we again consider the augmented space,

$$\log p(\mathcal{D}) \geq \mathbb{E}_{q(f,\mathbf{f})}[\log p(\mathcal{D}|f,\mathbf{f})] - \mathrm{KL}[q(f,\mathbf{f})\|p(f,\mathbf{f})]$$

where the variational posterior is,

$$q(f,\mathbf{f}) = p(f|\mathbf{f})q(\mathbf{f})$$

Variational Inference of GPs

- To specify the variational posterior for f , we again consider the augmented space,

$$\log p(\mathcal{D}) \geq \mathbb{E}_{q(f,\mathbf{f})}[\log p(\mathcal{D}|f,\mathbf{f})] - \mathrm{KL}[q(f,\mathbf{f})\|p(f,\mathbf{f})]$$

$$q(f,\mathbf{f}) = p(f|\mathbf{f})q(\mathbf{f})$$

- Then the ELBO can be rewritten as,

$$\begin{aligned}
\mathcal{L} &= \sum_{i=1}^{n} \mathbb{E}_{q(\mathbf{f}_i)}[\log p(y_i|\mathbf{f}_i)] - \mathbb{E}_{q(f,\mathbf{f})}[\log \frac{p(f|\mathbf{f})q(\mathbf{f})}{p(f|\mathbf{f})p(\mathbf{f})}] \\
&= \sum_{i=1}^{n} \mathbb{E}_{q(\mathbf{f}_i)}[\log p(y_i|\mathbf{f}_i)] - \mathbb{E}_{q(\mathbf{f})}[\log \frac{q(\mathbf{f})}{p(\mathbf{f})}] \\
&= \sum_{i=1}^{n} \mathbb{E}_{q(\mathbf{f}_i)}[\log p(y_i|\mathbf{f}_i)] - \mathrm{KL}[q(\mathbf{f})\|p(\mathbf{f})]
\end{aligned}$$

Variational Inference of GPs

- To specify the variational posterior for f , we again consider the augmented space,

$$\log p(\mathcal{D}) \geq \mathbb{E}_{q(f,\mathbf{f})}[\log p(\mathcal{D}|f,\mathbf{f})] - \mathrm{KL}[q(f,\mathbf{f})\|p(f,\mathbf{f})]$$

$$q(f,\mathbf{f}) = p(f|\mathbf{f})q(\mathbf{f})$$

- Then the ELBO can be rewritten as,

$$\mathcal{L} = \sum_{i=1}^{n} \mathbb{E}_{q(\mathbf{f}_i)}[\log p(y_i|\mathbf{f}_i)] - \mathrm{KL}[q(\mathbf{f})\|p(\mathbf{f})]$$

stochastic estimations ✔ cubic of n computations ✘

KL between Gaussians: $\frac{1}{2}\left[\log\frac{|\Sigma_2|}{|\Sigma_1|} - d + \mathrm{tr}\{\Sigma_2^{-1}\Sigma_1\} + (\mu_2-\mu_1)^T\Sigma_2^{-1}(\mu_2-\mu_1)\right]$

Variational Inference using Inducing Points

- It seems that we can never get around the cubic computations if we deal with $\mathbf{f}$

$$\log p(\mathcal{D}) \geq \mathbb{E}_{q(f,\mathbf{f})}[\log p(\mathcal{D}|f, \mathbf{f})] - \mathrm{KL}[q(f, \mathbf{f})\|p(f, \mathbf{f})]$$

$$q(f, \mathbf{f}) = p(f|\mathbf{f})q(\mathbf{f})$$

Variational Inference using Inducing Points

- It seems that we can never get around the cubic computations if we deal with $\mathbf{f}$

$$\log p(\mathcal{D}) \geq \mathbb{E}_{q(f,\mathbf{f})}[\log p(\mathcal{D}|f,\mathbf{f})] - \mathrm{KL}[q(f,\mathbf{f})\|p(f,\mathbf{f})]$$

$$q(f,\mathbf{f}) = p(f|\mathbf{f})q(\mathbf{f})$$

- Instead of $\mathbf{f} = f(\mathbf{x}_{1:n})$, we consider $\mathbf{u} = f(\mathbf{z}_{1:m})$. $\mathbf{z}_{1:m}$ are inducing points that try to summarize the dataset.

$$q(f,\mathbf{u}) = p(f|\mathbf{u})q(\mathbf{u})$$

$$\begin{aligned}\log p(\mathcal{D}) \geq& \mathbb{E}_{q(f,\mathbf{u})}[\log p(\mathcal{D}|f,\mathbf{u})] - \mathrm{KL}[q(f,\mathbf{u})\|p(f,\mathbf{u})] \\ =& \mathbb{E}_{q(f,\mathbf{u})}[\log p(\mathcal{D}|f,\mathbf{u})] - \mathrm{KL}[q(\mathbf{u})\|p(\mathbf{u})]\end{aligned}$$

stochastic estimations ✓ cubic of m computations ✓

Variational Inference using Inducing Points

- Stochastic Variational Gaussian Processes (SVGP) [1, 2]

$$\mathcal{L} = \mathbb{E}_{q(f,\mathbf{u})}[\log p(\mathcal{D}|f, \mathbf{u})] - \mathrm{KL}[q(\mathbf{u})\|p(\mathbf{u})]$$

Hyper-parameters

Kernels: s^2 l^2

Likelihoods: σ^2

Variational parameters

Inducing Points $\mathbf{z}_{1:m}$

Variational Distribution $q(\mathbf{u}) = \mathcal{N}(\mu, \mathbf{S})$

Variational Inference using Inducing Points

- SVGP adapts the inducing locations and the variational distributions.

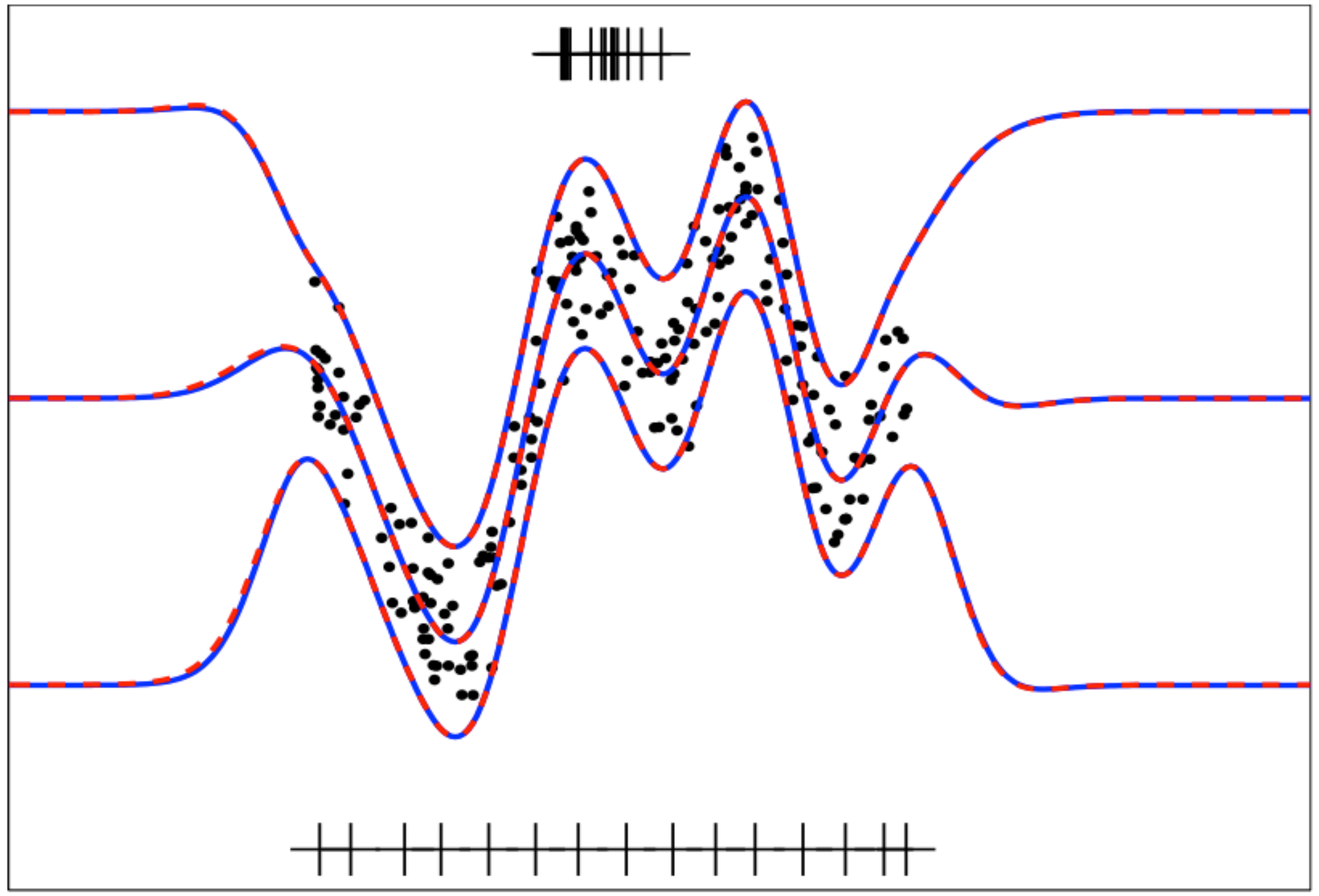


Image Courtesy of Titsias (2009)

Variational Inference using Inducing Points

- More inducing points approximates the true posterior better, without overfitting.

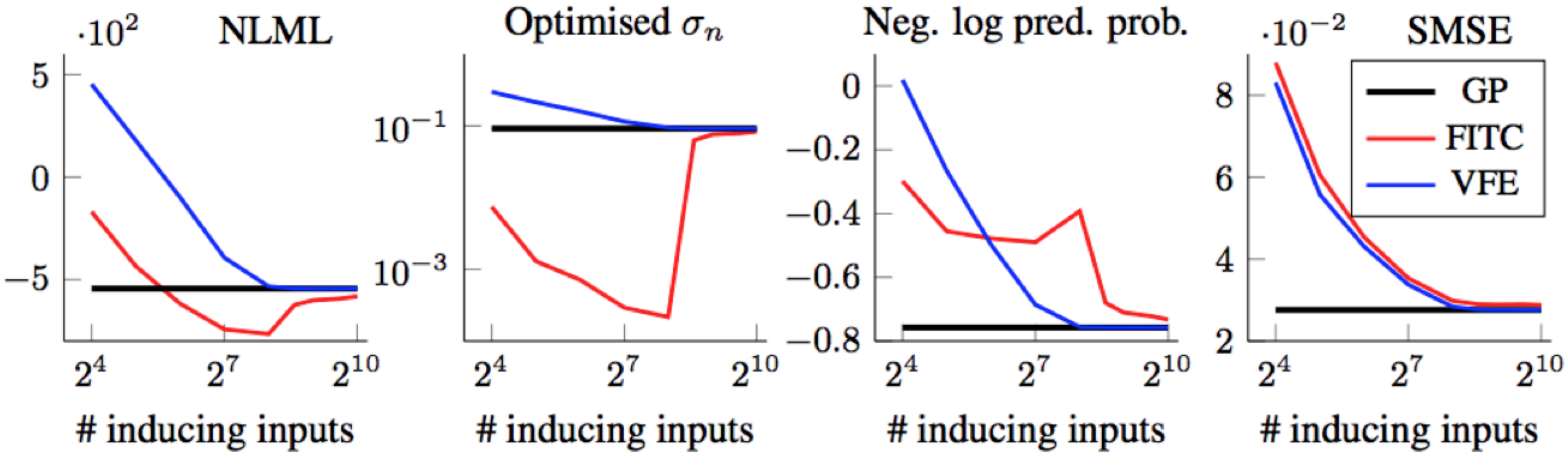


Image Courtesy of Bauer et. al., 2016

What are ongoing research directions?

- How to break the $\mathcal{O}(m^3)$ restriction to use more inducing points ?
  - Structured inducing points / Inter-domain inducing points
  - GPs, State-space models, Dynamic systems
  - Fast Numerical Solvers
- To approximate the model instead of approximate the posterior
  - (Structured) Kernel Interpolation
  - Random Fourier Features
- Online posterior inference for GPs
  - Streaming sparse GPs

Gaussian Processes

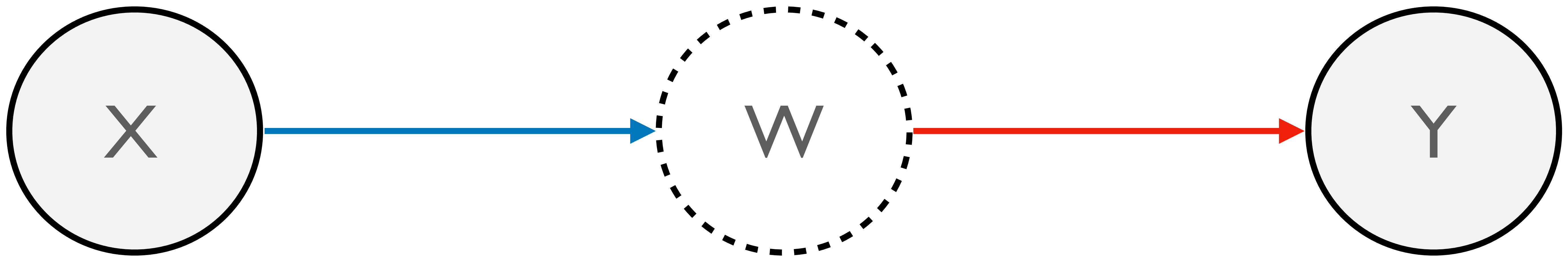
GP Inferences
using inducing points

Composite GPs

Inducing Points
Beyond GPs

The Composite of Gaussian processes

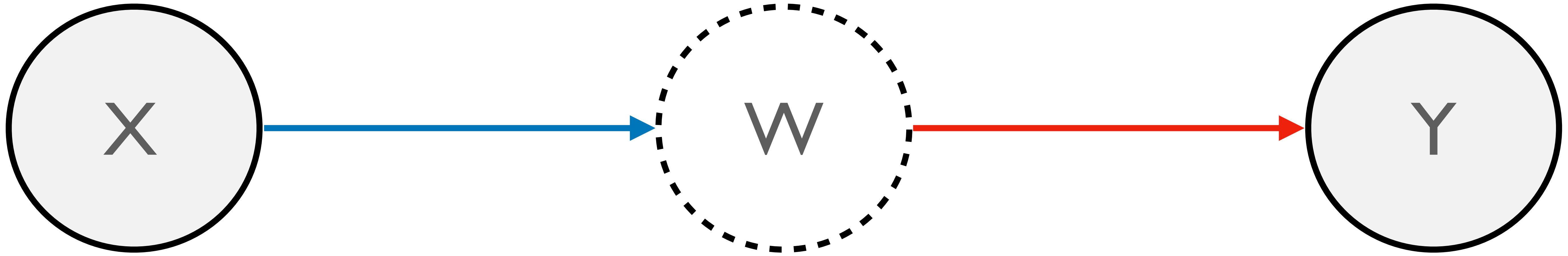
- We can composite multiple GPs for the connections between several variables.



- Assume the input X affects the output Y via the unobservable variable W,
- We use two Gaussian processes (blue and red) to model the connections.

$$f_w \sim \mathcal{GP}(0, k_w),\ f_y \sim \mathcal{GP}(0, k_y)$$

The Composite of Gaussian processes



- To approximate the posterior distribution $p(f_w, f_y|\mathcal{D})$, we introduce two sets of inducing points for two functions,

$$q(f_w, f_y, \mathbf{u}_w, \mathbf{u}_y) = p(f_w|\mathbf{u}_w)p(f_y|\mathbf{u}_y)q(\mathbf{u}_w, \mathbf{u}_y)$$

- The ELBO can be written as,

$$\begin{aligned}\log p(\mathcal{D}) &\geq \mathbb{E}_{q(f_w, f_y)}[\log p(\mathcal{D}|f_w, f_y)] - \mathrm{KL}[q(f_w, f_y, \mathbf{u}_w, \mathbf{u}_y)\|p(f_w, f_y, \mathbf{u}_w, \mathbf{u}_y)] \\ &= \mathbb{E}_{q(f_w, f_y)}[\log p(\mathcal{D}|f_w, f_y)] - \mathrm{KL}[q(\mathbf{u}_w, \mathbf{u}_y)\|p(\mathbf{u}_w)p(\mathbf{u}_y)]\end{aligned}$$

The Composite of Gaussian processes

- Gaussian processes can be composited in any non-cyclic graphical form,
- Each variable can be observable, partially observable, or hidden.

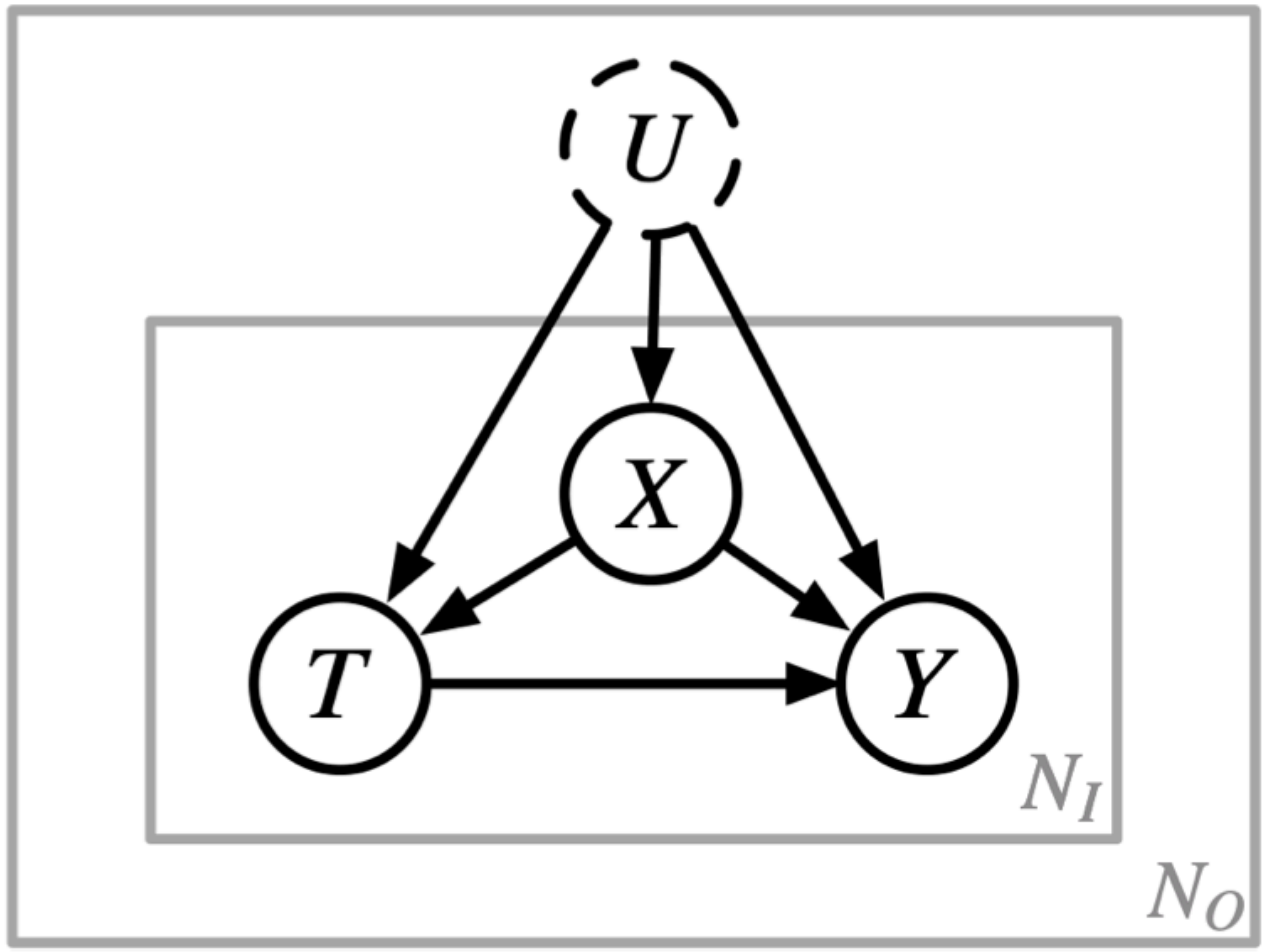
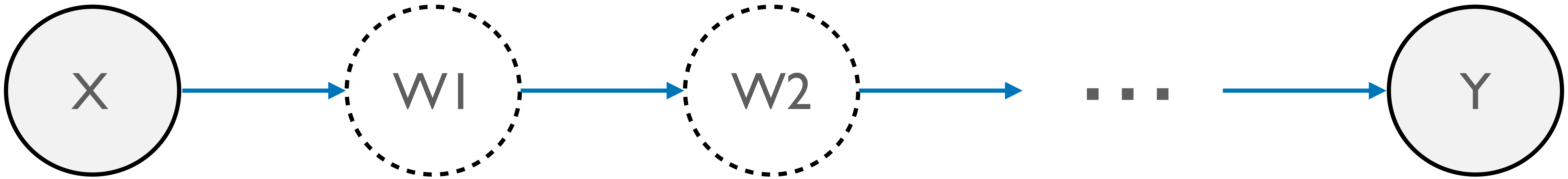


Image Courtesy of Witty et al., (2020)

Deep Gaussian processes

- Previous composite GPs are introduced to match variable relationships.

- Deep Gaussian processes composite a serial of GPs to increase the model flexibility.



What are ongoing research directions?

- How to efficiently characterize posterior correlations between GPs ?
  - Global inducing point variational posteriors
- Each GP in the composite usually has multiple outputs. How to design the multi-output GP and parameterize the multi-output variational posterior ?
  - Matrix-variate Gaussian posteriors

Gaussian Processes

GP Inferences
using inducing points

Composite GPs

Inducing Points
Beyond GPs

Data Summarizations

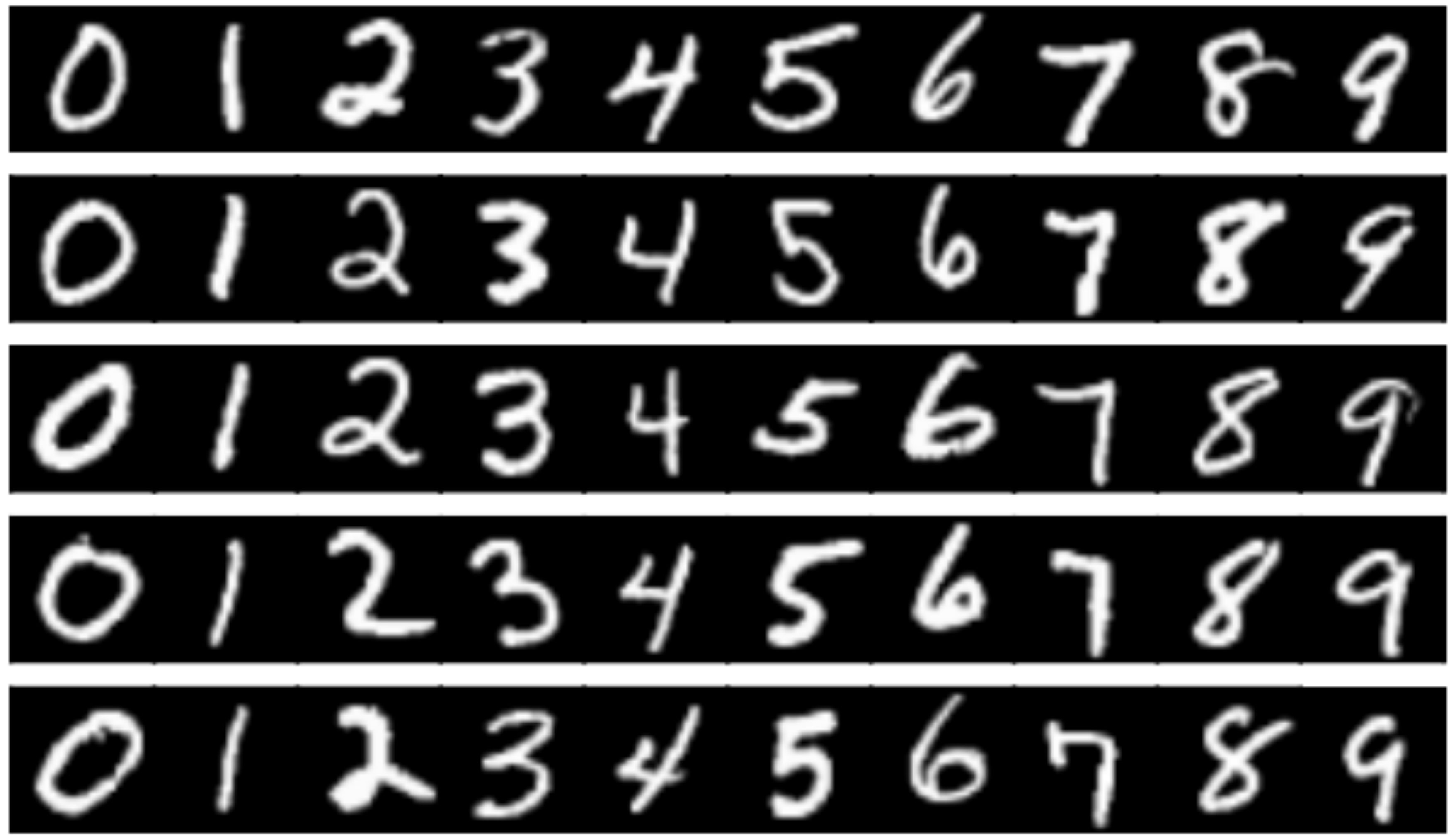
- Data summarization searches for a small set representative of a large dataset
  - Lower storage burden, Lower computational costs

- The GP interpretation naturally provides a criterion for data summarization: selecting the inducing points for the best posterior approximation.

$$\min_{\mathbf{Z} \in \mathcal{X}^m} \operatorname{trace}(\mathbf{K}(\mathbf{X}, \mathbf{X}) - \mathbf{K}(\mathbf{X}, \mathbf{Z})\mathbf{K}(\mathbf{Z}, \mathbf{Z})^{-1}\mathbf{K}(\mathbf{Z}, \mathbf{X}))$$

Data Summarizations

Random Points



Optimized Inducing Points

Image Courtesy of Titsias et al., (2019)

Function Approximations

- Function-space-distance regularization is an "impractical"golden-standard in continual learning, which regularizes the predictor's outputs on all seen data points.

$$\frac{1}{n}\sum_{i=1}^{n}\left(f(\mathbf{x}_i, \boldsymbol{\theta}) - f(\mathbf{x}_i, \boldsymbol{\theta}_0)\right)^2$$

- The storage constraint allows to keep a small set of points $\mathbf{Z} = \mathbf{z}_{1:m}$, then the function-space-distance is approximated by the subsampling estimation.

$$\frac{1}{m}\sum_{i=1}^{m}\left(f(\mathbf{z}_i; \boldsymbol{\theta}) - f(\mathbf{z}_i; \boldsymbol{\theta}_0)\right)^2$$

Function Approximations

- Assume the function is distributed as a Gaussian processes,

$$f(\mathbf{x};\boldsymbol{\theta}) \sim \mathcal{GP}(f(\mathbf{x};\boldsymbol{\theta}_0), k(\mathbf{x},\mathbf{x}'))$$

- The GP assumption allows to estimate $f(\mathbf{x};\boldsymbol{\theta})$ using $f(\mathbf{Z};\boldsymbol{\theta})$. Specifically, it is Gaussian distributed with the mean in the following expression,

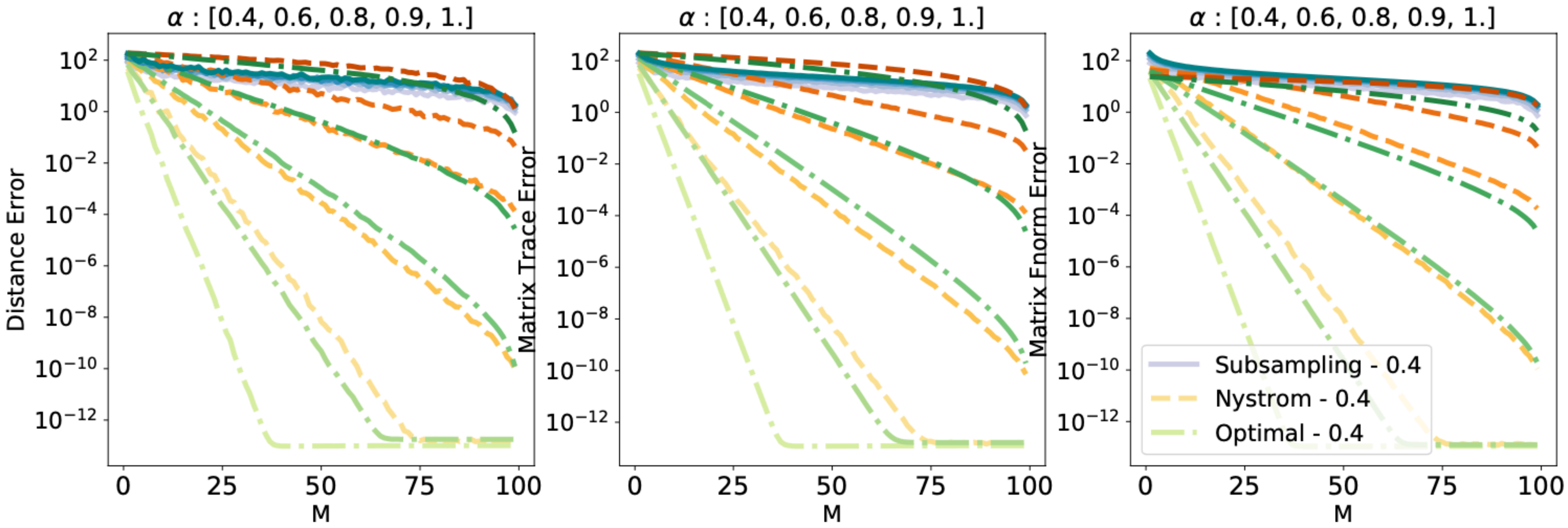
$$\hat{f}(\mathbf{x};\boldsymbol{\theta}) = f(\mathbf{x};\boldsymbol{\theta}_0) + k(\mathbf{x},\mathbf{Z})k(\mathbf{Z},\mathbf{Z})^{-1}\left(f(\mathbf{Z};\boldsymbol{\theta}) - f(\mathbf{Z};\boldsymbol{\theta}_0)\right)$$

- We can use $\hat{f}(\mathbf{x};\boldsymbol{\theta})$ to estimate the function-space-distance,

$$\frac{1}{n}\sum_{i=1}^{n}\left(f(\mathbf{x}_i;\theta) - f(\mathbf{x}_i;\theta_0)\right)^2 \approx \frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}(\mathbf{x}_i;\theta) - f(\mathbf{x}_i;\theta_0)\right)^2$$

$$= \left(f(\mathbf{Z};\theta) - f(\mathbf{Z};\theta_0)\right)^\top \mathbf{G}\left(f(\mathbf{Z};\theta) - f(\mathbf{Z};\theta_0)\right)$$

Function Approximations



How each method responds to the spectral decay of the input distribution?

A small set might contain a lot of information.

References

1. Titsias, M. (2009). Variational learning of inducing variables in sparse Gaussian processes. In Artificial Intelligence and Statistics, pages 567–574.

2. Hensman, J., Matthews, A., and Ghahramani, Z. (2015). Scalable variational Gaussian process classification. In *Artificial Intelligence and Statistics*, pages 351–360.

3. Hensman, J., Matthews, A. G. D. G., Filippone, M., & Ghahramani, Z. (2015). MCMC for variationally sparse Gaussian processes. *arXiv preprint arXiv:1506.04000*.

Appendix

MCMC using Inducing Points

- Can we similarly use inducing points for MCMC ?
- We look at the optimal variational distribution under inducing points.

$$q^\star \in \arg\min_q \mathrm{KL}[q(\mathbf{u})p(f|\mathbf{u}) \| p(f, \mathbf{u}|\mathcal{D})]$$

- The log density of the optimal variational distribution has the expression [3],

$$\log q^\star(\mathbf{u}) = \mathbb{E}_{p(\mathbf{u})p(f|\mathbf{u})}[\log p(\mathcal{D}|f, \mathbf{u})] + \log p(\mathbf{u}) + const$$

stochastic estimations ?   cubic of m computations ✓

MCMC using Inducing Points

- Can we similarly use inducing points for MCMC ?
- We look at the optimal variational distribution under inducing points.

$$q^\star \in \arg\min_q \mathrm{KL}[q(\mathbf{u})p(f|\mathbf{u}) \| p(f, \mathbf{u}|\mathcal{D})]$$

- The log density of the optimal variational distribution has the expression [3],

$$\log q^\star(\mathbf{u}) = \mathbb{E}_{p(\mathbf{u})p(f|\mathbf{u})}[\log p(\mathcal{D}|f, \mathbf{u})] + \log p(\mathbf{u}) + const$$

stochastic estimations ? cubic of m computations ✓

- We can obtain samples of $\mathbf{u}$ using MCMC.
- How to select/optimize the inducing locations $\mathbf{z}_{1:m}$ remains unclear.

Inferences using Inducing Points

	Variational Inference	Markov Chain Monte Carlo
Exact Posterior	✗	✗
Optimal Variational Distribution $q(\mathbf{u})$	✗	✓
Optimizing Inducing points $\mathbf{z}_{1:m}$	✓	?
Stochastic Optimizations	✓	?

MCMC using Inducing Points

- Can we similarly use inducing points for MCMC ?
- We look at the optimal variational distribution under inducing points.

$$q^\star \in \arg\min_q \mathrm{KL}[q(\mathbf{u})p(f|\mathbf{u})\|p(f,\mathbf{u}|\mathcal{D})]$$

$$\begin{aligned}
\mathrm{KL}[q(\mathbf{u})p(f|\mathbf{u})\|p(f,\mathbf{u}|\mathcal{D})] &= \mathbb{E}_{q(\mathbf{u})p(f|\mathbf{u})}[\log \frac{q(\mathbf{u})p(f|\mathbf{u})p(\mathcal{D})}{p(\mathbf{u})p(f|\mathbf{u})p(\mathcal{D}|f,\mathbf{u})}] \\
&= \mathbb{E}_{q(\mathbf{u})p(f|\mathbf{u})}[\log \frac{q(\mathbf{u})p(\mathcal{D})}{p(\mathbf{u})p(\mathcal{D}|f,\mathbf{u})}] \\
&= \mathbb{E}_{q(\mathbf{u})}[\log \frac{q(\mathbf{u})p(\mathcal{D})}{p(\mathbf{u})\exp\left(\mathbb{E}_{p(\mathbf{u})p(f|\mathbf{u})}[\log p(\mathcal{D}|f,\mathbf{u})]\right)}]
\end{aligned}$$

What are ongoing research directions?

- How to efficiently characterize posterior correlations between GPs ?
  - Global inducing point variational posteriors
- Each GP in the composite usually has multiple outputs. How to design the multi-output GP and parameterize the multi-output variational posterior ?
  - Matrix-variate Gaussian posteriors
- Running MCMC with inducing points requires computing the expected log likelihood and the KL divergence. For a single GP, the expected log likelihood can be approximated using Quadratures. For composite GPs, a serial of expectations are involved, how to estimate it accurately, or to enable stochastic estimations ?
  - Stochastic Gradient HMC

Connections to Neural Networks

- The predictive mean of a variational GP and a two-layer NN have similar expressions,

Predictive mean of Sparse GP

$$\mu(\mathbf{x}) = k(\mathbf{Z}, \mathbf{x})^\top \mathbf{K}_{\mathbf{zz}}^{-1} \mathbf{m}$$

Two-Layer Neural Networks

$$f(\mathbf{x}) = \sigma(\mathbf{W}\mathbf{x})^\top \mathbf{a}$$

Nonlinear — Linear

Connections to Neural Networks

- Interpreting each hidden unit of the NN as an inter-domain inducing point of the GP,

Predictive mean of Sparse GP

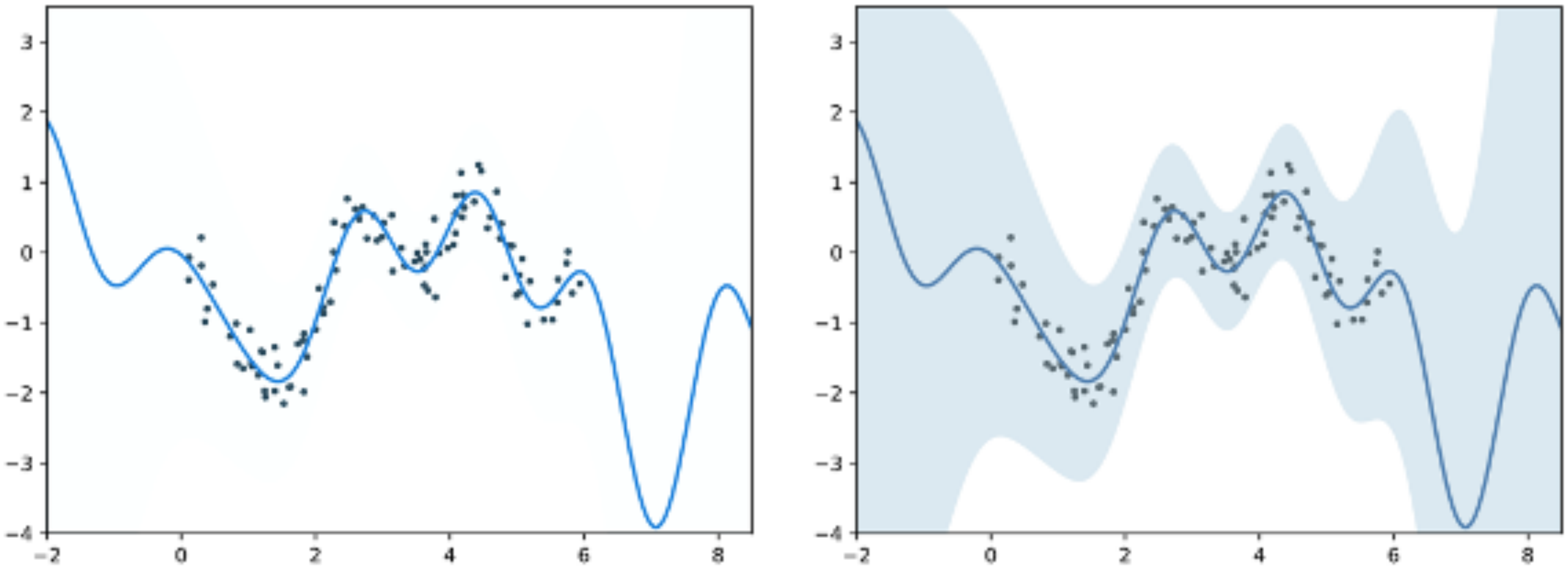
$$\mu(\mathbf{x}) = k(\mathbf{Z}, \mathbf{x})^\top \mathbf{K}_{\mathbf{zz}}^{-1} \mathbf{m}$$

$$\sigma(\mathbf{w}_i^\top \mathbf{x}) = k(\mathbf{z}_i, \mathbf{x})$$

Two-Layer Neural Networks

$$f(\mathbf{x}) = \sigma(\mathbf{W}\mathbf{x})^\top \mathbf{a}$$

Connections to Neural Networks



Generating uncertainty from a pos-trained deterministic neural network