Variational Inference for Gaussian processes

Hugh Salimbeni

4th year PhD with Marc

My research

NeurIPS 2017

Doubly Stochastic Variational Inference for Deep Gaussian Processes

Hugh Salimbeni
Imperial College London and PROWLER.io
hrs13@ic.ac.uk

Marc Peter Deisenroth
Imperial College London and PROWLER.io
m.deisenroth@imperial.ac.uk

AISTATS 2018

Natural Gradients in Practice: Non-Conjugate Variational Inference in Gaussian Process Models

Hugh Salimbeni
Imperial College London, PROWLER.io

Stefanos Eleftheriadis
PROWLER.io

James Hensman
PROWLER.io

All based on the material in this lecture

NeurIPS 2018

Gaussian Process Conditional Density Estimation

Vincent Dutordoir[*1] Hugh Salimbeni[*1,2] Marc Peter Deisenroth[1,2] James Hensman[1]
[1]PROWLER.io, Cambridge, UK [2]Imperial College London
{vincent, hugh, marc, james}@prowler.io

NeurIPS 2018

Orthogonally Decoupled Variational Gaussian Processes

Hugh Salimbeni[*]
Imperial College London
hrs13@ic.ac.uk

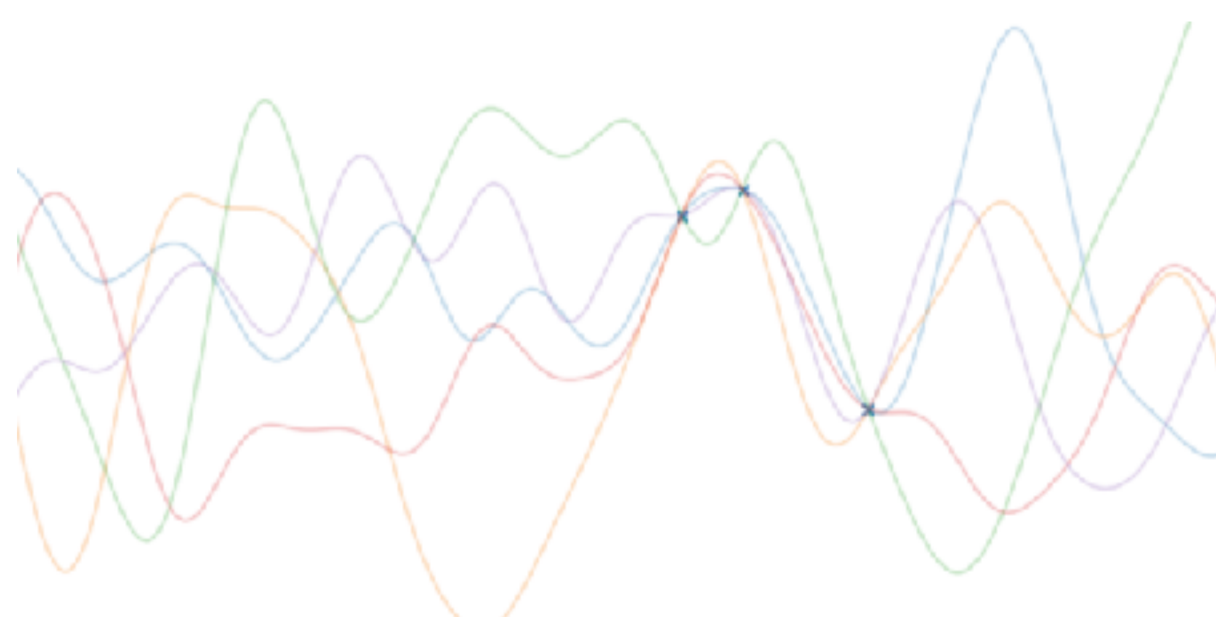
Ching-An Cheng[*]
Georgia Institute of Technology
cacheng@gatech.edu

Byron Boots
Georgia Institute of Technology
bboots@gatech.edu

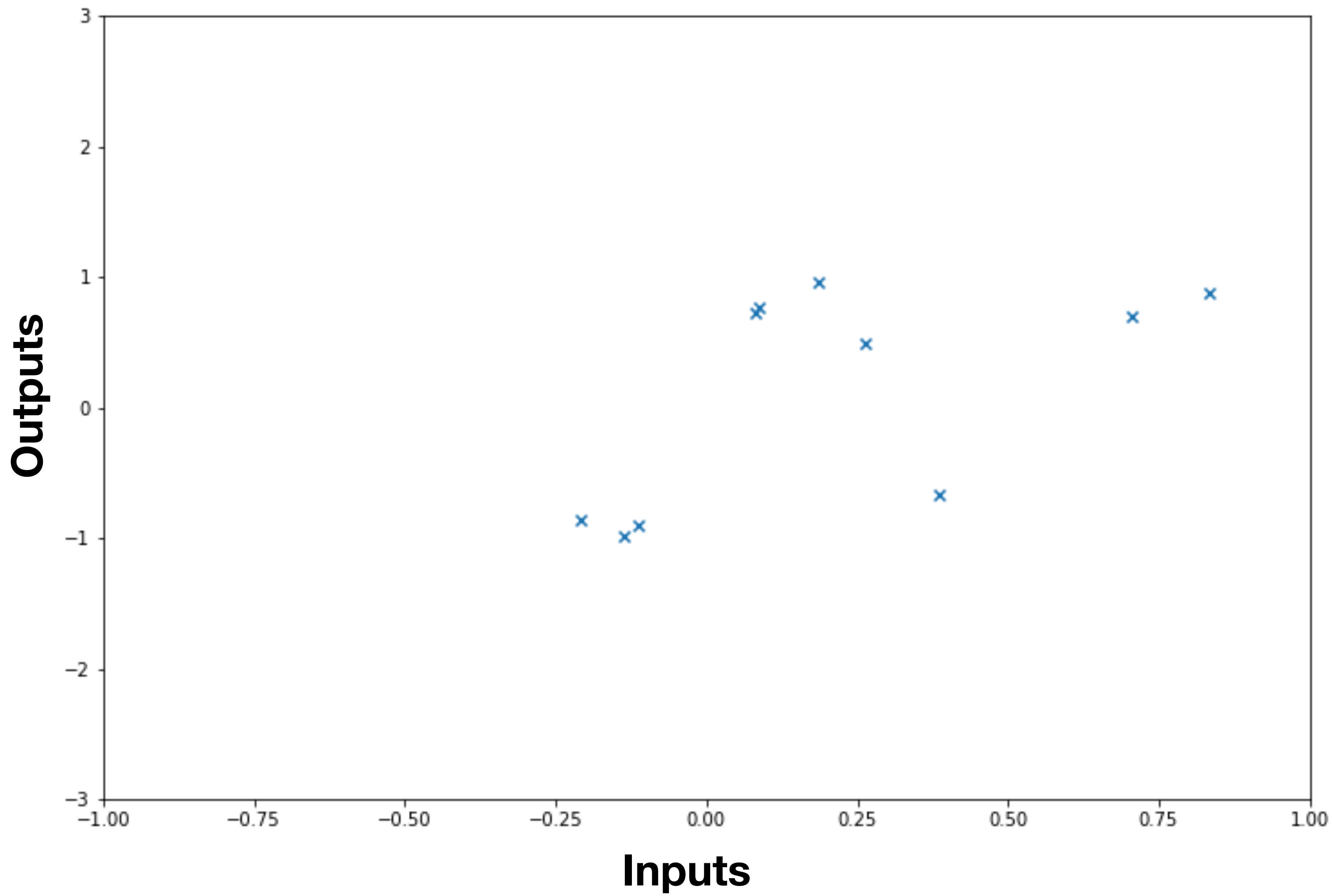
Marc Deisenroth
Imperial College London
mpd37@ic.ac.uk

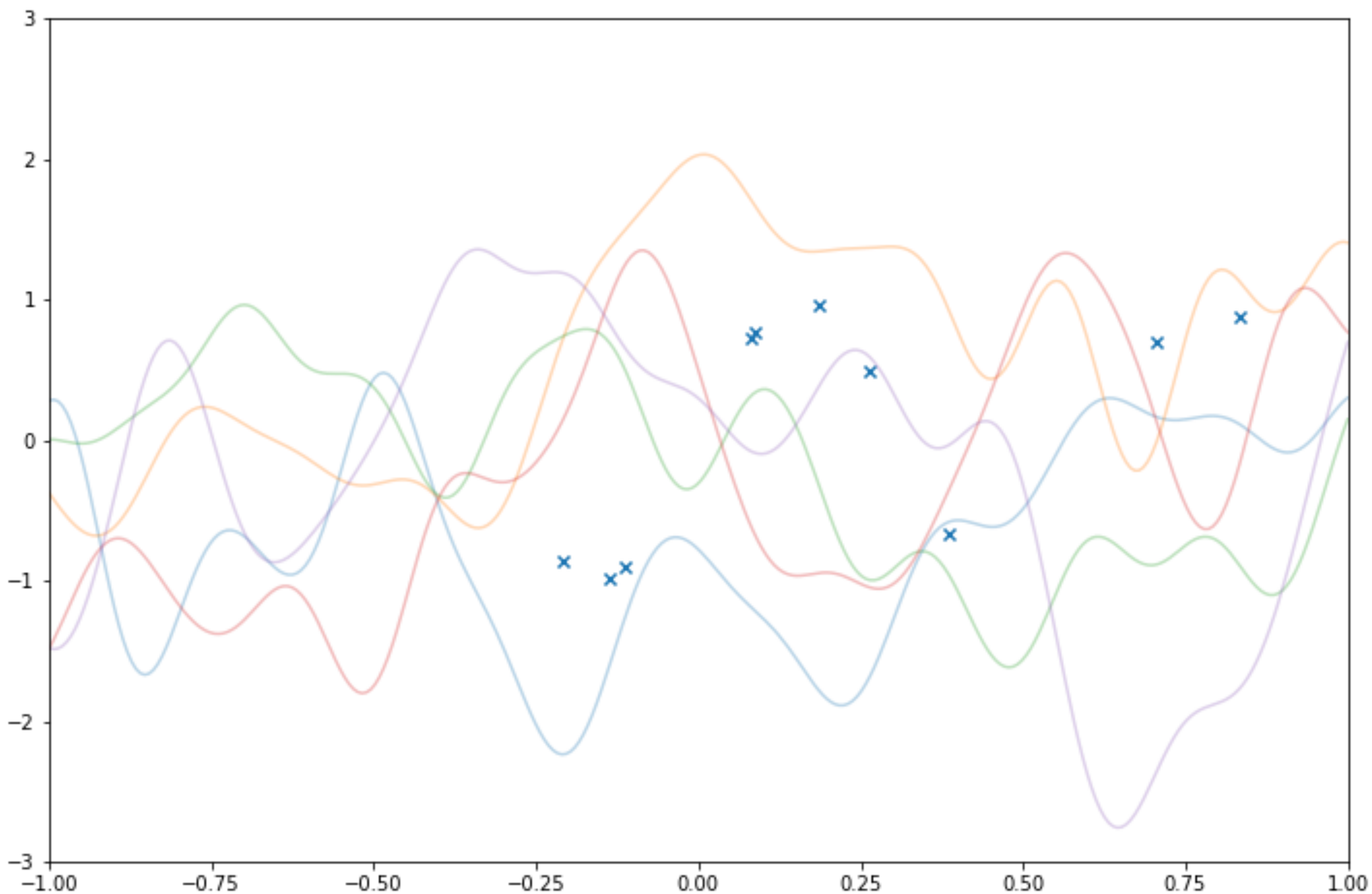
Overview

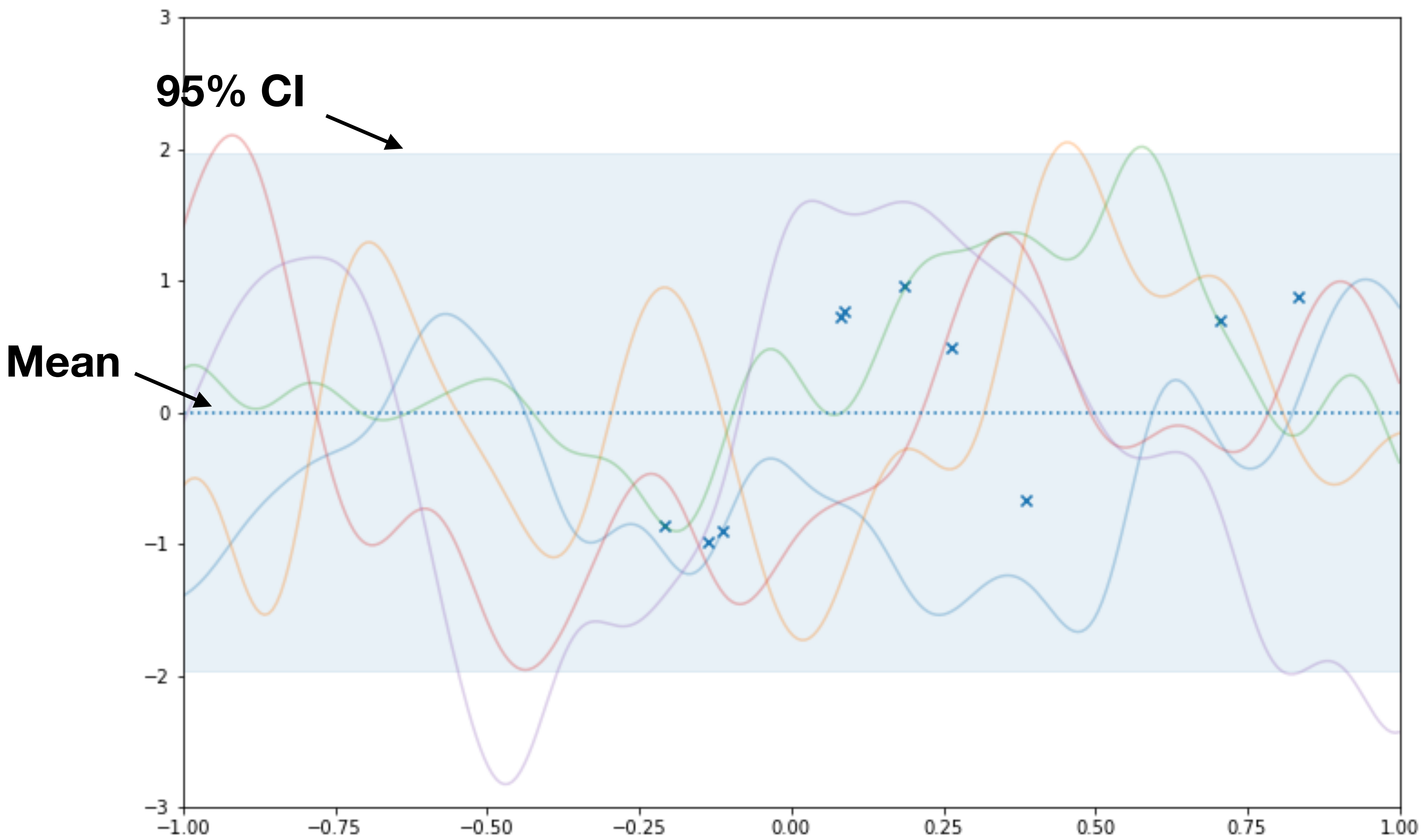
- **Review GPs and VI**
- Establish what problems we want to solve
- Discuss alternative approaches
- VI for GPs part 1 (conjugacy)
- VI for GPs part 2 (scalability)
- Deep GPs



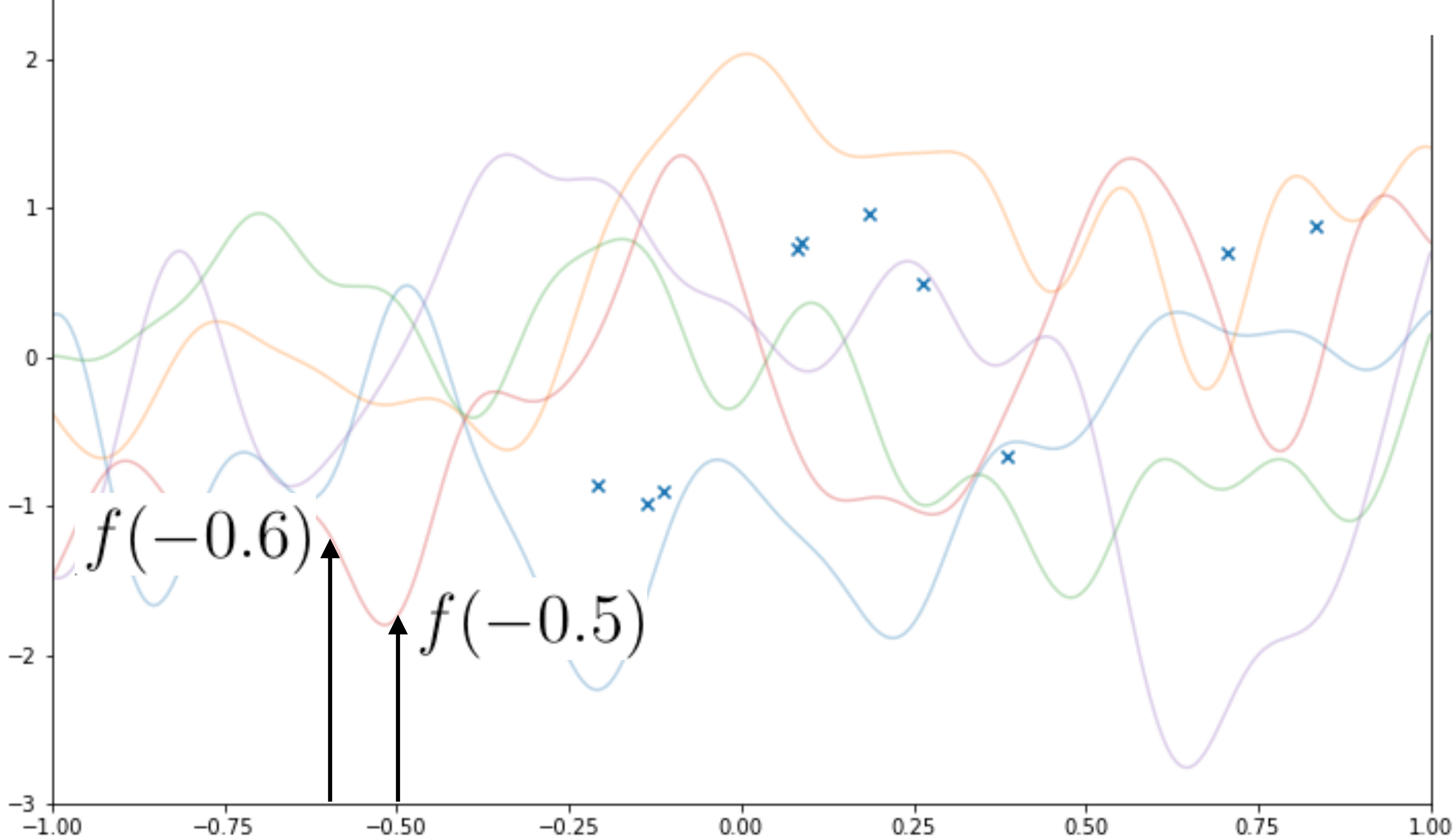
Recap: GPs



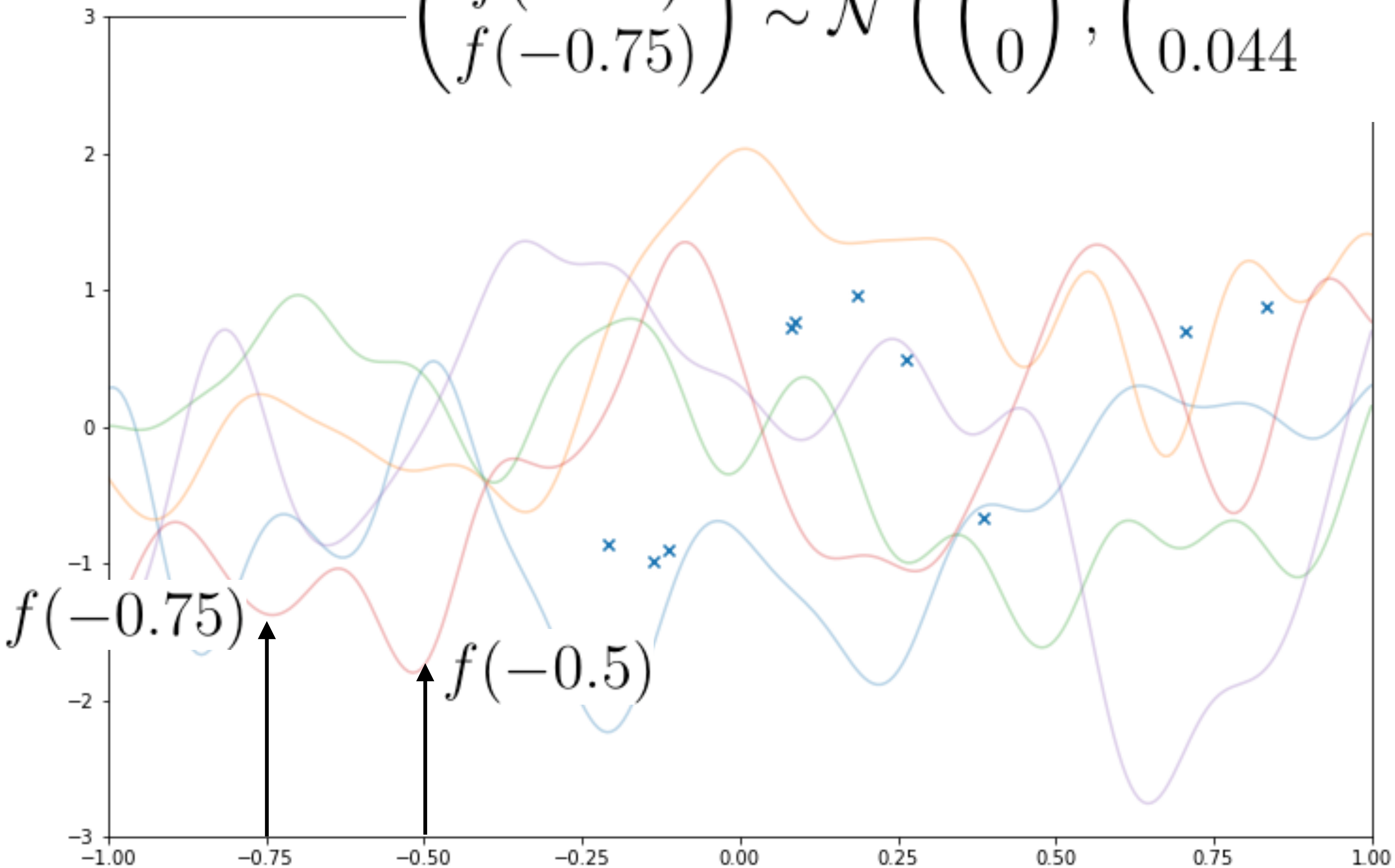

3
2
1
0
−1
−2
−3
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00


95% CI
Mean
3
2
1
0
−1
−2
−3
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00

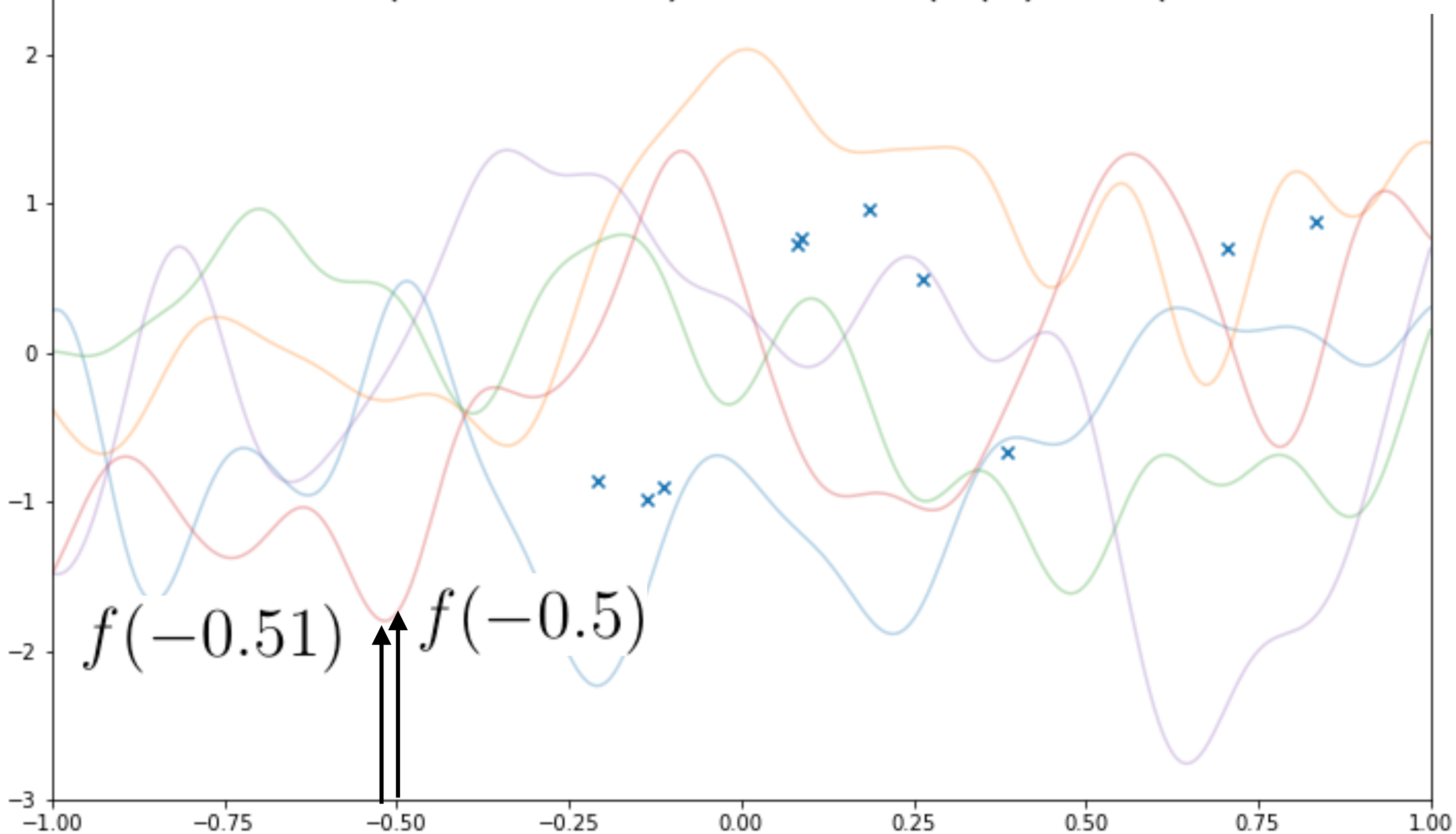
$$\begin{pmatrix} f(-0.5) \\ f(-0.6) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.61 \\ 0.61 & 1 \end{pmatrix} \right)$$



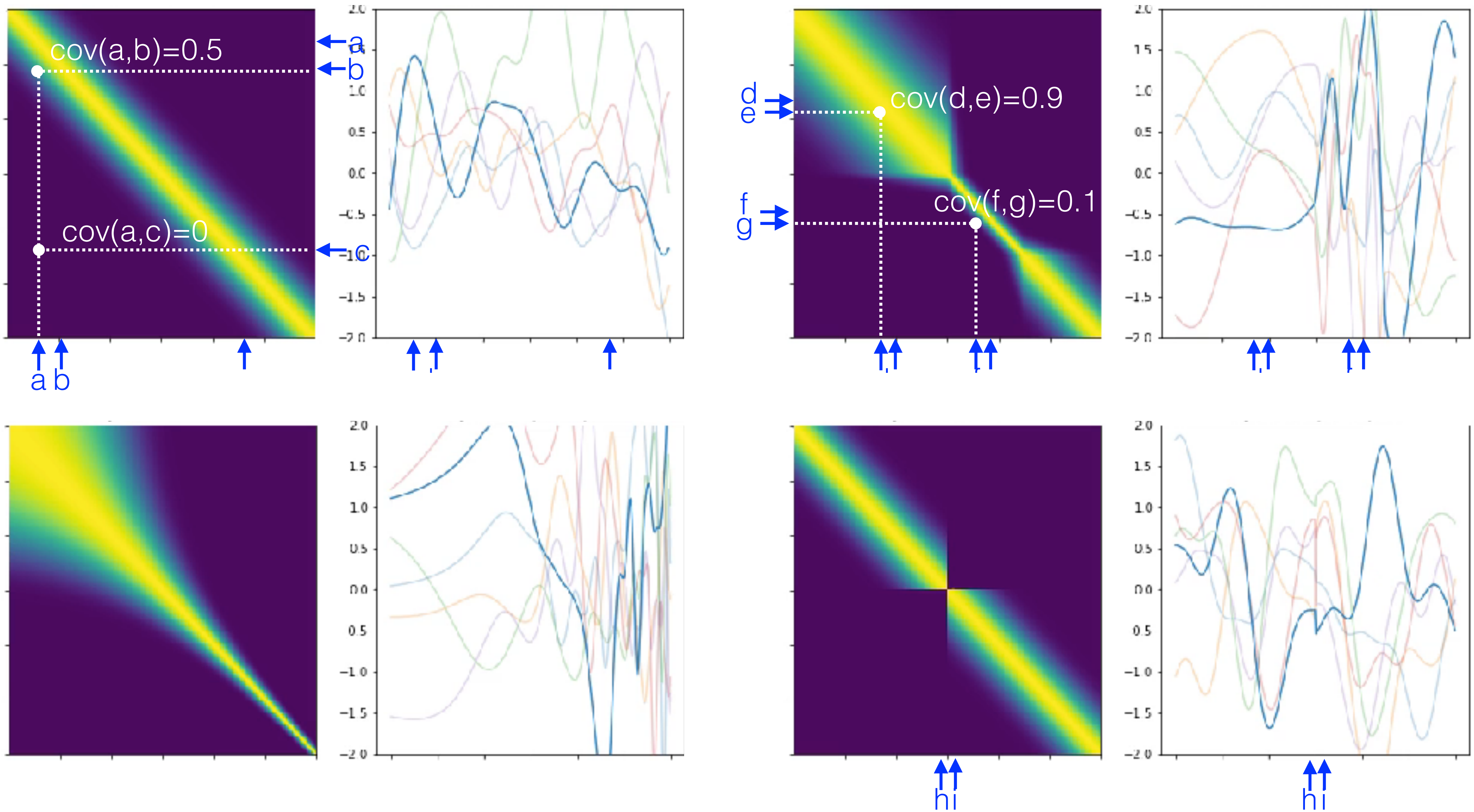
$$\begin{pmatrix} f(-0.5) \\ f(-0.75) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.044 \\ 0.044 & 1 \end{pmatrix} \right)$$



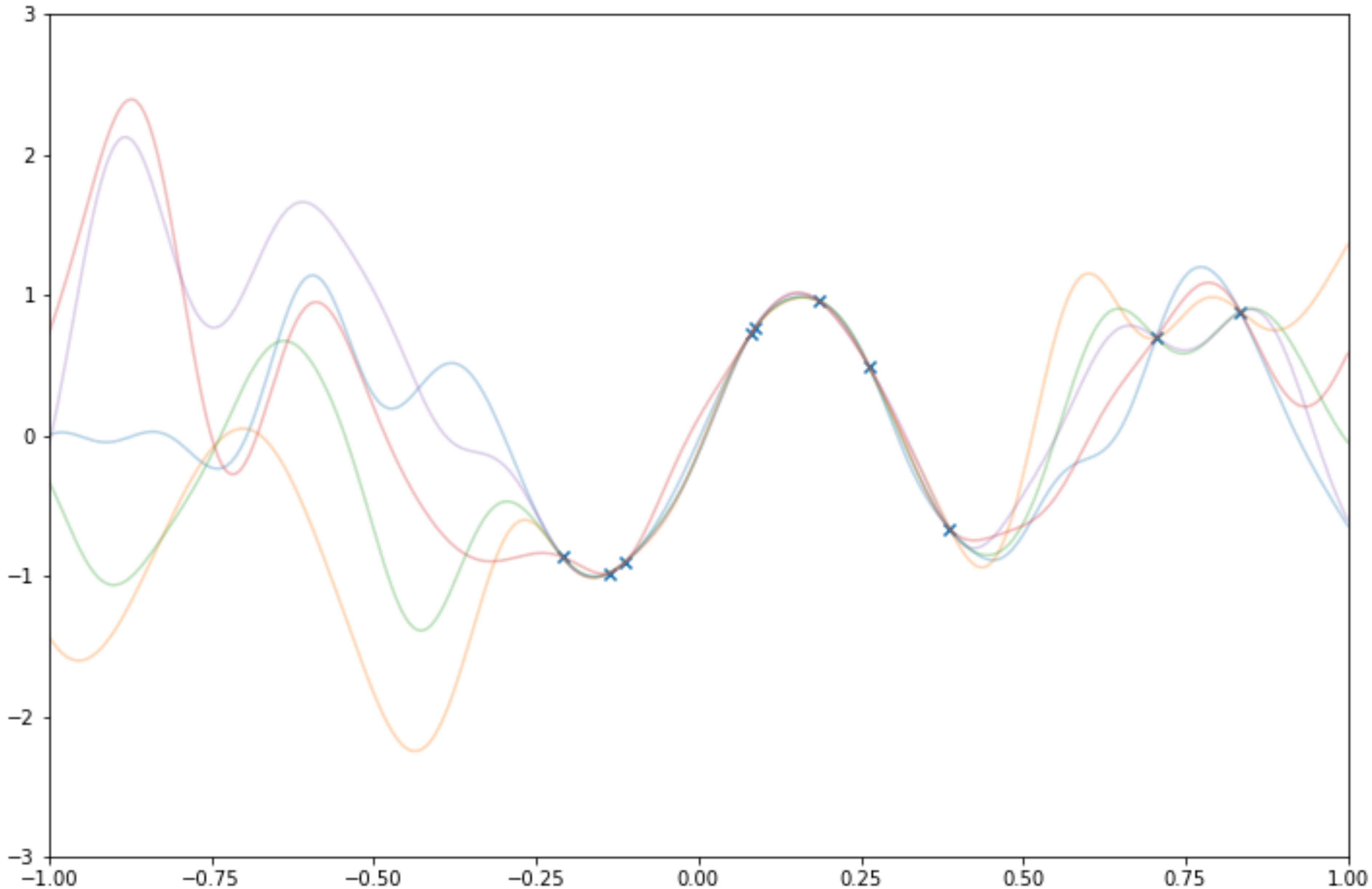
$$\begin{pmatrix} f(-0.5) \\ f(-0.51) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.995 \\ 0.995 & 1 \end{pmatrix} \right)$$

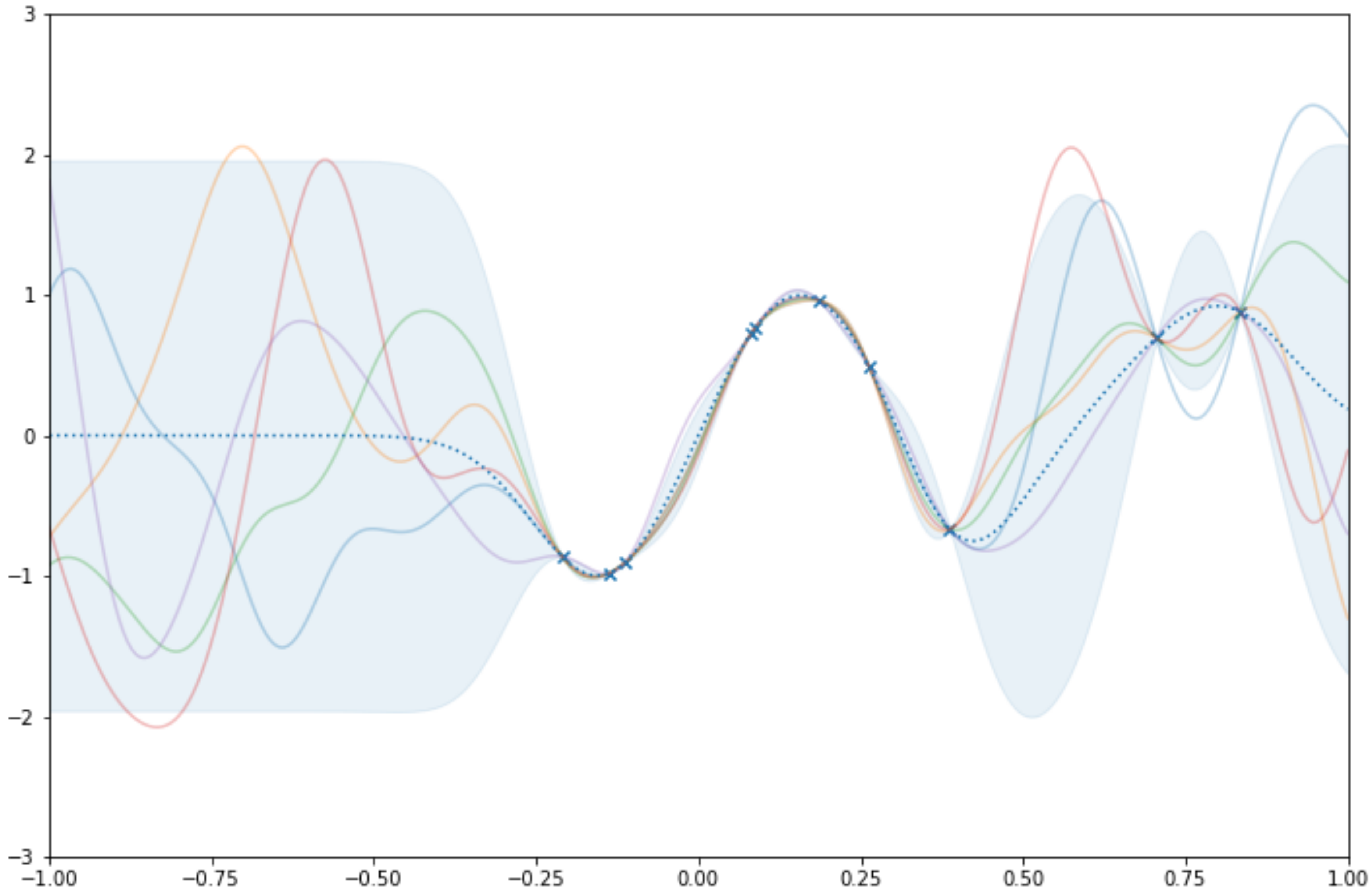


$$\begin{pmatrix} f(x_1) \\ f(x_2) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} m(x_1) \\ m(x_2) \end{pmatrix}, \begin{pmatrix} k(x_1, x_1) & k(x_1, x_2) \\ k(x_2, x_1) & k(x_2, x_2) \end{pmatrix} \right)$$



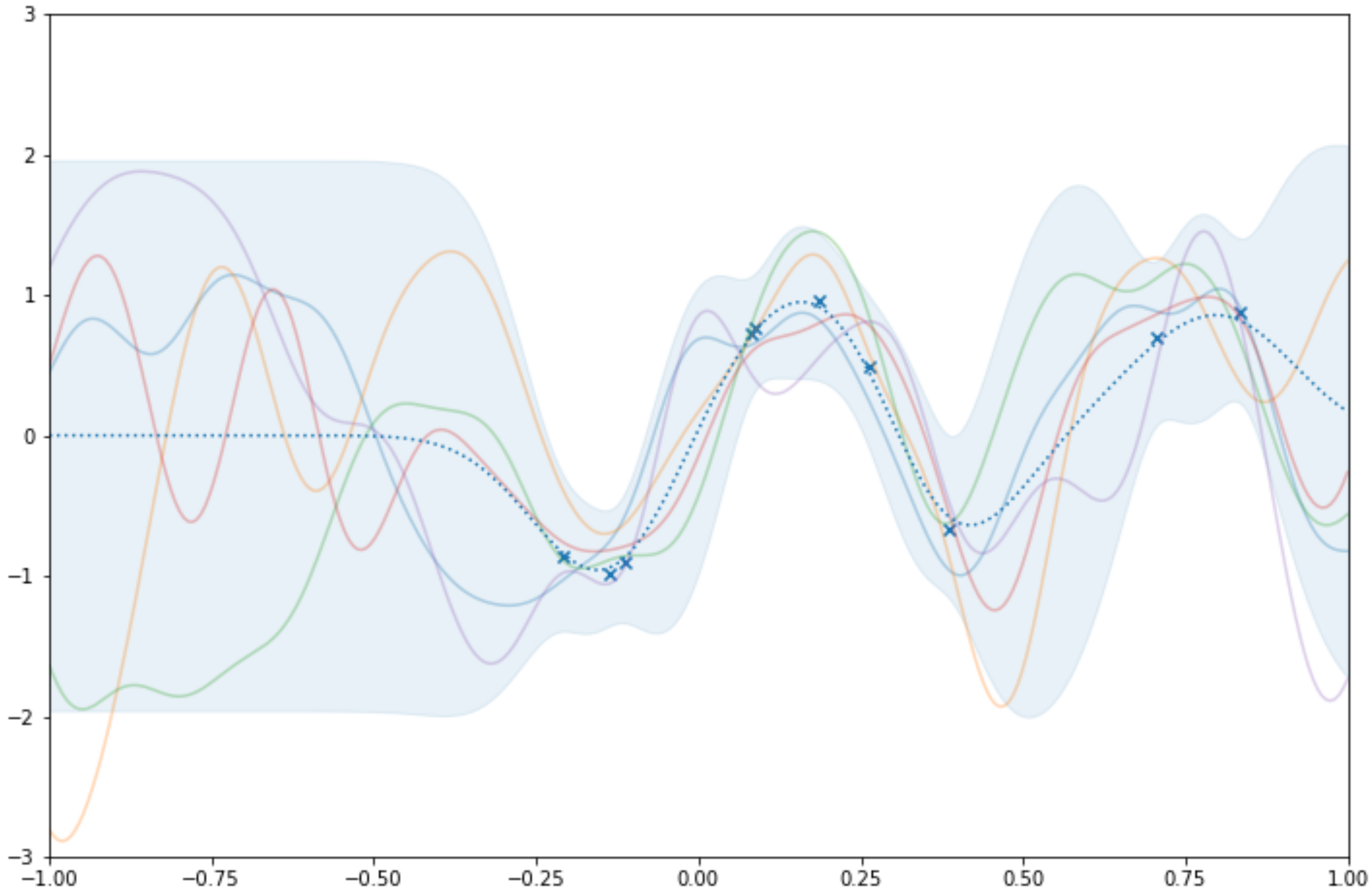
Posterior samples:





3
2
1
0
−1
−2
−3
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00

With noisy observations:



Deriving the posterior

Key ideas:

- Partition the prior

$$p(f) = p(f_* \mid \mathbf{f})p(\mathbf{f})$$

- Write the model as three terms, each of which is Gaussian

$$p(f, \mathbf{y}) = \underbrace{p(f_*|\mathbf{f})}_{\text{projection}} \underbrace{p(\mathbf{f})p(\mathbf{y}|\mathbf{f})}_{\text{data term}}$$

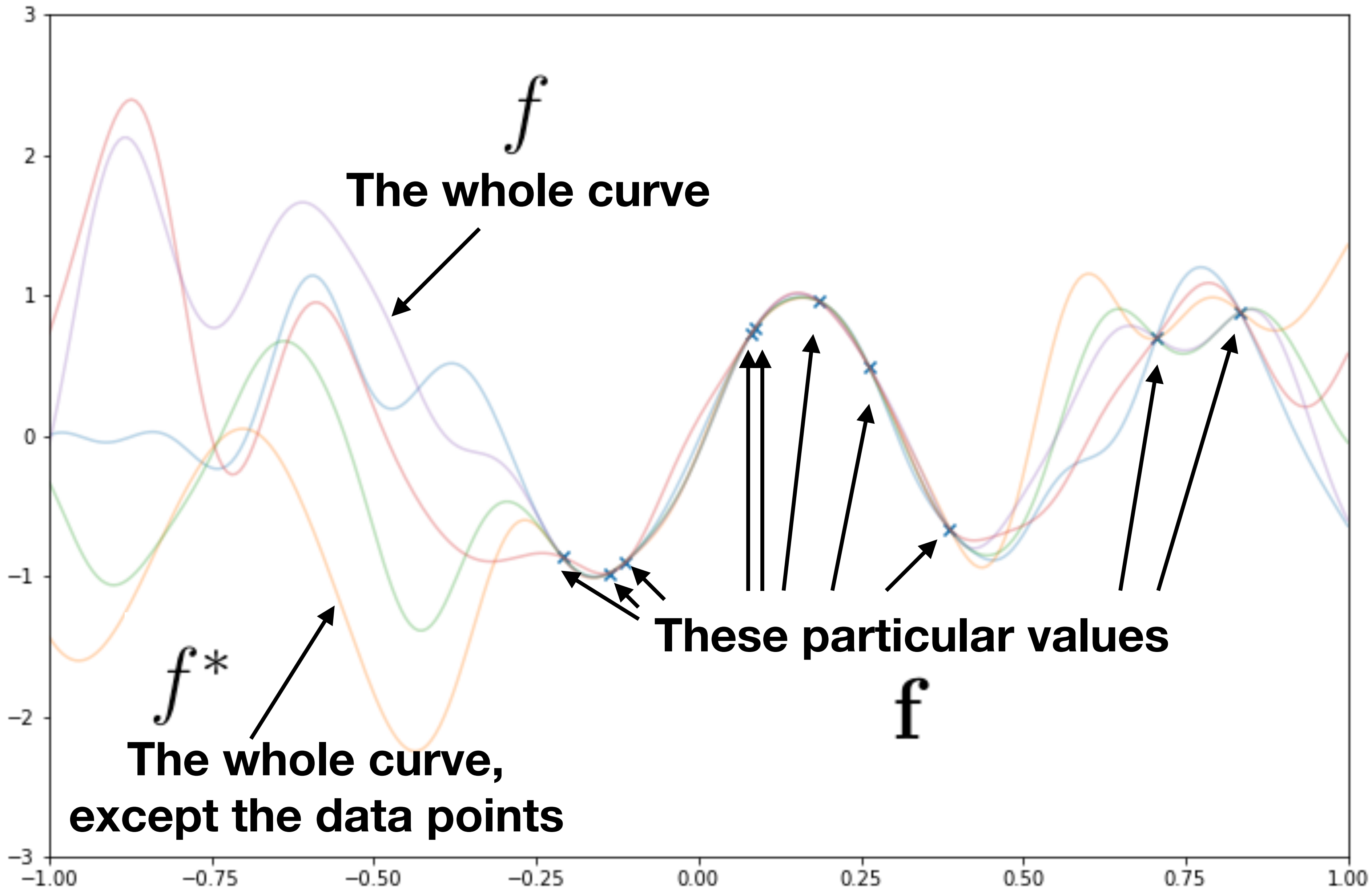
- Use standard results for products of Gaussians

- Integrate out the data variables

NB there are other equivalent ways to derive these results

Some notation

Symbol	Size	Equivalent to	Interpretation
$f(x)$	1	$f(x)$	A single function value
f	∞	$\{f(x) \mid x \in \mathbb{R}\}$	The entire function
$\mathbf{f}$	N	$\{f(x_n) \mid n = 1, \ldots, N\}$	The function values at the data x_n
f_*	∞	$f \setminus \mathbf{f}$	All the function values that are not in $\mathbf{f}$



Symbol	Num elements	Equivalent to	Interpretation
$f(x)$	1	$f(x)$	A single function value
f	∞	$\{f(x) \mid x \in \mathbb{R}\}$	The entire function
$\mathbf{f}$	N	$\{f(x_n) \mid n = 1, \ldots, N\}$	The function values at the data x_n
f^*	∞	$f \setminus \mathbf{f}$	All the function values that are not in $\mathbf{f}$

The model

Prior Likelihood

$$p(f, \{y_n, x_n\}_{n=1}^N) = p(f) \prod_{n=1}^N p(y_n \mid f(x_n))$$

Vector form for the likelihood $\prod_{n=1}^N p(y_n \mid f(x_n)) = p(\mathbf{y}|\mathbf{f}) = \mathcal{N}(\mathbf{y} \mid \mathbf{f}, \sigma^2 \mathbf{I})$

Vector form for the model $p(f, \mathbf{y}, \mathbf{x}) = p(f)p(\mathbf{y}|\mathbf{f})$

Variable partitions

$$p(f) = p(f_* \mid \mathbf{f})p(\mathbf{f})$$

$$p(\mathbf{f}) = \mathcal{N}(\mathbf{f}|\mathbf{0}, \mathbf{K})$$

$$\begin{aligned} p(f_*|\mathbf{f}) &= \mathcal{GP}(\mu, \Sigma) \\ \mu(x) &= \mathbf{k}(x)^\top \mathbf{K}^{-1}\mathbf{f} \\ \Sigma(x, x') &= k(x, x') - \mathbf{k}(x)^\top \mathbf{K}^{-1}\mathbf{k}(x') \end{aligned}$$

Symbol	Size	Equivalent to	Interpretation
$\mathbf{k}(x)$	N	$\{k(x, x_n) \mid n = 1, \ldots, N\}$	Covariance between a test point and the data
$\mathbf{K}$	N, N	$\{k(x_i, x_j) \mid i, j = 1, \ldots, N\}$	Covariance between data points

Standard result #1: conditioning

$$\mathcal{N}\left(\begin{pmatrix} a \\ b \end{pmatrix} \mid \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}, \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}\right) =$$

$$\mathcal{N}(a|\mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(b - \mu_b), \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}).\ \mathcal{N}(b|\mu_b, \Sigma_{bb})$$

Standard result #2a: product of two Gaussians

$$\mathcal{N}(a|\mu_a, \Sigma_a)\mathcal{N}(a|\mu_b, \Sigma_b) =$$

$$\mathcal{N}(a|\Lambda\left(\Sigma_a^{-1}\mu_a + \Sigma_b^{-1}\mu_b\right), \Lambda)\ \mathcal{N}(\mu_a|\mu_b, \Sigma_a + \Sigma_b)$$

$$\Lambda^{-1} = \Sigma_a^{-1} + \Sigma_b^{-1}$$

Standard result #2b: product of two Gaussians

$$\mathcal{N}(Aa|\mu_a, \Sigma_a)\mathcal{N}(a|\mu_b, \Sigma_b) =$$

$$\mathcal{N}(a|\Lambda\left(A^\top\Sigma_a^{-1}\mu_a + \Sigma_b^{-1}\mu_b\right), \Lambda)\mathcal{N}(\mu_a|A\mu_b, \Sigma_a + A\Sigma_b A^\top)$$

$$\Lambda^{-1} = A^\top\Sigma_a^{-1}A + \Sigma_b^{-1}$$

Variable partitions

$$p(f) = p(f_* \mid \mathbf{f})p(\mathbf{f})$$

$$p(\mathbf{f}) = \mathcal{N}(\mathbf{f}|\mathbf{0}, \mathbf{K})$$

$$\begin{aligned} p(f_*|\mathbf{f}) &= \mathcal{GP}(\mu, \Sigma) \\ \mu(x) &= \mathbf{k}(x)\mathbf{K}^{-1}\mathbf{f} \\ \Sigma(x, x') &= k(x, x') - \mathbf{k}(x)^\top \mathbf{K}^{-1}\mathbf{k}(x') \end{aligned}$$

Symbol	Size	Equivalent to	Interpretation
$\mathbf{k}(x)$	N	$\{k(x, x_n) \mid n = 1, \ldots, N\}$	Covariance between a test point and the data
$\mathbf{K}$	N, N	$\{k(x_i, x_j) \mid i, j = 1, \ldots, N\}$	Covariance between data points

Alternative partitions (for later)

$$p(f) = p(\tilde{f}_* \,|\, \tilde{\mathbf{f}})p(\tilde{\mathbf{f}})$$

$$p(\tilde{\mathbf{f}}) = \mathcal{N}(\tilde{\mathbf{f}} \,|\, \mathbf{0}, \tilde{\mathbf{K}})$$

$$p(\tilde{f}_* | \tilde{\mathbf{f}}) = \mathcal{GP}(\mu, \Sigma)$$

$$\tilde{\mu}(x) = \tilde{\mathbf{k}}(x)^\top \tilde{\mathbf{K}}^{-1} \tilde{\mathbf{f}}$$

$$\tilde{\Sigma}(x, x') = k(x, x') - \tilde{\mathbf{k}}(x)^\top \tilde{\mathbf{K}}^{-1} \tilde{\mathbf{k}}(x')$$

Symbol	Size	Equivalent to	Interpretation
$\boxed{\tilde{\mathbf{f}}}$	$\boxed{M}$	$\{f(\tilde{x}_m) \,|\, n = 1, \ldots, M\}$	Some $\boxed{\text{other function values}}$ we can choose
$\tilde{f}_*$	∞	$f \setminus \tilde{\mathbf{f}}$	All the function values that are not in $\tilde{\mathbf{f}}$
$\tilde{\mathbf{k}}(x)$	M	$\{k(x, \tilde{x}_m) \,|\, m = 1, \ldots, M\}$	Covariance between a test point and the pseudo-data
$\tilde{\mathbf{K}}$	M, M	$\{k(\tilde{x}_i, \tilde{x}_j) \,|\, i, j = 1, \ldots, M\}$	Covariance between pseudo-data

Back to the model

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

$$p(f, \mathbf{y}) = \boxed{p(f_*|\mathbf{f})\; p(\mathbf{f})}\, p(\mathbf{y}|\mathbf{f})$$

$$p(f_*|\mathbf{f}) = \mathcal{GP}(\mu, \Sigma)$$
$$\mu(x) = \mathbf{k}(x)\mathbf{K}^{-1}\mathbf{f}$$
$$\Sigma(x, x') = k(x, x') - \mathbf{k}(x)^\top \mathbf{K}^{-1}\mathbf{k}(x')$$

$$p(\mathbf{y}|\mathbf{f}) = \mathcal{N}(\mathbf{y} \mid \mathbf{f}, \sigma^2\mathbf{I})$$

$$p(\mathbf{f}) = \mathcal{N}(\mathbf{f}|\mathbf{0}, \mathbf{K})$$

$$p(f, \mathbf{y}) = \underbrace{p(f_* | \mathbf{f})}_{\text{projection}} \underbrace{p(\mathbf{f}) p(\mathbf{y}|\mathbf{f})}_{\text{data term}}$$

$$p(f, \mathbf{y}) = \mathcal{N}(\mathbf{a}_*^\top \mathbf{f} | f_*, \ldots) \boxed{\mathcal{N}(\mathbf{f}|\mathbf{0}, \mathbf{K}) \mathcal{N}(\mathbf{f}|\mathbf{y}, \sigma^2 \mathbf{I})}$$

$$\mathcal{N}(\mathbf{k}_*^\top \mathbf{K}^{-1} \mathbf{f} | f_*, k_{**} - \mathbf{k}_*^\top \mathbf{K}^{-1} \mathbf{k}_*)$$

$$\boxed{= \mathcal{N}(\mathbf{f}|\ldots, \ldots) \mathcal{N}(\ldots|\ldots, \ldots)}$$

$$\mathcal{N}(\mathbf{f}|\mathbf{0},\mathbf{K})\mathcal{N}(\mathbf{f}|\mathbf{y},\sigma^2\mathbf{I})$$

$$\mathcal{N}(\mathbf{f}\mid\bar{\mathbf{m}},\bar{\mathbf{S}})\mathcal{N}(\mathbf{y}|\mathbf{0},\mathbf{K}+\sigma^2\mathbf{I})$$

$$\begin{aligned}\bar{\mathbf{m}} &= \bar{\mathbf{S}}(\mathbf{K}^{-1}\mathbf{0}+\sigma^{-2}\mathbf{y})\\ \bar{\mathbf{S}} &= (\mathbf{K}^{-1}+\sigma^{-2}\mathbf{I})^{-1}\end{aligned}$$

$$\bar{\mathbf{m}} = \bar{\mathbf{S}}(\mathbf{K}^{-1}\mathbf{0} + \sigma^{-2}\mathbf{y}) = \sigma^{-2}(\mathbf{K}^{-1} + \sigma^{-2}\mathbf{I})^{-1}\mathbf{y}$$
$$= \mathbf{K}(\mathbf{K} + \sigma^{2}\mathbf{I})^{-1}\mathbf{y}$$

$$\bar{\mathbf{S}} = (\mathbf{K}^{-1} + \sigma^{-2}\mathbf{I})^{-1} = \mathbf{K} - \mathbf{K}(\mathbf{K} + \sigma^{2}\mathbf{I})^{-1}\mathbf{K}$$

(Woodbury)

The Woodbury matrix identity is[4]

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U\left(C^{-1} + VA^{-1}U\right)^{-1}VA^{-1}$$

$$p(f, \mathbf{y}) = \mathcal{N}(\mathbf{k}_*^\top \mathbf{K}^{-1}\mathbf{f} | f_*, k_{**} - \mathbf{k}_*^\top \mathbf{K}^{-1}\mathbf{k}_*)\mathcal{N}(\mathbf{f}|\bar{\mathbf{m}}, \bar{\mathbf{S}})\mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K} + \sigma^2\mathbf{I})$$

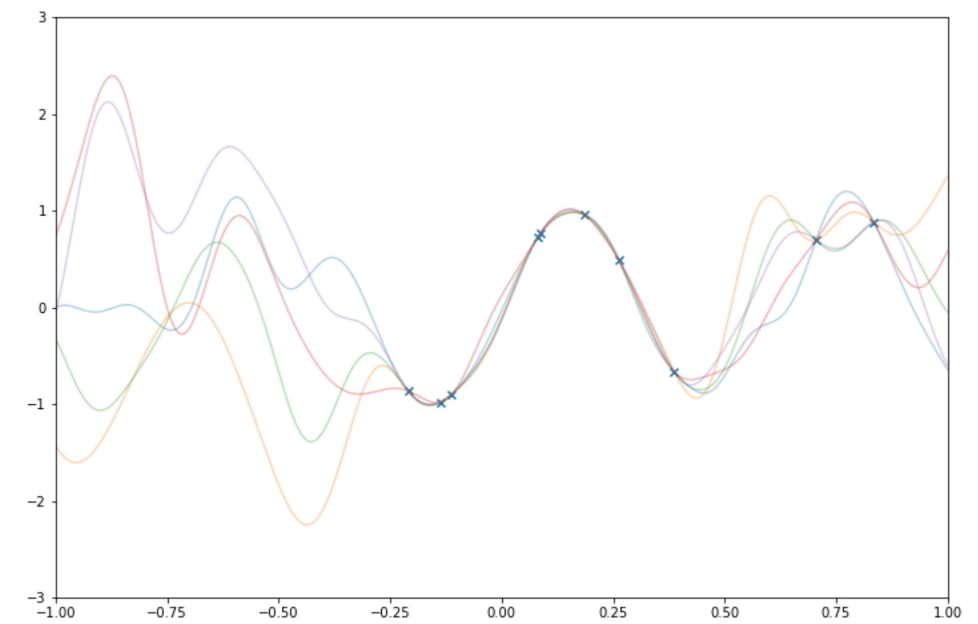
$$= \mathcal{N}(\mathbf{f}|..., ...)\mathcal{N}(f_*|\mathbf{k}_*^\top \mathbf{K}^{-1}\bar{\mathbf{m}}, k_{**} - \mathbf{k}_*^\top \mathbf{K}^{-1}\mathbf{k}_* + \mathbf{k}_*^\top \mathbf{K}^{-1}\bar{\mathbf{S}}\mathbf{K}^{-1}\mathbf{k}_*)\mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K} + \sigma^2\mathbf{I})$$

Posterior

$$\mathcal{N}(f_*|\mathbf{k}_*^\top \mathbf{K}^{-1}\bar{\mathbf{m}}, k_{**} - \mathbf{k}_*^\top \mathbf{K}^{-1}\mathbf{k}_* + \mathbf{k}_*^\top \mathbf{K}^{-1}\bar{\mathbf{S}}\mathbf{K}^{-1}\mathbf{k}_*)$$

$$\bar{\mathbf{m}} = \mathbf{K}(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{y}$$

$$\bar{\mathbf{S}} = \mathbf{K} - \mathbf{K}(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{K}$$



Or equivalently

$$\mathcal{N}(f_*|\mathbf{k}_*^\top(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\bar{\mathbf{y}}, k_{**} - \mathbf{k}_*^\top(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{k}_*)$$

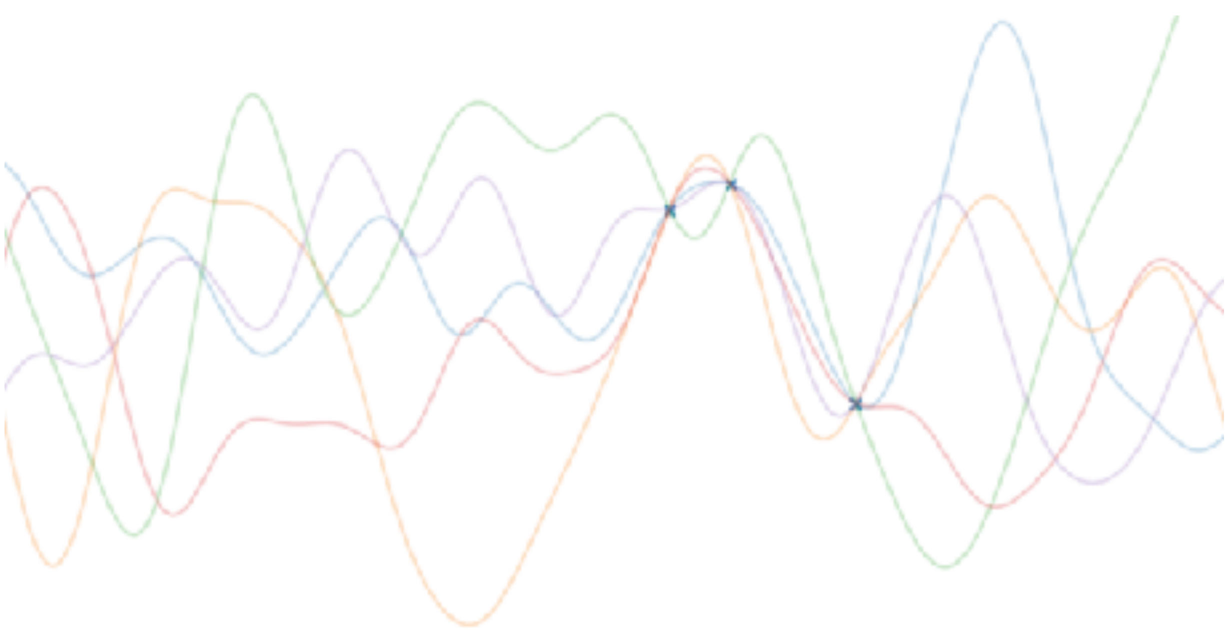
Marginal likelihood

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K} + \sigma^2\mathbf{I})$$

Everything here is N^2 memory and N^3 complexity

Overview

- ~~Review GPs~~ **and VI**
- Establish what problems we want to solve
- Discuss alternative approaches
- VI for GPs part 1 (conjugacy)
- VI for GPs part 2 (scalability)
- Deep GPs



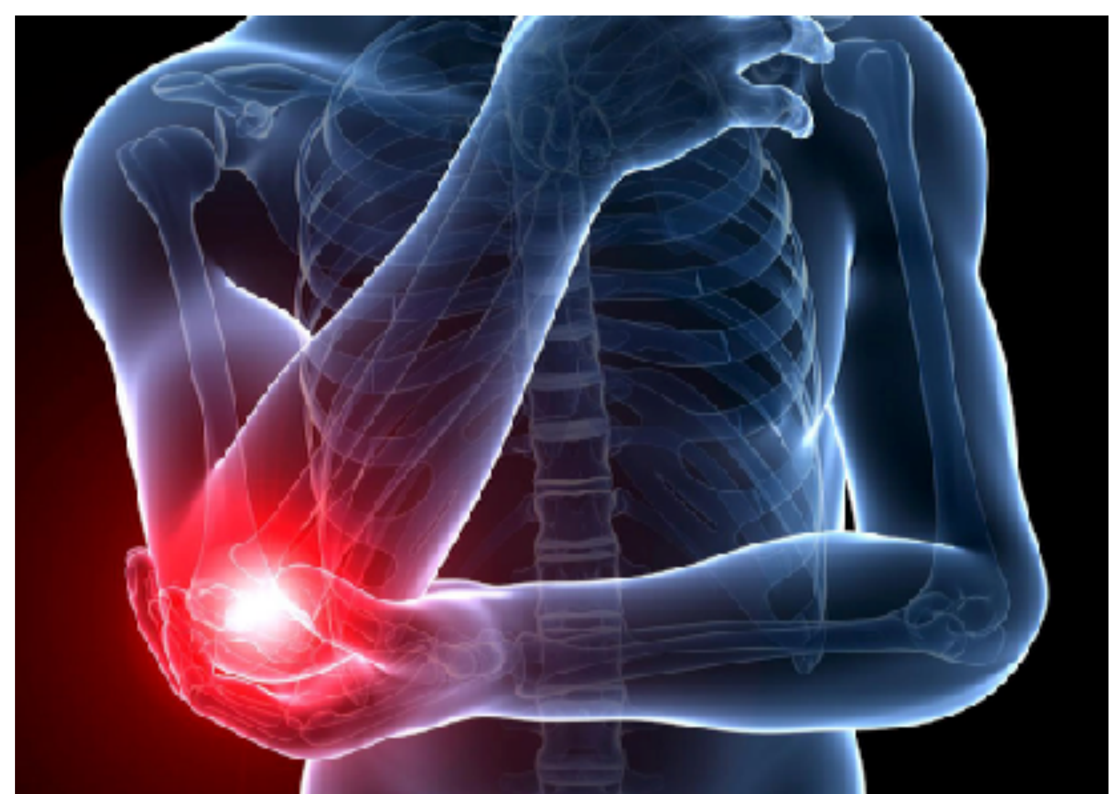
Recap: VI

Key points:

- Make an approximate posterior 'as close as possible' to the true posterior

- 'Closeness' is measured in KL divergence from the approximation to the true posterior

- Turns integration (*hard*) into optimization (*easy*)



Recap: VI (1)

$$\begin{aligned}
\log p(y) &= \mathbb{E}_{q(z)} \log \frac{p(y,z)}{q(z)} + \mathrm{KL}(q(z)||p(z|y)) \\
&= \mathbb{E}_{q(z)} \log \frac{p(y,z)}{q(z)} + \mathbb{E}_{q(z)} \log \frac{q(z)}{p(z|y)} \\
&= \mathbb{E}_{q(z)} \log \frac{p(y,z)}{q(z)} + \mathbb{E}_{q(z)} \left[\log q(z) - \log p(z|y)\right] \\
&= \mathbb{E}_{q(z)} \log \frac{p(y,z)}{q(z)} + \mathbb{E}_{q(z)} \left[\log q(z) - \log \frac{p(y,z)}{p(y)}\right] \\
&= \mathbb{E}_{q(z)} \left[\log p(y,z) + \log q(z) + \log q(z) - \log p(y,z) + \log p(y)\right] \\
&= \mathbb{E}_{q(z)} \log p(y) \\
&= \log p(y)
\end{aligned}$$

Fixed **ELBO** **KL divergence from approximate posterior to true posterior**

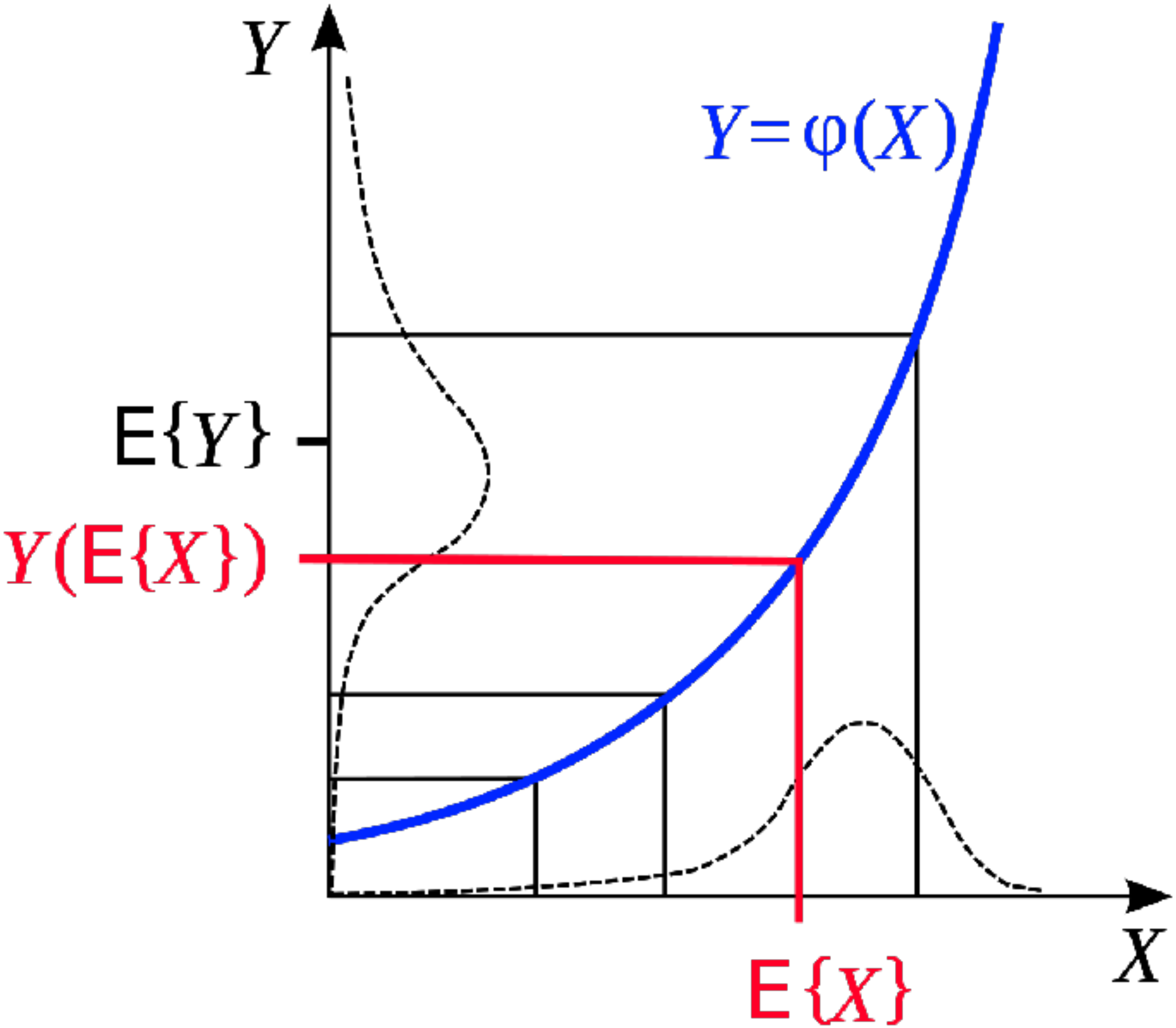
$$\log p(y) = \mathbb{E}_{q(z)} \log \frac{p(y,z)}{q(z)} + \mathrm{KL}(q(z)||p(z|y))$$

Maximize **Minimize**

Recap: VI (2)

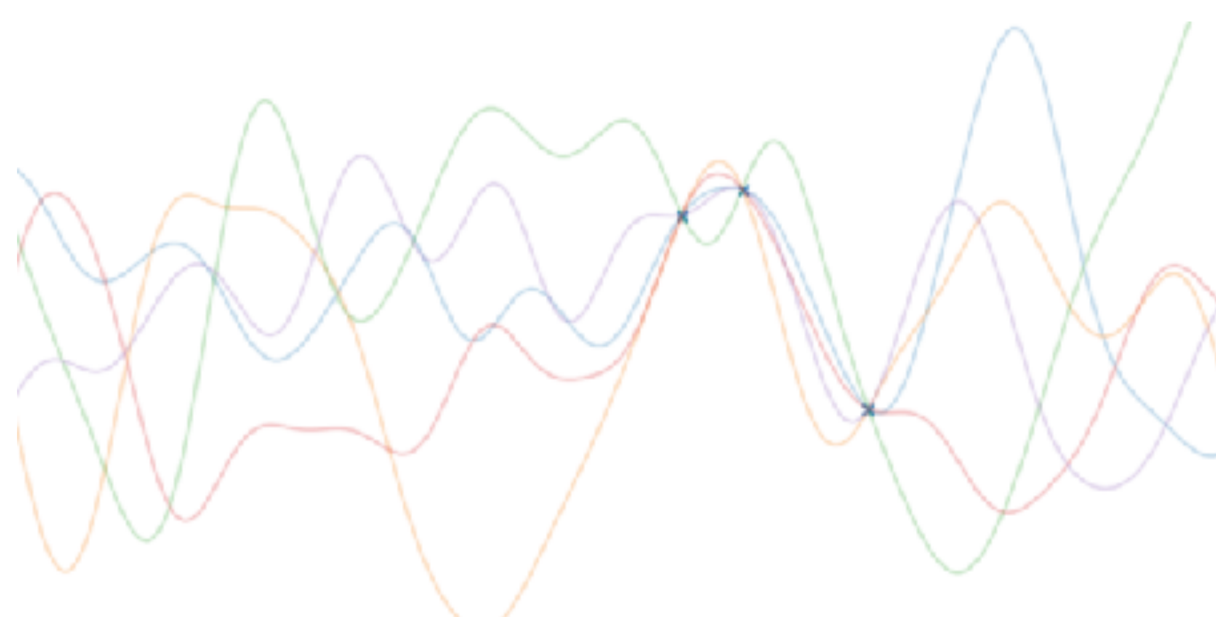
$$
\begin{aligned}
p(y) &= \mathbb{E}_{q(z)} \frac{p(y,z)}{q(z)} \\
\log p(y) &= \log \mathbb{E}_{q(z)} \frac{p(y,z)}{q(z)} \\
&\geq \mathbb{E}_{q(z)} \log \frac{p(y,z)}{q(z)}
\end{aligned}
$$

Recap: VI (2)



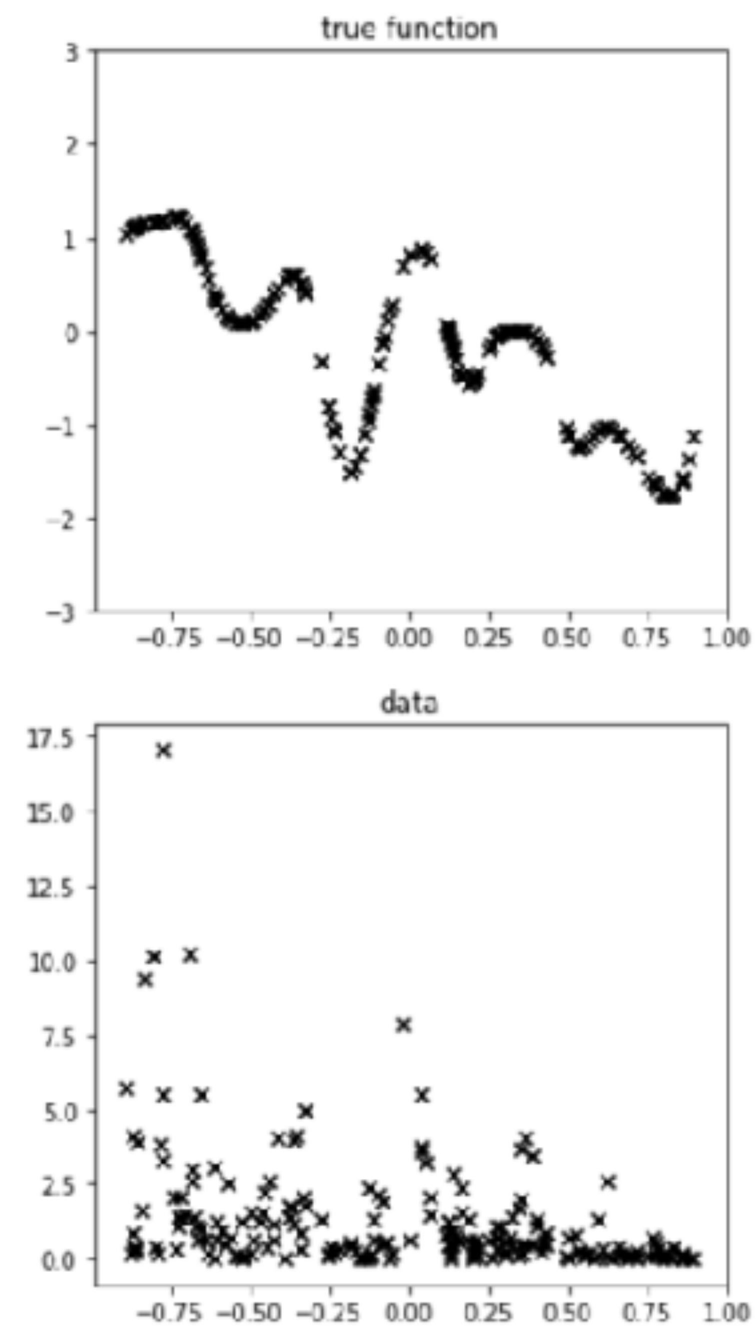
Overview

- ~~Review GPs and VI~~
- **Establish what problems we want to solve**
- Discuss alternative approaches
- VI for GPs part 1 (conjugacy)
- VI for GPs part 2 (scalability)
- Deep GPs



Problems to solve #1: conjugacy

- Exact approach only possible with Gaussian likelihood
- We want: classification models, heavy tailed likelihoods, models for positive quantities etc.
- We might include a GP as part of a larger model (e.g. Deep GP)



Problems to solve #1: conjugacy

Modelling a rate

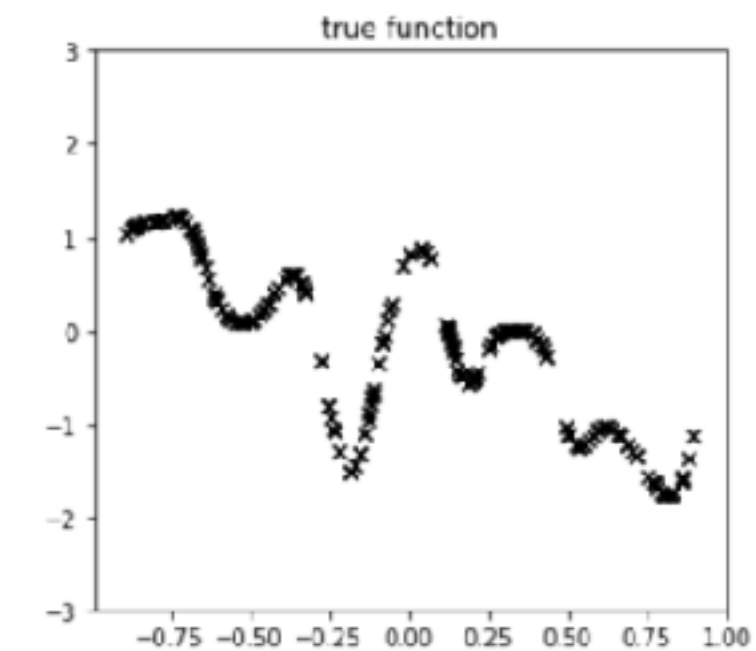
$p(y_n \mid f, x_n) = \lambda_n e^{-y_n \lambda_n}$	exponential distribution
$\lambda_n = e^{f(x_n)}$	exponential link function
$f \sim GP(m, k)$	Gaussian process prior

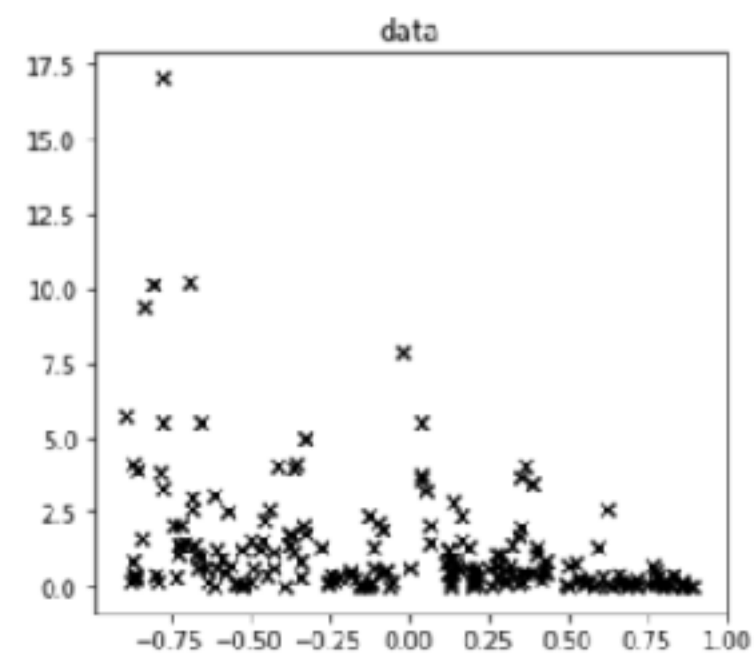
Classification

$p(y_n = 1 \mid f, x_n) = p_n$	Bernoulli distribution
$p_n = \sigma(f(x_n))$	logistic link function
$f \sim GP(m, k)$	Gaussian process prior

Hyperpriors

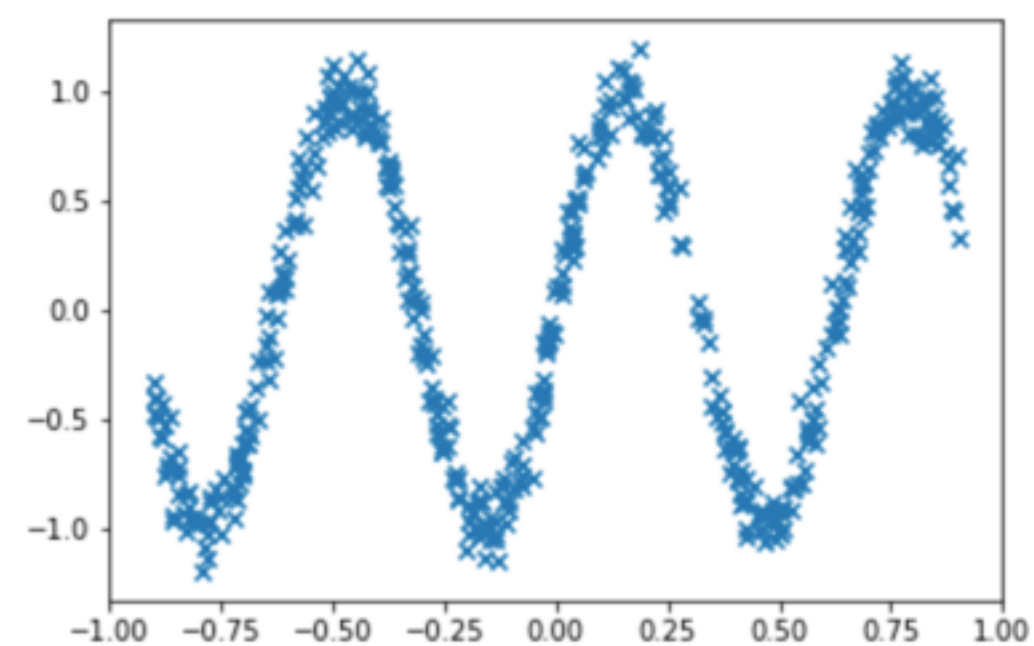
$p(y_n \mid f, x_n) = \mathcal{N}(y_n \mid f(x_n), \sigma^2)$	Gaussian likelihood
$f \sim GP(m, k_\theta)$	Gaussian process prior
$\theta \sim \Gamma(a, b)$	hyperprior





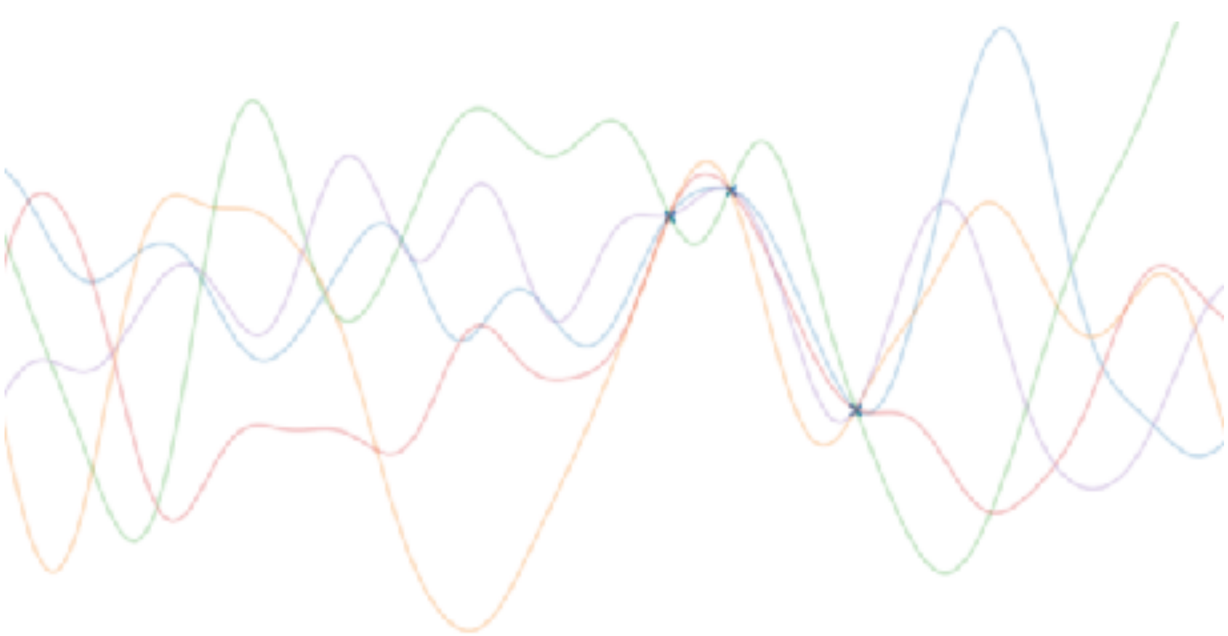
Problems to solve #2: scalability

- Exact approach incurs N^2 memory and N^3 complexity
- We want to deal with datasets larger than N=5000
- Ideally, we would like to deal with datasets that are too large to fit in memory

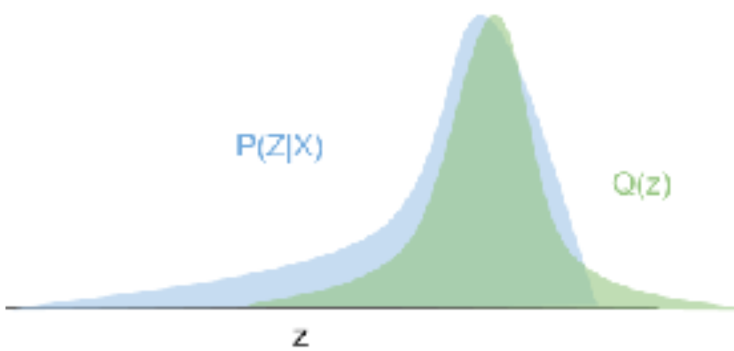


Overview

- ~~Review GPs and VI~~
- ~~Establish what problems we want to solve~~
- **Discuss alternative approaches**
- VI for GPs part 1 (conjugacy)
- VI for GPs part 2 (scalability)
- Deep GPs

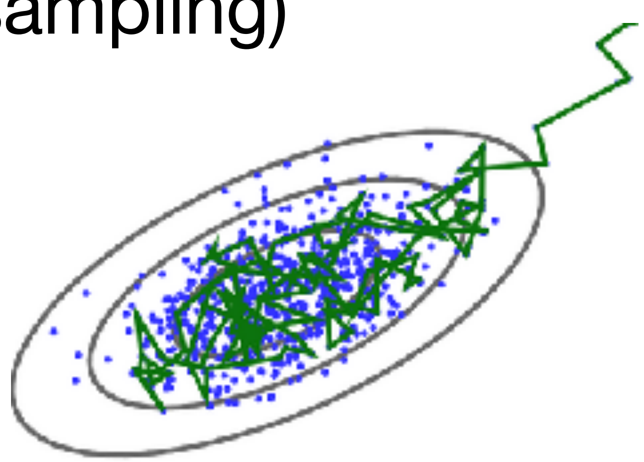


Alternative approaches: non-conjugacy

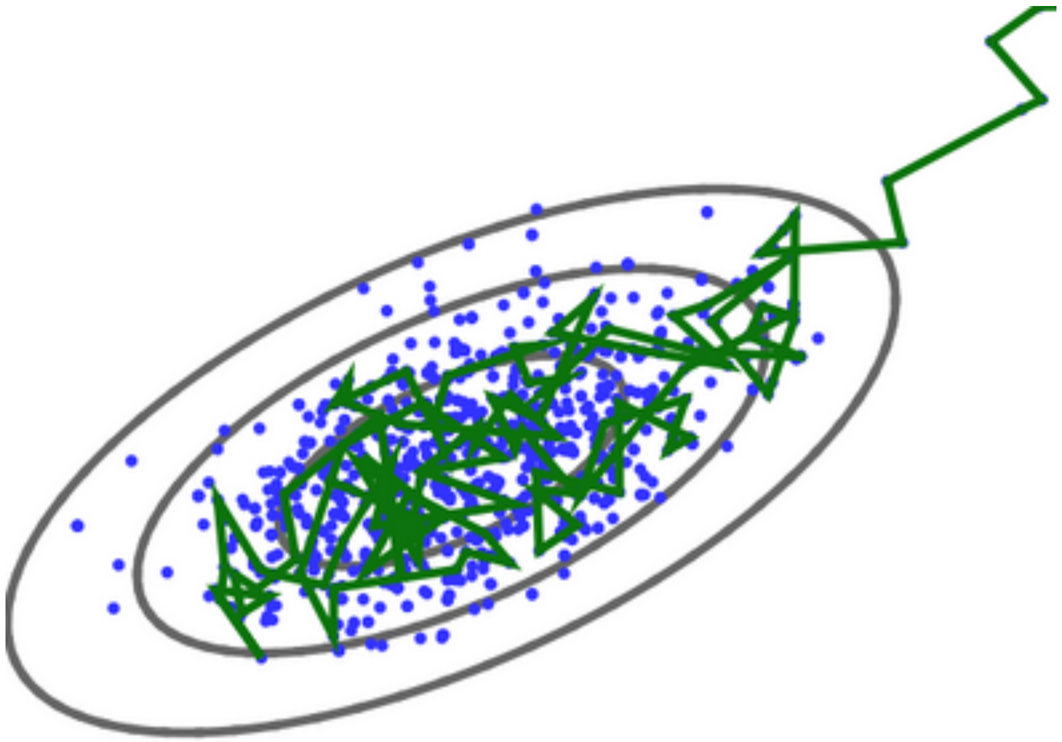


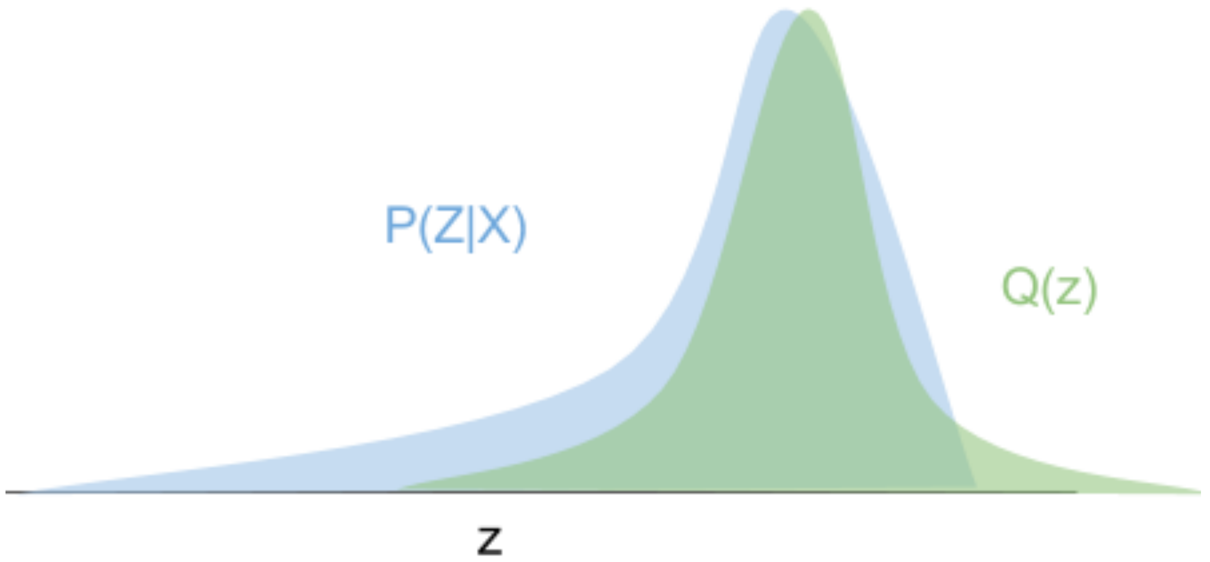
- Deterministic methods (MAP, Laplace, local variational methods, EP, VI, moment matching)

- Sampling methods (Gibbs sampling, HMC, Elliptical slice sampling)



Sampling vs deterministic





Asymptotically exact

Optimization problem
Can do model learning jointly with inference
(Might get a reasonable answer cheaply)

Can't tell when to stop
No marginal likelihood
(Might get a terrible answer given feasible compute)

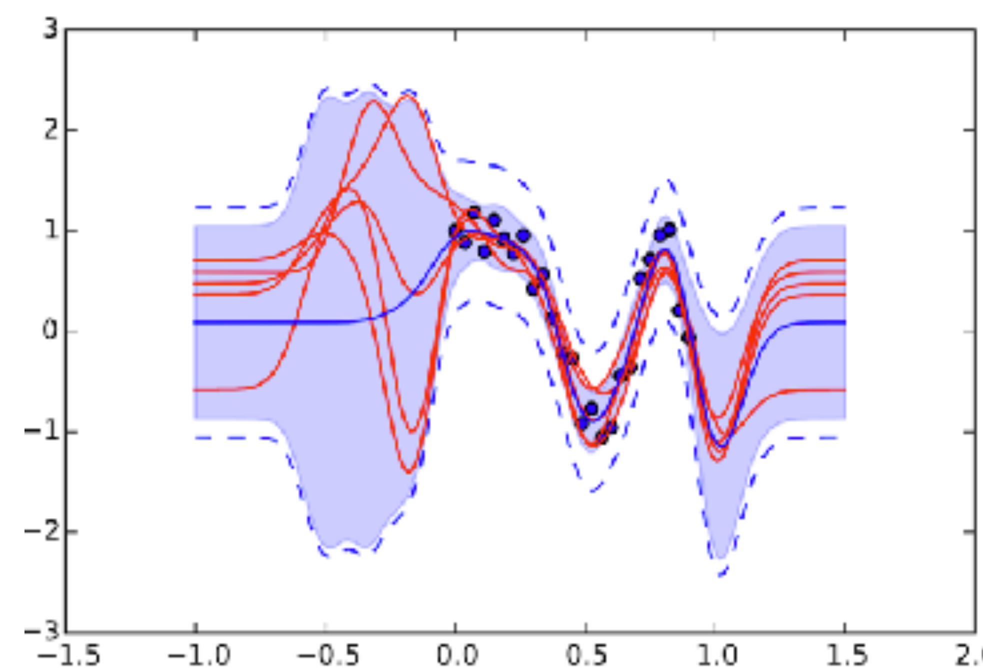
Inaccurate

A note on high dimensional MCMC algorithms

- Intuitions in low dimensions can be dangerously misleading in high dimensions
- High dimensional space is hard to navigate using naïve random walks - there are too many bad directions!
- See this excellent introduction for why HMC is a good idea in high dimensions: youtu.be/_fnDz2Bz3h8

Alternative approaches: scalability

- Approximate the model
- Approximate the algebra
- Approximate the posterior

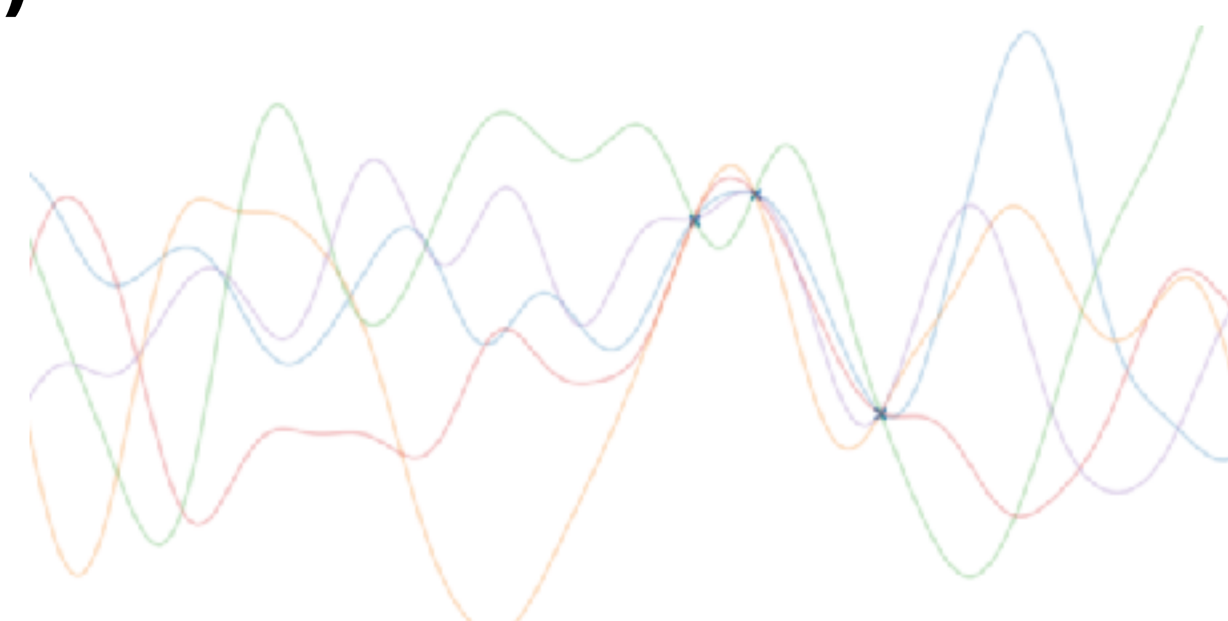


NB there are equivalences between methods

Distinction between approaches not always clear

Overview

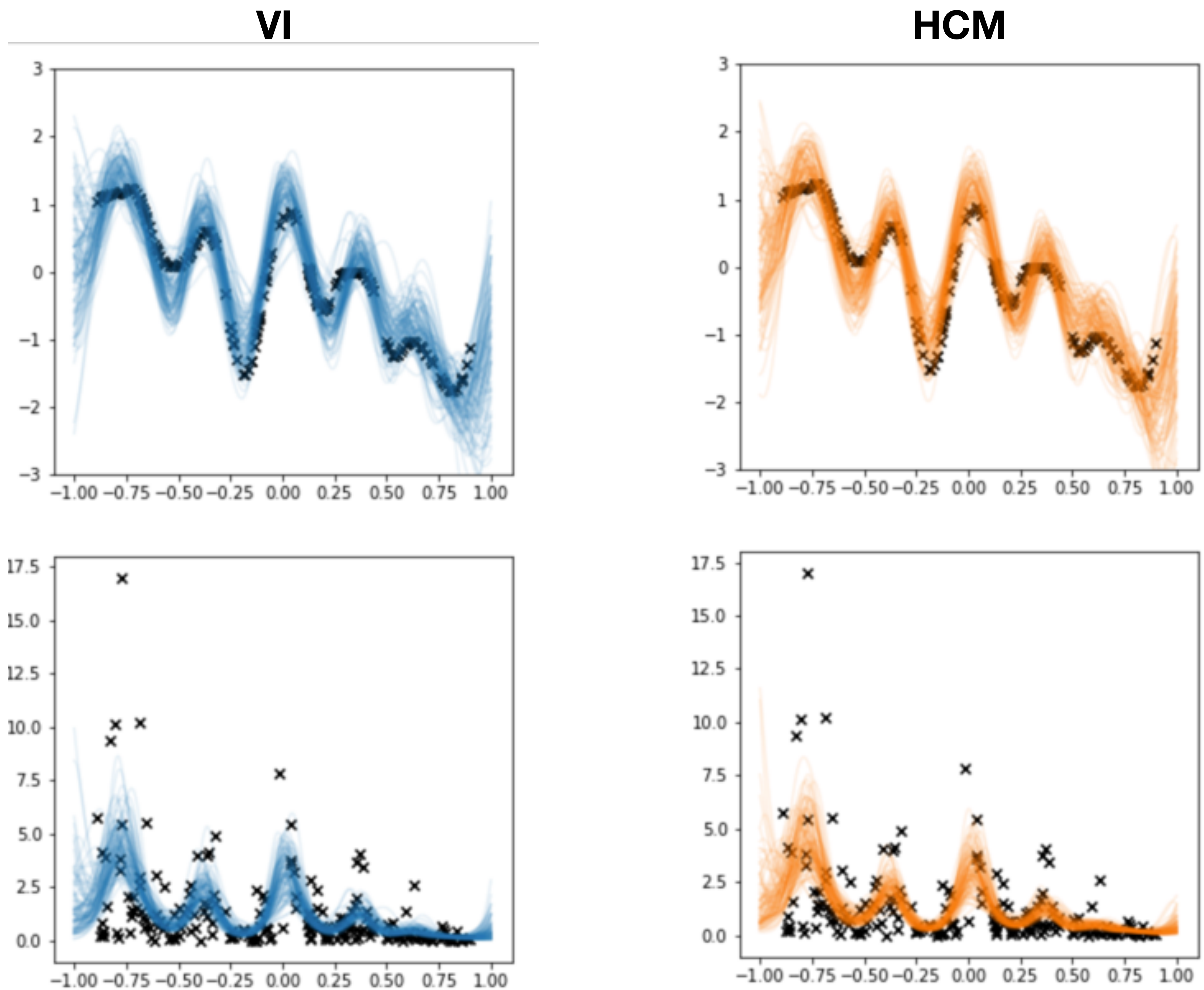
- ~~Review GPs and VI~~
- ~~Establish what problems we want to solve~~
- ~~Discuss alternative approaches~~
- **VI for GPs part 1 (conjugacy)**
- VI for GPs part 2 (scalability)
- Deep GPs

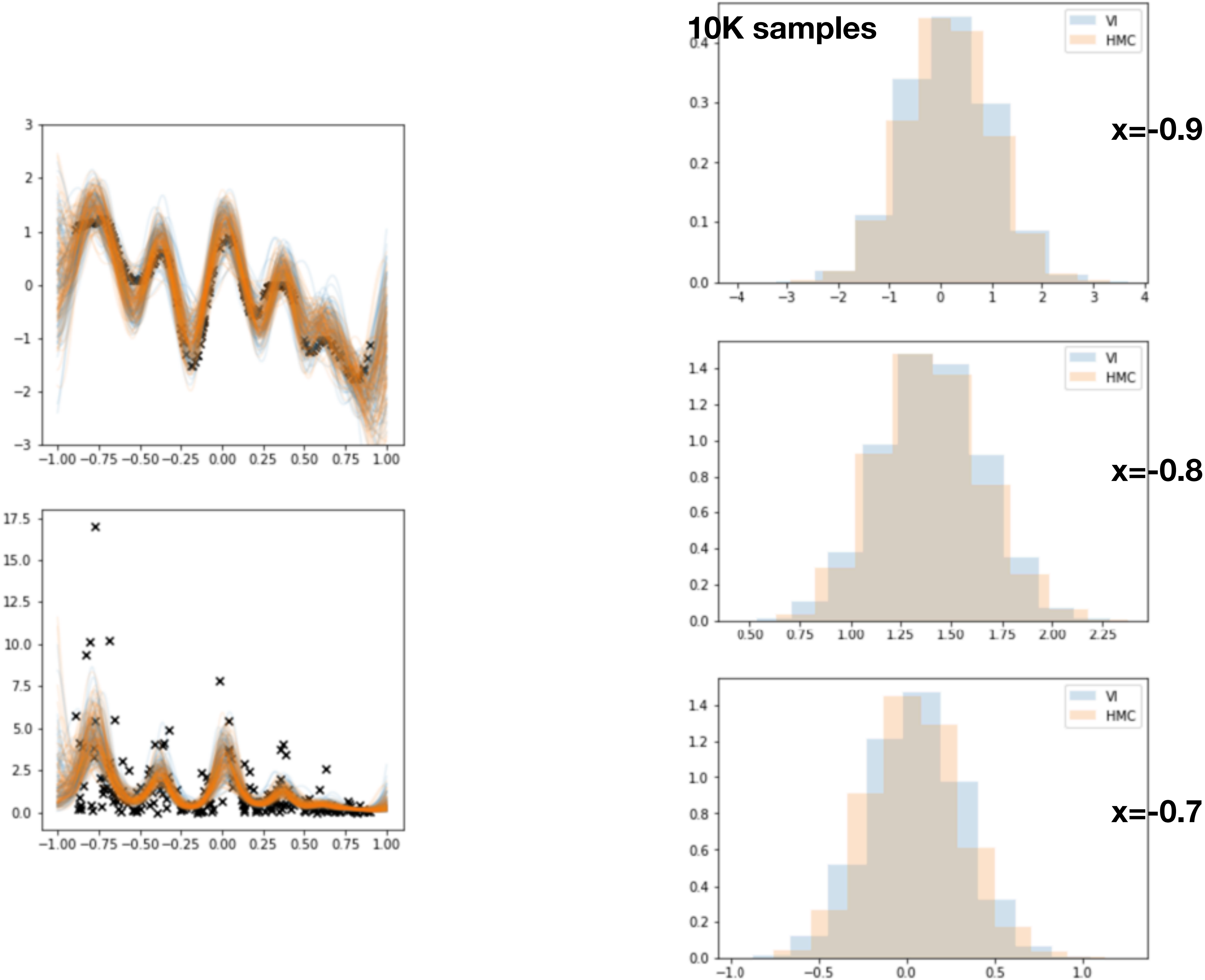


Key points

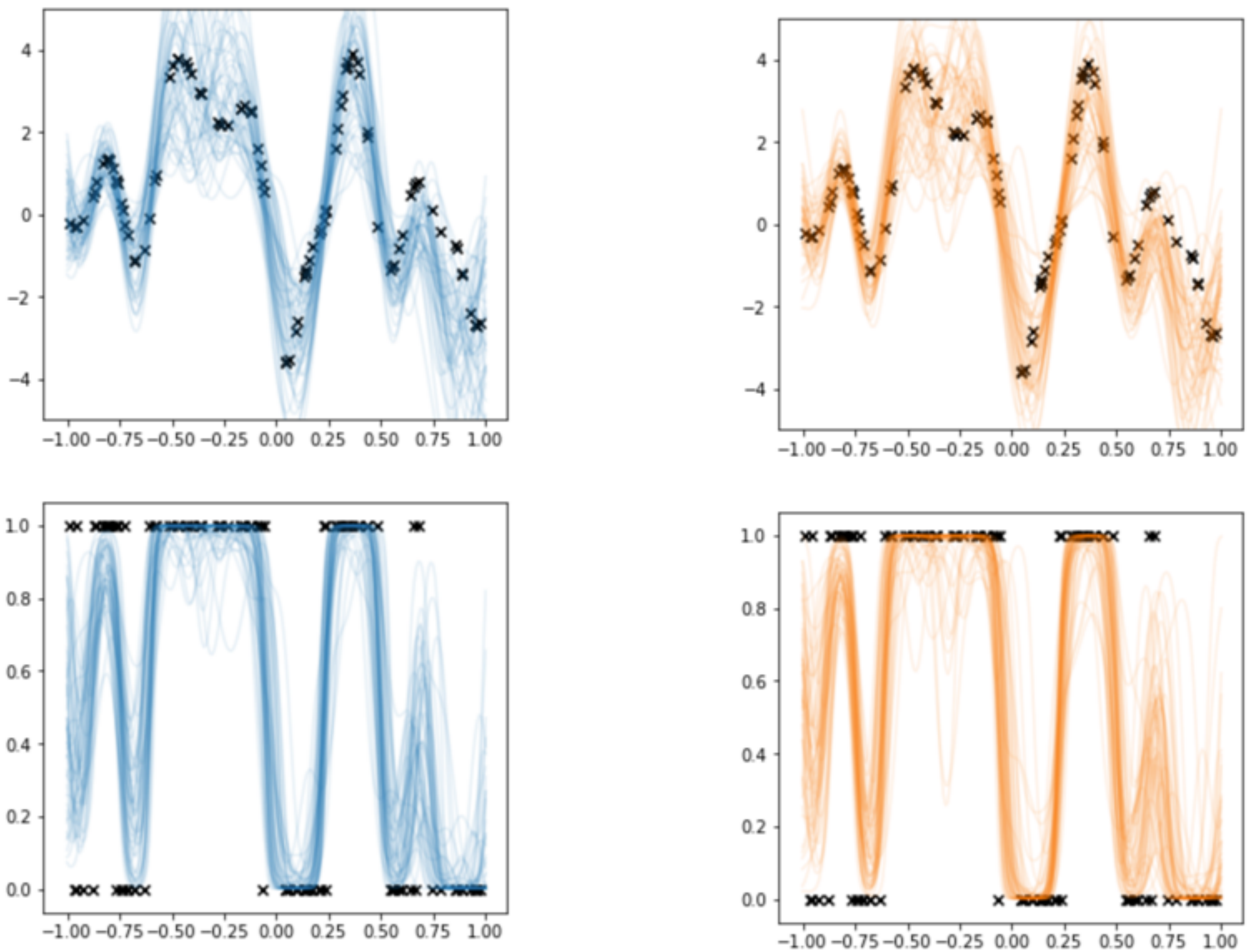
- Use a multivariate Gaussian for the data functions values
- ELBO is a sum of 1D expectations and a closed form KL
- Optimize with respect to variational parameters

$$
\begin{aligned}
\text{ELBO} &= \mathbb{E}_{q(f)} \log \frac{p(\mathbf{y}, \mathbf{f})}{q(\mathbf{f})} \\
&= \mathbb{E}_{q(f)} \log \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f})}{q(\mathbf{f})} \\
&= \mathbb{E}_{q(f)} \log p(\mathbf{y}|\mathbf{f}) + \mathbb{E}_{q(\mathbf{f})} \log \frac{p(\mathbf{f})}{q(\mathbf{f})} \\
&= \sum_n \mathbb{E}_{q(f(x_n))} \log p(y_n|f(x_n)) + \mathbb{E}_{q(\mathbf{f})} \log \frac{p(\mathbf{f})}{q(\mathbf{f})} \\
&= \sum_n \mathbb{E}_{q(f(x_n))} \log p(y_n|f(x_n)) - \mathrm{KL}(q(\mathbf{f})||p(\mathbf{f})) \\
q(\mathbf{f}) &= \mathcal{N}(\mathbf{f}|\mathbf{m}, \mathbf{S}) \\
p(\mathbf{f}) &= \mathcal{N}(\mathbf{f}|\mathbf{0}, \mathbf{K}) \\
\mathrm{KL}(q(\mathbf{f})||p(\mathbf{f})) &= \tfrac{1}{2}\left[\mathbf{m}^\top \mathbf{K}^{-1}\mathbf{m} + \mathrm{Tr}(\mathbf{K}^{-1}\mathbf{S}) - D + \log|\mathbf{K}| - \log|\mathbf{S}|\right] \\
q(f(x_n)) &= \mathcal{N}(m_n, S_{nn})
\end{aligned}
$$


VI
HCM



10K samples
VI
HMC
x=-0.9
x=-0.8
x=-0.7



4
2
0
−2
−4
−1.00 −0.75 −0.50 −0.25 0.00 0.25 0.50 0.75 1.00
1.0
0.8
0.6
0.4
0.2
0.0

VI pros and cons

$$\text{ELBO} = \sum_n \mathbb{E}_{q(f(x_n))} \log p(y_n|f(x_n)) - \text{KL}(q(\mathbf{f})||p(\mathbf{f}))$$

- Log likelihood is smooth (easy for accurate 1D integration
- KL is closed-form and computation is parallel
- Easy to optimize (can also use natural gradients)

- Could introduce error if using quadrature
- Only closed form if using a Gaussian posterior
- Requires N + N^2 memory* and N^3 computation

*** Possible to show the covariance has a special structure, reducing memory requirement to 2N.**

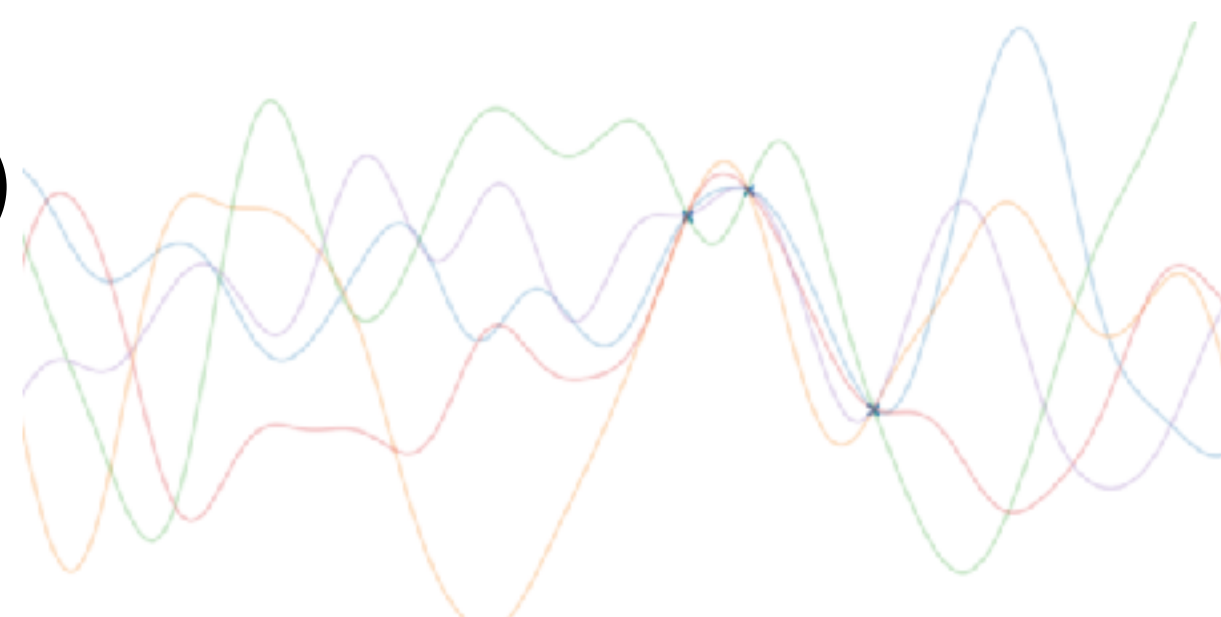
What about the full function?

$$\begin{aligned}
\text{ELBO} &= \mathbb{E}_{q(f)} \log \frac{p(\mathbf{y}, f)}{q(f)} \\
&= \mathbb{E}_{q(f)} \log \frac{p(\mathbf{y}|\mathbf{f})p(f)}{q(f)} \\
&= \mathbb{E}_{q(f)} \log p(\mathbf{y}|\mathbf{f}) + \mathbb{E}_{q(f)} \log \frac{p(f)}{q(f)} \\
&= \sum_n \mathbb{E}_{q(f(x_n))} \log p(y_n|f_n) + \mathbb{E}_{q(f)} \log \frac{p(f)}{q(f)}
\end{aligned}$$

$$\begin{aligned}
p(f) &= p(f_*|\mathbf{f})p(\mathbf{f}) \\
\text{ELBO} &= \sum_n \mathbb{E}_{q(f(x_n))} \log p(y_n|f_n) + \mathbb{E}_{q(f)} \log \frac{p(f_*|\mathbf{f})p(\mathbf{f})}{p(f_*|\mathbf{f})q(\mathbf{f})} \\
&= \sum_n \mathbb{E}_{q(f(x_n))} \log p(y_n|f_n) + \mathbb{E}_{q(f)} \log \frac{p(\mathbf{f})}{q(\mathbf{f})} \\
&= \sum_n \mathbb{E}_{q(f(x_n))} \log p(y_n|f_n) + \mathbb{E}_{q(\mathbf{f})} \log \frac{p(\mathbf{f})}{q(\mathbf{f})}
\end{aligned}$$

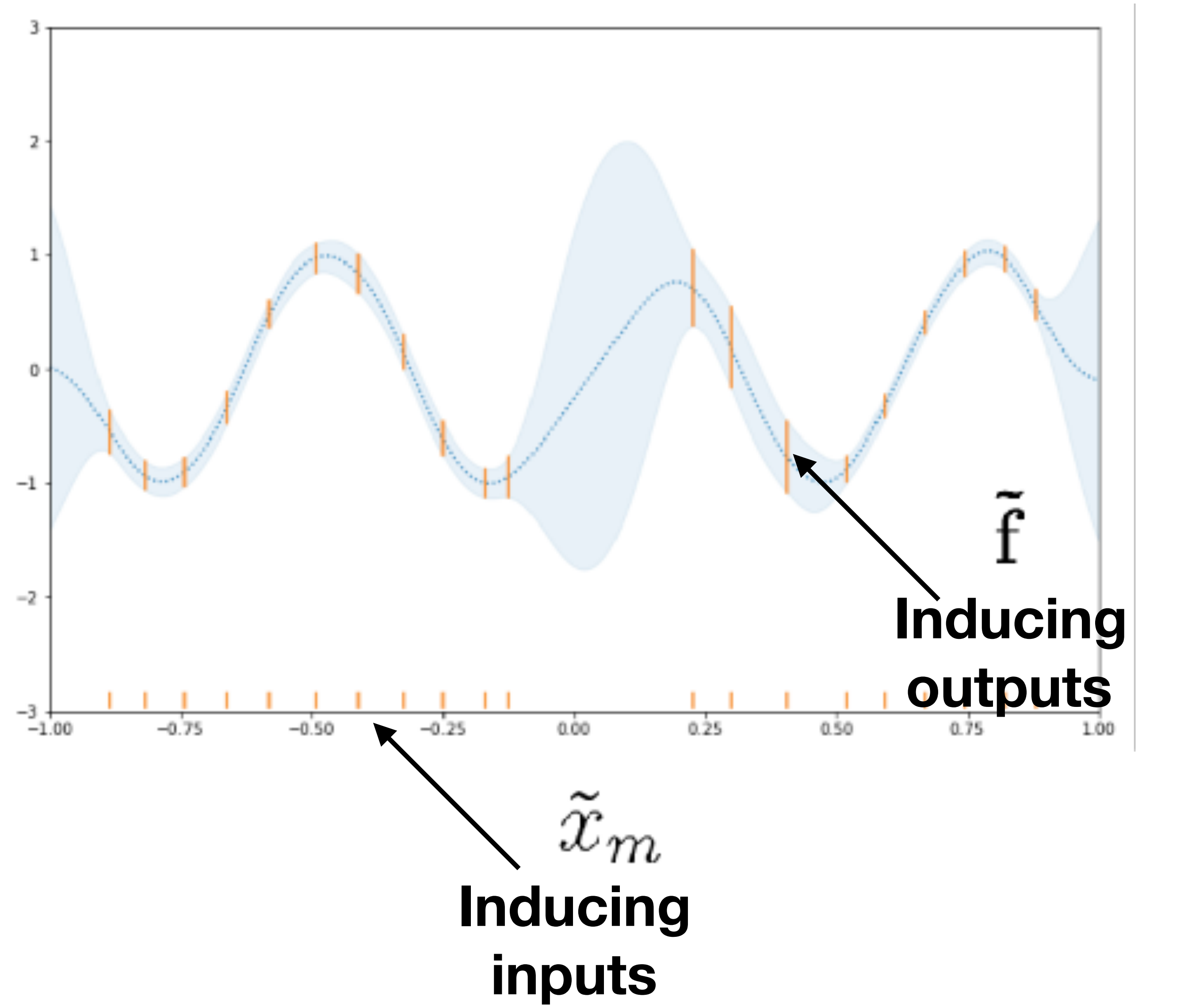
Overview

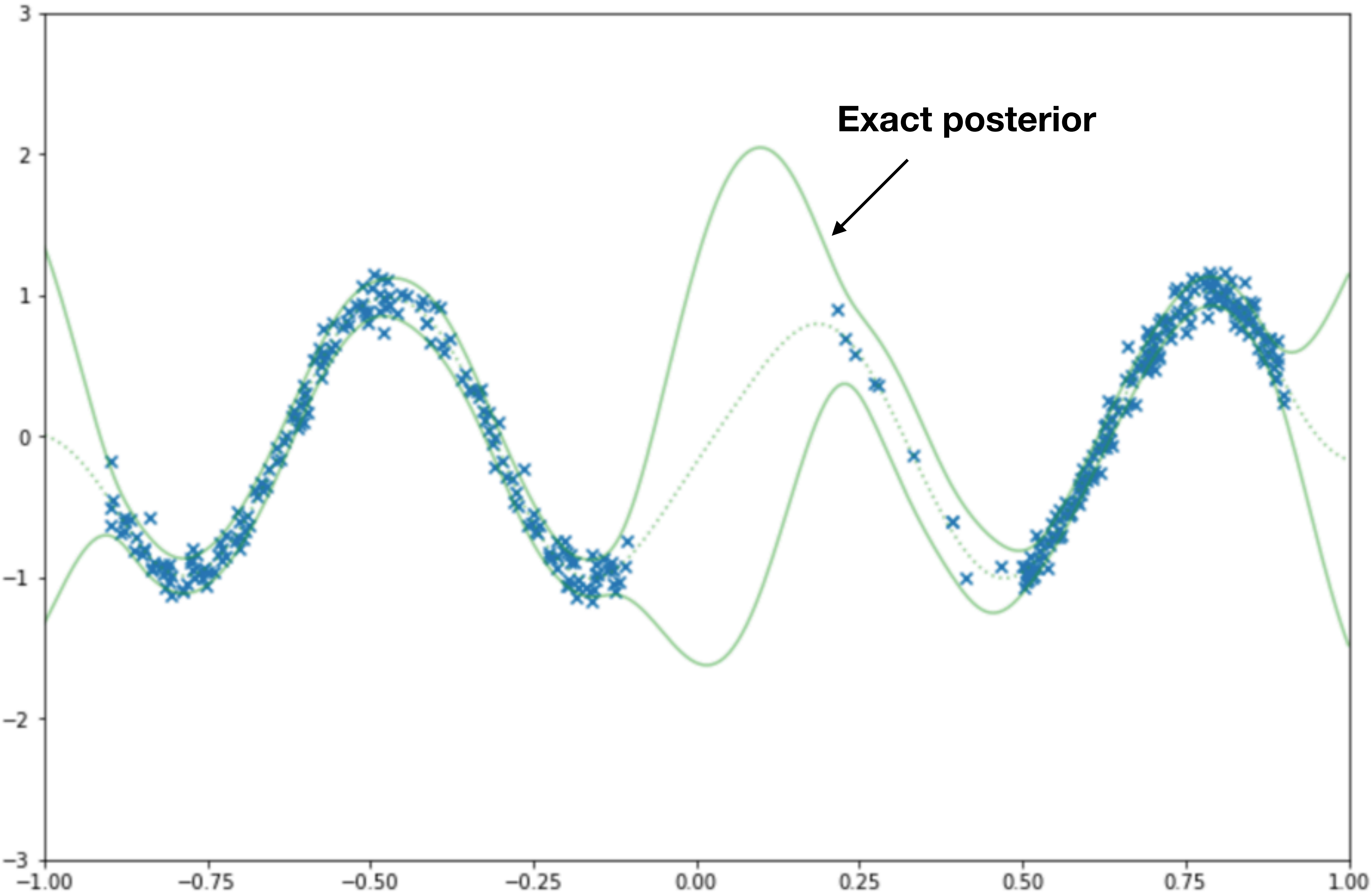
- ~~Review GPs and VI~~
- ~~Establish what problems we want to solve~~
- ~~Discuss alternative approaches~~
- ~~VI for GPs part 1 (conjugacy)~~
- **VI for GPs part 2 (scalability)**
- Deep GPs

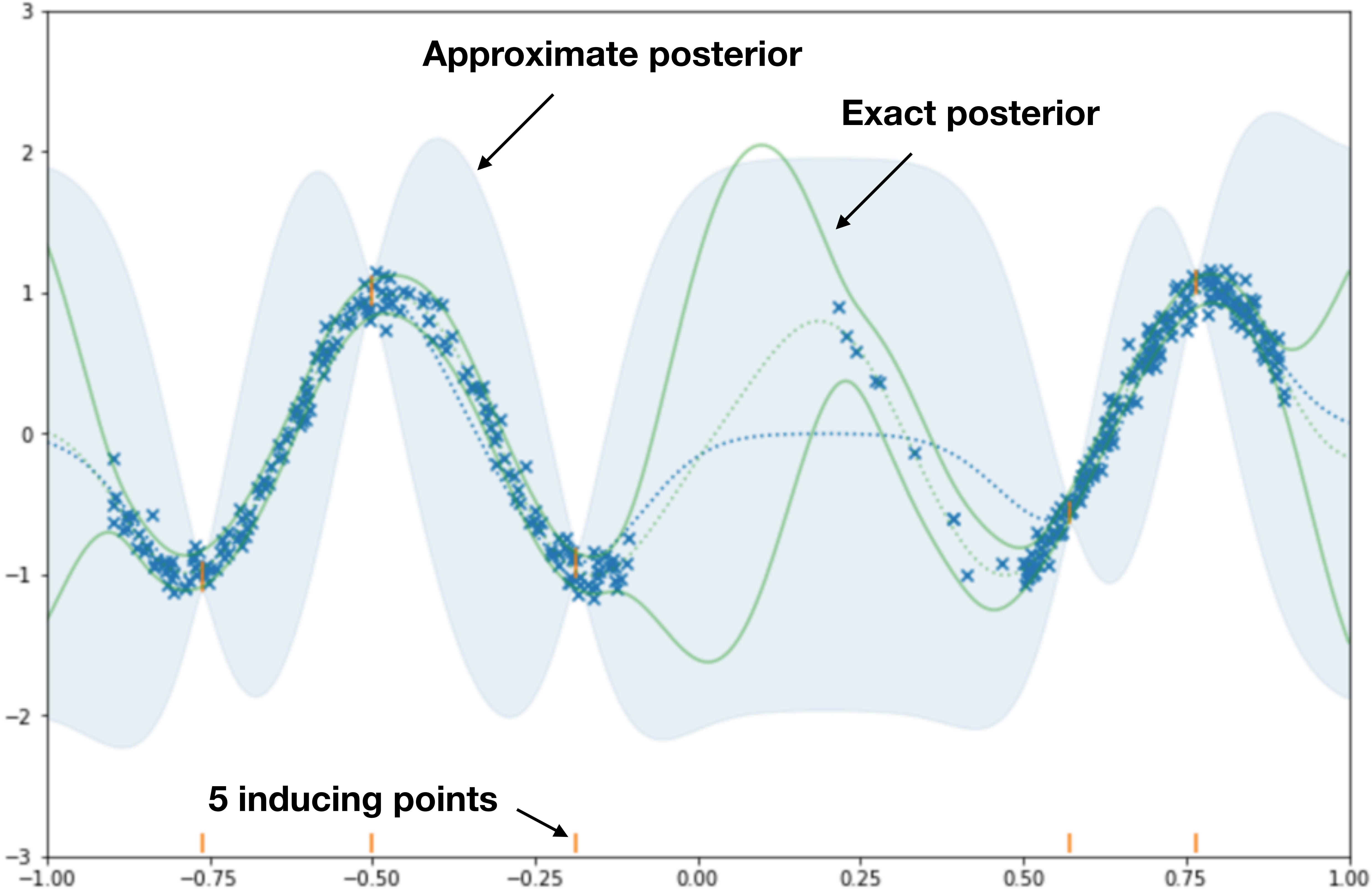


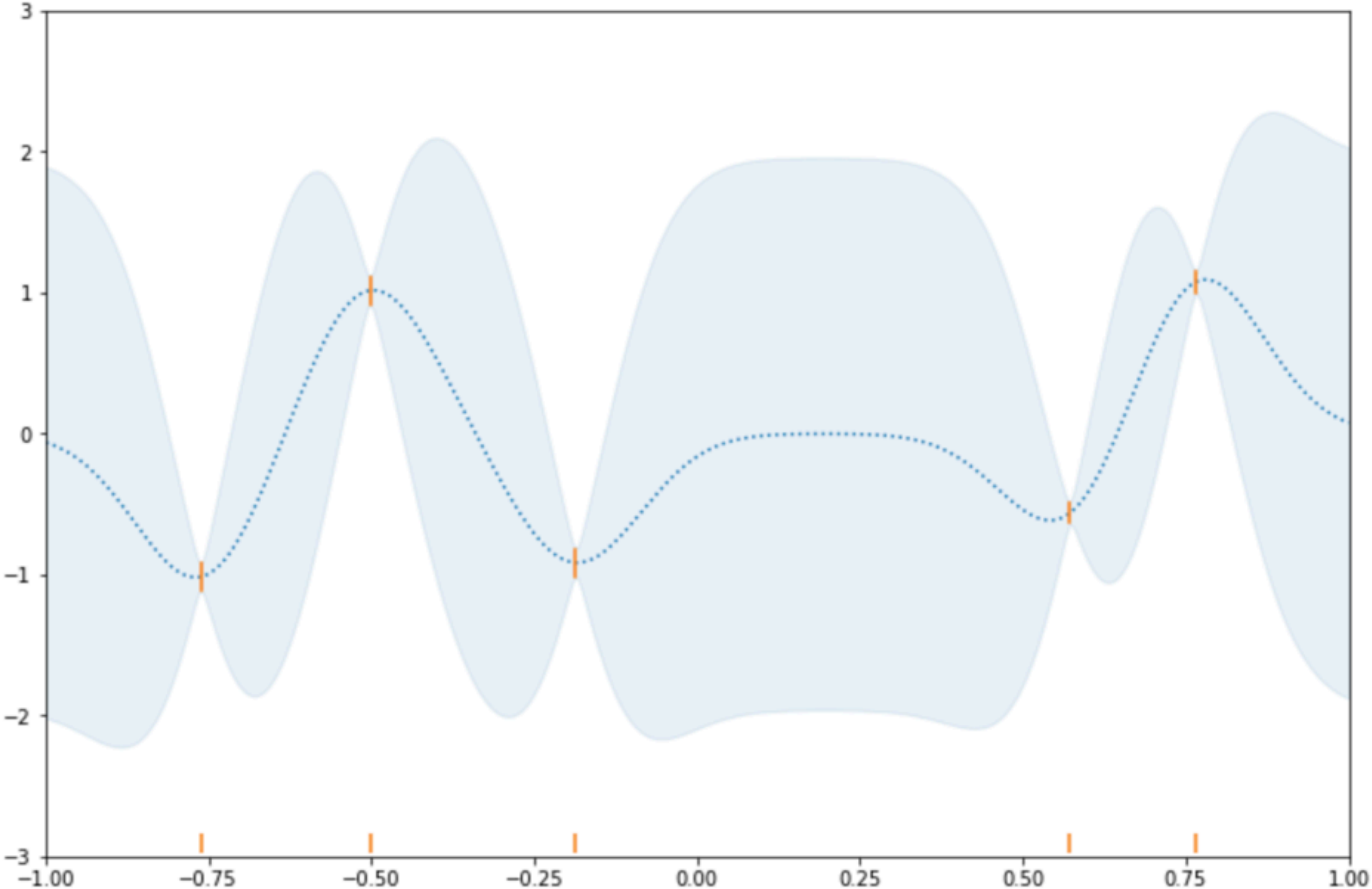
Key idea

- For a variational posterior by conditioning on a set of *inducing points* $\tilde{\mathbf{f}}$
- The KL simplifies, just as in the dense case
- The variational distribution has Gaussian compute marginals, if $q(\tilde{\mathbf{f}})$ is Gaussian. These marginals can be compute just as in the single layer case

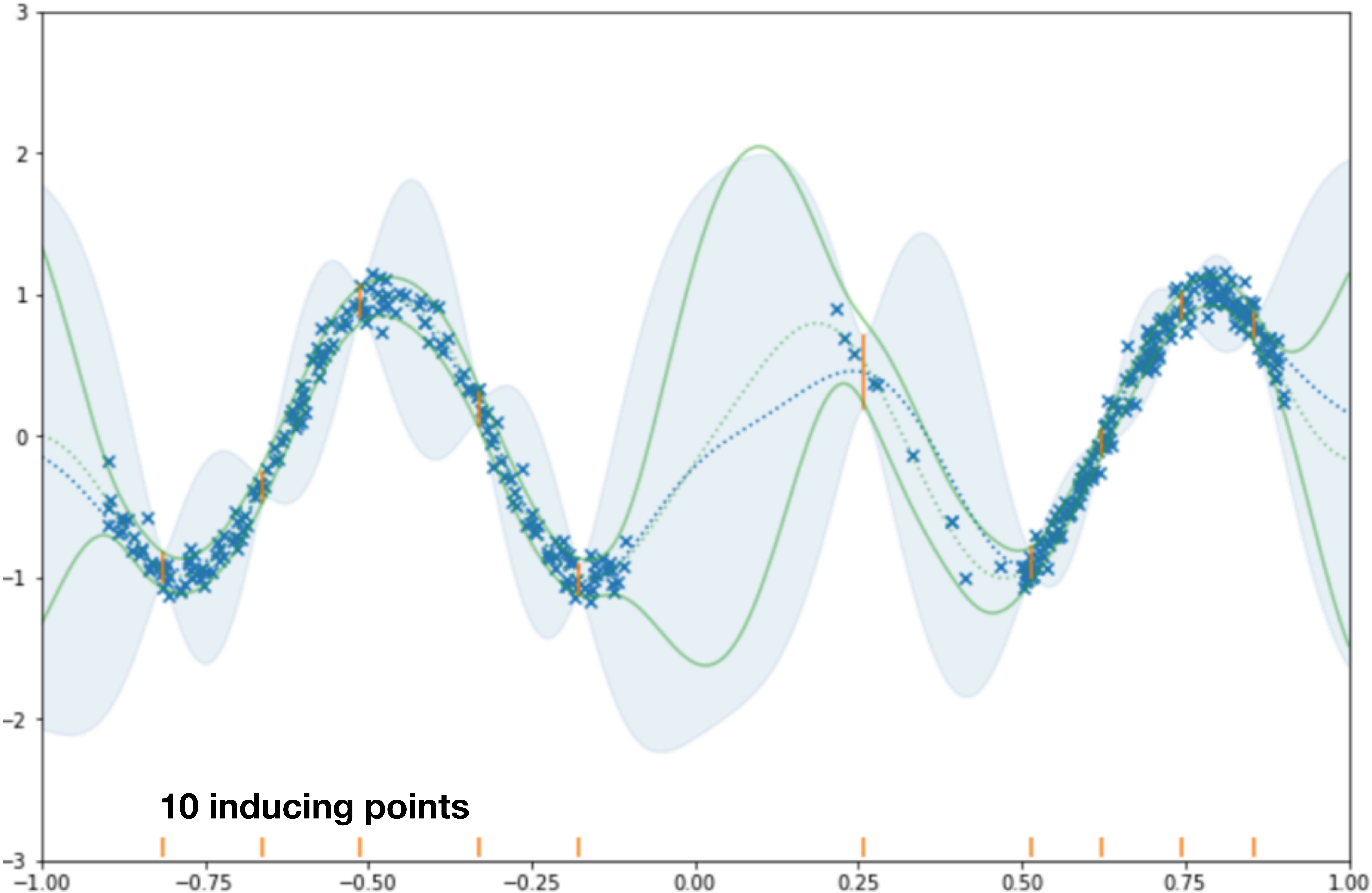


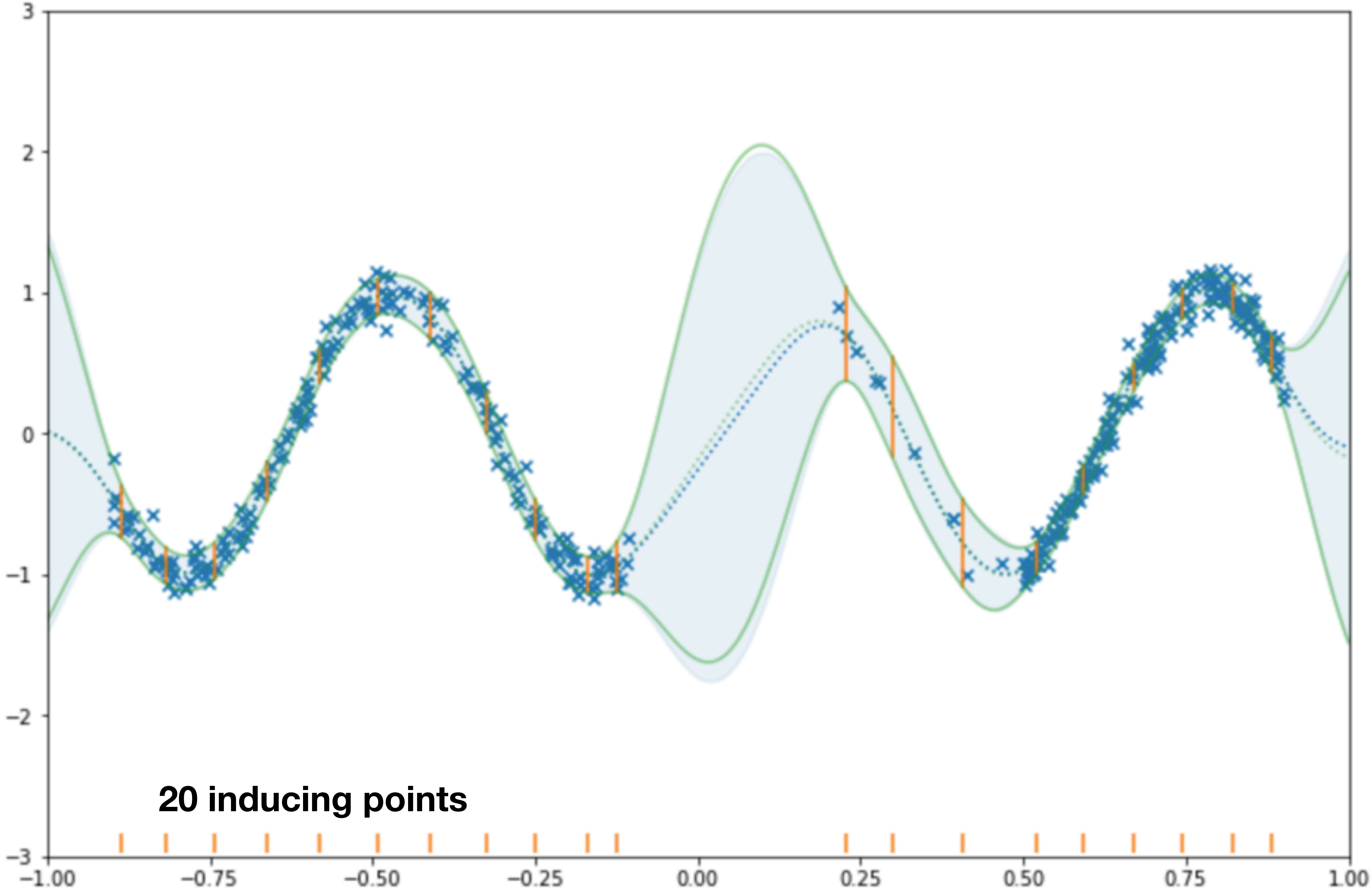

Exact posterior
3
2
1
0
−1
−2
−3
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00


Approximate posterior
Exact posterior
5 inducing points
3
2
1
0
−1
−2
−3
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00

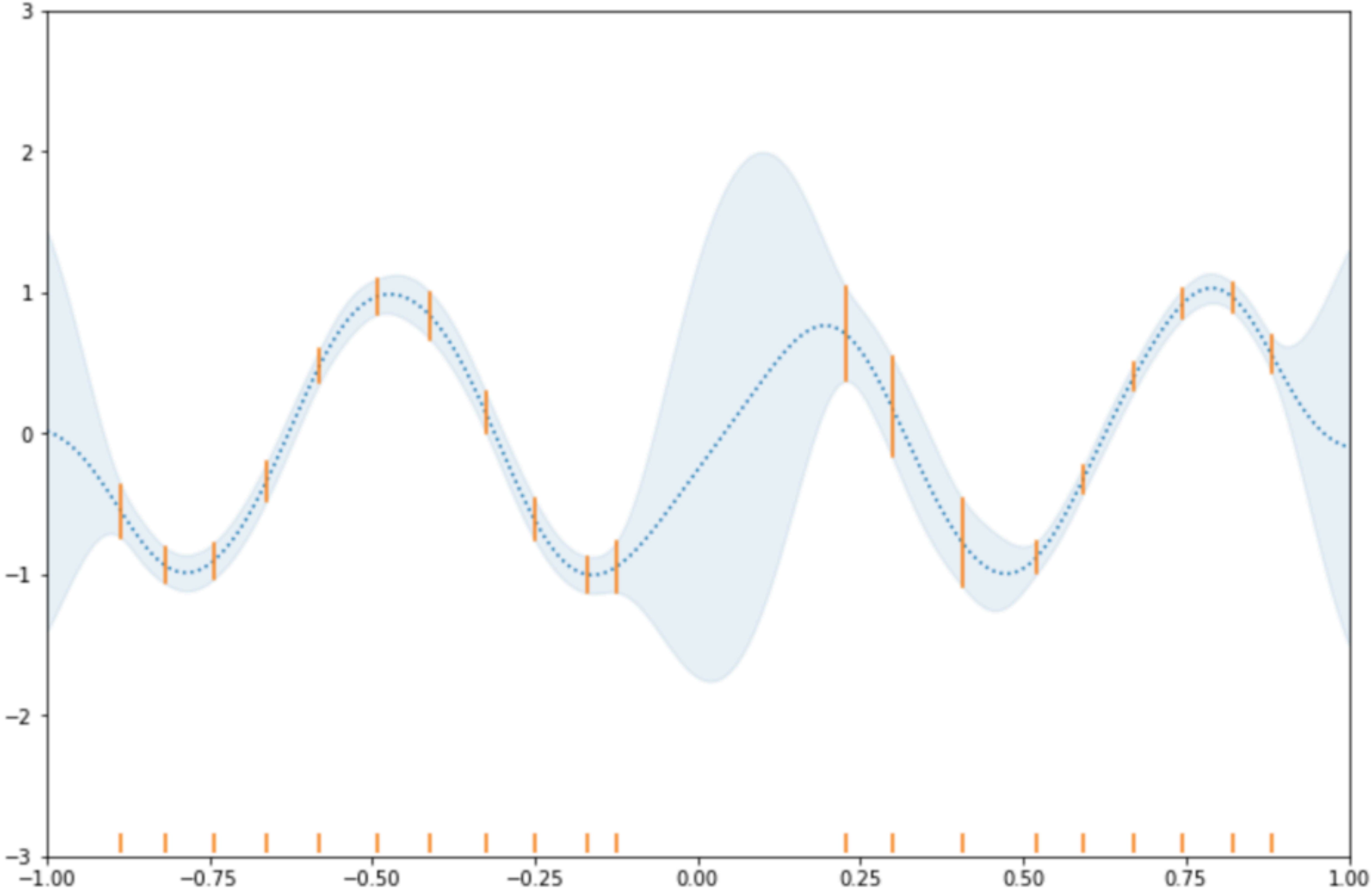


3
2
1
0
−1
−2
−3
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00

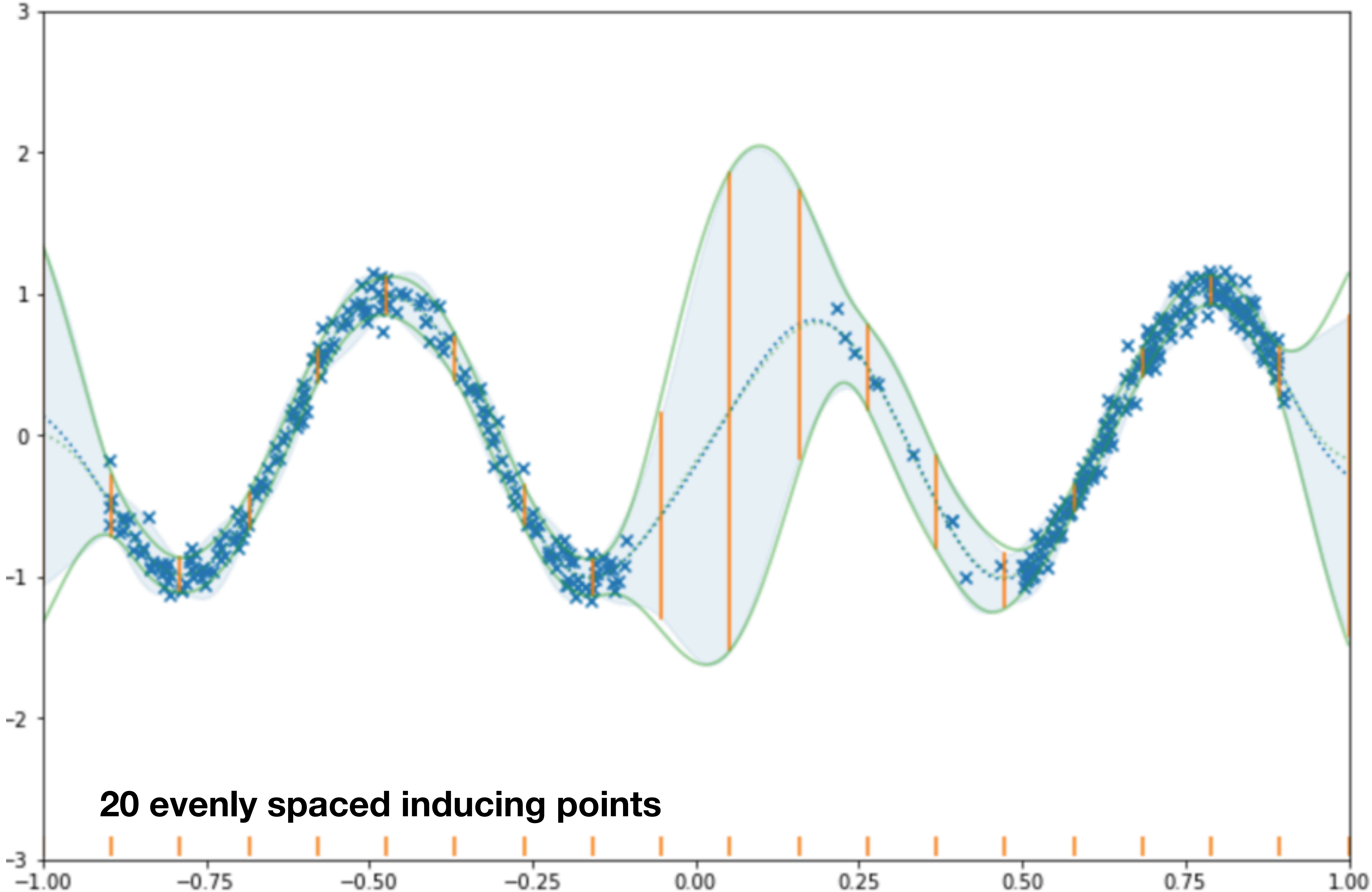

3
2
1
0
-1
-2
-3
-1.00
-0.75
-0.50
-0.25
0.00
0.25
0.50
0.75
1.00
10 inducing points



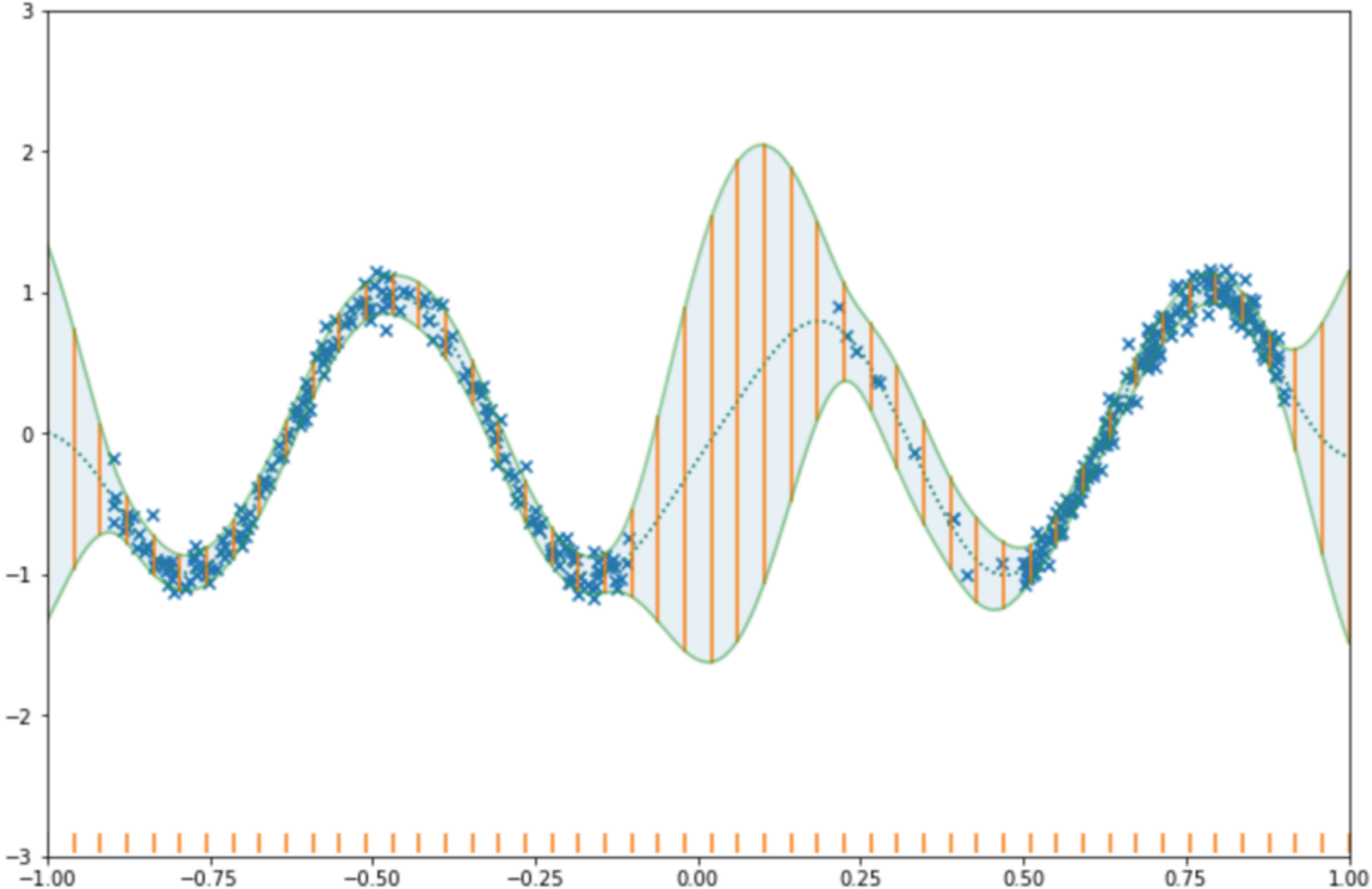
3
2
1
0
−1
−2
−3
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
20 inducing points



3
2
1
0
−1
−2
−3
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00



20 evenly spaced inducing points


3
2
1
0
−1
−2
−3
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00

Variable partitions

$$p(f) = p(\tilde{f}_* \,|\, \tilde{\mathbf{f}})p(\tilde{\mathbf{f}})$$

$$p(\tilde{\mathbf{f}}) = \mathcal{N}(\tilde{\mathbf{f}} \,|\, \mathbf{0}, \tilde{\mathbf{K}})$$

$$p(\tilde{f}_* | \tilde{\mathbf{f}}) = \mathcal{GP}(\mu, \Sigma)$$

$$\tilde{\mu}(x) = \tilde{\mathbf{k}}(x)^\top \tilde{\mathbf{K}}^{-1}\tilde{\mathbf{f}}$$

$$\tilde{\Sigma}(x, x') = k(x, x') - \tilde{\mathbf{k}}(x)^\top \tilde{\mathbf{K}}^{-1}\tilde{\mathbf{k}}(x')$$

Symbol	Size	Equivalent to	Interpretation
$\tilde{\mathbf{f}}$	M	$\{f(\tilde{x}_m) \,|\, n = 1, \ldots, M\}$	Some other function values we can choose
$\tilde{f}_*$	∞	$f \setminus \tilde{\mathbf{f}}$	All the function values that are not in $\tilde{\mathbf{f}}$
$\tilde{\mathbf{k}}(x)$	M	$\{k(x, \tilde{x}_m) \,|\, m = 1, \ldots, M\}$	Covariance between a test point and the pseudo-data
$\tilde{\mathbf{K}}$	M, M	$\{k(\tilde{x}_i, \tilde{x}_j) \,|\, i, j = 1, \ldots, M\}$	Covariance between pseudo-data

$$
\begin{aligned}
\text{ELBO} &= \mathbb{E}_{q(f)} \log \frac{p(\mathbf{y}, f)}{q(f)} \\
&= \mathbb{E}_{q(f)} \log \frac{p(\mathbf{y}|\mathbf{f})p(f)}{q(f)} \\
&= \mathbb{E}_{q(f)} \log p(\mathbf{y}|\mathbf{f}) + \mathbb{E}_{q(f)} \log \frac{p(f)}{q(f)} \\
&= \sum_n \mathbb{E}_{q(f(x_n))} \log p(y_n|f_n) + \mathbb{E}_{q(f)} \log \frac{p(f)}{q(f)}
\end{aligned}
$$

$\boxed{q(f) = p(f_*|\tilde{\mathbf{f}})q(\tilde{\mathbf{f}})}$ **Assumption 1**

$$
p(f) = p(f_*|\tilde{\mathbf{f}})p(\tilde{\mathbf{f}})
$$

$$
\begin{aligned}
\text{ELBO} &= \sum_n \mathbb{E}_{q(f(x_n))} \log p(y_n|f_n) + \mathbb{E}_{q(f)} \log \frac{p(f_*|\tilde{\mathbf{f}})p(\tilde{\mathbf{f}})}{p(f_*|\tilde{\mathbf{f}})q(\tilde{\mathbf{f}})} \\
&= \sum_n \mathbb{E}_{q(f(x_n))} \log p(y_n|f_n) + \mathbb{E}_{q(f)} \log \frac{p(\tilde{\mathbf{f}})}{q(\tilde{\mathbf{f}})} \\
&= \sum_n \mathbb{E}_{q(f(x_n))} \log p(y_n|f_n) + \mathbb{E}_{q(\tilde{\mathbf{f}})} \log \frac{p(\tilde{\mathbf{f}})}{q(\tilde{\mathbf{f}})}
\end{aligned}
$$

$$\sum_n \mathbb{E}_{q(f(x_n))} \log p(y_n|f_n) + \mathbb{E}_{q(\tilde{\mathbf{f}})} \log \frac{p(\tilde{\mathbf{f}})}{q(\tilde{\mathbf{f}})}$$

What is this??

Same as before

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

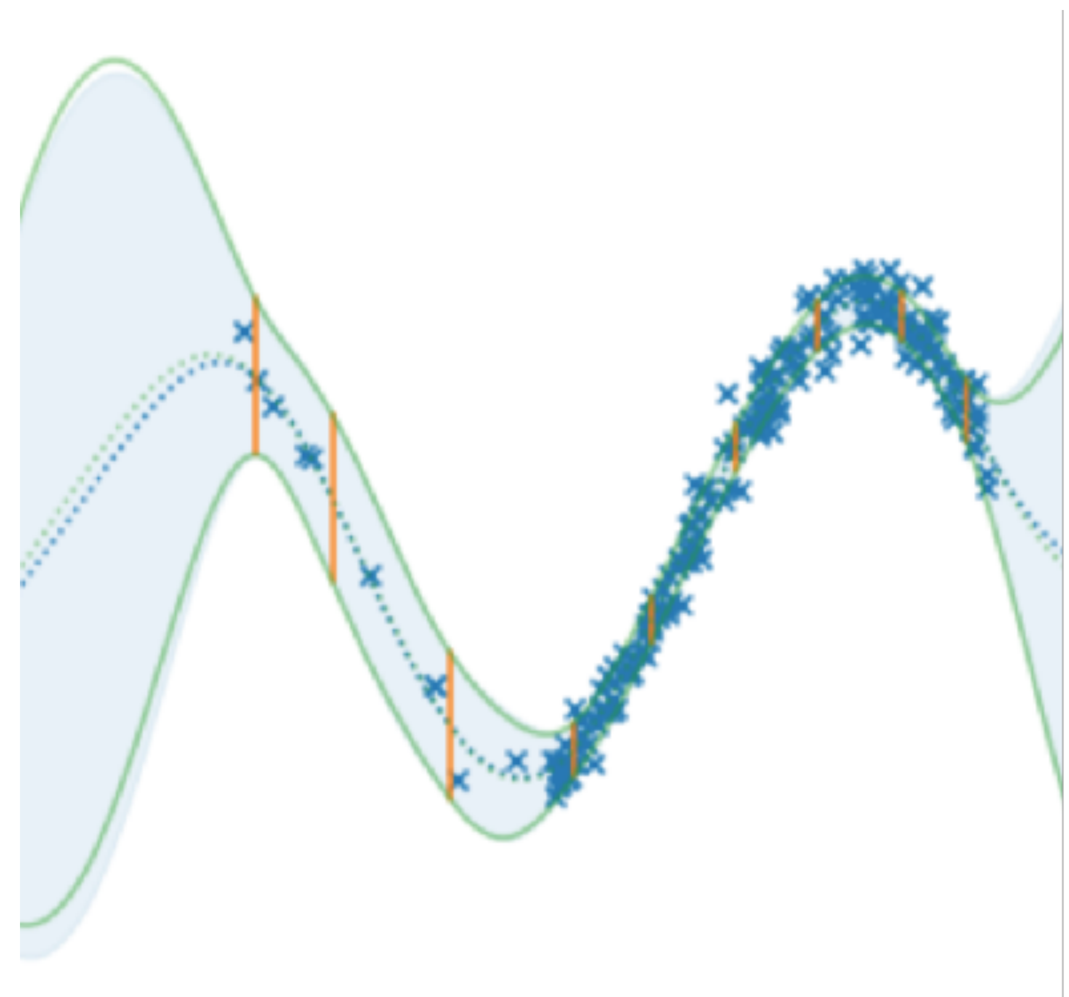
$$p(f(x_n)|\tilde{\mathbf{f}}) = \mathcal{N}(f(x_n)|\tilde{\mathbf{k}}(x)^\top \tilde{\mathbf{K}}^{-1}\tilde{\mathbf{f}}, k(x,x) - \tilde{\mathbf{k}}(x)^\top \tilde{\mathbf{K}}^{-1}\tilde{\mathbf{k}}(x))$$

$$\boxed{q(\tilde{\mathbf{f}}) = \mathcal{N}(\tilde{\mathbf{m}}, \tilde{\mathbf{S}})}$$ Assumption 2

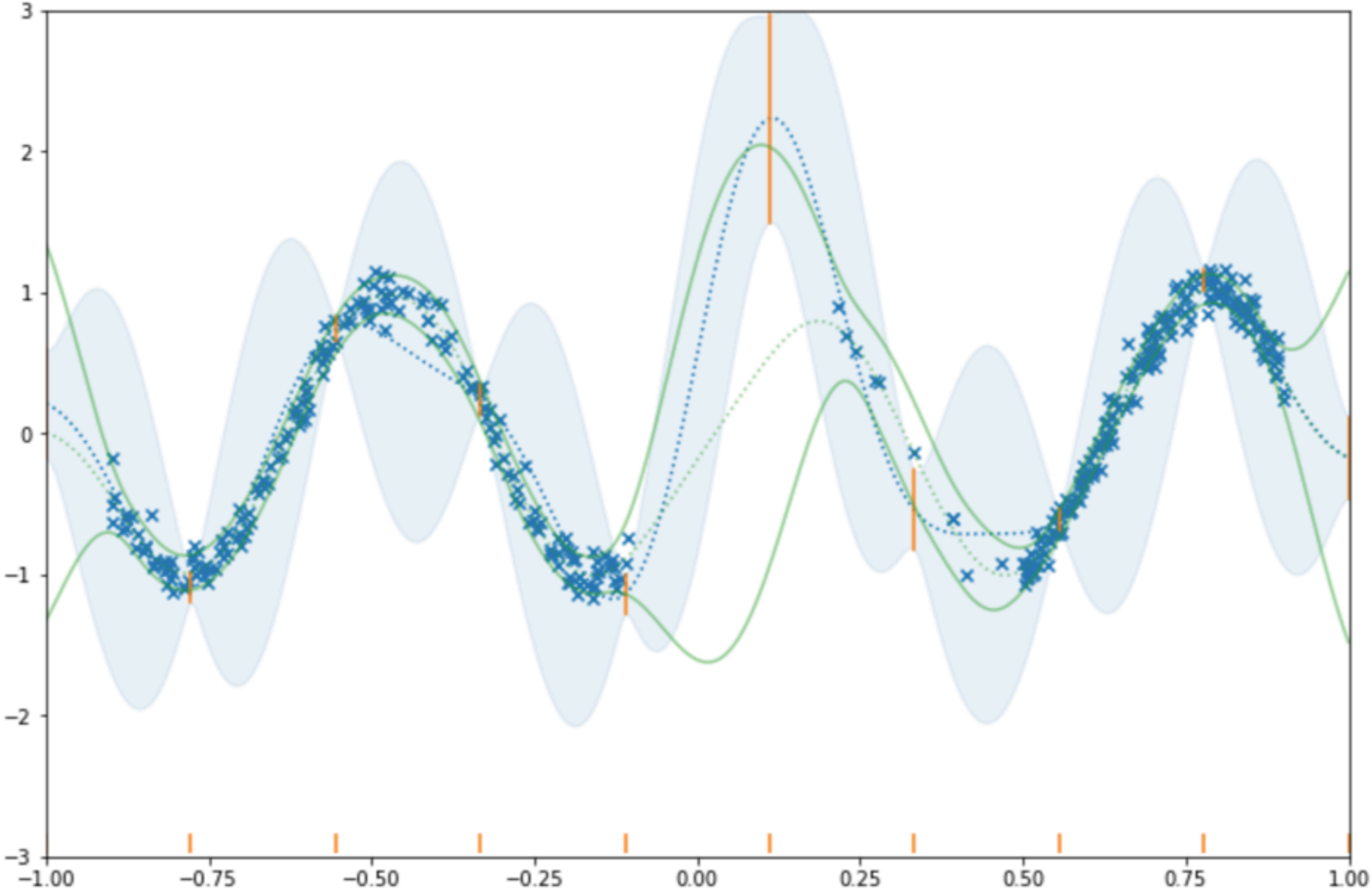
$$q(f(x_n)) = \mathcal{N}(f(x_n)|\tilde{\mathbf{k}}(x)^\top \tilde{\mathbf{K}}^{-1}\tilde{\mathbf{m}}, k(x,x) - \tilde{\mathbf{k}}(x)^\top \tilde{\mathbf{K}}^{-1}\tilde{\mathbf{k}}(x) + \tilde{\mathbf{k}}(x)^\top \tilde{\mathbf{K}}^{-1}\tilde{\mathbf{S}}\tilde{\mathbf{K}}^{-1}\tilde{\mathbf{k}}(x))$$

Interpretation

- ‘Compression’ of data into the inducing points
- ‘Sufficient statistics’
- ‘Pseudo-data’
- Very closely connected to other methods.
- VI has nice behaviour when the posterior is close to the true posterior
- Always safe to optimize inducing locations



Can still lead to bad results…

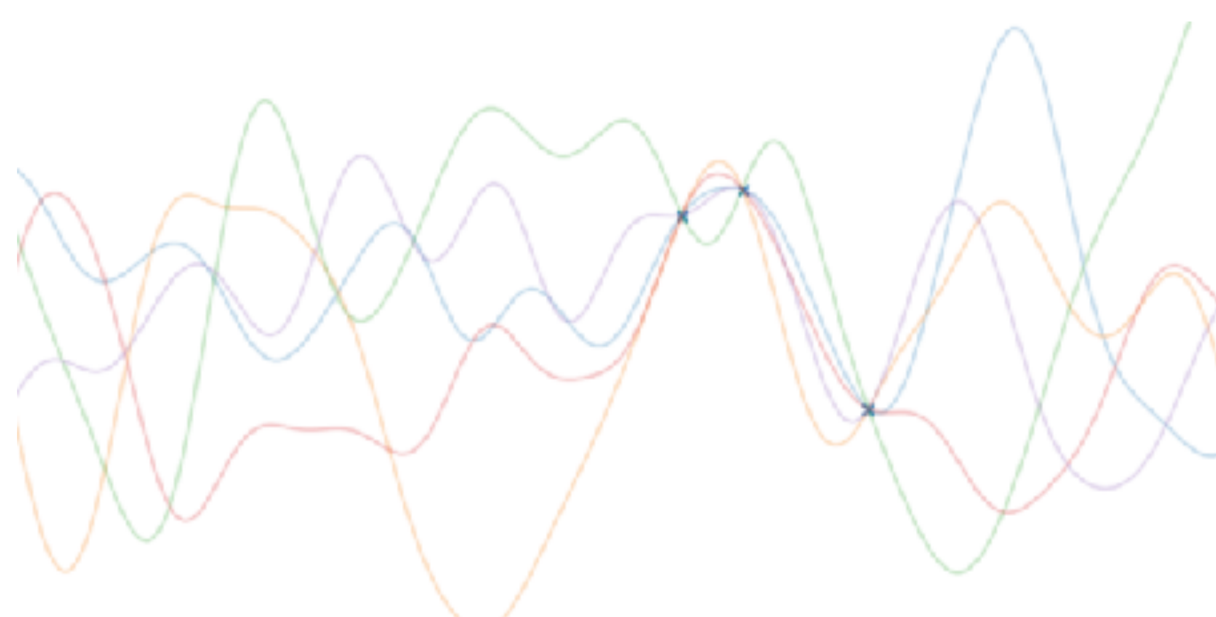


Further details:

- The data term is a sum - possible to subsample ('minibatch') data
- Special case of a Gaussian likelihood: closed form solution exist for **m**, **S**
- Natural gradients can be used, or alternatively direct optimization of the mean and square root of the covariance
- The same approach works for all likelihoods: deals with conjugacy and computation simultaneously.
- Posterior is 'full-rank' (not diagonal or degenerate)
- If inducing inputs are the data, then recover the non-conjugate approach from earlier
- Also possible to perform HMC over the inducing points in a hybrid approach.

Overview

- ~~Review GPs and VI~~
- ~~Establish what problems we want to solve~~
- ~~Discuss alternative approaches~~
- ~~VI for GPs part 1 (conjugacy)~~
- ~~VI for GPs part 2 (scalability)~~
- **Deep GPs**



Model

$$p(y, \{f^l\}_{l=1}^L) = \underbrace{\prod_{i=1}^N p(y_i | f^L(f^{L-1}(\dots f^1(x_i))))}_{\text{likelihood}} \underbrace{\prod_{l=1}^L p(f^l)}_{\text{prior}}$$

$$p(f^\ell) = \mathcal{GP}(m^\ell, k^\ell)$$

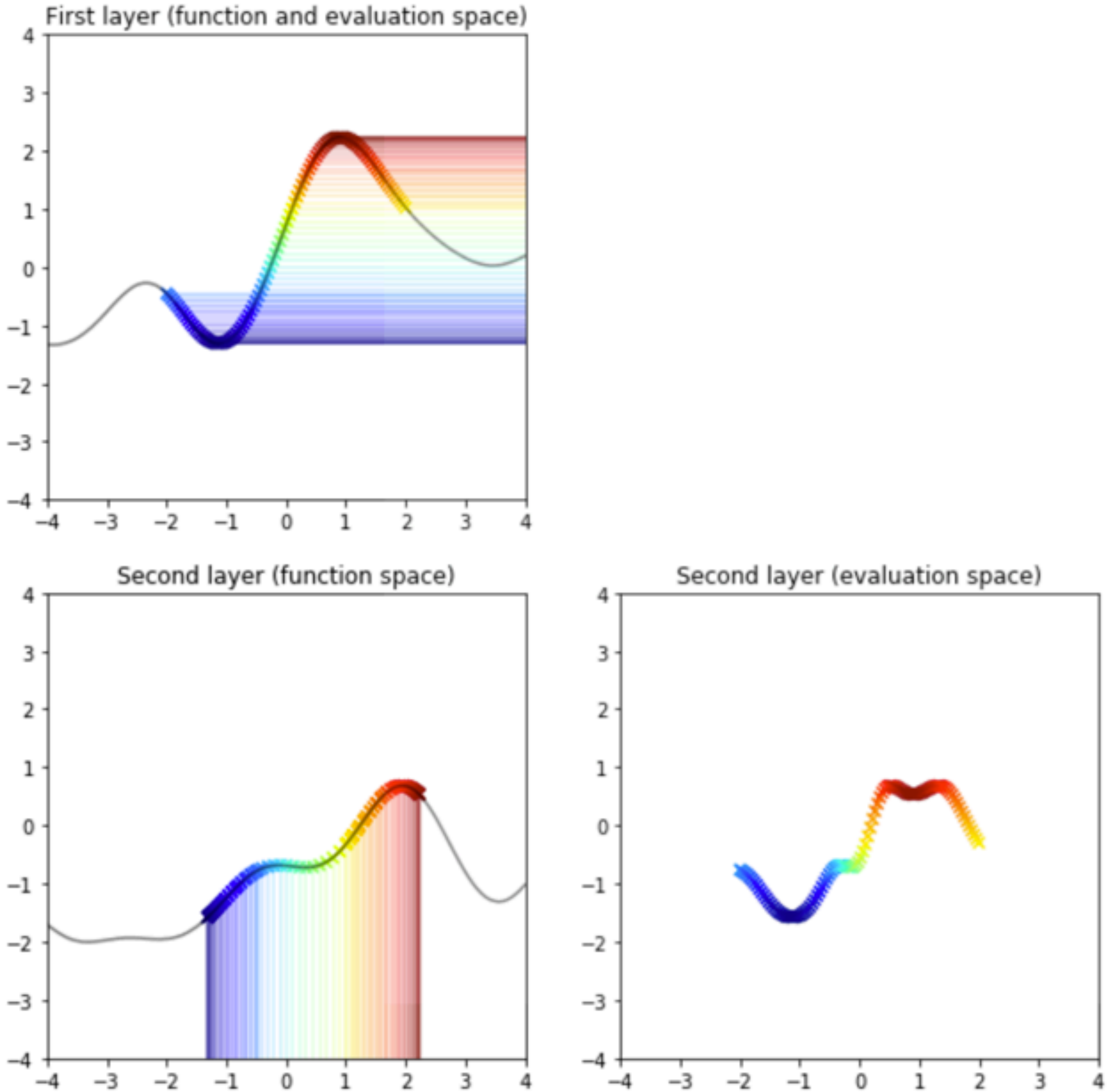
Two layer case

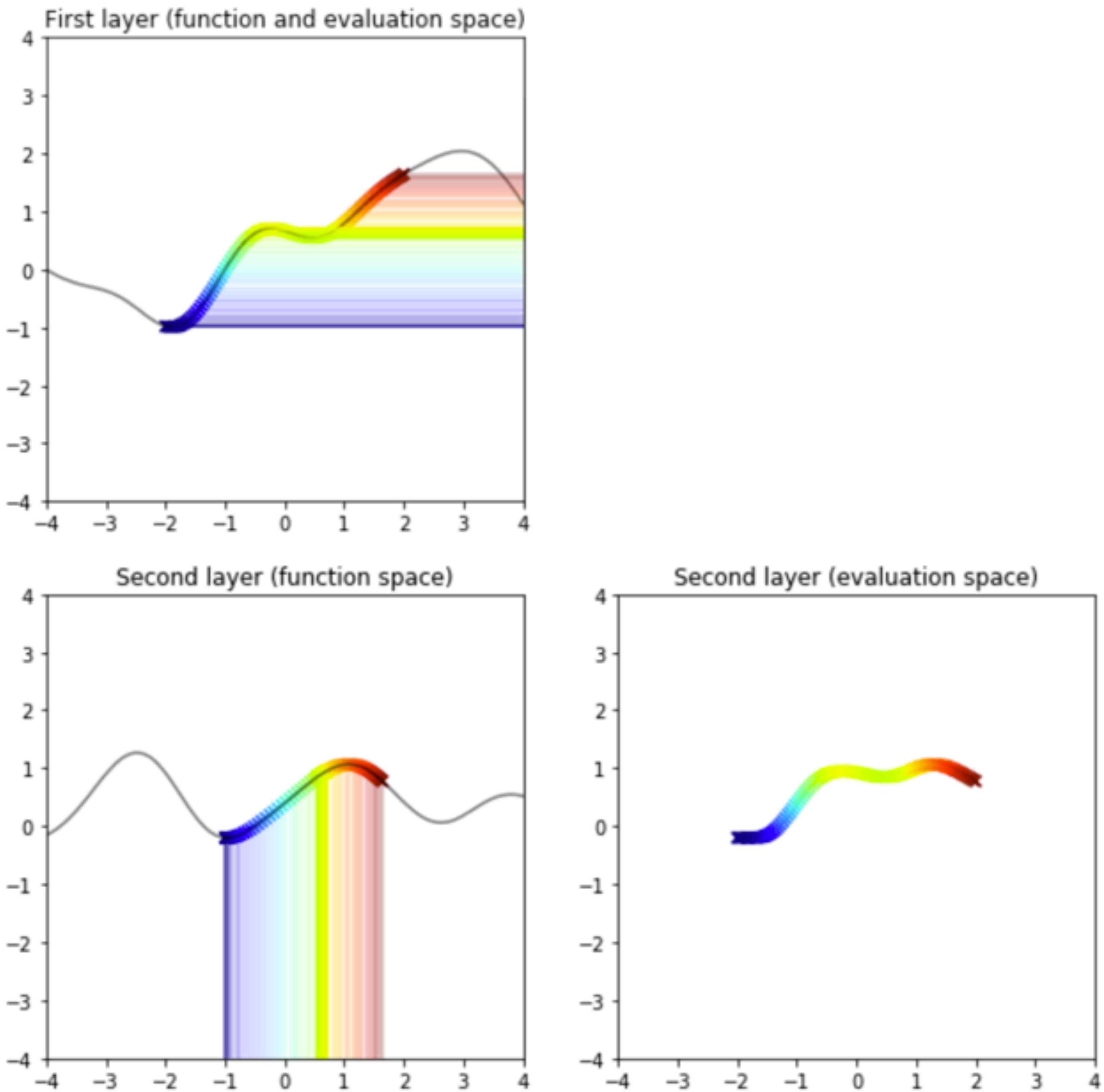
$$p(y, f^1, f^2) = \underbrace{\prod_{i=1}^{N} p\left(y_i | f^2(f^1(x_i))\right)}_{\text{likelihood}} \underbrace{p(f^1)p(f^2)}_{\text{prior}}$$

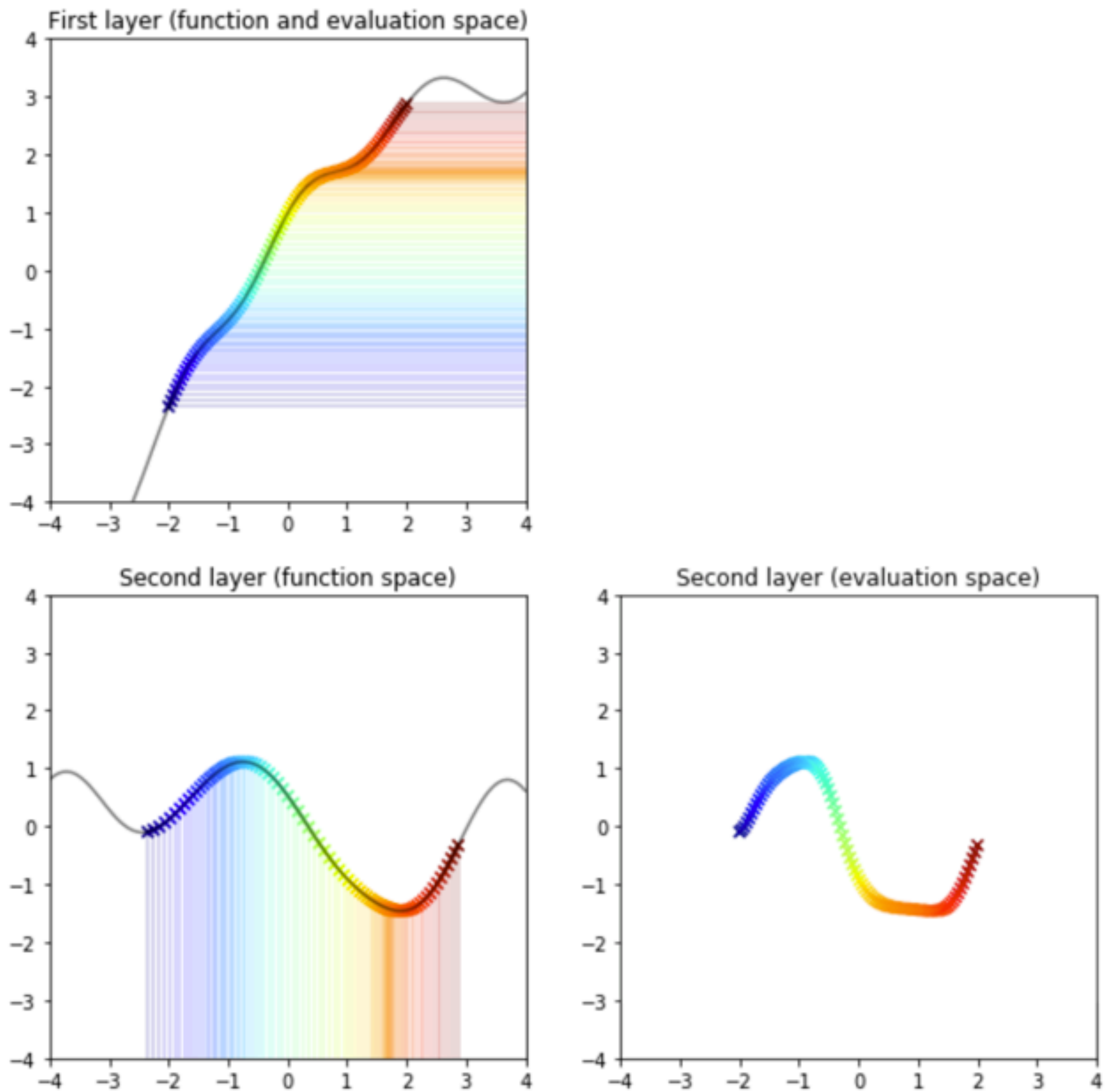
Variational posterior

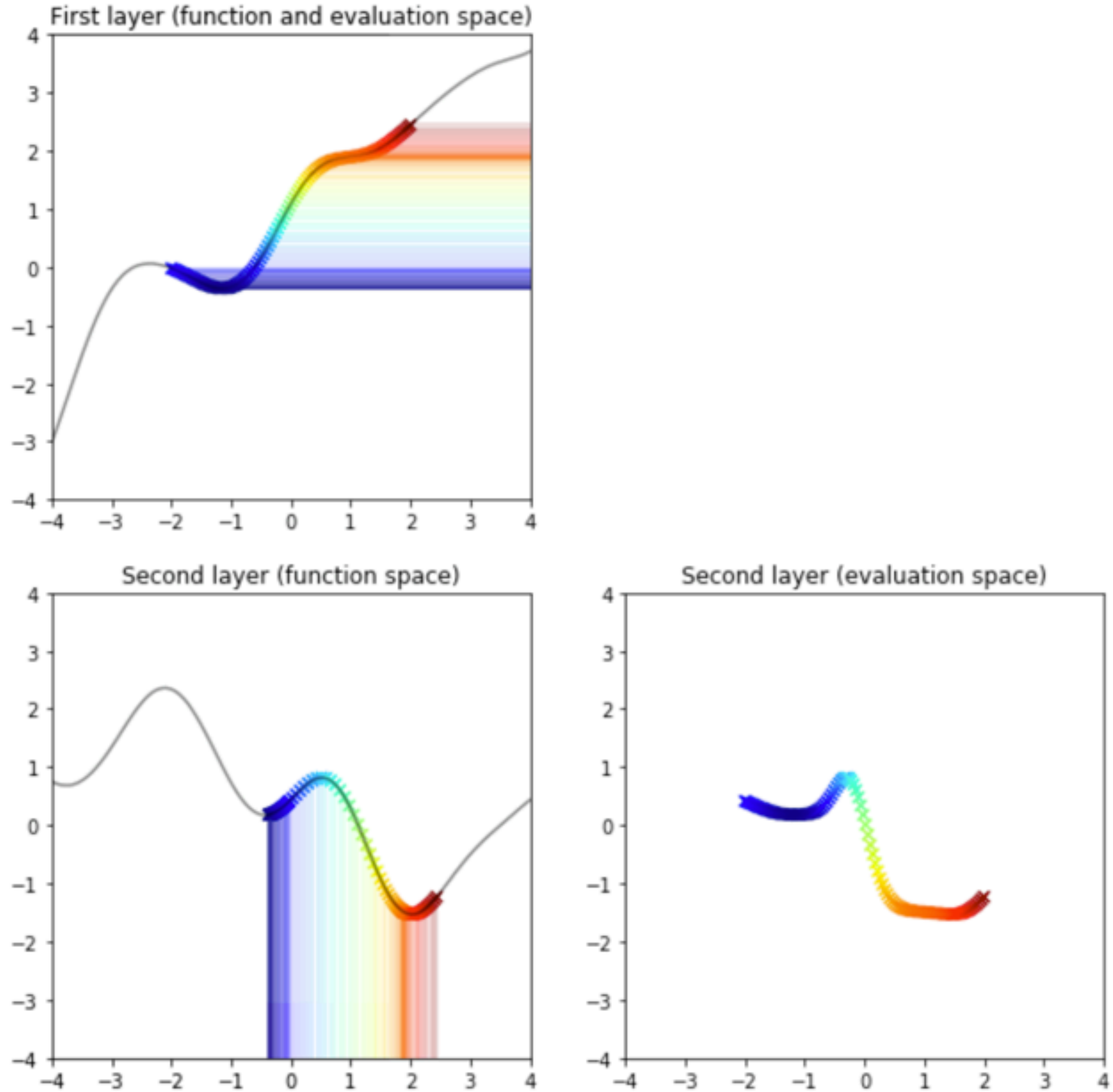
$$q(f^1, f^2) = q(f^1)q(f^2)$$

$$q(f^\ell) = p(f_*^\ell | \tilde{\mathbf{f}}^\ell) q(\tilde{\mathbf{f}}^\ell) \qquad q(\tilde{\mathbf{f}}^\ell) = \mathcal{N}(\mathbf{m}^\ell, \mathbf{S}^\ell)$$


First layer (function and evaluation space)
Second layer (function space)
Second layer (evaluation space)


First layer (function and evaluation space)
Second layer (function space)
Second layer (evaluation space)


First layer (function and evaluation space)
Second layer (function space)
Second layer (evaluation space)
4
3
2
1
0
−1
−2
−3
−4
−4 −3 −2 −1 0 1 2 3 4


First layer (function and evaluation space)
Second layer (function space)
Second layer (evaluation space)
4
3
2
1
0
−1
−2
−3
−4

As in the single layer case, we have

$$\begin{aligned}\mu_{\mathbf{m}^\ell} =& m^\ell(x) + \mathbf{k}^\ell(x)^\top \mathbf{K}^{\ell^{-1}} \mathbf{m}^\ell \\ \Sigma_{\mathbf{S}^\ell}(x, x') =& k(x, x') + \mathbf{k}^\ell(x)^\top \mathbf{K}^{\ell^{-1}} (\mathbf{S}^\ell - \mathbf{K}^\ell) \mathbf{K}^{\ell^{-1}} \mathbf{k}^\ell(x')\end{aligned}$$

The bound is

$$\mathcal{L}_q = \mathbb{E}_{q(f^1)q(f^2)} \log \prod_{n=1}^N p\left(y_i | f^2(f^1(x_n))\right) - \mathrm{KL}(q(f^1)||p(f^1)) - \mathrm{KL}(q(f^2)||p(f^2))$$

Which simplifies to

$$\mathcal{L}_q = \sum_{i=1}^N \underbrace{\mathbb{E}_{q(f^1)q(f^2)} \log p\left(y_i | f^2(f^1(x_i))\right)}_{=L_i} - \mathrm{KL}(q(\tilde{\mathbf{f}}^1)||p(\tilde{\mathbf{f}}^1)) - \mathrm{KL}(q(\tilde{\mathbf{f}}^2)||p(\tilde{\mathbf{f}}^2))$$

'Reparameterization trick'

$$\begin{aligned} L_i &= E_{q(f^2)q(f^1)} \log p\left(y_i | f^2(f^1(x_i))\right) \\ &= E_{q(f^2)p(f^1(x_i))} \log p\left(y_i | f^2(f^1(x_i))\right) \\ &= E_{q(f^2)p(\epsilon^1)} \log p\left(y_i | f^2(\mu_{\mathbf{m}^1}(x_i) + \epsilon^1 \sqrt{k_{\mathbf{S}^1}(x_i, x_i)}\right) \\ &= E_{q(f^2)p(\epsilon^1)} \log p\left(y_i | f^2(z_i(\epsilon^1))\right) \end{aligned}$$

$$\begin{aligned} L_i &= E_{q(f^2)p(\epsilon^1)} \log p\left(y_i | f^2(z_i(\epsilon^1))\right) \\ &= E_{q(f^2(z_i(\epsilon^1)))p(\epsilon^1)} \log p\left(y_i | f^2(z_i(\epsilon^1))\right) \\ &= E_{p(\epsilon^2)p(\epsilon^1)} \log p\left(y_i | f^2(z_i(\epsilon^1))\right) \\ &= E_{p(\epsilon^2)p(\epsilon^1)} \log p\left(y_i | \mu_{\mathbf{m}^2}(z_i(\epsilon^1)) + \epsilon^2 \sqrt{k_{\mathbf{S}^2}(z_i(\epsilon^1), z_i(\epsilon^1))}\right) \end{aligned}$$

Integral is now over 'white' Gaussian variables.
Can take the expectation through sampling.