Sparse GP Approximations

An Introduction to Pseudo-Input Methods

Markus Kaiser
markus.kaiser@siemens.com

5. April 2018

Siemens AG — CT RDA BAM LSY-DE
Technical University of Munich

Agenda

1. Standard GP Recap
2. Model Augmentation
3. SPGP: More Hyperparameters
4. SGPR: A Variational Approach
5. SVGP: Stochastic Optimization

Standard GP Recap

Gaussian Processes

Definition

A **Gaussian Process (GP)** is a collection of random variables $\{F_x\}$, any finite subset of which has a joint Gaussian distribution.

- Extension of Gaussians to (infinite) function spaces
- F_x models the function value $f(x)$

Mean and Kernel Functions

A GP is completely determined by two functions.

Mean function $\mu_f(x) = \mathbb{E}[f(x)]$

Kernel function $\mathcal{K}(x, x') = \mathrm{cov}[f(x), f(x')]$

GP Posterior

GP predictive posterior

The predictive posterior for a function $f \sim \mathcal{GP}(\mathbf{0}, \mathcal{K})$ with observations $\boldsymbol{y} = f(\boldsymbol{X}) + \mathcal{N}(\mathbf{0}, \sigma_n^2 \mathbf{I})$ is

$$\mathrm{p}(\boldsymbol{f}_* \mid \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{X}_*) = \mathcal{N}(\boldsymbol{f}_* \mid \boldsymbol{\mu}_*, \Sigma_*), \text{ with}$$

$$\boldsymbol{\mu}_* = \boldsymbol{K}_{*f} \left(\boldsymbol{K}_{ff} + \sigma_n^2 \mathbf{I}\right)^{-1} \boldsymbol{y}$$

$$\Sigma_* = \boldsymbol{K}_{**} - \boldsymbol{K}_{*f} \left(\boldsymbol{K}_{ff} + \sigma_n^2 \mathbf{I}\right)^{-1} \boldsymbol{K}_{f*}.$$

GP Posterior

GP predictive posterior

The predictive posterior for a function $f \sim \mathcal{GP}(\mathbf{0}, \mathcal{K})$ with observations $\boldsymbol{y} = f(\boldsymbol{X}) + \mathcal{N}(\mathbf{0}, \sigma_n^2 \mathbf{I})$ is

$$\begin{aligned} \mathrm{p}(\boldsymbol{f}_* \mid \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{X}_*) &= \mathcal{N}(\boldsymbol{f}_* \mid \boldsymbol{\mu}_*, \Sigma_*), \text{ with} \\ \boldsymbol{\mu}_* &= \boldsymbol{K}_{*f} \left(\boldsymbol{K}_{ff} + \sigma_n^2 \mathbf{I}\right)^{-1} \boldsymbol{y} \\ \Sigma_* &= \boldsymbol{K}_{**} - \boldsymbol{K}_{*f} \left(\boldsymbol{K}_{ff} + \sigma_n^2 \mathbf{I}\right)^{-1} \boldsymbol{K}_{f*}. \end{aligned}$$

- We always need to consider the **whole dataset**
- The matrix inversion is $\mathcal{O}(N^3)$ in time and $\mathcal{O}(N^2)$ in space
- With precomputation we still have to pay $\mathcal{O}(N^2)$ for Σ_*

GP Marginal Likelihood

GP marginal likelihood

We usually use the marginal likelihood $\mathcal{L}$ for model selection with maximum likelihood

$$
\begin{aligned}
\mathcal{L}^{\mathrm{GP}}(\boldsymbol{\theta}) &= -\log \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{\theta}) \\
&= -\log \int \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{f}, \boldsymbol{\theta})\, \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{f} \\
&= -\log \int \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{f}, \sigma_n^2 \mathbf{I})\, \mathcal{N}(\boldsymbol{f} \mid \mathbf{0}, \boldsymbol{K}_{\boldsymbol{f}\boldsymbol{f}})\, \mathrm{d}\boldsymbol{f} \\
&= -\log \mathcal{N}(\boldsymbol{y} \mid \mathbf{0}, \boldsymbol{K}_{\boldsymbol{f}\boldsymbol{f}} + \sigma_n^2 \mathbf{I}) \\
&= \frac{1}{2} \boldsymbol{y}^{\mathsf{T}} \left(\boldsymbol{K}_{\boldsymbol{f}\boldsymbol{f}} + \sigma_n^2 \mathbf{I}\right)^{-1} \boldsymbol{y} + \frac{1}{2} \log \left| \boldsymbol{K}_{\boldsymbol{f}\boldsymbol{f}} + \sigma_n^2 \mathbf{I} \right| + \frac{N}{2} \log(2\pi).
\end{aligned}
$$

GP Marginal Likelihood

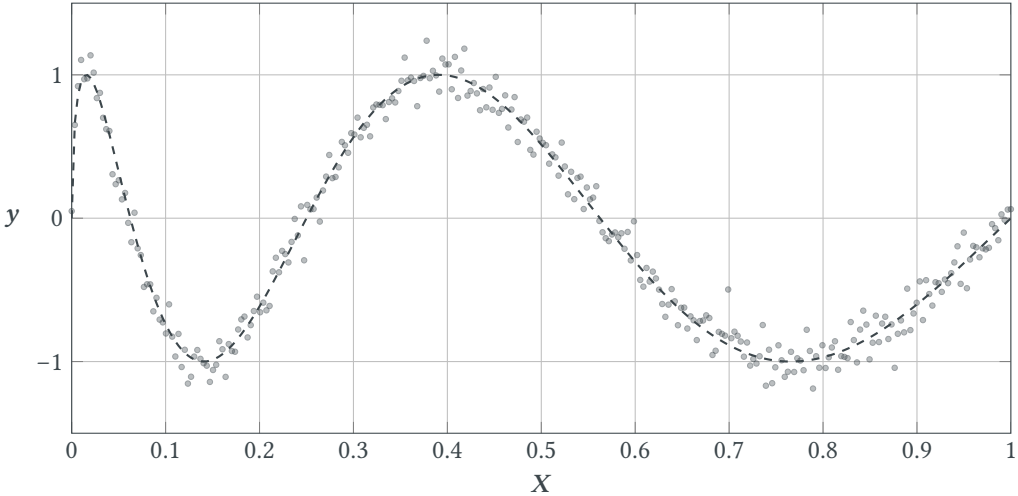
GP marginal likelihood

We usually use the marginal likelihood $\mathcal{L}$ for model selection with maximum likelihood

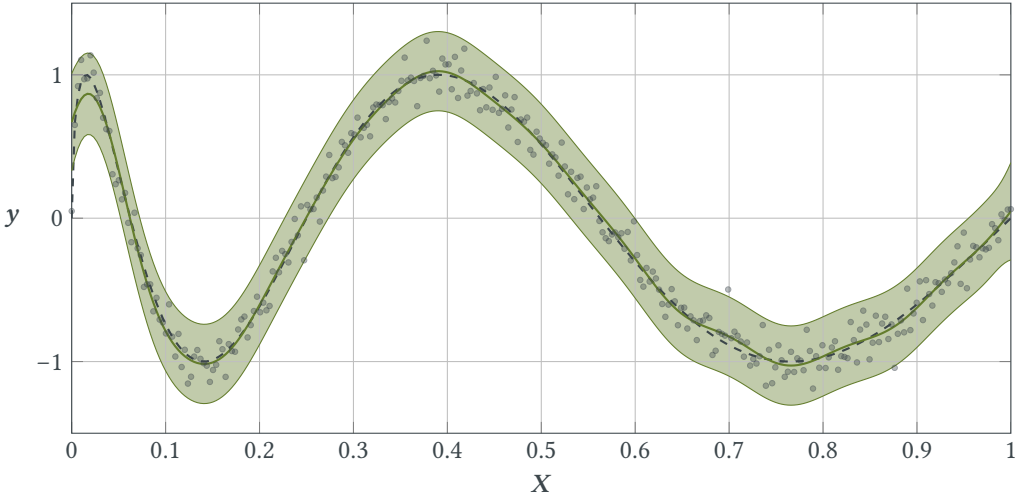
$$
\begin{aligned}
\mathcal{L}^{\mathrm{GP}}(\boldsymbol{\theta}) &= -\log \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{\theta}) \\
&= -\log \int \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{f}, \boldsymbol{\theta})\, \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{f} \\
&= -\log \int \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{f}, \sigma_n^2 \mathbf{I})\, \mathcal{N}(\boldsymbol{f} \mid \mathbf{0}, \boldsymbol{K}_{\boldsymbol{f}\boldsymbol{f}})\, \mathrm{d}\boldsymbol{f} \\
&= -\log \mathcal{N}(\boldsymbol{y} \mid \mathbf{0}, \boldsymbol{K}_{\boldsymbol{f}\boldsymbol{f}} + \sigma_n^2 \mathbf{I}) \\
&= \frac{1}{2} \boldsymbol{y}^{\mathsf{T}} \left(\boldsymbol{K}_{\boldsymbol{f}\boldsymbol{f}} + \sigma_n^2 \mathbf{I}\right)^{-1} \boldsymbol{y} + \frac{1}{2} \log \left| \boldsymbol{K}_{\boldsymbol{f}\boldsymbol{f}} + \sigma_n^2 \mathbf{I} \right| + \frac{N}{2} \log(2\pi).
\end{aligned}
$$

- Choose $\boldsymbol{\theta}^* \in \operatorname{argmin}_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta})$
- $\mathcal{L}$ is also $\mathcal{O}(N^3)$ in time and $\mathcal{O}(N^2)$ in space

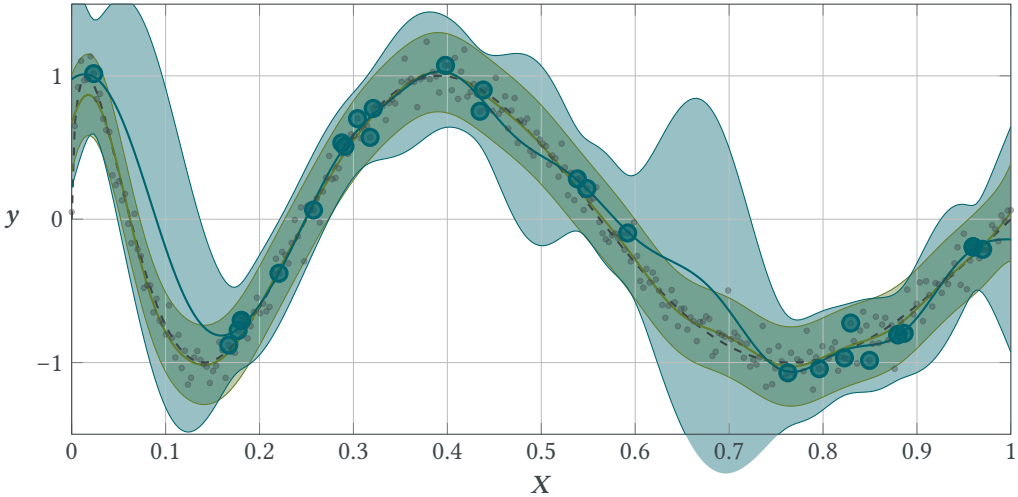
Reducing the Data Set



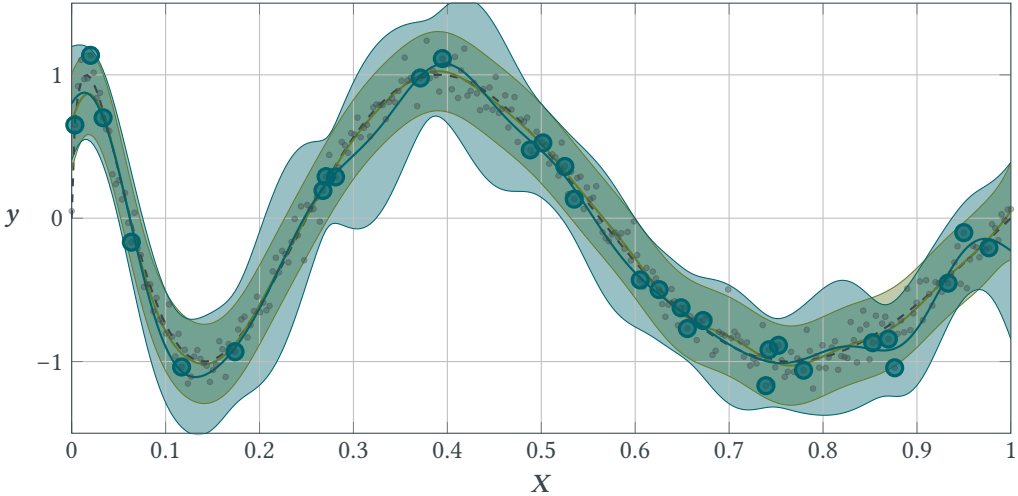
Reducing the Data Set



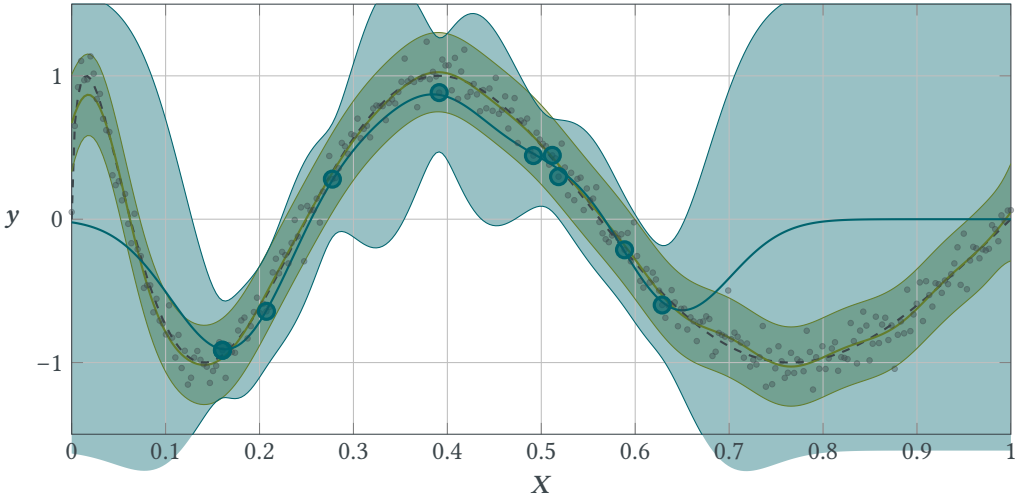
Reducing the Data Set



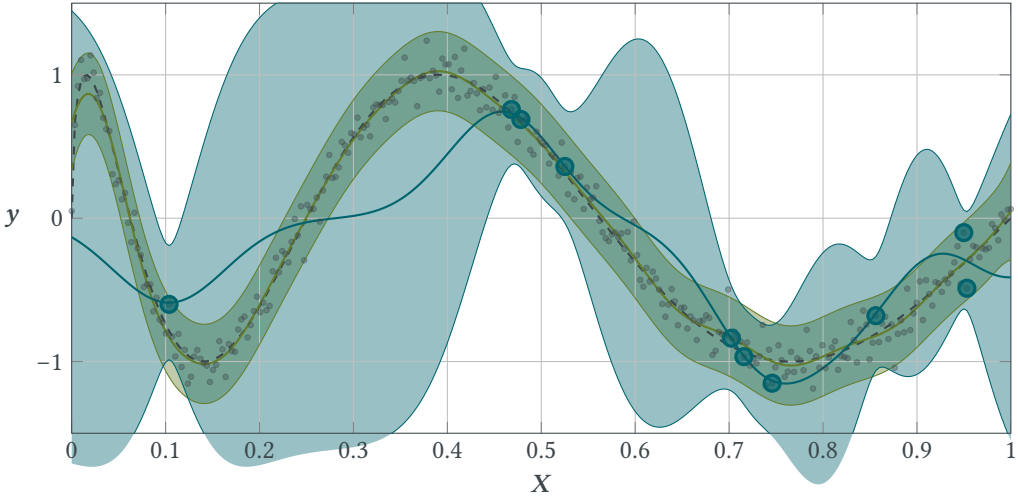
Reducing the Data Set



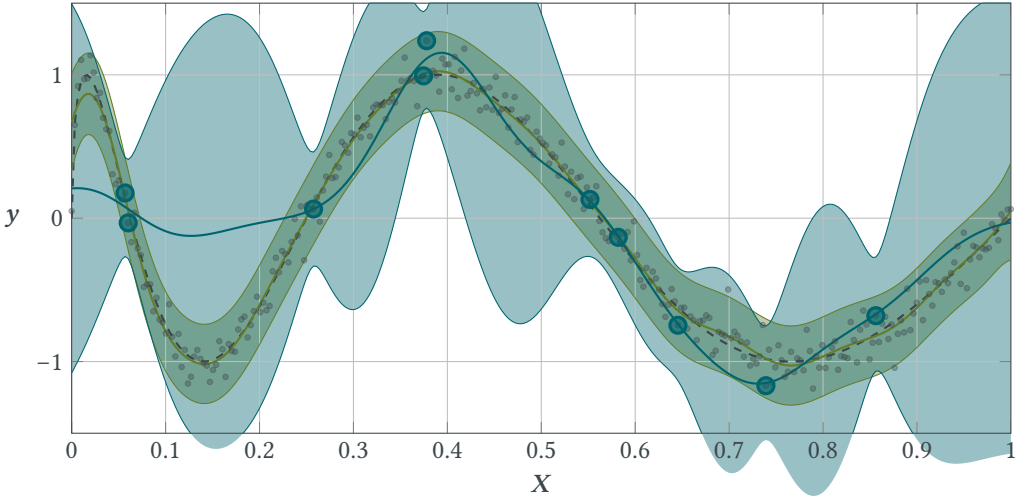
Reducing the Data Set



Reducing the Data Set



Reducing the Data Set



Can we somehow not ignore most of the data?

Model Augmentation

Pseudo Data

- Keep N original observations $\mathcal{D} = (X, \boldsymbol{y})$
- Assume M different **pseudo data points** $\bar{\mathcal{D}} = (Z, \boldsymbol{u})$
- $\boldsymbol{u} = f(Z)$ are **observations of the latent function**
- The $\bar{\mathcal{D}}$ are also called **inducing points**

Pseudo Data

- Keep N original observations $\mathcal{D} = (\boldsymbol{X}, \boldsymbol{y})$
- Assume M different **pseudo data points** $\bar{\mathcal{D}} = (\boldsymbol{Z}, \boldsymbol{u})$
- $\boldsymbol{u} = f(\boldsymbol{Z})$ are **observations of the latent function**
- The $\bar{\mathcal{D}}$ are also called **inducing points**

Augmented GP Joint

Due to the consistency of GPs, the data and pseudo data are **jointly Gaussian**.

$$\mathrm{p}(\boldsymbol{f}, \boldsymbol{u}) = \mathcal{N}\left(\begin{pmatrix} \boldsymbol{f} \\ \boldsymbol{u} \end{pmatrix} \middle| \, \boldsymbol{0}, \begin{pmatrix} \boldsymbol{K}_{\boldsymbol{ff}} & \boldsymbol{K}_{\boldsymbol{fu}} \\ \boldsymbol{K}_{\boldsymbol{uf}} & \boldsymbol{K}_{\boldsymbol{uu}} \end{pmatrix} \right)$$

- We usually drop the conditioning on $\boldsymbol{X}$, $\boldsymbol{Z}$ and $\boldsymbol{\theta}$

Conditioning on the Pseudo Data

Augmented GP conditional likelihood

Because $\mathrm{p}(\boldsymbol{f}, \boldsymbol{u})$ is Gaussian, we can calculate the conditional likelihood easily.

$$
\begin{aligned}
\mathrm{p}(\boldsymbol{y} \mid \boldsymbol{u}) &= \int \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{f})\, \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u})\, \mathrm{d}\boldsymbol{f} \\
&= \int \mathcal{N}\left(\boldsymbol{y} \,\middle|\, \boldsymbol{f}, \sigma_n^2 \mathbf{I}\right) \mathcal{N}\left(\boldsymbol{f} \,\middle|\, \boldsymbol{K}_{\boldsymbol{fu}} \boldsymbol{K}_{\boldsymbol{uu}}^{-1} \boldsymbol{u}, \boldsymbol{K}_{\boldsymbol{ff}} - \boldsymbol{K}_{\boldsymbol{fu}} \boldsymbol{K}_{\boldsymbol{uu}}^{-1} \boldsymbol{K}_{\boldsymbol{uf}}\right) \mathrm{d}\boldsymbol{f} \\
&= \mathcal{N}\left(\boldsymbol{y} \,\middle|\, \boldsymbol{K}_{\boldsymbol{fu}} \boldsymbol{K}_{\boldsymbol{uu}}^{-1} \boldsymbol{u}, \boldsymbol{K}_{\boldsymbol{ff}} - \boldsymbol{K}_{\boldsymbol{fu}} \boldsymbol{K}_{\boldsymbol{uu}}^{-1} \boldsymbol{K}_{\boldsymbol{uf}} + \sigma_n^2 \mathbf{I}\right) \\
&= \mathcal{N}\left(\boldsymbol{y} \,\middle|\, \boldsymbol{K}_{\boldsymbol{fu}} \boldsymbol{K}_{\boldsymbol{uu}}^{-1} \boldsymbol{u}, \boldsymbol{K}_{\boldsymbol{ff}} - \boldsymbol{Q}_{\boldsymbol{ff}} + \sigma_n^2 \mathbf{I}\right)
\end{aligned}
$$

- We call $\boldsymbol{Q}_{\boldsymbol{ff}}$ the low rank covariance

Conditioning on the Pseudo Data

Augmented GP conditional likelihood

Because $\mathrm{p}(\boldsymbol{f}, \boldsymbol{u})$ is Gaussian, we can calculate the conditional likelihood easily.

$$
\begin{aligned}
\mathrm{p}(\boldsymbol{y} \mid \boldsymbol{u}) &= \int \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{f})\, \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u})\, \mathrm{d}\boldsymbol{f} \\
&= \int \mathcal{N}\left(\boldsymbol{y} \,\middle|\, \boldsymbol{f}, \sigma_n^2 \mathbf{I}\right) \mathcal{N}\left(\boldsymbol{f} \,\middle|\, \boldsymbol{K}_{\boldsymbol{fu}} \boldsymbol{K}_{\boldsymbol{uu}}^{-1} \boldsymbol{u}, \boldsymbol{K}_{\boldsymbol{ff}} - \boldsymbol{K}_{\boldsymbol{fu}} \boldsymbol{K}_{\boldsymbol{uu}}^{-1} \boldsymbol{K}_{\boldsymbol{uf}}\right) \mathrm{d}\boldsymbol{f} \\
&= \mathcal{N}\left(\boldsymbol{y} \,\middle|\, \boldsymbol{K}_{\boldsymbol{fu}} \boldsymbol{K}_{\boldsymbol{uu}}^{-1} \boldsymbol{u}, \boldsymbol{K}_{\boldsymbol{ff}} - \boldsymbol{K}_{\boldsymbol{fu}} \boldsymbol{K}_{\boldsymbol{uu}}^{-1} \boldsymbol{K}_{\boldsymbol{uf}} + \sigma_n^2 \mathbf{I}\right) \\
&= \mathcal{N}\left(\boldsymbol{y} \,\middle|\, \boldsymbol{K}_{\boldsymbol{fu}} \boldsymbol{K}_{\boldsymbol{uu}}^{-1} \boldsymbol{u}, \boldsymbol{K}_{\boldsymbol{ff}} - \boldsymbol{Q}_{\boldsymbol{ff}} + \sigma_n^2 \mathbf{I}\right)
\end{aligned}
$$

- We call $\boldsymbol{Q}_{\boldsymbol{ff}}$ the low rank covariance
- This is just the predictive distribution again
- Assuming iid data, we can calculate the likelihood in (at most) $\mathcal{O}(N^2)$

But what use is it if we have no idea about $\tilde{D}$?

SPGP: More Hyperparameters

SPGP

Sparse Pseudo-input Gaussian Process (SPGP)

We interpret the augmented model as a new regression problem arising from $\tilde{\mathcal{D}}$.

- Assume **conditional independence of f (and y) given $\boldsymbol{u}$**
- We can choose Z and only have to handle $\boldsymbol{u}$

SPGP

Sparse Pseudo-input Gaussian Process (SPGP)

We interpret the augmented model as a new regression problem arising from $\tilde{\mathcal{D}}$.

- Assume conditional independence of f (and y) given $\boldsymbol{u}$
- We can choose Z and only have to handle $\boldsymbol{u}$

$$
\begin{aligned}
\mathrm{p}(\boldsymbol{y} \mid \boldsymbol{u}) &= \prod_n \mathrm{p}(y_n \mid \boldsymbol{u}) \\
&= \mathcal{N}\left(\boldsymbol{y} \,\middle|\, K_{fu} K_{uu}^{-1} \boldsymbol{u}, \operatorname{diag}(K_{ff} - Q_{ff}) + \sigma_n^2 \mathbf{I}\right)
\end{aligned}
$$

Using the low rank covariance

$$
Q_{ab} := K_{au} K_{uu}^{-1} K_{ub}
$$

SPGP Marginal Likelihood

SPGP Marginal Likelihood

$$
\begin{aligned}
\mathcal{L}^{\text{SPGP}}(\boldsymbol{\theta}) &= -\log \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{\theta}) \\
&= -\log \int \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{u})\, \mathrm{p}(\boldsymbol{u})\, \mathrm{d}\boldsymbol{u} \\
&= -\log \int \mathcal{N}\left(\boldsymbol{y} \,\middle|\, K_{fu} K_{uu}^{-1} \boldsymbol{u}, \operatorname{diag}(K_{ff} - Q_{ff}) + \sigma_n^2 \mathbf{I}\right) \mathcal{N}(\boldsymbol{u} \mid \mathbf{0}, K_{uu})\, \mathrm{d}\boldsymbol{u} \\
&= -\log \mathcal{N}\left(\boldsymbol{y} \,\middle|\, \mathbf{0}, Q_{ff} + \operatorname{diag}(K_{ff} - Q_{ff}) + \sigma_n^2 \mathbf{I}\right)
\end{aligned}
$$

- We do not know $\boldsymbol{u} = f(Z)$, so we **marginalize**
- Use the original **GP prior** $\mathrm{p}(\boldsymbol{u} \mid \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{u} \mid \mathbf{0}, K_{uu})$
- **Joint optimization of Z** with other hyperparameters
- Time-complexity is $\mathcal{O}(NM^2)$, **linear** in the data

Interpretation of the SPGP

$$\mathcal{L}^{\text{GP}}(\boldsymbol{\theta}) = -\log \mathcal{N}\left(\boldsymbol{y} \,\middle|\, \mathbf{0}, \boldsymbol{K}_{ff} + \sigma_n^2 \mathbf{I}\right)$$

$$\mathcal{L}^{\text{SPGP}}(\boldsymbol{\theta}) = -\log \mathcal{N}\left(\boldsymbol{y} \,\middle|\, \mathbf{0}, \boldsymbol{Q}_{ff} + \operatorname{diag}(\boldsymbol{K}_{ff} - \boldsymbol{Q}_{ff}) + \sigma_n^2 \mathbf{I}\right)$$

SPGP kernel

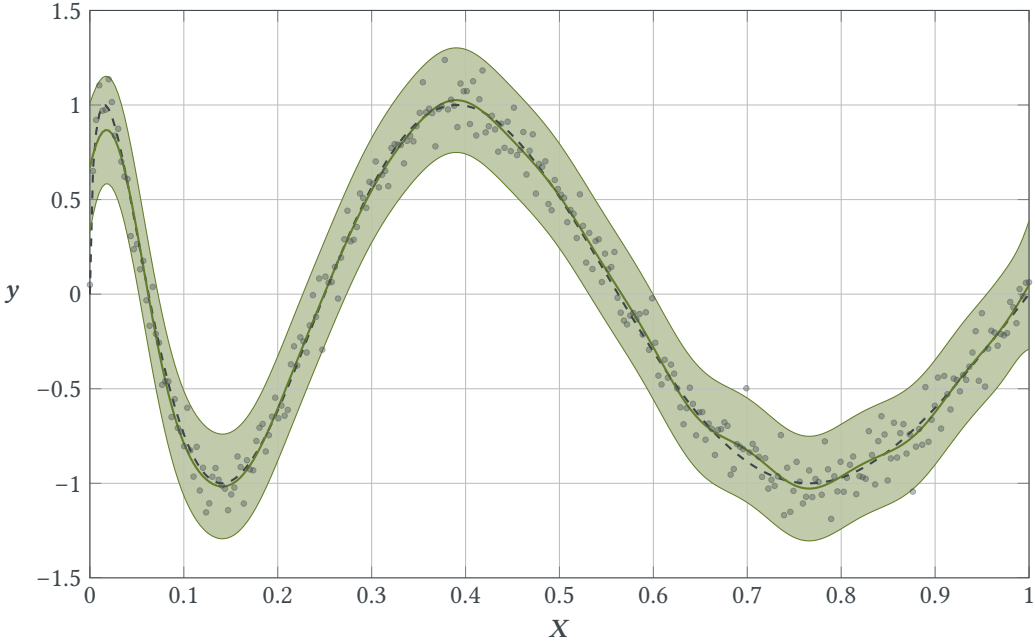
The SPGP derived from $\mathcal{GP}(\mathbf{0}, \mathcal{K})$ is a **new Gaussian Process $\mathcal{GP}(\mathbf{0}, \mathcal{K}^{\text{SPGP}})$** with

$$\mathcal{K}^{\text{SPGP}}(\boldsymbol{x}, \boldsymbol{x}') = \mathcal{Q}(\boldsymbol{x}, \boldsymbol{x}') + \delta_{\boldsymbol{x}\boldsymbol{x}'} \left(\mathcal{K}(\boldsymbol{x}, \boldsymbol{x}') - \mathcal{Q}(\boldsymbol{x}, \boldsymbol{x}')\right)$$

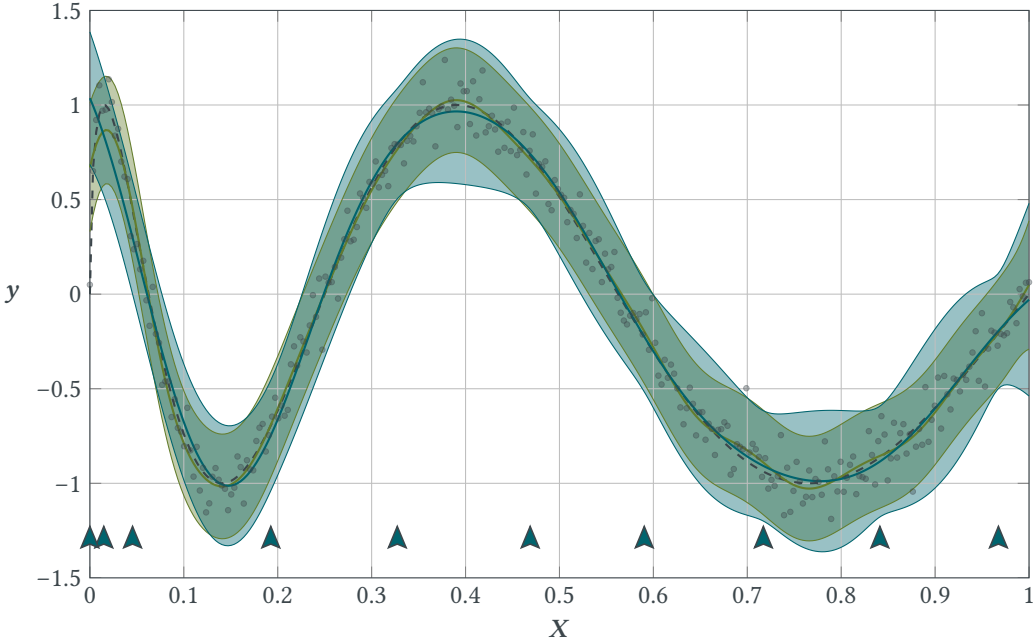
$$\mathcal{Q}(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{K}_{\boldsymbol{x}\boldsymbol{u}} \boldsymbol{K}_{\boldsymbol{u}\boldsymbol{u}}^{-1} \boldsymbol{K}_{\boldsymbol{u}\boldsymbol{x}'}$$

- $\mathcal{K}^{\text{SPGP}}$ is heavily dependent on $\boldsymbol{Z}$ and no longer interpretable
- It has a potentially **huge number of (hyper-)parameters**...
- ...which makes the model prone to **overfitting**

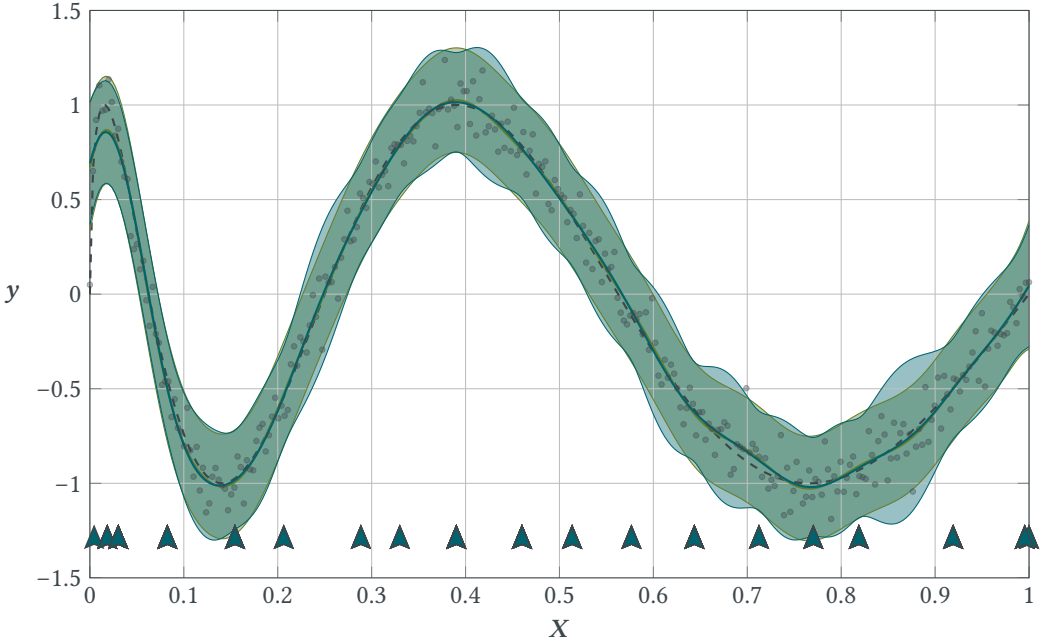
SPGP Showcase



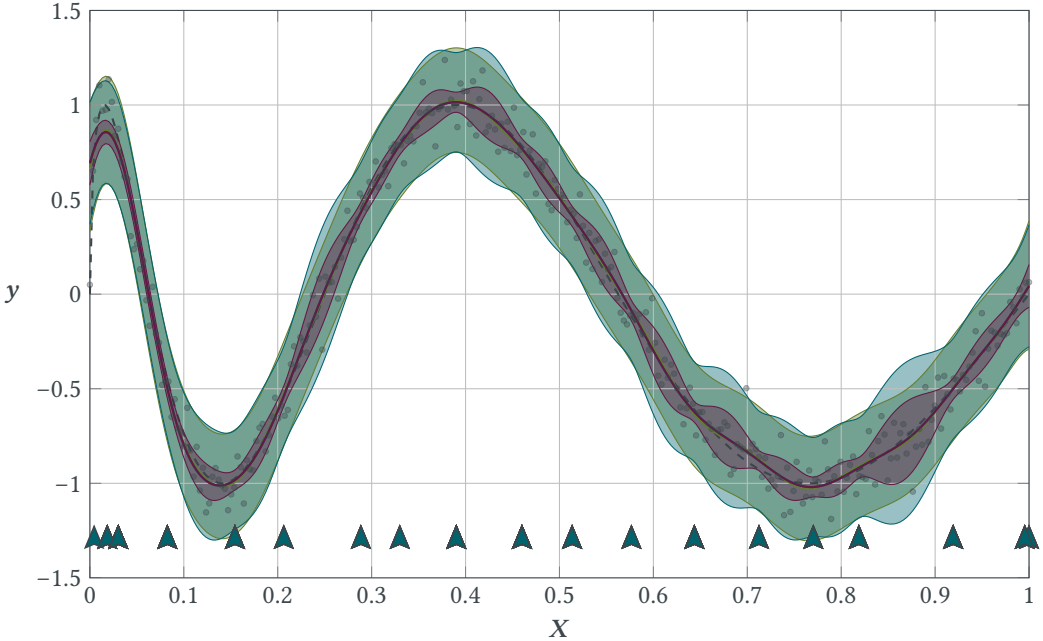
SPGP Showcase



SPGP Showcase



SPGP Showcase



SGPR: A Variational Approach

SGPR

Sparse Gaussian Process Regression (SGPR)

We want to variationally approximate the original Gaussian Process.

- Derive a variational lower bound on the original marginal likelihood $\mathcal{L}^{\mathrm{GP}}$
- Consider a custom variational scheme $\mathrm{q}(\boldsymbol{f}, \boldsymbol{u}) = \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u})\, \mathrm{q}(\boldsymbol{u})$

Sparse Gaussian Process Regression (SGPR)

We want to variationally approximate the original Gaussian Process.

- Derive a **variational lower bound** on the original marginal likelihood $\mathcal{L}^{\mathrm{GP}}$
- Consider a **custom variational scheme** $\mathrm{q}(\boldsymbol{f}, \boldsymbol{u}) = \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u})\, \mathrm{q}(\boldsymbol{u})$

- Inspect the augmented predictive posterior

$$\mathrm{p}(\boldsymbol{f}^* \mid \boldsymbol{y}) = \int \mathrm{p}(\boldsymbol{f}^* \mid \boldsymbol{f}, \boldsymbol{u})\, \mathrm{p}(\boldsymbol{f}, \boldsymbol{u} \mid \boldsymbol{y})\, \mathrm{d}\boldsymbol{f}\, \mathrm{d}\boldsymbol{u}$$

- **$\boldsymbol{u}$ is optimal** if $\boldsymbol{f}^*$ and $\boldsymbol{f}$ are **independent given $\boldsymbol{u}$**

$$\mathrm{p}(\boldsymbol{f}^* \mid \boldsymbol{f}, \boldsymbol{u}) = \mathrm{p}(\boldsymbol{f}^* \mid \boldsymbol{u})$$

$$\mathrm{p}(\boldsymbol{f}, \boldsymbol{u} \mid \boldsymbol{y}) = \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u})\, \mathrm{p}(\boldsymbol{u} \mid \boldsymbol{y})$$

Sparse Gaussian Process Regression (SGPR)

We want to variationally approximate the original Gaussian Process.

- Derive a **variational lower bound** on the original marginal likelihood $\mathcal{L}^{\mathrm{GP}}$
- Consider a **custom variational scheme** $\mathrm{q}(\boldsymbol{f}, \boldsymbol{u}) = \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u})\, \mathrm{q}(\boldsymbol{u})$

- Inspect the augmented predictive posterior

$$\mathrm{p}(\boldsymbol{f}^* \mid \boldsymbol{y}) = \int \mathrm{p}(\boldsymbol{f}^* \mid \boldsymbol{f}, \boldsymbol{u})\, \mathrm{p}(\boldsymbol{f}, \boldsymbol{u} \mid \boldsymbol{y})\, \mathrm{d}\boldsymbol{f}\, \mathrm{d}\boldsymbol{u}$$

- **$\boldsymbol{u}$ is optimal** if $\boldsymbol{f}^*$ and $\boldsymbol{f}$ are **independent given $\boldsymbol{u}$**

$$\mathrm{p}(\boldsymbol{f}^* \mid \boldsymbol{f}, \boldsymbol{u}) = \mathrm{p}(\boldsymbol{f}^* \mid \boldsymbol{u})$$
$$\mathrm{p}(\boldsymbol{f}, \boldsymbol{u} \mid \boldsymbol{y}) = \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u})\, \mathrm{p}(\boldsymbol{u} \mid \boldsymbol{y})$$

- We will approximate this situation with

$$\mathrm{q}(\boldsymbol{f}^*) = \int \mathrm{p}(\boldsymbol{f}^* \mid \boldsymbol{u})\, \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u})\, \mathrm{q}(\boldsymbol{u})\, \mathrm{d}\boldsymbol{f}\, \mathrm{d}\boldsymbol{u} = \int \mathrm{p}(\boldsymbol{f}^* \mid \boldsymbol{u})\, \mathrm{q}(\boldsymbol{u})\, \mathrm{d}\boldsymbol{u}$$

SGPR variational bound

$$
\begin{aligned}
\mathcal{L}^{\mathsf{SGPR}}(\boldsymbol{\theta}, Z, \mathrm{q}(\boldsymbol{u})) &= -\log \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{\theta}, Z, \boldsymbol{u}) \\
&= -\log \int \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{f})\, \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u})\, \mathrm{p}(\boldsymbol{u})\, \mathrm{d}\boldsymbol{f}\, \mathrm{d}\boldsymbol{u} \\
&= -\log \int \mathrm{q}(\boldsymbol{f}, \boldsymbol{u}) \frac{\mathrm{p}(\boldsymbol{y} \mid \boldsymbol{f})\, \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u})\, \mathrm{p}(\boldsymbol{u})}{\mathrm{q}(\boldsymbol{f}, \boldsymbol{u})}\, \mathrm{d}\boldsymbol{f}\, \mathrm{d}\boldsymbol{u} \\
&\geq -\int \mathrm{q}(\boldsymbol{f}, \boldsymbol{u}) \log \frac{\mathrm{p}(\boldsymbol{y} \mid \boldsymbol{f})\, \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u})\, \mathrm{p}(\boldsymbol{u})}{\mathrm{q}(\boldsymbol{f}, \boldsymbol{u})}\, \mathrm{d}\boldsymbol{f}\, \mathrm{d}\boldsymbol{u} \\
&= -\int \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u})\, \mathrm{q}(\boldsymbol{u}) \log \frac{\mathrm{p}(\boldsymbol{y} \mid \boldsymbol{f})\, \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u})\, \mathrm{p}(\boldsymbol{u})}{\mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u})\, \mathrm{q}(\boldsymbol{u})}\, \mathrm{d}\boldsymbol{f}\, \mathrm{d}\boldsymbol{u} \\
&= -\int \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u})\, \mathrm{q}(\boldsymbol{u}) \log \frac{\mathrm{p}(\boldsymbol{y} \mid \boldsymbol{f})\, \mathrm{p}(\boldsymbol{u})}{\mathrm{q}(\boldsymbol{u})}\, \mathrm{d}\boldsymbol{f}\, \mathrm{d}\boldsymbol{u} \\
&= -\int \mathrm{q}(\boldsymbol{u}) \left(\int \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u}) \log \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{f})\, \mathrm{d}\boldsymbol{f} + \log \frac{\mathrm{p}(\boldsymbol{u})}{\mathrm{q}(\boldsymbol{u})} \right) \mathrm{d}\boldsymbol{u}
\end{aligned}
$$

SGPR variational bound

$$
\begin{aligned}
\mathcal{L}^{\mathsf{SGPR}}(\boldsymbol{\theta}, Z, \mathrm{q}(\boldsymbol{u})) &= -\log \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{\theta}, Z, \boldsymbol{u}) \\
&= -\log \int \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{f})\, \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u})\, \mathrm{p}(\boldsymbol{u})\, \mathrm{d}\boldsymbol{f}\, \mathrm{d}\boldsymbol{u} \\
&= -\log \int \mathrm{q}(\boldsymbol{f}, \boldsymbol{u}) \frac{\mathrm{p}(\boldsymbol{y} \mid \boldsymbol{f})\, \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u})\, \mathrm{p}(\boldsymbol{u})}{\mathrm{q}(\boldsymbol{f}, \boldsymbol{u})}\, \mathrm{d}\boldsymbol{f}\, \mathrm{d}\boldsymbol{u} \\
&\geq -\int \mathrm{q}(\boldsymbol{f}, \boldsymbol{u}) \log \frac{\mathrm{p}(\boldsymbol{y} \mid \boldsymbol{f})\, \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u})\, \mathrm{p}(\boldsymbol{u})}{\mathrm{q}(\boldsymbol{f}, \boldsymbol{u})}\, \mathrm{d}\boldsymbol{f}\, \mathrm{d}\boldsymbol{u} \\
&= -\int \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u})\, \mathrm{q}(\boldsymbol{u}) \log \frac{\mathrm{p}(\boldsymbol{y} \mid \boldsymbol{f})\, \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u})\, \mathrm{p}(\boldsymbol{u})}{\mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u})\, \mathrm{q}(\boldsymbol{u})}\, \mathrm{d}\boldsymbol{f}\, \mathrm{d}\boldsymbol{u} \\
&= -\int \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u})\, \mathrm{q}(\boldsymbol{u}) \log \frac{\mathrm{p}(\boldsymbol{y} \mid \boldsymbol{f})\, \mathrm{p}(\boldsymbol{u})}{\mathrm{q}(\boldsymbol{u})}\, \mathrm{d}\boldsymbol{f}\, \mathrm{d}\boldsymbol{u} \\
&= -\int \mathrm{q}(\boldsymbol{u}) \left(\int \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u}) \log \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{f})\, \mathrm{d}\boldsymbol{f} + \log \frac{\mathrm{p}(\boldsymbol{u})}{\mathrm{q}(\boldsymbol{u})} \right) \mathrm{d}\boldsymbol{u}
\end{aligned}
$$

SGPR variational bound

$$
\begin{aligned}
\mathcal{L}^{\mathsf{SGPR}}(\boldsymbol{\theta}, Z, \mathrm{q}(\boldsymbol{u})) &= -\log \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{\theta}, Z, \boldsymbol{u}) \\
&= -\log \int \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{f})\, \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u})\, \mathrm{p}(\boldsymbol{u})\, \mathrm{d}\boldsymbol{f}\, \mathrm{d}\boldsymbol{u} \\
&= -\log \int \mathrm{q}(\boldsymbol{f}, \boldsymbol{u}) \frac{\mathrm{p}(\boldsymbol{y} \mid \boldsymbol{f})\, \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u})\, \mathrm{p}(\boldsymbol{u})}{\mathrm{q}(\boldsymbol{f}, \boldsymbol{u})}\, \mathrm{d}\boldsymbol{f}\, \mathrm{d}\boldsymbol{u} \\
&\geq -\int \mathrm{q}(\boldsymbol{f}, \boldsymbol{u}) \log \frac{\mathrm{p}(\boldsymbol{y} \mid \boldsymbol{f})\, \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u})\, \mathrm{p}(\boldsymbol{u})}{\mathrm{q}(\boldsymbol{f}, \boldsymbol{u})}\, \mathrm{d}\boldsymbol{f}\, \mathrm{d}\boldsymbol{u} \\
&= -\int \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u})\, \mathrm{q}(\boldsymbol{u}) \log \frac{\mathrm{p}(\boldsymbol{y} \mid \boldsymbol{f})\, \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u})\, \mathrm{p}(\boldsymbol{u})}{\mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u})\, \mathrm{q}(\boldsymbol{u})}\, \mathrm{d}\boldsymbol{f}\, \mathrm{d}\boldsymbol{u} \\
&= -\int \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u})\, \mathrm{q}(\boldsymbol{u}) \log \frac{\mathrm{p}(\boldsymbol{y} \mid \boldsymbol{f})\, \mathrm{p}(\boldsymbol{u})}{\mathrm{q}(\boldsymbol{u})}\, \mathrm{d}\boldsymbol{f}\, \mathrm{d}\boldsymbol{u} \\
&= -\int \mathrm{q}(\boldsymbol{u}) \left(\int \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u}) \log \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{f})\, \mathrm{d}\boldsymbol{f} + \log \frac{\mathrm{p}(\boldsymbol{u})}{\mathrm{q}(\boldsymbol{u})} \right) \mathrm{d}\boldsymbol{u}
\end{aligned}
$$

SGPR variational bound

$$
\begin{aligned}
\mathcal{L}^{\mathrm{SGPR}}(\boldsymbol{\theta}, Z, \mathrm{q}(\boldsymbol{u})) &= -\log \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{\theta}, Z, \boldsymbol{u}) \\
&= -\log \int \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{f})\, \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u})\, \mathrm{p}(\boldsymbol{u})\, \mathrm{d}\boldsymbol{f}\, \mathrm{d}\boldsymbol{u} \\
&= -\log \int \mathrm{q}(\boldsymbol{f}, \boldsymbol{u}) \frac{\mathrm{p}(\boldsymbol{y} \mid \boldsymbol{f})\, \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u})\, \mathrm{p}(\boldsymbol{u})}{\mathrm{q}(\boldsymbol{f}, \boldsymbol{u})}\, \mathrm{d}\boldsymbol{f}\, \mathrm{d}\boldsymbol{u} \\
&\geq -\int \mathrm{q}(\boldsymbol{f}, \boldsymbol{u}) \log \frac{\mathrm{p}(\boldsymbol{y} \mid \boldsymbol{f})\, \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u})\, \mathrm{p}(\boldsymbol{u})}{\mathrm{q}(\boldsymbol{f}, \boldsymbol{u})}\, \mathrm{d}\boldsymbol{f}\, \mathrm{d}\boldsymbol{u} \\
&= -\int \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u})\, \mathrm{q}(\boldsymbol{u}) \log \frac{\mathrm{p}(\boldsymbol{y} \mid \boldsymbol{f})\, \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u})\, \mathrm{p}(\boldsymbol{u})}{\mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u})\, \mathrm{q}(\boldsymbol{u})}\, \mathrm{d}\boldsymbol{f}\, \mathrm{d}\boldsymbol{u} \\
&= -\int \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u})\, \mathrm{q}(\boldsymbol{u}) \log \frac{\mathrm{p}(\boldsymbol{y} \mid \boldsymbol{f})\, \mathrm{p}(\boldsymbol{u})}{\mathrm{q}(\boldsymbol{u})}\, \mathrm{d}\boldsymbol{f}\, \mathrm{d}\boldsymbol{u} \\
&= -\int \mathrm{q}(\boldsymbol{u}) \left(\int \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u}) \log \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{f})\, \mathrm{d}\boldsymbol{f} + \log \frac{\mathrm{p}(\boldsymbol{u})}{\mathrm{q}(\boldsymbol{u})} \right) \mathrm{d}\boldsymbol{u}
\end{aligned}
$$

SGPR variational bound

$$
\begin{aligned}
\mathcal{L}^{\mathrm{SGPR}}(\boldsymbol{\theta}, Z, \mathrm{q}(\boldsymbol{u})) &= -\log \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{\theta}, Z, \boldsymbol{u}) \\
&= -\log \int \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{f})\, \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u})\, \mathrm{p}(\boldsymbol{u})\, \mathrm{d}\boldsymbol{f}\, \mathrm{d}\boldsymbol{u} \\
&= -\log \int \mathrm{q}(\boldsymbol{f}, \boldsymbol{u}) \frac{\mathrm{p}(\boldsymbol{y} \mid \boldsymbol{f})\, \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u})\, \mathrm{p}(\boldsymbol{u})}{\mathrm{q}(\boldsymbol{f}, \boldsymbol{u})}\, \mathrm{d}\boldsymbol{f}\, \mathrm{d}\boldsymbol{u} \\
&\geq -\int \mathrm{q}(\boldsymbol{f}, \boldsymbol{u}) \log \frac{\mathrm{p}(\boldsymbol{y} \mid \boldsymbol{f})\, \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u})\, \mathrm{p}(\boldsymbol{u})}{\mathrm{q}(\boldsymbol{f}, \boldsymbol{u})}\, \mathrm{d}\boldsymbol{f}\, \mathrm{d}\boldsymbol{u} \\
&= -\int \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u})\, \mathrm{q}(\boldsymbol{u}) \log \frac{\mathrm{p}(\boldsymbol{y} \mid \boldsymbol{f})\, \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u})\, \mathrm{p}(\boldsymbol{u})}{\mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u})\, \mathrm{q}(\boldsymbol{u})}\, \mathrm{d}\boldsymbol{f}\, \mathrm{d}\boldsymbol{u} \\
&= -\int \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u})\, \mathrm{q}(\boldsymbol{u}) \log \frac{\mathrm{p}(\boldsymbol{y} \mid \boldsymbol{f})\, \mathrm{p}(\boldsymbol{u})}{\mathrm{q}(\boldsymbol{u})}\, \mathrm{d}\boldsymbol{f}\, \mathrm{d}\boldsymbol{u} \\
&= -\int \mathrm{q}(\boldsymbol{u}) \left(\int \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u}) \log \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{f})\, \mathrm{d}\boldsymbol{f} + \log \frac{\mathrm{p}(\boldsymbol{u})}{\mathrm{q}(\boldsymbol{u})} \right) \mathrm{d}\boldsymbol{u}
\end{aligned}
$$

SGPR variational bound

$$
\begin{aligned}
\mathcal{L}^{\mathsf{SGPR}}(\boldsymbol{\theta}, Z, \mathrm{q}(\boldsymbol{u})) &= -\log \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{\theta}, Z, \boldsymbol{u}) \\
&= -\log \int \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{f})\, \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u})\, \mathrm{p}(\boldsymbol{u})\, \mathrm{d}\boldsymbol{f}\, \mathrm{d}\boldsymbol{u} \\
&= -\log \int \mathrm{q}(\boldsymbol{f}, \boldsymbol{u}) \frac{\mathrm{p}(\boldsymbol{y} \mid \boldsymbol{f})\, \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u})\, \mathrm{p}(\boldsymbol{u})}{\mathrm{q}(\boldsymbol{f}, \boldsymbol{u})}\, \mathrm{d}\boldsymbol{f}\, \mathrm{d}\boldsymbol{u} \\
&\geq -\int \mathrm{q}(\boldsymbol{f}, \boldsymbol{u}) \log \frac{\mathrm{p}(\boldsymbol{y} \mid \boldsymbol{f})\, \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u})\, \mathrm{p}(\boldsymbol{u})}{\mathrm{q}(\boldsymbol{f}, \boldsymbol{u})}\, \mathrm{d}\boldsymbol{f}\, \mathrm{d}\boldsymbol{u} \\
&= -\int \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u})\, \mathrm{q}(\boldsymbol{u}) \log \frac{\mathrm{p}(\boldsymbol{y} \mid \boldsymbol{f})\, \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u})\, \mathrm{p}(\boldsymbol{u})}{\mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u})\, \mathrm{q}(\boldsymbol{u})}\, \mathrm{d}\boldsymbol{f}\, \mathrm{d}\boldsymbol{u} \\
&= -\int \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u})\, \mathrm{q}(\boldsymbol{u}) \log \frac{\mathrm{p}(\boldsymbol{y} \mid \boldsymbol{f})\, \mathrm{p}(\boldsymbol{u})}{\mathrm{q}(\boldsymbol{u})}\, \mathrm{d}\boldsymbol{f}\, \mathrm{d}\boldsymbol{u} \\
&= -\int \mathrm{q}(\boldsymbol{u}) \left(\int \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u}) \log \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{f})\, \mathrm{d}\boldsymbol{f} + \log \frac{\mathrm{p}(\boldsymbol{u})}{\mathrm{q}(\boldsymbol{u})} \right) \mathrm{d}\boldsymbol{u}
\end{aligned}
$$

SGPR variational bound

$$
\begin{aligned}
\mathbb{E}_{\mathrm{p}(\boldsymbol{f}|\boldsymbol{u})}[\log \mathrm{p}(\boldsymbol{y}\,|\,\boldsymbol{f})] &= \int \mathrm{p}(\boldsymbol{f}\,|\,\boldsymbol{u}) \log \mathrm{p}(\boldsymbol{y}\,|\,\boldsymbol{f})\,\mathrm{d}\boldsymbol{f} \\
&= \int \log \mathcal{N}\left(\boldsymbol{y} \,\middle|\, \boldsymbol{f}, \sigma_n^2 \mathbf{I}\right) \mathcal{N}\left(\boldsymbol{f} \,\middle|\, \boldsymbol{K}_{\boldsymbol{f}\boldsymbol{u}} \boldsymbol{K}_{\boldsymbol{u}\boldsymbol{u}}^{-1} \boldsymbol{u}, \boldsymbol{K}_{\boldsymbol{f}\boldsymbol{f}} - \boldsymbol{Q}_{\boldsymbol{f}\boldsymbol{f}}\right) \mathrm{d}\boldsymbol{f} \\
&= \log \mathcal{N}\left(\boldsymbol{y} \,\middle|\, \boldsymbol{K}_{\boldsymbol{f}\boldsymbol{u}} \boldsymbol{K}_{\boldsymbol{u}\boldsymbol{u}}^{-1} \boldsymbol{u}, \sigma_n^2 \mathbf{I}\right) - \frac{1}{2\sigma_n^2} \operatorname{tr}\left(\boldsymbol{K}_{\boldsymbol{f}\boldsymbol{f}} - \boldsymbol{Q}_{\boldsymbol{f}\boldsymbol{f}}\right) \\
&= \log G(\boldsymbol{u}) - \frac{1}{2\sigma_n^2} \operatorname{tr}\left(\boldsymbol{K}_{\boldsymbol{f}\boldsymbol{f}} - \boldsymbol{Q}_{\boldsymbol{f}\boldsymbol{f}}\right)
\end{aligned}
$$

SGPR variational bound

$$
\begin{aligned}
\mathbb{E}_{\mathrm{p}(\boldsymbol{f}|\boldsymbol{u})}[\log \mathrm{p}(\boldsymbol{y}\,|\,\boldsymbol{f})] &= \int \mathrm{p}(\boldsymbol{f}\,|\,\boldsymbol{u}) \log \mathrm{p}(\boldsymbol{y}\,|\,\boldsymbol{f})\,\mathrm{d}\boldsymbol{f} \\
&= \int \log \mathcal{N}\left(\boldsymbol{y} \,\middle|\, \boldsymbol{f}, \sigma_n^2 \mathbf{I}\right) \mathcal{N}\left(\boldsymbol{f} \,\middle|\, \boldsymbol{K}_{\boldsymbol{f}\boldsymbol{u}} \boldsymbol{K}_{\boldsymbol{u}\boldsymbol{u}}^{-1} \boldsymbol{u}, \boldsymbol{K}_{\boldsymbol{f}\boldsymbol{f}} - \boldsymbol{Q}_{\boldsymbol{f}\boldsymbol{f}}\right) \mathrm{d}\boldsymbol{f} \\
&= \log \mathcal{N}\left(\boldsymbol{y} \,\middle|\, \boldsymbol{K}_{\boldsymbol{f}\boldsymbol{u}} \boldsymbol{K}_{\boldsymbol{u}\boldsymbol{u}}^{-1} \boldsymbol{u}, \sigma_n^2 \mathbf{I}\right) - \frac{1}{2\sigma_n^2} \operatorname{tr}\left(\boldsymbol{K}_{\boldsymbol{f}\boldsymbol{f}} - \boldsymbol{Q}_{\boldsymbol{f}\boldsymbol{f}}\right) \\
&= \log G(\boldsymbol{u}) - \frac{1}{2\sigma_n^2} \operatorname{tr}\left(\boldsymbol{K}_{\boldsymbol{f}\boldsymbol{f}} - \boldsymbol{Q}_{\boldsymbol{f}\boldsymbol{f}}\right)
\end{aligned}
$$

SGPR variational bound

$$
\begin{aligned}
\mathbb{E}_{\mathrm{p}(\boldsymbol{f}|\boldsymbol{u})}[\log \mathrm{p}(\boldsymbol{y}\,|\,\boldsymbol{f})] &= \int \mathrm{p}(\boldsymbol{f}\mid\boldsymbol{u}) \log \mathrm{p}(\boldsymbol{y}\,|\,\boldsymbol{f})\,\mathrm{d}\boldsymbol{f} \\
&= \int \log \mathcal{N}\left(\boldsymbol{y}\,\middle|\,\boldsymbol{f}, \sigma_n^2\mathbf{I}\right) \mathcal{N}\left(\boldsymbol{f}\,\middle|\,\boldsymbol{K}_{\boldsymbol{fu}}\boldsymbol{K}_{\boldsymbol{uu}}^{-1}\boldsymbol{u}, \boldsymbol{K}_{\boldsymbol{ff}} - \boldsymbol{Q}_{\boldsymbol{ff}}\right) \mathrm{d}\boldsymbol{f} \\
&= \log \mathcal{N}\left(\boldsymbol{y}\,\middle|\,\boldsymbol{K}_{\boldsymbol{fu}}\boldsymbol{K}_{\boldsymbol{uu}}^{-1}\boldsymbol{u}, \sigma_n^2\mathbf{I}\right) - \frac{1}{2\sigma_n^2}\operatorname{tr}\left(\boldsymbol{K}_{\boldsymbol{ff}} - \boldsymbol{Q}_{\boldsymbol{ff}}\right) \\
&= \log G(\boldsymbol{u}) - \frac{1}{2\sigma_n^2}\operatorname{tr}\left(\boldsymbol{K}_{\boldsymbol{ff}} - \boldsymbol{Q}_{\boldsymbol{ff}}\right)
\end{aligned}
$$

SGPR variational bound

$$
\begin{aligned}
\mathbb{E}_{\mathrm{p}(\boldsymbol{f}|\boldsymbol{u})}[\log \mathrm{p}(\boldsymbol{y}\,|\,\boldsymbol{f})] &= \int \mathrm{p}(\boldsymbol{f}\mid\boldsymbol{u}) \log \mathrm{p}(\boldsymbol{y}\,|\,\boldsymbol{f})\,\mathrm{d}\boldsymbol{f} \\
&= \int \log \mathcal{N}\left(\boldsymbol{y}\,\middle|\,\boldsymbol{f}, \sigma_n^2\mathbf{I}\right) \mathcal{N}\left(\boldsymbol{f}\,\middle|\,\boldsymbol{K}_{\boldsymbol{fu}}\boldsymbol{K}_{\boldsymbol{uu}}^{-1}\boldsymbol{u}, \boldsymbol{K}_{\boldsymbol{ff}} - \boldsymbol{Q}_{\boldsymbol{ff}}\right) \mathrm{d}\boldsymbol{f} \\
&= \log \mathcal{N}\left(\boldsymbol{y}\,\middle|\,\boldsymbol{K}_{\boldsymbol{fu}}\boldsymbol{K}_{\boldsymbol{uu}}^{-1}\boldsymbol{u}, \sigma_n^2\mathbf{I}\right) - \frac{1}{2\sigma_n^2}\operatorname{tr}\left(\boldsymbol{K}_{\boldsymbol{ff}} - \boldsymbol{Q}_{\boldsymbol{ff}}\right) \\
&= \log G(\boldsymbol{u}) - \frac{1}{2\sigma_n^2}\operatorname{tr}\left(\boldsymbol{K}_{\boldsymbol{ff}} - \boldsymbol{Q}_{\boldsymbol{ff}}\right)
\end{aligned}
$$

SGPR variational bound

$$
\begin{aligned}
\mathcal{L}^{\mathsf{SGPR}}(\boldsymbol{\theta}, \boldsymbol{Z}, \mathrm{q}(\boldsymbol{u})) &= -\int \mathrm{q}(\boldsymbol{u}) \left(\int \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u}) \log \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{f}) \, \mathrm{d}\boldsymbol{f} + \log \frac{\mathrm{p}(\boldsymbol{u})}{\mathrm{q}(\boldsymbol{u})} \right) \mathrm{d}\boldsymbol{u} \\
&= -\int \mathrm{q}(\boldsymbol{u}) \left(\log G(\boldsymbol{u}) - \frac{1}{2\sigma_n^2} \operatorname{tr}\left(\boldsymbol{K}_{\boldsymbol{f}\boldsymbol{f}} - \boldsymbol{Q}_{\boldsymbol{f}\boldsymbol{f}} \right) + \log \frac{\mathrm{p}(\boldsymbol{u})}{\mathrm{q}(\boldsymbol{u})} \right) \mathrm{d}\boldsymbol{u} \\
&= -\int \mathrm{q}(\boldsymbol{u}) \log \frac{G(\boldsymbol{u}) \, \mathrm{p}(\boldsymbol{u})}{\mathrm{q}(\boldsymbol{u})} \, \mathrm{d}\boldsymbol{u} + \frac{1}{2\sigma_n^2} \operatorname{tr}\left(\boldsymbol{K}_{\boldsymbol{f}\boldsymbol{f}} - \boldsymbol{Q}_{\boldsymbol{f}\boldsymbol{f}} \right)
\end{aligned}
$$

- Phew!
- We have not specified $\mathrm{q}(\boldsymbol{u})$. How should we choose it?

SGPR variational bound

$$
\begin{aligned}
\mathcal{L}^{\mathsf{SGPR}}(\boldsymbol{\theta}, \boldsymbol{Z}, \mathrm{q}(\boldsymbol{u})) &= -\int \mathrm{q}(\boldsymbol{u}) \left(\int \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u}) \log \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{f}) \, \mathrm{d}\boldsymbol{f} + \log \frac{\mathrm{p}(\boldsymbol{u})}{\mathrm{q}(\boldsymbol{u})} \right) \mathrm{d}\boldsymbol{u} \\
&= -\int \mathrm{q}(\boldsymbol{u}) \left(\log G(\boldsymbol{u}) - \frac{1}{2\sigma_n^2} \operatorname{tr}\left(\boldsymbol{K}_{ff} - \boldsymbol{Q}_{ff} \right) + \log \frac{\mathrm{p}(\boldsymbol{u})}{\mathrm{q}(\boldsymbol{u})} \right) \mathrm{d}\boldsymbol{u} \\
&= -\int \mathrm{q}(\boldsymbol{u}) \log \frac{G(\boldsymbol{u}) \, \mathrm{p}(\boldsymbol{u})}{\mathrm{q}(\boldsymbol{u})} \, \mathrm{d}\boldsymbol{u} + \frac{1}{2\sigma_n^2} \operatorname{tr}\left(\boldsymbol{K}_{ff} - \boldsymbol{Q}_{ff} \right)
\end{aligned}
$$

- Phew!
- We have not specified $\mathrm{q}(\boldsymbol{u})$. How should we choose it?

SGPR variational bound

$$
\begin{aligned}
\mathcal{L}^{\mathsf{SGPR}}(\boldsymbol{\theta}, \boldsymbol{Z}, \mathrm{q}(\boldsymbol{u})) &= -\int \mathrm{q}(\boldsymbol{u}) \left(\int \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u}) \log \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{f}) \, \mathrm{d}\boldsymbol{f} + \log \frac{\mathrm{p}(\boldsymbol{u})}{\mathrm{q}(\boldsymbol{u})} \right) \mathrm{d}\boldsymbol{u} \\
&= -\int \mathrm{q}(\boldsymbol{u}) \left(\log G(\boldsymbol{u}) - \frac{1}{2\sigma_n^2} \operatorname{tr}\left(\boldsymbol{K}_{ff} - \boldsymbol{Q}_{ff} \right) + \log \frac{\mathrm{p}(\boldsymbol{u})}{\mathrm{q}(\boldsymbol{u})} \right) \mathrm{d}\boldsymbol{u} \\
&= -\int \mathrm{q}(\boldsymbol{u}) \log \frac{G(\boldsymbol{u}) \, \mathrm{p}(\boldsymbol{u})}{\mathrm{q}(\boldsymbol{u})} \, \mathrm{d}\boldsymbol{u} + \frac{1}{2\sigma_n^2} \operatorname{tr}\left(\boldsymbol{K}_{ff} - \boldsymbol{Q}_{ff} \right)
\end{aligned}
$$

- Phew!
- We have not specified $\mathrm{q}(\boldsymbol{u})$. How should we choose it?

SGPR variational bound

$$
\begin{aligned}
\mathcal{L}^{\mathrm{SGPR}}(\boldsymbol{\theta}, \boldsymbol{Z}, \mathrm{q}(\boldsymbol{u})) &= -\int \mathrm{q}(\boldsymbol{u}) \left(\int \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u}) \log \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{f}) \, \mathrm{d}\boldsymbol{f} + \log \frac{\mathrm{p}(\boldsymbol{u})}{\mathrm{q}(\boldsymbol{u})} \right) \mathrm{d}\boldsymbol{u} \\
&= -\int \mathrm{q}(\boldsymbol{u}) \left(\log G(\boldsymbol{u}) - \frac{1}{2\sigma_n^2} \operatorname{tr}\left(\boldsymbol{K}_{ff} - \boldsymbol{Q}_{ff} \right) + \log \frac{\mathrm{p}(\boldsymbol{u})}{\mathrm{q}(\boldsymbol{u})} \right) \mathrm{d}\boldsymbol{u} \\
&= -\int \mathrm{q}(\boldsymbol{u}) \log \frac{G(\boldsymbol{u}) \, \mathrm{p}(\boldsymbol{u})}{\mathrm{q}(\boldsymbol{u})} \, \mathrm{d}\boldsymbol{u} + \frac{1}{2\sigma_n^2} \operatorname{tr}\left(\boldsymbol{K}_{ff} - \boldsymbol{Q}_{ff} \right)
\end{aligned}
$$

- Phew!
- We have not specified $\mathrm{q}(\boldsymbol{u})$. How should we choose it?

The optimal $\mathrm{q}(\boldsymbol{u})$

$$
\begin{aligned}
\min_{\mathrm{q}(\boldsymbol{u})} \mathcal{L}^{\mathsf{SGPR}}(\boldsymbol{\theta}, Z, \mathrm{q}(\boldsymbol{u})) &= \min_{\mathrm{q}(\boldsymbol{u})} - \int \mathrm{q}(\boldsymbol{u}) \log \frac{G(\boldsymbol{u})\,\mathrm{p}(\boldsymbol{u})}{\mathrm{q}(\boldsymbol{u})}\,\mathrm{d}\boldsymbol{u} + \frac{1}{2\sigma_n^2}\operatorname{tr}\left(\boldsymbol{K}_{ff} - \boldsymbol{Q}_{ff}\right) \\
&= \min_{\mathrm{q}(\boldsymbol{u})} - \int \mathrm{q}(\boldsymbol{u}) \log \frac{G(\boldsymbol{u})\,\mathrm{p}(\boldsymbol{u})}{\mathrm{q}(\boldsymbol{u})}\,\mathrm{d}\boldsymbol{u} \\
&= \min_{\mathrm{q}(\boldsymbol{u})} \mathrm{KL}(\mathrm{q}(\boldsymbol{u}) \,\|\, G(\boldsymbol{u})\,\mathrm{p}(\boldsymbol{u}))
\end{aligned}
$$

The optimal $\mathrm{q}(\boldsymbol{u})$

$$
\begin{aligned}
\min_{\mathrm{q}(\boldsymbol{u})} \mathcal{L}^{\mathsf{SGPR}}(\boldsymbol{\theta}, Z, \mathrm{q}(\boldsymbol{u})) &= \min_{\mathrm{q}(\boldsymbol{u})} - \int \mathrm{q}(\boldsymbol{u}) \log \frac{G(\boldsymbol{u})\,\mathrm{p}(\boldsymbol{u})}{\mathrm{q}(\boldsymbol{u})}\,\mathrm{d}\boldsymbol{u} + \frac{1}{2\sigma_n^2}\operatorname{tr}\left(\boldsymbol{K}_{ff} - \boldsymbol{Q}_{ff}\right) \\
&= \min_{\mathrm{q}(\boldsymbol{u})} - \int \mathrm{q}(\boldsymbol{u}) \log \frac{G(\boldsymbol{u})\,\mathrm{p}(\boldsymbol{u})}{\mathrm{q}(\boldsymbol{u})}\,\mathrm{d}\boldsymbol{u} \\
&= \min_{\mathrm{q}(\boldsymbol{u})} \mathrm{KL}(\mathrm{q}(\boldsymbol{u}) \,\|\, G(\boldsymbol{u})\,\mathrm{p}(\boldsymbol{u}))
\end{aligned}
$$

- Minimize the divergence by choosing

$$
\mathrm{q}^*(\boldsymbol{u}) \propto G(\boldsymbol{u})\,\mathrm{p}(\boldsymbol{u}) = \mathcal{N}\left(\boldsymbol{y} \,\middle|\, \boldsymbol{K}_{fu}\boldsymbol{K}_{uu}^{-1}\boldsymbol{u}, \sigma_n^2\mathrm{I}\right)\mathcal{N}(\boldsymbol{u} \mid \boldsymbol{0}, \boldsymbol{K}_{uu})
$$

The optimal $\mathrm{q}(\boldsymbol{u})$

$$
\begin{aligned}
\min_{\mathrm{q}(\boldsymbol{u})} \mathcal{L}^{\mathrm{SGPR}}(\boldsymbol{\theta}, Z, \mathrm{q}(\boldsymbol{u})) &= \min_{\mathrm{q}(\boldsymbol{u})} - \int \mathrm{q}(\boldsymbol{u}) \log \frac{G(\boldsymbol{u})\,\mathrm{p}(\boldsymbol{u})}{\mathrm{q}(\boldsymbol{u})}\,\mathrm{d}\boldsymbol{u} + \frac{1}{2\sigma_n^2}\operatorname{tr}\left(\boldsymbol{K}_{ff} - \boldsymbol{Q}_{ff}\right) \\
&= \min_{\mathrm{q}(\boldsymbol{u})} - \int \mathrm{q}(\boldsymbol{u}) \log \frac{G(\boldsymbol{u})\,\mathrm{p}(\boldsymbol{u})}{\mathrm{q}(\boldsymbol{u})}\,\mathrm{d}\boldsymbol{u} \\
&= \min_{\mathrm{q}(\boldsymbol{u})} \mathrm{KL}(\mathrm{q}(\boldsymbol{u}) \,\|\, G(\boldsymbol{u})\,\mathrm{p}(\boldsymbol{u}))
\end{aligned}
$$

- Minimize the divergence by choosing

$$
\mathrm{q}^*(\boldsymbol{u}) \propto G(\boldsymbol{u})\,\mathrm{p}(\boldsymbol{u}) = \mathcal{N}\left(\boldsymbol{y} \,\middle|\, \boldsymbol{K}_{fu}\boldsymbol{K}_{uu}^{-1}\boldsymbol{u}, \sigma_n^2\mathrm{I}\right)\mathcal{N}(\boldsymbol{u} \mid \boldsymbol{0}, \boldsymbol{K}_{uu})
$$

- This is Gaussian again

$$
\mathrm{q}^*(\boldsymbol{u}) = \frac{G(\boldsymbol{u})\,\mathrm{p}(\boldsymbol{u})}{\int G(\boldsymbol{u})\,\mathrm{p}(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u}} = \mathcal{N}(\boldsymbol{u} \mid \boldsymbol{\mu}_u, \Sigma_u)
$$

$$
\boldsymbol{\mu}_u = \sigma_n^{-2}\boldsymbol{K}_{mm}\boldsymbol{B}^{-1}\boldsymbol{K}_{mn}\boldsymbol{y} \qquad \Sigma_u = \boldsymbol{K}_{mm}\boldsymbol{B}^{-1}\boldsymbol{K}_{mm} \qquad \boldsymbol{B} = \boldsymbol{K}_{mm} + \sigma_n^{-2}\boldsymbol{K}_{mn}\boldsymbol{K}_{nm}
$$

(Slight abuse of notation)

SGPR variational bound

- Resubstituting q^* into $\mathcal{L}^{\mathsf{SGPR}}$ yields the final bound

$$
\begin{aligned}
\mathcal{L}^{\mathsf{SGPR}}(\boldsymbol{\theta}, Z) &\geq \mathcal{L}^{\mathsf{SGPR}}(\boldsymbol{\theta}, Z, \mathrm{q}^*(\boldsymbol{u})) \\
&= -\int \mathrm{q}^*(\boldsymbol{u}) \log \frac{G(\boldsymbol{u})\,\mathrm{p}(\boldsymbol{u})}{\mathrm{q}^*(\boldsymbol{u})}\,\mathrm{d}\boldsymbol{u} + \frac{1}{2\sigma_n^2}\operatorname{tr}\left(\boldsymbol{K}_{ff} - \boldsymbol{Q}_{ff}\right) \\
&= -\int \frac{G(\boldsymbol{u})\,\mathrm{p}(\boldsymbol{u})}{\int G(\boldsymbol{u})\,\mathrm{p}(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u}} \log \frac{G(\boldsymbol{u})\,\mathrm{p}(\boldsymbol{u})}{\frac{G(\boldsymbol{u})\,\mathrm{p}(\boldsymbol{u})}{\int G(\boldsymbol{u})\,\mathrm{p}(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u}}}\,\mathrm{d}\boldsymbol{u} + \frac{1}{2\sigma_n^2}\operatorname{tr}\left(\boldsymbol{K}_{ff} - \boldsymbol{Q}_{ff}\right) \\
&= -\log \int G(\boldsymbol{u})\,\mathrm{p}(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u} + \frac{1}{2\sigma_n^2}\operatorname{tr}\left(\boldsymbol{K}_{ff} - \boldsymbol{Q}_{ff}\right) \\
&= -\log \mathcal{N}\left(\boldsymbol{y} \,\middle|\, \boldsymbol{0}, \boldsymbol{Q}_{ff} + \sigma_n^2 \mathbf{I}\right) + \frac{1}{2\sigma_n^2}\operatorname{tr}\left(\boldsymbol{K}_{ff} - \boldsymbol{Q}_{ff}\right)
\end{aligned}
$$

SGPR variational bound

- Resubstituting q^* into $\mathcal{L}^{\mathrm{SGPR}}$ yields the final bound

$$
\begin{aligned}
\mathcal{L}^{\mathrm{SGPR}}(\boldsymbol{\theta}, Z) &\geq \mathcal{L}^{\mathrm{SGPR}}(\boldsymbol{\theta}, Z, \mathrm{q}^*(\boldsymbol{u})) \\
&= -\int \mathrm{q}^*(\boldsymbol{u}) \log \frac{G(\boldsymbol{u})\,\mathrm{p}(\boldsymbol{u})}{\mathrm{q}^*(\boldsymbol{u})}\,\mathrm{d}\boldsymbol{u} + \frac{1}{2\sigma_n^2}\operatorname{tr}\left(\boldsymbol{K}_{ff} - \boldsymbol{Q}_{ff}\right) \\
&= -\int \frac{G(\boldsymbol{u})\,\mathrm{p}(\boldsymbol{u})}{\int G(\boldsymbol{u})\,\mathrm{p}(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u}} \log \frac{G(\boldsymbol{u})\,\mathrm{p}(\boldsymbol{u})}{\frac{G(\boldsymbol{u})\,\mathrm{p}(\boldsymbol{u})}{\int G(\boldsymbol{u})\,\mathrm{p}(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u}}}\,\mathrm{d}\boldsymbol{u} + \frac{1}{2\sigma_n^2}\operatorname{tr}\left(\boldsymbol{K}_{ff} - \boldsymbol{Q}_{ff}\right) \\
&= -\log \int G(\boldsymbol{u})\,\mathrm{p}(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u} + \frac{1}{2\sigma_n^2}\operatorname{tr}\left(\boldsymbol{K}_{ff} - \boldsymbol{Q}_{ff}\right) \\
&= -\log \mathcal{N}\left(\boldsymbol{y} \,\middle|\, \boldsymbol{0}, \boldsymbol{Q}_{ff} + \sigma_n^2 \mathbf{I}\right) + \frac{1}{2\sigma_n^2}\operatorname{tr}\left(\boldsymbol{K}_{ff} - \boldsymbol{Q}_{ff}\right)
\end{aligned}
$$

SGPR variational bound

- Resubstituting q^* into $\mathcal{L}^{\mathrm{SGPR}}$ yields the final bound

$$
\begin{aligned}
\mathcal{L}^{\mathrm{SGPR}}(\boldsymbol{\theta}, Z) &\geq \mathcal{L}^{\mathrm{SGPR}}(\boldsymbol{\theta}, Z, \mathrm{q}^*(\boldsymbol{u})) \\
&= -\int \mathrm{q}^*(\boldsymbol{u}) \log \frac{G(\boldsymbol{u})\,\mathrm{p}(\boldsymbol{u})}{\mathrm{q}^*(\boldsymbol{u})}\,\mathrm{d}\boldsymbol{u} + \frac{1}{2\sigma_n^2}\operatorname{tr}\left(\boldsymbol{K}_{ff} - \boldsymbol{Q}_{ff}\right) \\
&= -\int \frac{G(\boldsymbol{u})\,\mathrm{p}(\boldsymbol{u})}{\int G(\boldsymbol{u})\,\mathrm{p}(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u}} \log \frac{G(\boldsymbol{u})\,\mathrm{p}(\boldsymbol{u})}{\frac{G(\boldsymbol{u})\,\mathrm{p}(\boldsymbol{u})}{\int G(\boldsymbol{u})\,\mathrm{p}(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u}}}\,\mathrm{d}\boldsymbol{u} + \frac{1}{2\sigma_n^2}\operatorname{tr}\left(\boldsymbol{K}_{ff} - \boldsymbol{Q}_{ff}\right) \\
&= -\log \int G(\boldsymbol{u})\,\mathrm{p}(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u} + \frac{1}{2\sigma_n^2}\operatorname{tr}\left(\boldsymbol{K}_{ff} - \boldsymbol{Q}_{ff}\right) \\
&= -\log \mathcal{N}\left(\boldsymbol{y} \,\middle|\, \boldsymbol{0}, \boldsymbol{Q}_{ff} + \sigma_n^2 \mathbf{I}\right) + \frac{1}{2\sigma_n^2}\operatorname{tr}\left(\boldsymbol{K}_{ff} - \boldsymbol{Q}_{ff}\right)
\end{aligned}
$$

SGPR variational bound

- Resubstituting q^* into $\mathcal{L}^{\mathrm{SGPR}}$ yields the final bound

$$
\begin{aligned}
\mathcal{L}^{\mathrm{SGPR}}(\boldsymbol{\theta}, Z) &\geq \mathcal{L}^{\mathrm{SGPR}}(\boldsymbol{\theta}, Z, \mathrm{q}^*(\boldsymbol{u})) \\
&= -\int \mathrm{q}^*(\boldsymbol{u}) \log \frac{G(\boldsymbol{u})\,\mathrm{p}(\boldsymbol{u})}{\mathrm{q}^*(\boldsymbol{u})}\,\mathrm{d}\boldsymbol{u} + \frac{1}{2\sigma_n^2}\operatorname{tr}\left(\boldsymbol{K}_{ff} - \boldsymbol{Q}_{ff}\right) \\
&= -\int \frac{G(\boldsymbol{u})\,\mathrm{p}(\boldsymbol{u})}{\int G(\boldsymbol{u})\,\mathrm{p}(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u}} \log \frac{G(\boldsymbol{u})\,\mathrm{p}(\boldsymbol{u})}{\frac{G(\boldsymbol{u})\,\mathrm{p}(\boldsymbol{u})}{\int G(\boldsymbol{u})\,\mathrm{p}(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u}}}\,\mathrm{d}\boldsymbol{u} + \frac{1}{2\sigma_n^2}\operatorname{tr}\left(\boldsymbol{K}_{ff} - \boldsymbol{Q}_{ff}\right) \\
&= -\log \int G(\boldsymbol{u})\,\mathrm{p}(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u} + \frac{1}{2\sigma_n^2}\operatorname{tr}\left(\boldsymbol{K}_{ff} - \boldsymbol{Q}_{ff}\right) \\
&= -\log \mathcal{N}\left(\boldsymbol{y} \,\middle|\, \boldsymbol{0}, \boldsymbol{Q}_{ff} + \sigma_n^2 \mathbf{I}\right) + \frac{1}{2\sigma_n^2}\operatorname{tr}\left(\boldsymbol{K}_{ff} - \boldsymbol{Q}_{ff}\right)
\end{aligned}
$$

Interpretation of SGPR

$$\mathcal{L}^{\mathrm{GP}}(\boldsymbol{\theta}) = -\log \mathcal{N}\left(\boldsymbol{y} \,\middle|\, \mathbf{0}, \boldsymbol{K}_{ff} + \sigma_n^2 \mathbf{I}\right)$$

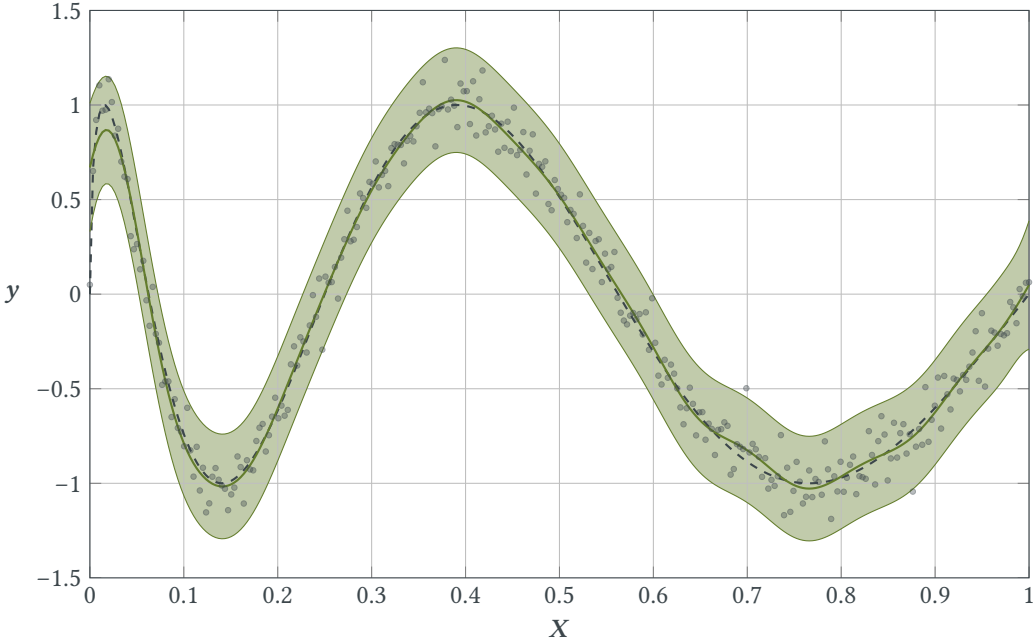
$$\mathcal{L}^{\mathrm{SPGP}}(\boldsymbol{\theta}) = -\log \mathcal{N}\left(\boldsymbol{y} \,\middle|\, \mathbf{0}, \boldsymbol{Q}_{ff} + \mathrm{diag}(\boldsymbol{K}_{ff} - \boldsymbol{Q}_{ff}) + \sigma_n^2 \mathbf{I}\right)$$

SGPR variational bound

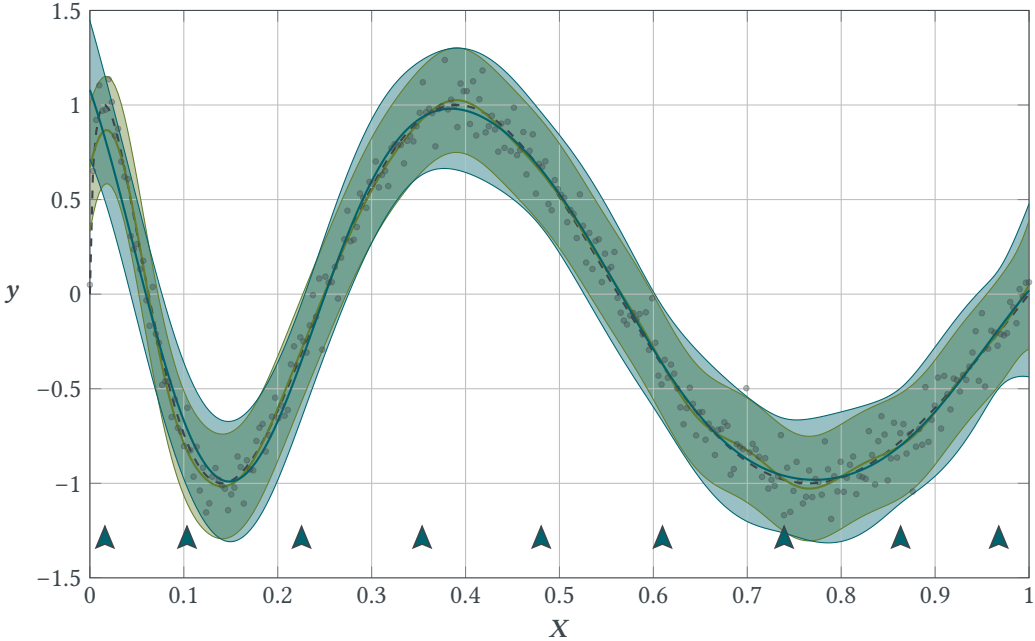
$$\mathcal{L}^{\mathrm{SGPR}}(\boldsymbol{\theta}, Z) \geq -\log \mathcal{N}\left(\boldsymbol{y} \,\middle|\, \mathbf{0}, \boldsymbol{Q}_{ff} + \sigma_n^2 \mathbf{I}\right) + \frac{1}{2\sigma_n^2} \mathrm{tr}\left(\boldsymbol{K}_{ff} - \boldsymbol{Q}_{ff}\right)$$

- This likelihood is $\mathcal{O}(NM^2)$, same with approximate predictions
- It does **not factorize** along the data, which makes stochastic optimization impossible
- The bound is tight if $\bar{\mathcal{D}} = \mathcal{D}$
- Choosing good $\bar{\mathcal{D}}$ does only make the bound tighter and cannot lead to overfitting

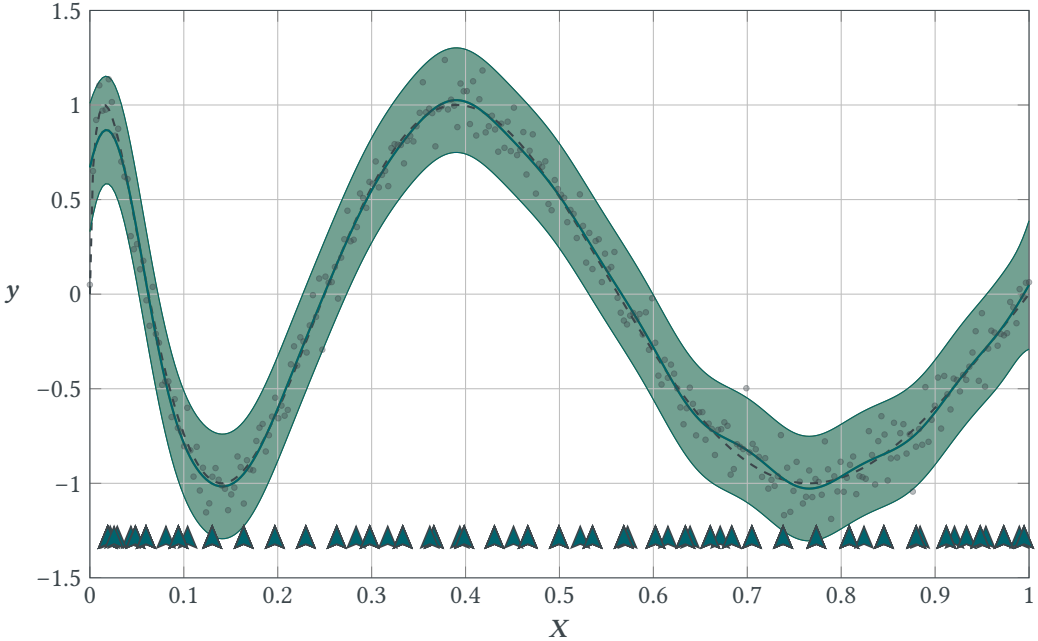
SGPR Showcase



SGPR Showcase



SGPR Showcase



SVGP: Stochastic Optimization

SVGP

Scalable Variational Gaussian Process (SVGP)

We want to variationally approximate the original Gaussian Process.

- Use the same strategy for the variational lower bound as SGPR
- Consider a custom variational scheme $\mathrm{q}(\boldsymbol{f}, \boldsymbol{u}) = \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u})\, \mathrm{q}(\boldsymbol{u})$
- Instead of choosing $\mathrm{q}^*(\boldsymbol{u})$, optimize $\mathrm{q}(\boldsymbol{u}) = \mathcal{N}(\boldsymbol{u} \mid \boldsymbol{m}, \boldsymbol{S})$

SVGP

Scalable Variational Gaussian Process (SVGP)

We want to variationally approximate the original Gaussian Process.

- Use the same strategy for the **variational lower bound** as SGPR
- Consider a **custom variational scheme** $\mathrm{q}(\boldsymbol{f}, \boldsymbol{u}) = \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{u})\, \mathrm{q}(\boldsymbol{u})$
- Instead of choosing $\mathrm{q}^*(\boldsymbol{u})$, optimize $\mathrm{q}(\boldsymbol{u}) = \mathcal{N}(\boldsymbol{u} \mid \boldsymbol{m}, \boldsymbol{S})$

- q^* is optimal but expensive to evaluate
- Instead we find a cheaper but worse approximation via optimization
- This hurts the bound, but does not alter the model
- $\boldsymbol{S}$ has **quadratically many parameters**

SVGP variational bound

$$\mathcal{L}^{\mathsf{SVGP}}(\boldsymbol{\theta}, Z, \boldsymbol{m}, S) \geq -\int \mathrm{q}(\boldsymbol{u}) \log \frac{G(\boldsymbol{u})\,\mathrm{p}(\boldsymbol{u})}{\mathrm{q}(\boldsymbol{u})}\,\mathrm{d}\boldsymbol{u} + \frac{1}{2\sigma_n^2}\,\mathrm{tr}\big(\boldsymbol{K}_{ff} - \boldsymbol{Q}_{ff}\big)$$

$$= -\int \mathrm{q}(\boldsymbol{u}) \log \frac{G(\boldsymbol{u})\,\mathrm{p}(\boldsymbol{u})}{\mathrm{q}(\boldsymbol{u})}\,\mathrm{d}\boldsymbol{u} + \frac{1}{2\sigma_n^2}\,\mathrm{tr}\big(\boldsymbol{K}_{ff} - \boldsymbol{Q}_{ff}\big)$$

$$= -\int \log G(\boldsymbol{u})\,\mathrm{q}(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u} + \mathrm{KL}(\mathrm{q}(\boldsymbol{u}) \parallel \mathrm{p}(\boldsymbol{u})) + \frac{1}{2\sigma_n^2}\,\mathrm{tr}\big(\boldsymbol{K}_{ff} - \boldsymbol{Q}_{ff}\big)$$

$$= -\,\mathbb{E}_{\mathrm{q}(\boldsymbol{u})}[\log G(\boldsymbol{u})] + \mathrm{KL}(\mathrm{q}(\boldsymbol{u}) \parallel \mathrm{p}(\boldsymbol{u})) + \frac{1}{2\sigma_n^2}\,\mathrm{tr}\big(\boldsymbol{K}_{ff} - \boldsymbol{Q}_{ff}\big)$$

SVGP variational bound

$$
\begin{aligned}
\mathcal{L}^{\mathsf{SVGP}}(\boldsymbol{\theta}, Z, \boldsymbol{m}, S) &\geq -\int \mathrm{q}(\boldsymbol{u}) \log \frac{G(\boldsymbol{u})\,\mathrm{p}(\boldsymbol{u})}{\mathrm{q}(\boldsymbol{u})}\,\mathrm{d}\boldsymbol{u} + \frac{1}{2\sigma_n^2}\,\mathrm{tr}\big(\boldsymbol{K}_{ff} - \boldsymbol{Q}_{ff}\big) \\
&= -\int \mathrm{q}(\boldsymbol{u}) \log \frac{G(\boldsymbol{u})\,\mathrm{p}(\boldsymbol{u})}{\mathrm{q}(\boldsymbol{u})}\,\mathrm{d}\boldsymbol{u} + \frac{1}{2\sigma_n^2}\,\mathrm{tr}\big(\boldsymbol{K}_{ff} - \boldsymbol{Q}_{ff}\big) \\
&= -\int \log G(\boldsymbol{u})\,\mathrm{q}(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u} + \mathrm{KL}(\mathrm{q}(\boldsymbol{u}) \parallel \mathrm{p}(\boldsymbol{u})) + \frac{1}{2\sigma_n^2}\,\mathrm{tr}\big(\boldsymbol{K}_{ff} - \boldsymbol{Q}_{ff}\big) \\
&= -\,\mathbb{E}_{\mathrm{q}(\boldsymbol{u})}[\log G(\boldsymbol{u})] + \mathrm{KL}(\mathrm{q}(\boldsymbol{u}) \parallel \mathrm{p}(\boldsymbol{u})) + \frac{1}{2\sigma_n^2}\,\mathrm{tr}\big(\boldsymbol{K}_{ff} - \boldsymbol{Q}_{ff}\big)
\end{aligned}
$$

SVGP variational bound

$$
\begin{aligned}
\mathcal{L}^{\mathsf{SVGP}}(\boldsymbol{\theta}, \boldsymbol{Z}, \boldsymbol{m}, \boldsymbol{S}) &\geq -\int \mathrm{q}(\boldsymbol{u}) \log \frac{G(\boldsymbol{u})\,\mathrm{p}(\boldsymbol{u})}{\mathrm{q}(\boldsymbol{u})}\,\mathrm{d}\boldsymbol{u} + \frac{1}{2\sigma_n^2}\operatorname{tr}\big(\boldsymbol{K}_{ff} - \boldsymbol{Q}_{ff}\big) \\
&= -\int \mathrm{q}(\boldsymbol{u}) \log \frac{G(\boldsymbol{u})\,\mathrm{p}(\boldsymbol{u})}{\mathrm{q}(\boldsymbol{u})}\,\mathrm{d}\boldsymbol{u} + \frac{1}{2\sigma_n^2}\operatorname{tr}\big(\boldsymbol{K}_{ff} - \boldsymbol{Q}_{ff}\big) \\
&= -\int \log G(\boldsymbol{u})\,\mathrm{q}(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u} + \mathrm{KL}(\mathrm{q}(\boldsymbol{u}) \parallel \mathrm{p}(\boldsymbol{u})) + \frac{1}{2\sigma_n^2}\operatorname{tr}\big(\boldsymbol{K}_{ff} - \boldsymbol{Q}_{ff}\big) \\
&= -\mathbb{E}_{\mathrm{q}(\boldsymbol{u})}[\log G(\boldsymbol{u})] + \mathrm{KL}(\mathrm{q}(\boldsymbol{u}) \parallel \mathrm{p}(\boldsymbol{u})) + \frac{1}{2\sigma_n^2}\operatorname{tr}\big(\boldsymbol{K}_{ff} - \boldsymbol{Q}_{ff}\big)
\end{aligned}
$$

SVGP variational bound

$$
\begin{aligned}
\mathcal{L}^{\mathsf{SVGP}}(\boldsymbol{\theta}, Z, \boldsymbol{m}, S) &\geq -\int \mathrm{q}(\boldsymbol{u}) \log \frac{G(\boldsymbol{u})\,\mathrm{p}(\boldsymbol{u})}{\mathrm{q}(\boldsymbol{u})}\,\mathrm{d}\boldsymbol{u} + \frac{1}{2\sigma_n^2}\operatorname{tr}\big(\boldsymbol{K}_{ff} - \boldsymbol{Q}_{ff}\big) \\
&= -\int \mathrm{q}(\boldsymbol{u}) \log \frac{G(\boldsymbol{u})\,\mathrm{p}(\boldsymbol{u})}{\mathrm{q}(\boldsymbol{u})}\,\mathrm{d}\boldsymbol{u} + \frac{1}{2\sigma_n^2}\operatorname{tr}\big(\boldsymbol{K}_{ff} - \boldsymbol{Q}_{ff}\big) \\
&= -\int \log G(\boldsymbol{u})\,\mathrm{q}(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u} + \mathrm{KL}(\mathrm{q}(\boldsymbol{u}) \parallel \mathrm{p}(\boldsymbol{u})) + \frac{1}{2\sigma_n^2}\operatorname{tr}\big(\boldsymbol{K}_{ff} - \boldsymbol{Q}_{ff}\big) \\
&= -\mathbb{E}_{\mathrm{q}(\boldsymbol{u})}[\log G(\boldsymbol{u})] + \mathrm{KL}(\mathrm{q}(\boldsymbol{u}) \parallel \mathrm{p}(\boldsymbol{u})) + \frac{1}{2\sigma_n^2}\operatorname{tr}\big(\boldsymbol{K}_{ff} - \boldsymbol{Q}_{ff}\big)
\end{aligned}
$$

SVGP variational bound

$$
\begin{aligned}
\mathbb{E}_{\mathrm{q}(\boldsymbol{u})}[\log G(\boldsymbol{u})] &= \int \log G(\boldsymbol{u})\, \mathrm{q}(\boldsymbol{u})\, \mathrm{d}\boldsymbol{u} \\
&= \int \log \mathcal{N}\left(\boldsymbol{y} \,\middle|\, \boldsymbol{K}_{\boldsymbol{f}\boldsymbol{u}} \boldsymbol{K}_{\boldsymbol{u}\boldsymbol{u}}^{-1} \boldsymbol{u}, \sigma_n^2 \mathrm{I}\right) \mathcal{N}(\boldsymbol{u} \mid \boldsymbol{m}, S)\, \mathrm{d}\boldsymbol{u} \\
&= \log \mathcal{N}\left(\boldsymbol{y} \,\middle|\, \boldsymbol{K}_{\boldsymbol{f}\boldsymbol{u}} \boldsymbol{K}_{\boldsymbol{u}\boldsymbol{u}}^{-1} \boldsymbol{m}, \sigma_n^2 \mathrm{I}\right) - \frac{1}{2\sigma_n^2} \operatorname{tr}\left(\boldsymbol{K}_{\boldsymbol{f}\boldsymbol{u}} \boldsymbol{K}_{\boldsymbol{u}\boldsymbol{u}}^{-1} S \boldsymbol{K}_{\boldsymbol{u}\boldsymbol{u}}^{-1} \boldsymbol{K}_{\boldsymbol{u}\boldsymbol{f}}\right)
\end{aligned}
$$

SVGP variational bound

$$
\begin{aligned}
\mathcal{L}^{\mathrm{SVGP}}(\boldsymbol{\theta}, Z, \boldsymbol{m}, S) &\geq -\mathbb{E}_{\mathrm{q}(\boldsymbol{u})}[\log G(\boldsymbol{u})] + \mathrm{KL}(\mathrm{q}(\boldsymbol{u}) \parallel \mathrm{p}(\boldsymbol{u})) + \frac{1}{2\sigma_n^2}\operatorname{tr}\left(\boldsymbol{K}_{ff} - \boldsymbol{Q}_{ff}\right) \\
&= -\log \mathcal{N}\left(\boldsymbol{y} \,\middle|\, \boldsymbol{K}_{fu}\boldsymbol{K}_{uu}^{-1}\boldsymbol{m}, \sigma_n^2 \mathrm{I}\right) + \mathrm{KL}(\mathrm{q}(\boldsymbol{u}) \parallel \mathrm{p}(\boldsymbol{u})) \\
&\quad + \frac{1}{2\sigma_n^2}\operatorname{tr}\left(\boldsymbol{K}_{ff} - \boldsymbol{Q}_{ff}\right) + \frac{1}{2\sigma_n^2}\operatorname{tr}\left(\boldsymbol{K}_{fu}\boldsymbol{K}_{uu}^{-1}\boldsymbol{S}\boldsymbol{K}_{uu}^{-1}\boldsymbol{K}_{uf}\right)
\end{aligned}
$$

SVGP interpretation

$$\mathcal{L}^{\mathsf{GP}}(\boldsymbol{\theta}) = -\log \mathcal{N}\left(\boldsymbol{y} \,\middle|\, \mathbf{0}, \boldsymbol{K}_{ff} + \sigma_n^2 \mathbf{I}\right)$$

$$\mathcal{L}^{\mathsf{SPGP}}(\boldsymbol{\theta}) = -\log \mathcal{N}\left(\boldsymbol{y} \,\middle|\, \mathbf{0}, \boldsymbol{Q}_{ff} + \operatorname{diag}(\boldsymbol{K}_{ff} - \boldsymbol{Q}_{ff}) + \sigma_n^2 \mathbf{I}\right)$$

$$\mathcal{L}^{\mathsf{SGPR}}(\boldsymbol{\theta}, Z) \geq -\log \mathcal{N}\left(\boldsymbol{y} \,\middle|\, \mathbf{0}, \boldsymbol{Q}_{ff} + \sigma_n^2 \mathbf{I}\right) + \frac{1}{2\sigma_n^2} \operatorname{tr}\left(\boldsymbol{K}_{ff} - \boldsymbol{Q}_{ff}\right)$$

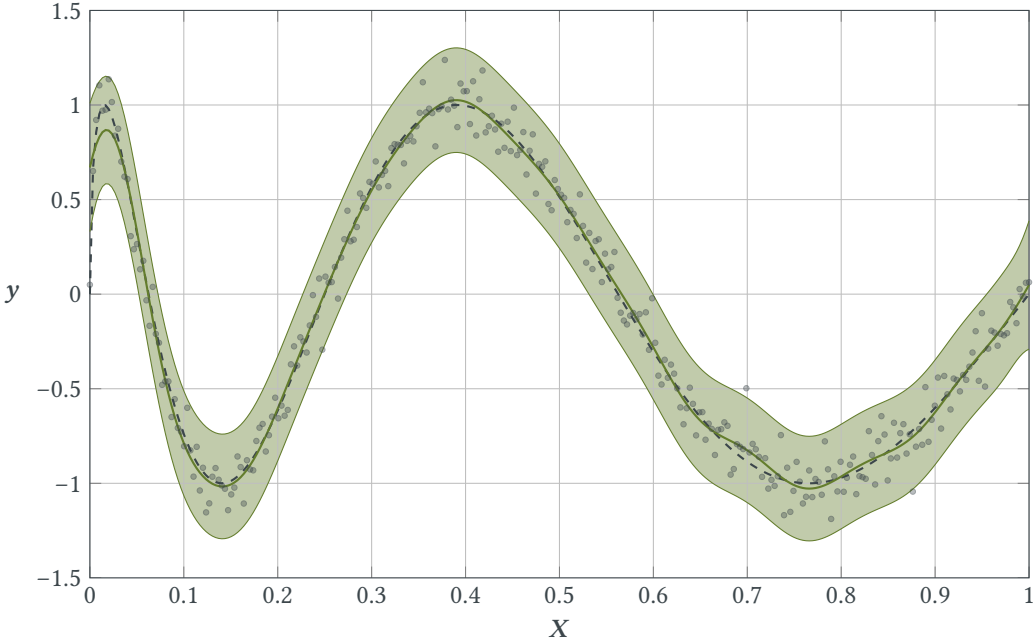
SVGP variational bound

$$\mathcal{L}^{\mathsf{SVGP}}(\boldsymbol{\theta}, Z, \boldsymbol{m}, S) \geq -\log \mathcal{N}\left(\boldsymbol{y} \,\middle|\, \boldsymbol{K}_{fu}\boldsymbol{K}_{uu}^{-1}\boldsymbol{m}, \sigma_n^2 \mathbf{I}\right) + \frac{1}{2\sigma_n^2} \operatorname{tr}\left(\boldsymbol{K}_{ff} - \boldsymbol{Q}_{ff}\right)$$

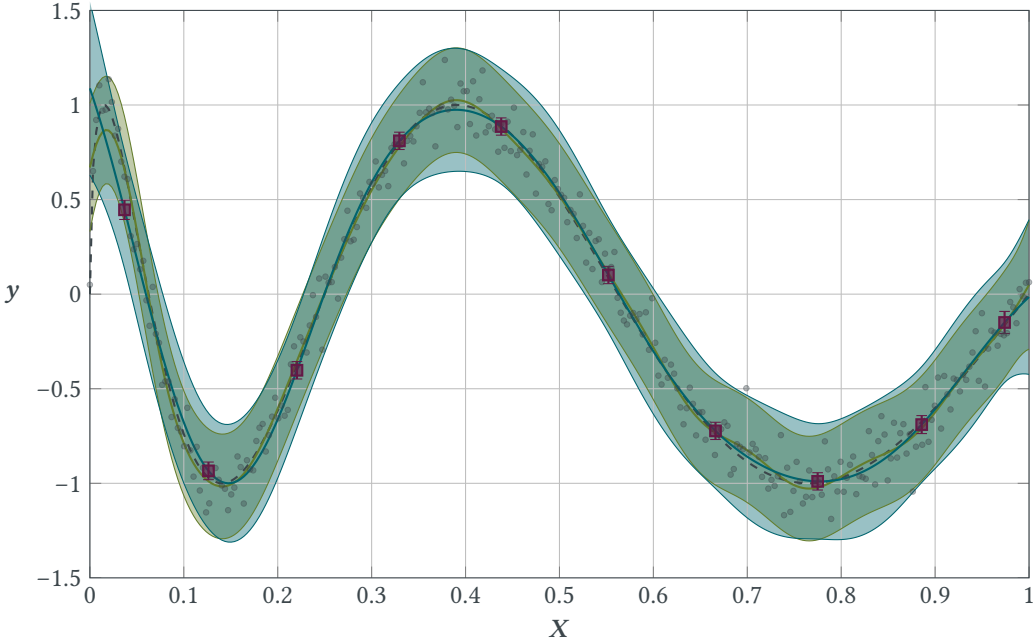
$$+ \operatorname{KL}(\mathrm{q}(\boldsymbol{u}) \,\|\, \mathrm{p}(\boldsymbol{u})) + \frac{1}{2\sigma_n^2} \operatorname{tr}\left(\boldsymbol{K}_{fu}\boldsymbol{K}_{uu}^{-1}\boldsymbol{S}\boldsymbol{K}_{uu}^{-1}\boldsymbol{K}_{uf}\right)$$

- The complete likelihood is still $\mathcal{O}(NM^2)$
- But it **factorizes** along the data - we can do stochastic optimization!
- A single prediction is only $\mathcal{O}(M^2)$ **and we do not need the data**

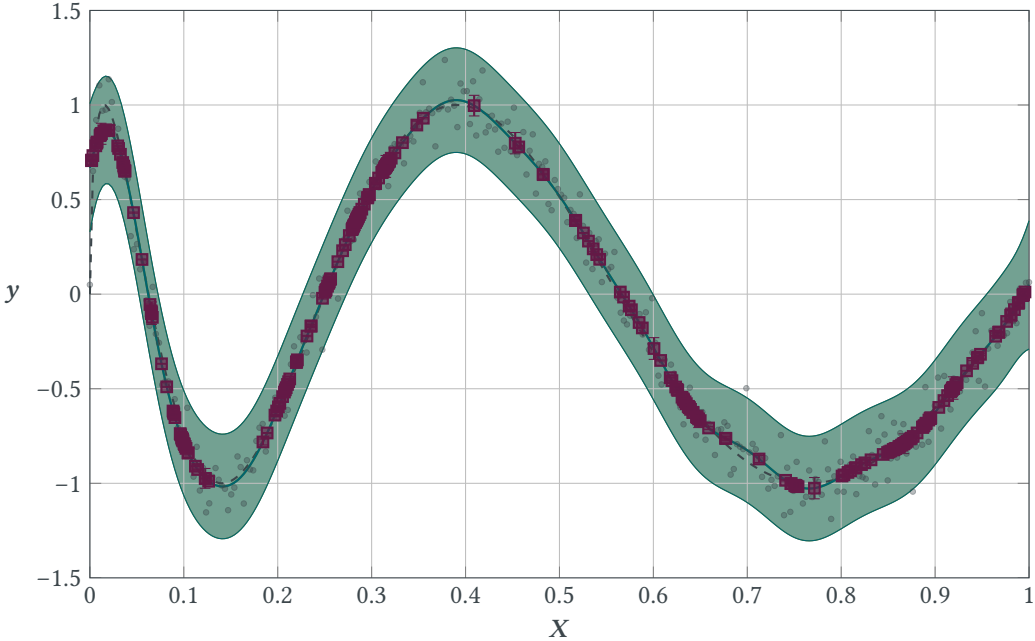
SVGP Showcase



SVGP Showcase



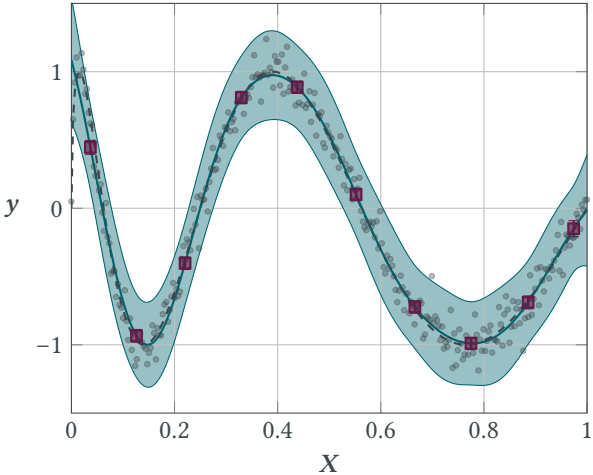
SVGP Showcase



Summary

We have made progress on seemingly fundamental restrictions of GPs:

- Scale to large data sets
- Scale to high dimensions (ish)
- Compact representations
- Acceptable speeds



Next time: Propagation of uncertainties?