

02477 – Bayesian Machine Learning: Lecture 11

Michael Riis Andersen

Technical University of Denmark,
DTU Compute, Department of Applied Math and Computer Science

Outline

- 1 The ELBO objective
- 2 Free-form vs fixed-form variational inference
- 3 Hyperparameter estimation in VI
- 4 Scaling Gaussian processes using variational inference

The ELBO objective

Variational inference: big picture

- Our goal is to approximate a posterior distribution of interest

$$p \equiv p(\mathbf{w}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{w})p(\mathbf{w})}{p(\mathbf{y})}$$

- Variational inference in three steps

1. Define collection of "simple" approximate probability distributions \mathcal{Q} (*the variational family*)
2. Define a measure of "distance" between probability distributions $\mathbb{D}[q||p]$
3. Search for the distribution $q \in \mathcal{Q}$ that resembles the exact posterior p as close as possible as measured by $\mathbb{D}[q||p]$

- The variational approximation q for the target distribution $p \approx q$ is defined as

$$q_* = \arg \min_{q \in \mathcal{Q}} \mathbb{D}[q||p]$$

Understanding the ELBO objective

- Suppose our model of interest is

$$p(\mathbf{y}, \mathbf{w}) = \prod_{n=1}^N p(y_n | \mathbf{w}) p(\mathbf{w})$$

- We optimize the ELBO \mathcal{L} to minimize the KL divergence (*between approximation and target*)

$$\mathcal{L}[q] = \mathbb{E}_q [\ln p(\mathbf{y}, \mathbf{w})] - \mathbb{E}_q [\ln q(\mathbf{w})]$$

Understanding the ELBO objective

- Suppose our model of interest is

$$p(\mathbf{y}, \mathbf{w}) = \prod_{n=1}^N p(y_n | \mathbf{w}) p(\mathbf{w})$$

- We optimize the ELBO \mathcal{L} to minimize the KL divergence (*between approximation and target*)

$$\mathcal{L}[q] = \mathbb{E}_q [\ln p(\mathbf{y}, \mathbf{w})] - \mathbb{E}_q [\ln q(\mathbf{w})]$$

- Re-writing the lower bound

$$\mathcal{L}[q] \equiv \mathbb{E}_q \left[\ln \prod_{n=1}^N p(y_n | \mathbf{w}) p(\mathbf{w}) \right] - \mathbb{E}_q [\ln q(\mathbf{w})]$$

Understanding the ELBO objective

- Suppose our model of interest is

$$p(\mathbf{y}, \mathbf{w}) = \prod_{n=1}^N p(y_n | \mathbf{w}) p(\mathbf{w})$$

- We optimize the ELBO \mathcal{L} to minimize the KL divergence (*between approximation and target*)

$$\mathcal{L}[q] = \mathbb{E}_q [\ln p(\mathbf{y}, \mathbf{w})] - \mathbb{E}_q [\ln q(\mathbf{w})]$$

- Re-writing the lower bound

$$\begin{aligned} \mathcal{L}[q] &\equiv \mathbb{E}_q \left[\ln \prod_{n=1}^N p(y_n | \mathbf{w}) p(\mathbf{w}) \right] - \mathbb{E}_q [\ln q(\mathbf{w})] \\ &= \sum_{n=1}^N \mathbb{E}_q [\ln p(y_n | \mathbf{w})] + \mathbb{E}_q [\ln p(\mathbf{w})] - \mathbb{E}_q [\ln q(\mathbf{w})] \end{aligned}$$

Understanding the ELBO objective

- Suppose our model of interest is

$$p(\mathbf{y}, \mathbf{w}) = \prod_{n=1}^N p(y_n | \mathbf{w}) p(\mathbf{w})$$

- We optimize the ELBO \mathcal{L} to minimize the KL divergence (*between approximation and target*)

$$\mathcal{L}[q] = \mathbb{E}_q [\ln p(\mathbf{y}, \mathbf{w})] - \mathbb{E}_q [\ln q(\mathbf{w})]$$

- Re-writing the lower bound

$$\begin{aligned} \mathcal{L}[q] &\equiv \mathbb{E}_q \left[\ln \prod_{n=1}^N p(y_n | \mathbf{w}) p(\mathbf{w}) \right] - \mathbb{E}_q [\ln q(\mathbf{w})] \\ &= \sum_{n=1}^N \mathbb{E}_q [\ln p(y_n | \mathbf{w})] + \mathbb{E}_q [\ln p(\mathbf{w})] - \mathbb{E}_q [\ln q(\mathbf{w})] \\ &= \sum_{n=1}^N \mathbb{E}_q [\ln p(y_n | \mathbf{w})] - \mathbb{E}_q \left[\ln \frac{q(\mathbf{w})}{p(\mathbf{w})} \right] \end{aligned}$$

Understanding the ELBO objective

- Suppose our model of interest is

$$p(\mathbf{y}, \mathbf{w}) = \prod_{n=1}^N p(y_n | \mathbf{w}) p(\mathbf{w})$$

- We optimize the ELBO \mathcal{L} to minimize the KL divergence (*between approximation and target*)

$$\mathcal{L}[q] = \mathbb{E}_q [\ln p(\mathbf{y}, \mathbf{w})] - \mathbb{E}_q [\ln q(\mathbf{w})]$$

- Re-writing the lower bound

$$\begin{aligned} \mathcal{L}[q] &\equiv \mathbb{E}_q \left[\ln \prod_{n=1}^N p(y_n | \mathbf{w}) p(\mathbf{w}) \right] - \mathbb{E}_q [\ln q(\mathbf{w})] \\ &= \sum_{n=1}^N \mathbb{E}_q [\ln p(y_n | \mathbf{w})] + \mathbb{E}_q [\ln p(\mathbf{w})] - \mathbb{E}_q [\ln q(\mathbf{w})] \\ &= \sum_{n=1}^N \mathbb{E}_q [\ln p(y_n | \mathbf{w})] - \mathbb{E}_q \left[\ln \frac{q(\mathbf{w})}{p(\mathbf{w})} \right] \\ &= \sum_{n=1}^N \mathbb{E}_q [\ln p(y_n | \mathbf{w})] - \text{KL}[q(\mathbf{w}) || p(\mathbf{w})] \end{aligned}$$

Understanding the ELBO objective

- Suppose our model of interest is

$$p(\mathbf{y}, \mathbf{w}) = \prod_{n=1}^N p(y_n | \mathbf{w}) p(\mathbf{w})$$

- We optimize the ELBO \mathcal{L} to minimize the KL divergence (*between approximation and target*)

$$\mathcal{L}[q] = \mathbb{E}_q [\ln p(\mathbf{y}, \mathbf{w})] - \mathbb{E}_q [\ln q(\mathbf{w})]$$

- Re-writing the lower bound

$$\begin{aligned} \mathcal{L}[q] &\equiv \mathbb{E}_q \left[\ln \prod_{n=1}^N p(y_n | \mathbf{w}) p(\mathbf{w}) \right] - \mathbb{E}_q [\ln q(\mathbf{w})] \\ &= \sum_{n=1}^N \mathbb{E}_q [\ln p(y_n | \mathbf{w})] + \mathbb{E}_q [\ln p(\mathbf{w})] - \mathbb{E}_q [\ln q(\mathbf{w})] \\ &= \sum_{n=1}^N \mathbb{E}_q [\ln p(y_n | \mathbf{w})] - \mathbb{E}_q \left[\ln \frac{q(\mathbf{w})}{p(\mathbf{w})} \right] \\ &= \sum_{n=1}^N \mathbb{E}_q [\ln p(y_n | \mathbf{w})] - \text{KL}[q(\mathbf{w}) || p(\mathbf{w})] \end{aligned}$$

- The first term (expected log likelihood) is a *data-fit* term, while the KL term encourages q to be close to the prior $p(\mathbf{w})$ (*regularization term*)

Free-form vs fixed-form variational inference

Free-form variational inference

- *Factorized approximation* for approximating the target distribution $p \equiv p(\mathbf{w}|\mathbf{y})$

$$q(\mathbf{w}) = \prod_{j=1}^J q(\mathbf{w}_j), \quad \text{where} \quad \mathbf{w} = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_J]$$

Free-form variational inference

- *Factorized approximation* for approximating the target distribution $p \equiv p(\mathbf{w}|\mathbf{y})$

$$q(\mathbf{w}) = \prod_{j=1}^J q(\mathbf{w}_j), \quad \text{where } \mathbf{w} = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_J]$$

- Optimality condition for mean-field families

$$\ln q^*(\mathbf{w}_k) = \mathbb{E}_{i \neq k} [\ln p(\mathbf{y}, \mathbf{w})] + K$$

Free-form variational inference

- *Factorized approximation* for approximating the target distribution $p \equiv p(\mathbf{w}|\mathbf{y})$

$$q(\mathbf{w}) = \prod_{j=1}^J q(\mathbf{w}_j), \quad \text{where } \mathbf{w} = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_J]$$

- Optimality condition for mean-field families

$$\ln q^*(\mathbf{w}_k) = \mathbb{E}_{i \neq k} [\ln p(\mathbf{y}, \mathbf{w})] + K$$

- Strategy for free-form variational inference
 1. Deduce optimal form for each factor and derive parameters
 2. Derive fixed-point algorithm based on the optimal parameters

Free-form variational inference

- *Factorized approximation* for approximating the target distribution $p \equiv p(\mathbf{w}|\mathbf{y})$

$$q(\mathbf{w}) = \prod_{j=1}^J q(\mathbf{w}_j), \quad \text{where } \mathbf{w} = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_J]$$

- Optimality condition for mean-field families

$$\ln q^*(\mathbf{w}_k) = \mathbb{E}_{i \neq k} [\ln p(\mathbf{y}, \mathbf{w})] + K$$

- Strategy for free-form variational inference

1. Deduce optimal form for each factor and derive parameters
2. Derive fixed-point algorithm based on the optimal parameters

- | | |
|---------------------------------------------|--------------------------------------------------|
| + Optimal functional form given assumptions | - Requires model-specific derivations |
| + Fast optimization | - Required integrals may be intractable |
| | - Optimal forms may not be "known" distributions |

Fixed-form variational inference

- *Fixed-form* variational inference as an alternative to *free-form*: We give up on the optimal functional form and simply *assume* some family of distributions q_ψ with parameters ψ
- We often refer to ψ as *variational parameters*

Fixed-form variational inference

- *Fixed-form* variational inference as an alternative to *free-form*: We give up on the optimal functional form and simply *assume* some family of distributions q_{ψ} with parameters ψ
- We often refer to ψ as *variational parameters*
- **Example:** Consider a model $p(\mathbf{y}, \mathbf{w})$, where $\mathbf{w} \in \mathbb{R}^D$

Fixed-form variational inference

- *Fixed-form* variational inference as an alternative to *free-form*: We give up on the optimal functional form and simply *assume* some family of distributions q_{ψ} with parameters ψ
- We often refer to ψ as *variational parameters*
- **Example:** Consider a model $p(\mathbf{y}, \mathbf{w})$, where $\mathbf{w} \in \mathbb{R}^D$
 - Full-rank Gaussians: $\mathbf{m} \in \mathbb{R}^D$, $\mathbf{V} \in \mathbb{R}^{D \times D}$
$$q(\mathbf{w}) = \mathcal{N}(\mathbf{w} | \mathbf{m}, \mathbf{V}),$$

Fixed-form variational inference

- *Fixed-form* variational inference as an alternative to *free-form*: We give up on the optimal functional form and simply *assume* some family of distributions q_{ψ} with parameters ψ
- We often refer to ψ as *variational parameters*
- **Example:** Consider a model $p(\mathbf{y}, \mathbf{w})$, where $\mathbf{w} \in \mathbb{R}^D$

- Full-rank Gaussians: $\mathbf{m} \in \mathbb{R}^D$, $\mathbf{V} \in \mathbb{R}^{D \times D}$

$$q(\mathbf{w}) = \mathcal{N}(\mathbf{w} | \mathbf{m}, \mathbf{V}),$$

- Low-rank Gaussians: $\mathbf{m} \in \mathbb{R}^D$, $\mathbf{B} \in \mathbb{R}^{D \times K}$, and $\mathbf{C} \in \mathbb{R}^{D \times D}$ is diagonal.

$$q(\mathbf{w}) = \mathcal{N}(\mathbf{w} | \mathbf{m}, \mathbf{B}\mathbf{B}^T + \mathbf{C}^2),$$

Fixed-form variational inference

- *Fixed-form* variational inference as an alternative to *free-form*: We give up on the optimal functional form and simply *assume* some family of distributions q_{ψ} with parameters ψ
- We often refer to ψ as *variational parameters*
- **Example:** Consider a model $p(\mathbf{y}, \mathbf{w})$, where $\mathbf{w} \in \mathbb{R}^D$

- Full-rank Gaussians: $\mathbf{m} \in \mathbb{R}^D$, $\mathbf{V} \in \mathbb{R}^{D \times D}$

$$q(\mathbf{w}) = \mathcal{N}(\mathbf{w} | \mathbf{m}, \mathbf{V}),$$

- Low-rank Gaussians: $\mathbf{m} \in \mathbb{R}^D$, $\mathbf{B} \in \mathbb{R}^{D \times K}$, and $\mathbf{C} \in \mathbb{R}^{D \times D}$ is diagonal.

$$q(\mathbf{w}) = \mathcal{N}(\mathbf{w} | \mathbf{m}, \mathbf{B}\mathbf{B}^T + \mathbf{C}^2),$$

Fixed-form variational inference

- *Fixed-form* variational inference as an alternative to *free-form*: We give up on the optimal functional form and simply *assume* some family of distributions q_{ψ} with parameters ψ

- We often refer to ψ as *variational parameters*

- **Example:** Consider a model $p(\mathbf{y}, \mathbf{w})$, where $\mathbf{w} \in \mathbb{R}^D$

- Full-rank Gaussians: $\mathbf{m} \in \mathbb{R}^D$, $\mathbf{V} \in \mathbb{R}^{D \times D}$

$$q(\mathbf{w}) = \mathcal{N}(\mathbf{w} | \mathbf{m}, \mathbf{V}),$$

- Low-rank Gaussians: $\mathbf{m} \in \mathbb{R}^D$, $\mathbf{B} \in \mathbb{R}^{D \times K}$, and $\mathbf{C} \in \mathbb{R}^{D \times D}$ is diagonal.

$$q(\mathbf{w}) = \mathcal{N}(\mathbf{w} | \mathbf{m}, \mathbf{B}\mathbf{B}^T + \mathbf{C}^2),$$

- Mean-field Gaussians: $\mathbf{m} = [m_1, \dots, m_D] \in \mathbb{R}^D$ and $\mathbf{v} = [v_1, \dots, v_D] \in \mathbb{R}^D$

$$q(\mathbf{w}) = \prod_{i=1}^D \mathcal{N}(w_i | m_i, v_i),$$

- Non-restricted to Gaussians: Gamma, Gauss, Beta, Dirichlet, Bernoulli etc

Fixed-form variational inference II

- *Fixed-form* variational inference as an alternative to *free-form*

- **Example:** Consider a model $p(\mathbf{y}, \mathbf{w})$, where $\mathbf{w} \in \mathbb{R}^D$

- Suppose we choose a full-rank Gaussian family

$$q(\mathbf{w}) = \mathcal{N}(\mathbf{w} | \mathbf{m}, \mathbf{V}),$$

such that the variational family \mathcal{Q} consists of all multivariate Gaussian distributions

- q_ψ is now *parametrized* by the *variational parameters* $\psi = \{\mathbf{m}, \mathbf{V}\}$

$$\begin{aligned} \mathcal{L}[q_\psi] &= \mathbb{E}_{q_\psi} [\ln p(\mathbf{y}, \mathbf{w})] - \mathbb{E}_{q_\psi} [\ln q_\psi(\mathbf{w})] \\ &= \mathbb{E}_{\mathcal{N}(\mathbf{w} | \mathbf{m}, \mathbf{V})} [\ln p(\mathbf{y}, \mathbf{w})] - \mathbb{E}_{\mathcal{N}(\mathbf{w} | \mathbf{m}, \mathbf{V})} [\ln \mathcal{N}(\mathbf{w} | \mathbf{m}, \mathbf{V})] \end{aligned}$$

- Fitting the approximation using gradient-based methods

$$q^* = \arg \max_{q \in \mathcal{Q}} \mathcal{L}[q] \quad \Longleftrightarrow$$

Fixed-form variational inference II

- *Fixed-form* variational inference as an alternative to *free-form*

- **Example:** Consider a model $p(\mathbf{y}, \mathbf{w})$, where $\mathbf{w} \in \mathbb{R}^D$

- Suppose we choose a full-rank Gaussian family

$$q(\mathbf{w}) = \mathcal{N}(\mathbf{w} | \mathbf{m}, \mathbf{V}),$$

such that the variational family \mathcal{Q} consists of all multivariate Gaussian distributions

- q_ψ is now *parametrized* by the *variational parameters* $\psi = \{\mathbf{m}, \mathbf{V}\}$

$$\begin{aligned} \mathcal{L}[q_\psi] &= \mathbb{E}_{q_\psi} [\ln p(\mathbf{y}, \mathbf{w})] - \mathbb{E}_{q_\psi} [\ln q_\psi(\mathbf{w})] \\ &= \mathbb{E}_{\mathcal{N}(\mathbf{w} | \mathbf{m}, \mathbf{V})} [\ln p(\mathbf{y}, \mathbf{w})] - \mathbb{E}_{\mathcal{N}(\mathbf{w} | \mathbf{m}, \mathbf{V})} [\ln \mathcal{N}(\mathbf{w} | \mathbf{m}, \mathbf{V})] \end{aligned}$$

- Fitting the approximation using gradient-based methods

$$q^* = \arg \max_{q \in \mathcal{Q}} \mathcal{L}[q] \quad \Longleftrightarrow$$

Fixed-form variational inference II

- *Fixed-form* variational inference as an alternative to *free-form*

- **Example:** Consider a model $p(\mathbf{y}, \mathbf{w})$, where $\mathbf{w} \in \mathbb{R}^D$

- Suppose we choose a full-rank Gaussian family

$$q(\mathbf{w}) = \mathcal{N}(\mathbf{w} | \mathbf{m}, \mathbf{V}),$$

such that the variational family \mathcal{Q} consists of all multivariate Gaussian distributions

- q_{ψ} is now *parametrized* by the *variational parameters* $\psi = \{\mathbf{m}, \mathbf{V}\}$

$$\begin{aligned} \mathcal{L}[q_{\psi}] &= \mathbb{E}_{q_{\psi}} [\ln p(\mathbf{y}, \mathbf{w})] - \mathbb{E}_{q_{\psi}} [\ln q_{\psi}(\mathbf{w})] \\ &= \mathbb{E}_{\mathcal{N}(\mathbf{w} | \mathbf{m}, \mathbf{V})} [\ln p(\mathbf{y}, \mathbf{w})] - \mathbb{E}_{\mathcal{N}(\mathbf{w} | \mathbf{m}, \mathbf{V})} [\ln \mathcal{N}(\mathbf{w} | \mathbf{m}, \mathbf{V})] \end{aligned}$$

- Fitting the approximation using gradient-based methods

$$q^* = \arg \max_{q \in \mathcal{Q}} \mathcal{L}[q] \quad \Longleftrightarrow \quad \psi^* = \arg \max_{\psi} \mathcal{L}[q_{\psi}] \quad \Longleftrightarrow$$

Fixed-form variational inference II

- *Fixed-form* variational inference as an alternative to *free-form*

- **Example:** Consider a model $p(\mathbf{y}, \mathbf{w})$, where $\mathbf{w} \in \mathbb{R}^D$

- Suppose we choose a full-rank Gaussian family

$$q(\mathbf{w}) = \mathcal{N}(\mathbf{w} | \mathbf{m}, \mathbf{V}),$$

such that the variational family \mathcal{Q} consists of all multivariate Gaussian distributions

- q_ψ is now *parametrized* by the *variational parameters* $\psi = \{\mathbf{m}, \mathbf{V}\}$

$$\begin{aligned} \mathcal{L}[q_\psi] &= \mathbb{E}_{q_\psi} [\ln p(\mathbf{y}, \mathbf{w})] - \mathbb{E}_{q_\psi} [\ln q_\psi(\mathbf{w})] \\ &= \mathbb{E}_{\mathcal{N}(\mathbf{w} | \mathbf{m}, \mathbf{V})} [\ln p(\mathbf{y}, \mathbf{w})] - \mathbb{E}_{\mathcal{N}(\mathbf{w} | \mathbf{m}, \mathbf{V})} [\ln \mathcal{N}(\mathbf{w} | \mathbf{m}, \mathbf{V})] \end{aligned}$$

- Fitting the approximation using gradient-based methods

$$q^* = \arg \max_{q \in \mathcal{Q}} \mathcal{L}[q] \quad \Longleftrightarrow \quad \psi^* = \arg \max_{\psi} \mathcal{L}[q_\psi] \quad \Longleftrightarrow \quad \mathbf{m}^*, \mathbf{V}^* = \arg \max_{\mathbf{m}, \mathbf{V}} \mathcal{L}[q_\psi]$$

Hyperparameter estimation in VI

Dealing with hyperparameters for variational inference

- Consider a model with data \mathbf{y} , parameters θ , and hyperparameters ξ

$$p(\theta|\mathbf{y}, \xi) = \frac{p(\mathbf{y}|\theta, \xi)p(\theta|\xi)}{p(\mathbf{y}|\xi)}$$

- **Examples**

1. Linear regression

$\theta = \mathbf{w}$ would be the regression weights

$\xi = \{\alpha, \beta\}$ would be prior and noise precision.

Dealing with hyperparameters for variational inference

- Consider a model with data \mathbf{y} , parameters θ , and hyperparameters ξ

$$p(\theta|\mathbf{y}, \xi) = \frac{p(\mathbf{y}|\theta, \xi)p(\theta|\xi)}{p(\mathbf{y}|\xi)}$$

■ Examples

1. Linear regression

$\theta = \mathbf{w}$ would be the regression weights

$\xi = \{\alpha, \beta\}$ would be prior and noise precision.

2. Gaussian process regression

$\theta = \mathbf{f}$ would be the latent function values

$\xi = \{\sigma^2, \kappa, \ell\}$ would be noise variance and kernel hyperparameters

Dealing with hyperparameters for variational inference

- Consider a model with data \mathbf{y} , parameters θ , and hyperparameters ξ

$$p(\theta|\mathbf{y}, \xi) = \frac{p(\mathbf{y}|\theta, \xi)p(\theta|\xi)}{p(\mathbf{y}|\xi)}$$

- **Examples**

1. Linear regression

$\theta = \mathbf{w}$ would be the regression weights

$\xi = \{\alpha, \beta\}$ would be prior and noise precision.

2. Gaussian process regression

$\theta = \mathbf{f}$ would be the latent function values

$\xi = \{\sigma^2, \kappa, \ell\}$ would be noise variance and kernel hyperparameters

- Hyperparameter estimation via the marginal likelihood (MLII/MAPII)

$$\hat{\xi} = \arg \max_{\xi} \log p(\mathbf{y}|\xi)$$

Dealing with hyperparameters for variational inference

- Consider a model with data \mathbf{y} , parameters θ , and hyperparameters ξ

$$p(\theta|\mathbf{y}, \xi) = \frac{p(\mathbf{y}|\theta, \xi)p(\theta|\xi)}{p(\mathbf{y}|\xi)}$$

- **Examples**

1. Linear regression

$\theta = \mathbf{w}$ would be the regression weights

$\xi = \{\alpha, \beta\}$ would be prior and noise precision.

2. Gaussian process regression

$\theta = \mathbf{f}$ would be the latent function values

$\xi = \{\sigma^2, \kappa, \ell\}$ would be noise variance and kernel hyperparameters

- Hyperparameter estimation via the marginal likelihood (MLII/MAPII)

$$\hat{\xi} = \arg \max_{\xi} \log p(\mathbf{y}|\xi)$$

- But what to do for non-conjugate model, where we cannot compute the marginal likelihood?

$$p(\theta|\mathbf{y}, \xi) \approx q(\theta)$$

Dealing with hyperparameters for variational inference

- Consider a model with data \mathbf{y} , parameters θ , and hyperparameters ξ

$$p(\theta|\mathbf{y}, \xi) = \frac{p(\mathbf{y}|\theta, \xi)p(\theta|\xi)}{p(\mathbf{y}|\xi)}$$

- Variational approximation $p(\theta|\mathbf{y}, \xi) \approx q(\theta)$ by optimizing the ELBO

$$\mathcal{L}_{\xi} [q] = \mathbb{E}_q [\ln p(\mathbf{y}, \theta|\xi)] - \mathbb{E}_q [\ln q(\theta)]$$

Dealing with hyperparameters for variational inference

- Consider a model with data \mathbf{y} , parameters θ , and hyperparameters ξ

$$p(\theta|\mathbf{y}, \xi) = \frac{p(\mathbf{y}|\theta, \xi)p(\theta|\xi)}{p(\mathbf{y}|\xi)}$$

- Variational approximation $p(\theta|\mathbf{y}, \xi) \approx q(\theta)$ by optimizing the ELBO

$$\mathcal{L}_{\xi} [q] = \mathbb{E}_q [\ln p(\mathbf{y}, \theta|\xi)] - \mathbb{E}_q [\ln q(\theta)]$$

- The ELBO is a lowerbound on the log marginal likelihood

$$\log p(\mathbf{y}|\xi) \geq \mathcal{L}_{\xi} [q]$$

Dealing with hyperparameters for variational inference

- Consider a model with data \mathbf{y} , parameters θ , and hyperparameters ξ

$$p(\theta|\mathbf{y}, \xi) = \frac{p(\mathbf{y}|\theta, \xi)p(\theta|\xi)}{p(\mathbf{y}|\xi)}$$

- Variational approximation $p(\theta|\mathbf{y}, \xi) \approx q(\theta)$ by optimizing the ELBO

$$\mathcal{L}_\xi [q] = \mathbb{E}_q [\ln p(\mathbf{y}, \theta|\xi)] - \mathbb{E}_q [\ln q(\theta)]$$

- The ELBO is a lowerbound on the log marginal likelihood

$$\log p(\mathbf{y}|\xi) \geq \mathcal{L}_\xi [q]$$

- and hence, we can do hyperparameter estimation via ELBO

$$\hat{\xi} = \arg \max_{\xi} p(\mathbf{y}|\xi) \approx \arg \max_{\xi} \mathcal{L}_\xi [q]$$

Dealing with hyperparameters for variational inference

- Consider a model with data \mathbf{y} , parameters θ , and hyperparameters ξ

$$p(\theta|\mathbf{y}, \xi) = \frac{p(\mathbf{y}|\theta, \xi)p(\theta|\xi)}{p(\mathbf{y}|\xi)}$$

- Variational approximation $p(\theta|\mathbf{y}, \xi) \approx q(\theta)$ by optimizing the ELBO

$$\mathcal{L}_\xi [q] = \mathbb{E}_q [\ln p(\mathbf{y}, \theta|\xi)] - \mathbb{E}_q [\ln q(\theta)]$$

- The ELBO is a lowerbound on the log marginal likelihood

$$\log p(\mathbf{y}|\xi) \geq \mathcal{L}_\xi [q]$$

- and hence, we can do hyperparameter estimation via ELBO

$$\hat{\xi} = \arg \max_{\xi} p(\mathbf{y}|\xi) \approx \arg \max_{\xi} \mathcal{L}_\xi [q]$$

- **Key take-away:** We can optimize the ELBO wrt. *variational parameters* ψ and *hyperparameters* ξ simultaneously

$$\hat{\psi}^*, \hat{\xi}^* = \arg \max_{\xi} \mathcal{L}_\xi [q_\psi]$$

Scaling Gaussian processes using variational inference

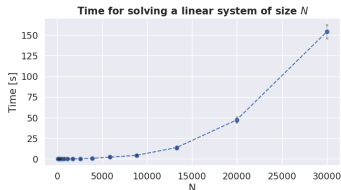
Speeding up Gaussian process inference

- Gaussian process priors, $f \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$ can be extremely powerful, but scales poorly: $\mathcal{O}(N^3)$ in computational complexity and $\mathcal{O}(N^2)$ in memory footprint

$$p(y^* | \mathbf{y}) = \mathcal{N}(y^* | \mu_{y^* | \mathbf{y}}, \sigma_{y^* | \mathbf{y}}^2)$$

$$\mu_{y^* | \mathbf{y}} = \mathbf{k} (\mathbf{K} + \beta^{-1} \mathbf{I})^{-1} \mathbf{y}$$

$$\sigma_{y^* | \mathbf{y}}^2 = c - \mathbf{k} (\mathbf{K} + \beta^{-1} \mathbf{I})^{-1} \mathbf{k}^T$$



Speeding up Gaussian process inference

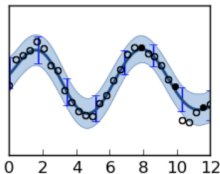
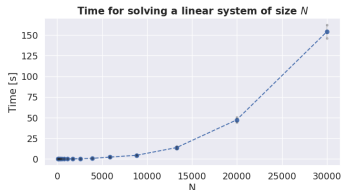
- Gaussian process priors, $f \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$ can be extremely powerful, but scales poorly: $\mathcal{O}(N^3)$ in computational complexity and $\mathcal{O}(N^2)$ in memory footprint

$$p(y^* | \mathbf{y}) = \mathcal{N}(y^* | \mu_{y^* | \mathbf{y}}, \sigma_{y^* | \mathbf{y}}^2)$$

$$\mu_{y^* | \mathbf{y}} = \mathbf{k} (\mathbf{K} + \beta^{-1} \mathbf{I})^{-1} \mathbf{y}$$

$$\sigma_{y^* | \mathbf{y}}^2 = c - \mathbf{k} (\mathbf{K} + \beta^{-1} \mathbf{I})^{-1} \mathbf{k}^T$$

- Obviously, using a subset of data would be faster, but can we do something more clever? Yes!



Hensman et al, 2019: Gaussian Processes for Big Data

Speeding up Gaussian process inference

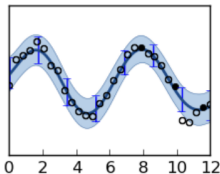
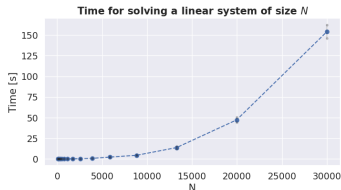
- Gaussian process priors, $f \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$ can be extremely powerful, but scales poorly: $\mathcal{O}(N^3)$ in computational complexity and $\mathcal{O}(N^2)$ in memory footprint

$$p(y^* | \mathbf{y}) = \mathcal{N}\left(y^* | \mu_{y^* | \mathbf{y}}, \sigma_{y^* | \mathbf{y}}^2\right)$$

$$\mu_{y^* | \mathbf{y}} = \mathbf{k} (\mathbf{K} + \beta^{-1} \mathbf{I})^{-1} \mathbf{y}$$

$$\sigma_{y^* | \mathbf{y}}^2 = c - \mathbf{k} (\mathbf{K} + \beta^{-1} \mathbf{I})^{-1} \mathbf{k}^T$$

- Obviously, using a subset of data would be faster, but can we do something more clever? Yes!
- We introduce *inducing points* \mathbf{z}_i for $i = 1, \dots, M$ for $M \ll N$



Hensman et al, 2019: Gaussian Processes for Big Data

Speeding up Gaussian process inference

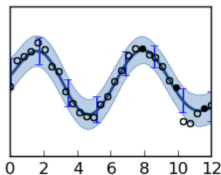
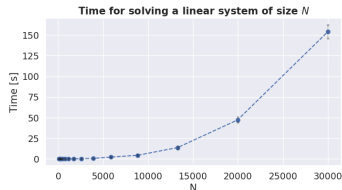
- Gaussian process priors, $f \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$ can be extremely powerful, but scales poorly: $\mathcal{O}(N^3)$ in computational complexity and $\mathcal{O}(N^2)$ in memory footprint

$$p(y^* | \mathbf{y}) = \mathcal{N}(y^* | \mu_{y^* | \mathbf{y}}, \sigma_{y^* | \mathbf{y}}^2)$$

$$\mu_{y^* | \mathbf{y}} = \mathbf{k} (\mathbf{K} + \beta^{-1} \mathbf{I})^{-1} \mathbf{y}$$

$$\sigma_{y^* | \mathbf{y}}^2 = c - \mathbf{k} (\mathbf{K} + \beta^{-1} \mathbf{I})^{-1} \mathbf{k}^T$$

- Obviously, using a subset of data would be faster, but can we do something more clever? Yes!
- We introduce *inducing points* \mathbf{z}_i for $i = 1, \dots, M$ for $M \ll N$
- Combining *variational inference* with *inducing points* allows us to approximate the posterior much faster, i.e. $\mathcal{O}(NM^2)$, and even faster, $\mathcal{O}(M^2)$, using *mini-batching* (next week)



Hensman et al, 2019: Gaussian Processes for Big Data

Introducing inducing points

- Suppose we have dataset $\mathcal{D} = \{\mathbf{x}_n, y_n\}$, where $\mathbf{x}_n \in \mathbb{R}^D$ and $y_n \in \mathbb{R}$

$$y_n = f(\mathbf{x}_n) + e_n$$

- Assuming Gaussian noise and a GP prior, i.e. $f(\mathbf{x}) \sim \mathcal{GP}(\mathbf{0}, k(\mathbf{x}, \mathbf{x}'))$ yields the joint distribution

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

- **Goal:** fast way compute $p(\mathbf{f}|\mathbf{y})$

$$f_n = f(\mathbf{x}_n)$$

Introducing inducing points

- Suppose we have dataset $\mathcal{D} = \{\mathbf{x}_n, y_n\}$, where $\mathbf{x}_n \in \mathbb{R}^D$ and $y_n \in \mathbb{R}$

$$y_n = f(\mathbf{x}_n) + e_n$$

- Assuming Gaussian noise and a GP prior, i.e. $f(\mathbf{x}) \sim \mathcal{GP}(\mathbf{0}, k(\mathbf{x}, \mathbf{x}'))$ yields the joint distribution

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

- **Goal:** fast way compute $p(\mathbf{f}|\mathbf{y})$

$$f_n = f(\mathbf{x}_n)$$

- We introduce M *inducing points* $\mathbf{z}_i \in \mathbb{R}^D$ for $m = 1, \dots, M$ such that

$$u_i = f(\mathbf{z}_i)$$

Introducing inducing points

- Suppose we have dataset $\mathcal{D} = \{\mathbf{x}_n, y_n\}$, where $\mathbf{x}_n \in \mathbb{R}^D$ and $y_n \in \mathbb{R}$

$$y_n = f(\mathbf{x}_n) + e_n$$

- Assuming Gaussian noise and a GP prior, i.e. $f(\mathbf{x}) \sim \mathcal{GP}(\mathbf{0}, k(\mathbf{x}, \mathbf{x}'))$ yields the joint distribution

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

- **Goal:** fast way compute $p(\mathbf{f}|\mathbf{y})$

$$f_n = f(\mathbf{x}_n)$$

- We introduce M *inducing points* $\mathbf{z}_i \in \mathbb{R}^D$ for $m = 1, \dots, M$ such that

$$u_i = f(\mathbf{z}_i)$$

- Extended joint distribution for data \mathbf{y} and latent function values for both \mathbf{f} and \mathbf{u}

$$p(\mathbf{y}, \mathbf{f}, \mathbf{u}) = p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{u})p(\mathbf{u}) = \mathcal{N}(\mathbf{y}|\mathbf{f}, \sigma^2 \mathbf{I})\mathcal{N}(\mathbf{f}|\mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{u}, \mathbf{K}_{ff} - \mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{K}_{fu})\mathcal{N}(\mathbf{u}|\mathbf{0}, \mathbf{K}_{uu})$$

Introducing inducing points

- Suppose we have dataset $\mathcal{D} = \{\mathbf{x}_n, y_n\}$, where $\mathbf{x}_n \in \mathbb{R}^D$ and $y_n \in \mathbb{R}$

$$y_n = f(\mathbf{x}_n) + e_n$$

- Assuming Gaussian noise and a GP prior, i.e. $f(\mathbf{x}) \sim \mathcal{GP}(\mathbf{0}, k(\mathbf{x}, \mathbf{x}'))$ yields the joint distribution

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

- **Goal:** fast way compute $p(\mathbf{f}|\mathbf{y})$

$$f_n = f(\mathbf{x}_n)$$

- We introduce M *inducing points* $\mathbf{z}_i \in \mathbb{R}^D$ for $m = 1, \dots, M$ such that

$$u_i = f(\mathbf{z}_i)$$

- Extended joint distribution for data \mathbf{y} and latent function values for both \mathbf{f} and \mathbf{u}

$$p(\mathbf{y}, \mathbf{f}, \mathbf{u}) = p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{u})p(\mathbf{u}) = \mathcal{N}(\mathbf{y}|\mathbf{f}, \sigma^2 \mathbf{I})\mathcal{N}(\mathbf{f}|\mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{u}, \mathbf{K}_{ff} - \mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{K}_{fu})\mathcal{N}(\mathbf{u}|\mathbf{0}, \mathbf{K}_{uu})$$

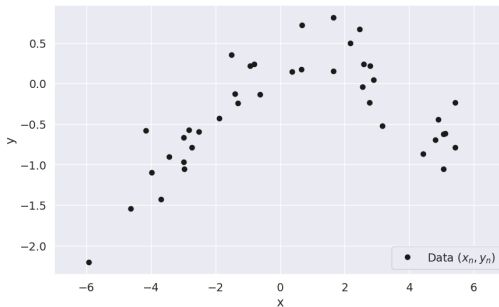
- We can always marginalize to get the original model back

$$p(\mathbf{y}, \mathbf{f}) = \int p(\mathbf{y}, \mathbf{f}, \mathbf{u})d\mathbf{u}$$

Intuition

■ Our extended joint distribution

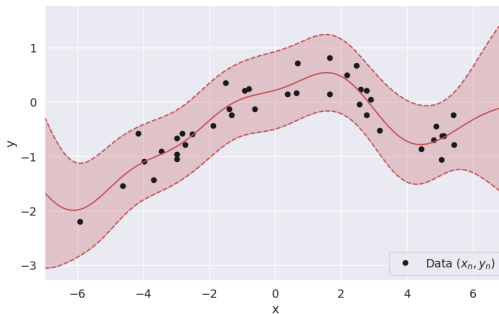
$$p(\mathbf{y}, \mathbf{f}, \mathbf{u}) = p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{u})p(\mathbf{u}) = \mathcal{N}(\mathbf{y}|\mathbf{f}, \sigma^2 \mathbf{I})\mathcal{N}(\mathbf{f}|\mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{u}, \mathbf{K}_{ff} - \mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{K}_{fu})\mathcal{N}(\mathbf{u}|\mathbf{0}, \mathbf{K}_{uu})$$



Intuition

■ Our extended joint distribution

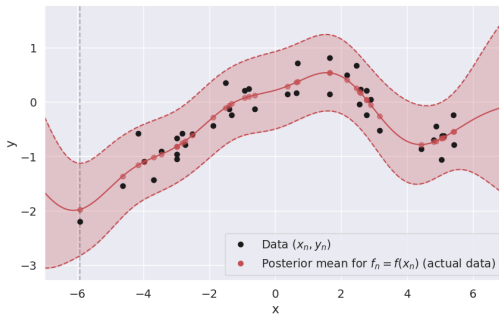
$$p(\mathbf{y}, \mathbf{f}, \mathbf{u}) = p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{u})p(\mathbf{u}) = \mathcal{N}(\mathbf{y}|\mathbf{f}, \sigma^2 \mathbf{I})\mathcal{N}(\mathbf{f}|\mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{u}, \mathbf{K}_{ff} - \mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{K}_{fu})\mathcal{N}(\mathbf{u}|\mathbf{0}, \mathbf{K}_{uu})$$



Intuition

■ Our extended joint distribution

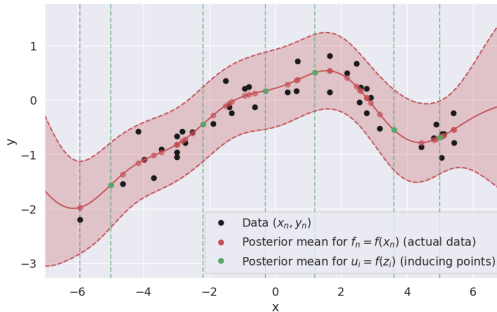
$$p(\mathbf{y}, \mathbf{f}, \mathbf{u}) = p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{u})p(\mathbf{u}) = \mathcal{N}(\mathbf{y}|\mathbf{f}, \sigma^2 \mathbf{I})\mathcal{N}(\mathbf{f}|\mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{u}, \mathbf{K}_{ff} - \mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{K}_{fu})\mathcal{N}(\mathbf{u}|\mathbf{0}, \mathbf{K}_{uu})$$



Intuition

■ Our extended joint distribution

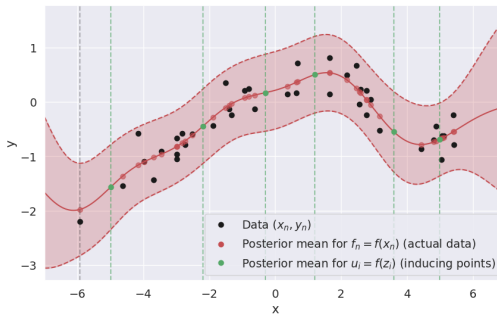
$$p(\mathbf{y}, \mathbf{f}, \mathbf{u}) = p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{u})p(\mathbf{u}) = \mathcal{N}(\mathbf{y}|\mathbf{f}, \sigma^2 \mathbf{I})\mathcal{N}(\mathbf{f}|\mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{u}, \mathbf{K}_{ff} - \mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{K}_{fu})\mathcal{N}(\mathbf{u}|\mathbf{0}, \mathbf{K}_{uu})$$



Intuition

■ Our extended joint distribution

$$p(\mathbf{y}, \mathbf{f}, \mathbf{u}) = p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{u})p(\mathbf{u}) = \mathcal{N}(\mathbf{y}|\mathbf{f}, \sigma^2 \mathbf{I})\mathcal{N}(\mathbf{f}|\mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{u}, \mathbf{K}_{ff} - \mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{K}_{fu})\mathcal{N}(\mathbf{u}|\mathbf{0}, \mathbf{K}_{uu})$$



■ Can we find a set of *inducing points* $\{\mathbf{z}_i\}_{i=1}^M$ and associated function values $u_i = f(\mathbf{z}_i)$ such that we can absorb all the information from the full data set?

Setting up the approximation and choosing a variational family

- Our extended joint distribution

$$p(\mathbf{y}, \mathbf{f}, \mathbf{u}) = p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{u})p(\mathbf{u}) = \mathcal{N}(\mathbf{y}|\mathbf{f}, \sigma^2 \mathbf{I})\mathcal{N}(\mathbf{f}|\mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{u}, \mathbf{K}_{ff} - \mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{K}_{fu})\mathcal{N}(\mathbf{u}|\mathbf{0}, \mathbf{K}_{uu})$$

Setting up the approximation and choosing a variational family

- Our extended joint distribution

$$p(\mathbf{y}, \mathbf{f}, \mathbf{u}) = p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{u})p(\mathbf{u}) = \mathcal{N}(\mathbf{y}|\mathbf{f}, \sigma^2 \mathbf{I})\mathcal{N}(\mathbf{f}|\mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{u}, \mathbf{K}_{ff} - \mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{K}_{fu})\mathcal{N}(\mathbf{u}|\mathbf{0}, \mathbf{K}_{uu})$$

- We will make a variational approximation: $p(\mathbf{f}, \mathbf{u}|\mathbf{y}) \approx q(\mathbf{f}, \mathbf{u})$

Setting up the approximation and choosing a variational family

- Our extended joint distribution

$$p(\mathbf{y}, \mathbf{f}, \mathbf{u}) = p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{u})p(\mathbf{u}) = \mathcal{N}(\mathbf{y}|\mathbf{f}, \sigma^2 \mathbf{I})\mathcal{N}(\mathbf{f}|\mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{u}, \mathbf{K}_{ff} - \mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{K}_{fu})\mathcal{N}(\mathbf{u}|\mathbf{0}, \mathbf{K}_{uu})$$

- We will make a variational approximation: $p(\mathbf{f}, \mathbf{u}|\mathbf{y}) \approx q(\mathbf{f}, \mathbf{u})$

- A clever choice for the variational family

$$q(\mathbf{f}, \mathbf{u}) = p(\mathbf{f}|\mathbf{u})q(\mathbf{u}) = \mathcal{N}(\mathbf{f}|\mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{u}, \mathbf{K}_{ff} - \mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{K}_{fu})\mathcal{N}(\mathbf{u}|\mathbf{m}_u, \mathbf{S}_u)$$

where

$\mathbf{m}_u \in \mathbb{R}^M$ and $\mathbf{S}_u \in \mathbb{R}^{M \times M}$ are variational parameters to be estimated

\mathbf{K}_{ff} is the prior covariance matrix for \mathbf{f}

\mathbf{K}_{uu} is the prior covariance matrix for \mathbf{u}

\mathbf{K}_{fu} is the prior covariance between \mathbf{f} and \mathbf{u}

Setting up the approximation and choosing a variational family

- Our extended joint distribution

$$p(\mathbf{y}, \mathbf{f}, \mathbf{u}) = p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{u})p(\mathbf{u}) = \mathcal{N}(\mathbf{y}|\mathbf{f}, \sigma^2 \mathbf{I})\mathcal{N}(\mathbf{f}|\mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{u}, \mathbf{K}_{ff} - \mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{K}_{fu})\mathcal{N}(\mathbf{u}|\mathbf{0}, \mathbf{K}_{uu})$$

- We will make a variational approximation: $p(\mathbf{f}, \mathbf{u}|\mathbf{y}) \approx q(\mathbf{f}, \mathbf{u})$

- A clever choice for the variational family

$$q(\mathbf{f}, \mathbf{u}) = p(\mathbf{f}|\mathbf{u})q(\mathbf{u}) = \mathcal{N}(\mathbf{f}|\mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{u}, \mathbf{K}_{ff} - \mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{K}_{fu})\mathcal{N}(\mathbf{u}|\mathbf{m}_u, \mathbf{S}_u)$$

where

$\mathbf{m}_u \in \mathbb{R}^M$ and $\mathbf{S}_u \in \mathbb{R}^{M \times M}$ are variational parameters to be estimated

\mathbf{K}_{ff} is the prior covariance matrix for \mathbf{f}

\mathbf{K}_{uu} is the prior covariance matrix for \mathbf{u}

\mathbf{K}_{fu} is the prior covariance between \mathbf{f} and \mathbf{u}

- Recall $p(\mathbf{f}|\mathbf{u})$ is the prior conditional distribution derived from the prior $p(\mathbf{f}, \mathbf{u})$.

The approximate posterior distribution

- Our choice of variational family:

$$q(\mathbf{f}, \mathbf{u}) = p(\mathbf{f}|\mathbf{u})q(\mathbf{u}) = \mathcal{N}(\mathbf{f}|\mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{u}, \mathbf{K}_{ff} - \mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{K}_{fu})\mathcal{N}(\mathbf{u}|\mathbf{m}_u, \mathbf{S}_u)$$

implies a marginal variational approximation for $q(\mathbf{f})$ (linear Gaussian system)

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

The approximate posterior distribution

- Our choice of variational family:

$$q(\mathbf{f}, \mathbf{u}) = p(\mathbf{f}|\mathbf{u})q(\mathbf{u}) = \mathcal{N}(\mathbf{f}|\mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{u}, \mathbf{K}_{ff} - \mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{K}_{fu})\mathcal{N}(\mathbf{u}|\mathbf{m}_u, \mathbf{S}_u)$$

implies a marginal variational approximation for $q(\mathbf{f})$ (linear Gaussian system)

$$\begin{aligned} q(\mathbf{f}) &= \int p(\mathbf{f}|\mathbf{u})q(\mathbf{u})d\mathbf{u} \\ &= \int \mathcal{N}(\mathbf{f}|\mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{u}, \mathbf{K}_{ff} - \mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{K}_{fu})\mathcal{N}(\mathbf{u}|\mathbf{m}_u, \mathbf{S}_u)d\mathbf{u} \end{aligned}$$

The approximate posterior distribution

- Our choice of variational family:

$$q(\mathbf{f}, \mathbf{u}) = p(\mathbf{f}|\mathbf{u})q(\mathbf{u}) = \mathcal{N}(\mathbf{f}|\mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{u}, \mathbf{K}_{ff} - \mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{K}_{fu})\mathcal{N}(\mathbf{u}|\mathbf{m}_u, \mathbf{S}_u)$$

implies a marginal variational approximation for $q(\mathbf{f})$ (linear Gaussian system)

$$\begin{aligned} q(\mathbf{f}) &= \int p(\mathbf{f}|\mathbf{u})q(\mathbf{u})d\mathbf{u} \\ &= \int \mathcal{N}(\mathbf{f}|\mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{u}, \mathbf{K}_{ff} - \mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{K}_{fu})\mathcal{N}(\mathbf{u}|\mathbf{m}_u, \mathbf{S}_u)d\mathbf{u} \\ &= \mathcal{N}(\mathbf{f}|\underbrace{\mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{m}_u}_{\mathbf{m}_f}, \underbrace{\mathbf{K}_{ff} + \mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}(\mathbf{S}_u - \mathbf{K}_{uu})\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf}}_{\mathbf{S}_f}) \end{aligned}$$

The approximate posterior distribution

- Our choice of variational family:

$$q(\mathbf{f}, \mathbf{u}) = p(\mathbf{f}|\mathbf{u})q(\mathbf{u}) = \mathcal{N}(\mathbf{f}|\mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{u}, \mathbf{K}_{ff} - \mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{K}_{fu})\mathcal{N}(\mathbf{u}|\mathbf{m}_u, \mathbf{S}_u)$$

implies a marginal variational approximation for $q(\mathbf{f})$ (linear Gaussian system)

$$\begin{aligned} q(\mathbf{f}) &= \int p(\mathbf{f}|\mathbf{u})q(\mathbf{u})d\mathbf{u} \\ &= \int \mathcal{N}(\mathbf{f}|\mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{u}, \mathbf{K}_{ff} - \mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{K}_{fu})\mathcal{N}(\mathbf{u}|\mathbf{m}_u, \mathbf{S}_u)d\mathbf{u} \\ &= \mathcal{N}(\mathbf{f}|\underbrace{\mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{m}_u}_{\mathbf{m}_f}, \underbrace{\mathbf{K}_{ff} + \mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}(\mathbf{S}_u - \mathbf{K}_{uu})\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf}}_{\mathbf{S}_f}) \\ &= \mathcal{N}(\mathbf{f}|\mathbf{m}_f, \mathbf{S}_f) \end{aligned}$$

The approximate posterior distribution

- Our choice of variational family:

$$q(\mathbf{f}, \mathbf{u}) = p(\mathbf{f}|\mathbf{u})q(\mathbf{u}) = \mathcal{N}(\mathbf{f}|\mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{u}, \mathbf{K}_{ff} - \mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{K}_{fu})\mathcal{N}(\mathbf{u}|\mathbf{m}_u, \mathbf{S}_u)$$

implies a marginal variational approximation for $q(\mathbf{f})$ (linear Gaussian system)

$$\begin{aligned} q(\mathbf{f}) &= \int p(\mathbf{f}|\mathbf{u})q(\mathbf{u})d\mathbf{u} \\ &= \int \mathcal{N}(\mathbf{f}|\mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{u}, \mathbf{K}_{ff} - \mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{K}_{fu})\mathcal{N}(\mathbf{u}|\mathbf{m}_u, \mathbf{S}_u)d\mathbf{u} \\ &= \mathcal{N}(\mathbf{f}|\underbrace{\mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{m}_u}_{\mathbf{m}_f}, \underbrace{\mathbf{K}_{ff} + \mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}(\mathbf{S}_u - \mathbf{K}_{uu})\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf}}_{\mathbf{S}_f}) \\ &= \mathcal{N}(\mathbf{f}|\mathbf{m}_f, \mathbf{S}_f) \end{aligned}$$

- ... and similarly we can make predictions for new input points to get $f^* = f(\mathbf{x}^*)$

$$\begin{aligned} p(f^*|\mathbf{y}) &= \int p(f^*|\mathbf{u})p(\mathbf{u}|\mathbf{y})d\mathbf{u} \\ &\approx \int p(f^*|\mathbf{u})q(\mathbf{u})d\mathbf{u} \\ &= \mathcal{N}(f^*|\underbrace{\mathbf{K}_{f^*u}\mathbf{K}_{uu}^{-1}\mathbf{m}_u}_{\mathbf{m}_{f^*}}, \underbrace{\mathbf{K}_{f^*f^*} + \mathbf{K}_{f^*u}\mathbf{K}_{uu}^{-1}(\mathbf{S}_u - \mathbf{K}_{uu})\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf^*}}_{\mathbf{S}_{f^*s}}) \end{aligned}$$

The approximate posterior distribution

- Our choice of variational family:

$$q(\mathbf{f}, \mathbf{u}) = p(\mathbf{f}|\mathbf{u})q(\mathbf{u}) = \mathcal{N}(\mathbf{f}|\mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{u}, \mathbf{K}_{ff} - \mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{K}_{fu})\mathcal{N}(\mathbf{u}|\mathbf{m}_u, \mathbf{S}_u)$$

implies a marginal variational approximation for $q(\mathbf{f})$ (linear Gaussian system)

$$\begin{aligned} q(\mathbf{f}) &= \int p(\mathbf{f}|\mathbf{u})q(\mathbf{u})d\mathbf{u} \\ &= \int \mathcal{N}(\mathbf{f}|\mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{u}, \mathbf{K}_{ff} - \mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{K}_{fu})\mathcal{N}(\mathbf{u}|\mathbf{m}_u, \mathbf{S}_u)d\mathbf{u} \\ &= \mathcal{N}(\mathbf{f}|\underbrace{\mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{m}_u}_{\mathbf{m}_f}, \underbrace{\mathbf{K}_{ff} + \mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}(\mathbf{S}_u - \mathbf{K}_{uu})\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf}}_{\mathbf{S}_f}) \\ &= \mathcal{N}(\mathbf{f}|\mathbf{m}_f, \mathbf{S}_f) \end{aligned}$$

- ... and similarly we can make predictions for new input points to get $f^* = f(\mathbf{x}^*)$

$$\begin{aligned} p(f^*|\mathbf{y}) &= \int p(f^*|\mathbf{u})p(\mathbf{u}|\mathbf{y})d\mathbf{u} \\ &\approx \int p(f^*|\mathbf{u})q(\mathbf{u})d\mathbf{u} \\ &= \mathcal{N}(f^*|\underbrace{\mathbf{K}_{f^*u}\mathbf{K}_{uu}^{-1}\mathbf{m}_u}_{\mathbf{m}_{f^*}}, \underbrace{\mathbf{K}_{f^*f^*} + \mathbf{K}_{f^*u}\mathbf{K}_{uu}^{-1}(\mathbf{S}_u - \mathbf{K}_{uu})\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf^*}}_{\mathbf{S}_{f^*s}}) \end{aligned}$$

- So all we need to do is to estimate the mean and covariance for $q(\mathbf{u}) = \mathcal{N}(\mathbf{u}|\mathbf{m}_u, \mathbf{S}_u)$

Calculating the ELBO

We substitute our model and variational approximation into the ELBO

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

Calculating the ELBO

We substitute our model and variational approximation into the ELBO

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

Calculating the ELBO

We substitute our model and variational approximation into the ELBO

$$\begin{aligned}\mathcal{L}[q] &= \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{y}, \mathbf{f}, \mathbf{u})] - \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log q(\mathbf{f}, \mathbf{u})] \\ &= \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{u})p(\mathbf{u})] - \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{f}|\mathbf{u})q(\mathbf{u})] \\ &= \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{y}|\mathbf{f})] + \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{f}|\mathbf{u})] + \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{u})] \\ &\quad - \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{f}|\mathbf{u})] - \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log q(\mathbf{u})]\end{aligned}$$

Calculating the ELBO

We substitute our model and variational approximation into the ELBO

$$\begin{aligned}
 \mathcal{L}[q] &= \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{y}, \mathbf{f}, \mathbf{u})] - \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log q(\mathbf{f}, \mathbf{u})] \\
 &= \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{u})p(\mathbf{u})] - \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{f}|\mathbf{u})q(\mathbf{u})] \\
 &= \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{y}|\mathbf{f})] + \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{f}|\mathbf{u})] + \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{u})] \\
 &\quad - \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{f}|\mathbf{u})] - \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log q(\mathbf{u})] \\
 &= \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{y}|\mathbf{f})] + \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{u})] - \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log q(\mathbf{u})]
 \end{aligned}$$

Calculating the ELBO

We substitute our model and variational approximation into the ELBO

$$\begin{aligned}
 \mathcal{L}[q] &= \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{y}, \mathbf{f}, \mathbf{u})] - \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log q(\mathbf{f}, \mathbf{u})] \\
 &= \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{u})p(\mathbf{u})] - \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{f}|\mathbf{u})q(\mathbf{u})] \\
 &= \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{y}|\mathbf{f})] + \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{f}|\mathbf{u})] + \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{u})] \\
 &\quad - \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{f}|\mathbf{u})] - \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log q(\mathbf{u})] \\
 &= \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{y}|\mathbf{f})] + \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{u})] - \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log q(\mathbf{u})] \\
 &= \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{y}|\mathbf{f})] + \mathbb{E}_{q(\mathbf{u})} [\log p(\mathbf{u})] - \mathbb{E}_{q(\mathbf{u})} [\log q(\mathbf{u})]
 \end{aligned}$$

Calculating the ELBO

We substitute our model and variational approximation into the ELBO

$$\begin{aligned}
 \mathcal{L}[q] &= \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{y}, \mathbf{f}, \mathbf{u})] - \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log q(\mathbf{f}, \mathbf{u})] \\
 &= \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{u})p(\mathbf{u})] - \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{f}|\mathbf{u})q(\mathbf{u})] \\
 &= \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{y}|\mathbf{f})] + \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{f}|\mathbf{u})] + \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{u})] \\
 &\quad - \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{f}|\mathbf{u})] - \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log q(\mathbf{u})] \\
 &= \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{y}|\mathbf{f})] + \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{u})] - \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log q(\mathbf{u})] \\
 &= \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{y}|\mathbf{f})] + \mathbb{E}_{q(\mathbf{u})} [\log p(\mathbf{u})] - \mathbb{E}_{q(\mathbf{u})} [\log q(\mathbf{u})] \\
 &= \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{y}|\mathbf{f})] - \text{KL}[q(\mathbf{u})||p(\mathbf{u})]
 \end{aligned}$$

Calculating the ELBO

We substitute our model and variational approximation into the ELBO

$$\begin{aligned}
 \mathcal{L}[q] &= \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{y}, \mathbf{f}, \mathbf{u})] - \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log q(\mathbf{f}, \mathbf{u})] \\
 &= \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{u})p(\mathbf{u})] - \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{f}|\mathbf{u})q(\mathbf{u})] \\
 &= \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{y}|\mathbf{f})] + \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{f}|\mathbf{u})] + \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{u})] \\
 &\quad - \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{f}|\mathbf{u})] - \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log q(\mathbf{u})] \\
 &= \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{y}|\mathbf{f})] + \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{u})] - \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log q(\mathbf{u})] \\
 &= \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{y}|\mathbf{f})] + \mathbb{E}_{q(\mathbf{u})} [\log p(\mathbf{u})] - \mathbb{E}_{q(\mathbf{u})} [\log q(\mathbf{u})] \\
 &= \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{y}|\mathbf{f})] - \text{KL}[q(\mathbf{u})||p(\mathbf{u})]
 \end{aligned}$$

Next, recall that $p(\mathbf{y}|\mathbf{f}) = \mathcal{N}(\mathbf{y}|\mathbf{f}, \sigma^2 \mathbf{I}) = \prod_{n=1}^N \mathcal{N}(y_n|f_n, \sigma^2)$

Calculating the ELBO

We substitute our model and variational approximation into the ELBO

$$\begin{aligned}
 \mathcal{L}[q] &= \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{y}, \mathbf{f}, \mathbf{u})] - \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log q(\mathbf{f}, \mathbf{u})] \\
 &= \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{u})p(\mathbf{u})] - \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{f}|\mathbf{u})q(\mathbf{u})] \\
 &= \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{y}|\mathbf{f})] + \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{f}|\mathbf{u})] + \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{u})] \\
 &\quad - \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{f}|\mathbf{u})] - \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log q(\mathbf{u})] \\
 &= \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{y}|\mathbf{f})] + \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{u})] - \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log q(\mathbf{u})] \\
 &= \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{y}|\mathbf{f})] + \mathbb{E}_{q(\mathbf{u})} [\log p(\mathbf{u})] - \mathbb{E}_{q(\mathbf{u})} [\log q(\mathbf{u})] \\
 &= \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{y}|\mathbf{f})] - \text{KL}[q(\mathbf{u})||p(\mathbf{u})]
 \end{aligned}$$

Next, recall that $p(\mathbf{y}|\mathbf{f}) = \mathcal{N}(\mathbf{y}|\mathbf{f}, \sigma^2 \mathbf{I}) = \prod_{n=1}^N \mathcal{N}(y_n|f_n, \sigma^2)$

$$\mathcal{L}[q] = \sum_{n=1}^N \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log \mathcal{N}(y_n|f_n, \sigma^2)] - \text{KL}[q(\mathbf{u})||p(\mathbf{u})]$$

Calculating the ELBO

We substitute our model and variational approximation into the ELBO

$$\begin{aligned}
 \mathcal{L}[q] &= \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{y}, \mathbf{f}, \mathbf{u})] - \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log q(\mathbf{f}, \mathbf{u})] \\
 &= \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{u})p(\mathbf{u})] - \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{f}|\mathbf{u})q(\mathbf{u})] \\
 &= \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{y}|\mathbf{f})] + \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{f}|\mathbf{u})] + \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{u})] \\
 &\quad - \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{f}|\mathbf{u})] - \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log q(\mathbf{u})] \\
 &= \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{y}|\mathbf{f})] + \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{u})] - \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log q(\mathbf{u})] \\
 &= \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{y}|\mathbf{f})] + \mathbb{E}_{q(\mathbf{u})} [\log p(\mathbf{u})] - \mathbb{E}_{q(\mathbf{u})} [\log q(\mathbf{u})] \\
 &= \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log p(\mathbf{y}|\mathbf{f})] - \text{KL}[q(\mathbf{u})||p(\mathbf{u})]
 \end{aligned}$$

Next, recall that $p(\mathbf{y}|\mathbf{f}) = \mathcal{N}(\mathbf{y}|\mathbf{f}, \sigma^2 \mathbf{I}) = \prod_{n=1}^N \mathcal{N}(y_n|f_n, \sigma^2)$

$$\begin{aligned}
 \mathcal{L}[q] &= \sum_{n=1}^N \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} [\log \mathcal{N}(y_n|f_n, \sigma^2)] - \text{KL}[q(\mathbf{u})||p(\mathbf{u})] \\
 &= \sum_{n=1}^N \mathbb{E}_{p(f_n|\mathbf{u})q(\mathbf{u})} [\log \mathcal{N}(y_n|f_n, \sigma^2)] - \text{KL}[q(\mathbf{u})||p(\mathbf{u})]
 \end{aligned}$$

Calculating the first term

$$\mathbb{E}_{p(f_n|\mathbf{u})q(\mathbf{u})} \left[\log \mathcal{N}(y_n|f_n, \sigma^2) \right] = \int \int p(f_n|\mathbf{u})q(\mathbf{u}) \log \mathcal{N}(y_n|f_n, \sigma^2) df_n d\mathbf{u}$$

Calculating the first term

$$\begin{aligned}\mathbb{E}_{p(f_n|\mathbf{u})q(\mathbf{u})} \left[\log \mathcal{N}(y_n|f_n, \sigma^2) \right] &= \int \int p(f_n|\mathbf{u})q(\mathbf{u}) \log \mathcal{N}(y_n|f_n, \sigma^2) df_n d\mathbf{u} \\ &= \int \int p(f_n|\mathbf{u})q(\mathbf{u}) d\mathbf{u} \log \mathcal{N}(y_n|f_n, \sigma^2) df_n \quad (\text{Recall: } q(\mathbf{f}) = \mathcal{N}(\mathbf{f}|\mathbf{m}_f, \mathbf{S}_f))\end{aligned}$$

Calculating the first term

$$\begin{aligned}\mathbb{E}_{p(f_n|u)q(u)} \left[\log \mathcal{N}(y_n|f_n, \sigma^2) \right] &= \int \int p(f_n|u)q(u) \log \mathcal{N}(y_n|f_n, \sigma^2) df_n du \\ &= \int \int p(f_n|u)q(u) du \log \mathcal{N}(y_n|f_n, \sigma^2) df_n && (\text{Recall: } q(\mathbf{f}) = \mathcal{N}(\mathbf{f}|\mathbf{m}_f, \mathbf{S}_f)) \\ &= \int \mathcal{N}(f_n|m_{f_n}, \sigma_{f_n}^2) \log \mathcal{N}(y_n|f_n, \sigma^2) df_n\end{aligned}$$

Calculating the first term

$$\begin{aligned}\mathbb{E}_{p(f_n|u)q(u)} \left[\log \mathcal{N}(y_n|f_n, \sigma^2) \right] &= \int \int p(f_n|u)q(u) \log \mathcal{N}(y_n|f_n, \sigma^2) df_n du \\ &= \int \int p(f_n|u)q(u) du \log \mathcal{N}(y_n|f_n, \sigma^2) df_n && (\text{Recall: } q(\mathbf{f}) = \mathcal{N}(\mathbf{f}|\mathbf{m}_f, \mathbf{S}_f)) \\ &= \int \mathcal{N}(f_n|m_{f_n}, \sigma_{f_n}^2) \log \mathcal{N}(y_n|f_n, \sigma^2) df_n \\ &= \mathbb{E}_{\mathcal{N}(f_n|m_{f_n}, \sigma_{f_n}^2)} \left[\log \mathcal{N}(y_n|f_n, \sigma^2) \right]\end{aligned}$$

Calculating the first term

$$\begin{aligned}
 \mathbb{E}_{p(f_n|u)q(u)} \left[\log \mathcal{N}(y_n|f_n, \sigma^2) \right] &= \int \int p(f_n|u)q(u) \log \mathcal{N}(y_n|f_n, \sigma^2) df_n du \\
 &= \int \int p(f_n|u)q(u) du \log \mathcal{N}(y_n|f_n, \sigma^2) df_n && (\text{Recall: } q(\mathbf{f}) = \mathcal{N}(\mathbf{f}|\mathbf{m}_f, \mathbf{S}_f)) \\
 &= \int \mathcal{N}(f_n|m_{f_n}, \sigma_{f_n}^2) \log \mathcal{N}(y_n|f_n, \sigma^2) df_n \\
 &= \mathbb{E}_{\mathcal{N}(f_n|m_{f_n}, \sigma_{f_n}^2)} \left[\log \mathcal{N}(y_n|f_n, \sigma^2) \right] \\
 &= \mathbb{E}_{\mathcal{N}(f_n|m_{f_n}, \sigma_{f_n}^2)} \left[-\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} (y_n - f_n)^2 \right]
 \end{aligned}$$

Calculating the first term

$$\begin{aligned}
 \mathbb{E}_{p(f_n|u)q(u)} \left[\log \mathcal{N}(y_n|f_n, \sigma^2) \right] &= \int \int p(f_n|u)q(u) \log \mathcal{N}(y_n|f_n, \sigma^2) df_n du \\
 &= \int \int p(f_n|u)q(u) du \log \mathcal{N}(y_n|f_n, \sigma^2) df_n && (\text{Recall: } q(\mathbf{f}) = \mathcal{N}(\mathbf{f}|\mathbf{m}_f, \mathbf{S}_f)) \\
 &= \int \mathcal{N}(f_n|m_{f_n}, \sigma_{f_n}^2) \log \mathcal{N}(y_n|f_n, \sigma^2) df_n \\
 &= \mathbb{E}_{\mathcal{N}(f_n|m_{f_n}, \sigma_{f_n}^2)} \left[\log \mathcal{N}(y_n|f_n, \sigma^2) \right] \\
 &= \mathbb{E}_{\mathcal{N}(f_n|m_{f_n}, \sigma_{f_n}^2)} \left[-\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma} (y_n - f_n)^2 \right] \\
 &= \mathbb{E}_{\mathcal{N}(f_n|m_{f_n}, \sigma_{f_n}^2)} \left[-\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma} (y_n^2 + f_n^2 - 2y_n f_n) \right]
 \end{aligned}$$

Calculating the first term

$$\begin{aligned}
 \mathbb{E}_{p(f_n|u)q(u)} \left[\log \mathcal{N}(y_n|f_n, \sigma^2) \right] &= \int \int p(f_n|u)q(u) \log \mathcal{N}(y_n|f_n, \sigma^2) df_n du \\
 &= \int \int p(f_n|u)q(u) du \log \mathcal{N}(y_n|f_n, \sigma^2) df_n && \text{(Recall: } q(\mathbf{f}) = \mathcal{N}(\mathbf{f}|\mathbf{m}_f, \mathbf{S}_f)\text{)} \\
 &= \int \mathcal{N}(f_n|m_{f_n}, \sigma_{f_n}^2) \log \mathcal{N}(y_n|f_n, \sigma^2) df_n \\
 &= \mathbb{E}_{\mathcal{N}(f_n|m_{f_n}, \sigma_{f_n}^2)} \left[\log \mathcal{N}(y_n|f_n, \sigma^2) \right] \\
 &= \mathbb{E}_{\mathcal{N}(f_n|m_{f_n}, \sigma_{f_n}^2)} \left[-\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma} (y_n - f_n)^2 \right] \\
 &= \mathbb{E}_{\mathcal{N}(f_n|m_{f_n}, \sigma_{f_n}^2)} \left[-\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma} (y_n^2 + f_n^2 - 2y_n f_n) \right] \\
 &= -\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma} (y_n^2 + \mathbb{E}_{\mathcal{N}(f_n|m_{f_n}, \sigma_{f_n}^2)} [f_n^2] - 2y_n \mathbb{E}_{\mathcal{N}(f_n|m_{f_n}, \sigma_{f_n}^2)} [f_n])
 \end{aligned}$$

Calculating the first term

$$\begin{aligned}
 \mathbb{E}_{p(f_n|u)q(u)} \left[\log \mathcal{N}(y_n|f_n, \sigma^2) \right] &= \int \int p(f_n|u)q(u) \log \mathcal{N}(y_n|f_n, \sigma^2) df_n du \\
 &= \int \int p(f_n|u)q(u) du \log \mathcal{N}(y_n|f_n, \sigma^2) df_n && \text{(Recall: } q(\mathbf{f}) = \mathcal{N}(\mathbf{f}|\mathbf{m}_f, \mathbf{S}_f)\text{)} \\
 &= \int \mathcal{N}(f_n|m_{f_n}, \sigma_{f_n}^2) \log \mathcal{N}(y_n|f_n, \sigma^2) df_n \\
 &= \mathbb{E}_{\mathcal{N}(f_n|m_{f_n}, \sigma_{f_n}^2)} \left[\log \mathcal{N}(y_n|f_n, \sigma^2) \right] \\
 &= \mathbb{E}_{\mathcal{N}(f_n|m_{f_n}, \sigma_{f_n}^2)} \left[-\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma} (y_n - f_n)^2 \right] \\
 &= \mathbb{E}_{\mathcal{N}(f_n|m_{f_n}, \sigma_{f_n}^2)} \left[-\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma} (y_n^2 + f_n^2 - 2y_n f_n) \right] \\
 &= -\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma} (y_n^2 + \mathbb{E}_{\mathcal{N}(f_n|m_{f_n}, \sigma_{f_n}^2)} [f_n^2] - 2y_n \mathbb{E}_{\mathcal{N}(f_n|m_{f_n}, \sigma_{f_n}^2)} [f_n]) \\
 &= -\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma} (y_n^2 + m_{f_n}^2 + \sigma_{f_n}^2 - 2y_n m_{f_n})
 \end{aligned}$$

Calculating the first term

$$\begin{aligned}
 \mathbb{E}_{p(f_n|u)q(u)} \left[\log \mathcal{N}(y_n|f_n, \sigma^2) \right] &= \int \int p(f_n|u)q(u) \log \mathcal{N}(y_n|f_n, \sigma^2) df_n du \\
 &= \int \int p(f_n|u)q(u) du \log \mathcal{N}(y_n|f_n, \sigma^2) df_n && \text{(Recall: } q(\mathbf{f}) = \mathcal{N}(\mathbf{f}|\mathbf{m}_f, \mathbf{S}_f)\text{)} \\
 &= \int \mathcal{N}(f_n|m_{f_n}, \sigma_{f_n}^2) \log \mathcal{N}(y_n|f_n, \sigma^2) df_n \\
 &= \mathbb{E}_{\mathcal{N}(f_n|m_{f_n}, \sigma_{f_n}^2)} \left[\log \mathcal{N}(y_n|f_n, \sigma^2) \right] \\
 &= \mathbb{E}_{\mathcal{N}(f_n|m_{f_n}, \sigma_{f_n}^2)} \left[-\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma} (y_n - f_n)^2 \right] \\
 &= \mathbb{E}_{\mathcal{N}(f_n|m_{f_n}, \sigma_{f_n}^2)} \left[-\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma} (y_n^2 + f_n^2 - 2y_n f_n) \right] \\
 &= -\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma} (y_n^2 + \mathbb{E}_{\mathcal{N}(f_n|m_{f_n}, \sigma_{f_n}^2)} [f_n^2] - 2y_n \mathbb{E}_{\mathcal{N}(f_n|m_{f_n}, \sigma_{f_n}^2)} [f_n]) \\
 &= -\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma} (y_n^2 + m_{f_n}^2 + \sigma_{f_n}^2 - 2y_n m_{f_n}) \\
 &= -\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma} (y_n^2 + m_{f_n}^2 - 2y_n m_{f_n}) - \frac{1}{2\sigma} \sigma_{f_n}^2
 \end{aligned}$$

Calculating the first term

$$\begin{aligned}
 \mathbb{E}_{p(f_n|u)q(u)} \left[\log \mathcal{N}(y_n|f_n, \sigma^2) \right] &= \int \int p(f_n|u)q(u) \log \mathcal{N}(y_n|f_n, \sigma^2) df_n du \\
 &= \int \int p(f_n|u)q(u) du \log \mathcal{N}(y_n|f_n, \sigma^2) df_n && (\text{Recall: } q(\mathbf{f}) = \mathcal{N}(\mathbf{f}|\mathbf{m}_f, \mathbf{S}_f)) \\
 &= \int \mathcal{N}(f_n|m_{f_n}, \sigma_{f_n}^2) \log \mathcal{N}(y_n|f_n, \sigma^2) df_n \\
 &= \mathbb{E}_{\mathcal{N}(f_n|m_{f_n}, \sigma_{f_n}^2)} \left[\log \mathcal{N}(y_n|f_n, \sigma^2) \right] \\
 &= \mathbb{E}_{\mathcal{N}(f_n|m_{f_n}, \sigma_{f_n}^2)} \left[-\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma} (y_n - f_n)^2 \right] \\
 &= \mathbb{E}_{\mathcal{N}(f_n|m_{f_n}, \sigma_{f_n}^2)} \left[-\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma} (y_n^2 + f_n^2 - 2y_n f_n) \right] \\
 &= -\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma} (y_n^2 + \mathbb{E}_{\mathcal{N}(f_n|m_{f_n}, \sigma_{f_n}^2)} [f_n^2] - 2y_n \mathbb{E}_{\mathcal{N}(f_n|m_{f_n}, \sigma_{f_n}^2)} [f_n]) \\
 &= -\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma} (y_n^2 + m_{f_n}^2 + \sigma_{f_n}^2 - 2y_n m_{f_n}) \\
 &= -\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma} (y_n^2 + m_{f_n}^2 - 2y_n m_{f_n}) - \frac{1}{2\sigma} \sigma_{f_n}^2 \\
 &= -\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma} (y_n - m_{f_n})^2 - \frac{1}{2\sigma} \sigma_{f_n}^2
 \end{aligned}$$

Calculating the first term

$$\begin{aligned}
 \mathbb{E}_{p(f_n|u)q(u)} \left[\log \mathcal{N}(y_n|f_n, \sigma^2) \right] &= \int \int p(f_n|u)q(u) \log \mathcal{N}(y_n|f_n, \sigma^2) df_n du \\
 &= \int \int p(f_n|u)q(u) du \log \mathcal{N}(y_n|f_n, \sigma^2) df_n && (\text{Recall: } q(\mathbf{f}) = \mathcal{N}(\mathbf{f}|\mathbf{m}_f, \mathbf{S}_f)) \\
 &= \int \mathcal{N}(f_n|m_{f_n}, \sigma_{f_n}^2) \log \mathcal{N}(y_n|f_n, \sigma^2) df_n \\
 &= \mathbb{E}_{\mathcal{N}(f_n|m_{f_n}, \sigma_{f_n}^2)} \left[\log \mathcal{N}(y_n|f_n, \sigma^2) \right] \\
 &= \mathbb{E}_{\mathcal{N}(f_n|m_{f_n}, \sigma_{f_n}^2)} \left[-\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma} (y_n - f_n)^2 \right] \\
 &= \mathbb{E}_{\mathcal{N}(f_n|m_{f_n}, \sigma_{f_n}^2)} \left[-\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma} (y_n^2 + f_n^2 - 2y_n f_n) \right] \\
 &= -\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma} (y_n^2 + \mathbb{E}_{\mathcal{N}(f_n|m_{f_n}, \sigma_{f_n}^2)} [f_n^2] - 2y_n \mathbb{E}_{\mathcal{N}(f_n|m_{f_n}, \sigma_{f_n}^2)} [f_n]) \\
 &= -\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma} (y_n^2 + m_{f_n}^2 + \sigma_{f_n}^2 - 2y_n m_{f_n}) \\
 &= -\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma} (y_n^2 + m_{f_n}^2 - 2y_n m_{f_n}) - \frac{1}{2\sigma} \sigma_{f_n}^2 \\
 &= -\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma} (y_n - m_{f_n})^2 - \frac{1}{2\sigma} \sigma_{f_n}^2 \\
 &= \log \mathcal{N}(y_n|m_{f_n}, \sigma^2) df_n - \frac{1}{2\sigma} \sigma_{f_n}^2
 \end{aligned}$$

The KL term

- The two terms in the lowerbound

$$\mathcal{L}[q] = \sum_{n=1}^N \mathbb{E}_{p(f_n|\mathbf{u})q(\mathbf{u})} [\log \mathcal{N}(y_n|f_n, \sigma^2)] - \text{KL}[q(\mathbf{u})||p(\mathbf{u})]$$

The KL term

- The two terms in the lowerbound

$$\mathcal{L}[q] = \sum_{n=1}^N \mathbb{E}_{p(f_n|\mathbf{u})q(\mathbf{u})} [\log \mathcal{N}(y_n|f_n, \sigma^2)] - \text{KL}[q(\mathbf{u})||p(\mathbf{u})]$$

- The KL divergence between two multivariate Gaussians can be computed in closed-form:

$$\begin{aligned} & \text{KL}[\mathcal{N}(\mathbf{u}|\mathbf{m}_0, \mathbf{\Sigma}_0)||\mathcal{N}(\mathbf{u}|\mathbf{m}_1, \mathbf{\Sigma}_1)] \\ &= \frac{1}{2} \left[\text{trace}(\mathbf{\Sigma}_1^{-1}\mathbf{\Sigma}_0) + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T \mathbf{\Sigma}_1^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) - D + \log \frac{|\mathbf{\Sigma}_1|}{|\mathbf{\Sigma}_0|} \right] \end{aligned}$$

The KL term

- The two terms in the lowerbound

$$\mathcal{L}[q] = \sum_{n=1}^N \mathbb{E}_{p(f_n|\mathbf{u})q(\mathbf{u})} [\log \mathcal{N}(y_n|f_n, \sigma^2)] - \text{KL}[q(\mathbf{u})||p(\mathbf{u})]$$

- The KL divergence between two multivariate Gaussians can be computed in closed-form:

$$\begin{aligned} & \text{KL}[\mathcal{N}(\mathbf{u}|\mathbf{m}_0, \Sigma_0)||\mathcal{N}(\mathbf{u}|\mathbf{m}_1, \Sigma_1)] \\ &= \frac{1}{2} \left[\text{trace}(\Sigma_1^{-1}\Sigma_0) + (\mu_1 - \mu_0)^T \Sigma_1^{-1}(\mu_1 - \mu_0) - D + \log \frac{|\Sigma_1|}{|\Sigma_0|} \right] \end{aligned}$$

- So for $q(\mathbf{u}) = \mathcal{N}(\mathbf{u}|\mathbf{m}_u, \mathbf{S}_u)$ and $p(\mathbf{u}) = \mathcal{N}(\mathbf{u}|\mathbf{0}, \mathbf{K}_{uu})$, we get

$$\text{KL}[q(\mathbf{u})||p(\mathbf{u})] = \frac{1}{2} \left[\text{trace}(\mathbf{K}_{uu}^{-1}\mathbf{S}_u) + \mathbf{m}_u^T \mathbf{K}_{uu}^{-1} \mathbf{m}_u - M + \log \frac{|\mathbf{K}_{uu}|}{|\mathbf{S}_u|} \right]$$

Combining everything

- We derived

$$\mathcal{L}[q] = \sum_{n=1}^N \mathbb{E}_{p(f_n|u)q(u)} \left[\log \mathcal{N}(y_n|f_n, \sigma^2) \right] - \text{KL}[q(u)||p(u)]$$

- ... and then showed that the first term simplifies to

$$\mathbb{E}_{p(f_n|u)q(u)} \left[\log \mathcal{N}(y_n|f_n, \sigma^2) \right] = \log \mathcal{N}(y_n|m_{f_n}, \sigma^2) df_n - \frac{1}{2\sigma} \sigma_{f_n}^2$$

- Combining yields

$$\begin{aligned} \mathcal{L}[q] &= \sum_{n=1}^N \left[\log \mathcal{N}(y_n|m_{f_n}, \sigma^2) df_n - \frac{1}{2\sigma} \sigma_{f_n}^2 \right] - \text{KL}[q(u)||p(u)] \\ &= \sum_{n=1}^N \log \mathcal{N}(y_n|m_{f_n}, \sigma^2) - \frac{1}{2\sigma} \sum_{n=1}^N \sigma_{f_n}^2 - \text{KL}[q(u)||p(u)] \\ &= \log \mathcal{N}(\mathbf{y}|\mathbf{m}_f, \sigma^2 \mathbf{I}) - \frac{1}{2\sigma} \text{trace}(\mathbf{S}_f) - \text{KL}[q(u)||p(u)] \end{aligned}$$

where

$$\begin{aligned} \mathbf{m}_f &= \mathbf{K}_{fu} \mathbf{K}_{uu}^{-1} \mathbf{m}_u \\ \mathbf{S}_f &= \mathbf{K}_{ff} + \mathbf{K}_{fu} \mathbf{K}_{uu}^{-1} (\mathbf{S}_u - \mathbf{K}_{uu}) \mathbf{K}_{uu}^{-1} \mathbf{K}_{uf} \end{aligned}$$

- We can optimize this bound wrt. the *variational parameters* \mathbf{m}_u and \mathbf{S}_u and wrt. the *hyperparameters* simultaneously!

One step more step

- It turns out we can optimize the bound wrt. \mathbf{m}_u and \mathbf{S}_u analytically

$$\begin{aligned}\mathcal{L}[q] &= \sum_{n=1}^N \left[\log \mathcal{N}(y_n | \mathbf{m}_{f_n}, \sigma^2) \right] - \text{KL}[q(\mathbf{u}) || p(\mathbf{u})] \\ &= \sum_{n=1}^N \log \mathcal{N}(y_n | \mathbf{m}_{f_n}, \sigma^2) - \frac{1}{2\sigma} \sum_{n=1}^N \sigma_{f_n}^2 - \text{KL}[q(\mathbf{u}) || p(\mathbf{u})] \\ &= \log \mathcal{N}(\mathbf{y} | \mathbf{m}_f, \sigma^2 \mathbf{I}) - \frac{1}{2\sigma} \text{trace}(\mathbf{S}_f) - \text{KL}[q(\mathbf{u}) || p(\mathbf{u})]\end{aligned}$$

- ... to get

$$\begin{aligned}\mathbf{S}_u^{-1} &= \frac{1}{\sigma^2} \mathbf{K}_{uu}^{-1} \mathbf{K}_{uf} \mathbf{K}_{fu} \mathbf{K}_{uu}^{-1} + \mathbf{K}_{uu}^{-1} \\ \mathbf{m}_u &= \frac{1}{\sigma^2} \mathbf{S}_u \mathbf{K}_{uu}^{-1} \mathbf{K}_{uf} \mathbf{y}\end{aligned}$$

- ... which leads to the *collapsed lowerbound*

$$\mathcal{L}[q] = \log \mathcal{N}(\mathbf{y} | \mathbf{0}, \mathbf{K}_{fu} \mathbf{K}_{uu} \mathbf{K}_{uf} + \sigma^2 \mathbf{I}) - \frac{1}{2\sigma} \text{trace}(\mathbf{K}_{ff} - \mathbf{K}_{fu} \mathbf{K}_{uu}^{-1} \mathbf{K}_{uf})$$

The big picture and how to use it in practice

Goal: fast way compute $p(\mathbf{f}|\mathbf{y})$

1. Choose a set of *inducing points* \mathbf{z}_i , where $u_i = f(\mathbf{z}_i)$

The big picture and how to use it in practice

Goal: fast way compute $p(\mathbf{f}|\mathbf{y})$

1. Choose a set of *inducing points* \mathbf{z}_i , where $u_i = f(\mathbf{z}_i)$
2. Optimize the *collapsed bound* wrt. our hyperparameters θ

$$\hat{\theta}^* = \arg \max_{\theta} \mathcal{L}_{\theta} [q] = \log \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K}_{fu}\mathbf{K}_{uu}\mathbf{K}_{uf} + \sigma^2\mathbf{I}) - \frac{1}{2\sigma} \text{trace}(\mathbf{K}_{ff} - \mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf})$$

The big picture and how to use it in practice

Goal: fast way compute $p(\mathbf{f}|\mathbf{y})$

1. Choose a set of *inducing points* \mathbf{z}_i , where $u_i = f(\mathbf{z}_i)$
2. Optimize the *collapsed bound* wrt. our hyperparameters θ

$$\hat{\theta}^* = \arg \max_{\theta} \mathcal{L}_{\theta} [q] = \log \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K}_{fu}\mathbf{K}_{uu}\mathbf{K}_{uf} + \sigma^2\mathbf{I}) - \frac{1}{2\sigma} \text{trace}(\mathbf{K}_{ff} - \mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf})$$

3. Compute the posterior mean and covariance for latent function values
 $q(\mathbf{u}) = \mathcal{N}(\mathbf{u}|\mathbf{m}_u, \mathbf{S}_u)$

$$\mathbf{S}_u^{-1} = \frac{1}{\sigma^2} \mathbf{K}_{uu}^{-1} \mathbf{K}_{uf} \mathbf{K}_{fu} \mathbf{K}_{uu}^{-1} + \mathbf{K}_{uu}^{-1}$$

$$\mathbf{m}_u = \frac{1}{\sigma^2} \mathbf{S}_u \mathbf{K}_{uu}^{-1} \mathbf{K}_{uf} \mathbf{y}$$

The big picture and how to use it in practice

Goal: fast way compute $p(\mathbf{f}|\mathbf{y})$

1. Choose a set of *inducing points* \mathbf{z}_i , where $u_i = f(\mathbf{z}_i)$
2. Optimize the *collapsed bound* wrt. our hyperparameters θ

$$\hat{\theta}^* = \arg \max_{\theta} \mathcal{L}_{\theta} [q] = \log \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K}_{fu}\mathbf{K}_{uu}\mathbf{K}_{uf} + \sigma^2\mathbf{I}) - \frac{1}{2\sigma} \text{trace}(\mathbf{K}_{ff} - \mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf})$$

3. Compute the posterior mean and covariance for latent function values
 $q(\mathbf{u}) = \mathcal{N}(\mathbf{u}|\mathbf{m}_u, \mathbf{S}_u)$

$$\mathbf{S}_u^{-1} = \frac{1}{\sigma^2} \mathbf{K}_{uu}^{-1} \mathbf{K}_{uf} \mathbf{K}_{fu} \mathbf{K}_{uu}^{-1} + \mathbf{K}_{uu}^{-1}$$

$$\mathbf{m}_u = \frac{1}{\sigma^2} \mathbf{S}_u \mathbf{K}_{uu}^{-1} \mathbf{K}_{uf} \mathbf{y}$$

4. Compute the approximate posterior distribution for $q(\mathbf{f}) = \mathcal{N}(\mathbf{f}|\mathbf{m}_f, \mathbf{S}_f)$

$$\mathbf{m}_f = \mathbf{K}_{fu} \mathbf{K}_{uu}^{-1} \mathbf{m}_u$$

$$\mathbf{S}_f = \mathbf{K}_{ff} + \mathbf{K}_{fu} \mathbf{K}_{uu}^{-1} (\mathbf{S}_u - \mathbf{K}_{uu}) \mathbf{K}_{uu}^{-1} \mathbf{K}_{uf}$$

The big picture and how to use it in practice

Goal: fast way compute $p(\mathbf{f}|\mathbf{y})$

1. Choose a set of *inducing points* \mathbf{z}_i , where $u_i = f(\mathbf{z}_i)$
2. Optimize the *collapsed bound* wrt. our hyperparameters θ

$$\hat{\theta}^* = \arg \max_{\theta} \mathcal{L}_{\theta} [q] = \log \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K}_{fu}\mathbf{K}_{uu}\mathbf{K}_{uf} + \sigma^2\mathbf{I}) - \frac{1}{2\sigma} \text{trace}(\mathbf{K}_{ff} - \mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf})$$

3. Compute the posterior mean and covariance for latent function values
 $q(\mathbf{u}) = \mathcal{N}(\mathbf{u}|\mathbf{m}_u, \mathbf{S}_u)$

$$\mathbf{S}_u^{-1} = \frac{1}{\sigma^2} \mathbf{K}_{uu}^{-1} \mathbf{K}_{uf} \mathbf{K}_{fu} \mathbf{K}_{uu}^{-1} + \mathbf{K}_{uu}^{-1}$$

$$\mathbf{m}_u = \frac{1}{\sigma^2} \mathbf{S}_u \mathbf{K}_{uu}^{-1} \mathbf{K}_{uf} \mathbf{y}$$

4. Compute the approximate posterior distribution for $q(\mathbf{f}) = \mathcal{N}(\mathbf{f}|\mathbf{m}_f, \mathbf{S}_f)$

$$\mathbf{m}_f = \mathbf{K}_{fu} \mathbf{K}_{uu}^{-1} \mathbf{m}_u$$

$$\mathbf{S}_f = \mathbf{K}_{ff} + \mathbf{K}_{fu} \mathbf{K}_{uu}^{-1} (\mathbf{S}_u - \mathbf{K}_{uu}) \mathbf{K}_{uu}^{-1} \mathbf{K}_{uf}$$

5. Use $p(\mathbf{f}|\mathbf{y}) \approx q(\mathbf{f})$ to make predictions

Airline delays dataset

- Flight arrival and departure times for every commercial flight in the USA from Jan. 2008 to April 2008
- 2 million flights: 700000 flight for training, 100000 for testing
- Target variable: Delay in minutes:
- $D = 8$ features: age of aircraft, flight distance, airtime, departure time, arrival time, day of the week, day of the month, month
- Squared exponential kernel with separate lengthscale $\ell_i > 0$ for each dimension

$$k(\mathbf{x}, \mathbf{x}') = \kappa^2 \exp \left[-\frac{1}{2} \sum_{i=1}^D \ell_i^{-1} |\mathbf{x}_i - \mathbf{x}'_i|^2 \right] + \tau$$

- Using $M = 1000$ inducing points

Hensman et al: Gaussian Processes for Big Data (2013)

Airline delays dataset

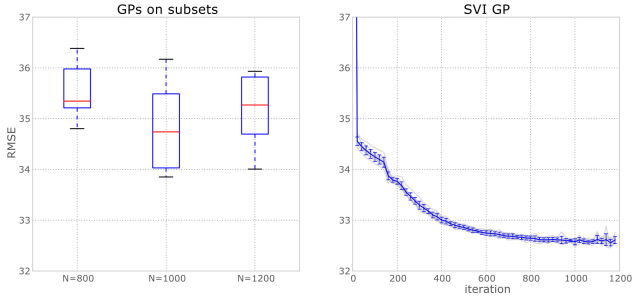


Figure 7: Root mean squared errors in predicting flight delays using information about the flight.

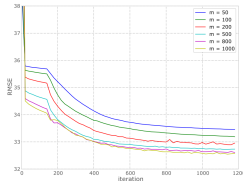


Figure 8: Root mean square errors for models with different numbers of inducing variables.