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

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

- 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 $\mathbb{D}\left[q||p\right]$

  3. Search for the distribution $q \in \mathcal{Q}$ that resembles the exact posterior p as close as possible as measured by $\mathbb{D}\left[q||p\right]$

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

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

Understanding the ELBO objective

- Suppose our model of interest is

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

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

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

- Re-writing the lower bound

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

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

Free-form vs fixed-form variational inference

Free-form variational inference

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

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

- Optimality condition for mean-field families

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

- 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

\+ Fast optimization

\- Requires model-specific derivations

\- 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 an the optimal functional form and simply *assume* some family of distributions $q_{\boldsymbol{\psi}}$ with parameters $\boldsymbol{\psi}$

- We often refer to $\boldsymbol{\psi}$ as *variational parameters*

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

  - Full-rank Gaussians: $\boldsymbol{m} \in \mathbb{R}^D$, $\boldsymbol{V} \in \mathbb{R}^{D\times D}$
$$q(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}, \boldsymbol{V}),$$

  - Low-rank Gaussians: $\boldsymbol{m} \in \mathbb{R}^D$, $\boldsymbol{B} \in \mathbb{R}^{D\times K}$, and $\boldsymbol{C} \in \mathbb{R}^{D\times D}$ is diagonal.
$$q(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}, \boldsymbol{B}\boldsymbol{B}^T + \boldsymbol{C}^2),$$

  - Mean-field Gaussians: $\boldsymbol{m} = \left[m_1, \ldots, m_D\right] \in \mathbb{R}^D$ and $\boldsymbol{v} = \left[v_1, \ldots, v_D\right] = \mathbb{R}^D$
$$q(\boldsymbol{w}) = \prod_{i=1}^{D} \mathcal{N}(w_i|m_i, v_i),$$

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

Fixed-form variational inference II

- *Fixed-form* variational inference as an alternative to *free-form*
- **Example**: Consider a model $p(\boldsymbol{y}, \boldsymbol{w})$, where $\boldsymbol{w} \in \mathbb{R}^D$
- Suppose we choose a full-rank Gaussian family

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

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

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

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

- Fitting the approximation using gradient-based methods

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

Hyperparameter estimation in VI

Dealing with hyperparameters for variational inference

- Consider a model with data $\boldsymbol{y}$, parameters $\boldsymbol{\theta}$, and hyperparameters $\boldsymbol{\xi}$

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

- **Examples**
  1. Linear regression
     $\boldsymbol{\theta} = \boldsymbol{w}$ would be the regression weights
     $\boldsymbol{\xi} = \{\alpha, \beta\}$ would be prior and noise precision.
  2. Gaussian process regression
     $\boldsymbol{\theta} = \boldsymbol{f}$ would be the latent function values
     $\boldsymbol{\xi} = \left\{\sigma^2, \kappa, \ell\right\}$ would be noise variance and kernel hyperparameters

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

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

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

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

Dealing with hyperparameters for variational inference

- Consider a model with data $\boldsymbol{y}$, parameters $\boldsymbol{\theta}$, and hyperparameters $\boldsymbol{\xi}$

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

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

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

- The ELBO is a lowerbound on the log marginal likelihood

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

- and hence, we can do hyperparameter estimation via ELBO

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

- **Key take-away**: We can optimize the ELBO wrt. *variational parameters* $\boldsymbol{\psi}$ and *hyperparameters* $\boldsymbol{\xi}$ simultaneously

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

Scaling Gaussian processes using variational inference

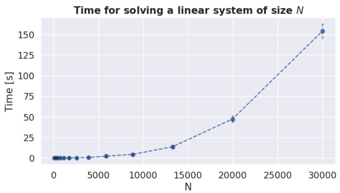
Speeding up Gaussian process inference

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

$$p(y^*|\boldsymbol{y}) = \mathcal{N}\left(y^* \middle| \mu_{y^*|\boldsymbol{y}}, \sigma^2_{y^*|\boldsymbol{y}}\right)$$

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

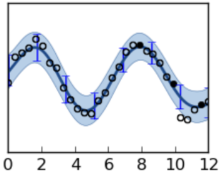
$$\sigma^2_{y^*|\boldsymbol{y}} = c - \boldsymbol{k}\left(\boldsymbol{K} + \beta^{-1}\boldsymbol{I}\right)^{-1}\boldsymbol{k}^T$$



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

- We introduce *inducing points* $\boldsymbol{z}_i$ for $i = 1, \ldots, M$ 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

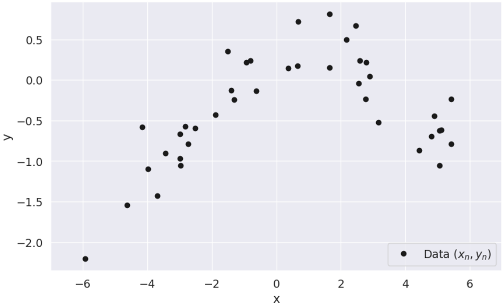
Introducing inducing points

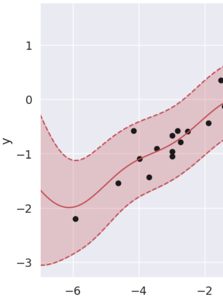
- Suppose we have dataset $\mathcal{D} = \{\boldsymbol{x}_n, y_n\}$, where $\boldsymbol{x}_n \in \mathbb{R}^D$ and $y_n \in \mathbb{R}$
$$y_n = f(\boldsymbol{x}_n) + e_n$$
- Assuming Gaussian noise and a GP prior, i.e. $f(\boldsymbol{x}) \sim \mathcal{GP}(\boldsymbol{0}, k(\boldsymbol{x}, \boldsymbol{x}'))$ yields the joint distribution
$$p(\boldsymbol{y}, \boldsymbol{f}) = p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{f}, \sigma^2\boldsymbol{I})\mathcal{N}(\boldsymbol{f}|\boldsymbol{0}, \boldsymbol{K}_{ff})$$
- **Goal**: fast way compute $p(\boldsymbol{f}|\boldsymbol{y})$
$$f_n = f(\boldsymbol{x}_n)$$
- We introduce M *inducing points* $\boldsymbol{z}_i \in \mathbb{R}^D$ for $m = 1, \ldots, M$ such that
$$u_i = f(\boldsymbol{z}_i)$$
- Extended joint distribution for data $\boldsymbol{y}$ and latent function values for both $\boldsymbol{f}$ and $\boldsymbol{u}$
$$p(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{u}) = p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f}|\boldsymbol{u})p(\boldsymbol{u}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{f}, \sigma^2\boldsymbol{I})\mathcal{N}(\boldsymbol{f}|\boldsymbol{K}_{fu}\boldsymbol{K}_{uu}^{-1}\boldsymbol{u}, \boldsymbol{K}_{ff} - \boldsymbol{K}_{fu}\boldsymbol{K}_{uu}^{-1}\boldsymbol{K}_{fu})\mathcal{N}(\boldsymbol{u}|\boldsymbol{0}, \boldsymbol{K}_{uu})$$
- We can always marginalize to get the original model back
$$p(\boldsymbol{y}, \boldsymbol{f}) = \int p(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{u})\mathrm{d}\boldsymbol{u}$$

Intuition

- Our extended joint distribution

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





- Can we find a set of *inducing points* $\{\boldsymbol{z}_i\}_{i=1}^{M}$ and associated function values $u_i = f(\boldsymbol{z}_i)$ 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

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

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

- A clever choice for the variational family

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

  where

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

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

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

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

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

The approximate posterior distribution

- Our choice of variational family:

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

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

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

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

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

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

Calculating the ELBO

We substitute our model and variational approximation into the ELBO

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

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

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

Calculating the first term

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

$$= \int\int p(f_n|\boldsymbol{u})q(\boldsymbol{u})\mathrm{d}\boldsymbol{u}\log\mathcal{N}(y_n|f_n,\sigma^2)\mathrm{d}f_n \qquad \text{(Recall: } q(\boldsymbol{f}) = \mathcal{N}(\boldsymbol{f}|\boldsymbol{m}_f,\boldsymbol{S}_f))$$

$$= \int\mathcal{N}\left(f_n|m_{f_n},\sigma^2_{f_n}\right)\log\mathcal{N}(y_n|f_n,\sigma^2)\mathrm{d}f_n$$

$$= \mathbb{E}_{\mathcal{N}\left(f_n|m_{f_n},\sigma^2_{f_n}\right)}\left[\log\mathcal{N}(y_n|f_n,\sigma^2)\right]$$

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

$$= \mathbb{E}_{\mathcal{N}\left(f_n|m_{f_n},\sigma^2_{f_n}\right)}\left[-\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma}(y_n^2 + f_n^2 - 2y_nf_n)\right]$$

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

$$= -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma}(y_n^2 + m_{f_n}^2 + \sigma^2_{f_n} - 2y_nm_{f_n})$$

$$= -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma}(y_n^2 + m_{f_n}^2 - 2y_nm_{f_n}) - \frac{1}{2\sigma}\sigma^2_{f_n}$$

$$= -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma}(y_n - m_{f_n})^2 - \frac{1}{2\sigma}\sigma^2_{f_n}$$

$$= \log\mathcal{N}(y_n|m_{f_n},\sigma^2)\mathrm{d}f_n - \frac{1}{2\sigma}\sigma^2_{f_n}$$

The KL term

- The two terms in the lowerbound

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

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

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

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

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

Combining everything

- We derived

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

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

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

- Combining yields

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

  where

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

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

One step more step

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

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

- ... to get

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

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

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

The big picture and how to use it in practice

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

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

2. Optimize the *collapsed bound* wrt. our hyperparameters $\boldsymbol{\theta}$

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

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

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

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

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

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

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

5. Use $p(\boldsymbol{f}|\boldsymbol{y}) \approx q(\boldsymbol{f})$ 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

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

- Using $M = 1000$ inducing points

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

Airline delays dataset

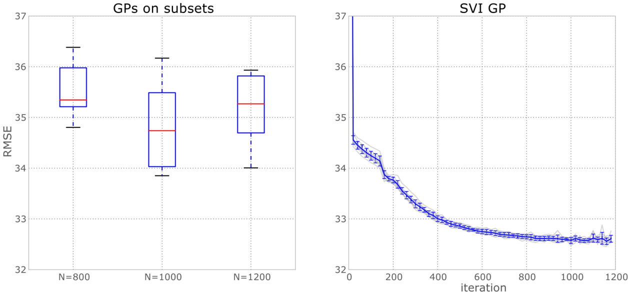


Figure 7: Root mean squared errors in predicting flight delays using information about the flight.

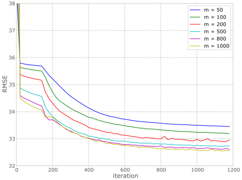


Figure 8: Root mean square errors for models with different numbers of inducing variables.