Stochastic variational Gaussian process regression

Constructing the ELBO

We are interested in maximizing

$$\log p_\theta(\boldsymbol{y}) = \log \int p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f})d\boldsymbol{f} \tag{1}$$

with respect to some set of generative parameters θ (note that we will omit the $\cdot_\theta$ subscript for notational clarity).

We start by introducing a set of inducing points $\boldsymbol{z} = \{z_m\}_{m=1}^M$ with corresponding function values (inducing variables) $\boldsymbol{u} = \{u_m\}_{m=1}^M$ which have a prior distribution $p(\boldsymbol{u}) = \mathcal{N}(0, \boldsymbol{K}_{uu})$ and a posterior $p(\boldsymbol{u}|\boldsymbol{y})$. We then introduce a variational distribution $q_\phi(\boldsymbol{f}, \boldsymbol{u}) = p(\boldsymbol{f}|\boldsymbol{u})q_\phi(\boldsymbol{u})$, where

$$q_\phi(\boldsymbol{u}) = \mathcal{N}(\boldsymbol{u}; \boldsymbol{m}, \boldsymbol{S}). \tag{2}$$

Here, we have introduced the variational parameters $\phi = \{\boldsymbol{m}, \boldsymbol{S}\}$, and we will drop the $\cdot_\phi$ subscript for notational clarity in what follows.

We start by constructing a variational lower bound (ELBO):

$$\log p(\boldsymbol{y}) \geq \mathbb{E}_{q(\boldsymbol{u},\boldsymbol{f})}\left[\log p(\boldsymbol{y}|\boldsymbol{u}, \boldsymbol{f})\right] - KL\left[q(\boldsymbol{u}, \boldsymbol{f})||p(\boldsymbol{u}, \boldsymbol{f})\right]. \tag{3}$$

Importantly, maximizing this ELBO is equivalent to minimizing

$$KL\left[q(\boldsymbol{f}, \boldsymbol{u})||p(\boldsymbol{f}, \boldsymbol{u}|\boldsymbol{y})\right] \tag{4}$$

with respect to the variational parameters ϕ, and this will be important when we compute our predictive distribution in Equation 23.

We then note that $p(\boldsymbol{y}|\boldsymbol{u}, \boldsymbol{f}) = p(\boldsymbol{y}|\boldsymbol{f})$. Defining a distribution

$$q(\boldsymbol{f}) = \int p(\boldsymbol{f}|\boldsymbol{u})q(\boldsymbol{u})d\boldsymbol{u}, \tag{5}$$

this allows us to integrate out $\boldsymbol{u}$ in the expectation:

$$\mathbb{E}_{q(\boldsymbol{u},\boldsymbol{f})}\left[\log p(\boldsymbol{y}|\boldsymbol{u}, \boldsymbol{f})\right] = \mathbb{E}_{q(\boldsymbol{f})}\left[\log p(\boldsymbol{y}|\boldsymbol{f})\right]. \tag{6}$$

We also note that

$$\begin{aligned} KL\left[q(\boldsymbol{u}, \boldsymbol{f})||p(\boldsymbol{u}, \boldsymbol{f})\right] &= \int q(\boldsymbol{u}, \boldsymbol{f}) \log \frac{q(\boldsymbol{u}, \boldsymbol{f})}{p(\boldsymbol{u}, \boldsymbol{f})} d\boldsymbol{f}d\boldsymbol{u} && (7)\\ &= \int q(\boldsymbol{u})p(\boldsymbol{f}|\boldsymbol{u}) \log \frac{q(\boldsymbol{u})p(\boldsymbol{f}|\boldsymbol{u})}{p(\boldsymbol{u})p(\boldsymbol{f}|\boldsymbol{u})} d\boldsymbol{f}d\boldsymbol{u} && (8)\\ &= \int q(\boldsymbol{u}) \log \frac{q(\boldsymbol{u})}{p(\boldsymbol{u})} d\boldsymbol{u} && (9)\\ &= KL\left[q(\boldsymbol{u})||p(\boldsymbol{u})\right]. && (10) \end{aligned}$$

Putting Equation 6 and Equation 7 together, our ELBO simplifies to

$$\log p(\boldsymbol{y}) \geq \mathbb{E}_{q(\boldsymbol{f})}\left[\log p(\boldsymbol{y}|\boldsymbol{f})\right] - KL\left[q(\boldsymbol{u})||p(\boldsymbol{u})\right]. \tag{11}$$

Computing the likelihood term

We now turn our attention to the likelihood term $\mathbb{E}_{q(\boldsymbol{f})}\left[\log p(\boldsymbol{y}|\boldsymbol{f})\right]$. To evaluate this, we need $q(\boldsymbol{f})$. We note that everything is jointly Gaussian with a prior given by our kernel

$$p\left(\begin{bmatrix}\boldsymbol{f}\\ \boldsymbol{u}\end{bmatrix}\right) = \mathcal{N}\left(0, \begin{bmatrix}\boldsymbol{K}_{ff} & \boldsymbol{K}_{fu}\\ \boldsymbol{K}_{uf} & \boldsymbol{K}_{uu}\end{bmatrix}\right). \tag{12}$$

From standard Gaussian identities, this gives rise to a conditional distribution

$$p(\boldsymbol{f}|\boldsymbol{u}) = \mathcal{N}(\boldsymbol{f}; \boldsymbol{K}_{fu}\boldsymbol{K}_{uu}^{-1}\boldsymbol{u}, \boldsymbol{K}_{ff} - \boldsymbol{K}_{fu}\boldsymbol{K}_{uu}^{-1}\boldsymbol{K}_{uf}). \tag{13}$$

We can now compute the marginal distribution $q(\boldsymbol{f})$ analytically:

$$q(\boldsymbol{f}) = \int \mathcal{N}(\boldsymbol{f}; \boldsymbol{K}_{fu}\boldsymbol{K}_{uu}^{-1}\boldsymbol{u}, \boldsymbol{K}_{ff} - \boldsymbol{K}_{fu}\boldsymbol{K}_{uu}^{-1}\boldsymbol{K}_{uf})\mathcal{N}(\boldsymbol{u}; \boldsymbol{m}, \boldsymbol{S})d\boldsymbol{u} \tag{14}$$

$$= \mathcal{N}(\boldsymbol{f}; \boldsymbol{K}_{fu}\boldsymbol{K}_{uu}^{-1}\boldsymbol{m}, \boldsymbol{K}_{fu}\boldsymbol{K}_{uu}^{-1}\boldsymbol{S}\boldsymbol{K}_{uu}^{-1}\boldsymbol{K}_{uf} + \boldsymbol{K}_{ff} - \boldsymbol{K}_{fu}\boldsymbol{K}_{uu}^{-1}\boldsymbol{K}_{uf}). \tag{15}$$

Assuming that the likelihood factorizes (i.e. $p(\boldsymbol{y}|\boldsymbol{f}) = \prod_i p(y_i|f_i)$), we can further simplify the likelihood term by writing

$$\mathbb{E}_{q(\boldsymbol{f})}\left[\log p(\boldsymbol{y}|\boldsymbol{f})\right] = \sum_i \mathbb{E}_{q(f_i)}\left[\log p(y_i|f_i)\right]. \tag{16}$$

We thus only need the marginal distributions $q(f_i)$, which can be computed efficiently (Equation 28). We can then evaluate the 1-dimensional expectations over f_i using either Monte Carlo samples from $q(f_i)$ or with Gauss-Hermite quadrature which has been described extensively elsewhere.

Computing the KL term

We now address the KL term $KL\left[q(\boldsymbol{u})||p(\boldsymbol{u})\right]$. This has a closed-form expression given by

$$KL\left[\mathcal{N}(\boldsymbol{u}; \boldsymbol{m}, \boldsymbol{S})||\mathcal{N}(\boldsymbol{u}; 0, \boldsymbol{K}_{uu})\right] = 0.5\left(tr(\boldsymbol{K}_{uu}^{-1}\boldsymbol{S}) + \boldsymbol{m}^T\boldsymbol{K}_{uu}^{-1}\boldsymbol{m} - k + \log|\boldsymbol{K}_{uu}| - \log|\boldsymbol{S}|\right). \tag{17}$$

Together with Equation 16, this allows us to evaluate the ELBO in a differentiable manner which in turn provides a framework for optimizing the generative parameters θ of our model $(\ell, \sigma^2_{signal}, \sigma^2_{noise}, ...)$.

Inference

To perform inference, we are also interested in the distribution over function values $p(\boldsymbol{f}^*)$ at some set of test locations $\boldsymbol{x}^*$. Recalling that our variational inference procedure gives rise to $q(\boldsymbol{u}, \boldsymbol{f}) \approx p(\boldsymbol{u}, \boldsymbol{f}|\boldsymbol{y})$, we can write

$$p(\boldsymbol{f}^*|\boldsymbol{y}) = \int p(\boldsymbol{f}^*, \boldsymbol{f}, \boldsymbol{u}|\boldsymbol{y})d\boldsymbol{f}d\boldsymbol{u} \tag{18}$$

$$= \int p(\boldsymbol{f}^*|\boldsymbol{f}\boldsymbol{u})p(\boldsymbol{f}, \boldsymbol{u}|\boldsymbol{y})d\boldsymbol{f}d\boldsymbol{u} \tag{19}$$

$$\approx \int p(\boldsymbol{f}^*|\boldsymbol{f}\boldsymbol{u})q(\boldsymbol{f}, \boldsymbol{u})d\boldsymbol{f}d\boldsymbol{u} \tag{20}$$

$$= \int p(\boldsymbol{f}^*|\boldsymbol{f}\boldsymbol{u})p(\boldsymbol{f}|\boldsymbol{u})q(\boldsymbol{u})d\boldsymbol{f}d\boldsymbol{u} \tag{21}$$

$$= \int p(\boldsymbol{f}^*, \boldsymbol{f}|\boldsymbol{u})q(\boldsymbol{u})d\boldsymbol{f}d\boldsymbol{u} \tag{22}$$

$$= \int p(\boldsymbol{f}^*|\boldsymbol{u})q(\boldsymbol{u})d\boldsymbol{u}. \tag{23}$$

Conveniently, we already know the form of this marginal distribution from Equation 15. It is interesting to note that this predictive distribution does not depend on the data $\boldsymbol{y}$ but is entirely summarized by the parameters ϕ of our variational distribution $q(\boldsymbol{u})$.

The 'whitened' parameterization

We can simplify the evaluation of our ELBO by parameterizing $q(\boldsymbol{u}) = \mathcal{N}(\boldsymbol{u}; \boldsymbol{K}_{uu}^{\frac{1}{2}}\boldsymbol{m}, \boldsymbol{K}_{uu}^{\frac{1}{2}}\boldsymbol{S}\left(\boldsymbol{K}_{uu}^T\right)^{\frac{1}{2}})$. Here, $\boldsymbol{K}_{uu}^{\frac{1}{2}}$ can be any matrix factorization with the usual choice being the lower triangular cholesky factor $\boldsymbol{L}_{uu}$ s.t. $\boldsymbol{L}_{uu}\boldsymbol{L}_{uu}^T = \boldsymbol{K}_{uu}$. Since the KL divergence is invariant to the affine transformation $\boldsymbol{x} \to \boldsymbol{L}^{-1}\boldsymbol{x}$, we can rewrite KL term from

$$KL\left[\mathcal{N}(\boldsymbol{u}; \boldsymbol{L}\boldsymbol{m}, \boldsymbol{L}\boldsymbol{S}\boldsymbol{L}^T)||\mathcal{N}(\boldsymbol{u}; 0, \boldsymbol{K}_{uu})\right] \tag{24}$$

to

$$KL\left[\mathcal{N}(\boldsymbol{u};\boldsymbol{m},\boldsymbol{S})||\mathcal{N}(\boldsymbol{u};0,\boldsymbol{I})\right] = 0.5\left(tr(\boldsymbol{S}) + \boldsymbol{m}^T\boldsymbol{m} - k - \log|\boldsymbol{S}|\right), \tag{25}$$

which gets rid of all terms involving $\boldsymbol{L}_{uu}$.

For this whitened parameterization, the distribution $q(\boldsymbol{f})$ is given by (Equation 15):

$$q(\boldsymbol{f}) = \mathcal{N}(\boldsymbol{f}; \boldsymbol{K}_{fu}(\boldsymbol{L}^{-1})^T\boldsymbol{m}, \boldsymbol{K}_{fu}(\boldsymbol{L}^{-1})^T\boldsymbol{S}\boldsymbol{L}^{-1}\boldsymbol{K}_{uf} + \boldsymbol{K}_{ff} - \boldsymbol{K}_{fu}\boldsymbol{K}_{uu}^{-1}\boldsymbol{K}_{uf}). \tag{26}$$

We then write $\boldsymbol{S} = \boldsymbol{S}_L\boldsymbol{S}_L^T$, denote $\boldsymbol{\Psi} = \boldsymbol{K}_{fu}(\boldsymbol{L}^{-1})^T$, and note that for any matrix $\boldsymbol{A}$

$$(\boldsymbol{A}\boldsymbol{A})_{ii} = \sum_j \boldsymbol{A}_{ij}\boldsymbol{A}_{ji}^T = \sum_j \boldsymbol{A}_{ij}^2. \tag{27}$$

This allows us to compute the marginal variance $\boldsymbol{\Sigma}_{ii}$ of $q(\boldsymbol{f})$ as

$$\boldsymbol{\Sigma}_{ii} = \boldsymbol{K}_{ff,ii} + \sum_j(\boldsymbol{\Psi}\boldsymbol{S}_L)_{ij}^2 - \sum_j \boldsymbol{\Psi}_{ij}^2. \tag{28}$$

Additionally, the mean parameters $\boldsymbol{\mu}_i$ are easily computed as $(\boldsymbol{\Psi}\boldsymbol{m})_i$.

Together, $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_{ii}$ fully specify the marginal distributions $q(f_i)$. We also recall from Equation 16 that this is all we need to compute the likelihood term in our ELBO, while the KL term is easily computed according to Equation 25. Finally, these computations are all differentiable, which allows us to perform gradient-based optimization of both the generative parameters θ and the variational parameters ϕ. After convergence, inference is performed according to Equation 23.