

# Notes: Generalised Variational Inference in Function Spaces for Image Data

**Jian Shu (James) Wu**

*Department of Computer Science  
University College London*

JIAN.WU.22@UCL.AC.UK

**Editor:**

## 1. Gaussian Processes for Classification

Notes taken from chapter 4 from Matthews (2017).

For Gaussian process regression (GPR), a class of models is defined:

$$f \sim \mathcal{GP}(0, K(\theta))$$

where  $f : X \rightarrow \mathbb{R}$ , mapping to the set of real numbers  $\mathbb{R}$  and  $K$  is the covariance function  $K : X \times X \rightarrow \mathbb{R}$  parameterised by  $\theta$ .

For binary Gaussian process classification (GPC), a mapping is defined:

$$g : \mathbb{R} \rightarrow [0, 1]$$

transforming a value on the real line to the unit interval to represent a probability. A bernoulli random variable  $\mathcal{B}$  can be defined such that:

$$f_c \sim \mathcal{B}(g(f))$$

where  $f_c : X \rightarrow \{0, 1\}$ , the desired binary classifier.

For multiclass classification of  $J$  different classes, models are defined:

$$f^{(j)} \sim \mathcal{GP}(0, K(\theta^{(j)}))$$

where  $j = 1, \dots, J$ , defining  $J$  i.i.d. Gaussian processes. Concatenating  $\mathbf{f} = [f_1 \dots f_J]^T$ , the classification operation can be defined:

$$\mathbf{f}_c \sim \text{Cat}(\mathcal{S}(\mathbf{f}))$$

where  $\mathbf{f}_c : X \rightarrow \{0, \dots, J\}$ , the desired multiclass classifier.  $\mathcal{S} : \mathbb{R}^J \rightarrow \Delta(J)$ , a mapping from a  $J$  dimensional real vector to a  $J$  dimensional probability simplex.  $\text{Cat}$  is the categorical distribution (generalisation of Bernoulli distribution for Categorical Data).

There are different possible choices for  $\mathcal{S}$ . The multiclass generalisation of the logit likelihood:

$$\mathcal{S}_{softmax}(\mathbf{f})_i = \frac{\exp(f^{(i)})}{\sum_{j=1}^J \exp(f^{(j)})}$$

The robust max function:

$$\mathcal{S}_{robust}^{(\epsilon)}(\mathbf{f})_i = \begin{cases} 1 - \epsilon, & \text{if } i = \arg \max(\mathbf{f}) \\ \epsilon, & \text{otherwise} \end{cases}$$

taking class label of the maximum value with probability of  $1 - \epsilon$  and probability  $\epsilon$  of picking one of the other classes uniformly at random, where  $\epsilon$  is chosen. This formulation provides robustness to outliers, as it only considers the ranking of the GPR models for each class.

A benefit of the robust max function is that the variational expectation is analytically tractable with respect to the normal CDF ( $q(\mathbf{f}) = \mathcal{N}(\mu, C)$ ,  $\mathbf{f} \in \mathbb{R}^J$ ) and one dimensional quadrature ( $\mathcal{S}_{robust}^{(\epsilon)}(\mathbf{f})_i \in \mathbb{R}$ ):

$$\int_{\mathbb{R}^J} q(\mathbf{f}) \log(\mathcal{S}_{robust}^{(\epsilon)}(\mathbf{f})_y) d\mathbf{f} = \log(1 - \epsilon)S + \log\left(\frac{\epsilon}{J - 1}\right)(1 - S)$$

where  $S$  is the probability that the function value corresponding to observed class  $y$  is larger than the other function values at that point:

$$S = \mathbb{E}_{\mathbf{f}^{(y)} \sim \mathcal{N}(\mathbf{f}^{(y)} | \mu^{(y)}, C^{(y)})} \left[ \prod_{i \neq y} \phi\left(\frac{\mathbf{f}^{(y)} - \mu^{(i)}}{\sqrt{C^{(i)}}}\right) \right]$$

where  $\phi$  is the standard normal CDF. This one dimensional integral can be evaluated using Gauss-Hermite quadrature.

## References

Alexander Graeme de Garis Matthews. *Scalable Gaussian process inference using variational methods*. PhD thesis, University of Cambridge, 2017.