Technical Report

Pairwise Judgements and Absolute Ratings with Gaussian Process Priors

- a collection of technical details

Revision: January 2014 (svn:1336)

Bjørn Sand Jensen and Jens Brehm Nielsen
{bjje,jenb}@dtu.dk

Technical University of Denmark
DTU Compute (formerly DTU Informatics)
Department of Applied Mathematics and Computer Science
2800 Lyngby
Denmark

History

Date	Version	Comment	Citation
1. Sep 2011	September 2011	Supplementary material to [11]	Bjørn Sand Jensen and Jens Brehm Nielsen, Pairwise Judgements and Absolute Ratings with Gaussian Process Priors, Technical Report, DTU Informatics, September 2011.
6. Februay 2014	January 20114	Reworked structure, some details on absolute case and inference	Bjørn Sand Jensen and Jens Brehm Nielsen, Pairwise Judgements and Absolute Ratings with Gaussian Process Priors - a collection of technical details, Technical Report, DTU Compute, January 2014.
Scheduled		Updated details on the absolute case (incl. truncated Gaussian) and examples	

Table 1: Document History

Please report any bugs and typos to either: bjje@dtu.dk or jenb@dtu.dk.

Contents

1 Notation

$\mathbb{R}$	The reals.
$\mathbb{Z}$	Integers.
$\mathbb{X}$	Domain of the input variable / input space.
$\mathbb{Y}$	Domain of the output variable / output space.
$\mathcal{X}$	A set. Typically of input instances from $\mathbb{X}$
$\mathcal{Y}$	A set. Typically of outputs from $\mathbb{Y}$
$\mathcal{D}$	A set. Typically a joint collection of inputs and outputs, so a dataset.
y, x $\mathbf{y}$, $\mathbf{x}$	A (random) variable or instance. A (random) vector, if not started otherwise a collum vector of dimension Dx1. If dealing with instances we will typically use non-bold notation for the variable even though it might be multi-dimensional in nature.
k	A particular experiment (note: always subscript)
n	total number of inputs, typically in $\mathcal{X}$ (not necessarily outputs)
m	total number of experiments/outputs typically in $\mathcal{Y}$ (not necessarily same number of inputs)
$P(x)$	A probability, $P(x) \in [0;1]$
$p(x\|\boldsymbol{\theta})$	A probability density (sometimes also likelihood) depended on parameters in $\boldsymbol{\theta}$
$p(x\|y)$	A conditional probability density where x is condition on another random variable
$\mathbb{E}_{p(x)}(x)$	Expectation under $p(x)$
$\mathbb{V}(x)$	Second order moment of x
$k(\mathbf{x}, \mathbf{x}')$	Covariance function evaluated for two inputs.
$\mathbf{k}(\mathbf{x}_t)$ or $\mathbf{k}_t$	
$\mathbf{K}$ or $\mathbf{K}(X, X)$	Covariance or Kernel Matrix
$\mathbf{x}_t$	A test input.
f	An abstract function
$f(\mathbf{x})$ or $\mathbf{f}$	The function evaluated at $\mathbf{x}$. If multiple inputs the vector, $\mathbf{f}$, contains multiple evaluations, i.e., $\mathbf{f} = [f(x_1), f(x_2), .., f(x_l)]^\top$
$\mathbf{f}_t$	The (predicted) functional value of x_t
$S(x)$	Entropy of the random variable x
∇	First derivative, Gradient
$\nabla\nabla$	Second derivative, Hessian
ω	Different use. Sometimes to indicate a specific parameter.
$\boldsymbol{\theta}$	A set of parameters or hyper-parameters
ν	Precision parameter describing subject consistency
$\mu(f)$	Parametrization of a distribution's mean
$\alpha(f), \beta(f)$	A parametrization of a distribution's shape parameter based on f
$\mathcal{N}(x\|\mu, \Sigma)$ $\mathcal{N}(\mu, \Sigma)$	Normal/Gaussian Distribution
$\boldsymbol{\Phi}(z)$	Cumulative Gaussian, with mean 0 and std. dev. 1
$\text{Beta}(\alpha, \beta)$	A standard two parameter beta distribution
$\text{B}(\alpha, \beta)$	beta function
$\log(x)$, $\ln(x)$	Natural logarithm
$\log_2(x)$	Base 2 logarithm

2 Introduction

This technical report contains a short overview and mathematical details in connection with preference learning using Gaussian Processes priors serving a coherent collection of supplementary material for a number of publications (by the same authors) on the subject, e.g. [11, 16, 10]

The report focuses on the observations, i.e. likelihood functions, placed in a standard Gaussian process framework. We thus consider different likelihoods and derive the required mathematical results needed for doing feasible Bayesian inference.

The main objective is to elicit human preference by robust and flexible statistical modeling for which apply a Bayesian framework:

$$p\left(\mathbf{f}, \boldsymbol{\theta} | \mathcal{Y}, \mathcal{X}\right) = \frac{p\left(\boldsymbol{\theta}_{\mathcal{GP}}\right) p\left(\mathbf{f} | \boldsymbol{\theta}_{\mathcal{GP}}, \mathcal{X}\right) p\left(\boldsymbol{\theta}_{\mathcal{L}}\right) p\left(\mathcal{Y} | \mathbf{f}, \boldsymbol{\theta}_{\mathcal{L}}\right)}{p\left(\mathcal{Y} | \mathcal{X}\right)}, \tag{1}$$

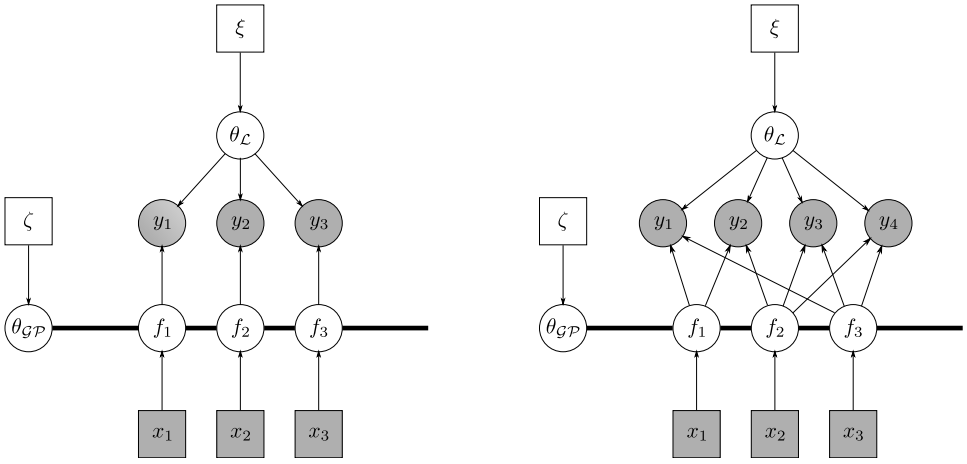
The Likelihood
The likelihood $p\left(\mathcal{Y} | \mathbf{f}, \boldsymbol{\theta}_{\mathcal{L}}\right)$ is the main focus of this report. We consider different likelihoods relevant to preference elicitation classified by the paradigm they implies

- Absolute (Section 3.2) - see Figure 1(a)
  - Continuous
    - Normal (standard regression case, not considered in detail)
    - Truncated Gaussian
    - Beta
    - (others exists, but not considered here)
  - Discrete
    - Probit (classification, not considered in detail)
    - (others exists, but not considered here)
- Relative (Section 3.1) - see Figure 1(b)
  - Continuous
    - Beta
    - (others exists, but not considered here)
  - Discrete
    - Probit (Thurstone)
    - (others exists, but not considered here)

The Priors
The particular choice of priors - in particular $p\left(\mathbf{f} | \boldsymbol{\theta}_{\mathcal{GP}}, \mathcal{X}\right)$ - is shortly described in Section 4.

The Inference - posterior
An conceptual overview of inference in GP based models is given in Section 5 with focus on the posterior $p\left(\mathbf{f}, \boldsymbol{\theta} | \mathcal{Y}, \mathcal{X}\right)$.

The Inference - predictions
The predictive distribution for for unseen instances, $p(\mathcal{Y}_t | \mathcal{X}_t)$ is considered in Section 6 for the various likelihood models models.



(a) Plate notation (general).

(b) Chain notation (for the pairwise case).

Figure 1: Graphical models of an absolute model (a) and a relative model (b) with all variables and parameters included. Round boxes indicate random variables, which are observed if the box is shaded and are hidden otherwise. Square boxes indicate deterministic variables and parameters, where a shading indicates that it is an explanatory variables and a non-shading indicates that the parameter is fixed. The thick line indicates a Gaussian field. Note, the special nature of the pairwise framework, in which each observed response depends on two functional values and thus two inputs.

3 Likelihoods

In a Bayesian framework the likelihood $p\left(\mathcal{Y}|\mathbf{f},\boldsymbol{\theta}_{\mathcal{L}}\right)$ gives an unique opportunity to transform any prior beliefs or assumptions concerning a given model f into posterior ditto through Bayes relation (see e.g. [1] or [13]). In this preference elicitation framework we assume that the likelihood factorizes over observations, i.e., each observation is drawn independently. This enables us to apply the product rule and construct specific *likelihood functions* $\mathcal{L}$ defined as the conditional density of one particular observation given the model. Consequently, the likelihood of a given dataset can be expressed by applying the product rule and multiplying together all the likelihood functions at the individual observations.

In this section we present different likelihood functions relevant for preference learning (pairwise) and attribute assessments (absolute). In section 3.1 we presents two likelihood functions suitable for pairwise responses and in section 3.2 we present a novel likelihood function for absolute bounded responses, e.g., a rating scale. Details concerning the likelihood functions can be found in appendix A.

3.1 Relative

Pairwise comparisons are easily motivated from a cognitive perspective, due to the inherent relative nature of perceptual evaluation, yet pairwise comparisons can be considered more broadly. Regardless, it is usually possible to describe any aspect of a pairwise comparison, such as preference, real difference, or perceived similarity in terms of a latent function [19].

In the following we model user preference of two distinct inputs, $u \in \mathcal{X}$ and $v \in \mathcal{X}$, in terms of the difference between two functional values, $f(u)$ and $f(v)$. This implies a function $f : \mathcal{X} \to \mathbb{R}$, which defines an internal, but latent absolute preference.

The general setup is as follows: We consider n distinct inputs $x_i \in \mathcal{X}$ denoted $\mathcal{X} = \{x_i | i = 1, ..., n\}$, and a set of m responses on pairwise comparisons between any two inputs in $\mathcal{X}$, denoted by

$$\mathcal{Y} = \{(y_k; u_k, v_k) | k = 1, ..., m\}, \tag{2}$$

where $y_k \in \mathbb{Y}$. The inputs $u_k \in \mathcal{X}$ and $v_k \in \mathcal{X}$ are option one and two in the k'th pairwise comparison, respectively. When considering preference learning the main idea is to adapt an efficient, i.e. fast and robust, learning framework. In this section we therefore present two different types of response variables:

- **binary** where $y_k = d_k, d_k \in \{-1, 1\}$
- **continuous and bounded** where $y_k = \pi_k, \pi_k \in]0, 1[$.

In both cases we consider y a stochastic variable, informally implying the definition of the conditional density as $p\left(y_k | f\left(u_k\right), f\left(v_k\right)\right)$, denoted by $p(y_k|\mathbf{f}_k)$ with $\mathbf{f}_k = \left[f\left(u_k\right), f\left(v_k\right)\right]^\top$.

3.1.1 Relative - Discrete - Binary Response (Probit)

When restricting the response variable to be a discrete, two-alternatives, forced choice, paired-comparison between the two presented options, we define the response variable as $d_k \in \{-1, 1\}$. A preference for either u_k or v_k is indicated by -1 or $+1$, respectively.

When considering noise on the forced decisions the resulting random variable can be modeled by a classic choice model such as the Logit or Probit [2, chapter 6]. Here, we restrict ourself to the Probit model mainly for analytical reasons.

Given a function, f, we can define the likelihood of observing a discrete choice $d_k = \{-1, 1\}$ directly as the conditional density. `Probit likelihood`

$$\mathcal{L}_{bin} \equiv p\left(d_k|\mathbf{f}_k\right) = \Phi\left(d_k \frac{f\left(v_k\right) - f\left(u_k\right)}{\sqrt{2}\sigma}\right), \tag{3}$$

where $\Phi(x)$ is the cumulative Gaussian (with zero mean and unity variance). This classic Probit likelihood is by no means a new invention and can be dated back to Thurstone and his fundamental definition of *The Law of Comparative Judgment*[19]. However, it was first considered with GPs in [7] and later in e.g. [4] and [6].

Example:
For completeness we show a few examples of the likelihood model as a function of the difference $\Delta(f)$ between $f(u)$ and $f(v)$.

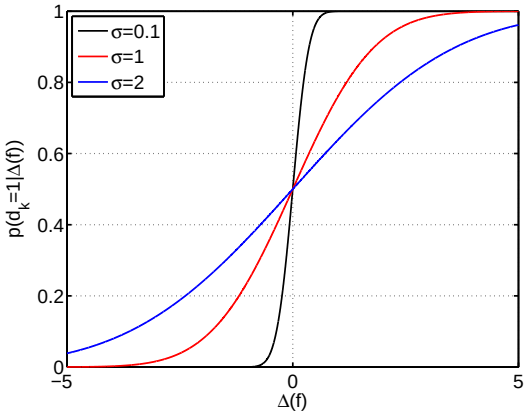


Figure 2: The Probit likelihood for different noise levels, σ and $\Delta(f)$

3.1.2 Relative - Continuous - Beta-Pairwise

We define a continuous but bounded response $\pi \in]0;1[$ observed when comparing u and v. The first option, u, is preferred for $\pi < 0.5$. The second option, v, is preferred for $\pi > 0.5$ and none is preferred for $\pi = 0.5$. The degree to which the prevailing options is preferred over the other is described by the distance to 0.5. Thus, the response captures both the choice between u and v and the degree of the preference.

The Probit is used as a mean function for the Beta distribution with parametrized shape parameters α and β, thus

$$p\left(\pi_k|\mathbf{f}_k\right) = \text{Beta}\left(\pi_k|\,\alpha(\mathbf{f}_k), \beta(\mathbf{f}_k)\right). \tag{4}$$

To express the shape parameters of the Beta distribution as a function of the Probit mean function $\mu(\mathbf{f}_k)$, we apply a well-known re-parametrization of the Beta distribution [8].

$$\alpha(\mathbf{f}_k) = \nu\mu(\mathbf{f}_k) \qquad \text{and} \qquad \beta(\mathbf{f}_k) = \nu(1-\mu(\mathbf{f}_k)), \tag{5}$$

where ν relates to the precision of the Beta distribution and is not parameterized by f. With this parameterization the distribution becomes the so-called *two-parameter, restricted Beta distribution*. Finally, our novel likelihood function depicted in Fig. 3.1.2 is described by

`Pairwise Beta likelihood`

$$\begin{aligned}\mathcal{L}_{cont} \equiv p\left(\pi_k|\mathbf{f}_k\right) &= \text{Beta}\left(\pi_k|\,\nu\mu(\mathbf{f}_k), \nu(1-\mu(\mathbf{f}_k))\right) &(6)\\ &= \frac{1}{\text{B}\left(\nu\mu\left(\mathbf{f}_k\right), \nu\left(1-\mu\left(\mathbf{f}_k\right)\right)\right)}\pi^{\nu\mu(\mathbf{f}_k)-1}(1-\pi)^{\nu(1-\mu(\mathbf{f}_k))-1} &(7)\end{aligned}$$

where $\text{B}(\alpha,\beta)$ is the beta function and the Beta mean function $\mu(\mathbf{f}_k)$ is given by

`Probit link-function`

$$\mu\left(\mathbf{f}_k\right) = \Phi\left(\frac{\mathbf{f}\left(v_k\right) - \mathbf{f}\left(u_k\right)}{\sqrt{2}\sigma}\right). \tag{8}$$

The precision term ν in Eq. (5) is inversely related to the observation noise on the continuous bounded responses. In general ν can be viewed *as a measure of how consistent the scale is used in a given comparison.*

Example:
For completeness we show a few examples of the likelihood model as a function of the difference $\Delta(f)$ between $f(u)$ and $f(v)$.

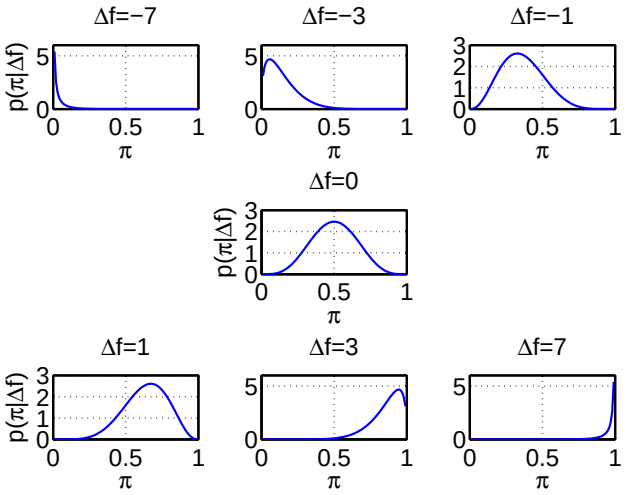
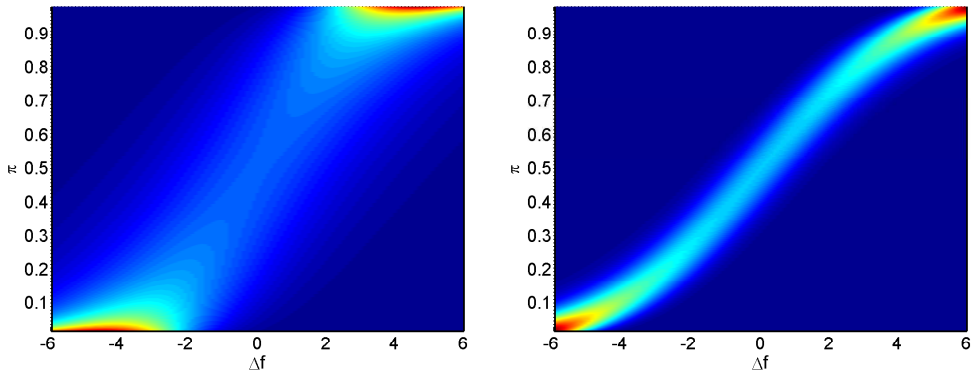
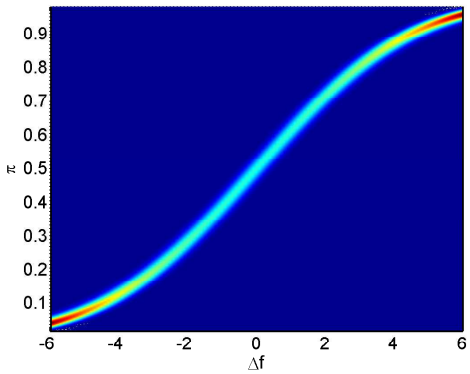


Figure 3: The Beta likelihood for specific $\Delta(f)$



(a) High Noise Level

(b) Medium noise level



(c) Low noise level (high consistency)

Figure 4: Illustration of the Beta likelihood for various noise levels (a), (b) and (c)

3.2 Absolute

Absolute ratings (in the form of bounded rating scales or categories) occur in a multitude of fields and standards, e.g., attribute assessments, quality determination (ITU P.835) etc.

The setup is similar to a standard regression setup where a set of inputs, $\mathcal{X} = \{x_k | k = 1, .., n\}$ has a corresponding output set of the same size, i.e., $\mathcal{Y} = \{y_k; x_k | k = 1, .., m\}$. Sometimes this is written as $\mathcal{D} = \{(y_k, x_k) | k = 1, .., n\}$, with $n = m$.

3.2.1 Absolute - Continuous - The Beta Likelihood

We consider a bounded response such that $y \in]0, 1[$ to a given instance $x \in \mathcal{X}$ is a stochastic random variable. The general idea of the Beta regression framework is to assume a latent general regression function $f : \mathcal{X} \to \mathbb{R}$. Given this function we can define our observation noise model as the likelihood of observing the bounded scale value y_k as

Absolute Beta likelihood

$$\mathcal{L}_{beta} \equiv p(y_k | f(x_k)) = \text{Beta}\left(y_k | \alpha\left(f\left(x_k\right)\right), \beta\left(f\left(x_k\right)\right)\right), \tag{9}$$

which is based on a Beta type distribution with shape parameters α and β dependent on the latent function f. We suggest to use a well-known re-parametrization of the beta distribution [8] (similar to section 3.1.2)

$$\alpha\left(f\left(x\right)\right) = \nu\mu\left(f\left(x\right)\right) \qquad \beta\left(f\left(x\right)\right) = \nu\left(1 - \mu\left(f\left(x\right)\right)\right), \tag{10}$$

where $\mu\left(f(x)\right)$ defines the mean of the beta distribution as a function of the latent regression function f and ν is related to the precision of the observation. Several mean link function would apply as long as it maps from an infinite interval to a bounded interval, is monotonic increasing and anti-symmetric around zero, as e.g., the probit or any other sigmoid function [2, chapter 6]. As mean link function we use the Probit defined by

Probit link-function

$$\mu\left(f\left(x\right)\right) = \Phi\left(\frac{f\left(x_k\right)}{\sqrt{2}\sigma}\right), \tag{11}$$

where σ is the slope parameter which in effect defines the scale of f. The Probit is chosen mainly for analytical reasons that will eventually by clear when the inference method is described in section 5.

4 The (Gaussian process) prior

So far we have not made any specific assumptions about f, hence to complete our Bayesian elicitation framework we need to define how to model f to complement any of the likelihood functions described in section 3. The only thing we have assumed so far regarding f is that we can treat it as a random variable and maintain a distribution over it.

There are in principle many parametric functions that can be applied in this case, also in a Bayesian setting. However, a Gaussian Process is an obvious and flexible choice by which our limited prior knowledge regarding f and its expected structure can be formulated intuitively without constraining the flexibility. In [7] Gaussian Processes were applied for the first time for the types of problems considered here and we continue along similar lines.

A Gaussian Process does in essence define a distribution over functions (with infinite flexibility) as required. Formally, a Gaussian Process is defined as [18, page 13]

GP definition

> A Gaussian process is a collection of random variables, any finite number of which have a (consistent) joint Gaussian distribution

It is typically denoted with a zero-mean function in the following way

GP notation

$$f(x) \sim \mathcal{GP}(0, k(x, \cdot)), \tag{12}$$

where $k(x, \cdot)$ refers to the *covariance function* (or kernel) defining the covariance between two random variables. We refer to [18] for an excellent treatment of Gaussian Process.

4.1 Covariance Function

A Gaussian Process is defined by a mean function (typically defined to be zero like above) and a covariance function. For a finite number of observed inputs a Gaussian Process can be described using a matrix notation, where the key (in the case of a zero-mean function) is the covariance matrix, which is determined by a particular choice of covariance function.

Covariance Function and Matrix

Hence, the main task (and constraint) is the definition of a suitable covariance function which must obey all the requirements for valid covariance functions/kernels (see [18]).

Example: The canonical example of a covariance function is the rather misnamed squared exponential function typically defined by

Squared Exponential

$$k(x_i, x_j) = \frac{1}{\sigma_{sf}^2} \exp\left(-\frac{1}{\sigma_l^2} \|x_i - x_j\|_2^2\right) \tag{13}$$

where σ_{sf}^2 is the variance and σ_l^2 is the length scale.

The main point to notice from the example is that most covariance functions contain free parameters and in a full, hierarchical Bayesian modeling framework it would be correct to formulate priors for each of these parameters. However, in this report we restrict ourselves to terminate the modeling at the first hierarchical level and will use only a point estimate of the parameter, estimated by an *evidence optimization* scheme [12]. This will be explained in greater detail in section 5.

4.1.1 Multi-Task Kernels

For preference learning (and in perceptual experiments in general) we are typically interested in evaluations from multiple subjects. Usually in classical statistics, it is assumed that subjects are related, produce similar responses and comes from the same population. From a machine learning point of view we also want to exploit subject differences and in particular subject similarities - and take advantage of these potential similarities in our model.

Incorporation of this type of collaborate modeling in a Gaussian Process setting is easily obtained using multi-task learning [5, 3, 4]. An especially elegant and effective formulation of multi-task learning uses a product covariance function $k(x, x')$, where each input x contains both stimuli specific features x_a and user specific features x_u, hence

collaboration

$$k(x, x') = k(x_a, x_a') \, k(x_u, x_u{}') \tag{14}$$

Now, the collaborate modeling problem is reduced to specification of meaningful covariance function "bricks" in the construction of the product kernel, which result in sufficient correlation structures for the problem at hand.

This formulation can directly be employed as is - yet the practical evaluation and specific/custom inference for multi/task kernels is left for future work.

Note also the link between the multitask formulation and matrix-variate Normal distributions.

5 The Inference - posterior

In the absolute **regression case** we have $\mathcal{Y} = \{(y_k; x_k) \,|k = 1, .., m\}$ and $p(\mathcal{Y}|\mathcal{X}) = \prod_{k=1}^{K} p(y_k|f(x_k), \boldsymbol{\theta}_{\mathcal{L}})$ Regression case

In the **relative/pairwise case** we have $\mathcal{Y} = \{(y_k; u_k, v_k) \,|k = 1, .., m, u_k \in \mathcal{X}, v_k \in \mathcal{X}\}$ and $p(\mathcal{Y}|\mathcal{X}) = \prod_{k=1}^{K} p(y_k|f(u_k), f(v_k), \boldsymbol{\theta}_{\mathcal{L}})$ Pairwise case

We first consider the full posterior in this general setup, i.e, Full posterior

$$p(\mathbf{f}, \boldsymbol{\theta}|\mathcal{Y}, \mathcal{X}) = \frac{p(\boldsymbol{\theta}_{\mathcal{GP}})\, p(\mathbf{f}|\boldsymbol{\theta}_{\mathcal{GP}}, \mathcal{X})\, p(\boldsymbol{\theta}_{\mathcal{L}})\, p(\mathcal{Y}|\mathbf{f}, \boldsymbol{\theta}_{\mathcal{L}})}{p(\mathcal{Y}|\mathcal{X})}, \tag{15}$$

with $p(\mathcal{Y}|\mathcal{X}) = \int\int\int p(\boldsymbol{\theta}_{\mathcal{GP}})\, p(\mathbf{f}|\boldsymbol{\theta}_{\mathcal{GP}}, \mathcal{X})\, p(\boldsymbol{\theta}_{\mathcal{L}})\, p(\mathcal{Y}|\mathbf{f}, \boldsymbol{\theta}_{\mathcal{L}})\, d\boldsymbol{\theta}_{\mathcal{GP}} d\boldsymbol{\theta}_{\mathcal{L}} d\mathbf{f}$ as the marginal likelihood or evidence.

Next, we will typically consider the covariance parameters and likelihood parameters constant point-estimates, thus we obtain Constant $\boldsymbol{\theta}$

$$p(\mathbf{f}|\mathcal{Y}, \mathcal{X}, \boldsymbol{\theta}_{\mathcal{GP}}, \boldsymbol{\theta}_{\mathcal{L}}) \quad = \quad p(\mathbf{f}|\boldsymbol{\theta}_{\mathcal{GP}}, \mathcal{X})\, p(\mathcal{Y}|\mathbf{f}, \boldsymbol{\theta}_{\mathcal{L}}) \,/\, p(\mathcal{Y}|\mathcal{X}, \boldsymbol{\theta}_{\mathcal{GP}}, \boldsymbol{\theta}_{\mathcal{L}}) \tag{16}$$

where $p(\mathcal{Y}|\mathcal{X}, \boldsymbol{\theta}_{\mathcal{GP}}, \boldsymbol{\theta}_{\mathcal{L}}) = \int p(\mathbf{f}|\mathcal{X}, \boldsymbol{\theta}_{\mathcal{GP}})\, p(\mathcal{Y}|\mathbf{f}, \boldsymbol{\theta}_{\mathcal{L}}) d\mathbf{f}$.

Due to notational convenience we often write the full posterior in equation (16) as $p(\mathbf{f}|\mathcal{Y}, \boldsymbol{\theta}_{\mathcal{GP}}, \boldsymbol{\theta}_{\mathcal{L}}) = p(\mathbf{f}|\boldsymbol{\theta}_{\mathcal{GP}})\, p(\mathcal{Y}|\mathbf{f}, \boldsymbol{\theta}_{\mathcal{L}}) \,/\, p(\mathcal{Y}|\boldsymbol{\theta}_{\mathcal{GP}}, \boldsymbol{\theta}_{\mathcal{L}})$ or simply $p(\mathbf{f}|\mathcal{Y}) = p(\mathbf{f})\, p(\mathcal{Y}|\mathbf{f}) \,/\, p(\mathcal{Y})$. Thus, we may at times choose to omit the conditioning on $\mathcal{X}$ and $\boldsymbol{\theta}$ for notation convenience. Notational convenience

In some cases the posterior $p(\mathbf{f}|\mathcal{Y}, \mathcal{X}, \boldsymbol{\theta}_{\mathcal{GP}}, \boldsymbol{\theta}_{\mathcal{L}})$ can be found analytically in case of conjugate priors. With the GP prior and the Probit or Beta Likelihood described in section 3 we are, however, not able to obtain a closed-form expression for the posterior. There are a number of ways to overcoming this issue.

- Reject the current model based on computational intractability.
- Sampling (MCMC methods)
- Analytical Approximations (Deterministic)
  - Laplace
  - Expectation Propagation [14].
  - Variational Approximation (EP can be considered a VB method (see [20])).
  - (possibly other options)

In this report we apply the well-known Laplace approximation, which is derived for the models described in previous sections. Left out details is included in appendix B, which will also be referenced during the short description in this section.

5.1 Laplace

5.1.1 Posterior Moment Matching

The main idea is to approximate the posterior by a single finite multivariate Gaussian distribution, such that

$$p\left(\mathbf{f}|\mathcal{Y},\mathcal{X},\boldsymbol{\theta}_{\mathcal{GP}},\boldsymbol{\theta}_{\mathcal{L}}\right)\approx q\left(\mathbf{f}\right)=\mathcal{N}\left(\mathbf{f}|\hat{\mathbf{f}},\mathbf{A}^{-1}\right). \tag{17}$$

Where $\hat{\mathbf{f}}$ is the mode of the posterior and $\mathbf{A}$ is the Hessian of the negative log-posterior at the mode. The mode is found as

$$\begin{aligned}\hat{\mathbf{f}} &= \arg\max_{\mathbf{f}} p\left(\mathbf{f}|\mathcal{Y}\right) &(18)\\ &= \arg\max_{\mathbf{f}} p\left(\mathcal{Y}|\mathbf{f}\right)p\left(\mathbf{f}\right) &(19)\\ &= \arg\max_{\mathbf{f}} \log\left[p\left(\mathcal{Y}|\mathbf{f}\right)p\left(\mathbf{f}\right)\right] &(20)\\ &= \arg\max_{\mathbf{f}} \log\left[\Psi\left(\mathbf{f}|\mathcal{Y}\right)\right] &(21)\end{aligned}$$

The general solution to the problem can thus be found by considering the unnormalized log-posterior $\Psi\left(\mathbf{f}|\mathcal{Y}\right)$ and the resulting cost function which is to be maximized in respect to $\mathbf{f}$, is given by

$$\Psi\left(\mathbf{f}|\mathcal{Y}\right)=\log p\left(\mathcal{Y}|\mathbf{f}\right)-\frac{1}{2}\mathbf{f}^{T}\mathbf{K}^{-1}\mathbf{f}-\frac{1}{2}\log\left|\mathbf{K}\right|-\frac{N}{2}\log 2\pi.$$

In order to provide a robust and fast approach for finding the mode we propose a damped Newton method with soft linesearch to maximize Eq. (22). In our case the basic damped Newton step can be calculated without inversion of the Hessian (see appendix C.1)

$$\mathbf{f}^{new}=\left(\mathbf{K}^{-1}+\mathbf{W}-\lambda\mathbf{I}\right)^{-1}\left[\left(\mathbf{W}-\lambda\mathbf{I}\right)\mathbf{f}+\nabla\log p(\mathcal{Y}|\mathbf{f},\boldsymbol{\theta}_{\mathcal{L}})\right], \tag{22}$$

where λ is the adaptive damping factor controlled via a standard Levenberg–Marquardt procedure as in [15].

Furthermore, $\mathbf{W}$ is the second order partial derivative of the log-likelihood with respect to the $f(x)'s$, i.e.

$$\mathbf{W}_{i,j}=-\sum_{k}\nabla\nabla_{i,j}\log p(y_k|\mathbf{f}_k,\boldsymbol{\theta}_{\mathcal{L}}) \tag{23}$$

where we have defined $\nabla\nabla_{i,j}=\frac{\partial^2}{\partial f(x_i)\partial f(x_j)}$.

For pairwise observations the individual likelihood terms $\nabla\nabla_{i,j}\log p(y_k|\mathbf{f}_k,\boldsymbol{\theta}_{\mathcal{L}})$ will be nonzero only when both x_i and x_j occur as either v_k or u_k in $\mathbf{f}_k$. Therefore, the Hessian (of the log-likelihood) $\mathbf{W}$ is not diagonal since it depends on two inputs, which makes the approximation slightly more involved (see appendix A.1).

For absolute likelihoods $\mathbf{W}$ becomes diagonal since each likelihood term only depends on a single $f(x)$. In classical GP classification this is exploited for an optimized implementation, however we maintain a general implementation for flexibility.

In all cases the resulting approximation after convergence can be shown to be

$$p\left(\mathbf{f}|\mathcal{Y},\mathcal{X},\boldsymbol{\theta}\right)\approx\mathcal{N}\left(\mathbf{f}|\hat{\mathbf{f}},\left(\mathbf{W}+\mathbf{K}^{-1}\right)^{-1}\right). \tag{24}$$

5.1.2 Hyper-parameters

The moment matching procedure is a relatively fast method to approximate $p(\mathbf{f}|\mathcal{Y})$, however, the solution does not include a Bayesian treatment of the uncertainty about the (hyper-)parameters in neither the likelihood nor the GP prior.

The full joint posterior is $p(\mathbf{f}, \boldsymbol{\theta}_{\mathcal{L}}, \boldsymbol{\theta}_c|\mathcal{Y}, \mathcal{X})$, and to get this we would require suitable priors on $\boldsymbol{\theta}_{\mathcal{L}}, \boldsymbol{\theta}_c$. In a full Bayesian treatment, we would then integrate out the hyper-parameters in order to obtain the posterior of interest, typically $p(\mathbf{f}|\mathcal{Y}, \mathcal{X})$ (which explicitly does not have any dependencies on any parameters). In practice though, this is often intractable to do analytically.

Instead, we resort to a simpler class of methods generally referred to as empirical Bayes. In Machine Learning this is typically known as evidence optimization or ML-II in which the marginal likelihood - or evidence - is maximized. We will extend this slightly to allow for a prior on the hyper-parameters which in turn leads to a regularized estimate of the hyper-parameters.

5.1.2.1 Empirical Bayes I: Evidence Optimization Given the Laplace approximation we update the marginal likelihood, denoted by $p(\mathcal{Y}|\mathcal{X}, \boldsymbol{\theta})$, with regards to the hyper-parameters collected in $\boldsymbol{\theta}$. The evidence is like the posterior analytically intractable, hence we apply the Laplace approximation to approximate the evidence as

$$p(\mathcal{Y}|\mathcal{X}, \boldsymbol{\theta}) \approx \log p\left(\mathcal{Y}|\hat{\mathbf{f}}\right) - \frac{1}{2}\hat{\mathbf{f}}^T\mathbf{K}^{-1}\hat{\mathbf{f}} - \frac{1}{2}\log|I + \mathbf{KW}|\,. \tag{25}$$

This expression depends both implicitly and explicitly on the covariance hyper-parameters: Implicitly through the dependence on the solution of $\hat{\mathbf{f}}$, and explicitly since directly used in the covariance function. Obviously, the expression also depends on the likelihood parameters through the log likelihood. The resulting approach can be considered in an EM-style update

---

1. Step:

$$p\left(\mathbf{f}|\mathcal{Y}, \mathcal{X}, \hat{\boldsymbol{\theta}}\right) = p\left(\mathbf{f}|\mathcal{X}, \hat{\boldsymbol{\theta}}_{\mathcal{GP}}\right) p\left(\mathcal{Y}|\mathbf{f}, \hat{\boldsymbol{\theta}}_{\mathcal{L}}\right) / p\left(\mathcal{Y}|\mathcal{X}, \hat{\boldsymbol{\theta}}\right) \tag{26}$$

$$\approx \mathcal{N}\left(\mathbf{f}, \left(\mathbf{K}^{-1} + \mathbf{W}\right)^{-1}\right) \tag{27}$$

via the Laplace approximation.

2. Step:

$$\hat{\boldsymbol{\theta}} = \text{argmax}_{\boldsymbol{\theta}} p(\mathcal{Y}|\mathcal{X}, \boldsymbol{\theta}) \tag{28}$$

... and iterate

By considering the gradient of Eq. (25) it is possibly to optimize it in step 2 with a standard gradient based optimization method, for which we apply a BFGS method found in [15]. This crude empirical estimation of the likelihood parameters is believe to work if parameters in the likelihood is closely linked to f, hence the prior on f will also regularize the likelihood parameters, e.g., for small data set. Any noise on the data also regularize the evidence procedure (as training with noise).

5.1.2.2 Empirical Bayes II: Regularized Approach (approximate MAP) We extend the evidence principle to allow for a simple regularization of the estimate by placing individual priors on the hyper-parameters. We still make the Laplace approximation with fixed parameters $\boldsymbol{\theta}$, but in the re-estimation step we consider the posterior of the parameter given the current Laplace approximation. So with a general prior over $\boldsymbol{\theta}$, $p(\boldsymbol{\theta}|\mathcal{X})$, Bayes gives us

$$p(\boldsymbol{\theta}|\mathcal{Y},\mathcal{X}) = p(\mathcal{Y}|\mathcal{X},\boldsymbol{\theta})\, p(\boldsymbol{\theta}|\mathcal{X})\, / p(\mathcal{Y}|\mathcal{X}) \tag{29}$$

or

$$\log p(\boldsymbol{\theta}|\mathcal{Y},\mathcal{X}) = \log p(\mathcal{Y}|\mathcal{X},\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}|\mathcal{X}) - \log p(\mathcal{Y}|\mathcal{X}), \tag{30}$$

where we notice that $p(\mathcal{Y}|\mathcal{X},\boldsymbol{\theta})$ is exactly the evidence term in equation (25), with the approximation already given. Furthermore, $\log p(\mathcal{Y}|\mathcal{X})$ is independent of $\boldsymbol{\theta}$, and the prior $\log p(\boldsymbol{\theta}|\mathcal{X})$ is typically considered independent of $\mathcal{X}$.

Note, that with a uniform prior, $\log p(\boldsymbol{\theta}) = \text{const}$, this reduces to the standard point-estimate of the evidence procedure, i.e., $\arg\max_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}|\mathcal{Y},\mathcal{X}) = \arg\max_{\boldsymbol{\theta}} \log p(\mathcal{Y}|\boldsymbol{\theta},\mathcal{X})$

Hence, the only change to the standard case is the extra term $\log p(\boldsymbol{\theta})$ which is to be differentiated with respect to the parameters. Thus, in an EM-style approach we then re-estimate $\hat{\boldsymbol{\theta}}$ in the second step by optimizing in regards to the approximate posterior of $\boldsymbol{\theta}$. The resulting optimization scheme becomes

---

1. Step:

$$\begin{aligned} p\left(\mathbf{f}|\mathcal{Y},\mathcal{X},\hat{\boldsymbol{\theta}}\right) &= p\left(\mathbf{f}|\mathcal{X},\hat{\boldsymbol{\theta}}_{\mathcal{GP}}\right) p\left(\mathcal{Y}|\mathbf{f},\hat{\boldsymbol{\theta}}_{\mathcal{L}}\right) / p\left(\mathcal{Y}|\mathcal{X},\hat{\boldsymbol{\theta}}\right) & (31)\\ &\approx \mathcal{N}\left(\mathbf{f},\left(\mathbf{K}^{-1}+\mathbf{W}\right)^{-1}\right) & (32)\end{aligned}$$

via the Laplace approximation.

2. Step:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}|\mathcal{Y},\mathcal{X}) = \arg\max_{\boldsymbol{\theta}} \{\log p(\mathcal{Y}|\mathcal{X},\boldsymbol{\theta}) + \log p(\boldsymbol{\theta})\} \tag{33}$$

... and iterate

Note, that the above approximation style can be extended to a full VB approach resulting in a approximate posteriors over the parameters.

It is of course possible to further extend the prior hierarchy, however, given the added complexity we typically need more advanced inference methods than the ones considered above.

The applied algorithm is listed and summarized in figure 5.2.

5.2 Summary

We have shown the main steps required in the Laplace approximation and provided a approach to regularize the estimate of the hyper-parameters.

Algorithm 1 Laplace Algorithm with $\boldsymbol{\theta}$

Require: $\mathcal{Y}$, $\boldsymbol{\theta}_{\text{init}}$, $\log p(\boldsymbol{\theta}|\mathcal{X})$, $p(y|\mathbf{f})$(likelihood function)

$\boldsymbol{\theta}^{\leftarrow \boldsymbol{\theta}_{\text{init}}}$
repeat
  Compute $\mathbf{K}$ with current $\boldsymbol{\theta}$
  $(\mathbf{f}, \mathbf{W}) \leftarrow FindMode(\mathcal{Y}, \boldsymbol{\theta})$ ▷ Basic Laplace Approximation
  $(p(\boldsymbol{\theta}|\mathcal{Y}, \mathcal{X}),\ \partial p(\boldsymbol{\theta}|\mathcal{Y}, \mathcal{X})/\partial \boldsymbol{\theta}) \leftarrow \text{EvaluateHyper}(\mathbf{f}, \mathbf{W}, \boldsymbol{\theta}, p(\boldsymbol{\theta}))$
  $\boldsymbol{\theta} \leftarrow BFGS(\boldsymbol{\theta}, \partial p(\boldsymbol{\theta}|\mathcal{Y}, \mathcal{X})/\partial \boldsymbol{\theta})$
until Converged
return $\mathbf{f}, p(\mathcal{Y}|\mathcal{X}), \boldsymbol{\theta}$

Figure 5: The pseudo code shows the required computational steps to find both the posterior $p(\mathbf{f}|\mathcal{Y})$ and the regularized estimate of the hyper-parameters $\boldsymbol{\theta}$ based on the (approximated) evidence $p(\mathcal{Y}|\mathcal{X})$ and prior.

6 The Inference - predictions

In this section we will describe the necessary machinery required for making predictions of the various responses described in section 3 for new inputs, specifically based on the predictive distributions.

In section 5 we described the (approximate) inference method of the latent variables based on the observations, in particular we considered $p(\mathbf{f}|\mathcal{Y}, \boldsymbol{\theta})$. Given the associated likelihood this (approximation) allows us to make predictions for new inputs.

6.1 Prediction of latent function

We will start out making predictions about the latent function f with the end goal to do predictions of the observable variable y for an unobserved example. In the absolute case the test input is a single instance, $r \in \mathcal{X}$, while in the pairwise case it is two instances $r \in \mathcal{X}$ and $s \in \mathcal{X}$ from which we need to make predictions (note that they can be - and often are - correlated). In the following we consider the general case of the pairwise response which easily reduces to the absolute regression case by omitting the s input.

Given the GP, the joint prior between test $\mathbf{f}_t = [f(r), f(s)]^T$ (or $\mathbf{f}_t = [f(r)]$ in the simpler case) and training $\mathbf{f}$ is given by a multivariate Gaussian distribution

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}_t \end{bmatrix} = \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{K} & \mathbf{k}_t \\ \mathbf{k}_t^T & \mathbf{K}_t \end{bmatrix} \right), \tag{34}$$

where $\mathbf{k}_t$ is a matrix with elements $\mathbf{k}_{i,1} = k(x_i, r)$ and $\mathbf{k}_{i,2} = k(x_i, s)$ with x_i being a training input. The conditional $p(\mathbf{f}_t|\mathbf{f})$ is obviously Gaussian as well and can directly be obtained from Eq. (34). The predictive distribution is given as

$$p(\mathbf{f}_t|\mathcal{Y}) = \int p(\mathbf{f}_t|\mathbf{f})\, p(\mathbf{f}|\mathcal{Y})\, d\mathbf{f}. \tag{35}$$

With $p(\mathbf{f}|\mathcal{Y})$ approximated as a finite Gaussian then $p(\mathbf{f}_t|\mathcal{Y})$ will be Gaussian too, given as $\mathcal{N}(\mathbf{f}_t|\mu^*, \mathbf{K}^*)$ with $\hat{\mathbf{f}}$ as the mode of the posterior and $\mathbf{W}$ as the negative Hessian.

The mean prediction is given by

$$\mu^* = [\mu_r^*, \mu_s^*]^T = \mathbf{k}_t{}^T \mathbf{K}^{-1} \hat{\mathbf{f}} \tag{36}$$

and the covariance matrix (or variance in the absolute case)

$$\mathbf{K}^* = \begin{bmatrix} \mathbf{K}_{rr}^* & \mathbf{K}_{rs}^* \\ \mathbf{K}_{sr}^* & \mathbf{K}_{ss}^* \end{bmatrix} = \mathbf{K}* = \mathbf{K}_t - \mathbf{k}_t{}^T (\mathbf{I} + \mathbf{W}\mathbf{K})^{-1} \mathbf{W}\, \mathbf{k}_t \tag{37}$$

Note that in the absolute regression case $\mathbf{K}^*$ reduces to a scalar namely the predictive variance of the single test point, r.

With the predictive distribution for $\mathbf{f}_t$, the predictive distribution for the observed variable is available as

Prediction of the observed response

$$p\left(y_t|\mathcal{Y}\right) = \int p\left(y_t|\mathbf{f}_t, \mathcal{Y}\right) p\left(\mathbf{f}_t|\mathcal{Y}\right) d\mathbf{f}_t. \tag{38}$$

This obviously depends on the applied likelihood and will in the following be treated separately for the different cases.

6.2 Relative

6.2.1 Relative - Discrete - Probit

The predictive preference is given by:

$$p\left(d^*|\mathcal{Y}\right) = \int p\left(d^*|\mathbf{f}_t, \mathcal{Y}\right) p\left(\mathbf{f}_t|\mathcal{Y}\right) d\mathbf{f}_t \tag{39}$$

If $p\left(\mathbf{f}_t|\mathcal{Y}\right)$ is Gaussian as assumed and given the Probit likelihood this integral can be evaluated in closed form. The result is:

$$P\left(r \succ s|\mathcal{Y}\right) = \Phi\left(\frac{\mu_r^* - \mu_s^*}{\sigma^*}\right) \tag{40}$$

with $(\sigma^*)^2 = 2\sigma^2 + \mathbf{K}_{rr}^* + \mathbf{K}_{ss}^* - \mathbf{K}_{rs}^* - \mathbf{K}_{sr}^*$.

6.2.2 Relative - Continuous - Beta Likelihood

In the Beta case the predictions are somewhat more complicated, since the observed variable is Π. Thus, we need to be careful when considering e.g., a binary preference deduced from this.

We first start out defining the predictive distribution over the observed variable

$$P\left(\pi^*|\mathcal{Y}\right) = \int p(\pi^*|\mathbf{f}_t, \mathcal{Y})\, P(\mathbf{f}_t|\mathcal{Y})\, d\mathbf{f}_t \tag{41}$$

In the Beta case this integral does not to our knowledge have an obviously approximation[1]. With the above predictive distribution for π, and the chosen mean function we can **define** the probability of the binary choice as.

$$P\left(r \succ s|\mathcal{Y}\right) = \int_{\pi=0}^{1/2} P\left(\pi|\mathcal{Y}\right)\, d\pi \tag{42}$$

[1] Obviously still looking for an approximation

The latter can be written in a more effective manner.

$$
\begin{aligned}
P\left(r \succ s|\mathcal{Y}\right) &= \int_{\pi=0}^{1/2} \int_{f_t} p(\pi|\mathbf{f}_t, \mathcal{Y})\, P(\mathbf{f}_t|\mathcal{Y})\, d\mathbf{f}_t d\pi && (43)\\
&= \int_{f_t} P(\mathbf{f}_t|\mathcal{Y}) \int_{\pi=0}^{1/2} p(\pi|\mathbf{f}_t, \mathcal{Y})\, d\pi d\mathbf{f}_t && (44)\\
&= \int_{f_t} P(\mathbf{f}_t|\mathcal{Y}) Betacdf\left(1/2, \alpha(\mathbf{f}_t), \beta(\mathbf{f}_t)\right) d\mathbf{f}_t && (45)\\
& && (46)
\end{aligned}
$$

where we have used Fubini's theorem. This integral can be evaluated using numerical methods - or simple Monte Carlo sampling if found appropriate.

6.3 Absolute Ratings

6.3.1 Absolute - Continious - Beta

The response is in this case a scalar value, on the scale, which is also the response to be predicted based on the test input x_t. The predictions of the latent function, $f(x_t)$, follows the same as for the pairwise, however, $\mathbf{f}_t$ only contains a single element, thus we now denote it by f_t.

The full predictive distribution over the response is given by Eq. 38, which can be evaluated using MCMC methods. We will although focus on an important consequence of the chosen mean parametrization by considering the predictive mean (i.e. the mean of the predictive distribution).

$$
\begin{aligned}
\mathbb{E}_{p(y_t|\mathcal{Y})}\left\{y_t\right\} &= \int y_t p\left(y_t|\mathcal{Y}\right) dy_t = \int y_t \int p\left(y_t|f_t\right) p\left(f_t|\mathcal{Y}\right) df_t dy_t && (47)\\
&= \int\int y_t p\left(y_t|f_t\right) p\left(f_t|\mathcal{Y}\right) df_t dy_t && (48)\\
&= \int\int y_t p\left(y_t|f_t\right) p\left(f_t|\mathcal{Y}\right) dy_t df_t && (49)\\
&= \int p\left(f_t|\mathcal{Y}\right) \int y_t p\left(y_t|f_t\right) dy_t df_t && (50)\\
&= \int p\left(f_t|\mathcal{Y}\right) \mathbb{E}_{p(y_t|f_t)}\left\{y_t\right\} df_t. && (51)
\end{aligned}
$$

where we have used Fubini's theorem.

Since $p\left(y_t|f_t\right)$ is defined by the likelihood function from section 3.2 with the probit mean

function $\mu(f)$, we get

$$\mathbb{E}_{p(y_t|\mathcal{Y})}\{y_t\} = \int p(f_t|\mathcal{Y})\,\mu(f_t)\,df_t \tag{52}$$

$$= \int p(f_t|\mathcal{Y})\,\Phi\left(\frac{f_t}{\sqrt{2}\sigma}\right)df_t = \Phi\left(\frac{\mu_t}{\sqrt{2\sigma^2+(\sigma^*)^2}}\right). \tag{53}$$

which of course provides a fast prediction of the mean response.

7 Summary

In this technical report/reference we have presented the required steps to model pairwise comparisons and absolute judgments using Gaussian Process priors.

References

[1] Christopher M Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006. 3

[2] R.D. Bock and JV Jones. The measurement and prediction of judgment and choice. 1968. 3.1.1, 3.2.1

[3] E. Bonilla, K.M. Chai, and C. Williams. Multi-task Gaussian process prediction. *Advances in Neural Information Processing Systems*, 20:153–160, 2008. 4.1.1

[4] Edwin Bonilla, Shengbo Guo, and Scott Sanner. Gaussian Process Preference Elicitation. In J Lafferty, C K I Williams, J Shawe-Taylor, R S Zemel, and A Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 262–270. 2010. 3.1.1, 4.1.1

[5] E.V. Bonilla, F.V. Agakov, and C.K.I. Williams. Kernel multi-task learning using task-specific features. *Proceedings of the 11th AISTATS*, 2007. 4.1.1

[6] Wei Chu and Z Ghahramani. Extensions of Gaussian processes for ranking: semi-supervised and active learning. In *Workshop Learning to Rank at Advances in Neural Information Processing Systems 18*, 2005. 3.1.1

[7] Wei Chu and Zoubin Ghahramani. Preference learning with Gaussian processes. *ICML 2005 - Proceedings of the 22nd International Conference on Machine Learning*, pages 137–144, 2005. 3.1.1, 4

[8] Silvia Ferrari and Francisco Cribari-Neto. Beta Regression for Modelling Rates and Proportions. *Journal of Applied Statistics*, 31(7):799–815, August 2004. 3.1.2, 3.2.1

[9] B S Jensen. Pairwise Comparisons using GP priors - review, implementation and extensions. Unpublished, techinal report. B

[10] B. S. Jensen, J. B. Nielsen, and J. Larsen. Bounded gaussian process regression. *Proceedings of IEEE International Workshop on Machine Learning for Signal Processing*, 2013. DOI 10.1109/MLSP.2013.6661916. 2

[11] Bjørn Sand Jensen, Jens Brehm Nielsen, and Jan Larsen. Efficient Preference Learning with Pairwise Continuous Observations and Gaussian Processes. *IEEE International Workshop on Machine Learning for Signal Processing*, 2011. (document), 2, B.1.2.2

[12] D. MacKay. The Evidence Framework Applied to Classification Networks. *Neural Computation*, 4(5):720–736, 1992. 4.1

[13] David J. C. MacKay. *INFORMATION THEORY, INFERENCE, AND LEARNING ALGORITHMS*. Cambridge University Press, 2003. 3

[14] Thomas Minka. *A family of algorithms for approximate Bayesian inference*. PhD thesis, MIT, 2001. 5

[15] Hans Bruun Nielsen. immoptibox - A MATLAB TOOLBOX FOR OPTIMIZATION AND DATA FITTING. 5.1.1, 5.1.2.1, B.3

[16] J. B. Nielsen, B. S. Jensen, T. J. Hansen, and J. Larsen. Personalized audio systems - a bayesian approach. *Proceedings of the 135th Audio Engineering Society (AES) Convention*, 2013. 2

[17] K B Petersen and M S Pedersen. The Matrix Cookbook, 2008. C.1

[18] Carl Edward Rasmussen and Christopher K I Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006. 4, 4, 4.1, B, B.1.2.1, B.3, C.1

[19] L L Thurstone. A law of comparative judgement. *Psychological Review*, 34, 1927. 3.1, 3.1.1

[20] Martin J. Wainwright and Michael I. Jordan. Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends in Machine Learning*, 1(12):1–305, 2008. 5

A Model Derivatives

A.1 Relative

A.1.1 Relative - Discrete - Probit

We define the argument to the probit such that

$$\mathrm{z}_k = \mathrm{y}_k \frac{\mathrm{f}\left(u_k\right) \text{ - } \mathrm{f}\left(v_k\right) \text{ - } \mathrm{b}}{\sqrt{2}\sigma}$$

where b is the bias term, typically not employed.

A.1.1.1 First derivative of log-likelihood :

$$\underbrace{\frac{d}{d\mathbf{f}\left(x^i\right)} \ln p\left(d_k|\mathbf{f}\left(u_k\right), \mathbf{f}\left(v_k\right)\right)}_{Nx1\ vector\ non-zero\ for\ x^i=u_k or x^i=v_k} = \frac{d}{d\mathbf{f}} \ln \int_{-\infty}^{z\left(f(u^k),f(v^k),\sigma,yk\right)} N\left(t|0,1\right)dt \tag{54}$$

$$= \frac{d}{d\mathbf{f}\left(x^i\right)} \ln \Phi\left(\frac{y_k\left(\mathbf{f}\left(u^k\right) - \mathbf{f}\left(v^k\right)\right)}{\sqrt{2}\sigma}\right) \tag{55}$$

$$= \frac{d}{d\mathbf{f}\left(x^i\right)} \ln \Phi\left(z_k\right) \tag{56}$$

$$= \mathbb{I}\left(x^i\right) \frac{1}{\Phi\left(z_k\right)} \frac{d}{d\mathbf{f}} \int_{-\infty}^{z\left(f(u^k),f(v^k),\sigma\right)} N\left(t|0,1\right)dt \tag{57}$$

$$= \mathbb{I}\left(x^i\right) y_k \frac{1}{\Phi\left(z_k\right)} \frac{1}{\sqrt{2}\sigma} N\left(z_k|0,1\right) \tag{58}$$

$$\mathbb{I}\left(s\right) = \begin{cases} 1 & if\ s = u^k \\ -1 & if\ s = v^k \\ 0 & otherwise \end{cases}$$

The indicator is due to

$$\frac{d}{d\mathbf{f}\left(u^k\right)} \ln p\left(d_k|\mathbf{f}\right) = -\frac{d}{d\mathbf{f}\left(v^k\right)} \ln p\left(d_k|\mathbf{f}\right)$$

A.1.1.2 Second derivative of log-likelihood :

$$\frac{d}{d\mathbf{f}\left(x^i\right)d\mathbf{f}\left(x^j\right)}\ln p\left(d_k|\mathbf{f}\left(u^k\right),\mathbf{f}\left(v^k\right)\right)= \tag{59}$$

$$= \frac{d}{d\mathbf{f}\left(x^j\right)}y_k\mathbb{I}\left(x^i\right)\mathbb{I}\left(x^j\right)\frac{1}{\sqrt{2}\sigma}\frac{1}{\Phi\left(z_k\right)}N\left(z_k|0,1\right) \tag{60}$$

$$= y_k\mathbb{I}\left(x^i\right)\mathbb{I}\left(x^j\right)\frac{1}{\sqrt{2}\sigma}\frac{d}{d\mathbf{f}\left(x^j\right)}\frac{N\left(z_k|0,1\right)}{\Phi\left(z_k\right)} \tag{61}$$

$$= y_k\mathbb{I}\left(x^i\right)\mathbb{I}\left(x^j\right)\frac{1}{\sqrt{2}\sigma}\left[\frac{\frac{d}{d\mathbf{f}\left(x^j\right)}N\left(z_k|0,1\right)}{\Phi\left(z_k\right)}-N\left(z_k|0,1\right)\frac{\frac{d}{d\mathbf{f}\left(x^j\right)}\Phi\left(z_k\right)}{\Phi^2\left(z_k\right)}\right] \tag{62}$$

$$= -\left(y_k{}^2\right)\mathbb{I}\left(x^i\right)\mathbb{I}\left(x^j\right)\frac{1}{2\sigma^2}\left[\frac{z_kN\left(z_k|0,1\right)}{\Phi\left(z_k\right)}+\frac{N^2\left(z_k|0,1\right)}{\Phi^2\left(z_k\right)}\right] \tag{63}$$

$$\mathbb{I}\left(s\right)=\begin{cases}1 & if\ s=u^k\\ -1 & if\ s=v^k\\ 0 & otherwise\end{cases} \tag{64}$$

A.1.1.3 Third derivative of log-likelihood :

$$\frac{\partial^3}{\partial f\left(x_i\right)\partial f\left(x_j\right)\partial f\left(x_k\right)}\log p\left(y_k|f(u_k),f(v_k)\right)=$$

$$= -\frac{1}{2\sigma^2}y_k{}^2\mathbb{I}\left(x_i\right)\mathbb{I}\left(x_j\right)\mathbb{I}\left(x_k\right)\left[\frac{y_k\mathcal{N}\left(z_k\right)}{\sqrt{2}\sigma\Phi\left(z_k\right)}-\frac{y_kz_k{}^2\mathcal{N}\left(z_k\right)}{\sqrt{2}\sigma\Phi\left(z_k\right)}-\frac{y_kz_k\mathcal{N}(z_k)^2}{\sqrt{2}\sigma\Phi(z_k)^2}-\frac{y_kz_k\sqrt{2}\mathcal{N}(z_k)^2}{\sigma\Phi(z_k)^2}-\frac{2y_k\mathcal{N}(z_k)^3}{\sqrt{2}\sigma\Phi(z_k)^2}\right]$$

$$= -\frac{1}{2\sqrt{2}\sigma^3}y_k{}^3\mathbb{I}\left(x_i\right)\mathbb{I}\left(x_j\right)\mathbb{I}\left(x_k\right)\left[\frac{\mathcal{N}\left(z_k\right)}{\Phi\left(z_k\right)}-\frac{z_k{}^2\mathcal{N}\left(z_k\right)}{\Phi\left(z_k\right)}-\frac{3z_k\mathcal{N}(z_k)^2}{\Phi(z_k)^2}-\frac{2\mathcal{N}(z_k)^3}{\Phi(z_k)^2}\right] \tag{65}$$

A.1.1.4 First Derivative of the Log-Likelihood With Respect to Parameters :

σ:

$$\frac{\partial}{\partial\sigma}\log p\left(y_k|f(u_k),f(v_k)\right)=-\frac{z_k}{\sigma}\frac{\mathcal{N}\left(z_k\right)}{\Phi\left(z_k\right)} \tag{66}$$

$$\frac{\partial}{\partial\log\left(\sigma\right)}\log p\left(y_k|f(u_k),f(v_k)\right)=-z_k\frac{\mathcal{N}\left(z_k\right)}{\Phi\left(z_k\right)} \tag{67}$$

b **(optional**:

$$\frac{\partial}{\partial b}\log p\left(y_k|f(u_k),f(v_k)\right)=\ -\ \frac{\mathrm{yk}\mathcal{N}\left(z_k\right)}{\sqrt{2}\sigma\Phi\left(z_k\right)} \tag{68}$$

A.1.1.5 First Derivative of Hessian With Respect to Parameters : σ:

$$\frac{\partial W}{\partial(\sigma)} = \frac{-\nabla\nabla \log p(D|\mathbf{f},\omega)}{\partial(\sigma)} \tag{69}$$

$$= -\mathbb{I}(x^i)\,\mathbb{I}(x^j)\left[\frac{yk^2}{\sigma^3}\left[\frac{z_k\mathcal{N}(z_k)}{\Phi(z_k)} + \frac{\mathcal{N}(z_k)^2}{\Phi(z_k)^2}\right]\right. \tag{70}$$

$$\left. -\frac{1}{2\sigma^3}yk^2 z_k\mathcal{N}(z_k)\frac{(-\Phi(z_k)^2 + {z_k}^2\Phi(z_k)^2 + 3z_k\mathcal{N}(z_k)\Phi(z_k) + 2\mathcal{N}(z_k)^2)}{\Phi(z_k)^3}\right]$$

b (bias, optional)

$$\frac{\partial W}{\partial b} = \frac{-\nabla\nabla \log p(D|\mathbf{f},\omega)}{\partial b} \tag{71}$$

$$= \mathbb{I}(x^i)\,\mathbb{I}(x^j)\,\frac{yk^3}{2\sqrt{2}\sigma^3}\,\mathcal{N}(z_k)\frac{(-\Phi(z_k)^2 + zk^2\Phi(z_k)^2 + 3z_k\mathcal{N}(z_k)(z_k|0,1)\,\Phi(z_k) + 2\mathcal{N}(z_k)^2)}{\Phi(z_k)^3}$$

A.1.2 Relative - Continuous - Beta

For ease of understanding and reading, the derivatives of the parameterized beta likelihood have been broken down into smaller pieces. First, the derivatives wrt. $f(x^i)$ of the log-beta were calculated including the shape parameters $\alpha(f)$ and $\beta(f)$ of the beta distribution explicitly as a function of f. Thereby, the derivatives of $\alpha(f)$ and $\beta(f)$ occur in the derivatives of the log-likelihood. Next, the derivatives of $\alpha(f)$ and $\beta(f)$ are found including the parametrization of the mean $\mu(f)$. Again, this means that the derivatives of $\alpha(f)$ and $\beta(f)$ are given by the derivatives of the mean function $\mu(f)$. Finally, the derivatives of the mean function wrt. $f(x^i)$ can be found. At the very end we collect the derivatives of the different terms and reduce the resulting equation to yield a compact solution.

A.1.2.1 First derivative of log-likelihood

$$\begin{aligned}
&\frac{d}{d\mathbf{f}(x^i)}\left[\log p(\pi_k \mid \mathbf{f}(u^k), \mathbf{f}(v^k))\right] \\
&\quad= \frac{d}{d\mathbf{f}(x^i)}\left[\log Beta(\pi_k \mid \alpha(f), \beta(f))\right] \\
&\quad= \frac{d\alpha(f)}{d\mathbf{f}(x^i)}\left[\log(\pi_k) - \psi(\alpha(f)) + \psi(\alpha(f)+\beta(f))\right] \\
&\quad+ \frac{d\beta(f)}{d\mathbf{f}(x^i)}\left[\log(1-\pi_k) - \psi(\beta(f)) + \psi(\alpha(f)+\beta(f))\right], \qquad (72)
\end{aligned}$$

where $\alpha(f)$ and $\beta(f)$ are given by equation (5), respectively, and ψ is the digamma function (corresponds to the polygamma function of order 0).

The first derivatives of the shape parameters are given by

$$\frac{d}{d\mathbf{f}(x^i)}\alpha(f) = \nu\frac{d}{d\mathbf{f}(x^i)}\mu(f) \qquad (73)$$

$$\frac{d}{d\mathbf{f}(x^i)}\beta(f) = -\nu\frac{d}{d\mathbf{f}(x^i)}\mu(f) = -\frac{d}{d\mathbf{f}(x^i)}\alpha(f), \qquad (74)$$

where the first derivative of the mean model $\mu(f)$ is given by

$$\frac{d}{d\mathbf{f}(x^i)}\mu(f) = I(x^i)\cdot\mathcal{N}\left(\frac{\mathbf{f}(v^k)-\mathbf{f}(u^k)}{\sqrt{2}\sigma}\middle|\, 0, 1\right) \qquad (75)$$

Inserting equation (74) and equation (75) into equation (72) yields

$$\begin{aligned}
&\frac{d}{d\mathbf{f}(x^i)}\left[\log p(\pi_k \mid \mathbf{f}(u^k), \mathbf{f}(v^k))\right] \\
&\quad= I(x^i)\cdot\nu\cdot\mathcal{N}\left(\frac{\mathbf{f}(v^k)-\mathbf{f}(u^k)}{\sqrt{2}\sigma}\middle|\, 0, 1\right) \\
&\quad\cdot\left(\log(\pi_k) - \log(1-\pi_k) - \psi(\alpha(f)) + \psi(\beta(f))\right) \qquad (76)
\end{aligned}$$

A.1.2.2 Second derivative of log-likelihood

$$
\begin{aligned}
&\frac{d^2}{d\mathbf{f}(x^i)d\mathbf{f}(x^j)}\left[\log p(\pi_k \mid \mathbf{f}(u^k), \mathbf{f}(v^k))\right] \\
&\quad = \frac{d^2\alpha(f)}{d\mathbf{f}(x^i)d\mathbf{f}(x^j)}\left[\log(\pi_k) - \psi(\alpha(f)) + \psi(\alpha(f)+\beta(f))\right] \\
&\quad + \frac{d^2\beta(f)}{d\mathbf{f}(x^i)d\mathbf{f}(x^j)}\left[\log(1-\pi_k) - \psi(\beta(f)) + \psi(\alpha(f)+\beta(f))\right] \\
&\quad + \frac{d\alpha(f)}{d\mathbf{f}(x^i)}\frac{d\alpha(f)}{d\mathbf{f}(x^j)}\left[\psi^{(1)}(\alpha(f)+\beta(f)) - \psi^{(1)}(\alpha(f))\right] \\
&\quad + \frac{d\beta(f)}{d\mathbf{f}(x^i)}\frac{d\beta(f)}{d\mathbf{f}(x^j)}\left[\psi^{(1)}(\alpha(f)+\beta(f)) - \psi^{(1)}(\beta(f))\right] \\
&\quad + \left(\frac{d\alpha(f)}{d\mathbf{f}(x^i)}\frac{d\beta(f)}{d\mathbf{f}(x^j)} + \frac{d\alpha(f)}{d\mathbf{f}(x^j)}\frac{d\beta(f)}{d\mathbf{f}(x^i)}\right)\psi^{(1)}\left(\alpha(f)+\beta(f)\right),
\end{aligned} \tag{77}
$$

where the first derivatives of the shape parameters are given by equation (73) and equation (74) and th second derivatives of the shape parameters are given by

$$
\frac{d^2}{d\mathbf{f}(x^i)d\mathbf{f}(x^j)}\alpha(f) = \nu\frac{d^2}{d\mathbf{f}(x^i)d\mathbf{f}(x^j)}\mu(f) \tag{78}
$$

$$
\frac{d^2}{d\mathbf{f}(x^i)d\mathbf{f}(x^j)}\beta(f) = -\nu\frac{d^2}{d\mathbf{f}(x^i)d\mathbf{f}(x^j)}\mu(f) = -\frac{d^2}{d\mathbf{f}(x^i)d\mathbf{f}(x^j)}\alpha(f), \tag{79}
$$

where the second derivative of the mean model $\mu(f)$ is given by

$$
\begin{aligned}
&\frac{d^2}{d\mathbf{f}(x^i)d\mathbf{f}(x^j)}\mu(f) \\
&\qquad = -I(x^i)I(x^j)\frac{\mathbf{f}(v^k)-\mathbf{f}(u^k)}{2\sigma^2}\cdot\mathcal{N}\left(\frac{\mathbf{f}(v^k)-\mathbf{f}(u^k)}{\sqrt{2}\sigma}\middle|\, 0, 1\right)
\end{aligned} \tag{80}
$$

Inserting equation (74), equation (75), equation (79) and equation (80) into equation (77) yields

$$
\begin{aligned}
&\frac{d^2}{d\mathbf{f}(x^i)d\mathbf{f}(x^j)}\left[\log p(\pi_k \mid \mathbf{f}(u^k), \mathbf{f}(v^k))\right] \\
&\qquad = -I(x^i)I(x^j)\cdot\nu^2\cdot\mathcal{N}\left(\frac{\mathbf{f}(v^k)-\mathbf{f}(u^k)}{\sqrt{2}\sigma}\middle|\, 0, 1\right) \\
&\qquad \cdot\left[\mathcal{N}\left(\frac{\mathbf{f}(v^k)-\mathbf{f}(u^k)}{\sqrt{2}\sigma}\middle|\, 0, 1\right)\cdot\left(\psi^{(1)}(\alpha(f)) + \psi^{(1)}(\alpha(f))\right)\right. \\
&\qquad \left. + \frac{\mathbf{f}(v^k)-\mathbf{f}(u^k)}{2\nu\sigma^2}\left(\log(\pi_k) - \log(1-\pi_k) - \psi(\alpha(f)) + \psi(\beta(f))\right)\right]
\end{aligned} \tag{81}
$$

A.1.2.3 Third derivative of log-likelihood

$$
\begin{aligned}
&\frac{d^3}{d\mathbf{f}(x^i)d\mathbf{f}(x^j)d\mathbf{f}(x^l)}\left[\log p(\pi_k \mid \mathbf{f}(u^k), \mathbf{f}(v^k))\right] \\
&= \frac{d^3\alpha\left(f\right)}{d\mathbf{f}(x^i)d\mathbf{f}(x^j)d\mathbf{f}(x^l)}\left[\log(\pi_k) - \psi(\alpha(f)) + \psi(\alpha(f) + \beta(f))\right] \\
&+ \frac{d^3\beta\left(f\right)}{d\mathbf{f}(x^i)d\mathbf{f}(x^j)d\mathbf{f}(x^l)}\left[\log(1-\pi_k) - \psi(\beta(f)) + \psi(\alpha(f) + \beta(f))\right] \\
&+ \left(\frac{d^2\alpha\left(f\right)}{d\mathbf{f}(x^i)d\mathbf{f}(x^j)}\frac{d\alpha\left(f\right)}{d\mathbf{f}(x^l)} + \frac{d^2\alpha\left(f\right)}{d\mathbf{f}(x^i)d\mathbf{f}(x^l)}\frac{d\alpha\left(f\right)}{d\mathbf{f}(x^j)} + \frac{d^2\alpha\left(f\right)}{d\mathbf{f}(x^j)d\mathbf{f}(x^l)}\frac{d\alpha\left(f\right)}{d\mathbf{f}(x^i)}\right) \\
&\cdot \left[\psi^{(1)}\left(\alpha(f) + \beta(f)\right) - \psi^{(1)}\left(\alpha(f)\right)\right] \\
&+ \left(\frac{d^2\beta\left(f\right)}{d\mathbf{f}(x^i)d\mathbf{f}(x^j)}\frac{d\beta\left(f\right)}{d\mathbf{f}(x^l)} + \frac{d^2\beta\left(f\right)}{d\mathbf{f}(x^i)d\mathbf{f}(x^l)}\frac{d\beta\left(f\right)}{d\mathbf{f}(x^j)} + \frac{d^2\beta\left(f\right)}{d\mathbf{f}(x^j)d\mathbf{f}(x^l)}\frac{d\beta\left(f\right)}{d\mathbf{f}(x^i)}\right) \\
&\cdot \left[\psi^{(1)}\left(\alpha(f) + \beta(f)\right) - \psi^{(1)}\left(\beta(f)\right)\right] \\
&+ \left(\frac{d^2\alpha\left(f\right)}{d\mathbf{f}(x^i)d\mathbf{f}(x^j)}\frac{d\beta\left(f\right)}{d\mathbf{f}(x^l)} + \frac{d^2\alpha\left(f\right)}{d\mathbf{f}(x^i)d\mathbf{f}(x^l)}\frac{d\beta\left(f\right)}{d\mathbf{f}(x^j)} + \frac{d^2\alpha\left(f\right)}{d\mathbf{f}(x^j)d\mathbf{f}(x^l)}\frac{d\beta\left(f\right)}{d\mathbf{f}(x^i)}\right) \\
&\cdot \left[\psi^{(1)}\left(\alpha(f) + \beta(f)\right)\right] \\
&+ \left(\frac{d^2\beta\left(f\right)}{d\mathbf{f}(x)d\mathbf{f}(x^j)}\frac{d\alpha\left(f\right)}{d\mathbf{f}(x^l)} + \frac{d^2\beta\left(f\right)}{d\mathbf{f}(x^i)d\mathbf{f}(x^l)}\frac{d\alpha\left(f\right)}{d\mathbf{f}(x^j)} + \frac{d^2\beta\left(f\right)}{d\mathbf{f}(x^j)d\mathbf{f}(x^l)}\frac{d\alpha\left(f\right)}{d\mathbf{f}(x^i)}\right) \\
&\cdot \left[\psi^{(1)}\left(\alpha(f) + \beta(f)\right)\right] \\
&+ \frac{d\alpha(f)}{d\mathbf{f}(x^i)}\frac{d\alpha(f)}{d\mathbf{f}(x^j)}\frac{d\alpha(f)}{d\mathbf{f}(x^l)}\left[\psi^{(2)}\left(\alpha(f) + \beta(f)\right) - \psi^{(2)}\left(\alpha(f)\right)\right] \\
&+ \frac{d\beta(f)}{d\mathbf{f}(x^i)}\frac{d\beta(f)}{d\mathbf{f}(x^j)}\frac{d\beta(f)}{d\mathbf{f}(x^l)}\left[\psi^{(2)}\left(\alpha(f) + \beta(f)\right) - \psi^{(2)}\left(\beta(f)\right)\right] \\
&+ \left(\frac{d\alpha(f)}{d\mathbf{f}(x^i)}\frac{d\alpha(f)}{d\mathbf{f}(x^j)}\frac{d\beta(f)}{d\mathbf{f}(x^l)} + \frac{d\alpha(f)}{d\mathbf{f}(x^i)}\frac{d\alpha(f)}{d\mathbf{f}(x^l)}\frac{d\beta(f)}{d\mathbf{f}(x^j)} + \frac{d\alpha(f)}{d\mathbf{f}(x^j)}\frac{d\alpha(f)}{d\mathbf{f}(x^l)}\frac{d\beta(f)}{d\mathbf{f}(x^i)}\right) \\
&\cdot \left[\psi^{(2)}\left(\alpha(f) + \beta(f)\right)\right] \\
&+ \left(\frac{d\beta(f)}{d\mathbf{f}(x^i)}\frac{d\beta(f)}{d\mathbf{f}(x^j)}\frac{d\alpha(f)}{d\mathbf{f}(x^l)} + \frac{d\beta(f)}{d\mathbf{f}(x^i)}\frac{d\beta(f)}{d\mathbf{f}(x^l)}\frac{d\alpha(f)}{d\mathbf{f}(x^j)} + \frac{d\beta(f)}{d\mathbf{f}(x^j)}\frac{d\beta(f)}{d\mathbf{f}(x^l)}\frac{d\alpha(f)}{d\mathbf{f}(x^i)}\right) \\
&\cdot \left[\psi^{(2)}\left(\alpha(f) + \beta(f)\right)\right], \qquad (82)
\end{aligned}
$$

where the first derivatives of the shape parameters are given by equation (73) and equation (74), the second derivatives of the shape parameters are given by equation (78) and equation (79) and the third derivatives of the shape parameters are given by

$$\frac{d^3}{d\mathbf{f}(x^i)d\mathbf{f}(x^j)d\mathbf{f}(x^l)}\alpha(f) = \nu\frac{d^3}{d\mathbf{f}(x^i)d\mathbf{f}(x^j)d\mathbf{f}(x^l)}\mu(f) \qquad (83)$$

$$\frac{d^3}{d\mathbf{f}(x^i)d\mathbf{f}(x^j)d\mathbf{f}(x^l)}\beta(f) = -\nu\frac{d^3}{d\mathbf{f}(x^i)d\mathbf{f}(x^j)d\mathbf{f}(x^l)}\mu(f) = -\frac{d^3}{d\mathbf{f}(x^i)d\mathbf{f}(x^j)d\mathbf{f}(x^l)}\alpha(f), \qquad (84)$$

where the third derivative of the mean model $\mu(f)$ is given by

$$\begin{aligned}
&\frac{d^3}{d\mathbf{f}(x^i)d\mathbf{f}(x^j)d\mathbf{f}(x^l)}\mu(f) \\
&\qquad = I(x^i)I(x^j)I(x^l)\frac{1}{2\sigma^2}\left(\frac{\left(\mathbf{f}(v^k)-\mathbf{f}(u^k)\right)^2}{2\sigma^2}-1\right) \\
&\qquad \cdot \mathcal{N}\left(\frac{\mathbf{f}(v^k)-\mathbf{f}(u^k)}{\sqrt{2}\sigma}\middle| 0,1\right)
\end{aligned} \tag{85}$$

Again by making use off all equations, the third derivative of the log likelihood reduces to

$$\begin{aligned}
&\frac{d^3}{d\mathbf{f}(x^i)d\mathbf{f}(x^j)d\mathbf{f}(x^l)}\left[\log p(\pi_k \mid \mathbf{f}(u^k),\mathbf{f}(v^k))\right] \\
&= I(x^j)I(x^i)I(x^l)\cdot\nu^3\cdot\mathcal{N}\left(\frac{f(v^k)-f(u^k)}{\sqrt{2}\sigma}\middle| 0,1\right) \\
&\cdot\left[\frac{1}{2\nu^2\sigma^2}\left(\frac{(\mathbf{f}(v^k)-\mathbf{f}(u^k))^2}{2\sigma^2}-1\right)(\log(\pi_k)-\log(1-\pi_k)-\psi(\alpha(f))+\psi(\beta(f))) \right. \\
&+\frac{3\cdot(\mathbf{f}(v^k)-\mathbf{f}(u^k))}{2\nu\sigma^2}\mathcal{N}\left(\frac{\mathbf{f}(v^k)-\mathbf{f}(u^k)}{\sqrt{2}\sigma}\middle| 0,1\right)\left(\psi^{(1)}(\alpha(f))+\psi^{(1)}(\beta(f))\right) \\
&\left.+\left[\mathcal{N}\left(\frac{\mathbf{f}(v^k)-\mathbf{f}(u^k)}{\sqrt{2}\sigma}\middle| 0,1\right)\right]^2\left(\psi^{(2)}(\beta(f))-\psi^{(2)}(\alpha(f))\right)\right]
\end{aligned} \tag{86}$$

A.1.2.4 First Derivative of the Log-Likelihood With Respect to parameters

ν:

$$\begin{aligned}
&\frac{d}{d\nu}\left[\log p(\pi_k \mid \mathbf{f}(u^k),\mathbf{f}(v^k))\right] \\
&\quad = \frac{d}{d\nu}\left[\log Beta(\pi_k \mid \alpha(f),\beta(f))\right] \\
&\quad = \frac{d\alpha(f)}{d\nu}\left[\log(\pi_k)-\psi(\alpha(f))+\psi(\alpha(f)+\beta(f))\right] \\
&\quad + \frac{d\beta(f)}{d\nu}\left[\log(1-\pi_k)-\psi(\beta(f))+\psi(\alpha(f)+\beta(f))\right],
\end{aligned} \tag{87}$$

where

$$\frac{d\alpha(f)}{d\nu} = \frac{d}{d\nu}\left[\nu\cdot\mu(f)\right] = \mu(f) \tag{88}$$

$$\frac{d\beta(f)}{d\nu} = \frac{d}{d\nu}\left[\nu\left(1-\mu(f)\right)\right] = 1-\mu(f) \tag{89}$$

σ:

$$\begin{aligned}
&\frac{d}{d\sigma}\left[\log p(\pi_k \mid \mathbf{f}(u^k),\mathbf{f}(v^k))\right] \\
&\quad = \frac{d}{d\sigma}\left[\log Beta(\pi_k \mid \alpha(f),\beta(f))\right] \\
&\quad = \frac{d\alpha(f)}{d\sigma}\left[\log(\pi_k)-\psi(\alpha(f))+\psi(\alpha(f)+\beta(f))\right] \\
&\quad + \frac{d\beta(f)}{d\sigma}\left[\log(1-\pi_k)-\psi(\beta(f))+\psi(\alpha(f)+\beta(f))\right],
\end{aligned} \tag{90}$$

where

$$\frac{d\alpha(f)}{d\sigma} = \frac{d}{d\sigma}\left[\nu \cdot \mu(f)\right] = \nu\frac{d\mu(f)}{d\sigma} \tag{91}$$

$$\frac{d\beta(f)}{d\sigma} = \frac{d}{d\sigma}\left[\nu\left(1-\mu(f)\right)\right] = -\nu\frac{d\mu(f)}{d\sigma} = -\frac{d\alpha(f)}{d\sigma}, \tag{92}$$

and

$$\frac{d\mu(f)}{d\sigma} = -\frac{\mathbf{f}(v^k)-\mathbf{f}(u^k)}{\sigma}\mathcal{N}\left(\frac{\mathbf{f}(v^k)-\mathbf{f}(u^k)}{\sqrt{2}\sigma}\middle|0,1\right) \tag{93}$$

Putting everything together yields

$$\begin{aligned}
&\frac{d}{d\sigma}\left[\log p(\pi_k \mid \mathbf{f}(u^k),\mathbf{f}(v^k))\right] \\
&\quad = -\frac{\mathbf{f}(v^k)-\mathbf{f}(u^k)}{\sigma}\mathcal{N}\left(\frac{\mathbf{f}(v^k)-\mathbf{f}(u^k)}{\sqrt{2}\sigma}\middle|0,1\right) \\
&\quad \cdot \left[\log(\pi_k) - \psi(\alpha(f)) - \log(1-\pi_k) + \psi(\beta(f))\right]
\end{aligned} \tag{94}$$

A.1.2.5 First Derivative of the Gradient of the Log-Likelihood with Respect to parameters ν:

$$\begin{aligned}
&\frac{d}{d\nu}\left[\frac{d\log p(\pi_k \mid \mathbf{f}(u^k),\mathbf{f}(v^k))}{d\mathbf{f}(x^i)}\right] \\
&\quad = I(x^i)\cdot\mathcal{N}\left(\frac{\mathbf{f}(v^k)-\mathbf{f}(u^k)}{\sqrt{2}\sigma}\middle|0,1\right) \\
&\quad \cdot\left[\left(\log(\pi_k) - \log(1-\pi_k) - \psi^{(0)}(\alpha(f)) + \psi^{(0)}(\beta(f))\right)\right. \\
&\quad \left. - \nu\cdot\left(\psi^{(1)}(\alpha(f))\mu(\mathbf{f}_k) - \psi^{(1)}(\beta(f))(1-\mu(\mathbf{f}_k))\right)\right]
\end{aligned} \tag{95}$$

σ:

$$\begin{aligned}
&\frac{d}{d\sigma}\left[\frac{d\log p(\pi_k \mid \mathbf{f}(u^k),\mathbf{f}(v^k))}{d\mathbf{f}(x^i)}\right] \\
&\quad = I(x^i)\nu\cdot\mathcal{N}\left(\frac{\mathbf{f}(v^k)-\mathbf{f}(u^k)}{\sqrt{2}\sigma}\middle|0,1\right) \\
&\quad \cdot\left[\left(\frac{\left(\mathbf{f}(v^k)-\mathbf{f}(u^k)\right)^2}{2\sigma^3} - \frac{1}{\sigma}\right)\cdot\left(\log(\pi_k) - \log(1-\pi_k) - \psi^{(0)}(\alpha(f)) + \psi^{(0)}(\beta(f))\right)\right. \\
&\quad \left. - \frac{d\alpha}{d\sigma}\left(\psi^{(1)}(\alpha(f)) + \psi^{(1)}(\beta(f))\right)\right]
\end{aligned} \tag{96}$$

A.1.2.6 First Derivative of Hessian With Respect to Parameters

ν:

$$\begin{aligned}
&\frac{d}{d\nu}\left[\frac{d^2\log p(\pi_k \mid \mathbf{f}(u^k),\mathbf{f}(v^k))}{d\mathbf{f}(x^i)d\mathbf{f}(x^j)}\right] \\
&\quad= \frac{d^2\mu(f)}{d\mathbf{f}(x^i)d\mathbf{f}(x^j)}\cdot\Big[\log(\pi_k)-\log(1-\pi_k)-\psi(\alpha(f))+\psi(\alpha(f))-\alpha(f)\cdot\psi^{(1)}(\alpha(f))+\beta(f)\cdot\psi^{(1)}(\beta(f))\Big] \\
&\quad- 2\nu\frac{d\mu(f)}{d\mathbf{f}(x^i)}\frac{d\mu(f)}{d\mathbf{f}(x^j)}\left[\psi^{(1)}(\alpha(f))+\frac{1}{2}\alpha(f)\cdot\psi^{(2)}(\alpha(f))+\psi^{(1)}(\beta(f))+\frac{1}{2}\beta(f)\cdot\psi^{(2)}(\beta(f))\right]
\end{aligned} \tag{97}$$

σ:

$$\begin{aligned}
&\frac{d}{d\sigma}\left[\frac{d^2\log p(\pi_k \mid \mathbf{f}(u^k),\mathbf{f}(v^k))}{d\mathbf{f}(x^i)d\mathbf{f}(x^j)}\right] \\
&\quad= \frac{d}{d\sigma}\left[\frac{d^2\mu(f)}{d\mathbf{f}(x^i)d\mathbf{f}(x^j)}\right]\cdot\nu\cdot[\log(\pi_k)-\log(1-\pi_k)-\psi(\alpha(f))+\psi(\beta(f))] \\
&\quad- \frac{d^2\mu(f)}{d\mathbf{f}(x^i)d\mathbf{f}(x^j)}\frac{d\mu(f)}{d\sigma}\cdot\nu^2\left[\psi^{(1)}(\alpha(f))+\psi^{(1)}(\beta(f))\right] \\
&\quad- \left(\frac{d}{d\sigma}\left[\frac{d\mu(f)}{d\mathbf{f}(x^i)}\right]\frac{d\mu(f)}{d\mathbf{f}(x^j)}+\frac{d}{d\sigma}\left[\frac{d\mu(f)}{d\mathbf{f}(x^j)}\right]\frac{d\mu(f)}{d\mathbf{f}(x^i)}\right)\nu^2\left[\psi^{(1)}(\alpha(f))+\psi^{(1)}(\beta(f))\right] \\
&\quad- \frac{d\mu(f)}{d\mathbf{f}(x^i)}\frac{d\mu(f)}{d\mathbf{f}(x^j)}\frac{d\mu(f)}{d\sigma}\cdot\nu^3\left[\psi^{(2)}(\alpha(f))-\psi^{(2)}(\beta(f))\right],
\end{aligned} \tag{98}$$

where

$$\begin{aligned}
&\frac{d}{d\sigma}\left[\frac{d^2\mu(f)}{d\mathbf{f}(x^i)d\mathbf{f}(x^j)}\right] \\
&\quad= I(x^i)I(x^j)\frac{\mathbf{f}(v^k)-\mathbf{f}(u^k)}{2\sigma^3}\left(3-\frac{\left(\mathbf{f}(v^k)-\mathbf{f}(u^k)\right)^2}{2\sigma^2}\right)\cdot\mathcal{N}\left(\frac{\mathbf{f}(v^k)-\mathbf{f}(u^k)}{\sqrt{2}\sigma}\middle|0,1\right),
\end{aligned} \tag{99}$$

$$\begin{aligned}
&\frac{d}{d\sigma}\left[\frac{d\mu(f)}{d\mathbf{f}(x^i)}\right] \\
&\quad= I(x^i)\frac{1}{\sigma}\left(\frac{\left(\mathbf{f}(v^k)-\mathbf{f}(u^k)\right)^2}{2\sigma^2}-1\right)\cdot\mathcal{N}\left(\frac{\mathbf{f}(v^k)-\mathbf{f}(u^k)}{\sqrt{2}\sigma}\middle|0,1\right),
\end{aligned} \tag{100}$$

and

$$\frac{d\mu(f)}{d\sigma} = -\frac{\mathbf{f}(v^k)-\mathbf{f}(u^k)}{\sigma}\cdot\mathcal{N}\left(\frac{\mathbf{f}(v^k)-\mathbf{f}(u^k)}{\sqrt{2}\sigma}\middle|0,1\right). \tag{101}$$

A.2 Absolute

A.2.1 Absolute - Continuous - Beta

For ease of understanding and reading, the derivatives of the parameterized log-beta likelihood have been broken down into smaller pieces following section A.1.2. However, since the likelihood function only depends on one input the need for an indicator function keeping track of the inputs, is avoided compared to the pairwise case from section A.1.2.

A.2.1.1 First derivative of log-likelihood

$$\begin{aligned}
&\frac{d}{d\mathbf{f}(x^k)}\left[\log p(\pi_k \mid \mathbf{f}(x^k))\right] \\
&\quad = \frac{d}{d\mathbf{f}(x^k)}\left[\log Beta(\pi_k \mid \alpha(f), \beta(f))\right] \\
&\quad = \frac{d\alpha(f)}{d\mathbf{f}(x^k)}\left[\log(\pi_k) - \psi(\alpha(f)) + \psi(\alpha(f)+\beta(f))\right] \\
&\quad + \frac{d\beta(f)}{d\mathbf{f}(x^k)}\left[\log(1-\pi_k) - \psi(\beta(f)) + \psi(\alpha(f)+\beta(f))\right], \qquad (102)
\end{aligned}$$

where $\alpha(f)$ and $\beta(f)$ are given by equation (10), and ψ is the digamma function (corresponds to the polygamma function of order 0).

The first derivatives of the shape parameters are given by

$$\frac{d}{d\mathbf{f}(x^k)}\alpha(f) = \nu\frac{d}{d\mathbf{f}(x^k)}\mu(f) \qquad (103)$$

$$\frac{d}{d\mathbf{f}(x^k)}\beta(f) = -\nu\frac{d}{d\mathbf{f}(x^k)}\mu(f) = -\frac{d}{d\mathbf{f}(x^k)}\alpha(f), \qquad (104)$$

where the first derivative of the mean model $\mu(f)$ is given by

$$\frac{d}{d\mathbf{f}(x^k)}\mu(f) = \cdot\mathcal{N}\left(\frac{\mathbf{f}(x^k)}{\sqrt{2}\sigma}\middle|\, 0, 1\right) \qquad (105)$$

Inserting equation (104) and equation (105) into equation (102) yields

$$\begin{aligned}
&\frac{d}{d\mathbf{f}(x^k)}\left[\log p(\pi_k \mid \mathbf{f}(x^k))\right] \\
&\quad = \cdot\nu\cdot\mathcal{N}\left(\frac{\mathbf{f}(x^k)}{\sqrt{2}\sigma}\middle|\, 0, 1\right) \\
&\quad \cdot\left(\log(\pi_k) - \log(1-\pi_k) - \psi(\alpha(f)) + \psi(\beta(f))\right) \qquad (106)
\end{aligned}$$

A.2.1.2 Second derivative of log-likelihood

$$
\begin{aligned}
&\frac{d^2}{d\mathbf{f}(x^k)^2}\left[\log p(\pi_k \mid \mathbf{f}(x^k))\right] \\
&\quad = \frac{d^2\alpha(f)}{d\mathbf{f}(x^k)^2}\left[\log(\pi_k) - \psi(\alpha(f)) + \psi(\alpha(f)+\beta(f))\right] \\
&\quad + \frac{d^2\beta(f)}{d\mathbf{f}(x^k)^2}\left[\log(1-\pi_k) - \psi(\beta(f)) + \psi(\alpha(f)+\beta(f))\right] \\
&\quad + \left(\frac{d\alpha(f)}{d\mathbf{f}(x^k)}\right)^2\left[\psi^{(1)}(\alpha(f)+\beta(f)) - \psi^{(1)}(\alpha(f))\right] \\
&\quad + \left(\frac{d\beta(f)}{d\mathbf{f}(x^k)}\right)^2\left[\psi^{(1)}(\alpha(f)+\beta(f)) - \psi^{(1)}(\beta(f))\right] \\
&\quad + 2\frac{d\alpha(f)}{d\mathbf{f}(x^k)}\frac{d\beta(f)}{d\mathbf{f}(x^k)}\psi^{(1)}\left(\alpha(f)+\beta(f)\right),
\end{aligned} \tag{107}
$$

where the first derivatives of the shape parameters are given by equation (10) and the second derivatives of the shape parameters are given by

$$
\frac{d^2}{d\mathbf{f}(x^k)^2}\alpha(f) = \nu\frac{d^2}{d\mathbf{f}(x^k)^2}\mu(f) \tag{108}
$$

$$
\frac{d^2}{d\mathbf{f}(x^k)^2}\beta(f) = -\nu\frac{d^2}{d\mathbf{f}(x^k)^2}\mu(f) = -\frac{d^2}{d\mathbf{f}(x^k)^2}\alpha(f), \tag{109}
$$

where the second derivative of the mean model $\mu(f)$ is given by

$$
\begin{aligned}
&\frac{d^2}{d\mathbf{f}(x^k)^2}\mu(f) \\
&\qquad = -\frac{\mathbf{f}(x^k)}{2\sigma^2}\cdot\mathcal{N}\left(\frac{\mathbf{f}(x^k)}{\sqrt{2}\sigma}\middle|\, 0, 1\right)
\end{aligned} \tag{110}
$$

Inserting equation (104), equation (105), equation (109) and equation (110) into equation (107) yields

$$
\begin{aligned}
&\frac{d^2}{d\mathbf{f}(x^k)^2}\left[\log p(\pi_k \mid \mathbf{f}(x^k))\right] \\
&\quad = -\cdot\nu^2\cdot\mathcal{N}\left(\frac{\mathbf{f}(x^k)}{\sqrt{2}\sigma}\middle|\, 0, 1\right) \\
&\quad \cdot\left[\mathcal{N}\left(\frac{\mathbf{f}(x^k)}{\sqrt{2}\sigma}\middle|\, 0, 1\right)\cdot\left(\psi^{(1)}(\alpha(f)) + \psi^{(1)}(\alpha(f))\right)\right. \\
&\quad \left. + \frac{\mathbf{f}(x^k)}{2\nu\sigma^2}\left(\log(\pi_k) - \log(1-\pi_k) - \psi(\alpha(f)) + \psi(\beta(f))\right)\right]
\end{aligned} \tag{111}
$$

A.2.1.3 Third derivative of log-likelihood

$$
\begin{aligned}
&\frac{d^3}{d\mathbf{f}(x^k)^3}\left[\log p(\pi_k \mid \mathbf{f}(x^k))\right] \\
&= \frac{d^3\alpha\,(f)}{d\mathbf{f}(x^k)^3}\left[\log(\pi_k) - \psi(\alpha(f)) + \psi(\alpha(f)+\beta(f))\right] \\
&+ \frac{d^3\beta\,(f)}{d\mathbf{f}(x^k)^3}\left[\log(1-\pi_k) - \psi(\beta(f)) + \psi(\alpha(f)+\beta(f))\right] \\
&+ 3\frac{d^2\alpha\,(f)}{d\mathbf{f}(x^k)^2}\frac{d\alpha\,(f)}{d\mathbf{f}(x^k)} \cdot \left[\psi^{(1)}\left(\alpha(f)+\beta(f)\right) - \psi^{(1)}\left(\alpha(f)\right)\right] \\
&+ 3\frac{d^2\beta\,(f)}{d\mathbf{f}(x^k)^2}\frac{d\beta\,(f)}{d\mathbf{f}(x^k)} \cdot \left[\psi^{(1)}\left(\alpha(f)+\beta(f)\right) - \psi^{(1)}\left(\beta(f)\right)\right] \\
&+ 3\frac{d^2\alpha\,(f)}{d\mathbf{f}(x^k)^2}\frac{d\beta\,(f)}{d\mathbf{f}(x^k)} \cdot \left[\psi^{(1)}\left(\alpha(f)+\beta(f)\right)\right] \\
&+ 3\frac{d^2\beta\,(f)}{d\mathbf{f}(x^k)^2}\frac{d\alpha\,(f)}{d\mathbf{f}(x^k)} \cdot \left[\psi^{(1)}\left(\alpha(f)+\beta(f)\right)\right] \\
&+ \left(\frac{d\alpha(f)}{d\mathbf{f}(x^k)}\right)^3 \left[\psi^{(2)}\left(\alpha(f)+\beta(f)\right) - \psi^{(2)}\left(\alpha(f)\right)\right] \\
&+ \left(\frac{d\beta(f)}{d\mathbf{f}(x^k)}\right)^3 \left[\psi^{(2)}\left(\alpha(f)+\beta(f)\right) - \psi^{(2)}\left(\beta(f)\right)\right] \\
&+ 3\left(\frac{d\alpha(f)}{d\mathbf{f}(x^k)}\right)^2 \frac{d\beta(f)}{d\mathbf{f}(x^k)} \cdot \left[\psi^{(2)}\left(\alpha(f)+\beta(f)\right)\right] \\
&+ 3\left(\frac{d\beta(f)}{d\mathbf{f}(x^k)}\right)^2 \frac{d\alpha(f)}{d\mathbf{f}(x^k)} \cdot \left[\psi^{(2)}\left(\alpha(f)+\beta(f)\right)\right], \qquad (112)
\end{aligned}
$$

where the first derivatives of the shape parameters are given by equation (73) and equation (74), the second derivatives of the shape parameters are given by equation (78) and equation (79) and the third derivatives of the shape parameters are given by

$$\frac{d^3}{d\mathbf{f}(x^k)^3}\alpha(f) = \nu\frac{d^3}{d\mathbf{f}(x^k)^3}\mu(f) \qquad (113)$$

$$\frac{d^3}{d\mathbf{f}(x^k)^3}\beta(f) = -\nu\frac{d^3}{d\mathbf{f}(x^k)^3}\mu(f) = -\frac{d^3}{d\mathbf{f}(x^k)^3}\alpha(f), \qquad (114)$$

where the third derivative of the mean model $\mu(f)$ is given by

$$
\begin{aligned}
&\frac{d^3}{d\mathbf{f}(x^k)^3}\mu(f) \\
&\qquad = \frac{1}{2\sigma^2}\left(\frac{\mathbf{f}(x^k)^2}{2\sigma^2} - 1\right) \\
&\qquad \cdot \mathcal{N}\left(\frac{\mathbf{f}(x^k)}{\sqrt{2}\sigma}\middle| 0, 1\right) \qquad (115)
\end{aligned}
$$

Again by making use off all equations, the third derivative of the log likelihood reduces to

$$\begin{aligned}
&\frac{d^3}{d\mathbf{f}(x^k)^3}\left[\log p(\pi_k \mid \mathbf{f}(x^k))\right] \\
&= \cdot\nu^3 \cdot \mathcal{N}\left(\frac{\mathbf{f}(x^k)}{\sqrt{2}\sigma}\middle| 0,1\right) \\
&\cdot \left[\frac{1}{2\nu^2\sigma^2}\left(\frac{\mathbf{f}(x^k)^2}{2\sigma^2}-1\right)(\log(\pi_k)-\log(1-\pi_k)-\psi(\alpha(f))+\psi(\beta(f))\right) \\
&+ \frac{3\cdot\mathbf{f}(x^k)}{2\nu\sigma^2}\mathcal{N}\left(\frac{\mathbf{f}(x^k)}{\sqrt{2}\sigma}\middle| 0,1\right)\left(\psi^{(1)}(\alpha(f))+\psi^{(1)}(\beta(f))\right) \\
&+ \left[\mathcal{N}\left(\frac{\mathbf{f}(x^k)}{\sqrt{2}\sigma}\middle| 0,1\right)\right]^2\left(\psi^{(2)}(\beta(f))-\psi^{(2)}(\alpha(f))\right)\Bigg] \qquad (116)
\end{aligned}$$

A.2.1.4 First Derivative of the Log-Likelihood With Respect to parameters

ν:

$$\begin{aligned}
&\frac{d}{d\nu}\left[\log p(\pi_k \mid \mathbf{f}(x^k))\right] \\
&\quad= \frac{d}{d\nu}\left[\log Beta(\pi_k \mid \alpha(f),\beta(f))\right] \\
&\quad= \frac{d\alpha(f)}{d\nu}\left[\log(\pi_k)-\psi(\alpha(f))+\psi(\alpha(f)+\beta(f))\right] \\
&\quad+ \frac{d\beta(f)}{d\nu}\left[\log(1-\pi_k)-\psi(\beta(f))+\psi(\alpha(f)+\beta(f))\right], \qquad (117)
\end{aligned}$$

where

$$\frac{d\alpha(f)}{d\nu} = \frac{d}{d\nu}\left[\nu\cdot\mu(f)\right] = \mu(f) \qquad (118)$$

$$\frac{d\beta(f)}{d\nu} = \frac{d}{d\nu}\left[\nu\left(1-\mu(f)\right)\right] = 1-\mu(f) \qquad (119)$$

σ:

$$\begin{aligned}
&\frac{d}{d\sigma}\left[\log p(\pi_k \mid \mathbf{f}(x^k))\right] \\
&\quad= \frac{d}{d\sigma}\left[\log Beta(\pi_k \mid \alpha(f),\beta(f))\right] \\
&\quad= \frac{d\alpha(f)}{d\sigma}\left[\log(\pi_k)-\psi(\alpha(f))+\psi(\alpha(f)+\beta(f))\right] \\
&\quad+ \frac{d\beta(f)}{d\sigma}\left[\log(1-\pi_k)-\psi(\beta(f))+\psi(\alpha(f)+\beta(f))\right], \qquad (120)
\end{aligned}$$

where

$$\frac{d\alpha(f)}{d\sigma} = \frac{d}{d\sigma}\left[\nu\cdot\mu(f)\right] = \nu\frac{d\mu(f)}{d\sigma} \qquad (121)$$

$$\frac{d\beta(f)}{d\sigma} = \frac{d}{d\sigma}\left[\nu\left(1-\mu(f)\right)\right] = -\nu\frac{d\mu(f)}{d\sigma} = -\frac{d\alpha(f)}{d\sigma}, \qquad (122)$$

and

$$\frac{d\mu(f)}{d\sigma} = -\frac{\mathbf{f}(x^k)}{\sigma}\mathcal{N}\left(\frac{\mathbf{f}(x^k)}{\sqrt{2}\sigma}\middle|0,1\right) \tag{123}$$

Putting everything together yields

$$\begin{aligned}
&\frac{d}{d\sigma}\left[\log p(\pi_k \mid \mathbf{f}(x^k))\right] \\
&\quad = -\frac{\mathbf{f}(x^k)}{\sigma}\mathcal{N}\left(\frac{\mathbf{f}(x^k)}{\sqrt{2}\sigma}\middle|0,1\right) \\
&\quad \cdot \left[\log(\pi_k) - \psi(\alpha(f)) - \log(1-\pi_k) + \psi(\beta(f))\right]
\end{aligned} \tag{124}$$

A.2.1.5 First Derivative of Hessian With Respect to Parameters :

ν:

$$\begin{aligned}
&\frac{d}{d\nu}\left[\frac{d^2\log p(\pi_k \mid \mathbf{f}(x^k))}{d\mathbf{f}(x^k)^2}\right] \\
&\quad = \frac{d^2\mu(f)}{d\mathbf{f}(x^k)^2}\cdot\left[\log(\pi_k) - \log(1-\pi_k) - \psi(\alpha(f)) + \psi(\alpha(f)) - \alpha(f)\cdot\psi^{(1)}(\alpha(f)) + \beta(f)\cdot\psi^{(1)}(\beta(f))\right] \\
&\quad - 2\nu\left(\frac{d\mu(f)}{d\mathbf{f}(x^k)}\right)^2\left[\psi^{(1)}(\alpha(f)) + \frac{1}{2}\alpha(f)\cdot\psi^{(2)}(\alpha(f)) + \psi^{(1)}(\beta(f)) + \frac{1}{2}\beta(f)\cdot\psi^{(2)}(\beta(f))\right]
\end{aligned} \tag{125}$$

σ:

$$\begin{aligned}
&\frac{d}{d\sigma}\left[\frac{d^2\log p(\pi_k \mid \mathbf{f}(u^k))}{d\mathbf{f}(x^k)^2}\right] \\
&\quad = \frac{d}{d\sigma}\left[\frac{d^2\mu(f)}{d\mathbf{f}(x^k)^2}\right]\cdot\nu\cdot\left[\log(\pi_k) - \log(1-\pi_k) - \psi(\alpha(f)) + \psi(\beta(f))\right] \\
&\quad - \frac{d^2\mu(f)}{d\mathbf{f}(x^k)^2}\frac{d\mu(f)}{d\sigma}\cdot\nu^2\left[\psi^{(1)}(\alpha(f)) + \psi^{(1)}(\beta(f))\right] \\
&\quad - 2\frac{d}{d\sigma}\left[\frac{d\mu(f)}{d\mathbf{f}(x^k)}\right]\frac{d\mu(f)}{d\mathbf{f}(x^k)}\nu^2\left[\psi^{(1)}(\alpha(f)) + \psi^{(1)}(\beta(f))\right] \\
&\quad - \left(\frac{d\mu(f)}{d\mathbf{f}(x^k)}\right)^2\frac{d\mu(f)}{d\sigma}\cdot\nu^3\left[\psi^{(2)}(\alpha(f)) - \psi^{(2)}(\beta(f))\right],
\end{aligned} \tag{126}$$

where

$$\begin{aligned}
&\frac{d}{d\sigma}\left[\frac{d^2\mu(f)}{d\mathbf{f}(x^k)^2}\right] \\
&\quad = \frac{\mathbf{f}(x^k)}{2\sigma^3}\left(3 - \frac{\mathbf{f}(x^k)^2}{2\sigma^2}\right)\cdot\mathcal{N}\left(\frac{\mathbf{f}(x^k)}{\sqrt{2}\sigma}\middle|0,1\right),
\end{aligned} \tag{127}$$

$$\begin{aligned}
&\frac{d}{d\sigma}\left[\frac{d\mu(f)}{d\mathbf{f}(x^k)}\right] \\
&\quad = \frac{1}{\sigma}\left(\frac{\mathbf{f}(x^k)^2}{2\sigma^2} - 1\right)\cdot\mathcal{N}\left(\frac{\mathbf{f}(x^k)}{\sqrt{2}\sigma}\middle|0,1\right),
\end{aligned} \tag{128}$$

and

$$\frac{d\mu(f)}{d\sigma} = -\frac{\mathbf{f}(x^k)}{\sigma}\cdot\mathcal{N}\left(\frac{\mathbf{f}(x^k)}{\sqrt{2}\sigma}\middle|0,1\right). \tag{129}$$

B Laplace Approximation - Details and Derivation

This presentation follows an initial technical report [9] and is based on the basic Laplace approximation for GP classifiers presented in e.g. [18].

B.1 Posterior Approximation

Posterior Approximatio

B.1.1 Moment Matching: Mode and Covariance

:

B.1.1.1 Prior

$$p\left(\mathbf{f}|\mathcal{X}\right)=\mathcal{N}\left(\mathbf{f}|0,\Sigma\left(\mathcal{X}\right)\right)=\frac{1}{\left(2\pi\right)^{n/2}|\Sigma|^{1/2}}e^{-\frac{1}{2}\mathbf{f}^{T}\Sigma^{-1}\mathbf{f}}=\frac{1}{Z}e^{-\frac{1}{2}\mathbf{f}^{T}\Sigma^{-1}\mathbf{f}} \tag{130}$$

B.1.1.2 Log-Prior

$$\begin{aligned}\log p\left(\mathbf{f}|\mathcal{X}\right) &= \log\mathcal{N}\left(\mathbf{f}|0,\Sigma\right) &(131)\\ &= \log e^{-\frac{1}{2}\mathbf{f}^{T}\Sigma^{-1}\mathbf{f}}+\log\frac{1}{Z} &(132)\\ &= -\frac{1}{2}\mathbf{f}^{T}\Sigma^{-1}\mathbf{f}-\frac{1}{2}\log|\Sigma|-\frac{N}{2}\log 2\pi &(133)\end{aligned}$$

B.1.1.3 Log-Posterior

$$\begin{aligned}\log p\left(\mathbf{f}|\mathcal{X}\right) &= \log\frac{p\left(\mathcal{Y}|\mathcal{X},\mathbf{f}\right)p(\mathbf{f}|\mathcal{X})}{p\left(\mathcal{Y}|\mathcal{X}\right)} &(134)\\ &= \log\frac{1}{Z}p\left(\mathcal{Y}|\mathcal{X},\mathbf{f}\right)p(\mathbf{f}|\mathcal{X}) &(135)\\ &= \log\frac{1}{Z}+\log p\left(\mathcal{Y}|\mathcal{X},\mathbf{f}\right)+\log p(\mathbf{f}|\mathcal{X}) &(136)\\ \log Z\,p\left(\mathbf{f}|\mathcal{X},\mathcal{Y}\right) &= \log p\left(\mathcal{Y}|\mathcal{X},\mathbf{f}\right)+\log p(\mathbf{f}|\mathcal{X}) &(137)\\ \psi\left(\mathbf{f}|\mathcal{D}\right) &= \psi\left(\mathbf{f}|\mathcal{X},\mathcal{Y}\right)=\log p\left(\mathcal{Y}|\mathcal{X},\mathbf{f}\right)+\log p(\mathbf{f}|\mathcal{X}) &(138)\end{aligned}$$

B.1.1.4 Cost Function

$$\psi\left(\mathbf{f}|\mathcal{X},\mathcal{Y}\right)=\underbrace{\log p\left(\mathcal{Y}|\mathbf{f}\right)}_{likelihood factorizes}\underbrace{-\frac{1}{2}\mathbf{f}^{T}\Sigma^{-1}\mathbf{f}-\frac{1}{2}\log|\Sigma|-\frac{N}{2}\log 2\pi}_{prior} \tag{139}$$

B.1.1.5 Cost Derivatives: First

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{f}}\psi\left(\mathbf{f}|\mathcal{D}\right) &= \left[\frac{\partial}{\partial f(x_1)}\psi\left(\mathbf{f}|\mathcal{Y}\right), \frac{\partial}{\partial f(x_2)}\psi\left(\mathbf{f}|\mathcal{Y}\right), .., \frac{\partial}{\partial f(x_n)}\psi\left(\mathbf{f}|\mathcal{Y}\right)\right]^{\top} && (140)\\
&= \frac{\partial}{\partial \mathbf{f}}\left(\log p\left(\mathcal{Y}|\mathbf{f}\right) - \frac{1}{2}\mathbf{f}^T\mathbf{\Sigma}^{-1}\mathbf{f} - \frac{1}{2}\log|\mathbf{\Sigma}| - \frac{N}{2}\log 2\pi\right) && (141)\\
&= \frac{\partial}{\partial \mathbf{f}}\log p\left(\mathcal{Y}|\mathbf{f}\right) - \frac{\partial}{\partial \mathbf{f}}\left(\frac{1}{2}\mathbf{f}^T\mathbf{\Sigma}^{-1}\mathbf{f}\right) - \frac{\partial}{\partial \mathbf{f}}\left(\frac{1}{2}\log|\mathbf{\Sigma}|\right) - \frac{\partial}{\partial \mathbf{f}}\left(\frac{N}{2}\log 2\pi\right) && (142)\\
&= \frac{\partial}{\partial \mathbf{f}}\log p\left(\mathcal{Y}|\mathbf{f}\right) - \frac{\partial}{\partial \mathbf{f}}\left(\frac{1}{2}\mathbf{f}^T\mathbf{\Sigma}^{-1}\mathbf{f}\right) && (143)\\
&= \frac{\partial}{\partial \mathbf{f}}\log p\left(\mathcal{Y}|\mathbf{f}\right) - \mathbf{\Sigma}^{-1}\mathbf{f} && (144)\\
&= \left[\frac{\partial}{\partial \mathbf{f}}\log\prod_{k=1}^{m} p\left(d_k|f\left(u_k\right), f\left(u_k\right)\right)\right] - \mathbf{\Sigma}^{-1}\mathbf{f} && (145)\\
&= \left[\frac{\partial}{\partial \mathbf{f}}\sum_{k=1}^{m}\log p\left(d_k|f\left(u_k\right), f\left(u_k\right)\right)\right] - \mathbf{\Sigma}^{-1}\mathbf{f} && (146)\\
&= \left[\sum_{k=1}^{m}\frac{\partial}{\partial \mathbf{f}}\log p\left(d_k|f\left(u_k\right), f\left(u_k\right)\right)\right] - \mathbf{\Sigma}^{-1}\mathbf{f} && (147)\\
&= \left[\sum_{k=1}^{m}\mathbf{z}_k\right] - \mathbf{\Sigma}^{-1}\mathbf{f} && (148)\\
&= \nabla\log p\left(\mathcal{Y}|\mathbf{f}\right) - \mathbf{\Sigma}^{-1}\mathbf{f} && (149)
\end{aligned}$$

B.1.1.6 Cost Derivative: Second

$$\begin{aligned}
\mathbf{f} &= \left[f\left(x_1\right), f\left(x_1\right), .., , f\left(x_n\right)\right] && (150)\\
\mathbf{f}_k &= \left[f\left(u_k\right), f\left(v_k\right)\right] && (151)\\
& && (152)
\end{aligned}$$

where the u_k and v_k can be any of the input instances $x \in \mathcal{X}$.

It is again easer to work directly with the full vectors and matrices such than an element in the final seconder Hessian (of the cost) is

$$\nabla_i\nabla_j\psi\left(\mathbf{f}|D\right) = \left[\frac{\partial^2}{\partial \mathbf{f}\partial \mathbf{f}}\psi\left(\mathbf{f}|\mathcal{D}\right)\right]_{i,j} = \frac{\partial^2}{\partial f\left(x^i\right)\partial f\left(x^j\right)}\psi\left(\mathbf{f}|\mathcal{D}\right)$$

Thus differentiation twice with respect to the vector f gives us.

$$\begin{aligned}
\nabla\nabla\psi\left(\mathbf{f}|D\right) &= \frac{\partial^2}{\partial\mathbf{f}\partial\mathbf{f}}\psi\left(\mathbf{f}|\mathcal{D}\right) && (153)\\
&= \frac{\partial}{\partial\mathbf{f}}\left[\frac{\partial}{\partial\mathbf{f}}\psi\left(\mathbf{f}|\mathcal{D}\right)\right] && (154)\\
&= \frac{\partial}{\partial\mathbf{f}}\left[\left[\sum_{k=1}^{m}\frac{\partial}{\partial\mathbf{f}}\log p\left(d_k|f\left(u_k\right),f\left(u_k\right)\right)\right]-\boldsymbol{\Sigma}^{-1}\mathbf{f}\right] && (155)\\
&= \frac{\partial}{\partial\mathbf{f}}\sum_{k=1}^{m}\frac{\partial}{\partial\mathbf{f}}\log p\left(d_k|f\left(u_k\right),f\left(u_k\right)\right)-\frac{\partial}{\partial\mathbf{f}}\boldsymbol{\Sigma}^{-1}\mathbf{f} && (156)\\
&= \sum_{k=1}^{m}\frac{\partial}{\partial\mathbf{f}}\left[\frac{\partial}{\partial\mathbf{f}}\log p\left(d_k|f\left(u_k\right),f\left(u_k\right)\right)\right]-\boldsymbol{\Sigma}^{-1} && (157)\\
&= \sum_{k=1}^{m}\frac{\partial}{\partial\mathbf{f}}\left[\mathbf{z}_k\right]-\boldsymbol{\Sigma}^{-1} && (158)\\
&= \sum_{k=1}^{m}\nabla\nabla\log p\left(d_k|\mathbf{f}_k\right)-\boldsymbol{\Sigma}^{-1} && (159)\\
&= \sum_{k=1}^{m}\mathbf{Z}_k-\boldsymbol{\Sigma}^{-1} && (160)\\
&= -\mathbf{W}-\boldsymbol{\Sigma}^{-1} && (161)
\end{aligned}$$

where we have defined

$$\begin{aligned}
\mathbf{W} &= -\sum_{k=1}^{m}\mathbf{Z}_k = -\sum_{k=1}^{m}\nabla\nabla\log p\left(d_k|\mathbf{f}_k\right) = -\nabla\nabla\log p\left(\mathcal{D}|\mathbf{f}\right) && (162)\\
\mathbf{W}_{i,j} &= -\sum_{k=1}^{m}\frac{\partial^2}{\partial f\left(x^i\right)\partial f\left(x^j\right)}\log p\left(d_k|f\left(u_k\right),f\left(v_k\right)\right) && (163)
\end{aligned}$$

† we note that the individual terms in the sum over the examples, $\frac{\partial^2}{\partial f(x^i)\partial f(x^j)}\log p\left(d_k|f\left(u_k\right),f\left(v_k\right)\right)$, will be non-zero only if x_i is the u_k or v_k - x_j is the u_k or v_k. Thus only four terms will (at most) be non-zero in $\mathbf{Z}_k$ (the derivatives in respect to each observation); two in the diagonal (when $x_i = x_j = u_k$ and $x_i = x_j = v_k$), and to off diagonal (when $x_i = u_k, x_j = v_k$ and $x_i = v_k, x_j = u_k$).

For implementation it is interesting to write out the matrix, i.e.

$$\mathbf{W} = -\begin{bmatrix}
\sum_{k=1}^{m}\frac{\partial^2}{\partial f(x^1)\partial f(x^1)}\log p\left(d_k|\mathbf{f}_k\right) & \sum_{k=1}^{m}\frac{\partial^2}{\partial f(x^1)\partial f(x^2)}\log p\left(d_k|\mathbf{f}_k\right) & & \\
\sum_{k=1}^{m}\frac{\partial^2}{\partial f(x^2)\partial f(x^1)}\log p\left(d_k|\mathbf{f}_k\right) & & & \\
 & & & \sum_{k=1}^{m}\frac{\partial^2}{\partial f(x^n)\partial f(x^n)}\log p\left(d_k|\mathbf{f}_k\right)
\end{bmatrix}$$

B.1.2 Optimization Method

B.1.2.1 Newton Given the required derivatives we can use a standard gradient approach. The standard Newton Step can be derived as (see e.g. [18])

$$\begin{aligned}
\mathbf{f}^{new} &= \mathbf{f} - (\nabla\nabla\psi)^{-1}\nabla\psi && (164)\\
&= \mathbf{f} - \frac{\nabla\psi}{\nabla\nabla\psi} && (165)\\
&= \mathbf{f} + \left(\mathbf{K}^{-1} + \mathbf{W}\right)^{-1}\left(\nabla \log p(\mathcal{D}|\mathbf{f}) - \mathbf{K}^{-1}\mathbf{f}\right) && (166)\\
&= \mathbf{f} + \left(\mathbf{K}^{-1} + \mathbf{W}\right)^{-1}\nabla \log p(\mathcal{D}|\mathbf{f}) - \left(\mathbf{K}^{-1} + \mathbf{W}\right)^{-1}\mathbf{K}^{-1}\mathbf{f} && (167)\\
&= \left(I - \left(\mathbf{K}^{-1} + \mathbf{W}\right)^{-1}\mathbf{K}^{-1}\right)\mathbf{f} + \left(\mathbf{K}^{-1} + \mathbf{W}\right)^{-1}\nabla \log p(\mathcal{D}|\mathbf{f}) && (168)\\
&= \left(\mathbf{K}^{-1} + \mathbf{W}\right)^{-1}\left(\left(\mathbf{K}^{-1} + \mathbf{W}\right) - \mathbf{K}^{-1}\right)\mathbf{f} + \left(\mathbf{K}^{-1} + \mathbf{W}\right)^{-1}\nabla \log p(\mathcal{D}|\mathbf{f}) && (169)\\
&= \left(\mathbf{K}^{-1} + \mathbf{W}\right)^{-1}(\mathbf{W})\,\mathbf{f} + \left(\mathbf{K}^{-1} + \mathbf{W}\right)^{-1}\nabla \log p(\mathcal{D}|\mathbf{f}) && (170)\\
&= \left(\mathbf{K}^{-1} + \mathbf{W}\right)^{-1}(\mathbf{W}\mathbf{f} + \nabla \log p(\mathcal{D}|\mathbf{f})) && (171)\\
&= \left(\mathbf{K}^{-1} + \mathbf{W}\right)^{-1}(\mathbf{W}\mathbf{f} + \nabla \log p(\mathcal{D}|\mathbf{f})) && (172)
\end{aligned}$$

B.1.2.2 Damped Newton However, it has been found optimal, i.e., more efficient and faster, to use a slightly more involved so-called damped Newton step, which follows the Levenberg-Marq. (LM) procedure for controlling the damping factor λ. This can in our case also be derived without direct inversion of the Hessian.

$$\begin{aligned}
\mathbf{f}^{new} &= \mathbf{f} - (\nabla\nabla\psi + \lambda I)^{-1}(\nabla\psi) && (173)\\
&= \mathbf{f} - (\nabla\nabla\psi + \lambda I)^{-1}(\nabla\psi) && (174)\\
&= \mathbf{f} - \left(-\mathbf{W} - \mathbf{K}^{-1} + \lambda I\right)^{-1}\left(\nabla \log p(\mathcal{D}|\mathbf{f}) - \mathbf{K}^{-1}\mathbf{f}\right) && (175)\\
&= \mathbf{f} + \left(\mathbf{K}^{-1} + \mathbf{W} - \lambda I\right)^{-1}\left(\nabla \log p(\mathcal{D}|\mathbf{f}) - \mathbf{K}^{-1}\mathbf{f}\right) && (176)\\
&= \mathbf{f} - \left(\mathbf{K}^{-1} + \mathbf{W} - \lambda I\right)^{-1}\mathbf{K}^{-1}\mathbf{f} + \left(\mathbf{K}^{-1} + \mathbf{W} - \lambda I\right)^{-1}\nabla \log p(\mathcal{D}|\mathbf{f}) && (177)\\
&= \left(\mathbf{I} - \left(\mathbf{K}^{-1} + \mathbf{W} - \lambda I\right)^{-1}\mathbf{K}^{-1}\right)\mathbf{f} + \left(\mathbf{K}^{-1} + \mathbf{W} - \lambda I\right)^{-1}\nabla \log p(\mathcal{D}|\mathbf{f}) && (178)
\end{aligned}$$

However, since

$$\begin{aligned}
\left(\mathbf{I} - \left(\mathbf{K}^{-1} + \mathbf{W} - \lambda I\right)^{-1}\mathbf{K}^{-1}\right)\mathbf{f} &= \left(\mathbf{K}^{-1} + \mathbf{W} - \lambda I\right)^{-1}\left(\left(\mathbf{K}^{-1} + \mathbf{W} - \lambda I\right) - \mathbf{K}^{-1}\right)\mathbf{f}\\
&= \left(\mathbf{K}^{-1} + \mathbf{W} - \lambda I\right)^{-1}(\mathbf{W} - \lambda I)\,\mathbf{f}, && (179)
\end{aligned}$$

we finally obtain

$$\mathbf{f}^{new} = \left(\mathbf{K}^{-1} + \mathbf{W} - \lambda\mathbf{I}\right)^{-1}\left((\mathbf{W} - \lambda I)\,\mathbf{f} + \nabla \log p(\mathcal{D}|\mathbf{f})\right). \quad (180)$$

In the most advanced case (used in [11]) we combine this method with a Linesearch along the damped Newton direction which results in very faster convergence, even for the more complicated Beta likelihood case.

B.2 Evidence Approximation and Derivatives

The covariance and likelihood function are potentially parameterized by a number of parameters typically referred to as hyper-parameters.

B.2.1 Evidence Approximation

We approximate the model evidence written conveniently as

$$p\left(\mathcal{Y}|\boldsymbol{\theta}\right) \equiv p\left(\mathcal{Y}|\mathcal{X},\boldsymbol{\theta}\right) = \int p\left(\mathcal{Y}|f\right) p\left(f\right) df$$

with a finite Gaussian, which is readily available from our previous approximation of the un-normalized (log) posterior, i.e. with the Laplace approximation

$$\begin{aligned}
p\left(\mathcal{Y}|\mathcal{X},\boldsymbol{\theta}\right) &= \int \exp\left\{\Psi\left(\mathbf{f}\right)\right\} d\mathbf{f}|_{\Psi(\mathbf{f})=\Psi\left(\widehat{\mathbf{f}}\right)-\frac{1}{2}\left(f-\hat{f}\right)A\left(f-\hat{f}\right)} && (181)\\
&= \int \exp\left\{\Psi\left(\widehat{\mathbf{f}}\right) - \frac{1}{2}\left(f-\hat{f}\right)A\left(f-\hat{f}\right)\right\} d\mathbf{f} && (182)\\
&= \exp\left\{\Psi\left(\widehat{\mathbf{f}}\right)\right\} \int -\frac{1}{2}\left(f-\hat{f}\right)A\left(f-\hat{f}\right) d\mathbf{f} && (183)\\
&= \exp\left\{\Psi\left(\widehat{\mathbf{f}}\right)\right\} \sqrt{\left(2\pi\right)^{n}\left|A^{-1}\right|} && (184)\\
&= \exp\left\{\Psi\left(\widehat{\mathbf{f}}\right)\right\} \left(2\pi\right)^{n/2}\left|A^{-1}\right|^{1/2} && (185)
\end{aligned}$$

Now taking the log,

$$\begin{aligned}
\log p\left(\mathcal{Y}|\boldsymbol{\theta}\right) &= \log\left[\exp\left\{\Psi\left(\widehat{\mathbf{f}}\right)\right\}\left(2\pi\right)^{N/2}|A|^{-1/2}\right] && (186)\\
&= \Psi\left(\widehat{\mathbf{f}}\right) + \log\left[\left(2\pi\right)^{N/2}\left|A^{-1}\right|^{1/2}\right] && (187)\\
&= \Psi\left(\widehat{\mathbf{f}}\right) + \frac{N}{2}\log\left(2\pi\right) + \frac{1}{2}\log\left|A^{-1}\right|^{1/2} && (188)\\
&= \log p\left(\mathcal{Y}|\hat{\mathbf{f}}\right) - \frac{1}{2}\hat{\mathbf{f}}^{T}\Sigma^{-1}\hat{\mathbf{f}} - \frac{1}{2}\log|\Sigma| - \frac{N}{2}\log 2\pi + \frac{N}{2}\log\left(2\pi\right) + \frac{1}{2}\log|A| && (189)\\
&= \log p\left(\mathcal{Y}|\hat{\mathbf{f}}\right) - \frac{1}{2}\hat{\mathbf{f}}^{T}\Sigma^{-1}\hat{\mathbf{f}} - \left(\frac{1}{2}\log|\Sigma| - \frac{1}{2}\log\left|A^{-1}\right|\right) && (190)\\
&= \log p\left(\mathcal{Y}|\hat{\mathbf{f}}\right) - \frac{1}{2}\hat{\mathbf{f}}^{T}\Sigma^{-1}\hat{\mathbf{f}} - \frac{1}{2}\log|\Sigma|\,|A| && (191)\\
&= \log p\left(\mathcal{Y}|\hat{\mathbf{f}}\right) - \frac{1}{2}\hat{\mathbf{f}}^{T}\Sigma^{-1}\hat{\mathbf{f}} - \frac{1}{2}\log|\Sigma|\,|A| && (192)\\
w/A = -\nabla\nabla\log p\left(\hat{\mathbf{f}}|\mathcal{Y}\right) &= -\nabla\nabla\log p\left(\mathcal{Y}|\hat{\mathbf{f}}\right) + \Sigma^{-1} = W + \Sigma^{-1} && (193)\\
&= \log p\left(\mathcal{Y}|\hat{\mathbf{f}}\right) - \frac{1}{2}\hat{\mathbf{f}}^{T}\Sigma^{-1}\hat{\mathbf{f}} - \frac{1}{2}\log|\Sigma|\left|W + \Sigma^{-1}\right| && (194)\\
&= \log p\left(\mathcal{Y}|\hat{\mathbf{f}}\right) - \frac{1}{2}\hat{\mathbf{f}}^{T}\Sigma^{-1}\hat{\mathbf{f}} - \frac{1}{2}\log|I_N + \Sigma W| && (195)
\end{aligned}$$

The idea is to apply a standard gradient based approach for finding the maximum of this i.r.t. to each hyperparameter. We therefore consider the derivative of $p(\mathcal{Y}|\theta)$ with regards

to each parameter, i.e. due to the dependence on the parameters we obtain two terms,

$$\frac{\partial \log p\left(\mathcal{Y}|\theta\right)}{\partial \theta_j} = \underbrace{\frac{\partial \log p\left(\mathcal{Y}|\theta\right)}{\partial \theta_j}|_{\exp licit}}_{Term\ A} + \underbrace{\sum_{i=1}^{N} \frac{\partial \log p\left(\mathcal{Y}|\theta\right)}{\partial \hat{f}_i} \frac{\partial \hat{f}_i}{\partial \theta_j}}_{Term\ B} \tag{196}$$

B.2.2 Evidence Derivatives - Covariance Function Parameters

:

B.2.2.1 Term A

$$\begin{aligned}
\frac{\partial \log p\left(\mathcal{Y}|\theta\right)}{\partial \theta_j}\Big|_{\exp licit} &= \frac{\partial}{\partial \theta_j}\left[\log p\left(\mathcal{Y}|\hat{\mathbf{f}}\right) - \frac{1}{2}\hat{\mathbf{f}}^T\sum\nolimits^{-1}\hat{\mathbf{f}} - \frac{1}{2}\log\left|I_N + \Sigma W\right|\right] && (197)\\
&\quad (first\ term\ does\ not\ excplicit\ depends\ on\ theta) && (198)\\
&= \frac{1}{2}\hat{\mathbf{f}}^T\mathbf{\Sigma}^{-1}\frac{\partial \mathbf{\Sigma}}{\theta_j}\mathbf{\Sigma}^{-1}\hat{\mathbf{f}} - \frac{\partial}{\partial \theta_j}\frac{1}{2}\log\left|I_N + \Sigma W\right| && (199)\\
&= \frac{1}{2}\hat{\mathbf{f}}^T\mathbf{\Sigma}^{-1}\frac{\partial \mathbf{\Sigma}}{\theta_j}\mathbf{\Sigma}^{-1}\hat{\mathbf{f}} - \frac{\partial}{\partial \theta_j}\frac{1}{2}\log\left|I_N + \Sigma W\right| && (200)\\
&= \frac{1}{2}\hat{\mathbf{f}}^T\mathbf{\Sigma}^{-1}\frac{\partial \mathbf{\Sigma}}{\theta_j}\mathbf{\Sigma}^{-1}\hat{\mathbf{f}} - \frac{1}{2}tr\left(\left(I_N + \Sigma W\right)^{-1}\frac{\partial\left(I_N + \Sigma W\right)}{\partial \theta_j}\right) && (201)\\
&= \frac{1}{2}\hat{\mathbf{f}}^T\mathbf{\Sigma}^{-1}\frac{\partial \mathbf{\Sigma}}{\theta_j}\mathbf{\Sigma}^{-1}\hat{\mathbf{f}} - \frac{1}{2}tr\left(\left(I_N + \Sigma W\right)^{-1}\left[\frac{\partial I_N}{\partial \theta_j} + \frac{\partial \Sigma W}{\partial \theta_j}\right]\right) && (202)\\
&= \frac{1}{2}\hat{\mathbf{f}}^T\mathbf{\Sigma}^{-1}\frac{\partial \mathbf{\Sigma}}{\theta_j}\mathbf{\Sigma}^{-1}\hat{\mathbf{f}} - \frac{1}{2}tr\left(\left(I_N + \Sigma W\right)^{-1}\frac{\partial \Sigma W}{\partial \theta_j}\right) && (203)\\
&= \frac{1}{2}\hat{\mathbf{f}}^T\mathbf{\Sigma}^{-1}\frac{\partial \mathbf{\Sigma}}{\theta_j}\mathbf{\Sigma}^{-1}\hat{\mathbf{f}} - \frac{1}{2}tr\left(\left(I_N + \Sigma W\right)^{-1}\frac{\partial \Sigma}{\partial \theta_j}W\right) && (204)
\end{aligned}$$

B.2.2.2 Term B Second factor:

$$\frac{\partial \hat{\mathbf{f}}}{\partial \theta_j} = \left(I_N + \Sigma W\right)^{-1}\frac{\partial \Sigma}{\partial \theta_j}\nabla \log p\left(\mathcal{Y}|\hat{\mathbf{f}}\right)\ \ (similar\ to\ gpml\ 5.24) \tag{205}$$

First factor:

$$\begin{aligned}
\frac{\partial \log p\left(\mathcal{Y}|\theta\right)}{\partial f_i} &= -\frac{1}{2}\frac{\partial \log |I_N + \Sigma W|}{\partial f_i} (\text{since only implicit terms}) && (206)\\
&= -\frac{1}{2} tr\left\{ (I_N + \Sigma W)^{-1} \frac{\partial \left(I + KW\right)}{\partial \widehat{f}_i} \right\} && (207)\\
&= -\frac{1}{2} tr\left\{ (I_N + \Sigma W)^{-1} \frac{\partial KW}{\partial \widehat{f}_i} \right\} && (208)\\
&= -\frac{1}{2} tr\left\{ (I_N + \Sigma W)^{-1} K \frac{\partial W}{\partial \widehat{f}_i} \right\} \text{ (note not as simpel as gpml 5.23)} && (209)
\end{aligned}$$

B.2.3 Evidence Derivatives - Likelihood Function Parameters

Only the explicit terms exists here so:

$$\begin{aligned}
\frac{\partial \log p\left(\mathcal{Y}|\phi,\theta\right)}{\partial \omega_j} &= \left.\frac{\partial \log p\left(\mathcal{Y}|\phi,\theta\right)}{\partial \omega_j}\right|_{\exp licit} && (210)\\
&= \frac{\partial}{\partial \omega_j}\left[\log p\left(\mathcal{Y}|\hat{\mathbf{f}}\right) - \frac{1}{2}\hat{\mathbf{f}}^T \sum^{-1} \hat{\mathbf{f}} - \frac{1}{2}\log |I_N + \Sigma W| \right] && (211)\\
&= \frac{\partial}{\partial \omega_j} \log p\left(\mathcal{Y}|\hat{\mathbf{f}}\right) - \underbrace{\frac{\partial}{\partial \omega_j}\frac{1}{2}\hat{\mathbf{f}}^T \Sigma^{-1} \hat{\mathbf{f}}}_{=0 \text{ since ker nel does not depend on param}} - \frac{\partial}{\partial \omega_j}\frac{1}{2}\log |I_N + \Sigma W| && (212)\\
&= \underbrace{\frac{\partial}{\partial \omega_j} \log p\left(\mathcal{Y}|\hat{\mathbf{f}}\right)}_{Term\,1} \underbrace{- \frac{\partial}{\partial \omega_j}\frac{1}{2}\log |I_N + \Sigma W|}_{Term\,2} && (213)
\end{aligned}$$

Term 1:

$$\frac{\partial}{\partial \omega_j} \log p\left(\mathcal{Y}|\hat{\mathbf{f}}\right) \tag{214}$$

Term 2:

$$\begin{aligned}
-\frac{\partial}{\partial \omega_j}\frac{1}{2}\log |I_N + \Sigma W| &= -\frac{1}{2} tr\left\{ (I_N + \Sigma W)^{-1} \frac{\partial \left(I_N + \Sigma W\right)}{\partial \omega_j} \right\} && (215)\\
&= -\frac{1}{2} tr\left\{ (I_N + \Sigma W)^{-1} \frac{\partial \left(I_N + \Sigma W\right)}{\partial \omega_j} \right\} && (216)\\
&= -\frac{1}{2} tr\left\{ (I_N + \Sigma W)^{-1} \left[\frac{\partial \left(I_N\right)}{\partial \omega_j} + \frac{\partial \Sigma W}{\partial \omega_j} \right] \right\} && (217)\\
&= -\frac{1}{2} tr\left\{ (I_N + \Sigma W)^{-1} \frac{\partial \Sigma W}{\partial \omega_j} \right\} && (218)\\
&= -\frac{1}{2} tr\left\{ (I_N + \Sigma W)^{-1} \left[\frac{\partial \Sigma}{\partial \omega_j} W + \Sigma \frac{\partial W}{\partial \omega_j} \right] \right\} && (219)\\
&= -\frac{1}{2} tr\left\{ (I_N + \Sigma W)^{-1} \Sigma \frac{\partial W}{\partial \omega_j} \right\} && (220)
\end{aligned}$$

with

$$\frac{\partial W}{\partial \omega_j} = \frac{-\nabla\nabla \log p\left(\mathcal{Y}|\mathbf{f}, \omega\right)}{\partial \omega_j}$$

B.3 Hyperparameter optimization Method

The optimization of the evidence/MAP estimate is performed using a BFGS method, implemented by Hans B. Nielsen and can be found in the matlab toolbox `immoptibox` [15]. This implementation has been compared with a state-of the-art method for GPs by Carl E. Rasmussens `minimize` [18] for a number of ad-hoc test problems and found slightly superior or similar (yet a more general comparison will be required to make a final conclusion).

C Mathematical Details and Implementation Notes

C.1 Matrix Identities

The Kailath Variant [17, p. 17] of the matrix inversion lemma:

$$(A + BC)^{-1} = A^{-1} - A^{-1}B(I + CA^{-1}B)^{-1}CA^{-1}$$

So

$$\left(K^{-1} + W\right)^{-1} = K - K(I + WK)^{-1}WK$$

(maybe the same trick as[18]). The eigenvalues of $(I + WK)$ are bounden below by 1 and thus provides a more stable version. Conjecture: The upper bound is probably bounded by $1 < |\lambda_i| < n^2 \max {a_{ij}}^2$.