

---

# Approximate Hyperparameter Marginalisation for Gaussian Processes

---

Anonymous Author(s)

Affiliation

Address

email

## Abstract

Jan wrote; In Bayesian methods a common problem is to choose a prior. In inference with Gaussian processes this task could involve choosing kernel function hyperparameters. In practice however, it is often unclear how to make this choice a priori. Therefore, most implementations deviate from proper Bayesian treatment by estimating the hyperparameters from the data via maximum likelihood methods. In contrast, we propose to add another hierarchy of inference on top of that. In particular, we propose to place a prior distribution over the hyperparameters. Its hyperparameters can in turn be estimated learned from the data. Since the resulting integrals of the marginalizations are non-analytic we use a Taylor expansion to yield a Gaussian process which approximates the correct process with marginalized hyperparameters. PERHAPS WRITE ABOUT INTEGRAL KERNEL AND RELATION TO RATIONAL QUADRATIC? We conduct experiments illustrating the benefits our approach on artificial as well as on real data.

## 1 Introduction

## 2 Gaussian Processes

Gaussian processes (GPs) constitute a powerful method for performing Bayesian inference about functions using a limited set of observations [3]. A GP is defined as a distribution over the functions  $f : \mathcal{X} \rightarrow \mathbb{R}$  such that the distribution over the possible function values on any finite set of  $\mathcal{X}$  is multi-variate Gaussian. A vector of observations  $\mathbf{y} = \{y_1, \dots, y_n\}$  could be viewed as a single point sampled from a  $n$ -variate Gaussian distribution.

A GP is completely defined by its first and second moments: a mean function  $\mu : \mathcal{X} \rightarrow \mathbb{R}$ , which describes the overall trend of the function, and a positive semidefinite covariance function, or kernel,  $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  which describes how function values are correlated as a function of their locations in the domain. Given a function  $f : \mathcal{X} \rightarrow \mathbb{R}$  about which we wish to perform inference and a set of input points  $\mathbf{x} \subseteq \mathcal{X}$ , the Gaussian process prior distribution over the function values  $\mathbf{f} = f(\mathbf{x})$  is given by:

$$\begin{aligned} p(\mathbf{f}|\mathbf{x}, \boldsymbol{\theta}, I) &:= \mathcal{N}(\mathbf{f}; \mu_{\boldsymbol{\theta}}(\mathbf{x}), K_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{x})) \\ &:= \frac{1}{\sqrt{\det 2\pi K_{\mathbf{f}}}} \exp\left(-\frac{1}{2}(\mathbf{f} - \mu_{\mathbf{f}})^{\top} K_{\mathbf{f}}^{-1}(\mathbf{f} - \mu_{\mathbf{f}})\right) \end{aligned} \quad (1)$$

where  $\boldsymbol{\theta}$  is a vector containing any parameters required by  $\mu$  and  $K$ : the *hyperparameters* of the model,  $I$ . Due to the ubiquity of  $I$  we henceforth drop it from explicit representation for notational convenience. There exist a wide variety of mean and covariance functions which can be chosen in order to reflect any prior knowledge available about the function of interest.



### 3.1 The Squared Exponential Kernel

As is the case in standard GP inference, a prior covariance function must be selected which is representative of any prior belief about the function of interest. In order to develop our method we first look at one of the most pervasive covariance functions used in Bayesian inference: the *Squared Exponential*, or *Gaussian*, kernel:

$$K_\theta(x, x') = h^2 \exp\left(-\frac{1}{2} \frac{\Delta^2}{L^2}\right) \quad (10)$$

where  $\Delta^2 = (x - x')^2$ , and  $h$  is a scaling parameter termed the output scale. The parameter  $L$ , the *length scale*, affects how closely points co-vary as a function of  $\Delta^2$ . The length parameter is perhaps the most crucial hyperparameter to train correctly, and is where we begin our exposition.

#### 3.1.1 Incorporating Uncertainty

The standard GP prior assumes that the value of  $L$  is deterministic; any distribution an individual has over  $L$  is purely the epistemic characterisation of their uncertainty in the ‘true’ value of  $L$ . We alter the prior such that we make  $L$  a random variable with its own generative distribution  $p(L)$ . The value of  $L$  should be strictly positive, hence we make the substitution  $L = \exp(\beta)$  and define a Gaussian distribution over  $\beta$  so that  $p(L)$  will have strictly positive support:

$$K_\theta(x, x') = h^2 \exp\left(-\frac{1}{2} \Delta^2 \exp(-2\beta)\right) \quad (11)$$

which we will write as  $K_\beta$ , and

$$\begin{aligned} p(\beta|\nu, \Lambda) &= \mathcal{N}(\beta; \nu, \Lambda) \\ &= \frac{1}{\sqrt{2\pi\Lambda}} \exp\left(-\frac{(\beta - \nu)^2}{2\Lambda}\right) \end{aligned} \quad (12)$$

where  $\nu$  and  $\Lambda$  form hyper-hyperparameters.

As stated before, we marginalise our the uncertainty in  $\beta$  (or  $L$ ), before fusing the prior with observations. The marginalised kernel is calculated by evaluating the following integral:

$$K_{\nu, \Lambda} = \int_{-\infty}^{+\infty} K_\beta p(\beta|\nu, \Lambda) d\beta \quad (13)$$

which is non-analytic given the current form of  $K_\beta$ . We approximate the solution to the integral by first approximating  $K_\beta$  using a second-order Taylor expansion around the mean of  $p(\beta)$  given by  $\nu$ . In order to make the integral analytic, we would like  $K_\beta$  to be in exponential form, hence we note the following identity and substitution:

$$\begin{aligned} K_\beta &= \exp\left(-\ln \frac{1}{K_\beta}\right) \\ &= \exp(-A) \end{aligned} \quad (14)$$

and we Taylor expand around  $A$ :

$$A \approx \ln \frac{1}{K_\beta} \Big|_\nu + (\beta - \nu) \frac{\partial A}{\partial \beta} \Big|_\nu + \frac{1}{2} (\beta - \nu)^2 \frac{\partial^2 A}{\partial^2 \beta} \Big|_\nu \quad (15)$$

Given (11), the derivatives of  $K_\beta$  can be calculated:

$$\begin{aligned} K'_\beta &= K_\beta (\Delta^2 \exp(-2\beta)) \\ K''_\beta &= K_\beta (\Delta^4 \exp(-4\beta) - \Delta^2 \exp(-2\beta)) \end{aligned}$$

which can be used for calculating the partial derivatives of (15):

$$A = \ln \frac{1}{K_\beta} \quad (16)$$

$$\frac{\partial A}{\partial \beta} = -\frac{K'_\beta}{K_\beta} = -\Delta^2 \exp(-2\beta) \quad (17)$$

$$\frac{\partial^2 A}{\partial^2 \beta} = -\frac{K''_\beta}{K_\beta} + \frac{K'^2_\beta}{K_\beta^2} = 2\Delta^2 \exp(-2\beta) \quad (18)$$

Using (14) and substituting (16-18) into (15) we arrive at the following expression for  $K_\beta$ :

$$\begin{aligned} K_\beta &\approx \exp \left( - \left( \ln \frac{1}{K_\nu} - (\beta - \nu)C + (\beta - \nu)^2 C \right) \right) \\ &= K_\nu \exp \left( ((\beta - \nu)C - (\beta - \nu)^2 C) \right) \end{aligned} \quad (19)$$

where  $C = \Delta^2 \exp(-2\nu)$

All that remains is to evaluate the integral in (13), which by using (12) and (19) is as follows:

$$\begin{aligned} K_{\nu,\Lambda} &\approx K_\nu \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi\Lambda}} \exp \left( \frac{-(\beta - \nu)^2}{2\Lambda} \right) ((\beta - \nu)C - (\beta - \nu)^2 C) \\ &= K_\nu \exp \left( \frac{\Lambda C^2}{2(1 + 2\Lambda C)} \right) \frac{1}{\sqrt{1 + 2\Lambda C}} \end{aligned} \quad (20)$$

This result is a new kernel which is the product of the original kernel  $K_\beta$  evaluated at  $\beta = \nu$ , and an expression which is a function of the squared distance  $\Delta^2$ .

### 3.1.2 Proof of Positive Semi-Definiteness

In order to confirm that our new kernel  $K_{\nu,\Lambda}$  is a legitimate covariance function, we must prove that it fulfils the necessary condition of kernels which is positive semi-definiteness. We first note that a sum of kernels is also a kernel, therefore:

$$K_{\nu,\Lambda} = \int K_\beta p(\beta) d\beta \quad (21)$$

should be a legitimate kernel if  $K_\beta$  is also a legitimate kernel as  $p(\beta)$  just weights the contents of the integral. Using the expression for  $K_\beta$  in (19) and completing the square we arrive at the following expression:

$$K_\beta = K_\nu \exp \left( -\frac{1}{2} \left( \beta - \nu - \frac{1}{2} \right)^2 C \right) \exp \left( \frac{1}{4} C \right) \quad (22)$$

If we substitute in  $K_\nu$ , where:

$$\begin{aligned} K_\nu &= h^2 \exp \left( -\frac{1}{2} \Delta^2 \exp(-2\nu) \right) \\ &= h^2 \exp \left( -\frac{C}{2} \right) \end{aligned} \quad (23)$$

we can write:

$$K_\beta = h^2 \exp \left( -\frac{C}{4} \right) \exp \left( -\frac{1}{2} \left( \beta - \nu - \frac{1}{2} \right)^2 C \right) \quad (24)$$

We note that the product of kernels is also a kernel. Both exponentials in (24) form legitimate kernels, which leaves us with the result that  $K_\beta$  is also a kernel and by definition is positive semi-definite as required.

## 4 Experiments

## 5 Related Work

## 6 Conclusion

## Acknowledgments

Do we have any? Aladdin / Orchid?

216  
217  
218  
219  
220  
221  
222  
223  
224  
225  
226  
227  
228  
229  
230  
231  
232  
233  
234  
235  
236  
237  
238  
239  
240  
241  
242  
243  
244  
245  
246  
247  
248  
249  
250  
251  
252  
253  
254  
255  
256  
257  
258  
259  
260  
261  
262  
263  
264  
265  
266  
267  
268  
269

## References

- [1] M.H. Chen, Q.M. Shao, and J.G. Ibrahim. *Monte Carlo methods in Bayesian computation*. Springer Verlag, 2000.
- [2] M.A. Osborne, R. Garnett, S.J. Roberts, C. Hart, S. Aigrain, N.P. Gibson, and S. Aigrain. Bayesian quadrature for ratios. *Fifteenth International Conference on Artificial Intelligence and Statistics*, 2012.
- [3] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press Cambridge, MA, USA, 2006.
- [4] C.E. Rasmussen and Z. Ghahramani. Bayesian monte carlo. *Advances in neural information processing systems*, 15:489–496, 2003.