Statistical Machine Learning Gaussian Processes

Luca Bortolussi

Data Science and Scientific Computing

1 Random functions and Bayesian regression

- Bayesian linear regression places a (Gaussian) prior over the weights vector, and computes the (Gaussian) posterior distribution over weights.
- What does this mean? Consider linear basis functions. In this case, the regression line is a *random line*, with the property that the output prediction at any point is a Gaussian random variable
- This concept can be generalised: taking linear combinations of basis functions with (Gaussian) random coefficients leads to a (Gaussian) random function

1.1 Random functions terminology

- A random function is an infinite collection of random variables indexed by the argument of the function
- A popular alternative name is a *stochastic process*
- When considering the random function evaluated at a (finite) set of points, we get a random vector
- The distribution of this random vector is called *finite dimensional marginal*

Exercise

Let $\phi_0(x), \dots, \phi_{M-1}(x)$ be a fixed set of functions, and let $f(x) = \sum w_i \phi_i(x)$. If $\mathbf{w} \sim \mathcal{N}(0, D)$ compute: $N(0, I)$, compute:

- 1. The single-point marginal distribution of $f(x)$
- 2. The two-point marginal distribution of $f(x_1)$, $f(x_2)$

1.2 The Gram matrix

- Generalising the exercise to more than two points, we get that *any* finite dimensional marginal of this process is multivariate Gaussian
- The covariance matrix of this function is given by evaluating a function of two variables at all possible pairs
- The function is defined by the set of basis functions

$$
k(x_i, x_j) = \boldsymbol{\phi}(x_i)^T \boldsymbol{\phi}(x_j)
$$

- The covariance matrix is often called *Gram matrix* and is (necessarily) symmetric and positive definite
- Bayesian prediction in regression then is essentially the same as computing conditionals for Gaussians (more later)

1.3 Main limitation of Bayesian regression

- Choice of basis functions inevitably impacts what can be predicted
- Suppose one wishes the basis functions to tend to zero as $x \to \infty$
- Then, necessarily, very large input values will have predicted outputs near zero with high confidence!
- Ideally, one would want a prior over functions which would have the same uncertainty everywhere

1.4 Function Space view

- In order to construct such priors, one possibility would be to construct a countable sequence of basis functions. We can partition the full \mathbb{R}^n in compact sets, and define a finite number of basis functions supported in each compact set so that the variance in each point of the state space is a constant (partition of unity).
- This approach, called the *weights space view*, is unpractical, but it demonstrates the existence of truly infinite dimensional Gaussian Processes.
- In general, it is more useful to take the dual point of view, and work with kernels rather than with basis functions.

2 Gaussian Processes

2.1 GP definition

• A Gaussian Process (GP) is a stochastic process indexed by a continuous variable *x* s.t. all finite dimensional marginals are multivariate Gaussian

• A GP is uniquely defined by its *mean* and *covariance* functions, denoted by $\mu(x)$ and $k(x, x')$:

$$
f \sim \mathcal{GP}(\mu, k) \leftrightarrow \mathbf{f} = (f(x_1), \dots, f(x_N)) \sim \mathcal{N}(\mu, K),
$$

$$
\mu = (\mu(x_1), \dots, \mu(x_N)), \quad K = (k(x_i, x_j))_{i,j}
$$

• The covariance function must satisfy some conditions (Mercer's theorem), essentially it needs to evaluate to a symmetric positive definite function for all sets of input points put points \mathbb{R}^n . I. Williams, Gaussian Processes for MIT Processes for MIT Press, 2006, 20

Example

Consider a 1-dimensional GP with mean function $\mu(x) \equiv 0$, and with Gaussian covariance function:

$$
k(x, x') = \exp\left[-\frac{1}{2}|x - x'|^2\right]
$$

 $(x) = 1$ If we consider a test set $V^* = x$, we then the $\mathbf{F} = (f(x_1), \dots, f(x_n))$ is The variance at each point *x* is $k(x, x) = 1$. If we consider a test set $X^* = x_1, \ldots, x_n$, then the distribution of $f^* = (f(x))$ is joint distribution of $f^* = (f(x_1), \ldots, f(x_n))$ is

$$
\mathbf{f}^* \sim \mathcal{N}(\mathbf{0}, K(X^*, X^*))
$$

 (X^*) is the Gram matrix. $K_{ij} = k(x_i, x_j)$, which is symmetric and positive definite. where $K(X^*, X^*)$ is the Gram matrix, $K_{ij} = k(x_i, x_j)$, which is symmetric and positive definite.

2.2 Noise-free prediction

- point wise \mathbf{r} is much and minus two times two times the standard deviation for each input value \mathbf{r} • Suppose now to observe the exact value of the GP at *N* different points, $X = x_1, \ldots, x_N$, with observations $\mathbf{f} = f(x_1)$ $f(x_N)$ with observations $\mathbf{f} = f(x_1), \ldots, f(x_N)$.
- Consider also the test points $X^* = x_1, \ldots, x_n$, with function values $\mathbf{f}^* = (f(x_1), \ldots, f(x_n))$
(unobserved to be estimated) (unobserved, to be estimated).

• The *joint prior distribution* of f on inputs X and test points X^* is

$$
\begin{bmatrix} \mathbf{f} \\ \mathbf{f}_{*} \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(X,X) & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{bmatrix}\right). \tag{2.18}
$$

If the observe the values at X, then we need to *condition* on these values. Hence the • If we observe the values at X, then we need to *condition* on these values. Hence the conditional f^* if is p_1 is and similarly for the other entries \mathbb{R}^n . conditional f ∗ |f is θ is distribution that distribution $\mathbf{f}^* \mathbf{f}$ is $\frac{1}{2}$ for $\frac{1}{2}$ for $\frac{1}{2}$ for $\frac{1}{2}$

$$
\mathbf{f}_{*}|X_{*}, X, \mathbf{f} \sim \mathcal{N}(K(X_{*}, X)K(X, X)^{-1}\mathbf{f},
$$

$$
K(X_{*}, X_{*}) - K(X_{*}, X)K(X, X)^{-1}K(X, X_{*})).
$$
 (2.19)

Fulction values for the standard formula families can be sampled of a G can since which is obtained by the standard formula for the conditional of a Gaussian.

Example C. E. Rasmussen & C. K. I. Williams, Gaussian Processes for Machine Learning, the MIT Press, 2006, te

 \mathbf{r}

Figure 2.2 (b) shows the results of the results of the results of the results of the final set of the Consider again the 1-dimensional GP with mean function $\mu(x) \equiv 0$, and with Gaussian covari- $\begin{bmatrix} 1 & 0 \end{bmatrix}$ ance function: 1 differential the interestional of which mean function $\mu(x) = 0$, and with Gaussian covariance ance function.

$$
k(x, x') = \exp\left[-\frac{1}{2}|x - x'|^2\right]
$$

2.3 Noisy predictions of y actual generators of y actual generators \mathbf{z} 2.3 Noisy predictions Assuming additions in distributions in distributions in \mathbf{C}

• Suppose we cannot observe the values f of a GP at points X , but a perturbed version on the key predictions for Gaussian process regressions for Gaussian process regressions $\mathcal{L}(\mathcal{A})$ • Suppose we cannot observe the values f of a GP at points X , but a perturbed version of p mean: q them: \mathcal{L} or cov(\mathcal{L}) \mathcal{L} \mathcal{L}

$$
y(\mathbf{x}) = f(\mathbf{x}) + \varepsilon,
$$

 $\frac{1}{2}$ where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ follows $\mathcal{E} \sim \mathcal{N}(0, \sigma^2)$

- $f(x, y)$
 \therefore $f(x, y)$ = $f(x, y)$ + 2 $f(x, y)$ + 2 $f(x, y)$ • Then the covariance of observations is $cov(y) = K(X, X) + \sigma^2 I$
- Then the covariance of observations is $\text{cov}(y) = K(x, x) + 0$. ⁸There are some situations where it is reasonable to assume that the observations are t the function values at the test location values \mathcal{L}_t and \mathcal{L}_t as the prior as the prior as the prior as

$$
\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(X,X) + \sigma_n^2 I & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{bmatrix}\right). \tag{2.21}
$$

• Conditioning on observations y, we get

$$
\mathbf{f}_{*}|X, \mathbf{y}, X_{*} \sim \mathcal{N}(\bar{\mathbf{f}}_{*}, \text{ cov}(\mathbf{f}_{*})), \text{ where } (2.22)
$$

$$
\overline{\mathbf{f}}_* \triangleq \mathbb{E}[\mathbf{f}_*|X, \mathbf{y}, X_*] = K(X_*, X)[K(X, X) + \sigma_n^2 I]^{-1} \mathbf{y}, \qquad (2.23)
$$

$$
cov(\mathbf{f}_*) = K(X_*, X_*) - K(X_*, X)[K(X, X) + \sigma_n^2 I]^{-1} K(X, X_*). \quad (2.24)
$$

2.4 Linear predictor

• For a single point x^* , the predictive distribution reads for a single point \mathbf{x}^* , the predictive distribution reads

$$
\bar{f}_* = \mathbf{k}_*^\top (K + \sigma_n^2 I)^{-1} \mathbf{y},\tag{2.25}
$$

$$
\mathbb{V}[f_*] = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^\top (K + \sigma_n^2 I)^{-1} \mathbf{k}_*.
$$
 (2.26)

where $\mathbf{k}_{*} = (k(\mathbf{x}^*, \mathbf{x}_1), \dots, k(\mathbf{x}^*, \mathbf{x}_N))$

 N is a linear prediction eq. (2.25) is a linear combination eq. (2.25) is a linear combination of observed α • It can be seen that the average prediction is a linear combination of the kernels evaluated \sum_{N} on the input points:

$$
\bar{f}(\mathbf{x}^*) = \sum_{i=1}^N \alpha_i k(\mathbf{x}^*, \mathbf{x}_i)
$$

where $\alpha = (K + \sigma^2 I)^{-1} \mathbf{y}$.

where $\alpha = (K + \sigma^2 I)^{-1}$ **y**.

2.5 Posterior GP

• It is easy to see that the posterior process $f|y$ is again a Gaussian process, with mean

$$
\mathbb{E}[f(\mathbf{x})|\mathbf{y}] = K(\mathbf{x}, X)(K + \sigma^2 I)^{-1}\mathbf{y}
$$

and covariance although the GP defines a joint Gaussian dis-

$$
k(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}') - K(\mathbf{x}, X)(K + \sigma^2 I)^{-1} K(X, \mathbf{x}')
$$

3 Kernel functions

3.1 Kernels 4.3 Eigenfunction $\frac{1}{2}$ Eigenfunction $\frac{1}{2}$

• The notion of kernel comes from the theory of integral operators on a space X with measure μ . A real kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defines an integral operator T_k (applied to integrable f) as: *f*) as: α . α

$$
(T_k f)(\mathbf{x}) = \int_X k(\mathbf{x}, \mathbf{y}) f(\mathbf{y}) d\mu(\mathbf{y})
$$

• A kernel is positive semidefinite if, for all $f \in L_2(\mathcal{X}, \mu)$:

$$
\int_{X\times X} k(\mathbf{x}, \mathbf{y}) f(\mathbf{x}) f(\mathbf{y}) d\mu(\mathbf{x}) d\mu(\mathbf{y}) \ge 0
$$

- Equivalently, a kernel is positive (semi)definite if for any collection of *n* points $\{x_i \mid i = \}$ 1, ..., *n*}, the Gram matrix *K*, $K_{ij} = l(\mathbf{x_i}, \mathbf{x_j})$ is positive (semi)definite (Mercer's theorem).
- The Gram matrix of a symmetric kernel, $k(x, y) = k(y, x)$, is symmetric.

3.2 Eigenfunctions

• An eigenfunction ϕ with eigenvalue λ of k satisfies \mathbf{c} experiment ϕ with eigenvalue λ of k satisfies

$$
\int k(\mathbf{x}, \mathbf{y}) \phi(\mathbf{x}) d\mu(\mathbf{x}) = \lambda \phi(\mathbf{y})
$$

- There can be an infinite number of eigenfunctions, which can be ordered w.r.t. decreasing eigenvalues, and they can be chosen orthogonal, i.e. such that $\int \phi_i(\mathbf{x})\phi_j(\mathbf{x})d\mu(\mathbf{x}) = \delta_{ij}$
	- A kernel can be decomposed using eigenfunctions:

Theorem 4.2 (Mercer's theorem). Let (X, μ) be a finite measure space and $k \in L_{\infty}(\mathcal{X}^2, \mu^2)$ be a kernel such that $T_k : L_2(\mathcal{X}, \mu) \to L_2(\mathcal{X}, \mu)$ is positive definite (see eq. (4.2)). Let $\phi_i \in L_2(\mathcal{X}, \mu)$ be the normalized eigenfunctions of T_k associated with the eigenvalues $\lambda_i > 0$. Then:

1. the eigenvalues $\{\lambda_i\}_{i=1}^{\infty}$ are absolutely summable

2.

$$
k(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^{\infty} \lambda_i \phi_i(\mathbf{x}) \phi_i^*(\mathbf{x}'), \qquad (4.37)
$$

holds μ^2 almost everywhere, where the series converges absolutely and uniformly μ^2 almost everywhere.

3.3 Reproducing Kernel Hilbert Spaces

Definition 6.1 (Reproducing kernel Hilbert space). Let H be a Hilbert space of real functions f defined on an index set X . Then H is called a reproducing kernel Hilbert space endowed with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ (and norm $||f||_{\mathcal{H}} =$ $\sqrt{\langle f, f \rangle_{\mathcal{H}}}$) if there exists a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with the following properties:

1. for every $\mathbf{x}, k(\mathbf{x}, \mathbf{x}')$ as a function of \mathbf{x}' belongs to \mathcal{H} , and

2. k has the reproducing property $\langle f(\cdot), k(\cdot, \mathbf{x}) \rangle_{\mathcal{H}} = f(\mathbf{x}).$

See e.g. Schölkopf and Smola [2002] and Wegman [1982]. Note also that as $k(\mathbf{x}, \cdot)$ and $k(\mathbf{x}', \cdot)$ are in H we have that $\langle k(\mathbf{x}, \cdot), k(\mathbf{x}', \cdot) \rangle_{\mathcal{H}} = k(\mathbf{x}, \mathbf{x}').$

The RKHS uniquely determines k , and vice versa, as stated in the following theorem:

Theorem 6.1 (Moore-Aronszajn theorem, Aronszajn [1950]). Let X be an index set. Then for every positive definite function $k(\cdot, \cdot)$ on $\mathcal{X} \times \mathcal{X}$ there exists a unique RKHS and vice versa a unique RKHS, and vice versa.

$\frac{1}{2}$ (which has the dot product here $\frac{1}{2}$ $\frac{1}{2}$ 3.3.1 RKHS and Eigenfunctions

- The functions belonging to the RKHS associated with a kernel *k* can be written as a linear are the analogues of the analogues of delta functions ϕ of $\ln f(x) = \sum f d(x)$, with $\sum f^2/1 \le \infty$ (the combination of the eigenfunctions ϕ_j of k: $f(\mathbf{x}) = \sum_j f_j \phi_j(\mathbf{x})$, with $\sum_j f_j^2 / \lambda_j < \infty$ (this is representer of the representation and its interest in the RKHS. a smoothness constraint).
- Such functions define an Hilbert space *H* with inner product $\langle f, g \rangle_H = \sum_j \frac{f_j g_j}{\lambda_j}$ λ _j and the show that the space *H* with inner product $\langle f, g \rangle_H = \sum_j \frac{\partial g_j}{\partial f_j}$
- This Hilbert space is the RKHS corresponding to kernel k :

$$
\langle f(\cdot), k(\cdot, \mathbf{x}) \rangle_{\mathcal{H}} = \sum_{i=1}^{N} \frac{f_i \lambda_i \phi_i(\mathbf{x})}{\lambda_i} = f(\mathbf{x}). \tag{6.2}
$$

 v_{minary} Similarly

$$
\langle k(\mathbf{x}, \cdot), k(\mathbf{x}', \cdot) \rangle_{\mathcal{H}} = \sum_{i=1}^{N} \frac{\lambda_i \phi_i(\mathbf{x}) \lambda_i \phi_i(\mathbf{x}')}{\lambda_i} = k(\mathbf{x}, \mathbf{x}'). \tag{6.3}
$$

• Furthermore, the norm of $k(x, \cdot)$ is $k(x, x) < \infty$: it belongs to H. • Furthermore, the norm of $k(x, \cdot)$ is $k(x, x) < \infty$: it belongs to *H*.

3.4 Classification of Kernel functions

 \sim 1 is \sim 1 is A kernel $k(x, y)$ can be classified w.r.t dependence on x and y.

- Stationary kernel: it is a function of $\mathbf{x} \mathbf{y}$ (invariant to translations).
- $\frac{1}{\text{F}}$ is a function of $\mathbf{A} = \mathbf{y}$ (invariant to differentially). • Isotropic kernel: it is a function of $\|\mathbf{x} - \mathbf{y}\|$ (invariant to rigid motions). • Isotropic kernel: it is a function of $\|\mathbf{x} - \mathbf{y}\|$ (invariant to rigid motions).
- Dot-product kernel: it is a function of $x^T y$ (invariant w.r.t. rotations with respect to the \Box). origin).

Continuity properties of the GPs and kernels k .

ty in mean square of a process f at \mathbf{x} : for each $\mathbf{x}_k \to \mathbf{x}$ it holds that $\mathbb{E}[\|f(\mathbf{x})\|]$ • Continuity in mean square of a process *f* at **x**: for each $x_k \to x$, it holds that $\mathbb{E}[\|f(x_k) - f(x_k)\|^2]$ $v \to 0$. $f(\mathbf{x})\|^2] \to 0.$

- A process is continuous in m.s. at **x** iff *k* is continuous at $k(x, x)$. For stationary kernels, *k* must be continuous at zero. must be continuous at zero. \sim $\frac{1}{2}$ c(x, x). For stationary learnals, $\frac{1}{2}$
- If *k* is 2*k*th differentiable, than *f* is *k*th differentiable (in m.s.). k is 2kth differentiable, than f is kth differentiable (in m.s.).

3.4.1 Gaussian kernel 2

• The Gaussian or Squared Exponential kernel is defined by qu red Exponential kernel is define \mathbf{b} v

$$
k(\mathbf{x}, \mathbf{y}) = \alpha \exp \left[-\frac{\|\mathbf{x} - \mathbf{y}\|^2}{\lambda^2}\right]
$$

- α is called the amplitude, it regulates the magnitude of variance at each point **x**. λ , instead, is the characteristic length-scale, which regulates the speed of decay of the correlation is the characteristic length-scale, which regulates the speed of decay of the correlation between points. regulates the magnitude of variance at each point is isotropic and among the most
- The Gaussian kernel is isotropic and among the most used in computational statistics, and its RKHS is dense in the space of continuous functions over a compact set in \mathbb{R}^n .
- The Automatic-Relevance Detection Gaussian Kernel generalises the GK as

$$
k(\mathbf{x}, \mathbf{y}) = \alpha \exp \left[-\sum_{j} \frac{|x_j - y_j|^2}{\lambda_j^2} \right]
$$

3.4.2 Matérn kernel

• The Matérn kernel is defined by

$$
k_{\text{Matern}}(r) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}r}{\ell}\right)^{\nu} K_{\nu} \left(\frac{\sqrt{2\nu}r}{\ell}\right), \tag{4.14}
$$

with positive parameters ν and ℓ , where K_{ν} is a modified Bessel function

- **■** If *ν* > *h*, then the process with Matérn kernel is *h* times differentiable (in m.s.) For *ν* → ∞, then the MK becomes the GK • Examples of Matern Kernel: then the MK becomes the GK.
-

3.4.3 Matérn and Exponential kernel

• Typical choice for MK is $v = p + \frac{1}{2}$, giving

$$
k_{\nu=p+1/2}(r) \; = \; \exp\left(-\frac{\sqrt{2\nu}r}{\ell}\right) \frac{\Gamma(p+1)}{\Gamma(2p+1)} \sum_{i=0}^{p} \frac{(p+i)!}{i!(p-i)!} \left(\frac{\sqrt{8\nu}r}{\ell}\right)^{p-i} . \tag{4.16}
$$

It is possible that the most interesting cases for machine learning are $\nu = 3/2$ and $\nu = 5/2$, for which

$$
k_{\nu=3/2}(r) = \left(1 + \frac{\sqrt{3}r}{\ell}\right) \exp\left(-\frac{\sqrt{3}r}{\ell}\right),
$$

\n
$$
k_{\nu=5/2}(r) = \left(1 + \frac{\sqrt{5}r}{\ell} + \frac{5r^2}{3\ell^2}\right) \exp\left(-\frac{\sqrt{5}r}{\ell}\right),
$$
\n(4.17)

• for $v = 1/2$, we get the Exponential Kernel

$$
k(\mathbf{x}, \mathbf{y}) = \exp\left[\|\mathbf{x} - \mathbf{y}\| / \lambda\right]
$$

which in one dimension corresponds to the Ornstein-Ulembeck process (the model of which in one dimension corresponds to the Oristem-Oremocek process (the moder or
velocity of a particle undergoing Brownian motion), which is continuous but nowhere differentiable. α -differentiable.

3.4.4 Polynomial kernel $\sum_{i=1}^{n}$

• Simple dot-products kernels are the polynomial kernel, for p integer: the polynomial Kerner, for p in \cdot

$$
k(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^T \mathbf{x}')^p
$$

• This corresponds to a kernel obtained by a set of polynomial basis functions: tained by a set of polynomial basis functions:

$$
k(\mathbf{x}, \mathbf{x}') = (\mathbf{x} \cdot \mathbf{x}')^p = \left(\sum_{d=1}^D x_d x'_d\right)^p = \left(\sum_{d_1=1}^D x_{d_1} x'_{d_1}\right) \cdots \left(\sum_{d_p=1}^D x_{d_p} x'_{d_p}\right)
$$

$$
= \sum_{d_1=1}^D \cdots \sum_{d_p=1}^D (x_{d_1} \cdots x_{d_p}) (x'_{d_1} \cdots x'_{d_p}) \triangleq \phi(\mathbf{x}) \cdot \phi(\mathbf{x}'). \tag{4.23}
$$

• The basis functions ϕ_m are given by all monomials of degree p, i.e. $\sum m_j = p$:

$$
\phi_{\mathbf{m}}(\mathbf{x}) = \sqrt{\frac{p!}{m_1! \cdots m_D!}} x_1^{m_1} \cdots x_D^{m_D}.
$$
\n(4.24)

$\mathbf{0}$, given by eq. (4.35). 3.5 Composition of Kernels

Kernels can be composed according to certain rules, giving rise to new kernels.

Techniques for Constructing New Kernels.

Given valid kernels $k_1(\mathbf{x}, \mathbf{x}')$ and $k_2(\mathbf{x}, \mathbf{x}')$, the following new kernels will also be valid:

where $c > 0$ is a constant, $f(\cdot)$ is any function, $q(\cdot)$ is a polynomial with nonnegative coefficients, $\phi(\mathbf{x})$ is a function from x to \mathbb{R}^M , $k_3(\cdot, \cdot)$ is a valid kernel in \mathbb{R}^M , **A** is a symmetric positive semidefinite matrix, \mathbf{x}_a and \mathbf{x}_b are variables (not necessarily disjoint) with $\mathbf{x} = (\mathbf{x}_a, \mathbf{x}_b)$, and k_a and k_b are valid kernel functions over their respective spaces.

complex kernels appropriate to specific applications. We require that the kernel 4 Hyperparameters $\frac{1}{\sqrt{2}}$

${\bf M}$ a few computed functions. For a more extensive discussed functions. For a more extensive discussed on ${\bf M}$ 4.1 Marginal likelihood

- In order to do model selection (e.g. between different kernels) we can use the marginal k elihood k likelihood.
- ! x^Tx′ + c "2 with c > 0, then the corresponding feature mapping φ(x) contains constant and line as well as the set the performance of the network. • This can be used also to set hyperparameters of the kernel functions, like the amplitude or the lengthscale of the Gaussian kernel or the lengthscale of the Gaussian kernel.
- For GP, we can compute the marginal likelihood analytically:

$$
\mathcal{L} = \log p(\mathbf{y}|X) = \log \int p(\mathbf{f}|X)p(\mathbf{y}|\mathbf{f}, X)d\mathbf{f}
$$

which gives

$$
\mathcal{L} = -\frac{1}{2} \mathbf{y}^T (K + \sigma^2 I)^{-1} \mathbf{y} - \frac{1}{2} \log |(K + \sigma^2 I)| - \frac{N}{2} \log 2\pi
$$

• This follows also by observing that $y \sim N(0, K + \sigma^2 I)$.

The log marginal likelihood

$$
\mathcal{L} = -\frac{1}{2} \mathbf{y}^T (K + \sigma^2 I)^{-1} \mathbf{y} - \frac{1}{2} \log |(K + \sigma^2 I)| - \frac{N}{2} \log 2\pi
$$

has three terms

- $-\frac{1}{2}y^T(K + \sigma^2 I)^{-1}y$ is the data fit.
- $-\frac{1}{2} \log |(K + \sigma^2 I)|$ is a complexity penalty.
- $-\frac{N}{2} \log 2\pi$ is a constant.

Data from 1dim example with Gaussian kernels

Figure 5.3: Panel (a) shows a decomposition of the log marginal likelihood into its constituents: data-fit and complexity penalty, as a function of the characteristic length-scale. The training data is drawn from a Gaussian process with SE covariance function and parameters $(\ell, \sigma_f, \sigma_n) = (1, 1, 0.1)$, the same as in Figure 2.5, and we are fitting only the length-scale parameter ℓ (the two other parameters have been set in accordance with the generating process). Panel (b) shows the log marginal likelihood accordance with the generating process). Tance (o) shows the log marginal intermodular as a function of the characteristic length-scale for different sizes of training sets. Also $\frac{1}{2}$ as a function of the enaracteristic reigni-scale for different sizes of training sets. This shown, are the 95% confidence intervals for the posterior length-scales.

Data from 1dim example with Gaussian kernels

Figure 5.4: Contour plot showing the log marginal likelihood as a function of the characteristic length-scale and the noise level, for the same data as in Figure 2.5 and Figure 5.3. The signal variance hyperparameter was set to $\sigma_f^2 = 1$. The optimum is close to the parameters used when generating the data. Note, the two ridges, one for small noise and length-scale $\ell = 0.4$ and another for long length-scale and noise $\sigma_n^2 = 1$. The contour lines spaced 2 units apart in log probability density. Figure 5.3. The signal variance hyperparameter was set to $\sigma_f^2 = 1$. The optimum is for small noise and length-scale $\ell = 0.4$ and another for long length-scale and noise $\sigma_n^2 = 1$. The contour lines spaced 2 units apart in log probability density.

Dete coming from a cample of a 1dim CD with Gaussian kernal and hyperparameters. Data coming from a sample of a 1dim GP with Gaussian kernel and hyperparameters $\lambda = 1$,
 $\tau = 0.1$ T , σ and σ $\alpha = 1, \sigma = 0.1.$

Figure 2.5: (a) Data is generated from a GP with hyperparameters $(\ell, \sigma_f, \sigma_n)$ = $(1, 1, 0.1)$, as shown by the $+$ symbols. Using Gaussian process prediction with these hyperparameters we obtain a 95% confidence region for the underlying function f (shown in grey). Panels (b) and (c) again show the 95% confidence region, but this time for hyperparameter values $(0.3, 1.08, 0.00005)$ and $(3.0, 1.16, 0.89)$ respectively. igure 2.5: (a) Data is generated from a GP with hyperparameters $(\ell, \sigma_f, \sigma_n)$ = b_1, b_2, b_3 as shown by the \pm symbols. Csing Gaussian process prediction with these vertex we obtain a 95% confidence region for the underlying function, f shown in grey). Panels (b) as

Hypernarameter ontimisation 4.2 Hyperparameter optimisation \mathbf{v} training data points the slope of the log marginal likelihood is very shallow is very shallow is very shallow is very shallow in \mathbf{r} .

• In order to set the hyperparameters, we can maximise the log marginal likelihood:

$$
\mathcal{L} = -\frac{1}{2} \mathbf{y}^T (K + \sigma^2 I)^{-1} \mathbf{y} - \frac{1}{2} \log |(K + \sigma^2 I)| - \frac{N}{2} \log 2\pi
$$

• Its derivative w.r.t. an hyperparameter θ is Its derivative w.r.t. an hyperparameter θ is

$$
\frac{\partial}{\partial \theta_j} \log p(\mathbf{y}|X, \theta) = \frac{1}{2} \mathbf{y}^\top K^{-1} \frac{\partial K}{\partial \theta_j} K^{-1} \mathbf{y} - \frac{1}{2} \operatorname{tr} (K^{-1} \frac{\partial K}{\partial \theta_j})
$$
\n
$$
= \frac{1}{2} \operatorname{tr} \left((\boldsymbol{\alpha} \boldsymbol{\alpha}^\top - K^{-1}) \frac{\partial K}{\partial \theta_j} \right) \text{ where } \boldsymbol{\alpha} = K^{-1} \mathbf{y}. \tag{5.9}
$$

The derivative is relatively cheap to compute, once we invert the matrix K . Hence we can • The derivative is relatively cheap to compute, once we invert the matrix K . Hence we can use gradient methods to optimise \mathcal{L} .

that they are parameters of a non-parameters of a non-parameters of accordance with the weight-space with the w

• Purely Bayesian methods (giving a prior on hyperparameters) are complicated by the in general complex functional form (no conjugate prior).

4.3 Non-constant prior mean \mathbf{E}

- The typical choice for the prior mean is the zero function. Data is processed by subtracting the sample mean from the observations. • The typical choice for the prior mean is the zero function. Data is processed by subtracting
- the sample mean from the observations.
• As an alternative, one can either use a deterministic function for the priori mean (and • As an anchiative, one can enter use a deterministic function for the priori mean (and subtract if from data, adding it back to predictions), or use a generalised linear model for the prior mean: \mathbf{a} subtract if from data, adding it back to predictions), or use a generalised linear model for α prot mean. t_{total} and t_{total} is a polynomial regression, i.e. t_{total} polynomial regression in α . When fitting the model, one could optimize over the parameters \mathbf{y}

$$
g(\mathbf{x}) = f(\mathbf{x}) + \mathbf{h}(\mathbf{x})^{\top} \boldsymbol{\beta}, \text{ where } f(\mathbf{x}) \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}')), \qquad (2.39)
$$

here f(x) is a zero mean GP, h(x) are a set of fixed basis functions, and are • If we put a Gaussian prior over coefficients β , we can treat them in a Bayesian way, and get a GP. get a GP. get a GP: get a GP: \hat{B} we put a Gaussian prior over coefficients β , we can treat them in a Bayesian way, and parameters. For \mathbf{F} : now with an added contribution in the covariance function caused by the un-

$$
g(\mathbf{x}) \sim \mathcal{GP}(\mathbf{h}(\mathbf{x})^{\top}\mathbf{b}, k(\mathbf{x}, \mathbf{x}') + \mathbf{h}(\mathbf{x})^{\top}B\mathbf{h}(\mathbf{x}')), \qquad (2.40)
$$

• In this way, we obtain the following predictive distribution at a point x^* : It this way, we obtain the following predictive distribution at a point x^* :

$$
\overline{\mathbf{g}}(X_*) = H_*^\top \overline{\boldsymbol{\beta}} + K_*^\top K_y^{-1} (\mathbf{y} - H^\top \overline{\boldsymbol{\beta}}) = \overline{\mathbf{f}}(X_*) + R^\top \overline{\boldsymbol{\beta}},
$$

\n
$$
cov(\mathbf{g}_*) = cov(\mathbf{f}_*) + R^\top (B^{-1} + HK_y^{-1}H^\top)^{-1} R,
$$
\n(2.41)

where the H matrix concets the $\mathbf{u}(\mathbf{x})$ vectors for an training (and H_* an ecst)
 \overline{Q} (D^{-1} + $H V^{-1} H^{\top}$) -1($H V^{-1}$), D^{-1}). After $H V^{-1} V$ cases, $\bar{\beta} = (B^{-1} + HK_y^{-1}H^{\top})^{-1}(HK_y^{-1}\mathbf{y} + B^{-1}\mathbf{b})$, and $R = H_* - HK_y^{-1}K_*$. where the H matrix collects the $h(x)$ vectors for all training (and H_* all test) $\mathcal{N}=\mathcal{N}$ $N = \frac{1}{2}$ interpretation of the mean expression, eq. (2.41) top lines: $\frac{1}{2}$ where the H matrix collects the $h(x)$ vectors for all training (and H_* all test)

- The new predictive distribution has mean $H_* p$ (from the finear moder) plus a term coming from the GP model of residuals. bution has mean $H^I_*\beta$ (from the linear model) plus a siduals. $\sigma =$ The new predictive distribution has mean $H_*^T \bar{\beta}$ (from the linear model) plus a term coming • The new predictive distribution has mean $H_*^T \overline{\beta}$ (from the linear model) plus a term coming
from the GP model of residuals The new predictive distribution has mean $H^*_* \beta$ (from the linear model) plus a term comi
- Taking a flat prior (limit for $B^-1 \rightarrow$ matrix of zeros): the sum of the usual covariance term and a new non-negative contribution. laking a flat prior (limit for B^{-} 1 \rightarrow matrix of zeros):

 p is independent of binary which is independent of binary which is independent of binary p

$$
\overline{\mathbf{g}}(X_*) = \overline{\mathbf{f}}(X_*) + R^\top \overline{\boldsymbol{\beta}},
$$

\n
$$
cov(\mathbf{g}_*) = cov(\mathbf{f}_*) + R^\top (HK_y^{-1}H^\top)^{-1}R,
$$
\n(2.42)

the data term and prior, and the predictive mean is simply the mean linear where the limiting $\beta = (HK_y^{-1}H^{\dagger})^{-1}HK_y^{-1}y$. Notice that predictions under $\ddot{B} = (HK^{-1}H^{\top})^{-1}HK^{-1}$ where the limiting $\bar{\beta} = (HK_y^{-1}H^{\top})^{-1}HK_y^{-1}\mathbf{y}$. Notice that predictions under $\frac{1}{2}$. Only by plugging the modified notation by plugging the modified name $\frac{1}{2}$

5 GP classification et classification ϵ the characteristic should not be implemented named not by plugging the modified named in the model of modified the model of model in the model of model of model in the model of model of model in the model of model of model

 $\frac{1}{2}$ and $\frac{1}{2}$ of $\frac{1}{2}$ and $\frac{1}{2}$ and $\frac{1}{2}$ in the limiting of $\frac{1}{2}$ and $\frac{1}{2}$ a the following model for the class conditionals: • The idea behind GP classification is to extend logistic (or probit) regression, by assuming the following model for the class conditionals: $\frac{1}{2}$ implementation. Instead eq. (2.42) must be used. Even if the used.

$$
\pi(\mathbf{x}) = p(C_1|\mathbf{x}) = \sigma(f(\mathbf{x})) \text{ where } f \sim GP(\mu, k)
$$

f is often call *latent function*. Note that π is a random function, as f is. • *f* is often call *latent function*. Note that π is a random function, as *f* is.

Figure 3.2: Panel (a) shows a sample latent function $f(x)$ drawn from a Gaussian process as a function of x. Panel (b) shows the result of squashing this sample function through the logistic logit function, $\lambda(z) = (1 + \exp(-z))$ probability $\pi(x) = \lambda(f(x)).$ Figure 3.2: Panel (a) shows a sample latent function $f(x)$ drawn from a Gaussian process as a function of x. Panel (b) shows the result of squashing this sample funcand we are not probability $\pi(x) = \lambda(f(x))$. numismum is the function $f(x)$ is the lattent function $f(x)$ and α is a numino of α numino $f(x)$ and α is α is a numino of α of α and β is a numino of α of α and β is a numino of α or α tion through the logistic logit function, $\lambda(z) = (1 + \exp(-z))^{-1}$ to obtain the class

• Let *X*, y the observations, with $y_i \in \{0, 1\}$. $\sum_{l=1}^{n}$ sections will be to remove $\sum_{l=1}^{n}$ for $\sum_{l=1}^{n}$ \mathbf{r} the experiment with $\mathbf{r} \in (0,1)$. γ the observations, with $y_i \in \{0, 1\}$.

f(x) function from the linear logistic model in eq. (3.6) by a Gaussian process, \mathbf{S} 5.1 GP classification combined it with smooth likelihood functions, such as the logistic or probit. The logistic or probit. The logistic or problem is a the logistic or problem in the logistic or problem in the logistic or problem. The logistic

- \bullet *f* is often call *latent* or *nuisance function*. It is not observed directly. We only observe at α values of α is the inputs α is the inputs α and the class labels α and the class labels α a point **x** the realisation of a Bernoulli random variable with probability $\pi(\mathbf{x})$. process in combination with a step-function likelihood. In particular, assuming
	- Inference at a test point x^* is done, as usual in a Bayesian setting, in two steps: formulation of the model, and the computational goal pursued in the coming ϵ at a test point \bf{x} is done, as usual in a Dayesian setting, in two st ce at a test point \bf{x} is done, as usual in a Bayesian setting, in two steps:
		- 1. Compute the posterior f^* of f at the prediction point x^* .

$$
p(f_*|X, \mathbf{y}, \mathbf{x}_*) = \int p(f_*|X, \mathbf{x}_*, \mathbf{f}) p(\mathbf{f}|X, \mathbf{y}) \, d\mathbf{f},\tag{3.9}
$$

vith $p(\mathbf{f}|X, \mathbf{y}) = p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|X)/p(\mathbf{y}/X)$ by Bayes theorem. with $p(\mathbf{f}|X, \mathbf{y}) = p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|X)/p(\mathbf{y}/X)$ by Bayes theorem.

2. Compute the predictive distribution at x^*

$$
\bar{\pi}_* \triangleq p(y_* = +1|X, \mathbf{y}, \mathbf{x}_*) = \int \sigma(f_*) p(f_*|X, \mathbf{y}, \mathbf{x}_*) df_*.
$$
 (3.10)

5.2 Laplace Approximation

- As in Bayesian logistic regression, the computation of the posterior $p(f|X, y)$ cannot be carried out analytically carried out analytically.
- However, we can do a Laplace approximation of the posterior around the MAP \hat{f} . The unnormalised log posterior is:

$$
\Psi(\mathbf{f}) \triangleq \log p(\mathbf{y}|\mathbf{f}) + \log p(\mathbf{f}|X)
$$

= $\log p(\mathbf{y}|\mathbf{f}) - \frac{1}{2}\mathbf{f}^{\top}K^{-1}\mathbf{f} - \frac{1}{2}\log|K| - \frac{n}{2}\log 2\pi.$ (3.12)
Differentiating eq. (3.12) w.r.t. **f** we obtain

 $\frac{2}{2}$ $\frac{2}{2}$ Differentiating eq. (3.12) w.r.t. **f** we obtain \mathcal{A} simple observation from eq. (3.21) is that positive training examples will be positive training examples will be positive training examples with \mathcal{A} Differentialing eq. (3.12) w.r.t. I we obtain

$$
\nabla \Psi(\mathbf{f}) = \nabla \log p(\mathbf{y}|\mathbf{f}) - K^{-1}\mathbf{f},\tag{3.13}
$$

$$
\nabla \nabla \Psi(\mathbf{f}) = \nabla \nabla \log p(\mathbf{y}|\mathbf{f}) - K^{-1} = -W - K^{-1}, \quad (3.14)
$$

 w_i is diagonal, as observations are i.i.d where W is diagonal, as observations are i.i.d.

• It can be optimised with a Newton-Rapson scheme: σ_{tot} is diagonal, as observations are maximum. n_{tot} with a rewton-Kapson scheme. 44.44 Classification in the contraction of \mathcal{A} classification in the contraction in the contraction of \mathcal{A} vector machine (see section 6.4). vector machine (see section 6.4). \mathcal{L} $W_{\rm eff} = 0.000$ compute V \sim $W_{\rm eff}$ y under the variance of f $W_{\rm eff}$

$$
\mathbf{f}^{\text{new}} = \mathbf{f} - (\nabla \nabla \Psi)^{-1} \nabla \Psi = \mathbf{f} + (K^{-1} + W)^{-1} (\nabla \log p(\mathbf{y}|\mathbf{f}) - K^{-1} \mathbf{f})
$$

= $(K^{-1} + W)^{-1} (W\mathbf{f} + \nabla \log p(\mathbf{y}|\mathbf{f})).$ (3.18)

log likelihoods used likelihood functions are the logistic, and the cumulative Gaussian, see • The Laplace approximation around the MAP f is a Gaussian q with mean To gain more intuition about this update, let us consider what happens to • The Laplace approximation around the MAP \hat{f} is a Gaussian q with mean In the behavior approximation around the MAF j is a Gaussian q with mean $\frac{1}{\sqrt{2}}$ The first term is due to the value of f \mathcal{N} if we condition on a particular value of \mathcal{N} $E_{\text{max}}(X, y) = E_{\text{max}}(X, y)$ $T_{\rm eff}$ is the variance of f \sim if we condition on a particular value of \sim

$$
\mathbb{E}_q[f_*|X, \mathbf{y}, \mathbf{x}_*] = \mathbf{k}(\mathbf{x}_*)^\top K^{-1} \hat{\mathbf{f}} = \mathbf{k}(\mathbf{x}_*)^\top \nabla \log p(\mathbf{y}|\hat{\mathbf{f}}).
$$
 (3.21)

 $\mathcal{C}_\mathbf{p}$ with the exact mean, given by Opper and Winther $\mathcal{C}_\mathbf{p}$ as a discrete positive and Winther [2000] as and variance on f and thus there is an additional term of K cov(f K) K under variance cover the Gaussian approximation cover $\mathcal{L}(\mathcal{X})$ on f and thus there is an additional term of k(x, y) K1 cov(f), y) K1 cov(f), y) K1 cov(f) K1 cov(f) (X, y)K1k(
K1k(x, y)K1k(x, y)K1 under variance cover the Gaussian approximation cover $\mathcal{L}(\mathcal{X})$

$$
\mathbb{V}_q[f_*|X, \mathbf{y}, \mathbf{x}_*] = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^{\top} K^{-1} \mathbf{k}_* + \mathbf{k}_*^{\top} K^{-1} (K^{-1} + W)^{-1} K^{-1} \mathbf{k}_*
$$

= $k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^{\top} (K + W^{-1})^{-1} \mathbf{k}_*,$ (3.24)

where the last line is obtained using the matrix inversion lemma eq. (A.9).

where the last line is obtained using the matrix inversion lemma eq. (A.9). The matrix inversion lemma eq. (A.9).

• The prediction π^* can be computed by the integral

$$
\bar{\pi}_* \simeq \mathbb{E}_q[\pi_*|X, \mathbf{y}, \mathbf{x}_*] = \int \sigma(f_*)q(f_*|X, \mathbf{y}, \mathbf{x}_*) df_*, \qquad (3.25)
$$

entirely the well-explained points, and f well-explained points new

h can be approximated with the same logit-probit-logit trick $\frac{1}{2}$. $\frac{1}{2}$ which can be approximated with the same logit-probit-logit trick used for Bayesian logis- μ regression. tic regression.

5.3 Expectation Propagation

- ϵ (better) alternative to Laplace approximation is to use a variational method, typically for the probit activation function. • A (better) alternative to Laplace approximation is to use a variational method, typically
- rower than the Hessian indicates, or it could be a skew peak, while the Laplace • A first option is to approximate the posterior distribution by a Gaussian q , minimising the (reversed) KL divergence $KL(q(\mathbf{f}|X, \mathbf{y}), p(\mathbf{f}|X, \mathbf{y}))$ (the minimisation of the KL divergence $KL(\mathbf{p}(\mathbf{f}|\mathbf{Y}, \mathbf{y}))$ *q*($\mathbf{f}|\mathbf{Y}, \mathbf{y})$) is intractable) $KL(p(\mathbf{f}|X, \mathbf{y}), q(\mathbf{f}|X, \mathbf{y}))$ is intractable).
- FIFULD (over obs *i*, until convergence) a Gaussian approximation of the posterior by ively (over obs *i*, until convergence) a Gaussian approximation of the posterior by • Alternatively, one can use the Expectation Propagation algorithm, which constructs itera
	- ery (over oos *t*, until convergence) a Gaussian approximation of the posterior by
1. taking the current Gaussian approximation and factoring out the term for the *i*-th likelihood $p(y_i|f_i)$, obtaining a distribution for all observations but the *i*-th one.
	- 2. multiplying the cavity by the exact likelihood of the *i*-th observation, and finding a Gaussian approximation by moment matching of such a (non-Gaussian) distribu- $\sum_{i=1}^{n}$ tion. tion.

• EP is more accurate than Laplace approximation, and provides also an approximation of the Marginal likelihood.

5.4 Pitfalls of GP prediction

- Addition of a new observation *always* reduces uncertainty at all points → vulnerable to outliers
- Optimisation of hyperparameters often tricky: works well if σ^2 is known, otherwise it can
be seriously multimodal be seriously multimodal
- MAIN PROBLEM: GP prediction relies on a matrix inversion which scales cubically with the number of points!
- Sparsification methods have been proposed but in high dimension GP regression is likely to be tricky nevertheless