Let us define the N training points as

$$\{(x_i, y_i)\}_{i=1,2,\dots,N}, \text{ where } y_i = F(x_i)$$

and the M test points as

$$\{(x_i^*, f(x_i^*)\}_{i=1,2,\dots M}$$

The fundamental GP assumption is:

$$\begin{bmatrix} y \\ f \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} K_{X,X} & K_{X,X^*} \\ K_{X^*,X} & K_{X^*,X^*} \end{bmatrix}\right)$$

where X represents the training data and X^* represents the test data.

To get a posterior predictive distribution, we need inference in jointly Gaussian distributions.

Lemma: Let M be a general partitioned matrix,

then $M = \begin{bmatrix} E & F \\ G & H \end{bmatrix}$

$$M^{-1} = \begin{bmatrix} (M/H)^{-1} & -(M/H)^{-1}FH^{-1} \\ -H^{-1}G(M/H)^{-1} & H^{-1} + H^{-1}G(M/H)^{-1}FH^{-1} \end{bmatrix}$$

where

$$M/H = E - FH^{-1}G$$

Proof: $\begin{bmatrix} I & -FH^{-1} \\ 0 & I \end{bmatrix}\begin{bmatrix} E & F \\ G & H \end{bmatrix} = \begin{bmatrix} E-FH^{-1}G & 0 \\ G & H \end{bmatrix}$ — ①

and

$$\begin{bmatrix} E-FH^{-1}G & 0 \\ G & H \end{bmatrix}\begin{bmatrix} I & 0 \\ -H^{-1}G & I \end{bmatrix} = \begin{bmatrix} E-FH^{-1}G & 0 \\ 0 & H \end{bmatrix}$$

— ②

Considering ① and ②

$$\underbrace{\begin{bmatrix} I & -FH^{-1} \\ 0 & I \end{bmatrix}}_{X}\underbrace{\begin{bmatrix} E & F \\ G & H \end{bmatrix}}_{M}\underbrace{\begin{bmatrix} I & 0 \\ -H^{-1}G & I \end{bmatrix}}_{Z} = \underbrace{\begin{bmatrix} E-FH^{-1}G & 0 \\ 0 & H \end{bmatrix}}_{W}$$

Taking inverse on both sides,

$$Z^{-1}M^{-1}X^{-1} = W^{-1}$$

hence

$$M^{-1} = ZW^{-1}X$$

$$\Rightarrow \begin{bmatrix} E & F \\ G & H \end{bmatrix}^{-1} = \begin{bmatrix} I & 0 \\ -H^{-1}G & I \end{bmatrix} \begin{bmatrix} (M/H)^{-1} & 0 \\ 0 & H^{-1} \end{bmatrix} \begin{bmatrix} I & -FH^{-1} \\ 0 & I \end{bmatrix}$$

$$= \begin{bmatrix} (M/H)^{-1} & 0 \\ -H^{-1}G(M/H)^{-1} & H^{-1} \end{bmatrix} \begin{bmatrix} I & -FH^{-1} \\ 0 & I \end{bmatrix}$$

$$= \begin{bmatrix} (M/H)^{-1} & -(M/H)^{-1}FH^{-1} \\ -H^{-1}G(M/H)^{-1} & H^{-1} + H^{-1}G(M/H)^{-1}FH^{-1} \end{bmatrix}$$

Now,
writing the standard bivariate distribution using matrix approach,

$$p(f^*, y) = e^{\frac{-1}{2}\begin{bmatrix} f-\mu_1 \\ y-\mu_2 \end{bmatrix} \begin{bmatrix} K_{x^*,x^*} & K_{x^*,x} \\ K_{x,x^*} & K_{x^*,x^*} \end{bmatrix}^{-1} \begin{bmatrix} f-\mu_1 \\ y-\mu_2 \end{bmatrix}}$$

Now, using the previous equation, we get

$$p(f,y) = e^{\frac{-1}{2}\begin{bmatrix} f-\mu_1 \\ y-\mu_2 \end{bmatrix} \begin{bmatrix} I & 0 \\ -K_{x,x}^{-1}K_{x,x^*} & I \end{bmatrix} \begin{bmatrix} (K/K_{x,x})^{-1} & 0 \\ 0 & K_{x,x}^{-1} \end{bmatrix}}$$

$$\cdot \begin{bmatrix} I & -K_{x^*,x}K_{x,x}^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} f-\mu_1 \\ y-\mu_2 \end{bmatrix}$$

where $K/K_{x,x} = K_{x^*,x^*} - (K_{x^*,x})(K_{x,x})^{-1}(K_{x,x^*})$

$$= e^{\left\{\frac{-1}{2}(f-\mu_1-(K_{x^*,x})(K_{x,x})^{-1}(y-\mu_2))^T (K/K_{x,x})^{-1} (f-\mu_1-(K_{x^*,x})K_{x,x}^{-1}(y-\mu_2))\right\}}$$

$$\cdot e^{\left\{\frac{-1}{2}(y-\mu_2)^T K_{x,x}^{-1}(y-\mu_2)\right\}}$$

$p(y)$

Date:

hence

We have written

$$p(f, y) = (\,) \cdot p(y)$$

$$\rightarrow (\,) = p\left(\frac{f}{y}\right)$$ [posterior distribution]

hence, we can write $\mu_{1/2}$ as $\mu_1 + (K_{X^*,X})(K_{X,X})^{-1}(y - \mu_2)$

we consider y to be distributed around μ_2

hence

$$\mu_{1/2} = \mu_1 + (K_{X^*,X})(K_{X,X})^{-1}(y)$$

if we take μ_1 as zero, we get

$$\boxed{\mu_f = (K_{X^*,X})(K_{X,X})^{-1}(y)}$$

Similarly, we get

$$\boxed{K_f = K/K_{X,X} = K_{X^*,X^*} - K_{X^*,X}(K_{X,X})^{-1}K_{X,X^*}}$$

hence, $\boxed{f \mid X^*, D \sim N(\mu_f, K_f)}$

D ↓ prior data

→ The kernel function needs to be selected on the basis of trends in the data.

Eg: If the data shows a periodic trend, then we can add a periodic kernel. The effect of the kernels can be stacked.

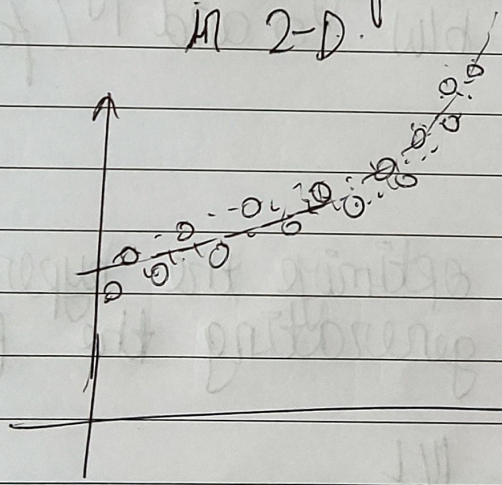
kernels can be added:

$$k_c(x, x' | \gamma_c) = k_a(x, x' | \gamma_a) + k_b(x, x' | \gamma_b)$$

and kernels multiplication:

$$k_c(x, x' | \gamma_c) \approx k_a(x, x' | \gamma_a) \cdot k_b(x, x' | \gamma_b)$$

Hence consider the following data distribution in 2-D.



The distribution can be thought of as periodic with a quadratic trend.

Hence, model kernel = $(v_1 x \cdot x') \cdot (v_2 x \cdot x') + e^{\left(\frac{2}{l^2}\right) \sin^2\left(\frac{\pi}{p}(x - x')\right)}$

$(v_1 x \cdot x')$ · $(v_2 x \cdot x')$ → (linear) × (linear)

$e^{\left(\frac{2}{l^2}\right) \sin^2\left(\frac{\pi}{p}(x - x')\right)}$ → periodic kernel

→ The utility of GP is the fact that it does not simply give us a prediction, but rather a predictive distribution. This predictive distribution can be used to asses the confidence of the prediction. If the confidence of the model goes up as new samples are encountered, we can be sure that it is extrapolating well.

→ [redacted]

→ Yes, we can adjust this system to perform classification. For example, we can use the regression output and squash it using the sigmoid function to limit its output b/w 0 and 1 (for binary classification)

→ We need to optimize the hyperparameters of the kernel generating the prior distribution.

We can use MLL

$\log(p(y|X,\theta))$, where X is the training data and θ represents the hyperparameters of the kernel.

Now $\log(p(y|X,\theta)) = \log(N(y|0, K_{X,X}))$

Since the prior samples are sampled from mean 0 and covariance matrix $K_{X,X}$.

Now, J is differentiable wrt RHS hence it can be optimized.