$$K(x, x') = \sigma^2 exp\left(-\frac{|x - x'|^2}{2l^2}\right) \tag{11}$$

$$K(x, x') = \sigma^2 exp\left(\sum_{i=1}^{k} -\frac{|x_i - x_i'|^2}{2l_i^2}\right) \tag{12}$$

$$\Sigma(x, x') = K(x, x') + I\sigma_y \tag{13}$$

Interesting idea ↓ explain rationale

Equation 12 shows the kernel function extended to multidimensional problem. for k dimensions there are k+3 hyper-parameters. l is the scale length or the horizontal scaling, effectively how quickly the correlation between two points decays for each dimension. σ is the vertical scaling. σ_y, shown in equation 13, is a representation of noise in the evaluations, this maintains some uncertainty around evaluation points. In hyper-parameter optimisation and other applications with noisy evaluations, Gaussian noise is added to the covariance matrix to avoid over-fitting. These hyper-parameters have a significant effect of the expression of the model and can be key to the final result, in particular l. Because of this, these settings are normally dealt with automatically rather than hand tuned. One approach used in [9] is to integrate over the hyper-parameters using Monte Carlo estimates, however this can be computationally expensive. Another common approach is to use a marginal likelihood estimates and optimise these settings via gradient descent.

$$p(f|\theta, D) = \mathcal{GP}(0, K(x, x')) \tag{14}$$

$$p(y|\theta, D) = \mathcal{GP}(0, K(x, x') + I\sigma_y) \tag{15}$$

Equations 14 and 15 show a definition of Gaussian processes over an noiseless and noisy function respectively. For a point of interested, y(x'), the Gaussian process can be considered a joint distribution over the y(x') and the query observation pair history D. Using the marginalisation property of gaussians this can be restructured as equation 16. An estimate of y at x' is simply $y(x')$ conditioned on $y(\boldsymbol{x})$, shown in equation 17

$$p(y(\boldsymbol{x}), y(x')) \sim \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}K(\boldsymbol{x}, \boldsymbol{x}) & K(\boldsymbol{x}, x')\\ K(\boldsymbol{x}, x')^T & K(x', x')\end{pmatrix}\right) \tag{16}$$

$$p(y(x')|y(\boldsymbol{x})) = \frac{p(\boldsymbol{x}, x')}{p(\boldsymbol{x})} \sim \mathcal{N}(\mu', \sigma') \tag{17}$$

$$\mu' = K(\boldsymbol{x}, x')K(\boldsymbol{x}, \boldsymbol{x})^{-1}y(\boldsymbol{x}) \tag{18}$$

$$\sigma' = K(x', x') - K(\boldsymbol{x}, x')K(\boldsymbol{x}, \boldsymbol{x})^{-1}K(\boldsymbol{x}, x')^T \tag{19}$$