3 Bayesian Optimisation

Bayesian Optimisation (BO) methods for hyper-parameter optimisation have become popular over the last decade due to the SOTA performance they can produce [30] [6]. One of the core weaknesses of many BO implementations is the lack of scalability, with models such as Gaussian Processes computationally scaling cubically with observations. The objective of BO and Sequential Model-Based Optimisation (SMBO) in general, in the context of hyper-parameter optimisation, can be described as trying to find $x^* = argmin\ f(x)$ where $x \in X$ and $X \subseteq \mathbf{R}^k$, X is a bounded and compact region and k is the dimensionality of the search space in our case the number of hyper-parameters. These methods assume correlation between observations and endeavour to use all of the available information to produce useful points for evaluating, x, by exploiting a model constructed based upon a set of function query-observations pairs $D = \{(x_n, y_n)\}_{n=1}^{N}$ where $y_n \sim \mathcal{N}(f(x), \sigma_n^2)$.

3.1 Posterior Model

A posterior model uses the observation history D often in combination with some prior to make estimation about the function f. These models generally model the value of f across the input space, either directly or indirectly, while also having some measure of uncertainty.

3.1.1 Gaussian Process

The most common model used in hyper-parameter optimisation is a Gaussian Process (GP)[9]. The Gaussian process is non-parametric model which is widely used due to its flexibility and simplicity, allowing many common acquisition functions to be described in a closed form, while having a well calibrated measure of uncertainty. A GP can be considered a generalisation of a multivariate Gaussian distribution to any finite number of variables. In our case this denotes all possible values within the bounded region X. A Gaussian process is fully defined, analogously to a Gaussian distribution, by a mean function, $m(x)$ and covariance function $K(x, x')$ shown in equation 9. It is common to use $m(x) = 0$ as the Gaussian process is generally robust to an arbitrary mean given sufficient data, which gives an equation for the GP as 10.

$$f(x) \sim \mathcal{N}(m(x), K(x, x')) \tag{9}$$

$$f(x) \sim \mathcal{N}(0, K(x, x')) \tag{10}$$

The covariance function or kernel function is used to generate the covariance matrices in a GP. The kernel is responsible for how both the prior and posterior are expressed. The most commonly used kernel function is the *Squared Exponential Kernel*, given in equation 11. The covariance between two points is a function of their separation scaled by the hyper-parameter l.