I will provide some notation here for myself, to help me formulate myself later.

- $\mathbb{E}[x]$ is the expected value of x.

- $\mathbf{G}$ is the GPR-model and $\mathbf{G}(x)$ is the predicted value at x.

- σ is the uncertainty of the GPR-model, and $\sigma(x)$ is the uncertainty at x

- f is the true function.

- $\mathbf{\Omega}$ is the configuration space.

- $\mathbf{X}$ are the sampled points of the configuration space.

- $f(\mathbf{X}) := \mathbf{Y}$ are the true values of said points.

Overview/Notation

The goal in global optimization of atomic structures, is to discover the configuration with the lowest energy, according to some potential V. The configuration space, which we shall call Ω, is generally of high dimension, as each additional atom in the configuration, introduces additional degress of freedom. Furthermore, precise potentials (eg. DFT) have high computational complexity. This means, that given some configuration $\mathbf{x} \in \Omega$, calculating $V(\mathbf{x})$ is a lengthy process. As a result, traditional means of optimization are infeasible.
A common way to overcome this, is to use a "surrogate model (G)" that approximates V. In this study we use *Gaussian Process Regression* to construct G, this shall process is described in the following section. The surrogate model is then used to guide the search for the global minimum. In order to decide how G is to be used, different aquisition functions can be employed. These determine how we use G to guide the search.

Gaussian Process Regression

The model G is constructed in the following way. Let us assume we have a set of configurations $X = \{\mathbf{x_1}, \ldots, \mathbf{x_k}\} \in \Omega$, where Ω is an N-dimensional configuration space. For these configurations we calculate their actual energy using the potential V such that $Y = V(X)$. In order to use this data to make predictions we make use of a gaussian process regression (GPR). A gaussian process $\mathcal{G}$ is an infinite collection of random variables $\{f_i, i \in \Omega\}$, where any finite subset of these follow a multivariate normal distribution, whose distribution is determined by a *kernel*. $\mathcal{G}$ thus assigns a random variable f_i to each point in the configuration space. A gaussian process is entirely defined by its *kernel* or alternatively *covariance funtion* often denoted $\sum$. A kernel takes two points $x, x' \in \Omega$, and outputs some positive number. It essentially measures how much the GP changes, as we move around in the configuration space. A common choice for kernel is the radial basis function (RBF) given by,

$$\sum (x, x') = \exp\left\{\|x - x'\|^2\right\}.$$

Importantly, the kernel determines the joint distribution, that the finite subsets follow. It does so, in the following way,

$$\begin{pmatrix} f_1 \\ \vdots \\ f_n \end{pmatrix} = Y \sim \mathcal{N}(0, K), \quad K_{ij} = \sum (x_i, x_j).$$

We can use this to make predictions. Let us assume that we observe some realization of $Y = \widetilde{Y}$. We wish to use this observation, to predict the value of y^* at some point x^*. As before, we can construct a joint probability distribution given by our gaussian process,

$$\begin{pmatrix} Y \\ y^* \end{pmatrix} \sim N \left(0, \widetilde{K} \right).$$

Given that we have observed $\widetilde{Y}$, we can construct the conditial probability distribution

$$\left| \begin{pmatrix} Y \\ y^* \end{pmatrix} \right| Y = \widetilde{Y}.$$

This is useful, due to the classic result that,

$$f(y) = f(x, y) f(x \mid y).$$

<div align="center">

Search Methods

</div>

Our method of modelling the 1d-function, involves updating the GPR-model iteratively. In each step, a point of the domain is sampled, evaluated and then added to the model. The point we choose to sample, is dictated by the search method. For this study I have chosen to implement 4 different search methods. These are; *bayesian search*, *aquisition functions*, *alternating search* and *expected improvement*. Besides these I have also implemented a purely random search, which is used for comparison. What follows is a brief overview of the methods, and a discussion of their benefits and drawbacks.

Bayesian

This method seeks to improve the accuracy of the model *globally*, by sampling the point where the model is most uncertain, i.e. where σ is maximized. While the method does yield a quickly improving model globally, it has some significant drawbacks. It discourages search in previously explored areas, even when the global minima is contained of these. As a consequence, it also scales poorly with the size of the configuration space, as the model is encouraged to explore areas far apart. The diagram below shows an example of such a search.

Aquisition Function

This method chooses as its next point, the minimum of a function of the following form,

$$F(p) = \hat{E}(p) - \kappa \cdot \sigma(p).$$

Where κ is some previously specified constant. This method leverages both the models current prediction of the energy and the uncertainty. It encourages, due to the $\hat{E}$, the model to search areas where it currently believes there might be a minima. However, the σ part prevents the model from becoming stuck in one area. This is due to the fact, that σ will decrease in these areas, as the model explores them. F will therefore tend to increase in explored areas. The search method has a beneficial general tendency of encouraging exploitation when σ is low and exploration when σ is low. The success of the method does however depend quite a bit on the choice of κ. A suboptimal choice of this parameter, will upset the balance of exploration and exploitation.

Expected Improvement

Of the 4 methods *expected improvement* is both the most complicated and the most relevant. Lets start with some notation

$$d = \arg\min D.$$

The method also takes a parameter ϵ, which we shall call the *required improvement*. For each in $p \in C$, we then calculate the following quantities,

$$P(p < d - \epsilon).$$

Which is the probability the model attributes to p being lower than our threshold. We also calculate,

$$\mu := \mathbb{E}\left(p \mid p < d - \epsilon\right).$$

Which is expected value at p, given that it is below the threshold.