Bayesian Optimization: Searching for the global minima

Erik Andersson, Sebastian Holmin

January 2021

In this project we will investigate the use of Gaussian Processes (GP) to model the potential energy surface (PES) for adding a Au atom to a Au slab, i.e. the difference in average energy per atom between the slab with and without the extra atom. To sample the energy we use an embedded medium theory (EMT) calculator provided in the `asap3` package. Sampling the energy this way is resource intensive, so GPs are likely well suited method for reducing the computational time for modeling the PES.

1 Local minimization

We begin by analysing the performance gain of Bayesian Optimization (BO) for finding the global minimum of the PES. As a reference we implement a simple global minimization technique where several random starting positions are used for the local minimizer `scipy.optimize.minimize`. In fig. 1 the true energy surface, sampled in a dense grid using EMT, together with the local minima are shown. From 500 local searches, the global minimum was found 13% of times, meaning that you need 17 local searches to have above a 90% chance of finding the global minimum. One local minimization took on average 39 EMT samples to compute, so for 90% accuracy you need roughly 620 EMT evaluations using this naive method.

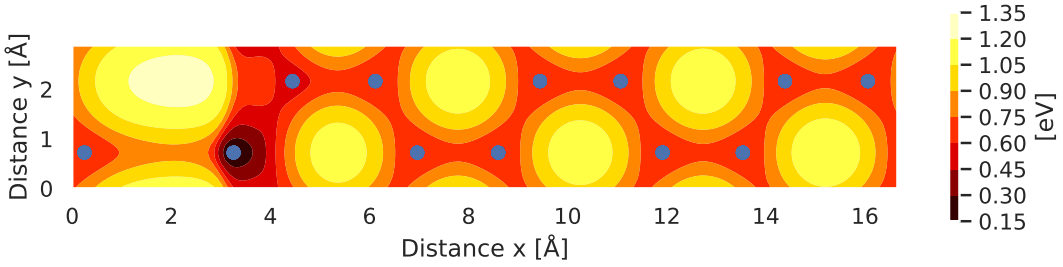
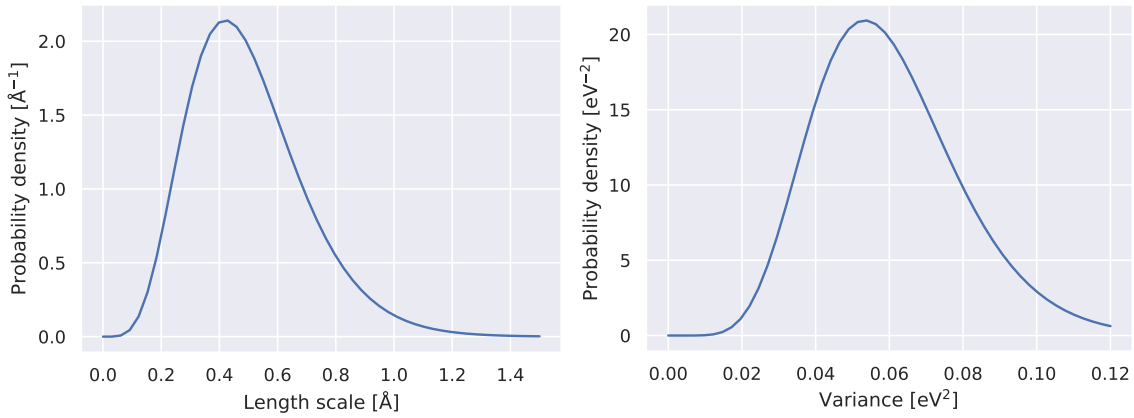


Figure 1: The potential energy surface (PES) for adding an atom of Au to a Au slab. The blue dots are local minima where such an atom might rest.



(a) (b)

Figure 2: Prior distributions for the length scale l (a), and variance σ^2 (b) of the GP.

2 Bayesian optimization

We now contrast this by implementing Bayesian optimization, which uses Gaussian Processes to estimate the surface from fewer EMT samples and finds the global minimum from this estimation.

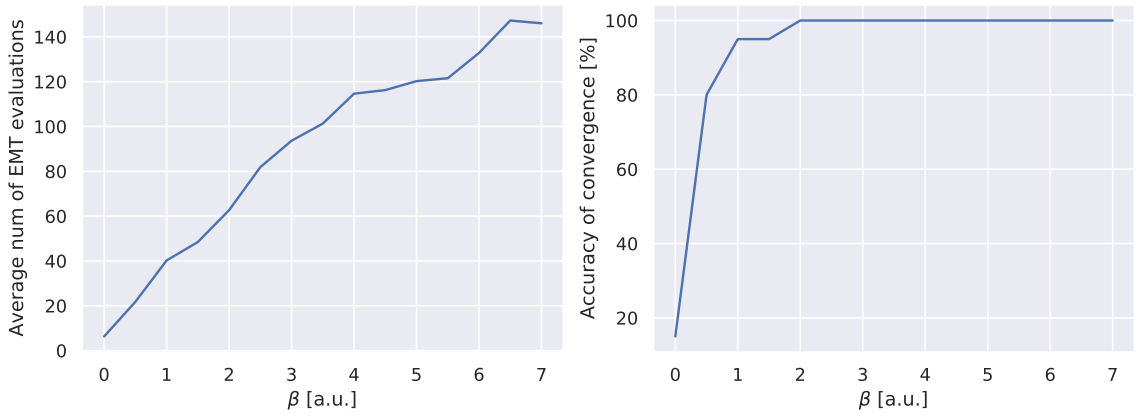
To get a suitable performance for the Bayesian optimization a number of parameters needs to be chosen properly. The first are the hyperpriors for the variance and length scale of the squared-exponential kernel (we also use a bias term). These will drastically impact the ability of the GP to estimate the surface and thus the stability of the minimization.

The hyperpriors chosen are shown in fig. 2, they are gamma functions based on analysis of the 'true' energy surface in fig. 1. The energy of the PES fluctuates in the range 1.35 eV to 0.15 eV. With our bias term, we can choose the variation of the kernel to be such that we have a three σ (99.7%) probability of generating values in this range (for points several length scales away from data points). That is, $(1.35 - 0.15)/2$ eV$= 3\sigma \Rightarrow \sigma^2 = 0.04$ eV^2. We set this value as the mean, and made the prior wide enough to give some leniency. The optimized posterior turns out to lie close to 0.04, regardless of the choice of prior, which hints that this choice was appropriate. The length scale is connected to how fast the function is expected to vary (points generated further away have small correlations) and the densely sampled PES seems to vary on the scale of 1 Å (from local maximum to closest local minimum) which suggests that the length scale should not be much longer than that and the mass of the prior should be around half that.

When doing Bayesian optimization we want to prioritize estimating the surface at points close to the global minimum while balancing exploration of the entire surface to find global minimum candidates. To do this we introduce an acquisition function with another parameter β as

$$A(x, y) = -\mu(x, y) + \beta\sigma(x, y) \tag{1}$$

where μ and σ are the mean and standard deviation from the GP estimation at the point (x, y). The PES is sampled at the maximum of $A(x, y)$ which is obtained by sampling A in a loose grid (20 by 20 points) followed by a local minimization starting from the best candidate.



(a) (b)

Figure 3: The average number of EMT evaluations needed to reach convergence (a), as well as the accuracy of the converged results for different seeds (b), plotted against the parameter β.

Given the hyperpriors above we ran the optimization for a range of β values, repeated 20 times with different random seeds (but the same set for every β), and investigated the convergence. The frequency of convergence to the global minimum (here the closest minimum was chosen) is shown in fig. 3b where we see that a β value of roughly 2.5, corresponding to on average 82 EMT samples, is necessary to get a stable convergence. In fig. 3a we plot the average amount of EMT evaluations used for the optimization as a function of β. We can see that a lower β, as expected, leads to a fewer necessary EMT samples. Note however that even for large β the amount of samples is significantly smaller than for the repeated local minimization we tested previously with a larger probability of success.

We now continue using $\beta = 2.5$ and rerun the Bayesian optimization. A detailed picture of the trained model can be seen in fig. 4. We can see that, after convergence, the GP roughly models the oscillations of fig. 1, but with reduced accuracy far away from the global minimum. As expected, we can see that the uncertainty is large in the spaces farthest from sampled positions. These regions generally coincide with regions were the PES is large, as expected since these are much less likely to contain the global minimum. We can also see that the acquisition function after convergence is largest at the global minimum, but also large in regions of high uncertainty, which is also to be expected.

Note also that there are a large amount of samples at the border of the area, this comes from bounds on the local optimization. This was necessary since $A(x, y)$ will not capture the periodicity of the PES. A more elegant solution might be to utilize a periodic kernel for the GP, which might also provide a more stable fit.

3 Transition paths barriers

We now aim to model global properties of the PES, instead of the global minimum. Specifically, we will model energy along a transition path from the global minimum to a local minimum at $(11, 2.1)$.

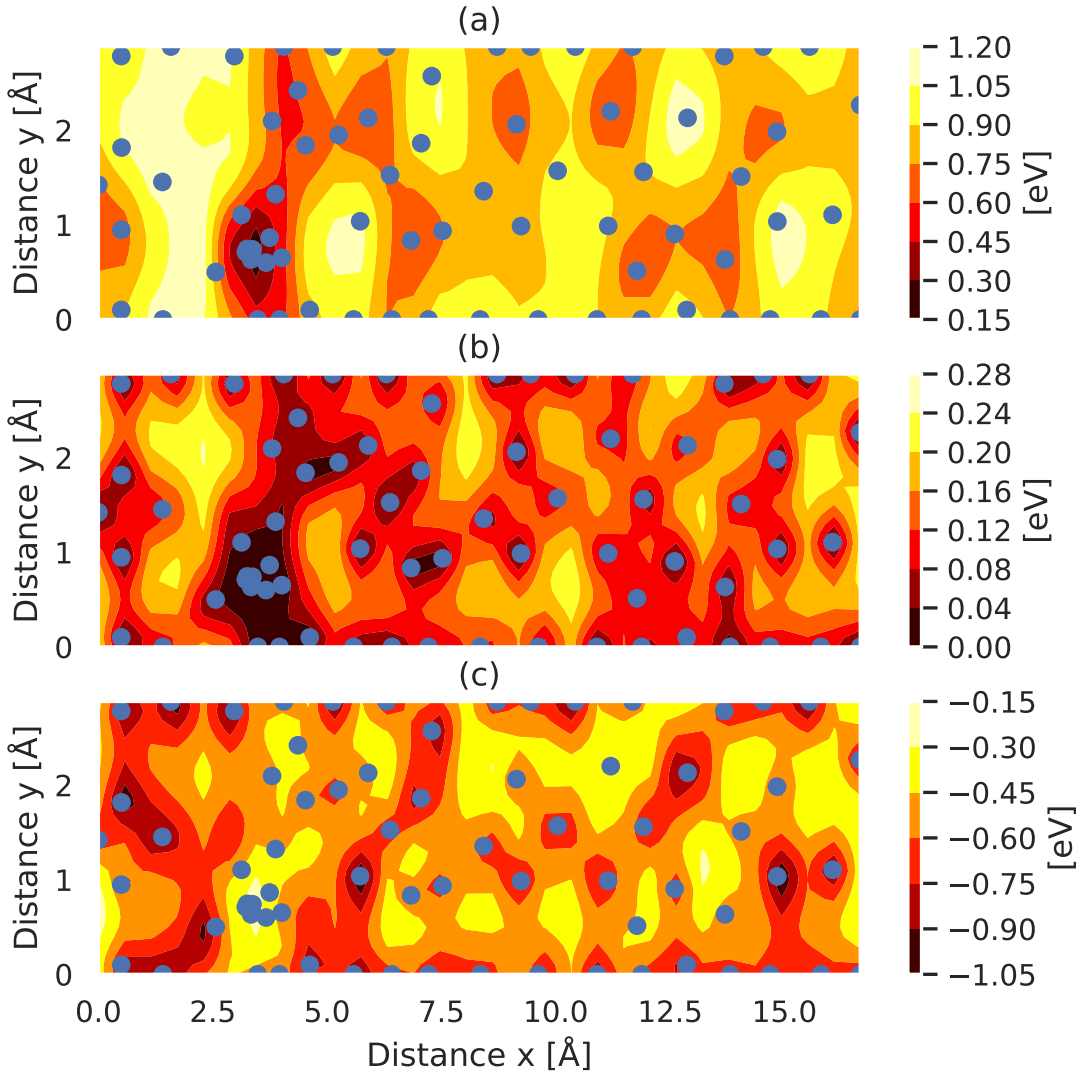
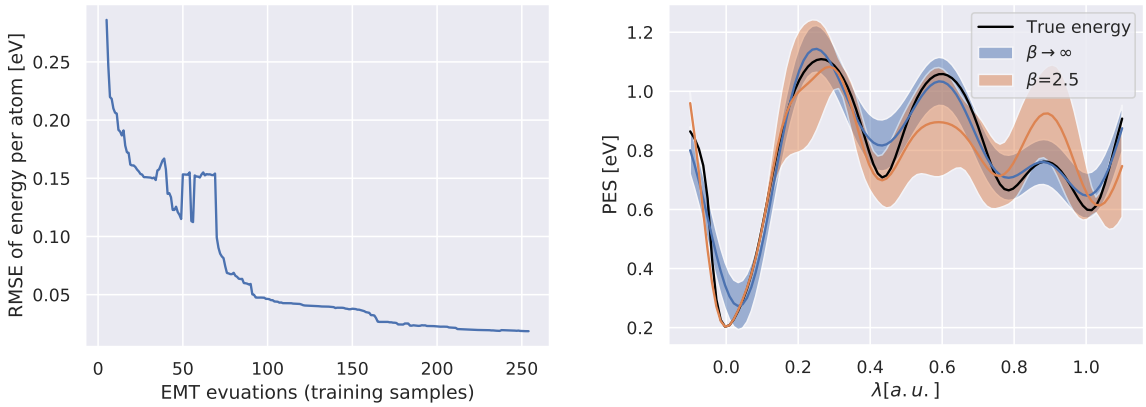


Figure 4: (a) The PES predicted by the GP after convergence. (b) The standard deviation σ after convergence. (c) The acquisition function after convergence. We can see that the uncertainty is large in regions where the PES is large.



(a) (b)

Figure 5: (a) Root mean squared error of the general purpose GP. (b) The predictions of the optimization GP ($\beta = 2.5$) and the general purpose GP ($\beta \to \infty$) together with the densely sampled 'true' PES.

To this end we replace the acquisition function with

$$A(x, y) = \sigma(x, y)$$

which now solely prioritizes exploration (it is equivalent to the $\beta \to$ inf limit of eq. (1)). We compare this model with the densely sampled data from fig. 1 and compute the RMSE as a function of training samples, which is shown in fig. 5a. We can see that at 100 samples the gain in accuracy is significantly reduced per training sample, this corresponds to an error of 0.05 eV which is less than 5% of the maximum fluctuation of energy in the PES. Whether this accuracy is sufficient in practice of course depends on the physical properties you are trying to estimate.

Another feature of fig. 5a worth discussing is the discontinuous behaviour around 50 samples. In this region the GP entirely fails to fit certain training points, including the points at $\lambda = 0$ and $\lambda = 1$, instead it predicts a slowly varying curve across the entire area. This behaviour is not straight forward to explain, but might be connected to how the training points are fairly evenly spaced apart, which limits the GPs ability to pick up on higher frequency components.

We now investigate the performance of this model at 100 samples compared to the previously trained one with acquisition function (1) using $\beta = 2.5$. The mean and standard deviation of the models are shown in fig. 5b together with the 'true' value (densely sampled using EMT) along the 1D path from the global minimum at $(3.26, 0.72)$ to the local minimum at $(11, 2.1)$. We can see that of course the general purpose GP does much better at large over the entire interval, which is not surprising since it was trained using more data points. Note however that it fails to predict the position of the global minimum. This is not very surprising since the global minimum is very narrow and the GP is unlikely to sample again close to the given training point at $\lambda = 0$. This could potentially be remedied by adding a term to the acquisition function which encourages sampling areas with large magnitude derivatives.