Chapter 2
Theory of Bayesian Optimization

In this chapter, we introduce the theory of Bayesian optimization procedure and illustrate its application to a simple problem. A more involved application of Bayesian optimization will be presented in Chap. 3.

2.1 Bayesian Interpretation of Probability

Consider rolling a die with k sides, labeled a_1, a_2, …, and a_k, respectively. Let $P(a_j)$ be the 'probability' that a particular side a_j appears after rolling the die. Before attempting to calculate $P(a_j)$, it is necessary to clarify the meaning of the word 'probability'. In other words, we must specify what the number $P(a_j)$ quantifies. In statistics, the concept of 'probability' is formally interpreted in one of two ways. In the *frequentist interpretation of probability*, $P(a_j)$ is the fraction of times that a_j appears in a very large number of die rolls. In the *Bayesian interpretation of probability*, $P(a_j)$ is the extent to which we believe that the number a_j will appear prior to rolling the die.

For the case of a die with k sides, there is little difference between the frequentist and Bayesian interpretation of probability. Given that the die is not biased in any way, we would set $P(a_j) = 1/k$ in both the frequentist and Bayesian interpretations. However, a major difference between the frequentist and Bayesian interpretation of probability arises when we consider so-called *learning-type problems*, in which new information on the system becomes available over time. For example, consider a robot whose job is to sort oranges from lemons. Suppose that the robot is presented with a fruit, and that the robot has no useful information to help distinguish between oranges and lemons. For example, the robot does not know that round fruit are more likely to be oranges rather than lemons. In this case, the robot would set $P(\text{o}) = 1/2$ and $P(\text{l}) = 1/2$, where $P(\text{o})$ and $P(\text{l})$ are the probabilities that the fruit is an orange or lemon, respectively. Now, suppose that new information is loaded into the robot's memory from an external source, namely

D. Packwood, *Bayesian Optimization for Materials Science*,
SpringerBriefs in the Mathematics of Materials 3,
https://doi.org/10.1007/978-981-10-6781-5_2

$$L(\mathrm{r}|\mathrm{o}) = 8, \tag{2.1}$$

and

$$L(\mathrm{r}|\mathrm{l}) = 1. \tag{2.2}$$

These numbers are called *likelihoods*, and result from measurements on different types of fruits by the external source. $L(\mathrm{r}|\mathrm{o})$ measures the 'likelihood' that an orange is round, and $L(\mathrm{r}|\mathrm{l})$ measures the 'likelihood' that a lemon is round. The precise physical meaning of 'likelihood' and it units are do not need to made so clear, providing that the values of the likelihoods are always measured in a consistent way. In order to utilize the information provided by the likelihoods, we employ a formula called *Bayes' rule*. Bayes' rule can be written as

$$P(\mathrm{o}|\mathrm{r}) \propto L(\mathrm{r}|\mathrm{o})P(\mathrm{o}) \tag{2.3}$$

$$P(\mathrm{l}|\mathrm{r}) \propto L(\mathrm{r}|\mathrm{l})P(\mathrm{l}), \tag{2.4}$$

where $P(\mathrm{o}|\mathrm{r})$ and $P(\mathrm{l}|\mathrm{r})$ are the probability that a round fruit is an orange, and the probability that a round fruit is a lemon, respectively. Substituting the numbers given above, we find that $P(\mathrm{o}|\mathrm{r}) \propto 8 \times 0.5 = 4$ and $P(\mathrm{l}|\mathrm{r}) \propto 1 \times 0.5 = 0.5$. Eliminating the proportionality constants then gives $P(\mathrm{o} \mid \mathrm{r}) = 4/(4 + 0.5) = 0.89$ and $P(\mathrm{o}|\mathrm{r}) = 0.5 / (4 + 0.5) = 0.11$. Thus, when presented with a round fruit, the robot will determine that there is a probability of 0.89 that the fruit is an orange and a 0.11 probability that the fruit is a lemon. With Bayes' rule, the robot is therefore able to improve its ability to classify fruit by incorporating information provided by an external source. This kind of process is not natural within the frequentist interpretation of probability, in which the probabilities $P(\mathrm{o})$, $P(\mathrm{r})$, $P(\mathrm{o}|\mathrm{r})$ and $P(\mathrm{l}|\mathrm{r})$ remain fixed for all time, regardless of any new information which may appear.

Within the Bayesian interpretation of probability, $P(\mathrm{o})$ and $P(\mathrm{l})$ are referred to as *prior probabilities* and $P(\mathrm{o}|\mathrm{r})$ and $P(\mathrm{l}|\mathrm{r})$ are referred to as *posterior probabilities*. Note that the likelihoods $L(\mathrm{r}|\mathrm{o})$ and $L(\mathrm{r}|\mathrm{l})$ in Eqs. (2.1) and (2.2) can be regarded as a function of the type of fruit. For this reason, L is referred to as a *likelihood function.*

As one might have guessed, Bayesian optimization makes use of the Bayesian interpretation of probability and Bayes' rule. We elaborate upon this point further in the following section.

2.2 Equilibrium Bond Lengths *Via* Bayesian Optimization

Consider the problem of estimating the equilibrium bond length r_0 of a diatomic molecule such as Br_2. By 'equilibrium bond length', we mean that the interatomic potential energy $u(r)$ is minimized when $r = r_0$. We suppose that the analytical form of $u(r)$ is unknown, i.e., that we cannot find r_0 by directly differentiating a simple

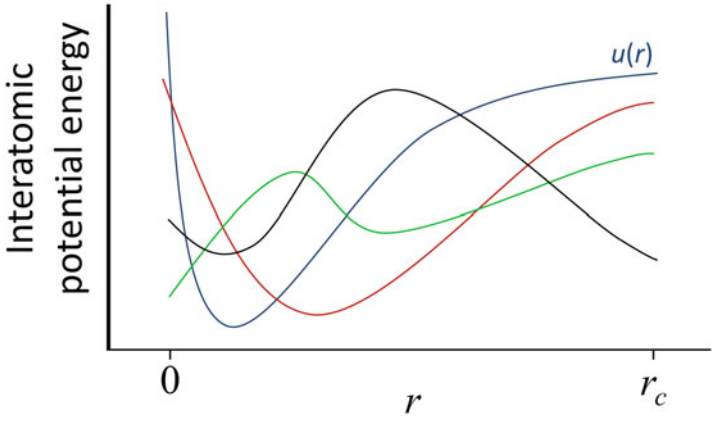

Fig. 2.1 Sketch of the sample space Ω (see main text). Only four candidate functions are shown. The function shown in blue corresponds to the true interatomic potential for the diatomic molecule. The sample space contains all candidate functions that are real-valued and differentiable between 0 and r_c. This includes functions which are physically unreasonable, such as the ones shown in green and black

formula. While the analytical form of $u(r)$ is unknown, it is a physical requirement that $u(r)$ be real-valued, continuous, and differentiable over the interval $0 \leq r \leq r_c$, where, r_c is the dissociation limit of the molecule. For simplicity, we assume that r_c is a well-defined and known constant. We define the *sample space* Ω as the collection of all real-valued functions which continuous and differentiable over the interval $0 \leq r \leq r_c$ (Fig. 2.1). In the present context, Ω can be thought of as a collection of 'candidate functions' for the interatomic potential, one of which corresponds to the true interatomic potential, *u*. Ω contains an infinite number of functions.

Estimation of the equilibrium bond length r_0 via Bayesian optimization runs according to the following steps. (i) Generate a random sample of interatomic separations and measure (or calculate, from first-principles) *u* for each case. (ii) Independently of (i), assign a prior probability to the functions in Ω. The prior probability measures our intuitive feeling about which functions in Ω correspond to the true interatomic potential. (iii) Use the sample data and Bayes' rule to calculate the posterior probability for the functions in Ω. (iv) Use the posterior distribution to estimate r_0. The estimated value of r_0 is denoted r^*. (*v*) Measure (or calculate from first-principles) $u(r^*)$, the interatomic potential at distance r^*, and add r^* and $u(r^*)$ to the sample data. (vi) Repeat steps (ii)–(v) until the global minimum of the interatomic potential is identified (i.e., when the minimum value of *u* in the sample remains unchanged over several iterations).

Note that, strictly speaking, Bayesian optimization assumes that all functions in the sample space are finite. This assumption is actually violated for the present system, because it is a physical requirement that $u(0) = \infty$ for the true interatomic potential *u*. In the present analysis, we will get around this issue by simply supposing that $u(0)$ is finite.

2.2.1 Prior Probability

In Bayesian optimization, we choose a multivariate Gaussian distribution for the prior probability distribution. In other words, choose s interatomic separations r_i, r_j, …, r_k (where $0 \leq r_l \leq r_d$ for $l = i, j, \ldots,$ or k). Let $u(r_i)$, $u(r_j)$, …, and $u(r_k)$ be the values of the true interatomic potential at points r_i, r_j, …, and r_k. The prior probability that the vector $(u(r_i), u(r_j), \ldots, u(r_k))$ is contained in an infinitesimal region of space centered at point $\mathbf{v} = (v_i, v_j, \ldots, v_j)$ is given by $g(v_i, v_j, .., v_k)dv_i dv_j \cdots dv_k$, where

$$g(v_i, v_j, \ldots, v_k) = \frac{1}{(2\pi)^{s/2}|\mathbf{K}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{v}-\boldsymbol{\mu})^T \mathbf{K}^{-1}(\mathbf{v}-\boldsymbol{\mu})\right) \tag{2.5}$$

is referred to as the *prior probability density*. In Eq. (2.5), the $s \times 1$ column matrix $\boldsymbol{\mu}$ is called the *mean vector*, the $s \times s$ matrix $\mathbf{K}$ is called the *covariance matrix*, and $|\mathbf{A}|$ denotes the determinant of matrix $\mathbf{A}$. In Eq. (2.5), $\mathbf{v}$ is treated as an $s \times 1$ column vector. If $g(v_i, v_j, \ldots, v_k)$ is particularly large, then it means that we have a strong intuitive feeling that $u(r_i) = v_i$, $u(r_j) = v_j$, …, and $u(r_k) = v_k$ for the true interatomic potential energy function u.

Our intuitive beliefs about the interatomic potential u are encoded into the prior distribution through the mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{K}$. The choice of $\boldsymbol{\mu}$ and (particularly) $\mathbf{K}$ has a large influence on the performance of Bayesian optimization, and therefore they should be carefully considered before applying Bayesian optimization to a physical problem.

If we write $\boldsymbol{\mu} = (\mu_i, \mu_j, \ldots, \mu_k)$ for the mean vector, then μ_i can be regarded as our intuitive guess for $u(r_i)$, the actual value of the interatomic potential at interatomic separation r_i. For example, we might choose the harmonic oscillator potential as an initial guess for u, and write

$$\mu_i = \frac{1}{2}C(r_i - R_0)^2, \tag{2.6}$$

where the parameters C and R_0 are guessed by considering literature data for similar diatomic molecules. There are no particular restrictions on the values of μ_i, however they must be finite.

Let us write $\mathbf{K} = [K_{ij}]_{s\times s}$ for the covariance matrix. If V is chosen at random from Ω according to the prior distribution, then K_{ij} measures the extent to which we believe $V(r_i)$ should be correlated with $V(r_j)$. To quantify this correlation, let L be an intuitive guess for the correlation length of the interatomic potential $u(r)$. Roughly speaking, L measures the length over which r must change in order for $u(r)$ to change significantly. Returning now to K_{ij}, we would expect for K_{ij} to be large when $|r_i - r_j| < L$, and moreover K_{ij} should decrease rapidly as $|r_i - r_j|$ increases beyond L. This behavior can be acquired by choosing a squared exponential function for K_{ij}, i.e.,

$$K_{ij} = a \exp\left(-\frac{\left| r_i - r_j \right|^2}{2L^2} \right), \tag{2.7}$$

where a is another constant. If a single point $V(r_i)$ is randomly generated from Ω in accordance with the prior distribution, then a is interpreted as the mean-square deviation of $V(r_i)$ from μ_i. This interpretation follows from the fact that the diagonal elements of $\mathbf{K}$ formally correspond to the variance of $V(r_i)$, when V is randomly generated from the prior distribution.

The constants a and L are referred to as *hyperparameters* and have a critical influence on the performance of Bayesian optimization. While a and L can be also chosen based on our intuitive feelings about the system, in practice it is preferable to choose them via a training procedure. We will describe this in Sect. 2.6 and in the following chapter.

Note that Bayesian optimization is not restricted to the covariance matrix defined in Eq. (2.7). Mathematically speaking, it is only necessary for the covariance matrix to be positive semidefinite. Some alternative forms of the covariance matrix are discussed in reference [1]. The advantage of Eq. (2.7) is that it physical interpretation is relatively straightforward.

2.2.2 *Likelihood Function and Posterior Distribution*

In Bayesian optimization, the likelihood function is assumed to be the same Gaussian density function as was used for the prior probability density in Eq. (2.5). To explain what is meant here, let us start by re-writing Eq. (2.5) as

$$\begin{aligned} &g\left(v_\alpha, v_\beta, \ldots, v_\gamma, v_i, v_j, \ldots, v_k\right) \\ &= \frac{1}{(2\pi)^{\frac{m+s}{2}} |\mathbf{K}|^{1/2}} \exp\left(-\frac{1}{2} \left[\begin{pmatrix} \mathbf{v}_{\alpha:\gamma} \\ \mathbf{v}_{i:k} \end{pmatrix} - \begin{pmatrix} \boldsymbol{\mu}_{\alpha:\gamma} \\ \boldsymbol{\mu}_{i:k} \end{pmatrix} \right]^T \begin{bmatrix} \mathbf{K}_{\alpha:\gamma,\alpha:\gamma} & \mathbf{K}_{\alpha:\gamma,i:k} \\ \mathbf{K}_{i:k,\alpha:\gamma} & \mathbf{K}_{i:k,i:k} \end{bmatrix}^{-1} \left[\begin{pmatrix} \mathbf{v}_{\alpha:\gamma} \\ \mathbf{v}_{i:k} \end{pmatrix} - \begin{pmatrix} \boldsymbol{\mu}_{\alpha:\gamma} \\ \boldsymbol{\mu}_{i:k} \end{pmatrix} \right] \right) \end{aligned} \tag{2.8}$$

In Eq. (2.8), $\mathbf{v}_{\alpha:\gamma} = (v_\alpha, v_\beta, \ldots, v_\gamma)$, $\mathbf{v}_{i:k} = (v_i, v_j, \ldots, v_k)$, m is the length of the vector $\mathbf{v}_{\alpha:\gamma}$, s is the length of the vector $\mathbf{v}_{i:k}$, $\mathbf{K}_{\alpha:\gamma,\alpha:\gamma}$ is the covariance matrix for points r_α, r_β, …, and r_γ, $\mathbf{K}_{i:k,i:k}$ as the covariance matrix for points r_i, r_j, …, and r_k, and

$$\mathbf{K}_{\alpha:\gamma,i:k} = \begin{bmatrix} K_{\alpha i} & K_{\alpha j} & \cdots & K_{\alpha k} \\ K_{\beta i} & K_{\beta j} & \cdots & K_{\beta k} \\ \vdots & \vdots & \ddots & \vdots \\ K_{\gamma i} & K_{\gamma j} & \cdots & K_{\gamma k} \end{bmatrix} \tag{2.9}$$

Note that $\mathbf{K}_{i:k,\alpha:\gamma}$ is the transpose of $\mathbf{K}_{\alpha:\gamma,i:k}$. The likelihood function used in Bayesian optimization is defined as

$$L\left(v_i, v_j, \ldots, v_k \middle| v_\alpha, v_\beta, \ldots, v_\gamma\right) = g\left(v_i, v_j, \ldots, v_k \middle| v_\alpha, v_\beta, \ldots, v_\gamma\right). \tag{2.10}$$

Here, $g(v_i, v_j, \ldots, v_k | v_a, v_\beta, \ldots, v_\gamma)$ is a so-called *conditional density*. It corresponds to the prior probability density in Eq. (2.8) calculated a point $(v_i, v_j, \ldots, v_k)$ with the values of v_α, v_β, …, and v_γ held fixed. An analytic formula for the conditional density can be written, however it turns out that this formula is not necessary for our purposes (see Appendix 2.1).

Let us now suppose that we are provided with a sample of s points $(r_i, u(r_i))$, $(r_j, u(r_j))$, …, $(r_k, u(r_k))$, where for $t = i, j, \ldots$, and k, we have $0 < r_t < r_d$ and the value of $u(r_t)$ is known exactly. In analogy to Eqs. (2.3) and (2.4), the posterior probability that the vector $(u(r_\alpha), u(r_\beta), \ldots, u(r_\gamma))$ is contained in an infinitesimal region of space centered at point $\mathbf{v} = (v_\alpha, v_\beta, \ldots, v_\gamma)$ is given by Bayes' rule, namely

$$\begin{aligned} &f\left(v_\alpha, v_\beta, \ldots, v_\gamma \middle| u_i, u_j, \ldots, u_k\right) dv_\alpha dv_\beta \cdots dv_\gamma \\ &\quad \propto L\left(u_i, u_j, \ldots, u_k \middle| v_\alpha, v_\beta, \ldots, v_\gamma\right) g\left(v_\alpha, v_\beta, \ldots, v_\gamma\right) dv_\alpha dv_\beta \cdots dv_\gamma, \end{aligned} \tag{2.11}$$

where we have used the notation $u_i = u(r_i)$. $f(v_\alpha, v_\beta, \ldots, v_\gamma | u_i, u_j, \ldots, u_k)$ is referred to as the *posterior probability density*. Substituting Eqs. (2.5) and (2.10) into Eq. (2.11) and performing various manipulations, we obtain (see Appendix 2.1),

$$f\left(v_\alpha, v_\beta, \ldots, v_\gamma \middle| u_i, u_j, \ldots, u_k\right) \propto \exp\left(-\frac{1}{2}\left[\mathbf{v}_{\alpha:\gamma} - \boldsymbol{\mu}^*_{\alpha:\gamma}\right]^T \left(\mathbf{K}^*_{\alpha:\gamma,\alpha:\gamma}\right)^{-1} \left[\mathbf{v}_{\alpha:\gamma} - \boldsymbol{\mu}^*_{\alpha:\gamma}\right]\right), \tag{2.12}$$

where

$$\boldsymbol{\mu}^*_{\alpha:\gamma} = \boldsymbol{\mu}_{\alpha:\gamma} - \mathbf{K}_{\alpha:\gamma,i:k}\mathbf{K}^{-1}_{i:k,i:k}(\mathbf{u}_{i:k} - \boldsymbol{\mu}_{i:k}), \tag{2.13}$$

and

$$\mathbf{K}^*_{\alpha:\gamma,\alpha:\gamma} = \mathbf{K}_{\alpha:\gamma,\alpha:\gamma} - \mathbf{K}_{\alpha:\gamma,i:k}\mathbf{K}^{-1}_{i:k,i:k}\mathbf{K}_{i:k,\alpha:\gamma}. \tag{2.14}$$

The right-hand side of Eq. (2.12) is actually the unnormalized multivariate Gaussian distribution. If we are interested in the posterior density a single point v_α, then Eq. (2.12) simplifies to

$$f\left(v_\alpha \middle| u_i, u_j, \ldots, u_k\right) = \frac{1}{\sqrt{2\pi K^*_{\alpha\alpha}}} \exp\left(-\frac{\left(v_\alpha - \mu^*_\alpha\right)^2}{2K^*_{\alpha\alpha}}\right), \tag{2.15}$$

where

$$\mu^*_\alpha = \mu_\alpha - \mathbf{K}_{\alpha,i:k}\mathbf{K}^{-1}_{i:k,i:k}(\mathbf{u}_{i:k} - \boldsymbol{\mu}_{i:k}), \tag{2.16}$$

$$K^*_{\alpha\alpha} = K_{\alpha\alpha} - \mathbf{K}_{\alpha,i:k}\mathbf{K}^{-1}_{i:k,i:k}\mathbf{K}_{i:k,\alpha}, \tag{2.17}$$

the row-vector $\mathbf{K}_{\alpha,i:k}$ is defined as $(K_{\alpha,i}, K_{\alpha,j}, \ldots, K_{\alpha,k})$, and $\mathbf{K}_{i:k,\alpha}$ is the transpose of $\mathbf{K}_{\alpha,i:k}$. In practice, we use Eqs. (2.15–2.17) in all calculations of the posterior distribution.

2.2.3 *Example Calculation of the Posterior Distribution*

Figure 2.2a–c plot the mean and variance of the posterior distribution from Eqs. (2.16) and (2.17) calculated from a sample of three interatomic displacements for an isolated Br_2 molecule. The red line represents the posterior mean, the thin black lines measure the posterior variance and correspond to the 95% confidence limits of the posterior distribution, i.e.,

$$\mu^*_\alpha \pm 1.96\sqrt{\mathbf{K}^*_{\alpha,\alpha}},$$

and the blue line shows the actual interatomic potential energy curve. The potential energy at each point was calculated from first principles using density functional theory (DFT). DFT calculations discussed here and elsewhere in this chapter were performed with the VASP code [2], using a plane wave basis set, projector-augmented wave (PAW) potentials, and the generalized gradient approximation (GGA).

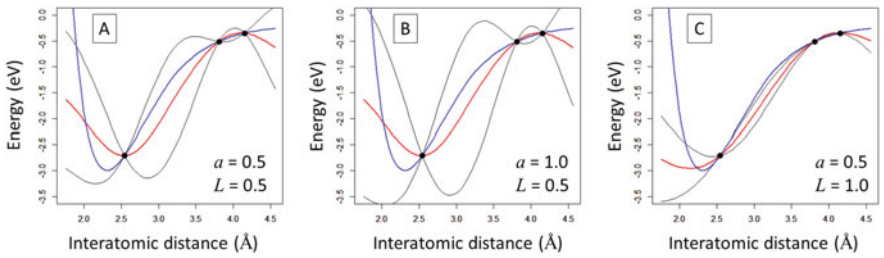

Fig. 2.2 Calculation of the posterior distribution for the interatomic potential energy for a Br_2 molecule, using a sample of three interatomic distances and the corresponding energies and various values of the hyperparameters a and L. The blue curve corresponds to the true interatomic potential energy, the red curve correspond to the mean of the posterior distribution (Eq. 2.16), and the black curves correspond to the 95% confidence limits of the posterior distribution. See Sect. 2.3 for details. Note that the y axis actually plots the total energy of the Br_2 molecule, which is equal to the interatomic potential energy plus a constant

By looking at the sample points in Fig. 2.2a–c, it is clear that the posterior mean interpolates the sample points exactly, and that the posterior variance is zero at the sample points. This is a general behavior of Bayesian optimization, and shows that the posterior distribution predicts the sample data exactly.

An important observation is that both the posterior mean and variance depend on the values of the hyperparameters a and L in the covariance matrix in Eq. (2.7). In general, the posterior variance grows and then shrinks back to zero as we move between successive sample points. As the hyperparameter L increases and the 'correlation length' of the system grows, the posterior variance grows more slowly between successive sample points and the posterior mean begins to resemble a linear interpolation between successive sample points. In this sense, when the correlation length in the system is assumed to be large, we become more confident that true potential energy curve can be obtained by a linear interpolation between successive points. In Fig. 2.2c, in which the correlation length is large, it can be seen that the true potential energy curve (blue) actually lies outside of the 95% confidence limits of the Gaussian distribution, showing that Bayesian optimization predicts a very small probability for the true interatomic potential when the correlation length L is large. In general, we should choose the hyperparameters so that the true interatomic potential lies within the 95% confidence limits of the Gaussian distribution. In this situation, the posterior density will be large for functions closely resembling the true interatomic potential, and the Bayesian optimization procedure will be able to accurately estimate the location of the minimum of the true potential curve.

2.2.4 The Expected Improvement

Having generated the posterior distribution from the sample data, we now need to predict the point which minimizes the interatomic potential energy. There are a variety of ways of predicting the position of the optimum using the posterior distribution. One of the most popular methods is involves the *expected improvement*, which we consider here.

The expected improvement at point r_α is defined as

$$EI(r_\alpha) = E_f\left[\max\left(u_{\min} - V(r_\alpha), 0\right)\right], \tag{2.18}$$

where E_f[A] is the expected value (average) of the random variable A with respect to the posterior distribution in Eq. (2.18), and $u_{\min}$ is the minimum interatomic potential energy in the sample. $V(r_\alpha)$ is the value of a function V evaluated at point r_α, where V has been randomly generated from the sample space according to the posterior distribution. The interatomic distance which minimizes the interatomic potential energy is then estimated as the value of r_α which maximizes Eq. (2.18). Thus, the expected improvement considers our current 'best guess' of the minimum interatomic potential, $u_{\min}$, and then determines the point which, on average, will improve upon that guess the most.

In order to calculate the expected improvement, the following formula may be used (see Appendix 2.2),

$$EI(r_\alpha) = \left(u_{\min} - \mu_\alpha^*\right)\Phi\left(\frac{u_{\min} - \mu_\alpha^*}{\sqrt{K_{\alpha\alpha}^*}}\right) + \sqrt{K_{\alpha\alpha}^*}\,\varphi\left(\frac{u_{\min} - \mu_\alpha^*}{\sqrt{K_{\alpha\alpha}^*}}\right). \tag{2.19}$$

In Eq. (2.19), $\Phi(x)$ and $\varphi(x)$ are the normal distribution function and normal density function, respectively, evaluated at point x. Both functions can be can be easily called in a statistical programming environment such as *R* [3].

2.2.5 Example Run of Bayesian Optimisation

In order to demonstrate the calculation of the expected improvement, and to show the Bayesian optimization procedure in action, we return to the example of the Br_2 molecule discussed at the end of the previous section. Figure 2.3a plots the posterior mean (red line) and confidence limits (black lines) from a sample of two interatomic distances, using hyperparameter values $a = 0.5$ and $L = 0.5$ and energies calculated via DFT. The green line represents the expected value calculated from Eq. (2.18). The peak of the expected improvement lies at 2.25 Å. The true interatomic potential energy $u(r)$ is then calculated for the interatomic distance $r = 2.25$ Å via DFT, and this data is added to the sample. Figure 2.3b plots the posterior mean, confidence limits, and expected improvement for the new sample. This time, the expected improvement peaks at 2.27 Å, and so the true interatomic potential energy is calculated at this interatomic displacement, and this data is added to the sample. After repeating this procedure only a few more times (Fig. 2.3c, d), the expected improvement peaks at 2.30 Å (Fig. 2.3e). This corresponds to the exact optimum interatomic bond length for the Br_2 molecule (within the accuracy of the present DFT method), showing that the calculation has converged to the global optimum within relatively few iterations of the Bayesian optimization procedure.

A close look Fig. 2.3a–e unveils a key feature of Bayesian optimization. Comparing the expected improvement calculated at successive rounds of the Bayesian optimization procedure, we see that the added sample points are widely scattered and are not localized at any particular point. For example, at the end of the first, second, and third rounds of Bayesian optimization (Fig. 2.3a–c), the expected value peaks at 2.58, 2.27, and 1.76 Å, respectively, and these values and the corresponding interatomic potential energy are added to the sample data. This shows that Bayesian optimization is a non-local search method, which is in contrast with conventional optimizers which rely on a local gradient. The reason for the non-locality of Bayesian optimization is that the posterior distribution in Eq. (2.15) is computed by utilizing *all* information in the sample, which may be scattered

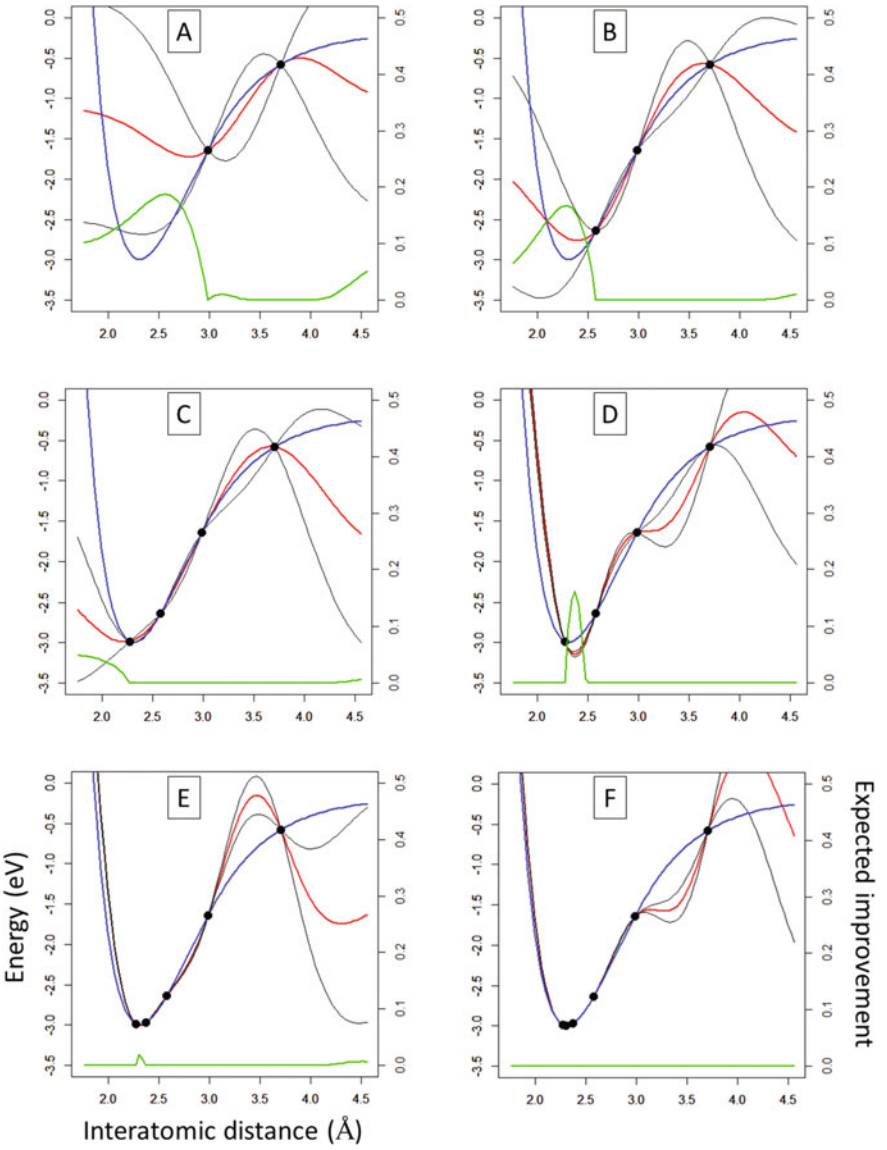

Fig. 2.3 Bayesian optimization procedure for estimating the interatomic displacement which minimizes the interatomic potential of a Br_2 molecule. Blue, red, black, and green curves correspond to the true potential energy curve, the mean of the posterior distribution, the 95% confidence limits of the posterior distribution, and the expected improvement, respectively. Starting with a sample of two interatomic displacements and their energies (**a**), sample data is successively added according to the maximum of the expected improvement. Note that the point added D is not shown, as it appears beyond the scale of these plots. Hyperparameter values of $a = 0.5$ and $L = 0.5$ were used. See main text for details

around the space. The non-locality of Bayesian optimization makes it less prone (but not immune) to getting trapped in local minima.

Figure 2.3a–e also demonstrate the so-called *exploitation versus exploration trade-off* concept [4]. The expected improvement tends to grow as the posterior mean decreases and the posterior variance increases. The former effect encourages investigation of 'promising' regions ('exploitation'), on the basis of information contained in the sample, whereas the latter effect encourages the exploration of regions in which we have little sample information ('exploration'). Exploitation is evident in Fig. 2.3f and e, in which the general region of the potential minimum becomes apparent and the search focuses on this region. Exploration is evident in Fig. 2.3c, in which the relatively small interatomic distance of 1.76 Å is suddenly added to the sample, causing the Bayeisan optimization procedure to gather information on the system at very small interatomic distance. The extent of exploration versus exploitation is determined by the posterior mean and variance, which in turn are strongly affected by the hyperparameters a and L. This shows once again the importance of the hyperparameters in determining the effectiveness of the Bayesian optimization procedure.

The performance of Bayesian optimization is further shown in Fig. 2.4. Figure 2.4a plots the minimum interatomic potential energy and optimal interatomic distance for the Br_2 molecule as a function of the sample size used in the calculation of the posterior distribution. The sample size corresponds to the number of DFT calculations. Because the initial sample contained 2 points, the number of iterations of Bayesian optimization is equal to the sample size −1. In Fig. 2.4b, the minimum interatomic potential energy and optimal interatomic distance for the case of random sampling from a grid of 83 interatomic displacements between 1.76 and 4.46 Å. For the latter calculations, the minimum interatomic potential energies and

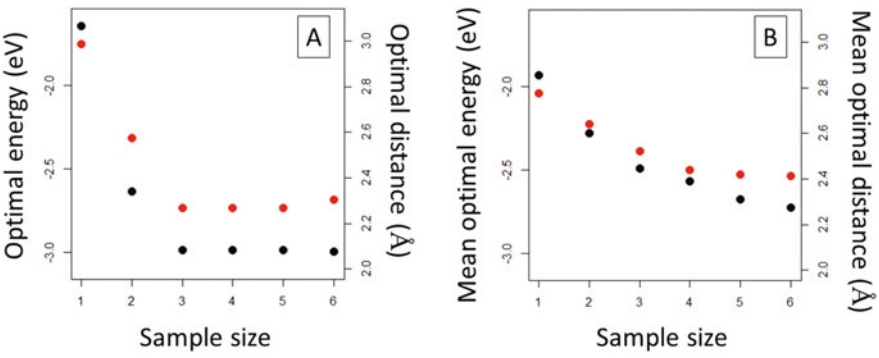

Fig. 2.4 **a** Minimum interatomic potential energy (black points) and corresponding interatomic distance (red points) predicted by the Bayesian optimization procedure in Fig. 2.3. The interatomic distance at the true minimum of the potential energy curve is 2.30 Å. **b** Minimum interatomic potential energy and corresponding interatomic distance predicted from random sampling interatomic displacements (see text for details). The data points in (b) have been averaged across 100 independent rounds of random sampling

optimal interatomic distances have been averaged across 100 rounds of sampling. While Bayesian optimization is able to find the optimal interatomic displacement (2.30 Å) within 5 iterations of the procedure, random sampling, on average, predicts a much larger value of around 2.42 Å for the same sample size.

2.2.6 Training

The discussion and the end of the previous section demonstrates that the performance of the Bayesian optimization procedure is heavily affected by the choice of parameters a and L for the covariance matrix. In the context of Bayesian optimization, *training* refers to a procedure for choosing the 'best' values of a and L for the prior distribution. Here, 'best' refers to the values of a and L which result in the quickest identification of the global optimum for a given sample of data. While the 'best' values of a and L can often be determined by physical intuition, it is common (although not necessarily more reliable) to choose these values by a more statistical approach. One of the most standard of these statistical approaches is referred to as *marginal likelihood maximization* (MLM), which we introduce here.

Continuing with the problem of finding the equilibrium bond distance in a diatomic molecule, suppose again that we have a sample of s interatomic bond distances and the corresponding energies, $(r_i, u(r_i))$, $(r_j, u(r_j))$, …, $(r_k, u(r_k))$. In the MLM approach, we find the values of the hyperparameters a and L which maximize the value of the prior distribution in Eq. (2.5) when calculated for the points in the sample data. More precisely, we wish to obtain the values of the hyperparameters which maximize $g(u_i, u_j, \ldots, u_k)$, where g is the prior distribution in Eq. (2.5) and $u_i = u(r_i)$, $u_j = u(r_j)$, …, and $u_k = u(r_k)$ are the sample data for the interaction potential.

For the special case of a constant prior mean $\boldsymbol{\mu}$ and a squared exponential function (as in Eq. (2.7)) for the covariance function, we can use the following equations to maximize the prior distribution, namely (see Appendix 2.3)

$$\log g\left(u_i, u_j, \ldots, u_k\right) = -\frac{s}{2}\log\left(\frac{1}{s}|\mathbf{R}|^{1/s}(\mathbf{u}-\boldsymbol{\mu})^T\mathbf{R}^{-1}(\mathbf{u}-\boldsymbol{\mu})\right) + c, \tag{2.20}$$

where c is a term which is independent of a and L, and

$$a = \frac{1}{s}(\mathbf{u}-\boldsymbol{\mu})^T\mathbf{R}^{-1}(\mathbf{u}-\boldsymbol{\mu}), \tag{2.21}$$

where $\mathbf{R} = [R_{ij}]_{s\times s}$ and

$$R_{ij} = \exp\left(-\frac{\left(r_i - r_j\right)^2}{2L^2}\right). \tag{2.22}$$

First, the value of L which maximizes the right-hand size of Eq. (2.20) is identified by numerically computing Eq. (2.20) across a grid of candidate values of L. This value of L is then substituted directly into Eq. (2.21) to obtain the optimum value of a.

For the special case of the Br_2 molecule studied here, the use of MLM approach to obtain the hyperparameters does not improve the performance of Bayesian optimization compared to the results discussed at the end of Sect. 2.5. However, the MLM approach can be useful for situations in which more difficult systems are studied and it is not possible to guess a value of a or a typical 'correlation distance' from intuition. In any case, because hyperparameters have a critical influence on the performance of Bayesian optimization, we strongly advocate for the use of a physically motivated procedure to estimate good values for a and L. Such a procedure is discussed in the following chapter.

2.3 Bayesian Optimization in the General Case

The above formalism can be immediately applied to systems beyond a simple diatomic molecule. In the general case, we have n *objects*, x_1, x_2, …, x_n, where object x_k has a *property* $h(x_k)$. We wish to identify the object whose property has the minimum value. These objects may be different types of materials or different configurations of molecules, and the properties may be material properties such as thermal conductivity or molecular properties such as HOMO energy level.

In order to apply the formalism above to the general case, we simply replace the interatomic distances r_i, r_j, …, with the objects x_i, x_j, …, and replace the interatomic potential energies $u(r_i)$, $u(r_j)$, …, with the properties $h(x_i)$, $h(x_j)$, …. The only major change to the above formalism is in the covariance function in Eq. (2.6). In place of Eq. (2.7), we must write

$$K_{ij} = a \exp\left(-\frac{d\left(x_i, x_j\right)^2}{2L^2} \right), \tag{2.23}$$

where $d(x_i, x_j)$ measures the degree of similarity between objects x_i and x_j. The specific definition of $d(x_i, x_j)$ is arbitrary, however for excellent performance of Bayesian optimization it is essential that $d(x_i, x_j)$ be chosen after giving very careful consideration to the physics of the problem under study. Usually, $d(x_i, x_j)$ is defined as

$$d\left(x_i, x_j\right) = \left|\left|\phi(x_i) - \phi\left(x_j\right)\right|\right|, \tag{2.24}$$

where $\phi(x_i)$ and $\phi(x_j)$ are referred to as *feature vectors* (or *descriptors*) for the objects x_i and x_j, respectively (see Eq. (1.4) for the definition of the $\| \ \|$ notation).

As discussed in the previous chapter, feature vectors are real-valued vectors which encode the key physics of the system of interest. In the example of the diatomic molecule, the feature vector for the distance r_i was simply set to $\phi(r_i) = r_i$. However, in most cases it is not so obvious which feature vectors one should use in order to achieve good performance with Bayesian optimization, and a great deal of physical intuition is needed to deduce such feature vectors. In any case, once such feature vectors are available, Bayesian optimization proceeds exactly as described in the sections above.

In this chapter, we have discussed Bayesian optimization in the context of minimization. However, Bayesian optimization can also be used to solve problems related to maximization as well. For the case where we wish to find the value of r which maximizes the value of some function u, the expected improvement in Eq. (2.18) must be re-written as

$$EI(r_\alpha) = E_f\left[\max\left(V(r_\alpha) - u_{\max}, 0\right)\right], \tag{2.25}$$

where u_{max} is the largest value of u in the sample data. Moreover, Eq. (2.19) must be replaced with

$$EI(r_\alpha) = \left(\mu_\alpha^* - u_{\max}\right)\Phi\left(\frac{\mu_\alpha^* - u_{\max}}{\sqrt{K_{\alpha\alpha}^*}}\right) + \sqrt{K_{\alpha\alpha}^*}\varphi\left(\frac{\mu_\alpha^* - u_{\max}}{\sqrt{K_{\alpha\alpha}^*}}\right) \tag{2.26}$$

Equation (2.26) can be proven by following similar steps to those shown in Appendix 2.2. Apart from the definition of the expected improvement, no changes to the theoretical framework developed above are necessary for solving maximization problems via Bayesian optimization.

2.4 *R* Code for Bayesian Optimization

One of the great advantages of Bayesian optimization is that it is relatively easy to implement in a computational environment. A program for calculating the expected improvement using an initial sample of Br-Br interatomic distances and the corresponding potential energies is available online at http://www.packwood.icems.kyoto-u.ac.jp/download/

This code is written in the *R* programming language, and can be executed within the *R* command line interface [3]. The *R* command line interface can be downloaded freely at https://www.r-project.org/, and numerous tutorials on *R* can be found online.

Successive applications of this code can be used to find the optimal distance between the Br atoms. Note that this code assumes that the energies have been pre-calculated for all points on a tight grid of bond distances, and that the initial

sample is drawn randomly from this pre-calculated data. Obviously there is no need to perform Bayesian optimization in this case, as the equilibrium distance could be identified by directly looking at the pre-calculated data. In realistic applications of Bayesian optimization, the code will need to interface with first-principles calculation software or an experimental apparatus in order to obtain the sample data and subsequent measurements.

Appendix 2.1

To prove Eq. (2.12) – (2.14), we first substitute Eqs. (2.10) into (2.11) to obtain

$$f\left(v_\alpha, v_\beta, \ldots, v_\gamma \middle| u_i, u_j, \ldots, u_k\right) \propto g\left(u_i, u_j, \ldots, u_k \middle| v_\alpha, v_\beta, \ldots, v_\gamma\right) g\left(v_\alpha, v_\beta, \ldots, v_\gamma\right) \tag{2.27}$$

Because the likelihood function and the prior density are Gaussian probability densities, Eq. (2.27) simplifies to (by the basic properties of conditional densities [5])

$$f\left(v_\alpha, v_\beta, \ldots, v_\gamma \middle| u_i, u_j, \ldots, u_k\right) \propto g\left(v_\alpha, v_\beta, \ldots, v_\gamma, u_i, u_j, \ldots, u_k\right), \tag{2.28}$$

or, by using Eq. (2.8),

$$f\left(v_\alpha, v_\beta, \ldots, v_\gamma \middle| u_i, u_j, \ldots, u_k\right) \propto \exp\left(-\frac{1}{2}\left[\begin{pmatrix}\mathbf{v}_{\alpha:\gamma}\\ \mathbf{v}_{i:k}\end{pmatrix} - \begin{pmatrix}\boldsymbol{\mu}_{\alpha:\gamma}\\ \boldsymbol{\mu}_{i:k}\end{pmatrix}\right]^T \begin{bmatrix}\mathbf{K}_{\alpha:\gamma,\alpha:\gamma} & \mathbf{K}_{\alpha:\gamma,i:k}\\ \mathbf{K}_{i:k,\alpha:\gamma} & \mathbf{K}_{i:k,i:k}\end{bmatrix}^{-1} \left[\begin{pmatrix}\mathbf{v}_{\alpha:\gamma}\\ \mathbf{v}_{i:k}\end{pmatrix} - \begin{pmatrix}\boldsymbol{\mu}_{\alpha:\gamma}\\ \boldsymbol{\mu}_{i:k}\end{pmatrix}\right]\right) \tag{2.29}$$

This expression can be simplified using an identity which applies to block matrices (see, Ref. [6])

$$\begin{bmatrix}\mathbf{K}_{\alpha:\gamma,\alpha:\gamma} & \mathbf{K}_{\alpha:\gamma,i:k}\\ \mathbf{K}_{i:k,\alpha:\gamma} & \mathbf{K}_{i:k,i:k}\end{bmatrix}^{-1} = \begin{bmatrix}\left(\mathbf{K}_{\alpha:\gamma,\alpha:\gamma} - \mathbf{K}_{\alpha:\gamma,i:k}\mathbf{K}_{i:k,i:k}^{-1}\mathbf{K}_{i:k,\alpha:\gamma}\right)^{-1} & \left(\mathbf{K}_{\alpha:\gamma,\alpha:\gamma} - \mathbf{K}_{\alpha:\gamma,i:k}\mathbf{K}_{i:k,i:k}^{-1}\mathbf{K}_{i:k,\alpha:\gamma}\right)^{-1}\mathbf{K}_{\alpha:\gamma,i:k}\mathbf{K}_{i:k,i:k}^{-1}\\ -\left(\mathbf{K}_{i:k,i:k} - \mathbf{K}_{i:k,\alpha:\gamma}\mathbf{K}_{\alpha:\gamma,\alpha:\gamma}^{-1}\mathbf{K}_{\alpha:\gamma,i:k}\right)^{-1}\mathbf{K}_{i:k,\alpha:\gamma}\mathbf{K}_{\alpha:\gamma,\alpha:\gamma}^{-1} & \left(\mathbf{K}_{i:k,i:k} - \mathbf{K}_{i:k,\alpha:\gamma}\mathbf{K}_{\alpha:\gamma,\alpha:\gamma}^{-1}\mathbf{K}_{\alpha:\gamma,i:k}\right)^{-1}\end{bmatrix} \tag{2.30}$$

Substituting Eqs. (2.30) into (2.29) and performing some tedious but straightforward algebraic manipulations yields Eqs. (2.12)–(2.14).

Appendix 2.2

To prove Eq. (2.19), we write

$$\begin{aligned} EI(r_\alpha) &= E_f[\min(u_{\min} - V(r_\alpha), 0)] \\ &= \int_{-\infty}^{u_{\min}} (u_{\min} - z) \frac{1}{\sqrt{2\pi K^*_{\alpha\alpha}}} e^{-\left(z-\mu^*_\alpha\right)^2 / 2K^*_{\alpha\alpha}} dz \\ &= \underbrace{u_{\min} \int_{-\infty}^{u_{\min}} \frac{1}{\sqrt{2\pi K^*_{\alpha\alpha}}} e^{-\left(z-\mu^*_\alpha\right)^2 / 2K^*_{\alpha\alpha}} dz}_{A} - \underbrace{\int_{-\infty}^{u_{\min}} \frac{z}{\sqrt{2\pi K^*_{\alpha\alpha}}} e^{-\left(z-\mu^*_\alpha\right)^2 / 2K^*_{\alpha\alpha}} dz}_{B}. \end{aligned} \tag{2.31}$$

The term A on the right-hand side of Eq. (2.31) is equal to

$$A = u_{\min} \Phi\left(\frac{u_{\min} - \mu^*_\alpha}{\sqrt{K^*_{\alpha\alpha}}}\right), \tag{2.32}$$

by the definition of the cumulative normal distribution. As for the term B in Eq. (2.31), we write

$$\begin{aligned} B &= \int_{-\infty}^{u_{\min}} \frac{\mu^*_\alpha + \left(z - \mu^*_\alpha\right)}{\sqrt{2\pi K^*_{\alpha\alpha}}} e^{-\left(z-\mu^*_\alpha\right)^2 / 2K^*_{\alpha\alpha}} dz. \\ &= \mu^*_\alpha \Phi\left(\frac{u_{\min} - \mu^*_\alpha}{\sqrt{K^*_{\alpha\alpha}}}\right) + \underbrace{\int_{-\infty}^{u_{\min}} \frac{\left(z - \mu^*_\alpha\right)}{\sqrt{2\pi K^*_{\alpha\alpha}}} e^{-\left(z-\mu^*_\alpha\right)^2 / 2K^*_{\alpha\alpha}} dz}_{C}. \end{aligned} \tag{2.33}$$

By substituting the variable

$$h = \frac{z - \mu^*_\alpha}{\sqrt{2K^*_{\alpha\alpha}}} \tag{2.34}$$

into the term C in Eq. (2.33) and performing the integration, we obtain

$$\begin{aligned} C &= \left(\frac{2K^*_{\alpha\alpha}}{\pi}\right)^{1/2} \int_{-\infty}^{u_{\min}} h e^{-h^2} dh \\ &= -\left(\frac{K^*_{\alpha\alpha}}{2\pi}\right)^{1/2} e^{-\left(u_{\max}-\mu^*_\alpha\right)^2 / 2K_{\alpha\alpha}} \\ &= -\sqrt{K^*_{\alpha\alpha}}\,\phi\left(\frac{u_{\min}-\mu^*_\alpha}{\sqrt{K^*_{\alpha\alpha}}}\right) \end{aligned} \tag{2.35}$$

where the definition of the standard normal probability density was used. We obtain the result after combining Eqs. (2.31), (2.32), (2.33) and (2.35).

Appendix 2.3

To prove Eqs. (2.20) and (2.21), we take the logarithm of the prior probability density in Eq. (2.5) to obtain

$$\begin{aligned} \log g\left(u_i, u_j, \ldots, u_k\right) &= -\frac{1}{2}(\mathbf{u}-\boldsymbol{\mu})^T (a\mathbf{R})^{-1}(\mathbf{u}-\boldsymbol{\mu}) - \frac{1}{2}\log|a\mathbf{R}| - \frac{s}{2}\log 2\pi \\ &= \underbrace{-\frac{1}{2}(\mathbf{u}-\boldsymbol{\mu})^T (a\mathbf{R})^{-1}(\mathbf{u}-\boldsymbol{\mu})}_{A} \underbrace{-\frac{s}{2}\log\left(a|\mathbf{R}|^{1/s}\right)}_{B} \underbrace{-\frac{s}{2}\log 2\pi}_{C}. \end{aligned} \tag{2.36}$$

In the first line of Eq. (2.36), we used the definition of the matrix $\mathbf{R}$ in Eq. (2.22) and the fact that $a\mathbf{R} = \boldsymbol{\Sigma}$. In the second line, we used the fact that $|a\mathbf{R}| = a^s|\mathbf{R}|$, which follows from the basic properties of determinants. Solving the equation $\partial \log g(u_i, u_j, \ldots, u_k)/\partial a = 0$ gives

$$a = \frac{1}{s}(\mathbf{u}-\boldsymbol{\mu})^T \mathbf{R}^{-1}(\mathbf{u}-\boldsymbol{\mu}), \tag{2.37}$$

which is simply Eq. (2.21). To obtain an expression for L, fist note that the term A reduces to

$$A = -s/2 \tag{2.38}$$

upon substituting Eq. (2.37). Substituting Eq. (2.37) into the term marked B gives

$$B = -\frac{s}{2}\log\left(\frac{1}{s}|\mathbf{R}|^{1/s}(\mathbf{u}-\boldsymbol{\mu})^T \mathbf{R}^{-1}(\mathbf{u}-\boldsymbol{\mu})\right). \tag{2.39}$$

Substituting Eqs. (2.38) and (2.39) into Eq. (2.34), and noting that the terms A and C are independent of a and L, gives Eq. (2.22). Finally, by noting that maximization of the logarithm of the prior probability density is equivalent to maximizing the prior probability density itself, we arrive at the result.

References

1. Rasmussen CE, Williams CKI. Gaussian processes for machine learning. MA, USA: The MIT Press; 2016 (Chapter 4).
2. Kresse G, Furthmuller J. Efficient iterative schemes for ab initio total energy calculations using a plane-wave basis set. Phys Rev B. 1996;54:11169–86.
3. R Core Team. R: a language and environment for statistical computing. R foundation for statistical computing. https://www.R-project.org/ (2017).
4. Frazier P, Wang J. Bayesian optimization for materials design. In: Lookman T, Alexander FJ, Rajan K, editors. Information science for materials discovery and design. Springer Series in Materials Science 225, Switzerland: Springer International Publishing; 2016.
5. Miller I, Miller M. John E. Freund's mathematical statistics with applications. 7th ed. Upper Saddle River, NJ, USA: Pearson Prentice-Hall; 2014.
6. Petersen KB, Pedersen MS. The matrix cookbook. http://matrixcookbook.com (Section 9.1.3). 15 Nov 2012.

http://www.springer.com/978-981-10-6780-8

Bayesian Optimization for Materials Science

Packwood, D.

2017, VIII, 42 p. 16 illus., 12 illus. in color., Softcover
ISBN: 978-981-10-6780-8