

Federal State Autonomous Educational Institution National Research University
"Higher School of Economics"

Faculty of Computer Science
Basic education programme:
Applied Mathematics and Informatics

Research Project Report on the Topic:

*Development of methods to optimise the data collection system for determining
the parameters of physical systems or for building machine learning models*

Fulfilled by:

Student of the group БПМИ209
Kemal Azamat



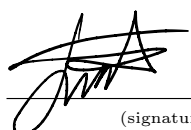
(signature)

19/05/2023

(date)

Coursework manager:

Tarakanov Aleksandr Aleksandrovich



(signature)

21/05/2023

(date)

Contents

1	Annotation	3
1.1	Annotation in Russian	3
2	Literature review	4
3	Introduction	5
3.1	Hermite polynomials as basis function	5
3.2	Constructing a design matrix	6
3.3	Calculating a quadratic form	6
3.4	Kullback-Leibler distance (KL-divergence)	7
3.5	Expected informational gain	7
4	Maximising the expected information gain. Optimiser.	9
5	Experiments. Visualisation of results.	10
6	Conclusion	13
7	References	14

1 Annotation

This term paper, "Developing Optimization Methods for Data Acquisition System to Build Machine Learning Models", discusses optimal Bayesian design methods for optimizing experiment parameters to maximize measurement value. A new approach is proposed to optimize the observational space and the experimental setup to obtain optimal formulas which reduce the need to calculate numerous integrals and reduce computational cost.

A surrogate model based on polynomial chaos expansion (PCE), which uses orthogonality of basis polynomials for efficient approximation, is presented. The concept of expected information gain as a target function for determining the optimal experiment parameters on a limited computational budget is considered. The proposed optimal Bayesian experimental design approach is evaluated using numerical test cases, demonstrating its advantages over traditional methods. The results indicate a significant improvement in the efficiency of data collection and the quality of the constructed machine learning models.

This work represents an important step in the development of optimal Bayesian experimental design methods, providing a practically applicable approach for improving the data collection process for machine learning models.

1.1 Annotation in Russian

В данной курсовой работе "Разработка методов оптимизации системы сбора данных для построения моделей машинного обучения" рассматриваются методы оптимального байесовского проектирования для оптимизации параметров эксперимента с целью максимизации значения измерений. Предлагается новый подход к оптимизации пространства наблюдения и экспериментальной установки для получения оптимальных формул, которые уменьшают необходимость вычисления многочисленных интегралов и снижают вычислительные затраты.

Представлена суррогатная модель на основе полиномиального расширения хаоса (PCE), которая использует ортогональность базисных полиномов для эффективной аппроксимации. Рассматривается концепция ожидаемого прироста информации как целевой функции для определения оптимальных параметров эксперимента при ограниченном вычислительном бюджете. Предложенный подход к оптимальному байесовскому проектированию экспериментов оценивается с помощью численных тестовых примеров, демонстрирующих его преиму-

щества перед традиционными методами. Результаты показывают значительное улучшение эффективности сбора данных и качества построенных моделей машинного обучения.

Данная работа представляет собой важный шаг в развитии методов оптимального байесовского экспериментального дизайна, обеспечивая практически применимый подход для улучшения процесса сбора данных для моделей машинного обучения. *Keywords: polynomials, normalization, design matrix, quadratic form, mean value, KL-divergence, expected information gain, gradient lift, quadrature, variation derivative, target function replacement, orthogonality of basis polynomials*

2 Literature review

I started my literature review with Jorge Nocedal and Stephen J. Wright, "Numerical Optimisation". It provides an overview of the theory and practice of numerical optimisation, including iterative methods of optimisation such as gradient lifting, which I will use, as well as second order methods such as Newton's method. I have identified for myself the gradient-lift method used in this book. Already applied numerical calculation method for gradient calculation. Each step can be indicated by the following formula:

$$x_{new} = x_{old} + \alpha \nabla f(x_{old}),$$

where x_{new} , x_{old} - new and old approaches, α - learning rate, $\nabla f(x_{old})$ - gradient of function f at the given point.

Optimal Bayesian Design (OBED) methods have become an important tool in the field of statistical machine learning. Panin and Marzouk (2018) discuss the importance of the relationship between data, a priori information and expected information gain in the context of OBED. The probabilistic interpretation of data and a priori information (distributions), as O'Hagan (1994) points out, is one of the key elements for a Bayesian approach. In modern years there has been considerable interest in the application of surrogate models, including polynomial chaos expansion (PCE). Marzouk et al. (2007) demonstrate how stochastic spectral methods can effectively solve inverse problems in the context of a Bayesian approach.

The importance of choosing the right Bayesian experiment design has been highlighted in Kaloner and Verdinelli (1995) and Panin (2017). They point out that the value of a measurement is determined by its contribution to the expected information gain, which is expressed through the Kullback-Leibler distance (KLD) between a priori and posterior probability densities.

The expected information gain can be expressed as:

$$EIG = \int \log \left(\frac{p(\theta|x, d)}{p(\theta)} \right) p(\theta|x, d) d\theta$$

The next source turned out to be Walter Gautschi, "Orthogonal Polynomials: Computation and Approximation". There has been a lot of research into orthogonal polynomials, including Hermite polynomials. They are known to have a number of properties that will prove useful in my research in the future. Some of the key properties that influenced my choice are: orthogonality, the ability to be used to approximate functions, and their relationship to Gaussian probability distributions. Decomposing a function into a series of Hermite polynomials can provide an exact or approximate representation of the function with a finite number of terms.

The article "Bayesian Experimental Design for Nonlinear Mixed-Effects Models with Application to HIV Dynamics" by Cong Han and Kathryn Chaloner presents a study that applies Bayesian methodology to experimental design for nonlinear models. In the case of this paper, these are mixed-effects models with application to HIV dynamics. The approaches and methods proposed for optimal data collection and optimality estimation gave a strong impetus to my research.

The study uses a Bayesian methodology to incorporate a priori information about the model parameters and obtain Bayesian estimates after observing the data. The incorporation of uncertainty and variability in the parameters was clearly demonstrated.

3 Introduction

3.1 Hermite polynomials as basis function

In today's world, where data volumes are growing at an incredible rate, data processing and analysis techniques play a crucial role in many scientific research and application areas. One of the key aspects of data processing is data collection. In this context, the task of optimal experiment design becomes particularly relevant. The coursework is devoted to the development of methods for optimizing the data collection system for building machine learning models. The paper considers an approach based on optimal Bayesian design methods. The main goal is to identify the most valuable points in the observational space to obtain the most informative data. The problem is to determine which experiment parameters should be changed and in what sequence in order to maximize the informativeness of the resulting data. This involves optimising various parameters, such as noise level and degree of polynomial chaos, and choosing the optimal experimental design

in a Bayesian context. In addition, the manuscript additionally presents a novel methodology to assess the anticipated information gain, thereby enabling enhanced accuracy in quantifying the worth of distinct measurements. A pivotal facet of this approach lies in the utilization of surrogate models, which effectively mitigate the computational overhead.

Central to this study is the use of Hermite polynomials as basis functions. Hermite polynomials, which are solutions of the Hermite differential equation, can be expressed through a recurrence relation:

$$\mathbf{H}_{n+1}(\mathbf{x}) = 2\mathbf{x} \cdot \mathbf{H}_n(\mathbf{x}) - 2n\mathbf{H}_{n-1}(\mathbf{x}),$$

where $\mathbf{H}_n(\mathbf{x})$ is a Hermite polynomial of degree n .

3.2 Constructing a design matrix

The next step is to construct a design matrix. The design matrix is used in machine learning to solve a linear regression problem. It is a matrix in which the columns are the values of the basis functions, in this case Hermite polynomials, calculated at the sample points. If we denote \mathbf{X} is the vector of input data and n is the maximum degree of the polynomial, the design matrix \mathbf{A} can be expressed as follows:

$$\mathbf{A}_{ij} = \mathbf{H}_j(\mathbf{X}_i),$$

where $\mathbf{H}_j(\mathbf{X}_i)$ is the Hermite polynomial of degree j calculated at point \mathbf{X}_i . We then compute a quadratic form for the posterior distribution. Recall that the posterior distribution plays a key role in the Bayesian approach to statistical inference.

3.3 Calculating a quadratic form

The quadratic form is an important part of the definition of the posterior distribution. It is calculated as the sum of two parts: the first part is based on the design matrix and the standard deviation of the probability, and the second part is a diagonal matrix obtained by dividing one by the square of the standard deviation of the a priori distribution.

The quadratic form \mathbf{Q} of the posterior distribution is calculated using the formula:

$$\mathbf{Q} = (\mathbf{A}^T \mathbf{A}) / \sigma_{\text{lh}}^2 + \mathbf{I} / \sigma_{\text{pr}}^2,$$

where σ_{lh} is a standard deviation of the likelihood function, \mathbf{A}^T is a transposed design matrix, \mathbf{I} is a unit matrix, σ_{pr} is a standard deviation of prior distribution,

Let us define the Kullback-Leibler divergence.

3.4 Kullback-Leibler distance (KL-divergence)

The Kullback-Leibler distance (KL-divergence) plays a crucial role in our study, as it serves as a key metric for determining the difference between probability distributions. It is a measure that quantifies the difference between two probability distributions. The Kullback-Leibler divergence $\mathbf{D}_{\text{KL}}(\mathbf{P}||\mathbf{Q})$ between two distributions \mathbf{P} and \mathbf{Q} is defined as follows:

$$\mathbf{D}_{\text{KL}}(\mathbf{P}||\mathbf{Q}) = \sum_{\mathbf{i}} \mathbf{P}(\mathbf{i}) \cdot \log \left(\frac{\mathbf{P}(\mathbf{i})}{\mathbf{Q}(\mathbf{i})} \right),$$

The formula is as follows for the continuous case:

$$\mathbf{D}_{\text{KL}}(\mathbf{P}||\mathbf{Q}) = \int \mathbf{P}(\mathbf{x}) \cdot \log \left(\frac{\mathbf{P}(\mathbf{x})}{\mathbf{Q}(\mathbf{x})} d\mathbf{x} \right),$$

One cannot ignore the property of KL-divergence, namely its non-negativity, i.e. $D_{KL}(P||Q) \geq 0$ for all P, Q , and $D_{KL}(P||Q) = 0$ if and only if \mathbf{P} and \mathbf{Q} are identical. However, despite these positive aspects, it is important to note that the KL divergence is not a metric in the strict sense, as it is not symmetrical, i.e. $D_{KL}(P||Q) \neq D_{KL}(Q||P)$

The Kullback-Leibler divergence is computed between the posterior distribution and the a priori distribution of the polynomial coefficients. Assuming \mathbf{c} is the vector of true values of the polynomial coefficients, \mathbf{e} is the noise vector, then the Kullback-Leibler divergence \mathbf{D}_{KL} can be calculated numerically, replacing the costly integration:

$$\mathbf{D}_{\text{KL}} = \frac{1}{2} \text{Tr}[(\Sigma_{\text{pr}}^{-1} \Sigma_{\text{post}}) - \mathbf{I} + \mu_0^T \Sigma_{\text{pr}}^{-1} \mu_0 + \log \frac{|\Sigma_{\text{pr}}|}{|\Sigma_{\text{post}}|}],$$

where μ_0 is a mean of posterior distribution, Tr is a trace of the matrix, Σ_{pr} is a covariance matrix of prior distribution, Σ_{post} is a covariance matrix of the posterior distribution, \mathbf{I} is a unit matrix, $|\cdot|$ is a matrix definition operation.

3.5 Expected informational gain

Expected Information Gain (EIG) is a fundamental concept in information theory and in our study serves as the main criterion for optimising the data collection process. EIG allows us to

evaluate the effectiveness of different data collection scenarios in terms of their ability to reduce uncertainty in the model. Our idea is to select data collection scenarios that maximise information gain, i.e. that most effectively reduce uncertainty with respect to the model.

The difference in entropy between the a prior and a posterior distributions is defined as the expected information gain:

$$EIG = H(P) - H(P|D)$$

where $H(P)$ is entropy of a prior distribution, $H(PD)$ is a entropy of a posterior distribution, and D is the data that was obtained from the experiment.

However, it is important to note that the calculation of EIG can be computationally intensive, especially for complex models and large amounts of data. Approximation methods have been developed that allow efficient estimation of EIG. In our case, we use the surrogate model method, which reduces the computational cost of calculating EIG.

Then, we compute the expected information gain of the Marzuk article, which is the average of the Kullback-Leibler divergence over the sample:

$$\mathbf{EIG} = \frac{1}{\mathbf{N}} \sum_{i=1}^{\mathbf{N}} \mathbf{D}_{\mathbf{KL}}(\mathbf{i}),$$

where \mathbf{N} is the sample size, $\mathbf{D}_{\mathbf{KL}}(\mathbf{i})$ is the Kullback-Leibler divergence value for the i -th element of the sample. To calculate the expected information gain in the context of our study, we use the following Monte Carlo approach.

First, we generate a sample from the a prior distribution of model parameters and the distribution of model errors. This allows us to model the different possible 'realities' that could occur.

Let us denote the vector of model parameters as \mathbf{c} and the vector of errors as \mathbf{e} . Then we generate \mathbf{c} and \mathbf{e} from the a prior distributions:

$$c \sim \mathcal{N}(0, \sigma_{pr}^2 I), \quad e \sim \mathcal{N}(0, \sigma_{lh}^2 I)$$

We then calculate the Kullback-Leibler divergence (KL divergence) between the a priori and posterior distributions for each generated reality. Finally, the expected information gain is calculated as the average of the KL divergence over all generated realities:

$$EIG = \frac{1}{N_{sample}} \sum_{i=1}^{N_{sample}} D_{KL}(c_i, e_i),$$

Where N_{sample} is a number of generated 'realities'.

Methods based on Kullback-Leibler divergence and expected information gain can effectively optimize the data collection process, thus improving the quality of machine learning model construction.

4 Maximising the expected information gain. Optimiser.

The next important step is to write an optimiser. The optimiser used in this study is based on the gradient lift method. It significantly helps in solving the problem of maximising the Expected Information Gain (EIG) to find the optimal values of the model parameters.

The gradient lift method is an iterative algorithm used for function optimisation. In general terms, it can be written as follows:

$$x_{n+1} = x_n + \mu \nabla F(x_n),$$

where x_n is the vector of parameters at the n -th step, μ is the learning rate, $\nabla F(x_n)$ is the gradient of the function F , which is computed at each point of x_n . The direction of the gradient indicates the direction of the largest increase of the function, so its use allows us to quickly find the local maximum of the function. A special feature of this approach is the calculation of the numerical gradient. Instead of calculating the derivative of a function analytically, an approximate method based on small changes in the input parameters is used. This makes it possible to handle functions that are difficult or impossible to differentiate analytically. The gradient of a function at a point is defined as a vector of partial derivatives of the function at that point. My implementation uses a numerical method to calculate the gradient, which approximates the value of the derivative using the central difference formula:

$$\frac{d(f(x))}{dx} \sim \frac{f(x+h) - f(x-h)}{2h}$$

where h is the small number used to approximate the derivative of the function. The gradient lift method repeats the parameter update process a given number of times ($step_{num}$), resulting in the parameter vector where the expected information gain function reaches a maximum.

The main advantage of the gradient lift method is its simplicity and versatility. This method can be used to optimise any function that can be differentiated. However, it is worth noting that the gradient lift method has a number of limitations and may not be effective in the

case of functions with many local maxima or poorly conditioned functions. To this end, we will conduct a study where we will change the key parameters. The results will be visualised in the next chapter.

In this study, the optimiser is used to find the maximum EIG, which allows us to identify the most informative experiments to improve the quality of our model.

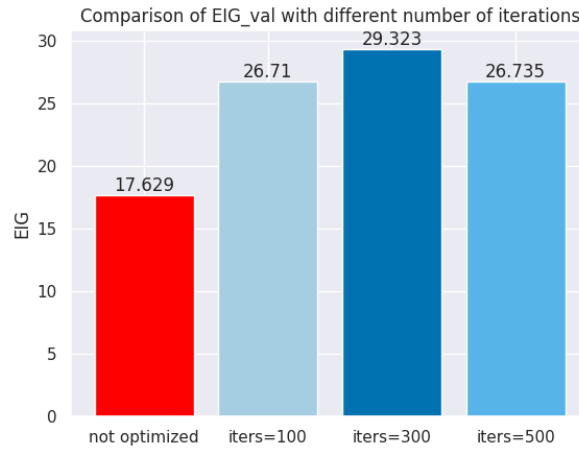
5 Experiments. Visualisation of results.

The following input parameters are defined at the start of this script:

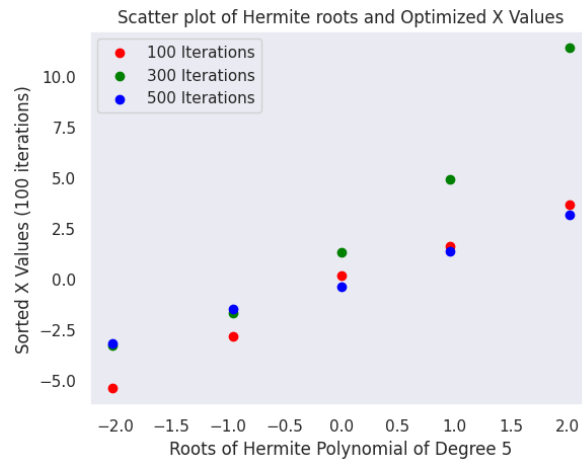
- **x** - `np.linspace`, which creates a one-dimensional array of size 21 with uniformly distributed values from -1 to 1 .
- **d** - degree of the Hermite polynomial to be used in this experiment.
- **xdata** - `np.linspace`, which creates a one-dimensional array of size 5 with uniformly distributed values between -1 and 1 .
- σ_{pr} - standard deviation of the a prior distribution. In this case it is 1 .
- σ_{lh} - standard deviation of the probability. In this case it is $1e - 2$.
- **deg_{max}** - maximum degree of the polynomial used in this experiment. In this case it is 3 .

In addition, a function f , which is a function to calculate the expected information gain and which is optimised using a gradient lift, is also given as an input. Once the points have been optimised, we can calculate the EIG value. Visualise and compare with the non-optimised data (*Figure 5.1*). Note that the optimised points will provide more "useful information". We have almost doubled the score of our main metric. The right graph shows the Gauss-Hermit quadrature X and the optimal optimized Y points according to our calculations. Note that increasing the number of iterations of the gradient lift noticeably improves the visibility of the "Hermite polynomial quadrature rule".

The following graphs (*Figure 5.2*) show the change in state (value) of each point from $xdata$ with increasing number of iterations and the change in EIG , also with increasing number of iterations. For the primary analysis, we have taken 5 evenly distributed points on the interval $[-1, 1]$.

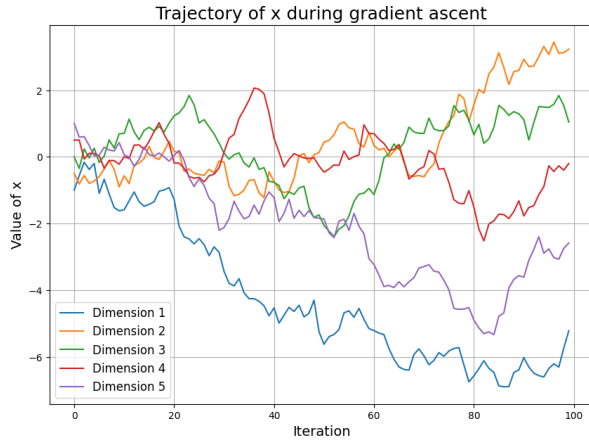


(a) EIG values

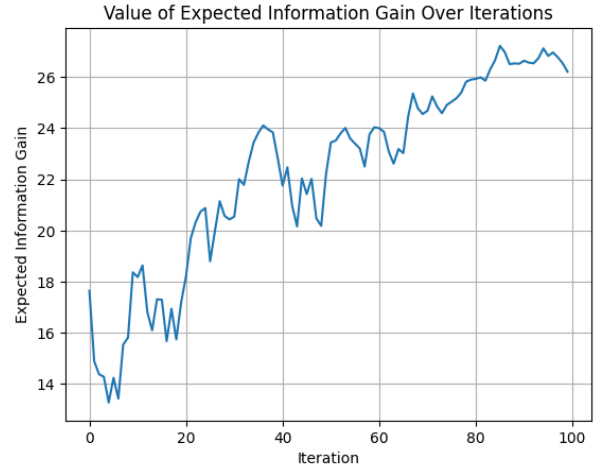


(b) Point distribution

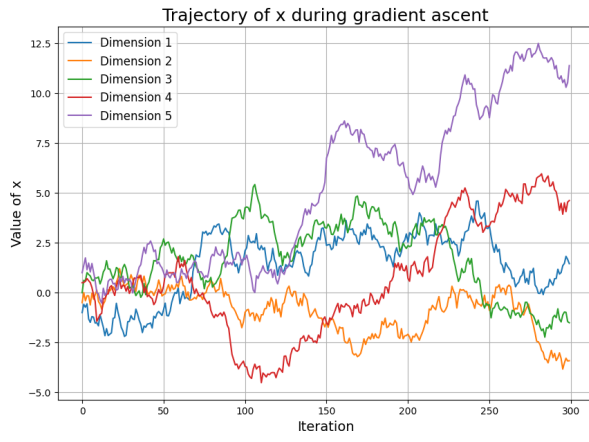
Figure 5.1:



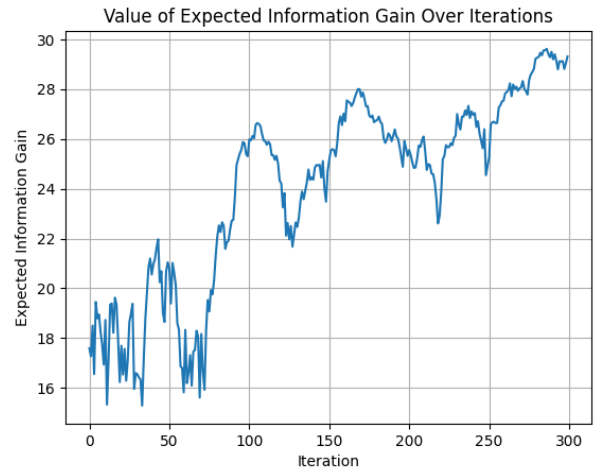
(a) iters num = 100



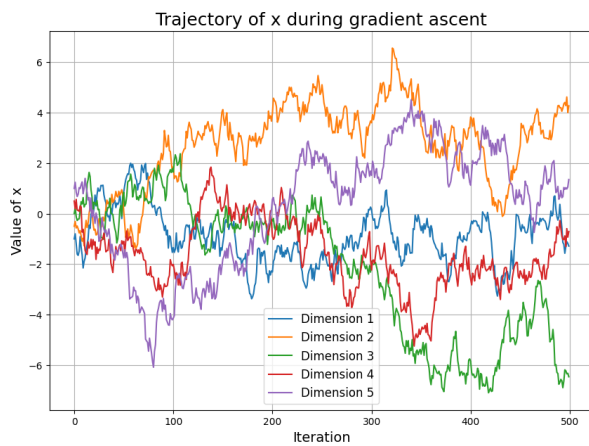
(b) EIG loss (iters num = 100)



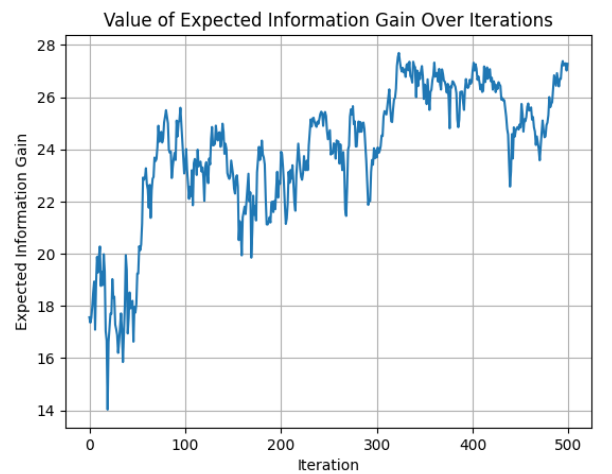
(c) iters num = 300



(d) EIG loss (iters num = 300)

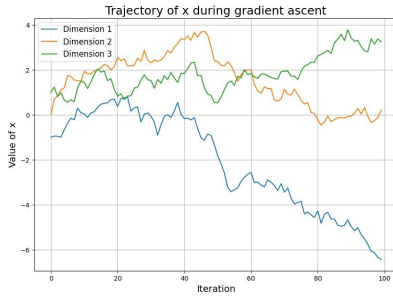


(e) iters num = 500

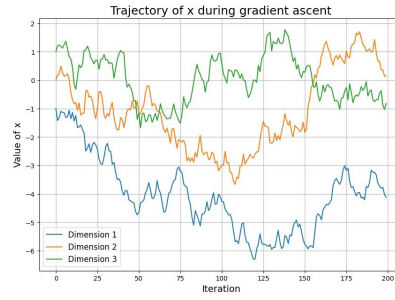


(f) EIG loss (iters num = 500)

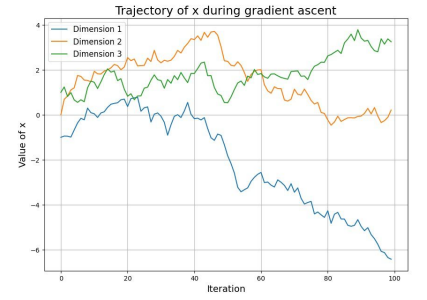
Figure 5.2:



(a) iters=100



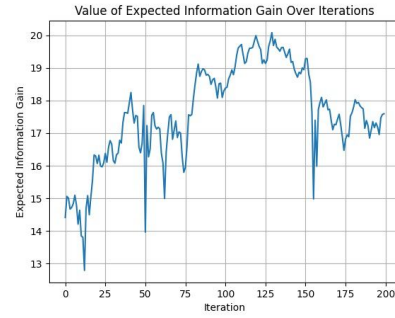
(b) iters=200



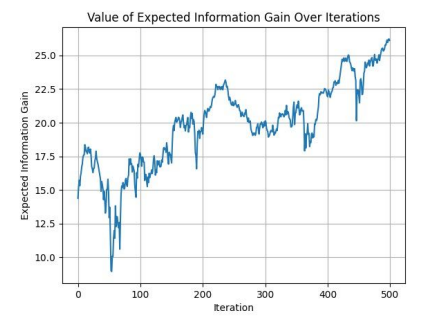
(c) iters=500



(d) iters=100



(e) iters=200



(f) iters=500

Figure 5.3:

Similarly, experiments were carried out for three points. The results showed the success of our model, which can be seen in the graph (Figure 5.3). It can be seen that in the initial iterations the quality drops off dramatically, but after 20-30 iterations it increases rapidly. This can be explained by the fact that the first steps of the numerical gradient calculation do not approximate our function well.

6 Conclusion

In conclusion, our careful work on the implementation of the algorithm has shown quite decent results. Our approach, which is based on the optimization of orthogonal Hermite polynomials using numerical optimization techniques, has shown its quality in numerical experiments.

Comparative analysis showed an increase in the key metric by almost a factor of 2. This shows the accuracy of our algorithm.

As a tool to visualize our results, informative and detailed graphs were created to demonstrate stepwise changes in the behavior of our algorithm.

But the research is not limited to this. I plan further work on this project, which will include modifying our target function. Instead of the current approach based on minimising the total square deviation, we intend to use a function based on variation in the derivative. I expect that this will improve the accuracy and robustness of our algorithm even further when considering more dynamic systems.

7 References

- [1] J. Nocedal, and S. J. Wright. Numerical Optimization. Springer, 2006.
- [2] W. Gautschi. Orthogonal Polynomials: Computation and Approximation. Oxford University Press, 2004.
- [3] K. L. Chung. A Course in Probability Theory. Academic Press, 2001.
- [4] F. Nielsen. A Gentle Introduction to Information Entropy and the Kullback-Leibler Divergence. Online publication, 2014. Available at: <http://www2.imm.dtu.dk/pubdb/p.php?3274>
- [5] A. G. Wilson, Z. Hu, R. R. Salakhutdinov, and E. P. Xing. Deep Kernel Learning. In Proceedings of the International Conference on Artificial Intelligence and Statistics. 2016.
- [6] T. P. Minka. Expectation Propagation for approximate Bayesian inference. In Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence, pages 362–369. 2001.
- J. C. Platt. Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. In Advances in Kernel Methods, pages 185–208. MIT Press, 1999.
- [8] K. P. Murphy. Machine Learning: A Probabilistic Perspective. MIT Press, 2012.
- [9] C. M. Bishop. Pattern Recognition and Machine Learning. Springer, 2006.
- [10] T. M. Cover and J. A. Thomas. Elements of Information Theory. John Wiley Sons, 2012.
- [11] R. O. Duda, P. E. Hart, and D. G. Stork. Pattern Classification. John Wiley Sons, 2001.
- [12] L. Wasserman. All of Statistics: A Concise Course in Statistical Inference. Springer, 2004.