Optimal Spectroscopic Measurement Design: Bayesian Framework for Rational Data Acquisition

Yusei Ito[1], **Yasuo Takeichi**[1], **Hideitsu Hino**[2,3], **Kanta Ono**[1]
[1]Department of Applied Physics, Osaka University
[2]The Institute of Statistical Mathematics, [3]RIKEN AIP
{yusei_ito, takeichi, ono}@ap.eng.osaka-u.ac.jp
hino@ism.ac.jp

Abstract

We have proposed an optimal experimental design method for spectroscopic measurement that can determine the appropriate number and placement of measurement points in a rational manner. Spectroscopic measurement is a fundamental experiment for material characterization. It is essential to determine the optimal experimental points automatically for autonomous experiments, however they have traditionally been decided by human expert. In this work, we have developed a method for extracting prior information from a standard spectra database and incorporating it into the Bayesian experimental design framework to determine the optimal measurement points automatically. We verified the proposed method by applying it to X-ray absorption spectrum measurements and evaluated its optimality by typical analysis. We found that only 70% of the measurement points used in previous studies were sufficient and also the determined points are consistent to the experts' intuition. The proposed method is expected to facilitate more efficient and fully automated experiments in the future.

1 Introduction

Spectroscopy is a fundamental multi-modal (image + spectra) measurement technique for material characterization that provides spectra reflecting the electronic or chemical states at each spatial point [1, 2, 3]. Although useful, it is time-consuming because it involves capturing spectral information with spatial information (2D, 3D), making it a multi-dimensional measurement.

In spectroscopy measurements, continuous spectra are often discretized for measurement, and interpretation is performed by interpolation. Therefore, determining which points and how many points to measure in the spectral dimension directly affects the measurement time and accuracy. Since the optimal conditions for these measurements vary depending on the sample and the measurement instruments, these conditions have traditionally been determined manually at each time. However, for fully automated experiments [4, 5], it is extremely important to automatically determine these conditions in a rational manner.

Ueno et al. proposed a method for adaptively determining the measurement points and the number of them [6, 7]. However, this method can only be used when spectra are measured at each spatial point individually. There are currently no established optimization methods that can be applied to cases where multiple spectra are measured simultaneously, such as 3D measurements [1].

In this paper, we propose a method for determining the optimal condition in general spectroscopic measurement case before conducting the experiment (not adaptively). Our method involves Bayesian experimental design to find the optimal experimental conditions based on prior information [8, 9, 10]. We use standard spectra database [11] to obtain the prior information and determine "how many"

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

and “which” energy points should be measured. Additionally, our approach enables determining the minimum number of measurement points by evaluating the bias, which is the mean of the expected measurement error when measurements are performed multiple times at specific energy points, and the variance of the measurement results.

2 Method

To determine the optimal measurement points, we define an evaluation function for measurement conditions called *expected loss* based on Bayesian experimental design [8, 9, 10]. The optimal conditions can be obtained by minimizing this function. First, we present the overall expression of the expected loss, followed by an explanation of each element in the formula. The relationship between a standard Bayesian experimental design method are explained in Appendix A.1.

2.1 Overall formulation

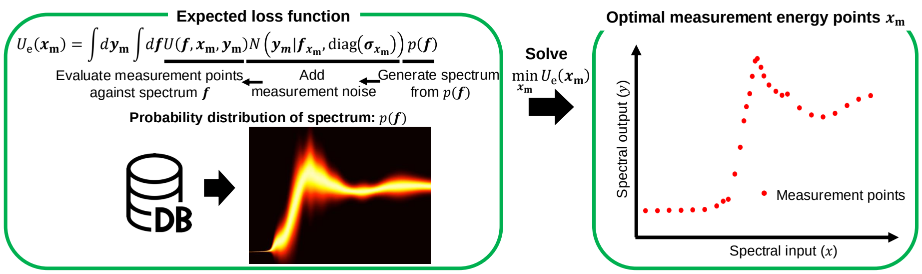

Figure 1: Construction of the evaluation function (expected loss). The expected loss can be calculated by determining the evaluation function of the measurement points $(\boldsymbol{x}_\mathrm{m}, \boldsymbol{y}_\mathrm{m})$ when measuring a certain spectrum f, and the probability of a certain spectrum being measured using a database. Then, minimizing the expected loss provides the optimal measurement point.

Spectra are generally continuous and modeled as a function $f(x)$, where x is an evaluation point and $f(x)$ is the value of the spectra curve. However, in computers and measurement devises, a spectrum $(x, f(x))$ is treated as a vector taking values at sufficiently finely discretized input points and their outputs. When the size of the discretized spectra is N, the spectrum is characterized by N dimensional vectors $\boldsymbol{x} = (x_1, \ldots, x_N)$ and $\boldsymbol{f} = (f(x_1), \ldots, f(x_N))$. Additionally, measurement noise is treated as zero-mean Gaussian with possibly varying standard deviation $\sigma(x)$ depending on the point x, and the corresponding N dimensional vector is denoted by $\boldsymbol{\sigma} \in \mathbb{R}^N$.

The overall formula for the expected loss function, which serves as the evaluation function for the measurement points $\boldsymbol{x}_\mathrm{m} \in \mathbb{R}^M, M < N$, can be represented as follows:

$$U_\mathrm{e}(\boldsymbol{x}_\mathrm{m}) = \int d\boldsymbol{y}_\mathrm{m} \int d\boldsymbol{f} U(\boldsymbol{f}, \boldsymbol{x}_\mathrm{m}, \boldsymbol{y}_\mathrm{m}) N(\boldsymbol{y}_\mathrm{m}|\boldsymbol{f}_{\boldsymbol{x}_\mathrm{m}}, \mathrm{diag}(\boldsymbol{\sigma}^2_{\boldsymbol{x}_\mathrm{m}})) p(\boldsymbol{f}), \tag{1}$$

where $\boldsymbol{f}_{\boldsymbol{x}_\mathrm{m}} \in \mathbb{R}^M$ is the subset of values in $\boldsymbol{f}$ and $\boldsymbol{\sigma}^2_{\boldsymbol{x}_\mathrm{m}} \in \mathbb{R}^M$ is the subset of values in $\boldsymbol{\sigma}$ at the measurement points $\boldsymbol{x}_\mathrm{m}$. Namely, we use $\boldsymbol{x}_\mathrm{m}$ as a indicator vector of length M which extracts subset of points from a vector of length N. The function $U(\boldsymbol{f}, \boldsymbol{x}_\mathrm{m}, \boldsymbol{y}_\mathrm{m})$ in the integrand of Eq. (1) is called the loss function, which is the evaluation function when a certain spectrum $\boldsymbol{f}$ is measured at the measurement point $\boldsymbol{x}_\mathrm{m}$ and then $\boldsymbol{y}_\mathrm{m}$ is obtained. Note that $\mathrm{diag}(\boldsymbol{\sigma}^2_{\boldsymbol{x}_\mathrm{m}}) \in \mathbb{R}^{M \times M}$ is a diagonal matrix with elements $\mathrm{diag}(\boldsymbol{\sigma}^2_{\boldsymbol{x}_\mathrm{m}})_{ii} = \sigma^2_{\boldsymbol{x}_\mathrm{m}i}, i = 1, \ldots, M$. This formula captures meaningful concept as shown in Fig. 1: It starts by generating a spectrum from a prior distribution, then obtains measurement points by adding noise, and finally evaluates the measurement points by calculating the loss considered the generated spectrum as ground truth of the measurement target. The formula can be considered as calculating the expected value of the loss function at $\boldsymbol{x}_\mathrm{m}$ with respect to the prior distribution of the spectral and the conditional distribution of the noise corrupted measurement. Note that the ground truth is sampled probabilistically because it is more natural than considering the ground truth as deterministic, since the object of measurement is unknown.

2.2 Determination of prior probability distribution

In this paper, we used the standard spectra database to determine the spectral prior probability distribution $p(\boldsymbol{f})$ [11]. To facilitate subsequent analysis, the prior distribution is assumed to be Gaussian distribution: $p(\boldsymbol{f}) = N(\boldsymbol{f}|\boldsymbol{\mu}, K)$. The mean and variance are those of the spectra contained within the spectra database, and the covariance is determined by setting the correlation of the points $x_i, x_j \in \boldsymbol{x}$ with the parameter c: $k(x_i, x_j) = \exp\left\{-(x_i - x_j)^2/c^2\right\}$. The parameter c that means correlation distances to other measurement points was determined by using the framework of type Ⅱ maximum likelihood estimation [12]. Details are provided in Appendix A.2.

2.3 Loss function and corresponding expected loss

We used the squared error between grand truth function (which is sampled from the prior distribution constructed by using a database) and mean function of posterior distribution for the loss function: $U(\boldsymbol{f}, \boldsymbol{x}_{\mathrm{m}}, \boldsymbol{y}_{\mathrm{m}}) = ||\boldsymbol{f} - \boldsymbol{\mu}_{\mathrm{post}}||^2$, where $\boldsymbol{\mu}_{\mathrm{post}}$ represents the mean value of posterior distribution $p(\boldsymbol{f}|\boldsymbol{x}_{\mathrm{m}}, \boldsymbol{y}_{\mathrm{m}})$. We then can obtain the following representation of the expected loss by substituting them into Eq. (1):

$$U_{\mathrm{e}}(\boldsymbol{x}_{\mathrm{m}}) = \sum_i [k(x_i, x_i) - \boldsymbol{k}_{\mathrm{M}}(x_i)^T C_{\mathrm{M}}^{-1} \boldsymbol{k}_{\mathrm{M}}(x_i)], \tag{2}$$

where $(\boldsymbol{k}_{\mathrm{M}}(x_i))_i = k(x_i, x_{\mathrm{m}i}), i = 1, 2, \ldots, M$. Details of calculations are given in Appendix A.3. Finally, we can determine the optimal measurement points by identifying those that minimizing the expected loss.

3 Experiment

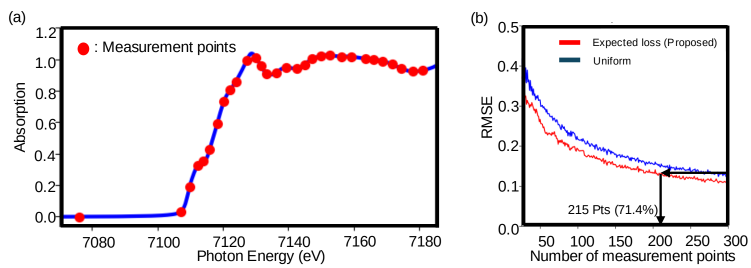

Figure 2: Result and evaluation of measurement points obtained by minimizing expected loss. (a) Optimal measurement points when the number of measurement points are 30. (b) The performance comparison between proposed method (red) and uniform step sampling (blue) by the accuracy of linear regression, a typical analysis method.

3.1 Application to XAS

We applied the proposed method to the Fe-K edge X-ray absorption spectrum (XAS) measurement, which measures the absorbance at each X-ray energy. We used the MDR XAFS Database [11] to obtain the prior distribution. We then obtained the measurement points that optimize the expected loss defined in Eq. (2) following the methodology described in the Appendix A.4. Experimental details are provided in the Appendix A.5. Figure 2(a) shows the 30 measurement points optimized by using the proposed method. Comparison between uniform step size sampling in various number of measurement points are provided in Appendix A.6. The measurement points obtained are sparse in the low-energy region, which is less informative, but are more densely sampled after the absorption edge, where they provide greater information.

3.2 Evaluation

In order to evaluate the obtained measurement points quantitatively, we evaluated the performance based on the accuracy of linear regression, which is a typical analysis method for XAS. We generated spectra by randomly selecting two spectra from the standard spectra dataset, weighting them with randomly generated coefficients, and adding noise. Then, we performed linear regression using least squares method on the generated spectrum using the standard spectra from which it was generated, and calculated the error in the coefficients. We used the average error value from 10,000 trials to assess the measurement points. Figure 2(b) shows the results of the performance evaluation at the optimal measurement point for each number of measurement points. For comparison, it also shows the results of the performance evaluation using equally spaced sampling. Appendix A.7 shows the cases of three, four, and five randomly selected spectra, respectively. We confirmed that highly efficient measurement points were realized, achieving the same level of accuracy as the conventionally used 301 points of equally interval sampling with only about 71.4 % of the 215 points.

4 Discussion

Optimal number of measurement points. In the proposed method, we considered the spectrum generated from the prior distribution as the ground truth spectrum of the measurement target and calculated the squared error when performing regression solely based on the information provided by the measurement points. The optimal number of measurement points can be determined by the following step: first, determine the optimal measurement points in conditions where the measurement points are fixed. Then, calculate the expected loss value for this configuration and compare it with the desired accuracy. If the expected accuracy is not satisfactory, we can gradually increase the number of measurement points to determine the optimal number of them.

Minimal number of measurement points. Expected loss can be decomposed into (squared) bias, which is the mean of the expected measurement error and its variance, which is the variability of the predicted results as shown in Appendix A.8. Figure 3 shows the results of calculating each for the measurement points. It can be observed that in regions with few measurement points, the contribution of bias is dominant, while in regions with many measurement points, the variance prevails. The minimal number of measurement points can be considered as the number of points that yield a small bias, allowing for accurate average predictions. Therefore, the intersection of bias and variance serves as an indicator of the minimal number of measurement points. In the application for XAS shown in the previous section, the number of measurement points at the intersecting points is 27, which is consistent with expert knowledge [13].

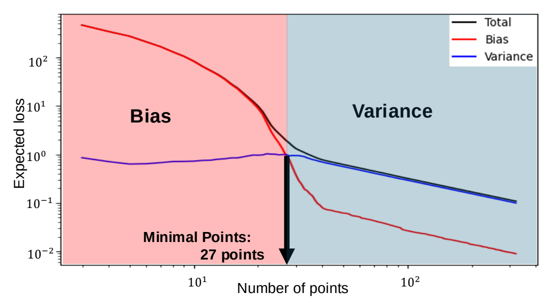

Figure 3: The result of bias-variance decomposition. The point where bias and variance intersect represents the minimal number of measurement points.

Limitations. Since our method assumes the existence of a database, it cannot be applied to measurement techniques that do not have accumulated data. This problem is expected to be solved by the expansion of the extensive simulation database that has been actively developed in recent years [14]. In addition, the mean and variance in the database are used to create the prior distribution. However, this approach may overlook small features, such as small peaks. A future task is to develop a more effective prior distribution that takes into account factors such as the rate of change.

5 Conclusion

In this work, we formulated an evaluation function for the measurement points by using Bayesian experimental design framework and demonstrated that efficient measurement points could be obtained by minimizing this function. Additionally, we determined the optimal number of measurement points under the given conditions and the minimal measurement points by discussing the evaluation function. Since this method can be broadly applied to spectroscopic measurements, we believe it can be used to determine optimal conditions in a wide range of automated experiments, contributing to the fully automated material discovery.

Acknowledgments and Disclosure of Funding

This work is partly supported by the JST-Mirai Program, Grant Number JPMJMI19G1, the MEXT Program: Data Creation and Utilization-Type Material Research and Development Project (Digital Transformation Initiative Center for Magnetic Materials) Grant Number JPMXP1122715503, the MEXT as “Developing a Research Data Ecosystem for the Promotion of Data-Driven Science”, the JSPS Grant-in-Aid for Transformative Research Areas (A) 22H05109, 23H04483, and the JST Moonshot R&D (Grant Number JPMJMS2236).

References

[1] Makoto Hirose, Nozomu Ishiguro, Kei Shimomura, Duong-Nguyen Nguyen, Hirosuke Matsui, Hieu Chi Dam, Mizuki Tada, and Yukio Takahashi. Oxygen-diffusion-driven oxidation behavior and tracking areas visualized by X-ray spectro-ptychography with unsupervised learning. *Commun. Chem.*, 2(1):1–7, 2019.

[2] Wiebke Jahr, Benjamin Schmid, Christopher Schmied, Florian O Fahrbach, and Jan Huisken. Hyperspectral light sheet microscopy. *Nat. Commun.*, 6(1):7990, 2015.

[3] Daniel C Fernandez, Rohit Bhargava, Stephen M Hewitt, and Ira W Levin. Infrared spectroscopic imaging for histopathologic recognition. *Nat. Biotechnol.*, 23(4):469–474, 2005.

[4] Eric Stach, Brian DeCost, A Gilad Kusne, Jason Hattrick-Simpers, Keith A Brown, Kristofer G Reyes, Joshua Schrier, Simon Billinge, Tonio Buonassisi, Ian Foster, Carla P Gomes, John M Gregoire, Apurva Mehta, Joseph Montoya, Elsa Olivetti, Chiwoo Park, Eli Rotenberg, Semion K Saikin, Sylvia Smullin, Valentin Stanev, and Benji Maruyama. Autonomous experimentation systems for materials development: A community perspective. *Matter*, 4(9):2702–2726, 2021.

[5] Martin Seifrid, Robert Pollice, Andrés Aguilar-Granda, Zamyla Morgan Chan, Kazuhiro Hotta, Cher Tian Ser, Jenya Vestfrid, Tony C Wu, and Alán Aspuru-Guzik. Autonomous chemical experiments: Challenges and perspectives on establishing a self-driving lab. *Acc. Chem. Res.*, 55(17):2454–2466, 2022.

[6] Tetsuro Ueno, Hideitsu Hino, Ai Hashimoto, Yasuo Takeichi, Masahiro Sawada, and Kanta Ono. Adaptive design of an X-ray magnetic circular dichroism spectroscopy experiment with gaussian process modelling. *Npj Comput. Mater.*, 4(1):1–8, 2018.

[7] Tetsuro Ueno, Hideaki Ishibashi, Hideitsu Hino, and Kanta Ono. Automated stopping criterion for spectral measurements with active learning. *Npj Comput. Mater.*, 7(1):1–9, 2021.

[8] Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Stat. Sci.*, 10(3):273–304, 1995.

[9] Ofir Harari and David M Steinberg. Optimal designs for gaussian process models |via spectral decomposition. *J. Stat. Plan. Inference*, 154:87–101, 2014.

[10] Bertrand Gauthier and Luc Pronzato. Spectral approximation of the IMSE criterion for optimal designs in kernel-based interpolation models. *SIAM/ASA J. Uncertain. Quantif.*, 2(1):805–825, 2014.

[11] Masashi Ishii, Hiroko Nagao, Kosuke Tanabe, Asahiko Matsuda, and Hideki Yoshikawa. MDR XAFS DB. Materials Data Repository. *National Institute for Materials Science*, 2021.

[12] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[13] Yasuo Takeichi, Reiko Murao, and Masao Kimura. Micromechanism of heterogeneous reduction of iron ore sinters investigated by synchrotron X-ray multimodal analysis. *ISIJ Int.*, 63(12):2017–2022, 2023.

[14] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin a. Persson. The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mater.*, 1(1):011002, 2013.

[15] Yusei Ito, Yasuo Takeichi, Hideitsu Hino, and Kanta Ono. Rational partitioning of spectral feature space for effective clustering of massive spectral image data. *Sci. Rep.*, 14(1):22549, 2024.

A Appendix

A.1 The relationship between previous Bayesian experimental design method

In this section, we derive the expected loss function from the Bayesian experimental design framework.

In accordance with Chaloner's review [8], optimal experimental design is achieved through the optimization of "expected loss". In the general case, when the experimental design is η, the values to be estimated is $\theta \in \Theta$, and the measurement results are $y \in \mathscr{Y}$, the optimal experimental design η^* is given by minimizing (or maximizing) the following expected loss (or utility) function:

$$U_{\mathrm{e}}(\eta) = \int_{\mathscr{Y}} dy \int_{\Theta} d\theta U(\theta, \eta, y) p(\theta, y|\eta), \tag{A.1}$$

where $U(\theta, \eta, y)$ represents the loss function when the ground truth parameter value is θ, the experimental design η and the measurement results are y, and $\mathscr{Y}, \Theta$ represent all possible measurement results and parameters.

We then consider the aforementioned formulation with respect to spectral measurements. In this paper, we replaced the spectrum measurement as the problem of estimating $f(x)$ under the assumption that $y = f(x) + \epsilon$, where x represents the parameter to be varied in the spectral measurements, y is the corresponding output and ϵ is the measurement noise. Therefore, the experimental design η is the set of measurement points $\boldsymbol{x}_{\mathrm{m}} \in \mathbb{R}^M$, and the parameter to be estimated, denoted as θ, is the function f that represents a spectrum. To simplify the analysis , we treat the function $f(x)$ instead as a pair of vectors $(\boldsymbol{x}, \boldsymbol{f})$ of sufficiently finely discretized input and function values. Then the optimal set of measurement points $\boldsymbol{x}^*_{\mathrm{m}}$ can be obtained by minimizing the following expected loss function $U_{\mathrm{e}}(\boldsymbol{x}_{\mathrm{m}})$:

$$U_{\mathrm{e}}(\boldsymbol{x}_{\mathrm{m}}) = \int d\boldsymbol{y}_{\mathrm{m}} \int d\boldsymbol{f} U(\boldsymbol{f}, \boldsymbol{x}_{\mathrm{m}}, \boldsymbol{y}_{\mathrm{m}}) N(\boldsymbol{y}_{\mathrm{m}}|\boldsymbol{f}_{\boldsymbol{x}_{\mathrm{m}}}, \mathrm{diag}(\sigma^2_{\boldsymbol{x}_{\mathrm{m}}})) p(\boldsymbol{f}), \tag{A.2}$$

where $\boldsymbol{f}_{\boldsymbol{x}_{\mathrm{m}}} \in \mathbb{R}^M$ is the subset of values in $\boldsymbol{f}$ and $\sigma^2_{\boldsymbol{x}_{\mathrm{m}}} \in \mathbb{R}^M$ is the subset of values in $\boldsymbol{\sigma}$ at the measurement points $\boldsymbol{x}_{\mathrm{m}}$. Here, we assumed that the measurement noise is Gaussian noise and its standard deviation be $\boldsymbol{\sigma}$, and used $p(\boldsymbol{f}, \boldsymbol{y}_{\mathrm{m}}|\boldsymbol{x}_{\mathrm{m}}) = p(\boldsymbol{y}_{\mathrm{m}}|\boldsymbol{f}_{\boldsymbol{x}_{\mathrm{m}}}, \boldsymbol{x}_{\mathrm{m}}) p(\boldsymbol{f}) = N(\boldsymbol{y}_{\mathrm{m}}|\boldsymbol{f}_{\boldsymbol{x}_{\mathrm{m}}}, \mathrm{diag}(\sigma^2_{\boldsymbol{x}_{\mathrm{m}}})) p(\boldsymbol{f})$, where $p(\boldsymbol{f})$ is the prior distribution of the spectrum to be measured and $\mathrm{diag}(\sigma^2_{\boldsymbol{x}_{\mathrm{m}}}) \in \mathbb{R}^{M \times M}$ is a diagonal matrix with elements $\mathrm{diag}(\sigma^2_{\boldsymbol{x}_{\mathrm{m}}})_{ii} = \sigma^2_{\boldsymbol{x}_{\mathrm{m}} i}, i = 1, \ldots, M$.

A.2 Type Ⅱ maximum likelihood estimation

Type Ⅱ maximum likelihood estimation method involves calculating the marginal likelihood by the measurement data and selecting parameter that maximizes it [12]. Instead, since the spectra of standard spectra database have no measurement noise, we calculated the expected marginal likelihood L:

$$\begin{aligned} L &= \sum_i \int d\boldsymbol{y} \ln(N(\boldsymbol{y}|\boldsymbol{\mu}, C)) N(\boldsymbol{y}|\boldsymbol{s_i}, \mathrm{diag}(\boldsymbol{\sigma^2})) \\ &= \sum_i \Big(-\frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln|C| - \frac{1}{2}\mathrm{Tr}(\Sigma C^{-1}) - \frac{1}{2}(\boldsymbol{s_i} - \boldsymbol{\mu})^T C^{-1} (\boldsymbol{s_i} - \boldsymbol{\mu}) \Big), \end{aligned} \tag{A.3}$$

where $\boldsymbol{s_i}$ is a spectrum of standard spectral dataset, $C = K + \mathrm{diag}(\boldsymbol{\sigma^2})$, $K_{ij} = k(x_i, x_j)$ and $\boldsymbol{\sigma^2}$ is the set of standard deviation of measurement noises at each point. We obtained optimal c by maximizing L by calculating a value for each c discretized sufficiently finely and setting it to the value that is the largest.

A.3 Details of calculations of expected utilities

Expected loss $U_{\mathrm{e}}(\boldsymbol{x}_{\mathrm{m}})$ are formulated as follows when the loss is the squared L_2 error of $\boldsymbol{f}$ and $\boldsymbol{\mu}_{\mathrm{post}}$:

$$\begin{aligned} U_{\mathrm{e}}(\boldsymbol{x}_{\mathrm{m}}) &= \int d\boldsymbol{y}_{\mathrm{m}} \int d\boldsymbol{f} ||\boldsymbol{f} - \boldsymbol{\mu}_{\mathrm{post}}||^2 N(\boldsymbol{y}_{\mathrm{m}}|\boldsymbol{f}_{\boldsymbol{x}_{\mathrm{m}}}, \sigma^2_{\boldsymbol{x}_{\mathrm{m}}}) p(\boldsymbol{f}) \\ &= \sum_i \int d\boldsymbol{y}_{\mathrm{m}} \int d\boldsymbol{f} (f_i - \mu_{\mathrm{post}i})^2 N(\boldsymbol{y}_{\mathrm{m}}|\boldsymbol{f}_{\boldsymbol{x}_{\mathrm{m}}}, \sigma^2_{\boldsymbol{x}_{\mathrm{m}}}) p(\boldsymbol{f}), \end{aligned} \tag{A.4}$$

where $\boldsymbol{\mu}_{\text{post}}$ represents the mean value of posterior distribution $p(\boldsymbol{f}|\boldsymbol{x}_{\text{m}}, \boldsymbol{y}_{\text{m}})$ and $p(\boldsymbol{f}) = N(\boldsymbol{f}|\boldsymbol{\mu}, K)$. By using the results of Gaussian process regression [12], $\boldsymbol{\mu}_{\text{post}}$ can be represented as $\mu_{\text{post}i} = \mu_i + \boldsymbol{k}_{\text{M}}^{\boldsymbol{T}}(x_i) C_{\text{M}}^{-1}(\boldsymbol{y}_{\text{m}} - \boldsymbol{\mu}_{\text{M}})$ where $(\boldsymbol{k}_{\text{M}}(x_i))_i = k(x_i, x_{\text{m}i}), i = 1, 2, \ldots, M$. By performing simple calculation, we have

$$U_{\text{e}}(\boldsymbol{x}_{\text{m}}) = \sum_i [k(x_i, x_i) - \boldsymbol{k}_{\text{M}}(x_i)^T C_{\text{M}}^{-1} \boldsymbol{k}_{\text{M}}(x_i)]. \tag{A.5}$$

A.4 Optimization method for expected loss function

It is difficult to analytically find the measurement points that minimizes expected loss function defined by Eq. (2). The optimization was performed using the greedy method. First, the initial sampling is done, and then the next points that will decrease the expected loss the most are sampled one after another, as shown in Algorithm 1. This process results in a measurement points that approximately optimizes the expected loss. In the case of XAS application, the computational time is about 15 minutes by using a laptop with Apple M2 CPU (16GB RAM).

Algorithm 1 Optimization expected loss U_{e}

Input: M : Number of measurement points
$\boldsymbol{x}_{\text{i}}$: Initial measuring points
$\boldsymbol{x}$: Grid point set
Output: $\boldsymbol{x}_{\text{m}}$: Optimal measurement points
1: $\boldsymbol{x}^* \leftarrow \boldsymbol{x}_{\text{i}}$
2: **for** t = 1 to M **do**
3: Calculate $U_{\text{e}}(\boldsymbol{x}^* \cup x)$
4: Sampling the most beneficial point
$x_{\text{t}} = \text{argmin}_{x \in \boldsymbol{x}} U_{\text{e}}(\boldsymbol{x}^* \cup x)$
5: $\boldsymbol{x}^* \leftarrow \boldsymbol{x}^* \cup x_{\text{t}}$
6: **end for**
7: **return** $\boldsymbol{x}^*$

A.5 Experimental details

We used 61 Fe-K edge XAS from the MDR XAFS Database [11], as of May 2022, for our standard spectra database. The continuous spectra were discretized in steps of 0.1 eV over the range from 7076.2 eV to 7181.2 eV and treated as vectors. While the measurement noise varies depending on the measurement method and instrument, we adopted the noise quantities reported by Ito et al. [15] as an example. To determine the parameter c in the prior distribution, we calculated the expected marginal likelihood using Eq. (A.3) by varying the parameter c from 4 eV to 5 eV in increments of 0.05 eV as shown in Fig. A.1. The optimal value of parameter c, which maximizes the expected marginal likelihood, was determined to be 4.35 eV. When using the method in Appendix A.4 to obtain the optimal measurement points, the initial points were set at 7076.2 eV and 7181.2 eV at each end.

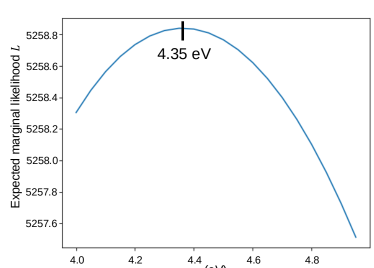

Figure A.1: Expected marginal likelihood versus parameter c. The optimal parameter c was set to the value that maximizes the expected marginal likelihood.

A.6 Comparison between uniform sampling in various number of measurement points

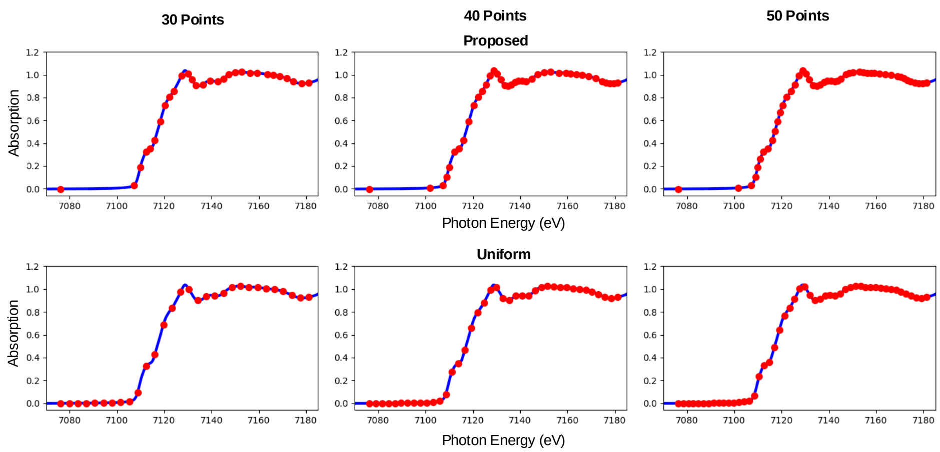

Figure A.2: Measurement points obtained by optimizing expected loss (upper row) and equal intervals (lower row). From left to right, the number of measurement points are 30 points, 40 points and 50 points. For comparison, the lower row shows the case of equal interval sampling.

A.7 Evaluation by linear regression when more than 3 components

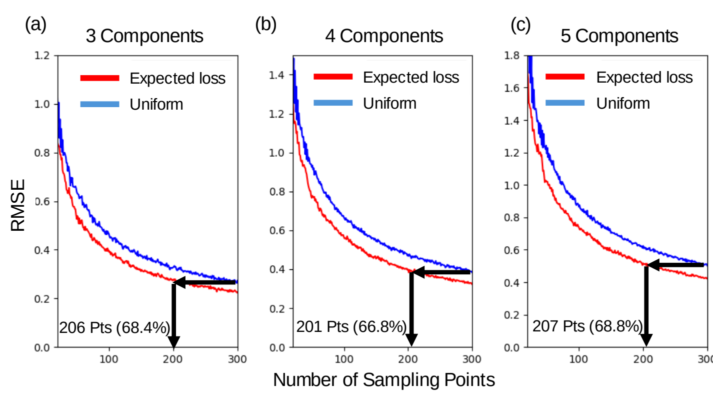

Figure A.3: Evaluation of measurement points obtained by optimizing expected loss. The performance was evaluated by the accuracy of linear regression, a typical analysis method. Panel (a) shows the linear regression with a number of Components of 3. Panel (b, c) show when the number of Components is 4, 5 respectably.

A.8 Calculation of bias variance decomposition

By performing simple calculation, we can decompose it into bias and variance terms as described below:

$$
\begin{aligned}
U_{\mathrm{e}}(\boldsymbol{x}_{\mathrm{m}}) &= \int d\boldsymbol{y}_{\mathrm{m}} \int d\boldsymbol{f} ||\boldsymbol{f} - \mathbb{E}_{\boldsymbol{y}_{\mathrm{m}}}[\boldsymbol{\mu}_{\mathrm{post}}]||^2 N(\boldsymbol{y}_{\mathrm{m}}|\boldsymbol{f}_{\boldsymbol{x}_{\mathrm{m}}}, \sigma^2_{\boldsymbol{x}_{\mathrm{m}}}) p(\boldsymbol{f}) \\
&+ \int d\boldsymbol{y}_{\mathrm{m}} \int d\boldsymbol{f} ||\boldsymbol{\mu}_{\mathrm{post}} - \mathbb{E}_{\boldsymbol{y}_{\mathrm{m}}}[\boldsymbol{\mu}_{\mathrm{post}}]||^2 N(\boldsymbol{y}_{\mathrm{m}}|\boldsymbol{f}_{\boldsymbol{x}_{\mathrm{m}}}, \sigma^2_{\boldsymbol{x}_{\mathrm{m}}}) p(\boldsymbol{f}) \\
&= \sum_i [k(x_{\mathrm{m}i}, x_{\mathrm{m}i}) + \boldsymbol{k}_{\mathrm{M}}(x_{\mathrm{m}i})^T \{C_{\mathrm{M}}^{-1} K_{\mathrm{M}} C_{\mathrm{M}}^{-1} - 2C_{\mathrm{M}}^{-1}\} \boldsymbol{k}_{\mathrm{M}}(x_{\mathrm{m}i})] \\
&+ \sum_i [\boldsymbol{k}_{\mathrm{M}}(x_{\mathrm{m}i})^T \{C_{\mathrm{M}}^{-1} \mathrm{diag}(\sigma^2_{\boldsymbol{x}_{\mathrm{m}}}) C_{\mathrm{M}}^{-1}\} \boldsymbol{k}_{\mathrm{M}}(x_{\mathrm{m}i})]
\end{aligned} \tag{A.6}
$$

The first term represents the error between the mean of the prediction results and the grand truth spectrum generated from the prior distribution $p(\boldsymbol{f})$, i.e., bias, while the second term represents the variability of the prediction results, i.e., variance.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We stated our main claims in the both the Abstract and the Introduction accurately.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of our work in the last paragraph of the Section 4.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We provide the full set of assumptions and a complete proof in the Section 2 and Appendix A.2, A.3, A.8.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provided the full source code and standard spectra dataset for reproducing our results in the supplemental material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The data and code are not currently publicly available since we are preparing an extended version of this work for a journal. We plan to release the data and code after publication of the journal article. We provided the full source code to reviewers.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We stated the used standard spectra database for understanding the results in the Section 3. We also provided the full source code including all of the parameters in the supplemental materials.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Our optimization algorithm described in the section 2 is deterministic. Although our evaluation method 3.2 is a probabilistic, we did not consider this a problem as we evaluated the linear regression accuracy at any number of measurement points and the result is almost smooth.

Guidelines:

- The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provided our computational resource using for this research in the Appendix A.4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We confirmed that our research is conducted in the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

Justification: Our proposed method is an automatic measurement optimisation method and has no impact on society. This is because the method does not lead to new measurement methods, but is combined with existing measurement methods.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our proposed method has no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We used the public database "MDR XAS Database" [11] in the Section 3. We cited it appropriately. Spectra that we used in this research are provided in the supplemental material. We confirmed that the license of this database is CC-BY-NC-SA.

Guidelines:

- The answer NA means that the paper does not use existing assets.

- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provided the implementation of our proposed method in the Supplementary Materials, including the documentation file (README.md).

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: We confirmed that this paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

Justification: We confirmed that this paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.