Bayesian Sequential Experimental Design for a Partially Linear Model with a Gaussian Process Prior

Shunsuke Horii
Waseda University
s.horii@waseda.jp

Abstract

We study the problem of sequential experimental design to estimate the parametric component of a partially linear model with a Gaussian process prior. We consider an active learning setting where an experimenter adaptively decides which data to collect to achieve their goal efficiently. The experimenter's goals may vary, such as reducing the classification error probability or improving the accuracy of estimating the parameters of the data generating process. This study aims to improve the accuracy of estimating the parametric component of a partially linear model. Under some assumptions, the parametric component of a partially linear model can be regarded as a causal parameter, the average treatment effect (ATE) or the average causal effect (ACE). We propose a Bayesian sequential experimental design algorithm for a partially linear model with a Gaussian process prior, which is also considered as a sequential experimental design tailored to the estimation of ATE or ACE. We show the effectiveness of the proposed method through numerical experiments based on synthetic and semi-synthetic data.

1 Introduction

Statistical inference and machine learning can be described as the problem of deriving some decisions from data. For example, the goal of many statistical inference problems is to accurately estimate some feature of the distribution of the data-generating mechanism, while the purpose of prediction problems is to construct a predictor with good prediction performance. Many statistical inference methods and machine learning algorithms first gather a large amount of data randomly sampled from the underlying data-generating distribution. Then they output an estimation of an estimand or a predictor. However, in some cases, we have limited resources for collecting such data. In such cases, it is valuable to determine how we can use these resources as much as possible. In statistics, sequential experimental design studies ways to sample adaptively to improve estimation efficiency [1, 2, 3]. In machine learning, the study of sequentially obtaining labeled data from unlabeled data is called active learning [4, 5, 6]. The study on a sequential design strategy for global optimization of black-box functions is also called Bayesian optimization [7]. In a sequential experimental design, the sample point to be obtained is determined adaptively so that the estimation accuracy of the estimand is as good as possible. A typical estimand is the regression coefficients of a linear regression model [1, 3]. In this study, we consider a sequential experimental design in which estimand is a causal parameter of the data-generating distribution. The problem of causal inference, which is the estimation of the effect of changing the value of a treatment variable on the outcome variable, is one of the most important problems in data science. Although there are other studies of sequential experimental design for causal inference, these studies aim at estimating the structure of the causal graph [8, 9]. On the other hand, our study aims at estimating a causal parameter known as the average treatment effect (ATE) or the average causal effect (ACE).

2022 NeurIPS Workshop on Gaussian Processes, Spatiotemporal Modeling, and Decision-making Systems.

In this study, the relationship between variables is modeled by a partially linear model. The study of a partially linear model dates back to Engle et al [10]. The partially linear model combines the linear and nonparametric regression models. Under some assumptions, the linear component of the model can be regarded as a causal parameter, and Double/Debiased machine learning (DML) [11], which has been actively studied in recent years, can be seen as one of the estimation algorithms of it. In this study, we consider the partially linear model proposed in [12], which assumes a Gaussian process prior distribution for the nonlinear component. We define an information gain based on the posterior distribution of the causal parameter. We can efficiently estimate the causal parameter by adaptively obtaining data with the largest information gain.

2 Partially Linear Model with Gaussian Process Prior

This study investigates the causal effect of the binary treatment variable $T \in \{0, 1\}$ on the outcome variable $Y \in \mathbb{R}$. In addition to T, Y, we assume that there are some other covariates $\boldsymbol{X} \in \mathbb{R}^d$. We model the relation between $T, Y, \boldsymbol{X}$ by the following model.

$$Y_i = \theta T_i + f(\boldsymbol{X}_i) + \varepsilon_i, \qquad \mathrm{E}\left[\varepsilon_i | T_i, \boldsymbol{X}_i\right] = 0, \tag{1}$$

where $f(\cdot)$ is an unknown nonlinear function from $\mathbb{R}^d$ to $\mathbb{R}$. The model (1) is an example of partially linear models since the right-hand side is linear in T_i and nonlinear in $\boldsymbol{X}_i$. Furthermore, we assume that the nonlinear function f follows a Gaussian process prior $\mathcal{GP}(\boldsymbol{0}, C(\cdot, \cdot; \boldsymbol{\omega}))$ defined by the covariance function $C(\cdot, \cdot, \boldsymbol{\omega})$ with parameter $\boldsymbol{\omega}$, and the error term ε_i is i.i.d. with a Gaussian distribution $\mathcal{N}(0, s_\varepsilon^{-1})$. This is an instance of the model proposed in [12].

We are interested in the estimation of θ since it can be considered as a causal parameter $\mathrm{E}[Y^{(1)} - Y^{(0)}]$ or $\mathrm{E}_{\mathrm{do}(T=1)}[Y] - \mathrm{E}_{\mathrm{do}(T=0)}[Y]$ under some assumptions. See the appendix A for the relation between θ and causal parameters. Among various possible methods for estimating θ, we consider an estimation method based on the posterior distribution of θ. For this purpose, we assume that θ is a random variable with a Gaussian prior $\mathcal{N}(\mu_{\theta,0}, s_{\theta,0}^{-1})$. Given the observations $\boldsymbol{t} = [t_1, \ldots, t_n]^\top$, $\boldsymbol{y} = [y_1, \ldots, y_n]^\top$, and $\boldsymbol{X} = [\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_n^\top]^\top$, the posterior distribution of θ is given by

$$p(\theta | \boldsymbol{t}, \boldsymbol{y}, \boldsymbol{X}) = \mathcal{N}(\mu_{\theta,n}, s_{\theta,n}^{-1}), \tag{2}$$

$$\mu_{\theta,n} = s_{\theta,n}(s_{\theta,0}\mu_{\theta,0} + \boldsymbol{t}^\top (s_\varepsilon^{-1}\boldsymbol{I}_n + \boldsymbol{\Phi})^{-1}\boldsymbol{y}), \tag{3}$$

$$s_{\theta,n} = s_{\theta,0} + \boldsymbol{t}^\top (s_\varepsilon^{-1}\boldsymbol{I}_n + \boldsymbol{\Phi})^{-1}\boldsymbol{t}, \tag{4}$$

where $\boldsymbol{I}_n$ is the n-dimensional identity matrix and $\boldsymbol{\Phi}$ is the covariance matrix of $[f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n)]^\top$ given by

$$\boldsymbol{\Phi} = \begin{bmatrix} C(\boldsymbol{x}_1, \boldsymbol{x}_1; \boldsymbol{\omega}) & C(\boldsymbol{x}_1, \boldsymbol{x}_2; \boldsymbol{\omega}) & \cdots & C(\boldsymbol{x}_1, \boldsymbol{x}_n; \boldsymbol{\omega}) \\ C(\boldsymbol{x}_2, \boldsymbol{x}_1; \boldsymbol{\omega}) & C(\boldsymbol{x}_2, \boldsymbol{x}_2; \boldsymbol{\omega}) & \cdots & C(\boldsymbol{x}_2, \boldsymbol{x}_n; \boldsymbol{\omega}) \\ \vdots & \vdots & \ddots & \vdots \\ C(\boldsymbol{x}_n, \boldsymbol{x}_1; \boldsymbol{\omega}) & C(\boldsymbol{x}_n, \boldsymbol{x}_2; \boldsymbol{\omega}) & \cdots & C(\boldsymbol{x}_n, \boldsymbol{x}_n; \boldsymbol{\omega}) \end{bmatrix}. \tag{5}$$

The posterior mean $\mu_{\theta,n}$ is the Bayes optimal estimator in the sense that it minimizes the Bayes risk function for the squared error loss.

The model (1) has the hyperparameters $\mu_\theta, s_\theta, \boldsymbol{\omega}, s_\varepsilon$. One way to handle them is to assume a prior distribution for them. In that case, however, the analytic form of the posterior distribution of θ is no longer available, so we have to rely on numerical computation methods such as Markov Chain Monte Carlo (MCMC) method. These calculation methods are not the subject of our study, we refer the reader to [12] for details.

3 Sequential Experimental Design based on Information Gain

In this section, we first describe the problem of the sequential experimental design. Let $\mathcal{X}_{\mathrm{pool}}$ be the set of d-dimensional candidate covariate vectors to sample. Our problem of sequential experimental design is to determine $\boldsymbol{x}_{n+1} \in \mathcal{X}_{\mathrm{pool}}$ and $t_{n+1} \in \{0, 1\}$ given $\boldsymbol{t}, \boldsymbol{y}$, and $\boldsymbol{X}$. Inspired by the information-theoretic methods for active data collection [1, 3, 6, 7], we define the information gain for our problem as follows.

$$g(\boldsymbol{x}, t) = \mathrm{H}\left[p(\theta | \boldsymbol{t}, \boldsymbol{y}, \boldsymbol{X})\right] - \mathrm{E}_{p(y|\boldsymbol{t}, \boldsymbol{y}, \boldsymbol{X}, t, \boldsymbol{x})}\left[\mathrm{H}\left[p(\theta | \boldsymbol{t}, \boldsymbol{y}, \boldsymbol{X}, t, y, \boldsymbol{x})\right]\right], \tag{6}$$

where $\mathrm{H}[p(\theta)] = -\int p(\theta) \ln p(\theta) \mathrm{d}\theta$ represents the differential entropy. This means that we consider data that reduces the entropy of the posterior distribution of θ the most informative. The optimal design is to select $(\boldsymbol{x}_{n+1}, t_{n+1})$ that maximizes the information gain, that is,

$$(\boldsymbol{x}_{n+1}, t_{n+1}) = \underset{\boldsymbol{x} \in \mathcal{X}_{\mathrm{pool}}, t \in \{0,1\}}{\arg\max} \; g(\boldsymbol{x}, t). \tag{7}$$

Since the posterior distribution is Gaussian, its entropy is given by

$$\mathrm{H}[p(\theta | \boldsymbol{t}, \boldsymbol{y}, \boldsymbol{X})] = \frac{1}{2}(1 + \ln 2\pi s_{\theta,n}^{-1}). \tag{8}$$

Noting that the right-hand side of (8) does not depend on $\boldsymbol{y}$ and the first term in (6) does not depend on $(\boldsymbol{x}, t)$, (7) can be written as

$$(\boldsymbol{x}_{n+1}, t_{n+1}) = \underset{\boldsymbol{x} \in \mathcal{X}_{\mathrm{pool}}, t \in \{0,1\}}{\arg\min} \; \frac{1}{2} \left(1 + \ln 2\pi (s_{\theta,n+1}(\boldsymbol{x}, t))^{-1}\right), \tag{9}$$

where

$$s_{\theta,n+1}(\boldsymbol{x}, t) = s_{\theta,0} + [\boldsymbol{t}^\top, t](s_\varepsilon^{-1} \boldsymbol{I}_{n+1} + \boldsymbol{\Phi}'(\boldsymbol{x}))^{-1}[\boldsymbol{t}^\top, t]^\top, \tag{10}$$

$$\boldsymbol{\Phi}'(\boldsymbol{x}) = \begin{bmatrix} C(\boldsymbol{x}_1, \boldsymbol{x}_1; \boldsymbol{\omega}) & C(\boldsymbol{x}_1, \boldsymbol{x}_2; \boldsymbol{\omega}) & \cdots & C(\boldsymbol{x}_1, \boldsymbol{x}_n; \boldsymbol{\omega}) & C(\boldsymbol{x}_1, \boldsymbol{x}; \boldsymbol{\omega}) \\ C(\boldsymbol{x}_2, \boldsymbol{x}_1; \boldsymbol{\omega}) & C(\boldsymbol{x}_2, \boldsymbol{x}_2; \boldsymbol{\omega}) & \cdots & C(\boldsymbol{x}_2, \boldsymbol{x}_n; \boldsymbol{\omega}) & C(\boldsymbol{x}_2, \boldsymbol{x}; \boldsymbol{\omega}) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ C(\boldsymbol{x}_n, \boldsymbol{x}_1; \boldsymbol{\omega}) & C(\boldsymbol{x}_n, \boldsymbol{x}_2; \boldsymbol{\omega}) & \cdots & C(\boldsymbol{x}_n, \boldsymbol{x}_n; \boldsymbol{\omega}) & C(\boldsymbol{x}_n, \boldsymbol{x}; \boldsymbol{\omega}) \\ C(\boldsymbol{x}, \boldsymbol{x}_1; \boldsymbol{\omega}) & C(\boldsymbol{x}, \boldsymbol{x}_2; \boldsymbol{\omega}) & \cdots & C(\boldsymbol{x}, \boldsymbol{x}_n; \boldsymbol{\omega}) & C(\boldsymbol{x}, \boldsymbol{x}; \boldsymbol{\omega}) \end{bmatrix}. \tag{11}$$

The proposed optimal design (7) is similar to the methods proposed in [1, 3, 6, 7]. However, the objective of those studies is to estimate the function that represents the relationship between inputs and outputs, or the whole data-generating distribution, whereas the objective of our study is to estimate the causal parameter θ. In our problem, the nonlinear function f is a nuisance parameter, that is, it is noise for estimating θ. In [13], it is pointed out that plug-in estimation, in which the data-generating distribution is estimated and then the causal parameter is estimated, may not have the desired properties. They call targeted learning a learning method that is tailored to an estimand such as a causal parameter, so we can call our method a targeted sequential experimental design.

4 Experiments

In this section, we illustrate the performance of the optimal design on synthetic data and semi-synthetic data.

4.1 Experiments on Synthetic Data

We illustrate the behavior of the optimal design using synthetic data. In this experiment, we generate θ according to the prior distribution $\mathcal{N}(\mu_{\theta,0}, s_{\theta,0}^{-1})$, where the hyperparameters are set to $\mu_{\theta,0} = 0, s_{\theta,0} = 1.0$. We assume that the covariance function $C(\cdot, \cdot, \boldsymbol{\omega})$ of the Guassian process prior for the nonlinear function f is the RBF kernel

$$C(\boldsymbol{x}, \boldsymbol{x}'; \omega) = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{2\omega}\right), \tag{12}$$

with the hyperparameter value $\omega = 1.0$. We construct the set of candidate covariate vectors $\mathcal{X}_{\mathrm{pool}}$ by generating each element from $\mathcal{N}(0, 1.0^2)$. The dimension of a covariate vector is $d = 10$ and $\mathcal{X}_{\mathrm{pool}}$ contains 400 sample. We compare the optimal design, which determines the drawing sample according to (7), with the random design, which randomly draws $(\boldsymbol{x}_{n+1}, t_{n+1})$ from $\mathcal{X}_{\mathrm{pool}} \times \{0, 1\}$. The outcome variable Y_i is generated according to the model (1), where the precision parameter of the noise $s_\varepsilon = 1.0$. In both designs, the first 10 samples were randomly selected from $\mathcal{X}_{\mathrm{pool}} \times \{0, 1\}$, and after that, the samples were selected according to each rule. The estimator outputs the posterior mean $\mu_{\theta,n}$, and we examined the absolute error $|\mu_{\theta,n} - \theta|$. Figure 1 shows the absolute error curves of the Bayes estimator for the random design case and the optimal design case as the functions of the sample size. The depicted absolute error is the average of 1000 experiments. We can see that the estimation error of the Bayes estimator in the optimal design case is much smaller than in the random design case, especially when the sample size is small.

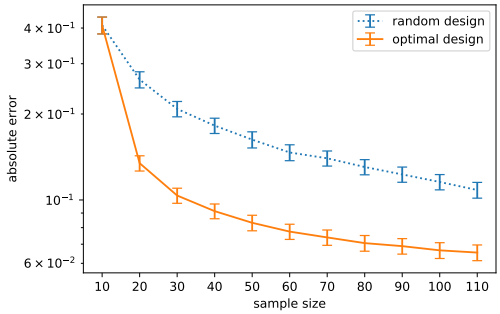

Figure 1: Experimental results on synthetic data. Average$\pm 95\%$ confidence intervals are depicted.

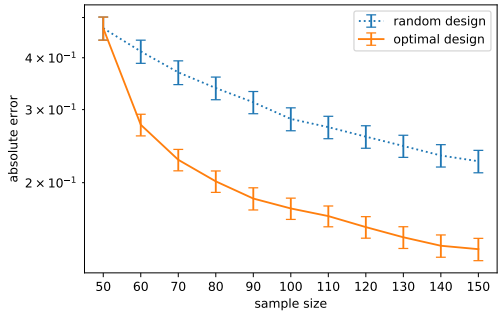

Figure 2: Experimental results on semi-synthetic data. Average$\pm 95\%$ confidence intervals are depicted.

4.2 Experiments on Semi-synthetic Data

In real-world applications, ground truth causal effects are rarely available. Thus, we evaluate the optimal design through an experiment on semi-synthetic data. For the evaluation, we use pre-established benchmarking data for causal inference. The dataset is constructed from the Infant Health and Development Program (IHDP) [14]. Each observation consists of 25 covariates, an indicator variable indicating whether the infant received special care, and an outcome variable representing the final cognitive test score. The dataset has 747 observations. Since the dataset also contains counterfactual data, we can calculate the treatment effect for each observation. We assume that the average of them is the true value of the ATE, and it is 4.021. We conducted almost the same experiment as for synthetic data, with a few minor modifications to the experimental conditions. The hyperparameters are set to $\mu_{\theta,0} = 0, s_{\theta,0} = 1.0$. The dot product kernel

$$C(\boldsymbol{x}, \boldsymbol{x}', \omega) = \omega + \boldsymbol{x}^\top \boldsymbol{x}', \tag{13}$$

with the hyperparameter value $\omega = 1.0$ is used for the Gaussian process prior. The Bayes optimal estimator requires the value of s_ε. As a rough estimator of s_ε, we used the inverse of the sample variance of $\boldsymbol{y}$. As in the previous experiment, We compare the optimal design and random design. The set of candidate covariate vectors $\mathcal{X}_{\text{pool}}$ is constructed by randomly choosing 400 observations from the total of 747 observations. In both designs, the first 50 samples were randomly selected from $\mathcal{X}_{\text{pool}}$, and after that, the samples were chosen according to each rule. We examined the absolute error of the posterior mean $|\mu_{\theta,n} - \text{ATE}|$. Figure 2 shows the absolute error curves of the Bayes estimator for the random case and the optimal design case as the functions of the sample size. The depicted absolute error is the average of 1000 experiments. As in the results of synthetic data, we can see that the estimation error of the Bayes estimator in the optimal design case is much smaller than in the random design case, especially when the sample size is small.

5 Conclusion

In this paper, we proposed a Bayesian sequential experimental design algorithm tailored to estimate the parametric component of a partially linear model with a Gaussian process prior. Under some assumptions, the parametric component of a partially linear model can be regarded as a causal parameter, the average treatment effect (ACE) or the average causal effect (ACE). Therefore, we can say that the proposed method is a sequential experimental design tailored to the estimation of a causal parameter. In general, causal parameters can be a function of the data generating distribution. ATE or ACE is one of the causal parameters, and there are various other causal parameters, such as conditional average treatment effect (CATE) or effect modification. Our study assumed a partially linear model for the data generating mechanism. However, various modeling approaches are possible for the data generating mechanism. We believe that our proposed method can be applied to other combinations of causal parameters and data-generating mechanisms, which is a subject for future work.

References

[1] Matthias Seeger, Florian Steinke, and Koji Tsuda. Bayesian inference and optimal design in the sparse linear model. In *Artificial Intelligence and Statistics*, pages 444–451. PMLR, 2007.

[2] Elizabeth G Ryan, Christopher C Drovandi, James M McGree, and Anthony N Pettitt. A review of modern computational algorithms for bayesian optimal design. *International Statistical Review*, 84(1):128–154, 2016.

[3] Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, pages 273–304, 1995.

[4] Burr Settles. Active learning literature survey. 2009.

[5] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *ACM computing surveys (CSUR)*, 54(9):1–40, 2021.

[6] David JC MacKay. Information-based objective functions for active data selection. *Neural computation*, 4(4):590–604, 1992.

[7] José Miguel Hernández-Lobato, Matthew W Hoffman, and Zoubin Ghahramani. Predictive entropy search for efficient global optimization of black-box functions. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[8] Yang-Bo He and Zhi Geng. Active learning of causal networks with intervention experiments and optimal designs. *Journal of Machine Learning Research*, 9(Nov):2523–2547, 2008.

[9] Alain Hauser and Peter Bühlmann. Two optimal strategies for active learning of causal models from interventional data. *International Journal of Approximate Reasoning*, 55(4):926–939, 2014. Special issue on the sixth European Workshop on Probabilistic Graphical Models.

[10] Robert F Engle, Clive WJ Granger, John Rice, and Andrew Weiss. Semiparametric estimates of the relation between weather and electricity sales. *Journal of the American statistical Association*, 81(394):310–320, 1986.

[11] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018.

[12] Taeryon Choi and Yoonsung Woo. A partially linear model using a gaussian process prior. *Communications in Statistics-Simulation and Computation*, 44(7):1770–1786, 2015.

[13] Mark J Van der Laan, Sherri Rose, et al. *Targeted learning: causal inference for observational and experimental data*, volume 10. Springer, 2011.

[14] Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

[15] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[16] Judea Pearl. Causality: Models, reasoning, and inference. 2000.

A Partially Linear Model and Causal Inference

In this section, we show that the linear component of a partially linear model can be regarded as a causal parameter.

In the Neyman-Rubin causal model, the existence of potential outcomes $Y^{(0)}, Y^{(1)}$ is assumed [15]. It is also assumed that $Y = TY^{(1)} + (1-T)Y^{(0)}$. The average treatment effect (ATE) is defined as

$$\mathrm{ATE} = \mathrm{E}[Y^{(1)}] - \mathrm{E}[Y^{(0)}]. \tag{14}$$

When the condition $Y^{(0)}, Y^{(1)} \perp\!\!\!\perp T | \boldsymbol{X}$ holds, the treatment assignment is called strongly ignorable. When the strongly ignorable condition holds,

$$\begin{aligned}
\mathrm{E}[Y^{(1)}] - \mathrm{E}[Y^{(0)}] &= \mathrm{E}_{\boldsymbol{X}}[\mathrm{E}[Y^{(1)}|\boldsymbol{X}]] - \mathrm{E}_{\boldsymbol{X}}[\mathrm{E}[Y^{(0)}|\boldsymbol{X}]] && (15)\\
&= \mathrm{E}_{\boldsymbol{X}}[\mathrm{E}[Y^{(1)}|\boldsymbol{X}, T=1]] - \mathrm{E}_{\boldsymbol{X}}[\mathrm{E}[Y^{(0)}|\boldsymbol{X}, T=0]] && (16)\\
&= \mathrm{E}_{\boldsymbol{X}}[\mathrm{E}[Y|\boldsymbol{X}, T=1]] - \mathrm{E}_{\boldsymbol{X}}[\mathrm{E}[Y|\boldsymbol{X}, T=0]]. && (17)
\end{aligned}$$

In our model (1), $\mathrm{E}[Y|\boldsymbol{X}, T] = \theta T + f(\boldsymbol{X})$. Thus,

$$\mathrm{E}_{\boldsymbol{X}}[\mathrm{E}[Y|\boldsymbol{X}, T=1]] - \mathrm{E}_{\boldsymbol{X}}[\mathrm{E}[Y|\boldsymbol{X}, T=0]] = \mathrm{E}_{\boldsymbol{X}}[\theta + f(\boldsymbol{X})] - \mathrm{E}_{\boldsymbol{X}}[f(\boldsymbol{X})] \tag{18}$$

$$= \theta. \tag{19}$$

Next, we show that θ can also be regarded as a causal parameter in Pearl's structural causal inference framework [16]. We assume that the data is generated according to the following structural equation model.

$$\boldsymbol{X} = \boldsymbol{\varepsilon}_{\boldsymbol{X}}, \tag{20}$$

$$Y = \theta T + f(\boldsymbol{X}) + \varepsilon_Y, \qquad \mathrm{E}[\varepsilon_Y] = 0, \tag{21}$$

$$T = \psi(\boldsymbol{X}, \varepsilon_T), \tag{22}$$

where ψ is some nonlinear function, and $\boldsymbol{\varepsilon}_{\boldsymbol{X}}, \varepsilon_Y$ and ε_T are independent. These structural equations specify probability distributions $p(y|t, \boldsymbol{x}), p(t|\boldsymbol{x})$ and $p(\boldsymbol{x})$. In this case, the distribution of $\boldsymbol{X}$ and Y after an intervention $T = t$ is

$$p_{\mathrm{do}(T=t)}(\boldsymbol{x}, y) = p(y|t, \boldsymbol{x})p(\boldsymbol{x}). \tag{23}$$

Marginalizing this distribution with respect to $\boldsymbol{X}$, we obtain

$$p_{\mathrm{do}(T=t)}(y) = \int p(y|t, \boldsymbol{x})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \tag{24}$$

The average causal effect is defined as

$$\mathrm{ACE} = \mathrm{E}_{\mathrm{do}(T=1)}[Y] - \mathrm{E}_{\mathrm{do}(T=0)}[Y], \tag{25}$$

where

$$\mathrm{E}_{\mathrm{do}(T=t)}[Y] = \int y \cdot p_{\mathrm{do}(T=t)}(y)\mathrm{d}y. \tag{26}$$

If the order of integration can be exchanged, $\mathrm{E}_{\mathrm{do}(T=t)}[Y] = \theta t + \mathrm{E}[f(\boldsymbol{X})]$, and then $\mathrm{ACE} = \theta$.

B Derivation of the Posterior Distribution of θ

Let $\boldsymbol{z} = [z_1, \ldots, z_n]^\top = [f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n)]^\top$. From the assumption, $\boldsymbol{z}$ follows a Gaussian distribution $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Phi})$. The exponent of $p(\boldsymbol{y}, \boldsymbol{z}|\theta, \boldsymbol{t}, \boldsymbol{X}) = p(\boldsymbol{y}|\boldsymbol{z}, \theta, \boldsymbol{t}, \boldsymbol{X})p(\boldsymbol{z})$ can be written as

$$-\frac{1}{2}\left(\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{z} \end{bmatrix} - \boldsymbol{A}^{-1}\boldsymbol{b}\right)^\top \boldsymbol{A}\left(\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{z} \end{bmatrix} - \boldsymbol{A}^{-1}\boldsymbol{b}\right) + \text{const.}, \tag{27}$$

where

$$\boldsymbol{A} = \begin{bmatrix} s_\varepsilon \boldsymbol{I}_n & -s_\varepsilon \boldsymbol{I}_n \\ -s_\varepsilon \boldsymbol{I}_n & s_\varepsilon \boldsymbol{I}_n + \boldsymbol{\Phi}^{-1} \end{bmatrix}, \tag{28}$$

$$\boldsymbol{b} = \begin{bmatrix} s_\varepsilon \theta \boldsymbol{t} \\ -s_\varepsilon \theta \boldsymbol{t} \end{bmatrix}. \tag{29}$$

The inverse matrix of $\boldsymbol{A}$ is given by

$$\boldsymbol{A}^{-1} = \begin{bmatrix} s_\varepsilon^{-1}\boldsymbol{I}_n + \boldsymbol{\Phi} & \boldsymbol{\Phi} \\ \boldsymbol{\Phi} & \boldsymbol{\Phi} \end{bmatrix}. \tag{30}$$

Thus, $p(\boldsymbol{y}|\theta, \boldsymbol{t}, \boldsymbol{X})$ is a Gaussian with mean $\theta\boldsymbol{t}$ and covariance matrix $s_\varepsilon^{-1}\boldsymbol{I}_n + \boldsymbol{\Phi}$. The exponent of $p(\boldsymbol{\theta}|\boldsymbol{t}, \boldsymbol{y}, \boldsymbol{X}) \propto p(\boldsymbol{y}|\theta, \boldsymbol{t}, \boldsymbol{X})p(\theta)$ can be written as

$$-\frac{1}{2}\left(\left(s_\theta + \boldsymbol{t}^T(s_\varepsilon^{-1}\boldsymbol{I}_n + \boldsymbol{\Phi})^{-1}\boldsymbol{t}\right)\theta^2 - 2\left(s_\theta\mu_\theta + \boldsymbol{t}^T(s_\varepsilon^{-1}\boldsymbol{I}_n + \boldsymbol{\Phi})^{-1}\boldsymbol{y}\right)\theta\right) + \text{const.} \tag{31}$$

Thus, $p(\boldsymbol{\theta}|\boldsymbol{t}, \boldsymbol{y}, \boldsymbol{X})$ is a Gaussian with mean $\mu_{\theta,n}$ and variance $s_{\theta,n}^{-1}$ given in (3) and (4), respectively.