Technische Universität München

Department of Mathematics

# Active Bayesian Causal Discovery for Gaussian Process Networks

Master's Thesis

Stefan Kienle

| | |
|---|---|
| Supervisor: | Prof. Dr. Mathias Drton |
| Advisor: | David Strieder |
| Submission Date: | 13.09.2022 |

I hereby declare that this thesis is my own work and that no other sources have been used except those clearly indicated and referenced.

Munich, 13.09.2022

# Acknowledgements

I would like to thank Professor Mathias Drton for selecting a very interesting research topic that revealed many interrelations between different lines of research to me and for his interesting ideas and suggestions. Further, I would like to thank David Strieder for a very good supervision providing great feedback, fruitful discussions and well thought out suggestions.

Finally, I am deeply grateful for the support from my parents, Hannah Lange and all my friends during the time of my studies.

# German Abstract

Aktuelle Ergebnisse der Forschung zur Kausalinferenz haben gezeigt, dass die kausale Struktur eines Systems von Zufallsvariablen oft aus (reinen) Beobachtungsdaten bestimmt werden kann. Prinzipiell kann die Kausalstruktur auch aus wenigen Interventionsbeaobachtungen abgeleitet werden, falls es möglich ist diese zu erheben. In Situationen, in denen Beobachtungen sehr teuer sind, wir jedoch prinzipiell Interventionen an den Systemvariablen vornehmen können, stellt sich die Frage, was die informativste Intervention zur Bestimmung der Kausalstruktur ist.

In dieser Arbeit betrachten wir einen aktiven Bayes'schen Ansatz für das Lernen von Kausalstrukturen. Er wurzelt in der Schnittmenge von statistischer Forschung zur Kausalinferenz und Forschung zum aktiven Lernen im Bereich der Künstlichen Intelligenz. Der Ansatz wurde von [49, von Kügelgen et al.] vorgeschlagen und in dieser Arbeit präsentieren wir zum ersten Mal numerische Ergebnisse der vollständigen Implementierung. Dazu modellieren wir ein System von Zufallsvariablen mit Hilfe von strukturellen Kausalmodellen (SCM), die einen gerichteten azyklischen Graphen (DAGs) implizieren, und schätzen die darin enthaltenen funktionalen Beziehungen mittels Gauss Prozess (GP) Regression. Wir nehmen an, dass perfekte Interventionen in die Systemvariablen möglich sind. In einem sequenziellen Verfahren wählen wir in jedem Schritt das informativste Interventionsexperiment (unter Verwendung der Bayes'schen Versuchsplanung) und erhalten einen Interventionsdatenpunkt aus der Durchführung des Experiments. Dann berechnen wir die Posterior-Verteilung der DAGs (bzw. SCMs, die die kausale Struktur kodieren) und aktualisieren anschließend die GP-Fits in den SCMs. Wir beenden das Verfahren, sobald die Posterior-Wahrscheinlichkeit für ein DAG deutlich gößer ist als für alle anderen. Wir haben das Verfahren für den Fall von zwei Zufallsvariablen implementiert und den Algorithmus an synthetischen Daten getestet. Dabei hat der Algorithmus die richtige kausale Struktur mit großer Zuverlässigkeit erkannt. In den Experimenten hat die Implementierung versucht, wahrscheinliche kausale Beziehungen zu bestätigen. Dafür is es am informativsten auf dem jeweiligen Grund der Kausalbeziehung zu intervenieren; immer in einer gleichbleibenden kleinen Region, in der wir bereits einige anfängliche Beobachtungen haben und die Schätzung der funktionalen Beziehung zwischen Grund und Effekt einen gewissen Grad an nichtlinearer Krümmung aufweist. Außerdem wird beschrieben, wie die Implementierung auf den Fall von vier Variablen verallgemeinert werden kann. Basierend auf den Beobachtungen über das Verhalten des Algorithmus im bivariaten Fall schlagen wir heuristisch ein Verfahren vor, das durch Ideen aus der Bayes'schen Optimierung motiviert ist, und den Rechenaufwand verringern kann.

Der Flaschenhals des Verfahrens ist das Optimierungsverfahren, welches schon bei vier Variablen zu zeitaufwendige Rechnungen benötigt. Die Beobachtung, dass Interventionen immer in dem selben Bereich gewählt werden deutet jedoch darauf hin, dass die Optimierung sich vereinfachen lässt und es ein alternatives Verfahren mit vergleichbaren Ergebnissen gibt.

# English Abstract

In Causal Discovery we aim to learn the causal structure from a set of system random variables based on purely observational data. Recent results showed, that this is possible in many cases. If interventions are possible, we can in principle infer the causal structure from comparably little interventional observations. In situations where observations are very expensive but interventions can be done, it is natural to ask, what would be the most informative intervention to determine the underlying causal structure. In this thesis, we focus on learning causal relations using interventions and implement an algorithm that simulates an intelligent agent which repeatedly interacts with its environment.

In this work we consider an Active Bayesian approach for causal structure learning. It is rooted in the intersection of statistical research on Causal Discovery and research on Active Learning in the area of Artificial Intelligence. The approach was proposed by [49, von Kügelgen et al.] and in this paper we present numerical results of the full implementation for the first time. We model an environment of random variables using Structural Causal Models (SCM) that imply corresponding Directed Acyclic Graphs (DAGs) and estimate the functional relations therein using Gaussian Process (GP) regression. We assume that we can perform perfect interventions on the environment variables. In each step iteration of a sequential procedure we choose the most informative intervention experiment (using Bayesian Experimental Design) and obtain one interventional data point from performing the experiment. Then, we calculate the posterior distribution of the DAGs (resp. SCMs, which encode the causal structure) and afterwards update the GP fits in the SCM. We terminate the procedure as soon as the posterior probability for one DAG is significantly larger than for all others.

We implement the procedure for the case of two random variables. We test the algorithm on synthetic data from which the true causal relation is know. The implemented algorithm finds the true causal relation with great reliability for many different parameter choices. In the experiments we find indication that it is best to confirm likely causal relations by intervening on the respective causes; always on the same support region, where we already have some initial observations and where the estimate of the functional relation has some degree of nonlinear curvature. We also describe how to generalize the implementation to the case of four variables, for which computations become very time consuming. Based on the observations about the behavior of the algorithm in the bivariate case, we heuristically propose a procedure how to overcome the computational burden (motivated by ideas from Bayesian Optimization).

The bottleneck of the method is the optimization procedure, which already requires too time-consuming calculations in the case of four variables. However, the observation that interventions are always chosen in the same region indicates that the optimization can be simplified and that there exists is an alternative procedure with comparable results.

# Contents

# List of Symbols

The next list describes several symbols that will be later used within the body of the document

| | |
|---|---|
| $\mathbb{E}_X$ | Expectation operator w.r.t. the distribution of the random variable $X$ (we omit $X$ when it is clear from the context) |
| $\Gamma(\cdot)$ | Gamma function; $\Gamma : \mathbb{R}_+ \to \mathbb{R}_+; x \mapsto \int_0^\infty t^{x-1}e^{-t}\,dt$ |
| $\lambda$ | Lebesgue Measure |
| $\mathbb{1}$ | Indicator function |
| $\mathcal{B}(X)$ | Borel-$\sigma$-algebra of $X$, where $(X, \tau)$ is a topological space |
| $\mathcal{F}[\cdot]$ | Fourier Transf. ;$\mathcal{F}[f](\omega) := (2\pi)^{-d/2}\int f(x)e^{-i\langle x,\omega\rangle}\,dx$,$\omega \in \mathbb{R}^d, f \in L^1(\mathbb{R}^d)$ |
| $\mathcal{O}(\cdot)$ | Landau symbol |
| $\mathcal{O}^*(\cdot)$ | Landau symbol for boundedness in probability ("tight") |
| $\mathcal{U}[a, b]$ | Uniform distribution over the real interval $[a, b]$ |
| $\Phi(\cdot)$ | Cumulative density function of a standard normal random variable, $\Phi : \mathbb{R} \to \mathbb{R}; x \mapsto \int_{-\infty}^x e^{-\frac{t^2}{2}}\,dt$ |
| $\phi(\cdot)$ | Probability density function of a standard normal random variable, $\phi : \mathbb{R} \to \mathbb{R}; t \mapsto e^{-\frac{t^2}{2}}$ |
| $\mathbf{P}_X$ | Distribution of the random variable $X$ |
| $p_X$ | Probability density function (w.r.t. Lebesgue or Counting Measure) of the random variable $X$ |
| $Py$ | Orthogonal projection from a Hilbert space $H$ onto a subspace $U$, $Py := \operatorname{argmin}_{x \in U} \|y - x\|$ |
| $\mathbb{K}$ | Either $\mathbb{R}$ or $\mathbb{C}$ |
| $\mathcal{X}$ | Respective set of interest |
| $A \Subset B$ | $A$ is a compact subset of $B$ |
| $C_c^\infty(\Omega)$ | $:= \{\varphi \in C^\infty : supp(\varphi) \Subset \Omega\}$ |
| $L^1_{loc}(\Omega)$ | $:= \{f : \Omega \to \mathbb{K} \text{ measurable} : \forall K \Subset \Omega \text{ it holds } f\|_K \in L^1(K)\}$ |

# List of Figures

# List of Tables

# 1. Introduction

Causal relations are of central interest in almost all sciences. We start learning them in early childhood, for example when throwing a ball. In principle we have two methods to reveal such relations: manipulating features of our environment to see what other features do or do not change; and observing the variation of features of our environment without manipulating it. We call such a manipulation an intervention (experiment) and if we can perform interventions it is possible to reveal the true causal structure [15, 16]. But in many cases interventions are impossible to perform, for example we cannot change the blood type of a patient and observe its influence on some medication. In such cases we must try to infer causal structures from purely observational data, which is known as causal discovery or causal structure learning (see [31, 17],[23, IV Causal inference] for an overview).
In this work we focus on learning causal relations using interventions, by simulating an intelligent agent which repeatedly interacts with its environment. The approach of learning causal structures of some environment through experimenting, observing evidence and subsequently updating our hypothesis is related to Artificial Intelligence research which aims to build human like models [20] and closely follows the idea termed the "child as a scientist" [18]. In context of causal discovery the most interesting feature of the approach is that it may provide guidance for the selection of next experiment. Especially when performing interventions is very expensive or extremely time consuming, it is of importance to find the most informative intervention experiment. The approach discussed in this thesis can provide guidance for selecting the most beneficial next experiment to infer the true causal structure of the environment variables.

Causal relations can be modeled by structural causal models which imply a corresponding graph that allows us to read off conditional independencies and a factorization of the joint distribution [29, 26]. Further, Pearl [29, Theorem 3.4.1] developed the *do*-calculus which provides a framework for interventions. We will utilize these advances in graphical modeling that have been established over the past decades and connect them to Bayesian Experimental Design [21] (an information theoretic concept based on the seminal work of Shannon [39]), Gaussian Process regression [34] (a flexible non parametric regression method) and Bayesian Optimization [24, 5] (a global optimization algorithm for black box functions). Throughout this work, we will explore some interrelationships between these lines of research.

This thesis considers a sequential active Bayesian approach for causal discovery proposed in [49], where we want to find the most informative intervention (in an information theoretic sense) at each step, then observe the respective interventional data point and update the estimated structural equations according to a given graph. The structural equations are estimated using Gaussian Process regression. In our setup we start with very little knowledge about the system, i.e. a few purely observational data points and try to find the most informative interventional data point, that allows us to decide on a causal relation with high probability. Since we will decide on a hypothesis (DAG) based on Bayesian inference we must pay attention on choosing the prior distribution (over DAGs) and about the hypothesis space (DAG space) (refer to Lindley's paradox [22]). For the calculation of the posterior we must calculate likelihoods for all possible graphs what implies a significant computational burden because of the super-exponential number of directed acyclic graphs [38]. In the bivariate case the approach performs well and provides very interesting insight regarding the role of interventional data and the choice of an intervention value among infinitely many possible values. The case of four random variables already calls for sampling approaches such as Monte Carlo Markov Chain (MCMC) or efficient cluster calculations. We will only discuss an implementation and heuristically point out alternative approaches that may be worth considering for future research in this direction, but do not provide numerical examples.

This work is structured as follows: In Chapter 2 we review a detailed description of the theoretical concepts we will use. For many results the derivations and proofs are included because they provide interesting insights about the interrelationships between the concepts. We start with GP regression, argue why we restrict on the Matérn kernel as covariance function and give a full description of the resulting function space from which we infer the regression function. Then, we introduce the information theoretic concepts we will need for Bayesian Optimization and introduce the special case of Bayesian Experimental design we will use. Further we consider theory and a convergence result for Bayesian Optimization. For graphical modelling we provide a detailed introduction and present the most important results for our approach. In Chapter 3 we introduce the approach proposed in [49] in much detail. We then start with the case of only two variables where the interventional data enters the model independent of the initial observations. This already shows significant performance. After establishing a solid understanding of the procedure we add the step to find the optimal intervention and perform the most informative experiment to obtain the next intervention data point. But this implies a dependence between the data points. We cope with this problem by assuming a sequential structure, which allows to calculate the likelihood in closed form consisting of Gaussian density functions. We then present numerical examples related to existing research about additive noise models (a subclass of structural equation models). Then, we turn to the fourvariate case, present the necessary calculation steps one would need to perform, point out the computational challenges and present heuristic ideas how one could possibly overcome the computational challenges. In Chapter 4 we conclude the thesis.

# 2. Theoretical Concepts

## 2.1. Gaussian Process Regression

This section provides a detailed derivation of the concepts of Gaussian process (GP) regression, that we will use in the proceeding of this work. Gaussian processes are used to describe a distribution over functions in a function space corresponding to a so called kernel. We infer the regression function as the mean function of the posterior process given some observations. The derivation loosely follows [34, Chapter 2] and the most relevant results are summarized in theorem (2.6) at the end of this section.
At first we need the notion of a kernel, which will be used to specify the covariance between any two points of our space under consideration, $\mathcal{X}$. The following definition is taken from [19].

**Definition 2.1** (Kernel)**.** Let $\mathcal{X}$ be a nonempty set. A symmetric function $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ is called a **positive definite kernel**, if for any $n \in \mathbb{N}$, $(c_1, \ldots, c_n) \subset \mathbb{R}$ and $(x_1, \ldots, x_n) \subset \mathcal{X}$,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j k(x_i, x_j) \geq 0.$$

To simplify notation, we treat kernel as a synonym for positive definite kernel.

For any finite collection of points, $\mathbf{x} = (x_1, \ldots, x_n) \in \mathcal{X}$, we denote by $k_{XX}$ the matrix with entries $[k_{XX}]_{ij} = k(x_i, x_j)$. The above definition implies that $k_{XX}$ is a positive semi definite matrix. In the literature $k_{XX}$ is called kernel matrix or Gram matrix.
The following definition of Gaussian processes closely follows [14] and [19].

**Definition 2.2** (Gaussian processes)**.** Let $\mathcal{X}$ be a nonempty set, $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a positive definite kernel and $m : \mathcal{X} \to \mathbb{R}$ be any real-valued function. Then, a random function $f : \mathcal{X} \to \mathbb{R}$ is said to be a **Gaussian Process (GP)** with mean function $m$ and covariance kernel $k$, denoted by $f \sim \mathcal{GP}(m, k)$, if the following holds: For any finite set $X = (x_1, \ldots, x_n) \in \mathcal{X}$ of any size $n \in \mathbb{N}$, the random vector

$$f_X = (f(x_1), \ldots, f(x_n))^\intercal \in \mathbb{R}^n$$

follows the multivariate normal distribution $\mathcal{N}(m_X, k_{XX})$ with covariance matrix $k_{XX} = (k(x_i, x_j))_{i,j=1}^{n} \in \mathbb{R}^{n\times n}$ and mean vector $m_X = (m(x_1), \ldots, m(x_n))^\intercal$.

**Remark 2.3.** In [14, Theorem 12.1.3. on p. 443] it is proven, that there is a one-to-one correspondence between Gaussian processes $f \sim \mathcal{GP}(m, k)$ and pairs $(m, k)$ of mean function $m$ and kernel $k$. This relation follows from an application of Kolmogorovs extension theorem.

Consider a situation where we have $N$ i.i.d. samples, $(x_i, y_i) \in \mathcal{Z} := \mathcal{X} \times \mathcal{Y}$ for $1 \leq i \leq N$, drawn from an unknown probability distribution $\mathbb{P}_{\mathcal{Z}}$ on $\mathcal{Z}$ of the random vector $Z = (X, Y)^\intercal$, where $\mathcal{X}$ denotes the input space and $\mathcal{Y}$ denotes the output space. Assume that there exists a functional relation of the form

$$y_i = f(x_i) + \epsilon_i,$$

where $\epsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma_n^2)$ for $1 \leq i \leq N$ and jointly independent of $\mathbf{x} = (x_1, \ldots, x_N)$. If $f \sim \mathcal{GP}(0, k)$, then $\tilde{f} = f + m \sim \mathcal{GP}(m, k)$ for a deterministic mean function $m$. Therefore we can for simplicity assume our prior believes to have mean function zero, i.e., assume $f \sim \mathcal{GP}(0, k)$ for some fixed kernel. A zero mean function is not necessarily a drastic limitation in practical applications, since it does not restrict the mean function of the posterior process to be zero.

We want to infer a function on some domain $\tilde{X}_* \subset \mathcal{X}$. For this purpose we discretize the domain in a suitable way such that our task reduces to infer the function values at the points $\mathbf{x}_* = \{x_*^{(1)}, \ldots, x_*^{(n)}\} \in \tilde{X}_*$. Denote the respective function values by $\mathbf{f}_{X_*} = (f(x_*^{(1)}), \ldots, f(x_*^{(n)}))^\intercal$. According to our assumptions, we thus have

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_{X_*} \end{bmatrix} |\mathbf{x}{,}\mathbf{x}_* \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} k_{XX} + \sigma_n^2 I & k_{XX_*} \\ k_{X_*X} & k_{X_*X_*} \end{bmatrix}\right).$$

**Remark 2.4** (Conditional Distribution of Multivariate Gaussian Vector)**.** Let $N = (X_1, \ldots, X_n, Y_1, \ldots, Y_l)$ be multivariate normal with density $f_N(\cdot)$. By definition, for any $\alpha \in \mathbb{R}^{n+l}$ it holds that $\alpha^\intercal N$ is normally distributed, thus, the vector $(Y_1, \ldots, Y_l)$ is also normally distributed with density $f_{\mathbf{Y}}$. The conditional distribution of $\mathbf{X}$ given $\mathbf{Y}$, $f_{\mathbf{X}|\mathbf{Y}}(\cdot, \mathbf{x}) = \frac{f_N(\cdot)}{f_{\mathbf{Y}}(\cdot)}$, is again normally distributed, what can be proven by actually performing the division and use some inversion lemma for matrices to arrive at a normal form of the distribution function.

Since a multivariate normal distribution is fully specified by its mean and variance, we can obtain the posterior distribution by calculating $\mathbb{E}\left[\mathbf{f}_{X_*}|\mathbf{y}{,}\mathbf{x}{,}\mathbf{x}_*\right]$ and $\operatorname{Var}\left(\mathbf{f}_{X_*}|\mathbf{y}{,}\mathbf{x}{,}\mathbf{x}_*\right)$. For this purpose we define $A := -k_{X_*X}(k_{XX} + \sigma_n^2 I)^{-1}$ and an auxiliary normally distributed random variable $\mathbf{z} := \mathbf{f}_{X_*} + A\mathbf{y}$ that is orthogonal to $\mathbf{y}$, since

$$\operatorname{Cov}(\mathbf{z}, \mathbf{y}) = \operatorname{Cov}(\mathbf{f}_{X_*}, \mathbf{y}) + \operatorname{Cov}(A\mathbf{y}, \mathbf{y}) = k_{X_*X} + A(k_{XX} + \sigma_n^2 I) = k_{X_*X} - k_{X_*X} = 0,$$

and zero covariance between two Gaussians is equivalent to independence. We have

$$\begin{aligned} \mathbb{E}\left[\mathbf{f}_{X_*}|\mathbf{y}{,}\mathbf{x}{,}\mathbf{x}_*\right] &= \mathbb{E}\left[\mathbf{z} - A\mathbf{y}|\mathbf{y}{,}\mathbf{x}{,}\mathbf{x}_*\right] = \mathbb{E}\left[\mathbf{z}|\mathbf{y}{,}\mathbf{x}{,}\ \mathbf{x}_*\right] - \mathbb{E}\left[A\mathbf{y}|\mathbf{y}{,}\mathbf{x}{,}\mathbf{x}_*\right] \\ &\overset{\mathbf{y}\perp\mathbf{z}}{=} \mathbb{E}\left[\mathbf{z}|\mathbf{x}{,}\mathbf{x}_*\right] - A\mathbf{y} = -A\mathbf{y} = k_{X_*X}(k_{XX} + \sigma_n^2 I)^{-1}\mathbf{y} \end{aligned} \tag{2.1}$$

and

$$
\begin{aligned}
\operatorname{Var}(\mathbf{f}_{X_*}|\mathbf{y},\mathbf{x},\ \mathbf{x}_*) &= \operatorname{Var}(\mathbf{z} - A\mathbf{y}|\mathbf{y},\mathbf{x},\mathbf{x}_*)\\
&= \operatorname{Cov}(\mathbf{z} - A\mathbf{y}, \mathbf{z} - A\mathbf{y}|\mathbf{y},\mathbf{x},\mathbf{x}_*)\\
&= \operatorname{Var}(\mathbf{z}|\mathbf{y},\mathbf{x},\mathbf{x}_*) + \underbrace{\operatorname{Var}(A\mathbf{y}|\mathbf{y},\mathbf{x},\mathbf{x}_*)}_{=0} - 2A\underbrace{\operatorname{Cov}(\mathbf{y},\mathbf{z}|\mathbf{y},\mathbf{x},\mathbf{x}_*)}_{=0}\\
&\overset{\mathbf{y}\perp\mathbf{z}}{=} \operatorname{Var}(\mathbf{z}|\mathbf{x},\ \mathbf{x}_*)\\
&= \operatorname{Var}(\mathbf{f}_{X_*} + A\mathbf{y}|\mathbf{x},\mathbf{x}_*)\\
&= \operatorname{Var}(\mathbf{f}_{X_*}|\mathbf{x},\mathbf{x}_*) + A\operatorname{Var}(\mathbf{y}|\mathbf{x},\mathbf{x}_*)A^\intercal + 2A\operatorname{Cov}(\mathbf{f}_{X_*},\mathbf{y}|\mathbf{x},\mathbf{x}_*)\\
{\scriptstyle (k_{XX}+\sigma_n^2 I)\ \text{symmetric}} &= k_{X_*X_*} + k_{X_*X}(k_{XX}+\sigma_n^2 I)^{-1}k_{XX_*} - 2k_{X_*X}(k_{XX}+\sigma_n^2 I)^{-1}k_{XX_*}\\
&= k_{X_*X_*} - k_{X_*X}(k_{XX}+\sigma_n^2 I)^{-1}k_{XX_*}.
\end{aligned}
\tag{2.2}
$$

To summarize, the predictive posterior distribution is,

$$
\mathbf{f}_{X_*}|\mathbf{y},\mathbf{x},\mathbf{x}_* \sim \mathcal{N}\left(k_{X_*X}(k_{XX}+\sigma_n^2 I)^{-1}\mathbf{y}, k_{X_*X_*} - k_{X_*X}(k_{XX}+\sigma_n^2 I)^{-1}k_{XX_*}\right). \tag{2.3}
$$

Another useful property in GP regression is that we can compute the marginal likelihood, $p(\mathbf{y}|\mathbf{x}) = \int_{\mathbb{R}^n} p(\mathbf{y}|\mathbf{f}_X,\mathbf{x})p(\mathbf{f}_X|\mathbf{x})\,d\mathbf{f}_X$, in closed form as below. Under the Gaussian process model we have $\mathbf{f}_X|\mathbf{x} \sim \mathcal{N}(0, k_{XX})$ and $\mathbf{y}|\mathbf{f}_X,\mathbf{x} \sim \mathcal{N}(\mathbf{f}_X, \sigma_n^2 I_N)$. First we investigate the product of the two densities

$$
\begin{aligned}
p(\mathbf{y}|\mathbf{f}_X,x)p(\mathbf{f}_X|\mathbf{x}) &= (2\pi)^{-\frac{N}{2}}\det(k_{XX})^{-\frac{1}{2}}(2\pi)^{-\frac{N}{2}}\det(\sigma_n^2 I_N)^{-\frac{1}{2}}\\
&\qquad \exp\left[-\frac{1}{2}\mathbf{f}_X{}^\intercal k_{XX}^{-1}\mathbf{f}_X\right]\exp\left[-\frac{1}{2}(\mathbf{y}-\mathbf{f}_X)^\intercal(\sigma_n^2 I_N)^{-1}(\mathbf{y}-\mathbf{f}_X)\right]\\
&= (2\pi)^{-N}\det(k_{XX})^{-\frac{1}{2}}\det(\sigma_n^2 I_N)^{-\frac{1}{2}}\\
&\qquad \exp\left[-\frac{1}{2}\mathbf{f}_X^\intercal k_{XX}^{-1}\mathbf{f}_X - \frac{1}{2}(\mathbf{y}-\mathbf{f}_X)^\intercal(\sigma_n^2 I_N)^{-1}(\mathbf{y}-\mathbf{f}_X)\right]\\
&= (2\pi)^{-N}\det(k_{XX})^{-\frac{1}{2}}\det(\sigma_n^2 I_N)^{-\frac{1}{2}}\\
&\qquad \exp\left[-\frac{1}{2}\mathbf{f}_X^\intercal\underbrace{(k_{XX}^{-1}+(\sigma_n^2 I_N)^{-1})}_{:=\Sigma^{-1}}\mathbf{f}_X - \frac{1}{2\sigma_n^2}\mathbf{y}^\intercal\mathbf{y} + \frac{1}{\sigma_n^2}\mathbf{y}^\intercal\mathbf{f}_X\right]\\
&\overset{(*)}{=} (2\pi)^{-N}\det(k_{XX})^{-\frac{1}{2}}\det(\sigma_n^2 I_N)^{-\frac{1}{2}}\\
&\qquad \exp\left[-\frac{1}{2}(\mathbf{f}_X - \frac{1}{\sigma_n^2}\Sigma\mathbf{y})^\intercal\Sigma^{-1}(\mathbf{f}_X - \frac{1}{\sigma_n^2}\Sigma\mathbf{y}) - \frac{1}{2}\mathbf{y}^\intercal(k_{XX}+\sigma_n^2 I_N)^{-1}\mathbf{y}\right]\\
&\overset{(**)}{=} (2\pi)^{-\frac{N}{2}}\det((k_{XX}+\sigma_n^2 I_N)^{-1})^{\frac{1}{2}}\exp\left[-\frac{1}{2}\mathbf{y}^\intercal(k_{XX}+\sigma_n^2 I_N)^{-1}\mathbf{y}\right]\\
&\qquad (2\pi)^{-\frac{N}{2}}\det(\Sigma)^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(\mathbf{f}_X - \frac{1}{\sigma_n^2}\Sigma\mathbf{y}^\intercal)\Sigma^{-1}(\mathbf{f}_X - \frac{1}{\sigma_n^2}\Sigma\mathbf{y})\right]
\end{aligned}
$$

where we used that

$$
\begin{aligned}
(*) :& (\mathbf{f}_X - \frac{1}{\sigma_n^2}\Sigma\mathbf{y})^\intercal\Sigma^{-1}(\mathbf{f}_X - \frac{1}{\sigma_n^2}\Sigma\mathbf{y}) = \mathbf{f}_X{}^\intercal\Sigma^{-1}\mathbf{f}_X - \frac{2}{\sigma_n^2}\mathbf{y}^\intercal\mathbf{f}_X + \frac{1}{\sigma_n^4}\mathbf{y}^\intercal\Sigma\mathbf{y} \\
&\overset{A.9}{=} \mathbf{f}_X{}^\intercal\Sigma^{-1}\mathbf{f}_X - \frac{2}{\sigma_n^2}\mathbf{y}^\intercal\mathbf{f}_X + \frac{1}{\sigma_n^2}\mathbf{y}^\intercal\mathbf{y} - \mathbf{y}^\intercal(k_{XX} + \sigma_n^2 I_N)^{-1}\mathbf{y}
\end{aligned}
$$

and

$$
\begin{aligned}
(**) : \det(\Sigma) &\overset{A.9}{=} \det(k_{XX} - k_{XX}(k_{XX} + \sigma_n^2 I_N)^{-1}k_{XX}) \\
&\overset{A.11}{=} \det(k_{XX})\det((k_{XX} + \sigma_n^2 I_N)^{-1})\det(k_{XX} + \sigma_n^2 I_N - k_{XX}k_{XX}^{-1}k_{XX}) \\
&= \det(k_{XX})\det((k_{XX} + \sigma_n^2 I_N)^{-1})\det(\sigma_n^2 I_N).
\end{aligned}
$$

Thus for the integral we calculate

$$
\begin{aligned}
p(\mathbf{y}|\mathbf{x}) &= \int_{\mathbb{R}^n} p(\mathbf{y}|\mathbf{f}_X, x)p(\mathbf{f}_X|\mathbf{x})d\mathbf{f}_X \\
&= (2\pi)^{-\frac{N}{2}}\det((k_{XX} + \sigma_n^2 I_N)^{-1})^{\frac{1}{2}}\exp\left[-\frac{1}{2}\mathbf{y}^\intercal(k_{XX} + \sigma_n^2 I_N)^{-1}\mathbf{y}\right] \\
&\quad \int_{\mathbb{R}^n}(2\pi)^{-\frac{N}{2}}\det(\Sigma^{-1})^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(\mathbf{f}_X - \frac{1}{\sigma_n^2}\Sigma\mathbf{y})^\intercal\Sigma^{-1}(\mathbf{f}_X - \frac{1}{\sigma_n^2}\Sigma\mathbf{y})\right]d\mathbf{f}_X \\
&= (2\pi)^{-\frac{N}{2}}\det(k_{XX} + \sigma_n^2 I_N)^{-\frac{1}{2}}\exp\left[-\frac{1}{2}\mathbf{y}^\intercal(k_{XX} + \sigma_n^2 I_N)^{-1}\mathbf{y}\right].
\end{aligned}
\tag{2.4}
$$

**Remark 2.5.** Another way to derive the marginal likelihood is to use (A.8) and immediately arrive at the result. But the longer proof provides useful tools to work with multivariate normal distributions, which is why it is presented above.

We summarize the above results in the following theorem.

**Theorem 2.6.** *Under our Gaussian process model defined above ($y = f(x) + \epsilon$), the marginal likelihood of observing the (noisy) targets $\boldsymbol{y}$ given the covariates $\boldsymbol{x}$ is the normal density*

$$
\boldsymbol{y}|\boldsymbol{x} \sim \mathcal{N}\left(\boldsymbol{0}, k_{XX} + \sigma_n^2 I_N\right). \tag{2.5}
$$

*The predictive posterior distribution of the function values, $\boldsymbol{f}_{X_*}$, at a set of covariates, $\boldsymbol{x}_* = \{x_*^{(1)}, \dots, x_*^{(n)}\}$, is*

$$
\boldsymbol{f}_{X_*}|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{x}_* \sim \mathcal{N}\left(k_{X_*X}(k_{XX} + \sigma_n^2 I)^{-1}\boldsymbol{y}, k_{X_*X_*} - k_{X_*X}(k_{XX} + \sigma_n^2 I)^{-1}k_{XX_*}\right). \tag{2.6}
$$

*Proof.* See the two calculations above. □

**Remark 2.7.** Since (2.6) holds for any set of points $X^* \in \mathcal{X}^m$ of any size $m \in \mathbb{N}$, Kolmogorovs extension theorem [14, Theorems 12.1.2] and the definition of GPs imply that the process $f \sim \mathcal{GP}(0, k)$ *conditioned on* the training data $\mathbf{x}, \mathbf{y}$ is a draw from $\mathcal{GP}(\bar{m}, \bar{k})$, with

$$
\begin{aligned}
\bar{m}(x) &= k_{xX}(k_{XX} + \sigma_n^2 I)^{-1}\mathbf{y}, \quad x \in \mathcal{X}, \\
\bar{k}(x, x') &= k(x, x') - k_{xX}(k_{XX} + \sigma_n^2 I)^{-1}k_{Xx'}, \quad x, x' \in \mathcal{X}.
\end{aligned}
\tag{2.7}
$$

## 2.2. Matérn Kernel and its Function Space

In the previous section we have seen, that the GP regression heavily relies on the chosen kernel. In the proceeding work we will only use the so called Matérn kernels, as it is suggested in [42, Section 1.6]. The following section provides an overview of the class of Matérn kernels and in particular their properties and the resulting function (sample) space.

### Matérn Class

**Definition 2.8.** The **Matérn class of kernels** (or covariance functions) is given by

$$k_{\nu,\gamma}(x, x') = \frac{2^{1-\nu}}{\Gamma(\nu)} \left( \frac{\sqrt{2\nu}||x - x'||_2}{\gamma} \right)^{\nu} K_{\nu} \left( \frac{\sqrt{2\nu}||x - x'||_2}{\gamma} \right), \tag{2.8}$$

with $\nu, \gamma \in \mathbb{R}_+$ and $K_\nu$ a modified Bessel function of the second kind [46], i.e., for $\nu \in \mathbb{R}$ and $x > 0$,

$$K_\nu(x) = \int_0^\infty \cosh(\nu t) \exp(-x \cosh(t)) \, dt.$$

**Remark 2.9.** In general, the definition works with any norm. Since we are working in $\mathbb{R}^n$ and therefore have norm equivalence, we choose the euclidean norm in the definition.

In the following the most important properties of the Matérn kernel, stated in [34, Chapter 4], are summarized.

1. Matérn kernels are isotropic (and therefore also stationary) as functions of $||x - x'|| =: r$ and thus invariant to all rigid motions in the input space. In stochastic process theory, processes with constant mean and covariance function invariant to translations are called weakly stationary.

2. The process $f \sim \mathcal{GP}(m, k)$ is $l$-times mean square differentiable if and only if $\nu > l$.

3. If $\nu$ is half integer, that is $\nu = p + \frac{1}{2}$, with $p \in \mathbb{N}_+$; then the Matérn kernel is a product of an exponential and a polynomial of order $p$ and can be written as

$$k_{\nu=p+\frac{1}{2},\gamma} = \exp\left( -\frac{\sqrt{2\nu}r}{\gamma} \right) \frac{\Gamma(p+1)}{\Gamma(2p+1)} \sum_{i=0}^{p} \frac{(p+i)!}{i!(p-i)!} \left( \frac{\sqrt{8\nu}r}{\gamma} \right)^{p-i}. \tag{2.9}$$

4. For $\nu = \frac{1}{2}$ we obtain the **exponential (or Gaussian) kernel**

$$k_E(r) = \exp(-\frac{r}{\gamma}).$$

The corresponding process is mean square continuous but not mean square differentiable. If the input space is one dimensional this is the covariance function of the Ornstein-Uhlenbeck process [47].

5. For $\nu \to \infty$ we obtain the **squared exponential kernel**

$$k_{SE}(r) = \exp(-\frac{r^2}{2\gamma^2})$$

(often this kernel is called radial-basis function (RBF) kernel).

As a practical guide it is suggested in [34, Section 4.2.1] to choose $\nu \in \{\frac{3}{2}, \frac{5}{2}\}$, because for lower values the processes becomes very rough (like a standard Brownian Motion) and for higher values the existence of higher order derivatives should be justified by prior knowledge.

### GP Sample Space for Matérn kernels

For readability we state only two main results and give some intuition on how to obtain these facts. The required terminology and most important theorems are summarized in (C). The below facts hold for the setup of this work. In particular we assume that $\mathcal{X} \subset \mathbb{R}^d$ is a compact metric space.

1. The posterior mean function of the GP regression using a Matérn kernel, $k_{\nu,\gamma}$, is an element of the RKHS $\mathcal{H}_{k_{\nu,\gamma}}$ (see C.7 for an explicit characterization).

2. For a given $f \sim \mathcal{GP}(0, k_{\nu,\gamma})$, there exists a version (see C.8 for a definition) $\tilde{f}$ such that $\tilde{f} \in \mathcal{H}_{k_{\nu',\gamma'}}$ with probability 1 for all $\nu', \gamma' > 0$ satisfying $\nu > \nu' + d/2 \in \mathbb{N}$, where $\mathcal{H}_{k_{\nu',\gamma'}}$ is the RKHS of the Matérn kernel $k_{\nu',\gamma'}$ with parameters $\nu'$ and $\gamma'$. Further the RKHS $\mathcal{H}_{k_{\nu',\gamma'}}$ is norm equivalent to the Sobolev space $W(\mathcal{X})^{(\nu'+d/2),2}$ (see C.13, C.5 and [19, Section 4] for details).

The first assertion follows from the result, that the posterior mean of a GP regression with kernel $k$ is equal to the result of a kernel ridge regression with the exact same kernel [19, Proposition 3.6]. From the latter one it is known to be a member of the RKHS $\mathcal{H}_k$ [19, Section 3.2].

The second fact relies on a couple of significant theorems from functional analysis and requires a rich language before one can formally prove them. Therefore we only provide a very short informal summary of the argumentation.

At first one can show that GP sample space for a given kernel $k$ is equivalent to the $\theta$-power of the RKHS $\mathcal{H}_k$ (see C.10 for a definition). This means that the GP sample space is an, with rougher functions, enlarged version of $\mathcal{H}_k$ (see C.11). Second, in case of Matérn kernels, the resulting enlarged Hilbert space $\mathcal{H}^{\theta}_{k_{\nu,\gamma}}$ is equivalent to the RKHS $\mathcal{H}_{k_{\nu',\gamma'}}$ of a modified Matérn kernel with parameters $\nu', \gamma'$, and thus equivalent to a Sobolev space of order $(\nu' + d/2)$, where $\nu > \nu' + d/2 \in \mathbb{N}$ (see C.5). In total our GP sample space is a Sobolev space which has approximately order $\nu - d/2$ and therefore also includes functions that have "less" weak derivatives than in the RKHS $\mathcal{H}_{k_{\nu,\gamma}}$.

## 2.3. Information Theory

This section provides some basics of information theory that are needed to define the information gain. In the computer science literature information gain often means the mutual information between two random variables, which is the Kullback-Leibler (KL) divergence between the joint distribution and the product of the two marginal distributions. We start with classical definitions and lemmas taken from [11], that we need for Bayesian Optimization and then describe the idea of designing an experiment which gains us most information (in expectation) about the true parameter value in a Bayesian sense (following [21]).

**Definition 2.10.** The **differential entropy** of a set $X_1, X_2, \ldots, X_n$ of random variables with density $p(x_1, x_2, \ldots, x_n)$ is defined as

$$h(X_1, X_2, \ldots, X_n) = -\int p(\mathbf{x}) \log p(\mathbf{x}) \, d\mathbf{x}. \tag{2.10}$$

**Definition 2.11.** If $X, Y$ have a joint density function $p(x, y)$, we can define the **conditional differential entropy** $h(X|Y)$ as

$$h(X|Y) = -\iint p(x, y) \log p(x|y) \, dx \, dy. \tag{2.11}$$

Since in general $p(x|y) = p(x, y)/p(y)$ with $p(y) = \int p(x, y) \, dx$, we can also write (assuming any of the differential entropies are finite)

$$h(X|Y) = h(X, Y) - h(Y). \tag{2.12}$$

**Theorem 2.12** (Entropy of a multivariate normal distribution)**.** *Let* $\boldsymbol{X} \sim \mathcal{N}_n(\mu, \Sigma)$*, i.e. has multivariate normal distribution (A.5). Then*

$$h(\boldsymbol{X}) = \frac{1}{2} \log \left( (2\pi e)^n \det(\Sigma) \right). \tag{2.13}$$

*Proof.* A calculation of (2.10) with the multivariate normal density yields the assertion (see [11, Theorem 8.4.1]). □

**Definition 2.13.** The **mutual information** $I(X; Y)$ between two random variables with joint density $p(x, y)$ is defined as

$$I(X; Y) = \iint p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \, dx \, dy. \tag{2.14}$$

From the definition it is clear that

$$I(X; Y) = h(X) - h(X|Y) = h(Y) - h(Y|X) = h(X) + h(Y) - h(X, Y), \tag{2.15}$$

and that

$$I(X; Y) = D_{KL} \left( p(x, y) \, || \, p(x)p(y) \right), \tag{2.16}$$

where $D_{KL}$ denotes the **Kullback-Leibler** Divergence.

**Remark 2.14.** In the setup of our GP regression model $\mathbf{y} = \mathbf{f}_X + \boldsymbol{\epsilon}$ (2.6) we have, using (2.12) and (2.13),

$$\begin{aligned} I(\mathbf{y};\mathbf{f}_X) &= h(\mathbf{y}) - h(\mathbf{y}|\mathbf{f}_X) \\ &= \frac{1}{2}\log\left((2\pi e)^n \det(k_{XX} + \sigma_n^2 I)\right) - \frac{1}{2}\log\left((2\pi e)^n \det(\sigma_n^2 I)\right) \\ &= \frac{1}{2}\log\left(\frac{(2\pi e)^n \sigma_n^{2n} \det(\sigma_n^{-2} k_{XX} + I)}{(2\pi e)^n \sigma_n^{2n} \det(I)}\right) \\ &= \frac{1}{2}\log\left(\det(\sigma_n^{-2} k_{XX} + I)\right). \end{aligned} \tag{2.17}$$

## 2.3.1. Bayesian Experimental Design using Information Gain

For a general overview of Bayesian Experimental Design see [8]. Since we will only use the information gain as the quantity of interest, it improves readability to restrict on this case right away.

In Bayesian experimental design, the maximum information will be obtained when the posterior distribution is concentrated on a single parameter value. We define the amount of information provided by an experiment as the difference between posterior information and prior information [21]. Note that this quantity is large if the prior has a large variance and the posterior has little variance (resp. is concentrated on a single value).

**Definition 2.15.** The **amount of information** provided by an experiment $\mathcal{E}$ (implying a distribution for the observable outcome $X_{\mathcal{E}}$), with prior knowledge $p(\theta)$, when the observation is $x_{\mathcal{E}}$, is

$$g(\mathcal{E}, p(\theta), x_{\mathcal{E}}) = \int p(\theta|x_{\mathcal{E}}) \log(p(\theta|x_{\mathcal{E}}))\, d\theta - \int p(\theta) \log(p(\theta))\, d\theta. \tag{2.18}$$

**Remark 2.16.** The expression depends on $x_{\mathcal{E}}$ and some results are more informative than others. This does not need to cause concern when we consider the average information provided by an experiment.

**Definition 2.17.** The **average amount of information** (or information gain in $\theta$) provided by the experiment $\mathcal{E}$, with prior knowledge $p(\theta)$, is

$$g(\mathcal{E}, p(\theta)) = \mathbb{E}_{X_{\mathcal{E}}}\left[\int p(\theta|x) \log(p(\theta|x))\, d\theta - \int p(\theta) \log(p(\theta))\, d\theta\right] \tag{2.19}$$

**Remark 2.18.** The expected information gain in $\theta$ from conducting the experiment $\mathcal{E}$ is the mutual information between $\theta$ and the outcome of the experiment $X_{\mathcal{E}}$, i.e.

$$g(\mathcal{E}, p(\theta)) = I(\theta; X_{\mathcal{E}}) = D_{KL}\left(p_{\mathcal{E}}(\theta, x)\,||\,p(\theta)p_{\mathcal{E}}(x)\right). \tag{2.20}$$

With the notation $p_{\mathcal{E}}$ we stress that the distribution depends on the chosen experiment.

*Proof.* Since we have $p(x|y) = p(x,y)/p(y)$ with $p(y) = \int p(x,y)\,dx$, and Fubini's theorem [13, Satz 3.10] is applicable, it holds

$$\begin{aligned} g(\mathcal{E}, p(\theta)) &= \iint p_{\mathcal{E}}(\theta, x) \log\left(\frac{p_{\mathcal{E}}(\theta, x)}{p_{\mathcal{E}}(x)}\right) d\theta\, dx - \iint p_{\mathcal{E}}(\theta, x) \log\left(p(\theta)\right)\, dx\, d\theta \\ &= \iint p_{\mathcal{E}}(\theta, x) \log \frac{p_{\mathcal{E}}(\theta, x)}{p_{\mathcal{E}}(x) p(\theta)}\, dx\, d\theta. \end{aligned}$$

□

In this work we want to find the experimental design $\mathcal{E}$, that maximizes the average amount of information that we can gain from conducting the experiment, i.e. we try to find

$$\mathcal{E}^* = \underset{\mathcal{E}}{\operatorname{argmax}}\, g(\mathcal{E}, p(\theta)) = \underset{\mathcal{E}}{\operatorname{argmax}}\, D_{KL}\left(p_{\mathcal{E}}(\theta, x)\, ||\, p(\theta) p_{\mathcal{E}}(x)\right). \tag{2.21}$$

## 2.4. Bayesian Optimization

We consider a general definition of the Bayesian Optimization algorithm (see [5] for a gentle introduction) and then turn to a specific version where we will state a convergence result provided in [41]. Since the base proof for a simplified case of the convergence contains useful insights, it can be found in (B).

### Optimization Problem

Consider an optimization problem for a (nonlinear) function $f : \mathcal{X}_{ad} \to \mathbb{R}$ over an admissible set $\mathcal{X}_{ad} \subset \mathbb{R}^d$ ($d \in \mathbb{N}$), defined via box constraints,

$$\min_{x \in \mathcal{X}_{ad}} f(x). \tag{2.22}$$

For this type of problem there exists an enormous body of literature and for all cases where the objective function can be evaluated quickly there most probably exist more efficient algorithms than Bayesian optimization. But suppose we can only approximate the function value with some costly Monte Carlo simulation, then we are in a setting for which Bayesian Optimization is designed for, namely, optimization of a costly to evaluate objective function, of which we have little knowledge.

### 2.4.1. Algorithm

The numerical experiments in this work use an implementation of the Bayesian Optimization algorithm provided in the Python package "scikit-optimization". Therefore the pseudo code below corresponds to that implementation.
In each iteration the algorithm fits a new GP model for the objective function using all data points obtained up to this iteration and then finds the next data point by optimizing a simple acquisition function. The three main acquisition functions proposed in

the literature are the expected improvement (EI), lower confidence bound (LCB) and probability of improvement (PI). Proper definitions of the acquisition functions can be found after the pseudo code because we will define some required quantities therein.
Since we are in a setting with box constraints, it intuitively makes sense, that we will eventually find an optimal value if we just try at sufficiently many points (of course assuming some smoothness of $f$). The idea is to develop a strategy for picking points that, efficiently leads to an acceptable solution. In the literature this is often referred to as a continuum armed bandit problem [41] and usually the essence of these types of problems is to find approaches that balance exploration and exploitation suitably. In our setting exploration approximately means to pick a point somewhere in the domain where we have little knowledge and exploitation means to pick a point somewhere near the best value of the points we picked so far.

**Algorithm 1** Bayesian Optimization

```
Require: $f(\cdot)$, box constraints and $n$        ▷ method to evaluate the objective function
  $x_0$ = Initial Point Generator()                  ▷ by default random choice
  $y_0 = f(x_0)$                                     ▷ get a possibly noisy function evaluation
  $\mathcal{D}_0 = (x_0, y_0)$
  $\pi = \mathcal{GP}(0, k_{\text{Matérn}})$                       ▷ prior for GP regression
  for $i = 1, 2, \dots, n$ do
      Fit a GP model on $\mathcal{D}_{i-1}$ with prior $\pi$.             ▷ hyperparameters tuned
      Get the posterior distribution of the estimate $\hat{f} \sim \mathcal{GP}(\boldsymbol{\mu}_{i-1}, \boldsymbol{\sigma}^2_{i-1})$
      Find $\mathbf{x}_i = \operatorname{argmin}_{x \in \mathcal{X}_{ad}} u(\boldsymbol{\mu}_{i-1}(\cdot), \boldsymbol{\sigma}_{i-1}(\cdot), x_i)$   ▷ optimize acquisition function $u$
      Sample $y_i = f(x_i)$
      $\mathcal{D}_i = \{\mathcal{D}_{i-1}, (x_i, y_i)\}$
      $j^* = \operatorname{argmin}_{j \in \{0,\dots,i\}} y_j$                 ▷ store current best choice
      $x_i^* = x_{j^*}$; $y_i^* = y_{j^*}$
  end for
  return $x_i^*$, $y_i^*$
```

Note that the choice of the Matérn kernel implies that we model our objective function $f$ as a stationary Gaussian process. The following choices of acquisition functions are readily used and also implemented in the Python package ”scikit-optimization”.

**Definition 2.19** (**Lower Confidence Bound**)**.** Let $\kappa > 0$.

$$LCB(x_i) = \mu_{i-1}(x_i) - \kappa\sigma_{i-1}(x_i) \tag{2.23}$$

**Remark 2.20.** The parameter $\kappa$ is the so called trade-off (between exploitation and exploration) parameter. While we consider minimization of the acquisition function in the algorithm, literature often considers maximization and in that case one maximizes the **upper confidence bound (UCB)** instead of minimizing the LCB.

$$UCB(x_i) = \mu_{i-1}(x_i) + \kappa\sigma_{i-1}(x_i). \tag{2.24}$$

We point this out here because the convergence result in [41] considers a Bayesian optimization algorithm with maximization of UCB acquisition function.

**Definition 2.21** ((Modified) **Probability of Improvement).** Let $\kappa > 0$.

$$PI(x_i) = P(\hat{f}_{i-1}(x_i) \leq y_{i-1} - \kappa) = \Phi\left(\frac{y^*_{i-1} - \kappa - \mu_{i-1}(x_i)}{\sigma_{i-1}(x_i)}\right) \tag{2.25}$$

**Remark 2.22.** Using the probability of improvement without the trade-off parameter $\kappa$ as acquisition function enhances pure exploitation, because points that have a high probability of being infinitesimally greater than $y^*_{i-1}$ will be drawn over points that offer larger gains but less certainty.

Using our knowledge about GP regression, one can establish a closed form of the expected improvement. By $(\cdot)^-$ we denote the function that is the identity if the argument is negative and zero otherwise.

**Definition 2.23** (**Expected Improvement).** Assume that $\sigma_{i-1}(x_i) > 0$, otherwise we set $EI(x_i) = 0$. Integration by parts yields

$$\begin{aligned} EI(x_i) &= \mathbb{E}[(y^*_{i-1} - \hat{f}(x_i))^-] \\ &= y^*_{i-1}P(\hat{f}_{i-1}(x_i) \leq y^*_{i-1}) - \int_{-\infty}^{\frac{y^*_{i-1}-\mu_{i-1}(x_i)}{\sigma_{i-1}(x_i)}} (z\sigma_{i-1}(x_i) + \mu_{i-1}(x_i)) \frac{e^{-\frac{1}{2}z^2}}{\sqrt{2\pi}\sigma_{i-1}(x_i)} \, dz \\ &= (y^*_{i-1} - \mu_{i-1}(x_i))\Phi\left(\frac{y^*_{i-1} - \mu_{i-1}(x_i)}{\sigma_{i-1}(x_i)}\right) - \sigma_{i-1}(x_i)\phi\left(\frac{y^*_{i-1} - \mu_{i-1}(x_i)}{\sigma_{i-1}(x_i)}\right) \end{aligned} \tag{2.26}$$

### 2.4.2. Convergence of GP optimization

We consider a convergence results for Bayesian Optimization with LCB (UCB) acquisition function which combines the two necessary results in [41, Theorem 3, Theorem 5] for our setting. A full proof of a simplified case can be found in the appendix. The details of lifting the (simple) proof to the general result can be found in [41, Appendix B]. Therein a martingale concentration inequality is established to bound the distance of the GP regression estimate of $f$ to $f$. The remainder of the proof is then similar to the one in (B). A prove of convergence for Bayesian optimization with EI acquisition function is established in [6].

In the following GP-UCB algorithm is the same as Algorithm 1 but with taking UCB acquisition function and the argmax instead of the argmin. In the following $t$ indicates the current iteration and $T$ is the last iterate. The below theorem bounds the **cumulative regret** $R_T$ after $T$ iterations, which is the sum of instantaneous regrets $r_t = f(x^*) - f(x_t)$, where $x^* = \text{argmax}_{x \in D} f(x)$ (note that $x^*$ does not have to be unique). The asymptotic property we opt for is called **no-regret**, $\lim_{T\to\infty} R_T/T = 0$. Intuitively the no-regret property means that on average we do not regret in the long run, what is a very strong statement because the instantaneous regret is always larger than zero. Thus an algorithm satisfying this property converges to a (possibly local) optimum. If we are able to find a function of $T$ that bounds $R_T$, we can also quantify the speed of convergence. Note that $r_t$ and $R_T$ are theoretic quantities which usually are not known and they are never revealed to the algorithm. Since the algorithm picks the iterates based on a GP fit that changes in every iteration it is clear that the $r_t$ quantities are neither independent nor identically distributed. Thus, it is already very good if we can establish results that bound the cumulative regret (with a quantity that grows slower than the identity) with high probability.

**Theorem 2.24.** *Let $\delta \in (0,1)$, $D \subset \mathbb{R}^d$ be compact and convex, $d \in \mathbb{N}$ and $\nu > 1$. Assume that the true underlying $f$ lies in the RKHS $\mathcal{H}_{k_{Matérn}}(D)$ corresponding to the kernel $k_{\nu,\tilde{\gamma}}$, and that the noise $\epsilon_t$ has zero mean conditioned on the history and is bounded by $\sigma$ almost surely. In particular, assume $\|f\|^2_{k_{\nu,\tilde{\gamma}}} \leq B$ and let $\beta_t = 2B + 300\gamma_t \log^3(t/\delta)$. Running GP-UCB with $\beta_t$, prior $\mathcal{GP}(0, k_{\nu,\tilde{\gamma}})$ and noise model $\mathcal{N}(0,\sigma^2)$, we obtain a regret bound of $\mathcal{O}^*(\sqrt{T}(B\sqrt{\gamma_T} + \gamma_T))$ with high probability (over the noise). Precisely,*

$$P(R_T \leq \sqrt{C_1 T \beta_T \gamma_T} \quad \forall T \geq 1) \geq 1 - \delta,$$

*where $C_1 = \frac{8}{\log(1+\sigma^{-2})}$ and $\gamma_T = \mathcal{O}(T^{\frac{d(d+1)}{2\nu+d(d+1)}} \log(T))$.*

*Proof.* See [41, Appendix B]. □

## 2.5. Causal Models

This section provides an introduction to the graphical concepts and models we develop our causal discovery approach upon. We start with an introduction of directed acyclic graphs (DAGs) and describe concepts which allow us to read of causalities from the respective graphs. Then, we relate distributions generated by Structural Causal Models (SCMs) to the corresponding graphs and obtain that they satisfy the necessary conditions, such that we can read off causal relations. We then turn to Additive Noise Models (ANMs), where we can infer the correct DAG from purely observational data for most distributions [32, Proposition 21]. We will later assume an additive noise structure but also include interventional data, despite some (pure) observations. It will be interesting to ask how the approach performs in cases where ANMs are non-identifiable from purely observational data [32, 52]. Lastly we consider interventions, introduce the do notation [28] and a procedure to calculate intervention distributions given a SCM. We mainly follow [31, Chapter 6] and [32, Section 1].

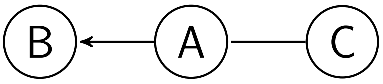
### 2.5.1. Graphical Concepts

The first part of this section very closely follows [31, Section 6.1] where the definitions are literally quoted.

Consider a $\mathbb{N} \ni p$-dimensional random vector $\mathbf{X} = (X_1, ..., X_p)$ with joint distribution

$\mathbf{P}_{\mathbf{X}}$, density $p(\mathbf{x})$ and index set $\mathcal{V} := \{1, \dots, p\}$. A **graph** $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ consists of (finitely many) **nodes** (or **vertices**) and **edges** $\mathcal{E} \subseteq \mathcal{V}^2$ with $(\nu, \nu) \notin \mathcal{E}$ for any $\nu \in \mathcal{V}$.

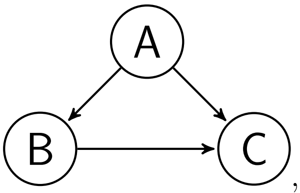
**Example 2.25** (Graph)**.** Let $\mathcal{V} = \{A, B, C\}$ and $\mathcal{E} = \{(A, B), (A, C), (C, A)\}$. Then $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ takes the form,



.

A graph $\mathcal{G}_1 = (\mathcal{V}_1, \mathcal{E}_1)$ is called a **subgraph** of $\mathcal{G}$ if $\mathcal{V} = \mathcal{V}_1$ and $\mathcal{E}_1 \subseteq \mathcal{E}$; we then write $\mathcal{G}_1 \leq \mathcal{G}$. If additionally, $\mathcal{E}_1 \neq \mathcal{E}$, then $\mathcal{G}_1$ is a **proper subgraph** of $\mathcal{G}$. A node $i$ is called a **parent** of $j$ if $(i, j) \in \mathcal{E}$ and $(j, i) \neq \mathcal{E}$ and a **child** if $(j, i) \in \mathcal{E}$ and $(i, j) \neq \mathcal{E}$. The set of parents of $j$ is denoted by $\mathbf{PA}_j^{\mathcal{G}}$, and the set of its children by $\mathbf{CH}_j^{\mathcal{G}}$.

**Example 2.26** (Parents and Children)**.** Let $\mathcal{V} = \{A, B, C\}$ and $\mathcal{E} = \{(A, B), (B, C), (A, C)\}$. Then $\mathcal{G}$ takes the form,

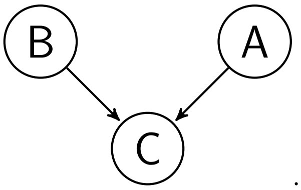


,

and $\mathbf{PA}_B^{\mathcal{G}} = \{A\}$, $\mathbf{PA}_C^{\mathcal{G}} = \{A, B\}$, $\mathbf{CH}_B^{\mathcal{G}} = \{C\}$ and $\mathbf{CH}_A^{\mathcal{G}} = \{C, B\}$.

Two nodes $i$ and $j$ are **adjacent** if either $(j, i) \in \mathcal{E}$ or $(i, j) \in \mathcal{E}$. We call $\mathcal{G}$ **fully connected** if all pairs of nodes are adjacent. We say that there is an **undirected edge** between two adjacent nodes $i$ and $j$ if $(j, i) \in \mathcal{E}$ and $(i, j) \in \mathcal{E}$. An edge between two adjacent nodes is **directed** if it is not undirected (notation: $i \to j$ for $(i, j) \in \mathcal{E}$). We call $\mathcal{G}$ **directed** if all its edges are directed. Three nodes are called an **immorality** or a **v-structure** if one node is a child of the two others that themselves are not adjacent.

**Example 2.27** (V-structure (immorality))**.** Let $\mathcal{V} = \{A, B, C\}$ and $\mathcal{E} = \{(B, C), (A, C)\}$. Then $\mathcal{G}$ takes the form,



.

The **skeleton** of $\mathcal{G}$ does not take the directions of the edges into account. This corresponds to a graph where $\mathcal{E}$ is augmented by the other direction $(j, i)$ of every edge $(i, j)$, if it is not an element already.

A **path** in $\mathcal{G}$ is a sequence of (at least two) distinct vertices $i_1, \dots, i_m$, such that there is an edge between $i_k$ and $i_{k+1}$ for all $k = 1, \dots, m-1$. If $i_{k-1} \to i_k$ and $i_{k+1} \to i_k$, $i_k$ is called a **collider relative to this path**. If $i_k \to i_{k+1}$ for all $k$, we speak of a **directed**

**path** from $i_1$ to $i_m$ and call $i_1$ an **ancestor** of $i_m$ and $i_m$ a descendant of $i_1$. The set of all ancestors of i is denoted by $\mathbf{AN}_i^{\mathcal{G}}$ and $i$ is not an ancestor of itself. Further, $i$ is neither a descendant nor a non-descendant of itself. We denote all descendants of $i$ by $\mathbf{DE}_i^{\mathcal{G}}$, and all non-descendants of $i$, excluding $i$ and including parents of $i$ in graph $\mathcal{G}$, by $\mathbf{ND}_i^{\mathcal{G}}$. A node without parents is called a **source node**, a node without children a **sink node**. A permutation $\pi$, that is a bijective function $\pi : \{1, \dots, d\} \to \{1, \dots, d\}$ is called a **topological** or **causal ordering** if it satisfies $\pi(i) < \pi(j)$ if $j \in \mathbf{DE}_i^{\mathcal{G}}$. A graph $\mathcal{G}$ is called a **partially directed acyclic graph (PDAG)** if there is no directed cycle, that is, if there is no pair $(j, k)$ with directed paths from $j$ to $k$ and from $k$ to $j$. $\mathcal{G}$ is called a **directed acyclic graph (DAG)** if it is a PDAG and all edges are directed.

**Definition 2.28** (Pearl's d-separation)**.** In a DAG $\mathcal{G}$, a path between nodes $i_1$ and $i_m$ is **blocked by a set S** (with neither $i_1$ nor $i_m$ in $\mathbf{S}$) whenever there is a node $i_k$, such that one of the following two possibilities holds:

(i) $i_k \in S$

$$\begin{aligned} & i_{k-1} \to i_k \to i_{k+1} \\ \text{or} \quad & i_{k-1} \leftarrow i_k \leftarrow i_{k+1} \\ \text{or} \quad & i_{k-1} \leftarrow i_k \to i_{k+1} \end{aligned}$$

(ii) neither $i_k$ nor any of its descendants is in $\mathbf{S}$, i.e., $(\{i_k\} \cup \mathbf{DE}_{i_k}) \cap \mathbf{S} = \emptyset$, and

$$i_{k-1} \to i_k \leftarrow i_{k+1}.$$

Furthermore, in a DAG $\mathcal{G}$, we say that two disjoint subsets of vertices $\mathbf{A}$ and $\mathbf{B}$ are **d-separated** by a third (also disjoint) subset $\mathbf{S}$ if every path between nodes in $\mathbf{A}$ and $\mathbf{B}$ is blocked by $\mathbf{S}$. We then write

$$\mathbf{A} \perp\!\!\!\perp_{\mathcal{G}} \mathbf{B} \mid \mathbf{S}.$$

**Definition 2.29** (Markov property)**.** Given a DAG $\mathcal{G}$ and a joint distribution $\mathbf{P_X}$, this distribution is said to satisfy the **(global) Markov property** with respect to the DAG $\mathcal{G}$ if

$$\mathbf{A} \perp\!\!\!\perp_{\mathcal{G}} \mathbf{B} \mid \mathbf{C} \Rightarrow \mathbf{A} \perp\!\!\!\perp \mathbf{B} \mid \mathbf{C}$$

for all disjoint vertex sets $\mathbf{A}, \mathbf{B}, \mathbf{C}$.

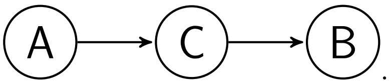
**Remark 2.30** (Equivalent formulations of the Markov property)**.** According to [31, Theorem 6.22] the above definition of the Markov property is equivalent to the following two definitions, if $\mathbf{P_X}$ has a density $p$.
Let $\mathcal{G}$ be a DAG and $\mathbf{P_X}$ be a joint distribution.

(i) $\mathbf{P_X}$ is **local Markov** w.r.t. $\mathcal{G}$ if each variable is independent of its non-descendants (without its parents) given its parents.

(ii) $\mathbf{P_X}$ satisfies the **Markov factorization property** w.r.t. $\mathcal{G}$ if

$$p(\mathbf{x}) = p(x_1, \ldots, x_d) = \prod_{j=1}^{d} p(x_j | \mathbf{pa}_j^{\mathcal{G}}).$$

**Example 2.31** (Illustrative example)**.** Consider $\mathcal{V} = \{\text{A,B,C}\}$ and $\mathcal{E} = \{(\text{A,C}), (\text{C,B})\}$, i.e., $\mathcal{G}$ takes the form,



It holds that $\text{A} \perp\!\!\!\perp_{\mathcal{G}} \text{B} \mid \text{C}$, since A and B are blocked by C. Assume $(\text{A,B,C}) \sim \mathbf{P_X}$ is Markov w.r.t $\mathcal{G}$, then we have $\text{A} \perp\!\!\!\perp \text{B} \mid \text{C}$. If $\mathbf{P_X}$ has a density $p$, this implies

$$\begin{aligned} p(a|c)p(b|c) &= p(a,b|c) = \frac{p(a,b,c)}{p(c)} \\ \Leftrightarrow\ p(a,b,c) &= p(a|c)p(b|c)p(c) \\ \Leftrightarrow\ p(a,b,c) &= p(a)p(c|a)p(b|c), \end{aligned}$$

i.e., the Markov factorization property. Further, in this toy example, the equivalence between the local and global Markov property is straight forward.

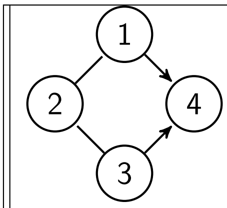
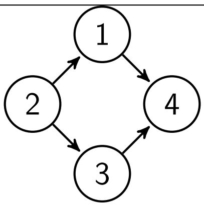
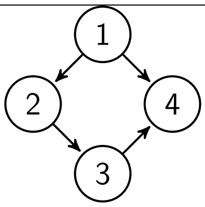
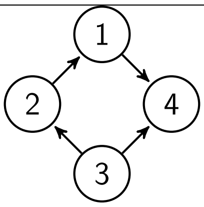
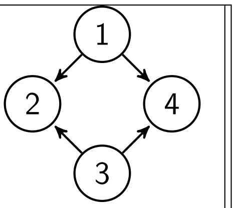
**Definition 2.32** (Markov equivalence of graphs)**.** We denote by $\mathcal{M}(\mathcal{G}) := \{\mathbf{P_X} : \mathbf{P_X}$ (global) Markov w.r.t. $\mathcal{G}\}$, the set of all distributions that are Markov with respect to $\mathcal{G}$. Two DAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ are **Markov equivalent** if $\mathcal{M}(\mathcal{G}_1) = \mathcal{M}(\mathcal{G}_2)$ and the set of all DAGs that are Markov equivalent to some DAG, $\mathcal{G}$, is called **Markov equivalence class** of $\mathcal{G}$.

**Remark 2.33.** Markov equivalence is defined in a way such that two graphs are Markov equivalent if and only if they satisfy the same set of $d$-separations, i.e., the same set of (conditional) independence conditions. This characterisation can become arbitrarily tedious to check. Verma and Pearl provide a convenient alternative characterisation in [48, Theorem 1], which is quoted as a lemma below.

**Lemma 2.34** (Characterisation of Markov equivalence)**.** Two DAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ are Markov equivalent if and only if they have the same skeleton and the same v-structures.

**Remark 2.35.** An Markov equivalence class can be uniquely represented by a completed PDAG. This graph has the property that $(i,j) \in \mathcal{E}$ if and only if one member of the Markov equivalence class has the same edge.

**Example 2.36** (Illustrative example)**.** In the below table all blue colored graphs belong to the same Markov equivalence class. The red colored DAG is not in this Markov equivalence class.

| CPDAG | DAG 1 | DAG 2 | DAG 3 | DAG 4 |
|---|---|---|---|---|
|  |  |  |  |  |

The Markov property is a very common assumption and one key success factor of DAG models in causal inference. It enables us to efficiently exploit (conditional) independencies of the distribution. In this work we will employ structural causal models for causal discovery, which imply the Markov property directly. They will be introduced just after this section.
We need one more assumption such that every directed edge in our graph encodes a causality in a meaningful way. The reason is that typically any distribution $\mathbf{P_X}$ is Markov w.r.t. all fully connected acyclic graphs, i.e., we are in danger to have a bunch of redundant directed edges in our graph. In earlier days, efficient algorithms were proposed which use the notion of faithfulness to resolve this issue. More recent approaches typically assume causal minimality because it is a weaker assumption than faithfulness and at the same time sufficient for the purpose of causal inference. We state here both definitions following [31, Definition 6.33].

**Definition 2.37** (Faithfulness and causal minimality)**.** Consider a distribution $\mathbf{P_X}$ and a DAG $\mathcal{G}$.

(i) $\mathbf{P_X}$ is **faithful** to the DAG $\mathcal{G}$ if

$$\mathbf{A} \perp\!\!\!\perp \mathbf{B} \mid \mathbf{C} \Rightarrow \mathbf{A} \perp\!\!\!\perp_{\mathcal{G}} \mathbf{B} \mid \mathbf{C}.$$

(ii) A distribution satisfies **causal minimality** w.r.t. $\mathcal{G}$ if it is Markovian w.r.t. $\mathcal{G}$, but not to any proper subgraph of $\mathcal{G}$.

We state the most important results about causal minimality taken from [31, Section 6.5.3] below and then end this section.

**Lemma 2.38.** Any two nodes in a DAG $\mathcal{G}$ that are not directly connected by an edge can be $d$-separated.

*Proof.* Let $\mathcal{G}$ be a DAG and $i$ and $j$ nodes that are not directly connected by an edge. Assume we cannot $d$-separate the two nodes. If there is no path between the nodes they can be $d$-separated by the empty set. Therefore, there must be at least one path from $i$ to $j$ with at least one node in between $h$. The only possibility we can not $d$-separate by $h$ is, if it is a collider in one path and it satisfies one of the possibilities in [2.28, $(i)$]. But this would imply a cyclic structure what is a contradiction. □

**Lemma 2.39** (Faithfulness implies causal minimality)**.** If $\mathbf{P_X}$ is faithful and Markovian w.r.t. $\mathcal{G}$, then causal minimality is satisfied.

*Proof.* If $\mathbf{P_X}$ is Markovian w.r.t. a proper subgraph $\tilde{\mathcal{G}}$ of $\mathcal{G}$, there are two nodes that are directly connected in $\mathcal{G}$ but not in $\tilde{\mathcal{G}}$. Thus, they can be $d$-separated in $\tilde{\mathcal{G}}$ but not in $\mathcal{G}$. The Markov condition implies a corresponding conditional independence statement in $\mathbf{P_X}$ but then $\mathbf{P_X}$ can not be faithful w.r.t. $\mathcal{G}$. □

**Lemma 2.40** (Equivalence of causal minimality)**.** Consider the random vector $\mathbf{X} = (X_1, ..., X_p)$ and assume that the joint distribution has a density w.r.t. a product measure. Suppose that $\mathbf{P_X}$ is Markovian w.r.t. $\mathcal{G}$. Then $\mathbf{P_X}$ satisfies causal minimality w.r.t. $\mathcal{G}$ if and only if $\forall X_j \forall Y \in \mathbf{PA}_j^{\mathcal{G}}$ we have that $X_j \perp\!\!\!\perp \mathbf{PA}_j^{\mathcal{G}} \setminus \{Y\}$.

*Proof.* See [31, Appendix C.6]. □

### 2.5.2. Structural Causal Models (SCMs)

SCMs provide a framework which enables us to formalize causal discovery and causal learning. They entail among others

- an observational distribution,
- a causal graph,
- and an intervention distribution.

A good introduction is provided in [31, Section 6.4]. We already covered the graphical part but must connect it to SCM still. Intervention distributions are discussed in great detail in the next section.
For the definition of SCMs we quote [31, Definition 6.2] and also present the example provided therein. Then, the most important results for SCMs are presented.

**Definition 2.41** (Structural Causal Models)**.** A **structural causal model (SCM)** $\mathcal{C} := (\mathbf{S}, \mathbf{P_N})$ consists of a collection $\mathbf{S}$ of $d$ **(structural) assignments**

$$X_j := f_j(\mathbf{PA}_j, N_j), \qquad j = 1, \ldots, d, \tag{2.27}$$

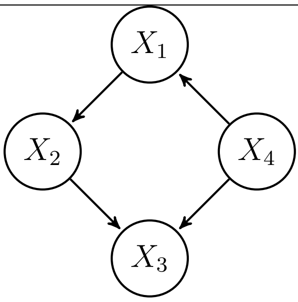
where $\mathbf{PA}_j \subseteq \{X_1, \ldots, X_d\} \setminus \{X_j\}$ are called **parents of** $X_j$; and a joint distribution $\mathbf{P_N} = \mathbf{P}_{N_1,\ldots,N_d}$ over the noise variables, which we require to be jointly independent; that is, $\mathbf{P_N}$ is a product distribution.
The graph $\mathcal{G}$ of an SCM is obtained by creating one vertex for each $X_j$ and drawing edges from each parent $\mathbf{PA}_j$ to $X_j$, that is, from each variable $X_k$ occurring on the right-hand side of equation (2.27) to $X_j$. We henceforth assume this graph to be acyclic. We sometimes call the elements of $\mathbf{PA}_j$ not only parents but also **direct causes** of $X_j$, and we call $X_j$ a **direct effect** of each of its direct causes.

In order to avoid unintuitive structural assignments like $X_2 := 0 \times X_1 + N_{X_2}$, we add the requirement of **structural minimality**. That is, whenever there is a random variable with index $k \in \{1, \ldots, k\}$ and a function $g$ such that

$$f_k(\mathbf{pa}_k, n_k) = g(\mathbf{pa}_k^*, n_k), \qquad \forall \mathbf{pa}_k, \forall n_k \text{ with } p(n_k) > 0, \tag{2.28}$$

where $\mathbf{PA}_k^* \subsetneq \mathbf{PA}_k$, we choose the latter representation. This is not a restrictive assumption, since there is a unique representation in which each function has a minimal number of inputs [31, Remark 6.6].

**Example 2.42** (Illustrative Example)**.** An example to illustrate the connection of an SCM to its graph $\mathcal{G}$.

| SCM | Graph $\mathcal{G}$ | Assumptions of SCM |
|---|---|---|
| $X_1 := f_1(X_3, N_1)$<br>$X_2 := f_2(X_1, N_2)$<br>$X_3 := f_3(N_3)$<br>$X_4 := f_4(X_2, X_3, N_4)$ |  | • $N_1, \ldots, N_4$ jointly independent<br>• $\mathcal{G}$ is acyclic |

We can efficiently sample from an SCM, by generating an i.i.d. sample of $\mathbf{N}^1, \ldots, \mathbf{N}^n \sim \mathbf{P_N}$ and then subsequently use the structural assignments to generate an i.i.d. sample of the joint distribution of all variables. The following lemma quotes [31, Prop. 6.3].

**Lemma 2.43** (Entailed distribution)**.** A SCM $\mathcal{C}$ defines a unique distribution over the variables $X_1, \ldots, X_d$: any $X_1, \ldots, X_d, N_1, \ldots, N_d$ satisfying $X_j = f_j(\mathbf{PA}_j, N_j)$ almost surely, where $(N_1, \ldots, N_d)$ has the desired distribution, induce the same distribution over $\mathbf{X} = (X_1, \ldots, X_d)$. We refer to it as the entailed distribution $\mathbf{P}_\mathbf{X}^\mathcal{G}$ and sometimes write $\mathbf{P_X}$.

*Proof.* See [31, Appendix C.2.] □

The next result relates SCMs to their graphs. It states [29, Theorem 1.4.1].

**Theorem 2.44.** *The law $\mathbf{P_X}$ generated by an SCM with graph $\mathcal{G}$ is Markov w.r.t. its graph.*

Lastly we need to know how distributions, observed in practice, relate to our modeling approach. The following result is taken from [32, Prop. 9].

**Lemma 2.45.** Consider $X_1, \ldots, X_p$ and let $\mathbf{P_X}$ have a strictly positive density w.r.t. the Lebesgue measure and assume it is Markov w.r.t. $\mathcal{G}$. Then there exists an SEM with graph $\mathcal{G}$ that generates $\mathbf{P_X}$.

*Proof.* Let $N_1, \ldots, N_p \sim \mathcal{U}[0, 1]$ and define $X_j = f_j(\mathbf{PA}_j, N_j)$ with

$$f_j(x_{\mathbf{PA}_j}, n_j) = F^{\leftarrow}_{X_j|\mathbf{PA}_j = x_{\mathbf{PA}_j}}(n_j),$$

where $F^{\leftarrow}_{X_j|\mathbf{PA}_j = x_{\mathbf{PA}_j}}$ is the generalized inverse of the distribution function of $X_j$ given $\mathbf{PA}_j$. The assertion follows from the probability integral transform. □

### 2.5.3. Additive Noise Models (ANMs)

Recall (2.6). If we assume a normally distributed additive noise for the regression model we have a closed form for the marginal likelihood, what tremendously improves numerical tractability. This motivates the use use of additive noise models for our approach.

**Definition 2.46** (Additive Noise Model)**.** An **additive noise model (ANM)** is a tuple $(\mathbf{S}, \mathbf{P_N})$ that consists of a collection $\mathbf{S}$ of $d$ equations

$$X_j := f_j(\mathbf{PA}_j) + N_j, \qquad j = 1, \dots, d, \tag{2.29}$$

where $\mathbf{P_N}$ is a product distribution and the corresponding graph is acyclic.

It can be shown that for such models causal minimality reduces to the condition that each function $f_j$ is not constant in any of its arguments [32, Proposition 17]. With the additive noise assumption we can infer the underlying DAG from purely observational data in the majority of the cases [32, Theorem 28]. But there are rare non-identifiable cases where it is possible to find a backward model [32, Proposition 23]. Next, we define what we mean with backward model for the bivariate case and then state one example that is non-identifiable what will serve as example to test the approach upon.

**Definition 2.47** ((Bivariate) Backward Model)**.** Let $\mathcal{C}$ be an ANM with $X_1 := N_1$ and $X_2 := f_2(X_1) + N_2$. We denote by $p_{N_1}$ the probability density of $N_1$ and use the same notation for all other random variables. Assume that all densities are well defined densities on $\mathbb{R}$. The joint distribution of $\mathcal{C}$ factorizes, $p_{\mathcal{C}}(x_1, x_2) = p_{N_1}(x_1)p_{N_2}(x_2 - f_2(x_1))$. We call the ANM $\tilde{\mathcal{C}}$, with $X_2 := \tilde{N}_2$ and $X_1 := g_1(X_2) + \tilde{N}_1$, a (valid) **backward model** if it satisfies $p_{\mathcal{C}}(x_1, x_2) = p_{\tilde{N}_2}(x_2)p_{\tilde{N}_1}(x_1 - g_1(x_2)) = p_{\tilde{\mathcal{C}}}(x_1, x_2)$.

**Example 2.48.** Consider $X_j = f_j(X_i) + N_j$ with fully supported noise variable $N_j$ that is independent of $X_i$. If $X_i$ is Gaussian, $N_j$ is Gaussian and $f_j$ is linear we can find a valid backward model, $X_i = g_i(X_j) + M_i$ with $M_i$ independent of $X_j$. Thus, it is impossible to infer a causal direction between the variables from purely observational data.

*Proof.* Since $X_j$ is a sum of two independent Gaussians it is also Gaussian, therefore we can consider $X_j$ and $X_i$ as elements of the Hilbert space $L^2(\mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda)$. Since $X_i$ is a closed convex subspace of $L^2(\mathbb{R})$ there exists a unique orthogonal projection $P$ that maps $X_j$ onto $X_i$. We have that $X_i = PX_j + X_i - PX_j$ where $\langle PX_j, X_i - PX_j \rangle_{L^2} = 0$. Since uncorrelatedness implies independence for Gaussians we can always find a valid backward model. □

Consider an explicit example. Let $X_1 \sim \mathcal{N}(0, \sigma^2)$ and $X_2 := aX_1 + N_2$ with $N_2 \sim \mathcal{N}(0, \tau^2)$, $X_1 \perp\!\!\!\perp N_2$, and $\sigma^2, \tau^2, a \in \mathbb{R}$ be the true data generating process. Then, for $\tilde{a} = \frac{a\sigma^2}{(\tau^2 + a^2\sigma^2)}$ we have $\tilde{a}X_2 \sim \mathcal{N}(0, \frac{a^2\sigma^4}{\tau^2 + a^2\sigma^2})$, $X_1 - \tilde{a}X_2 \sim \mathcal{N}(0, (1 - \tilde{a}a)^2\sigma^2 + \tilde{a}^2\tau^2)$, $\mathbb{E}[\tilde{a}X_2(X_1 - \tilde{a}X_2)] = 0$ and $\tilde{a}X_2 + (X_1 - \tilde{a}X_2) \sim \mathcal{N}(0, \sigma^2)$. Thus, $X_2 \sim \mathcal{N}(0, \tau^2 + a^2\sigma^2)$ and $X_1 := \tilde{a}X_2 + N_1$ with $N_1 \sim \mathcal{N}(0, \frac{\sigma^2\tau^2}{\tau^2 + a^2\sigma^2})$ defines a valid backward model for the true data generating distribution.

## 2.6. Interventions

Observing enough interventions enables us to exactly identify the causal directions between random variables [15, 16]. Though, it is sometimes only a theoretical concept (for example we cannot change the blood type of a patient), there are many cases where we can perform interventions which may provide us with necessary information to do causal inference in a meaningful way. The purpose of this section is to provide a summary on definitions and results about intervention distributions, total causal effects and methods to calculate them given we know the SCM. We follow [31, Sections 6.3, 6.6, 6.7] and quote the definitions and lemmas thereof.

### 2.6.1. Intervention Distribution

**Definition 2.49** (Intervention distribution)**.** Consider a SCM $\mathcal{C} := (\mathbf{S}, \mathbf{P_N})$ and its entailed distribution $\mathbf{P}_{\mathbf{X}}^{\mathcal{G}}$. We replace one (or several) of the structural assignments to obtain a new SCM $\tilde{\mathcal{C}}$. Assume that we replace the assignment for $X_k$ by

$$X_k := \tilde{f}(\tilde{\mathbf{PA}}_k, \tilde{N}_k).$$

We then call the entailed distribution of the new SCM an intervention distribution and say that the variables whose structural assignment we have replaced have been **intervened** on. We denote the new distribution by

$$\mathbf{P}_{\mathbf{X}}^{\tilde{\mathcal{C}}} =: \mathbf{P}_{\mathbf{X}}^{\mathcal{C};do(X_k:=\tilde{f}(\tilde{\mathbf{PA}}_k,\tilde{N}_k))}.$$

The set of noise variables in $\tilde{\mathcal{C}}$ now contains both some "new" $\tilde{N}'s$ and some "old" $N$'s, all of which are required to be jointly independent.
When $\tilde{f}(\tilde{\mathbf{PA}}_k, \tilde{N}_k)$ puts a point mass on a real value $a$, we simply write $\mathbf{P}_{\mathbf{X}}^{\mathcal{C};do(X_k:=a)}$ and call this an **atomic** (ideal) intervention. An intervention with $\tilde{\mathbf{PA}}_k = \mathbf{PA}_k$, that is, where direct causes remain direct causes, is called **imperfect**. We require that the new SCM $\tilde{\mathcal{C}}$ have an acyclic graph; the set of allowed interventions thus depends on the graph induced by $\mathcal{C}$.

An (perfect) intervention on $X_k$ graphically corresponds to remove all incoming edges to $X_k$ from the graph. We say that a total causal effect form $X$ to $Y$ exists, if we can measure dependence between the two random variables after intervening on $X$.

**Definition 2.50** (Total causal effect)**.** Given an SCM $\mathcal{C}$, there is a total causal effect from $X$ to $Y$ if and only if

$$X \not\perp\!\!\!\perp Y \qquad \text{in } \mathbf{P}_{\mathbf{X}}^{\mathcal{C};do(X:=\tilde{N}_X)}$$

for some random variable.

**Lemma 2.51** (Characterisation of total causal effect)**.** Given an SCM $\mathcal{C}$, the following statements are equivalent:

(i) There is a total causal effect from $X$ to $Y$.

(ii) There are $x^{\triangle}$ and $x^{\square}$ such that $\mathbf{P}_Y^{\mathcal{C};do(X:=x^{\triangle})} \neq \mathbf{P}_Y^{\mathcal{C};do(X:=x^{\square})}$.

(iii) There is $x^{\triangle}$ such that $\mathbf{P}_Y^{\mathcal{C};do(X:=x^{\triangle})} \neq \mathbf{P}_Y^{\mathcal{C}}$.

(iv) $X \not\!\perp\!\!\!\perp Y$ in $\mathbf{P}_{X,Y}^{\mathcal{C};do(X:=\tilde{N}_X)}$ for any $\tilde{N}_X$ whose distribution has full support.

*Proof.* See [31, Appendix  C.4.]. □

**Lemma 2.52** (Graphical criteria for total causal effects)**.** Assume we are given a SCM $\mathcal{C}$ with corresponding graph $\mathcal{G}$.

(i) If there is no directed path from $X$ to $Y$, then there is no total causal effect.

(ii) Sometimes there is a directed path but no total causal effect.

*Proof.* (i) See [31, Appendix  C.5.].

(ii) Consider the SCM $X_1 := N_1, X_2 := aX + N_2, X_3 := bX_2 + cX + N_3$, for $a, b, c \in \mathbb{R}$ with Gaussian noise. If $ab + c = 0$, then we have $X_1 \perp\!\!\!\perp X_3$. Further, in every intervention distribution, where we perfectly intervene on $X_1$, we also have $X_1 \perp\!\!\!\perp X_3$. Thus, there is no total causal effect from $X_1$ to $X_3$.
Note that, because we require structural minimality, this can only happen if there is more than one directed path between the respective nodes.

□

### 2.6.2. Calculating Intervention Distributions

Given a SCM $\mathcal{C}$, and writing $pa(j) := \mathbf{PA}_j^{\mathcal{G}}$, we have that

$$X_j | X_{pa(j)} = f^{(j)}(x_{pa(j)}, N_j) \equiv f^{(j)}(N_j).$$

For any SCM $\tilde{\mathcal{C}}$ that is constructed from $\mathcal{C}$ by intervening on (some) $X_k$ but not on $X_j$, the parents and the noise of $X_j$ are the same in $\tilde{\mathcal{C}}$ and $\mathcal{C}$. Therefore all these SCMs $\tilde{\mathcal{C}}$ satisfy a very powerful invariance statement,

$$p^{\tilde{\mathcal{C}}}(x_j | x_{pa(j)}) = p^{\mathcal{C}}(x_j | x_{pa(j)}). \tag{2.30}$$

Now consider $\tilde{\mathcal{C}} := \mathcal{C}; do(X_k := \tilde{N}_k)$, where $\tilde{N}_k$ allows for a density $\tilde{p}$. The Markov property together with (2.30) yields

$$\begin{aligned} p^{\mathcal{C};do(X_k:=\tilde{N}_k)}(x_1, \dots, x_d) &= \prod_{j \neq k} p^{\mathcal{C};do(X_k:=\tilde{N}_k)}(x_j | x_{pa(j)}) p^{\mathcal{C};do(X_k:=\tilde{N}_k)}(x_k) \\ &= \prod_{j \neq k} p^{\mathcal{C}}(x_j | x_{pa(j)}) \tilde{p}(x_k). \end{aligned} \tag{2.31}$$

This result allows us to compute an interventional statement (left-hand side) from observational quantities (right-hand side). This is a very important result, which became known under three different names: G-computation formula [37], truncated factorization [27] and manipulation theorem [40].

In later considerations we will mainly use the following special case of ideal intervention,

$$p^{\mathcal{C};do(X_k:=a)}(x_1,\ldots,x_d) = \begin{cases} \prod_{j\neq k} p^{\mathcal{C}}(x_j|x_{pa(j)}) & \text{if } x_k = a \\ 0 & \text{otherwise.} \end{cases} \tag{2.32}$$

# 3. Active Bayesian Causal Discovery

Motivated by the active Bayesian approach, that uses interventional evidence for causal discovery described in [49], we consider a setting where we have

- a low number of initial observations and
- the possibility to perform (perfect) interventions on every variable in the system under consideration.

In all numerical experiments we chose five initial data points and include the (optimal) intervention data points in a sequential way. The joint distribution factorizes the following way

$$p(x_{\text{init.}}, x^1_{\text{inter.}}, \dots, x^n_{\text{inter.}}) = p(x_{\text{init.}})p(x^1_{\text{inter.}}|x_{\text{init.}}) \cdots p(x^n_{\text{inter.}}|x_{\text{init.}}, x^1_{\text{inter.}}, \dots, x^{n-1}_{\text{inter.}}). \tag{3.1}$$

Since optimal interventional data depends on the initial observations, joint distributions would become very complex if we do not assume sequentiality. First we consider a summary of the approach proposed in [49] and sketch a possible pseudo code. Then we investigate the bivariate case in much detail and finally turn to the fourvariate case, where we state the calculation steps needed and point out computational challenges. We conclude this chapter with heuristic proposals to overcome these computational challenges.

## 3.1. General Approach

### Problem setting

Consider a SCM over a set of $d$ real-valued observable variables $\mathbf{X} = \{X_1, \dots, X_d\} \sim \mathbf{P_X}$ with a corresponding DAG $G^*$, i.e., we have structural assignments of the form

$$X_j := f_j(\mathbf{PA}_j^{G^*}, N_j), \qquad j = 1, \dots, d, \tag{3.2}$$

where the noise distribution, $\mathbf{P_N} = \mathbf{P}_{N_1, \dots, N_d}$, is jointly independent and $\mathbf{P_X}$ factorises according to the DAG $G^*$. Further, we require structural minimality.
We restrict ourselves to ANMs, where we additionally assume that the structural assignments are of the form

$$X_j := f_j(\mathbf{PA}_j^{G^*}) + N_j, \qquad j = 1, \dots, d. \tag{3.3}$$

This assumption is also crucial to have closed forms of the likelihoods. But we here additionally gain that causal minimality reduces to the condition that each function $f_j$ is not constant in any of its arguments [32, Prop. 17] and identifiability based on purely observational data, if the conditioned bivariate sub-models satisfy a certain Ordinary Differential Equation [32, Theorem 28].

### Active Bayesian causal discovery

Let $\mathcal{G}$ be the set of all DAGs over $d$ variables and $\mathbf{P}_G$ a prior distribution over possible causal graphs $\tilde{G} \in \mathcal{G}$. We denote the parameters of each graph with $\theta_{\tilde{G}} \in \Theta_G$ and place another prior distribution $\mathbf{P}_{\theta_G|G}$ over graph parameters given the graph. The pair $(G, \theta_G)$ encodes a causal model which describes how to generate data $\tilde{\mathbf{D}} \in \mathcal{X}$ and we define the likelihood function in the causal model by $p_{\mathbf{D}|G,\theta_G}(\tilde{\mathbf{D}}|\tilde{G}, \theta_{\tilde{G}})$. According to the law of total probability, the evidence of $\tilde{G}$ after observing data $\tilde{\mathbf{D}}$ (marginal likelihood) is then given by

$$p_{\mathbf{D}|G}(\tilde{\mathbf{D}}|\tilde{G}) = \int_{\Theta_G} p_{\mathbf{D}|G,\theta_G}(\tilde{\mathbf{D}}|\tilde{G}, \theta_{\tilde{G}}) p_{\theta_G|G}(\theta_{\tilde{G}}|\tilde{G}) \, d\theta_{\tilde{G}} \tag{3.4}$$

and the posterior distributions over graphs $G$ and parameters are respectively given by

$$\mathbf{P}_{G|\mathbf{D}} \propto \mathbf{P}_G \mathbf{P}_{\mathbf{D}|G}, \qquad \text{and} \qquad \mathbf{P}_{\theta_G|\mathbf{D},G} \propto \mathbf{P}_{\theta_G|G} \mathbf{P}_{\mathbf{D}|\theta_G,G}. \tag{3.5}$$

In our approach, we want to perform the one intervention of the form $do(X_j = x)$, that gains us most information about $(G, \theta_G)$ from observing the remaining variables (denoted by) $\mathbf{X}_{-j}$. In order to achieve this, we turn to Bayesian Experimental design [21, 8], where we select an experiment, $do(X_j = x)$, aiming to maximise a given utility function $U(\mathbf{X}_{-j}|do(X_j = x))$. Given the current model specified by a prior $\mathbf{P}_{(G,\theta_G)}$ and a likelihood $\mathbf{P}_{\mathbf{X}_{-j}|(G,\theta_G),do(X_j=x)}$, the optimal experiment is the one which maximises expected utility,

$$(j^*, x^*) = \operatorname*{argmax}_{j \in \{1,\dots,d\},\ x \in \mathcal{X}_j} \int U(\mathbf{x}_{-j}|do(X_j = x)) p_{\mathbf{X}_{-j}|do(X_j=x)}(\mathbf{x}_{-j}|x) \, d\mathbf{x}_{-j}. \tag{3.6}$$

As utility we use the information gain in $(G, \theta_G)$ from performing $do(X_j = x)$ and observing $\mathbf{X}_{-j}$,

$$\begin{aligned} U(\mathbf{X}_{-j}|do(X_j = x)) = &\int_{(\mathcal{G},\Theta_G)} p_{(G,\theta_G)|\mathbf{X}_{-j},do(X_j=x)} \log(p_{(G,\theta_G)|\mathbf{X}_{-j},do(X_j=x)}) \, d(\tilde{G}, \theta_{\tilde{G}}) \\ &- \int_{(\mathcal{G},\Theta_G)} p_{(G,\theta_G)} \log(p_{(G,\theta_G)}) \, d(\tilde{G}, \theta_{\tilde{G}}). \end{aligned}$$

Since the ladder part is independent of $do(X_j = x)$ it can be disregarded in optimization. Because we are interested in learning the causal structure, we use the information gain in $G$ as utility, i.e.,

$$U(\mathbf{X}_{-j}|do(X_j = x)) = \sum_{\tilde{G} \in \mathcal{G}} p_{G|\mathbf{X}_{-j},do(X_j=x)} \log(p_{G|\mathbf{X}_{-j},do(X_j=x)}). \tag{3.7}$$

If we plug (3.7) in (3.6) and rearrange terms using Bayes theorem we arrive at

$$\begin{aligned}(j^*, x^*) &= \underset{j\in\{1,\dots,d\},\; x\in\mathcal{X}_j}{\operatorname{argmax}} \sum_{\tilde{G}\in\mathcal{G}} p_G \int p_{\mathbf{x}_{-j}|G,do(X_j=x)} \log(p_{G|\mathbf{x}_{-j},do(X_j=x)})\, d\mathbf{x}_{-j} \\ &= \underset{j\in\{1,\dots,d\},\; x\in\mathcal{X}_j}{\operatorname{argmax}} \mathbb{E}_G \left[\mathbb{E}_{\mathbf{x}_{-j}|G,do(X_j=x)} \left[\log\left(p_{G|\mathbf{x}_{-j},do(X_j=x)}(G|\mathbf{x}_{-j}, do(X_j=x))\right)\right]\right].\end{aligned} \tag{3.8}$$

If we employ (3.3), use GPs as priors over the functions $f_i$ and assume $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$, the marginal likelihood (3.4) and posteriors (3.5) are available in closed form. Thus we can efficiently sample $\mathbf{x}_{-j}$ from the interventional distribution $\mathbf{P}_{\mathbf{x}_{-j}|G,do(X_j=x)}$ implied by a graph $G$. This enables us to use a Monte Carlo estimator for our objective,

$$(j^*, x^*) = \underset{j\in\{1,\dots,d\},\; x\in\mathcal{X}_j}{\operatorname{argmax}} \sum_{\tilde{G}\in\mathcal{G}} p_G(\tilde{G}) \frac{1}{M} \sum_{m=1}^{M} \log(p_{G|\mathbf{x}_{-j},do(X_j=x)}(\tilde{G}|\mathbf{x}_{-j}^{(m)}, do(X_j = x))). \tag{3.9}$$

To solve the optimization problem it is very convenient to use a derivative free optimization algorithm such as Bayesian Optimization. It is designed for situations, where the objective function is costly to evaluate and the gradient is not available.

### Pseudo Code

The following pseudo code is rather coarse and leaves out many calculation steps of the procedure. It summarizes the most important steps and supplements the previous section in terms of an algorithmic view on the proposed approach.

**Algorithm 2** Causal Discovery with Optimal Interventions

```
Require: $\mathcal{D}$, $G_{\text{list}}$, $n_{\text{interventions}}$        ▷ data, list of graphs with corr. parameters
  $\mathcal{D}_0 = \mathcal{D}$
  $d$ = number of nodes
  Compute $p(\mathcal{D}_0|G)$ for all $G$ in $G_{\text{list}}$ and store.
  Compute $p(G|\mathcal{D}_0)$ for all $G$ in $G_{\text{list}}$ and store.
  for $i = 1, \dots, n_{\text{interventions}}$ do
      $(j^*, x_{j^*})$ = optimal intervention given $\mathcal{D}_{i-1}$        ▷ perform Bayesian Opt.
      Get sample $D_i = (x_{j^*}, \mathbf{x}_{-j^*})$        ▷ perform intervention experiment
      Compute $p(D_i|\mathcal{D}_{i-1}, G)$ for all $G$ in $G_{\text{list}}$ and store        ▷ likelihood based on
                                                                    estimates of $\mathcal{D}_{i-1}$
      Store $p(D_i, \mathcal{D}_{i-1}|G) = p(D_i|\mathcal{D}_{i-1}, G)p(\mathcal{D}_{i-1}|G)$ for all $G$ in $G_{\text{list}}$
      Compute $p(G|\mathcal{D}_i)$ for all $G$ in $G_{\text{list}}$ and $i = 0, 1, \dots, n_{\text{interventions}}$ and store
      $\mathcal{D}_i = \{D_i, \mathcal{D}_{i-1}\}$
  end for
  return $p(G|\mathcal{D}_i)$ for all $G$ in $G_{\text{list}}$ and $i = 0, 1, \dots, n_{\text{interventions}}$; $\mathcal{D}_i$ for $i =$
  $1, \dots, n_{\text{interventions}}$
```

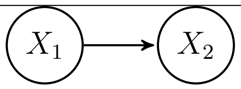
## 3.2. Bivariate Case

At first we consider the bivariate setting, where we do perform random interventions that are independent of the initial observations. This setup is close to the numerical example given in [49] and produces very similar results. It is an interesting "base case", because it already works very well and extremely efficient with respect to computation time. In the second section we will then consider the necessary theory for the case of optimal interventions, discuss and interpret numerical properties from it. In this work we always assume to have no prior believes about the true underlying DAG and therefore model the DAG prior using a uniform distribution. Further, we always use $\gamma = 1.75$ as parameter for the Matérn kernel, because it implies reasonable fits for our examples. The implementation is coded in Python and uses the "scikit-learn" and the "scikit-optimize" package.

### 3.2.1. Methodology for Numerical Experiments

In the numerical experiments presented in this thesis we always assume that we have chosen model parameters for error variances and kernel parameters. The implementation of the procedure allows to choose different variance parameters for all error terms in all possible graphs. If we assume "known variance" (because we have a known measurement error of the random variables) we must use these known variances in all possible ANMs (DAGs). When we assume the variances to be known we will mainly work with the following example, that assumes equal error variances. This assumption is discussed (in a different context) for example in [9, 30].

**Example 3.1** (Known Variance)**.** Assume we are given five i.i.d. observations of the true data generating process,

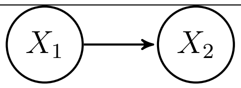
| | Graph | ANM |
|---|---|---|
| $G_{true}$ |  | $X_1 := \epsilon_1,$<br>$X_2 := 2\tanh(X_1) + \epsilon_2,$ |

where $\epsilon_1 \sim \mathcal{N}(0, 1)$ and $\epsilon_2 \sim \mathcal{N}(0, 1)$. And we choose the true variances in every possible graph.

If we do not know the variances and only have some initial observations, it is not clear how to come up with the model parameters. We could estimate sample variances for source nodes and we could estimate residual variances for child nodes. But methodologically this is rather difficult, because we would define the ANMs based on our data. What would be the most appropriate methodology in the case of unknown variance depends on the specific use case.

The implementation was tested in many different scenarios and is able to detect the true causal relation with great reliability. We will consider numerical examples with parameter choices that are, from a statistical point of view, not methodologically appropriate, but, from a numerical point of view, insightful to understand the behavior of the implemented algorithm. Below we introduce the example we will use for this purpose.

**Example 3.2.** Assume we are given five i.i.d. observations of the true data generating process,

| | Graph | ANM |
|---|---|---|
| $G_{true}$ |  | $X_1 := \epsilon_1,$<br>$X_2 := 2\tanh(X_1) + \epsilon_2,$ |

where $\epsilon_1 \sim \mathcal{N}(0,1)$ and $\epsilon_2 \sim \mathcal{N}(0,0.1)$. In the support area of [-1,1] the $2\tanh$-function is very close to a linear relation of the two variables. We have seen, that a linear relation in Gaussian ANMs always allows a valid backward model (2.48). Thus, this example is a hard problem for algorithms that infer causal directions from purely observational data. It will become even more difficult if we sample $X_1$ from a Gaussian that is more concentrated around zero than a standard Gaussian.
Let $G_1 = G_{true}$, $G_2$ be the graph where the edge is reversed and $G_3$ be the empty graph. In the implementation we must specify a variance for $X_1$ and $X_2$ in all three graphs. For $G_1$ we take the true parameters. Since the $2\tanh$-function is very linear in the support area of [-1,1] with a slope of 2, we can compute an approximate backward model for this area. If we would choose the parameters for $G_2$ according to this backward model and if we would only have sampling points where $X_1 \in [-1,1]$, the posterior probabilities of $G_1$ resp. $G_2$ should be very similar and we would have two very plausible competing models (for the initial data). It will be interesting to see what the most informative intervention is in these cases, because in principle $G_2$ can be cast out due to the fact that the observed values of $X_1$ from interventions on $X_2$ are not having the variance specified in $G_2$.

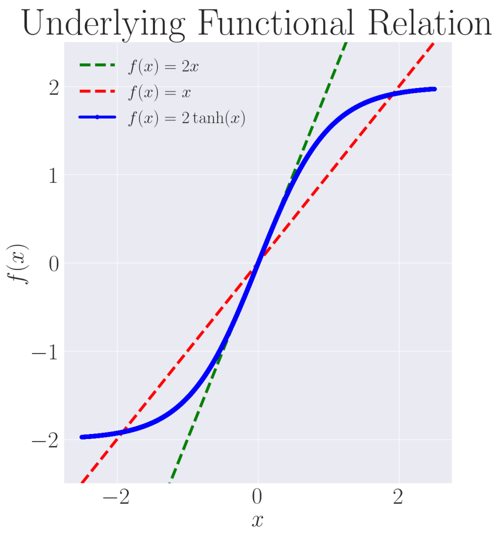


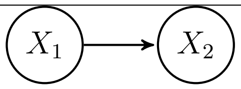
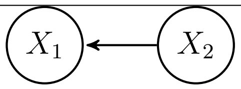
Figure 3.1.: Plot of $2\tanh$ fct. and linear approx.. (Example (3.2).

This motivates to approximate the true ANM by the following linear Gaussian ANM, $X_1 \sim \mathcal{N}(0,1)$ and $X_2 := X_1 + N_2$ with $N_2 \sim \mathcal{N}(0,0.1)$. The corresponding backward ANM (2.48) is $\tilde{X}_2 \sim \mathcal{N}(0,1.1)$ and $\tilde{X}_1 := \frac{10}{11}\tilde{X}_2 + \tilde{N}_1$ with $\tilde{N}_1 \sim \mathcal{N}(0,1/11)$. Thus, when we choose $\epsilon_1 \sim \mathcal{N}(0,0.1)$ and $\epsilon_2 \sim \mathcal{N}(0,1)$ in $G_2$ we define a very plausible, competitive hypothesis about the true causal relation (for the initial data). In $G_3$ we choose both variances to be $\mathcal{N}(0,1)$.

We will always present numerical results for (3.1) first and if it is more illustrative we will present numerical results for (3.2) and study the behavior of the implementation thereof.

### 3.2.2. Causal Discovery with Random Independent Interventions

Let $x_1$ and $x_2$ denote vectors of $N$ i.i.d. observations of a bivariate ANM, where the noise terms are mutually independent. We assume acyclicity and causal sufficiency. The task of causal discovery comes down to decide between two models (listed in the table below).

| | Graph | ANM |
|---|---|---|
| $G_1$ |  | $X_1 := \epsilon_1,$<br>$X_2 := f^{(2)}(X_1) + \epsilon_2$ |
| $G_2$ |  | $X_1 := f^{(1)}(X_2) + \epsilon_1,$<br>$X_2 := \epsilon_2$ |

For each graph we assume that we have chosen specific values for each $\sigma_i^2$ in $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ and that we have chosen specific kernel parameters for $k$ in $\hat{f}^{(\cdot)} \sim \mathcal{GP}(0, k)$. In order to perform model selection with a Bayesian approach we need to calculate the (joint) marginal likelihoods for observational data and interventional data. According to (2.6) we can calculate the likelihood for observational data, $(x_1, x_2)$, given the graph, e.g. $G_1$, in closed form,

$$\begin{aligned} p(x_1, x_2 | G_1) &= p(x_2 | x_1, G_1) p(x_1 | G_1) \\ &= \underbrace{\int_{\mathbb{R}^N} p(x_2 | f^{(2)}_{x_1}, x_1, G_1) p(f^{(2)}_{x_1} | x_1, G_1) \, df^{(2)}_{x_1}}_{\mathcal{N}(0, k_{X_1 X_1} + \sigma_2^2 I_N)} \; p(x_1 | G_1) \\ &= (2\pi)^{-\frac{N}{2}} \det(k_{X_1 X_1} + \sigma_2^2 I_N)^{-\frac{1}{2}} \exp\left[ -\frac{1}{2} \mathbf{x}_2^\intercal (k_{X_1 X_1} + \sigma_2^2 I_N)^{-1} \mathbf{x}_2 \right] p(x_1 | G_1). \end{aligned}$$

Assume we choose an intervention $X_1 = x_1^{(X_1)}$, independent from our initial data and observe the interventional data point $(x_1^{(X_1)}, x_2^{(X_1)})$. Let $\tilde{x}_1 := (x_1^\intercal, x_1^{(X_1)})^\intercal$ and $\tilde{x}_2 := (x_2^\intercal, x_2^{(X_1)})^\intercal$. Based on our assumptions, the joint marginal likelihood of the available data is

$$\begin{aligned}
&p(x_1, x_2, x_1^{(X_1)}, x_2^{(X_1)}|G_1) \\
&\quad = p(x_1, x_2|G_1)p(x_1^{(X_1)}, x_2^{(X_1)}|x_1, x_2, G_1) \\
&\quad = p(x_2|x_1, G_1)p(x_2^{(X_1)}|x_1^{(X_1)}, x_1, x_2, G_1)p(x_1|G_1)\underbrace{p(x_1^{(X_1)}|x_1, x_2, G_1)}_{=\delta_{x_1^{(X_1)}}\left(x_1^{(X_1)}\right)} \\
&\quad = \underbrace{\int_{\mathbb{R}^{N+1}} p(\tilde{x}_2|f_{\tilde{x}_1}^{(2)}, \tilde{x}_1, G_1)p(f_{\tilde{x}_1}^{(2)}|\tilde{x}_1, G_1)\, df_{\tilde{x}_1}^{(2)}}_{\mathcal{N}(0, k_{\tilde{X}_1\tilde{X}_1} + \sigma_2^2 I_{N+1})} \; p(x_1|G_1) \\
&\quad = (2\pi)^{-\frac{N+1}{2}} \det(k_{\tilde{X}_1\tilde{X}_1} + \sigma_2^2 I_{N+1})^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\tilde{\mathbf{x}_2}^{\intercal}(k_{\tilde{X}_1\tilde{X}_1} + \sigma_2^2 I_{N+1})^{-1}\tilde{\mathbf{x}}_{\mathbf{2}}\right] p(x_1|G_1).
\end{aligned}$$

Thus, we can calculate the joint marginal likelihood by augmenting the initial data with the (given the graph) relevant interventional data, fit a GP model on the augmented data (i.e. data that contain new information about the functional relation), obtain the marginal likelihood thereof and multiply it with the marginal likelihood of the initial source node observations. Further interventional data points can be included analogously.

We can include a data point $(x_1^{(X_2)}, x_2^{(X_2)})$ obtained by an intervention on $X_2$ as follows.

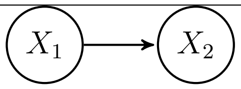
$$\begin{aligned}
&p(x_1, x_2, x_1^{(X_2)}, x_2^{(X_2)}|G_1) \\
&\quad = p(x_1, x_2|G_1)p(x_1^{(X_2)}, x_2^{(X_2)}|x_1, x_2, G_1) \\
&\quad = p(x_2|x_1, G_1)p(x_1|G_1)\underbrace{p(x_2^{(X_2)}|x_1^{(X_2)}, x_1, x_2, G_1)}_{=\delta_{x_2^{(X_2)}}\left(x_2^{(X_2)}\right)} p(x_1^{(X_2)}|x_1, x_2, G_1) \\
&\quad = \int_{\mathbb{R}^N} p(x_2|f_{x_1}^{(2)}, x_1, G_1)p(f_{x_1}^{(2)}|x_1, G_1)df_{x_1}^{(2)} \; p(x_1|G_1)p(x_1^{(X_2)}|x_1, x_2, G_1) \\
&\quad = (2\pi)^{-\frac{N}{2}} \det(k_{X_1X_1} + \sigma_2^2 I_N)^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\mathbf{x}_2{}^{\intercal}(k_{X_1X_1} + \sigma_2^2 I_N)^{-1}\mathbf{x}_2\right] p(x_1|G_1)p(x_1^{(X_2)}|G_1).
\end{aligned}$$

Here we do not have to fit another GP model since we do not have any new relevant information about the functional relationship given our graph. Instead we just multiply the likelihood of obtaining $x_1^{(X_2)}$ to the marginal likelihood of our initial data. Every further interventional data point can be included analogously.

We can also easily combine both types of interventions and obtain a closed form like

$$\begin{aligned}
&p(x_1, x_2, , x_1^{(X_1)}, x_2^{(X_1)}, x_1^{(X_2)}, x_2^{(X_2)}|G_1) \\
&= (2\pi)^{-\frac{N+1}{2}} \det(k_{\tilde{X}_1\tilde{X}_1} + \sigma_2^2 I_{N+1})^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\tilde{\mathbf{x}}_2^{\intercal}(k_{\tilde{X}_1\tilde{X}_1} + \sigma_2^2 I_{N+1})^{-1}\tilde{\mathbf{x}}_2\right] p(x_1|G_1)p(x_1^{(X_2)}|G_1).
\end{aligned}$$

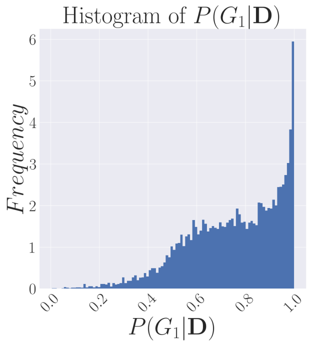
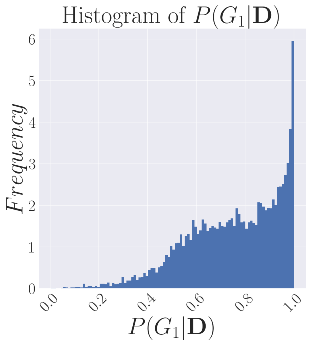
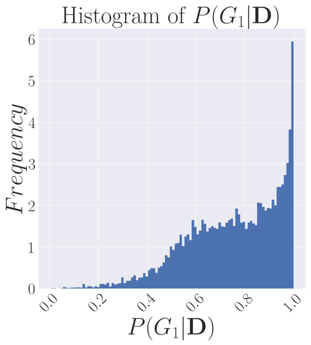
**Example 3.3** (Independent Random Interventions (3.1))**.** Consider a situation, where we are given five i.i.d. observations of the true data generating process,

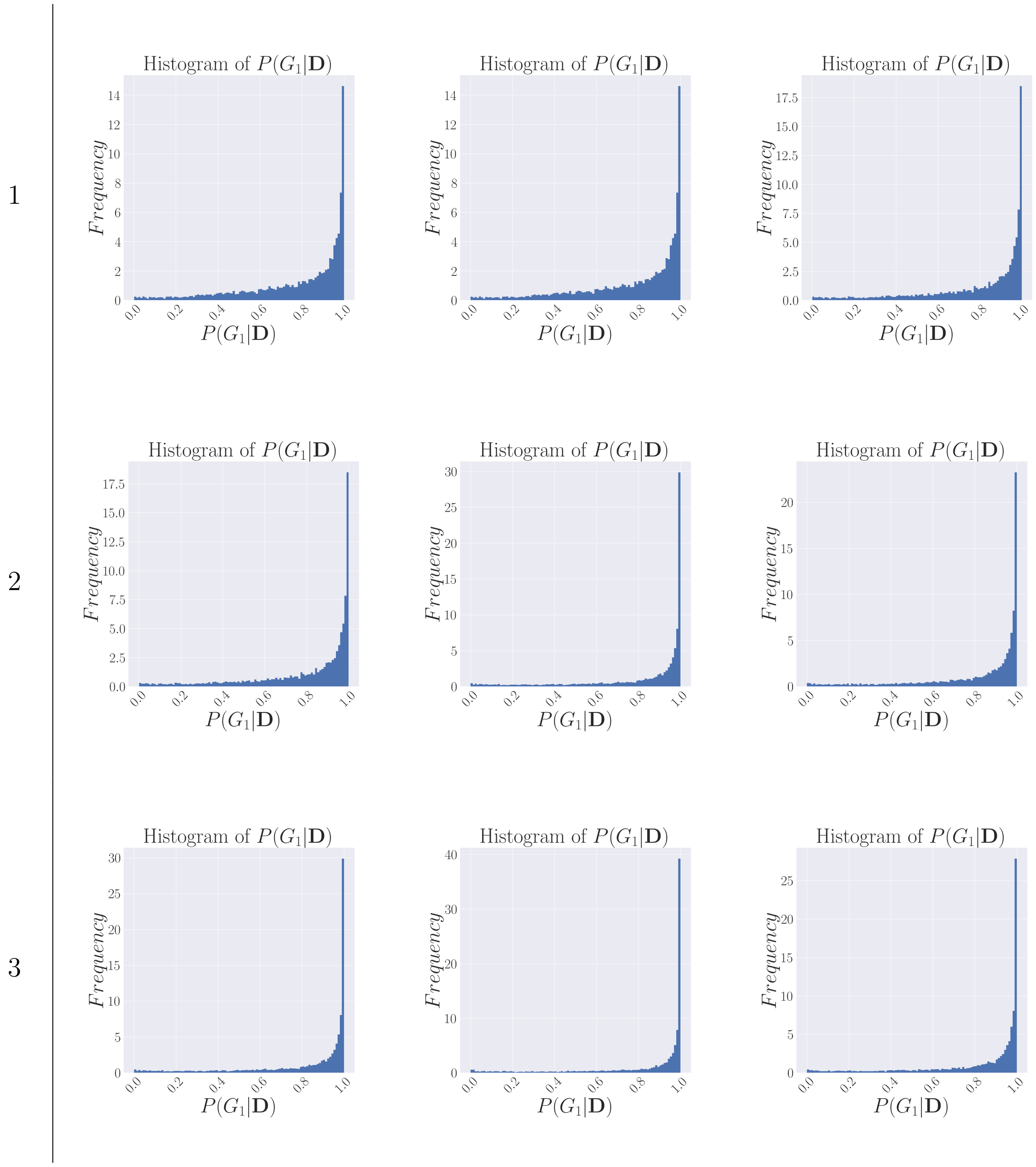
<table>
<tr><th></th><th>Graph</th><th>ANM</th></tr>
<tr><td>$G_{true}$</td><td>
</td><td>$X_1 := \epsilon_1,$<br>$X_2 := 2\tanh(X_1) + \epsilon_2,$</td></tr>
</table>

where $\epsilon_1 \sim \mathcal{N}(0,1)$ and $\epsilon_2 \sim \mathcal{N}(0,1)$. Assume that we specify for $G_1$, $\epsilon_1 \sim \mathcal{N}(0,1)$ and $\epsilon_2 \sim \mathcal{N}(0,1)$ and for $G_2$, $\epsilon_1 \sim \mathcal{N}(0,1)$ and $\epsilon_2 \sim \mathcal{N}(0,1)$ (see 3.1). Further assume that we decided on an interval where we uniformly draw interventions for each random variable. In this example we draw $x_1^{(X_1)}$ from $\mathcal{U}[-3,3]$ and $x_2^{(X_2)}$ from $\mathcal{U}[-2.1,2.1]$ and observe 10 intervention samples alternating between interventions on $X_1$ and $X_2$; starting with an intervention on $X_1$. Note that we chose another interval for interventions on $X_2$ because the true deterministic function is bounded in $[-2,2]$. We choose the Squared Exponential (SE) kernel with $\gamma = 1.75$ for each graph and calculate $P(G_i|D)$ in the following way,

$$P(G_i|D) = \frac{p(x_1, x_2, , x_1^{(X_1)}, x_2^{(X_1)}, x_1^{(X_2)}, x_2^{(X_2)}|G_i)P(G_i)}{\sum_{j=1}^{2} p(x_1, x_2, , x_1^{(X_1)}, x_2^{(X_1)}, x_1^{(X_2)}, x_2^{(X_2)}|G_j)P(G_j)}.$$

The experiment was repeated $10,000$ times with different random seeds. Further, a similar experiment where only interventions on $X_1$ were performed and a third experiment where only interventions on $X_2$ were performed. In (3.1) histograms of $10,000$ observations of $P(G_{true}|\mathbf{D})$ for each respective experiment and iteration are shown.

<table>
<tr><th>It.</th><th>Alternating Random Interventions</th><th>Only Random Interventions on $X_1$</th><th>Only Random Interventions on $X_2$</th></tr>
<tr><td>0</td><td>
</td><td>
</td><td>
</td></tr>
</table>


1
2
3
Histogram of $P(G_1|\mathbf{D})$
Frequency
$P(G_1|\mathbf{D})$
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
12
14
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
0
5
10
15
20
25
30
0
10
20
30
40

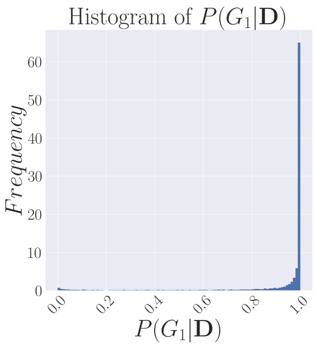
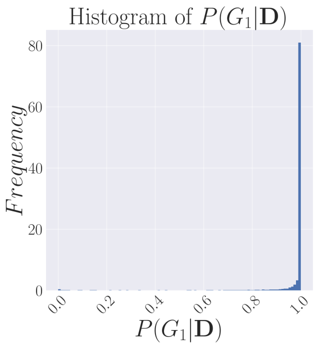
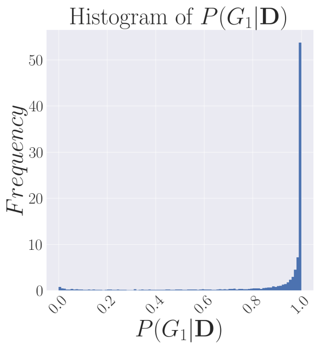
| | | | |
|---|---|---|---|
| 10 |  |  |  |

Table 3.1.: Comparison of the empirical distributions of the posterior probability of the true data generating graph, $P(G_1|\mathbf{D})$, for the three different experiments assuming known variance. (Example 3.3).

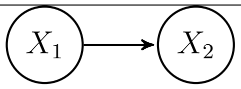
In the following table the most important numbers rounded to four decimals are summarized.

| | Mean | Std. Deviat | $P(G_1|\mathbf{D}) \geq 0.95$ | $P(G_1|\mathbf{D}) \leq 0.5$ |
|---|---|---|---|---|
| "Alternating Rnd. Interv." | 0.9087 | 0.2217 | 0.7815% | 0.0745% |
| "Only Rnd. Interv. on $X_1$" | 0.9528 | 0.1624 | 0.8836% | 0.0372% |
| "Only Rnd. Interv. on $X_2$" | 0.8810 | 0.2426 | 0.707% | 0.0963% |

The results indicate that the best strategy to reveal the underlying causal relation is to intervene on the cause, i.e. on $X_1$ in the example.

Next, we consider a similar example where the behavior of the implementation can be illustrated well.

**Example 3.4** (Independent Random Interventions (3.2))**.** Consider a situation, where we are given five i.i.d. observations of the true data generating process,

| | Graph | ANM |
|---|---|---|
| $G_{true}$ |  | $X_1 := \epsilon_1,$<br>$X_2 := 2\tanh(X_1) + \epsilon_2,$ |

where $\epsilon_1 \sim \mathcal{N}(0,1)$ and $\epsilon_2 \sim \mathcal{N}(0,0.1)$. Assume that we specify for $G_1$, $\epsilon_1 \sim \mathcal{N}(0,1)$ and $\epsilon_2 \sim \mathcal{N}(0,0.1)$ and for $G_2$, $\epsilon_1 \sim \mathcal{N}(0,0.1)$ and $\epsilon_2 \sim \mathcal{N}(0,1)$ (see 3.2). Further assume that we decided on an interval where we uniformly draw interventions for each random variable. In this example we draw $x_1^{(X_1)}$ from $\mathcal{U}[-3,3]$ and $x_2^{(X_2)}$ from $\mathcal{U}[-2.1,2.1]$ and observe 10 intervention samples alternating between interventions on $X_1$ and $X_2$; starting with an intervention on $X_1$. Note that we chose another interval for interventions on $X_2$ because the true deterministic function is bounded in $[-2,2]$. We choose the Squared

Exponential (SE) kernel with $\gamma = 1.75$ for each graph and calculate $P(G_i|D)$ in the following way,

$$P(G_i|D) = \frac{p(x_1, x_2, , x_1^{(X_1)}, x_2^{(X_1)}, x_1^{(X_2)}, x_2^{(X_2)}|G_i)P(G_i)}{\sum_{j=1}^{2} p(x_1, x_2, , x_1^{(X_1)}, x_2^{(X_1)}, x_1^{(X_2)}, x_2^{(X_2)}|G_j)P(G_j)}.$$

A typical result of this experiment is presented in the figures below.

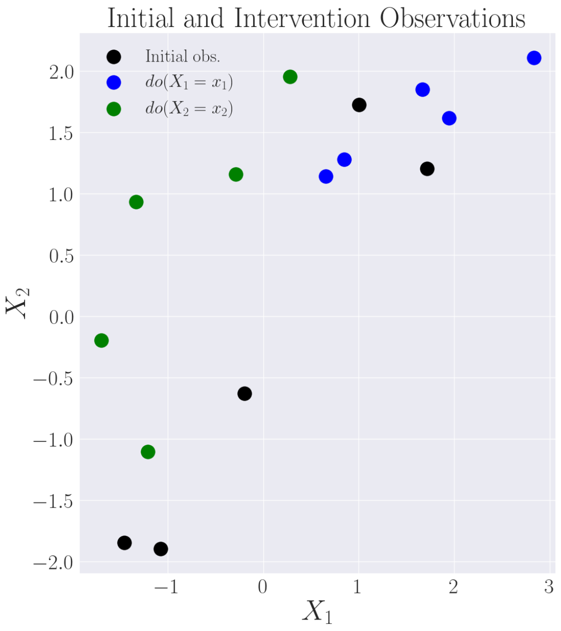


Figure 3.2.: Obs. from initial-, $do(x_1)$- and $do(x_2)$-distribution. (Example 3.4)

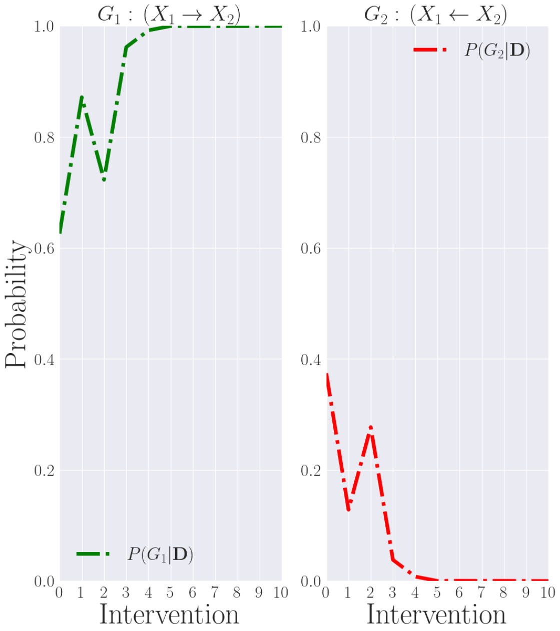


Figure 3.3.: DAG posterior probability, $P(G|\mathbf{D})$. $k_{SE}$ with $\gamma = 1.75$. (Example 3.4)

While the “intervention on $X_1$” data fits quite good to the initial observations, the “intervention on $X_2$” data does not. Since we calculate the GP regression for graph $G_2$ by augmenting the initial data with the $do(x_2)$-data, it becomes clear, that the marginal likelihood of the data $P(G_2|\mathbf{D})$ should become low relative to $P(G_1|\mathbf{D})$ (where we include the $do(x_1)$-data for GP regression). In other words, it is much more difficult to find a regression function for the initial data combined with the $do(x_2)$-data, than for the initial data combined with the $do(x_1)$-data. This explains the result in Figure 3.3, where less than ten interventions are needed to ”converge” to the true graph, $G_1$. In the beginning, the occurrence of a sawblade pattern is typical in this setup, because we have only little data and the GP regression finds well fitting function describing the wrong direction $X_2 \to X_1$. The purpose of Figures 3.4 and 3.5 is to provide some more intuition on the impact of the GP regressions’ marginal likelihood on the quantity of interest $P(G|\mathbf{D})$. The 95% confidence interval depicted in the plots relates to the GP fit. Every domain point inferred with GP regression is normally distributed. The blue line is the mean of every such point and the orange area is the 95% confidence interval based on the variance at each respective point. So the confidence band contains information on how sure we are about our estimate in the corresponding area.

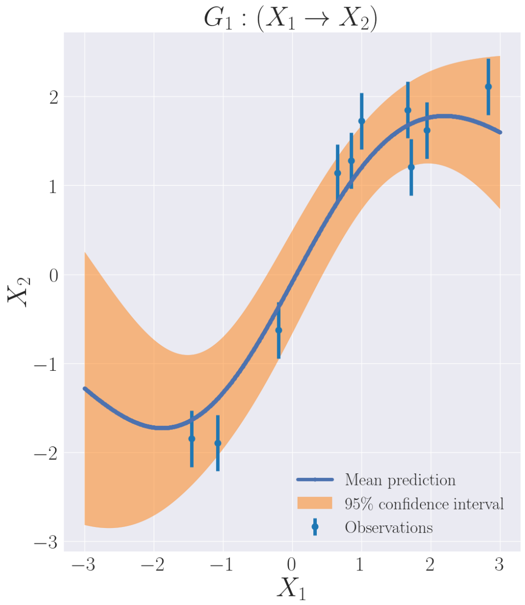


Figure 3.4.: GP fit ($k_{SE,\gamma=1.75}$) on initial- and $do(x_1)$-data. (Example 3.4).

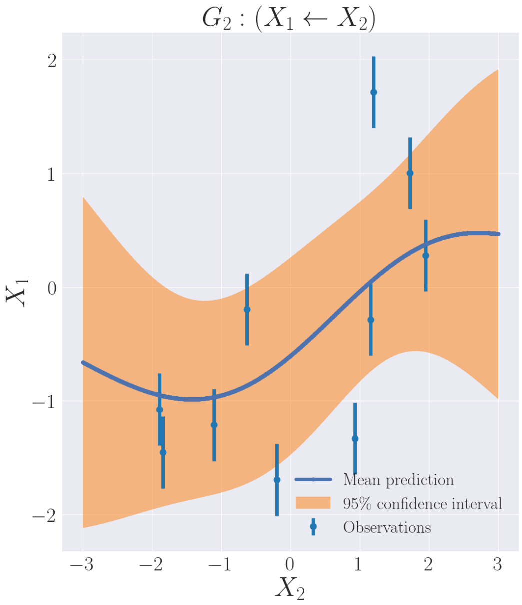
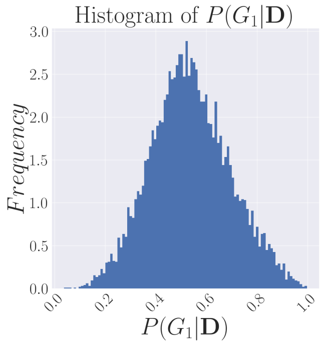
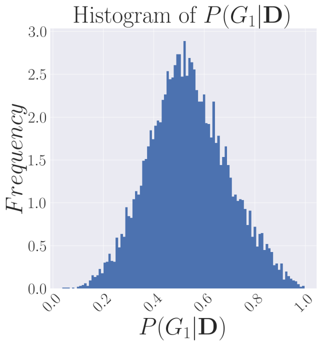
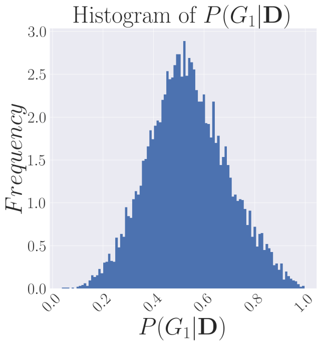
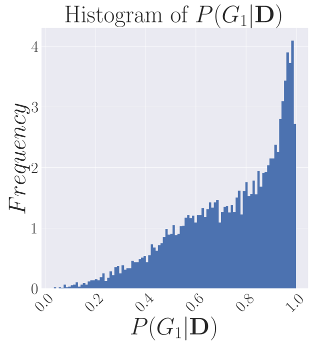
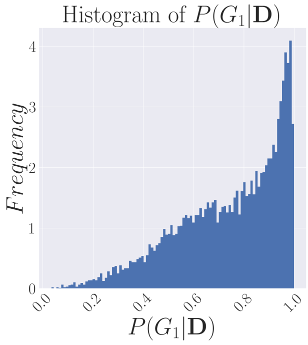
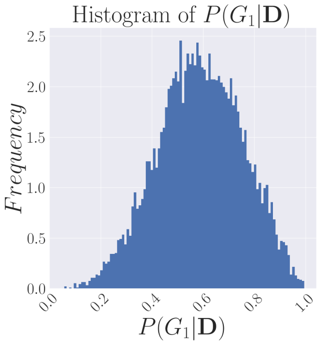
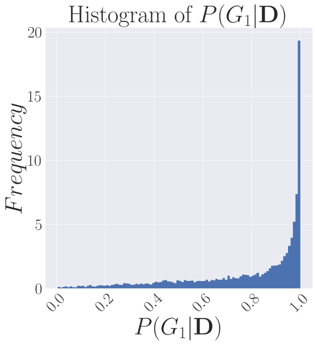
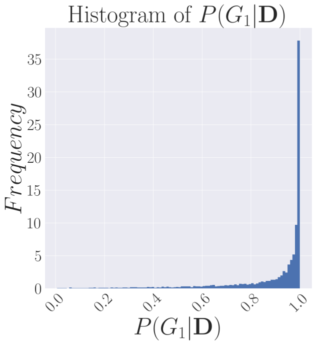
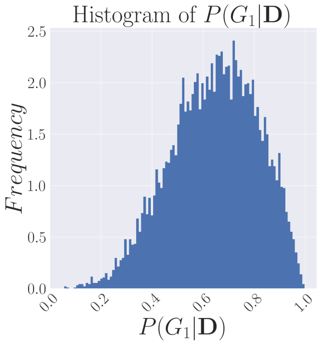


Figure 3.5.: GP fit ($k_{SE,\gamma=1.75}$) on initial- and $do(x_2)$-data. (Example 3.4).

In order to evaluate the performance of this approach, $10,000$ repetitions of the same experiment were repeated with different random seeds. The mean of $P(G_1|\mathbf{D})$ after the tenth intervention was $\approx 0.9913$ with a standard deviation of $\approx 0.0614$. If we decided for graph $G_1$ only if $P(G_1|\mathbf{D}) \geq 0.95$, the success rate was 0.9719. If we defined failure as $P(G_1|\mathbf{D}) \leq 0.5$, the failure rate was 0.0052.

Since adding "intervention on $X_1$" data is very similar to augmenting the initial data set with more i.i.d. samples from the data generating process, it is a reasonable question to ask at this point, if interventions make a difference. To answer this question we change the experiment such that we only do interventions on $X_1$. And further do another experiment where we, instead of doing interventions on $X_1$, in each step augment our data by one more i.i.d observation. In the "only interventions on $X_1$" case, the mean of $P(G_1|\mathbf{D})$ after $10,000$ experiments was $\approx 0.9953$, while in the "i.i.d. union" case it was $\approx 0.9375$ after the tenth observation has been added. Thus, for this example we have numerical support that interventional data is more informative than purely observational data. To compare the three different methods for adding a new data point, it is insightful to consider how the histograms of the posterior distribution from the true graph given the respective data evolves over the iterations. In (3.2) histograms of $10,000$ observations of $P(G_{true}|\mathbf{D})$ for each respective experiment and iteration is shown. We will later consider a similar study for the case of optimal interventions. Histograms of the case, where we only perform interventions on $X_2$ are not presented. The corresponding histograms are very similar to the first two columns of the below table.

| It. | Alternating Random Interventions | Only Random Interventions on $X_1$ | Additional i.i.d. Sample |
|---|---|---|---|
| 0 |  |  |  |
| 1 |  |  |  |
| 2 |  |  |  |

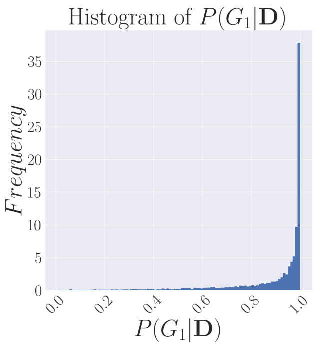
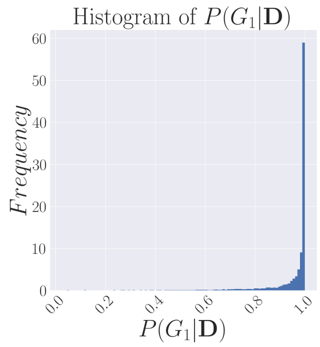
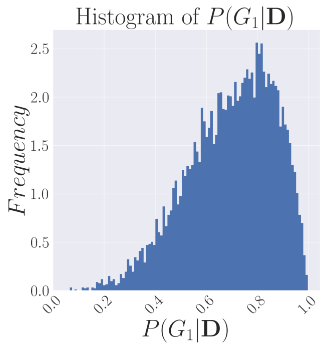
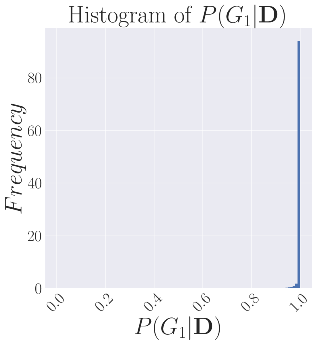
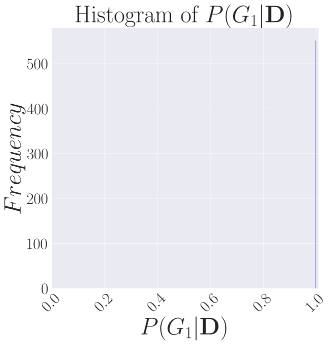
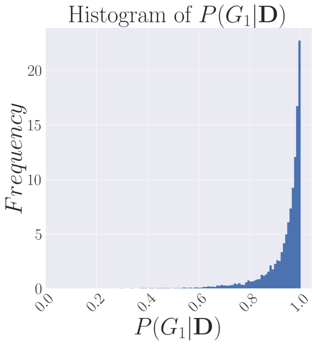
| | | | |
|---|---|---|---|
| 3 |  |  |  |
| 10 |  |  |  |

Table 3.2.: Comparison of the empirical distributions of the posterior probability of the true data generating DAG, $P(G_1|\mathbf{D})$, for the three different experiments. (Example 3.4)

For this example we illustrate the properties of Matérn kernels, when changing the parameter $\nu$. We keep the setting from above fixed and only change the SE kernel to Matérn kernels with different $\nu$'s but always the same $\gamma$. We propose to use the fixed value $\nu = 2.5$ for all settings since the results for causal discovery are similar in all cases.

**Case $\nu = 2.5$ :**
We can expect a smooth fit of the regression function and the confidence interval to be rather smooth too. An important aspect of this example is that the fit is very similar to the case of the SE kernel, which is infinitely many often mean square differentiable compared to the employed Matérn kernel which is only two times mean square differentiable.

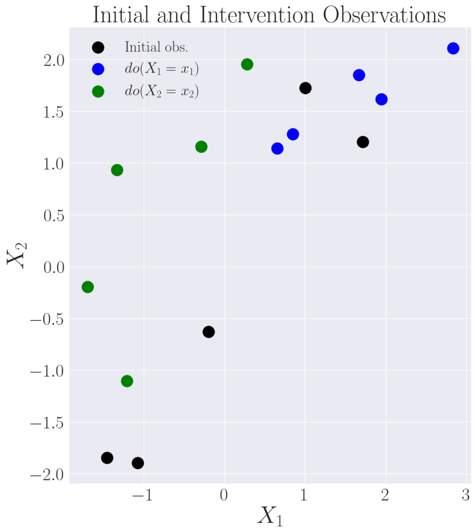


Figure 3.6.: Obs. from initial-, $do(x_1)$- and $do(x_2)$-distribution. (Example 3.4, $\nu = 2.5$).

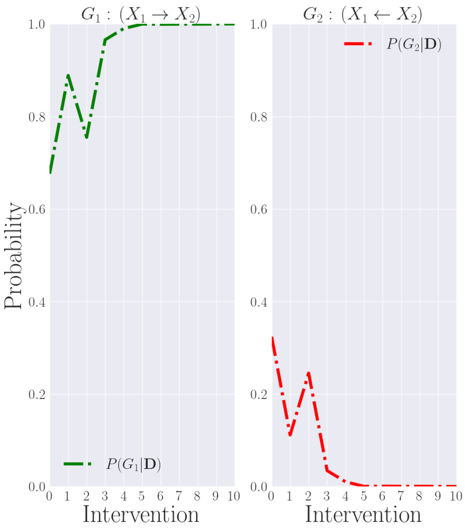


Figure 3.7.: DAG posterior probability, $P(G|\mathbf{D})$. (Example 3.4, $\nu = 2.5$).

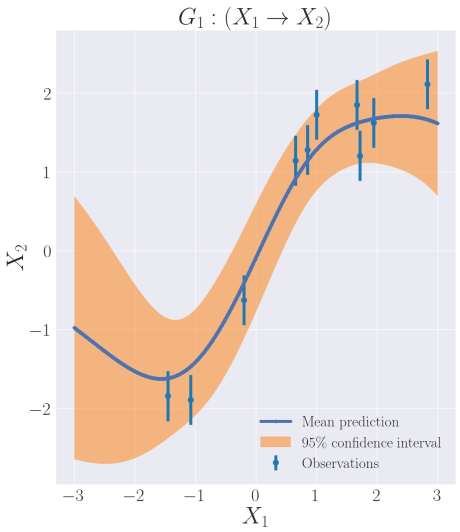


Figure 3.8.: GP fit ($k_{\nu=2.5,\gamma=1.75}$) on initial- and $do(x_1)$-data. (Example 3.4, $\nu = 2.5$).

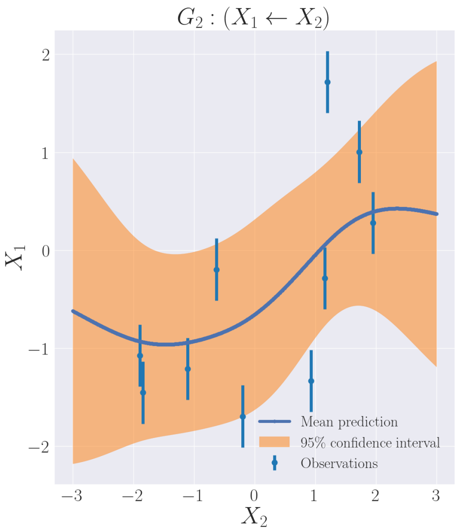


Figure 3.9.: GP fit ($k_{\nu=2.5,\gamma=1.75}$) on initial- and $do(x_2)$-data. (Example 3.4, $\nu = 2.5$).

**Case $\nu = 1.5$ :**
We can expect a rougher fit compared to the previous case, what also implies wider confidence intervals.

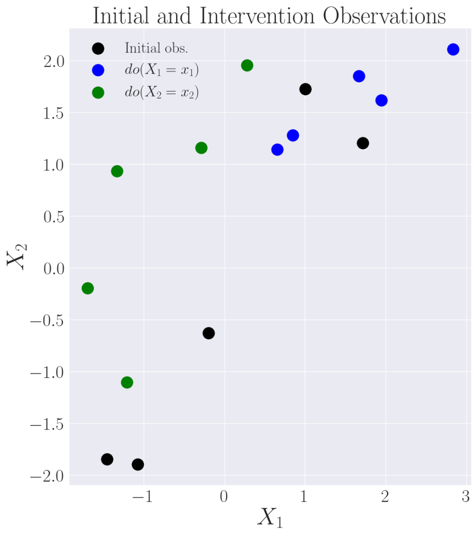


Figure 3.10.: Obs. from initial-, $do(x_1)$- and $do(x_2)$-distribution. (Example 3.4, $\nu = 1.5$).

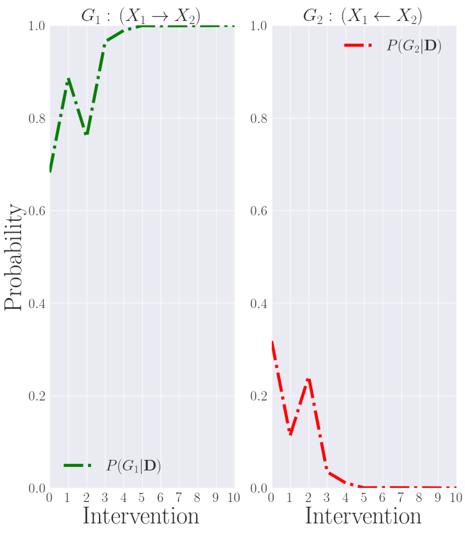


Figure 3.11.: DAG posterior probability, $P(G|\mathbf{D})$. (Example 3.4, $\nu = 1.5$).

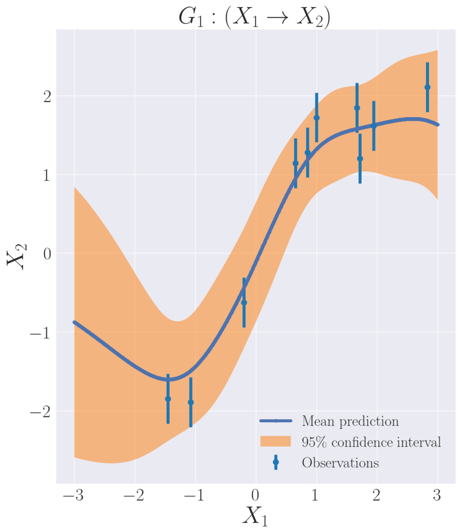


Figure 3.12.: GP fit ($k_{\nu=1.5,\gamma=1.75}$) on initial- and $do(x_1)$-data. (Example 3.4, $\nu = 1.5$).

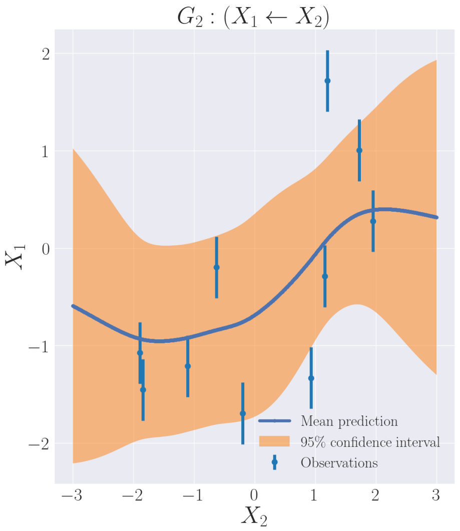


Figure 3.13.: GP fit ($k_{\nu=1.5,\gamma=1.75}$) on initial- and $do(x_2)$-data. (Example 3.4, $\nu = 1.5$).

**Case $\nu = 0.5$ :**
We can expect a rough fit compared to the previous cases, what also implies heavily inflated confidence bands in areas where we do not have any observations.

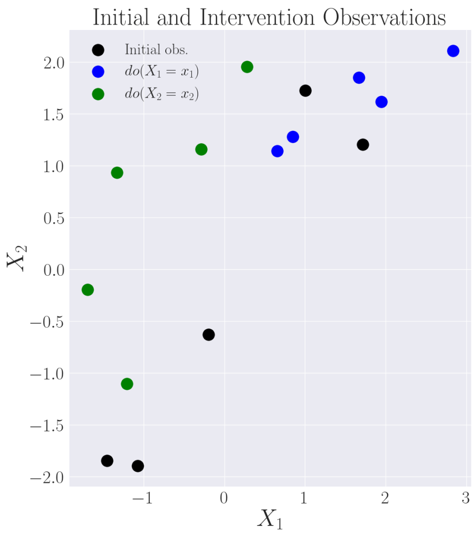


Figure 3.14.: Obs. from initial-, $do(x_1)$- and $do(x_2)$-distribution. (Example 3.4, $\nu = 0.5$).

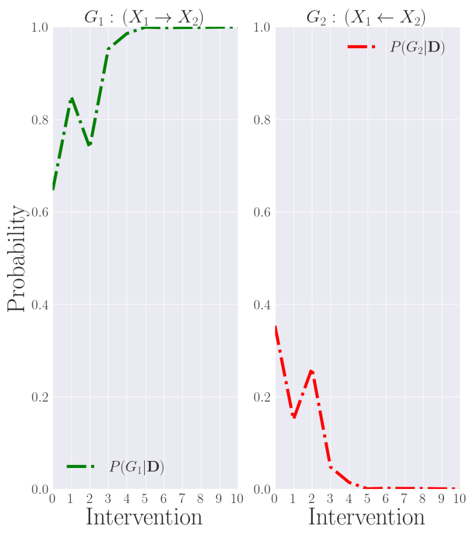


Figure 3.15.: DAG posterior probability, $P(G|\mathbf{D})$. (Example 3.4, $\nu = 0.5$).

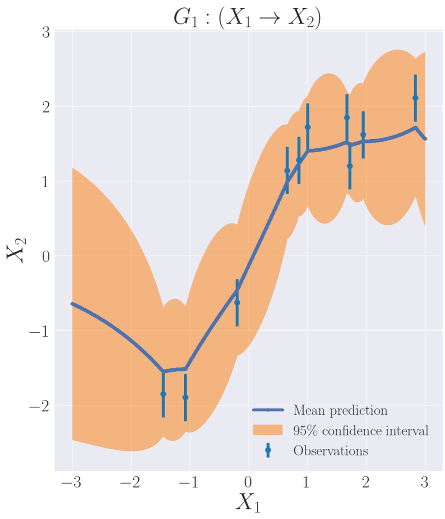


Figure 3.16.: GP fit $(k_{\nu=0.5,\gamma=1.75})$ on initial- and $do(x_1)$-data. (Example 3.4, $\nu = 0.5$).

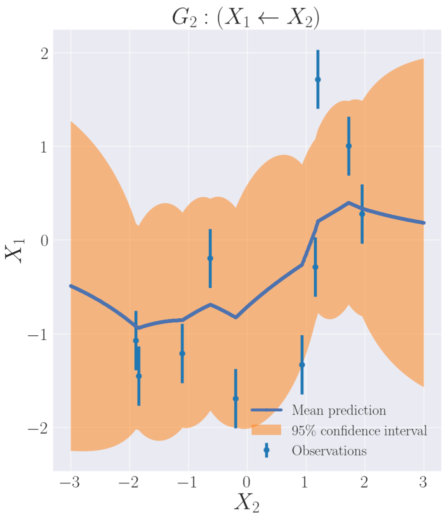


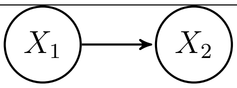
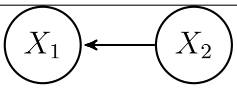
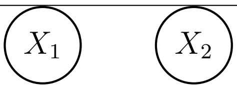
Figure 3.17.: GP fit $(k_{\nu=0.5,\gamma=1.75})$ on initial- and $do(x_2)$-data. (Example 3.4, $\nu = 0.5$).

### 3.2.3. Causal Discovery with Optimal Interventions

Let $x_1$ and $x_2$ denote vectors of $N$ i.i.d. observations of a bivariate ANM, where the noise terms are mutually independent and assume acyclicity and causal sufficiency, as before. In the following we include the empty graph as a possible model choice, compared to the previous section, where we only had the two graphs containing a directed edge. This is motivated because we want to avoid situations where the hypothesis space is too narrow. For example, $G_1$ is a far better explanation for the observations than $G_2$ but in reality both are poor explanations because the truth is, that the variables are not dependent at all. If we do not include the empty graph as a possible hypothesis, we would have a high posterior probability for $G_1$ and would make wrong inference. Especially, since we will use flat priors for the graphs, this may prevent us to end up in such situation (similar to what is known as Lindley's paradox [22]).When we compare the posteriors in both cases the latter one has one summand more in the denominator (the DAGs are defined just below)

$$P(G_1|D) = \frac{p(D|G_1)P(G_1)}{p(D|G_1)P(G_1) + p(D|G_2)P(G_2) + p(D|G_3)P(G_3)}$$

Thus, we decide between the following models

| | Graph | ANM |
|---|---|---|
| $G_1$ |  | $X_1 := \epsilon_1,$<br>$X_2 := f^{(2)}(X_1) + \epsilon_2$ |
| $G_2$ |  | $X_1 := f^{(1)}(X_2) + \epsilon_1,$<br>$X_2 := \epsilon_2$ |
| $G_3$ |  | $X_1 := \epsilon_1,$<br>$X_2 := \epsilon_2$ |

For each graph we assume that we have chosen specific values for each $\sigma_i^2$ in $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ and that we have chosen specific kernel parameters for $k$ in $\hat{f}^{(\cdot)} \sim \mathcal{GP}(0, k)$. The (joint) marginal likelihoods for the observational data can be calculated using (2.6) and are given by

$$\underbrace{p(x_1, x_2|G_1)}_{=:C_{G_1}} = \underbrace{(2\pi)^{-\frac{N}{2}} \det(k_{X_1X_1} + \sigma_2^2 I_N)^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\mathbf{x_2}^\intercal (k_{X_1X_1} + \sigma_2^2 I_N)^{-1}\mathbf{x_2}\right]}_{\mathcal{N}(0, k_{X_1X_1} + \sigma_2^2 I_N)} p(x_1|G_1),$$

$$\underbrace{p(x_1, x_2|G_2)}_{=:C_{G_2}} = \underbrace{(2\pi)^{-\frac{N}{2}} \det(k_{X_2X_2} + \sigma_1^2 I_N)^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\mathbf{x_1}^\intercal (k_{X_2X_2} + \sigma_1^2 I_N)^{-1}\mathbf{x_1}\right]}_{\mathcal{N}(0, k_{X_2X_2} + \sigma_1^2 I_N)} p(x_2|G_2) \text{ and}$$

$$\underbrace{p(x_1, x_2|G_3)}_{=:C_{G_3}} = p(x_1|G_3)p(x_2|G_3).$$

The posterior probabilities are

$$P(G_i|D) = \frac{p(x_1, x_2|G_i)P(G_i)}{\sum_{j=1}^{3} p(x_1, x_2|G_j)P(G_j)} \quad \text{for } i \in \{1, 2, 3\}.$$

Next, we want to answer the question what is the most informative intervention that we can perform, i.e., we try to find $(j^*, x^*)$ from (3.8). For this we need to calculate $P(\mathbf{x}_{-j}|G, do(X_j = x))$ and $P(G|\mathbf{x}_{-j}, do(X_j = x))$, given our initial observations. To ease the notational burden a little bit, we denote $x_1^1$ a value of an intervention on $X_1$ and $x_2^2$ an intervention value on $X_2$. At this stage of our sequential approach, the estimator for the functional value at a respective intervention value is

$$f_{X_1}^{(2)}(x_1^1)|x_1, x_2, x_1^1 \sim \mathcal{N}(\underbrace{k_{x_1^1 X_1}(k_{X_1 X_1} + \sigma_{2_{G_1}}^2 I_N)^{-1} x_2}_{:=\tilde{\mu}_{G_1}(x_1^1)}, \underbrace{k_{x_1^1 x_1^1} - k_{x_1^1 X_1}(k_{X_1 X_1} + \sigma_{2_{G_1}}^2 I_N)^{-1} k_{X_1 x_1^1}}_{:=\tilde{\sigma}_{G_1}^2(x_1^1)}),$$

$$f_{X_2}^{(1)}(x_2^2)|x_1, x_2, x_2^2 \sim \mathcal{N}(\underbrace{k_{x_2^2 X_2}(k_{X_2 X_2} + \sigma_{1_{G_2}}^2 I_N)^{-1} x_1}_{:=\tilde{\mu}_{G_2}(x_2^2)}, \underbrace{k_{x_2^2 x_2^2} - k_{x_2^2 X_2}(k_{X_2 X_2} + \sigma_{1_{G_2}}^2 I_N)^{-1} k_{X_2 x_2^2}}_{:=\tilde{\sigma}_{G_2}^2(x_2^2)}).$$

For the calculation of $P(G|\mathbf{x}_{-j}, do(X_j = x))$ define $D_1 := x_2^1, x_1^1, x_1, x_2$ and $D_2 := x_1^2, x_2^2, x_1, x_2$. Below, we do an exemplary calculation of $P(D_1|G_1)$.

$$\begin{aligned} p(D_1|G_1) &= p(x_1, x_2|G_1)p(x_2^1, x_1^1|x_1, x_2, G_1) \\ &= p(x_2^1|x_1^1, x_1, x_2, G_1)p(x_1^1|x_1, x_2, G_1)p(x_1, x_2|G_1) \\ &= \underbrace{\int_{\mathbb{R}} p(x_2^1|f_{X_1}^{(2)}(x_1^1), x_1, x_2, x_1^1, G_1)p(f_{X_1}^{(2)}(x_1^1)|x_1, x_2, x_1^1, G_1)\, df_{X_1}^{(2)}(x_1^1)}_{\mathcal{N}\left(\tilde{\mu}_{G_1}(x_1^1), \sigma_{2_{G_1}}^2 + \tilde{\sigma}_{G_1}^2(x_1^1)\right)} \; p(x_1, x_2|G_1) \\ &= (2\pi)^{-\frac{1}{2}}(\sigma_{2_{G_1}}^2 + \tilde{\sigma}_{G_1}^2(x_1^1))^{-\frac{1}{2}} \exp\left(\frac{(x_2^1 - \tilde{\mu}_{G_1}(x_1^1))^2}{2(\sigma_{2_{G_1}}^2 + \tilde{\sigma}_{G_1}^2(x_1^1))}\right) C_{G_1} \end{aligned} \tag{3.10}$$

The other likelihoods are given by

$$\begin{aligned} p(D_1|G_2) &= \underbrace{(2\pi)^{-\frac{1}{2}}\sigma_{2_{G_2}}^{-1} \exp\left(\frac{(x_2^1)^2}{2\sigma_{2_{G_2}}^2}\right)}_{\mathcal{N}(0, \sigma_{2_{G_2}}^2)} C_{G_2}, \\ p(D_1|G_3) &= \underbrace{(2\pi)^{-\frac{1}{2}}\sigma_{2_{G_3}}^{-1} \exp\left(\frac{(x_2^1)^2}{2\sigma_{2_{G_3}}^2}\right)}_{\mathcal{N}(0, \sigma_{2_{G_3}}^2)} C_{G_3}, \end{aligned} \tag{3.11}$$

$$
\begin{aligned}
p(D_2|G_1) &= \underbrace{(2\pi)^{-\frac{1}{2}}\sigma^{-1}_{2_{G_1}}\exp\left(\frac{(x_1^2)^2}{2\sigma^1_{2_{G_1}}}\right)}_{\mathcal{N}(0,\sigma^1_{2_{G_1}})} C_{G_1}, \\
p(D_2|G_2) &= \underbrace{(2\pi)^{-\frac{1}{2}}(\sigma^2_{1_{G_2}} + \tilde{\sigma}^2_{G_2}(x_2^2))^{-\frac{1}{2}}\exp\left(\frac{(x_1^2 - \tilde{\mu}_{G_2}(x_2^2))^2}{2(\sigma^2_{1_{G_2}} + \tilde{\sigma}^2_{G_2}(x_2^2))}\right)}_{\mathcal{N}(\tilde{\mu}_{G_2}(x_2^2),\sigma_1^2+\tilde{\sigma}^2_{G_2}(x_2^2))} C_{G_2}, \\
p(D_2|G_3) &= \underbrace{(2\pi)^{-\frac{1}{2}}\sigma^{-1}_{2_{G_3}}\exp\left(\frac{(x_1^2)^2}{2\sigma^1_{2_{G_3}}}\right)}_{\mathcal{N}(0,\sigma^1_{2_{G_3}})} C_{G_3},
\end{aligned}
\tag{3.12}
$$

and the respective posterior quantities can be computed as follows

$$
P(G_i|D_k) = \frac{p(D_k|G_i)P(G_i)}{\sum_{j=1}^3 p(D_k|G_j)P(G_j)} \quad \text{for } i \in \{1,2,3\},\ k \in \{1,2\}.
$$

Since this calculation already revealed how to sample from $P(\mathbf{x}_{-j}|G, do(X_j = x))$, we have all necessary ingredients in place to solve our optimization problem. Before we let the Bayesian Optimization algorithm work for us, it is worth to consider the problem more closely to simplify the optimization.

Define $g_j(x) := \sum_{G\in\mathcal{G}} P(G) \int P(x_{-j}|G, do(X_j = x)) \log P(G|x_{-j}, do(X_j = x))\, dx_{-j}$. In case of $j = 1$, we have

$$
\begin{aligned}
&g_1(x_1^1) = \\
&\quad P(G_1)\int_{\mathbb{R}} \frac{\exp\left(\frac{(x_2^1-\tilde{\mu}_{G_1}(x_1^1))^2}{2(\sigma^2_{2_{G_1}}+\tilde{\sigma}^2_{G_1}(x_1^1))}\right)}{\sqrt{(2\pi)(\sigma^2_{2_{G_1}} + \tilde{\sigma}^2_{G_1}(x_1^1))}} \Bigg[\log\left(\frac{P(G_1)C_{G_1}\exp\left(\frac{(x_2^1-\tilde{\mu}_{G_1}(x_1^1))^2}{2(\sigma^2_{2_{G_1}}+\tilde{\sigma}^2_{G_1}(x_1^1))}\right)}{\sqrt{(2\pi)(\sigma^2_{2_{G_1}} + \tilde{\sigma}^2_{G_1}(x_1^1))}}\right) \\
&\qquad\qquad\qquad\qquad\qquad\qquad - \log\left(\sum_{j=1}^3 p(D_1|G_j)P(G_j)\right)\Bigg]\, dx_2^1 \\
&\quad + P(G_2)\int_{\mathbb{R}} \frac{1}{\sigma_{2_{G_2}}\sqrt{(2\pi)}}\exp\left(\frac{(x_2^1)^2}{2\sigma^2_{2_{G_2}}}\right)\Bigg[\log\left(\frac{P(G_2)C_{G_2}}{\sigma_{2_{G_2}}\sqrt{(2\pi)}}\exp\left(\frac{(x_2^1)^2}{2\sigma^2_{2_{G_2}}}\right)\right) \\
&\qquad\qquad\qquad\qquad\qquad\qquad - \log\left(\sum_{j=1}^3 p(D_1|G_j)P(G_j)\right)\Bigg]\, dx_2^1 \\
&\quad + P(G_3)\int_{\mathbb{R}} \frac{1}{\sigma_{2_{G_3}}\sqrt{(2\pi)}}\exp\left(\frac{(x_2^1)^2}{2\sigma^2_{2_{G_3}}}\right)\Bigg[\log\left(\frac{P(G_2)C_{G_3}}{\sigma_{2_{G_3}}\sqrt{(2\pi)}}\exp\left(\frac{(x_2^1)^2}{2\sigma^2_{2_{G_3}}}\right)\right) \\
&\qquad\qquad\qquad\qquad\qquad\qquad - \log\left(\sum_{j=1}^3 p(D_1|G_j)P(G_j)\right)\Bigg]\, dx_2^1.
\end{aligned}
$$

We can use the linearity of the integral, the fact that additive constants are irrelevant for optimization and that, e.g., the following holds

$$P(G_1)\int_{\mathbb{R}} \frac{\exp\left(\frac{(x_2^1-\tilde{\mu}_{G_1}(x_1^1))^2}{2(\sigma^2_{2_{G_1}}+\tilde{\sigma}^2_{G_1}(x_1^1))}\right)}{\sqrt{(2\pi)(\sigma^2_{2_{G_1}}+\tilde{\sigma}^2_{G_1}(x_1^1))}} \log\left(\frac{P(G_1)C_{G_1}\exp\left(\frac{(x_2^1-\tilde{\mu}_{G_1}(x_1^1))^2}{2(\sigma^2_{2_{G_1}}+\tilde{\sigma}^2_{G_1}(x_1^1))}\right)}{\sqrt{(2\pi)(\sigma^2_{2_{G_1}}+\tilde{\sigma}^2_{G_1}(x_1^1))}}\right) dx_2^1$$
$$=\frac{P(G_1)}{2}\,\mathbb{E}_{X_2\sim\mathcal{N}(\tilde{\mu}_{G_1}(x_1^1),(\sigma^2_{2_{G_1}}+\tilde{\sigma}^2_{G_1}(x_1^1)))}\left[\left(\frac{X_2-\tilde{\mu}_{G_1}(x_1^1)}{\sqrt{(\sigma^2_{2_{G_1}}+\tilde{\sigma}^2_{G_1}(x_1^1))}}\right)^2\right]$$
$$+P(G_1)\log(P(G_1))+P(G_1)\log(C_{G_1})-P(G_1)\log\left(\sqrt{(2\pi)(\sigma^2_{2_{G_1}}+\tilde{\sigma}^2_{G_1}(x_1^1))}\right),$$

to obtain an equivalent objective function

$$\begin{aligned}\tilde{g}_1(x_1^1) = &-P(G_1)\log\left(\sqrt{(2\pi)({}^{G_1}\sigma_2^2+\tilde{\sigma}^2_{G_1}(x_1^1))}\right) - P(G_1)\,\mathbb{E}_{x_2^1\sim\mathcal{N}(\tilde{\mu}_{G_1}(x_1^1),(\sigma^2_{2_{G_1}}+\tilde{\sigma}^2_{G_1}(x_1^1)))}\left[A_{x_1^1}\right] \\ &- P(G_2)\,\mathbb{E}_{x_2^1\sim\mathcal{N}(0,\sigma^2_{2_{G_2}})}\left[A_{x_1^1}\right] - P(G_3)\,\mathbb{E}_{x_2^1\sim\mathcal{N}(0,\sigma^2_{2_{G_3}})}\left[A_{x_1^1}\right],\end{aligned} \tag{3.13}$$

with

$$A_{x_1^1} := \log\left(\sum_{j=1}^{3} p(D_1|G_j)P(G_j)\right).$$

The solution of the optimization problem for $j=1$ can thus be written as

$$^{*}x_1^1 = \underset{x_1^1\in\mathcal{X}_1}{\operatorname{argmax}}\, g_1(x_1^1) = \underset{x_1^1\in\mathcal{X}_1}{\operatorname{argmax}}\, \tilde{g}_1(x_1^1). \tag{3.14}$$

We can efficiently calculate $\tilde{g}_1(x_1^1)$ using a Monte Carlo method and then use the derivative free Bayesian Optimization to find an approximation of $^{*}x_1^1$.
For $j=2$ we analogously get that,

$$^{*}x_2^2 = \underset{x_2^2\in\mathcal{X}_2}{\operatorname{argmax}}\, g_2(x_2^2) = \underset{x_2^2\in\mathcal{X}_2}{\operatorname{argmax}}\, \tilde{g}_2(x_2^2), \tag{3.15}$$

where

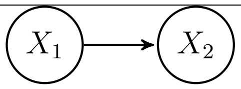
$$\begin{aligned}\tilde{g}_2(x_2^2) = &-P(G_2)\log\left(\sqrt{(2\pi)(\sigma^2_{1_{G_2}}+\tilde{\sigma}^2_{G_2}(x_2^2))}\right) - P(G_2)\,\mathbb{E}_{x_1^2\sim\mathcal{N}(\tilde{\mu}_{G_2}(x_2^2),(\sigma^2_{1_{G_2}}+\tilde{\sigma}^2_{G_2}(x_2^2)))}\left[B_{x_2^2}\right] \\ &- P(G_1)\,\mathbb{E}_{x_1^2\sim\mathcal{N}(0,\sigma^2_{1_{G_1}})}\left[B_{x_2^2}\right] - P(G_3)\,\mathbb{E}_{x_1^2\sim\mathcal{N}(0,\sigma^2_{1_{G_3}})}\left[B_{x_2^2}\right]\end{aligned}$$

with

$$B_{x_2^2} := \log\left(\sum_{j=1}^{3} p(D_2|G_j)P(G_j)\right).$$

Having the optimizers from above, we decide for the $j^*$, that is the index of $\max\{g_2(^*x_2^2), g_1(^*x_1^1)\}$. If $j^* = 1$, we perform the optimal intervention $^*x_1^1$. Suppose we obtain the interventional data sample $(^*x_1^1, x_2^1)$. For this data point we can compute the likelihood given our current believes. Based on these likelihoods we can compute the posterior probability of the graphs. Since the sample stems from the true intervened data generating process, it does, depending on the DAG, contain new relevant information about $f^{(1)}$ or $f^{(2)}$. Therefore, we update the GP fits in all DAGs accordingly at the end of each iteration . For the next iteration we will use the new fit exactly the same way and perform the calculations analogously to them described above. Further, we can make use of the sequential setup and safe the $p(D_j, G_i)$'s after every iteration and compute only the one new factor for every quantity in (3.11).
Next, we re-consider the 2 tanh-example from above and study the behavior of the implemented algorithm in detail.

**Example 3.5** (Sequential Optimal Interventions (3.1))**.** Similar as in (3.4), consider a situation, where we are given five i.i.d. observations of the true data generating process,

| | Graph | ANM |
|---|---|---|
| $G_{true}$ |  | $X_1 := \epsilon_1,$<br>$X_2 := 2\tanh(X_1) + \epsilon_2,$ |

where $\epsilon_1 \sim \mathcal{N}(0,1)$ and $\epsilon_2 \sim \mathcal{N}(0,1)$.
Assume that we specify for $G_1$, $\epsilon_1 \sim \mathcal{N}(0,1)$ and $\epsilon_2 \sim \mathcal{N}(0,1)$; for $G_2$, $\epsilon_1 \sim \mathcal{N}(0,1)$ and $\epsilon_2 \sim \mathcal{N}(0,1)$ and for $G_3$, $\epsilon_1 \sim \mathcal{N}(0,1)$ and $\epsilon_2 \sim \mathcal{N}(0,1)$ (see 3.1). Further, for the kernel we choose the Matérn kernel with $\nu = 2.5$ and $\gamma = 1.75$ in both models $G_1$ and $G_2$. A typical result of this experiment is summarized in the following figures.

The implemented algorithm is able to arrive at the true graph very quickly and achieves very high posterior probability of the true DAG after only 3 interventions. It did no intervention on $X_2$. Thus, in our example it is most informative about the causal directions, to test if our beliefs about the functional relation $f^{(2)}$ are true. The algorithm chooses a location, where we have observations and where the data points imply a nonlinear curvature of the GP fit of $f^{(2)}$. Further, it is remarkable that it chooses interventions always in the same region of the function support. To investigate this behavior further, we consider the objective functions subject to optimization. Below we plot the objective functions for an intervention on $X_1$ (left) and for an intervention on $X_2$ (right). The overall minimum is obtained by comparing the two minima $g_1(x_1^{1^*})$ and $g_2(x_2^{2^*})$. We also report the corresponding interventional data point, that was sampled based on the intervention value of the optimization procedure. The true objectives and the simplified objectives fit together very well up to scaling and have a almost quadratic shaped unique minimum (inside the box constraints). In the following table the first row corresponds to the first optimization inside the procedure and the second row to the second. If we compare the figures row wise, we can see that that the minimizer is approximately equal in both iterations. This also holds for the remaining optimizations of the procedure. This is a clear hint, that we can find an easier procedure to obtain the next intervention variable and intervention value.

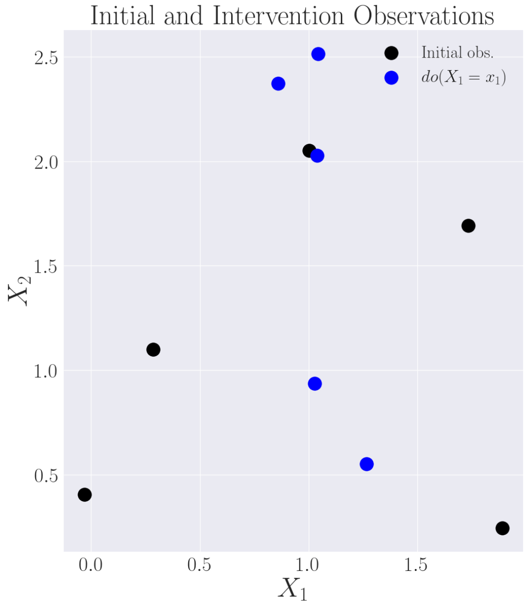


Figure 3.18.: Obs. from initial- and $do(x_1)$-distribution. (Example 3.5).

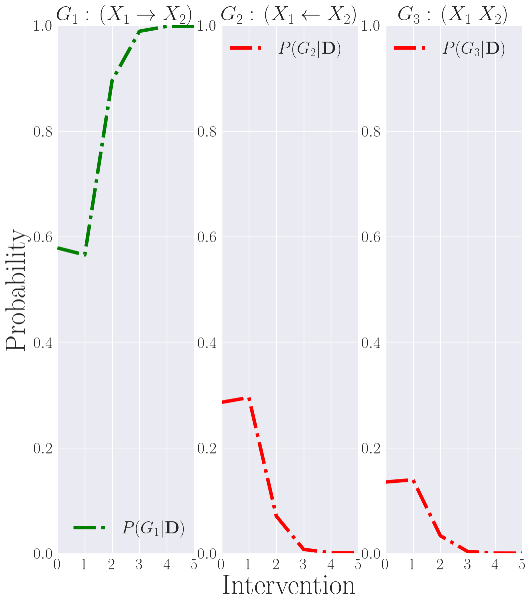
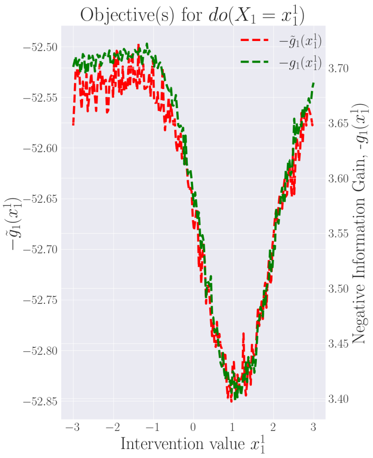
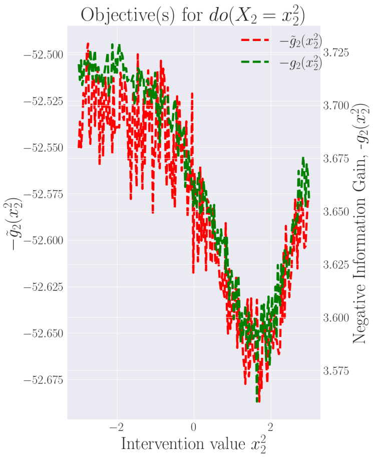


Figure 3.19.: DAG posterior probability, $P(G|\mathbf{D})$. (Example 3.5). $k_{\nu=2.5,\gamma=1.75}$ (Matérn).

| Intervention Sample | Objective function(s) for an intervention on $X_1$ | Objective function(s) for an intervention on $X_2$ |
|---|---|---|
| $(x_1^1, x_2^1) = (1.27, 0.55)$ |  |  |

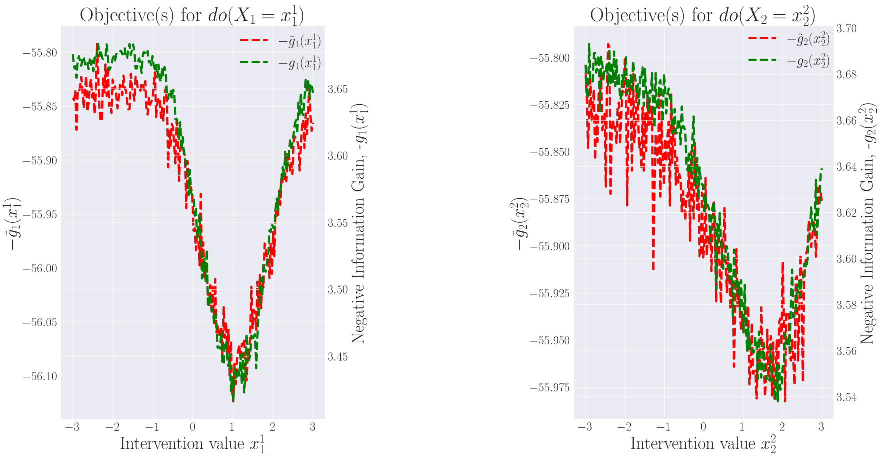
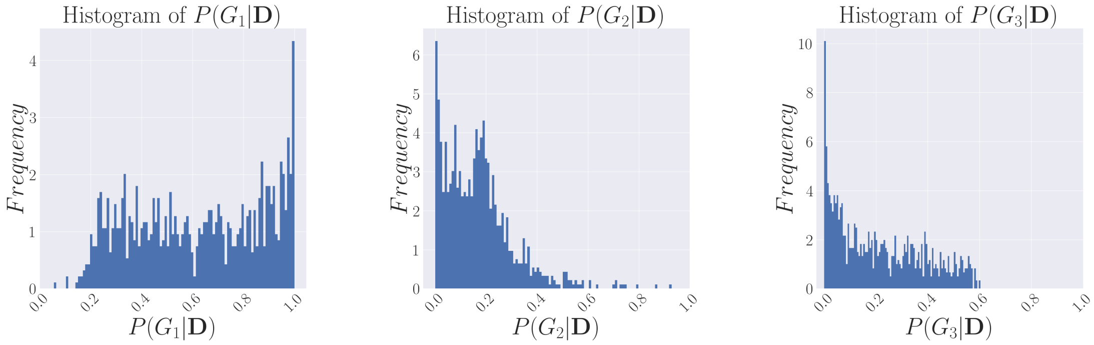
| | |
|---|---|
| $(x_1^1, x_2^1) = (0.86, 2.37)$ |  |

Table 3.3.: Detailed depiction of the optimization procedure for the bivariate case in the model $G_{true}$ : $X_1 \to X_2$, where $X_1 \sim \mathcal{N}(0, 1)$ and $X_2 = 2\tanh(X_1) + \epsilon_2$ with $\epsilon_2 \sim \mathcal{N}(0, 1)$. For the Monte Carlo approximation of the integrals, 5000 sampling points were used.

Next, we illustrate the performance of the implementation on the problem similar as before using histograms of the posterior probabilities of the DAGs. The histograms are build from 1,000 repetitions of the above described experiment.

| | |
|---|---|
| 0 |  |

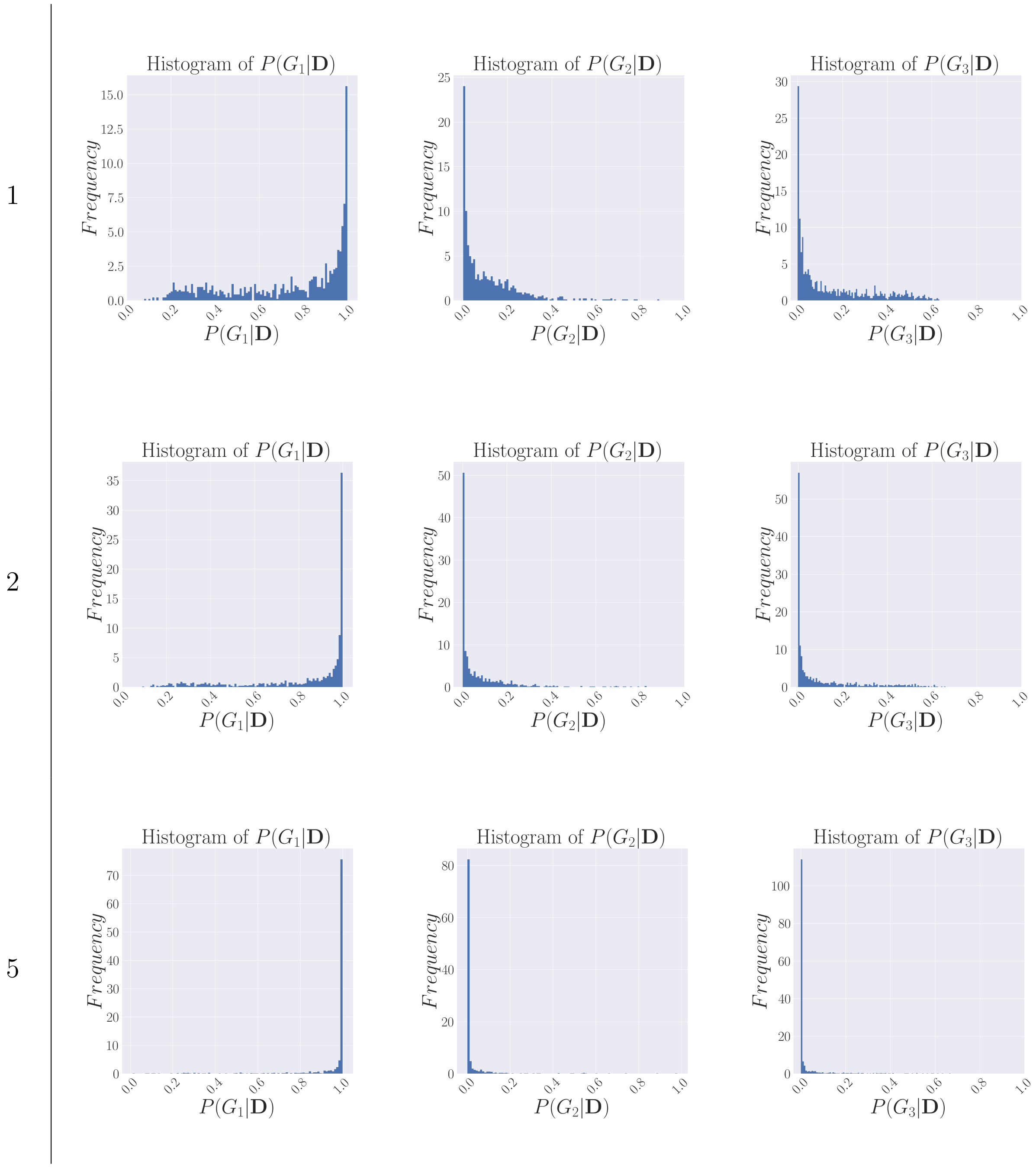
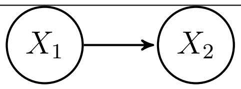


Table 3.4.: Comparison of the empirical distributions of the posterior probabilities of the three different DAGs. (Example 3.6).

In this experiment, the mean of $P(G_1|\mathbf{D})$ after the tenth intervention was $\approx 0.9505$ with a standard deviation of $\approx 0.1467$. If we decided for graph $G_1$ only if $P(G_1|\mathbf{D}) \geq 0.95$, the success rate was 0.843. If we defined failure as $P(G_1|\mathbf{D}) \leq 0.5$, the failure rate was 0.032. The results are similar to the results of the best performing approach in (3.3) but already after the fifth intervention rather than the tenth intervention.

Now, we first re-consider the experimental setup from the previous section (3.4) and afterwards summarize the most important observations of both presented examples.

**Example 3.6** (Sequential optimal interventions (3.4))**.** Similar as in (3.4), consider a situation, where we are given five i.i.d. observations of the true data generating process,

| | Graph | ANM |
|---|---|---|
| $G_{true}$ |  | $X_1 := \epsilon_1,$<br>$X_2 := 2\tanh(X_1) + \epsilon_2,$ |

where $\epsilon_1 \sim \mathcal{N}(0,1)$ and $\epsilon_2 \sim \mathcal{N}(0,0.1)$.
Assume that we specify for $G_1$, $\epsilon_1 \sim \mathcal{N}(0,1)$ and $\epsilon_2 \sim \mathcal{N}(0,0.1)$; for $G_2$, $\epsilon_1 \sim \mathcal{N}(0,0.1)$ and $\epsilon_2 \sim \mathcal{N}(0,1)$ and for $G_3$, $\epsilon_1 \sim \mathcal{N}(0,1)$ and $\epsilon_2 \sim \mathcal{N}(0,1)$ (see 3.2). Further, for the kernel we choose the Matérn kernel with $\nu = 2.5$ and $\gamma = 1.75$ in both models $G_1$ and $G_2$. A typical result of this experiment is summarized in the following figures.

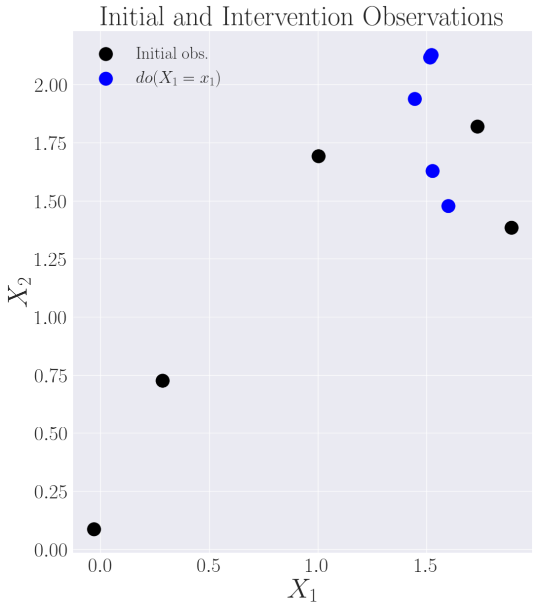


Figure 3.20.: Obs. from initial- and $do(x_1)$-distribution. (Example 3.6).

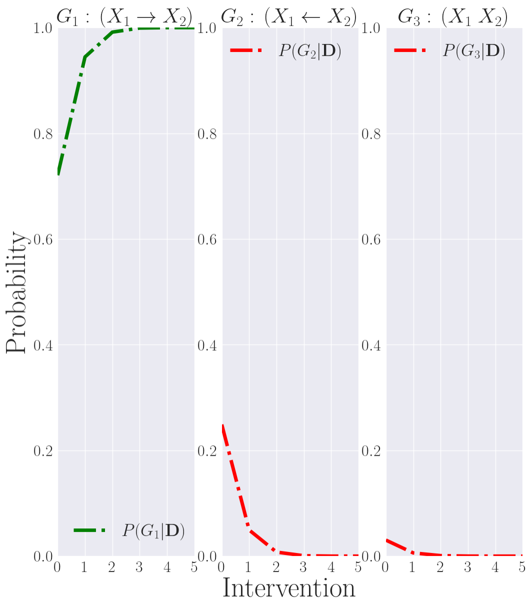


Figure 3.21.: DAG posterior probability, $P(G|\mathbf{D})$. (Example 3.6). $k_{\nu=2.5,\gamma=1.75}$ (Matérn).

The algorithm is able to arrive at the true graph very quickly and achieves very high probability after only 2 interventions. It did no intervention on $X_2$ like in the previous example (3.3). So it seems that in our example it is most informative about the causal directions, to test if our beliefs about the functional relation $f^{(2)}$ are indeed correct. The algorithm chooses a location for the intervention values, where we have observations, the regression function suggests some nonlinearity and it seems to be close to the maximum of the regression function (3.22). It is remarkable, that the implementation intervenes

inside a small region of the $X_1$ domain in every iteration, because with the chosen model variance parameters it would also be “easy” to cast out $G_2$ as the chosen model variance and the observed variance from an intervention on $X_2$ data point would be far off (see 3.2). This can be interpreted as further evidence that it is most informative to confirm the correct functional relation.
We can also see in (3.23) that the confidence band of the ”wrong” functional relation is wider compared to the one in (3.22), what is one reason for the higher likelihood of the true graph. Further we can see, that the empty graph has very little posterior probability already before the first intervention happened. Overall the GP fits and the posterior probabilities are plausible given the available data.

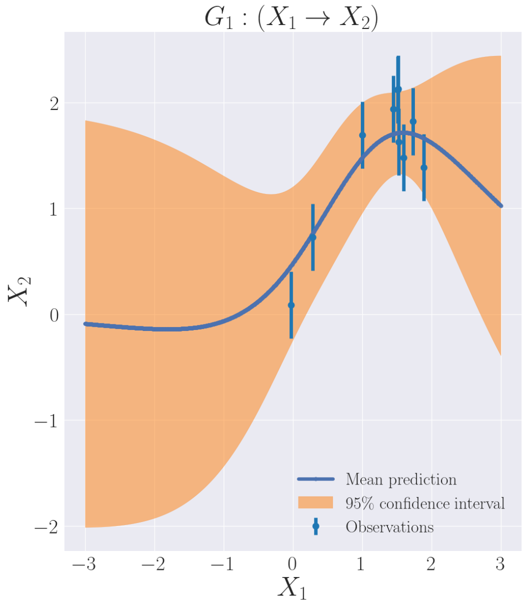


Figure 3.22.: GP fit ($k_{\nu=2.5,\gamma=1.75}$) on initial- and $do(x_1)$-data. (Example 3.6).

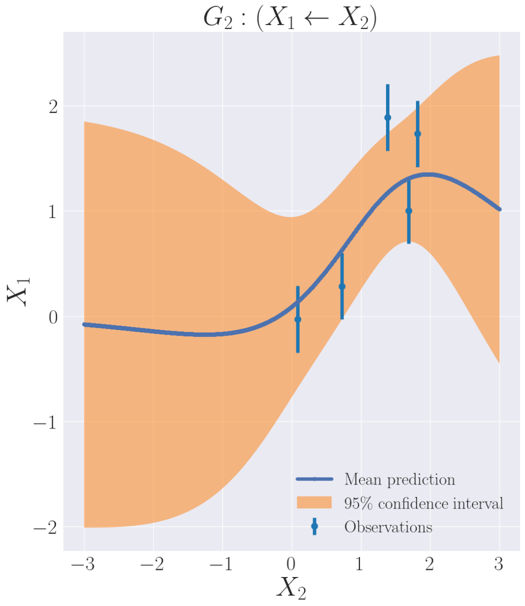


Figure 3.23.: GP fit ($k_{\nu=2.5,\gamma=1.75}$) on initial- and $do(x_2)$-data. (Example 3.6).

We can investigate the optimization procedure more closely. Below we plot the objective functions for an intervention on $X_1$ (left) and for an intervention on $X_2$ (right). The overall minimum is obtained by comparing the two minima $g_1(x_1^{1^*})$ and $g_2(x_2^{2^*})$. We also report the corresponding interventional data point, that was sampled based on the intervention value of the optimization procedure. We can see that the true objective and the simplified objective approximately have the same minimizer. Further they agree very well up to some scaling. Similar as in (3.5), the objective from an intervention on $X_1$ does not change much during the process and we always intervene at the same $X_1$ domain point. One possible explanation for this behavior is that we can efficiently decide on a causal direction based on good knowledge of the regression function on a small region of the domain. The figures also suggest, that there might be a more efficient procedure to find the optimal intervention than approximating the information gain via

Monte Carlo simulations, since the objective functions are extremely similar and have their minimum always in the same region. The following figures are ordered, i.e. the first row of the table corresponds to the optimization based on only the initial observations, the second row corresponds to the optimization based on the initial observations and the first interventional observation and so on.

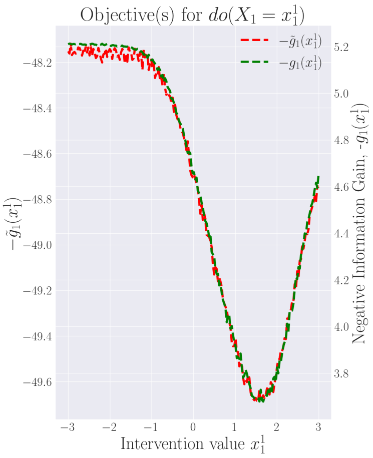
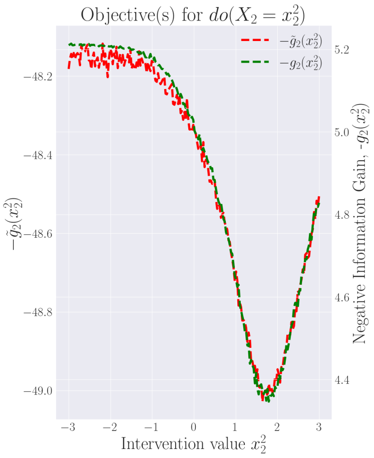
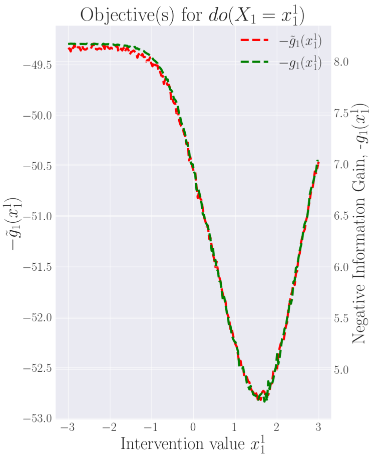
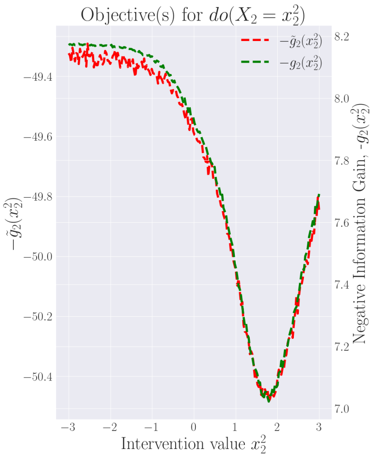
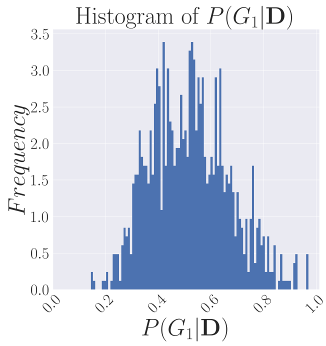
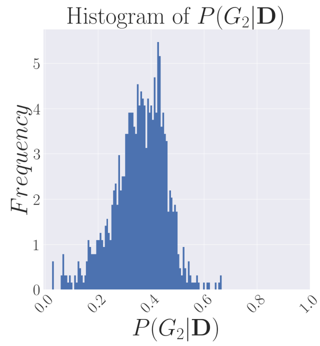
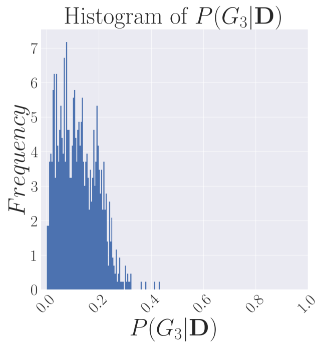
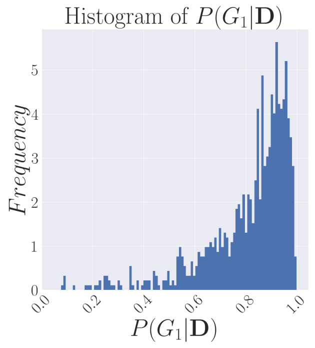
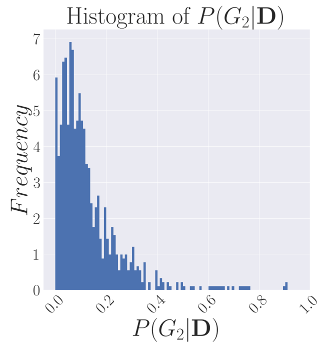
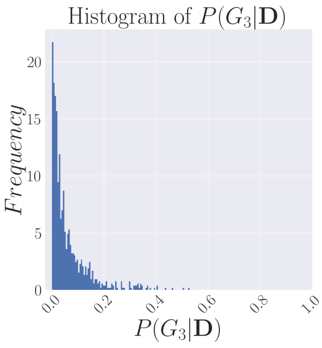
| Intervention Sample | Objective function(s) for an intervention on $X_1$ | Objective function(s) for an intervention on $X_2$ |
|---|---|---|
| $(x_1^1, x_2^1) = (1.63, 1.49)$ |  |  |
| $(x_1^1, x_2^1) = (1.6, 2.15)$ |  |  |

Table 3.5.: Detailed depiction of the optimization procedure for the bivariate case in the model $G_{true} : X_1 \to X_2$, where $X_1 \sim \mathcal{N}(0, 1)$ and $X_2 = 2\tanh(X_1) + \epsilon_2$ with $\epsilon_2 \sim \mathcal{N}(0, 0.1)$. For the Monte Carlo approximation of the integrals, 5000 sampling points were used.

The algorithm was tested repeating the experiment described above 1,000 times and report the posterior probabilities for each graph after each optimal intervention. We already have very concentrated posterior probability for the true DAG, after the second intervention. If we decided for graph $G_1$ only if $P(G_1|\mathbf{D}) \geq 0.95$, the success rate was $\approx 0.98$ after the fifth intervention. Note in (3.2) we had $\approx 0.9714$ after ten alternating interventions. If we consider it to be a failure, if the posterior probability of a graph is less than 0.5, then we would have failed in four out of thousand cases after the fifth intervention. The overall mean of the posterior distribution of the true graph after the fifth intervention was $\approx 0.9928$ with a standard deviation of 0.0554. After the second intervention we had a 95%-success rate of 0.729, a mean (of the posterior distribution of the true graph) of $\approx 0.9382$ and a standard deviation of $\approx 0.1191$. The results are summarized in the following table.

| | | | |
|---|---|---|---|
| 0 |  |  |  |
| 1 |  |  |  |

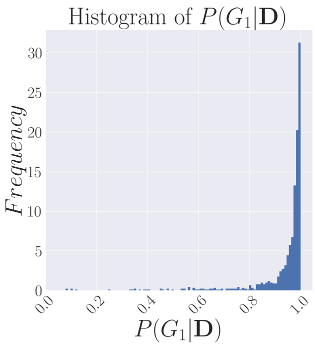
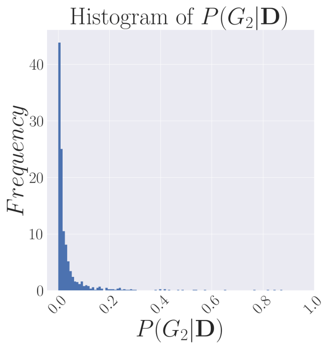
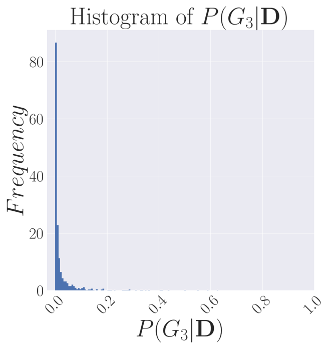
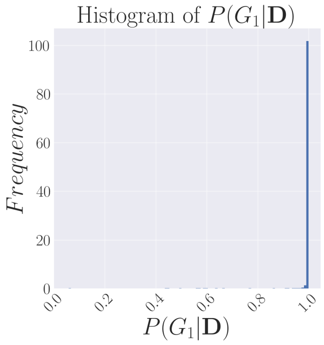
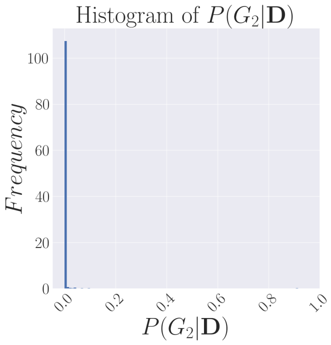
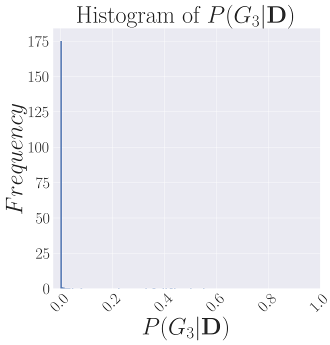
| | | | |
|---|---|---|---|
| 2 |  |  |  |
| 5 |  |  |  |

Table 3.6.: Comparison of the empirical distributions of the posterior probabilities for the three different DAGS. (Example 3.6).

The implementation also kept track of the confidence bound violations of the true functional relation estimate, i.e. for each experiment 100 $X_1$-support points were checked on how often the true function, $2\tanh$, lies outside of the 95% confidence bands of the respective estimate $\hat{f}^{(2)}$. The plot of the confidence bound violations has two peaks that are somewhat symmetric around zero and one minimum at zero. This is the case because the algorithm intervenes always somewhere around one of the two peaks. Since the true function has rather high curvature in these two areas, the regression fit, where we do not intervene, is rather poor. The minimum at zero can be explained because it is near enough at the intervention spots in both cases. The following plot shows the confidence bound violations. The result is another hint that the most efficient way of determining causal relations is to focus on a small domain area, rather than have a good estimate of the functional relation for the whole domain.

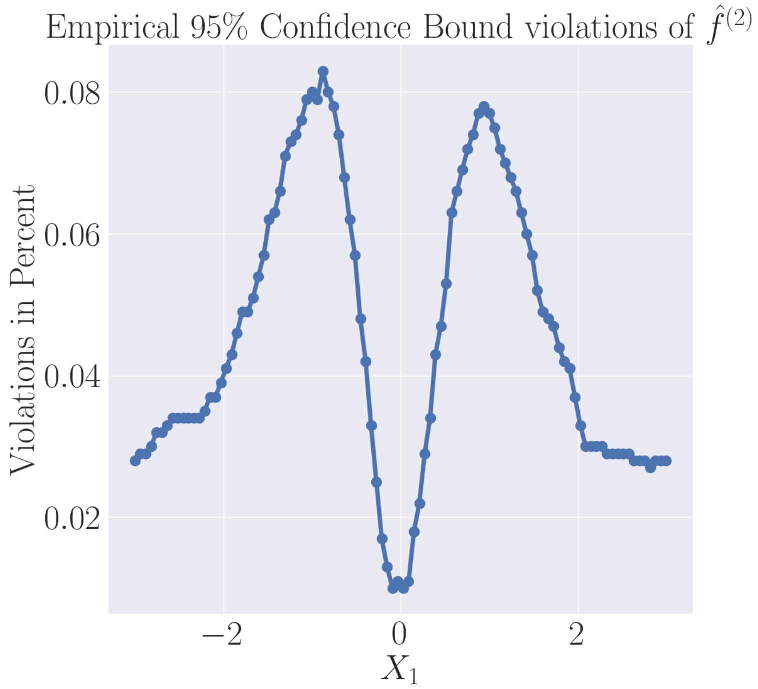


Figure 3.24.: Empirical 95% confidence bound violations of $\hat{f}^{(2)}$ in example (3.6).

But what is still an open question which does not become clear from the example so far, when does the algorithm perform interventions on $X_2$. In the following we present one of the rare cases, where the algorithm performed interventions on $X_2$.

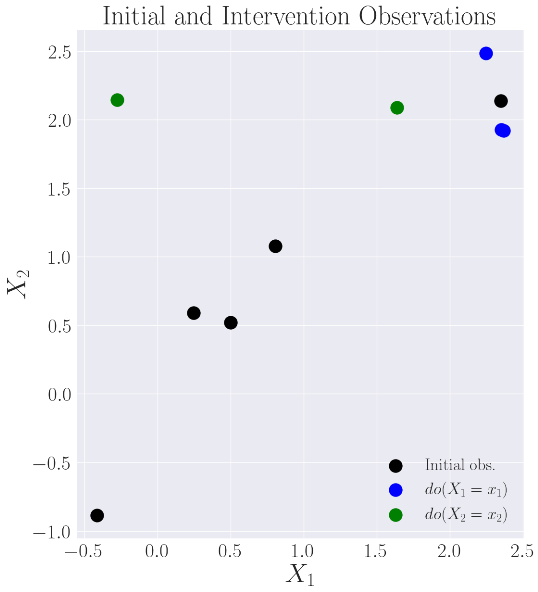


Figure 3.25.: Obs. from initial-, $do(x_1)$- and $do(x_2)$-distribution. (Example 3.6).

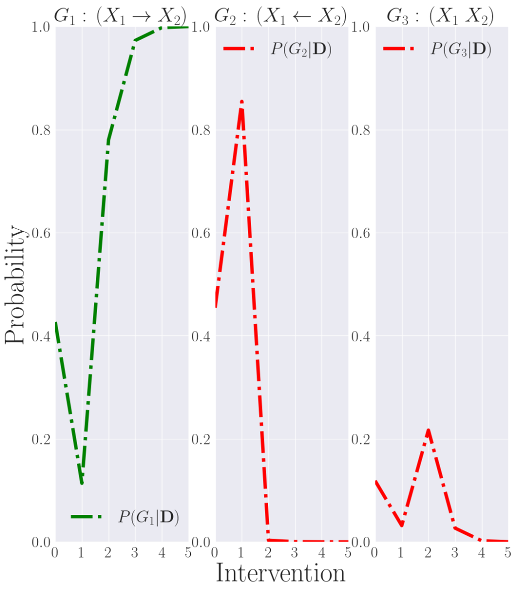


Figure 3.26.: DAG posterior probability, $P(G|\mathbf{D})$. (Example 3.6). $k_{\nu=2.5,\gamma=1.75}$ (Matérn).

The initial observations seem to be perfectly linear and thus (for the initial data) there exists a plausible backward model. The posterior probability given the initial observations suggest, that $G_2$ is the most plausible model at the initial stage of the algorithm. Then, the green point in line with the (linear) black points was sampled from an intervention on $X_2$. This seems to confirm that $G_2$ is the correct model. But then the green point that is far of was sampled and suddenly $G_2$ became completely implausible. But for this example the implementation finds the true graph based on the "miss specified" variance parameter in $G_2$. The two data point outside the confidence band in (3.28) nicely illustrate this.

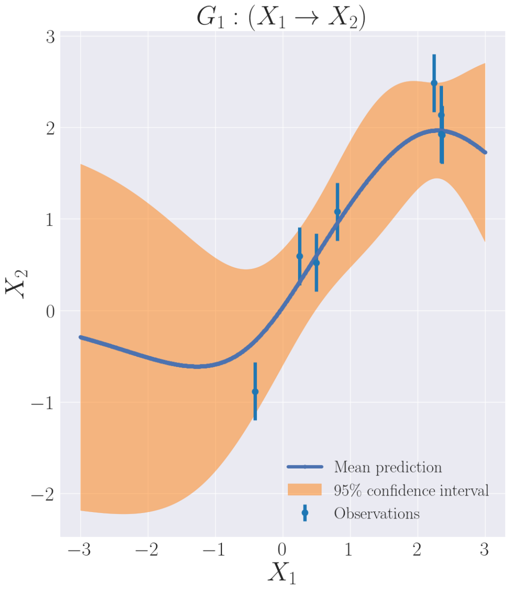


Figure 3.27.: GP fit ($k_{\nu=2.5,\gamma=1.75}$) on initial- and $do(x_1)$-data. (Example 3.6).

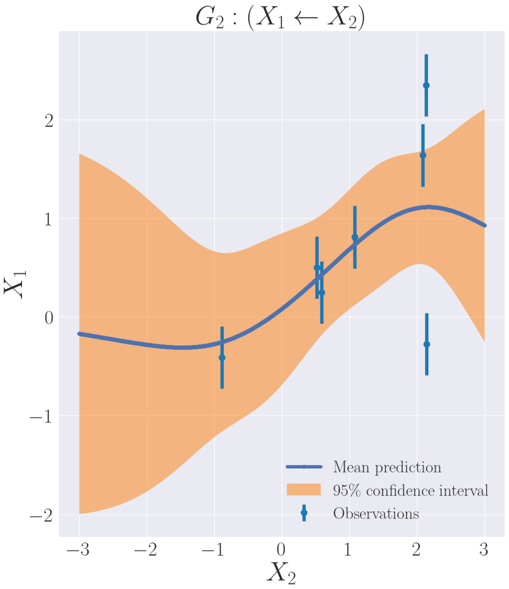


Figure 3.28.: GP fit ($k_{\nu=2.5,\gamma=1.75}$) on initial- and $do(x_2)$-data. (Example 3.6).

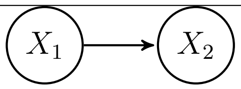
To conclude the two above considered examples, we summarize the most important findings.

- When using optimal interventions, the algorithm needs five interventions to have a very concentrated posterior probability on the true data generating DAG. In the case of random independent interventions it needed 10 interventions for a similar result.
- The objective functions of the procedure are very similar in every iteration.
- The algorithm chooses intervention values mostly on one and the same variable in a small domain region, where the regression function has nonlinear curvature.
- The algorithm chooses intervention values at locations where we already have at least one observation. But it does not necessarily choose the location where we have most observations.

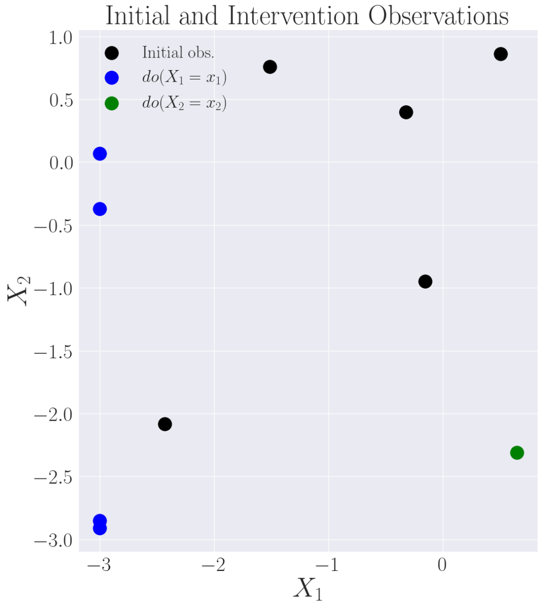
Based on the theory on additive noise models, it is possible, that interventions are most informative if they are chosen in areas of the function support, where the true underlying function is shaped very nonlinear. The idea is, that the algorithm chooses intervention values at locations where it currently believes no backward model is possible. This would explain why in the previous example it is enough to intervene only on one variable always at the same domain area. This would imply that for every ANM which is identifiable from purely observational data, it should be enough to intervene only on parent (cause) nodes.

We now consider a linear Gaussian Case which is non-identifiable from purely observational data (2.48). This case is in particular interesting because we can see how the algorithm chooses interventional observations that provide the little bit of more information, compared to purely observational data, we need to infer the correct causal direction.

**Example 3.7** (Linear Gaussian)**.** Consider a situation, where we are given five i.i.d. observations of the true data generating process,

| | Graph | ANM |
|---|---|---|
| $G_{true}$ |  | $X_1 := \epsilon_1,$<br>$X_2 := \frac{X_1}{2} + \epsilon_2,$ |

where $\epsilon_1 \sim \mathcal{N}(0,1)$ and $\epsilon_2 \sim \mathcal{N}(0,1)$.
Assume that we specify for $G_1$, $\epsilon_1 \sim \mathcal{N}(0,1)$ and $\epsilon_2 \sim \mathcal{N}(0,1)$; for $G_2$, $\epsilon_1 \sim \mathcal{N}(0,1)$ and $\epsilon_2 \sim \mathcal{N}(0,1)$ and for $G_3$, $\epsilon_1 \sim \mathcal{N}(0,1)$ and $\epsilon_2 \sim \mathcal{N}(0,1)$ (see 3.1). Further for the kernel used in the models $G_1$ and $G_2$ we choose the Matérn kernel with $\nu = 2.5$ and $\gamma = 1.75$. A result of this experiment is summarized in the following figures.
The implemented algorithm tends to the correct graph with high probability. In the plots of the regression estimates, we can see that the success relies on balancing between “rule out that it is DAG $G_2$” (by intervening at a value of $X_2$ where we are rather sure about the functional relation) and “make sure the functional relation in the first DAG is correct” (by testing the regression function). We can also see that the goal of the optimization is to determine the causal direction and not to get a good estimate of the underlying functional relation.



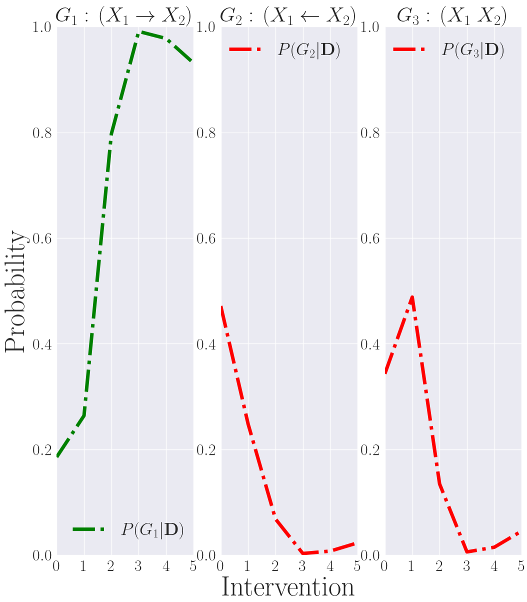
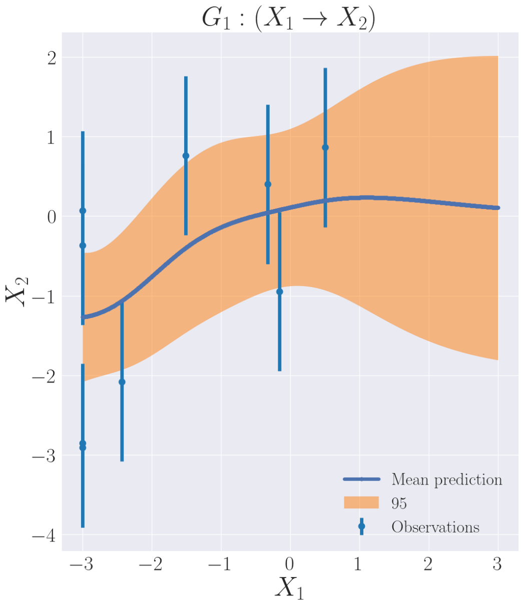


Figure 3.29.: Obs. from initial-, $do(x_1)$- and $do(x_2)$-distribution. (Example 3.7).

Figure 3.30.: DAG posterior probability, $P(G|\mathbf{D})$. (Example 3.7). $k_{\nu=2.5,\gamma=1.75}$ (Matérn).



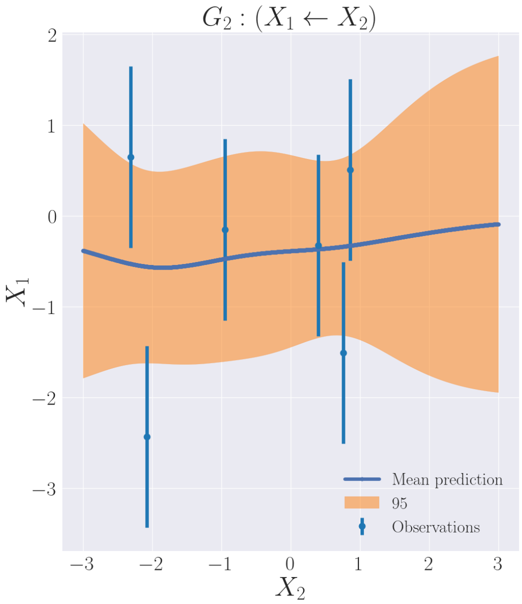


Figure 3.31.: GP fit ($k_{\nu=2.5,\gamma=1.75}$) on initial- and $do(x_1)$-data. (Example 3.7).

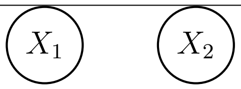
Figure 3.32.: GP fit ($k_{\nu=2.5,\gamma=1.75}$) on initial- and $do(x_2)$-data. (Example 3.7).

To conclude the example, we augment the summary of the most important finding about the behavior of the implementation.

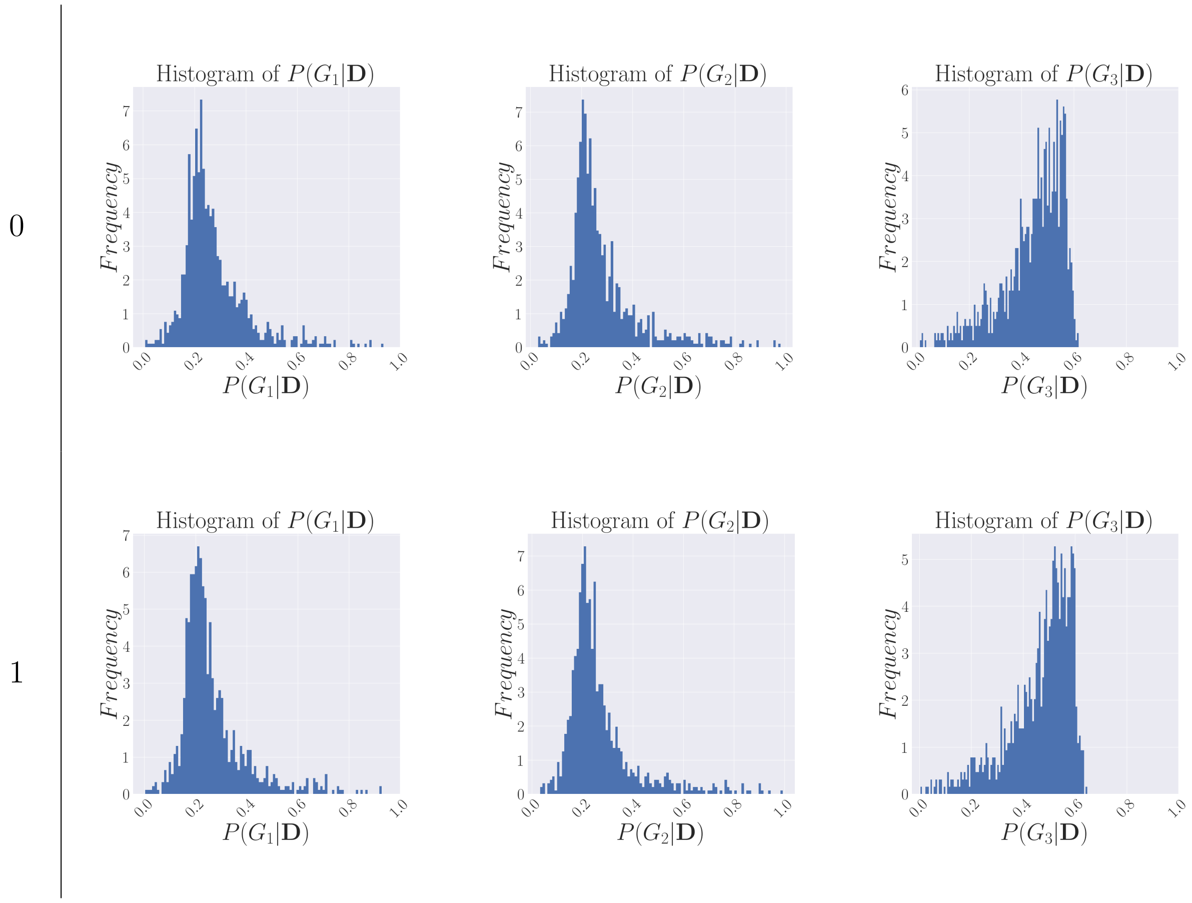
- Whenever there is a valid backward model for a parent child relation of two random variables, it is beneficial (for the implemented algorithm) to intervene on both.

We have evaluated the algorithm on its ability to find functional relations between random variables. What is left is the ability to detect independence between the random variables. The following examples tries to close this gap.

**Example 3.8** (Independent Random Variables)**.** Consider a situation, where we are given five i.i.d. observations of the true data generating process,

| | Graph | ANM |
|---|---|---|
| $G_{true}$ |  | $X_1 := \epsilon_1,$<br>$X_2 := \epsilon_2,$ |

where $\epsilon_1 \sim \mathcal{N}(0,1)$ and $\epsilon_2 \sim \mathcal{N}(0,1)$ independent.
Assume that we specify for each DAG all noise terms to be $\mathcal{N}(0,1)$. Further for the kernel used in the models $G_1$ and $G_2$ we choose the Matérn kernel with $\nu = 2.5$ and $\gamma = 1.75$. We conducted the same simulation study as in (3.6) to check if the algorithm can identify when there is no relation between the two random variables.



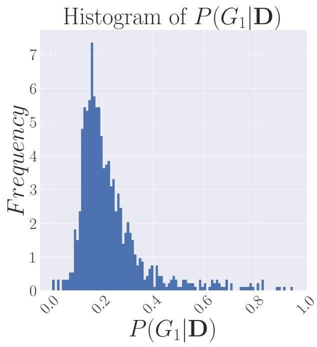
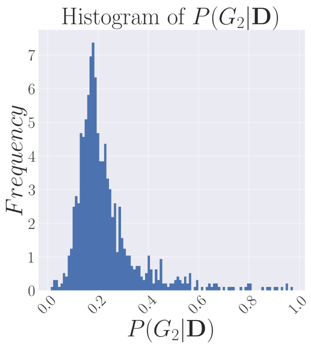
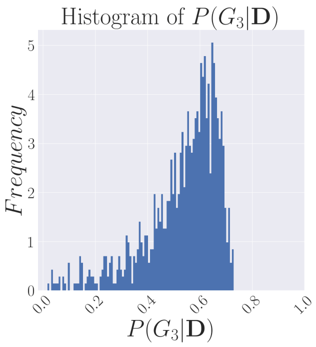
| | | | |
|---|---|---|---|
| 5 |  |  |  |

Table 3.7.: Comparison of the empirical distributions of the posterior probabilities for the three different DAGs. (Example 3.6).

The implementation tends to detect that the empty graph is likely to be the data generating model. But it does perform significantly worse than in examples where we had a functional relation between the random variables. The mean of the posterior distribution of the true DAG after the fifth intervention was only $\approx 0.2323$ with a standard deviation of $\approx 0.1344$. There seems to be an upper bound for the posterior probability. If we apply this method in practice we should always use another method to double-check for independence (for example the one described in [33]).

We conclude the section on the bivariate case and summarize the most important findings. Despite for the information gain we have closed forms for all quantities we need in the implementation. If interventional observations are possible to perform, we can infer the causal direction based on a comparably low number of observations. In the section on random independent interventions we found hints, that it is most informative to intervene on causes. We found more experimental evidence for it in the section on optimal random interventions. Based on the considered examples, the implemented algorithm tries to intervene on the causes in areas where it already has observational evidence; it tries to confirm likely causal relations (in areas where no backward models are possible). If there exist valid backward models it tends to intervene on parent and child. We also observed that the objective functions of the optimization in each iteration are very similar. This can be seen as a hint that there is an equivalent but much simpler way to choose the optimal intervention value, than optimizing the information gain. We have also seen that the algorithm only performs okay in detecting independencies.

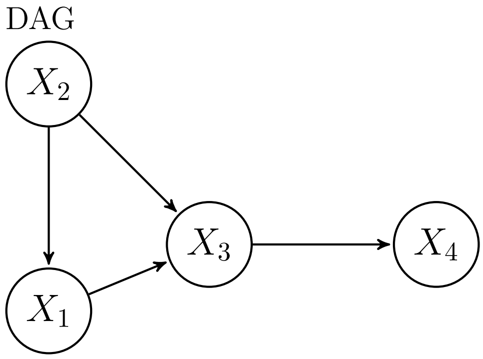
## 3.3. Fourvariate Case

In this section we generalize the procedure from the bivariate case to the fourvariate case. We go through all necessary computation steps of the procedure by considering an example and a corresponding pseudo code. The pseudo code can easily be transferred to settings with more than four variables and it provides a solid foundation for the basic idea of the approach. We conclude the section by describing the calculations for the optimization procedure in detail. From these considerations the computational burden becomes clear which is a primer to the last section of this Chapter, that deals with strategies to overcome the computational burden.

### 3.3.1. Generate a list of DAGs

Before we can start with the computations in our approach we must have a procedure that generates us a list with all 543 possible DAGs containing four variables [31, Appendix B]. We do so naively, by sampling a candidate, check if it is a valid DAG and include it to our list, if the candidate is not already included. This works just fine in the case of only four variables.

We can depict a DAG in a matrix representation (usually called adjacency matrix). The variables are enumerated, the first row stands for the first variable and so on, the same for the columns. We encode incoming edges by a 1 and read the matrix row wise, i.e. if we have a 1 in the second column of the first row, then the corresponding DAG has a directed edge from variable 2 to variable 1. The following shows an example.



Matrix

$$\begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

Matrices corresponding to valid DAGs are permutation similar to a lower triangular binary matrix (having ones on the lower triangular). They have a zero diagonal, must not be symmetric, must have at least one source node (row of zeros) and at least one sink node (column of zeros). Further, any sub-matrix we obtain from deleting the $i^{\text{th}}$ column and row must also satisfy the previously mentioned conditions. Based on this simple rules we can already form our control agent algorithm, who checks the four conditions for every sub-matrix.

**Algorithm 3** DAG Control Agent

```
Require: $G_{candidate}$, $G_{\text{list}}$    ▷ candidate matrix and list of already accepted matrices
  for $G$ in $\{G_{candidate} \cup$ all sub-matrices$\}$ do
      if $G$ is symmetric return false    ▷ Check symmetry
      if sum of $tr(G) \neq 0$ return false    ▷ Check diagonal entries
      if all row sums of $G$ are $> 0$ return false    ▷ Check for a source node
      if all column sums of $G$ are $> 0$ return false    ▷ Check for a sink node
      if $G \in G_{\text{list}}$ return false    ▷ Check if the candidate is already in our list
  end for
  return true
```

As previously mentioned, it was possible to generate all graphs by randomly generating candidates and check if they are valid proposals. The following random matrix generator exploits the characterizations of DAG matrices and that there are more graphs having a relatively low number of edges than graphs having a high number of edges. While there are more efficient methods exploiting permutation matrices, the proposed procedure can be easily implemented and is sufficient for the low dimension case.

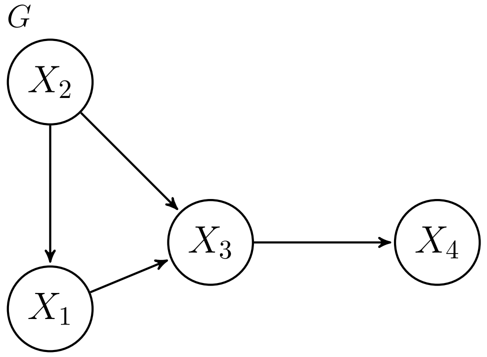
**Algorithm 4** Generate DAG Candidate

```
Require: $h_1$, $h_2$    ▷ Parameters for the Bernoulli distribution
  $G_{candidate} = \mathbf{0} \in \mathbb{R}^{4\times4}$
  $ind$ = list of index tuples, $(i_1, i_2)$, of lower triangular matrix
  count-ones = 0
  for $i$ in $ind$ do
      sample = $Bernoulli\left(\frac{6-\text{count-ones}}{h_1}\right)$
      count-ones = sample + count-ones
      $G_{candidate}(i_1, i_2)$ = sample
      if sample = 1 then
          $G_{candidate}(i_2, i_1) = |\text{sample} - 1|$
      else
          sample = $Bernoulli\left(\frac{6-\text{count-ones}}{h_2}\right)$
          count-ones = sample + count-ones
          $G_{candidate}(i_2, i_1)$ = sample
      end if
  end for
return $G_{candidate}$
```

Because we know the total number of DAGs for a given number of variables, we can sample a new candidate graph and check if we can include it to our list in each iteration and while loop until we reach the respective total number of possible DAGs. To speed up the procedure a little bit we can start the procedure with a list of graphs containing the extreme cases, i.e. the empty graph and the two graphs having ones on the upper resp. lower triangular matrix.

### 3.3.2. Likelihood given a DAG

Before we start with an example DAG and compute the likelihood thereof, note that we can infer regression functions with any (finite) number of inputs easily using GP regression. More precisely for our setting, for any finite natural number $l$ we can infer functions of the form $f : \mathbb{R}^l \mapsto \mathbb{R}$. The relevant information we input into the Matérn kernel is the distance between the input data points. Close points have high covariance and distant points have a low covariance. Therefore, lifting the functional estimation part from two to any number of variables is straight forward.
Consider the DAG form the previous section, define it as $G$ and consider the following ANM, where noise terms are normally distributed, i.e. $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ for $i \in \{1, \dots, 4\}$.



ANM

$$
\begin{aligned}
X_1 &:= f^{(1)}(X_2) + \epsilon_1 \\
X_2 &:= \epsilon_2 \\
X_3 &:= f^{(3)}(X_1, X_2) \\
X_4 &:= f^{(4)}(X_3) + \epsilon_4
\end{aligned}
$$

Suppose we have $N$ i.i.d. (initial) observations of the above ANM, $x = (x_1, x_2, x_3, x_4) \in \mathbb{R}^{4 \times N}$. Similar as before we can very efficiently compute the likelihood as a product of Gaussian likelihoods,

$$
p(x_1, x_2, x_3, x_4|G) = \underbrace{p(x_2|G)}_{\mathcal{N}(0,\sigma_2^2)} \underbrace{p(x_1|x_2, G)}_{\mathcal{N}(0,k_{X_2X_2}+\sigma_1^2 I_N)} \underbrace{p(x_3|x_1, x_2, G)}_{\mathcal{N}(0,k_{(X_1,X_2)(X_1,X_2)}+\sigma_3^2 I_N)} \underbrace{p(x_4|x_3, G)}_{\mathcal{N}(0,k_{X_3X_3}+\sigma_4^2 I_N)} .
$$

Now consider an intervention on $X_2$ where we observed the data point from the intervention distribution, $x^2 = (x_1^2, x_2^2, x_3^2, x_4^2)$. Because we assume to obtain data sequentially, the likelihood of the intervention data can be computed efficiently since we know the distributions of the GP fits. The fits are estimated from all previously obtained data that is relevant for the respective functional relation. We consider an intervention data point relevant for the estimation of the regression function, if it contains information about the functional relation. For one specific node $j$, that are all observations obtained from intervention distributions except from the distribution where we intervene on $j$. In the above example, for $f^{(3)}$, this would be all data points obtained from intervention distributions except $do(X_3 = x_3^3)$ samples. We signal that a functional estimate depends on all currently available relevant data by a tilde, i.e., $f^{(4)}_{\tilde{X}_3}$. Coming back to our example, for a given $x_2^2$ we have the following distribution of the functional estimates

$$
\begin{aligned}
&f^{(1)}_{\tilde{X}_2}(x_2^2)|\tilde{x}_1, \tilde{x}_2, x_2^2 \sim \mathcal{N}(\underbrace{k_{x_2^2 \tilde{X}_2}(k_{\tilde{X}_2 \tilde{X}_2} + \sigma_1^2 I_{\tilde{N}})^{-1}\tilde{x}_1}_{:=\tilde{\mu}_{G^{(1)}}(x_2^2)}, \underbrace{k_{x_2^2 x_2^2} - k_{x_2^2 \tilde{X}_2}(k_{\tilde{X}_2 \tilde{X}_2} + \sigma_1^2 I_{\tilde{N}})^{-1} k_{\tilde{X}_1 x_2^2}}_{:=\tilde{\sigma}^2_{G^{(1)}}(x_2^2)}), \\
&f^{(3)}_{(\tilde{X}_1, \tilde{X}_2)}(x_1^2, x_2^2)|\tilde{x}_1, \tilde{x}_2, \tilde{x}_3, x_2^2, x_1^2 \sim \mathcal{N}\left(\tilde{\mu}_{G^{(3)}}(x_1^2, x_2^2), \tilde{\sigma}^2_{G^{(3)}}(x_1^2, x_2^2)\right), \\
&f^{(4)}_{\tilde{X}_3}(x_3^2)|\tilde{x}_3, \tilde{x}_4, x_3^2 \sim \mathcal{N}\left(\tilde{\mu}_{G^{(4)}}(x_3^2), \tilde{\sigma}^2_{G^{(4)}}(x_3^2)\right).
\end{aligned}
$$

And we can calculate the likelihood via

$$
p(x^2, \tilde{x}|G) = \underbrace{p(x_1^2|\tilde{x}_1, \tilde{x}_2, x_2^2, G)}_{\mathcal{N}\left(\tilde{\mu}_{G^{(1)}}(x_2^2), \tilde{\sigma}^2_{G^{(1)}}(x_2^2) + \sigma_1^2\right)} \underbrace{p(x_3^2|\tilde{x}_1, \tilde{x}_2, \tilde{x}_3, x_2^2, x_1^2, G)}_{\mathcal{N}\left(\tilde{\mu}_{G^{(3)}}(x_1^2, x_2^2), \tilde{\sigma}^2_{G^{(3)}}(x_1^2, x_2^2) + \sigma_3^2\right)} \underbrace{p(x_4^2|\tilde{x}_3, \tilde{x}_4, x_3^2, G)}_{\mathcal{N}\left(\tilde{\mu}_{G^{(4)}}(x_3^2), \tilde{\sigma}^2_{G^{(4)}}(x_3^2) + \sigma_4^2\right)} \; p(\tilde{x}|G).
$$

Since we assume the sequential factorization of the likelihood, we can store the current likelihood after each iteration for each graph and use it to compute the next likelihood. Thus, given a graph, in each step of the procedure, we fit a GP model based on one observation more than the previous fit, compute the likelihoods of the new data point (which is not yet included for the GP fit) and obtain the overall likelihood as a product of the respective previous likelihood with the likelihood (given the functional relation estimate based on the previous observations) of the new data point. Since all likelihoods involved are Gaussian and the GP prediction relies on function evaluations and solving one matrix equation, the computation is very efficient. The following pseudo code is one possibility to implement the procedure. We work with graph objects, that contain the matrix representing the DAG, the relevant parameters for the ANM (error variances and parameters for the Matérn kernel). Further, for each possible intervention we must store the corresponding samples separately. During the process we keep track of the likelihoods up to the current iteration with a list $C_G$ containing the likelihoods of the respective iteration. All this information is stored in the graph object.

**Algorithm 5** Likelihood $p(D_j, \tilde{D}|G)$

```
Require: $x^j$, $j$, $G_{\text{object}}$  ▷ intervention data point, intervention variable & graph object
  $G$ = graph                                                  ▷ stored as matrix in $G_{\text{object}}$
  $C_G$ = likelihood for the graph before seeing $x^j$            ▷ stored in $G_{\text{object}}$
  $ind$ = list of variable indices
  $res$ = list of zeros with same length as $ind$
  for $i$ in $ind$ do
      $dim$ = row sum of $i^{th}$ row of $G$
      if $dim > 0 \vee i \neq j$ then
          covariates = relevant data for GP fit given $G$
          $x^j_{\text{pred}}$ = intervention observations of same variables as in covariates
          target = relevant observations of $i^{th}$ variable for GP fit given $G$
          Fit GP model with covariates and target
          $\tilde{\mu}_{G^{(i)}}(x^j_{\text{pred}}), \tilde{\sigma}^2_{G^{(i)}}(x^j_{\text{pred}})$ = GP prediction of $x^j_{\text{pred}}$
          $res[i]$ = likelihood of $x^j_i$ w.r.t. $\mathcal{N}\left(\tilde{\mu}_{G^{(i)}}(x^j_{\text{pred}}), \tilde{\sigma}^2_{G^{(i)}}(x^j_{\text{pred}}) + \sigma_i^2\right)$
      else
          if $i \neq j$ then
              $res[i] = 1$      ▷ perfect intervention has Dirac density, could be generalized
          else
              $res[i]$ = likelihood of $x^j_i$ w.r.t. $\mathcal{N}(0, \sigma_i^2)$          ▷ Source node density
          end if
      end if
  end for
  $p(D_j, \tilde{D}|G) = \text{prod}(res) \cdot C_G$
  return $G_{\text{object}}$ augmented with $p(D_j, \tilde{D}|G)$ and $x^j$
```

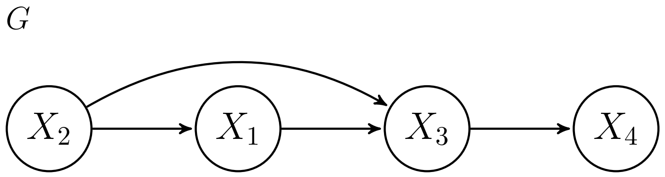
Recall the posterior has the form,

$$p(G|D_j, \tilde{D}) = \frac{p(D_j, \tilde{D}|G)p(G)}{\sum_{\tilde{G} \in \mathcal{G}} p(D_j, \tilde{D}|\tilde{G})p(\tilde{G})}. \tag{3.16}$$

In this work we always assume to have no prior believes about the true underlying graph and therefore model the graph prior using a uniform distribution. Thus, we must compute the above likelihood for all graphs. This still runs fairly efficient and is not causing any computational trouble in the four variable case. But because the number of graphs grows super exponentially with the number of nodes this approach becomes computationally intractable quickly. At this point we may realize that it could be possible to obtain a closed form of the posterior when we use a normal distribution as prior for the graphs (because we know the likelihood is a product of Gaussians). But since the graph space is discrete this will most probably cause some serious measurability problems. Though, it may be possible to develop some closed form for the posterior starting from this idea. Up to this point the procedure is computationally very efficient and very well suitable for causal discovery. We can efficiently include intervention data and compute posteriors quickly, even for larger systems.

### 3.3.3. Sample from the Intervention Distribution

We can approximate the information gain in graphs using Monte Carlo simulations. For this we must sample from the intervention distribution $\mathbf{P}_{\mathbf{X}_{-j}|G_{\text{obj.}},do(X_j=x)}$. In our approach we have (updated) beliefs about the underlying ANMs in every iteration. Each graph implies a topological ordering of the random variables [3, Proposition 1.4.3.], [10, Topological sort 22.4]. For the example from the previous section we have

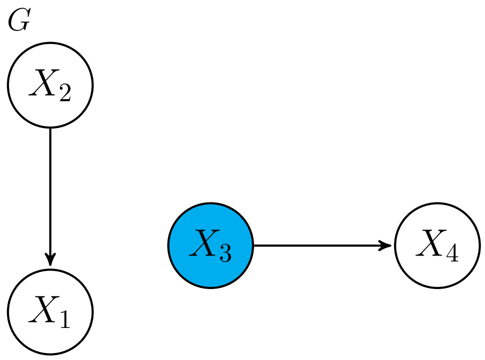


Topological Ordering

$$\tau = [2, 1, 3, 4]$$

We can sample from the underlying ANM by first simulating source nodes, then simulate the nodes that have incoming edges only from source nodes using our current beliefs about the functional relation and the respective error term. With these observations of the first order we proceed through the topological order until we reach the sink nodes. When we sample from a (perfect) intervention distribution (given a graph), we can use the same procedure, but with a manipulated graph that has all incoming edges removed from the intervention variable and setting the value of the intervention variable to the chosen value of $x$.

Consider again the setting from above. The following example presents how to sample from the intervention distribution $do(X_3 = x)$. The data points are sampled sequentially top to bottom.



Sample Distribution

$$\begin{aligned}
&x_2^3 \sim \mathcal{N}\left(0, \sigma_2^2\right)\\
&x_1^3 \sim \mathcal{N}\left(\tilde{\mu}_{G^{(1)}}(x_2^2), \tilde{\sigma}^2_{G^{(1)}}(x_2^2) + \sigma_1^2\right)\\
&x_3^3 = x\\
&x_4^3 \sim \mathcal{N}\left(\tilde{\mu}_{G^{(4)}}(x), \tilde{\sigma}^2_{G^{(4)}}(x) + \sigma_4^2\right)
\end{aligned}$$

Note that we need the same mean and variance functions that shape the distributions as in the previous section. There we have used the corresponding distributions to compute likelihoods. Here we use the corresponding distributions to sample data points. In the following, one possible way to implement the procedure is presented.

**Algorithm 6** Sample from $\mathbf{P}_{\mathbf{X}_{-j}|G_{\text{obj.}},do(X_j=x)}$

```
Require: $x$, $j$, $G_{\text{object}}$    ▷ intervention value, intervention variable & graph object
  $G$ = graph    ▷ stored as matrix in $G_{\text{object}}$
  $ind$ = list of variable indices
  $x^j_{\text{sample}}$ = list of zeros with same length as $ind$
  $x^j_{\text{sample}}[j] = x$
  $dim$ = row sum of $i^{th}$ row of $G$
  $\tau$ = topological ordering of $G$
  for $i$ in $\tau$ do
    if $dim[i] = 0 \vee i \neq j$ then
      $x^j_{\text{sample}}[i]$ = sample from $\mathcal{N}(0, \sigma_i^2)$
    else
      if $i \neq j$ then
        cov = indices from incoming edges of variable $i$
        $\tilde{\mu}_{G^{(i)}}(x^j_{\text{sample}}[\text{cov}]), \tilde{\sigma}^2_{G^{(i)}}(x^j_{\text{sample}}[\text{cov}])$ = GP prediction of $x^j_{\text{sample}}[\text{cov}]$
        $x^j_{\text{sample}}[i]$ = sample from $\mathcal{N}\left(\tilde{\mu}_{G^{(i)}}(x^j_{\text{sample}}[\text{cov}]), \tilde{\sigma}^2_{G^{(i)}}(x^j_{\text{sample}}[\text{cov}]) + \sigma_i^2\right)$
      end if
    end if
  end for
  return $x^j_{\text{sample}}$
```

Note that sampling from the intervention distribution is very tractable for our approach since we only sample from Gaussian distributions, for which there exist very efficient algorithms readily implemented.

### 3.3.4. Optimization

In the previous sections we prepared all ingredients necessary for the optimization procedure. Recall, that we want to solve

$$(j^*, x^*) = \underset{j \in \{1,\dots,d\},\, \in \mathcal{X}_j}{\text{argmax}} \; \mathbb{E}_G \left[ \mathbb{E}_{\mathbf{X}_{-j}|G,do(X_j=x)} \left[ \log \left( p_{G|\mathbf{X}_{-j},do(X_j=x)}(G|\mathbf{x}_{-j}, do(X_j = x)) \right) \right] \right].$$

We approximate the nested expectation with a Monte Carlo simulation, sampling $m$ samples $D_j^{(m)}$ from $\mathbf{P}_{\mathbf{X}_{-j}|G_{\text{obj.}},do(X_j=x)}$ using (6) and calculate, for a given graph $\tilde{G}$,

$$\frac{1}{M} \sum_{m=1}^{M} \log(p_{G|\mathbf{X}_{-j},do(X_j=x)}(\tilde{G}|D_j^{(m)}, \tilde{D}))$$

using (5). The evaluation of $p(\tilde{G}|D_j^{(m)}, \tilde{D}))$ requires the computation of $p(D_j^{(m)}, \tilde{D}|G)$ for all graphs in our hypothesis space. And we must repeat the computation for every single sample $D_j^{(m)}$. To have a reasonable estimate we should use a relatively large number of samples $m$. In total the computational burden of a single evaluation of the objective function is already immense. In the case of only four variables it requires 543*$m$ times the execution of (6) and 543*$m$*543 times the execution of (5). The procedure can be parallelized, but still my computer was not able to perform an optimization of the objective function in acceptable time.

The results that were presented for the bivariate case suggest, that there may be a procedure which is approximately equivalent to maximizing the information gain and can be obtained much cheaper in terms of required computations. Since the only computational problem of the procedure really is the optimization part, there is hope to get the approach working for a large number of variables. The next section presents strategies to overcome the computational challenges.

## 3.4. Heuristic Strategies to Overcome Computational Challenges

We here discuss possible strategies to overcome the computational challenges pointed out in (3.3.4). The following ideas evolved during the work on the thesis and are presented in a heuristic way. The first strategy is motivated by the literature on Bayesian Optimization and tries to exploit the information nested in the regression fits. The second strategy is to use a well known and widely used statistical technique for computing expectations, Monte Carlo Markov Chain integration (MCMC). The third strategy is more practical and does almost the same computations as the original approach but restricts the DAG space in the optimization. The general approach itself and the proposed strategies to improve the computation time are hard to implement since they not only require lengthy computations but also require an efficient way to keep track of the computed quantities and the interventional data points. In the implementation we used Python dictionaries. This can work for larger DAGs as well using sparse matrices.

### 3.4.1. Strategy based on GP fits

This strategy is motivated by some observations of the behaviour of the optimization in the bivariate case (3.6). Namely,

- interventions are most informative in regions where we have observations already,
- it is more informative to confirm good estimates in areas with high nonlinearity than rejecting bad ones (i.e. if we currently belief that one parent child relation is likely to be there, it is more informative to intervene on the parent node than on the child node), and
- sometimes (like in the linear Gaussian case) we must rule out a possible backward model (i.e. intervene on a child node and hopefully obtain a sample which allows to confirm that the child actually is a child of its parent).

We try to pour these observations into a procedure using the information of the GP regression fits, based on similar ideas as in the Bayesian Optimization. To confirm good functional relations in areas where we already have observations, we minimize the variance of the functional relation estimate w.r.t. the prediction value, i.e. the variance of a GP prediction at some point in the input domain. For the next intervention value we choose the minimizer across all graphs and functional relations. Fortunately, this does not increase the computational burden, because we must calculate GP estimates in every iteration anyway (for the computation of the current likelihood of the data in each DAG). This procedure evolves around optimizing a quantity that has, given a DAG $G$, the form

$$\tilde{\sigma}^2_{G^{(j)}}(x) = \underbrace{k(x,x)}_{=\text{constant}} - k_{x\tilde{X}_{pa(j)}}(k_{\tilde{X}_{pa(j)}\tilde{X}_{pa(j)}} + \sigma_j^2 I)^{-1} k_{\tilde{X}_{pa(j)}x}, \quad x \in \mathcal{X}_{pa(j)}. \tag{3.17}$$

Since this is a quadratic chained with kernel evaluations, we can calculate closed forms for the derivative of $\sigma^2_{G^{(j)}}(x)$ w.r.t. $x$ and apply common minimization algorithms (a nonlinear conjugate gradient method [25, Chapter 5.2] suits well to the problem) to find a minimizer efficiently. If we have more than one input for the function, we take the node which is closest to the source node in the topological ordering (this can be an advantage, if the chosen intervention node is a parent of another input of the function).
To include the curvature information, we aim to maximize the absolute value of the second derivative of the inferred regression function. This works extremely well because the GP fit is very smooth such that a simple finite difference approximation already approximates extremely good. This calculation comes at very little computational cost. We define our objective the following way

$$f^{(j)}_{\text{obj.}}(x) := \sigma^2_{G^{(j)}}(x) - |\frac{\partial^2}{\partial x^2}\tilde{\mu}_{G^{(j)}}(x)| \tag{3.18}$$

Up to this point this would be a greedy strategy, because we focus to confirm one graph. We have seen in the linear Gaussian Case, that it can be beneficial to challenge current beliefs by testing competing graphs where the causal relation is reversed. Therefore, we sometimes switch roles in the parent child relation and minimize the objective function over the subset of graphs, for which the child node, that currently corresponds to the overall minimizer, is closer to the source node in topological order, than its parents. Since we observed in the bivariate case that three interventions are approximately enough to have a relatively concentrated posterior distribution, we also restrict our procedure to do at most three interventions on one and the same node. Thus we do interventions at most three times the number of nodes. A pseudo code for the procedure, without restricting the number of interventions on each variable, could look as follows.

**Algorithm 7** Strategy based on GP fits

```
Require: $G_{object}$, $\kappa$
  sample = $Bernoulli(\kappa)$
  $(G^*, j^*, x^*) = \text{argmin}_{G \in \mathcal{G}, j \in \{1,\dots,d\}, x \in \mathcal{X}_{pa(X_j)}} f^{(j)}_{\text{obj.}}(x)$
  if sample = 0 then
    Get $(j^*, x_{j^*})$ from $(G^*, j^*, x^*)$
  else
    $G_{sub}$ = Graphs where $pa(j^*)$ have lower topological order than $j^*$
    $(\bar{G}, \bar{j}, \bar{x}) = \text{argmin}_{G \in G_{sub}, j \in \{1,\dots,d\}, x \in \mathcal{X}_{pa(X_j)}} f^{(j)}_{\text{obj.}}(x)$
    Get $(j^*, x_{j^*})$ from $(\bar{G}, \bar{j}, \bar{x})$
  end if
  return $(j^*, x_{j^*})$
```

### 3.4.2. Monte Carlo Markov Chain

In the following, we quickly summarize the idea of MCMC integration and then point out how this technique can help to overcome our computational challenges. We follow [35, Markov Chain Monte Carlo Integration 11.1.2], where also a detailed description of the famous Metropolis-Hastings Sampler can be found.
Recall that the Monte Carlo estimator of $\mathbb{E}_X[g(X)]$ for a general function $g$ is

$$\bar{g} = \frac{1}{m} \sum_{i=1}^{m} g(x_i), \tag{3.19}$$

where $x_1, \dots, x_m$ is a sample from the distribution of $X$. If the sample is independent $\bar{g}$ converges in probability to $\mathbb{E}_X[g(X)]$ by the law of large numbers. The following generalization of the strong law of large numbers can be proven: If $\{X_0, X_1, X_2, \dots\}$ is a realization of an irreducible, ergodic Markov Chain with stationary distribution $\pi$, then

$$\bar{g(X)}_m = \frac{1}{m} \sum_{t=0}^{m} g(X_t) \overset{a.s.}{\to} \mathbb{E}_\pi[g(X)] \quad \text{as } m \to \infty, \tag{3.20}$$

where $X$ has the stationary distribution $\pi$.
In Bayesian statistics we are often faced with the problem of finding the posterior distribution and we usually know that the posterior distribution $p(x)$ is proportional to some likelihood times prior function $f(x)$, i.e. $p(x) \propto f(x)$. The idea of MCMC methods is to design a Markov Chain with stationary distribution $p(x)$. To generate the sample we need for the approximation of the integral we start with some initial point and then sample a candidate for the next point according to some proposal distribution, from which we can sample efficiently. We accept the new point if it satisfies the so called "detailed balance condition", which the stationary distribution satisfies. Inside this condition our problem, that we only know $p(x)$ up to proportionality, disappears because we have the same proportionality factor on both sides of the detailed balance equation. We should

only use later sampling points and ignore early observations (sometimes called ”burn in phase”) because they are very likely not distributed according to our target distribution. When the Markov chain converged to its stationary distribution, we can take all later sampled points for the estimation of the integral. In practice often visual criterion are used to determine, when the chain has converged.
Recall, that we want to calculate

$$(j^*, x^*) = \underset{j\in\{1,\dots,d\},x\in\mathcal{X}_j}{\operatorname{argmax}} \mathbb{E}_G\left[\mathbb{E}_{\mathbf{X}_{-j}|G,do(X_j=x)}\left[\log\left(p_{G|\mathbf{X}_{-j},do(X_j=x)}(G|\mathbf{x}_{-j}, do(X_j = x))\right)\right]\right].$$

In our approach we could use the MCMC method for the computation of

$$p_{G|\mathbf{X}_{-j},do(X_j=x)}(G|D_j^{(m)}, \tilde{D}),$$

what would at least avoid the computation of all graph posterior probabilities for every MC sampling point $D_j^{(m)}$. Thus it would reduce the number of executions of (5) from 543*$m$*543 to 543*$m$*$l$, where $l$ is the number of different DAGs visited by the Markov chain. This approach is only worth considering when $l$ is lower than 543 and we have an efficient procedure to move around in the DAG space. Thus, for four variables it might be, that an MCMC approach is not worth considering yet. If we would employ the MCMC method, we would consider $p_{G|\mathbf{X}_{-j},do(X_j=x)}(G|D_j^{(m)}, \tilde{D})$ as $\mathbb{E}_{G|D_j^{(m)},\tilde{D}}[\mathbb{1}_G]$ and apply the MCMC method for the latter expectation. This technique has proven itself successful in many practical applications and there is a huge body of literature about it. See for example [50, 7, 36].

### 3.4.3. Narrow Down DAG Space

This approach is the most practical one. The idea is, that after we calculated the posterior probabilities for the DAGs, we restrict our optimization on a subset of graphs that are most likely to be the data generating ones. The larger the subset, the longer the optimization procedure will take. If we intervene based on the maximized information gain for the subset of graphs, we can use the interventional data to compute likelihoods for all graphs again (because we can do it very efficiently) and in the next step choose a new subset based on the updated posterior probabilities. It might be wise to always include the empty graph to take into account the independence hypothesis. Note that this approach has only computational advantage if we calculate the posterior over graphs based on the subset, i.e. in the optimization procedure we calculate different posteriors than in the main loop, namely, for DAG $G$ in the subset $G_{\text{sub}}$,

$$p(G|D_j, \tilde{D}) = \frac{p(D_j, \tilde{D}|G)p(G)}{\sum_{\tilde{G}\in G_{\text{sub}}} p(D_j, \tilde{D}|\tilde{G})p(\tilde{G})}.$$

It might be possible to choose the subsets in a way such that we approximately do the same as maximizing information gain over the whole graph space with high probability. This idea is left for future research. In the following, we present a possible pseudo code to summarize the procedure. Note that the following pseudo code is almost the same as in (2) but restricts the DAG space in the beginning of the for loop.

**Algorithm 8** Narrow Down DAG Space

```
Require: $\mathcal{D}$, $G_{\text{list}}$, $n_{\text{interventions}}$        ▷ data, list of graphs with corr. parameters
  $\mathcal{D}_0 = \mathcal{D}$
  $d$ = number of nodes
  Compute $p(\mathcal{D}_0|G)$ for all $G$ in $G_{\text{list}}$ and store.
  Compute $p(G|\mathcal{D}_0)$ for all $G$ in $G_{\text{list}}$ and store.
  for $i = 1, \ldots, n_{\text{interventions}}$ do

      Method to get subset $G_{\text{sub}}$ based on $p(G|\mathcal{D}_{i-1})$
      $(j^*, x_{j^*}) = \text{argmax}_{j \in \{1,\ldots,d\}, x \in \mathcal{X}_j} \sum_{\tilde{G} \in G_{\text{sub}}} p_G(\tilde{G}) \frac{1}{M} \sum_{m=1}^{M} \log(p(\tilde{G}|\mathbf{x}_{-j}^{(m)}, do(X_j = x)))$

      Get sample $D_i = (x_{j^*}, \mathbf{x}_{-j^*})$        ▷ perform intervention experiment
      Compute $p(D_i|\mathcal{D}_{i-1}, G)$ for all $G$ in $G_{\text{list}}$ and store        ▷ likelihood based on
                                                                                              estimates of $\mathcal{D}_{i-1}$
      Store $p(D_i, \mathcal{D}_{i-1}|G) = p(D_i|\mathcal{D}_{i-1}, G)p(\mathcal{D}_{i-1}|G)$ for all $G$ in $G_{\text{list}}$
      Compute $p(G|\mathcal{D}_i)$ for all $G$ in $G_{\text{list}}$ and $i = 0, 1, \ldots, n_{\text{interventions}}$ and store
      $\mathcal{D}_i = \{D_i, \mathcal{D}_{i-1}\}$
  end for
  return $p(G|\mathcal{D}_i)$ for all $G$ in $G_{\text{list}}$ and $i = 0, 1, \ldots, n_{\text{interventions}}$; $\mathcal{D}_i$ for $i = 1, \ldots, n_{\text{interventions}}$
```

# 4. Conclusion

In this thesis we considered an active Bayesian causal discovery algorithm that optimizes the information gained on the causal structure from performing an intervention experiment and then updates the current beliefs. The strength of this approach is that we have closed forms for the likelihoods without imposing restrictive assumptions. The idea was proposed in [49]. Therein the authors provided numerical results for the bivariate case with independent interventions. We were able to reproduce similar results and extended the results to the case of optimal interventions. The transition from independent interventions to optimal interventions is not obvious, because optimal interventions lead to complex dependencies between initial observations and interventional observations. We assume a sequential way of generating the data, that allows us to factorize the likelihood and maintain the closed form.

The great potential of including interventional data for causal inference was illustrated in (3.7). This motivates to find the interventional experiment that is most informative about the true underlying causal structure. We studied the behaviour of the implemented algorithm for the bivariate case in much detail. We found that the objective functions of the optimization in each step are very similar. This can be seen as a hint that there is an equivalent but much simpler way to choose the optimal intervention value, than optimizing the information gain. Through a detailed study of many examples we gained some plausible behavioral characteristics of the algorithm. Based on the considered examples, the algorithm always intervenes in areas where we already have observational evidence, it tries to confirm likely causal relations in regions where no backward models are possible. If there exist valid backward models it can be beneficial to intervene on parent and child to rule out one direction. Depending on the data generating distribution this may take some intervention experiments until one very implausible data point for one causal direction is sampled. We have also seen that the algorithm only performs okay in detecting independencies. We conclude that optimizing the information gain in the DAG space is well suited for causal discovery in the considered Bayesian setup. But we have also seen that there may be an easier way to get a approximately equivalent procedure.
In the four variable case, the super exponentially growing number of DAGs already made the optimization too time consuming. We anyway presented pseudo code for the procedure because really only the optimization part is too time consuming but the computation of the posterior probabilities is still very quick. The considerations of this work are concluded by heuristic proposals to overcome the computational challenges. The first proposal is motivated by the findings about the behavior of the implementation in the bivariate case. We draw the connection to the Bayesian Optimization procedure.

In Bayesian optimization we estimate an unknown function using GP regression and choose the next point at which we evaluate the function based on the mean and covariance function obtained from the regression. In the considered causal discovery approach we estimate the functional relations between the variables with GP regression and calculate posterior probabilities based on them. Then we aim to choose the next intervention value such that the posterior distribution is most concentrated. The proposed idea is that we can find the next intervention value based on the information nested in the GP fits. This has the potential of saving lots of computation time.
The second and third proposals aim at saving computation time inside the general logic of the algorithm. They both aim to approximate the logarithmic posterior DAG probability. The MCMC proposal approximates the posterior distribution. The third proposal basically is a subspace optimization. For larger systems probably a combination of both is required.

Overall, we gained very interesting insight about interactive experimental causal structure learning, in this work. The idea proposed in [49] to use the Gaussian Process Regression and optimize information gain in the DAG space seem to lead into the right direction. But still there are many open questions and computational obstacles that must be overcome.

# Appendices

# A. Basic Statistics

At first we state Bayes theorem, following the lectures notes of "Foundations of Mathematical Statistics" taught by Prof. Mathias Drton. The Theorem was first mentioned in [4] and follows from the definition of conditional probabilities and the law of total probability.

**Theorem A.1** (Bayes Theorem)**.** *Consider an observation modeled as $X \sim \boldsymbol{P}_\theta, \theta \in \Theta \subset \mathbb{R}^k$. Suppose prior distribution has density $\pi$ w.r.t. a measure $\nu$ and $\boldsymbol{P}_\theta \ll \nu' \; \forall \theta$ with densities $p_\theta(x) = p(x|\theta)$. Then the posterior distribution has density (w.r.t. $\nu$) :*

$$\left( \frac{p(x,\theta)}{p(x)} = \right) p(\theta|x) = \frac{p(x|\theta)\pi(\theta)}{p(x)} \tag{A.1}$$

*where $p(x) = \int_\Theta p(x|\theta)\pi(\theta)\, d\nu(\theta)$ is prior predictive density of $X$.*

**Remark A.2.** The posterior density is proportional to the likelihood of the i.i.d. data $x \in \mathbb{R}^n$ times the prior distribution, i.e. $p(\theta|x) \propto L_x(\theta)\pi(\theta)$ where $L_x(\theta) = \prod_{i=1}^n p_\theta(x_i)$. Bayes estimators of $\theta$ are obtained as characteristics of the posterior distribution. Most frequently, the posterior mean is used:

$$\hat{\theta} = \mathbb{E}[\theta|X = x] = \int \theta p(\theta|x)\, d\nu(\theta).$$

Next we describe mean square continuity and differentiability of stochastic processes following [34, Section 4.1.1], [1].

**Definition A.3.** Let $\mathbf{x}_1, \mathbf{x}_2, \ldots$ be a sequence of points and $\mathbf{x}_*$ be a fixed point in $\mathbb{R}^d$ such that $\|\mathbf{x}_k - \mathbf{x}_*\| \to 0$ as $k \to \infty$. then a process $f(\mathbf{x})$ is continuous in mean square at $\mathbf{x}_*$ if $\mathbb{E}[\|f(\mathbf{x}_k) - f(\mathbf{x}_*)\|^2]$ as $k \to \infty$. A random field is continuous in mean square at $\mathbf{x}_*$ if and only if its covariance function (kernel) $k(\mathbf{x}, \mathbf{x}')$ is continuous at the point $\mathbf{x} = \mathbf{x}' = \mathbf{x}_*$.
If the mean square derivative of $f(\mathbf{x})$ in the $i^{\text{th}}$ direction exists, it is implicitly defined as

$$\frac{\partial f(\mathbf{x})}{\partial x_i} := \lim_{h \to 0} \mathbb{E}\left[ \left\| \frac{f(\mathbf{x} + h\mathbf{e}_i) - f(\mathbf{x})}{h} - \frac{\partial f(\mathbf{x})}{\partial x_i} \right\|^2 \right] = 0, \tag{A.2}$$

where $\mathbf{e}_i$ is the unit vector in $i^{\text{th}}$ direction. The covariance function (kernel) of $\frac{\partial f(\mathbf{x})}{\partial x_i}$ is given by $\frac{\partial^2 k(\mathbf{x},\mathbf{x}')}{\partial x_i \partial x_i'}$.

Because in Bayesian Experimental design it can happen that we mix discrete and continuous distributions, we make a short note on the product measures of Counting and Lebesgue measure. In the following we denote by $\mathcal{P}(\cdot)$ the power set of the argument. Recall, that $(\mathbb{N}, \mathcal{P}(\mathbb{N}), \nu)$ with $\nu : \mathcal{P}(\mathbb{N}) \to [0, \infty]; A \mapsto \sum_{k \in \mathbb{N}} \mathbb{1}_A$ is a measurable space and $\nu$ is $\sigma$-finite on $(\mathbb{N}, \mathcal{P}(\mathbb{N}))$.
Define the half open intervals in $\mathbb{R}^n$ as $(a, b] := \{x \in \mathbb{R}^n | a_j < x_j \le b_j \text{ for } j = 1, \dots, n\}$. Recall, that $(\mathbb{R}^n, \bigotimes_{i=1}^n \mathcal{B}(\mathbb{R}), \lambda)$ with $\lambda : \bigotimes_{i=1}^n \mathcal{B}(\mathbb{R}) \to [0, \infty]; (a, b] \mapsto \prod_{j=1}^n (b_j - a_j)$ is a measurable space and $\lambda$ is $\sigma$-finite on $(\mathbb{R}^n, \bigotimes_{i=1}^n \mathcal{B}(\mathbb{R}))$.

**Lemma A.4.** Let $(X_i, \mathcal{A}_i, \mu_i), i = 1, 2,$ be $\sigma$-finite measurable spaces. Then there exists a unique product measure $\mu_1 \otimes \mu_2$ on $\mathcal{A}_1 \otimes \mathcal{A}_2$. It holds for all $A \in \mathcal{A}_1 \otimes \mathcal{A}_2$ that

$$(\mu_1 \otimes \mu_2)(A) = \int \left( \int \mathbb{1}_A(x, \cdot) \, d\mu_2 \right) d\mu_1(x) = \int \left( \int \mathbb{1}_A(\cdot, y) \, d\mu_1 \right) d\mu_1(y).$$

*Proof.* See [13, Hilfssatz 3.7.]. □

Thus, $(\nu \otimes \lambda)$ is a unique well defined product measure on $\mathcal{P}(\mathbb{N}) \bigotimes \mathcal{B}(\mathbb{R}^n)$ with $(\nu \otimes \lambda)(A_1 \times A_2) = \nu(A_1)\lambda(A_2)$.

## A.1. Multivariate Normal (Gaussian) Distribution

The Definition A.5 and Theorem A.6 closely follow the ones in the lecture notes of “Stochastic Analysis” taught by Prof. Nina Gantert.

**Definition A.5.** A real-valued random variable $Z_1$ is Gaussian or normal if it has the density

$$f(z_1) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{(z_1 - m_1)^2}{2\sigma^2} \tag{A.3}$$

for some $m_1 \in \mathbb{R}$ and $\sigma > 0$ $[Z_1 \sim \mathcal{N}(m_1, \sigma^2)]$. $Z_1$ is generalized Gaussian if either $Z_1$ is Gaussian or $\mathbb{P}(Z_1 = m_1) = 1$ for some $m_1 \in \mathbb{R}$, i.e., ”$Z_1 \sim \mathcal{N}(m_1, 0)$”. Let $X = (X_1, X_2, \dots, X_n)$ be a random variable with values in $\mathbb{R}^n$. We say that $X$ is a (multivariate) Gaussian random variable if for every vector $\alpha = (\alpha_1, \dots, \alpha_n) \in \mathbb{R}^n$, the real-valued random variable $X_\alpha = \langle \alpha, X \rangle = \sum_{k=1}^n \alpha_k X_k$ is generalized Gaussian. We write $m(X) = \mathbb{E}[X] = (\mathbb{E}[X_1], \dots, \mathbb{E}[X_n])$ and denote by $\Sigma = \Sigma(X)$ the covariance matrix given by $\Sigma_{ij} = \mathrm{Cov}(X_i, X_j), 1 \le i, j \le n$. The random variable $X \sim \mathcal{N}_n(m, \Sigma)$ has the density

$$p(x) = (2\pi)^{-\frac{n}{2}} \det(\Sigma)^{-\frac{1}{2}} \exp\left( -\frac{1}{2}(x - m)^\intercal \Sigma^{-1}(x - m) \right) \tag{A.4}$$

for some $m \in \mathbb{R}^n$ and positive semidefinite $\Sigma \in \mathbb{R}^{n \times n}$.

**Theorem A.6.** *(i) For every vector $m = (m_1, \dots, m_n)$ and every positive semidefinite matrix $\Sigma \in \mathbb{R}^{n \times n}$, there exists a multivariate Gaussian random variable $X$ such that $\mathbb{E}[X] = (m_1, \dots, m_n)$ and $\Sigma(X) = \Sigma$.*

*(ii) If $X$ and $Y$ are multivariate Gaussian random variables and $\mathbb{E}[X] = \mathbb{E}[Y]$ and $\Sigma(X) = \Sigma(Y)$, then $X$ and $Y$ have the same distribution.*

*Proof.* (i) For every symmetric positive semidefinite matrix $\Sigma$ we can find a symmetric matrix such that $\Sigma = A^2$. Let $Z_1, Z_2, \ldots, Z_n$ be i.i.d. with law $\mathcal{N}(0,1)$ and let $Z = (Z_1, Z_2, \ldots, Z_n)$. Take $X = AZ + m$. Then $X$ is multivariate Gaussian and $\mathbb{E}[X] = m$. We calculate

$$\begin{aligned}\Sigma(X)_{ij} = Cov(X_i, X_j) &= \mathbb{E}[(AZ)_i(AZ)_j] \\ &= \mathbb{E}\left[\sum_{k=1}^{n} A_{ik}Z_k \sum_{l=1}^{n} A_{jl}Z_l\right] \\ &= \sum_{k=1}^{n}\sum_{l=1}^{n} A_{ik}A_{jl}\,\mathbb{E}[Z_kZ_l] \\ &= \sum_{k=1}^{n} A_{ik}A_{jk} = \sum_{k=1}^{n} A_{ik}A_{kj} = (A^2)_{ij}.\end{aligned}$$

(ii) Since $X$ and $Y$ are multivariate Gaussian wit same mean and variance, their characteristic functions agree for any $t \in \mathbb{R}^n$.

□

**Remark A.7.** The characteristic function of a multivariate normal distribution, $X \sim \mathbb{N}(\mu, \Sigma)$ is given by

$$\varphi_X(t) = e^{i\langle \mu, t\rangle - \frac{1}{2}\langle t, \Sigma t\rangle}. \tag{A.5}$$

**Lemma A.8.** Let $X \sim \mathbb{N}(\mu, \Sigma_1)$ and $\mu \sim \mathbb{N}(\tilde{\mu}, \Sigma_2)$ for $\tilde{\mu} \in \mathbb{R}^d$ and positive semi-definite $\Sigma_1, \Sigma_2 \in \mathbb{R}^{n\times n}$ with $d \in \mathbb{N}$. Then $X \sim \mathbb{N}(\tilde{\mu}, \Sigma_1 + \Sigma_2)$.

*Proof.*

$$\varphi_X(t) = \mathbb{E}\left[\mathbb{E}\left[e^{i\langle t, X\rangle}|\mu\right]\right] = \mathbb{E}\left[e^{i\langle \mu, t\rangle - \frac{1}{2}\langle t, \Sigma_1 t\rangle}\right] = e^{i\langle \tilde{\mu}, t\rangle - \frac{1}{2}\langle t, (\Sigma_1+\Sigma_2)t\rangle}. \tag{A.6}$$

□

## Matrix Algebra

**Lemma A.9.** Let $A$,$B \in \mathbb{R}^{n\times n}$ invertible matrices for some $n \in \mathbb{N}$. Then it holds

$$(A^{-1} + B^{-1})^{-1} = A - A(A+B)^{-1}A = B - B(A+B)^{-1}B \tag{A.7}$$

*Proof.* Basic linear algebra yields:

$$\begin{aligned}&A^{-1} + B^{-1} = B^{-1}(A+B)A^{-1} \\ \Leftrightarrow &(A^{-1} + B^{-1})^{-1} = A(A+B)^{-1}B = A(A+B)^{-1}((A+B) - A) \\ \Leftrightarrow &(A^{-1} + B^{-1})^{-1} = A - A(A+B)^{-1}A.\end{aligned}$$

Switching roles of $A$ and $B$ provides the last equality of the lemma. □

**Lemma A.10.** For $n$,$m \in \mathbb{N}$ let $Z \in \mathbb{R}^{n\times n}$, $W \in \mathbb{R}^{m\times m}$ and $U, V \in \mathbb{R}^{n\times m}$, it holds

$$(Z + UWV^\intercal)^{-1} = Z^{-1} - Z^{-1}U(W^{-1} + V^\intercal Z^{-1}U)^{-1}V^\intercal Z^{-1}, \tag{A.8}$$

given the respective inverses exist.

*Proof.* Basic linear algebra yields:

$$\begin{aligned}
&(Z + UWV^\intercal)(Z^{-1} - Z^{-1}U(W^{-1} + V^\intercal Z^{-1}U)^{-1}V^\intercal Z^{-1})\\
&= I - U(W^{-1} + V^\intercal Z^{-1}U)^{-1}V^\intercal Z^{-1} + UWV^\intercal Z^{-1} - UWV^\intercal Z^{-1}U(W^{-1} + V^\intercal Z^{-1}U)^{-1}V^\intercal Z^{-1}\\
&= I + UWV^\intercal Z^{-1} - (U + UWV^\intercal Z^{-1}U)((W^{-1} + V^\intercal Z^{-1}U)^{-1}V^\intercal Z^{-1})\\
&= I + UWV^\intercal Z^{-1} - UW(W^{-1} + V^\intercal Z^{-1}U)((W^{-1} + V^\intercal Z^{-1}U)^{-1}V^\intercal Z^{-1})\\
&= I + UWV^\intercal Z^{-1} - UWV^\intercal Z^{-1}.
\end{aligned}$$

□

**Lemma A.11.** For $n$,$m \in \mathbb{N}$ let $Z \in \mathbb{R}^{n\times n}$, $W \in \mathbb{R}^{m\times m}$ and $U, V \in \mathbb{R}^{n\times m}$, it holds

$$\det(Z + UWV^\intercal) = \det(Z)\det(W)\det(W^{-1} + V^\intercal Z^{-1}U), \tag{A.9}$$

given the respective inverses exist.

*Proof.* In order to prove the claim, we need the so called "Weinstein-Aronszajn identity", which states that

$$\det(I_m + V^\intercal U) = \det(I_n + UV^\intercal). \tag{A.10}$$

To see this consider a matrix $M$ consisting of the four blocks $B$, $-A$, $I_m$ and $I_n$ and note that the identity matrix is invertible, such that we can use the formula for the determinant of a block matrix:

$$\begin{aligned}
(i) : \det\begin{pmatrix} I_m & -V^\intercal \\ U & I_n \end{pmatrix} &= \det(I_m)\det(I_n - UI_m^{-1}(-V^\intercal)) = \det(I_n + UV^\intercal),\\
(ii) : \det\begin{pmatrix} I_m & -V^\intercal \\ U & I_n \end{pmatrix} &= \det(I_n)\det(I_m - (-V^\intercal)I_n^{-1}U) = \det(I_m + V^\intercal U).
\end{aligned}$$

Further note that we have

$$\det(Z + UV^\intercal) = \det(Z(I_n + Z^{-1}UV^\intercal)) = \det(Z)\det(I_n + Z^{-1}UV^\intercal).$$

An application of the Weinstein-Aronszajn identity yields

$$\det(Z + UV^\intercal) = \det(Z)\det(I_m + V^\intercal Z^{-1}U).$$

Now we replace $U$ by $UW$ in the above equation and obtain

$$\begin{aligned}
\det(Z + UWV^\intercal) &= \det(Z)\det(I_m + V^\intercal Z^{-1}UW)\\
&= \det(Z)\det((W^{-1} + V^\intercal Z^{-1}U)W))\\
&= \det(Z)\det(W)\det(W^{-1} + V^\intercal Z^{-1}U).
\end{aligned}$$

□

# B. (Simple) Bayesian Optimization Regret Bound

The theorem and proof presented here closely follow [41, Theorem 1, Appendix A]. Consider the setting of (2.4.2). But we restrict to the case where $|D| < \infty$. Before establishing the main proof we work through some preliminary results.
The informativeness of a set of sampling points $A \subset D$ about $f$ is measured by the information gain, i.e. the mutual information between $f$ and observations $y_A = f_A + \epsilon_A$ at these points, where $f_A := (f(x))_{x \in A}$ and $\epsilon_A \sim \mathcal{N}(0, \sigma^2 I_{|A|})$. If we define by $K_A$ the Gram matrix of the points in $A$ and use (2.17) we obtain for the information gain

$$I(y_A; f_A) = \frac{1}{2} \log \left( \det(I_{|A|} + \sigma^{-2} K_A) \right). \tag{B.1}$$

This quantity will show up in the regret bound, but we need to express it in terms of the predictive variances in order to be able to use it later.

**Lemma B.1.** Let $T \geq 1$ denote the index of the last point we already visited with our algorithm. The information gain for the points selected can be expressed in term of the predictive variances. If $f_T = (f(x_t))_{t \in \{1,\dots,T\}} \in \mathbb{R}^T$, then

$$I(y_T; f_T) = \frac{1}{2} \sum_{t=1}^{T} \log(1 + \sigma^{-2} \sigma_{t-1}^2(x_t)). \tag{B.2}$$

*Proof.* Note, that for $T \geq 1$ it holds $y_T | y_{T-1} \sim \mathcal{N}(\mu_{T-1}(x_T), \sigma^2 + \sigma_{T-1}^2(x_T))$ (see (3.10). For readability we write only $\sigma_{T-1}^2$. We have

$$\begin{aligned}
(i) : I(y_T; f_T) &= h(y_T) - \frac{1}{2} \log(\det(2\pi e \sigma^2 I_T)) = h(y_T) - \frac{1}{2} \log((2\pi e)^T \sigma^{2T}) \\
(ii) : h(y_T) &= h(y_{T-1}) + h(y_T | y_{T-1}) = h(y_{T-1}) + \frac{1}{2} \log(2\pi e(\sigma^2 + \sigma_{T-1}^2(x_T)) \\
&= \dots = \frac{1}{2} \log((2\pi e)^T \prod_{t=1}^{T} (\sigma^2 + \sigma_{t-1}^2(x_t))),
\end{aligned}$$

and thus,

$$I(y_T; f_T) = \frac{1}{2} \log \left( \frac{(2\pi e)^T \prod_{t=1}^{T} (\sigma^2 + \sigma_{t-1}^2(x_t))}{(2\pi e)^T \sigma^{2T}} \right) = \frac{1}{2} \sum_{t=1}^{T} \log(1 + \sigma^{-2} \sigma_{t-1}^2(x_t)).$$

□

We also need the following inequality.

**Lemma B.2.** Let $r \sim \mathcal{N}(0,1)$ and $c > 0$. Then it holds that

$$P(r > c) \leq \frac{1}{2} e^{-\frac{c^2}{2}}. \tag{B.3}$$

*Proof.* Let $r \sim \mathcal{N}(0,1)$ and $c > 0$. First note that

$$\begin{aligned} -\frac{r^2}{2} &= -\frac{r^2}{2} + rc - \frac{c^2}{2} - rc + c^2 - \frac{c^2}{2} \\ &= -\frac{(r-c)^2}{2} - c(r-c) - \frac{c^2}{2}. \end{aligned}$$

Using this, we can establish the assertion,

$$\begin{aligned} P(r > c) &= (2\pi)^{-\frac{1}{2}} \int_c^\infty e^{-\frac{r^2}{2}}\, dr \\ &= e^{-\frac{c^2}{2}} (2\pi)^{-\frac{1}{2}} \int_c^\infty e^{-\frac{(r-c)^2}{2} - c(r-c)}\, dr \\ &= e^{-\frac{c^2}{2}} (2\pi)^{-\frac{1}{2}} \int_0^\infty e^{-\frac{r^2}{2} - cr}\, dr \\ &\leq e^{-\frac{c^2}{2}} P(r > 0) = \frac{1}{2} e^{-\frac{c^2}{2}}. \end{aligned}$$

□

We can now establish a convergence result in the case of a finite function domain.

**Theorem B.3.** *Let $\delta \in (0,1)$ and $\beta_t = 2\log(|D|t^2\pi^2/6\delta)$. Running GP-UCB with $\beta_t$ for a sample $f \sim \mathcal{GP}(0,k)$, we obtain a regret bound of $\mathcal{O}^*(\sqrt{T\gamma_T \log(|D|)}$ with high probability. Precisely,*

$$P(R_T \leq \sqrt{C_1 T \beta_T \gamma_T} \quad \forall T \geq 1) \geq 1 - \delta, \tag{B.4}$$

*where $C_1 = \frac{8}{\log(1+\sigma^{-2})}$.*

*Proof.* Let $\delta \in (0,1)$ and set $\beta_t = 2\log(|D|\frac{\pi_t}{\delta})$, where $\pi_t$ can be any sequence satisfying $(\pi_t)_{t\in\mathbb{N}} > 0$ and $\sum_{t\geq 1} \pi_t^{-1} = 1$. Fix $t \geq 1$ and $x \in D$. Conditioned on $\mathbf{y}_{t-1}$ (and $x, \mathbf{x}_{t-1}$) we have, according to theorem (2.6), $f(x) \sim \mathcal{N}(\mu_{t-1}(x), \sigma_{t-1}^2(x))$. Define $r := \frac{f(x)-\mu_{t-1}(x)}{\sigma_{t-1}(x)}$, use (B.3) with $c = \sqrt{\beta_t}$ to bound $r$ and $-r$, and apply Boole's inequality to obtain

$$P(|f(x) - \mu_{t-1}(x)| > \sqrt{\beta_t}\sigma_{t-1}(x) \quad \forall x \in D) \leq |D| 2 \frac{1}{2} e^{-\frac{\beta_t}{2}} = \frac{|D|\delta}{|D|\pi_t} = \frac{\delta}{\pi_t}. \tag{B.5}$$

We again union bound this quantity over the iterations $t$ and obtain

$$P(|f(x) - \mu_{t-1}(x)| \leq \sqrt{\beta_t}\sigma_{t-1}(x) \quad \forall x \in D \quad \forall t \geq 1) \geq 1 - \sum_{t\geq 1} \frac{\delta}{\pi_t} = 1 - \delta. \tag{B.6}$$

The choice of $\pi_t = \pi^2 t^2/6$ we use in the theorem is valid, since $\sum_{t\geq 1} \frac{1}{t^2} = \frac{\pi^2}{6}$ [12, Folgerung 9.35.].
For the following let $x^*$ denote an optimizer of $f$ and again fix $t \geq 1$. By definition of $x_t$, (B.5) and the above result it holds with probability $1 - \frac{\delta}{\pi_t}$ that

$$\begin{aligned} &\mu_{t-1}(x_t) + \sqrt{\beta_t}\sigma_{t-1}(x_t) \geq \mu_{t-1}(x^*) + \sqrt{\beta_t}\sigma_{t-1}(x^*) \geq f(x^*) \\ \Leftrightarrow \quad & r_t = f(x^*) - f(x_t) \leq \mu_{t-1}(x_t) - f(x_t) + \sqrt{\beta_t}\sigma_{t-1}(x_t) \leq 2\sqrt{\beta_t}\sigma_{t-1}(x_t). \end{aligned}$$

Thus, we can bound the squared instantaneous regret with high probability

$$P(r_t^2 \leq 4\beta_t\sigma_{t-1}^2(x_t) \quad \forall t \geq 1) \geq 1 - \delta. \tag{B.7}$$

Since $\beta_t$ is non-decreasing in $t$ we have

$$\begin{aligned} 4\beta_t\sigma_{t-1}^2(x_t) &\leq 4\beta_T\sigma^2(\sigma^{-2}\sigma_{t-1}^2(x_t)) \\ &\leq 4\beta_T\sigma^2 C_2 \log(1 + \sigma^{-2}\sigma_{t-1}^2(x_t)), \end{aligned}$$

with $C_2 = \frac{\sigma^{-2}}{\log(1+\sigma^{-2})} \geq 1$, since $s^2 \leq C_2 \log(1 + s^2)$ for $s \in [0, \sigma^{-2}]$, and $\sigma^{-2}\sigma_{t-1}^2(x_{t-1}) \leq \sigma^{-2}k(x_t, x_t) \leq \sigma^{-2}$.
Note that $C_1 = \frac{8}{\log(1+\sigma^{-2})} = 8\sigma^2 C_2$. Using (B.2) we can bound the sum of squared instantaneous regrets with probability larger than $1 - \delta$, for all $T \geq 1$

$$\begin{aligned} \sum_{t=1}^T r_t^2 &\leq \beta_T 8\sigma^2 C_2 \frac{1}{2} \sum_{t=1}^T \log(1 + \sigma^{-2}\sigma_{t-1}^2(x_t)) \\ &= \beta_T C_1 I(y_T; f_T) \leq \beta_T C_1 \gamma_T, \end{aligned} \tag{B.8}$$

where $\gamma_T$ will be specified later.
Putting all together and using the Cauchy-Schwarz inequality (in $\mathbb{R}^T$) we have

$$R_T^2 = (\sum_{t=1}^T r_t)^2 \leq T(\sum_{t=1}^T r_t^2),$$

so that we have

$$P(R_T \leq \sqrt{C_1 T \beta_T \gamma_T} \quad \forall T \geq 1) \geq 1 - \delta.$$

□

The only thing left to specify is how to bound the maximum information gain after $T$ iterations, $\gamma_T$. Below a result for the Matérn kernel is stated from [41, Theorem 5].

**Theorem B.4.** *Let $D \subset \mathbb{R}^d$ be compact and convex, $d \in \mathbb{N}$. Assume the kernel function satisfies $k(x, x') \leq 1$. For Matérn kernels with $\nu > 1$:*

$$\gamma_T = \mathcal{O}(T^{\frac{d(d+1)}{2\nu+d(d+1)}} \log(T)). \tag{B.9}$$

*Proof.* See [41, Appendix C.1.]. □

# C. Kernel Theory

## C.1. Sobolev spaces

**Definition C.1** (Weak Differentiability)**.** Let $\Omega \subset \mathbb{R}^d$, $f \in L^1_{loc}(\Omega)$ and $\alpha \in \mathbb{N}_0^d$ a multi index. We say that $f$ has a weak derivative of order $\alpha$ in $L^1_{loc}(\Omega)$, if there exists a function $g \in L^1_{loc}(\Omega)$ such that for all $\varphi \in C_c^\infty(\Omega)$ it holds

$$\int_\Omega f \partial^\alpha \varphi \, dx = (-1)^{|\alpha|} \int_\Omega g \varphi \, dx. \tag{C.1}$$

We call $g$ the weak derivative of $f$ of order $\alpha$ and we write $\partial^\alpha f := g$.

**Definition C.2** (Sobolev Space)**.** Let $\Omega \subset \mathbb{R}^d$ open, $1 \leq p \leq \infty$ and $k \in \mathbb{N}$. The Sobolev space $W^{k,p}(\Omega)$ is defined as the set of all functions $f \in L^p(\Omega)$ that for all $\alpha \in \mathbb{N}_0^k$ with $|\alpha| \leq k$ have a weak derivative $\partial^\alpha f \in L^1_{loc}(\Omega)$. We define the Sobolev norm $||f||_{k,p}$ as

$$||f||_{k,p} := \left( \sum_{\alpha \in \mathbb{N}_0^k : |\alpha| \leq k} ||\partial^\alpha f||_p^p \right)^{\frac{1}{p}}$$

for $1 \leq p < \infty$ and

$$||f||_{k,\infty} := \sup_{\alpha \in \mathbb{N}_0^k : |\alpha| \leq k} ||\partial^\alpha f||_\infty .$$

## C.2. Integral Operator and its Eigensystem

We introduce an integral operator and investigate its eigensystem. We will follow the simplified considerations in [19, Section 4.1.1], where we assume that $\mathcal{X} \subset \mathbb{R}^d$ is a compact metric space and $k$ is a continuous kernel on $\mathcal{X}$.
Let $\mu$ be a finite Borel measure on $\mathcal{X}$ and $L_2(\mu)$ be the Hilbert space of square-integrable functions w.r.t. $\mu$. Define an integral operator with the kernel $k$ and the measure $\mu$ as

$$T_k : L_2(\mu) \to L_2(\mu); \quad f \mapsto \int k(\cdot, x) f(x) \, d\mu(x). \tag{C.2}$$

Since we only consider positive definite kernels and consider the simplified setting where $\mathcal{X}$ is compact, it is well known that $T_k$ is a compact, positive and self-adjoint operator. Thus we can apply the spectral theorem (see [43, Theorem A.5.13]) which guarantees us an eigen-decomposition of $T_k$ in the form

$$T_k f = \sum_{i \in I} \lambda_i \langle \phi_i, f \rangle_{L_2(\mu)} \phi_i, \tag{C.3}$$

where the convergence is in $L_2(\mu)$, $I \subset \mathbb{N}$ is a set of indices and $(\phi_i, \lambda_i)_{i \in I} \subset L_2(\mu) \times (0, \infty)$ are (countable) eigenfunctions and the associated eigenvalues of $T_k$ such that $\lambda_1 \geq \lambda_2 \geq \cdots > 0$:

$$T_k \phi_i = \lambda_i \phi_i, \quad i \in I.$$

Further it holds that $\langle \phi_i, \phi_j \rangle_{L_2(\mu)} = \delta_{ij}$, with $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ otherwise, i.e., the eigenfunctions $(\phi_i)_{i \in \mathbb{N}}$ form an orthonormal system in $L_2(\mu)$.

## C.3. Reproducing Kernel Hilbert Space (RKHS)

We here present the results relevant for this work from [19, Section 2.3]. That is, a definition of (RKHS), a way to construct the function space for a given kernel and as an example the RKHS when using a Matérn kernel.

**Definition C.3** (RKHS)**.** Let $\mathcal{X}$ be a nonempty set and $k$ be a positive definite kernel on $\mathcal{X}$. A Hilbert space $\mathcal{H}_k$ of functions on $\mathcal{X}$ equipped with an inner-product $\langle \cdot, \cdot \rangle_{\mathcal{H}_k}$ is called a reproducing kernel Hilbert space (RKHS) with reproducing kernel $k$, if the following are satisfied:

1. For all $x \in \mathcal{X}$, we have $k(\cdot, x) \in \mathcal{H}_k$;
2. For all $x \in \mathcal{X}$ and for all $f \in \mathcal{H}_k$,

$$f(x) = \langle f, k(\cdot, x) \rangle_{\mathcal{H}_k} \quad \text{(Reproducing Property)}.$$

**Remark C.4.** For each kernel $k$ there exists a uniquely associated RKHS and vice versa (see Moore-Aronszajn theorem [2]).

### Construction of RKHS $\mathcal{H}_k$ given a kernel k

Let $k$ be a positive definite kernel on $\mathcal{X}$. Then,

$$\mathcal{H}_0 := \operatorname{span}\{k(\cdot, x) : x \in \mathcal{X}\} = \left\{ f = \sum_{i=1}^{n} c_i k(\cdot, x_i) : n \in \mathbb{N}, c_1, \ldots, c_n \in \mathbb{R}, x_1, \ldots, x_n \in \mathcal{X} \right\},$$

is a pre-Hilbertspace, if we define the inner-product as

$$\langle f, g \rangle_{\mathcal{H}_0} := \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j k(x_i, y_j).$$

The RKHS $\mathcal{H}_k$ associated with $k$ is defined as the closure of $\mathcal{H}_0$ with respect to the norm $||f||_{\mathcal{H}_k} = \sqrt{\langle f, f\rangle_{\mathcal{H}_0}}$, i.e. $\mathcal{H}_k := \overline{\mathcal{H}_0}$. That is,

$$\mathcal{H}_k = \Bigg\{ f = \sum_{i=1}^{\infty} c_i k(\cdot, x_i) : (c_1, c_2, \dots) \subset \mathbb{R}, (x_1, x_2, , \dots) \subset \mathcal{X}, \text{ such that} \\ ||f||^2_{\mathcal{H}_k} := \lim_{n\to\infty} \left\| \sum_{i=1}^{n} c_i k(\cdot, x_i) \right\|^2_{\mathcal{H}_k} = \sum_{i,j=1}^{\infty} c_i c_j k(x_i, x_j) < \infty \Bigg\}. \tag{C.4}$$

We now state a result for Matérn kernels [19, Example 2.6].

**Example C.5.** Let $k_{\nu,\gamma}$ be the Matérn kernel on $\mathcal{X} \subset \mathbb{R}^d$ with Lipschitz boundary with parameters $\nu > 0$ and $h > 0$ such that $s := \nu + \frac{d}{2}$ is an integer. Then the RKHS $\mathcal{H}_{k_{\nu,\gamma}}$ of $k_{\nu,\gamma}$ is norm-equivalent to the Sobolev space $W(\mathcal{X})^{s,2}$ of order $s$. That is, we have $\mathcal{H}_{k_{\nu,\gamma}} = W(\mathcal{X})^{s,2}$ as a set of functions, and there exist constants $c_1, c_2 > 0$ such that

$$c_1 \|f\|_{W(\mathcal{X})^{s,2}} \leq \|f\|_{\mathcal{H}_{k_{\nu,\gamma}}} \leq c_2 \|f\|_{W(\mathcal{X})^{s,2}}, \quad \forall f \in \mathcal{H}_{k_{\nu,\gamma}}. \tag{C.5}$$

*Proof.* See Wendland [51, Corollary 10.48] and [34, Eq. 4.15]. □

We below state [19, Theorem 2.4], which provides us with an explicit characterization of the RKHS in terms of Fourier transforms in the case of shift-invariant kernels.

**Theorem C.6.** *Let $k$ be a shift-invariant kernel on $\mathcal{X} = \mathbb{R}^d$ such that $k(x, y) := \Phi(x - y)$ for $\Phi \in C(\mathbb{R}^d) \cap L_1(\mathbb{R}^d)$. Then the RKHS $\mathcal{H}_k$ of $k$ is given by*

$$\mathcal{H}_k = \left\{ f \in L_2(\mathbb{R}^d) \cap C(\mathbb{R}^d) : \|f\|^2_{\mathcal{H}_k} = \frac{1}{(2\pi)^{d/2}} \int \frac{|\mathcal{F}[f](\omega)|^2}{\mathcal{F}[\Phi](\omega)} d\omega < \infty \right\}, \tag{C.6}$$

*with the inner-product being*

$$\langle f, g\rangle_{\mathcal{H}_k} = \frac{1}{(2\pi)^{d/2}} \int \frac{\mathcal{F}[f](\omega)\overline{\mathcal{F}[g](\omega)}}{\mathcal{F}[\Phi](\omega)} d\omega, \quad f, g \in \mathcal{H}_k,$$

*where $\overline{\mathcal{F}[g](\omega)}$ denotes the complex conjugate of $\mathcal{F}[g](\omega)$.*

*Proof.* See Wendland [51, Theorem 10.12]. □

Finally we sate [19, Example 2.8], a full characterization of the RKHS of Matérn kernels.

**Example C.7** (RKHS of Matérn kernels)**.** Let $k_{\nu,\gamma}$ be the Matérn kernel on $\mathbb{R}^d$ with parameters $\nu > 0$ and $\gamma > 0$, and let $\mathcal{H}_{k_{\nu,\gamma}}$ of $k_{\nu,\gamma}$ be the associated RKHS. Then

$k_{\nu,\gamma}(x,y) = \Phi_{\nu,\gamma}(r)$ with $r := \|x-y\|_2$ and $\Phi_{\nu,\gamma}(r) := \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}r}{\gamma}\right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}r}{\gamma}\right)$ and the Fourier transform of $\Phi_{\nu,\gamma}$ is given by

$$\mathcal{F}[\Phi_{\nu,\gamma}](\omega) = C_{\nu,\gamma,d} \left(\frac{2\nu}{\gamma^2} + 4\pi^2 \|\omega\|_2^2\right)^{-\left(\nu+\frac{d}{2}\right)}, \quad \omega \in \mathbb{R}^d, \tag{C.7}$$

where $C_{\nu,\gamma,d} := \frac{2^d \pi^{d/2} \Gamma(\nu+d/2)(2\nu)^\nu}{\Gamma(\nu)\gamma^{2\nu}}$ [34, Eq. 4.15]. Therefore the RKHS $\mathcal{H}_{k_{\nu,\gamma}}$ can be written as

$$\begin{aligned}\mathcal{H}_{k_{\nu,\gamma}} = \Bigg\{ & f \in L_2(\mathbb{R}^d) \cap C(\mathbb{R}^d) : \\ & \|f\|^2_{\mathcal{H}_{k_{\nu,\gamma}}} = \frac{1}{(2\pi)^{d/2} C_{\nu,\gamma,d}} \int |\mathcal{F}[f](\omega)|^2 \left(\frac{2\nu}{\gamma^2} + 4\pi^2 \|\omega\|_2^2\right)^{\left(\nu+\frac{d}{2}\right)} d\omega < \infty \Bigg\},\end{aligned}$$

which shows that, for any $f \in \mathcal{H}_{k_{\nu,\gamma}}$, the magnitude of its Fourier transform $|\mathcal{F}[f](\omega)|$ decays polynomially fast as $|\omega| \to \infty$, and the speed of decay gets quicker as $\nu$ increases. Moreover, from C.7 and [51, Corollary 10.48], it follows that $\mathcal{H}_{k_{\nu,\gamma}}$ is norm-equivalent to the Sobolev space of order $\nu + d/2$.

## C.4. GP Sample Space

We first give three definitions and then present the central theorem, that allows us to characterize GP sample spaces with slightly modified versions of their respective RKHSs.

**Definition C.8** (A Version of a GP)**.** Let $f \sim \mathcal{GP}(m,k)$ be a Gaussian process with mean function $m : \mathcal{X} \to \mathbb{R}$ and covariance kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, where $\mathcal{X}$ is a nonempty set. Then a stochastic process $\tilde{f}$ on $\mathcal{X}$ is called a version of $f$, if $f(x) = \tilde{f}(x)$ holds with probability 1 for all $x \in \mathcal{X}$.

**Definition C.9** (Interior cone Condition)**.** A set $\mathcal{X} \subset \mathbb{R}^d$ is said to satisfy an interior cone condition if there exist an angle $\theta \in (0, 2\pi)$ and a radius $R > 0$ such that every $x \in \mathcal{X}$ is associated with a unit vector $\xi(x)$ so that the cone $C(x, \xi(x), \psi, R)$ is contained in $\mathcal{X}$, where

$$C(x, \xi(x), \psi, R) := \{x + ay : y \in \mathbb{R}^d, \|y\| = 1, \langle y, \xi(x) \rangle \geq \cos\psi, s \in [0, R]\}.$$

**Definition C.10** (Powers of RKHSs and kernels)**.** Let $\mathcal{X}$ be a compact metric space, $k$ be a continuous kernel on $\mathcal{X}$ with $\mathcal{H}_k$ being its RKHS, and $\nu$ be a finite Borel measure whose support is $\mathcal{X}$. Let $0 < \theta \leq 1$ be a constant, and assume that $\sum_{i \in I} \lambda_i^\theta \phi_i^2(x) < \infty$ holds for all $x \in \mathcal{X}$, where $(\lambda_i, \phi_i)_{i \in I}$ is the eigensystem of integral operator in (C.2). Then the $\theta$-th power of RKHS $\mathcal{H}_k$ is defined as

$$\mathcal{H}_k^\theta := \left\{ f = \sum_{i \in I} a_i \lambda_i^{\theta/2} \phi_i : \sum_{i \in I} a_i^2 < \infty \right\}, \tag{C.8}$$

where the inner-product is given by

$$\langle f, g\rangle_{\mathcal{H}_k^\theta} = \sum_{i\in I} \alpha_i\beta_i \quad \text{for} \quad f := \sum_{i\in I} \alpha_i \lambda_i^{\theta/2}\phi_i \in \mathcal{H}_k, \quad g := \sum_{i\in I} \beta_i \lambda_i^{\theta/2}\phi_i \in \mathcal{H}_k.$$

The $\theta$-th power of kernel $k$ is a function $k^\theta : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined by

$$k^\theta(x, y) := \sum_{i\in I} \lambda_i^\theta \phi_i(x)\phi_i(y), \quad x, y \in \mathcal{X}. \tag{C.9}$$

For a more intuitive understanding of the power parameter in the definition, we state [19, Remark 4.11] below.

**Remark C.11.** The power of the RKHS is an intermediate space (or more precisely, an interpolation space) between $L_2(\nu)$ and $\mathcal{H}_k$, and the constant $0 < \theta \le 1$ determines how close $\mathcal{H}_k^\theta$ is to $\mathcal{H}_k$ [45, Theorem 4.6]. For instance, if $\theta = 1$ we have $\mathcal{H}_k^\theta = \mathcal{H}_k$, and $\mathcal{H}_k^\theta$ approaches $L_2(\nu)$ as $\theta \to +0$. Indeed, $\mathcal{H}_k^\theta$ is nesting with respect to $\theta$:

$$\mathcal{H}_k = \mathcal{H}_k^1 \subset \mathcal{H}_k^\theta \subset \mathcal{H}_k^{\theta'} \subset L_2(\nu), \quad \text{for all } 0 < \theta' < \theta < 1.$$

In other words, $\mathcal{H}_k^\theta$ gets larger as $\theta$ decreases. If $\mathcal{H}_k$ is an RKHS consisting of smooth functions (such as Sobolev spaces), then $\mathcal{H}_k^\theta$ contains less smooth functions than those in $\mathcal{H}_k$.

Next we quote [19, Theorem 4.12] which is a special case of [44, Theorem 5.2]. It provides us with a characterization of GP sample spaces.

**Theorem C.12.** *Let $\mathcal{X}$ be a compact metric space, $k$ be a continuous kernel on $\mathcal{X}$ with $\mathcal{H}_k$ being its RKHS, and $\nu$ be a finite Borel measure whose support is $\mathcal{X}$. Let $0 < \theta < 1$ be a constant, and assume that $\sum_{i\in I} \lambda_i^2 \phi_i^2(x) < \infty$ holds for all $x \in \mathcal{X}$, where $(\lambda_i, \phi_i)_{i\in I}$ is the eigensystem of integral operator in (C.2). Consider $f \sim \mathcal{GP}(0, k)$. Then the following statements are equivalent.*

1. *$\sum_{i\in I} \lambda_i^{1-\theta} < \infty$.*
2. *The inclusion operator $I_{kk^\theta} : \mathcal{H}_k \to \mathcal{H}_k^\theta$ is Hilbert-Schmidt.*
3. *There exists a version $\tilde{f}$ of $f$ such that $\tilde{f} \in \mathcal{H}_k^\theta$ with probability 1.*

*Proof.* See [19, Theorem 4.12]. □

Finally we can quote [19, Corollary 4.15], which provides us with a characterization of the sample path properties for Matérn kernels.

**Lemma C.13** (Sample path properties for Matérn kernels)**.** Let $\mathcal{X} \subset \mathbb{R}^d$ be a bounded open set such that the boundary is Lipschitz and an interior cone condition is satisfied, and $k_{\nu,\gamma}$ be the Matérn kernel on $\mathcal{X}$ with parameters $\nu > 0$ and $\gamma > 0$ such that $\nu + d/2 \in \mathbb{N}$. Then for a given $f \sim \mathcal{GP}(0, k_{\nu,\gamma})$, there exists a version $\tilde{f}$ such that $\tilde{f} \in \mathcal{H}_{k_{\nu',\gamma'}}$ with probability 1 for all $\nu', \gamma' > 0$ satisfying $\nu > \nu' + d/2 \in \mathbb{N}$, where $\mathcal{H}_{k_{\nu',\gamma'}}$ is the RKHS of the Matérn kernel $k_{\nu',\gamma'}$ with parameters $\nu'$ and $\gamma'$.

*Proof.* See [19, Corollary 4.15]. □

# Bibliography

[1] Robert J Adler. *The Geometry of Random Fields*, volume 62. SIAM, 1981.

[2] Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

[3] Jørgen Bang-Jensen and Gregory Z Gutin. *Digraphs: theory, algorithms and applications.* Springer Science & Business Media, 2008.

[4] Thomas Bayes. Lii. an essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, frs communicated by mr. price, in a letter to john canton, amfr s. *Philosophical transactions of the Royal Society of London*, 53:370–418, 1763.

[5] Eric Brochu, Vlad M Cora, and Nando De Freitas. A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599*, 2010.

[6] Adam D Bull. Convergence rates of efficient global optimization algorithms. *Journal of Machine Learning Research*, 12(10), 2011.

[7] Federico Castelletti and Alessandro Mascaro. Bcdag: An r package for bayesian structure and causal learning of gaussian dags. *arXiv preprint arXiv:2201.12003*, 2022.

[8] Kathryn Chaloner and Isabella Verdinelli. Bayesian Experimental Design: A Review. *Statistical Science*, 10(3):273 – 304, 1995.

[9] Wenyu Chen, Mathias Drton, and Y Samuel Wang. On causal discovery with an equal-variance assumption. *Biometrika*, 106(4):973–980, 2019.

[10] Thomas H Cormen, Charles E Leiserson, Ronald L Rivest, and Clifford Stein. *Introduction to algorithms.* MIT press, 2022.

[11] T. M. Cover and J. A. Thomas. *Differential Entropy*, chapter 8, pages 243–261. John Wiley and Sons, Ltd, 2005.

[12] Robert Denk and Reinhard Racke. *Kompendium der ANALYSIS-Ein kompletter Bachelor-Kurs von Reellen Zahlen zu Partiellen Differentialgleichungen: Band 1: Differential-und Integralrechnung, Gewöhnliche Differentialgleichungen.* Springer-Verlag, 2011.

[13] Robert Denk and Reinhard Racke. *Kompendium der ANALYSIS-Ein kompletter Bachelor-Kurs von Reellen Zahlen zu Partiellen Differentialgleichungen: Band 2: Maß-und Integrationstheorie, Funktionentheorie, Funktionalanalysis, Partielle Differentialgleichungen*, volume 2. Springer-Verlag, 2012.

[14] R. M. Dudley. *Stochastic Processes*, page 439–486. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2 edition, 2002.

[15] Frederick Eberhardt. Almost optimal intervention sets for causal discovery. *arXiv preprint arXiv:1206.3250*, 2012.

[16] Frederick Eberhardt, Clark Glymour, and Richard Scheines. On the number of experiments sufficient and in the worst case necessary to identify all causal relations among n variables. *arXiv preprint arXiv:1207.1389*, 2012.

[17] Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

[18] Alison Gopnik. The scientist as child. *Philosophy of Science*, 63(4):485–514, 1996.

[19] M. Kanagawa, P. Hennig, D. Sejdinovic, and B. K. Sriperumbudur. Gaussian processes and kernel methods: A review on connections and equivalences. *Arxiv e-prints*, arXiv:1805.08845v1 [stat.ML], 2018.

[20] Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253, 2017.

[21] David Lindley. On a measure of the information provided by an experiment. *Annals of Mathematical Statistics*, 27:986–1005, 1956.

[22] Dennis V Lindley. A statistical paradox. *Biometrika*, 44(1/2):187–192, 1957.

[23] Marloes Maathuis, Mathias Drton, Steffen Lauritzen, and Martin Wainwright. *Handbook of graphical models.* CRC Press, 2018.

[24] Jonas Močkus. On bayesian methods for seeking the extremum. In *Optimization techniques IFIP technical conference*, pages 400–404. Springer, 1975.

[25] Jorge Nocedal and Stephen J Wright. *Numerical optimization.* Springer, 1999.

[26] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.

[27] Judea Pearl. Belief networks revisited. *Artificial intelligence in perspective*, pages 49–56, 1994.

[28] Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

[29] Judea Pearl. *Causality.* Cambridge university press, 2009.

[30] Jonas Peters and Peter Bühlmann. Identifiability of gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2014.

[31] Jonas Peters, Dominik Janzing, and Bernhard Schlkopf. *Elements of Causal Inference: Foundations and Learning Algorithms.* The MIT Press, 2017.

[32] Jonas Peters, Joris M. Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *J. Mach. Learn. Res.*, 15(1):2009–2053, jan 2014.

[33] Niklas Pfister, Peter Bühlmann, Bernhard Schölkopf, and Jonas Peters. Kernel-based tests for joint independence, 2016.

[34] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning).* The MIT Press, 2005.

[35] Maria L Rizzo. *Statistical computing with R.* Chapman and Hall/CRC, 2019.

[36] Christian P Robert, George Casella, and George Casella. *Monte Carlo statistical methods*, volume 2. Springer, 1999.

[37] James Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, 7(9-12):1393–1512, 1986.

[38] Robert W Robinson. Counting labeled acyclic digraphs. *New directions in the theory of graphs*, pages 239–273, 1973.

[39] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[40] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search.* MIT press, 2000.

[41] Niranjan Srinivas, Andreas Krause, Sham M. Kakade, and Matthias W. Seeger. Information-theoretic regret bounds for gaussian process optimization in the bandit setting. *IEEE Transactions on Information Theory*, 58(5):3250–3265, may 2012.

[42] Michael L Stein. *Interpolation of spatial data: some theory for kriging.* Springer Science & Business Media, 1999.

[43] I. Steinwart and A. Christmann. *Support Vector Machines.* Information Science and Statistics. Springer New York, 2008.

[44] Ingo Steinwart. Convergence types and rates in generic karhunen-loeve expansions with applications to sample path properties. *Potential Analysis*, 51(3):361–395, 2019.

[45] Ingo Steinwart and Clint Scovel. Mercer's theorem on general domains: On the interaction between measures, kernels, and rkhss. *Constructive Approximation*, 35:363–417, 2012.

[46] Takashi Takekawa. Fast parallel calculation of modified bessel function of the second kind and its derivatives. *SoftwareX*, 17:100923, 2022.

[47] G. E. Uhlenbeck and L. S. Ornstein. On the theory of the brownian motion. *Phys. Rev.*, 36:823–841, Sep 1930.

[48] Thomas S Verma and Judea Pearl. Equivalence and synthesis of causal models. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pages 221–236. 2022.

[49] Julius von Kügelgen, Paul K Rubenstein, Bernhard Schölkopf, and Adrian Weller. Optimal experimental design via bayesian optimization: active causal structure learning for gaussian process networks. *arXiv preprint arXiv:1910.03962*, 2019.

[50] Chunyi Wang and Radford M Neal. Mcmc methods for gaussian process models using fast approximations for the likelihood. *arXiv preprint arXiv:1305.2235*, 2013.

[51] Holger Wendland. *Scattered data approximation*, volume 17. Cambridge university press, 2004.

[52] Kun Zhang and Aapo Hyvarinen. On the identifiability of the post-nonlinear causal model. *arXiv preprint arXiv:1205.2599*, 2012.