Contents

List of Figures

columnwidth 421.10046pt

1 Introduction

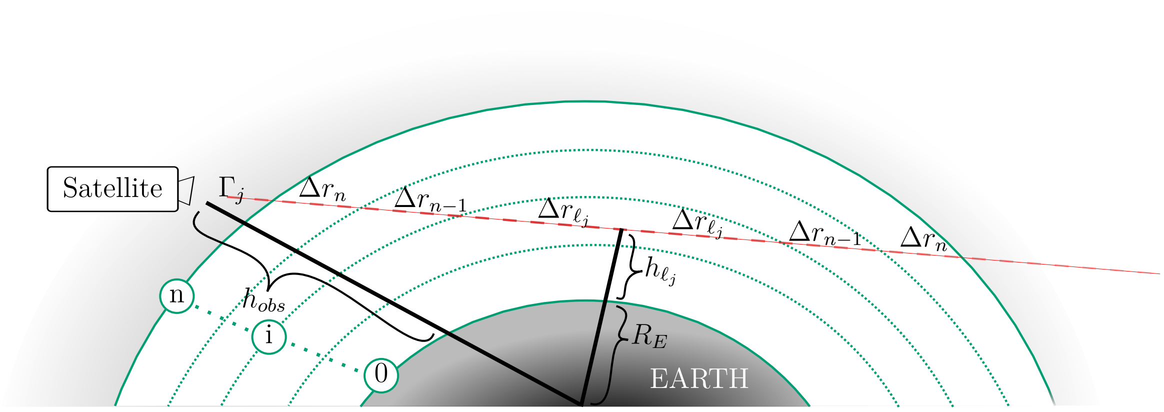
1.1 What is going on?, 3 facts, What is new in this thesis?

- hierachical Bayesian model, sampling to TT approx
- RTE as an example
- nonLinear to Linear Affine funciton (affine RTO)

1.2 Thesis Outline

2
Theoretical and Technical Background

2.1 Forward Model



In this section we describe the forward model which we use to simulate data and base the Bayesian inference on.

As shown in Figure **??**, one measurement of a stationary satellite can be describes as the path integral along the line of sight Γ_j for $j = 1, 2, \dots, m$. For each measurement we can define a tangent height h_{ℓ_j} as the shortest distance along the line of sight to the earth.

The j^{th} measurement, taken on line of sight Γ_j is modelled by the the radiative transfer equation (RTE) [1]

$$y_j = \int_{\Gamma_j} B(\nu,T) k(\nu,T) \frac{\boldsymbol{p}(T)}{k_{\mathrm{B}} \boldsymbol{T}(r)} \boldsymbol{x}(r) \tau(r) \mathrm{d}r + \eta_j \tag{2.1}$$

$$\tau(r) = \exp\left\{ - \int_{r_{\mathrm{obs}}}^{r} k(\nu,T) \frac{\boldsymbol{p}(T)}{k_B \boldsymbol{T}(r')} \boldsymbol{x}(r') \mathrm{d}r' \right\} \tag{2.2}$$

where the path from the satellite along the line-of-sight of the j^{th} pointing direction is Γ_j and the ozone concentration$\boldsymbol{x}(r)$ at distance r from the radiometer. The factor $\tau(r) \leq 1$ accounts for re-absorption of the radiation along the line-of-sight, which makes the RTE non linear. The noise η_j is added to each path integral, where the noise vector $\boldsymbol{\eta} \sim \mathcal{N}(0, \gamma^{-1}\mathbf{I})$ is normally distributed around zero with the noise precision γ. The absorption constant $k(\nu,T)$ for a single gas molecule at a specific wavenumber ν is given by the HITRAN database [2] and acts as a source function when multiplied with the black body radiation $B(\nu,T)$, given by Planck‘s law. Within the stratosphere the number density $p(T)/(k_{\mathrm{B}}T(r))$ of molecules is dependent on the pressure $p(T)$, the temperature $T(r)$, and the Boltzmann constant k_{B}. For fundamentals on the Radiative transfer equation we recommend 79BOOKRadiativeProcess.

We parametrize the ozone profile as a function of height, discretized into the n values in each of n layers of the discretized stratosphere where the i^{th} layer is defined by two spheres of radii $h_{i-1} < h_i$, $i = 1, \dots, n$, with h_0 and h_n. In between the heights h_{i-1} and h_i, each of the ozone concentration x_i, the pressure p_i, the temperature T_i, and thermal radiation is assumed to be constant. Above h_n and below h_0, the ozone concentration is set to zero, so no signal can be obtained. Then depending on the parameter of interest, which is either the ozone volume mixing ratio $\boldsymbol{x} = \{x_1, x_2, \dots, x_n\} \in \mathbb{R}^n$ or the fraction of pressure and temperature $\boldsymbol{p/T} = \{p_1/T_1, p_2/T_2, \dots, p_n/T_n\} \in \mathbb{R}^n$, we can rewrite the integral in Eq. (2.2) as e.g. $\boldsymbol{A_j}(\boldsymbol{x}, \boldsymbol{p}, \boldsymbol{T})\, \boldsymbol{x}$, where the absorption $\tau(r)$ induces non-linearity. Here, the row vector $\boldsymbol{A_j}(\boldsymbol{x}, \boldsymbol{p}, \boldsymbol{T}) \in \mathbb{R}^n$ defines a Kernel for each measurement so that the data vector

$$\boldsymbol{y} = \boldsymbol{A}(\boldsymbol{x}, \boldsymbol{p}, \boldsymbol{T})\, \boldsymbol{x} + \boldsymbol{\eta} = \boldsymbol{A}(\boldsymbol{x}, \boldsymbol{p}, \boldsymbol{T})\, \frac{\boldsymbol{p}}{\boldsymbol{T}} + \boldsymbol{\eta}\,. \tag{2.3}$$

can be written as a matrix vector multiplication, where the matrix $\boldsymbol{A}(\boldsymbol{x}, \boldsymbol{p}, \boldsymbol{T}) \in \mathbb{R}^{m\times n}$ and the noise vector $\boldsymbol{\eta} \in \mathbb{R}^{m}$.

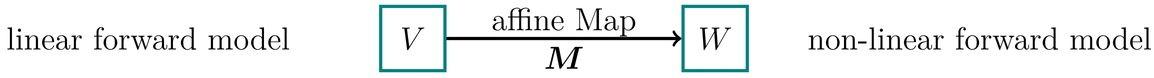
Since the absorption $\tau(r)$ reduces measurements by of order 1%, or less, making the inverse problem only weakly non-linear. We use that to approximate the non-linear forward model $\boldsymbol{A}(\boldsymbol{x}, \boldsymbol{p}, \boldsymbol{T})$ with a map $\boldsymbol{M}$ so that $\boldsymbol{A}(\boldsymbol{x}, \boldsymbol{p}, \boldsymbol{T}) \approx \boldsymbol{M}\boldsymbol{A}_L$ Where each row $\boldsymbol{A}_{L,j}$ of matrix as $\boldsymbol{A}_L \in \mathbb{R}^{m\times n}$ is defined by the linear forward model, where absorption is neglected, e.g. $\tau = 1$. Then $\boldsymbol{A}_{L,j}$ is either defined by $B(\nu, T)S(\nu, T)\frac{\boldsymbol{p}(T)}{k_{\mathrm{B}}\boldsymbol{T}(r)}\mathrm{d}r$ or $B(\nu, T)S(\nu, T)\frac{\boldsymbol{x}}{k_{\mathrm{B}}}\mathrm{d}r$, as in Eq.. (2.2), depending on the parameter of interest. This poses a linear inverse problem with the forward map defined by the matrix $\boldsymbol{A} = \boldsymbol{M}\boldsymbol{A}_L$, where $\boldsymbol{M}$ is, more specifically, an affine map.

2.2 Affine Map

To approximate the non-linear forward model we use an affine map $M : \boldsymbol{A}_L\boldsymbol{x} \rightarrow \boldsymbol{A}(\boldsymbol{x}, \boldsymbol{p}, \boldsymbol{T}))\boldsymbol{x}$, which maps the linear forward model $\boldsymbol{A}_L\boldsymbol{x}$ onto the non-linear forward model $\boldsymbol{A}(\boldsymbol{x}, \boldsymbol{p}, \boldsymbol{T})\boldsymbol{x}$.

An affine map is any linear map in between two vector spaces is or affine spaces, where in affine space does not need to have a zero origin. 2.3.1. PROPOSITION AND DEFINITIOn Berge book[]. In other words an affine map does not need to preserve the origin, or is a linear map on vector spaces including translation, or in the words of my supervisor, C. F., an affine map is a Taylor series of first order. For more information on affine spaces and maps we refer to [**two books**]

We generate two affine subspaces spaces
$V = \left\{\boldsymbol{A}(\boldsymbol{x}^{(1)}, \boldsymbol{p}, \boldsymbol{T}), \dots, \boldsymbol{A}(\boldsymbol{x}^{(m)}, \boldsymbol{p}, \boldsymbol{T})\right\}$ and $W = \left\{\boldsymbol{A}\boldsymbol{x}^{(1)}, \dots, \boldsymbol{A}\boldsymbol{x}^{(m)}\right\}$ over the same field, with fixed $\boldsymbol{p}, \boldsymbol{T}$. The parameter $\boldsymbol{x}$ is distributed as the so-called posterior distribution $\left\{\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(m)}\right\} \sim \pi(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$, with hyper-parameters $\boldsymbol{\theta}$, according to a Bayesian hierarchical model.



$$A(x,p,T)x \approx MA_L x = Ax$$

Figure 2.1: Schematics of Affine Map, which approximates the linear forward model to the non-linear forward model.

2.3 Bayesian Inference

In this this section we give a short introduction to Bayesian inference for a general parameter $\boldsymbol{x}$ given some data $\boldsymbol{y}$, later in section **??** we set up a more sophisticated Bayesian framework applied to the forward model in section **??**.

We can visualise the correlation structure of a measurement process through a hierarchiallly ordered directed acyclic graph (DAG), see Figure 2.2. As an observatory process naturally includes some random noise we include that in our DAG and classify the noise as a hyper-parameter in $\boldsymbol{\theta}$. Other hyper-parameters influence the parmeters $\boldsymbol{x}$ detemernistaclly, which are then mapped through the forward model onto the space of all measurables $\boldsymbol{u}$, from which we observe some data $\boldsymbol{y}$ including noise as previously mentioned. Drawing a DAG can help us to dependences within the measurement and modelling process. Given some data we inferer the distribution of the underling parameters and hyper-parameters by following the arrows in Figure **??** backwards and set up a Bayesain hierachlly ordered model.

Within a linear Bayesian hierarchial model we need to define a likelihood function as well as distribution over the unknown parameters $\boldsymbol{x}$ and hyper-parameters $\boldsymbol{\theta}$.

$$\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{A}\boldsymbol{x}, \boldsymbol{\Sigma}(\boldsymbol{\theta})) \tag{2.4a}$$

$$\boldsymbol{x}|\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{Q}^{-1}(\boldsymbol{\theta})) \tag{2.4b}$$

$$\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta}), \tag{2.4c}$$

with the noise covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, so that $\boldsymbol{\eta} \sim \mathcal{N}(0, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$ as in Eq. **??**, the prior precision matrix $\boldsymbol{Q}(\boldsymbol{\theta})$, prior mean $\boldsymbol{\mu}$ and some prior distribution over the hyper-parameters $\pi(\boldsymbol{\theta})$. Through sensibly choosing the prior distributions

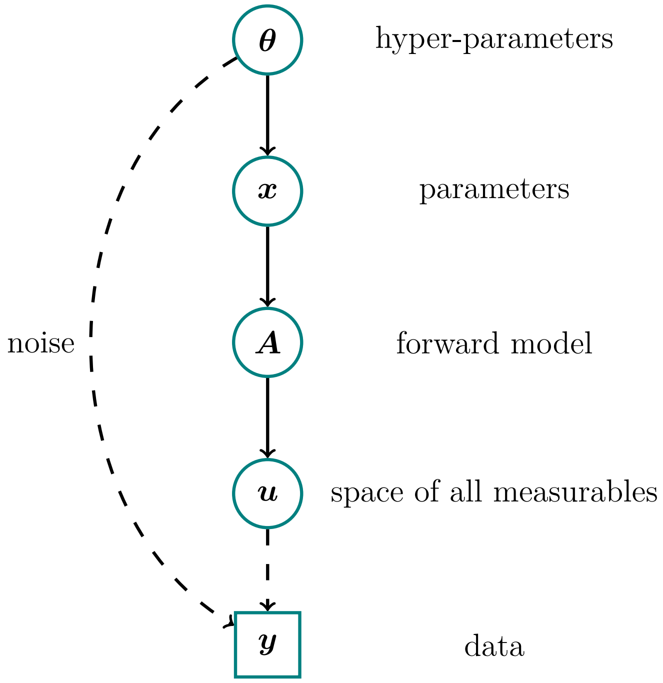
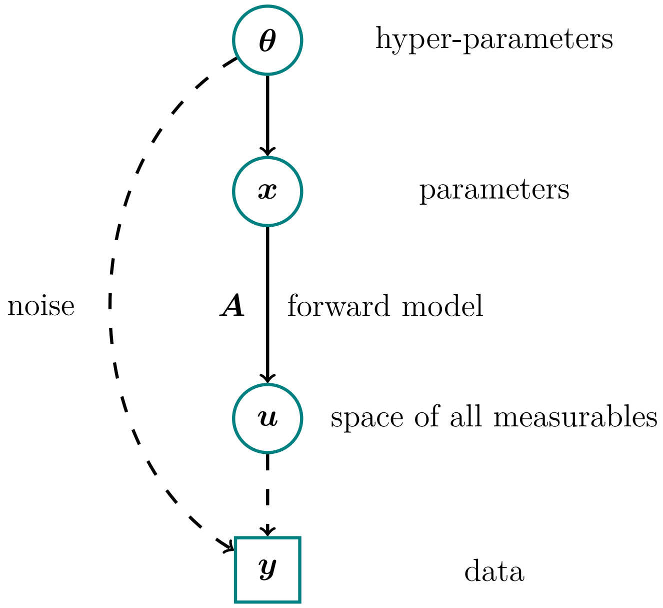


Figure 2.2: The directed acyclic graph (DAG) for a typical linear inverse problem visualises forward dependencies as solid line arrows for deterministic dependencies and dotted arrows for statistical dependencies. Naturally the data $\boldsymbol{y}$ has some noise described through included in some hyper-parameters $\boldsymbol{\theta}$. The parameters $\boldsymbol{x}$ have some dependency of those hyper-parameters $\boldsymbol{\theta}$. The parameter $\boldsymbol{x}$ is mapped onto the space of all measurables $\boldsymbol{u}$ through the linear forward model $\boldsymbol{A}$, so that $\boldsymbol{Ax}$ is a linear operation. From the space of all measurables we can observe some data $\boldsymbol{y}$, statistically, where as prevoiusly mentioned some random noise is added. We set up a more sophisticated Bayesian model in chapter **??** explicitly including all hyper-parameters and parameters of interest according to the forward model in section **??**.



$\pi(\boldsymbol{x}|\boldsymbol{y})$ as well as the hyper-parameters $\boldsymbol{\theta}$ and their prior distribution $\pi(\boldsymbol{\theta})$, we can incorporate functional dependencies as well physical properties of the parameters. The likelhood function $\pi(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})$ is a measure of how the parameters and hyper-parameters fit to the data according to our forward model, including information of the measurement process.

With a normally distributed prior and likelihood function this becomes a linear-Gaussian Bayesian hierarchical model. For more detailed Bayesian analysis we recommend [].

The posterior distribution, the function of interest,

$$\pi(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y}) = \frac{\pi(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})\pi(\boldsymbol{x},\boldsymbol{\theta})}{\pi(\boldsymbol{y})}\,, \tag{2.5}$$

is given according to Bayes' theorem [], with the prior distribution $\pi(\boldsymbol{x},\boldsymbol{\theta})$ and the normalising constant $\pi(\boldsymbol{y})$. If the normalising constant is finite and non-zero we can approximate the posterior distribution

$$\pi(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y}) \propto \pi(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})\pi(\boldsymbol{x},\boldsymbol{\theta})\,. \tag{2.6}$$

Then the expectation of any a function $h(\boldsymbol{x},\boldsymbol{\theta})$ can be described as

$$\mathrm{E}_{\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y}}[h(\boldsymbol{x})] = \int\int h(\boldsymbol{x},\boldsymbol{\theta})\,\pi(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{\theta}\,, \tag{2.7}$$

which is usually a high dimensional integral and computationally not feasible to solve.

One way to work around the high dimensionality is to parameterise $\boldsymbol{x}$ using hyper-parameters $\boldsymbol{\theta}$ so that $\boldsymbol{x}(\boldsymbol{\theta})$. Another way is to seperate the posterior distribution over latent field $\boldsymbol{x}$ and the hyper-parameters $\boldsymbol{\theta}$. This is particular benefitial, when $\boldsymbol{x}$ is high dimensional, e.g. $\boldsymbol{x} \in \mathbb{R}^n$ with $n = 45$ and can not be parametrised, and $\boldsymbol{\theta}$ is low dimensional, e.g. two dimensional.

2.3.1 Marginal and then Conditional

The marginal and then conditional (MTC) method factorises the full posterior distribution

$$\pi(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y}) = \pi(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})\pi(\boldsymbol{\theta}|\boldsymbol{y}) \tag{2.8}$$

into the marginal posterior distribtion $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ and conditional posterior distribution $\pi(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})$.

For the in Eq. **??** specified linear-Gaussian Bayesian hierarchical model the marginal posterior distribution is given as

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) = \int \pi(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})\,\mathrm{d}\boldsymbol{x} \tag{2.9}$$

$$\propto \sqrt{\frac{\det(\boldsymbol{\Sigma}^{-1})\,\det(\boldsymbol{Q})}{\det(\boldsymbol{Q}+\boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{A})}} \times \exp\Big[-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu})^T\boldsymbol{Q}_{\boldsymbol{\theta}|\boldsymbol{y}}(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu})\Big]\pi(\boldsymbol{\theta})\,, \tag{2.10}$$

with

$$\boldsymbol{Q}_{\boldsymbol{\theta}|\boldsymbol{y}} = \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{A}+\boldsymbol{Q})^{-1}\boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}\,. \tag{2.11}$$

See lemma [].

Then conditioned on the hyper-parameters $\boldsymbol{\theta}$ we can draw samples of the conditional posterior distribution

$$\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta} \sim \mathcal{N}\Big(\boldsymbol{\mu} + (\boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{A}+\boldsymbol{Q})^{-1}\boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu}), (\boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{A}+\boldsymbol{Q})^{-1}\Big)\,, \tag{2.12}$$

see section **??** or calculate weighted expectations of a function $h(\boldsymbol{x})$

$$\mathrm{E}_{\boldsymbol{x}|\boldsymbol{y}}[h(\boldsymbol{x})] = \int \mathrm{E}_{\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y}}[h(\boldsymbol{x})]\,\pi(\boldsymbol{\theta}|\boldsymbol{y})\,\mathrm{d}\boldsymbol{\theta}\,, \tag{2.13}$$

with weights given by $\pi(\boldsymbol{\theta}|\boldsymbol{y})$. [] Note that the noise covariance $\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})$ and the prior precision $\boldsymbol{Q} = \boldsymbol{Q}(\boldsymbol{\theta})$ are depending on hyper-parameters $\boldsymbol{\theta}$.

In this thesis we will use sampling and deterministic methods to characterise the posterior distribution over the hyper-parameters and present the basics of those in the following sections.

2.4 Sampling Methods

In this section we present the sampling based methods used in this thesis to generate an ergodic Markov-Chain $(\boldsymbol{x},\boldsymbol{\theta})^{(0)},\dots,(\boldsymbol{x},\boldsymbol{\theta})^{(k)},\dots,(\boldsymbol{x},\boldsymbol{\theta})^{(N)} \sim \pi(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})$, where the samples $(\boldsymbol{x},\boldsymbol{\theta})^{(k)}$ are distributed as $\pi(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})$. Ergodicity is implied if an aperiodic and irreduce chain **??** proves to be reversible, then the chain converges and has a unique equilibrium distribution. In other words if from a state in the chain we can reach every other state in the sampling space and the previous state, and we do not get stuck in periodic loop, then it converges. Instead of proving ergodicity we can in practise look e.g. at the output samples and their trace $\pi(\boldsymbol{x}^{(k)}, \boldsymbol{\theta}^{(k)}|\boldsymbol{y})$ to see convergence.

Then for large enough N the samples based estimate of Eq. **??** and of any function $h(\boldsymbol{x}, \boldsymbol{\theta})$ is

$$\mathrm{E}_{\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y}}[h(\boldsymbol{x}, \boldsymbol{\theta})] \approx \frac{1}{N} \sum_{k=1}^{N} h(\boldsymbol{x}^{(k)}, \boldsymbol{\theta}^{(k)}) \,. \tag{2.14}$$

In practise of this thesis we use Markoc-chain Monte-Carlo (MCMC) methods on target distributions such as $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ and so we will illustrate the sampling procedures for the following three subsections on that distribution.

2.4.1 Metropolis

The Metropolis algorithm is special case of the Metropolis-Hastings algorithm, with a symmetric proposal distribution $q(i|j) = q(j|i)$ [].

The Metropolis-Hastings algorithm starts with a initial sample $\boldsymbol{\theta}^{(t)}$ at at $t = 0$. We propose a new sample $\boldsymbol{\theta} \sim q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t-1)})$ according to the proposal distribution $q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t-1)})$ given the previous state. Then accept and set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}$ with

$$\alpha(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t-1)}) = \min \left\{ 1, \frac{\pi(\boldsymbol{\theta}|\boldsymbol{y}) q(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{\theta})}{\pi(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{y}) q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t-1)})} \right\} \tag{2.15}$$

or reject and keep $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)}$, which we do by comparing α to a uniform random number $u \sim \mathcal{U}(0, 1)$. Note that symmetrical proposal distribution $q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t-1)}) = q(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{\theta})$ cancel, then we call that a Metropolis-algorithm []. To assure convergence independent of the intial sample we discard samples after the so-called burn-in period and effectivally generate a Markov-Chain of length $N - N_{\text{burn-in}}$.

Since the proposal distribution is symmetric, Metropolis chains are reversible and as you can see in $\alpha(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t-1)})$ converge towards $\pi(\boldsymbol{\theta}|\boldsymbol{y})$. A more mathematical prove of ergodicity for the general Metropolis-Hastings algorithm can be found in [].

Algorithm 1: Metropolis

```
Initialize $\boldsymbol{\theta}^{(0)}$
for  $k = 1, \dots, N$ do
    Propose $\boldsymbol{\theta} \sim q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t-1)}) = q(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{\theta})$
    Compute
        $\alpha(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t-1)}) = \min\left\{1, \frac{\pi(\boldsymbol{\theta}|\boldsymbol{y})\cancel{q(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{\theta})}}{\pi(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{y})\cancel{q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t-1)})}}\right\}$
    Draw $u \sim \mathcal{U}(0, 1)$
    if $\alpha \geq u$  then
        Accept and set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}$
    else
        Reject and keep $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)}$
    end if
end for
Output: $\boldsymbol{\theta}^{(0)}, \dots, \boldsymbol{\theta}^{(k)}, \dots, \boldsymbol{\theta}^{(N)} \sim \pi(\boldsymbol{\theta}|\boldsymbol{y})$
```

2.4.2 Gibbs Sampling

Gibb sampling is a special case of the metropolis hastings algorihtm with the acceptance probaility of 1. In Gibbs sampling according to [] we move one directional by fixing all other direction.

Assume $\boldsymbol{\theta} \in \mathbb{R}^d$ is d-dimesnional and $\{\boldsymbol{\theta}_{<j}, \boldsymbol{\theta}_{>j}\} = \{\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{j-1}, \boldsymbol{\theta}_{j+1}, \dots, \boldsymbol{\theta}_d\}$ denotes all dimesnions except $\boldsymbol{\theta}_j$. Then to draw a new sample $\boldsymbol{\theta}^{(t)}$, we iterate through each dimension to draw a sample $\boldsymbol{\theta}_j^{(t)} \sim \pi(\boldsymbol{\theta}_j|\boldsymbol{\theta}_{<j}, \boldsymbol{\theta}_{>j}, \boldsymbol{y})$ from the conditional distribution.

Algorithm 2: Gibbs

```
Initialize $\boldsymbol{\theta}^{(0)} = \{\boldsymbol{\theta}_1^{(0)}, \dots, \boldsymbol{\theta}_j^{(0)}, \dots, \boldsymbol{\theta}_d^{(0)}\}$.
for  $k = 1, \dots, N$ do
    for  $j = 1, \dots, d$ do
        Draw $\boldsymbol{\theta}_j^{(t)} \sim \pi(\boldsymbol{\theta}_j|\boldsymbol{\theta}_{<j}, \boldsymbol{\theta}_{>j}, \boldsymbol{y})$
    end for
end for
Output: $\boldsymbol{\theta}^{(0)}, \dots, \boldsymbol{\theta}^{(k)}, \dots, \boldsymbol{\theta}^{(N)} \sim \pi(\boldsymbol{\theta}|\boldsymbol{y})$
```

If the target distribution is well behaved $\pi(\boldsymbol{\theta}_j|\boldsymbol{\theta}_{<j}, \boldsymbol{\theta}_{>j}, \boldsymbol{y}) > 0$ in the sampling space the chain produced by the gibbs sampling algorithm is aperiodic and ireducabile. We can also show that the criterium, the detailed balance condition, for reveribility is met []. Then a Gibbs sampler prodices an ergodic markov chain.

2.4.3 t-walk

We use the t-walk sampler developed By Christens and Fox as a black box sampler[]. Within the t-walk there are three differnt moves in the sampling space availbe and by construction of the t-walk is ergodic, see section [].

2.4.4 Draw a sample from a multivariate normal distribution

for Linear Gaussian Bayesian hierarchical model We need to draw a sample from the multivariate normal distribution. We can do this using the radomize then optimise (RTO) method [], with a perturb exponential.

In this thesis we use the RTO method to draw a sample from the full conditional distribution $\pi(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})$, see section **??**, which is a normal distribution we can rewrite to:

$$\begin{aligned} \pi(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}) &\propto \pi(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta}) && (2.16)\\ &= \exp\|\hat{\boldsymbol{A}}\boldsymbol{x} - \hat{\boldsymbol{y}}\|^2 \,, && (2.17) \end{aligned}$$

where

$$\hat{\boldsymbol{A}} = \begin{bmatrix} \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\theta})\boldsymbol{A} \\ \boldsymbol{Q}^{1/2}(\boldsymbol{\theta}) \end{bmatrix} , \quad \hat{\boldsymbol{y}} = \begin{bmatrix} \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\theta})\boldsymbol{y} \\ \boldsymbol{Q}^{1/2}(\boldsymbol{\theta})\boldsymbol{\mu} \end{bmatrix} []. \tag{2.18}$$

Then one sample can be computed by minimising the following equation with respect to $\hat{\boldsymbol{x}}$:

$$\boldsymbol{x}_i = \arg\min_{\hat{\boldsymbol{x}}} \|\hat{\boldsymbol{A}}\hat{\boldsymbol{x}} - (\hat{\boldsymbol{y}} + \boldsymbol{\eta})\|^2, \quad \boldsymbol{\eta} \sim \mathcal{N}(\boldsymbol{0}, \mathbf{I}) \,, \tag{2.19}$$

where we add a randomised perturbation $\boldsymbol{\eta}$. Next, we substitute $-\hat{\boldsymbol{A}}^T\boldsymbol{\eta} = \boldsymbol{v}_1 + \boldsymbol{v}_2$ we can rewrite the argument of Eq. 2.18 to

$$(\boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})\boldsymbol{A} + \boldsymbol{Q}(\boldsymbol{\theta}))\boldsymbol{x}_i = \boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})\boldsymbol{y} + \boldsymbol{Q}(\boldsymbol{\theta})\boldsymbol{\mu} + \boldsymbol{v}_1 + \boldsymbol{v}_2 \,, \tag{2.20}$$

where $\boldsymbol{v}_1 \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})\boldsymbol{A})$ and $\boldsymbol{v}_2 \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}(\boldsymbol{\theta}))$ are independent random variables [].

2.5 Numerical Approxiamtion Methods - Tensor Train

Using the tensor train format to approximate multidimensional functions π enables us to compute marginal posterior probability distribution cheaply. As the name suggest the tensor train format is a train of tensors, more specifically two and three dimensional tensors which we call cores π_k and which are connected through a rank r. As in Figure **??** displayed, each tensor $\pi_k \in \mathbb{R}^{r_{k-1} \times n \times r_k}$ of a d - dimensional space is basically a three dimesnional matrix, for $k = 1, \ldots, d$. gridsize n In the first and last dimension we have two dimensional tensor where $r_0 = r_d = 1$.

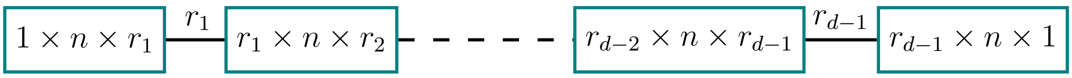

$1 \times n \times r_1$
r_1
$r_1 \times n \times r_2$
$r_{d-2} \times n \times r_{d-1}$
r_{d-1}
$r_{d-1} \times n \times 1$

Figure 2.3: text

Figure 2.4: nice matries picture

Then we find the marginal function by integrating over each of the cores

$$\begin{aligned} f_{X_k}(x_k) = \Bigg| \left(\int_{\mathbb{R}} \pi_1(x_1) \mathrm{d}x_1 \right) \cdots \left(\int_{\mathbb{R}} \pi_{k-1}(x_{k-1}) \mathrm{d}x_{k-1} \right) \\ \pi_k(x_k) \left(\int_{\mathbb{R}} \pi_{k+1}(x_{k+1}) \mathrm{d}x_{k+1} \right) \cdots \left(\int_{\mathbb{R}} \pi_d(x_d) \mathrm{d}x_d \right) \Bigg| . \end{aligned} \tag{2.21}$$

define

For more numerical stability we approximate the square root of $\sqrt{\pi}$, where the tt-cross [] gives us the cores $\boldsymbol{A}_k$. We follow the noitation of [] closely.

and $\boldsymbol{A}_k \in \mathbb{R}^{r_{k-1} \times n_k \times r_k}$ is the associated k-th coefficient tensor. For the k-th set of basis functions, we define the mass matrix $\boldsymbol{M}_k \in \mathbb{R}^{n_k \times n_k}$ by

$$\boldsymbol{M}_k[i,j] = \int_{X_k} \phi_k^{(i)}(x_k) \phi_k^{(j)}(x_k) \lambda(x_k)\, dx_k, \quad i = 1, ..., n_k, \quad j = 1, ..., n_k, \tag{2.22}$$

where $\{\phi_k^{(i)}(x_k)\}_{i=1}^{n_k}$ is the set of basis functions for the k-th coordinate. In the case of cartesian basis with $\lambda = 1$ M is the identity matrix, which we need to compute the marginals. For a boundend Parameter space we consider the weihtin funcitn $\lambda(x) = 1$ lesbegue measure

2.5.1 Marginal Functions

We defineγ'

The marginal PDF of X_1 can be expressed as

$$f_{X_1}(x_1) = \frac{1}{z}\left(\gamma' \prod_{i=2}^{d} \lambda_i(X_i) + \sum_{l_1=1}^{r_1}\left(\sum_{i=1}^{n_1} \phi_1^{(i)}(x_1)\boldsymbol{D}_1[i,l_1]\right)^2\right)\lambda_1(x_1), \tag{2.23}$$

where $\boldsymbol{D}_1[i,l_1] = \boldsymbol{B}_1[\alpha_0, i, l_1]$ and $\alpha_0 = 1$.

The marginal PDF of X_n can be expressed as

$$f_{X_n}(x_n) = \frac{1}{z}\left(\gamma' \prod_{i=1}^{d-1} \lambda_i(X_i) + \sum_{l_{n-1}=1}^{r_{n-1}}\left(\sum_{i=1}^{n_1} \phi_1^{(i)}(x_1)\boldsymbol{D}_n[l_{n-1},i]\right)^2\right)\lambda_n(x_n), \tag{2.24}$$

where $\boldsymbol{D}_n[l_{n-1},i] = \boldsymbol{B}_{pre,n}[l_{n-1},i,\alpha_n]$ and $\alpha_n = 1$.

We caclute the amtrixes $\boldsymbol{B}_k$ through backward marginalisation and $\boldsymbol{B}_{pre,n}$

Through process called orthogonalisation [] as used in [].

This subsection to caculte $\boldsymbol{B}_k$ is taken from [].

Proposition 1 (Backward Marginalisation)**:** Starting with the last coordinate $k = d$, we set $\boldsymbol{B}_d = \boldsymbol{A}_d$. The following procedure can be used to obtain the coefficient tensor $\boldsymbol{B}_{k-1} \in \mathbb{R}^{r_{k-2}\times n_{k-1}\times r_{k-1}}$, which we need for defining the marginal function $f_{X_k}(x_k)$:

1. Use the Cholesky decomposition of the mass matrix, $\boldsymbol{L}_k\boldsymbol{L}_k^\top = \boldsymbol{M}_k \in \mathbb{R}^{n_k\times n_k}$, to construct a tensor $\boldsymbol{C}_k \in \mathbb{R}^{r_{k-1}\times n_k\times r_k}$:

$$\boldsymbol{C}_k[\alpha_{k-1},\tau,l_k] = \sum_{i=1}^{n_k} \boldsymbol{B}_k[\alpha_{k-1},i,l_k]\boldsymbol{L}_k[i,\tau]. \tag{2.25}$$

2. Unfold $\boldsymbol{C}_k$ along the first coordinate and compute the thin QR decomposition, so that $\boldsymbol{C}_k^{(R)} \in \mathbb{R}^{r_{k-1}\times(n_k r_k)}$:

$$\boldsymbol{Q}_k\boldsymbol{R}_k = \left(\boldsymbol{C}_k^{(R)}\right)^\top. \tag{2.26}$$

3. Compute the new coefficient tensor:

$$\boldsymbol{B}_{k-1}[\alpha_{k-2},i,l_{k-1}] = \sum_{\alpha_{k-1}=1}^{r_{k-1}} \boldsymbol{A}_{k-1}[\alpha_{k-2},i,\alpha_{k-1}]\boldsymbol{R}_k[l_{k-1},\alpha_{k-1}]. \tag{2.27}$$

Then we need to do this the other way as well.

In addiotn we also have to do forward marginalistion starig with the first dimension

Proposition 2 (Forward Marginalistaion)**:** Starting with the first coordinate $k = 1$, we set $\boldsymbol{B}_{pre,1} = \boldsymbol{A}_1$. The following procedure can be used to obtain the coefficient tensor $\boldsymbol{B}_{pre,k+1} \in \mathbb{R}^{r_k \times n_{k+1} \times r_{k+1}}$ for defining the marginal function $f_{X_k}(x_k)$:

1. Use the Cholesky decomposition of the mass matrix, $\boldsymbol{L}_k \boldsymbol{L}_k^\top = \boldsymbol{M}_k \in \mathbb{R}^{n_k \times n_k}$, to construct a tensor $\boldsymbol{C}_k \in \mathbb{R}^{r_{k-1} \times n_k \times r_k}$:

$$\boldsymbol{C}_{pre,k}[\alpha_{k-1}, \tau, l_k] = \sum_{i=1}^{n_k} \boldsymbol{L}_k[i, \tau] \boldsymbol{B}_{pre,k}[\alpha_{k-1}, i, l_k]. \tag{2.28}$$

2. Unfold $\boldsymbol{C}_{pre,k}$ along the first coordinate and compute the thin QR decomposition, so that $\boldsymbol{C}_{pre,k}^{(R)} \in \mathbb{R}^{(r_{k-1} n_k) \times r_k}$:

$$\boldsymbol{Q}_{pre,k} \boldsymbol{R}_{pre,k} = (\boldsymbol{C}_{pre,k}^{(R)}). \tag{2.29}$$

3. Compute the new coefficient tensor $\boldsymbol{B}_{pre,k+1} \in \mathbb{R}^{r_{k-1} \times n_k \times r_k}$:

$$\boldsymbol{B}_{pre,k+1}[l_{k+1}, i, \alpha_{k+1}] = \sum_{\alpha_k = 1}^{r_k} \boldsymbol{R}_{pre,k}[l_{k+1}, \alpha_k] \boldsymbol{A}_{k+1}[\alpha_k, i, \alpha_{k+1}]. \tag{2.30}$$

The marginal PDF of X_k can be expressed as

$$f_{X_k}(x_k) = \frac{1}{z}\left(\gamma' \prod_{i=1}^{k-1} \lambda_i(X_i) \prod_{i=k+1}^{d} \lambda_i(X_i) + \sum_{l_{k-1}=1}^{r_{k-1}} \sum_{l_k=1}^{r_k} \left(\sum_{i=1}^{n_k} \phi_k^{(i)}(x_k) \boldsymbol{D}_k[l_{k-1}, i, l_k]\right)^2\right) \lambda_k(x_k), \tag{2.31}$$

where $\boldsymbol{D}_k \in \mathbb{R}^{r_{k-1} \times n_k \times r_k}$ and $\boldsymbol{R}_{pre,k-1} \in \mathbb{R}^{r_{k-1} \times r_{k-1}}$ and $\boldsymbol{B}_k \in \mathbb{R}^{r_{k-1} \times n_k \times r_k}$

$$\boldsymbol{D}_k[l_{k-1}, i, l_k] = \sum_{\alpha_{k-1}=1}^{r_{k-1}} \boldsymbol{R}_{pre,k-1}[l_{k-1}, \alpha_{k-1}] \boldsymbol{B}_k[\alpha_{k-1}, i, l_k]. \tag{2.32}$$

Appendices

References

[1] *Handbook for the Montreal protocol on substances that deplete the ozone layer.* Nairobi: The Secretariat of The Vienna Convention for the Protection of the Ozone Layer and The Montreal Protocol on Substances that Deplete the Ozone Layer, United Nations Environment Programme, 2006.

[2] Iouli E Gordon et al. “The HITRAN2020 molecular spectroscopic database”. In: *Journal of Quantitative Spectroscopy and Radiative Transfer* 277 (2022), p. 107949.