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

In this chapter, we provide a brief introduction to the methods used in this thesis. We keep the discussion as general as possible, as more specific details will be presented in the results section **??**. We begin by introducing the forward model in section 2.1, which we use to simulate the data. Since the forward model is weakly non-linear, we employ an affine transformation, see section 2.2, to project the linear model onto the non-linear one, allowing us to treat the problem as a linear inverse problem. This enables the application of Bayesian inference in section 2.3, where we formulate a hierarchical linear-Gaussian model to define and structure the posterior distribution. For comparison, we briefly present the Tikhonov regularization approach, see section 2.4. In section 2.5, we introduce Markov Chain Monte Carlo (MCMC) methods to sample from the posterior distribution. Finally, in section 2.6, instead of sampling, we can approximate the posterior distribution using the tensor train format.

2.1 Forward Model

The forward model is based on a satellite measuring through the amtosphere, known as limb sounding, as shown in Figure 2.1. One measurement y_j of a stationary satellite can be describes as the path integral through the atmosphere along the

line of sight, for $j = 1, 2, \ldots, m$. For each measurement we can define a tangent height h_{ℓ_j} as the shortest distance along the line of sight to the earth.

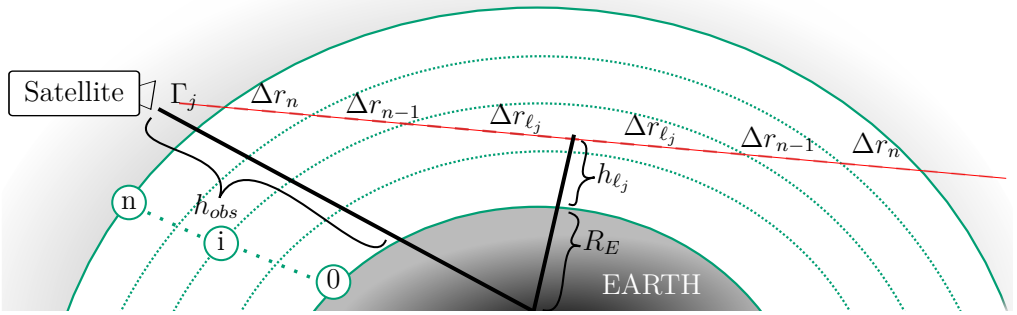

Figure 2.1: This figure illustrates a limb sounding measurement setup, specifically how the line of sight of a satellite at altitude h_{obs} is partitioned according to a discretized atmospheric model. The atmosphere is divided into n layers, allowing the line of sight Γ_j to be discretized into segments Δr_i for $i = \ell_j, \ldots, n$. Here, $\ell_j \in \mathbb{N}$ denotes the index corresponding to the tangent height h_{ℓ_j} relative to the Earth's radius R_E. This setup forms the basis for the numerical solution of the integral in Eq. 2.1, known as the radiative transfer equation.

The j^{th} measurement, is modelled by the radiative transfer equation (RTE) [1]

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

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

where the path from the satellite along the line-of-sight of the j^{th} pointing direction is Γ_j and the ozone concentration at distance r from the radiometer is $x(r)$ plus some noise η_j. Within the stratosphere the number density $p(T)/(k_{\text{B}} T(r))$ of molecules is dependent on the pressure $p(T)$, the temperature $T(r)$, and the Boltzmann constant k_{B}. 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 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 [3]. For fundamentals on the Radiative transfer equation we recommend [3, Chapter 1].

To enable Matrix-Vector multiplication, we parametrise the ozone profile as a function of height, discretised into the n values for each of n layers of the discretised atmosphere where the i^{th} layer is defined by two spheres of radii $h_{i-1} < h_i$, $i = 1, \ldots, n$, 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 $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\} \in \mathbb{R}^n$ or the fraction of pressure and temperature $\boldsymbol{p/T} = \{p_1/T_1, p_2/T_2, \ldots, p_n/T_n\} \in \mathbb{R}^n$, we can rewrite the integral in Eq. (2.1) as e.g. as a vector multiplication $\boldsymbol{A_j}(\boldsymbol{x}, \boldsymbol{p}, \boldsymbol{T})\, \boldsymbol{x}$, where the non-linear absorption $\tau(r)$ is included in $\boldsymbol{A_j}(\boldsymbol{x}, \boldsymbol{p}, \boldsymbol{T})$. 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 measurement process includes absorption $\tau(r)$ reducing measurements slighty and making the inverse problem weakly non-linear. Hence, we can approximate the non-linear forward model $\boldsymbol{A}(\boldsymbol{x}, \boldsymbol{p}, \boldsymbol{T})$ with a map $\boldsymbol{M}$ and the linear forward model $\boldsymbol{A}_L$, so that $\boldsymbol{A}(\boldsymbol{x}, \boldsymbol{p}, \boldsymbol{T}) \approx \boldsymbol{M}\boldsymbol{A}_L$. Here, $\boldsymbol{A}_{L,j}$ 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 each entry in the row vector $\boldsymbol{A}_{L,j}$ is either defined by $B(\nu, T)S(\nu, T)\frac{\boldsymbol{p}(T)}{k_{\text{B}}\boldsymbol{T}(r)}\mathrm{d}r$ or $B(\nu, T)S(\nu, T)\frac{\boldsymbol{x}}{k_{\text{B}}}\mathrm{d}r$, as in Eq. (2.1), 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

An affine map is any linear map in between two vector spaces or affine spaces, where an affine space does not need to preserve a zero origin, see Def. 2.3.1. in [4]. In other words an affine map does not need to map to the origin of the associated

vector space, 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 the books [4, 5]

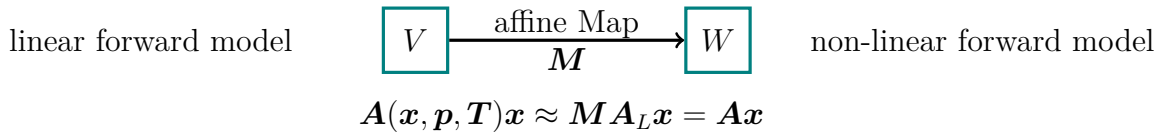

Figure 2.2: Schematics of Affine Map, which approximates the linear forward model to the non-linear forward model.

Consequently, we introduce an affine map $\boldsymbol{M} : \boldsymbol{A}_L\boldsymbol{x} \to \boldsymbol{A}(\boldsymbol{x}, \boldsymbol{p}, \boldsymbol{T})\boldsymbol{x}$, which maps the linear forward model $\boldsymbol{A}_L\boldsymbol{x}$ onto the non-linear forward model $\boldsymbol{A}(\boldsymbol{x}, \boldsymbol{p}, \boldsymbol{T})\boldsymbol{x}$. Then the non linear forward model $\boldsymbol{A}(\boldsymbol{x}, \boldsymbol{p}, \boldsymbol{T}) \approx \boldsymbol{M}\boldsymbol{A}_L$ is approximated by the affine map $\boldsymbol{M}$ and the linear forward model $\boldsymbol{A}_L$. In practise we generate two affine subspaces spaces $V = \left\{\boldsymbol{A}(\boldsymbol{x}^{(1)}, \boldsymbol{p}, \boldsymbol{T}), \ldots, \boldsymbol{A}(\boldsymbol{x}^{(m)}, \boldsymbol{p}, \boldsymbol{T})\right\}$ and $W = \left\{\boldsymbol{A}_L\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{A}_L\boldsymbol{x}^{(m)}\right\}$ over the same field, with fixed $\boldsymbol{p}, \boldsymbol{T}$ and find the mapping in between those. Here, the parameter $\boldsymbol{x}$ is distributed as the posterior distribution $\left\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\right\} \sim \pi(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$ conditioned on the hyper-parameters $\boldsymbol{\theta}$, according to our Bayesian hierarchical model.

2.3 Bayesian Inference

In this section, we introduce the basics of Bayesian inference for a general unknown parameter $\boldsymbol{x}$ given observed data

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\eta}, \tag{2.4}$$

based on a linear forward model $\boldsymbol{A}$ and additive noise $\boldsymbol{\eta}$. A more advanced Bayesian framework, applied/tailored to the previously introduced forward model, will be developed in Section **??**.

We can visualise the correlation structure between parameters as well as how distributions progress in a measurement process, using a hierarchically ordered directed acyclic graph (DAG), see Figure 2.3. Since any observational process

naturally involves random noise, we include this in the DAG and classify the noise variance as a hyper-parameter within $\boldsymbol{\theta}$ [6]. Other hyper-parameters, to which we assign a hyper-prior distribution $\pi(\boldsymbol{\theta})$, may influence the parameters $\boldsymbol{x}$ either statistically (indicated by solid arrows), as in Figure 2.3, or deterministically (indicated by dashed arrows). Here we incorporate prior knowledge of $\boldsymbol{\theta}$ and the parameter $\boldsymbol{x}$ by defining $\pi(\boldsymbol{\theta})$ and the prior distribution $\pi(\boldsymbol{x}|\boldsymbol{\theta})$ according to their receptive physical properties or functional dependences. This is one of the great strength of Bayesian modelling compared to e.g. regularisation, see section 2.4. Then the parameter $\boldsymbol{x}$ is mapped deterministically through the forward model onto the space of all measurable quantities $\boldsymbol{u}$. From this space, we statistically observe the actual data $\boldsymbol{y}$, which includes random (statistical) noise as mentioned above. The distribution of the data conditioned on the hyper-parameters $\boldsymbol{\theta}$ and the parameters $\boldsymbol{x}$ is called the likelihood function $\pi(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{x})$, which includes information about the measurement process through the forward model. Then given some observed data, we like to characterise the posterior distribution $\pi(\boldsymbol{\theta}, \boldsymbol{x}|\boldsymbol{y})$ of the underlying parameters and hyper-parameters by reversing the arrows in Figure 2.3.

The posterior distribution, our the function of interest, is defined by Bayes theorem

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

with the prior distribution $\pi(\boldsymbol{x}, \boldsymbol{\theta}) = \pi(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$ and the normalising constant $\pi(\boldsymbol{y})$. If the normalising constant is finite and non-zero we approximate the posterior distribution

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

and the expectation of any a function $h(\boldsymbol{x}_{\boldsymbol{\theta}})$, where $\boldsymbol{x}$ may depend on $\boldsymbol{\theta}$, is described as

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

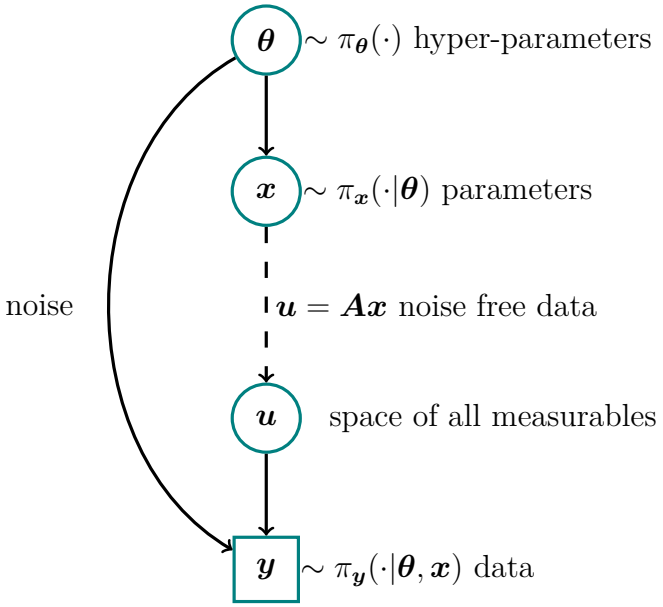

Figure 2.3: The directed acyclic graph (DAG) for a linear inverse problem visualises statistical dependencies as solid line arrows and deterministic dependencies as dotted arrows. The parameters $\boldsymbol{x}$ have some statistical dependency of those hyper-parameters $\boldsymbol{\theta}$, which are distributed as $\pi(\boldsymbol{\theta})$. Then a parameter $\boldsymbol{x} \sim \pi_{\boldsymbol{x}}(\cdot|\boldsymbol{\theta})$ is mapped onto the space of all measurables $\boldsymbol{u} = \boldsymbol{A}\boldsymbol{x}$ deterministically through the linear forward model $\boldsymbol{A}$. From the space of all measurables we observe some data $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\eta}$, statistically, so that $\boldsymbol{y} \sim \pi_{\boldsymbol{y}}(\cdot|\boldsymbol{\theta}, \boldsymbol{x})$, with naturally some random noise $\boldsymbol{\eta} \sim \pi_{\boldsymbol{\eta}}(\cdot|\boldsymbol{\theta})$.

which is may a high dimensional integral and computationally not feasible to solve. Then the sample based Monte Carlo estimate

$$\mathrm{E}_{\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y}}[h(\boldsymbol{x}_{\boldsymbol{\theta}})] \approx \underbrace{\frac{1}{N}\sum_{k=1}^{N} h(\boldsymbol{x}_{\boldsymbol{\theta}}^{(k)})}_{\boldsymbol{\mu}_{\text{samp}}}, \tag{2.8}$$

for large enough N (law of large numbers [7, Chapter 17]) is unbiased [8], where the samples $\{\boldsymbol{x}^{(k)}, \boldsymbol{\theta}^{(k)}\} \sim \pi_{\boldsymbol{x},\boldsymbol{\theta}}(\cdot|\boldsymbol{y})$, for $k = 1, \ldots, N$, form a sample set $\mathcal{M} = \{(\boldsymbol{x}, \boldsymbol{\theta})^{(1)}, \ldots, (\boldsymbol{x}, \boldsymbol{\theta})^{(N)}\}$. Furthermore the central limit theorem states that the samples means $\boldsymbol{\mu}_{samp}^{(i)}$, of independent samples sets $\mathcal{M}_i$ for $i = 1, \ldots, n$ of any distribution, converge in distribution to a normal distribution so that

$$\sqrt{n}(\boldsymbol{\mu}_{\text{samp}}^{(i)} - \boldsymbol{\mu}_{\text{int}}) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2)[9], \tag{2.9}$$

and if $\sigma^2 < \infty$ the Monte Carlo error $\boldsymbol{\mu}_{\text{samp}}^{(i)} - \boldsymbol{\mu}_{\text{int}}$ is bounded.

Generating those sample set from the posterior distribution proves to be the next problem. Since the hyper-parameters and parameters tend to be highly

correlated, see Rue and Held in [10] and Appendix A. One way to work around that is to parameterise $\boldsymbol{x}$ using hyper-parameters $\boldsymbol{\theta}$ so that $\boldsymbol{x}(\boldsymbol{\theta})$ or to factorise the posterior distribution

$$\pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y}) = \pi(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})\pi(\boldsymbol{\theta}|\boldsymbol{y}) \tag{2.10}$$

into the conditional posterior $\pi(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$ over latent field $\boldsymbol{x}$ and the marginal posterior $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ of the hyper-parameters $\boldsymbol{\theta}$. This is particular beneficial, when $\boldsymbol{x}$ is high dimensional, e.g. $\boldsymbol{x} \in \mathbb{R}^n$ with $n = 45$, and $\boldsymbol{\theta}$ is low dimensional, e.g. two dimensional. Then by the law of total expectation [11] Eq. 2.7 becomes

$$\mathrm{E}_{\boldsymbol{x}|\boldsymbol{y}}[h(\boldsymbol{x})] = \mathrm{E}_{\boldsymbol{\theta}|\boldsymbol{y}}\Big[\mathrm{E}_{\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y}}[h(\boldsymbol{x}_{\boldsymbol{\theta}})]\Big] = \int \mathrm{E}_{\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y}}[h(\boldsymbol{x}_{\boldsymbol{\theta}})]\, \pi(\boldsymbol{\theta}|\boldsymbol{y})\, \mathrm{d}\boldsymbol{\theta}\,, \tag{2.11}$$

where in the case of a linear-Gaussian Bayesian hierarchical model $\mathrm{E}_{\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y}}[h(\boldsymbol{x}_{\boldsymbol{\theta}})]$ is well defined.

Due to some Gaussian noise $\boldsymbol{\eta} \sim \mathcal{N}(0, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$ we can set up a linear-Gaussian Bayesian hierarchical model

$$\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{A}\boldsymbol{x}, \boldsymbol{\Sigma}(\boldsymbol{\theta})) \tag{2.12a}$$

$$\boldsymbol{x}|\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{Q}^{-1}(\boldsymbol{\theta})) \tag{2.12b}$$

$$\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})\,, \tag{2.12c}$$

with normally distributed likelihood function $\pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})$ and prior distribution $\pi(\boldsymbol{x}|\boldsymbol{\theta})$, specified by the noise covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ and prior precision matrix $\boldsymbol{Q}(\boldsymbol{\theta})$, modelled via the hyper-prior distribution $\pi(\boldsymbol{\theta})$ [6]. The prior mean is $\boldsymbol{\mu}$. This setup allows us to factorise the posterior distribution efficiently and we call this the marginal and then conditional method.

2.3.1 Marginal and then Conditional

For the in Eq. 2.12 specified linear-Gaussian Bayesian hierarchical model the marginal posterior distribution is given as

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) = \int \pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})\,\mathrm{d}\boldsymbol{x} \tag{2.13}$$

$$\propto \sqrt{\frac{\det(\boldsymbol{\Sigma}^{-1})\,\det(\boldsymbol{Q})}{\det(\boldsymbol{Q}+\boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{A})}} \times \exp\Big[-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu})^T\boldsymbol{Q}_{\boldsymbol{\theta}|\boldsymbol{y}}(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu})\Big]\pi(\boldsymbol{\theta})\,, \tag{2.14}$$

with

$$\boldsymbol{Q}_{\boldsymbol{\theta}|\boldsymbol{y}} = \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{A}+\boldsymbol{Q})^{-1}\boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}\,, \tag{2.15}$$

see Lemma 2 in [6]. Conditioned on the hyper-parameters $\boldsymbol{\theta}$ we can draw samples from the conditional posterior distribution

$$\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y} \sim \mathcal{N}\Big(\underbrace{\boldsymbol{\mu}+(\boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{A}+\boldsymbol{Q})^{-1}\boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{\mu})}_{\boldsymbol{\mu}_{\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y}}}, \underbrace{(\boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{A}+\boldsymbol{Q})^{-1}}_{\boldsymbol{\Sigma}_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}}}\Big) \tag{2.16}$$

using the Randomis then optimize (RTO) method, see section 2.5.2, or calculate weighted expectations e.g. $\mathrm{E}_{\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y}}[\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y}] = \boldsymbol{\mu}_{\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y}}$ or use Eq. 2.11 to calculate weighted expectations of $\mathrm{E}_{\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y}}[\mathrm{Var}(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})] = \boldsymbol{\Sigma}_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}}$ with weights given by $\pi(\boldsymbol{\theta}|\boldsymbol{y})$. Note that the noise covariance $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$ and the prior precision $\boldsymbol{Q} = \boldsymbol{Q}(\boldsymbol{\theta})$ are depending on hyper-parameters $\boldsymbol{\theta}$.

2.4 Regularisation

Another method to find a solution to a linear inverse problem as in Eq. 2.4 is to find a solution $\boldsymbol{x}_\lambda$ according to a data misfit norm and a regularisation semi-norm as in [6]. We will discuss the case of Tikhonov regularisation [12, 13] as this is the most similar to a linear-Gaussian hierarchial Bayesian model.

For a parameter $\boldsymbol{x}$ a linear forward model matrix $\boldsymbol{A}$ and some data $\boldsymbol{y}$ the data misfit norm

$$\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\| \tag{2.17}$$

gives a measure of how good data fits to a mapped parameter $\boldsymbol{Ax}$ and the regularisation semi norm

$$\lambda \|\boldsymbol{Tx}\| \tag{2.18}$$

penalises $\boldsymbol{x}$ according to $\boldsymbol{T}$ and the regularisation parameter $\lambda > 0$. Given λ a regularised solution

$$\boldsymbol{x}_\lambda = \arg\min_{\boldsymbol{x}} \|\boldsymbol{y} - \boldsymbol{Ax}\|^2 + \lambda \|\boldsymbol{Tx}\|^2 \tag{2.19}$$

can be found by the derivativ

$$\nabla_{\boldsymbol{x}}\Big\{(\boldsymbol{y} - \boldsymbol{Ax}_\lambda)^T(\boldsymbol{y} - \boldsymbol{Ax}_\lambda) + \lambda \boldsymbol{x}_\lambda^T \boldsymbol{T}^T \boldsymbol{T} \boldsymbol{x}_\lambda\Big\} = 0 \tag{2.20}$$

$$\iff \quad \nabla_{\boldsymbol{x}}\Big\{\boldsymbol{y}^T\boldsymbol{y} + \boldsymbol{x}_\lambda^T \boldsymbol{A}^T \boldsymbol{A} \boldsymbol{x}_\lambda - \boldsymbol{y}^T \boldsymbol{A} \boldsymbol{x}_\lambda - \boldsymbol{x}_\lambda^T \boldsymbol{A}^T \boldsymbol{y} + \lambda \boldsymbol{x}_\lambda^T \boldsymbol{T}^T \boldsymbol{T} \boldsymbol{x}_\lambda\Big\} = 0 \tag{2.21}$$

$$\iff \quad 2\boldsymbol{A}^T \boldsymbol{A} \boldsymbol{x}_\lambda - 2\boldsymbol{A}^T \boldsymbol{y} + 2\lambda \boldsymbol{T}^T \boldsymbol{T} \boldsymbol{x}_\lambda = 0. \tag{2.22}$$

Then a regularised solution is given as:

$$(\boldsymbol{A}^T \boldsymbol{A} + \lambda \boldsymbol{L})^{-1} \boldsymbol{A}^T \boldsymbol{y} = \boldsymbol{x}_\lambda, \tag{2.23}$$

where we can set $\boldsymbol{T}^T \boldsymbol{T} = \boldsymbol{L}$, which is typically matrix approximation of the nth derivative [13]. In practise $\boldsymbol{x}_\lambda$ is calculated for a range of λ, and is evaluated by the data-misfit norm with respect to the regularised semi-norm so that the best $\boldsymbol{x}_\lambda$ lays on the point of maximum curvature of a so-called L-Curve [14], which we will show in section **??**.

2.5 Sampling Methods

In this section we present the sampling methods used in this thesis and show that the methods we use actually draw samples $\mathcal{M} = \{(\boldsymbol{x}, \boldsymbol{\theta})^{(1)}, \ldots, (\boldsymbol{x}, \boldsymbol{\theta})^{(k)}, \ldots, (\boldsymbol{x}, \boldsymbol{\theta})^{(N)}\} \sim \pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})$ from the desired target distribution so that we can use sample based estimates as in Eq. 2.8. Here $\mathcal{M}$ is a Markov-Chain, where each sample $(\boldsymbol{x}, \boldsymbol{\theta})^{(k)}$ is only depending on the previous sample $(\boldsymbol{x}, \boldsymbol{\theta})^{(k-1)}$ []. Markov-chain Monte Carlo methods generates such a chain $\mathcal{M}$ with random (Monte Carlo) proposals $(\boldsymbol{x}, \boldsymbol{\theta})^{(k)} \sim q(\cdot|(\boldsymbol{x}, \boldsymbol{\theta})^{(k-1)})$ according to a proposal distribution, where ergodicity of

$\mathcal{M}$ is a sufficient criterium to use samples based estimates [8, 13]. The ergodicy theorem in [13] states that, if an aperiodic and irreducible Markov chain $\mathcal{M}$ is reversible then it converges towards a stationary unique equilibirum distribution $\pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})$. In other words if from any state in the chain we can reach any other state in the sampling space and the previous state, and we do not get stuck in periodic loop, then the chain converges towards a stationary distribution. In practise one can look at the trace $\pi(\boldsymbol{x}^{(k)}, \boldsymbol{\theta}^{(k)}|\boldsymbol{y})$ for $k = 1, \ldots, N$ of the samples and eyeball ergodicity.

The sampling methods used in this thesis have proven ergodic properties, so we will cite and refer the reader to the respective documents.

2.5.1 Metropolis- within Gibbs sampling

As introduced in section 2.3.1 when using the MTC method will sample seperatly from $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ and $\pi(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$. To sample from $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ we use a Metropolis-within-Gibbs sampler as in [6] and discuss the 2 dimesnional case, as used in this thesis, with $\boldsymbol{\theta} = (\theta_1, \theta_2)$, where we do a metroplis step in θ_1 deirection and a gibbs stepp in θ_2 direction. Ergodicity is proven here [15].

The Metropolis-within-Gibbs algorithm starts with a initial guess $\boldsymbol{\theta}^{(t)}$ at $t = 0$. Then, we propose a new sample $\theta_1 \sim q(\theta_1|\theta_1^{(t-1)})$ conditioned on the previous state according to a symmetric proposal distribution $q(\theta_1|\theta_1^{(t-1)}) = q(\theta_1^{(t-1)}|\theta_1)$, which is a special case of the Metropolis-Hastings algorithm [15] and cancels when computing the acceptance probability α. We accept and set $\theta_1^{(t)} = \theta_1$ with

$$\alpha(\theta_1|\theta_1^{(t-1)}) = \min\left\{1, \frac{\pi(\theta_1|\theta_2^{(t-1)}, \boldsymbol{y})q(\theta_1^{(t-1)}|\theta_1)}{\pi(\theta_1^{(t-1)}|\theta_2^{(t-1)}, \boldsymbol{y})q(\theta_1|\theta_1^{(t-1)})}\right\} \tag{2.24}$$

or reject and keep $\theta_1^{(t)} = \theta_1^{(t-1)}$, which we do by comparing α to a uniform random number $u \sim \mathcal{U}(0, 1)$. Next, we do a Gibbs step in θ_2 direction, where Gibbs sampling is a special case of the Metropolis-Hastings algorithm with acceptance probaility of 1, and draw the next sample $\theta_2^{(t)} \sim \pi(\cdot|\theta_1^{(t)}, \boldsymbol{y})$ conditioned on $\theta_1^{(t)}$ at step t. We reapet this N times and assure convergence independent of the intial sample (irreducibility) as we discard samples after the so-called burn-in period so that we produce a Markov-Chain of length $N - N_{\text{burn-in}}$.

Algorithm 1: Metropolis within Gibbs

1: Initialize and suppose two dimensional vector $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)})$
2: **for** $k = 1, \ldots, N$ **do**
3: Propose $\theta_1 \sim q(\cdot|\theta_1^{(t-1)}) = q(\theta_1^{(t-1)}|\cdot)$
4: Compute

$$\alpha(\theta_1|\theta_1^{(t-1)}) = \min\left\{1, \frac{\pi(\theta_1|\theta_2^{(t-1)}, \boldsymbol{y})\cancel{q(\theta_1^{(t-1)}|\theta_1)}}{\pi(\theta_1^{(t-1)}|\theta_2^{(t-1)}, \boldsymbol{y})\cancel{q(\theta_1|\theta_1^{(t-1)})}}\right\}$$

5: Draw $u \sim \mathcal{U}(0,1)$
6: **if** $\alpha \geq u$ **then**
7: Accept and set $\theta_1^{(t)} = \theta_1$
8: **else**
9: Reject and keep $\theta_1^{(t)} = \theta_1^{(t-1)}$
10: **end if**
11: Draw $\theta_2^{(t)} \sim \pi(\cdot|\theta_1^{(t)}, \boldsymbol{y})$
12: **end for**
13: Output: $\boldsymbol{\theta}^{(0)}, \ldots, \boldsymbol{\theta}^{(k)}, \ldots, \boldsymbol{\theta}^{(N)} \sim \pi(\boldsymbol{\theta}|\boldsymbol{y})$

2.5.2 Draw a sample from a multivariate normal distribution

after sampling from $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ we draw samples from $\pi(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$ within the MTC scheme. For Linear Gaussian Bayesian hierarchical model we can draw a sample $\boldsymbol{x}$ from the multivariate normal distribution$\pi(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$ using the radomize then optimise (RTO) method [16].

In doing so we can rewrite the full conditional normal distribution $\pi(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})$ to:

$$\pi(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}) \propto \pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta}) \tag{2.25}$$

$$= \exp\|\hat{\boldsymbol{A}}\boldsymbol{x} - \hat{\boldsymbol{y}}\|^2\,, \tag{2.26}$$

where

$$\hat{\boldsymbol{A}} = \begin{bmatrix} \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\theta})\boldsymbol{A} \\ \boldsymbol{Q}^{1/2}(\boldsymbol{\theta}) \end{bmatrix}, \quad \hat{\boldsymbol{y}} = \begin{bmatrix} \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\theta})\boldsymbol{y} \\ \boldsymbol{Q}^{1/2}(\boldsymbol{\theta})\boldsymbol{\mu} \end{bmatrix} \text{ [17].} \tag{2.27}$$

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

$$\boldsymbol{x}_i = \arg\min_{\hat{\boldsymbol{x}}} \|\hat{\boldsymbol{A}}\hat{\boldsymbol{x}} - (\hat{\boldsymbol{y}} + \boldsymbol{b})\|^2, \quad \boldsymbol{b} \sim \mathcal{N}(\boldsymbol{0}, \mathbf{I})\,, \tag{2.28}$$

where we add a randomised perturbation $\boldsymbol{b}$. Similarly as in section **??** we can rewrite the argument of Eq. 2.27 to

$$(\boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})\boldsymbol{A} + \boldsymbol{Q}(\boldsymbol{\theta}))\boldsymbol{x}_i = \boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})\boldsymbol{y} + \boldsymbol{Q}(\boldsymbol{\theta})\boldsymbol{\mu} + \boldsymbol{v}_1 + \boldsymbol{v}_2\,, \tag{2.29}$$

where we substitute $-\hat{\boldsymbol{A}}^T\boldsymbol{b} = \boldsymbol{v}_1 + \boldsymbol{v}_2$ so that $\boldsymbol{v}_1 \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})\boldsymbol{A})$ and $\boldsymbol{v}_2 \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}(\boldsymbol{\theta}))$ are independent random variables [6, 16].

If within the MTC scheme the Markov chain from the marginal posterior is ergodic then with independent samples $\boldsymbol{x}^{(k)}$ from the full conditional $\pi(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$ the combined chain $\{(\boldsymbol{x}, \boldsymbol{\theta})^{(1)}, \ldots, (\boldsymbol{x}, \boldsymbol{\theta})^{(k)}, \ldots, (\boldsymbol{x}, \boldsymbol{\theta})^{(N)}\} \sim \pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})$ is ergodic as well [18].

2.5.3 t-walk

If there is a functional dependency of the parameters $\boldsymbol{x}$ and the hyper-parameters $\boldsymbol{\theta}$ so that $\boldsymbol{x}(\boldsymbol{\theta})$ we can use the t-walk algorithm by Christens and Fox on $\pi(\boldsymbol{\theta}|\boldsymbol{y})$. We use the t-walk as a black box sampler, where convergence is guaranteed by construction [19].

2.6 Numerical Approxiamtion Methods - Tensor Train

First we will derive how we find marginal probality distribution funcitons and then give a brief introduction into tensor train format. Asumme that the triple $(\Omega, \mathcal{F}, \mathbb{P})$ is called a probability space over the whole sample (or parameter) space Ω with the collection $\mathcal{F}$ of very countable subset $\{A_n\}_{n\in\mathbb{N}}$ in Ω, so that $A_n \subseteq \Omega$, and $\mathbb{P}$ is a measure on $\mathcal{F}$. We assume $\mathbb{P}$ fullfils the requirements of a probality measure and $\mathcal{F}$ is a σ-algebra see Appendix B. We call $\mathbb{P}(A)$ the probaility of an event $A \subseteq \mathcal{F}$

$$\mathbb{P}(A) = \int_A \mathrm{d}\mathbb{P}[]. \tag{2.30}$$

We change variables using the Radon-Nikodym theroem [20]

$$\mathbb{P}(A) = \int_A \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}x}\,\mathrm{d}x = \int_A \pi(x)\,\mathrm{d}x \tag{2.31}$$

where $\mathrm{d}x$ is a measure on the same probality space, also known as the lebesgue measure and $\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}x}$ is often called Radon-Nikodym derivativ of $\mathbb{P}$ with repect to x. We can say that $\mathbb{P}$ has a density $\pi(x)$, with repect to x [21, Chapter 10]. Next, we can define $x : \Omega \longrightarrow X$ as a random variable connecting to the measurable space $(X, \mathcal{X})$ through the probability density function $\pi(x)$ [20]. For a d-dimensional x we call the cartesioan product $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_d$ the parameter space in X. If $X = \mathbb{R}^d$ every subset $\mathcal{X}_k \subseteq \mathbb{R}$, for $k = 1, \ldots, d$, and $x = (x_1, \ldots, x_d)$, with $x_k \in \mathcal{X}_k$, $\pi(x)$ is the joint probaility density. The marginal function

$$f_{X_k}(x_k) = \int_{\mathcal{X}_1} \cdots \int_{\mathcal{X}_d} \boldsymbol{\lambda}(x)\pi(x)\,\mathrm{d}x_1 \cdots \mathrm{d}x_{k-1}\,\mathrm{d}x_{k+1} \cdots \mathrm{d}x_d\,, \tag{2.32}$$

of an event x_k in $X_k \subseteq X$ is calculated by integrating the probility desnity funciton $\pi(x)$ over the parameter space in all other dimensions. Here we intrdouce a weight function $\boldsymbol{\lambda}(x)$ [22], which can be helpful for qraudrature rules. Cui et al. [23] call $\boldsymbol{\lambda}(x)$ the ' product-from Lebsegue measurable weighting functions" and define it as $\boldsymbol{\lambda}(\mathcal{X}) = \prod_{i=1}^{d} \boldsymbol{\lambda}_i(\mathcal{X}_i)$ and $\boldsymbol{\lambda}_i(\mathcal{X}_i) = \int_{\mathcal{X}_i} \boldsymbol{\lambda}_i(x_i)\mathrm{d}x_i$.

Using the tensor train (TT) format we can approximate d-dimensional function $\pi(x)$ and compute marginal probailty functions distributions cheaply. In doing so we have to define a d-dimesnional discrete univariate grid in the parameter space $\mathcal{X}$ with n grid points in each direction. Then, as the name suggest the tensor train format is a train of tensors which represent this d-dimesnional grid. More specifically each tensor is a two and three dimensional matrix, which we call core, $\pi_k \in \mathbb{R}^{r_{k-1} \times n \times r_k}$ and connected with the neighbouring tensor through ranks r_k and r_{k-1} for each $k = 1, \ldots, d$, as in Figure 2.4 displayed. For the first and last dimensional the core outer ranks are $r_0 = r_d = 1$, so that for $x = x_1, \ldots, x_d$ the function value $\pi(x) = \Big(\pi_1(x_1) \ldots \pi_d(x_d)\Big) \in \mathbb{R}$ is a vector-matrix-vector multiplication and each core $\pi_k(x_k)$ at a fixed x_k on the approximated grid has dimensions $r_{k-1} \times 1 \times r_k = r_{k-1} \times r_k$.

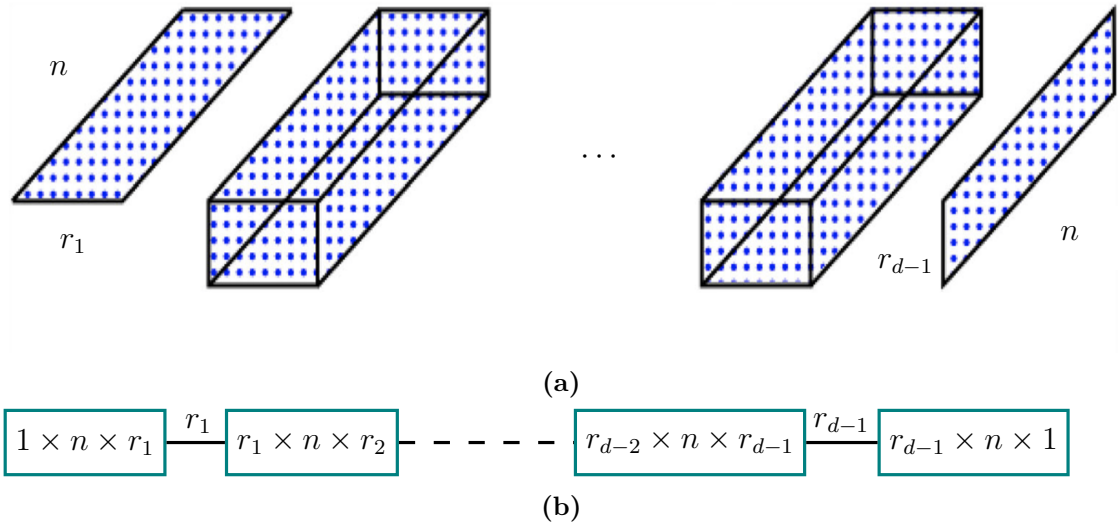

Figure 2.4: Here we visualize the tensor train cores as two and three dimesnional matrices. Each matrix has a length n according to the gridsize and ther cores are connected through ranks r. More specifiaclly a core π_k has dimesnions $r_{k-1} \times n \times r_k$, where $r_0 = r_d = 1$. Figure (a) is taken from [24].

Consequently, using a tensor train approxiamtion, the marginal target function

$$f_{X_k}(x_k) = \frac{1}{z} \Big| \left(\int_{\mathbb{R}} \boldsymbol{\lambda}_1(x_1)\pi_1(x_1)\mathrm{d}x_1 \right) \cdots \left(\int_{\mathbb{R}} \boldsymbol{\lambda}_{k-1}(x_{k-1})\pi_{k-1}(x_{k-1})\mathrm{d}x_{k-1} \right) \boldsymbol{\lambda}_k(x_k)\pi_k(x_k) \left(\int_{\mathbb{R}} \boldsymbol{\lambda}_{k+1}(x_{k+1})\pi_{k+1}(x_{k+1})\mathrm{d}x_{k+1} \right) \cdots \left(\int_{\mathbb{R}} \boldsymbol{\lambda}_d(x_d)\pi_d(x_d)\mathrm{d}x_d \right) \Big| \tag{2.33}$$

is given by integration over each core [25] including some normalising constant z [23].

From here we follow the notation and procedure mostly from Cui et al. [23] and approximate the square root

$$\sqrt{\pi(x)} \approx g(x) = \boldsymbol{G}_1(x_1), \ldots, \boldsymbol{G}_k(x_k), \ldots, \boldsymbol{G}_d(x_d) \tag{2.34}$$

for numerical stability, where each TT-core

$$G_k^{(\alpha_{k-1},\alpha_k)}(x_k) = \sum_{i=1}^{n_k} \phi_k^{(i)}(x_k)\boldsymbol{A}_k[\alpha_{k-1}, i, \alpha_k], \quad k = 1, ..., d, \tag{2.35}$$

is associated kth coefficient tensor $\boldsymbol{A}_k \in \mathbb{R}^{r_{k-1} \times n_k \times r_k}$ and the k-th basis functions $\phi_k^{(i)}(x_k)$. We assume the function

$$\pi(x) \approx \gamma' + g^2(x)\,, \tag{2.36}$$

is approximated through the TT decomposition $g(x)$, where the error γ' assures positivity and is chosen according to the L_2 norm

$$\gamma' \leq \frac{1}{\boldsymbol{\lambda}(\mathcal{X})} \|g - \sqrt{\pi}\|_2^2 \,. \tag{2.37}$$

Then the normalised target function is

$$f_X(x) = \frac{1}{z}\pi(x)\boldsymbol{\lambda}(x) = \frac{1}{z}(\gamma'\boldsymbol{\lambda}(x) + g^2(x)\boldsymbol{\lambda}(x)) \,. \tag{2.38}$$

Given the tensor train approximation of the sqared rooted function $\sqrt{\pi}$ can be expressed as

$$\begin{aligned} f_{X_k}(x_k) = \frac{1}{z}\Bigg(\gamma' \prod_{i=1}^{k-1} \boldsymbol{\lambda}_i(\mathcal{X}_i) \prod_{i=k+1}^{d} \boldsymbol{\lambda}_i(\mathcal{X}_i) \\ + \left(\int_{\mathbb{R}} \boldsymbol{G}_1^2(x_1)\boldsymbol{\lambda}_1(x_1)\mathrm{d}x_1\right) \cdots \left(\int_{\mathbb{R}} \boldsymbol{G}_{k-1}^2(x_{k-1})\boldsymbol{\lambda}_{k-1}(x_{k-1})\mathrm{d}x_{k-1}\right) \\ \boldsymbol{G}_k^2(x_k)\boldsymbol{\lambda}_k(x_k) \\ \left(\int_{\mathbb{R}} \boldsymbol{G}_{k+1}^2(x_{k+1})\boldsymbol{\lambda}_{k+1}(x_{k+1})\mathrm{d}x_{k+1}\right) \cdots \left(\int_{\mathbb{R}} \boldsymbol{G}_d^2(x_d)\boldsymbol{\lambda}_d(x_d)\mathrm{d}x_d\right)\Bigg) \,. \end{aligned} \tag{2.39}$$

To efficiently calculate these marginals on can use a procedure similar to something that is called left and right orthogonalisation of cores [26]. To do so we define the mass matrix $\boldsymbol{M}_k \in \mathbb{R}^{n_k \times n_k}$ by

$$\boldsymbol{M}_k[i,j] = \int_{X_k} \phi_k^{(i)}(x_k)\phi_k^{(j)}(x_k)\boldsymbol{\lambda}(x_k)\, dx_k, \quad i = 1, ..., n_k, \quad j = 1, ..., n_k, \tag{2.40}$$

where $\{\phi_k^{(i)}(x_k)\}_{i=1}^{n_k}$ is the set of basis functions for the k-th coordinate.

2.6.1 Marginal Functions

We calculate the marginal functions through procedures, which we call backward marginalisation [23] and forward marginalisation. We gain the coefficient matrices $\boldsymbol{B}_k$ through backward marginalisation and the coefficient matrices $\boldsymbol{B}_{pre,n}$ through forward marginalisation, which enables us to calculate marginal function similar to [23]. The proposition 1 to caculte $\boldsymbol{B}_k$ is taken from [23].

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 $\boldsymbol{C}_k \in \mathbb{R}^{r_{k-1} \times n_k \times r_k}$:

$$\boldsymbol{C}_k[\alpha_{k-1}, \tau, l_k] = \sum_{i=1}^{n_k} \boldsymbol{B}_k[\alpha_{k-1}, i, l_k] \boldsymbol{L}_k[i, \tau]. \tag{2.41}$$

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

$$\boldsymbol{Q}_k \boldsymbol{R}_k = (\boldsymbol{C}_k^{(R)})^\top. \tag{2.42}$$

3. Compute the new coefficient tensor:

$$\boldsymbol{B}_{k-1}[\alpha_{k-2}, i, l_{k-1}] = \sum_{\alpha_{k-1}=1}^{r_{k-1}} \boldsymbol{A}_{k-1}[\alpha_{k-2}, i, \alpha_{k-1}] \boldsymbol{R}_k[l_{k-1}, \alpha_{k-1}]. \tag{2.43}$$

We start the forward marginalisation with the first dimension as in Proposition 2.

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 $\boldsymbol{C}_k \in \mathbb{R}^{r_{k-1} \times n_k \times r_k}$:

$$\boldsymbol{C}_{pre,k}[\alpha_{k-1}, \tau, l_k] = \sum_{i=1}^{n_k} \boldsymbol{L}_k[i, \tau] \boldsymbol{B}_{pre,k}[\alpha_{k-1}, i, l_k]. \tag{2.44}$$

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

$$\boldsymbol{Q}_{pre,k} \boldsymbol{R}_{pre,k} = (\boldsymbol{C}_{pre,k}^{(R)}). \tag{2.45}$$

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

$$\boldsymbol{B}_{pre,k+1}[l_{k+1}, i, \alpha_{k+1}] = \sum_{\alpha_k=1}^{r_k} \boldsymbol{R}_{pre,k}[l_{k+1}, \alpha_k] \boldsymbol{A}_{k+1}[\alpha_k, i, \alpha_{k+1}]. \tag{2.46}$$

The marginal PDF of X_k can be expressed as

$$f_{X_k}(x_k) = \frac{1}{z}\left(\gamma' \prod_{i=1}^{k-1} \boldsymbol{\lambda}_i(X_i) \prod_{i=k+1}^{d} \boldsymbol{\lambda}_i(X_i) + \sum_{l_{k-1}=1}^{r_{k-1}} \sum_{l_k=1}^{r_k} \left(\sum_{i=1}^{n} \phi_k^{(i)}(x_k) \boldsymbol{D}_k[l_{k-1}, i, l_k]\right)^2\right) \boldsymbol{\lambda}_k(x_k), \tag{2.47}$$

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

$$\boldsymbol{D}_k[l_{k-1}, i, l_k] = \sum_{\alpha_{k-1}=1}^{r_{k-1}} \boldsymbol{R}_{pre,k-1}[l_{k-1}, \alpha_{k-1}] \boldsymbol{B}_k[\alpha_{k-1}, i, l_k]. \tag{2.48}$$

In the special case for the first dimension, $f_{X_1}(x_1)$ can be expressed as

$$f_{X_1}(x_1) = \frac{1}{z}\left(\gamma' \prod_{i=2}^{d} \boldsymbol{\lambda}_i(\mathcal{X}_i) + \sum_{l_1=1}^{r_1} \left(\sum_{i=1}^{n} \phi_1^{(i)}(x_1) \boldsymbol{D}_1[i, l_1]\right)^2\right) \boldsymbol{\lambda}_1(x_1), \tag{2.49}$$

where $\boldsymbol{D}_1[i, l_1] = \boldsymbol{B}_1[\alpha_0, i, l_1]$ and $\alpha_0 = 1$, and similiarly in the last dimesnion

$$f_{X_d}(x_d) = \frac{1}{z}\left(\gamma' \prod_{i=1}^{d-1} \boldsymbol{\lambda}_i(\mathcal{X}_i) + \sum_{l_{n-1}=1}^{r_{d-1}} \left(\sum_{i=1}^{n} \phi_1^{(i)}(x_1) \boldsymbol{D}_d[l_{n-1}, i]\right)^2\right) \boldsymbol{\lambda}_d(x_d), \tag{2.50}$$

where $\boldsymbol{D}_d[l_{n-1}, i] = \boldsymbol{B}_{pre,d}[l_{n-1}, i, \alpha_{n+1}]$ and $\alpha_{d+1} = 1$.

Appendices

A
Correlation Structure

In the book Gaussian Markov Random Fields, Rue and Held show the correlation structure between the hyper-parameter μ and the latent field $\boldsymbol{x}$, which slows down convergence especially when using Gibbs samplers. They consider the hierarchichal sturcture

$$\mu \sim \mathcal{N}(0,1) \tag{A.1}$$

$$\boldsymbol{x}|\mu \sim \mathcal{N}(\mu\mathbf{1}, \boldsymbol{Q}^{-1})\,, \tag{A.2}$$

and use a Gibbs sampler over the full conditionals

$$\mu^{(k)}|\boldsymbol{x}^{(k)} \sim \mathcal{N}\left(\frac{\mathbf{1}^T\boldsymbol{Q}\boldsymbol{x}^{(k-1)}}{1+\mathbf{1}^T\boldsymbol{Q}\mathbf{1}}, (1+\mathbf{1}^T\boldsymbol{Q}\mathbf{1})^{-1}\right) \tag{A.3}$$

$$\boldsymbol{x}^{(k)}|\mu^{(k)} \sim \mathcal{N}(\mu^{(k)}\mathbf{1}, \boldsymbol{Q}^{-1})\,. \tag{A.4}$$

In Figure A.1 one can clearly see that when steps only in the μ-direction, x-axis, or in the $\boldsymbol{x}$-direction, vertical-(y)-axis, are allowed it takes a lot of steps to explore the parameter space due to high correlation in between μ and $\boldsymbol{x}$. A solution is to update (μ, x) jointly, where, since μ is one dimensional, effectively only marginal density of μ, by integrating out $\boldsymbol{x}$ of the joint density $\pi(\mu, x)$, is needed.

$$\mu^\star \sim q(\mu^\star|\mu^{(k-1)}) \tag{A.5}$$

$$\boldsymbol{x}^{(k)}|\mu^\star \sim \mathcal{N}(\mu^\star\mathbf{1}, \boldsymbol{Q}^{-1}) \tag{A.6}$$

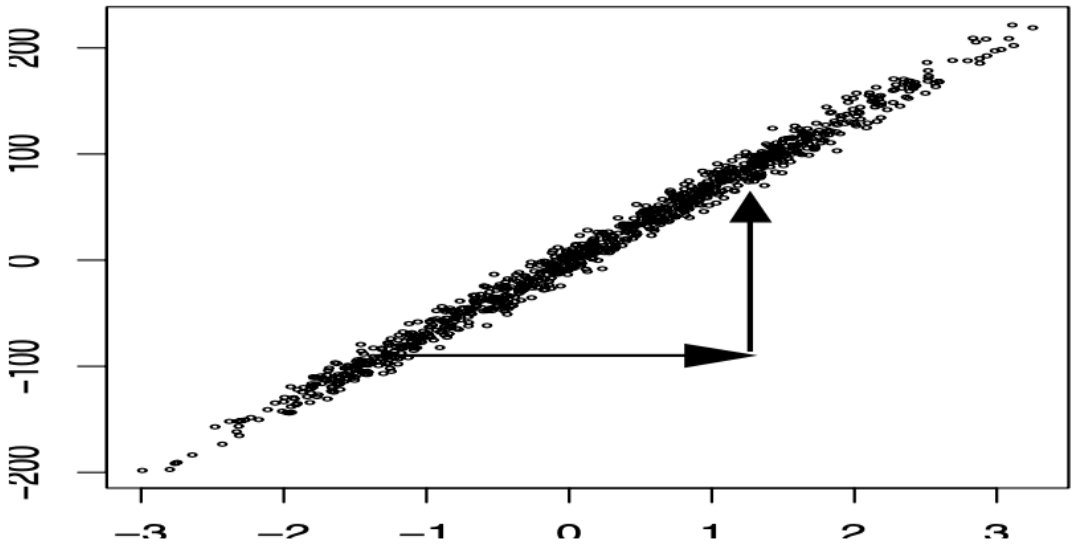

Figure A.1: Figure 4.1 Figure (a) shows the marginal chain for µ over 1000 iterations of the marginal chain for the hyperparmeter with a specuifuc autoregressive proces defined in $\boldsymbol{Q}$ The algorithm updates successively μ and x from their full conditionals. Figure (b) displays the pairs ($\mu(k)$, $1^T\boldsymbol{Q}x(k)$), with$\mu(k)$ on the horizontal axis. The slow mixing (and convergence) of $\mu$is due to the strong dependence with $1^T\boldsymbol{Q}x(k)$ as only horizontal and vertical moves are allowed. The arrows illustrate how a joint update can improve the mixing (and convergence)

With a simple MCMC algorithm on μ one can explore the sample space of μ and only draw a sample $\boldsymbol{x}$ from its full conditional after e.g. $\mu^\star$ is accepted.

B
Mesure theroy

Recall that the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where we call Ω the samples spae with a collection $\mathcal{F}$, which is a σ-algebra, of very countable subset $\{A_n\}_{n\in\mathbb{N}}$. We call A_n an Event in Ω, $A_n \subseteq \Omega$, and a map $\mathbb{P} : \mathcal{F} \longrightarrow \mathbb{R}$ a probavlity measure. Now we like to define a σ algebra and a prpbailty measure.

B.1 sigma alrgbea

A collections of subsets $\mathcal{F}$ is called sigma algebra if

- $\emptyset, \Omega \in \mathcal{F}$
- if $A \in \mathcal{F}$ then $A^C \coloneqq A/\Omega \in A$
- if $A_1, A_2, \cdots \in \mathcal{F}$ then $\bigcup_{j\in\mathbb{N}} A_j \in \mathcal{F}$

In other words the empty set $\emptyset$ and the whole sample space Ω should always lay in $\mathcal{F}$ If we take any subset A in $\mathcal{F}$ the complement A^C, which is the sample space without A, A/Ω, has to lay in $\mathcal{F}$ as well. So if we able to calculate the probability $\mathbb{P}(A)$ we have to calculate the probability of not A,$\mathbb{P}(A^C)$. Finally, if the collection of countable subsets $A_1, A_2, \ldots$ lays in $\mathcal{F}$ then the union $\bigcup_{j\in\mathbb{N}} A_j$ also has to lay in $\mathcal{F}$.

B.2 probailty measure

For a probability measure we require

- $\mathbb{P}(\Omega) = 1$ and $\mathbb{P}(\emptyset) = 0$
- $\mathbb{P}(A) \in [0, 1]$
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$ if A, B are disjoint or $A \cap B = \emptyset$
- $\mathbb{P}(\bigcup_{j\in\mathbb{N}} A_j) = \sum_{j\in\mathbb{N}} \mathbb{P}(A_j)$ if we have pairwise disjoint sets or $A_i \cap A_j = \emptyset$ for $i \neq j$

In other words the probability over the whole sample space should be equal to one and the probability over the empty set is zero. So for every subset A of the sample space Ω the probability $\mathbb{P}(A)$ lays in between zero and one. If we have two subsets A and B with no overlap then the probability of the union of those two subset , $\mathbb{P}(A \cup B)$,is equal to the sum of the probability of each of those subsets, $\mathbb{P}(A) + \mathbb{P}(B)$. This has to hold for the more general case of all countable unions of subsets $\bigcup_{j\in\mathbb{N}} A_j$.

See [27] [20]

References

[1] C. Readings. *Envisat MIPAS An Instrument for Atmospheric Chemistry and Climate Research*. Noordwijk: ESA Publications Division, 2000.

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

[3] George B. Rybicki and Alan P. Lightman. *Radiative processes in Astrophysics*. Weinheim: Wiley-VCH, 2004.

[4] Marcel Berger. *Geometry I. 4th Edition*. Berlin Heidelberg: Springer-Verlag, 2009.

[5] Katsumi Nomizu and Takeshi Sasaki. *Affine differential geometry*. Cambridge: Cambridge University Press, 1994.

[6] Colin Fox and Richard A Norton. "Fast sampling in a linear-Gaussian inverse problem". In: *SIAM/ASA Journal on Uncertainty Quantification* 4.1 (2016), pp. 1191–1218.

[7] S. P. Meyn and R.L. Tweedie. *Markov Chains and Stochastic Stability. 2nd Edition*. New York: Cambridge University Press, 2009.

[8] Gareth O. Roberts and Jeffrey S Rosenthal. "General state space Markov chains and MCMC algorithms". In: *Probability Surveys* 1 (2004), pp. 20–71.

[9] Charles J Geyer. "Practical markov chain monte carlo". In: *Statistical science* (1992), pp. 473–483.

[10] Havard Rue and Leonhard Held. *Gaussian Markov random fields: theory and applications*. London: CRC press, 2005.

[11] Charles W. Champ and Andrew V. Sills. "The Generalized Law of Total Covariance". In: *preprint* (2022). URL: `https://arxiv.org/abs/2205.14525`.

[12] Jari P. Kaipio and Erkki Somersalo. *Statistical and Computational Inverse Problems*. New York: Springer-Verlag New York, 2005.

[13] Sze M Tan, Colin Fox, and Geoff K. Nicholls. *Course notes for ELEC 445 – Inverse Problems and Imaging*. `https://coursesupport.physics.otago.ac.nz/wiki/pmwiki.php/ELEC445/HomePage`. [Online; accessed 10/12/23]. 2016.

[14] Per Christian Hansen and Dianne Prost O'Leary. "The use of the L-curve in the regularization of discrete ill-posed problems". In: *SIAM Journal on Scientific Computing* 14.6 (1993), pp. 1487–1503.

[15] Gareth O. Roberts and Jeffrey S Rosenthal. "Harris recurrence of Metropolis-within-Gibbs and trans-dimensional Markov chains". In: *The Annals of Applied Probability* 16.4 (2006), 2123–2139.

[16] Johnathan M Bardsley. "MCMC-based image reconstruction with uncertainty quantification". In: *SIAM Journal on Scientific Computing* 34.3 (2012), A1316–A1332.

[17] Johnathan M Bardsley et al. "Randomize-then-optimize: A method for sampling from posterior distributions in nonlinear inverse problems". In: *SIAM Journal on Scientific Computing* 36.4 (2014), A1895–A1910.

[18] Felipe Acosta, Mark L Huber, and Galin L Jones. "Markov chain Monte Carlo with linchpin variables". In: *preprint* (2014). URL: https://arxiv.org/abs/2205.14525.

[19] J. Andrés Christen and Colin Fox. "A general purpose sampling algorithm for continuous distributions (the t-walk)". In: *Bayesian Analysis* 5.2 (2010), pp. 263 –281. URL: https://doi.org/10.1214/10-BA603.

[20] M. Capiński and P.E. Kopp. *Measure, Integral and Probability. Springer Undergraduate Mathematics Series*. London: Springer-Verlag London, 2004.

[21] M. Simonnet. *Measures and Probabilities*. New York: Springer-Verlag, 1996.

[22] Philip J Davis and Philip Rabinowitz. *Methods of numerical integration*. San Diego, CA: Academic Press, Inc., 1984.

[23] Tiangang Cui and Sergey Dolgov. "Deep composition of tensor-trains using squared inverse rosenblatt transports". In: *Foundations of Computational Mathematics* 22.6 (2022), pp. 1863–1922.

[24] Colin Fox et al. "Grid methods for Bayes-optimal continuous-discrete filtering and utilizing a functional tensor train representation". In: *Inverse Problems in Science and Engineering* 29.8 (2021), pp. 1199–1217.

[25] Sergey Dolgov et al. "Approximation and sampling of multivariate probability distributions in the tensor train decomposition". In: *Statistics and Computing* 30 (2020), pp. 603–625.

[26] Ivan V Oseledets. "Tensor-train decomposition". In: *SIAM Journal on Scientific Computing* 33.5 (2011), pp. 2295–2317.

[27] Greg Lawler. *Notes on probability*. https://www.math.uchicago.edu/~lawler/probnotes.pdf. [Online; accessed 10/04/25]. 2016.