Contents

List of Figures

421.10046pt

1 Introduction

1.1 Motivation

- ozone coverage
- regularisation approach in atmospheric physics citation
- hierarchical modelling

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

- physical based hierarchical Bayesian model, sampling to TT approx
- RTE as an example
- non-linear to linear affine approximation

1.3 Thesis Outline

Note the following: In this case the best fit to data is not the best fit to parameters. In under- graduate statistics courses you would learn the more general notion that: "Conditioning on estimates gives poor predictive densities". [1]

421.10046pt

2 Theoretical and Technical Background

In this chapter, we provide a brief introductions and derivations to the methods used in this thesis as well as references for more details, non interested readers may skip this chapter and move on to the next one. We keep it as general as possible, as the expressions tailored towards specifically the forward map will be presented in the results Chapter 4 but without derivations. We begin by introducing a general hierarchical Bayesian approach to a linear inverse problem. Then we provide some background information on affine maps and the Tikhonov regularisation method. Next, we provide a small introduction into Sampling methods, more specifically the essentials of Markov-Chain monte Carlo methods. Lastly, we explain how we approximate functions using a Tensor-Train (TT) approach, which enables us to calculate marginal from the posterior distribution cheaply.

2.1 Bayesian Inference

Assume we observe some data

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\eta}, \tag{2.1}$$

based on a linear forward model $\boldsymbol{A}$, a unknown parameter $\boldsymbol{x}$ and some additive random noise $\boldsymbol{\eta}$. Naturally due to the noise we have some uncertainty which we include in the modelling process through the likelihood function $\pi(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{x})$ as well as other relevant information about the measurement process. We read $\pi(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{x})$ as the distribution over $\boldsymbol{y}$ conditioned on $\boldsymbol{x}$ and the hyper-parameter $\boldsymbol{\theta}$. Here $\boldsymbol{\theta}$ may account for multiple variables and is e.g. describing the distribution of the noise vector $\boldsymbol{\eta} \sim \pi_{\boldsymbol{\eta}}(\cdot|\boldsymbol{\theta})$ as well as the prior distribution $\pi(\boldsymbol{x}|\boldsymbol{\theta})$, which accounts for physical properties or functional dependences of $\boldsymbol{x}$. Consequently we define a hyper-prior distribution $\pi(\boldsymbol{\theta})$, where

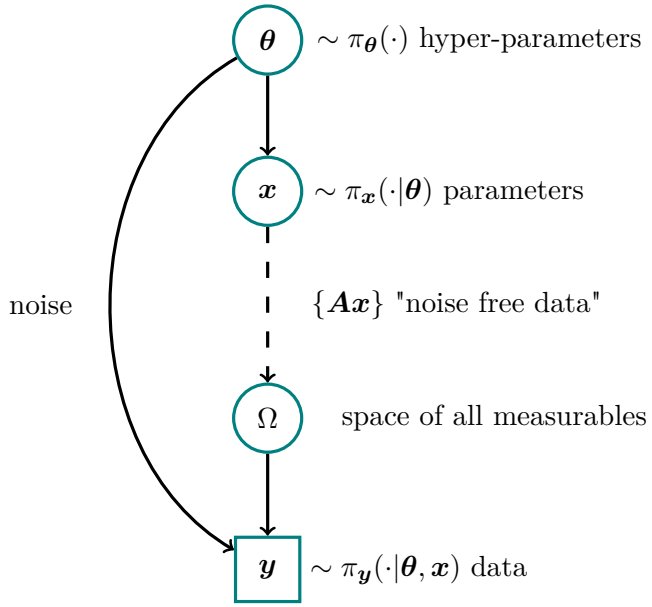

Figure 2.1: The directed acyclic graph (DAG) for a linear inverse problem visualises statistical dependencies as solid line arrows and deterministic dependencies as dotted arrows. The hyper-parameters $\boldsymbol{\theta}$ are distributed as the hyper-prior distribution $\pi(\boldsymbol{\theta})$. The prior distribution $\pi_{\boldsymbol{x}}(\cdot|\boldsymbol{\theta})$ for the parameter $\boldsymbol{x}$ and the noise are statsitically dependent on have some on those hyper-parameters. Then a parameter $\boldsymbol{x} \sim \pi_{\boldsymbol{x}}(\cdot|\boldsymbol{\theta})$ is mapped onto the space of all measurables $\boldsymbol{u} = \boldsymbol{Ax}$ deterministically through the linear forward model $\boldsymbol{A}$. From the space of all measurable noise free data we observe a data set $\boldsymbol{y} = \boldsymbol{Ax} + \boldsymbol{\eta}$ with some random noise $\boldsymbol{\eta} \sim \pi_{\boldsymbol{\eta}}(\cdot|\boldsymbol{\theta})$, which determines the likelihood function $\pi(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{x})$.

$\pi(\boldsymbol{x}, \boldsymbol{\theta}) = \pi(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$. Choosing these prior distribution is a delicate topic as it shall not affect the posterior distribution

$$\pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y}) = \frac{\pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})\pi(\boldsymbol{x}, \boldsymbol{\theta})}{\pi(\boldsymbol{y})} \propto \pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})\pi(\boldsymbol{x}, \boldsymbol{\theta})\,, \tag{2.2}$$

which according to Bayes theorem gives us a distribution of $\boldsymbol{x}$ and $\boldsymbol{\theta}$ given (conditioned) on some data. We can visualise theses hierarchically ordered correlation structure between parameters as well as how distributions progress through a measurement process, using a directed acyclic graph (DAG), see Figure 2.1.

Obviously we are interested in the posterior distribution as the expectation of any a function $h(\boldsymbol{x}_{\boldsymbol{\theta}})$, where $\boldsymbol{x}$ may depend on $\boldsymbol{\theta}$, is described as

$$\mathrm{E}_{\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y}}[h(\boldsymbol{x}_{\boldsymbol{\theta}})] = \underbrace{\int\int h(\boldsymbol{x}_{\boldsymbol{\theta}})\,\pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{\theta}}_{\boldsymbol{\mu}_{\text{int}}}\,, \tag{2.3}$$

which may be a high dimensional integral and computationally not feasible to solve. Therefore the unbiased sample based Monte Carlo estimate [2]

$$\mathrm{E}_{\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y}}[h(\boldsymbol{x}_{\boldsymbol{\theta}})] \approx \underbrace{\frac{1}{N}\sum_{k=1}^{N} h(\boldsymbol{x}_{\boldsymbol{\theta}}^{(k)})}_{\boldsymbol{\mu}_{\text{samp}}}\,, \tag{2.4}$$

for large enough N (law of large numbers [3, Chapter 17]) is often used. Here, 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 $\mathcal{M} = \{(\boldsymbol{x}, \boldsymbol{\theta})^{(1)}, \ldots, (\boldsymbol{x}, \boldsymbol{\theta})^{(N)}\}$. Generating a representative sample sets quickly from the posterior distribution, often presents a significant challenge. This is mainly due to the strong correlations that usually exist between the parameters and hyper-parameters, as discussed by Rue and Held in [4] and illustrated in Appendix B. If $\boldsymbol{x}$ can not be parametrised directly in terms of the hyper-parameters $\boldsymbol{\theta}$, i.e., $\boldsymbol{x}(\boldsymbol{\theta})$, it is beneficial to factorise the posterior distribution as

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

into the conditional posterior $\pi(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$ over the latent field $\boldsymbol{x}$ and the marginal posterior

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

over the hyper-parameters $\boldsymbol{\theta}$. This approach, known as the marginal and then conditional (MTC) method, is particularly advantageous when $\boldsymbol{x}$ is high-dimensional (e.g., $\boldsymbol{x} \in \mathbb{R}^n$ with $n \geq 45$), while $\boldsymbol{\theta}$ is low-dimensional (e.g., two-dimensional). Applying the law of total expectation [5], Eq. (2.3) becomes

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

where, in the case of a linear-Gaussian Bayesian hierarchical model, both the marginal distribution and the inner expectation $\mathbb{E}_{\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y}}\left[h(\boldsymbol{x}_{\boldsymbol{\theta}})\right]$ are well defined see Result chapter. Furthermore, the central limit theorem states that the samples mean $\boldsymbol{\mu}^{(i)}_{\text{samp}}$, of independent samples sets $\mathcal{M}_i$ for $i = 1, \ldots, n$ of any distribution, converge in distribution to a normal distribution so that

$$\sqrt{n}(\boldsymbol{\mu}^{(i)}_{\text{samp}} - \boldsymbol{\mu}_{\text{int}}) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2)[6], \tag{2.8}$$

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

On the Monte-Carlo Error and Integrated Autocorrelation time

To asses the error σ^2 we ignoring systematic error due to initialisation bias (burin in period) but we have to take into account that samples porduces by any system or algorithm are correlated. In general the error of a Monte-Carlo based estimate from a sample set $\mathcal{M}_i$ is:

$$(\sigma^{(i)})^2 = \text{var}(\boldsymbol{\mu}^{(i)}_{\text{samp}}) = \text{var}(\text{E}_{\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y}}[h(\boldsymbol{x}_{\boldsymbol{\theta}})]) = \left(\frac{1}{N}\sum_{k=1}^{N} h(\boldsymbol{x}^{(k)}_{\boldsymbol{\theta}}) - \boldsymbol{\mu}^{(i)}\right)^2. \tag{2.9}$$

Expanding this summation we see that

$$(\sigma^{(i)})^2 = \frac{1}{N^2}\sum_{k,s=1}^{N} C(k-s) \tag{2.10}$$

with the auto correlation coefficient $C(k-s) = (h(\boldsymbol{x}_{\boldsymbol{\theta}}^{(k)}) - \boldsymbol{\mu}^{(i)})(h(\boldsymbol{x}_{\boldsymbol{\theta}}^{(s)}) - \boldsymbol{\mu}^{(i)})$ and define the sample auto correlation function

$$\frac{C(0)}{N} \sum_{k,s=1}^{N} \frac{C(k-s)}{C(0)} \approx \mathrm{var}(h(\boldsymbol{x}_{\boldsymbol{\theta}})) \sum_{t=-\infty}^{\infty} \rho(t) \tag{2.11}$$

with the normalised auto correlation coefficient $\rho(k-s) = C(k-s)/C(0)$ at lag $k-s$, where $C(0) = \mathrm{var}(h(\boldsymbol{x}_{\boldsymbol{\theta}}))$ for $k = s$. Then the an estimate for the Monte-Carlo error is:

$$(\sigma^{(i)})^2 \approx \frac{\mathrm{var}(h(\boldsymbol{x}_{\boldsymbol{\theta}}))}{N} \underbrace{\sum_{t=-\infty}^{\infty} \rho(t)}_{2\tau_{\mathrm{int}}} = \mathrm{var}(h(\boldsymbol{x}_{\boldsymbol{\theta}})) \frac{2\tau_{\mathrm{int}}}{N}\,, \tag{2.12}$$

where we define the integrated autocorrelation time (IACT) τ_{int} as in [7] and [**<empty citation>**], which provides a good estimate on how many steps the sampling algorithm needs to take to produce one independent sample. More specifically, the effective sample size $\frac{2\tau_{\mathrm{int}}}{N}$ gives an estimate of on how efficient a sampler is.

2.2 Sampling Methods

In this section we present the sampling methods used in this thesis and show how these methods 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 apply sample-based estimates as in Eq. 2.4. Here, $\mathcal{M}$ denotes a Markov chain, where each new sample $(\boldsymbol{x},\boldsymbol{\theta})^{(k)}$ is only affected by the previous one, $(\boldsymbol{x},\boldsymbol{\theta})^{(k-1)}$. Markov chain Monte Carlo (MCMC) methods generate such a chain $\mathcal{M}$ using random (Monte Carlo) proposals $(\boldsymbol{x},\boldsymbol{\theta})^{(k)} \sim q(\cdot|(\boldsymbol{x},\boldsymbol{\theta})^{(k-1)})$ according to a proposal distribution conditioned on the previous sample (Markov), where ergodicity of the chain $\mathcal{M}$ is a sufficient criterion for using sample-based estimates [1, 2].

The ergodicity theorem in [1] states that, if a Markov chain $\mathcal{M}$ is aperiodic, irreducible, and reversible, then it converges to a unique stationary equilibrium distribution. In other words, if the chain can reach any state from any other state (irreducibility), is not stuck in periodic cycles (aperiodicity), and is reversible (detailed balance condition [1]), then it will converge to the desired target distribution with $\mathcal{M} \sim \pi(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})$. In practice, one can inspect the trace $\pi(\boldsymbol{x}^{(k)},\boldsymbol{\theta}^{(k)}|\boldsymbol{y})$ for $k = 1, \ldots, N$ and visually assess convergence and mixing properties of the chain to evaluate ergodicity. The sampling methods used in this thesis possess proven ergodic properties, and we therefore refer the reader to the corresponding literature for further details. Nevertheless, we will give a brief overview of the smapling algorithm used.

2.2.1 Sampling from the marginal posterior

As in Eq. 2.5, when using the MTC method we sample from $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ first and then determine the full conditional $\pi(\boldsymbol{x}|\boldsymbol{y})$ as in Eq. 2.7. To sample from $\pi(\boldsymbol{\theta}|\boldsymbol{y})$, we use a Metropolis-within-Gibbs (MWG) sampler as described in [8]. We apply the MWG sample for the two-dimensional case only, with $\boldsymbol{\theta} = (\theta_1, \theta_2)$, where we perform a Metropolis step in the θ_1 direction and a Gibbs step in the θ_2 direction. Ergodicity for this approach is proven in [9].

The Metropolis-within-Gibbs algorithm begins with an initial guess $\boldsymbol{\theta}^{(t)}$ at $t = 0$. We then propose a new sample $\theta_1 \sim q(\theta_1|\theta_1^{(t-1)})$, conditioned on the previous state, using 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 [9]. We accept and set $\theta_1^{(t)} = \theta_1$ with the acceptance probability

$$\alpha(\theta_1|\theta_1^{(t-1)}) = \min\left\{1, \frac{\pi(\theta_1|\theta_2^{(t-1)}, \boldsymbol{y})\, \cancel{q(\theta_1^{(t-1)}|\theta_1)}}{\pi(\theta_1^{(t-1)}|\theta_2^{(t-1)}, \boldsymbol{y})\, \cancel{q(\theta_1|\theta_1^{(t-1)})}}\right\} \tag{2.13}$$

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 perform a Gibbs step in the θ_2 direction, where Gibbs sampling is again a special case of the Metropolis-Hastings algorithm with acceptance probability equal to one, and draw the next sample $\theta_2^{(t)} \sim \pi(\cdot|\theta_1^{(t)}, \boldsymbol{y})$, conditioned on the current value $\theta_1^{(t)}$.

We repeat this procedure N' times and ensure convergence independently of the initial sample (irreducibility) by discarding the initial $N_{\text{burn-in}}$ samples after a so-called burn-in period, resulting in a Markov chain of length $N = N' - N_{\text{burn-in}}$.

Algorithm 1: Metropolis within Gibbs

1: Initialise 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: $\boldsymbol{\theta}^{(0)}, \ldots, \boldsymbol{\theta}^{(k)}, \ldots, \boldsymbol{\theta}^{(N)} \sim \pi(\boldsymbol{\theta}|\boldsymbol{y})$

2.2.2 t-walk sampler as black box

If the parameters $\boldsymbol{x}$ are functionally dependent on the hyper-parameters $\boldsymbol{\theta}$, i.e., $\boldsymbol{x} = \boldsymbol{x}(\boldsymbol{\theta})$, we can sample directly from the marginal posterior $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ using the t-walk algorithm by Christen and Fox [10]. The t-walk is employed as a black-box sampler, requiring the specification of the number of samples, burn-in period, support region, and the target distribution. Convergence to the target distribution is guaranteed by construction of the algorithm.

2.3 Numerical Approxiamtion Methods - Tensor Train

Instead of sampling from a target distribution $\pi(\boldsymbol{x})$ we can approximate that distribution on a d-dimensional grid with far fewer function evaluation compared to sampling methods using a tensor train (TT) approximation $\tilde{\pi}(\boldsymbol{x}) \approx \pi(\boldsymbol{x})$, with $\boldsymbol{x} \in \mathbb{R}^d$. First, we provide a short overview of probability spaces and their associate measures, as a foundation for calculating marginal probability distribution from the tensor train format. Note, that we follow the notation of Cui et al. [11] to introduce this methodology.

Assume that the triple $(\Omega, \mathcal{F}, \mathbb{P})$ defines a probability space, where Ω denotes the complete sample space, $\mathcal{F}$ is a σ-algebra consisting of a collection of countable subsets $\{A_n\}_{n\in\mathbb{N}}$ with $A_n \subseteq \Omega$, and $\mathbb{P}$ is a probability measure defined on $\mathcal{F}$. The formal conditions for $\mathbb{P}$ to be a probability measure, and for $\mathcal{F}$ to be a σ-algebra over Ω, are given in Appendix C. We denote

$$\mathbb{P}(A) = \int_A \mathrm{d}\mathbb{P} \tag{2.14}$$

as the probability of an event $A \in \mathcal{F}$. By applying the Radon-Nikodym theorem [12], we can change variables

$$\mathbb{P}(A) = \int_A \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\boldsymbol{x}}\,\mathrm{d}\boldsymbol{x} = \int_A \pi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}, \tag{2.15}$$

where $\mathrm{d}\boldsymbol{x}$ is a reference measure on the same probability space, commonly referred to as the Lebesgue measure. The Radon-Nikodym derivative $\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\boldsymbol{x}}$ of $\mathbb{P}$ with respect to $\boldsymbol{x}$ is often interpreted as the probability density function (PDF) $\pi(\boldsymbol{x})$. Thus, we say that $\mathbb{P}$ has a density $\pi(\boldsymbol{x})$ with respect to $\boldsymbol{x}$ [13, Chapter 10].

Now, let $X : \Omega \longrightarrow \mathbb{R}^d$ be a d-dimensional random variable mapping from the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to the measurable space $(\mathbb{R}^d, \mathcal{X})$, where $\mathcal{X}$ is a collection of subsets in $\mathbb{R}^d$. Then the associated PDF $\pi(\boldsymbol{x})$, is a joint density of X, induced by the probability measure on Ω [12, 14]. As in [11], we can define the parameter space as the Cartesian product $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_d$ with $x_k \in \mathcal{X}_k \subseteq \mathbb{R}$ and $\boldsymbol{x} = (x_1, \ldots, x_k, \ldots, x_d)$. The marginal density function for the k-th component is then given by

$$f_{X_k}(x_k) = \int_{\mathcal{X}_1} \cdots \int_{\mathcal{X}_d} \boldsymbol{\lambda}(\boldsymbol{x})\,\pi(\boldsymbol{x})\,\mathrm{d}x_1 \cdots \mathrm{d}x_{k-1}\,\mathrm{d}x_{k+1} \cdots \mathrm{d}x_d, \tag{2.16}$$

where we integrate over all dimensions except the k-th. Here, we introduce a weight function $\boldsymbol{\lambda}(x)$, which can be useful for quadrature rules [15], to which [11] refer to as a "product-form Lebesgue-measurable weighting function" and define it as

$$\boldsymbol{\lambda}(\mathcal{X}) = \prod_{i=1}^{d} \boldsymbol{\lambda}_i(\mathcal{X}_i), \quad \text{where} \quad \boldsymbol{\lambda}_i(\mathcal{X}_i) = \int_{\mathcal{X}_i} \boldsymbol{\lambda}_i(x_i)\,\mathrm{d}x_i.$$

In the tensor train (TT) format the integral in Eq. 2.16 for the marginal probability can be computed at a low computational cost as $\pi(\boldsymbol{x})$ is approximated by

$$\tilde{\pi}(\boldsymbol{x}) = \tilde{\pi}_1(x_1)\tilde{\pi}_2(x_2)\cdots\tilde{\pi}_d(x_d) \in \mathbb{R},$$

which is a sequence of matrix multiplications, with $\tilde{\pi}_k(x_k) \in \mathbb{R}^{r_{k-1}\times r_k}$ for a fixed grid point $\boldsymbol{x} = (x_1, \ldots, x_d)$ on a d-dimensional discrete univariate grid over the parameter space $\mathcal{X}$ bounded by the outer ranks $r_0 = r_d = 1$. We call $\tilde{\pi}_k \in \mathbb{R}^{r_{k-1}\times n\times r_k}$ a TT-core with ranks r_{k-1} and r_k, representing each dimension on n grid points. This enables us to approximate $\pi(\mathcal{X}) \approx \tilde{\pi}_1\tilde{\pi}_2\cdots\tilde{\pi}_d \in \mathbb{R}^d$ using $2nr + (d-2)nr^2$ evaluation points, as illustrated in Figure 2.2, instead of n^d function evaluation. Consequently, the marginal target distribution

$$\begin{aligned} f_{X_k}(x_k) = \frac{1}{z}\Big| &\left(\int_{\mathbb{R}} \boldsymbol{\lambda}_1(x_1)\tilde{\pi}_1(x_1)\,\mathrm{d}x_1\right)\cdots\left(\int_{\mathbb{R}} \boldsymbol{\lambda}_{k-1}(x_{k-1})\tilde{\pi}_{k-1}(x_{k-1})\,\mathrm{d}x_{k-1}\right) \\ &\boldsymbol{\lambda}_k(x_k)\tilde{\pi}_k(x_k) \\ &\left(\int_{\mathbb{R}} \boldsymbol{\lambda}_{k+1}(x_{k+1})\tilde{\pi}_{k+1}(x_{k+1})\,\mathrm{d}x_{k+1}\right)\cdots\left(\int_{\mathbb{R}} \boldsymbol{\lambda}_d(x_d)\tilde{\pi}_d(x_d)\,\mathrm{d}x_d\right)\Big| \end{aligned} \tag{2.17}$$

is computed by integrating over all TT cores except π_k, as in [16], including a normalisation constant z [11].

In practice, tensor train approximations may suffer from numerical instability, in particularly because it is not advantageous to approximate the target function $\pi(\boldsymbol{x})$ in e.g. the logarithmic space. Hence, Cui et al. [11] approximate the square root of the probability density

$$\sqrt{\pi(\boldsymbol{x})} \approx \sqrt{\tilde{\pi}} = \boldsymbol{G}_1(x_1), \ldots, \boldsymbol{G}_k(x_k), \ldots, \boldsymbol{G}_d(x_d)\,, \tag{2.18}$$

which insures positivity. Here, each TT-core is given by

$$G_k^{(\alpha_{k-1},\alpha_k)}(x_k) = \sum_{i=1}^{n_k} \phi_k^{(i)}(x_k)\boldsymbol{A}_k[\alpha_{k-1}, i, \alpha_k], \quad k = 1, \ldots, d, \tag{2.19}$$

where $\boldsymbol{A}_k \in \mathbb{R}^{r_{k-1}\times n_k\times r_k}$ is the k-th coefficient tensor and $\{\phi_k^{(i)}(x_k)\}_{i=1}^{n_k}$ are the basis functions corresponding to the k-th coordinate. The approximated density is written as:

$$\pi(\boldsymbol{x}) \approx \xi + (\sqrt{\tilde{\pi}})^2(\boldsymbol{x}), \tag{2.20}$$

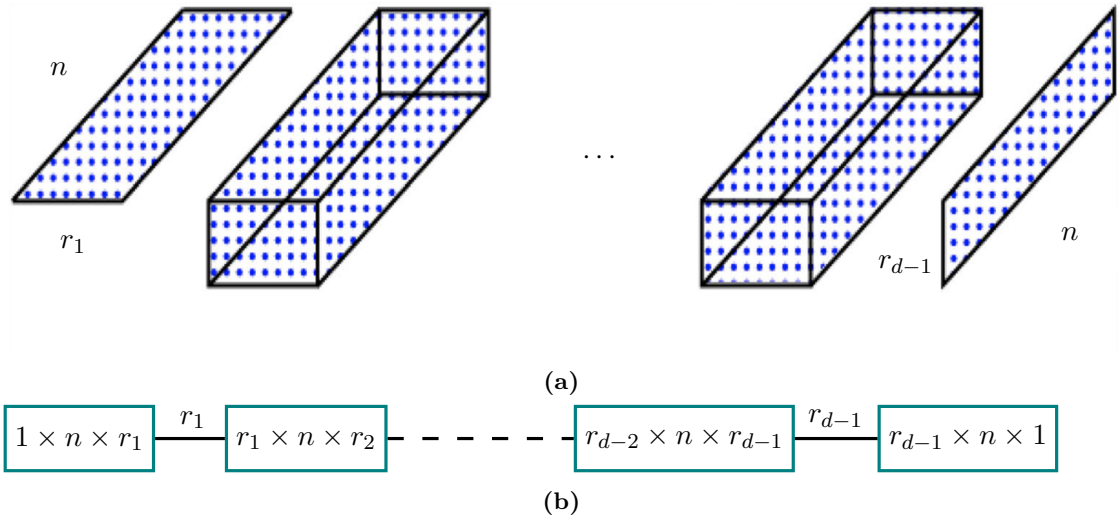

Figure 2.2: Here, we visualise the tensor train cores as two- and three-dimensional matrices. Each core has a length n, corresponding to the number of grid points in one dimension, and the cores are connected through ranks r_k. More specifically, a core $\tilde{\pi}_k$ has dimensions $r_{k-1} \times n \times r_k$, with outer ranks $r_0 = r_d = 1$. Using the TT-format enables us to represent a d-dimensional grid with only dnr^2 evaluation points instead of n^d grid points. Figure (a) is adapted from [17].

where ξ is a positive constant added according to the absolute error and the Lebesgue weighting, see Eq. 2.17, such that

$$0 \leq \xi \leq \frac{1}{\boldsymbol{\lambda}(\mathcal{X})} \|\sqrt{\tilde{\pi}} - \sqrt{\pi}\|_2^2. \tag{2.21}$$

This leads to the normalised target function

$$f_X(x) \approx \frac{1}{z} \left(\boldsymbol{\lambda}(x)\xi + \boldsymbol{\lambda}(x)\tilde{\pi}(x)\right), \tag{2.22}$$

where z is the normalisation constant. Given the tensor train approximation of $\sqrt{\pi}$, the marginal function $f_{X_k}(x_k)$ can be expressed as

$$\begin{aligned} f_{X_k}(x_k) \approx \frac{1}{z} \Bigg(& \xi \prod_{i=1}^{k-1} \lambda_i(\mathcal{X}_i) \prod_{i=k+1}^{d} \lambda_i(\mathcal{X}_i) \\ & + \left(\int_{\mathbb{R}} \lambda_1(x_1) \boldsymbol{G}_1^2(x_1)\, \mathrm{d}x_1\right) \cdots \left(\int_{\mathbb{R}} \lambda_{k-1}(x_{k-1}) \boldsymbol{G}_{k-1}^2(x_{k-1})\, \mathrm{d}x_{k-1}\right) \\ & \lambda_k(x_k) \boldsymbol{G}_k^2(x_k) \\ & \left(\int_{\mathbb{R}} \lambda_{k+1}(x_{k+1}) \boldsymbol{G}_{k+1}^2(x_{k+1})\, \mathrm{d}x_{k+1}\right) \cdots \left(\int_{\mathbb{R}} \lambda_d(x_d) \boldsymbol{G}_d^2(x_d)\, \mathrm{d}x_d\right) \Bigg). \end{aligned} \tag{2.23}$$

2.3.1 Marginal Functions

To compute these marginals efficiently, one can use a procedure similar to left and right orthogonalisation of TT-cores [18]. Cui et al. [11] referred to this backward marginalisation, see Prop. 2, to which I add the forward marginalisation, see Prob. 1. The

backward marginalisation provides us with the coefficient matrices $\boldsymbol{B}_k$, while the forward marginalisation gives the coefficient matrices $\boldsymbol{B}_{\text{pre},k}$. These matrices enable the efficient evaluation of marginal functions since they integrate over the coordinates either left or right of the k-th dimension, as in [11]. For this, we define the mass matrix $\boldsymbol{M}_k \in \mathbb{R}^{n_k \times n_k}$ as

$$\boldsymbol{M}_k[i,j] = \int_{\mathcal{X}_k} \phi_k^{(i)}(x_k)\phi_k^{(j)}(x_k)\boldsymbol{\lambda}(x_k)\,\mathrm{d}x_k, \quad i,j = 1,\ldots,n_k, \tag{2.24}$$

where $\{\phi_k^{(i)}(x_k)\}_{i=1}^{n_k}$ denotes the set of basis functions for the k-th coordinate. The proposition used to compute $\boldsymbol{B}_k$, stated in Proposition 1, is adapted directly from [11].

Proposition 1 (Backward Marginalisation as in [11])**:** 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 $\boldsymbol{C}_k^{(R)} \in \mathbb{R}^{r_{k-1} \times (n_k r_k)}$:

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

3. Compute the new coefficient tensor:

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

Proposition 2 (Forward Marginalisation)**:** Starting with the first coordinate $k = 1$, we set $\boldsymbol{B}_{\text{pre},1} = \boldsymbol{A}_1$. The following procedure can be used to obtain the coefficient tensor $\boldsymbol{B}_{\text{pre},k+1} \in \mathbb{R}^{r_k \times n_{k+1} \times r_{k+1}}$ 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 $\boldsymbol{C}_k \in \mathbb{R}^{r_{k-1} \times n_k \times r_k}$:

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

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

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

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

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

After computing the coefficient tensors $\boldsymbol{B}_{\text{pre},k+1}$ as in Prop. 2 and $\boldsymbol{B}_{k+1}$ from Prop. 1, the marginal PDF of k-th dimension can be expressed as

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

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

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

For the first dimension, $f_{X_1}(x_1)$ can be expressed as

$$f_{X_1}(x_1) \approx \frac{1}{z} \left(\xi \prod_{i=2}^{d} \lambda_i(\mathcal{X}_i) + \sum_{l_1=1}^{r_1} \left(\sum_{i=1}^{n} \phi_1^{(i)}(x_1) \boldsymbol{D}_1[i, l_1] \right)^2 \right) \lambda_1(x_1), \tag{2.33}$$

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

$$f_{X_d}(x_d) \approx \frac{1}{z} \left(\xi \prod_{i=1}^{d-1} \lambda_i(\mathcal{X}_i) + \sum_{l_{n-1}=1}^{r_{d-1}} \left(\sum_{i=1}^{n} \phi_1^{(i)}(x_1) \boldsymbol{D}_d[l_{n-1}, i] \right)^2 \right) \lambda_d(x_d), \tag{2.34}$$

where $\boldsymbol{D}_d[l_{n-1}, i] = \boldsymbol{B}_{\text{pre},d}[l_{n-1}, i, \alpha_{n+1}]$ and $\alpha_{d+1} = 1$. Note that we calculate the normalisation numerically within the process of finding the marginals so that $\sum f_{X_k}(x_k) = 1$.

2.4 Affine Map

The forward map, which we introduce in Ch. 3, poses a weakly non-linear forward problem, which could tackle by treating the problem as a linear problem and then iteratively updating the non-linear part after each parameter sample. Instead we chose to approximate the non-linear model using an affine map $\boldsymbol{M} : \boldsymbol{A}_L\boldsymbol{x} \to \boldsymbol{A}_{NL}\boldsymbol{x}$, which approximates the non-linear model using the linear model. Here we give a brief introduction into affine maps and present our approach to calculate the affine map deterministically, alternatively one can also determine this map using other e.g. machine learning methods [**<empty citation>**].

An affine map is any linear map between two vector spaces or affine spaces, where an affine space does not need to preserve a zero origin, see [19, Def. 2.3.1]. 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 a 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 [19, 20]

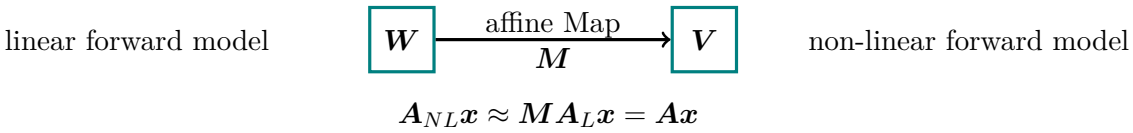

Figure 2.3: This Figure shows the schematic representation of the affine map $\boldsymbol{M}$, which approximates the non-linear forward model from the linear forward model. Here, V contains values produced by the linear forward model, and W contains the corresponding values from the non-linear forward model. Both V and W are affine subspaces over the same field. The affine map $\boldsymbol{M}$ projects elements from the linear forward model space V onto their counterparts in the non-linear forward model space W.

Consequently, to map in between the linear and non-linear forward map we generate two affine subspaces V and W over the same field. Assume we have noise free data vector $A_{NL}\boldsymbol{x} \in \mathbb{R}^m$, then the subspace associated with the linear forward model is

$$\boldsymbol{W} = \begin{bmatrix} | & & | & & | \\ \boldsymbol{A}_L\boldsymbol{x}^{(1)} & \cdots & \boldsymbol{A}_L\boldsymbol{x}^{(j)} & \cdots & \boldsymbol{A}_L\boldsymbol{x}^{(m)} \\ | & & | & & | \end{bmatrix} \in \mathbb{R}^{m\times m} \tag{2.35}$$

and with the non-linear forward model is

$$\boldsymbol{V} = \begin{bmatrix} | & & | & & | \\ \boldsymbol{A}_{NL}\boldsymbol{x}^{(1)} & \cdots & \boldsymbol{A}_{NL}\boldsymbol{x}^{(j)} & \cdots & \boldsymbol{A}_{NL}\boldsymbol{x}^{(m)} \\ | & & | & & | \end{bmatrix} = \begin{bmatrix} — & v_1 & — \\ & \vdots & \\ — & v_j & — \\ & \vdots & \\ — & v_m & — \end{bmatrix} \in \mathbb{R}^{m\times m}, \tag{2.36}$$

Here $\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\}$ are m different parameters.

Then we find the affine map

$$\boldsymbol{V}\boldsymbol{W}^{-1} = \boldsymbol{M} = \begin{bmatrix} — & r_0 & — \\ & \vdots & \\ — & r_j & — \\ & \vdots & \\ — & r_m & — \end{bmatrix} \in \mathbb{R}^{m \times m}, \tag{2.37}$$

row wise by solving $v_j = r_j \boldsymbol{W}$ for r_j, so that $\boldsymbol{A} = \boldsymbol{M}\boldsymbol{A}_{NL} \approx \boldsymbol{A}_{NL}$. Alternately one could also compute $\boldsymbol{M}$ using the inverse $\boldsymbol{W}^{-1}$.

2.5 Regularisation

As mentioned in the introduction the currently most used method to analyse any data in atmospheric physics is regularisation. Since we want to show that our methods is at computationally comparable if not faster and provides more information than regularisation we choose a regularise closet to our linear-Gaussian Bayesian framework, see section 4.2.

The Tikhonov approach in provides one solution $\boldsymbol{x}_\lambda$ that minimises both the data misfit norm

$$\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\| \tag{2.38}$$

and a regularisation semi-norm

$$\lambda \|\boldsymbol{T}\boldsymbol{x}\| \tag{2.39}$$

, as described in [8], with a linear forward model matrix $\boldsymbol{A}$, the data $\boldsymbol{y}$ and a regularisation operator $\boldsymbol{T}$ and the regularisation parameter $\lambda > 0$ which penalise $\boldsymbol{x}$ accordingly. For a fixed λ, the regularised solution

$$\boldsymbol{x}_\lambda = \arg\min_{\boldsymbol{x}} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|^2 + \lambda \|\boldsymbol{T}\boldsymbol{x}\|^2 \tag{2.40}$$

is obtained by taking the derivative with respect to $\boldsymbol{x}$ of the objective function:

$$\nabla_{\boldsymbol{x}} \left\{ (\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x})^T (\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}) + \lambda \boldsymbol{x}^T \boldsymbol{T}^T \boldsymbol{T}\boldsymbol{x} \right\} = 0 \tag{2.41}$$

$$\iff \quad \nabla_{\boldsymbol{x}} \left\{ \boldsymbol{y}^T\boldsymbol{y} + \boldsymbol{x}^T \boldsymbol{A}^T \boldsymbol{A}\boldsymbol{x} - 2\boldsymbol{y}^T \boldsymbol{A}\boldsymbol{x} + \lambda \boldsymbol{x}^T \boldsymbol{T}^T \boldsymbol{T}\boldsymbol{x} \right\} = 0 \tag{2.42}$$

$$\iff \quad 2\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} - 2\boldsymbol{A}^T\boldsymbol{y} + 2\lambda \boldsymbol{T}^T\boldsymbol{T}\boldsymbol{x} = 0, \tag{2.43}$$

or equivalently the "regularised normal equations" $\boldsymbol{A}^T\boldsymbol{y} = \boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} + \lambda \boldsymbol{T}^T\boldsymbol{T}\boldsymbol{x}$ [21]. Solving this equation yields the regularised solution

$$\boldsymbol{x}_\lambda = (\boldsymbol{A}^T\boldsymbol{A} + \lambda \boldsymbol{L})^{-1} \boldsymbol{A}^T\boldsymbol{y}\,, \tag{2.44}$$

where we define $\boldsymbol{L} := \boldsymbol{T}^T\boldsymbol{T}$, which typically represents a discrete matrix approximation of a differential operator choice [1].

In practice, $\boldsymbol{x}_\lambda$ is computed for a range of λ-values and evaluated based on the trade-off between the data misfit and the regularisation norm. The optimal value of λ is often chosen as the point of maximum curvature on the so-called L-curve [22], which we plot in Fig. 4.10. Additionaly one can thnk about it .... in [8]

3
Forward Model

In this chapter we present the forward model to which we apply all our methodology on. We follow the MIPAS handbook [23] and simulate data according to a cloud-free atmosphere in local thermodynamic equilibrium and assume a measurement instrument with infinite spectral resolution and no pointing errors.

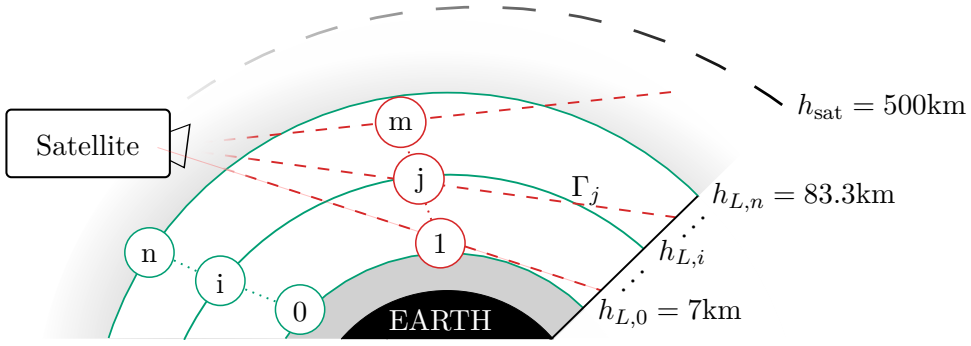

Figure 3.1: Schematic of measurement and analysis geometry, not to scale. The stationary satellite, at a constant height h_{sat} above Earth, takes $m = 41$ measurements along its line-of-sight defining by the line Γ_j. Each measurement has a limb height ℓ_j, $j = 1, 2, \ldots, m$ defined as the closest distance of Γ_j to the Earth surface. Between $h_{L,0} = 7\text{km}$ and $h_{L,n} = 83.3\text{km}$, the stratosphere is discretised into $n = 44$ layers as illustrated by the solid green lines.

A satellite at a constant height h_{sat} points through the atmosphere (limb-sounding) and measures thermal radiation of gas molecules along its line of sight, see Figure 3.1. One measurement of the thermal radiation if we target one specific molecule, in our case ozone denoted by the ozone volume mixing ratio $x(r)$ at distance r from the satellite,

of at the wave number ν is given by the path integral

$$y_j = \int_{\Gamma_j} B(\nu, T) k(\nu, T) \frac{p(T)}{k_{\mathrm{B}} T(r)} x(r) \tau(r) \mathrm{d}r + \eta_j \tag{3.1}$$

$$\tau(r) = \exp\Big\{ - \int_{r_{\mathrm{obs}}}^{r} k(\nu, T) \frac{p(T)}{k_B T(r')} x(r') \mathrm{d}r' \Big\}, \tag{3.2}$$

which is the radiative transfer equation (RTE) [23] where we define a tangent height h_{ℓ_j} and a pointing direction is Γ_j for each $j = 1, 2, \ldots, m$ measurement of the data vector $\boldsymbol{y} \in \mathbb{R}^m$ including some noise η_j. Within the atmosphere 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}. 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

$$k(\nu, T) = L(\nu, T_{\mathrm{ref}}) \frac{Q(T_{\mathrm{ref}})}{Q(T)} \frac{\exp\{-c_2 E''/T\}}{\exp\{-c_2 E''/T_{\mathrm{ref}}\}} \frac{1 - \exp\{-c_2 \nu / T\}}{1 - \exp\{-c_2 \nu / T_{\mathrm{ref}}\}} \tag{3.3}$$

is depend on the line intensity $L(\nu, T_{\mathrm{ref}})$ at reference temperature $T_{\mathrm{ref}} = 296K$, the lower-state energy of the transition E'', the second radiation constant $c2 = 1.4387769\mathrm{cmK}$ all provided by the HITRAN database [24]. The total internal partition function for the lower-state energy is

$$Q(T) = g'' \exp\{-\frac{c_2 E''}{T}\}, \tag{3.4}$$

with the statistical weight g'' (also called the degeneracy factor) accounting for the molecules non-rotational and rotational energy states, see [25]. Under the assumption of local thermodynamic equilibrium (LTE) the black body radiation act as a source function

$$B(\nu, T) = \frac{2hc^2\nu^3}{\exp\{\frac{hc\nu}{k_B T}\} - 1}, \tag{3.5}$$

with Planck's constant h and velocity of light c [**<empty citation>**]. For fundamentals on the Radiative transfer equation we recommend [26, Chapter 1].

To enable matrix-vector multiplication, we discretise the atmosphere in n layers, where the i^{th} layer is defined by two spheres of radii $h_{L,i-1} < h_{L,i}$, for $i = 1, \ldots, n$, with $h_{L,0}$ and $h_{L,n}$. Then we can discretise the ozone, pressure and temperature profiles as a function of height, where in between the heights $h_{L,i-1}$ and $h_{L,i}$, each of the ozone concentration x_i, the pressure p_i, the temperature T_i, as well as the thermal radiation is assumed to be constant. Above $h_{L,n}$ and below $h_{L,0}$, the ozone concentration is set to zero, so no signal can be obtained. 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 rewrite the integral in Eq. (3.1) for one noise free measurement using the trapezoidal rule as a vector-vector

multiplication $\boldsymbol{A_j}(\boldsymbol{x}, \boldsymbol{p}, \boldsymbol{T})\,\boldsymbol{x}$ or $\boldsymbol{A_j}(\boldsymbol{x}, \boldsymbol{p}, \boldsymbol{T})\,\boldsymbol{p}/\boldsymbol{T}$, where the non-linear absorption $\tau(r)$ is included in $\boldsymbol{A_j}(\boldsymbol{x}, \boldsymbol{p}, \boldsymbol{T}) \in \mathbb{R}^n$ which is the j-th row of the matrix $\boldsymbol{A}(\boldsymbol{x}, \boldsymbol{p}, \boldsymbol{T})$. Then given a noise vector $\boldsymbol{\eta} \in \mathbb{R}^m$ the data vector

$$\boldsymbol{y} = \boldsymbol{A}_{NL}\,\boldsymbol{x} + \boldsymbol{\eta} = \boldsymbol{A}_{NL}\,\frac{\boldsymbol{p}}{\boldsymbol{T}} + \boldsymbol{\eta} \tag{3.6}$$

is based on a matrix-vector multiplication, where we define $\boldsymbol{A}(\boldsymbol{x}, \boldsymbol{p}, \boldsymbol{T}) \equiv \boldsymbol{A}_{NL} \in \mathbb{R}^{m \times n}$ for simplicity so that $\boldsymbol{A}_{NL}\boldsymbol{x}$ or $\boldsymbol{A}_{NL}\boldsymbol{p}/\boldsymbol{T}$ implies the construction of $\boldsymbol{A}_{NL}$. If we neglect the absorption, e.g. set $\tau = 1$ in Eq. (3.2), this problem becomes a linear problem with the forward model given by $\boldsymbol{A}_L\boldsymbol{x}$ or $\boldsymbol{A}_L\boldsymbol{p}/\boldsymbol{T}$. Further, we classify the inverse problem as weakly non-linear, see e.g. Fig. 4.9, as neglecting the absorption changes the measurement only slightly.

4
Results and Conclusions

In this chapter we use the forward model to generate data given an underlying ground truth and then guide the reader towards obtaining the posterior distributions. Once we simulated some data we set up an Bayesian framework in Sec. 4.2, where we discuss the choice of prior distributions and formulate the posterior distributions for ozone and pressure over temperature respectively. Since our forward model is weakly non-linear we like to approximates the non-linear forward model $\boldsymbol{A}_{NL} \approx \boldsymbol{M}\boldsymbol{A}_{NL}$ with an affine map $\boldsymbol{M}$, see Sec. 4.3. In doing so we sample from the marginal posterior for ozone and compare that to the tensor-train (TT) approximation. Then we calculate mean and variance of the conditional posterior and use the obtained posterior ozone samples to create two affine subspaces, to which we map in between. Finally we calculate posterior distribution for ozone and pressure over temperature with the updated forward map $\boldsymbol{A} = \boldsymbol{M}\boldsymbol{A}_{NL}$ and compare to a ground truth. Lastly we evaluate on some errors occurring during the process. All programming and analysis is done in python on a MacBook Pro from 2019 with 2.4 Ghz quadcore intel core i5 processor.

4.1 Simulate Data based on a ground truth

We take a ground truth ozone profile generated from some data [27] of the microwave limb sounder on the aura satellite in the Antarctic region with a peak in high altitude (to show that the data is uninformative in those regions), see Fig. 4.8. The ozone profile from [27] provides ozone volume mixing ratios versus pressure, so we recursively calculate the geometric height with the hydrostatic equilibrium equation

$$\frac{\mathrm{d}p}{p} = \frac{-gM}{R^*T}\mathrm{d}h\,, \tag{4.1}$$

subscript i	geometric height h_i in km	gradient a_i
0	0	-6.5
1	11	0
2	20.1	1
3	32.2	2.8
4	47.4	0
5	51.4	-2.8
6	71.8	-2

Table 4.1: Definition of height depending temperature gradients.

with the acceleration due to gravity

$$g = g_0 \left(\frac{r_0}{r_0 + h} \right), \tag{4.2}$$

where the polar radius pf the earth is $r_0 \approx 6356\,\text{km}$, the gravitation at sea level is $g_0 \approx 9.81\text{m/s}^2$, $R^* = 8.31432 \times 10^{-3}\text{Nm/kmol/K}$ and the mean molecular weight of the air is $M = 28.97\text{kg/kmol}$ [28]. This holds up to a geometric height of 86km, where ignore a 0.04% change in M from 80km to 86km in geometric altitude.

Following [28] we form a temperature function

$$T(h) = \begin{cases} T_0 & , \quad h = 0 \\ T_0 + a_0 h & , \quad 0 \leq h < h_1 \\ T_0 + a_0 h_1 & , \quad h_1 \leq h < h_2 \\ T_0 + a_0 h_1 + a_1(h_2 - h_1) + a_2(h - h_2) & , \quad h_2 \leq h < h_3 \\ T_0 + a_0 h_1 + a_1(h_2 - h_1) + a_2(h_3 - h_2) + a_3(h - h_3) & , \quad h_3 \leq h < h_4 \\ T_0 + a_0 h_1 + a_1(h_2 - h_1) + a_2(h_3 - h_2) + a_3(h_4 - h_3) & \\ \quad + a_4(h - h_4) & , \quad h_4 \leq h < h_5 \\ T_0 + a_0 h_1 + a_1(h_2 - h_1) + a_2(h_3 - h_2) & \\ \quad + a_3(h_4 - h_3) + a_4(h_5 - h_4) + a_5(h - h_5) & , \quad h_5 \leq h < h_6 \\ T_0 + a_0 h_1 + a_1(h_2 - h_1) + a_2(h_3 - h_2) & \\ \quad + a_3(h_4 - h_3) + a_4(h_5 - h_4) + a_5(h_6 - h_5) & , \quad h_6 \leq h \lesssim 86 \end{cases}$$

with gradient and height values provided by [28], see Tab. 4.1, which act as the ground truth temperature, see Fig. 4.4.

Then we can compute a data vector $\boldsymbol{y}$, with $m = 42$ measurements according to the radiative transfer equation (RTE), see Eq. 3.1, determined by the satellite pointing accuracy of 150arcsec as requested by the internal report of the proposed cube-satellite

[29], within an atmosphere $h_{L,0} = 7\text{km}$ and $h_{L,n} = 83.3\text{km}$ with $n = 45$ layers. The height values $h_{L,i}$ for each layer $i = 0, \dots, n$ are defined by the ozone profile from [27] and its pressure values. We target thermal radiation at a wave number $\nu = 7.86\text{cm}^{-1}$, equal to a frequency of roughly 235GHz, where we assume that ozone is the only emitter at that frequency, see [**<empty citation>**], and calculate the absorption constant $k(\nu, T)$ as in Eq. 3.2, following the *HITRAN* database [24], which provides the line intensity $L(\nu, T_{\text{ref}})$ for the isotopologue $^{16}O_3$ with the AFGL Code 666. Lastly we add normally distributed $\boldsymbol{\nu} \sim \mathcal{N}(0, \gamma^{-1}\boldsymbol{I})$ noise so that we have a voltage Signal-to-Noise (SNR) of 60, similar to THz module on the MLS aura satellite [30].

inverse problem

4.2 Set up the Bayesian framework

Since the forward model described in Ch. 3 is weakly non-linear we will set up a linear Bayesian hierarchical framework first based on the linear forward model $\boldsymbol{A}_L$ and then later the approximated version $\boldsymbol{A}_{NL}\boldsymbol{M}\boldsymbol{A}_L$. Given some data we are aiming to recover an ozone profile and a pressure over temperature profile. In doing so we first draw a directed acyclic graph (DAG) to visualise the measurement and modelling process and determine hyper-parameters and correlations in between parameters. Then we define prior distribution over all parameters as well as a likelihood function so that we can formulate the posterior distribution.

We draw a DAG for the measurement and modelling process, where the hyper-hyper-parameters $\theta_\gamma, \theta_\delta, \theta_{p_0}, \theta_b, \theta_{\boldsymbol{h}}, \theta_{T_0}, \theta_{\boldsymbol{a}}$ in the top row of Fig. 4.1 determine the hyper-prior distributions $\pi(\gamma, \delta, p_0, b, \boldsymbol{h}_T, \boldsymbol{T}_0, \boldsymbol{a})$ statistically (solid line). Then the hyper-parameters determine the parameters $\boldsymbol{p}/\boldsymbol{T}$ deterministically. The temperature function $\boldsymbol{T} = (T_0, \boldsymbol{a}, \boldsymbol{h}_T)$, Eq. 4.3, is determined through $\boldsymbol{a}$ the temperature gradients at heights $\boldsymbol{h}$, see Tab. 4.1, where h_0 is set to zero as we model temperature variability at the sea-level temperature trough the an additional input T_0. Note that we define an exponential pressure function, Eq. 4.16, later in Sec. **??** so that $\boldsymbol{p}(p_0, b)$ is defined through the hyper-parameters p_0 (pressure at sea-level) and b (exponential gradient). Since we do not parametrise the ozone profile we assume a certain smoothness defined through the smoothness hyper-parameter δ and a precision matrix $\boldsymbol{Q}(\delta)$ which statistically defines a distribution over $\boldsymbol{x}$ (solid lines). The parameters $\boldsymbol{x}, \boldsymbol{p}, \boldsymbol{T}$ progress deterministically, see RTE in Eq. 3.1, into the forward model $\boldsymbol{A}_{NL}$ and generate a space of all possible noise free data $\boldsymbol{\Omega}$. From that space of all measurable $\boldsymbol{\Omega}$ we pick one data set to which we add some noise $\boldsymbol{\eta} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$, which is modelled through the hyper-parameter γ and the precision

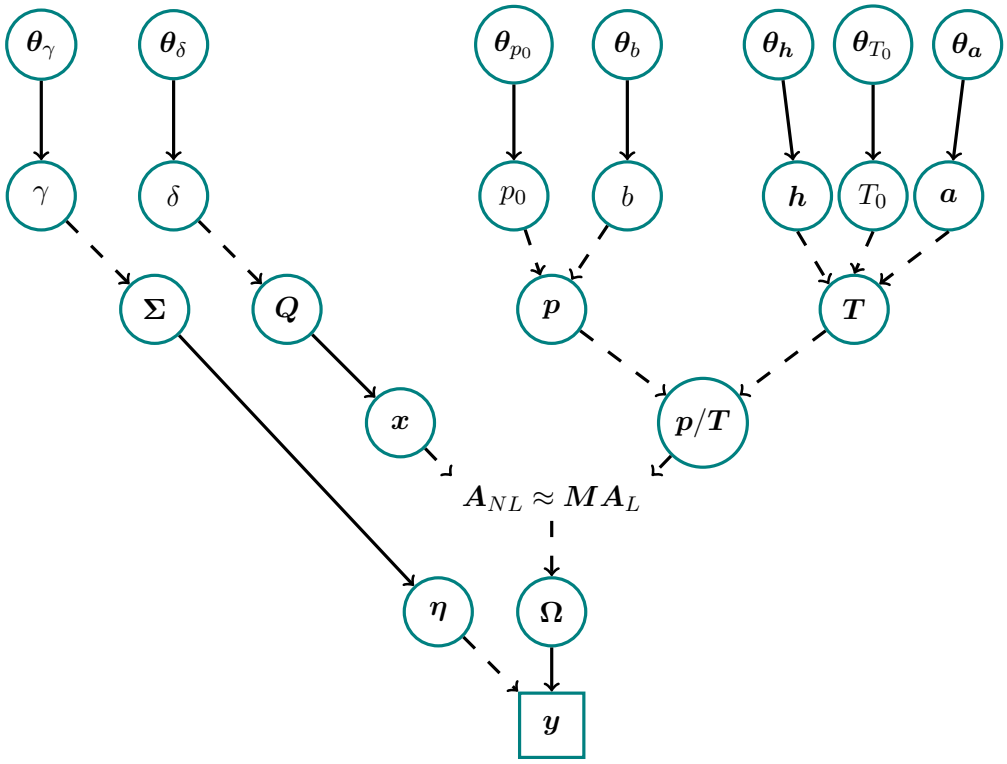

Figure 4.1: Complete directed acyclic graph of the forward model. The hyper-parameters at the top deterministically (dotted line) describe the parameters $(\boldsymbol{p}/\boldsymbol{T})$ or the noise covariance $\boldsymbol{\Sigma} = \gamma^{-1}\boldsymbol{I}$ of the random (solid line) noise $\boldsymbol{\eta} \sim \mathcal{N}(0, \gamma^{-1}\boldsymbol{I})$ and precision matrix $\boldsymbol{Q} = \delta\boldsymbol{L}$ of the distribution of $\boldsymbol{x} \sim \mathcal{N}(0, \delta\boldsymbol{L})$, where $\boldsymbol{L}$ is a graph Laplacian as in Eq. 4.6. We can group the noise precision γ and the smoothness parameter δ to define the marginal posterior over those hyper-parameters and then condition on them for the conditional posterior distribution,for further details see Fig. A.1. In this whole process where we condition on the pressure $\boldsymbol{p}$ and temperature $\boldsymbol{T}$, which we retrieve separately, see Fig. **??**. The hyper-parameters h_0, p_0, b deterministically describe the pressure function in Eq. 4.16, note that we only need three parameters here since $h_0 < h_{L,0}$ and $\boldsymbol{h} = \{h_1, h_2, h_3, h_4, h_5, h_6\}$, $\boldsymbol{a} = \{a_0, a_1, a_2, a_3, a_4\}$ and T_0 determine the temperature function. The parameters $\boldsymbol{x}$ and $\boldsymbol{p}/\boldsymbol{T}$ determine the space of all measurable noise free data $\boldsymbol{\Omega}$ through the forward model $\boldsymbol{A}(\boldsymbol{x}, \boldsymbol{p}, \boldsymbol{T})$ from which we randomly observe data set plus some random noise.

matrix $\boldsymbol{\Sigma} = \gamma^{-1}\boldsymbol{I}$ so that we obtain the noisy data vector $\boldsymbol{y}$. Since the noise is normally distributed, so is the likelihood function $\pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{p}, \boldsymbol{T})$. Then the joint posterior distribution

$$\pi(p_0, b, \boldsymbol{h_T}, \boldsymbol{a}, \delta, \gamma, \boldsymbol{x}|\boldsymbol{y}) \propto \pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{p}, \boldsymbol{T})\pi(p_0, b, \boldsymbol{h_T}, \boldsymbol{a}, \delta, \gamma) \tag{4.3}$$

over all 17 hyper-parameters and the parameter $\boldsymbol{x} \in \mathbb{R}^{45}$ is 62 dimensional. Ideally we characterise the joint posterior but this is computationally not feasible so we will factorise the posterior into

$$\pi(p_0, b, \boldsymbol{h_T}, \boldsymbol{a}, \delta, \gamma, \boldsymbol{x}|\boldsymbol{y}) = \pi(\delta, \gamma, \boldsymbol{x}|p_0, b, \boldsymbol{h_T}, \boldsymbol{a}, \boldsymbol{y})\pi(p_0, b, \boldsymbol{h_T}, \boldsymbol{a}|\delta, \gamma, \boldsymbol{x}, \boldsymbol{y})\,, \tag{4.4}$$

where we either condition on ozone $\boldsymbol{x}$ and the smoothness hyper-parameter δ as well as the noise hyper-parameter γ or on the fraction $\boldsymbol{p}/\boldsymbol{T}$, pressure over temperature, and

		TT bounds			
model parameters	priors	lower	upper	τ_{int}	Context
γ	$\mathcal{T}(1, 10^{-10})$	$5\,10^{-8}$	$4.5\,10^{-7}$		$\boldsymbol{y}$
δ	$\mathcal{T}(1, 10^{-10})$	-	-		$\boldsymbol{x}$
λ	-	500	7000		$\boldsymbol{x}$
$\boldsymbol{x}$	$\mathcal{N}(0, \delta\boldsymbol{L})$	-	-		$\boldsymbol{x}$
h_0	$\mathcal{N}(5.5, 0.5)$	4.76	5.74		$\boldsymbol{p}/\boldsymbol{T}$
p_0	$\mathcal{N}(500, 6)$	479	519		$\boldsymbol{p}/\boldsymbol{T}$
b	$\mathcal{N}(0.167, 7\,10^{-4})$	0.165	0.170		$\boldsymbol{p}/\boldsymbol{T}$
h_1	$\mathcal{N}(11, 0.1)$	10.6	11.3		$\boldsymbol{p}/\boldsymbol{T}$
h_2	$\mathcal{N}(20.1, 0.9)$	16.7	22.8		$\boldsymbol{p}/\boldsymbol{T}$
h_3	$\mathcal{N}(32.3, 3)$	23.8	43.6		$\boldsymbol{p}/\boldsymbol{T}$
h_4	$\mathcal{N}(47.4, 0.5)$	45.5	49.3		$\boldsymbol{p}/\boldsymbol{T}$
h_5	$\mathcal{N}(51.4, 0.5)$	49.5	53.3		$\boldsymbol{p}/\boldsymbol{T}$
h_6	$\mathcal{N}(71.8, 3)$	60.6	83.1		$\boldsymbol{p}/\boldsymbol{T}$
a_0	$\mathcal{N}(-6.5, 0.01)$	-6.54	-6.46		$\boldsymbol{p}/\boldsymbol{T}$
a_1	$\mathcal{N}(1, 0.01)$	0.96	1.04		$\boldsymbol{p}/\boldsymbol{T}$
a_2	$\mathcal{N}(2.8, 0.1)$	2.43	3.18		$\boldsymbol{p}/\boldsymbol{T}$
a_3	$\mathcal{N}(-2.8, 0.1)$	-3.18	-2.43		$\boldsymbol{p}/\boldsymbol{T}$
a_4	$\mathcal{N}(-2, 0.01)$	-2.04	-1.96		$\boldsymbol{p}/\boldsymbol{T}$
T_0	$\mathcal{N}(288.15, 2)$	281.8	294.5		$\boldsymbol{p}/\boldsymbol{T}$

Table 4.2: Summary of relevant parameter characteristics, bounds and sampling statistics. We denote $\mathcal{N}(\mu, \sigma)$ as the Gaussian and $\mathcal{T}(\alpha = \text{scale}, \beta = \text{rate})$ as the gamma distribution.

its hyper-parameters. Again as in Sec. 3 for brevity we write $\pi(p_0, b, \boldsymbol{h_T}, \boldsymbol{a}|\gamma, \boldsymbol{y})$ and $\pi(\delta, \gamma, \boldsymbol{x}|\boldsymbol{y})$, which implies that we conditioned on $\boldsymbol{x}$ or $\boldsymbol{p}$ and $\boldsymbol{T}$. Next we need to specify the prior distribution, which we summarise in Tab. 4.2, in order to formulate the posterior distributions.

4.2.1 Ozone conditioned on pressure and temperature

In this section we set the priors and describe the approach to evaluate the posterior distribution for ozone $\pi(\delta, \gamma, \boldsymbol{x}|\boldsymbol{y})$, including the noise hyper-parameter γ. Assuming Gaussian noise $\boldsymbol{\eta} \sim \mathcal{N}(0, \gamma^{-1}\boldsymbol{I})$, we define a linear-Gaussian Bayesian hierarchical model [8]

$$\boldsymbol{y}|\boldsymbol{x}, \gamma \sim \mathcal{N}(\boldsymbol{A}\boldsymbol{x}, \gamma^{-1}\boldsymbol{I}) \tag{4.5a}$$

$$\boldsymbol{x}|\delta \sim \mathcal{N}(0, \delta\boldsymbol{L}) \tag{4.5b}$$

$$\delta, \gamma \sim \pi(\delta, \gamma), \tag{4.5c}$$

with a normally distributed likelihood $\pi(\boldsymbol{y}|\boldsymbol{x},\gamma)$ including the forward model matrix $\boldsymbol{A}$ and prior distributions $\pi(\boldsymbol{x}|\delta)$ and $\pi(\delta,\gamma)$, the noise covariance matrix $\gamma^{-1}\boldsymbol{I}$, the prior precision matrix $\delta\boldsymbol{L}$ and the prior mean set to $\mathbf{0}$. We choose this model as it is very similar to a regularisation problem and we like to show that we can apply Bayesian methodology to receive results including uncertainties.

Prior Modelling

To complete the Bayesian framework we have define prior distribution for the hyper-parameters and parameters. Ideally we define the prior distributions as uninformative as possible, but include functional dependencies and physical properties.

First we set the precision matrix of the prior distribution for $\boldsymbol{x}$ to

$$\delta\boldsymbol{L} = \delta \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix} \tag{4.6}$$

which is the 1D Graph Laplacian as in [8, 31] with Dirichlet boundary condition, which will also act as the regulariser later in the Regularisation section, see Sec. 2.5. For δ and γ we pick gamma distributions so that $\gamma \sim \mathcal{T}(\boldsymbol{\theta_\gamma})$ and $\delta \sim \mathcal{T}(\boldsymbol{\theta_\delta})$, where $\boldsymbol{\theta_\gamma} = \boldsymbol{\theta_\delta} = (1, 10^{-10})$. These are relatively uninformative distributions see Fig. 4.11 and Fig. A.2, where we plot ozone profiles according to $\boldsymbol{x} \sim \mathcal{N}(0, \delta\boldsymbol{L})$ and like to note that we should not include negative ozone values but are currently not able to include e.g. a truncated multivariate normal prior distribution for $\boldsymbol{x}$. These gamma distributions have another advantage when sampling from the marginal posterior distribution $\pi(\gamma,\delta|\boldsymbol{y})$, where $\pi(\gamma|\lambda,\boldsymbol{y}) \sim \mathcal{T}(\cdot)$ with the regularisation parameter $\lambda = \delta/\gamma$.

posterior distribution into marginal and conditional posterior distribution

As noted in Sec. 2.1 we will factorise the posterior

$$\pi(\boldsymbol{x},\gamma,\delta|\boldsymbol{y}) \propto \pi(\boldsymbol{y}|\boldsymbol{x},\gamma,\delta)\pi(\boldsymbol{x},\gamma,\delta) \tag{4.7}$$

into

$$\pi(\boldsymbol{x},\gamma,\delta|\boldsymbol{y}) = \pi(\boldsymbol{x}|\gamma,\delta,\boldsymbol{y})\pi(\gamma,\delta|\boldsymbol{y}) \tag{4.8}$$

the marginal posterior $\pi(\gamma,\delta|\boldsymbol{y})$ and conditional posterior $\pi(\boldsymbol{x}|\gamma,\delta,\boldsymbol{y})$. Fox and Norton call this method the marginal and then conditional method (MTC) [8], where we break the correlation structure between $\boldsymbol{x}$ and γ,δ as illustrated in Fig. A.1 and Fig. B.1

by marginalising over $\boldsymbol{x}$ and evaluating this marginal posterior first and *then* the conditional posterior.

For the linear-Gaussian Bayesian hierarchical model specified in Eq. 4.17, the marginal posterior distribution over the hyper-parameters is given by

$$\pi(\lambda, \gamma|\boldsymbol{y}) \propto \lambda^{n/2}\gamma^{m/2} \exp\left\{-\frac{1}{2}g(\lambda) - \frac{\gamma}{2}f(\lambda)\right\}\pi(\lambda, \gamma), \tag{4.9}$$

with $\lambda = \delta/\gamma$, and

$$f(\lambda) = \boldsymbol{y}^T\boldsymbol{y} - (\boldsymbol{A}^T\boldsymbol{y})^T(\boldsymbol{A}^T\boldsymbol{A} + \lambda\boldsymbol{L})^{-1}(\boldsymbol{A}^T\boldsymbol{y})\,, \tag{4.10a}$$

$$\text{and } g(\lambda) = \log\det(\boldsymbol{A}^T\boldsymbol{A} + \lambda\boldsymbol{L})\,, \tag{4.10b}$$

see [8, Lemma 2]. When considering $\boldsymbol{x}$ and $\boldsymbol{y}$ as a joint multivariate normal distribution or a joint Gaussian Markov random field $(\boldsymbol{x}^T, \boldsymbol{y}^T)^T$, then $\boldsymbol{x}$ conditioned on the hyper-parameters γ, δ and the data $\boldsymbol{y}$ is the normally distributed conditional posterior distribution

$$\boldsymbol{x}|\delta, \gamma, \boldsymbol{y} \sim \mathcal{N}\big(\underbrace{(\boldsymbol{A}^T\boldsymbol{A} + \delta/\gamma\boldsymbol{L})^{-1}\boldsymbol{A}^T\boldsymbol{y}}_{\boldsymbol{x}_\lambda}, \underbrace{(\gamma\boldsymbol{A}^T\boldsymbol{A} + \delta\boldsymbol{L})}_{\gamma\boldsymbol{B}_\lambda}{}^{-1}\big)\,, \tag{4.11}$$

see [4, 8, 32] for more information. In this thesis we compute the mean

$$\mu_{\boldsymbol{x}|\boldsymbol{y}} = \int \boldsymbol{x}_\lambda \pi(\lambda|\boldsymbol{y})\mathrm{d}\lambda \approx \sum \boldsymbol{x}_{\lambda_i}\pi(\lambda_i|\boldsymbol{y})\,, \tag{4.12}$$

and covariance

$$\Sigma_{\boldsymbol{x}|\boldsymbol{y}} = \int \gamma^{-1}\pi(\gamma|\boldsymbol{y})\,\mathrm{d}\gamma \int \boldsymbol{B}_\lambda^{-1}\,\pi(\lambda|\boldsymbol{y})\,\mathrm{d}\lambda \approx \sum {\gamma_i}^{-1}\pi(\gamma_i|\boldsymbol{y}) \sum \boldsymbol{B}_{\lambda_i}^{-1}\pi(\lambda_i|\boldsymbol{y}) \tag{4.13}$$

as weighted expectations, by quadrature [35, Sec. 2.1], with $\sum\pi(\lambda_i|\boldsymbol{y}) = \sum\pi(\gamma_i|\boldsymbol{y}) = 1$. If that is too costly the randomise-then-optimise (RTO) [8, 33] may be a feasible alternative to sample from Eq. 4.11.

Most of the computational effort lays in the function $f(\lambda)$ and $g(\lambda)$ from the marginal posterior in Eq. 4.9. In Fig. 4.2 we see that $f(\lambda)$ and $g(\lambda)$ are well behaved within the region of interest. Consequently we approximate $f(\lambda) \approx \tilde{f}(\lambda)$ with 3rd order Taylor series around the mode λ_0 of $\pi(\lambda, \gamma|\boldsymbol{y})$. We also note that $\tilde{g}(\lambda) \approx g(\lambda)$ is behaves linear around λ_0 in the log-space. As a result of these observations the approximations are implicitly given by

$$f^{(r)}(\lambda_0) = (-1)^{r+1}r!(\boldsymbol{A}^T\boldsymbol{y})^T(\boldsymbol{B}_0^{-1}\boldsymbol{L})^r\boldsymbol{B}_0^{-1}\boldsymbol{A}_L^T\boldsymbol{y} \tag{4.14}$$

$$\text{and } \log\tilde{g}(\lambda) = (\log\lambda - \log\lambda_0)\frac{\log g(\lambda_{\max}) - \log g(\lambda_0)}{\log\lambda_{\max} - \log\lambda_0} + \log g(\lambda_0) \tag{4.15}$$

with $\boldsymbol{B}_0 = \boldsymbol{A}^T\boldsymbol{A} + \lambda_0\boldsymbol{L}$. We plot the approximations in Fig. 4.2 and elaborate on approximation errors in sec 4.7

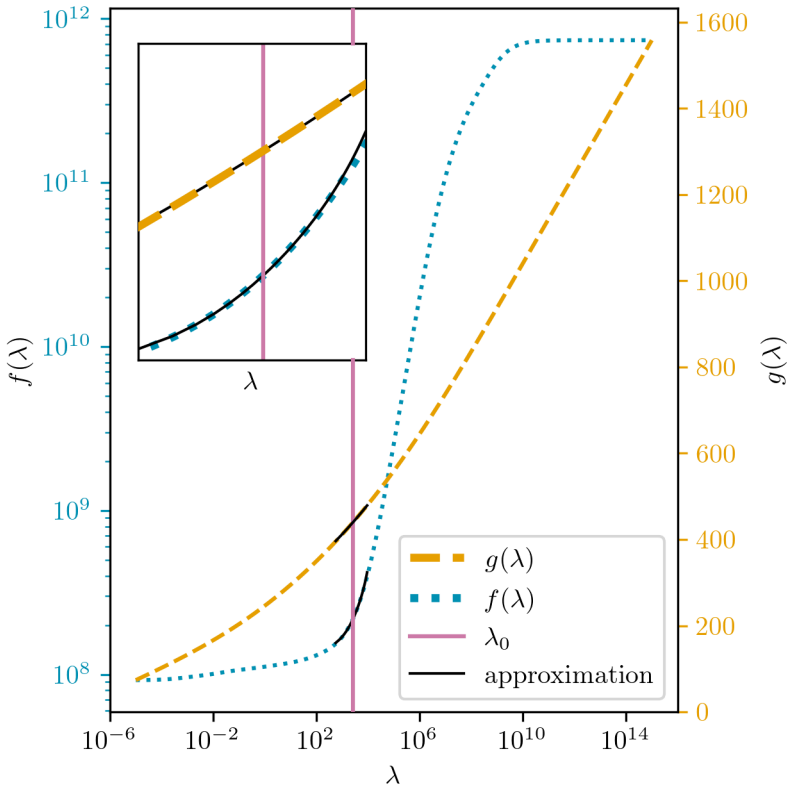

Figure 4.2: Plot of the functions $f(\lambda)$ and $g(\lambda)$ from the marginal posterior for a wide range of $\lambda = \delta/\gamma$. We plot the third order Taylor series in black around the mode of the marginal posterior (vertical line) for the sampling range of λ within the MWG algorithm.

4.2.2 Pressure over temperature conditioned on noise and ozone

First we obsevre that we can describe the pressure values in between $h_{L,0} = 7$km and $h_{L,n} = 83.3$km with an exponential function

$$p(h) = \exp\{-b\,h\}\,p_0 \quad , h_{L,n} \leq h \leq h_{L,0} \tag{4.16}$$

so that we parametrize the pressure $\boldsymbol{p}$ with the hyperparameters p_0, b. Then, within the hierachical Bayesian framwork

$$\boldsymbol{y}|\boldsymbol{p}, \boldsymbol{T}, \gamma \sim \mathcal{N}(\boldsymbol{A}\,\boldsymbol{p}/\boldsymbol{T}, \gamma^{-1}\boldsymbol{I}) \tag{4.17a}$$
$$\boldsymbol{a} \sim \mathcal{N}(\boldsymbol{\mu_a}, \boldsymbol{\Sigma_a}) \tag{4.17b}$$
$$\boldsymbol{h_T} \sim \mathcal{N}(\boldsymbol{\mu_T}, \boldsymbol{\Sigma}_{\boldsymbol{h}_T}) \tag{4.17c}$$
$$T_0 \sim \mathcal{N}(\mu_{T_0}, \sigma_{T_0}) \tag{4.17d}$$
$$p_0 \sim \mathcal{N}(\mu_{p_0}, \sigma_{p_0}) \tag{4.17e}$$
$$b \sim \mathcal{N}(\mu_b, \sigma_b) \tag{4.17f}$$

we define a normally distributed likelihood (due to Gaussian noise) and priors, where the hyper-prior mean and variances relate to the DAG in Fig. 4.1 so that $\boldsymbol{\theta_a} = (\boldsymbol{\mu_a}, \boldsymbol{\Sigma_a})$,

$\boldsymbol{\theta}_{\boldsymbol{h}_T} = (\boldsymbol{\mu}_{\boldsymbol{T}}, \boldsymbol{\Sigma}_{\boldsymbol{h}_T})$, $\boldsymbol{\theta}_{T_0} = (\mu_{T_0}, \sigma_{T_0})$, $\boldsymbol{\theta}_{p_0} = (\mu_{p_0}, \sigma_{p_0})$, and $\boldsymbol{\theta}_b = (\mu_b, \sigma_b)$. Note that we don not include h_0, from Tab. 4.1, in $\boldsymbol{h}_T$ since we model temperature variablilty at sea level through T_0.

Prior modelling

We summarise the mean and variance in Tab. 4.2 and plot the prior distributions samples against the ground truth for presssure $\boldsymbol{p}$ and temperature $\boldsymbol{T}$ seperately in Fig. 4.5 and 4.4 and jointly as $\boldsymbol{p}/\boldsymbol{T}$ in Fig. 4.4. We plot the prior distributions samples against the ground truth for $1/\boldsymbol{K}$ in Fig. A.4.

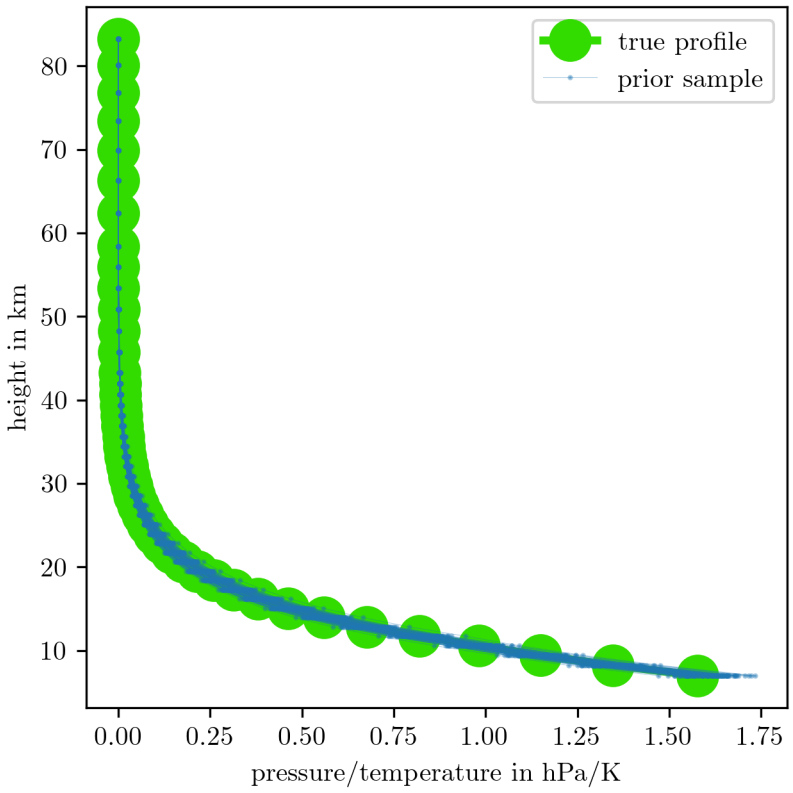

Figure 4.3: We draw samples from the hyper-prior distribution of $h_0, b, p_0, h_1, h_2, h_3, h_4, h_5, h_6, a_0, a_1, a_2, a_3, a_4$ and T_0 as defined in table 4.2 and then calculate $\boldsymbol{p}/\boldsymbol{T}$ according to the functions in Eq. 4.16 and 4.3.

We carefully choose the hyper-prior distributions $\boldsymbol{h}_{\boldsymbol{T}}$ so that the individal distributions for heights $h_1, h_2, h_3, h_4, h_5, h_6$ do not overlap, see Fig. A.3. Additionally we define the sampling space and the grid for the TT approximation accordingly. We remark that, we can already observe in Fig. 4.3 that $\boldsymbol{p}/\boldsymbol{T}$ inherits the structure of pressure function.

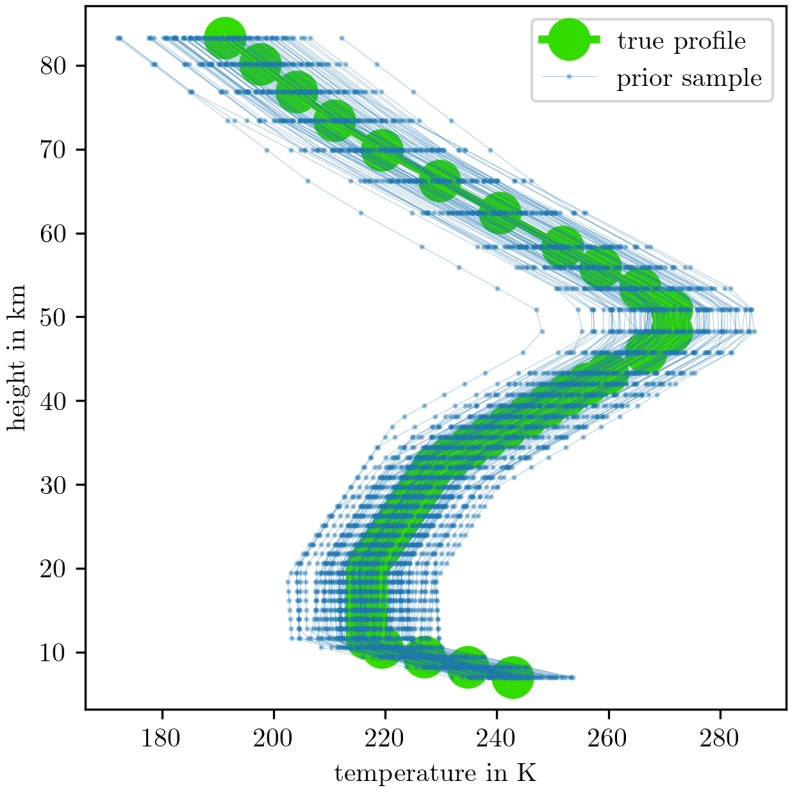

Figure 4.4: We draw samples from the hyper-prior distribution of $h_1, h_2, h_3, h_4, h_5, h_6, a_0, a_1, a_2, a_3, a_4$ and T_0 as defined in table 4.2 and then calculate $\boldsymbol{T}$ according to the function in Eq. 4.3.

posterior distribution

Then we can define the posterior distrubtion

$$\pi(p_0, b, \boldsymbol{h_T}, \boldsymbol{c_T}, \boldsymbol{a_T} | \boldsymbol{y}, \gamma, \boldsymbol{x}) \propto \exp\Big\{ -\frac{\gamma}{2} \Big\| \boldsymbol{y} - \boldsymbol{A}\frac{\boldsymbol{p}}{\boldsymbol{T}} \Big\|^2 \Big\} \pi(p_0, b, \boldsymbol{h_T}, \boldsymbol{c_T}, \boldsymbol{a_T}) \,, \tag{4.18}$$

which is, conditioned on the noise hyper-parameter γ, the ozone profile $\boldsymbol{x}$ and the smoothness hyper-parameter δ, a 16 dimesnional distribution.

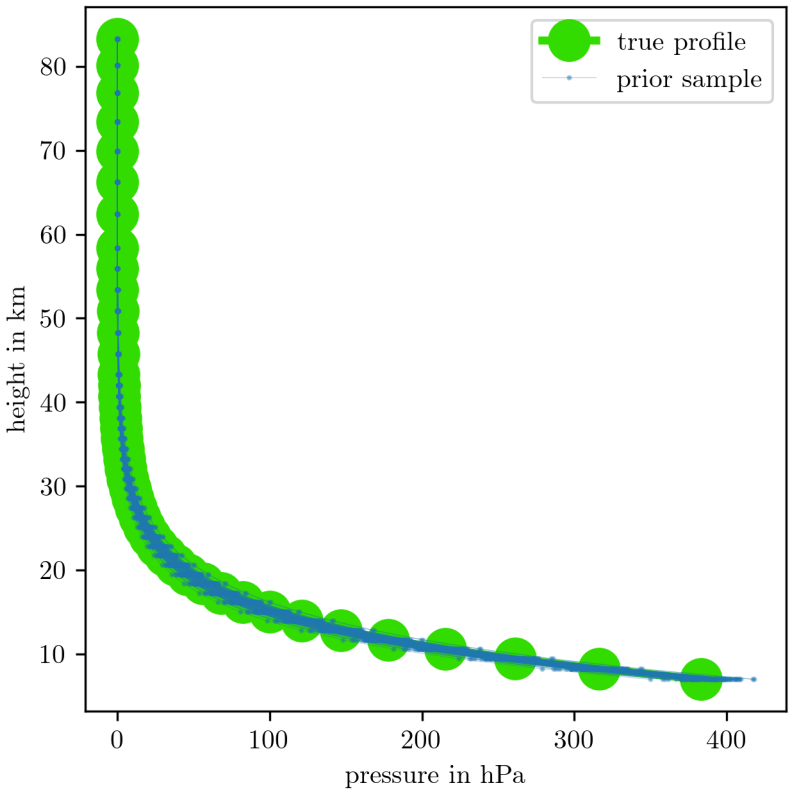

Figure 4.5: We draw samples from the hyper-prior distribution of h_0, b and p_0 as defined in table 4.2 and then calculate $\boldsymbol{p}$ according to the function in Eq. 4.16.

4.3 Approximate non-linear forward model with affine Map

With the posterior distributions formulated we now want to find an affine map $\boldsymbol{M}$ to approximate the non-linear forward model, we summarized the strategy for doing so in Fig. 4.6. We focus on the posteroir distribution of ozone profiles, by conditioning on pressure and temperature, as this is a quick process when using the MTC method. We approximate and sample from the marginal posterior $\pi(\gamma, \delta|\boldsymbol{y})$ and then charcterise the full conditional posterior distribution $\pi(\boldsymbol{x}|\boldsymbol{y})$ based on the linear forward model $\boldsymbol{A}_L$ which neglects absorption, see Eq. 3.1. Given samples $\boldsymbol{x} \sim \pi(\boldsymbol{x}|\boldsymbol{y})$ from the full conditional posterior distribution we can generate two affine subspaces based on the linear and non-linear model and find the mapping in between those.

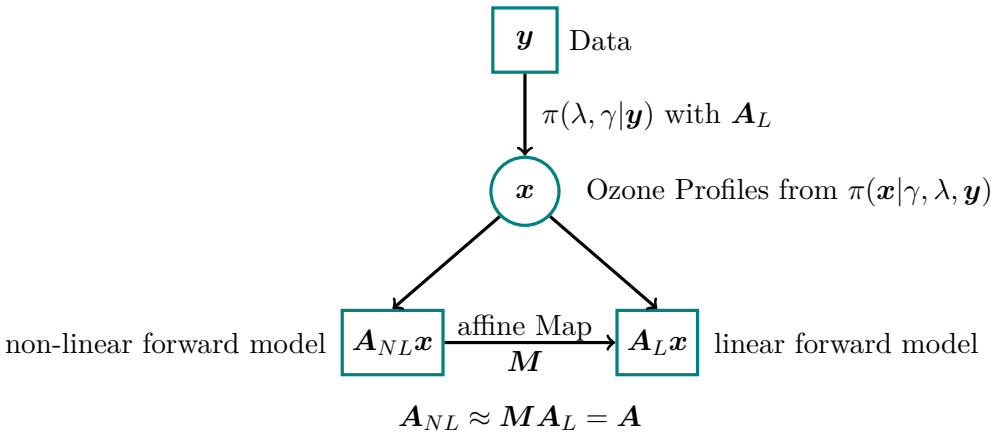

Figure 4.6: The strategy to find the affine map consist of evaluating the marginal posterior for ozone using the linear forward model. Then we draw ozone samples from the conditional posterior and calculate noise free data based on the linear and non-linear forward model. Next we find a mapping in between those two space so that we can approximate the non-linear forward model using an affine map and the linear forward model.

4.3.1 Sample from marginal posterior distribution for ozone - linear model

We set $\boldsymbol{A} = \boldsymbol{A}_L$ and characterise the marginal posterior $\pi(\lambda, \gamma|\boldsymbol{y})$ as in Eq. 4.9 by employing a Metropolis within Gibbs (MWG) algorithm. More specifically, we implement a Metropolis random walk on the full conditional

$$\pi(\lambda|\gamma, \boldsymbol{y}) \propto \lambda^{n/2+\alpha_\delta-1} \exp\left\{-\frac{1}{2}g(\lambda) - \frac{\gamma}{2}f(\lambda) - \beta_\delta\gamma\lambda\right\} \tag{4.19}$$

and do a Gibbs steps on

$$\gamma|\lambda, \boldsymbol{y} \sim \Gamma\left(\frac{m}{2} + \alpha_\delta + \alpha_\gamma, \frac{1}{2}f(\lambda) + \beta_\gamma + \beta_\delta\lambda\right) \tag{4.20}$$

to generate marginal posterior samples $(\lambda,\gamma)^{(1)},\dots,(\lambda,\gamma)^{(N)} \sim \pi(\lambda,\gamma|\boldsymbol{y})$. Note that, when changing variables from $\delta = \lambda\gamma$ to λ the hyper-prior distribution changes to $\pi(\lambda) \propto \lambda^{\alpha_\delta - 1}\gamma^{\alpha_\delta}\exp(-\beta_\delta\lambda\gamma)$, due to $\mathrm{d}\delta/\mathrm{d}\lambda = \gamma$.

Hence we run a Metropolis random walk on $\pi(\lambda|\gamma,\boldsymbol{y})$, the proposal distribution $q(\lambda'|\lambda^{(k)}) \sim \mathcal{N}(\lambda^{(k)}, 0.8\lambda_0)$ conditioned on the previous sample $\lambda^{(k)}$, with $k = 1,\dots,N$ is symmetric. Then, we accept or reject a new λ' sample by comparing the acceptance ratio

$$\log\left\{\frac{\pi(\lambda'|\gamma^{(k)},\boldsymbol{y})}{\pi(\lambda^{(k)}|\gamma^{(k)},\boldsymbol{y})}\right\} = \log\{\pi(\lambda'|\gamma^{(k)},\boldsymbol{y})\} - \log\{\pi(\lambda^{(k)}|\gamma^{(k)},\boldsymbol{y})\} \tag{4.21}$$

$$= \frac{n}{2}(\log\{\lambda'\} - \log\{\lambda^{(t-1)}\}) + \frac{1}{2}\Delta g + \frac{\gamma^{(t-1)}}{2}\Delta f + \beta_\delta\gamma^{(t-1)}\Delta\lambda\,, \tag{4.22}$$

where $\Delta\lambda = \lambda' - \lambda^{(k)}$ to a random uniform number in between 0 and 1. Note that since we calculate the acceptance ratio in the log space $\Delta f \approx \tilde{f}(\lambda') - \tilde{f}(\lambda^{(k)}) = \sum f^{(r)}(\lambda_0)\Delta\lambda' - \Delta\lambda^{(k)}$ is a 3rd order taylor approximaton ,see Fig. 4.2, where $\Delta\lambda' = \lambda' - \lambda_0$ and $\Delta\lambda^{(k)} = \lambda^{(k)} - \lambda_0$. Similarly we approximate $\Delta g \approx \exp\log\tilde{g}(\lambda') - \exp\log\tilde{g}(\lambda^{(k)})$ as in Eq. 4.10. Lastly, a Gibbs step provides a new $\gamma^{(k+1)} \sim \gamma|\lambda^{(k+1)},\boldsymbol{y}$, see Equation (4.20). See Algorithmic Box **??** for summary of the general version.

We initialise the MWG at the mode $(\lambda^{(0)},\gamma^{(0)}) = (\lambda_0,\gamma_0)$ and take for $N = 20000$ plus $N_{\text{burn-in}} = 100$ steps in less than one second. The standard deviation of the normal proposal distribution is set to $\sigma_\lambda = 0.8\lambda_0$ so that the acceptance rate is ≈ 0.5 as suggested in [**<empty citation>**]. The samples are plotted in Fig. 4.7 as a 2D scatter plot, as well as the trace of the MwG to show ergodicity. We calculate the integrated autocorrelation time (IACT) with the python implemnaton of [**<empty citation>**], which gives us $\tau_{\text{int},\gamma}$ = and $\tau_{\text{int},\delta}$ =.

4.3.2 Tensor-train approximation of the marginal posterior distribution for ozone

Alternatively we can approximate the marginal posterior with a tensor-train (TT) of the square root of marginal posterior on a predefined grid. We define a grid similar to the sampling region of the MWG sampler with 25 grid points in each dimension and use the `tt.cross.rectcross.rect_cross.cross` function from the `ttpy` python package, based on the rect cross algorithm in [**<empty citation>**], to caclulate the cores of each dimesnion in less than 0.1s. We set the number of ranks to a constant value $r = 4$ and optimse over those rankes with one sweep. To avoid underflow we have to add a 'normalisation' constant $c = 460$ so that $\pi(\lambda|\gamma,\boldsymbol{y}) = \exp\{\log\pi(\lambda|\gamma,\boldsymbol{y}) + c\}$. Then we calculate the marginals $\pi(\lambda|\boldsymbol{y})$ and $\pi(\gamma|\boldsymbol{y})$ as described in section **??**, assuming an absolute error of 1 the constant $\xi = 1/\lambda(\mathcal{X})$. We plot the TT approximation as a colour map on top of the obtained samples in the scatter plot in Fig. 4.7.

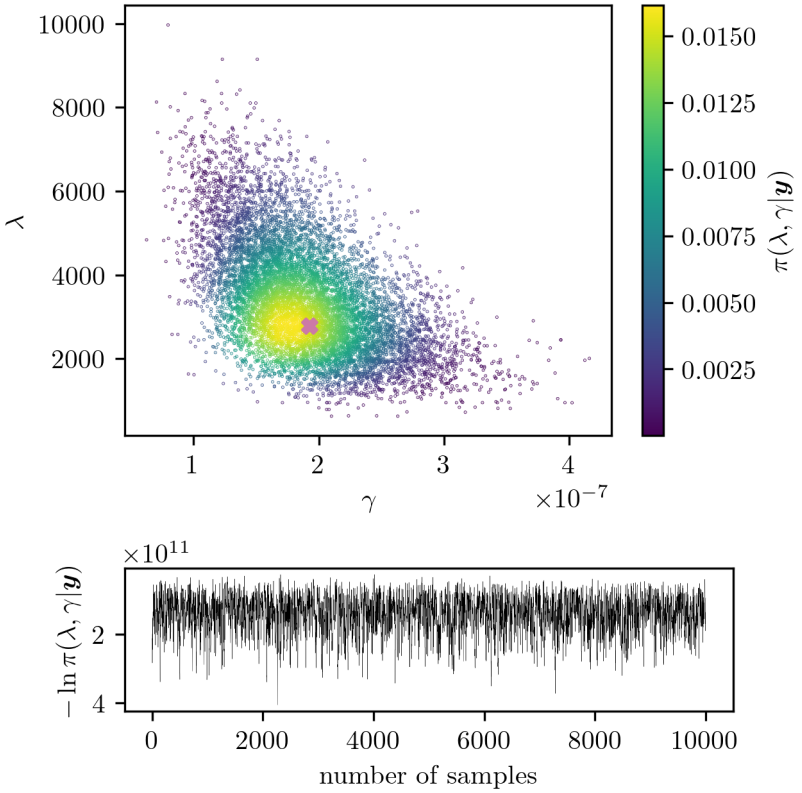

Figure 4.7: We scatter plot the samples of $\lambda = \delta/\gamma$ and γ from the marginal posterior $\pi(\lambda, \gamma|\boldsymbol{y})$ and colour code the samples using the TT approximation of $\pi(\lambda, \gamma|\boldsymbol{y})$. The mode of (λ_0, γ_0) of $\pi(\lambda, \gamma|\boldsymbol{y})$ provided by `scipy.optimize.fmin` is marked with the cross. To show ergodicity we plot the trace of the samples of the Metropolis-within-Gibbs sampler below.

4.3.3 Calculate mean and variance of the conditional posterior for ozone

Based on the marginal posteror distribution $\pi(\gamma, \delta|\boldsymbol{y})$ we calclulate the weighted mean and covariance of the conditional posterior $\pi(\boldsymbol{x}|\gamma, \delta, \boldsymbol{y})$ by quadrature as in Eq. 4.12 and Eq. 4.13.

By binning the samples from the MWG algorithm, see Fig. 4.7, into a normalised histogram with 25 bins we obtain function values for the marginal posterior. With the height of the bars as quadrature weights, e.g. $\pi(\lambda_i|\boldsymbol{y})$, where λ_i is at the centre of each bin we caculate the full conditional mean $\boldsymbol{\mu}_{\boldsymbol{x}|\boldsymbol{y}}$ and covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}|\boldsymbol{y}}$ as weighted expecations.

Alternatively we use the marginal distrinbutiuons $\pi(\delta|\boldsymbol{y})$ and $\pi(\gamma|\boldsymbol{y})$ from the TT approximation of $\sqrt{\pi(\delta, \gamma|\boldsymbol{y})}$ to caculate weighted expecations of $\boldsymbol{\mu}_{\boldsymbol{x}|\boldsymbol{y}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{x}|\boldsymbol{y}}$.

In practise, we have to invert $\boldsymbol{B}_\lambda$ and caclulate $\boldsymbol{x}_\lambda$, see Eq. 4.11 25 times (TT grid size and number of bins). A feasablie method is the Cholesky forward and backward substitution. It takles his takes roughly 1s to compute the mean and variance, which

we plot in Fig. 4.8 including samples. Note that we reject unphysical samples from the conditional posterior with negative ozone values.

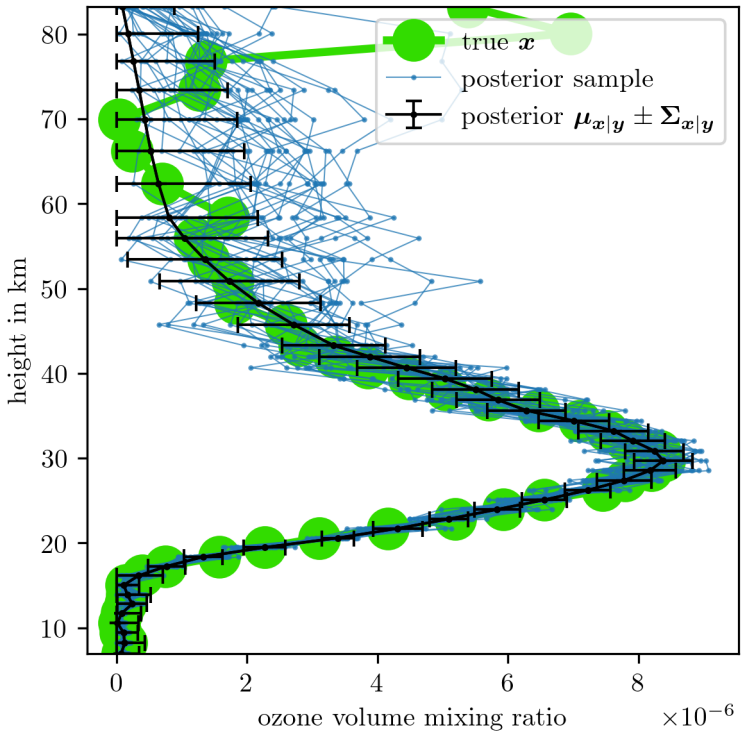

Figure 4.8: We draw samples from the conditional posterior distribution $\pi(\boldsymbol{x}|\lambda, \gamma, \boldsymbol{y})$ after characterising the marginal posterior $\pi(\lambda, \gamma|\boldsymbol{y})$ through sampling or TT approximation using the linear forward map $\boldsymbol{A}_L$. Note that we reject samples with unphysical negative values and effectively treat the conditional posterior as a truncated multivariate normal distribution. We will use those samples to find the affine map $\boldsymbol{M}$, see section 4.3

4.3.4 Asses approximated forward map

Given m samples $\boldsymbol{x}^j \sim \pi\boldsymbol{x}|\boldsymbol{y}$ for $j = 1, \ldots, m$ from the full conditional, as plotted in4.8, we finally are able approximate the non-linear forward model

$$\boldsymbol{A}_{NL} \approx \boldsymbol{M}\boldsymbol{A}_L = \boldsymbol{A}\,, \tag{4.23}$$

with the affine map $\boldsymbol{M}$ and the linear forward model $\boldsymbol{A}_L$. In doing so, we can generate two affine subspaces $\boldsymbol{W} = \{\boldsymbol{A}_L\boldsymbol{x}^1, \ldots, \boldsymbol{A}_L\boldsymbol{x}^m\}$ and $\boldsymbol{V} = \{\boldsymbol{A}_{NL}\boldsymbol{x}^1, \ldots, \boldsymbol{A}_{NL}\boldsymbol{x}^m\}$. Finally we use the `numpy.linalg.solve` python function to solve $\boldsymbol{M}\boldsymbol{W} = \boldsymbol{V}$ for each row of $\boldsymbol{M}$, see Sec. 2.4 for more details.

We asses the affine map using one of the samples $\boldsymbol{x} \sim \pi(\boldsymbol{x}|, \gamma, \lambda, \boldsymbol{y})$ from the conditional posterior and calculate the relative error $||\boldsymbol{M}\boldsymbol{W} - \boldsymbol{V}||/||\boldsymbol{M}\boldsymbol{W}||$ between the mapped noise

free data and the noise free data of non-linear forward model. We display the approximatin for one $\boldsymbol{x}$ sample in Fig. 4.9 we can approximate the non-linear forward model well within the relative difference between the noisy data and noise free non-linear data, which is approximately 1.7%. Consquently, from here on will use the approximated forward map.

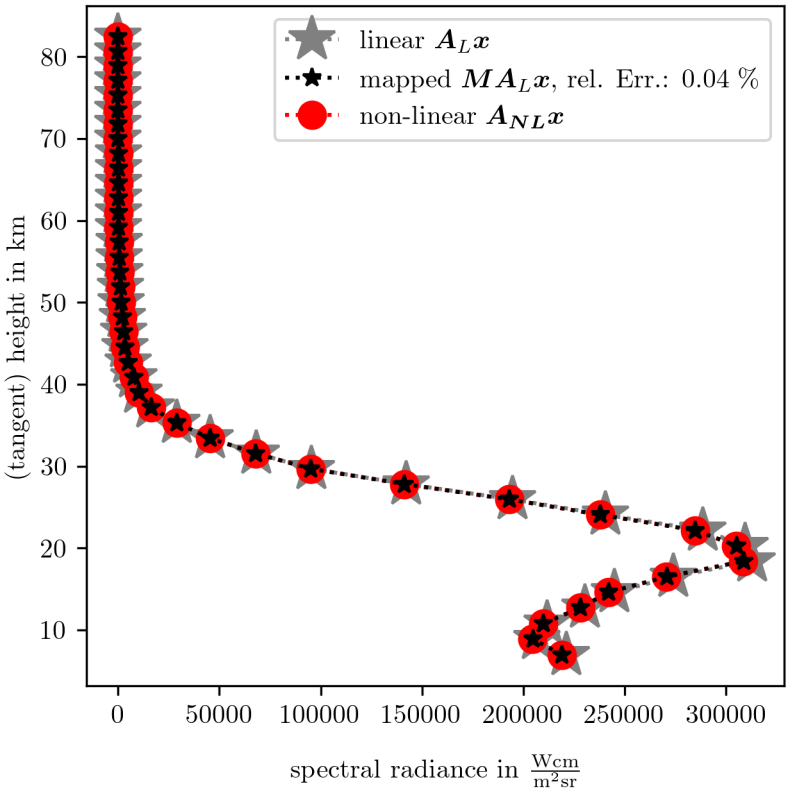

Figure 4.9: We asses how good we can map the linear forward model onto the non-linear forward model using the previous calculated affine map. The gray stars represent noise free linear data, where as the red circles present noise free non-linear data. Then we map the linear noise free data onto the non-linear noise free data and give the relative error in between the mapped noise free data and the non-linear data.

4.4 Solution by regularisation

Since we like to compare the MTC method to regularisation methods, we calculate a solution by Tikhonov regularisation as this is most similar to our chosen linear-Gaussian Bayesian framework. The Tikhonov regularised solution is defined as [36]

$$\boldsymbol{x}_\lambda = \arg\min_{\boldsymbol{x}} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_2^2 + \lambda \boldsymbol{x}^T \boldsymbol{L}\boldsymbol{x}\,, \tag{4.24}$$

with the regularisation parameter $\lambda = \delta/\gamma$. The regularised solution is typically calculated by solving the normal equations, see Sec. 2.5,

$$\boldsymbol{x}_\lambda = (\boldsymbol{A}^T\boldsymbol{A} + \lambda \boldsymbol{L})^{-1}\boldsymbol{A}^T\boldsymbol{y}\,. \tag{4.25}$$

To find the best regularisated solution we use the L-curve method [22]. Within this method we compute $\boldsymbol{x}_\lambda$, for 200 different λ values in between 1 to 10^7 and plot the solution semi norm $\sqrt{\boldsymbol{x}_\lambda^T \mathbf{L} \boldsymbol{x}_\lambda}$ against the data misfit norm $\|\boldsymbol{A}\boldsymbol{x}_\lambda - \boldsymbol{x}\|$, see Figure 4.10. The best regularised solution corresponding to the corner of the L-curve is located at the point of maximum curvature, see triangle in Fig. 4.10, which we find with the kneedle algorithm [37] using the python function `kneed.KneeLocator`. This takes roughly 2 seconds on a MacBook Pro from 2019 with 2.4 Ghz quadcore intel core i5 processor.

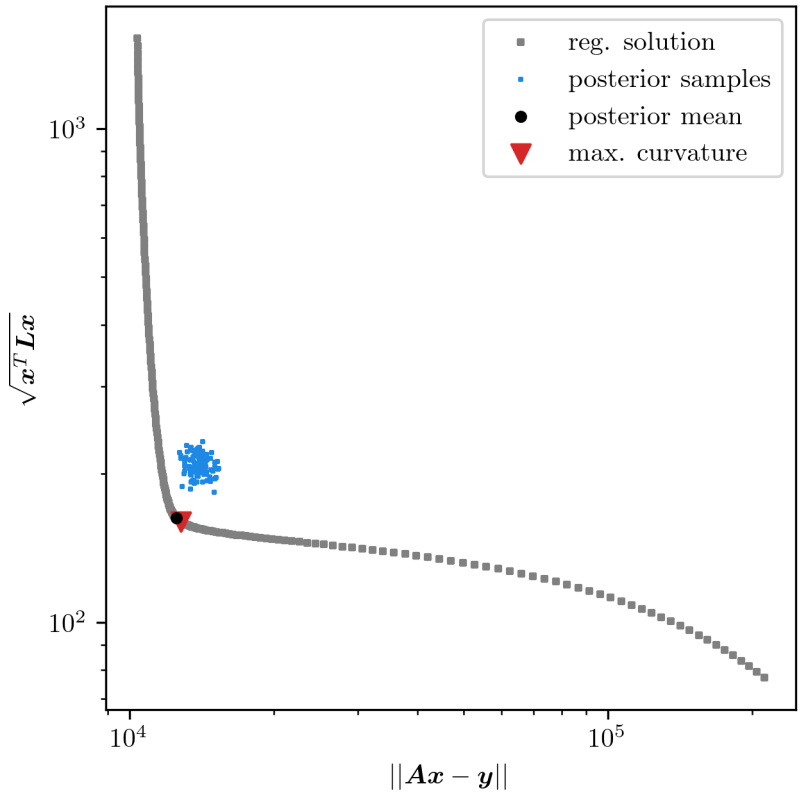

Figure 4.10: We calculate regularised solution as in Eq. **??** and plot the regularised semi norm $\sqrt{\boldsymbol{x}^T \boldsymbol{L} \boldsymbol{x}}$ against the data misfit norm $||\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}||$ to find the regularised solution at the point of maximum curvature of the so-called L-Curve. Additionally we calculate the data misfit norm and the regularised norm for the ozone posterior and for samples of the conditional posterior distribution. make box around Kneedle reagion

4.5 Characterise the posterior distribution of ozone with approximated non-linear model

From here on we use the approximation

$$\boldsymbol{A} = \boldsymbol{M}\boldsymbol{A}_L \tag{4.26}$$

of the non-linear forward map and use the exact same setup as in Sec. **??** and 4.3.3 but with the approximated forward map.

4.5.1 Hyper-parameters samples from and TT approximation of the marginal posterior distribution

The marginal posterior is defined as in Eq. **??** but with $\boldsymbol{A} = \boldsymbol{M}\boldsymbol{A}_L$. We run the MWG algorithm for $N = 20000$ plus $N_{\text{burn-in}} = 100$ and plot the samples in Fig. 4.11 as well as the marginal approximations provided by the TT decomposition, where we use the same setup as in Sec. 4.3.2.

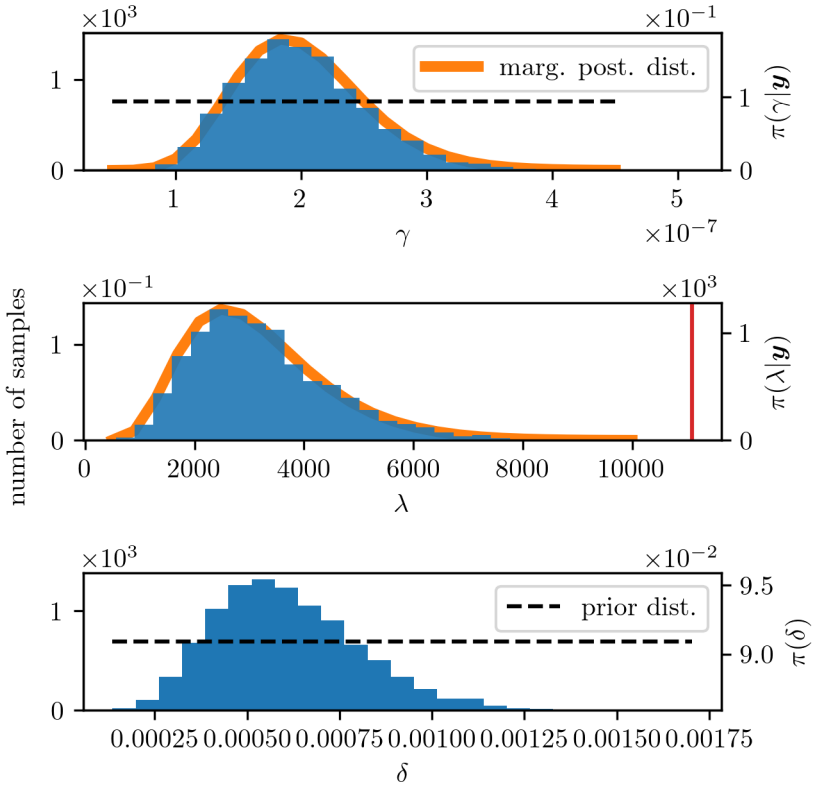

Figure 4.11: We plot the TT approximation of marginal posterior in orange and the samples as a histogram as well as the prior distribution with a dotted line. Note that we sample λ and γ using the Metropolis-within-Gibbs sampler and can calculate δ for every sample of the marginal posterior, we can not do this for the TT approximation. The regularised parameter corresponding to the regularised solution is marked thought the red vertical line at $\lambda_{\text{reg}} =$.

4.5.2 Conditional posterior variance and mean compared to regularised solution

Next, we charcterise the conditional posterior $\pi(\boldsymbol{x}|\gamma, \delta, \boldsymbol{y})$ as in Eq. 4.11. Agian we calculate the full conditional mean 4.12 and full conditinal covarince matrix 4.13 as weighted expextaion over a 25 point grid provided by either the marginal TT approxiamtions or the histogram of samples. We plot the conditional mean and variance in Fig. 4.12 and the regularised solution and one sample from the posterior.

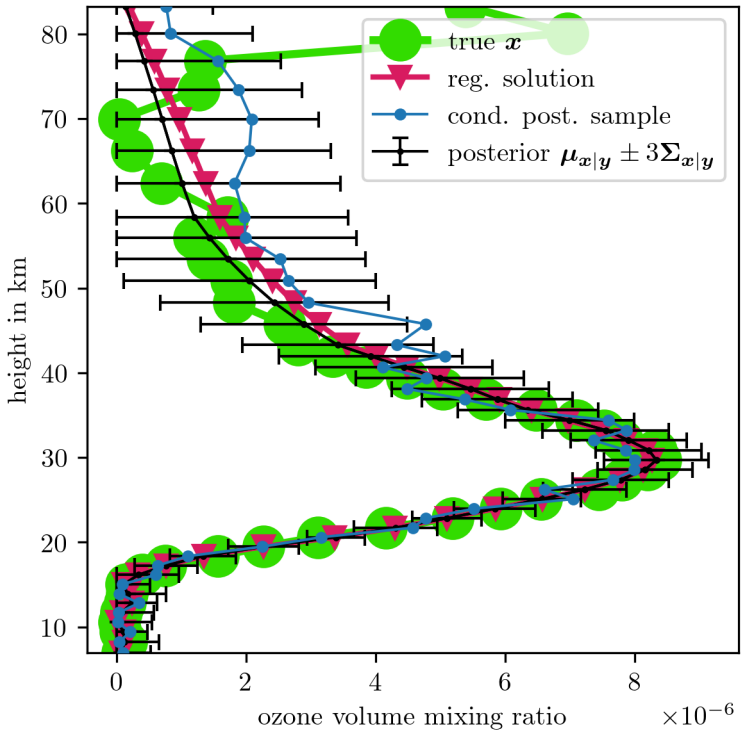

Figure 4.12: We plot the conditional posterior mean and variance in black and the regularised solution on top of the ground truth ozone profile in green. We use the updated forward map $\boldsymbol{MA}_L$

4.6 Posterior distribution for pressure/temperature with approximated non-linear model

The aim now is to characterise the posterior

$$\pi(p_0, b, T_0, \boldsymbol{h_T}, \boldsymbol{a_T}|\boldsymbol{y}, \gamma, \boldsymbol{x}) \propto \exp\Big\{ -\frac{\gamma}{2}\left\|\boldsymbol{y} - \boldsymbol{A}\frac{\boldsymbol{p}}{\boldsymbol{T}}\right\|^2 + \ln \pi(p_0, b, T_0, \boldsymbol{h_T}, \boldsymbol{a_T})\Big\}, \tag{4.27}$$

we condition on the ozone sample in Fig. 4.12, a γ sample from the marginal posterior and use the approximated forward model $\boldsymbol{A} = \boldsymbol{M}\boldsymbol{A}_L$. To validate the approximation we sample from the posterior using the t-walk [10] implementaion in python [38].

Again we define a grid with 25 grid points in each dimesnion, which also act as the sampling space. Since we appproximate a 16 dimensional function we have to carefully choose a grid as we do not want to approximate regions with low probabilty and like to keep the number of grid points low as this increses computaion time. We find the grid, see Tab. 4.2, iteratively by running the t-walk and then computing marginal distriubtions. Note that we bound the sampling space of the t-walk by TT-grid. We run the `tt.cross.rectcross.rect_cross.cross` function from the `ttpy` python package [**<empty citation>**] with constan rank $r = 16$, equal to dimesinon of the posterior. Next, we introduce a constant c in the posterior, which acts as a normalisation constant and is needed when approximating the sqaure root of the posterior for pressure and temperature with a tensor-train (TT) to avoid underflow. Then the posterior becomes:

$$\pi(p_0, b, T_0, \boldsymbol{h_T}, \boldsymbol{a_T}|\boldsymbol{y}, \gamma, \boldsymbol{x}) \propto \exp\Big\{ -\frac{\gamma}{2}\left\|\boldsymbol{y} - \boldsymbol{A}\frac{\boldsymbol{p}}{\boldsymbol{T}}\right\|^2 + \ln \pi(p_0, b, T_0, \boldsymbol{h_T}, \boldsymbol{a_T}) + c\Big\}. \tag{4.28}$$

To find the constant c we evaluate the logarithmic of posterior on 5000 random points and calcualte the maximum $c_{\max} < 0$ of those 5000 points. Then we set the constant to a value which pushes the posterior close to the upper numerical limit of our machine, which is approximately e^{700}. Since we approximate the sqare root, we conseravatively set the constant to $c = -c_{\text{diff}} + 325$. It takes roughly 2min for 15 sweep by the `cross` the to find the optimal tensors. Then we can compute the marginal as in Sec. 2.3, where we set $\xi = 1/\lambda\mathcal{X}$.

For comparison we run the t-walk on the posterior as defined in Eq. 4.27 for 5×10^6 steps plus a burn-in period of 10000, which takes around 7 mins on the same laptop. We plot the resulting histograms in Fig. 4.13 to 4.17, additionally we plot the trace of the samples in Fig. A.5. The integrated autocorrelaton times (IACT) for the hyper-parameters range from 0 to 1000 and are summarized in Tab. 4.2

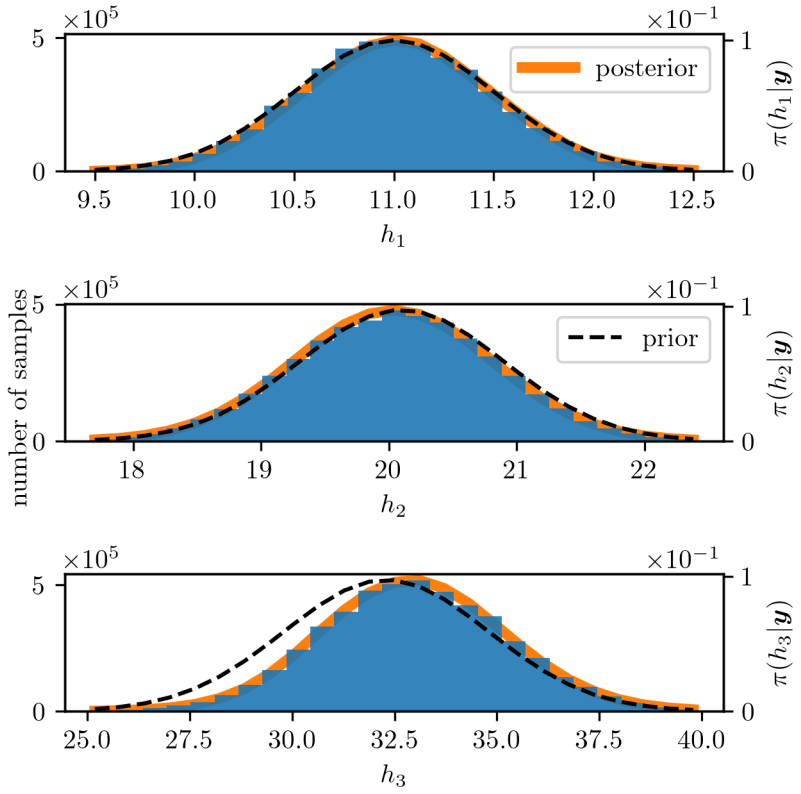

Figure 4.13: We plot the TT approximation of marginal posterior in orange and the samples as a histogram as well as the prior distribution with a dotted line.

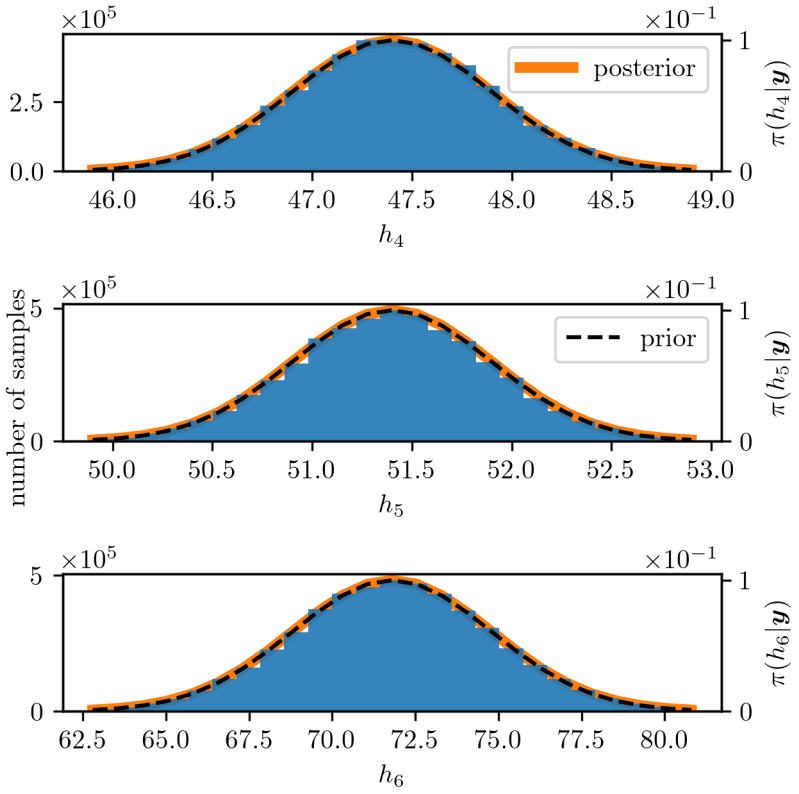

Figure 4.14: We plot the TT approximation of marginal posterior in orange and the samples as a histogram as well as the prior distribution with a dotted line.

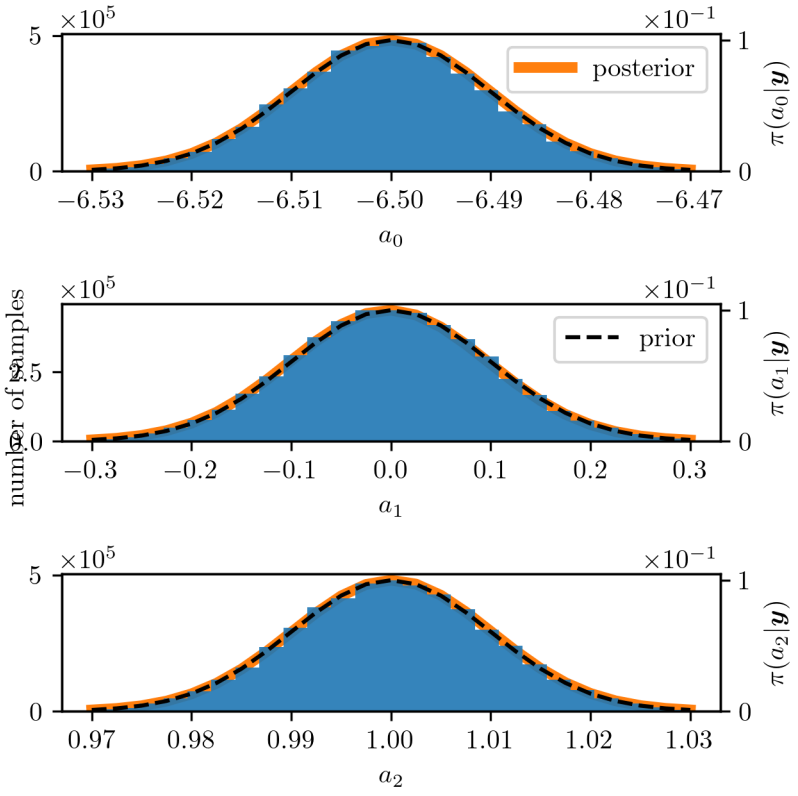

Figure 4.15: We plot the TT approximation of marginal posterior in orange and the samples as a histogram as well as the prior distribution with a dotted line.

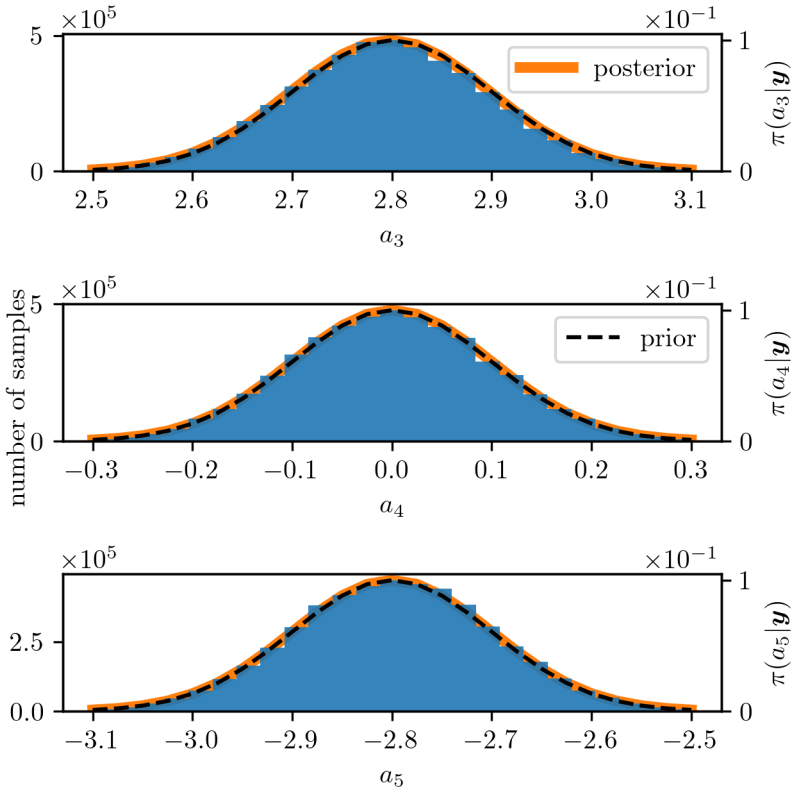

Figure 4.16: We plot the TT approximation of marginal posterior in orange and the samples as a histogram as well as the prior distribution with a dotted line.

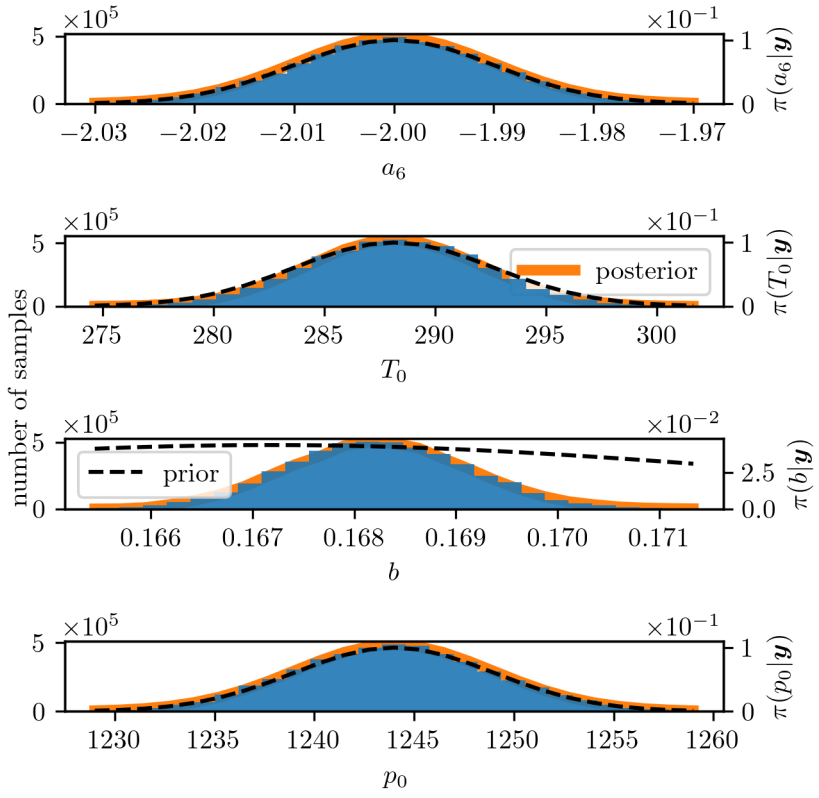

Figure 4.17: We plot the TT approximation of marginal posterior in orange and the samples as a histogram as well as the prior distribution with a dotted line.

To obtain temperature and pressure profiles we can either take samples from the output of the t-walk or by generating random values between 0 and 1 and comparing it to the cumulative distribution functions. We plot the posterior temperatire und pressure profiles in Fig. 4.18 and Fig. 4.19.

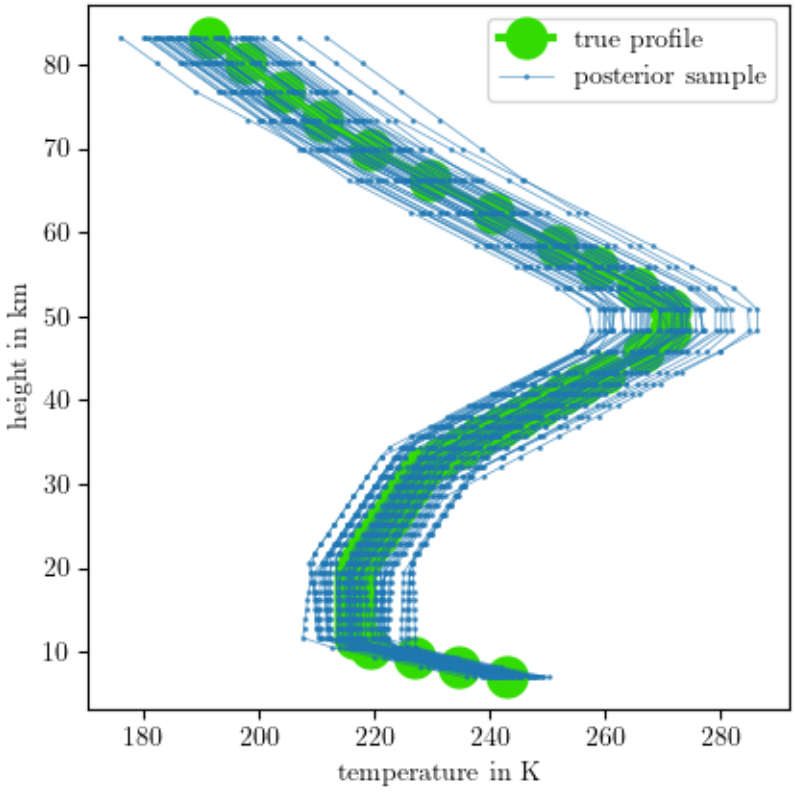

Figure 4.18: We take samples from the posterior distribution, as plotted in Figures 4.13 to 4.16 and plot the corresponding temperature function, see Eq: 4.3.

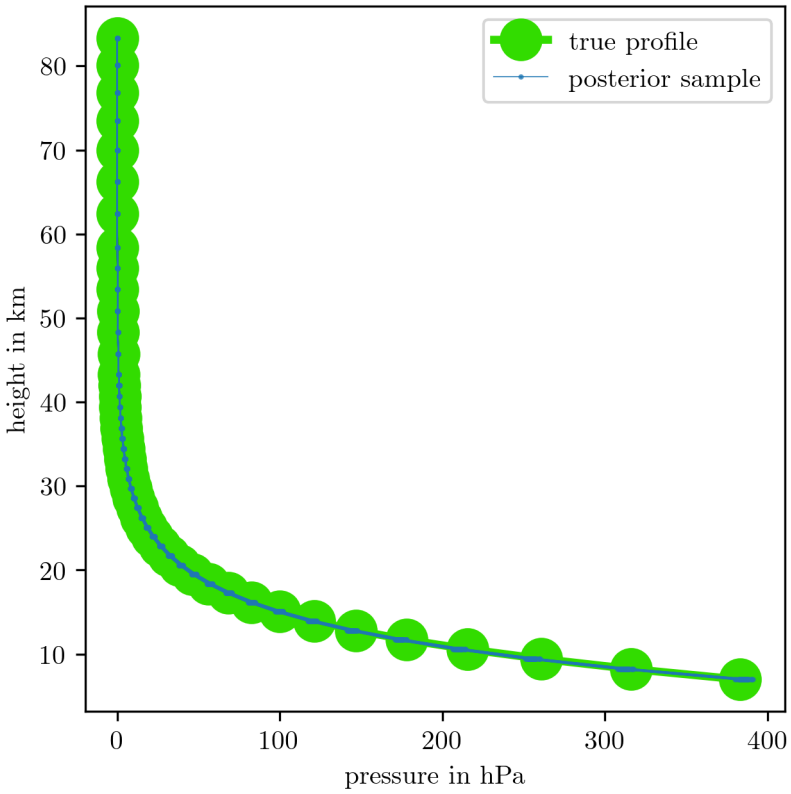

Figure 4.19: We take samples from the posterior distribution, as plotted in Fig. 4.17 and plot the corresponding pressure function, see Eq: 4.16.

4.7 Error analysis

In this section we try to estimate errors due to the function approximations of $f(\lambda)$ and $g(\lambda)$ and how these errors propagate to the marginal posterior. Additionally we discuss errors of the TT-approximation as well as Monte-Carlo errors when binning up the samples.

Error due to approximation of f and g

When approximating the functions $f(\lambda)$ and $g(\lambda)$ we find that the 3rd order Taylor series of $f(\lambda)$ and a linear approximation of $g(\lambda)$ in log-space gives the smallest error. The Taylor series truncation error of $f(\lambda)$ is bounded by the fourth order Taylor series $E_f = \arg\max_{\lambda} f^{(4)}(\lambda_0)/4!\,(\lambda-\lambda_0)^4$ and corresponds to an relative error bounded by 20%. Since the maximum absolute error of the approximation $\arg\max_{\lambda} |\tilde{g}(\lambda) - g(\lambda)| \approx 1$ corresponds to an relative error of approximately 0.3% and is small compared to $E_f \approx 1e8$ we ignore the approximation error of $g(\lambda)$. Then the maximum relative propagation error $\arg\max_{\lambda,\gamma} 0.5\gamma E_f / \log \pi(\lambda, \gamma|\boldsymbol{y})$ is bound by approximately 5%.

Tensor-train approximation error for the marginal posterior

When approximating the marginal posterior the maximum relative propagation error $\arg\max_{\lambda,\gamma} |\tilde{\pi}(\lambda,\gamma|\boldsymbol{y}) - \pi(\lambda,\gamma|\boldsymbol{y})|/|\pi(\lambda,\gamma|\boldsymbol{y})|$ is approximately 100% at $\gamma_{\max}$ and $\lambda_{\max}$, which are the maximum values of the λ and γ samples and lay in regions with very low probability. We consider this error negligible because the absolute error at $\gamma_{\max}$ and $\lambda_{\max}$ is smaller than $10^{-24} \approx 0$.

Note that one can reduce the maximum errors when approximation $f(\lambda)$ at the mean of $\pi(\lambda, \gamma|\boldsymbol{y})$ instead of the modes since $\pi(\lambda|\boldsymbol{y})$ is skewed, but we don't see noticeable differences in the conditional posterior $\pi(\boldsymbol{x}|\lambda, \gamma, \boldsymbol{y})$ when doing so. We consider these errors as tolerable.

Error on the number of sample bins

When we calculate the conditional mean and variance we have to bin up the samples or use a TT approximation on a predefined grid with a certain number of grid points, we like to give an estimate for this error as well. In doing we bin up samples and use the height $\tilde{\pi}(\boldsymbol{\theta}_d^{(k)})$ for a bin $k = 1, \ldots, \mathrm{N}_b$ to calculate the mean $\tilde{\mu}_d = \sum_{\mathrm{N}_b} \tilde{\pi}(\boldsymbol{\theta}_d^{(k)})$. We compare to the sample mean $\boldsymbol{\mu}_d = \sum_{k=1}^{N} \boldsymbol{\theta}_d^{(k)}/N$ and calculate the relative error $||\boldsymbol{\mu}_{\text{samp}} - \boldsymbol{\mu}_{\text{distr}}||/||\boldsymbol{\mu}_{\text{samp}}||$ where $\boldsymbol{\mu}_{\text{samp}} = (\tilde{\mu}_1, \ldots, \tilde{\mu}_D)$ and equivalently $\boldsymbol{\mu}_{\text{distr}} = (\tilde{\mu}_1, \ldots, \tilde{\mu}_D)$. Here d refers to the $D = 16$ hyper-parameters $\gamma, \lambda, h_1, h_2, h_3, h_4, h_5, h_6, a_0, a_1, a_2, a_3, a_4, T_0, p_0, b$.

When plotting the relative error, see Fig. 4.20, we see that it behaves proportional to $1/N$, as in Eq. 2.12 and we consider a relative error less than 0.1% good enough. This

happens roughly at a bin size of 25 and we choose this as our TT grid size. Note that we exclude the error due to τ_{int} the IACT and that we choose the grid according to the sampled values so that the sampling regions is the same as the region we approximate the posterior distributions on.

. we choose the grid size for the tensor-train approximation acordingly and calculte

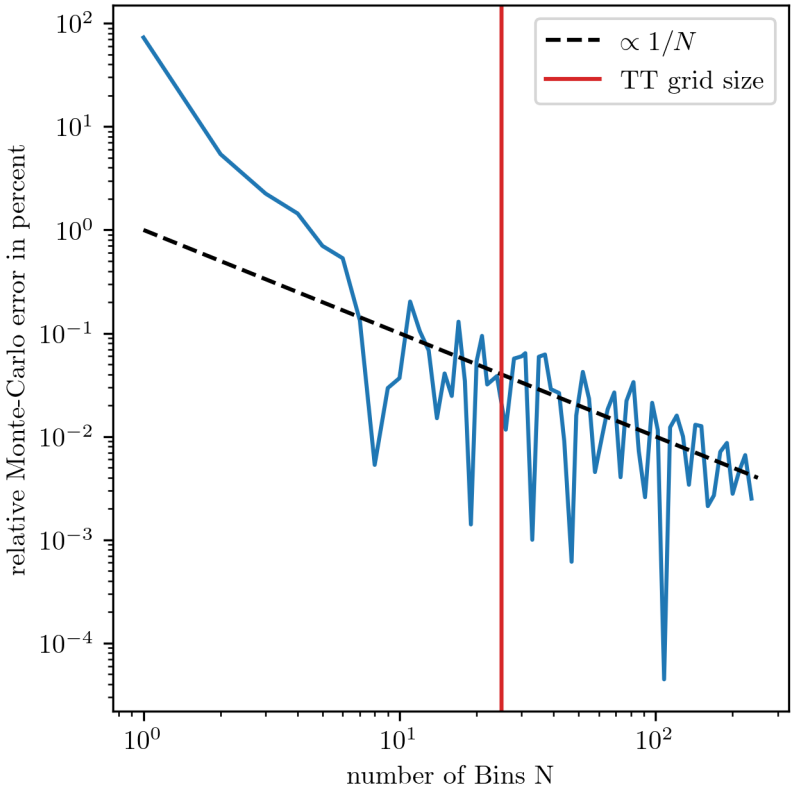

Figure 4.20: Assessment of Monte-Carlo error, where we calculate the relative error of the mean due to binning up the samples compared to the sample mean $||\boldsymbol{\mu}_{\text{samp}} - \boldsymbol{\mu}_{\text{distr}}||/||\boldsymbol{\mu}_{\text{samp}}||$.

the error of the tt approximation with the wasserstein distance.

Error due to tensor-train approximation

5 Conclusions

5.1 Methods

5.1.1 Regularisation vs MTC

- distribution vs one solution
- hierarchial model model lambda
- This maximises the full conditional distribution for $\boldsymbol{x}|\boldsymbol{y}$, so is not, as often erroneously stated, the maximum a posteriori (MAP) estimate which includes at the hyper-parameters.
- Sol Reg vs Mean, vs samples, see L-Curv Fig

5.1.2 Sampling vs TT

- time
- number of function evaluations
- numerical limits (python package)

5.2 Atmospheric Physics

- Data sensitive informative uninformative, Ozone in higher altitude, pressure temp
- truncated mean

6
Summary and Outlook

6.1 Atmospheric Physics

- Data sensitive informative uninformative, Ozone in higher altitude, pressure , temperature
- SNR v s Pointing accuracy from experience
- include pointing accuracy, weighted mean for pointing accuracy
- nadir geometry for higher altitudes citation

6.2 Methods

- graph Laplacian
- calculating the covariance can be expensive and if that is the case the RTO methods is the preferred choice.
- Through exploratory analysis we found that instead of increasing the ranks optimising the tensors by sweeping over them gives better approximations and is faster, which is crucial in higher dimensions as in section
- TT other bases
- speed gridsize intital ranks
- Machine learning or other methods for affine map
- regularised vs posterior

The TT alforith with has nkber of funciton evaluaation set with constant rank r $((D-2)r \times n \times r + 2 \times n \times r)2 \times n_{sweep}$ 400 for TT marg

Appendices

A Additional Figures

A.1 Ozone

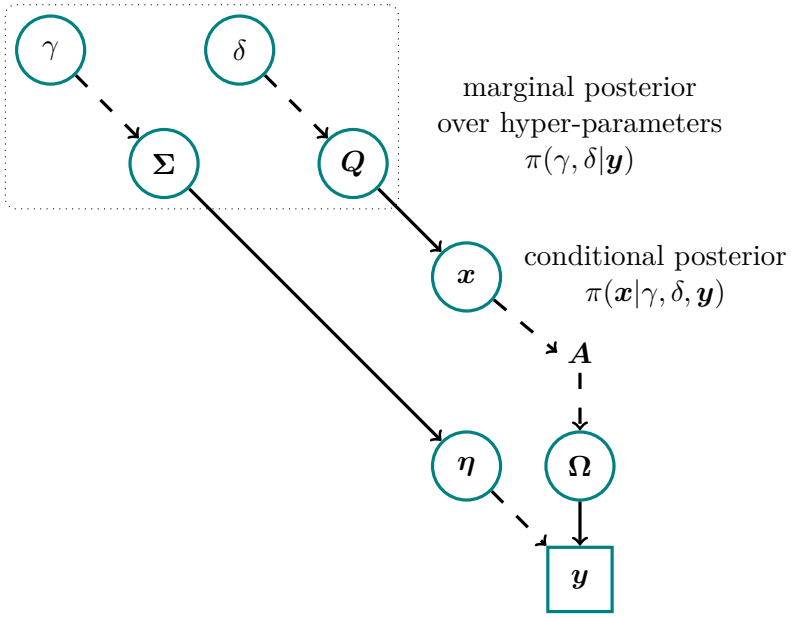

Figure A.1: Directed acyclic graph for modelling and measuring process of ozone highlighting the marginal and then conditional (MTC) scheme. The hyper-parameters δ and γ determine the noise covariance $\boldsymbol{\Sigma} = \gamma^{-1}\boldsymbol{I}$ for the random noise vector $\boldsymbol{\eta} \sim \mathcal{N}(0, \gamma^{-1}\boldsymbol{I})$ and the prior precision matrix $\boldsymbol{Q} = \delta\boldsymbol{L}$ for the normal distribution over $\boldsymbol{x} \sim \mathcal{N}(0, \delta\boldsymbol{L})$, where $\boldsymbol{L}$ is a graph Laplacian, see Eq. 4.6. In the MTC scheme we evaluate the marginal posterior over the hyper-parameters $\pi(\gamma, \delta|\boldsymbol{y})$ as in Eq. **??** first and then the conditional posterior $\pi(\boldsymbol{x}|\gamma, \delta, \boldsymbol{y})$ as in Eq. 4.11. The MTC scheme allows to evaluate the marginal posterior distribution over the hyper-parameters δ, γ independent of $\boldsymbol{x}$, breaking the correlation structure. Through the forward model $\boldsymbol{A}_{NL} \approx \boldsymbol{M}\boldsymbol{A}_L$ and the parameter $\boldsymbol{x}$ we generate a space of all measurable from which we randomly observe a data set $\boldsymbol{y}$ including random noise $\boldsymbol{\eta}$.

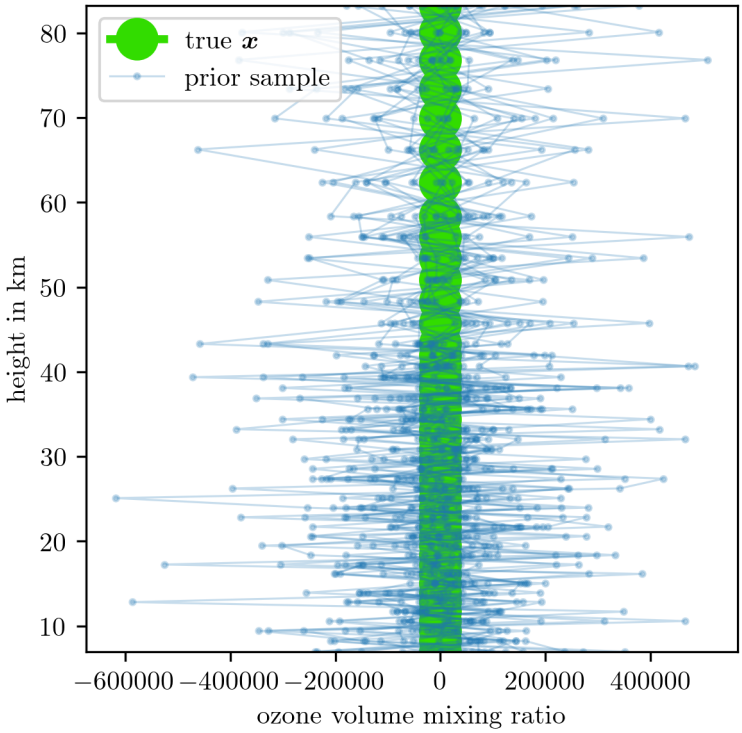

Figure A.2: We draw samples from ozone prior distribution $\boldsymbol{x} \sim \mathcal{N}(0, \delta \boldsymbol{L})$ after generating a sample from the hyper-prior distribution $\delta \sim \mathcal{T}(1, 10^{-10})$. Note that since the variance of prior samples is very large compared to the ozone volume mixing ratios, the ozone profile appears to be constant, which is not the case, see e.g. Fig. 4.8.

A.2 Pressure over Temperature

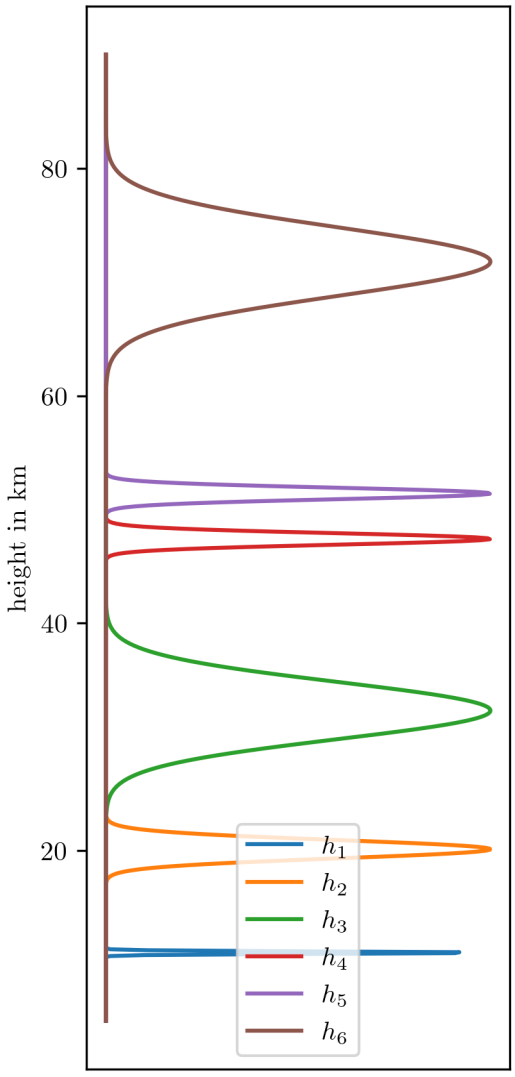

Figure A.3: Prior distributions $\pi(h_T)$, which we choose so that they do not overlap and not conflict with the temperature function 4.3

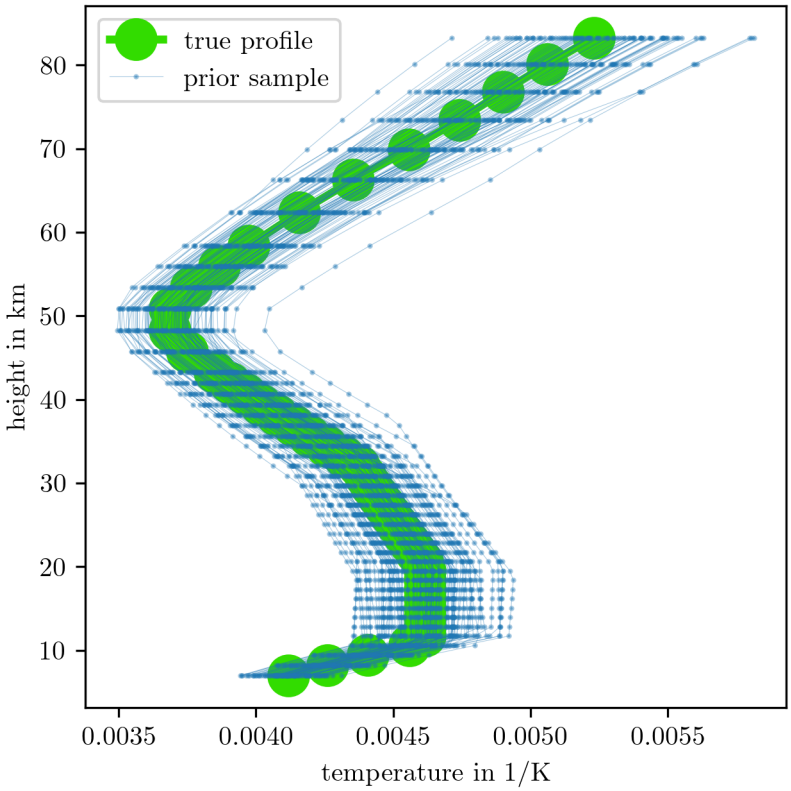

Figure A.4: Prior samples of the inverted temperature profile.

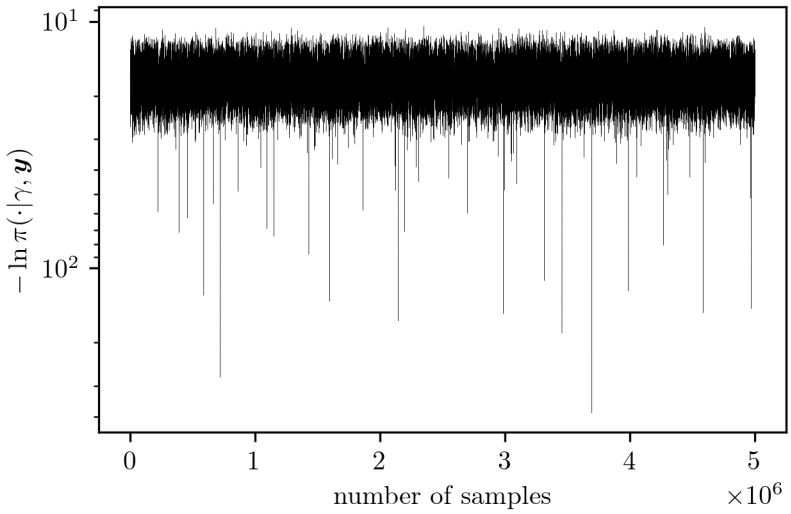

Figure A.5: Output trace of the t-walk on the posterior distribution $\pi(p_0, b, \boldsymbol{h_T}, \boldsymbol{a}|\gamma, \boldsymbol{y})$.

B
Correlation Structure

In the book Gaussian Markov Random Fields [4], Rue and Held demonstrate that a strong correlation between the hyper-parameter μ and the latent field $\boldsymbol{x}$ can significantly slow down convergence when using samplers, in particularly Gibbs samplers. They consider the hierarchical model

$$\mu \sim \mathcal{N}(0, 1) \tag{B.1a}$$
$$\boldsymbol{x}|\mu \sim \mathcal{N}(\mu\mathbf{1}, \boldsymbol{Q}^{-1}), \tag{B.1b}$$

and apply a Gibbs sampler based on the full conditional distributions

$$\mu^{(k)}|\boldsymbol{x}^{(k)} \sim \mathcal{N}\left(\frac{\mathbf{1}^T\boldsymbol{Q}\boldsymbol{x}^{(k-1)}}{1+\mathbf{1}^T\boldsymbol{Q}\mathbf{1}}, \left(1+\mathbf{1}^T\boldsymbol{Q}\mathbf{1}\right)^{-1}\right) \tag{B.2}$$
$$\boldsymbol{x}^{(k)}|\mu^{(k)} \sim \mathcal{N}(\mu^{(k)}\mathbf{1}, \boldsymbol{Q}^{-1}). \tag{B.3}$$

As illustrated in Figure B.1, when the sampler is restricted to steps only in the μ-direction (horizontal axis) or the $\boldsymbol{x}$-direction (vertical axis), it requires many iterations to adequately explore the parameter space. This inefficiency arises from the high correlation between μ and $\boldsymbol{x}$, visible in Figure B.1 as a 'squeeze' of the distribution.

A solution to the slow mixing problem is to update $(\mu, \boldsymbol{x})$ jointly. Since here μ is one dimensional, effectively only marginal density of μ is needed.

$$\mu^\star \sim q(\mu^\star|\mu^{(k-1)}) \tag{B.4}$$
$$\boldsymbol{x}^{(k)}|\mu^\star \sim \mathcal{N}(\mu^\star\mathbf{1}, \boldsymbol{Q}^{-1}) \tag{B.5}$$

With a simple MCMC algorithm targeting μ one can explore the sample space efficiently and only draw a corresponding sample for $\boldsymbol{x}$ from its full conditional once, for instance, the proposal $\mu^\star$ has been accepted.

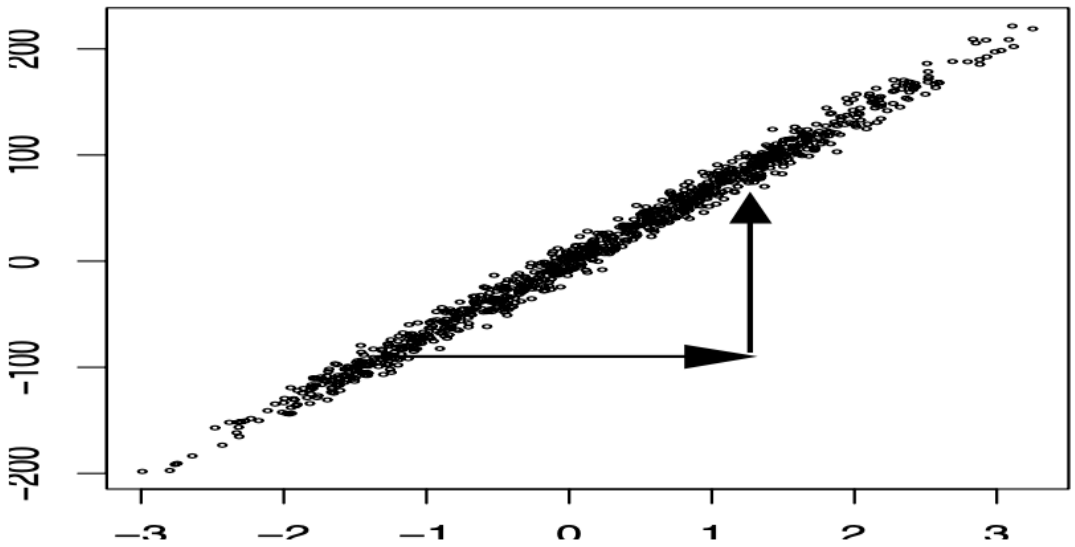

Figure B.1: The figure taken from [4, Figure 4.1 (b)], shows samples from a marginal chain for μ and $\mathbf{1}^T\boldsymbol{Q}\boldsymbol{x}^{(k)}$ over 1000 iterations, based on the hierarchical model in Eq. B.1, with an autoregressive process encoded in $\boldsymbol{Q}$. The algorithm updates μ and $\boldsymbol{x}$ successively from their full conditional distributions. The plot displays $(\mu^{(k)}, \mathbf{1}^T\boldsymbol{Q}\boldsymbol{x}^{(k)})$, with $\mu^{(k)}$ on the horizontal axis and $\mathbf{1}^T\boldsymbol{Q}\boldsymbol{x}^{(k)}$ on the vertical axis. The slow mixing and convergence of μ result from its strong dependence on $\mathbf{1}^T\boldsymbol{Q}\boldsymbol{x}^{(k)}$, while the sampler permits only axis-aligned (horizontal and vertical) and does not allow diagonal moves, as illustrated by the arrows.

C
Mesure theroy

Recall the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where Ω denotes the sample space, and $\mathcal{F}$ is a collection of countable subsets $\{A_n\}_{n\in\mathbb{N}}$ of Ω. Each $A_n \subseteq \Omega$ is called an event, and a map $\mathbb{P} : \mathcal{F} \to \mathbb{R}$ is referred to as a measure. In the following, we describe the conditions required for $\mathcal{F}$ to be a σ-algebra, and for $\mathbb{P}$ to qualify as a probability measure. We refer to [39] [12] for further reading.

C.1 probailty measure

For a probability measure, we require:

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

In other words, the probability assigned to the entire sample space must be equal to one, $\mathbb{P}(\Omega) = 1$, and the probability of the empty set must be zero, $\mathbb{P}(\emptyset) = 0$. For any subset $A \subseteq \Omega$, the probability $\mathbb{P}(A)$ must lie between zero and one, i.e., $\mathbb{P}(A) \in [0, 1]$. If e.g. two subsets A and B are disjoint (i.e., $A \cap B = \emptyset$), then the probability of their union satisfies $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$. This property must also hold for a countable sequence of disjoint sets $\{A_j\}_{j\in\mathbb{N}}$, such that $\mathbb{P}\left(\bigcup_{j\in\mathbb{N}} A_j\right) = \sum_{j\in\mathbb{N}} \mathbb{P}(A_j)$.

C.2 σ-algebra

A collections of subsets $\mathcal{F}$ is called σ-algebra if:

- $\emptyset, \Omega \in \mathcal{F}$,
- if $A \in \mathcal{F}$ then $A^C \coloneqq A/\Omega \in \mathcal{F}$
- 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 entire sample space Ω must always be elements of $\mathcal{F}$. If a set $A \in \mathcal{F}$, then its complement $A^C = \Omega \setminus A$ must also be in $\mathcal{F}$. If, in terms of a probability measure, we are able to assign a probability $\mathbb{P}(A)$ to an event A, we must also be able to assign a probability to the event "not A", i.e., $\mathbb{P}(A^C)$. Finally, if a countable collection of sets $A_1, A_2, \cdots \in \mathcal{F}$, then their union $\bigcup_{j\in\mathbb{N}} A_j$ must also be in $\mathcal{F}$. These three properties define the requirements for $\mathcal{F}$ to be a σ-algebra.

References

[1] 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.

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

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

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

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

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

[7] A. Sokal. "Monte Carlo Methods in Statistical Mechanics: Foundations and New Algorithms". In: *Functional Integration: Basics and Applications*. Ed. by Cecile DeWitt-Morette, Pierre Cartier, and Antoine Folacci. Boston, MA: Springer US, 1997, pp. 131–192.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

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

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

[14] Vesa Kaarnioja. *Inverse Problems. Eighth lecture*. `https://vesak90.userpage.fu-berlin.de/ip23/week8.pdf`. [Online; accessed 10/04/25]. 2023.

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

[16] 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.

[17] 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.

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

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

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

[21] Per Christian Hansen. “The L-Curve and its Use in the Numerical Treatment of Inverse Problems”. English. In: *Computational Inverse Problems in Electrocardiology*. Ed. by P. Johnston. WIT Press, 2001, pp. 119–142.

[22] 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.

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

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

[25] Marie Šimečková et al. “Einstein A-coefficients and statistical weights for molecular absorption transitions in the HITRAN database”. In: *Journal of Quantitative Spectroscopy and Radiative Transfer* 98.1 (2006), pp. 130–155.

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

[27] Schwartz M. et al. *MLS/Aura Level 2 Ozone (O3) Mixing Ratio V005*. `https://disc.gsfc.nasa.gov/datasets/ML2O3_005/summary?keywords=mls%20o3`. [Online; accessed 25/04/24]. 2020.

[28] *U.S. Standard Atmosphere, 1976*. Washington, D.C.: United States. National Oceanic and Atmospheric Administration, United States Committee on Extension to the Standard Atmosphere, 1976.

[29] Australian National Concurrent Design Facility. *CubeSat Microwave Radiometer Mission to Support Global Ozone Layer Monitoring. Concept Study - Summary Report*. unpublished, internal report. Canberra BC: UNSW Canberra Space, 2023.

[30] H.M. Pickett. “Microwave Limb Sounder THz module on Aura”. In: *IEEE Transactions on Geoscience and Remote Sensing* 44.5 (2006), pp. 1122–1130.

[31] Yu-Xiang Wang et al. “Trend Filtering on Graphs”. In: *Journal of Machine Learning Research* 17.105 (2016), pp. 1–41.

[32] Daniel Simpson, Finn Lindgren, and Håvard Rue. “Think continuous: Markovian Gaussian models in spatial statistics”. In: *Spatial Statistics* 1 (2012), pp. 16–29. URL: `https://www.sciencedirect.com/science/article/pii/S2211675312000048`.

[33] Johnathan M Bardsley et al. “Randomize-then-optimize for sampling and uncertainty quantification in electrical impedance tomography”. In: *SIAM/ASA Journal on Uncertainty Quantification* 3.1 (2015), pp. 1136–1158.

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

[35] Josef Dick, Frances Y. Kuo, and Ian H. Sloan. “High-dimensional integration: The quasi-Monte Carlo way”. In: *Acta Numerica* 22 (2013), 133–288.

[36] Per Christian Hansen. *Discrete Inverse Problems: Insight and Algorithms*. Philadelphia: SIAM, 2010.

[37] Ville Satopää et al. “Finding a "Kneedle" in a Haystack: Detecting Knee Points in System Behavior”. In: *2011 31st International Conference on Distributed Computing Systems Workshops*. IEEE. 2011, pp. 166–171.

[38] J. Andrés Christen and Colin Fox. *The t-walk software.* `https://www.cimat.mx/~jac/twalk/`. [Online; accessed 25/11/24].

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