

# Contents

<b>List of Figures</b>	<b>iii</b>
<b>1 Introduction</b>	<b>3</b>
1.1 Motivation . . . . .	3
1.2 What is going on?, 3 facts, What is new in this thesis? . . . . .	3
1.3 Thesis Outline . . . . .	3
<b>2 Forward Model</b>	<b>5</b>
<b>3 Theoretical and Technical Background</b>	<b>9</b>
3.1 Affine Map . . . . .	9
3.2 Bayesian Inference . . . . .	10
3.2.1 Marginal and conditional posterior distribution . . . . .	13
3.3 Regularisation . . . . .	13
3.4 Sampling Methods . . . . .	14
3.4.1 Metropolis- within Gibbs sampling . . . . .	15
3.4.2 Draw a sample from a multivariate normal distribution . . . . .	16
3.4.3 t-walk sampler as black box . . . . .	16
3.5 Numerical Approxiamtion Methods - Tensor Train . . . . .	17
3.5.1 Marginal Functions . . . . .	20
<b>4 Results</b>	<b>23</b>
4.1 Simulate Data . . . . .	23
4.2 Set up Bayesian model . . . . .	24
4.2.1 DAG . . . . .	25
4.2.2 Prior Modelling . . . . .	25
4.3 Posterior distributions with Linear model for Ozone – MTC . . . . .	31
4.3.1 Hyper-parameters samples from the marginal posterior distribution	32
4.3.2 Ozone samples from the conditional posterior . . . . .	34
4.4 Approximate non-linear forward model with affine Map . . . . .	36
4.5 Posterior distributions with approximated non-linear model for Ozone – MTC . . . . .	38
4.5.1 Hyper-parameters samples from the marginal posterior distribution	38

4.5.2	Ozone samples from the conditional posterior and regularised solution	39
4.5.3	Regularized Solution . . . . .	39
4.6	Solution by regularization . . . . .	40
4.6.1	Regularization . . . . .	41
4.7	Posterior Pressure and Temperature . . . . .	41

## Appendices

<b>A</b>	<b>Additional MCMC analysis</b>	<b>53</b>
<b>B</b>	<b>Correlation Structure</b>	<b>55</b>
<b>C</b>	<b>Mesure theroy</b>	<b>57</b>
C.1	probailty measure . . . . .	57
C.2	$\sigma$ -algebra . . . . .	58
<b>D</b>	<b>prior modelling</b>	<b>59</b>
	<b>References</b>	<b>65</b>

# List of Figures

2.1	Schematic of measurement and analysis geometry.	5
3.1	Schematics of the affine map	10
3.2	Bayesian Inference DAG	11
3.3	Visualisation of a tensor train	18
4.1	Complete directed acyclic graph of the forward model.	25
4.2	Samples from ozone prior distribution.	27
4.3	Prior Samples of $\mathbf{p}/\mathbf{T}$ according to the respective hyper-prior distribution.	28
4.4	Prior Samples of $\mathbf{T}$ according to the respective hyper-prior distribution.	29
4.5	Prior Samples of $\mathbf{p}$ according to the respective hyper-prior distribution.	30
4.6	Directed acyclic graph for ozone retrieval and MTC scheme.	31
4.7	Plot of the functions $f(\lambda)$ and $g(\lambda)$ for marginal posterior.	33
4.8	Scatter plot of samples from marginal posterior, including weighting from TT approximation; additional trace plot of the marginal posterior samples.	35
4.9	Ozone samples of the conditional posterior.	37
4.10	Strategy to find affine map.	37
4.11	Assessment of affine map.	39
4.12	Marginal posterior histograms and TT approximation as well as hyper-prior distribution.	40
4.13	Plot of the L-curve to find the regularised solution.	41
4.14	Ozone posterior mean and variance and the regularised solution compared to the ground truth.	42
4.15	Directed acyclic Graph for pressure and temperature.	43
4.16	Histograms and TT approximation of posterior distribution as well as hyper-prior distribution.	43
4.17	Histograms and TT approximation of posterior distribution as well as hyper-prior distribution.	44
4.18	Histograms and TT approximation of posterior distribution as well as hyper-prior distribution.	45
4.19	Histograms and TT approximation of posterior distribution as well as hyper-prior distribution.	46

4.20 Histograms and TT approximation of posterior distribution as well as hyper-prior distribution. . . . .	47
4.21 Temperature posterior samples. . . . .	48
4.22 Pressure posterior samples. . . . .	49
B.1 Correlation structure in between parameters and hyper-parameters . . . . .	56

421.10046pt



# 1

## Introduction

### 1.1 Motivation

- ozone coverage
- regularisation approach in atmospheric physics
- 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]

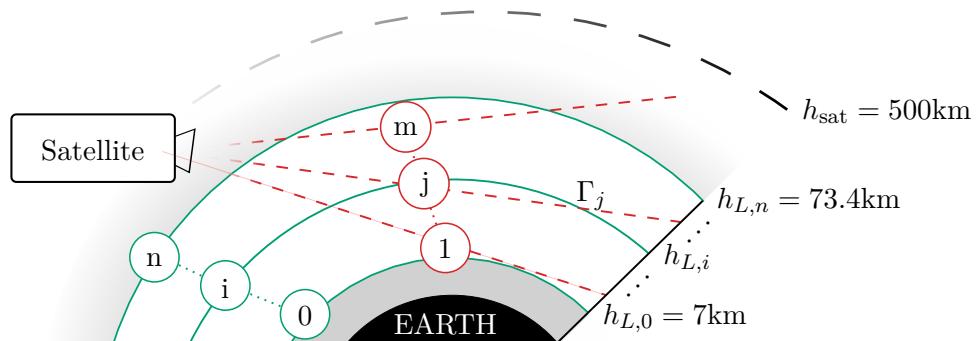


# 2

## Forward Model

In this chapter we present the forward model we apply all our methodology on. We follow the MIPAS handbook [2] and simulate data according to a atmosphere in local thermodynamic and and assume a measurement instrument with infinite spectral resolution and no pointing errors.

The forward model is based on a satellite measuring thermal radiation of gas molecules along its line of sight by pointing through the atmosphere to the edge(limb) of the atmosphere, known as limb sounding, as shown in Figure 2.1. One measurement  $y_j$ , of a stationary satellite is given by the path integral through the atmosphere along the line of sight. For each measurement  $j = 1, 2, \dots, m$  of a data set, we can define a tangent height  $h_{\ell_j}$  as the shortest distance along the line of sight to the earth.



**Figure 2.1:** Schematic of measurement and analysis geometry, not to scale. The stationary satellite, at a constant height  $h_{\text{sat}}$  above Earth, takes  $m = 41$  measurements along its line-of-sight defining by the line  $\Gamma_j$ . Each measurement has a limb height  $\ell_j$ ,  $j = 1, 2, \dots, m$  defined as the closest distance of  $\Gamma_j$  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.

Targeting the thermal radiation at one wave number  $\nu$  of one specific molecule, the  $j^{\text{th}}$  measurement, is modelled by the radiative transfer equation (RTE) [2]

$$y_j = \int_{\Gamma_j} B(\nu, T) k(\nu, T) \frac{p(T)}{k_B T(r)} x(r) \tau(r) dr + \eta_j \quad (2.1)$$

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

where the path from the satellite along the line-of-sight of the  $j^{\text{th}}$  pointing direction is  $\Gamma_j$  and the ozone concentration at distance  $r$  from the radiometer is  $x(r)$  plus some noise  $\eta_j$ . Within the atmosphere the number density  $p(T)/(k_B T(r))$  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

$$k(\nu, T) = L(\nu, T_{\text{ref}}) \frac{Q(T_{\text{ref}})}{Q(T)} \frac{\exp \{-c_2 E''/T\}}{\exp \{-c_2 E''/T_{\text{ref}}\}} \frac{1 - \exp \{-c_2 \nu/T\}}{1 - \exp \{-c_2 \nu/T_{\text{ref}}\}} \quad (2.3)$$

is depend on the line intensity  $L(\nu, T_{\text{ref}})$  at reference temperature  $T_{\text{ref}} = 296K$ , the lower-state energy of the transition  $E''$ , the second radiation constant  $c_2 = 1.4387769 \text{ cmK}$  all provided by the HITRAN database [3]. The total internal partition function for the lower-state energy is

$$Q(T) = g'' \exp \left\{ -\frac{c_2 E''}{T} \right\}, \quad (2.4)$$

with the statistical weight  $g''$  (also called the degeneracy factor) accounting for the molecules non-rotational and rotational energy states, see [4]. Under the assumption of local thermodynamic equilibrium (LTE) the black body radiation act as a source function

$$B(\nu, T) = \frac{2hc^2\nu^3}{\exp \left\{ \frac{hc\nu}{k_B T} \right\} - 1}, \quad (2.5)$$

with Planck's constant  $h$  and velocity of light  $c$  [ ]. For fundamentals on the Radiative transfer equation we recommend [5, Chapter 1].

To enable matrix-vector multiplication, we discretise the atmosphere in  $n$  layers, where the  $i^{\text{th}}$  layer is defined by two spheres of radii  $h_{L,i-1} < h_{L,i}$ , for  $i = 1, \dots, n$ , with  $h_0$  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  $\mathbf{x} = \{x_1, x_2, \dots, x_n\} \in \mathbb{R}^n$  or the fraction of pressure and temperature  $\mathbf{p}/\mathbf{T} = \{p_1/T_1, p_2/T_2, \dots, p_n/T_n\} \in \mathbb{R}^n$ , we solve the integral

in Eq. (2.1) using the trapezoidal rule so that we can rewrite the integral to a vector multiplication  $\mathbf{A}_j(\mathbf{x}, \mathbf{p}, \mathbf{T}) \mathbf{x}$  or  $\mathbf{A}_j(\mathbf{x}, \mathbf{p}, \mathbf{T}) \mathbf{p}/\mathbf{T}$ , where the non-linear absorption  $\tau(r)$  is included in  $\mathbf{A}_j(\mathbf{x}, \mathbf{p}, \mathbf{T})$ . Here, the row vector  $\mathbf{A}_j(\mathbf{x}, \mathbf{p}, \mathbf{T}) \in \mathbb{R}^n$  defines a Kernel for each measurement so that the data vector

$$\mathbf{y} = \mathbf{A}(\mathbf{x}, \mathbf{p}, \mathbf{T}) \mathbf{x} + \boldsymbol{\eta} = \mathbf{A}(\mathbf{x}, \mathbf{p}, \mathbf{T}) \frac{\mathbf{p}}{\mathbf{T}} + \boldsymbol{\eta}. \quad (2.6)$$

can be written as a matrix-vector multiplication, with the matrix  $\mathbf{A}(\mathbf{x}, \mathbf{p}, \mathbf{T}) \in \mathbb{R}^{m \times n}$  and the noise vector  $\boldsymbol{\eta} \in \mathbb{R}^m$ . Note, for simplicity, we do not explicitly specify whether  $\mathbf{A}(\mathbf{x}, \mathbf{p}, \mathbf{T})$  is constructed conditioned on  $\mathbf{x}$  or  $\mathbf{p}/\mathbf{T}$  and we ignore temperature dependencies of the pressure, absorption constant or the black body radiation.

Since the measurement process includes absorption  $\tau(r)$ , which reduces measurements only slightly, we classify the inverse problem as weakly non-linear. Hence, we can approximate the non-linear forward model  $\mathbf{A}(\mathbf{x}, \mathbf{p}, \mathbf{T})$  with a map  $\mathbf{M}$  and the linear forward model  $\mathbf{A}_L$ , so that  $\mathbf{A}(\mathbf{x}, \mathbf{p}, \mathbf{T}) \approx \mathbf{M}\mathbf{A}_L$ . Here,  $\mathbf{A}_{L,j}$  of matrix  $\mathbf{A}_L \in \mathbb{R}^{m \times n}$  is defined by the linear forward model, where absorption is neglected, e.g. set  $\tau = 1$  in Eq. (2.2). Then each entry in the row vector  $\mathbf{A}_{L,j}$  is either defined by  $B(\nu)k(\nu)\frac{\mathbf{p}}{k_B T} dr$  or  $B(\nu)k(\nu)\frac{\mathbf{x}}{k_B} dr$ , as in Eq. (2.1). This poses a linear inverse problem with the forward map defined by the matrix  $\mathbf{A} = \mathbf{M}\mathbf{A}_L$ , where  $\mathbf{M}$  is, more specifically, an affine map.

421.10046pt

# 3

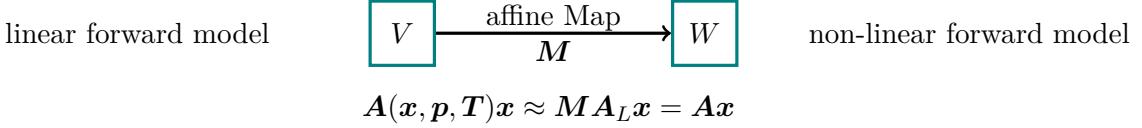
## Theoretical and Technical Background

In this chapter, we provide a brief introduction to the methods used in this thesis. We keep it as general as possible, as more specific details will be presented in the results Chapter 4. We begin by introducing the forward model in Section ??, which we use to simulate the data. Since the forward model is weakly non-linear, we employ an affine transformation, see Section 3.1, to project the linear model onto the non-linear one, allowing us to treat the problem as a linear inverse problem. This enables the application of Bayesian inference in Section 3.2, where we formulate a hierarchical linear-Gaussian model to define and structure the posterior distribution. For comparison, we briefly present the Tikhonov regularization approach, see Section 3.3. In Section 3.4, we introduce Markov Chain Monte Carlo (MCMC) methods to sample from the posterior distribution. Finally, in Section 3.5, instead of sampling, we can approximate the posterior distribution using the tensor train (TT) format.

### 3.1 Affine Map

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 [6, 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 [6, 7]

Consequently, we introduce an affine map  $\mathbf{M} : \mathbf{A}_L \mathbf{x} \rightarrow \mathbf{A}(\mathbf{x}, \mathbf{p}, \mathbf{T}) \mathbf{x}$ , which maps the linear forward model  $\mathbf{A}_L \mathbf{x}$  onto the non-linear forward model  $\mathbf{A}(\mathbf{x}, \mathbf{p}, \mathbf{T}) \mathbf{x}$ . Then the non-linear forward model matrix is approximated by  $\mathbf{A}(\mathbf{x}, \mathbf{p}, \mathbf{T}) \approx \mathbf{M} \mathbf{A}_L$ . In practise



**Figure 3.1:** This Figure shows the schematic representation of the affine map  $\mathbf{M}$ , 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  $\mathbf{M}$  projects elements from the linear forward model space  $V$  onto their counterparts in the non-linear forward model space  $W$ .

we generate two affine subspaces spaces  $V = \{\mathbf{A}(\mathbf{x}^{(1)}, \mathbf{p}, \mathbf{T}), \dots, \mathbf{A}(\mathbf{x}^{(m)}, \mathbf{p}, \mathbf{T})\}$  and  $W = \{\mathbf{A}_L\mathbf{x}^{(1)}, \dots, \mathbf{A}_L\mathbf{x}^{(m)}\}$  over the same field, with fixed  $\mathbf{p}, \mathbf{T}$  and find the mapping in between those. Here, the parameter  $\mathbf{x}$  is distributed as  $\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(m)}\} \sim \pi(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$ , where the posterior distribution  $\pi(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$  is conditioned on the hyper-parameters  $\boldsymbol{\theta}$  and defined according to a Bayesian hierarchical model.

## 3.2 Bayesian Inference

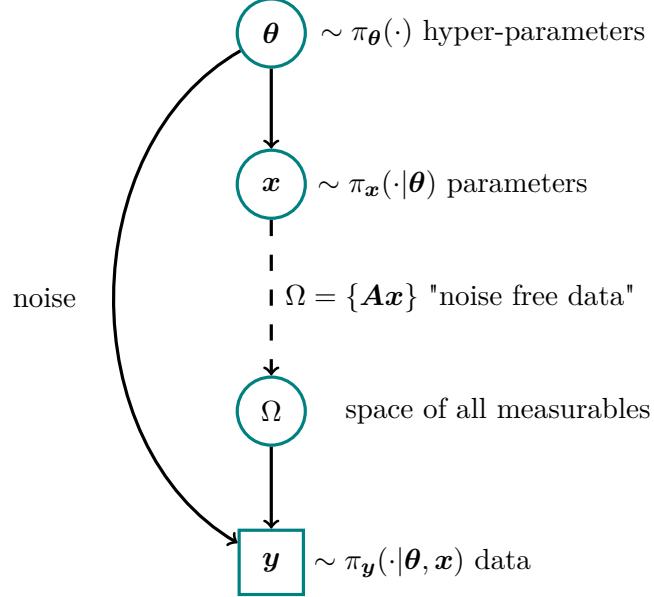
In this section, we introduce the basics of Bayesian inference for an unknown parameter  $\mathbf{x}$  given observed data

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \boldsymbol{\eta}, \quad (3.1)$$

based on a linear forward model  $\mathbf{A}$  and some additive noise  $\boldsymbol{\eta}$ . A more sophisticated Bayesian framework specifically applied to the previously introduced forward model and some simulated data will be developed in Section ??.

We can visualise the correlation structure between parameters as well as how distributions progress in a measurement process, using a hierarchically ordered directed acyclic graph (DAG), see Figure 3.2. Since any observational process naturally involves random noise, we include this in the DAG and classify the noise variance as a hyper-parameter within  $\boldsymbol{\theta}$  [8]. Other hyper-parameters, to which we assign a hyper-prior distribution  $\pi(\boldsymbol{\theta})$ , may influence the parameters  $\mathbf{x}$  either statistically (indicated by solid arrows), as in Figure 3.2, or deterministically (indicated by dashed arrows) if functional dependent on each other. Here we can incorporate prior knowledge of  $\boldsymbol{\theta}$  and the parameter  $\mathbf{x}$  by defining  $\pi(\boldsymbol{\theta})$  and the prior distribution  $\pi(\mathbf{x}|\boldsymbol{\theta})$  according to their physical properties or functional dependences. This is one of the great strength of Bayesian modelling compared to e.g. regularisation, see section 3.3. Then the parameter  $\mathbf{x}$  is mapped deterministically through the forward model onto the space of all measurables  $\Omega$ . From this space, we statistically observe the actual data  $\mathbf{y}$ , which includes random (statistical)

noise as mentioned above. The distribution of the data conditioned on the hyper-parameters  $\boldsymbol{\theta}$  and the parameters  $\mathbf{x}$  is called the likelihood function  $\pi(\mathbf{y}|\boldsymbol{\theta}, \mathbf{x})$ , which includes information about the measurement process through the forward model. Then given some observed data, we like to characterise the posterior distribution  $\pi(\boldsymbol{\theta}, \mathbf{x}|\mathbf{y})$  of the underlying parameters and hyper-parameters by reversing the arrows in Figure 3.2.



**Figure 3.2:** The directed acyclic graph (DAG) for a linear inverse problem visualises statistical dependencies as solid line arrows and deterministic dependencies as dotted arrows. The parameters  $\mathbf{x}$  have some statistical dependency of those hyper-parameters  $\boldsymbol{\theta}$ , which are distributed as  $\pi(\boldsymbol{\theta})$ . Then a parameter  $\mathbf{x} \sim \pi_{\mathbf{x}}(\cdot|\boldsymbol{\theta})$  is mapped onto the space of all measurables  $\mathbf{u} = \mathbf{A}\mathbf{x}$  deterministically through the linear forward model  $\mathbf{A}$ . From the space of all measurables we observe some data  $\mathbf{y} = \mathbf{A}\mathbf{x} + \boldsymbol{\eta}$ , statistically, so that  $\mathbf{y} \sim \pi_{\mathbf{y}}(\cdot|\boldsymbol{\theta}, \mathbf{x})$ , with naturally some random noise  $\boldsymbol{\eta} \sim \pi_{\boldsymbol{\eta}}(\cdot|\boldsymbol{\theta})$ .

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

$$\pi(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}) = \frac{\pi(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})\pi(\mathbf{x}, \boldsymbol{\theta})}{\pi(\mathbf{y})}, \quad (3.2)$$

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

$$\pi(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}) \propto \pi(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})\pi(\mathbf{x}, \boldsymbol{\theta}). \quad (3.3)$$

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

$$\mathbb{E}_{\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}}[h(\mathbf{x}_\theta)] = \underbrace{\int \int h(\mathbf{x}_\theta) \pi(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}) d\mathbf{x} d\boldsymbol{\theta}}_{\mu_{\text{int}}}, \quad (3.4)$$

which may be a high dimensional integral and computationally not feasible to solve. Therefore the unbiased [9] sample based Monte Carlo estimate

$$\mathbb{E}_{\mathbf{x}, \boldsymbol{\theta} | \mathbf{y}}[h(\mathbf{x}_{\boldsymbol{\theta}})] \approx \underbrace{\frac{1}{N} \sum_{k=1}^N h(\mathbf{x}_{\boldsymbol{\theta}}^{(k)})}_{\boldsymbol{\mu}_{\text{samp}}}, \quad (3.5)$$

for large enough  $N$  (law of large numbers [10, Chapter 17]) is often used. Here, the samples  $\{\mathbf{x}^{(k)}, \boldsymbol{\theta}^{(k)}\} \sim \pi_{\mathbf{x}, \boldsymbol{\theta}}(\cdot | \mathbf{y})$ , for  $k = 1, \dots, N$ , form a sample set  $\mathcal{M} = \{(\mathbf{x}, \boldsymbol{\theta})^{(1)}, \dots, (\mathbf{x}, \boldsymbol{\theta})^{(N)}\}$ . Furthermore, the central limit theorem states that the samples mean  $\boldsymbol{\mu}_{\text{samp}}^{(i)}$ , of independent samples sets  $\mathcal{M}_i$  for  $i = 1, \dots, n$  of any distribution, converge in distribution to a normal distribution so that

$$\sqrt{n}(\boldsymbol{\mu}_{\text{samp}}^{(i)} - \boldsymbol{\mu}_{\text{int}}) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2) [11], \quad (3.6)$$

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

Generating a representative sample set from the posterior distribution presents a significant challenge. This is also due to the strong correlations that often exist between the parameters and hyper-parameters, as discussed by Rue and Held in [12] and illustrated in Appendix B. If  $\mathbf{x}$  can not be parametrised directly in terms of the hyper-parameters  $\boldsymbol{\theta}$ , i.e.,  $\mathbf{x}(\boldsymbol{\theta})$ , it is beneficial to factorise the posterior distribution as

$$\pi(\mathbf{x}, \boldsymbol{\theta} | \mathbf{y}) = \pi(\mathbf{x} | \boldsymbol{\theta}, \mathbf{y}) \pi(\boldsymbol{\theta} | \mathbf{y}), \quad (3.7)$$

into the conditional posterior  $\pi(\mathbf{x} | \boldsymbol{\theta}, \mathbf{y})$  over the latent field  $\mathbf{x}$  and the marginal posterior  $\pi(\boldsymbol{\theta} | \mathbf{y})$  over the hyper-parameters  $\boldsymbol{\theta}$ . This approach, known as the marginal and then conditional (MTC) method, is particularly advantageous when  $\mathbf{x}$  is high-dimensional (e.g.,  $\mathbf{x} \in \mathbb{R}^n$  with  $n = 45$ ), while  $\boldsymbol{\theta}$  is low-dimensional (e.g., two-dimensional) and one can deterministically work out the marginal distribution. Applying the law of total expectation [13], Eq. (3.4) becomes

$$\mathbb{E}_{\mathbf{x} | \mathbf{y}}[h(\mathbf{x})] = \mathbb{E}_{\boldsymbol{\theta} | \mathbf{y}} \left[ \mathbb{E}_{\mathbf{x} | \boldsymbol{\theta}, \mathbf{y}}[h(\mathbf{x}_{\boldsymbol{\theta}})] \right] = \int \mathbb{E}_{\mathbf{x} | \boldsymbol{\theta}, \mathbf{y}}[h(\mathbf{x}_{\boldsymbol{\theta}})] \pi(\boldsymbol{\theta} | \mathbf{y}) d\boldsymbol{\theta}, \quad (3.8)$$

where, in the case of a linear-Gaussian Bayesian hierarchical model, both the marginal distribution and the inner expectation  $\mathbb{E}_{\mathbf{x} | \boldsymbol{\theta}, \mathbf{y}}[h(\mathbf{x}_{\boldsymbol{\theta}})]$  are well defined.

Assuming Gaussian noise  $\boldsymbol{\eta} \sim \mathcal{N}(0, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$ , we define a linear-Gaussian Bayesian hierarchical model [8]

$$\mathbf{y} | \mathbf{x}, \boldsymbol{\theta} \sim \mathcal{N}(\mathbf{A}\mathbf{x}, \boldsymbol{\Sigma}(\boldsymbol{\theta})) \quad (3.9a)$$

$$\mathbf{x} | \boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{Q}^{-1}(\boldsymbol{\theta})) \quad (3.9b)$$

$$\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta}), \quad (3.9c)$$

with a normally distributed likelihood  $\pi(\mathbf{y} | \mathbf{x}, \boldsymbol{\theta})$  Why? other likelihoods? and prior distributions  $\pi(\mathbf{x} | \boldsymbol{\theta})$  and  $\pi(\boldsymbol{\theta})$ , the noise covariance matrix  $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ , the prior precision matrix  $\mathbf{Q}(\boldsymbol{\theta})$  and the prior mean  $\boldsymbol{\mu}$ . This model enables efficient factorisation of the posterior distribution and application of the MTC method.

### 3.2.1 Marginal and conditional posterior distribution

For the linear-Gaussian Bayesian hierarchical model specified in Eq. 3.9, the marginal posterior distribution over the hyper-parameters is given by

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \int \pi(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}) d\mathbf{x} \quad (3.10)$$

$$\propto \sqrt{\frac{\det(\boldsymbol{\Sigma}^{-1}) \det(\mathbf{Q})}{\det(\mathbf{Q} + \mathbf{A}^T \boldsymbol{\Sigma}^{-1} \mathbf{A})}} \times \exp \left[ -\frac{1}{2} (\mathbf{y} - \mathbf{A}\boldsymbol{\mu})^T \mathbf{Q}_{\boldsymbol{\theta}|\mathbf{y}} (\mathbf{y} - \mathbf{A}\boldsymbol{\mu}) \right] \pi(\boldsymbol{\theta}), \quad (3.11)$$

with

$$\mathbf{Q}_{\boldsymbol{\theta}|\mathbf{y}} = \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1} \mathbf{A} \left( \mathbf{A}^T \boldsymbol{\Sigma}^{-1} \mathbf{A} + \mathbf{Q} \right)^{-1} \mathbf{A}^T \boldsymbol{\Sigma}^{-1}, \quad (3.12)$$

see [8, Lemma 2]. Conditioned on the hyper-parameters  $\boldsymbol{\theta}$ , we can draw samples from the normal conditional posterior distribution

$$\mathbf{x}|\boldsymbol{\theta}, \mathbf{y} \sim \mathcal{N} \left( \underbrace{\boldsymbol{\mu} + \left( \mathbf{A}^T \boldsymbol{\Sigma}^{-1} \mathbf{A} + \mathbf{Q} \right)^{-1} \mathbf{A}^T \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \mathbf{A}\boldsymbol{\mu})}_{\boldsymbol{\mu}_{\mathbf{x}|\boldsymbol{\theta}, \mathbf{y}}}, \underbrace{\left( \mathbf{A}^T \boldsymbol{\Sigma}^{-1} \mathbf{A} + \mathbf{Q} \right)^{-1}}_{\boldsymbol{\Sigma}_{\mathbf{x}|\boldsymbol{\theta}, \mathbf{y}}} \right), \quad (3.13)$$

using the Randomise-then-Optimise (RTO) method (see Section 3.4.2), or compute weighted expectations, as in Eq. 3.8, of the conditional mean and covariance matrix, where the weights are given by  $\pi(\boldsymbol{\theta}|\mathbf{y})$ . Note that both the noise covariance  $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$  and the prior precision matrix  $\mathbf{Q} = \mathbf{Q}(\boldsymbol{\theta})$  depend on the hyper-parameters  $\boldsymbol{\theta}$ .

## 3.3 Regularisation

Another method for obtaining a solution to the linear inverse problem in Eq. 3.1 is regularisation. In this approach, we seek a solution  $\mathbf{x}_\lambda$  that minimises both the data misfit norm and a regularisation semi-norm, as described in [8]. Here we focus on a regularisation semi-norm for the case of Tikhonov regularisation [1, 14], which is closest to a linear-Gaussian hierarchical Bayesian model, as introduced in Eq. 3.9.

Given a parameter vector  $\mathbf{x}$ , a linear forward model matrix  $\mathbf{A}$ , and data  $\mathbf{y}$ , the data misfit norm

$$\|\mathbf{y} - \mathbf{A}\mathbf{x}\| \quad (3.14)$$

quantifies how well the noise-free data  $\mathbf{A}\mathbf{x}$  matches the observed data. The regularisation semi-norm

$$\lambda \|\mathbf{T}\mathbf{x}\| \quad (3.15)$$

penalises the solution according to the regularisation operator  $\mathbf{T}$  and the regularisation parameter  $\lambda > 0$ . For a fixed  $\lambda$ , the regularised solution

$$\mathbf{x}_\lambda = \arg \min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|^2 + \lambda \|\mathbf{T}\mathbf{x}\|^2 \quad (3.16)$$

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

$$\nabla_{\mathbf{x}} \left\{ (\mathbf{y} - \mathbf{A}\mathbf{x})^T (\mathbf{y} - \mathbf{A}\mathbf{x}) + \lambda \mathbf{x}^T \mathbf{T}^T \mathbf{T}\mathbf{x} \right\} = 0 \quad (3.17)$$

$$\iff \nabla_{\mathbf{x}} \left\{ \mathbf{y}^T \mathbf{y} + \mathbf{x}^T \mathbf{A}^T \mathbf{A}\mathbf{x} - 2\mathbf{y}^T \mathbf{A}\mathbf{x} + \lambda \mathbf{x}^T \mathbf{T}^T \mathbf{T}\mathbf{x} \right\} = 0 \quad (3.18)$$

$$\iff 2\mathbf{A}^T \mathbf{A}\mathbf{x} - 2\mathbf{A}^T \mathbf{y} + 2\lambda \mathbf{T}^T \mathbf{T}\mathbf{x} = 0. \quad (3.19)$$

Solving this equation yields the regularised solution

$$\mathbf{x}_\lambda = (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{L})^{-1} \mathbf{A}^T \mathbf{y}, \quad (3.20)$$

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

In practice,  $\mathbf{x}_\lambda$  is computed for a range of  $\lambda$ -values and evaluated based on the trade-off between the data misfit and the regularisation norm. The optimal value of  $\lambda$  is often chosen as the point of maximum curvature on the so-called L-curve [15], which we plot in Section ??.

## 3.4 Sampling Methods

In this section we present the sampling methods used in this thesis and show how these methods draw samples  $\mathcal{M} = \{(\mathbf{x}, \boldsymbol{\theta})^{(1)}, \dots, (\mathbf{x}, \boldsymbol{\theta})^{(k)}, \dots, (\mathbf{x}, \boldsymbol{\theta})^{(N)}\} \sim \pi(\mathbf{x}, \boldsymbol{\theta} | \mathbf{y})$  from the desired target distribution, so that we can apply sample-based estimates as in Eq. 3.5. Here,  $\mathcal{M}$  denotes a Markov chain, where each new sample  $(\mathbf{x}, \boldsymbol{\theta})^{(k)}$  is only affected by the previous one,  $(\mathbf{x}, \boldsymbol{\theta})^{(k-1)}$ . Markov chain Monte Carlo (MCMC) methods generate such a chain  $\mathcal{M}$  using random (Monte Carlo) proposals  $(\mathbf{x}, \boldsymbol{\theta})^{(k)} \sim q(\cdot | (\mathbf{x}, \boldsymbol{\theta})^{(k-1)})$  according to a proposal distribution conditioned on the previous sample (Markov), where ergodicity of  $\mathcal{M}$  is a sufficient criterion for using sample-based estimates [1, 9].

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  $\pi(\mathbf{x}, \boldsymbol{\theta} | \mathbf{y})$ . In practice, one can inspect the trace  $\pi(\mathbf{x}^{(k)}, \boldsymbol{\theta}^{(k)} | \mathbf{y})$  for  $k = 1, \dots, N$  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.

### 3.4.1 Metropolis- within Gibbs sampling

As introduced in Section 3.2.1, when using the MTC method we sample separately from  $\pi(\boldsymbol{\theta}|\mathbf{y})$  and  $\pi(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$ . To sample from  $\pi(\boldsymbol{\theta}|\mathbf{y})$ , we use a Metropolis-within-Gibbs sampler as described in [8]. In this thesis, the sampler is applied to the two-dimensional case only, with  $\boldsymbol{\theta} = (\theta_1, \theta_2)$ , where we perform a Metropolis step in the  $\theta_1$  direction and a Gibbs step in the  $\theta_2$  direction. Ergodicity for this approach is proven in [16].

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 [16]. We accept and set  $\theta_1^{(t)} = \theta_1$  with the acceptance probability

$$\alpha(\theta_1|\theta_1^{(t-1)}) = \min \left\{ 1, \frac{\pi(\theta_1|\theta_2^{(t-1)}, \mathbf{y}) q(\theta_1^{(t-1)}|\theta_1)}{\pi(\theta_1^{(t-1)}|\theta_2^{(t-1)}, \mathbf{y}) q(\theta_1|\theta_1^{(t-1)})} \right\} \quad (3.21)$$

or reject and keep  $\theta_1^{(t)} = \theta_1^{(t-1)}$ , which we do by comparing  $\alpha$  to a uniform random number  $u \sim \mathcal{U}(0, 1)$ .

Next, we perform a Gibbs step in the  $\theta_2$  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  $\theta_2^{(t)} \sim \pi(\cdot|\theta_1^{(t)}, \mathbf{y})$ , 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  $k = 1, \dots, N'$  do
3:   Propose  $\theta_1 \sim q(\cdot|\theta_1^{(t-1)}) = q(\theta_1^{(t-1)}|\cdot)$ 
4:   Compute

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

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  $\theta_2^{(t)} \sim \pi(\cdot|\theta_1^{(t)}, \mathbf{y})$ 
12: end for
13: Output:  $\boldsymbol{\theta}^{(0)}, \dots, \boldsymbol{\theta}^{(k)}, \dots, \boldsymbol{\theta}^{(N)} \sim \pi(\boldsymbol{\theta}|\mathbf{y})$ 
```

### 3.4.2 Draw a sample from a multivariate normal distribution

As part of the MTC scheme, we only draw samples from the conditional distribution  $\pi(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$  after sampling from the marginal posterior  $\pi(\boldsymbol{\theta}|\mathbf{y})$ . For linear-Gaussian Bayesian hierarchical models, samples from the multivariate normal distribution  $\pi(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$  can be efficiently generated using the Randomise-then-Optimise (RTO) method [17].

The full conditional distribution can be rewritten as

$$\pi(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}) \propto \pi(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) \pi(\mathbf{x}|\boldsymbol{\theta}) \quad (3.22)$$

$$= \exp\left(-\|\hat{\mathbf{A}}\mathbf{x} - \hat{\mathbf{y}}\|^2\right), \quad (3.23)$$

where

$$\hat{\mathbf{A}} = \begin{bmatrix} \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\theta})\mathbf{A} \\ \mathbf{Q}^{1/2}(\boldsymbol{\theta}) \end{bmatrix}, \quad \hat{\mathbf{y}} = \begin{bmatrix} \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\theta})\mathbf{y} \\ \mathbf{Q}^{1/2}(\boldsymbol{\theta})\boldsymbol{\mu} \end{bmatrix} \quad [18]. \quad (3.24)$$

A sample  $\mathbf{x}_i$  can be computed by minimising the following equation with respect to  $\hat{\mathbf{x}}$ :

$$\mathbf{x}_i = \arg \min_{\hat{\mathbf{x}}} \|\hat{\mathbf{A}}\hat{\mathbf{x}} - (\hat{\mathbf{y}} + \mathbf{b})\|^2, \quad \mathbf{b} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad (3.25)$$

where we add a randomised perturbation  $\mathbf{b}$ . Similar to Section 3.3, this expression can be rewritten as

$$(\mathbf{A}^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}) \mathbf{A} + \mathbf{Q}(\boldsymbol{\theta})) \mathbf{x}_i = \mathbf{A}^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}) \mathbf{y} + \mathbf{Q}(\boldsymbol{\theta}) \boldsymbol{\mu} + \mathbf{v}_1 + \mathbf{v}_2, \quad (3.26)$$

where the term  $-\hat{\mathbf{A}}^T \mathbf{b}$  is decomposed as  $\mathbf{v}_1 + \mathbf{v}_2$ , with  $\mathbf{v}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{A}^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}) \mathbf{A})$  and  $\mathbf{v}_2 \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}(\boldsymbol{\theta}))$ , representing independent Gaussian random variables [8, 17].

If the Markov chain over the marginal posterior  $\pi(\boldsymbol{\theta}|\mathbf{y})$  is ergodic, and the conditional samples  $\mathbf{x}^{(k)} \sim \pi(\mathbf{x}|\boldsymbol{\theta}^{(k)}, \mathbf{y})$  are drawn independently, then the resulting joint chain  $\{(\mathbf{x}, \boldsymbol{\theta})^{(1)}, \dots, (\mathbf{x}, \boldsymbol{\theta})^{(N)}\} \sim \pi(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})$  is also ergodic [19].

### 3.4.3 t-walk sampler as black box

If the parameters  $\mathbf{x}$  are functionally dependent on the hyper-parameters  $\boldsymbol{\theta}$ , i.e.,  $\mathbf{x} = \mathbf{x}(\boldsymbol{\theta})$ , we can sample directly from the marginal posterior  $\pi(\boldsymbol{\theta}|\mathbf{y})$  using the t-walk algorithm by Christen and Fox [20]. The t-walk is employed as a black-box sampler, requiring only the specification of the number of samples, burn-in period, support region, and the sampling distribution. Convergence to the target distribution is guaranteed by construction of the algorithm.

### 3.5 Numerical Approxiamtion Methods - Tensor Train

Explain how to find normalisation constant and say that due to aproxiatming the sqaure root we ensure posivyt later when squaring it First, we provide a short overview of probability spaces and their associated measures, as a foundation for deriving marginal probability distribution, and then we give a brief introduction to the tensor train format. The motivation to use the tensor train format is that we can approximate a d-dimensional grid with far fewer data points compared to the total number of grid points.

Assume that the triple  $(\Omega, \mathcal{F}, \mathbb{P})$  defines a probability space, where  $\Omega$  denotes the complete sample space,  $\mathcal{F}$  is a  $\sigma$ -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  $\sigma$ -algebra over  $\Omega$ , are given in Appendix C. We denote

$$\mathbb{P}(A) = \int_A d\mathbb{P} \quad (3.27)$$

as the probability of an event  $A \in \mathcal{F}$ . By applying the Radon-Nikodym theorem [21], we can change variables

$$\mathbb{P}(A) = \int_A \frac{d\mathbb{P}}{dx} dx = \int_A \pi(x) dx, \quad (3.28)$$

where  $dx$  is a reference measure on the same probability space, commonly referred to as the Lebesgue measure. The Radon-Nikodym derivative  $\frac{d\mathbb{P}}{dx}$  of  $\mathbb{P}$  with respect to  $x$ , and is often interpreted as the probability density function (PDF)  $\pi(x)$ . Thus, we say that  $\mathbb{P}$  has a density  $\pi(x)$  with respect to  $x$  [22, Chapter 10].

Now, let  $X : \Omega \rightarrow \mathbb{R}^d$  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  $\pi(x)$ , is a joint density of  $X$ , induced by the probability measure on  $\Omega$  [21, 23]. As by Cui et al. [24], we can define the parameter space as the Cartesian product  $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \dots \times \mathcal{X}_d$  with  $x_k \in \mathcal{X}_k \subseteq \mathbb{R}$  and  $x = (x_1, \dots, x_k, \dots, x_d)$ . The marginal density function for the  $k$ -th component is then given by

$$f_{X_k}(x_k) = \int_{\mathcal{X}_1} \dots \int_{\mathcal{X}_d} \lambda(x) \pi(x) dx_1 \dots dx_{k-1} dx_{k+1} \dots dx_d, \quad (3.29)$$

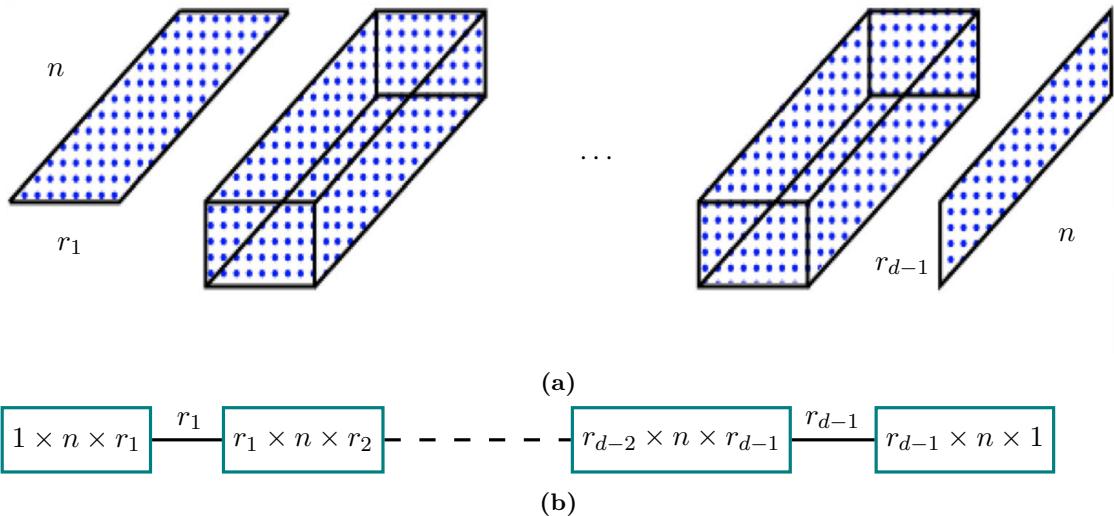
where we integrate over all dimensions except the  $k$ -th. Here, we introduce a weight function  $\lambda(x)$  [25], which can be useful for quadrature rules?. Cui et al. [24] refer to  $\lambda(x)$  as a "product-form Lebesgue-measurable weighting function" and define it as

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

Using the tensor train (TT) format, we can efficiently approximate a  $d$ -dimensional function  $\pi(x)$  and compute marginal probability distributions at low computational cost. To do so, we first define a  $d$ -dimensional discrete univariate grid over the parameter space  $\mathcal{X}$ , with  $n$  grid points in each dimension. In the tensor train format we can represent the function over this  $d$ -dimensional grid as a product train of 2D matrices (rank-2 tensor) and 3D matrices (rank-3 tensors), which we call TT-cores, see Fig. 3.3. More specifically each core  $\pi_k \in \mathbb{R}^{r_{k-1} \times n \times r_k}$  has ranks  $r_{k-1}$  and  $r_k$ , for  $k = 1, \dots, d$ , connecting it with its neighbouring cores, as illustrated in Figure 3.3. For the first and last cores, the outer ranks are set to  $r_0 = r_d = 1$ . This enables us to write the value  $\pi(x)$ , for a fixed point  $x = (x_1, \dots, x_d)$  on the grid, as a sequence of matrix multiplications

$$\pi_1(x_1)\pi_2(x_2) \cdots \pi_d(x_d) = \pi(x) \in \mathbb{R},$$

where each core  $\pi_k(x_k)$ , becomes a matrix of size  $r_{k-1} \times r_k$ . Clearly this shows that we only need  $dnr^2$  evaluation points instead of  $n^d$  grid points to approximate the whole parameter space. Consequently, with a tensor train approximation, the marginal target function



**Figure 3.3:** 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  $\pi_k$  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 [26].

$$f_{X_k}(x_k) = \frac{1}{z} \left| \left( \int_{\mathbb{R}} \lambda_1(x_1) \pi_1(x_1) dx_1 \right) \cdots \left( \int_{\mathbb{R}} \lambda_{k-1}(x_{k-1}) \pi_{k-1}(x_{k-1}) dx_{k-1} \right) \lambda_k(x_k) \pi_k(x_k) \right| \\ \left( \int_{\mathbb{R}} \lambda_{k+1}(x_{k+1}) \pi_{k+1}(x_{k+1}) dx_{k+1} \right) \cdots \left( \int_{\mathbb{R}} \lambda_d(x_d) \pi_d(x_d) dx_d \right) \right| \quad (3.30)$$

is computed by integrating over all TT cores except  $\pi_k$ , as in [27], including a normalisation constant  $z$  [24].

In practice, tensor train approximations may suffer from numerical instability, particularly because it is not advantageous to approximate the target function  $\pi(x)$  in for example, the logarithmic space. To address this, we follow the notation and procedure of Cui et al. [24] and instead approximate the square root of the probability density

$$\sqrt{\pi(x)} \approx g(x) = \mathbf{G}_1(x_1), \dots, \mathbf{G}_k(x_k), \dots, \mathbf{G}_d(x_d). \quad (3.31)$$

Here, each TT-core is given by

$$G_k^{(\alpha_{k-1}, \alpha_k)}(x_k) = \sum_{i=1}^{n_k} \phi_k^{(i)}(x_k) \mathbf{A}_k[\alpha_{k-1}, i, \alpha_k], \quad k = 1, \dots, d, \quad (3.32)$$

where  $\mathbf{A}_k \in \mathbb{R}^{r_{k-1} \times n_k \times r_k}$  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:

$$\pi(x) \approx \gamma' + g^2(x), \quad (3.33)$$

where  $\gamma'$  is a small positive constant added to ensure positivity and is chosen such that

$$\gamma' \leq \frac{1}{\lambda(\mathcal{X})} \|g - \sqrt{\pi}\|_2^2. \quad (3.34)$$

This leads to the normalised target function

$$f_X(x) = \frac{1}{z} \lambda(x) \pi(x) = \frac{1}{z} (\lambda(x) \gamma' + \lambda(x) g^2(x)), \quad (3.35)$$

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

$$\begin{aligned} f_{X_k}(x_k) &= \frac{1}{z} \left( \gamma' \prod_{i=1}^{k-1} \lambda_i(\mathcal{X}_i) \prod_{i=k+1}^d \lambda_i(\mathcal{X}_i) \right. \\ &\quad + \left( \int_{\mathbb{R}} \lambda_1(x_1) \mathbf{G}_1^2(x_1) dx_1 \right) \cdots \left( \int_{\mathbb{R}} \lambda_{k-1}(x_{k-1}) \mathbf{G}_{k-1}^2(x_{k-1}) dx_{k-1} \right) \\ &\quad \left. \lambda_k(x_k) \mathbf{G}_k^2(x_k) \right. \\ &\quad \left. \left( \int_{\mathbb{R}} \lambda_{k+1}(x_{k+1}) \mathbf{G}_{k+1}^2(x_{k+1}) dx_{k+1} \right) \cdots \left( \int_{\mathbb{R}} \lambda_d(x_d) \mathbf{G}_d^2(x_d) dx_d \right) \right). \end{aligned} \quad (3.36)$$

To compute these marginals efficiently, one can use a procedure similar to left and right orthogonalisation of TT-cores [28]. For this, we define the mass matrix  $\mathbf{M}_k \in \mathbb{R}^{n_k \times n_k}$  as

$$\mathbf{M}_k[i, j] = \int_{\mathcal{X}_k} \phi_k^{(i)}(x_k) \phi_k^{(j)}(x_k) \lambda(x_k) dx_k, \quad i, j = 1, \dots, n_k, \quad (3.37)$$

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

### 3.5.1 Marginal Functions

We compute the marginal functions using two procedures, referred to as backward

marginalisation [24] and forward marginalisation. The backward marginalisation provides

us with the coefficient matrices  $\mathbf{B}_k$ , while the forward marginalisation gives the coefficient

matrices  $\mathbf{B}_{\text{pre},n}$ . These matrices enable the efficient evaluation of marginal functions,

similar to [24]. The proposition used to compute  $\mathbf{B}_k$ , stated in Proposition 1, is

adapted directly from [24].

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

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

$$\mathbf{C}_k[\alpha_{k-1}, \tau, l_k] = \sum_{i=1}^{n_k} \mathbf{B}_k[\alpha_{k-1}, i, l_k] \mathbf{L}_k[i, \tau]. \quad (3.38)$$

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

$$\mathbf{Q}_k \mathbf{R}_k = (\mathbf{C}_k^{(R)})^\top. \quad (3.39)$$

3. Compute the new coefficient tensor:

$$\mathbf{B}_{k-1}[\alpha_{k-2}, i, l_{k-1}] = \sum_{\alpha_{k-1}=1}^{r_{k-1}} \mathbf{A}_{k-1}[\alpha_{k-2}, i, \alpha_{k-1}] \mathbf{R}_k[l_{k-1}, \alpha_{k-1}]. \quad (3.40)$$

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

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

$$\mathbf{C}_{\text{pre},k}[\alpha_{k-1}, \tau, l_k] = \sum_{i=1}^{n_k} \mathbf{L}_k[i, \tau] \mathbf{B}_{\text{pre},k}[\alpha_{k-1}, i, l_k]. \quad (3.41)$$

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

$$\mathbf{Q}_{\text{pre},k} \mathbf{R}_{\text{pre},k} = (\mathbf{C}_{\text{pre},k}^{(R)}). \quad (3.42)$$

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

$$\mathbf{B}_{\text{pre},k+1}[l_{k+1}, i, \alpha_{k+1}] = \sum_{\alpha_{k-1}=1}^{r_k} \mathbf{R}_{\text{pre},k}[l_{k+1}, \alpha_k] \mathbf{A}_{k+1}[\alpha_k, i, \alpha_{k+1}]. \quad (3.43)$$

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

$$f_{X_k}(x_k) = \frac{1}{z} \left( \gamma' \prod_{i=1}^{k-1} \lambda_i(X_i) \prod_{i=k+1}^d \lambda_i(X_i) + \sum_{l_{k-1}=1}^{r_{k-1}} \sum_{l_k=1}^{r_k} \left( \sum_{i=1}^n \phi_k^{(i)}(x_k) \mathbf{D}_k[l_{k-1}, i, l_k] \right)^2 \right) \lambda_k(x_k), \quad (3.44)$$

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

$$\mathbf{D}_k[l_{k-1}, i, l_k] = \sum_{\alpha_{k-1}=1}^{r_{k-1}} \mathbf{R}_{\text{pre},k-1}[l_{k-1}, \alpha_{k-1}] \mathbf{B}_k[\alpha_{k-1}, i, l_k]. \quad (3.45)$$

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

$$f_{X_1}(x_1) = \frac{1}{z} \left( \gamma' \prod_{i=2}^d \lambda_i(\mathcal{X}_i) + \sum_{l_1=1}^{r_1} \left( \sum_{i=1}^n \phi_1^{(i)}(x_1) \mathbf{D}_1[i, l_1] \right)^2 \right) \lambda_1(x_1), \quad (3.46)$$

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

$$f_{X_d}(x_d) = \frac{1}{z} \left( \gamma' \prod_{i=1}^{d-1} \lambda_i(\mathcal{X}_i) + \sum_{l_{n-1}=1}^{r_{d-1}} \left( \sum_{i=1}^n \phi_d^{(i)}(x_d) \mathbf{D}_d[l_{n-1}, i] \right)^2 \right) \lambda_d(x_d), \quad (3.47)$$

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



# 4

## Results

In this chapter we explain how we use the forward model to generate data and how we set up a Bayesian framework to then give distribution of solutions. Here we make it explicit and provide all information to be able to replicate our results. In this work we simulate data with an ozone profile from [1]. We follow the U.S. standard atmosphere, 1976 [2] to relate pressure and temperature and height and the HITRANonline [3] database to calculate the measured signal. We follow the mipas instrument description from the forward model [4].

All programming is done in python

### 4.1 Simulate Data

Ground truth We take an ozone profile generated from some MLS Data [29] in the Antarctic region. This gives us pressure values and Ozone volume mixing ratio values. We connect pressure and height values with the hydrostatic equilibrium equation

$$\frac{dp}{p} = \frac{-gM}{R^*T} dh, \quad (4.1)$$

which holds up to a geometric height of 84.356km, which is also the the boarer of the atmospher we set.

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

with  $r_0 \approx 6356$  km (also known as the polar radius of the earth) and  $g_0 \approx 9.81\text{m/s}^2$  and [5]. with the universal gas constant  $R^* \approx 8.314 \text{ N m/mol/K}$ . The molecules number ... is  $M = M_0 \approx 28.97 \text{ kg/mol}$  for altitudes below 79km [6]. We ignore the difference of .0004%

We take Temparute values

$$T(h) = \begin{cases} T_0, & h = 0 \\ T_0 + a_0 h, & 0 \leq h < h_1 \\ T_0 + a_0 h_1, & h_1 \leq h < h_2 \\ T_0 + a_0 h_1 + a_1(h - h_2), & h_2 \leq h < h_3 \\ T_0 + a_0 h_1 + a_1(h_3 - h_2) + a_2(h - h_3), & h_3 \leq h < h_4 \\ T_0 + a_0 h_1 + a_1(h_3 - h_2) + a_2(h_4 - h_3), & h_4 \leq h < h_5 \\ T_0 + a_0 h_1 + a_1(h_3 - h_2) + a_2(h_4 - h_3) + a_3(h - h_5), & h_5 \leq h < h_6 \\ T_0 + a_0 h_1 + a_1(h_3 - h_2) + a_2(h_4 - h_3) + a_3(h_6 - h_{T5}) + a_4(h - h_6), & h_6 \leq h \lesssim 85 \end{cases} \quad (4.3)$$

from the [30], We can formulate a temperature function from [30] where the geometric height height values are  $h_1 = 11\text{km}$   $h_2 = 20.1\text{km}$   $h_3 = 32.2\text{km}$   $h_4 = 47.4\text{km}$   $h_5 = 51.4\text{km}$   $h_6 = 71.8\text{km}$   $a_0 = -6.5\text{K/km}$   $a_1 = 1\text{K/km}$   $a_2 = 2.8\text{K/km}$   $a_3 = -2.8\text{K/km}$   $a_4 = -2\text{K/km}$  depending on 12 parameters with values provided by [], also see table ???. We set this as our true temperature profile and refer to [] for temperature values above 79 km.

[31] [29] [4] [32] [33] [2] [3] [30]

Then we can calculate data vector with a pointing accuracy 0.0075 which is equal ot arc as request by [31]. Then we get  $m =$  in between an atmosphere of  $h_{L,0} =$  and  $h_{L,n} =$ , where the layers are dictated by the MLS data. at wave number  $\nu =$  where ozone is most dominant According to Equation With voltage SNR = 60.

Given the calculte pressure value we can fit an exponentila within the rehion our interest the atmospheric layers

The pressure function,

$$p(h) = \exp \{-b(h - h_0)\} p_0, \quad (4.4)$$

is depending on three different parameters, the gradient  $b$  and the tuple  $(p_0, h_0)$ .

Where the ground truth values for that are

## 4.2 Set up Bayesian model

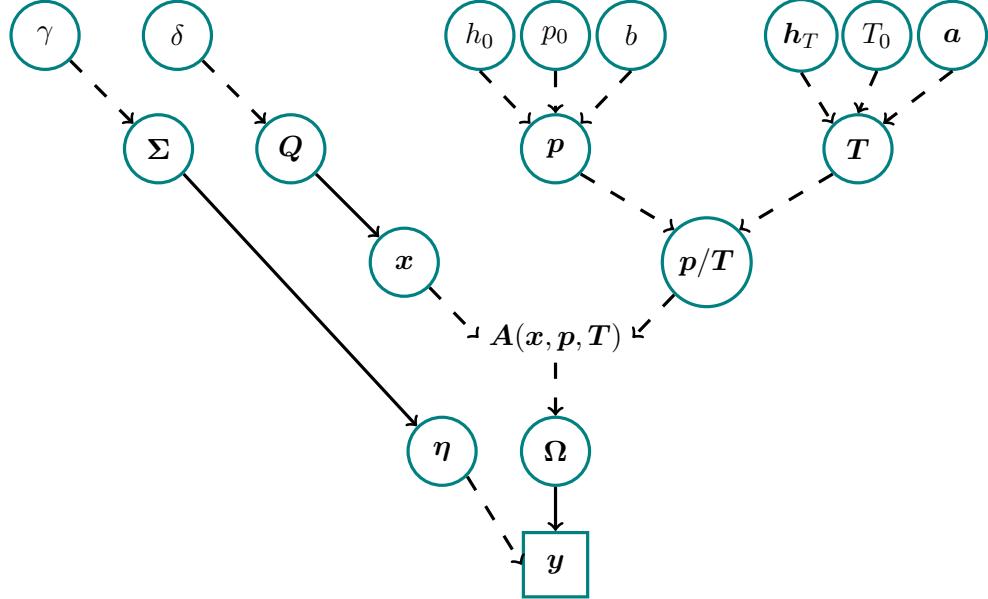
- fill in what we descriube beofre
- describe parametres and hyper paraemrets set up model to make dag
- then prior modelling

In this section we present the explicit Bayesian we will use to recover pressure temperature and ozone values. On big parts of bayesina modeling is prior modelling where we set up the model figure out correeltaion with DAG and choose priors as uninformativae as we can and accordfing to physical propoeritesd if we can

### 4.2.1 DAG

Again we draw a DAG and specify prior distribution over those

- correlation structure between ozone and temperature and pressure



**Figure 4.1:** Complete directed acyclic graph of the forward model. The hyper-parameters at the top deterministically (dotted line) describe the parameters  $(\mathbf{p}/\mathbf{T})$  or the noise covariance  $\Sigma = \gamma^{-1} \mathbf{I}$  of the random (solid line) noise  $\boldsymbol{\eta} \sim \mathcal{N}(0, \gamma^{-1} \mathbf{I})$  and precision matrix  $\mathbf{Q} = \delta \mathbf{L}$  of the distribution of  $\mathbf{x} \sim \mathcal{N}(0, \delta \mathbf{L})$ , where  $\mathbf{L}$  is a graph Laplacian as in Eq. 4.5. We can group the noise precision  $\gamma$  and the smoothness parameter  $\delta$  to define the marginal posterior over those hyper-parameters and then condition on them for the conditional posterior distribution, for further details see Fig. 4.6. In this whole process where we condition on the pressure  $\mathbf{p}$  and temperature  $\mathbf{T}$ , which we retrieve separately, see Fig. 4.15. The hyper-parameters  $h_0, p_0, b$  deterministically describe the pressure function in Eq. 4.4, note that we only need three parameters here since  $h_0 < h_{L,0}$  and  $\mathbf{h} = \{h_1, h_2, h_3, h_4, h_5, h_6\}$ ,  $\mathbf{a} = \{a_0, a_1, a_2, a_3, a_4\}$  and  $T_0$  determine the temperature function. The parameters  $\mathbf{x}$  and  $\mathbf{p}/\mathbf{T}$  determine the space of all measurable noise free data  $\Omega$  through the forward model  $\mathbf{A}(\mathbf{x}, \mathbf{p}, \mathbf{T})$  from which we randomly observe data set plus some random noise.

### 4.2.2 Prior Modelling

- physical meaning by choosing priors
- eventuel dependencies
- draw samples from priors should be as loose as possible

In practise we separate OPzone and condition on temperature and pressure or vice versa.

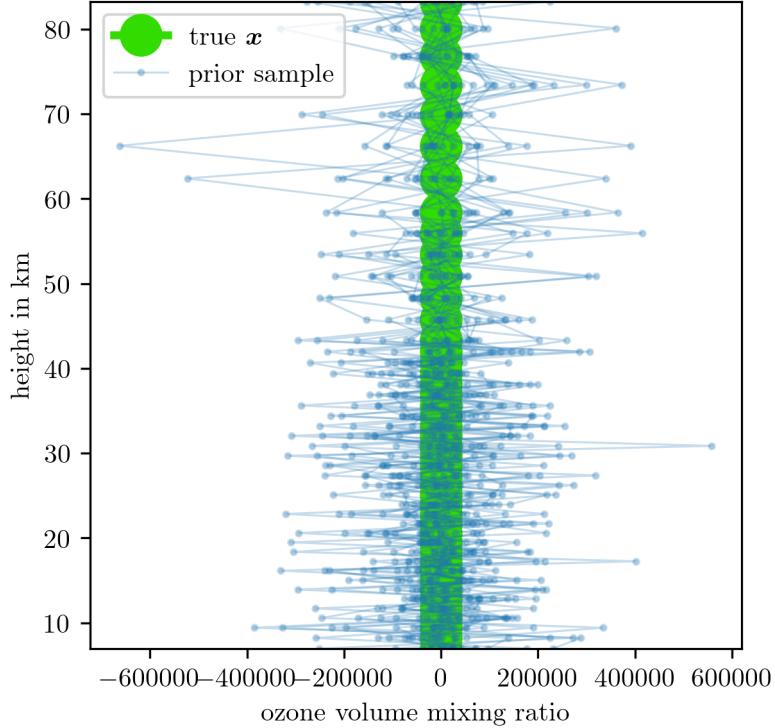
model parameters	priors	TT bounds		Context
		lower	upper	
$\gamma$	$\mathcal{T}(1, 10^{-10})$	$5 \cdot 10^{-8}$	$4.5 \cdot 10^{-7}$	$\mathbf{y}$
$\delta$	$\mathcal{T}(1, 10^{-10})$	-	-	$\mathbf{x}$
$\lambda$	-	500	7000	$\mathbf{x}$
$\mathbf{x}$	$\mathcal{N}(0, \delta \mathbf{L})$	-	-	$\mathbf{x}$
$h_0$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

**Table 4.1:** Gaussian  $\mathcal{N}(\mu, \sigma)$  and gamma distribution  $\mathcal{T}(\alpha = \text{scale}, \beta = \text{rate})$  Bounds for t and p 2.8 times the variance around the mean round pressure approx and test if would work with previous gamma prior or fix gamma prior with set values

## Ozone

graph lacoiacln wigthing in between similiar to a ciuop;led osscialtors what heught is change of weights we choose gamma distribution for easier sampling later on

$$\mathbf{Q} = \delta \mathbf{L} = \delta \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 6 & -5 \\ & & & \ddots & \ddots & \ddots \\ & & & & -5 & 10 & -5 \\ & & & & & -5 & 10 \end{bmatrix} \quad (4.5)$$



**Figure 4.2:** We draw samples from ozone prior distribution  $\mathbf{x} \sim \mathcal{N}(0, \delta \mathbf{L})$  after generating a sample from the hyper-prior distribution  $\delta \sim \mathcal{T}(1, 10^{-10})$ . Note that since the spread/variance of prior samples is very large compared to the ozone volume mixing ratios, the ozone profile appears to be constant, which it is not, as seen e.g. in Fig. 4.9.

### pressure over temperature

We parametrize the pressure by fitting one exponential to the pressure values  $p$  related to the height  $h$  by Eq. ???. We have to be carefull when we choose priors for the tmeprature and pressure

Height priors for temperature don ot overlap also grid for TT is chosen that way

Then pressure prior can not overlap the lowest atmospheric layer as then we would need a second gradient for the exponential

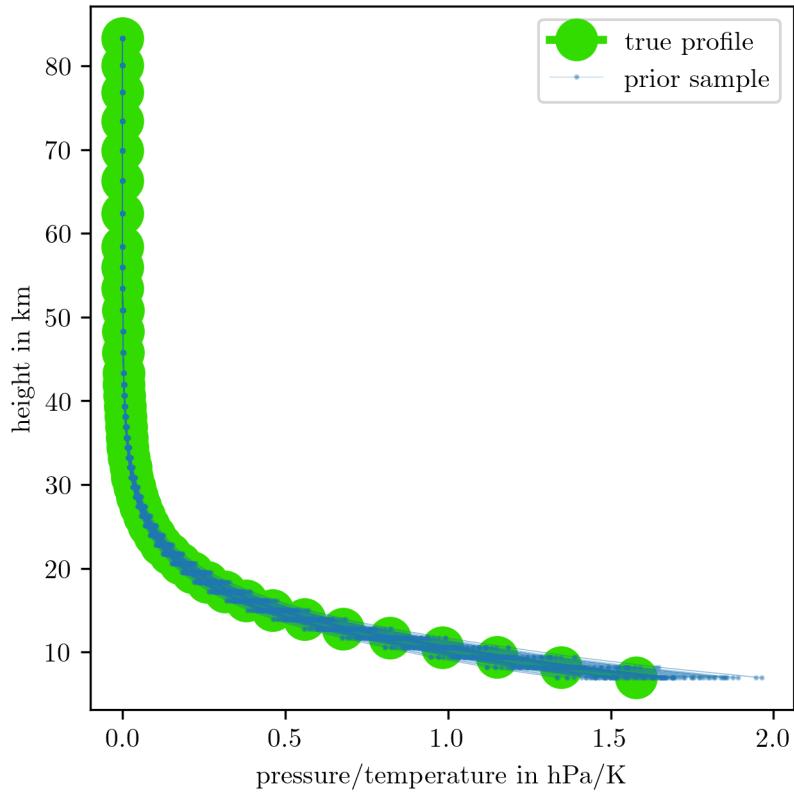
we are also limited by numerical issues of TT and essentially we can choose uniform priors but then the t-walk would run forever limit parameter space In case of uninformative data the priors can help us limit the parameter space.

We also show in the appendix that the temperature does not influence the pressure , even though they are correlated

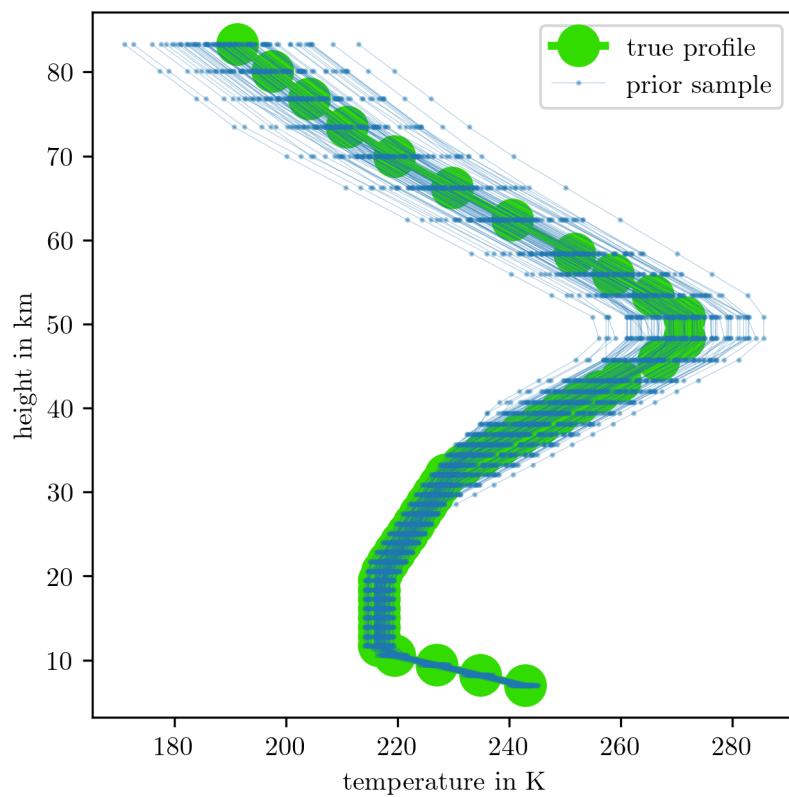
- show that pressure is dominant
- correlation structure

- appendix

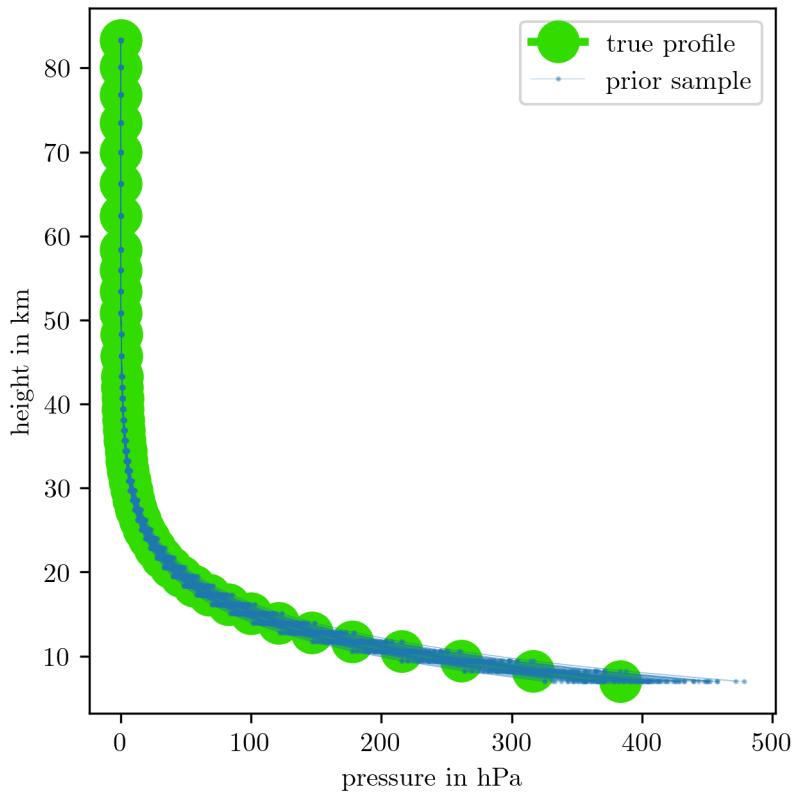
PriorTempOverPostMeanSigm



**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 4.1 and then calculate  $p/T$  according to the functions in Eq. 4.4 and 4.3.



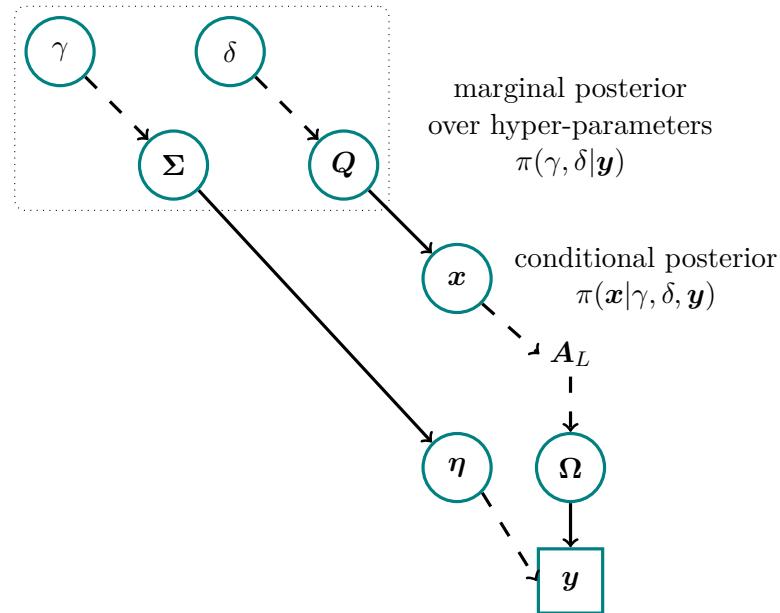
**Figure 4.4:** We draw samples from the hyper-prior distribution of  $h_1, h_2, h_3, h_4, h_5, h_6, a_0, a_1, a_2, a_3, a_4$  and  $T_0$  as defined in table 4.1 and then calculate  $\mathbf{T}$  according to the function in Eq. 4.3.



**Figure 4.5:** We draw samples from the hyper-prior distribution of  $h_0, b$  and  $p_0$  as defined in table 4.1 and then calculate  $\mathbf{p}$  according to the function in Eq. 4.4.

### 4.3 Posterior distributions with Linear model for Ozone – MTC

In this section we calculate the posterior marginal and then conditional (MTC) posterior distribution for ozone conditioned on the ground truth temperature and pressure profiles using the linear forward model  $\mathbf{A}_L$ . This is faster than the other way round (finding temperature over pressure conditioning on ozone) and temperature and pressure are well defined within the atmosphere so it is easier to just condition on a temperature and pressure profile out of a text book. We employ a so-called Metropolis within Gibbs (MWG) algorithm on the marginal posterior as summarised in the algorithmic Box 2 or use a Tensor-Train (TT) approximation to calculate marginal posterior values. Then we can either sample from the conditional posterior using the randomise then optimise (RTO) method or calculate conditional mean and variance using quadrature.



**Figure 4.6:** Directed acyclic graph for ozone retrieval and MTC scheme as described in Fig. 4.1. The hyper-parameters  $\delta$  and  $\gamma$  determine the noise covariance  $\Sigma$  for the random noise vector  $\eta \sim \mathcal{N}(0, \gamma^{-1} \mathbf{I})$  and the prior precision matrix  $Q = \delta \mathbf{L}$  for the distribution over  $x \sim \mathcal{N}(0, \delta \mathbf{L})$ , where  $\mathbf{L}$  is the graph Laplacian, see Eq. 4.5. In the MTC scheme we evaluate the marginal posterior over the hyper-parameters  $\pi(\gamma, \delta | \mathbf{y})$  as in Eq. ?? first and then conditional posterior  $\pi(x | \gamma, \delta, \mathbf{y})$  as in Eq. ???. The parameter  $x$  determine the space of all measurable noise free data  $\Omega$  through the forward model  $\mathbf{A}(x, p, T)$  from which we randomly observe a data set plus some random noise. Note that once we found an affine map we update the forward model to  $\mathbf{MA}_L$ .

The DAG in Fig. 4.6 visualises that process and we can show explicitly that we group the hyper-parameters  $\delta, \gamma$  together to determine the marginal posterior  $\pi(\gamma, \delta | \mathbf{y})$ . Here  $\gamma$ , the noise parameter, determines the noise precision  $\Sigma = \gamma^{-1} \mathbf{I}$  and  $\delta$ , the smoothness parameter, the precision matrix  $Q = \delta \mathbf{L}$  of the prior distribution for  $x$ . Then conditioned on the hyper-parameters the conditional posterior  $\pi(x | \gamma, \delta, \mathbf{y})$  gives the distribution of

posterior ozone profiles. Note that we use the linear model  $A_L$  here as we do not have an approximation to the non-linear model yet and all prior distributions are defined in Table 4.1. The full posterior  $\pi(\mathbf{x}, \gamma, \delta | \mathbf{y}) = \pi(\mathbf{x} | \gamma, \delta, \mathbf{y})\pi(\gamma, \delta | \mathbf{y})$  is given by multiplication of the marginal and conditional posterior densities.

#### 4.3.1 Hyper-parameters samples from the marginal posterior distribution

The marginal posterior distribution

$$\pi(\lambda, \gamma | \mathbf{y}) \propto \lambda^{n/2} \gamma^{m/2} \exp \left\{ -\frac{1}{2}g(\lambda) - \frac{\gamma}{2}f(\lambda) \right\} \pi(\lambda, \gamma), \quad (4.6)$$

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

$$f(\lambda) = \mathbf{y}^T \mathbf{y} - (\mathbf{A}_L^T \mathbf{y})^T (\mathbf{A}_L^T \mathbf{A}_L + \lambda \mathbf{L})^{-1} (\mathbf{A}_L^T \mathbf{y}), \quad (4.7a)$$

$$\text{and } g(\lambda) = \log \det(\mathbf{A}_L^T \mathbf{A}_L + \lambda \mathbf{L}), \quad (4.7b)$$

for the linear model  $A_L$ , see Sec. 3.2.1 for the derivation. To calculate function values more efficiently we approximate the function  $f(\lambda)$  and  $g(\lambda)$  with 3rd order Taylor series around the mode  $\lambda_0$  of  $\pi(\lambda, \gamma | \mathbf{y})$ , since the functions are well behaved over a large range of  $\lambda$  see Fig. 4.7. The derivatives for the Taylor series are

$$f^{(r)}(\lambda_0) = (-1)^{r+1} r! (\mathbf{A}_L^T \mathbf{y})^T (\mathbf{B}_0^{-1} \mathbf{L})^r \mathbf{B}_0^{-1} \mathbf{A}_L^T \mathbf{y} \quad (4.8)$$

$$\text{and } g^{(r)}(\lambda_0) = (-1)^{r+1} \text{tr}((\mathbf{B}_0^{-1} \mathbf{L})^r) \quad (4.9)$$

with  $\mathbf{B}_0 = \mathbf{A}_L^T \mathbf{A}_L + \lambda_0 \mathbf{L}$ . We find the mode at the minimum of  $-\log\{\pi(\lambda, \gamma | \mathbf{y})\}$  using `scipy.optimize.fmin` function and limit the number of function evaluation to 25 and use Cholesky back and forward substitution to calculate values of  $g(\lambda)$  and  $f(\lambda)$ . Additionally, we calculate  $\mathbf{B}_0^{-1} \mathbf{L}$  and  $\mathbf{B}_0^{-1} \mathbf{A}_L^T \mathbf{y}$  once more at  $\lambda_0$  and plot the Taylor approximation within the sampling region in Fig. 4.7.

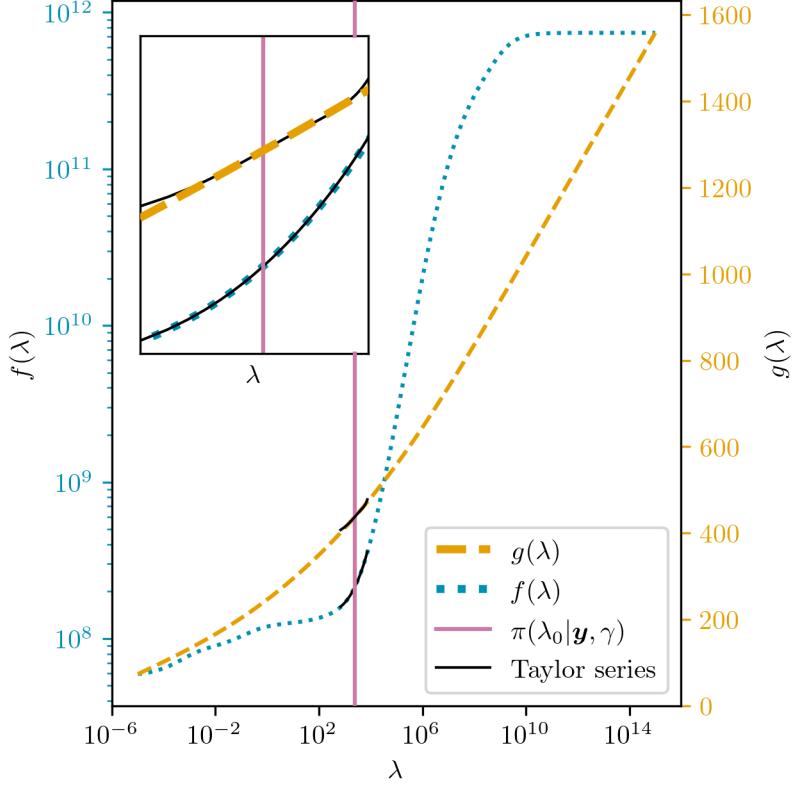
#### Error Analysis?

To characterise the marginal posterior function we employ a Metropolis within Gibbs (MWG) algorithm on  $\pi(\lambda, \gamma | \mathbf{y})$ . In doing so, one may implement a Metropolis random walk on the full conditional

$$\pi(\lambda | \gamma, \mathbf{y}) \propto \lambda^{n/2 + \alpha_\delta - 1} \exp \left\{ -\frac{1}{2}g(\lambda) - \frac{\gamma}{2}f(\lambda) - \beta_\delta \gamma \lambda \right\} \quad (4.10)$$

and do a Gibbs steps on

$$\gamma | \lambda, \mathbf{y} \sim \Gamma \left( \frac{m}{2} + \alpha_\delta + \alpha_\gamma, \frac{1}{2} f(\lambda) + \beta_\gamma + \beta_\delta \lambda \right) \quad (4.11)$$



**Figure 4.7:** 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  $\lambda$  within the MWG algorithm.

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

To run a Metropolis random walk on  $\pi(\lambda|\gamma, \mathbf{y})$  we choose a symmetric proposal distribution  $q(\lambda'|\lambda^{(k)}) \sim \mathcal{N}(\lambda^{(k)}, w_\lambda)$  conditioned on the previous sample  $\lambda^{(k)}$ , with  $k = 1, \dots, N$ . Here  $N$  is the length of the output chain, but to insure that there are no biases due to the initialisation at the mode  $(\lambda^{(0)}, \gamma^{(0)}) = (\lambda_0, \gamma_0)$  of the marginal posterior we discard the samples after a burn-in period  $N_{\text{burn-in}}$ . So the effective output is  $N - N_{\text{burn-in}}$ . To accept or reject a new sample  $\lambda$  we draw a random uniform number in between 0 and 1 and compare with the acceptance ratio

$$\log \left\{ \frac{\pi(\lambda|\gamma^{(t-1)}, \mathbf{y})}{\pi(\lambda^{(t-1)}|\gamma^{(t-1)}, \mathbf{y})} \right\} = \log\{\pi(\lambda|\gamma^{(t-1)}, \mathbf{y})\} - \log\{\pi(\lambda^{(t-1)}|\gamma^{(t-1)}, \mathbf{y})\} \quad (4.12)$$

$$= \frac{n}{2} (\log\{\lambda\} - \log\{\lambda^{(t-1)}\}) + \frac{1}{2} \Delta g + \frac{\gamma^{(t-1)}}{2} \Delta f + \beta_\delta \gamma^{(t-1)} \Delta \lambda, \quad (4.13)$$

which we calculate in the log space. Here  $\Delta f = f(\lambda') - f(\lambda^{(k)}) = \sum f^{(k)}(\lambda_0) \Delta \lambda' - \Delta \lambda^{(k)}$  and  $\Delta g = g(\lambda') - g(\lambda^{(k)}) = \sum g^{(k)}(\lambda_0) \Delta \lambda' - \Delta \lambda^{(k)}$ , where  $\Delta \lambda' = \lambda' - \lambda_0$  and  $\Delta \lambda^{(k)} = \lambda^{(k)} - \lambda_0$ .

Lastly, a Gibbs step provides a new  $\gamma^{(k+1)} \sim \gamma|\lambda^{(k+1)}, \mathbf{y}$ , see Equation (4.11). See Algorithmic Box 2 for summarised version.

**Algorithm 2:** Metropolis within Gibbs for  $\pi(\lambda, \gamma|\mathbf{y})$ 

```

1: Initialise  $\boldsymbol{\theta}^{(0)} = (\lambda^{(0)}, \gamma^{(0)})$  and set burn-in  $N_{\text{burn-in}}$ 
2: for  $k = 1, \dots, N'$  do
3:   Propose  $\lambda \sim \mathcal{N}(\lambda^{(t-1)}, 0.8\lambda_0)$ 
4:   Compute

$$\alpha(\lambda|\lambda^{(t-1)}) = \min \left\{ 1, \frac{\pi(\lambda|\gamma^{(t-1)}, \mathbf{y})}{\pi(\lambda^{(t-1)}|\gamma^{(t-1)}, \mathbf{y})} \right\}$$

5:   Draw  $u \sim \mathcal{U}(0, 1)$ 
6:   if  $\alpha \geq u$  then
7:     Accept and set  $\lambda^{(t)} = \lambda$ 
8:   else
9:     Reject and keep  $\lambda^{(t)} = \lambda^{(t-1)}$ 
10:  end if
11:  Draw  $\gamma^{(t)}|\lambda^{(t)}, \mathbf{y} \sim \text{Gamma}(0.5m + 2, 0.5f(\lambda^{(t)}) + 10^{-10}(1 + \lambda^{(t)}))$ 
12: end for
13: Output:  $(\lambda, \gamma)^{(N_{\text{burn-in}})}, \dots, (\lambda, \gamma)^{(k)}, \dots, (\lambda, \gamma)^{(N)} \sim \pi(\lambda, \gamma|\mathbf{y})$ 
```

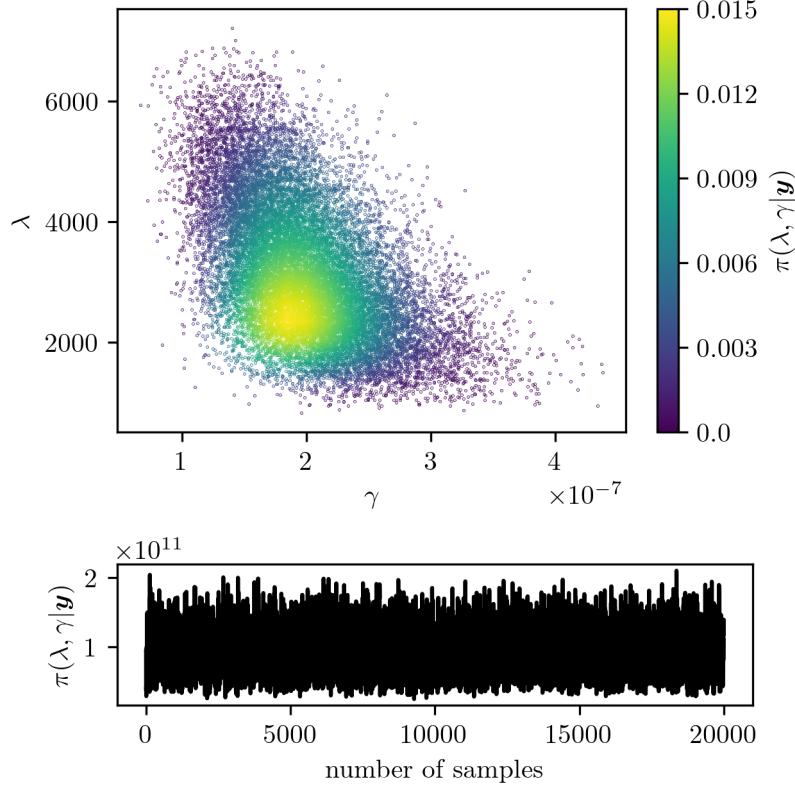
We run the MwG for  $N = 20000$  plus  $N_{\text{burn-in}} = 100$  steps and set the standard deviation of the normal proposal distribution to  $\sigma_\lambda = 0.8\lambda_0$  so that the acceptance rate is  $\approx 0.5$  as suggested in []. The samples are plotted in Fig. 4.8 as a 2D scatter plot, as well as the trace of the MwG to show ergodicity.

Alternatively we can approximate the square root of marginal posterior with a Tensor-Train (TT) on a grid with 40 grid points in each dimension. Here we use the `rect_cross.cross` as a black box algorithm from the `ttypy` python package, based on the rect cross algorithm in []. We set the number of ranks to a constant value equal to four and optimise over those ranks with one sweep. For numerical reasons, to avoid underflow, we have to add a constant  $c = 460$ , such as  $\pi(\lambda|\gamma, \mathbf{y}) = \exp\{\log \pi(\lambda|\gamma, \mathbf{y}) + c\}$ . We calculate the marginals  $\pi(\lambda|\mathbf{y})$  and  $\pi(\gamma|\mathbf{y})$  as in section ??, with a constant  $\gamma =$ , and plot the TT approximation as a colour code on top of the obtained samples in the scatter plot in Fig. 4.8.

On a MacBook Pro from 2019 with 2.4 Ghz quadcore intel core i5 processor it takes  $\lesssim 0.1s$  to find the Tensors to approximate the marginal posterior. In comparisons the to run the MwG takes  $\approx 0.7s$  for  $N = 20000$  effective samples.

### 4.3.2 Ozone samples from the conditional posterior

As part of the MTC scheme we draw ozone samples after determining the marginal posterior distribution. In this section we present two ways to draw ozone samples



**Figure 4.8:** We scatter plot the samples of  $\lambda = \delta/\gamma$  and  $\gamma$  from the marginal posterior  $\pi(\lambda, \gamma|\mathbf{y})$  and colour code the samples using the TT approximation of  $\pi(\lambda, \gamma|\mathbf{y})$ . To show ergodicity we plot the trace of the samples of the Metropolis-within-Gibbs sampler below.

from the conditional posterior

$$\mathbf{x}|\delta, \gamma, \mathbf{y} \sim \mathcal{N}\left(\underbrace{(\mathbf{A}_L^T \mathbf{A}_L + \delta/\gamma \mathbf{L})^{-1} \mathbf{A}_L^T \mathbf{y}}_{\mathbf{x}_\lambda}, \underbrace{(\gamma \mathbf{A}_L^T \mathbf{A}_L + \delta \mathbf{L})^{-1}}_{\gamma \mathbf{B}_\lambda}\right). \quad (4.14)$$

First we present the Randomize-than-Optimize (RTO) method [8, 17, 34], as previously described in Sec. 3.4.2. Alternatively we can integrate over the marginal posterior and calculate the mean and variance of the multivariate normal conditional posterior distribution. Note that calculating the covariance can be expensive and if that is the case the RTO methods is the preferred choice.

### Randomize then optimize – RTO

For the RTO method we start by drawing an independent hyper-parameter sample  $(\delta, \gamma) \sim \pi(\delta, \gamma|\mathbf{y})$  from the samples of the MwG. Then we generate two independent Gaussian random variables  $\mathbf{v}_1 \sim \mathcal{N}(\mathbf{0}, \gamma \mathbf{A}_L^T \mathbf{A}_L)$  and  $\mathbf{v}_2 \sim \mathcal{N}(\mathbf{0}, \delta \mathbf{L})$ . Here can use Cholesky factorisation of  $\mathbf{L}$  and the multiplication rule for normal distributions so that

$\mathbf{v}_1 \sim \sqrt{\delta} \mathbf{L}_C \mathcal{N}(0, \mathbf{I})$  and  $\mathbf{v}_2 \sim \sqrt{\gamma} \mathbf{A}_L^T \mathcal{N}(0, \mathbf{I})$ . Then we solve

$$(\gamma \mathbf{A}_L^T \mathbf{A}_L + \delta \mathbf{L}) \mathbf{x} = \gamma \mathbf{A}_L^T \mathbf{y} + \mathbf{v}_1 + \mathbf{v}_2, \quad (4.15)$$

using Cholesky back and forward substitution, for  $\mathbf{x}$  and obtain one independent sample of  $\pi(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$ . See Fig. 4.9, where we plot  $m$  = samples of the conditional posterior.

### Weighted Mean

Alternatively, we can calculate the mean

$$\mu_{\mathbf{x}|\mathbf{y}} = \int \mathbf{x}_\lambda \pi(\lambda|\mathbf{y}) d\lambda \approx \sum \mathbf{x}_{\lambda_i} \pi(\lambda_i|\mathbf{y}), \quad (4.16)$$

and covariance

$$\Sigma_{\mathbf{x}|\mathbf{y}} = \int \gamma^{-1} \pi(\gamma|\mathbf{y}) d\gamma \int \mathbf{B}_\lambda^{-1} \pi(\lambda|\mathbf{y}) d\lambda \approx \sum \gamma_i^{-1} \pi(\gamma_i|\mathbf{y}) \sum \mathbf{B}_{\lambda_i}^{-1} \pi(\lambda_i|\mathbf{y}) \quad (4.17)$$

of the conditional posterior  $\pi(\mathbf{x}|\delta, \gamma, \mathbf{y})$ , see Eq. 4.14, by quadrature [35, Sec. 2.1] We get function values for the marginal posterior either by binning the samples in Fig. 4.8 into a normalised histogram and use the height of the bars as quadrature weights, e.g.  $\pi(\lambda_i|\mathbf{y})$ , where  $\lambda_i$  is at the centre of each bin, or from the TT approximation of  $\sqrt{\pi(\gamma, \delta|\mathbf{y})}$ , see also Fig. 4.12. The integral can be interpreted as the weighted average, with  $\sum \pi(\lambda_i|\mathbf{y}) = 1$  for normalised marginal functions.

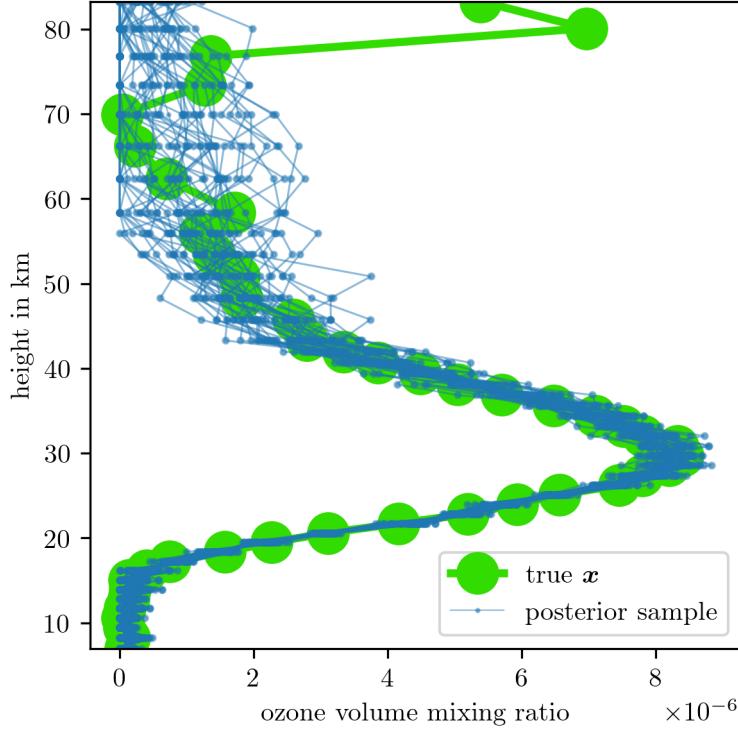
If using the samples from the MWG algorithm, we start by binning the samples into 3 bins and stop increasing the number of bins by one if the relative error between the previous and the current conditional posterior mean is less than 0.1%. This happens at a bin number of 5 and gives a total number of  $3 + 4 + 5 = 12$  solves of  $x_\lambda$  to find the conditional mean  $\mu_{\mathbf{x}|\mathbf{y}}$ . To calculate calculate the covariance matrix  $(\gamma \mathbf{B}_\lambda)^{-1}$ , we use Cholesky forward and backward substitution to invert  $\mathbf{B}_\lambda$ , solve the integral in Eq. 4.17 5 times.

Using the TT approximation we consider every second grid point of the total 40 grid points of  $\pi(\lambda|\mathbf{y})$  and  $\pi(\gamma|\mathbf{y})$ , which means we solve the integrals in Eq. 4.16 and 4.17 20 times. We plot the mean and variance in Figure 4.9.

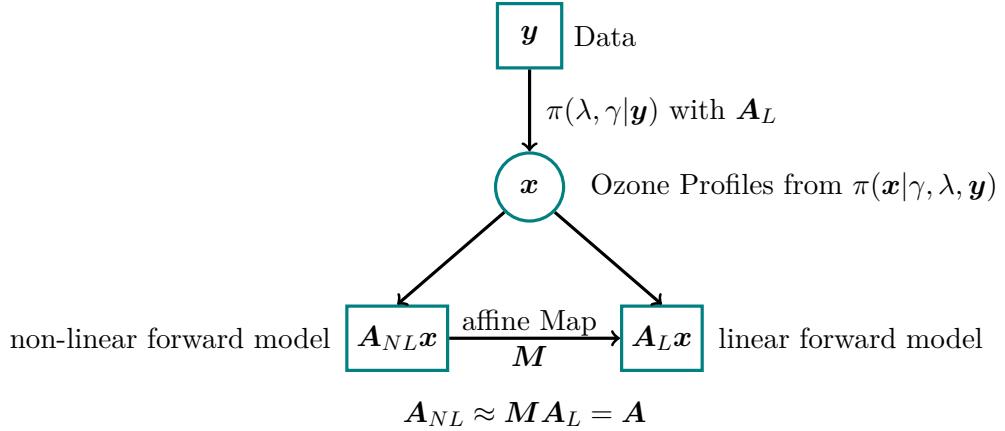
## 4.4 Approximate non-linear forward model with affine Map

With the samples  $\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(m)}\} \sim \pi(\mathbf{x}|\gamma, \lambda, \mathbf{y})$  from the conditional posterior we generate two affine spaces

$$\mathbf{V} = \begin{bmatrix} | & | & | \\ \mathbf{A}_{NL}(\mathbf{x}^{(1)}) & \cdots & \mathbf{A}_{NL}(\mathbf{x}^{(j)}) & \cdots & \mathbf{A}_{NL}(\mathbf{x}^{(m)}) \\ | & & | & & | \end{bmatrix} = \begin{bmatrix} | & v_1 & | \\ & \vdots & \\ | & v_j & | \\ & \vdots & \\ | & v_m & | \end{bmatrix}, \quad (4.18)$$



**Figure 4.9:** We draw samples from the conditional posterior distribution  $\pi(\mathbf{x}|\lambda, \gamma, \mathbf{y})$  after characterising the marginal posterior  $\pi(\lambda, \gamma|\mathbf{y})$  through sampling or TT approximation using the linear forward map  $\mathbf{A}_L$ . We will use those samples to find the affine map  $\mathbf{M}$ , see section 4.4



**Figure 4.10:** 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.

using the non-linear forward model, and

$$\mathbf{W} = \begin{bmatrix} | & | & | \\ \mathbf{A}_L\mathbf{x}^{(1)} & \dots & \mathbf{A}_L\mathbf{x}^{(j)} & \dots & \mathbf{A}_L\mathbf{x}^{(m)} \\ | & | & | \end{bmatrix} \quad (4.19)$$

*DRAFT Printed on May 27, 2025*

based on the linear forward model, which are matrices in  $\mathbb{R}^{m \times m}$ . To find the affine map

$$\mathbf{M} = \begin{bmatrix} - & r_0 & - \\ & \vdots & \\ - & r_j & - \\ & \vdots & \\ - & r_m & - \end{bmatrix} \in \mathbb{R}^{m \times m}, \quad (4.20)$$

we use the `numpy.linalg.solve` python function to solve

$$v_j = r_j \mathbf{W} \quad (4.21)$$

where  $r_j$  is the j-th row of  $\mathbf{M}$ .

Finally, we approximate the non-linear forward model

$$\mathbf{A}_{NL} \approx \mathbf{M} \mathbf{A}_L = \mathbf{A}, \quad (4.22)$$

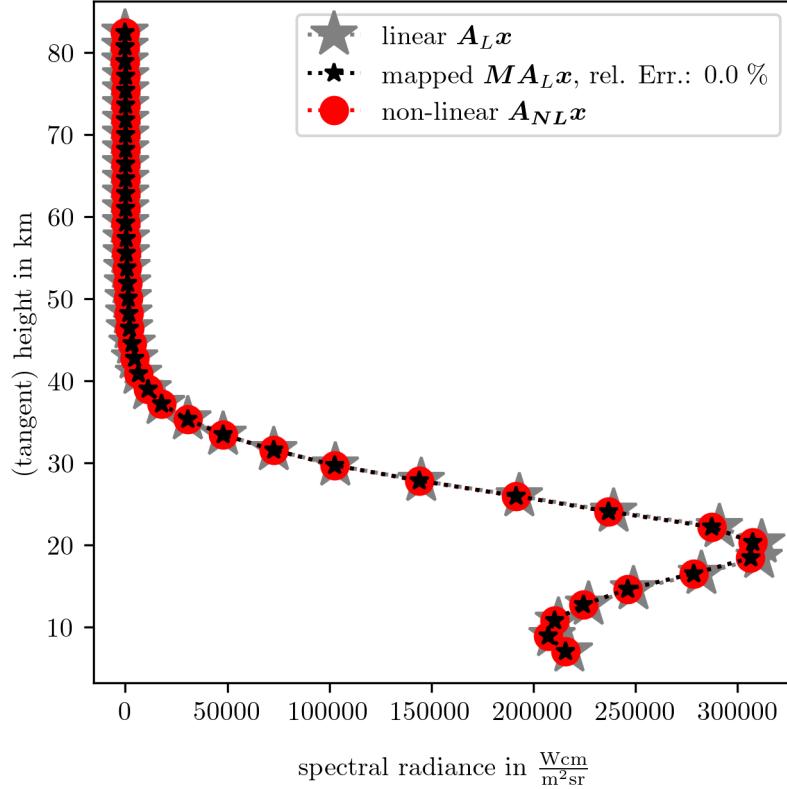
with the affine map  $\mathbf{M}$  and the linear forward model  $\mathbf{A}_L$ . In Fig. 4.11 we asses the affine map using one of the samples  $\mathbf{x} \sim \pi(\mathbf{x} | \gamma, \lambda, \mathbf{y})$  from the conditional posterior and calculate the relative error  $\|\mathbf{M} \mathbf{A}_L \mathbf{x} - \mathbf{A}_{NL} \mathbf{x}\| / \|\mathbf{M} \mathbf{A}_L \mathbf{x}\|$  in percent between the mapped noise free data and the noise free data based on the non-linear forward. As displayed in Fig. 4.11 we can approximate the non-linear forward model well **within the noise, which is??** and from here on will use the approximated forward map.

## 4.5 Posterior distributions with approximated non-linear model for Ozone – MTC

From now on updaeted forward map

### 4.5.1 Hyper-parameters samples from the marginal posterior distribution

- make table with set up for TT and sampling , number of samples
- normalize in evrey step
- we use RTO to generate samples
- samples vs calc values
- taylor expansion



**Figure 4.11:** 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.

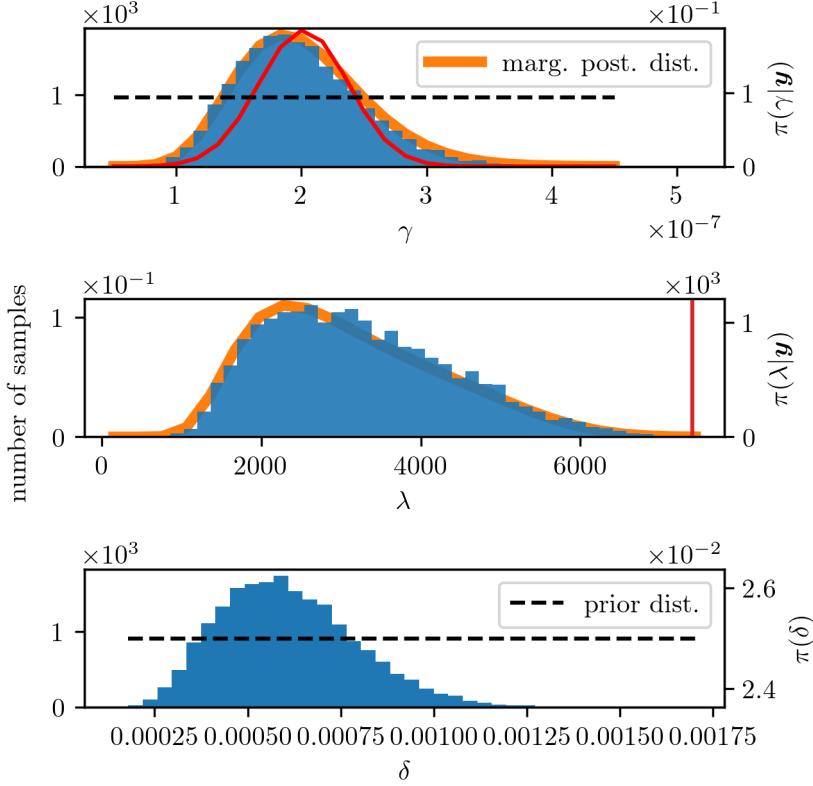
#### 4.5.2 Ozone samples from the conditional posterior and regularised solution

- draw samples using RTO no need to calc variance yet, when variance is hard to calculate, we use Cholesky defactorization (maybe put in appendix)
- or calc mean and variance using see other section

#### 4.5.3 Regularized Solution

- similar to MTC model
- picture of L-Curve and samples
- include in reg paraemter in histograms

It takes us less than 0.1s to draw 10000 samples from the marginal posterior and to calculate the conditional posterior mean or to solve for 200 different  $x_\lambda$  and to find the



**Figure 4.12:** 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  $\lambda$  and  $\gamma$  using the Metropolis-within-Gibbs sampler and can calculate  $\delta$  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  $\lambda_{\text{reg}} =$ .

regularised solution at the point of maximum curvature using the kneedle algorithm.

## 4.6 Solution by regularization

The Tikhonov regularized solution is defined as [36]

$$\mathbf{x}_\lambda = \arg \min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 + \lambda \mathbf{x}^T \mathbf{L} \mathbf{x}, \quad (4.23)$$

which is equivalent to minimizing the term  $T(\mathbf{x})$  in Equation (??) with  $\lambda = \delta/\gamma$ . This maximizes the full conditional distribution for  $\mathbf{x}$ , so is not, as often erroneously stated, the maximum a posteriori (MAP) estimate. The regularized solution is typically calculated by solving the normal equations

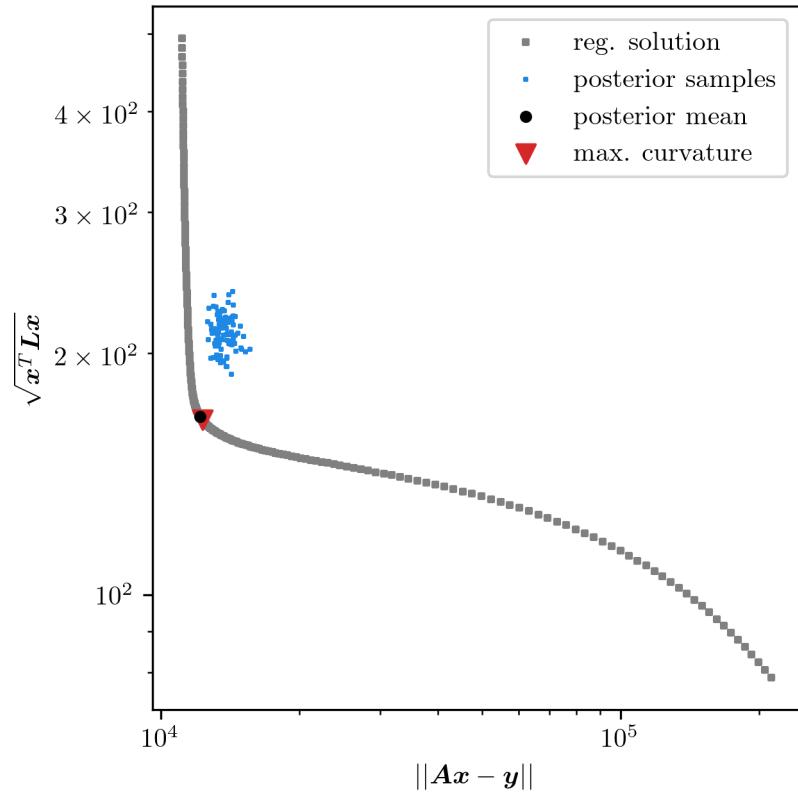
$$\mathbf{x}_\lambda = (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{L})^{-1} \mathbf{A}^T \mathbf{y} \quad (4.24)$$

with a separate procedure used to determine the regularizing parameter  $\lambda$ ; in Section 4.6.1 we demonstrate the L-curve method [15].

### 4.6.1 Regularization

According to [15], the regularized solution corresponds to the corner on the L-curve, the point of maximum curvature. As done in [15], we compute  $\mathbf{x}_\lambda$ , see Equation (4.24), for 200 different  $\lambda$  values and plot the solution semi norm  $\sqrt{\mathbf{x}_\lambda^T \mathbf{L} \mathbf{x}_\lambda}$  against the data misfit norm  $\|\mathbf{A}\mathbf{x}_\lambda - \mathbf{y}\|$ , see Figure ???. We find the point of maximum curvature with the kneedle algorithm [37] using the python function `kneed.KneeLocator()` on the zoomed-in region in Figure ???. The point of maximum curvature is marked with a triangle in Figure ?? and represents the regularized solution corresponding to  $\lambda_{\text{opt}} = 7316$ .

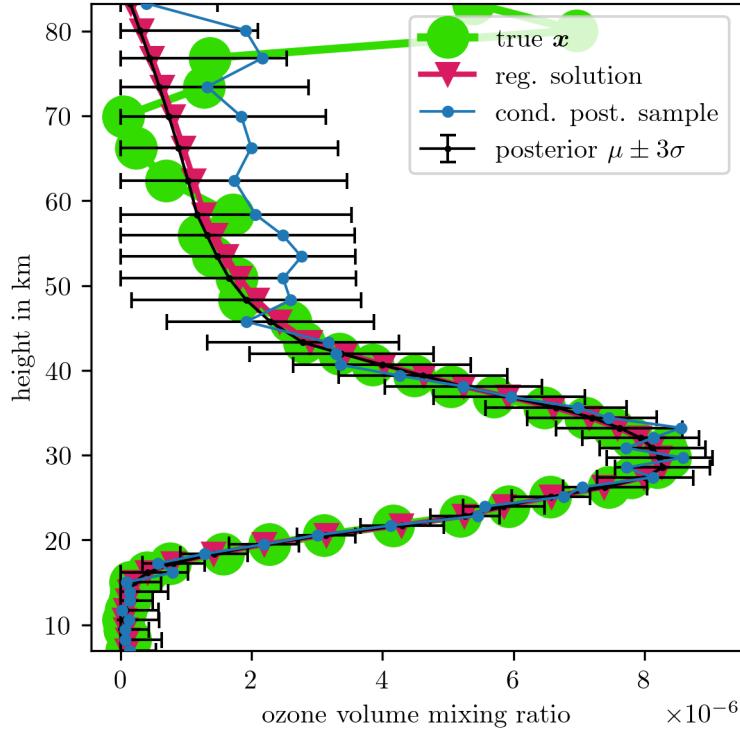
range of lambda Kneedle algorithm



**Figure 4.13:** We calculate regularised solution as in Eq. ?? and plot the regularised semi norm  $\sqrt{\mathbf{x}^T \mathbf{L} \mathbf{x}}$  against the data misfit norm  $\|\mathbf{A}\mathbf{x} - \mathbf{y}\|$  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**

## 4.7 Posterior Pressure and Temperature

- make table with set up for TT and sampling t-walk

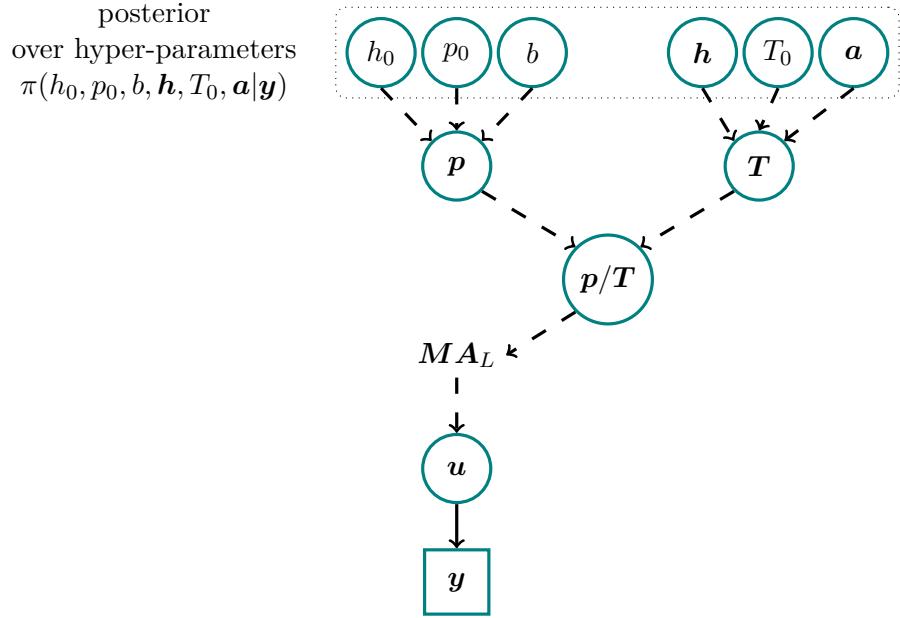


**Figure 4.14:** We plot the conditional posterior mean and variance in black and the regularised solution on top of the ground truth ozone profile in green. We use the updated forward map  $\mathcal{M}_L$

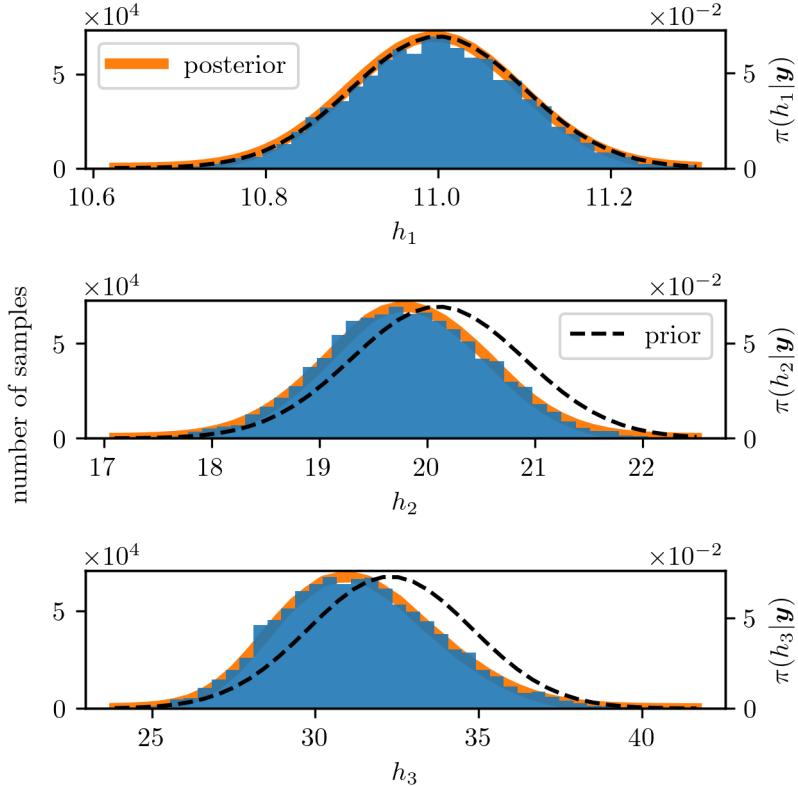
On a MacBook Pro from 2019 with 2.4ghz quadcore intel core i5 it takes TT takes  $\approx 2$  to max  $4\min \epsilon = 1.5e - 1$  or max number of sweeps 11 where one sweep is left to right to left conditioned on an ozone sample t-walk burnIn 1000 samples samples 1000000 takes 5 – 6 mins

$$\begin{aligned} \pi(h_1, p_0, b_1, b_2, \mathbf{h}_T, \mathbf{c}_T, \mathbf{a}_T | \mathbf{y}, \gamma, \mathbf{x}) &\propto \exp \left\{ -\frac{\gamma}{2} \left\| \mathbf{y} - \mathbf{A} \frac{\mathbf{p}}{T} \right\|_2^2 + \ln \pi(h_1, p_0, b_1, b_2, \mathbf{h}_T, \mathbf{c}_T, \mathbf{a}_T) \right\} \\ \pi(h_1, p_0, b_1, b_2, \mathbf{h}_T, \mathbf{c}_T, \mathbf{a}_T | \mathbf{y}) &\propto \pi(\mathbf{y} | h_1, p_0, b_1, b_2, \mathbf{h}_T, \mathbf{c}_T, \mathbf{a}_T) \pi(h_1, p_0, b_1, b_2, \mathbf{h}_T, \mathbf{c}_T, \mathbf{a}_T) \end{aligned}$$

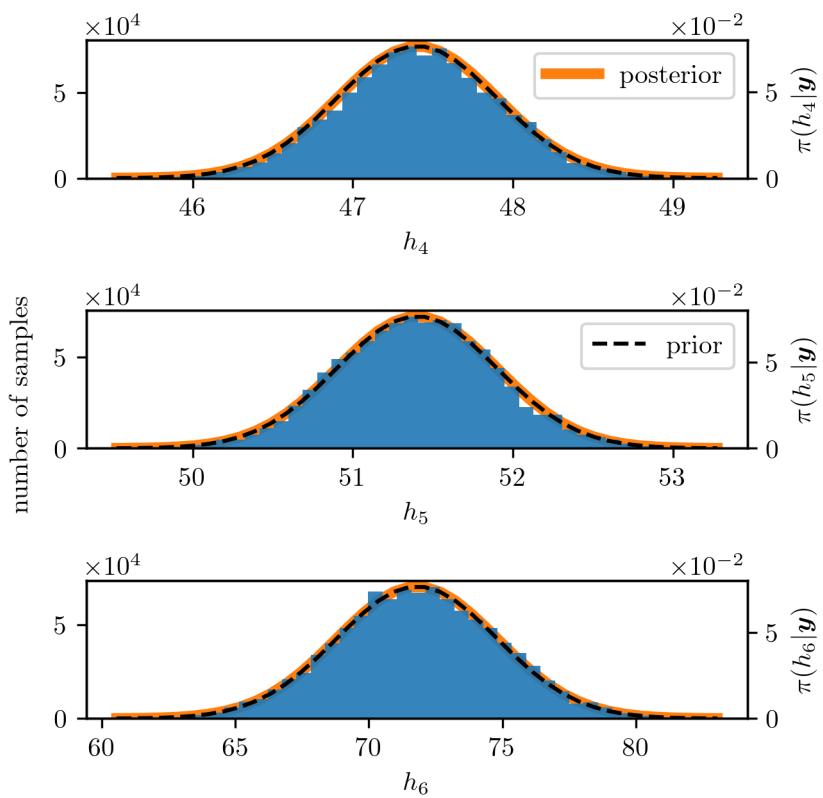
TT set up



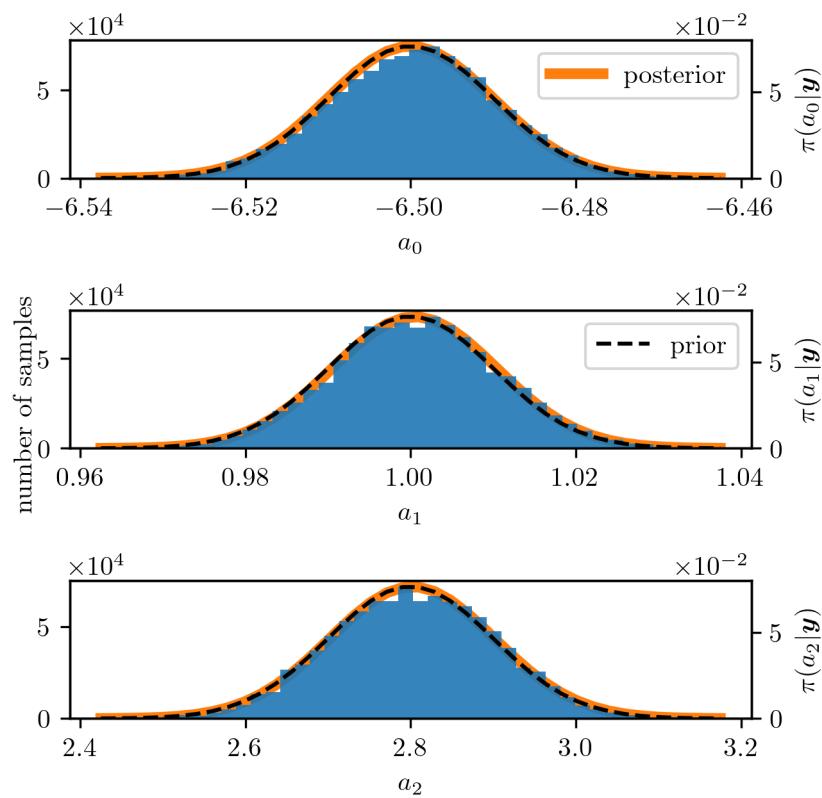
**Figure 4.15:** Conditioned on an ozone profile the posterior of the hyper-parameters describing pressure and temperature is given as in Eq. ???. Since pressure and temperature go into the forward model as  $p/T$  they are highly correlated but the pressure is the dominant parameter, see Fig. 4.3 and D.5. Note that here we use the updated forward model  $MA_L$  and conditioned on a  $\gamma$  sample from the previously evaluated marginal posterior see Fig. 4.12.



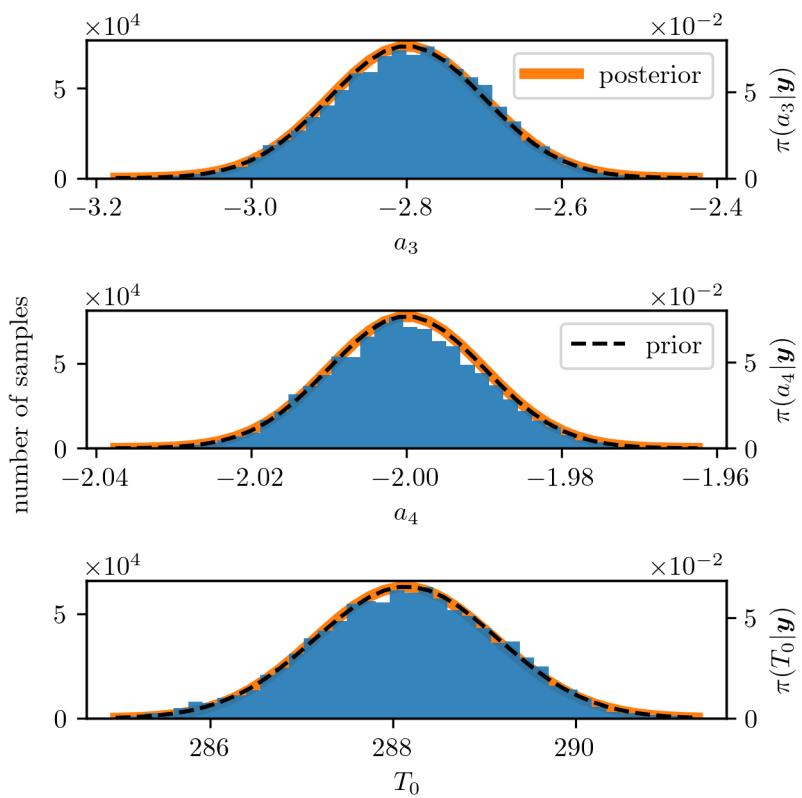
**Figure 4.16:** We plot the TT approximation of marginal posterior in orange and the samples as a histogram as well as the prior distribution with a dotted line.



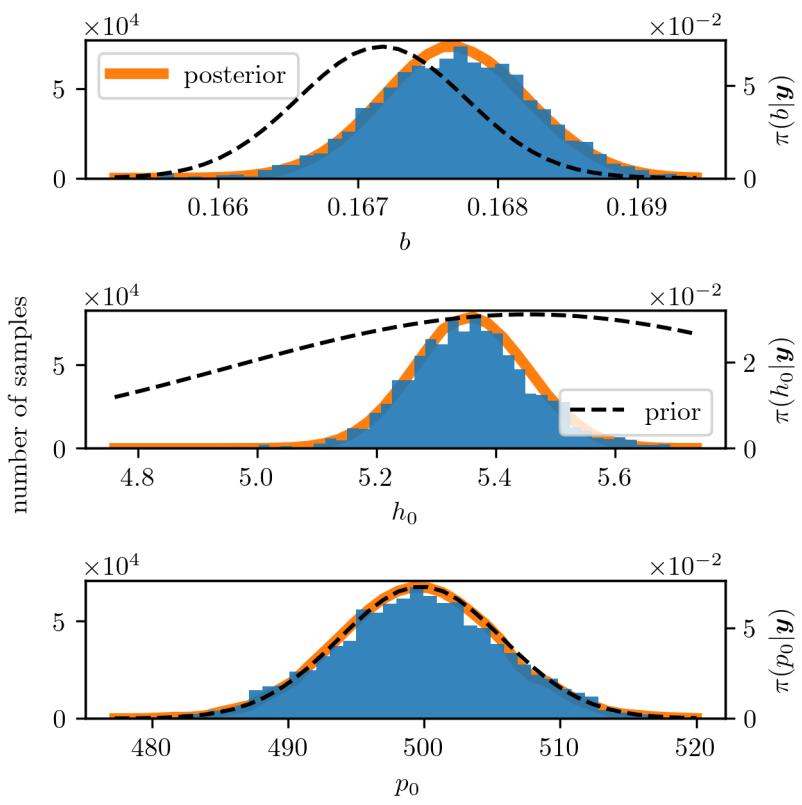
**Figure 4.17:** We plot the TT approximation of marginal posterior in orange and the samples as a histogram as well as the prior distribution with a dotted line.



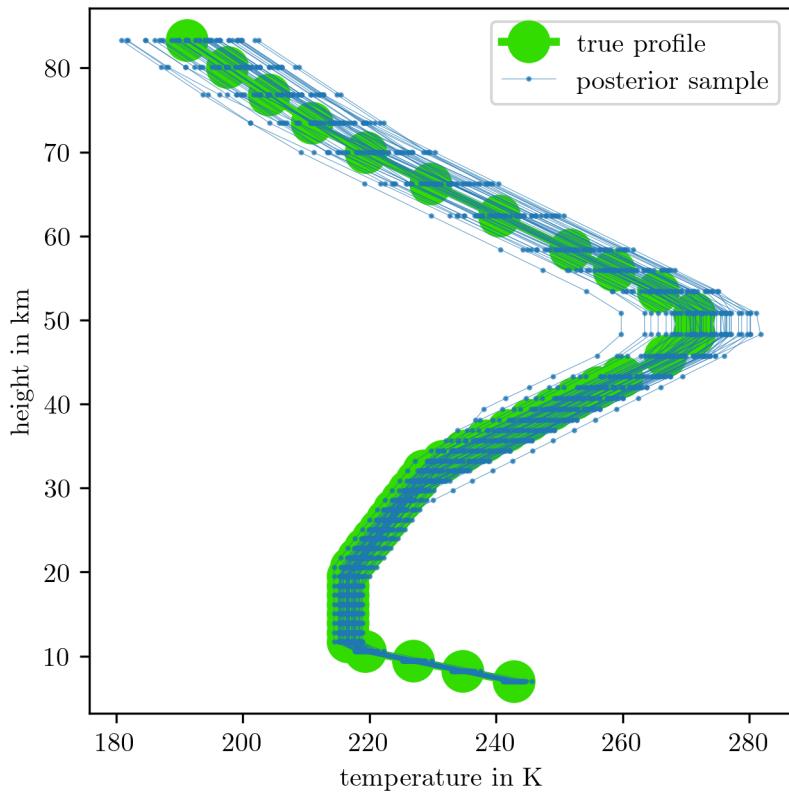
**Figure 4.18:** We plot the TT approximation of marginal posterior in orange and the samples as a histogram as well as the prior distribution with a dotted line.



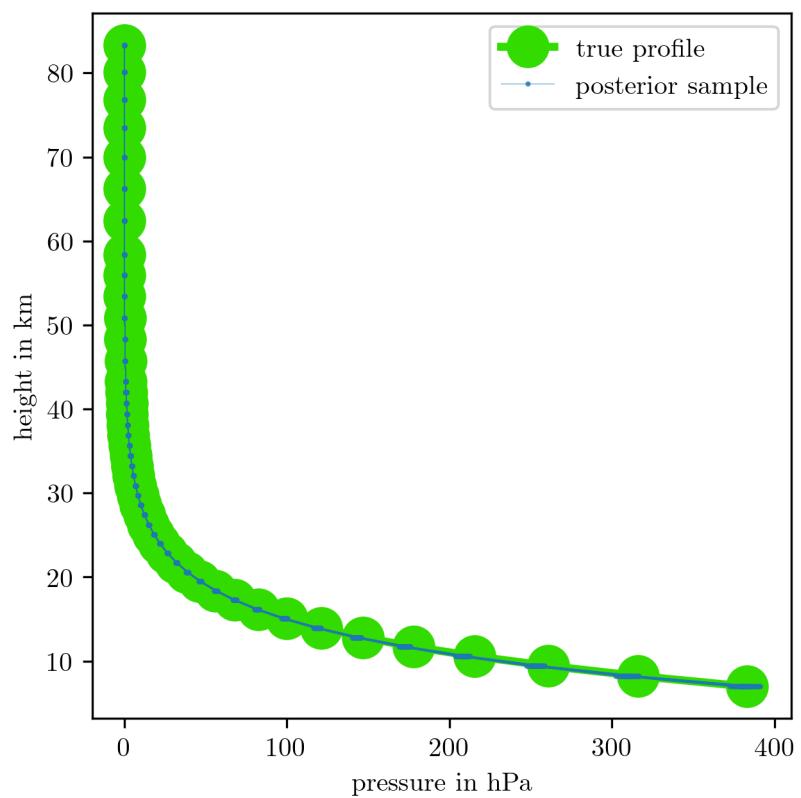
**Figure 4.19:** We plot the TT approximation of marginal posterior in orange and the samples as a histogram as well as the prior distribution with a dotted line.



**Figure 4.20:** 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 4.21:** We take samples from the posterior distribution, as plotted in Figures 4.16 to 4.19 and plot the corresponding temperature function, see Eq: 4.3.



**Figure 4.22:** We take samples from the posterior distribution, as plotted in Fig. 4.20 and plot the corresponding pressure function, see Eq: 4.4.



# Appendices



# A

## Additional MCMC analysis

- integrated autocorrelation time, cite Ulli Wolf [38] python and matlab
- acceptance rate for one D and multi D, cite roberts



# B

## Correlation Structure

In the book Gaussian Markov Random Fields [12], Rue and Held demonstrate that a strong correlation between the hyper-parameter  $\mu$  and the latent field  $\mathbf{x}$  can significantly slow down convergence when using samplers, in particular Gibbs samplers. They consider the hierarchical model

$$\mu \sim \mathcal{N}(0, 1) \quad (\text{B.1a})$$

$$\mathbf{x}|\mu \sim \mathcal{N}(\mu \mathbf{1}, \mathbf{Q}^{-1}), \quad (\text{B.1b})$$

and apply a Gibbs sampler based on the full conditional distributions

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

$$\mathbf{x}^{(k)} | \mu^{(k)} \sim \mathcal{N}(\mu^{(k)} \mathbf{1}, \mathbf{Q}^{-1}). \quad (\text{B.3})$$

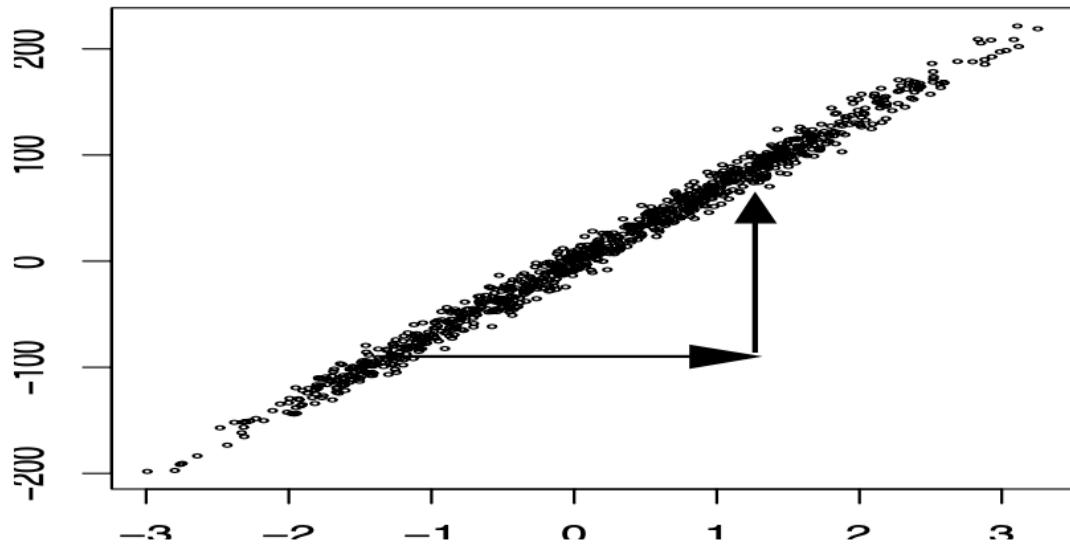
As illustrated in Figure B.1, when the sampler is restricted to steps only in the  $\mu$ -direction (horizontal axis) or the  $\mathbf{x}$ -direction (vertical axis), it requires many iterations to adequately explore the parameter space. This inefficiency arises from the high correlation between  $\mu$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

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

$$\mu^* \sim q(\mu^* | \mu^{(k-1)}) \quad (\text{B.4})$$

$$\mathbf{x}^{(k)} | \mu^* \sim \mathcal{N}(\mu^* \mathbf{1}, \mathbf{Q}^{-1}) \quad (\text{B.5})$$

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



**Figure B.1:** The figure taken from [12, Figure 4.1 (b)], shows samples from a marginal chain for  $\mu$  and  $\mathbf{1}^T \mathbf{Q} \mathbf{x}^{(k)}$  over 1000 iterations, based on the hierarchical model in Eq. B.1, with an autoregressive process encoded in  $\mathbf{Q}$ . The algorithm updates  $\mu$  and  $\mathbf{x}$  successively from their full conditional distributions. The plot displays  $(\mu^{(k)}, \mathbf{1}^T \mathbf{Q} \mathbf{x}^{(k)})$ , with  $\mu^{(k)}$  on the horizontal axis and  $\mathbf{1}^T \mathbf{Q} \mathbf{x}^{(k)}$  on the vertical axis. The slow mixing and convergence of  $\mu$  result from its strong dependence on  $\mathbf{1}^T \mathbf{Q} \mathbf{x}^{(k)}$ , 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  $\Omega$  denotes the sample space, and  $\mathcal{F}$  is a collection of countable subsets  $\{A_n\}_{n \in \mathbb{N}}$  of  $\Omega$ . Each  $A_n \subseteq \Omega$  is called an event, and a map  $\mathbb{P} : \mathcal{F} \rightarrow \mathbb{R}$  is referred to as a measure. In the following, we describe the conditions required for  $\mathcal{F}$  to be a  $\sigma$ -algebra, and for  $\mathbb{P}$  to qualify as a probability measure. We refer to [39] [21] 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 $\sigma$ -algebra

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

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

In other words, the empty set  $\emptyset$  and the entire sample space  $\Omega$  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  $A_1, A_2, \dots \in \mathcal{F}$ , 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  $\sigma$ -algebra.

# D

prior modelling

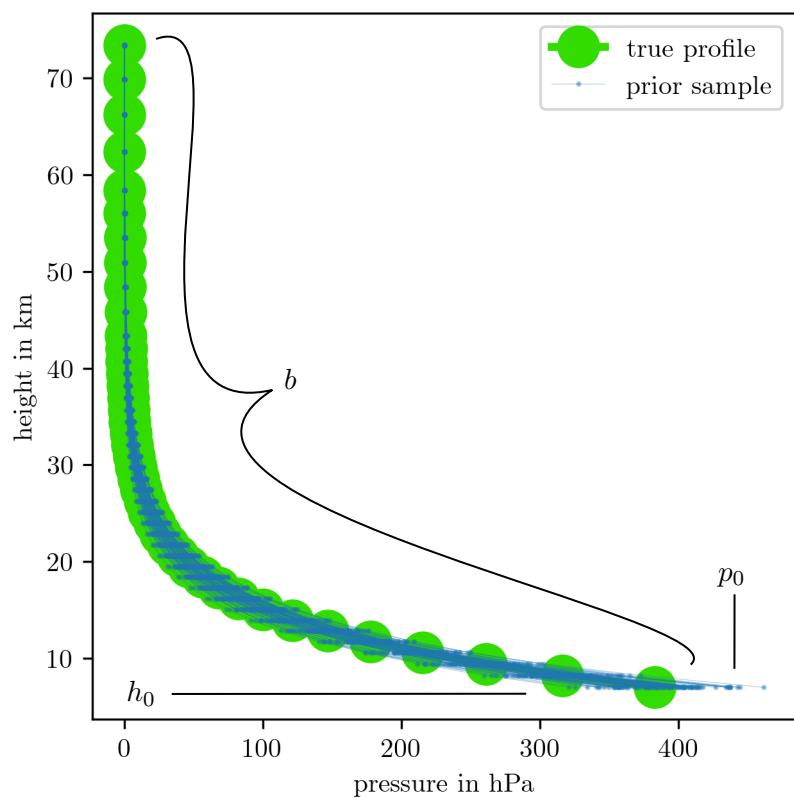


Figure D.1

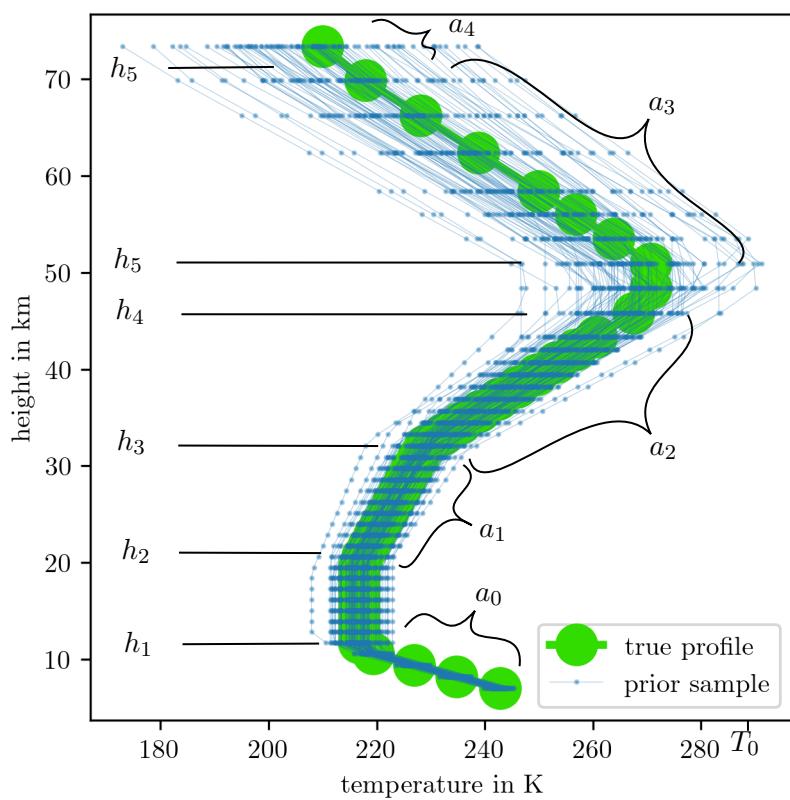


Figure D.2

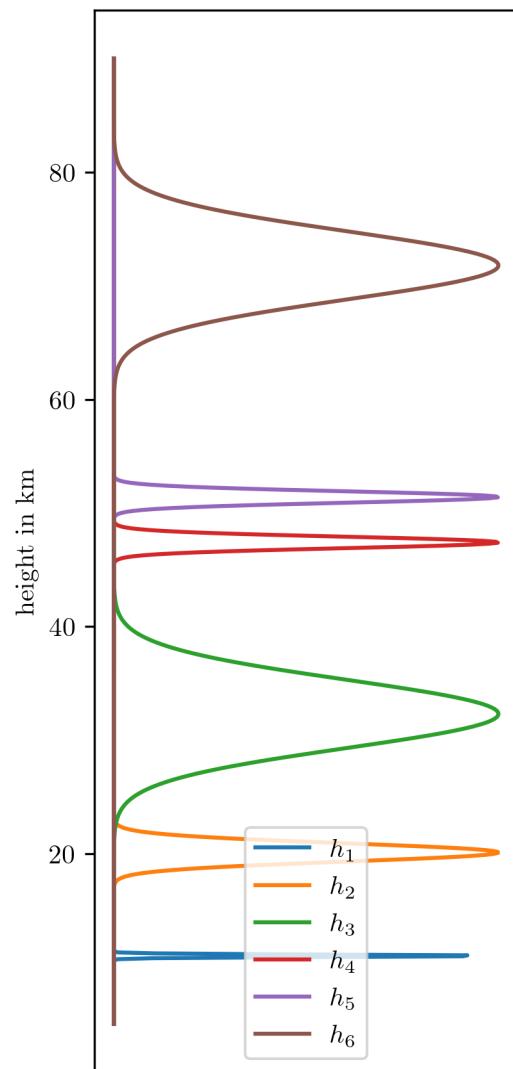
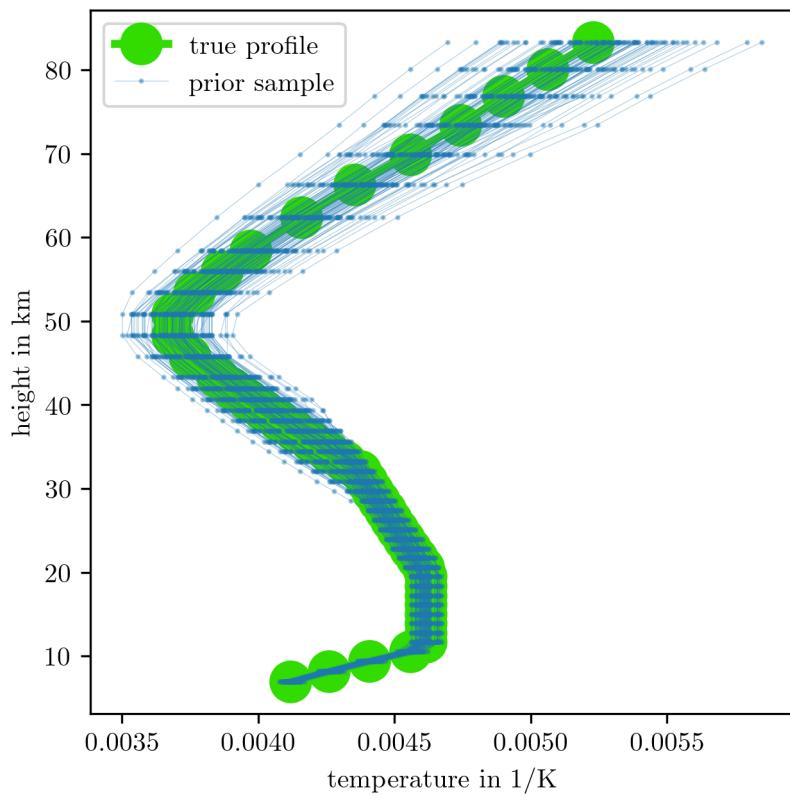


Figure D.3

**Figure D.4**

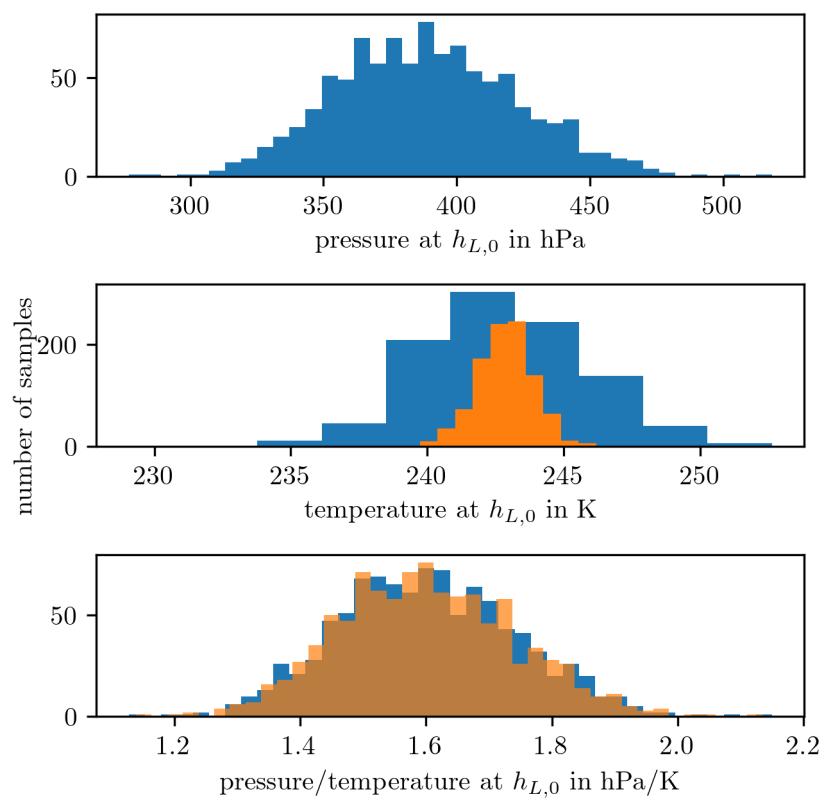


Figure D.5



## 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] C. Readings. *Envisat MIPAS An Instrument for Atmospheric Chemistry and Climate Research*. Noordwijk: ESA Publications Division, 2000.
- [3] Ioulia E Gordon et al. “The HITRAN2020 molecular spectroscopic database”. In: *Journal of Quantitative Spectroscopy and Radiative Transfer* 277 (2022), p. 107949.
- [4] 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.
- [5] George B. Rybicki and Alan P. Lightman. *Radiative processes in Astrophysics*. Weinheim: Wiley-VCH, 2004.
- [6] Marcel Berger. *Geometry I. 4th Edition*. Berlin Heidelberg: Springer-Verlag, 2009.
- [7] Katsumi Nomizu and Takeshi Sasaki. *Affine differential geometry*. Cambridge: Cambridge University Press, 1994.
- [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. “General state space Markov chains and MCMC algorithms”. In: *Probability Surveys* 1 (2004), pp. 20–71.
- [10] S. P. Meyn and R.L. Tweedie. *Markov Chains and Stochastic Stability. 2nd Edition*. New York: Cambridge University Press, 2009.
- [11] Charles J Geyer. “Practical markov chain monte carlo”. In: *Statistical science* (1992), pp. 473–483.
- [12] Havard Rue and Leonhard Held. *Gaussian Markov random fields: theory and applications*. London: CRC press, 2005.
- [13] Charles W. Champ and Andrew V. Sills. “The Generalized Law of Total Covariance”. In: *preprint* (2022). URL: <https://arxiv.org/abs/2205.14525>.
- [14] Jari P. Kaipio and Erkki Somersalo. *Statistical and Computational Inverse Problems*. New York: Springer-Verlag New York, 2005.
- [15] 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.
- [16] 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.
- [17] Johnathan M Bardsley. “MCMC-based image reconstruction with uncertainty quantification”. In: *SIAM Journal on Scientific Computing* 34.3 (2012), A1316–A1332.
- [18] Johnathan M Bardsley et al. “Randomize-then-optimize: A method for sampling from posterior distributions in nonlinear inverse problems”. In: *SIAM Journal on Scientific Computing* 36.4 (2014), A1895–A1910.

- [19] Felipe Acosta, Mark L Huber, and Galin L Jones. "Markov chain Monte Carlo with linchpin variables". In: *preprint* (2014). URL: <https://arxiv.org/abs/2205.14525>.
- [20] 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.
- [21] M. Capiński and P.E. Kopp. *Measure, Integral and Probability. Springer Undergraduate Mathematics Series*. London: Springer-Verlag London, 2004.
- [22] M. Simonnet. *Measures and Probabilities*. New York: Springer-Verlag, 1996.
- [23] Vesa Kaarnioja. *Inverse Problems. Eighth lecture*. <https://vesak90.userpage.fu-berlin.de/ip23/week8.pdf>. [Online; accessed 10/04/25]. 2023.
- [24] 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.
- [25] Philip J Davis and Philip Rabinowitz. *Methods of numerical integration*. San Diego, CA: Academic Press, Inc., 1984.
- [26] 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.
- [27] 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.
- [28] Ivan V Oseledets. "Tensor-train decomposition". In: *SIAM Journal on Scientific Computing* 33.5 (2011), pp. 2295–2317.
- [29] Schwartz M. et al. *MLS/Aura Level 2 Ozone (O3) Mixing Ratio V005*. [https://disc.gsfc.nasa.gov/datasets/ML203\\_005/summary?keywords=mls%20o3](https://disc.gsfc.nasa.gov/datasets/ML203_005/summary?keywords=mls%20o3). [Online; accessed 25/04/24]. 2020.
- [30] *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.
- [31] 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.
- [32] J. Andrés Christen and Colin Fox. *The t-walk software*. <https://www.cimat.mx/~jac/twalk/>. [Online; accessed 25/11/24].
- [33] Colin Fox. "Sampling and modelling of LARGE Gaussian processes". 9th World Meeting of International Society for Bayesian Analysis, 21st to 25th July 2008. Hamilton Island, Queensland, Australia. 2008. URL: [urlinktotalkabstractifany](http://urlinktotalkabstractifany).
- [34] 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.
- [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] Ulli Wolff. *UWerr.m Version6*. <https://www.physik.hu-berlin.de/de/com/ALPHAssoft>. [Online; accessed 5/11/23]. 2004.

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