

Hierarchical Bayesian Modelling and Posterior  
Inference Applied to an Atmospheric  
Limb-Sounder Measuring Ozone



University  
of Otago

ŌTĀKOU WHAKAIHU WAKA

NEW ZEALAND

Lennart Golks  
Department of Physics

A thesis submitted for the degree of  
*Doctor of Philosophy*

October 2025



## Acknowledgements



# Abstract

In this thesis, we develop a hierarchical Bayesian model based on the radiative transfer equation for an atmospheric limb-sounder targeting ozone, as described in [39]. To ensure effective measurements, we briefly assess the informativity of the forward model for different measurement approaches and adapt the data collection accordingly. Following [17], we utilise the marginal and then conditional scheme to provide posterior distributions of hyper-parameters and ozone profiles, and compare with a regularisation approach. After approximating the non-linear forward model with an affine map, we extend our model and the marginal and then conditional scheme to jointly infer posterior pressure, temperature and ozone.

The main contribution of this work is the application of tensor-trains to approximate high-dimensional posterior distributions [8, 11]. This enables us to generate samples from the target distribution with far fewer function evaluations compared to the t-walk sampling algorithm [6]. Tensor-train methods require a predefined grid and a “normalisation constant” so that function outputs are within computer precision, but once defined, they reduce the function evaluations per independent sample significantly. Another advantage of the tensor-train format is that marginal distributions, useful for characterisation of integrals via quadrature, can be calculated at a low computational cost, without any sampling. To further improve tensor-train methods, we suggest future work should focus on lowering tensor ranks, calculating “normalisation constants” to avoid numerical issues and reducing correlation structures between parameters automatically, all of which we currently have to do by exploratory analysis. Additionally, choosing accurate interpolation schemes between grid points is crucial to improving the effectiveness of the approximation.

Our results show that a hierarchical Bayesian approach, which quantifies posterior mean and variance of the parameter (ozone), provides more information than a regularisation approach at comparable computational time. In regions where the signal strength is low and the data is noise-dominated, we can not recover ozone structures from the ground truth. When including pressure and temperature describing hyper-parameters within our hierarchical Bayesian model, we find a strong correlation between ozone and pressure, whereas the model and data are uninformative about temperature. For future work, we recommend developing a more physically informed parametrised model for ozone within the atmosphere, incorporating atmospheric chemistry and other important processes.



# Contents

<b>List of Figures</b>	<b>ix</b>
<b>List of Abbreviations</b>	<b>xiii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Motivation . . . . .	1
1.2 Research Gap and Contribution . . . . .	2
1.3 Thesis Structure . . . . .	3
<b>2 Theoretical and Technical Background</b>	<b>5</b>
2.1 Hierarchical Bayesian Inference . . . . .	5
2.1.1 Marginal and then Conditional Method . . . . .	7
2.2 Sample-Based Estimates via Markov Chain Monte-Carlo Methods . . . . .	10
2.3 Numerical Approximation – Tensor-Train (TT) . . . . .	10
2.3.1 Marginal Probability Distributions . . . . .	13
2.3.2 Sampling from a TT Approximation . . . . .	15
2.3.3 Error of the TT Approximation . . . . .	17
2.4 Regularisation Approach . . . . .	18
<b>3 The Forward Model</b>	<b>21</b>
3.1 Radiative Transfer Equation . . . . .	22
3.2 Simulate Data Based on a Ground Truth . . . . .	23
3.2.1 Understanding the Forward Model . . . . .	25
<b>4 Linear Bayesian vs. Regularisation – Ozone</b>	<b>33</b>
4.1 Hierarchical Bayesian Framework for Ozone . . . . .	33
4.1.1 Prior Modelling . . . . .	35
4.2 Posterior Distribution . . . . .	35
4.2.1 Marginal Posterior . . . . .	36
4.2.2 Full Conditional Posterior – Ozone . . . . .	40
4.3 Solution by Regularisation . . . . .	43
<b>5 Affine Approximation of the Non-Linear Model</b>	<b>47</b>
5.1 Finding an Affine Approximation . . . . .	47
5.2 Marginal and Then Conditional Posterior – Ozone . . . . .	49

<b>6 Joint Retrieval of Ozone, Pressure and Temperature</b>	<b>53</b>
6.1 Hierarchical Bayesian Model – Ozone, Pressure, Temperature . . . . .	53
6.1.1 Prior Modelling . . . . .	56
6.2 Posterior Distribution . . . . .	59
6.2.1 Marginal Posterior Distribution – Pressure and Temperature . . .	59
6.2.2 Full Conditional Posterior Distribution – Ozone . . . . .	71
<b>7 Summary and Outlook</b>	<b>75</b>
7.1 Regularisation Solution vs. Hierarchical Bayesian Approach . . . . .	75
7.2 Sampling Methods vs. TT Approximation . . . . .	76
7.3 Atmospheric Physics . . . . .	77
<b>References</b>	<b>79</b>
<b>Appendices</b>	
<b>A Theoretical and Technical Background</b>	<b>85</b>
A.1 Correlation Structure . . . . .	85
A.2 Monte-Carlo Error and Integrated Autocorrelation Time . . . . .	86
A.3 Python Code . . . . .	88
<b>B Additional Figures</b>	<b>91</b>
B.1 Ozone . . . . .	92
B.1.1 Ozone Prior . . . . .	92
B.1.2 Integrated Autocorrelation Time . . . . .	93
B.1.3 Eigenvectors of Full Conditional Posterior Precision Matrix . . .	96
B.2 Pressure and Temperature . . . . .	98
B.2.1 Priors . . . . .	98
B.2.2 T-walk Trace . . . . .	100
B.2.3 Integrated Autocorrelation Time . . . . .	101

## List of Figures

2.1	Hierarchical Bayesian Model . . . . .	6
2.2	Visualisation of a tensor train . . . . .	12
3.1	Schematic of measurement and analysis geometry. . . . .	21
3.2	Tangent heights for different sequences of measurements. . . . .	26
3.3	Singular values of linear forward model matrix for different sequences of measurements. . . . .	28
3.4	First 10 right singular vectors of forward model. . . . .	29
3.5	Right singular vectors 11 to 19 of forward model. . . . .	30
3.6	Last 10 right singular vectors of forward model. . . . .	31
3.7	Logarithmic plot of data points at different tangent height. . . . .	32
4.1	Directed acyclic graph for ozone retrieval and MTC scheme. . . . .	34
4.2	Functions $f(\lambda)$ and $g(\lambda)$ of 2D marginal posterior. . . . .	36
4.3	IACT of $\lambda \sim \pi(\cdot \mathbf{y})$ , for linear model. . . . .	39
4.4	Samples from marginal posterior and TT approximation; trace plot of the MWG for $\pi(\lambda, \gamma \mathbf{y})$ . . . . .	41
4.5	Ozone samples of the full posterior. . . . .	42
4.6	Eigenvalues of the posterior precision matrix . . . . .	43
4.7	Relative Error of full posterior mean and covariance. . . . .	43
4.8	Plot of the L-curve to find the regularised solution. . . . .	45
5.1	Strategy to find affine map. . . . .	47
5.2	Assessment of affine map. . . . .	49
5.3	Marginal posterior histograms and TT approximation as well as hyper-prior distribution. . . . .	50
5.4	Full posterior mean and variance of ozone and the regularised solution compared to the ground truth. . . . .	51
6.1	Directed acyclic graph of Bayesian model for ozone, pressure and temperature. .	54
6.2	Prior Samples of $\mathbf{T}$ according to the respective hyper-prior distribution. .	56
6.3	Prior Samples of $\mathbf{p}$ according to the respective hyper-prior distribution. .	57
6.4	Prior Samples of $\mathbf{p}/\mathbf{T}$ according to the respective hyper-prior distribution. .	58
6.5	Correlation plot of samples from TT approximation . . . . .	61

6.6	Optimal rank and number of grid points for TT approximation . . . . .	63
6.7	Histograms and TT approximation of posterior distribution as well as hyper-prior distribution. . . . .	65
6.8	Histograms and TT approximation of posterior distribution as well as hyper-prior distribution. . . . .	66
6.9	Histograms and TT approximation of posterior distribution as well as hyper-prior distribution. . . . .	67
6.10	Histograms and TT approximation of posterior distribution as well as hyper-prior distribution. . . . .	68
6.11	Histograms and TT approximation of posterior distribution as well as hyper-prior distribution. . . . .	69
6.12	Histograms and TT approximation of posterior distribution as well as hyper-prior distribution. . . . .	70
6.13	Temperature posterior samples. . . . .	71
6.14	Pressure posterior samples. . . . .	72
6.15	Pressure posterior samples. . . . .	73
A.1	Correlation structure in between parameters and hyper-parameters . . . . .	86
B.1	Samples from ozone prior distribution. . . . .	92
B.2	IACT and autocorrelation of samples $\gamma \sim \pi(\cdot \mathbf{y})$ , for linear model. . . . .	93
B.3	IACT and autocorrelation of samples $\lambda \sim \pi(\cdot \mathbf{y})$ , for approximated model.	94
B.4	IACT and autocorrelation of samples $\gamma \sim \pi(\cdot \mathbf{y})$ , for approximated model.	95
B.5	First 11 eigenvectors of conditional precision matrix. . . . .	96
B.6	Last 23 eigenvectors of conditional precision matrix. . . . .	97
B.7	Prior distributions $\pi(\mathbf{h}_T)$ . . . . .	98
B.8	Prior samples of $1/T$ . . . . .	99
B.9	T-walk trace . . . . .	100
B.10	IACT and autocorrelation function of samples $\gamma \sim \pi(\cdot \mathbf{y})$ , for approximated model. . . . .	101
B.11	IACT and autocorrelation function of samples $\lambda \sim \pi(\cdot \mathbf{y})$ , for approximated model. . . . .	102
B.12	IACT and autocorrelation function of samples $b \sim \pi(\cdot \mathbf{y})$ , for approximated model. . . . .	103
B.13	IACT and autocorrelation function of samples $h_{T,1} \sim \pi(\cdot \mathbf{y})$ , for approximated model. . . . .	104
B.14	IACT and autocorrelation function of samples $T_0 \sim \pi(\cdot \mathbf{y})$ , for approximated model. . . . .	105
B.15	IACT and autocorrelation function of samples $p_0 \sim \pi(\cdot \mathbf{y})$ , for approximated model. . . . .	106

B.16 IACT and autocorrelation function of samples $h_{T,3} \sim \pi(\cdot \mathbf{y})$ , for approximated model. . . . .	107
B.17 IACT and autocorrelation function of samples $a_1 \sim \pi(\cdot \mathbf{y})$ , for approximated model. . . . .	108
B.18 IACT and autocorrelation function of samples $h_{T,2} \sim \pi(\cdot \mathbf{y})$ , for approximated model. . . . .	109
B.19 IACT and autocorrelation function of samples $a_0 \sim \pi(\cdot \mathbf{y})$ , for approximated model. . . . .	110
B.20 IACT and autocorrelation function of samples $a_2 \sim \pi(\cdot \mathbf{y})$ , for approximated model. . . . .	111
B.21 IACT and autocorrelation function of samples $a_3 \sim \pi(\cdot \mathbf{y})$ , for approximated model. . . . .	112
B.22 IACT and autocorrelation function of samples $h_{T,4} \sim \pi(\cdot \mathbf{y})$ , for approximated model. . . . .	113
B.23 IACT and autocorrelation function of samples $a_4 \sim \pi(\cdot \mathbf{y})$ , for approximated model. . . . .	114
B.24 IACT and autocorrelation function of samples $h_{T,5} \sim \pi(\cdot \mathbf{y})$ , for approximated model. . . . .	115
B.25 IACT and autocorrelation function of samples $a_5 \sim \pi(\cdot \mathbf{y})$ , for approximated model. . . . .	116
B.26 IACT and autocorrelation function of samples $h_{T,6} \sim \pi(\cdot \mathbf{y})$ , for approximated model. . . . .	117
B.27 IACT and autocorrelation function of samples $a_6 \sim \pi(\cdot \mathbf{y})$ , for approximated model. . . . .	118



## List of Abbreviations

<b>CDF</b>	.....	Cumulative Distribution Function
<b>DAG</b>	.....	Directed Acyclic Graph
<b>HITRAN</b>	.....	High Resolution Transmission
<b>IACT</b>	.....	Integrated Autocorrelation Time
<b>IRT</b>	.....	Inverse Rosenblatt Transform
<b>L</b>	.....	Linear
<b>MCMC</b>	.....	Markov Chain Monte-Carlo
<b>MH</b>	.....	Metropolis-Hastings
<b>MIPAS</b>	.....	Michelson Interferometer for Passive Atmospheric Sounding
<b>MLS</b>	.....	Microwave Limb Sounder
<b>MTC</b>	.....	Marginal and Then Conditional
<b>MVN</b>	.....	Multivariate Normal
<b>MWG</b>	.....	Metropolis Within Gibbs
<b>NASA</b>	.....	National Aeronautics and Space Administration
<b>NL</b>	.....	Non-Linear
<b>RMS</b>	.....	Root Mean Square
<b>RTE</b>	.....	Radiative Transfer Equation
<b>RTO</b>	.....	Randomise Then Optimise
<b>SIRT</b>	.....	Squared Inverse Rosenblatt Transform
<b>STD</b>	.....	Standard Deviation
<b>SVD</b>	.....	Singular Value Decomposition
<b>TT</b>	.....	Tensor-Train
<b>VMR</b>	.....	Volume Mixing Ratio

*xiv*

# 1

## Introduction

Here, we briefly describe the currently used standard to retrieve atmospheric trace gas, e.g. ozone, concentrations from limb-sounding measurements and what motivates us to employ a hierarchical Bayesian framework addressing this inverse problem. We explain how our approach contributes to and improves upon existing methods. Lastly, we provide the reader with the thesis structure.

### 1.1 Motivation

The only currently operating ozone limb sounder, the Microwave Limb Sounder (MLS) on NASA’s Aura satellite, is gradually drifting from its orbit and scheduled to be phased out by 2026 [13]. A group led by Harald Schwefel has proposed an alternative approach to fill this observational gap using a much smaller platform such as a 6U CubeSat (roughly 28cm × 15cm × 9cm) [58]. The proposed system includes a disk-shaped resonator targeting a narrow frequency band and converting the thermal radiation emitted by ozone molecules from the terahertz region to the optical domain [54, 50]. This frequency conversion offers a cost-effective and energy-efficient solution, as it circumvents the need for large, energy-hungry cooling devices that are traditionally required to capture terahertz signals. Instead, signal acquisition in the optical domain can be implemented by using compact, cheap, and low-power photonic technologies.

Currently, the inverse problem to retrieve any trace gas from limb-sounding data is approached by the atmospheric physics community using optimisation and regularisation techniques developed in the 1970s [43, 31]. These methods focus on finding the “best fit to data but not the best fit to parameters” [55]. Instead, we employ a hierarchically structured Bayesian framework to provide a distribution of ozone profiles, which represents

a range of feasible solutions to some given data. This probabilistic approach allows us to determine meaningful estimates and uncertainties of parameters.

## 1.2 Research Gap and Contribution

As already mentioned, currently the MLS retrieval algorithm [30] is based on the “optimal estimation” method from [43]. This approach provides a point estimate by fitting parameters to some data and iteratively minimising a squared residual norm, penalised against a chosen regularisation. This does not provide comprehensive information about the parameters’ underlying correlation structures can lead to unphysical results, e.g. negative ozone concentration values [49], and biased solutions, where the bias is then removed based on empirical decisions [32, 20]. Errors are provided by a local derivative of the forward map at the optimal solution, which is inherently highly sensitive to that specific point in the parameter space. Furthermore, these regularisation methods condition on external point estimates of other parameters, such as temperature or pressure [30].

In Bayesian modelling all unknown parameters are treated as random variables [27]. Further a hierarchically ordered model incorporates unknown hyper-parameters, which e.g. model the noise and control the smoothness of the ozone profile, and includes hyper-parameters and parameters in the retrieval process. Hence, a hierarchical Bayesian framework is able to model conditional dependences between parameters and hyper-parameters and provides posterior distributions over those, which present a range of feasible solutions for some given data. Livesey et al. [30] report “unexpected spectrally correlate noise” on the MLS aura, so here is another real reason why one should include noise in the model.

In this thesis the marginal and then conditional (MTC) method [17] is utilised and a hierarchical Bayesian model based on the radiative transfer equation (RTE) to provide posterior distributions of ozone, as well as temperature and pressure profiles is set up. The MTC scheme is a relatively new method within the Bayesian community, and is, to the best of our knowledge, the first application to a forward model based on the radiative transfer equation (RTE). First, we neglect the non-linearity of the RTE and employ a linear-Gaussian hierarchical Bayesian to give a posterior distribution of ozone profiles, and compare to a regularisation approach. Since the RTE is weakly non-linear, we use these results to find an affine map that approximates the non-linear forward model, which seems to be another novelty in the field of atmospheric remote sensing. Lastly, the hierarchical Bayesian framework is extended to jointly provide posterior ozone, pressure, and temperature profiles.

Moreover, instead of using sampling algorithms to characterise those posterior distributions we approximate posterior distributions directly utilising the tensor-train (TT)

format [8]. This allows us to generate independent samples from a TT approximation similar to the inverse Rosenblatt transform (IRT) [11] with far fewer function evaluations compared to conventional samplers. Further, in the TT format one can calculate marginal distributions of each hyper-parameter and evaluate integrals via quadrature without any sampling.

### 1.3 Thesis Structure

In Chapter 2, we give a brief overview of the key methods used along with references for further reading. Chapter 3 introduces the simplified forward model based on the RTE. Here we test a few measurement cases assess via a singular value decomposition (SVD) to understand the forward model and to determine how one can measure most effectively. and discusses strategies for measuring effectively. Based on those results and a ground truth, we simulate noisy data for an idealised limb sounder within a simplified atmosphere. Then, in Chapter 4, linear-Gaussian hierarchical Bayesian model based on the linearised RTE is constructed. We discuss some prior modelling choices, apply the MTC scheme and compare posterior ozone profiles and a regularisation approach to a ground truth ozone profile. Given those results we find an affine map in Chapter 5 to approximate the non-linear forward model. In Chapter 6, the previously built hierarchical Bayesian model is extended to include hyper-parameters corresponding to pressure and temperature and touch on some prior modelling choices. Again, using the MTC method joint estimates of posterior ozone, pressure and temperature profiles are provided and important aspects for improving the effectiveness and stability of TT approximations are highlighted. Lastly, in Chapter 7 we summarise and discuss our results and provide an outlook for future work. All programming and analysis in this thesis is done in Python, and the reported computation times correspond to a MacBook Pro from 2019 with a 2.4 GHz quad-core Intel Core i5 processor.



# 2

## Theoretical and Technical Background

In this Chapter, we give an introduction into the hierarchical Bayesian approach to inverse problems, some fundamentals of Markov chain Monte Carlo (MCMC) methods, and tensor-trains approximations of high-dimensional probability distributions. We keep it as general as possible and will not introduce specific sampling algorithms, as those are tailored towards the forward map and the applied problem and hence will be presented when used.

### 2.1 Hierarchical Bayesian Inference

Assume we observe some data

$$\mathbf{y} = \mathbf{A}(\mathbf{x}) + \boldsymbol{\eta}, \quad (2.1)$$

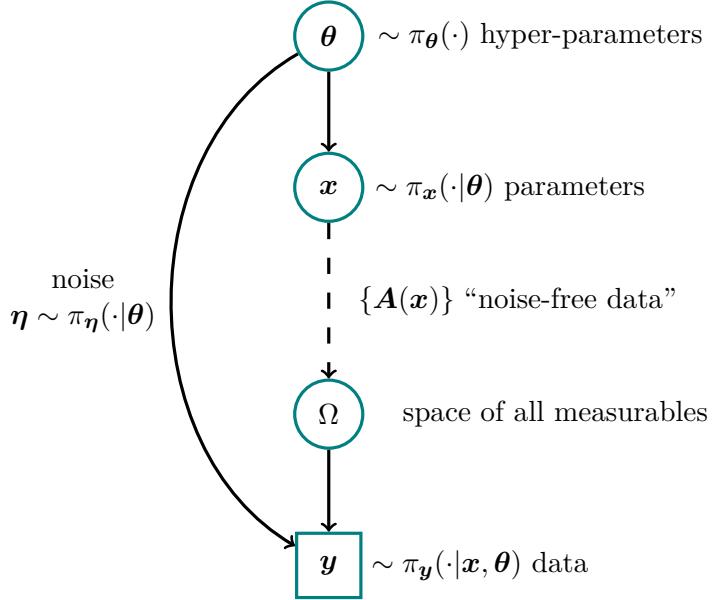
based on a forward model  $\mathbf{A}(\mathbf{x})$ , which may be non-linear, an unknown parameter vector  $\mathbf{x}$  and some additive random noise  $\boldsymbol{\eta}$ .

Naturally, due to the noise, the observation process in Eq. 2.1 is a random process. Hence, in Bayesian modelling, the aim is to determine a probability distribution over the parameter  $\mathbf{x}$  given some data  $\mathbf{y}$ . Further, a hierarchical Bayesian model incorporates unknown (auxiliary) hyper-parameters  $\boldsymbol{\theta}$ , and treats hyper-parameters and parameters as random variables [27, Chapter 3].

According to Bayes' theorem, the joint posterior distribution over the parameters  $\mathbf{x}$  and the hyper-parameter  $\boldsymbol{\theta}$  is given as

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

with finite and non-zero  $\pi(\mathbf{y})$ . The likelihood function  $\pi(\mathbf{y} | \mathbf{x}, \boldsymbol{\theta})$  is defined by the nature of the noise and the noise-free data  $\mathbf{A}(\mathbf{x})$ , which we read as the distribution over  $\mathbf{y}$  conditioned



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

on  $x$  and  $\theta$ . Here  $\theta$  may account for multiple hyper-parameters, e.g. modelling the noise vector  $\eta \sim \pi_\eta(\cdot|\theta)$ , where  $\sim$  reads as “is distributed as”, and describing physical properties or functional dependencies of  $x$  such as the smoothness of  $x$ . Consequently  $\theta$  is described by the hyper-prior distribution  $\pi(\theta)$ , where  $\pi(x, \theta) = \pi(x|\theta)\pi(\theta)$  with the prior distribution  $\pi(x|\theta)$ . Choosing these prior distributions is ultimately a modeller’s choice and is crucial, as those shall be as uninformative as possible for regions in hyper-parameter and parameter space where the data is informative. If the data is uninformative, the prior distributions can be informative and represent a rather restrictive range of (physically) feasible hyper-parameters and parameters.

Figure 2.1 visualises the conditional dependencies between hyper-parameters and parameters as well as how distributions progress through to a observation (square box) using a directed acyclic graph (DAG). We plot statistical dependencies as solid arrows and deterministic dependencies as dotted arrows.

Usually, the objective is to calculate the expectation of a function  $h(x)$ , which is defined as

$$\mathbb{E}_{x, \theta | y}[h(x)] = \underbrace{\int \int h(x) \pi(x, \theta | y) dx d\theta}_{\bar{h}}. \quad (2.3)$$

If that is a high-dimensional integral and computationally not feasible to solve, we approximate

$$\mathbb{E}_{\boldsymbol{x}, \boldsymbol{\theta} | \boldsymbol{y}}[h(\boldsymbol{x})] \approx \underbrace{\frac{1}{N} \sum_{k=1}^N h(\boldsymbol{x}^{(k)})}_{\bar{h}_N}, \quad (2.4)$$

with an unbiased sample-based Monte-Carlo estimate [41] for large enough  $N$  (law of large numbers [33, Chapter 17]). Here, the posterior samples  $\{\boldsymbol{x}^{(k)}, \boldsymbol{\theta}^{(k)}\} \sim \pi_{\boldsymbol{x}, \boldsymbol{\theta}}(\cdot | \boldsymbol{y})$ , for  $k = 1, \dots, N$ , form a sample set  $\mathcal{M} = \{(\boldsymbol{x}, \boldsymbol{\theta})^{(1)}, \dots, (\boldsymbol{x}, \boldsymbol{\theta})^{(N)}\}$ . The central limit theorem states that the sample mean  $\bar{h}_N^{(i)}$  of independent sample sets  $\mathcal{M}^{(i)} \sim \pi(\boldsymbol{x}, \boldsymbol{\theta} | \boldsymbol{y})$ , for  $i = 1, \dots, n$  from a distribution, converges to be normally distributed, so that

$$\sqrt{n}(\bar{h}_N^{(i)} - \bar{h}) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2) [21], \quad (2.5)$$

and if  $\sigma^2 < \infty$  the Monte-Carlo error  $\bar{h}_N^{(i)} - \bar{h}$  is bounded. In practice, we approximate the Monte-Carlo error from a sample set  $\mathcal{M}^{(i)}$  as

$$(\sigma^{(i)})^2 = \text{Var}(\mathbb{E}_{\boldsymbol{x}, \boldsymbol{\theta} | \boldsymbol{y}}[h(\boldsymbol{x})]) \approx \frac{\text{Var}(h(\boldsymbol{x}))}{N} \left( 1 + 2 \underbrace{\sum_{t=1}^W \frac{\Gamma(t)}{\Gamma(0)}}_{:= \tau_{\text{int}}} \right) = \text{Var}(h(\boldsymbol{x})) \frac{\tau_{\text{int}}}{N}, \quad (2.6)$$

where we have to take into account that the samples generated by any system or algorithm are correlated. We define the integrated autocorrelation time (IACT)  $\tau_{\text{int}}$  as in [17], which is twice the value of the IACT in [64, pp. 103-105]. Here the autocorrelation coefficient  $\Gamma(t) \propto \exp\{-|t|/\tau\} \rightarrow 0$  for  $t \rightarrow \infty$  at lag  $t$  decays exponentially and  $\Gamma(0) = \text{Var}(h(\boldsymbol{x}))$ . Choosing the summation window  $W$  is crucial because it has to be large compared to the decay time  $\tau$ , but for too large  $t$  the autocorrelation coefficient  $\Gamma(t)$  is noise-dominated. U. Wolff [62] (and the Python implementation by D. Hesse [26]) provide a way to not only calculate the IACT safely but also to quantify the errors of the estimated IACT.

The IACT provides a good estimate of the number of steps the sampling algorithm needs to take to produce one independent sample. According to the IACT, we define the effective sample size as  $\tau_{\text{int}}/N$ . We point out that for uncorrelated samples  $\tau_{\text{int}} = 1$  the error  $(\sigma^{(i)})^2$  is a typical Monte-Carlo estimate. See Appendix A.2 and [53, 62, 64] for a more detailed derivation.

### 2.1.1 Marginal and then Conditional Method

Quickly generating a representative sample set from the posterior distribution often presents a significant challenge. This is mainly due to the strong correlations that usually exist between the parameters and hyper-parameters, as discussed by Rue and Held in [45] and illustrated in Appendix A.1. If  $\boldsymbol{x}$  cannot be parametrised directly

in terms of the hyper-parameters  $\boldsymbol{\theta}$ , so that  $\mathbf{x}(\boldsymbol{\theta})$  is function of  $\boldsymbol{\theta}$ , it is beneficial to factorise the posterior distribution as **this is unclear. Of course  $\mathbf{x}(\text{theta})$  is a function of theta. What are you trying to say?** If  $\mathbf{x}$  cannot be parametrised directly in terms of the hyper-parameters  $\boldsymbol{\theta}$ , and only a low level representation of  $\mathbf{x}$  is available it is beneficial to factorise the posterior distribution as

$$\pi(\mathbf{x}, \boldsymbol{\theta} | \mathbf{y}) = \pi(\mathbf{x} | \boldsymbol{\theta}, \mathbf{y}) \pi(\boldsymbol{\theta} | \mathbf{y}), \quad (2.7)$$

into the full 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 hyper-parameter  $\boldsymbol{\theta}$ . This approach, known as the MTC method, is particularly advantageous when  $\mathbf{x} \in \mathbb{R}^n$  is high-dimensional, while  $\boldsymbol{\theta}$  is low-dimensional and the evaluation of  $\pi(\boldsymbol{\theta} | \mathbf{y})$  is relatively cheap. If a high level model of  $\mathbf{x}$  available we can represent the parameters space of  $\mathbb{R}^n$  with a number of hyper-parameters much smaller than  $n$ . That is when Bayesian modelling is especially powerful because we can impose prior knowledge through those hyper-parameters much more precisely and the dimensionality of the posterior may be drastically reduced. The marginal posterior over the hyper-parameter  $\boldsymbol{\theta}$ , that can be calculated by,

$$\pi(\boldsymbol{\theta} | \mathbf{y}) = \frac{\pi(\mathbf{y} | \mathbf{x}, \boldsymbol{\theta}) \pi(\mathbf{x} | \boldsymbol{\theta}) \pi(\boldsymbol{\theta})}{\pi(\mathbf{x} | \boldsymbol{\theta}, \mathbf{y}) \pi(\mathbf{y})} \propto \frac{\pi(\mathbf{y} | \mathbf{x}, \boldsymbol{\theta}) \pi(\mathbf{x} | \boldsymbol{\theta}) \pi(\boldsymbol{\theta})}{\pi(\mathbf{x} | \boldsymbol{\theta}, \mathbf{y})}, \quad (2.8)$$

as in [17, Lemma 2].

Applying the law of total expectation [5], Eq. (2.3) becomes

$$\mathbb{E}_{\mathbf{x}, \boldsymbol{\theta} | \mathbf{y}}[h(\mathbf{x})] = \int \int h(\mathbf{x}) \pi(\mathbf{x} | \boldsymbol{\theta}, \mathbf{y}) d\mathbf{x} \pi(\boldsymbol{\theta} | \mathbf{y}) d\boldsymbol{\theta} \quad (2.9)$$

$$= \int \mathbb{E}_{\mathbf{x} | \boldsymbol{\theta}, \mathbf{y}}[h(\mathbf{x})] \pi(\boldsymbol{\theta} | \mathbf{y}) d\boldsymbol{\theta} \quad (2.10)$$

$$= \mathbb{E}_{\boldsymbol{\theta} | \mathbf{y}} \left[ \mathbb{E}_{\mathbf{x} | \boldsymbol{\theta}, \mathbf{y}}[h(\mathbf{x})] \right]. \quad (2.11)$$

In the case of a linear-Gaussian hierarchical Bayesian model, both the marginal distribution  $\pi(\boldsymbol{\theta} | \mathbf{y})$  and the inner expectation  $\mathbb{E}_{\mathbf{x} | \boldsymbol{\theta}, \mathbf{y}}[h(\mathbf{x})]$  are well defined (see next subsection). If the integral in Eq. 2.10 is expensive to calculate, we use sample-based methods to produce a Markov chain  $\{(\mathbf{x}, \boldsymbol{\theta})^{(1)}, \dots, (\mathbf{x}, \boldsymbol{\theta})^{(N)}\} \sim \pi(\mathbf{x}, \boldsymbol{\theta} | \mathbf{y})$  and sample from  $\pi(\boldsymbol{\theta} | \mathbf{y})$  first and then draw samples from the full conditional posterior  $\pi(\mathbf{x} | \boldsymbol{\theta}, \mathbf{y})$ , e.g via the RTO method (see Sec. 6.2.2).

### Linear-Gaussian hierarchical Bayesian model

In case of normally distributed noise  $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$  with zero mean and covariance  $\boldsymbol{\Sigma}(\boldsymbol{\theta})$  and a linear forward model matrix  $\mathbf{A}$  the Eq. 2.1 simplifies to

$$\mathbf{y} = \mathbf{Ax} + \boldsymbol{\eta}. \quad (2.12)$$

Then we can obtain the marginal and full conditional posterior distribution explicitly. Our hierarchical linear-Gaussian Bayesian model is defined as

$$\mathbf{y}|\boldsymbol{x}, \boldsymbol{\theta} \sim \mathcal{N}(\mathbf{A}\boldsymbol{x}, \boldsymbol{\Sigma}(\boldsymbol{\theta})) \quad (2.13a)$$

$$\boldsymbol{x}|\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{Q}(\boldsymbol{\theta})^{-1}) \quad (2.13b)$$

$$\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta}), \quad (2.13c)$$

with a Gaussian likelihood function  $\pi(\mathbf{y}|\boldsymbol{x}, \boldsymbol{\theta})$ , a normally distributed prior  $\pi(\boldsymbol{x}|\boldsymbol{\theta})$ , with prior mean  $\boldsymbol{\mu}$  and prior precision  $\boldsymbol{Q}(\boldsymbol{\theta})$ , and a hyper-prior distribution  $\pi(\boldsymbol{\theta})$ . For derivation of the marginal posterior and the full conditional posterior distribution, consider the joint multivariate Gaussian distribution

$$\begin{pmatrix} \boldsymbol{x} \\ \mathbf{y} \end{pmatrix} \sim \mathcal{N} \left[ \begin{pmatrix} \boldsymbol{\mu} \\ \mathbf{A}\boldsymbol{\mu} \end{pmatrix}, \begin{pmatrix} \boldsymbol{Q}(\boldsymbol{\theta}) + \mathbf{A}^T \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} \mathbf{A} & -\mathbf{A}^T \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} \\ \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} \mathbf{A} & \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} \end{pmatrix}^{-1} \right], \quad (2.14)$$

with the joint precision matrix as in [52] (see also [45, 17]). Immediately, the full conditional posterior distribution can be formulated as

$$\boldsymbol{x}|\boldsymbol{\theta}, \mathbf{y} \sim \mathcal{N} \left( \underbrace{\boldsymbol{\mu} + (\boldsymbol{Q}(\boldsymbol{\theta}) + \mathbf{A}^T \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} \mathbf{A})^{-1} (\mathbf{y} - \mathbf{A}\boldsymbol{\mu})}_{\boldsymbol{\mu}_{\boldsymbol{x}|\boldsymbol{\theta}, \mathbf{y}}}, \underbrace{(\boldsymbol{Q}(\boldsymbol{\theta}) + \mathbf{A}^T \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} \mathbf{A})^{-1}}_{\boldsymbol{\Sigma}_{\boldsymbol{x}|\boldsymbol{\theta}, \mathbf{y}}} \right). \quad (2.15)$$

Then the marginal posterior distribution over the hyper-parameters is derived as in Eq. 2.8, where, as noted by Fox and Norton [17], the parameter  $\boldsymbol{x}$  cancels and we arrive at

$$\begin{aligned} \pi(\boldsymbol{\theta}|\mathbf{y}) \propto & \sqrt{\frac{\det(\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}) \det(\boldsymbol{Q}(\boldsymbol{\theta}))}{\det(\boldsymbol{Q}(\boldsymbol{\theta}) + \mathbf{A}^T \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} \mathbf{A})}} \exp \left\{ -\frac{1}{2} (\mathbf{y} - \mathbf{A}\boldsymbol{\mu})^T \right. \\ & \left. [\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} - \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} \mathbf{A} (\boldsymbol{Q}(\boldsymbol{\theta}) + \mathbf{A}^T \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} \mathbf{A})^{-1} \mathbf{A}^T \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}] (\mathbf{y} - \mathbf{A}\boldsymbol{\mu}) \right\} \pi(\boldsymbol{\theta}). \end{aligned} \quad (2.16)$$

Having the marginal posterior distribution  $\pi(\boldsymbol{\theta}|\mathbf{y})$  (independent of  $\boldsymbol{x}$ ) available breaks up the correlation structure between  $\boldsymbol{x}$  and  $\boldsymbol{\theta}$  and makes the MTC approach very efficient [17] (see Appendix A.1). Within this scheme, we evaluate the marginal posterior first and then either condition on hyper-parameters to draw full conditional posterior samples  $\boldsymbol{x} \sim \pi(\boldsymbol{x}|\boldsymbol{\theta}, \mathbf{y})$  (see Sec. 6.2.2) or evaluate the mean

$$\boldsymbol{\mu}_{\boldsymbol{x}|\mathbf{y}} = \int \boldsymbol{\mu}_{\boldsymbol{x}|\boldsymbol{\theta}, \mathbf{y}} \pi(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} \quad (2.17)$$

and covariance matrix

$$\boldsymbol{\Sigma}_{\boldsymbol{x}|\mathbf{y}} = \int \boldsymbol{\Sigma}_{\boldsymbol{x}|\boldsymbol{\theta}, \mathbf{y}} \pi(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} \quad (2.18)$$

of the posterior  $\pi(\boldsymbol{x}|\mathbf{y})$  by some quadrature rule.

## 2.2 Sample-Based Estimates via Markov Chain Monte-Carlo Methods

We use Markov chain Monte-Carlo (MCMC) methods to calculate sample-based estimates of  $\mathbb{E}_{\mathbf{x}, \boldsymbol{\theta} | \mathbf{y}}[h(\mathbf{x})]$  as in Eq. 2.4. Within the MTC scheme one may draw samples  $\{\boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(k)}, \dots, \boldsymbol{\theta}^{(N)}\} \sim \pi(\boldsymbol{\theta} | \mathbf{y})$  from the marginal posterior first and then evaluate the full conditional posterior  $\pi(\mathbf{x} | \boldsymbol{\theta}, \mathbf{y})$ . In doing so we generate a Markov chain  $\mathcal{M} = \{\boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(k)}, \dots, \boldsymbol{\theta}^{(N)}\}$ , where every new sample  $\boldsymbol{\theta}^{(k)}$  is proposed according to a random proposal  $\boldsymbol{\theta}^{(k)} \sim q(\cdot | \boldsymbol{\theta}^{(k-1)})$  and only affected by the previous one  $\boldsymbol{\theta}^{(k-1)}$ . For large enough  $N$  this chain of random variables can be used to calculate Monte-Carlo estimates, where ergodicity of  $\mathcal{M}$  is a sufficient criterion to do so [55, 41].

The ergodicity theorem in [55] 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, the chain can reach any state from any other state (irreducibility), is not stuck in periodic cycles (aperiodicity), and satisfies the detailed balance condition [55] (reversibility). Then the samples in that chain  $\mathcal{M} \sim \pi(\boldsymbol{\theta} | \mathbf{y})$  are samples from the desired target distribution. In practice, one can inspect the trace  $\pi(\boldsymbol{\theta}^{(k)} | \mathbf{y})$  for  $k = N_{\text{burn-in}}, \dots, N$  after a “burn-in” period  $N_{\text{burn-in}}$  which removes initialisation bias and visually assess if the chain is consistent with ergodicity. The specific sampling methods in this thesis possess proven ergodic properties, and we therefore provide the reader with corresponding literature for further details when the methods are introduced.

If the instance  $\boldsymbol{\theta}^{(k)}$  of an ergodic Markov chain represents an independent sample of the marginal posterior  $\pi(\boldsymbol{\theta} | \mathbf{y})$  and if a full conditional posterior sample  $\mathbf{x}^{(k)} \sim \pi(\mathbf{x} | \boldsymbol{\theta}^{(k)}, \mathbf{y})$  is drawn independently, as in Sec. 6.2.2, then the resulting sample  $(\mathbf{x}^{(k)}, \boldsymbol{\theta}^{(k)})$  is an independent sample from the joint posterior  $\pi(\mathbf{x}, \boldsymbol{\theta} | \mathbf{y})$  [17, 59]. Repeating this procedure gives the chain  $\{(\mathbf{x}, \boldsymbol{\theta})^{(1)}, \dots, (\mathbf{x}, \boldsymbol{\theta})^{(k)}, \dots, (\mathbf{x}, \boldsymbol{\theta})^{(N)}\} \sim \pi(\mathbf{x}, \boldsymbol{\theta} | \mathbf{y})$  of independent samples from the joint posterior.

## 2.3 Numerical Approximation – Tensor-Train (TT)

Instead of relying on sampling-based methods to explore a target function  $\pi(\mathbf{x})$ , we can approximate this function on a  $d$ -dimensional grid using a tensor-train (TT) approximation  $\tilde{\pi}(\mathbf{x}) \approx \pi(\mathbf{x})$ , where  $\mathbf{x} \in \mathbb{R}^d$ , with far fewer function evaluation compared to conventional sampling methods. In the following, we describe how to compute marginal functions from a probability density approximated in TT format and how to generate samples using a scheme similar to the inverse Rosenblatt transform (IRT) as in [11]. In this section we follow the notation and procedure introduced in [8].

As in [8], we can define the parameter space as the product space  $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_d$  with  $x_k \in \mathcal{X}_k \subseteq \mathbb{R}$ . The marginal function for the  $k$ -th component is then given by

$$f_{X_k}(x_k) = \frac{1}{z} \int_{\mathcal{X}_1} \cdots \int_{\mathcal{X}_{k-1}} \int_{\mathcal{X}_{k+1}} \cdots \int_{\mathcal{X}_d} \lambda(\mathbf{x}) \pi(\mathbf{x}) dx_1 \cdots dx_{k-1} dx_{k+1} \cdots dx_d, \quad (2.19)$$

where we integrate over all dimensions except the  $k$ -th, and  $z$  is a normalisation constant. Here, Cui and Dolgov [8] refer to  $\lambda(x)$  as the “product-form Lebesgue-measurable weighting function”, which can be useful for quadrature rules [9], and define it as

$$\lambda(\mathcal{X}) = \prod_{i=1}^d \lambda_i(\mathcal{X}_i), \quad \text{where } \lambda_i(\mathcal{X}_i) = \int_{\mathcal{X}_i} \lambda_i(x_i) dx_i. \quad (2.20)$$

To approximate a target function in the TT format one has to predefined  $d$ -dimensional discrete univariate grid over the parameter space  $\mathcal{X}$  with  $n$  grid points in each direction. Then the integral in Eq. 2.19 for the marginal function can be computed at a low computational cost, as  $\pi(\mathbf{x})$  is approximated by

$$\tilde{\pi}(\mathbf{x}) = \tilde{\pi}_1(x_1)\tilde{\pi}_2(x_2) \cdots \tilde{\pi}_d(x_d),$$

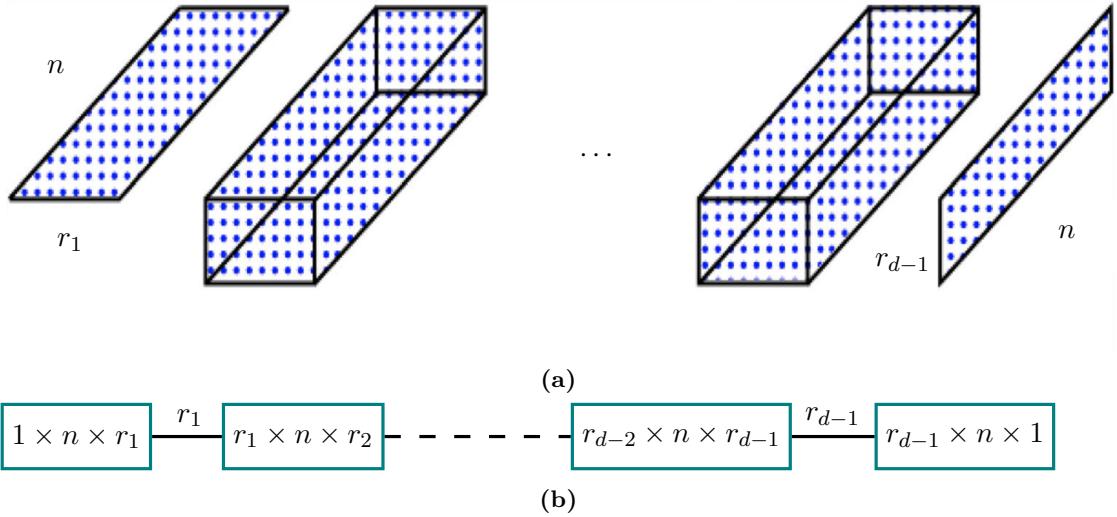
which is a sequence of matrix multiplications with  $\tilde{\pi}_k(x_k) \in \mathbb{R}^{r_{k-1} \times r_k}$  for a fixed grid point  $\mathbf{x} = (x_1, \dots, x_d)$ . A TT core  $\tilde{\pi}_k \in \mathbb{R}^{r_{k-1} \times n \times r_k}$  has ranks  $r_{k-1}$  and  $r_k$ , which are connecting the TT core to its respective neighbouring dimensions. A TT has outer ranks  $r_0 = r_d = 1$ . This enables us to approximate  $\pi(\mathcal{X}) \approx \tilde{\pi}_1 \tilde{\pi}_2 \cdots \tilde{\pi}_d$  over a discrete parameter space  $\mathcal{X}$  using  $2nr + (d-2)nr^2$  evaluation points for fixed ranks  $r = r_{k-1} = r_k$ , as illustrated in Figure 2.2, instead of  $n^d$  function evaluation. Consequently, the marginal function

$$\begin{aligned} f_{X_k}(x_k) &\approx \frac{1}{z} \left| \left( \int_{\mathcal{X}_1} \lambda_1(x_1) \tilde{\pi}_1(x_1) dx_1 \right) \cdots \left( \int_{\mathcal{X}_{k-1}} \lambda_{k-1}(x_{k-1}) \tilde{\pi}_{k-1}(x_{k-1}) dx_{k-1} \right) \right. \\ &\quad \left. \lambda_k(x_k) \tilde{\pi}_k(x_k) \right. \\ &\quad \left. \left( \int_{\mathcal{X}_{k+1}} \lambda_{k+1}(x_{k+1}) \tilde{\pi}_{k+1}(x_{k+1}) dx_{k+1} \right) \cdots \left( \int_{\mathcal{X}_d} \lambda_d(x_d) \tilde{\pi}_d(x_d) dx_d \right) \right| \end{aligned} \quad (2.21)$$

is computed by integrating over all TT cores except the  $k$ -th core  $\pi_k$ , as in [11], and normalised by the constant  $z$  [8].

In practice, TT approximations may suffer from numerical instability. In particular when non-negative functions (probability distributions) are approximated the TT approximation can have negative values in regions where true function values are very small. One way to ensure positivity is to square the target function, hence [8] approximate the square root of the target function and define the approximation as [8, Eq. 18]

$$\sqrt{\pi(\mathbf{x})} \approx \tilde{g}(\mathbf{x}) = \mathbf{G}_1(x_1), \dots, \mathbf{G}_k(x_k), \dots, \mathbf{G}_d(x_d). \quad (2.22)$$



**Figure 2.2:** Here, we visualise the TT cores as a train of two- and three-dimensional matrices. Each core has a length  $n$ , corresponding to the number of grid points in each dimension, and the cores are connected through ranks  $r_k$ . More specifically, a core  $\tilde{\pi}_k$  has dimensions  $r_{k-1} \times n \times r_k$ , with outer ranks  $r_0 = r_d = 1$ . Using the TT format enables us to represent a  $d$ -dimensional grid with only  $2nr + (d - 2)nr^2$  evaluation points instead of  $n^d$  grid points. Figure (a) is adapted from [19].

Here, as in each TT core is given by [8, Eq. 21]

$$G_k^{(\alpha_{k-1}, \alpha_k)}(x_k) = \sum_{i=1}^{n_k} \phi_k^{(i)}(x_k) \mathbf{A}_k[\alpha_{k-1}, i, \alpha_k], \quad k = 1, \dots, d, , \quad (2.23)$$

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 unnormalised function (probability density) is written as [8, Eq. 19]:

$$\pi(\mathbf{x}) \approx \xi + \tilde{g}(\mathbf{x})^2, \quad (2.24)$$

where  $\xi$  is a positive constant added according to the ratio of the Lebesgue weighted L2-norm error and the Lebesgue weighting (see Eq. 2.20 and [8, Eq. 35]) such that

$$0 \leq \xi \leq \frac{1}{\lambda(\mathcal{X})} \|\tilde{g} - \sqrt{\pi}\|_{L_\lambda^2(\mathcal{X})}^2. \quad (2.25)$$

This leads to the normalised target function [8, Eq. 19]

$$f_X(\mathbf{x}) \approx \frac{1}{z} \left( \lambda(\mathbf{x})\xi + \lambda(\mathbf{x})\tilde{g}(\mathbf{x})^2 \right), \quad (2.26)$$

with the normalisation constant  $z = \int_{\mathcal{X}} f_X(\mathbf{x}) d\mathbf{x}$ . Given the tensor train approximation

of  $\sqrt{\pi}$ , the marginal function  $f_{X_k}(x_k)$  can be expressed as

$$\begin{aligned} f_{X_k}(x_k) &\approx \frac{1}{z} \left( \xi \prod_{i=1}^{k-1} \lambda_i(\mathcal{X}_i) \prod_{i=k+1}^d \lambda_i(\mathcal{X}_i) \right. \\ &\quad + \left( \int_{\mathcal{X}_1} \lambda_1(x_1) \mathbf{G}_1^2(x_1) dx_1 \right) \cdots \left( \int_{\mathcal{X}_{k-1}} \lambda_{k-1}(x_{k-1}) \mathbf{G}_{k-1}^2(x_{k-1}) dx_{k-1} \right) \\ &\quad \lambda_k(x_k) \mathbf{G}_k^2(x_k) \\ &\quad \left. \left( \int_{\mathcal{X}_{k+1}} \lambda_{k+1}(x_{k+1}) \mathbf{G}_{k+1}^2(x_{k+1}) dx_{k+1} \right) \cdots \left( \int_{\mathcal{X}_d} \lambda_d(x_d) \mathbf{G}_d^2(x_d) dx_d \right) \right). \end{aligned} \quad (2.27)$$

### 2.3.1 Marginal Probability Distributions

The marginal functions  $f_{X_k}(x_k)$  of the target function  $f_X(\mathbf{x})$  are computed by a procedure to which Cui et al. [8] refer to as backwards marginalisation, see Prop. 2, and to which we add the forward marginalisation, see Prop. 1. This is similar to the left and right orthogonalisation of TT cores [35, 34]. The backwards marginalisation provides the coefficient matrices  $\mathbf{B}_k$ , while the forward marginalisation gives the coefficient matrices  $\mathbf{R}_{\text{pre},k}$ . These matrices enable the efficient evaluation of marginal functions since they integrate over the coordinates either left or right of the  $k$ -th dimension, as in [8]. In doing so, the mass matrix  $\mathbf{M}_k \in \mathbb{R}^{n_k \times n_k}$  is defined as in [8, Eq. 22]

$$\mathbf{M}_k[i, j] = \int_{\mathcal{X}_k} \phi_k^{(i)}(x_k) \phi_k^{(j)}(x_k) \lambda(x_k) dx_k, \quad i, j = 1, \dots, n_k, , \quad (2.28)$$

where  $\{\phi_k^{(i)}(x_k)\}_{i=1}^{n_k}$  denotes the set of basis functions for the  $k$ -th coordinate. The proposition used to compute  $\mathbf{B}_k$ , stated in Prop. 1, is adapted directly from [8].

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

$$f_{X_k}(x_k) \approx \frac{1}{z} \left( \xi \prod_{i=1}^{k-1} \lambda_i(X_i) \prod_{i=k+1}^d \lambda_i(X_i) + \sum_{l_{k-1}=1}^{r_{k-1}} \sum_{l_k=1}^{r_k} \left( \sum_{i=1}^n \phi_k^{(i)}(x_k) \mathbf{D}_k[l_{k-1}, i, l_k] \right)^2 \right) \lambda_k(x_k), \quad (2.29)$$

where  $\mathbf{D}_k \in \mathbb{R}^{r_{k-1} \times n \times r_k}$  and given as

$$\mathbf{D}_k[l_{k-1}, i, l_k] = \sum_{\alpha_{k-1}=1}^{r_{k-1}} \mathbf{R}_{\text{pre},k-1}[l_{k-1}, \alpha_{k-1}] \mathbf{B}_k[\alpha_{k-1}, i, l_k], \quad (2.30)$$

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

For the first dimension,  $f_{X_1}(x_1)$  can be expressed as [8, Eq. 30]

$$f_{X_1}(x_1) \approx \frac{1}{z} \left( \xi \prod_{i=2}^d \lambda_i(\mathcal{X}_i) + \sum_{l_1=1}^{r_1} \left( \sum_{i=1}^n \phi_1^{(i)}(x_1) \mathbf{D}_1[i, l_1] \right)^2 \right) \lambda_1(x_1) , \quad (2.31)$$

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

$$f_{X_d}(x_d) \approx \frac{1}{z} \left( \xi \prod_{i=1}^{d-1} \lambda_i(\mathcal{X}_i) + \sum_{l_{d-1}=1}^{r_{d-1}} \left( \sum_{i=1}^n \phi_1^{(i)}(x_1) \mathbf{D}_d[l_{d-1}, i] \right)^2 \right) \lambda_d(x_d), \quad (2.32)$$

where  $\mathbf{D}_d[l_{d-1}, i] = \mathbf{B}_{\text{pre}, d}[l_{d-1}, i, \alpha_{d+1}]$  and  $\alpha_{d+1} = 1$ . Note that in practice we calculate

$z$  numerically within the process of computing the marginals so that  $\sum f_{X_k}(x_k) = 1$  and

for Cartesian basis  $\mathbf{M}_k = \text{diag}(\lambda_k(\mathcal{X}_k))$  with  $\lambda(x) = 1$ .

**Proposition 1** (Backwards Marginalisation as in [8]): 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  $\mathbf{B}_k \in \mathbb{R}^{r_{k-1} \times n_k \times r_k}$ , which is needed for defining the marginal function  $f_{X_k}(x_k)$  or to draw samples from  $\tilde{\pi}(\mathbf{x})$  via the squared IRT scheme (see Alg. Box 1):

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  $\mathbf{C}_k \in \mathbb{R}^{r_{k-1} \times n_k \times r_k}$  [8, Eq. (27)]:

$$\mathbf{C}_k[\alpha_{k-1}, \tau, l_k] = \sum_{i=1}^{n_k} \mathbf{B}_k[\alpha_{k-1}, i, l_k] \mathbf{L}_k[i, \tau]. \quad (2.33)$$

2. Unfold  $\mathbf{C}_k$  along the first coordinate and compute the thin QR decomposition, so that  $\mathbf{C}_k^{(R)} \in \mathbb{R}^{r_{k-1} \times (n_k r_k)}$  [8, Eq. 28]:

$$\mathbf{Q}_k \mathbf{R}_k = (\mathbf{C}_k^{(R)})^\top. \quad (2.34)$$

3. Compute the new coefficient tensor [8, Eq. 29]:

$$\mathbf{B}_{k-1}[\alpha_{k-2}, i, l_{k-1}] = \sum_{\alpha_{k-1}=1}^{r_{k-1}} \mathbf{A}_{k-1}[\alpha_{k-2}, i, \alpha_{k-1}] \mathbf{R}_k[l_{k-1}, \alpha_{k-1}]. \quad (2.35)$$

**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  $\mathbf{R}_{\text{pre},k-1} \in \mathbb{R}^{r_{k-1} \times r_{k-1}}$  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  $\mathbf{C}_k \in \mathbb{R}^{r_{k-1} \times n_k \times r_k}$ :

$$\mathbf{C}_{\text{pre},k}[\alpha_{k-1}, \tau, l_k] = \sum_{i=1}^{n_k} \mathbf{L}_k[i, \tau] \mathbf{B}_{\text{pre},k}[\alpha_{k-1}, i, l_k]. \quad (2.36)$$

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

$$\mathbf{Q}_{\text{pre},k} \mathbf{R}_{\text{pre},k} = (\mathbf{C}_{\text{pre},k}^{(R)}). \quad (2.37)$$

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

$$\mathbf{B}_{\text{pre},k+1}[l_{k+1}, i, \alpha_{k+1}] = \sum_{\alpha_k=1}^{r_k} \mathbf{R}_{\text{pre},k}[l_{k+1}, \alpha_k] \mathbf{A}_{k+1}[\alpha_k, i, \alpha_{k+1}]. \quad (2.38)$$

### 2.3.2 Sampling from a TT Approximation

Instead of evaluating marginal functions for quadrature, the inverse Rosenblatt transform (IRT) provide a scheme to draw samples from an approximated function in the TT format [11]. Since the square root of the target function is approximated, Cui and Dolgov [8] call that the squared inverse Rosenblatt transform (SIRT).

#### Algorithm 1: Squared Inverse Rosenblatt Transform (SIRT)

- 1: **Input:** seeds  $\{\mathbf{u}^{(1)}, \dots, \mathbf{u}^{(N)}\} \sim \mathcal{U}(0, 1)^d$  and  $\mathbf{B}_1, \dots, \mathbf{B}_d$  from Prop. 1
- 2: **for**  $s = 1, \dots, N$  **do**
- 3:   **for**  $k = 1, \dots, d$  **do**
- 4:     compute normalised PDF  $f_{X_k|X_{<k}}(x_k | x_{k-1}^{(s)}, \dots, x_1^{(s)})$ , Eq. 2.40
- 5:     compute cumulative distribution function  $F_{X_k|X_{<k}}(x_k)$ , Eq. 2.39,
- 6:     project sample  $x_k^{(s)} = F_{X_k|X_{<k}}^{-1}(u_k^{(s)})$
- 7:     interpolate  $\mathbf{G}_k(x_k^{(s)})$ , Eq. 2.41
- 8:     update  $\mathbf{G}_{\leq k}(x_{\leq k}^{(s)}) = \mathbf{G}_{<k}(x_{<k}^{(s)}) \mathbf{G}_k(x_k^{(s)})$
- 9:   **end for**
- 10: **end for**
- 11: **Output:** samples  $\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}\}$ , where each  $\mathbf{x}^{(s)} \in \mathbb{R}^d$  for  $s = 1, \dots, N$

Given the Backward marginals  $\mathbf{B}_1, \dots, \mathbf{B}_d$  as in Prop. 1, draw  $N$  uniformly distributed seeds  $\{\mathbf{u}^{(1)}, \dots, \mathbf{u}^{(N)}\} \sim \mathcal{U}(0, 1)^d$ , where each  $\mathbf{u}^{(s)}$  is  $d$ -dimensional for  $s = 1, \dots, N$ . Then calculate the first marginal  $f_{X_1}(x_1)$  as in Eq. 2.31 and normalise with  $z = \int_{\mathcal{X}_1} f_{X_1}(x_1) dx_1$ .

Next, compute the cumulative distribution function (CDF)  $F_{X_1}(x_k) = \int_{-\infty}^{x_k} f_{X_1}(\hat{x}_1) d\hat{x}_1$ , which for the general case is given as [8, Eq. 17]:

$$F_{X_k|X_{<k}}(x_k) = \int_{-\infty}^{x_k} f_{X_k|X_{<k}}(\hat{x}_k|x_{k-1}, \dots, x_1) d\hat{x}_k ; \quad (2.39)$$

Now project the seed  $u_k^{(s)}$  on the parameter space, and generate the sample  $x_k^{(s)} = F_{X_k|X_{<k}}^{-1}(u_k^{(s)})$ . The “conditional marginal” is given as [8, Eq. 31]:

$$\begin{aligned} f_{X_k|X_{<k}}(x_k|x_{k-1}^{(s)}, \dots, x_1^{(s)}) &\approx \frac{1}{z} \left( \xi \prod_{i=k+1}^d \lambda_i(X_i) + \right. \\ &\quad \left. \sum_{l_k=1}^{r_k} \left( \sum_{i=1}^n \phi_k^{(i)}(x_k^{(s)}) \left( \sum_{\alpha_{k-1}=1}^{r_{k-1}} \mathbf{G}_{<k}^{(\alpha_{k-1})}(x_{<k}^{(s)}) \mathbf{B}_k[\alpha_{k-1}, i, l_k] \right) \right)^2 \right) \lambda_k(x_k) , \end{aligned} \quad (2.40)$$

where we marginalise over the dimensions  $k+1, \dots, d$  via  $\mathbf{B}_k$  and condition on the previous  $k-1$  samples through the product  $\mathbf{G}_k(x_k^{(s)}) \in \mathbb{R}^{1 \times r_{k-1}}$  to preserve the correlation structure. Function values between grid points  $i$  and  $i+1$  are approximated with a piecewise polynomial interpolation

$$\mathbf{G}_k(x_k^{(s)}) \approx \frac{x_k^{(s)} - x_k^{(i)}}{x_k^{(i+1)} - x_k^{(i)}} \mathbf{G}_k(x_k^{(i+1)}) + \frac{x_k^{(i+1)} - x_k^{(s)}}{x_k^{(i+1)} - x_k^{(i)}} \mathbf{G}_k(x_k^{(i)}) , \quad (2.41)$$

for  $x_k^{(i)} \leq x_k^{(s)} \leq x_k^{(i+1)}$  as in [11] for the next “conditional marginal”.

The procedure is repeated for each  $u_k^{(s)} \in \mathbf{u}^{(s)}$  to produce the samples  $\mathbf{x}^{(s)} \sim f_X(\mathbf{x})$ , as summarised in Alg. Box 1.[Function or distribution](#)

### Metropolis–Hastings – correction step

Since the samples by the SIRT scheme are generated from an approximation, it is sensible to correct those using a Metropolis–Hastings (MH) importance step. In doing so, compute the acceptance probability  $\alpha = \min(w^{(s+1)}/w^{(s)}, 1)$ , where

$$w(\mathbf{x}) = \frac{\pi(\mathbf{x})}{f_X(\mathbf{x})} = \frac{\pi(\mathbf{x})}{\xi + \tilde{g}(\mathbf{x})^2} \quad (2.42)$$

is the importance ratio. Note the normalising constants in the ratio  $w^{(s+1)}/w^{(s)}$  cancel. In practice, we calculate the importance ratio in the log space, where  $\log f_X(\mathbf{x}) = \log f_{X_1}(x_1) + \log f_{X_2|X_1}(x_2|x_1) + \dots + \log f_{X_k|X_{<k}}(x_k|x_{k-1}, \dots, x_1)$  is given as in Eq. 2.40, see [11]. We refer to this as the SIRT-MH scheme, which provides the corrected chain  $\{\mathbf{x}_{\text{MH}}^{(1)}, \dots, \mathbf{x}_{\text{MH}}^{(N)}\} \sim \pi(\mathbf{x})$ .

**Algorithm 2:** MH correction step

```

1: Input: samples  $\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N+1)}\}$ , where each  $\mathbf{x}^{(s)} \in \mathbb{R}^d$  for  $s = 1, \dots, N + 1$ 
2: for  $s = 1, \dots, N$  do
3:   compute MH ratio  $\frac{w^{(s+1)}}{w^{(s)}} = \frac{\pi(\mathbf{x}^{(s+1)})}{\pi(\mathbf{x}^{(s)})} \frac{f_X(\mathbf{x}^{(s)})}{f_X(\mathbf{x}^{(s+1)})}$ 
4:   compute acceptance probability  $\alpha = \min(w^{(s+1)}/w^{(s)}, 1)$ 
5:   Draw  $u \sim \mathcal{U}(0, 1)$ 
6:   if  $\alpha \geq u$  then
7:     Accept and set  $\mathbf{x}_{\text{MH}}^{(s+1)} = \mathbf{x}^{(s+1)}$ 
8:   else
9:     Reject and keep  $\mathbf{x}_{\text{MH}}^{(s+1)} = \mathbf{x}^{(s)}$ 
10:  end if
11: end for
12: Output: corrected sample chain  $\{\mathbf{x}_{\text{MH}}^{(1)}, \dots, \mathbf{x}_{\text{MH}}^{(N)}\}$ , where each  $\mathbf{x}_{\text{MH}}^{(s)} \in \mathbb{R}^d$  for  $s = 1, \dots, N$ 
```

**2.3.3 Error of the TT Approximation**

A straightforward way to assess an average error from the TT approximation is to calculate the relative root mean squared (RMS) error

$$\left( \frac{\int_{\mathcal{X}} (\pi(\mathbf{x}) - (\xi + \tilde{g}(\mathbf{x})^2))^2 \lambda(\mathbf{x}) d\mathbf{x}}{\int_{\mathcal{X}} \pi(\mathbf{x})^2 \lambda(\mathbf{x}) d\mathbf{x}} \right)^{1/2} = \frac{\|\pi(\mathbf{x}) - (\xi + \tilde{g}(\mathbf{x})^2)\|_{L_\lambda^2(\mathcal{X})}}{\|\pi(\mathbf{x})\|_{L_\lambda^2(\mathcal{X})}}. \quad (2.43)$$

The RMS is approximated by

$$\left( \frac{1}{N} \sum_{i=1}^N \left( \pi(\mathbf{x}^{(i)}) - (\xi + \tilde{g}(\mathbf{x}^{(i)})^2) \right)^2 \lambda(\mathbf{x}^{(i)}) \right)^{1/2} \approx \left( \int_{\mathcal{X}} (\pi(\mathbf{x}) - (\xi + \tilde{g}(\mathbf{x})^2))^2 \lambda(\mathbf{x}) d\mathbf{x} \right)^{1/2} \quad (2.44)$$

and similarly  $\int_{\mathcal{X}} \pi(\mathbf{x})^2 \lambda(\mathbf{x}) d\mathbf{x}$ .

**Absolute error bound**

If large errors occur in regions with low probability the RMS is sensible to those, whereas the Wasserstein distance weights the distance measures according to their respective probability values.

The Wasserstein distance is the infimum over all couplings between two probability distributions with respect to some distance measure. The Kantorovich-Rubinstein duality, as in [56, 1], says that the 1-Wasserstein distance is equal to the supremum of differences in expectations over all 1-Lipschitz functions  $h$  between two probability distributions. So the 1-Wasserstein distance provides an upper absolute error bound and is defined as

$$W_1(\pi, \tilde{\pi}) = \inf_{\nu \in \Pi(\pi, \tilde{\pi})} \int_{\mathcal{X} \times \mathcal{X}} c_{\mathcal{X}}(\mathbf{x}, \tilde{\mathbf{x}}) \nu(\mathbf{x}, \tilde{\mathbf{x}}) d\mathbf{x} d\tilde{\mathbf{x}}, \quad (2.45)$$

where  $\nu$  couples  $\mathbf{x}$  and  $\tilde{\mathbf{x}}$  so that the integral over the distance  $c_{\mathcal{X}}(\mathbf{x}, \tilde{\mathbf{x}})$  weighted by the probability measures  $\pi$  and  $\tilde{\pi}$  is the greatest lower bound of all integrals with respect to  $\nu$  in the set of all couplings  $\Pi(\pi, \tilde{\pi})$ . Often  $\nu$  is the transport plan, where  $c_{\mathcal{X}}(\mathbf{x}, \tilde{\mathbf{x}})$  is the (ground) cost function, and  $\nu(\mathbf{x}, \tilde{\mathbf{x}})$  is related to the mass which has to be transported and the 1-Wasserstein distance is the earth mover distance. On the other hand (Kantorovich-Rubinstein duality) the 1-Wasserstein distance

$$W_1(\pi, \tilde{\pi}) = \sup_{h(\mathbf{x}); c_{\mathcal{Y}}(h(\mathbf{x}), h(\tilde{\mathbf{x}})) \leq c_{\mathcal{X}}(\mathbf{x}, \tilde{\mathbf{x}})} \left\{ \int_{\mathcal{X}} h(\mathbf{x}) d\pi(\mathbf{x}) - \int_{\mathcal{X}} h(\tilde{\mathbf{x}}) d\tilde{\pi}(\tilde{\mathbf{x}}) \right\} \quad (2.46)$$

$$= \sup_{h(\mathbf{x}); c_{\mathcal{Y}}(h(\mathbf{x}), h(\tilde{\mathbf{x}})) \leq c_{\mathcal{X}}(\mathbf{x}, \tilde{\mathbf{x}})} \left\{ \mathbb{E}_{\mathbf{x} \sim \pi}[h(\mathbf{x})] - \mathbb{E}_{\tilde{\mathbf{x}} \sim \tilde{\pi}}[h(\tilde{\mathbf{x}})] \right\}. \quad (2.47)$$

is the lowest upper bound of differences in expectations over all 1-Lipschitz function  $h(\mathbf{x}) : \mathcal{X} \rightarrow \mathcal{Y}$  in between the two distributions  $\pi$  and  $\tilde{\pi}$ , with the distance measure  $c_{\mathcal{X}}$  on the set  $\mathcal{X}$  forming the metric space  $(\mathcal{X}, c_{\mathcal{X}})$  and similarly the metric space  $(\mathcal{Y}, c_{\mathcal{Y}})$ .

For two sample sets  $\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}\} \sim \pi$  and  $\{\tilde{\mathbf{x}}^{(1)}, \dots, \tilde{\mathbf{x}}^{(M)}\} \sim \tilde{\pi}$  the calculation of the Wasserstein distance becomes an optimisation problem that is to find the best coupling of samples weighted by their distribution value according to an appropriate distance measure [15], which we set to  $c_{\mathcal{X}}(\mathbf{x}, \tilde{\mathbf{x}}) = \|\mathbf{x} - \tilde{\mathbf{x}}\|_{L^2}$ . More specifically,

$$W_1(\pi, \tilde{\pi}) = \min_{\nu \in \Pi(\pi, \tilde{\pi})} \sum_{j=1}^M \sum_{i=1}^N \nu_{ij} \|\mathbf{x}^{(i)} - \tilde{\mathbf{x}}^{(j)}\|_{L^2}, \quad (2.48)$$

where the transport plan  $\nu \in \mathbb{R}_{\geq 0}^{N \times M}$  defines the coupling  $\nu_{ij} \in \nu$  as  $\nu_{ij} := \pi(\mathbf{x}^{(i)})\tilde{\pi}(\tilde{\mathbf{x}}^{(j)})$  similar to [15, Eq. 3.166]. Additionally it is required that  $\sum_{i=1}^N \pi(\mathbf{x}^{(i)}) = \sum_{j=1}^M \tilde{\pi}(\tilde{\mathbf{x}}^{(j)}) = 1$ . This gives us an upper bound of the absolute error between the expected value of any 1-Lipschitz function  $h$ .

## 2.4 Regularisation Approach

The currently most used method to analyse data in atmospheric physics is regularisation-based. Since we want to show that Bayesian methods provide more information than regularisation at a similar computational cost, the chosen regularisation approach is the closest equivalent to the linear-Gaussian Bayesian framework [17] in Sec. 4.1.

For a linear forward model matrix  $\mathbf{A}$ , data  $\mathbf{y}$  and a regularisation operator  $\mathbf{T}$ , the regularisation approach provides one solution  $\mathbf{x}$  that minimises both the data misfit norm

$$\|\mathbf{y} - \mathbf{Ax}\|_{L^2} \quad (2.49)$$

and a regularisation norm

$$\|\mathbf{T}\mathbf{x}\|_{L^2}. \quad (2.50)$$

For a fixed regularisation parameter  $\lambda > 0$ , the regularised solution as in [23, 17, 55] is given by  $\mathbf{x}_\lambda$  that minimises the weighted sum

$$\mathbf{x}_\lambda = \arg \min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_{L^2}^2 + \lambda \|\mathbf{T}\mathbf{x}\|_{L^2}^2 , \quad (2.51)$$

which can be calculated by taking the derivative with respect to  $\mathbf{x}$ :

$$\nabla_{\mathbf{x}} \left\{ (\mathbf{y} - \mathbf{A}\mathbf{x})^T (\mathbf{y} - \mathbf{A}\mathbf{x}) + \lambda \mathbf{x}^T \mathbf{T}^T \mathbf{T} \mathbf{x} \right\} = 0 \quad (2.52)$$

$$\iff \nabla_{\mathbf{x}} \left\{ \mathbf{y}^T \mathbf{y} + \mathbf{x}^T \mathbf{A}^T \mathbf{A} \mathbf{x} - 2 \mathbf{y}^T \mathbf{A} \mathbf{x} + \lambda \mathbf{x}^T \mathbf{T}^T \mathbf{T} \mathbf{x} \right\} = 0 \quad (2.53)$$

$$\iff 2 \mathbf{A}^T \mathbf{A} \mathbf{x} - 2 \mathbf{A}^T \mathbf{y} + 2\lambda \mathbf{T}^T \mathbf{T} \mathbf{x} = 0. \quad (2.54)$$

Eq. 2.54 yields the regularised solution

$$\mathbf{x}_\lambda = (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{L})^{-1} \mathbf{A}^T \mathbf{y} , \quad (2.55)$$

where we define  $\mathbf{L} := \mathbf{T}^T \mathbf{T}$ . Typically  $\mathbf{L}$  represents a discrete matrix approximation of a derivative operator [55]. For example

$$\mathbf{T} = \frac{1}{h} \begin{bmatrix} -1 & 1 \\ 0 & -1 & 1 \\ & \ddots & \ddots & \ddots \\ & & 0 & -1 & 1 \\ & & & 0 & -1 \end{bmatrix} , \quad (2.56)$$

is the first order forward difference operator with equal spacing  $h$  as in [55] that approximates the first derivative operator. Then

$$\mathbf{T}^T \mathbf{T} = \frac{1}{h^2} \begin{bmatrix} 1 & -1 \\ -1 & 2 & -1 \\ & \ddots & \ddots & \ddots \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{bmatrix} , \quad (2.57)$$

is a discrete approximation to the second derivative with Neumann boundary conditions [60].

If  $\lambda$  is large, then the effect of the data on the solution  $\mathbf{x}_\lambda$  is small or negligible and dominated by the regulariser resulting in an under fitted  $\mathbf{x}_\lambda$ . If  $\lambda$  is small, the solution  $\mathbf{x}_\lambda$  will be dominated by the data misfit norm. Then  $\mathbf{x}_\lambda$  is sensitive to noise resulting in an over fitted solution. We refer to [24] and [55] for a more comprehensive analysis on the effects of the regularisation parameter on the solution, e.g. due to small singular values of the forward model.

In practice,  $\mathbf{x}_\lambda$  is computed for a range of  $\lambda$ -values and the data misfit norm versus the regularisation norm is plotted in log-space to form an L-curve (see Fig. 4.8). Based on the trade-off between the data misfit and the regularisation norm the regularisation solution corresponds to the point of maximum curvature of the L-curve [25].

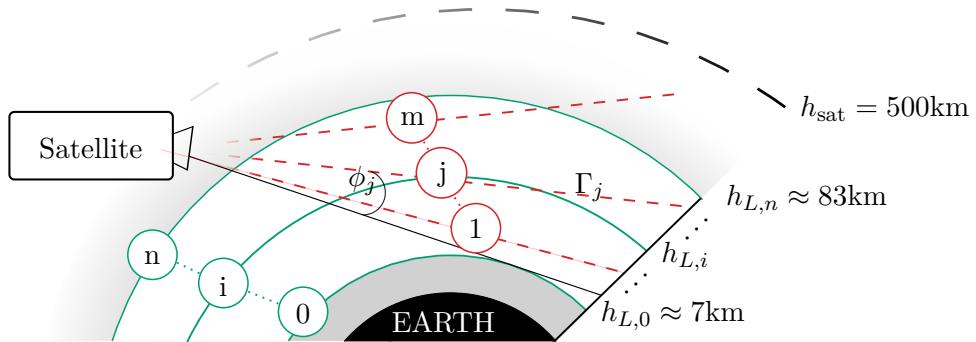
Alternatively one can introduce a Lagrangian  $\mathcal{L}(\mathbf{x}, \lambda) := \lambda \mathbf{x}^T \mathbf{L} \mathbf{x} + \|\mathbf{y} - \mathbf{A} \mathbf{x}\|_{L^2}^2$  similar to [29]. Then a solution  $\mathbf{x}_\lambda = \arg \min_{\mathbf{x}} \mathbf{x}^T \mathbf{L} \mathbf{x}$  minimises the squared regularisation norm with respect to constant  $\|\mathbf{y} - \mathbf{A} \mathbf{x}\|_{L^2}^2$  (see [17, fn. 6] and [47, Fig. 2.13]). So every solution  $\mathbf{x}_\lambda$  is extremely regularised for a given data misfit and the L-Curve presents the limit to on an open set of  $\mathbf{x}$ . Hence every sample of the posterior, which represents a feasible solution given the data, lays above the L-Curve because it is less regularised and has a higher  $\mathbf{x}^T \mathbf{L} \mathbf{x}$  value. So regularised solutions are representative of the posterior parameter space.

say on or above - do any points in the posterior actually lie in the L-Curve, I don't think so. Are you claiming that the regularised solutions are in the posterior? The L-curve could be the limit of an open set.

# 3

## The Forward Model

In this Chapter, we present the forward model to which we apply our entire methodology and conduct a singular value analysis for different measurement scenarios to understand the forward model and to determine a sensible way to measure ozone. We follow the Michelson interferometer for passive atmospheric sounding (MIPAS) handbook [39] and simulate data according to an idealised cloud-free atmosphere in local thermodynamic equilibrium, assuming a measurement instrument with infinite spectral resolution and no pointing errors. This is a simplified forward model and we do not include any other instrument-specific details, such as sensor area or antenna response, as they are not available to us.



**Figure 3.1:** Schematic of measurement and analysis geometry, not to scale. The stationary satellite, at a constant height  $h_{\text{sat}}$  above Earth, takes  $m$  measurements along its line-of-sight defined by the line  $\Gamma_j$ . Each measurement has a pointing angle  $\phi_j$  and a tangent height  $h_{\ell,j}$ ,  $j = 1, 2, \dots, m$  defined as the closest distance of  $\Gamma_j$  to the Earth's surface. Between  $h_{L,0} \approx 7\text{km}$  and  $h_{L,n} \approx 83\text{km}$ , the atmosphere is discretised into  $n$  layers as illustrated by the solid green lines.

### 3.1 Radiative Transfer Equation

A satellite at a constant height  $h_{\text{sat}}$  points through the atmosphere (limb-sounding) and measures thermal radiation of gas molecules along its straight line of sight  $\Gamma_j$ , see Figure 3.1. One measurement of the thermal radiation of one specific molecule, in our case ozone, denoted by the ozone volume mixing ratio (VMR)  $x(r)$  at distance  $r$  from the satellite, at the wave number  $\nu$ , is given by the path integral

$$y_j = \int_{\Gamma_j} B(\nu, T) k(\nu, T) \frac{p(T)}{k_B T(r)} x(r) \tau(r) dr + \eta_j \quad (3.1)$$

$$\tau(r) = \exp \left\{ - \int_{r_{\text{obs}}}^r k(\nu, T) \frac{p(T)}{k_B T(r')} x(r') dr' \right\}, \quad (3.2)$$

which is the radiative transfer equation (RTE) [39]. For more information on the processes within the atmosphere for ozone, we refer to [28]. We define a tangent height  $h_{\ell_j}$  and  $\Gamma_j$  for each  $j = 1, 2, \dots, m$ , so that the data vector  $\mathbf{y} \in \mathbb{R}^m$  including some additive noise  $\eta_j$ . Additionally, the pointing angle  $0 \leq \phi_j < \phi_{\max}$  is defined so that if  $\phi = 0$  arc sec the satellite points at  $h_{L,0}$  and for a pointing angle  $\phi_{\max}$  at  $h_{L,n}$ . 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

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

is dependent on the line intensity  $L(\nu, T_{\text{ref}})$  at reference temperature  $T_{\text{ref}} = 296K$ , the lower-state energy  $E''$  in  $\text{cm}^{-1}$  of the targeted transition and the second radiation constant  $c_2 := hc/k_B \approx 1.44\text{cmK}$  as in the HITRAN database [22], with Planck's constant  $h$  and speed of light  $c$ . Since we assume that the measurement device has a negligible frequency window, we neglect line broadening around  $\nu$  for the calculations of  $L(\nu, T_{\text{ref}})$ , which would normally be modelled as a convolution of the normalised Lorentz profile (collisional/pressure broadening) and the normalised Doppler (thermal broadening) profile [39]. Additionally, we target one specific molecule and calculate  $k(\nu, T)$  accordingly, which usually would involve summing the individual absorption constants for multiple radiating molecules weighted by their respective VMR [39]. The total internal partition function is given as:

$$Q(T) = g' \exp \left\{ -\frac{c_2 E'}{T} \right\} + g'' \exp \left\{ -\frac{c_2 E''}{T} \right\}, \quad (3.4)$$

with the statistical weight  $g''$  for the lower and  $g'$  for the upper energy state (also called the degeneracy factors) accounting for the molecule's non-rotational and rotational energy

states (see [51]), and the upper state energy  $E' = E'' + \nu$ . Under the assumption of local thermodynamic equilibrium (LTE), the black body radiation acts as a source function

$$B(\nu, T) = \frac{2hc^2\nu^3}{\exp\left\{\frac{c_2\nu}{T}\right\} - 1}. \quad (3.5)$$

For fundamentals on the RTE, we recommend [46, Chapter 1], and for a more comprehensive model, we refer to [38].

### 3.2 Simulate Data Based on a Ground Truth

To calculate the integrals in Eq. 3.1 and Eq. 3.2 numerically, we discretise the atmosphere in  $n$  layers and define height values  $h_{L,i-1} < h_{L,i}$  with respect to the surface of the earth, for  $i = 1, \dots, n$ . The  $i$ -th layer is defined by two spheres around the centre of the earth with radii  $r_0 + h_{L,i-1}$  and  $r_0 + h_{L,i}$ . Then the ozone VMR  $\mathbf{x} = \{x_1, x_2, \dots, x_n\} \in \mathbb{R}^n$ , pressure  $\mathbf{p} = \{p_1, p_2, \dots, p_n\} \in \mathbb{R}^n$  and temperature  $\mathbf{T} = \{T_1, T_2, \dots, T_n\} \in \mathbb{R}^n$ , as well as all other height dependent parameters, are discretised profiles with constant values between the heights  $h_{L,i-1} \leq h < h_{L,i}$ . Above  $h_{L,n}$  and below  $h_{L,0}$ , the ozone VMR is set to zero, so no signal can be obtained. We evaluate the integral in Eq. (3.1) for one noise-free measurement  $A_j(\mathbf{p}, \mathbf{T}, \mathbf{x})$ , using the trapezoidal rule. Here, each entry  $A_j$  of  $\mathbf{A}(\mathbf{p}, \mathbf{T}, \mathbf{x}) \in \mathbb{R}^m$  includes multiple evaluations of the integral in Eq. 3.2 to calculate the absorption  $\tau(r)$ . The data vector

$$\mathbf{y} = \mathbf{A}(\mathbf{x}) + \boldsymbol{\eta} \quad (3.6)$$

includes an additive noise vector  $\boldsymbol{\eta} \in \mathbb{R}^m$ , where we define the non-linear forward model as  $\mathbf{A}(\mathbf{x}) := \mathbf{A}(\mathbf{x}, \mathbf{p}, \mathbf{T}) \in \mathbb{R}^m$  for brevity. Similarly, we define  $\mathbf{A}_L \in \mathbb{R}^{m \times n}$ , which denotes the linear forward model matrix and neglects absorption (e.g. set  $\tau = 1$  in Eq. (3.2)) and enables matrix-vector multiplication  $\mathbf{A}_L \mathbf{x}$  to compute noise-free linear data. Further, we classify the inverse problem as a weakly non-linear inverse problem, because neglecting the absorption changes the measurements only slightly (about 1%, see Chapter 5).

As the ground truth for our methodology, we consider an ozone profile at distinct pressure values generated from some data [49] of the MLS on the Aura satellite within the Antarctic region. This ozone profile has a peak in the middle atmosphere and a second peak at higher altitudes, see Fig. 4.5, which seems to be a typical nighttime profile [28].

We recursively relate pressure  $p$  to geometric height  $h$  with the hydrostatic equilibrium equation

$$\frac{dp}{p} = \frac{-gM}{R^*T} dh, \quad (3.7)$$

starting with a pressure of 1013.25hPa at sea level. The acceleration due to gravity is

$$g = g_0 \left( \frac{r_0}{r_0 + h} \right), \quad (3.8)$$

where the polar radius of the earth is  $r_0 \approx 6356$  km, the gravitation at sea level is  $g_0 \approx 9.81$  m/s<sup>2</sup>,  $R^* = 8.31432 \times 10^{-3}$  Nm/kmol/K and the mean molecular weight of the air is  $M = 28.97$  kg/kmol [57]. This holds up to a geometric height of 86km, where we ignore a 0.04% non-linear change in  $M$  from 80km to 86km.

Following [57] we form the temperature function

$$T(h) = \begin{cases} T_0 & , h = 0 \\ T_0 + a_0 h & , 0 \leq h < h_{T,1} \\ T_0 + a_0 h_{T,1} & , h_{T,1} \leq h < h_{T,2} \\ T_0 + a_0 h_{T,1} + a_1 (h_{T,2} - h_{T,1}) + a_2 (h - h_{T,2}) & , h_{T,2} \leq h < h_{T,3} \\ T_0 + a_0 h_{T,1} + a_1 (h_{T,2} - h_{T,1}) \\ + a_2 (h_{T,3} - h_{T,2}) + a_3 (h - h_{T,3}) & , h_{T,3} \leq h < h_{T,4} \\ T_0 + a_0 h_{T,1} + a_1 (h_{T,2} - h_{T,1}) \\ + a_2 (h_{T,3} - h_{T,2}) + a_3 (h_{T,4} - h_{T,3}) + a_4 (h - h_{T,4}) & , h_{T,4} \leq h < h_{T,5} \\ T_0 + a_0 h_{T,1} + a_1 (h_{T,2} - h_{T,1}) \\ + a_2 (h_{T,3} - h_{T,2}) + a_3 (h_{T,4} - h_{T,3}) + a_4 (h_{T,5} - h_{T,4}) \\ + a_5 (h - h_{T,5}) & , h_{T,5} \leq h < h_{T,6} \\ T_0 + a_0 h_{T,1} + a_1 (h_{T,2} - h_{T,1}) \\ + a_2 (h_{T,3} - h_{T,2}) + a_3 (h_{T,4} - h_{T,3}) + a_4 (h_{T,5} - h_{T,4}) \\ + a_5 (h_{T,6} - h_{T,5}) + a_6 (h - h_{T,6}) & , h_{T,6} \leq h \lesssim 86 \end{cases} \quad (3.9)$$

with gradient and height values provided by [57] (see Tab. 3.1) which acts as the ground truth temperature profile (see Fig. 6.2).

We calculate one measurement according to the RTE as in Eq. 3.1 and Eq. 3.2 using the trapezoidal integration rule. We assume an atmosphere between  $h_{L,1} = 6.9$  km and  $h_{L,n} = 83.3$  km with  $n = 45$  equidistant layers and a satellite at a fixed height of  $h_{\text{sat}} = 500$  km (see Fig. 3.1). The height value  $h_{L,i}$  for each layer  $i = 1, \dots, n$  is defined by the pressure values from [49] and the hydrostatic equilibrium equation, see Eq. 3.7. We target ozone at a frequency of 235.71GHz, which lies within the region where the MLS observes ozone [32, 61]. The corresponding wave number is  $\nu = 7.86\text{cm}^{-1}$ . We calculate the absorption constant  $k(\nu, T)$  as in Eq. 3.2, following the high resolution transmission (HITRAN) database [22], which provides the line intensity  $L(\nu, T_{\text{ref}})$  for the isotopologue  $^{16}\text{O}_3$  with the AFGL Code 666. Lastly, we add independent and identically-distributed

subscript $i$	geometric height $h_{T,i}$ in km	gradient $a_i$
0	0	-6.5
1	11	0
2	20.1	1
3	32.2	2.8
4	47.4	0
5	51.4	-2.8
6	71.8	-2

**Table 3.1:** Definition of height depending temperature gradients.

Gaussian noise  $\boldsymbol{\eta} \sim \mathcal{N}(0, \gamma^{-1} \mathbf{I})$  so that the SNR = 150 (see Eq. 3.11) similar to [20], where a signal with a maximal spectral intensity of around 100K and a noise range of 0.4 to 1.6K is reported. We note that the methods used in this thesis will work with different SNRs or other frequencies. Before computing a data vector  $\mathbf{y} = \mathbf{A}(\mathbf{x}) + \boldsymbol{\eta}$  we decide for a measurement strategy with respect to the information provided by the forward model.

### 3.2.1 Understanding the Forward Model

By choosing a measurement strategy, we provide a quick and intuitive way of assessing if the data collection is effective, how much information is passed through the forward model, and how the signal-to-noise ratio (SNR) and the measurement strategy affect that information. One way of doing this is via a singular value decomposition (SVD) of the linear forward model matrix  $\mathbf{A}_L \in \mathbb{R}^{m \times n}$ , which is given as

$$\mathbf{A}_L = \sum_{i=1}^r \mathbf{u}_i \sigma_i \mathbf{v}_i^T = \mathbf{U} \Sigma \mathbf{V}^T \quad (3.10)$$

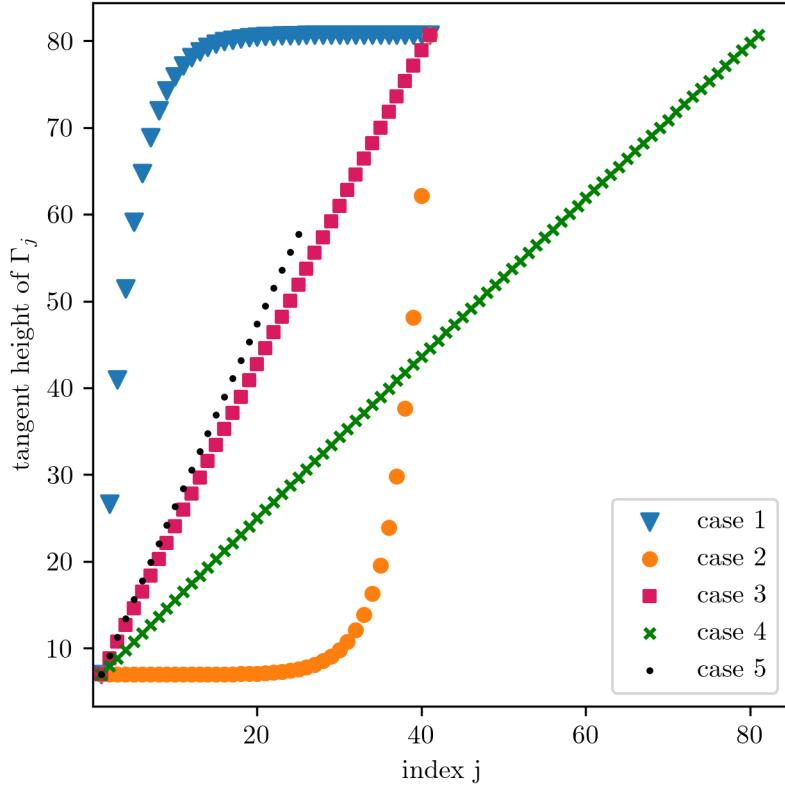
with  $r = \min\{m, n\}$ . Our main objective is to measure ozone  $\mathbf{x}$ , so our forward model  $\mathbf{A}_L$  includes temperature and pressure, the latter is dominant, see Fig. 6.4, decreases exponentially in height and hence does affect the information passed through the model. If the pressure is high, the signal is large. If the pressure is low, the signal is low, and the data tends to be noise-dominated.

The SVD of the forward model provides information on how  $\mathbf{A}_L$ , more specifically the right singular vectors  $\mathbf{v}_i$ , act on the parameter  $\mathbf{x}$  since a noise-free measurement vector is given as  $\mathbf{A}_L \mathbf{x}$ . The singular values  $\sigma_i$ , ordered in size from the largest  $\sigma_1$  to the smallest  $\sigma_r$ , weigh that information from the right singular vectors to the left singular vectors  $\mathbf{u}_i$ . Then the left singular vectors project  $\sigma_i \mathbf{v}_i^T \mathbf{x}$  onto the data space. For a large singular value, we can say that the forward model is informative about parameter structures according to the corresponding right singular vector. For a small singular value, we can

say that the forward model is uninformative about parameter structures according to the corresponding right singular vector. Further, if we define the SNR as

$$\text{SNR} := \frac{\max(y)}{\text{STD noise}} = \frac{\text{peak signal}}{\text{RMS noise}}, \quad (3.11)$$

we can roughly assume that the maximum singular value  $\max(y) \approx \sigma_1$  and the information transmitted through the forward model corresponds roughly to the singular values  $\sigma_i \gtrsim \max(y)/\text{SNR}$  [18]. Then, for very small singular values  $\sigma_i \ll \sigma_1/\text{SNR}$  below the RMS noise level or the noise standard deviation (STD), an effective rank  $r_{\text{eff}} \leq r$  is introduced. The parameter space spanned by  $\{\mathbf{v}_{r_{\text{eff}}+1}, \dots, \mathbf{v}_r\}$  is noise-dominated in the corresponding data space, see Figure 3.6. For singular values around the SNR, we expect reconstruction to have larger uncertainties, as that is where the noise is starting to influence the data. For large singular values above the SNR, we expect reconstructions to be close to the ground truth as the corresponding data is hardly influenced by the noise. See [55] for a more comprehensive analysis.



**Figure 3.2:** Tangent heights for five different sequences of measurements.

We test five different measurement strategies and plot the tangent heights corresponding to the pointing angles in Fig. 3.2. The measurement test cases are:

- **Case 1** includes 42 measurements between heights of  $\approx 7\text{km}$  and  $\approx 83\text{km}$  with pointing angles

$$\phi_j = \left( \frac{-1}{1.25^{j-1}} + 1 \right) \phi_{\max}, \quad \text{for } j = 1, \dots, 42.$$

- **Case 2** includes 42 measurements between heights of  $\approx 7\text{km}$  and  $\approx 83\text{km}$  with pointing angles

$$\phi_j = \frac{1.25^{j-1}}{1.25^{m-1}} 0.99 \phi_{\max}, \quad \text{for } j = 1, \dots, 42.$$

- **Case 3** includes 42 measurements between heights of  $\approx 7\text{km}$  and  $\approx 83\text{km}$  with pointing accuracy 150arc sec and pointing angles

$$\phi_j = (j - 1)150\text{arc sec}, \quad \text{for } j = 1, \dots, 42.$$

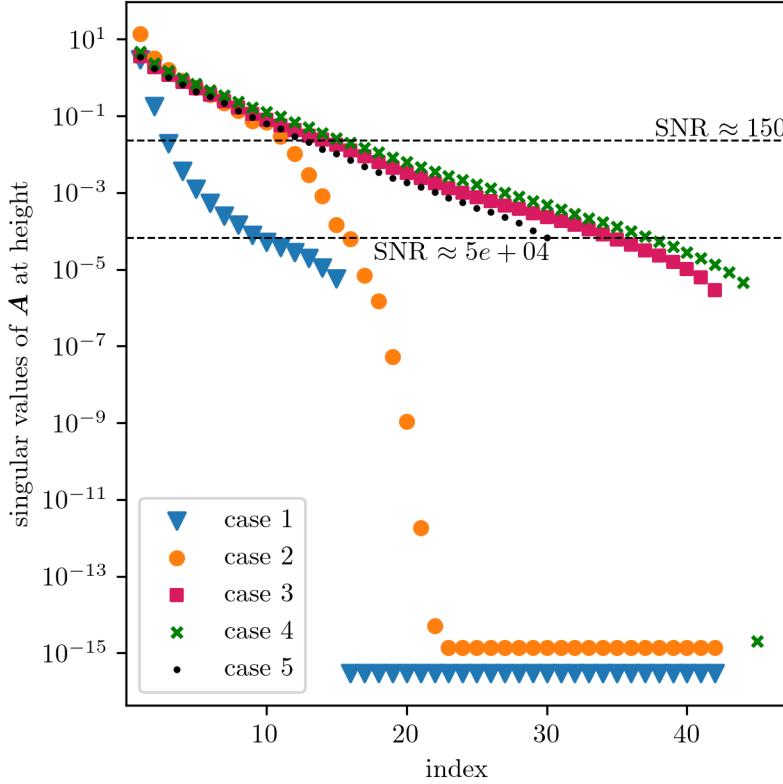
- **Case 4** includes 83 measurements between heights of  $\approx 7\text{km}$  and  $\approx 83\text{km}$  with pointing accuracy 77.5arc sec and pointing angles

$$\phi_j = (j - 1)77.5\text{arc sec}, \quad \text{for } j = 1, \dots, 83.$$

- **Case 5** includes 30 measurements between heights of  $\approx 7\text{km}$  and  $\approx 68\text{km}$  with pointing accuracy 175arc sec and pointing angles

$$\phi_j = (j - 1)175\text{arc sec}, \quad \text{for } j = 1, \dots, 30.$$

Case 1 collects more data in low signal regions at high altitudes. Case 2 collects more data in high signal regions at low altitudes. Case 3, case 4, and case 5 measure at equidistantly spaced pointing angles corresponding to different pointing accuracies. The pointing accuracy determines how well the satellite can point in a certain direction and, hence, the spacing of tangent heights in the atmosphere for a stable satellite at  $h_{\text{sat}}$ . Case 1, case 2, case 3, and case 4 measure in between heights  $h_{L,1} = 6.9\text{km}$  and  $h_{L,n} = 83.3\text{km}$ , case 5 does not collect measurements in high altitude regions. The pointing accuracy for case 3 in Fig. 3.2 of 150arc sec was given to us by the team of the University of New South Wales Canberra Space [14]. Case 4 has half the pointing accuracy of case 3, and case 5 has a slightly larger pointing accuracy than case 3. We visually assess the effective rank and how the singular values behave to determine which of the test cases is most effective. More specifically, if the singular values decay fast and only a few singular values are above an SNR of 150, the forward map is rather uninformative. If the singular values decay less fast and more singular values are above an SNR of 150, we classify the forward map as informative.



**Figure 3.3:** Singular values of the forward model matrix for different sequences of measurements. The corresponding tangent heights of the test cases are plotted in Fig. 3.2. The dotted vertical line marks an SNR according to  $\sigma_1$  of measurement case 5.

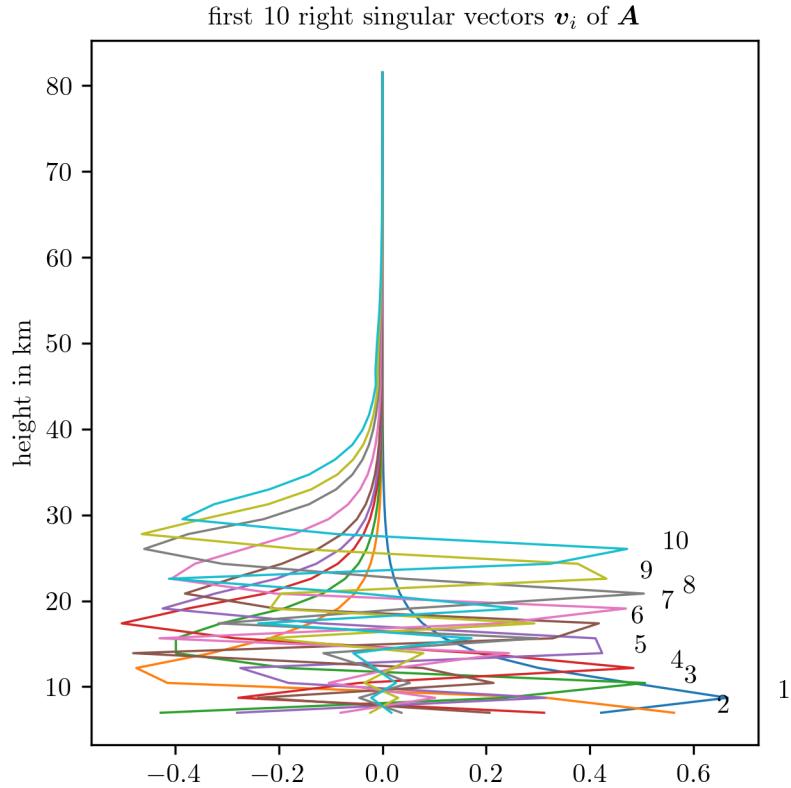
In Fig. 3.3, we plot the singular values for each of the measurement cases. The dotted lines in Fig. 3.3 correspond to an SNR of roughly 150 with respect to the largest singular value of case 5 and an SNR according to the lowest singular value of case 5, which would be required to reconstruct all information provided by the forward model. The largest singular value of case 5 has approximately the same value as the largest singular values of case 1, case 3, and case 4, so the SNR line is indicative for those cases as well.

Fig. 3.3 shows that case 2 has the largest singular value of all cases, but its singular values decrease faster than the singular values of cases 3, 4, and 5 in the region where noise is becoming dominant. In contrast, case 1 does obtain singular values which are decreasing the fastest of all cases and the largest number of singular values below an SNR of 150. Additionally, cases 1 and 2 have a large number of very low singular values, and the smallest effective ranks of all cases, so we conclude that cases 1 nor 2 is effective.

The singular values of cases 3, 4, and 5 with equidistantly spaced pointing angles behave similarly and have larger effective ranks than case 1 or 2. Case 4 measures almost twice as much compared to case 3, but does not provide much more information, which would justify the engineering effort required to achieve such pointing accuracy. The slightly

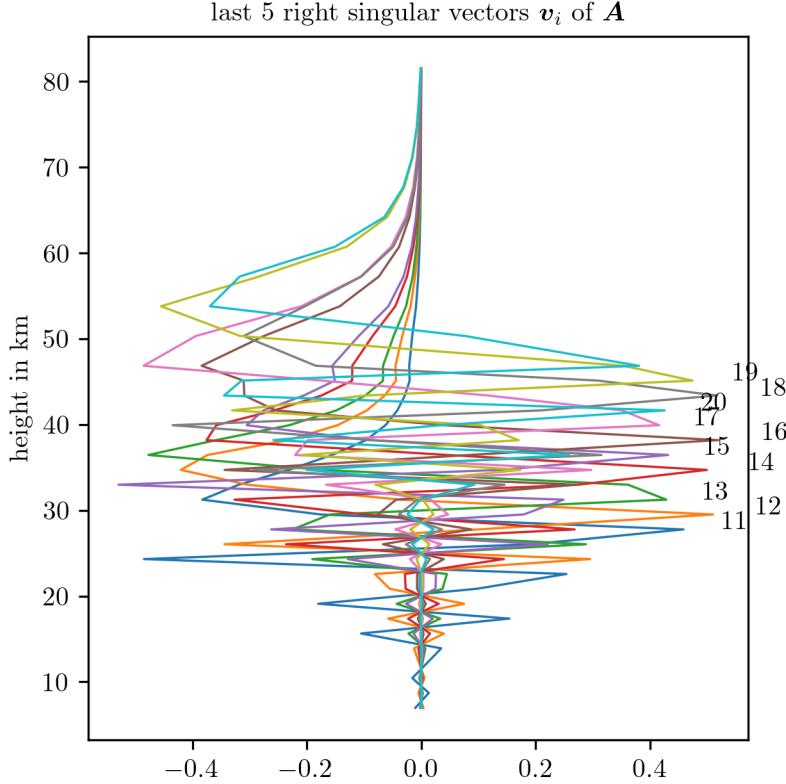
larger pointing accuracy in case 5 compared to case 3 provides similar information. The last 5 to 10 singular values of case 3 are so small that the information will be completely covered by the noise. Hence, we do not measure in noise-dominated regions and case 5 measures only up to a height of  $\approx 68\text{km}$  instead of  $\approx 83\text{km}$  without losing too much crucial information above the SNR. We note that if one wanted to obtain all information provided by the forward model, we would need an SNR of roughly  $10^4$ .

In principle, we show that it does matter how one measures, but one cannot get much more information by measuring more in regions where the information content is low or high. The test cases show that an efficient measurement strategy may consist of equidistantly spaced pointing angles and does stop measuring when the singular values (information) are too low. Consequently, we proceed with case 5 and plot the right singular vectors of the forward model versus height in the atmosphere to see to which parameter structures our model is sensitive.



**Figure 3.4:** First 10 right singular vectors of the forward model matrix for measurements case 5 in Fig. 3.2. These singular vectors correspond to high singular values of the forward model in Fig. 3.3.

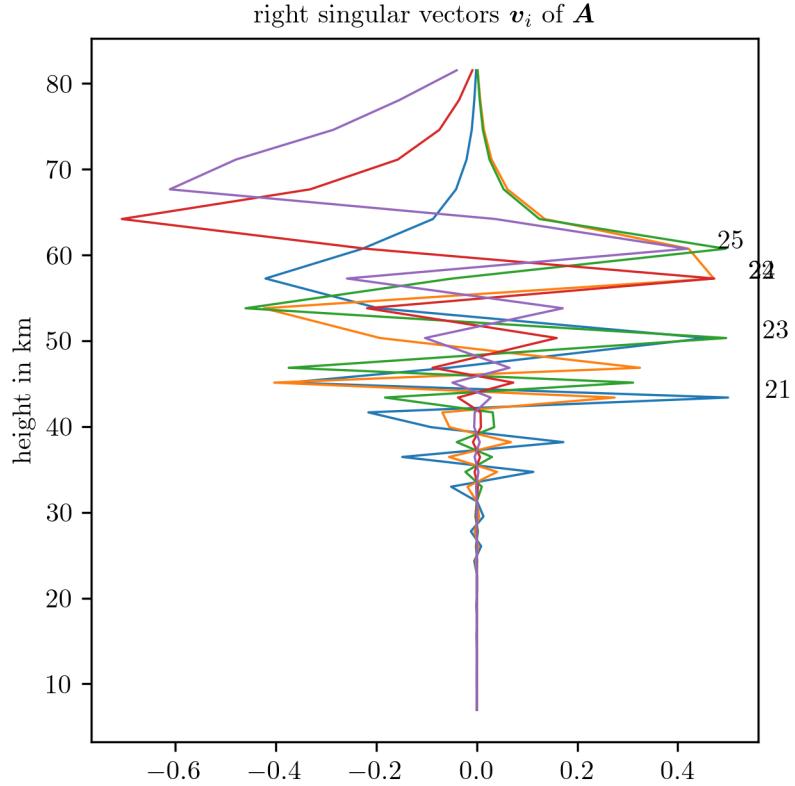
The parameter space of  $\mathbf{A}_L$  spanned by the first 10 right singular vectors plotted in Fig. 3.4 corresponds to the 10 largest singular values in Fig. 3.3 and represents parameter structures in the lower atmospheric regions. So we can assume that, given some data,



**Figure 3.5:** Right singular vectors with index  $i = 11, \dots, 19$  of the forward model matrix for measurements case 5 in Fig. 3.2. These singular vectors correspond to singular values in Fig. 3.3, where the noise level is similar to the data.

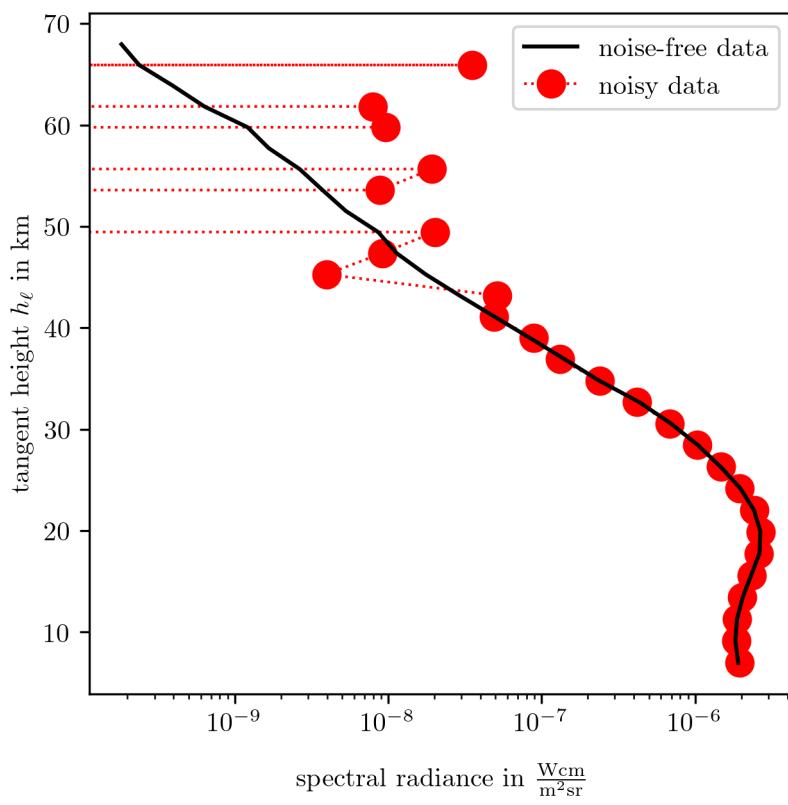
we will be able to provide good reconstructions of the parameter in lower altitudes up to  $\approx 30$ km. The right singular vectors in Fig. 3.5 correspond to the singular values  $\sigma_j$  for  $j = 11, \dots, 20$  around the SNR of 150, in Fig. 3.3. This is roughly where the noise starts to dominate the data, and the parameter space spanned by those right singular vectors represents parameter values in the middle atmosphere. Consequently, we expect an increasing uncertainty of reconstructed parameter values at heights between  $\approx 20$ km and  $\approx 55$ km. The singular vectors in Fig. 3.6 corresponding to the smallest 10 singular values and span structures in parameter space at higher altitudes. Since the singular values are very small, we will not be able to reconstruct parameter values from the ground truth above  $\approx 55$ km and the data is noise-dominated. More specifically, the retrieved parameter values at higher altitudes will be mostly determined by the prior or, in the case of regularisation, by the regulariser [55].

Now we can compute a data vector  $\mathbf{y} = \mathbf{A}(\mathbf{x}) + \boldsymbol{\eta}$ , with  $m = 30$  measurements determined by the satellite pointing accuracy of 175arc sec (see case 5 in Fig. 3.2), according to the RTE as in Eq. 3.1 and Eq. 3.2 using the trapezoidal integration rule. As already mentioned, we set the SNR to 150 and plot the data in Fig. 3.7, which is



**Figure 3.6:** Last 10 right singular vectors of the forward model matrix for measurements case 5 in Fig. 3.2. These singular vectors correspond to small singular values of the forward model in Fig. 3.3, where the data is noise-dominated.

noise-dominated in higher altitudes. Now, given the data, we like to determine the posterior distributions over ozone  $\mathbf{x}$ , pressure  $\mathbf{p}$  and temperature  $\mathbf{T}$ .



**Figure 3.7:** Logarithmic plot of data points at different tangent height. Note that negative values are not plotted, and noise is dominating at higher altitudes.

# 4

## Linear Bayesian vs. Regularisation – Ozone

In this Chapter we set up an hierarchiacal Bayesian model and apply the MTC method to provide posterior ozone profiles. In this chapter, we guide the reader through the process of setting up a Bayesian framework and ultimately obtaining the posterior distributions of hyper-parameters and parameters of interest, DAGs to visualise hierarchical and conditional depdenseies of the parameters establish a choice of prior distributions within our Bayesian model and formulate the posterior distributions. We treat this inverse problem as a linear inverse problem by neglecting the absorption term in RTE (see Eq. 3.1). We employ a Metrolpis within Gibbs sampler and a TT approxiamtion to characterise the marginal posterior and elaborate on some errors. Then we calculate the mean and the covariance matrix ofposterior for ozone, Lastly, we compare to a regualrisaiotn approach.

### 4.1 Hierarchical Bayesian Framework for Ozone

In this section, we set up the hierarchically-ordered linear-Gaussian Bayesian framework to determine the ozone posterior distribution, conditioned on ground truth temperature and pressure. Where, for now, we define the forward model matrix  $\mathbf{A} := \mathbf{A}_L$  and define the distributions of that Bayesian model, similarly to a regularisation approach, as:

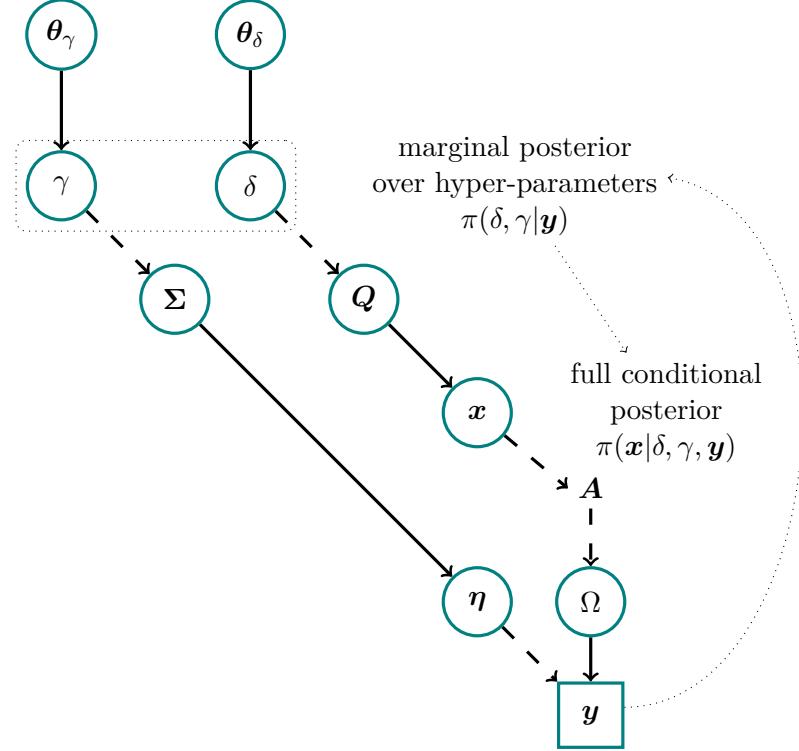
$$\mathbf{y}|\mathbf{x}, \gamma, \delta \sim \mathcal{N}(\mathbf{A}\mathbf{x}, \gamma^{-1}\mathbf{I}) \quad (4.1a)$$

$$\mathbf{x}|\delta \sim \mathcal{N}(\mathbf{0}, (\delta\mathbf{L})^{-1}) \quad (4.1b)$$

$$\delta \sim \Gamma(\alpha_\delta, \beta_\delta) \quad (4.1c)$$

$$\gamma \sim \Gamma(\alpha_\gamma, \beta_\gamma). \quad (4.1d)$$

Assuming Gaussian noise  $\boldsymbol{\eta} \sim \mathcal{N}(0, \gamma^{-1}\mathbf{I})$ , the likelihood function is a normal distribution with mean  $\mathbf{Ax}$  and covariance matrix  $\gamma^{-1}\mathbf{I}$ . We define the normal prior-distribution



**Figure 4.1:** DAG for visualisation of hierarchical modelling and measuring process of ozone, including the MTC scheme. The hyper-parameter  $\gamma$  deterministically (dotted line) sets the noise covariance  $\Sigma = \gamma^{-1} \mathbf{I}$  and hence the random (solid line) noise vector  $\eta \sim \mathcal{N}(0, \gamma^{-1} \mathbf{I})$ . The hyper-parameter  $\delta$  determines (dotted line) the prior precision matrix  $Q = \delta \mathbf{L}$  for the normally distributed (solid line) prior  $x | \delta \sim \mathcal{N}(0, \delta \mathbf{L})$ , where  $\mathbf{L}$  is a graph Laplacian, see Eq. 4.2. The hyper-prior distributions (solid line)  $\pi(\delta, \gamma)$  are defined by  $\theta_\gamma$  and  $\theta_\delta$ . Through a linear forward model  $\mathbf{A}$ , we generate a space of all measurable noise-free data  $\mathbf{Ax}$  from which we randomly observe a data set  $\mathbf{y}$  including some added noise  $\eta$ . Within the MTC scheme, we evaluate the marginal posterior over the hyper-parameters  $\pi(\gamma, \delta | \mathbf{y})$  first and then the full conditional posterior  $\pi(x | \delta, \gamma, \mathbf{y})$ . This breaks the correlation structure of  $x$  and  $\delta$  and  $\gamma$ , and allows us to evaluate the marginal posterior independent of  $x$ .

$\pi(x | \delta)$  with zero mean and precision matrix  $\delta \mathbf{L}$ , where  $\delta$  is a smoothness hyper-parameter and  $\mathbf{L}$  is the second order discrete derivate operator (see Eq. 4.2). Here the hyper-prior distributions  $\pi(\delta)$  and  $\pi(\gamma)$  are gamma distributions with shape  $\alpha$  and rate  $\beta$ .

We can visualise this hierarchical structure and the correlations between different hyper-parameters and parameters through a DAG, as in Fig. 4.1. The hyper-parameter  $\gamma$  sets the noise covariance deterministically (dotted line), but is itself statistically (solid line) defined by the hyper-prior distribution  $\pi(\gamma)$ . This is a gamma distribution, where  $\theta_\gamma$  determines the shape and rate of  $\pi(\gamma)$ . Similarly  $\theta_\delta$  defines  $\pi(\delta)$ , where  $\delta$  accounts for smoothness of the ozone profile and sets the prior precision  $Q(\delta)$ . Then  $\mathbf{Ax}$  determines the space of all measurable noise-free data sets  $\Omega$  through the linear forward model, from which we observe a data set  $\mathbf{y}$  including some noise  $\eta$ . Given that data, we “reverse the arrows” to determine the posterior distribution  $\pi(x, \theta | \mathbf{y})$  over the parameter  $x$  and the hyper-parameters  $\theta$ . Usually, due to underlying correlation structures, evaluating this posterior

poses a significant challenge. The MTC scheme breaks this correlation and provides the marginal posterior  $\pi(\delta, \gamma|\mathbf{y})$  first and then the full conditional posterior  $\pi(\mathbf{x}|\delta, \gamma, \mathbf{y})$ .

#### 4.1.1 Prior Modelling

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

By choosing a normally distributed prior  $\pi(\mathbf{x}|\delta)$  with zero mean and no other restrictions, it is clear that our model does not take into account that ozone values cannot be negative. As already mentioned, we set the precision matrix of that prior distribution to

$$\delta\mathbf{L} = \delta \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix} \quad (4.2)$$

which is a 1-dimensional Graph Laplacian as in [60, 17] with Dirichlet boundary condition. This matrix will also act as the regulariser later in Sec. ???. We reduce the dimension of  $\mathbf{x}$  from 45 to 34 by discarding every second ozone VMR over a height of  $\approx 47\text{km}$ . Doing that, while not changing  $\mathbf{L}$ , we effectively induce a larger correlation between points at higher altitude. We plot the corresponding prior ozone profiles according to  $\mathbf{x} \sim \mathcal{N}(0, (\delta\mathbf{L})^{-1})$  in Fig. B.1.

For  $\delta$  and  $\gamma$  we pick relatively uninformative gamma distributions so that  $\gamma \sim \mathcal{T}(\boldsymbol{\theta}_\gamma) \propto \gamma^{\alpha_\gamma-1} \exp(-\beta_\gamma \gamma)$  and  $\delta \sim \mathcal{T}(\boldsymbol{\theta}_\delta)$ , where  $\boldsymbol{\theta}_\gamma = \{\alpha_\gamma, \beta_\gamma\} = \{\alpha_\delta, \beta_\delta\} = \boldsymbol{\theta}_\delta = (1, 10^{-35})$  (see Fig. 5.3) similar to [17]. Those gamma distributions have another advantage when using the MWG algorithm to sample from the marginal posterior distribution  $\pi(\delta, \gamma|\mathbf{y})$ , where then  $\pi(\gamma|\lambda, \mathbf{y}) \sim \mathcal{T}(\cdot)$  is a gamma distribution with  $\lambda = \delta/\gamma$ , and easy to sample from.

## 4.2 Posterior Distribution

As explained in Sec. 2.1.1, we factorise the posterior

$$\pi(\mathbf{x}, \delta, \gamma|\mathbf{y}) \propto \pi(\mathbf{y}|\mathbf{x}, \delta, \gamma)\pi(\mathbf{x}, \delta, \gamma) \quad (4.3)$$

into

$$\pi(\mathbf{x}, \delta, \gamma|\mathbf{y}) = \pi(\mathbf{x}|\delta, \gamma, \mathbf{y})\pi(\delta, \gamma|\mathbf{y}) \quad (4.4)$$

the marginal posterior  $\pi(\delta, \gamma|\mathbf{y})$  and full conditional posterior  $\pi(\mathbf{x}|\delta, \gamma, \mathbf{y})$  (see Eq. 2.7). As discussed in Sec. 2.1.1, for the linear-Gaussian case,  $\mathbf{x}$  cancels in the marginal posterior over the hyper-parameters. Following the MTC scheme, we characterise the marginal posterior first and then the full conditional posterior.

### 4.2.1 Marginal Posterior

Consequently, for the hierarchical model specified in Eq. 4.1, the marginal posterior distribution over the hyper-parameters is given by

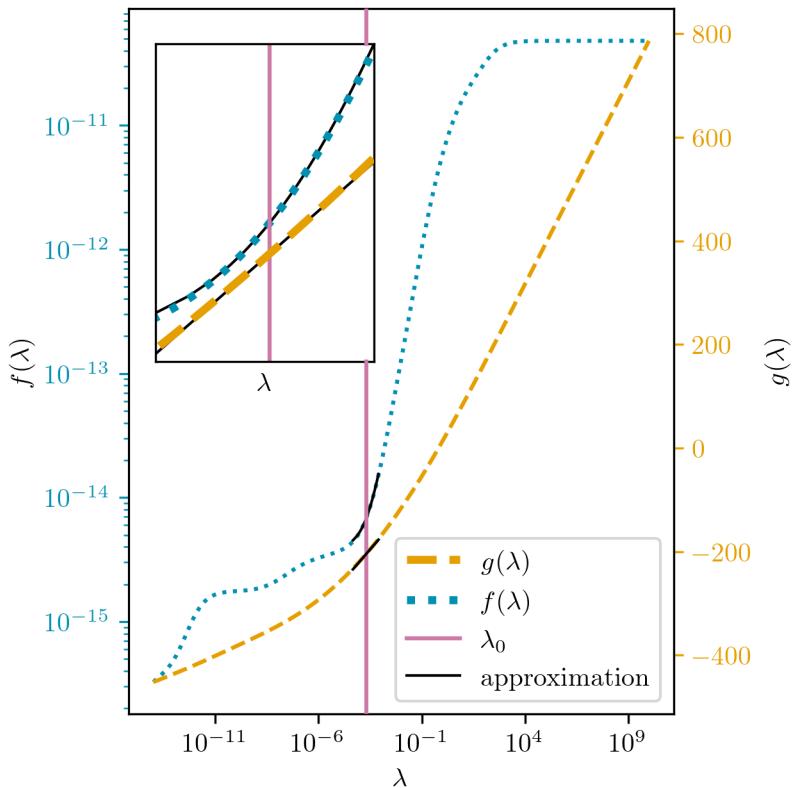
$$\pi(\lambda, \gamma | \mathbf{y}) \propto \lambda^{n/2 + \alpha_\delta - 1} \gamma^{m/2 + \alpha_\delta + \alpha_\gamma - 1} \exp \left\{ -\frac{1}{2} g(\lambda) - \frac{\gamma}{2} f(\lambda) - \beta_\delta \lambda \gamma - \beta_\gamma \gamma \right\}, \quad (4.5)$$

with the introduced regularisation parameter  $\lambda = \delta/\gamma$ , and

$$f(\lambda) = \mathbf{y}^T \mathbf{y} - (\mathbf{A}^T \mathbf{y})^T (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{L})^{-1} (\mathbf{A}^T \mathbf{y}), \quad (4.6a)$$

$$\text{and } g(\lambda) = \log \det(\mathbf{A}^T \mathbf{A} + \lambda \mathbf{L}). \quad (4.6b)$$

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  $d\delta/d\lambda = \gamma$ . For each evaluation of the marginal



**Figure 4.2:** Functions  $f(\lambda)$  and  $g(\lambda)$  from the marginal posterior in Eq. 4.5 for a wide range of  $\lambda = \delta/\gamma$ . We plot the approximations (see Eq. 4.9 and Eq. 4.10) in black around the mode of the marginal posterior (vertical line) for the sampling range of  $\lambda$  within the MWG algorithm.

posterior most of the computational effort lies in the calculation of  $f(\lambda)$  and  $g(\lambda)$ , which we evaluate using the Cholesky decomposition  $\mathbf{A}^T \mathbf{A} + \lambda \mathbf{L} = \mathbf{C}_\lambda \mathbf{C}_\lambda^T$ . More specifically we solve for  $(\mathbf{A}^T \mathbf{A} + \lambda \mathbf{L})^{-1} (\mathbf{A}^T \mathbf{y})$  via `scipy.linalg.cho_solve` and calculate the  $g(\lambda) = 2 \sum \log \text{diag}(\mathbf{C}_\lambda)$ . In Fig. 4.2 we see that  $f(\lambda)$  and  $g(\lambda)$  are well behaved within the

region of interest. Because of this, we approximate  $f(\lambda) \approx \tilde{f}(\lambda)$  with a Taylor series and  $\tilde{g}(\lambda) \approx g(\lambda)$  with a linear approximation in  $\lambda$  log-space around the mode  $\lambda_0$  of  $\pi(\lambda, \gamma | \mathbf{y})$ . The Taylor series coefficient of  $f(\lambda)$  is given by

$$f^{(r)}(\lambda_0) = (-1)^{r+1} (\mathbf{A}^T \mathbf{y})^T (\mathbf{B}_0^{-1} \mathbf{L})^r \mathbf{B}_0^{-1} \mathbf{A}_L^T \mathbf{y} \quad (4.7)$$

$$(4.8)$$

with  $\mathbf{B}_0 = \mathbf{A}^T \mathbf{A} + \lambda_0 \mathbf{L}$ . Note that usually a Taylor series includes a factor  $(r!)^{-1}$  [17], which in this case cancels in  $f^{(r)}(\lambda_0)$  so that  $f(\lambda)$  is approximated as

$$\tilde{f}(\lambda) = \sum_{r=0}^1 f^{(r)}(\lambda_0) (\lambda - \lambda_0)^r. \quad (4.9)$$

By exploratory analysis we find that the approximation

$$\tilde{g}(\lambda) = g(\lambda_0) + (\log \lambda - \log \lambda_0) \frac{g(1.25\lambda_0) - g(0.75\lambda_0)}{\log 1.25\lambda_0 - \log 0.75\lambda_0} \quad (4.10)$$

is sufficient. Note that  $g(\lambda)$  can be approximated with a Taylor series as well (see [17]). We plot the function  $f(\lambda)$  and  $g(\lambda)$  and their approximations in Fig. 4.2 and elaborate on the approximation errors in the section below.

### Error due to approximation of $f$ and $g$

To assess the approximation error, we lay a 100-point grid over the sampling region in each dimension and compare the approximations of  $f(\lambda)$ ,  $g(\lambda)$  and  $\pi(\lambda, \gamma | \mathbf{y})$  with their true function values.

Compared to a 2nd, 3-rd or 4-th order Taylor approximation, we find that the 1-st order Taylor approximation of  $f(\lambda)$  gives the smallest relative RMS error of  $\approx 9\%$  for  $\lambda = [10^{-5}, 8 \times 10^{-4}]$  (TT grid) and a maximum absolute error of  $\approx 3 \times 10^{-16}$ . Additionally, we find that a linear approximation of  $g(\lambda)$  is sufficient with relative RMS  $\approx 3\%$  and a maximum absolute error of  $\approx 5$ .

These errors then propagate into the marginal posterior  $\pi(\lambda, \gamma | \mathbf{y})$  so that the relative RMS error is  $\approx 6\%$  over the whole grid. When sampling, we evaluate the acceptance ratio in the log-space, so we report a relative RMS error of  $\approx 0.1\%$  for  $\log \pi(\lambda | \gamma, \mathbf{y})$ . We consider this good enough.

### Sample from marginal posterior – Metropolis within Gibbs

Using these approximations, we employ a Metropolis within Gibbs (MWG) sampler summarised in Alg. Box 3 to characterise  $\pi(\lambda, \gamma | \mathbf{y})$  as in [17]. More specifically, we implement a Metropolis random walk on

$$\pi(\lambda | \gamma, \mathbf{y}) \propto \lambda^{n/2 + \alpha_\delta - 1} \exp \left\{ -\frac{1}{2} g(\lambda) - \frac{\gamma}{2} f(\lambda) - \beta_\delta \gamma \lambda \right\} \quad (4.11)$$

and do a Gibbs step on

$$\gamma^{(k+1)} | \lambda^{(k+1)}, \mathbf{y} \sim \Gamma\left(\frac{m}{2} + \alpha_\delta + \alpha_\gamma, \frac{1}{2}f(\lambda^{(k+1)}) + \beta_\gamma + \beta_\delta \lambda^{(k+1)}\right). \quad (4.12)$$

Ergodicity for this approach is proven in [42].

**Algorithm 3:** Metropolis within Gibbs

```

1: Initialise  $(\lambda_1^{(0)}, \gamma_2^{(0)}) = (\lambda_0, \gamma_0)$ 
2: for  $k = 0, \dots, N - 1$  do
3:   Propose  $\lambda' \sim q(\cdot | \lambda^{(k)}) = q(\lambda^{(k)} | \cdot)$ 
4:   Compute

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

5:   Draw  $u \sim \mathcal{U}(0, 1)$ 
6:   if  $\alpha \geq u$  then
7:     Accept and set  $\lambda^{(k+1)} = \lambda'$ 
8:   else
9:     Reject and keep  $\lambda^{(k+1)} = \lambda^{(k)}$ 
10:  end if
11:  Draw  $\gamma^{(k+1)} \sim \pi(\cdot | \lambda^{(k+1)}, \mathbf{y})$ 
12: end for
13: Output:  $(\lambda, \gamma)^{(0)}, \dots, (\lambda, \gamma)^{(k)}, \dots, (\lambda, \gamma)^{(N)} \sim \pi(\boldsymbol{\theta} | \mathbf{y})$ 
```

The MWG algorithm starts at the initial guess  $(\lambda^{(0)}, \gamma^{(0)})$  at  $k = 0$ . We then propose a new sample  $\lambda' \sim q(\cdot | \lambda^{(k)})$ , conditioned on the previous state, using a symmetric proposal distribution  $q(\lambda' | \lambda^{(k)}) = q(\lambda^{(k)} | \lambda')$ , which is a Metropolis step and a special case of the Metropolis-Hastings algorithm [42]. We accept and set  $\lambda^{(k+1)} = \lambda'$  with the acceptance probability

$$\alpha(\lambda' | \lambda^{(k)}) = \min \left\{ 1, \frac{\pi(\lambda' | \gamma^{(k)}, \mathbf{y}) q(\lambda^{(k)} | \lambda')}{\pi(\lambda^{(k)} | \gamma^{(k)}, \mathbf{y}) q(\lambda' | \lambda^{(k)})} \right\}, \quad (4.13)$$

otherwise reject and keep  $\lambda^{(k+1)} = \lambda^{(k)}$ . In practice, we calculate the acceptance ratio in log space, so that

$$\log \left\{ \frac{\pi(\lambda' | \gamma^{(k)}, \mathbf{y})}{\pi(\lambda^{(k)} | \gamma^{(k)}, \mathbf{y})} \right\} = \log\{\pi(\lambda' | \gamma^{(k)}, \mathbf{y})\} - \log\{\pi(\lambda^{(k)} | \gamma^{(k)}, \mathbf{y})\} \quad (4.14)$$

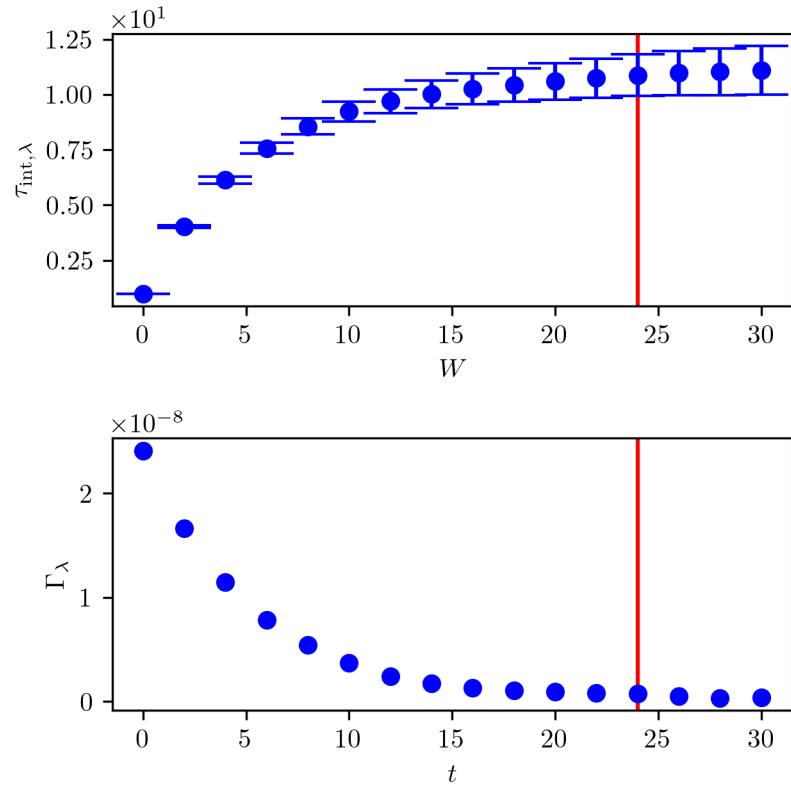
$$= \frac{n}{2}(\log\{\lambda'\} - \log\{\lambda^{(k)}\}) + \frac{1}{2}\Delta g + \frac{\gamma^{(k)}}{2}\Delta f + \beta_\delta \gamma^{(k)} \Delta \lambda, \quad (4.15)$$

where  $\Delta \lambda = \lambda' - \lambda^{(k)}$  and  $\Delta f \approx \tilde{f}(\lambda') - \tilde{f}(\lambda^{(k)}) = f^{(1)}(\lambda_0)\Delta \lambda - \Delta \lambda^{(k)}$ , with  $\Delta \lambda' = \lambda' - \lambda_0$  and  $\Delta \lambda^{(k)} = \lambda^{(k)} - \lambda_0$ . Similarly we approximate  $\Delta g \approx \tilde{g}(\lambda') - \tilde{g}(\lambda^{(k)})$ .

Next, we perform a Gibbs step on  $\pi(\gamma^{(k+1)} | \lambda^{(k+1)}, \mathbf{y})$  conditioned on the previously drawn  $\lambda^{(k+1)}$ . Gibbs sampling is again a special case of the Metropolis-Hastings algorithm

with acceptance probability equal to one. Repeating this  $N$  time give us marginal posterior samples  $(\lambda, \gamma)^{(1)}, \dots, (\lambda, \gamma)^{(N)} \sim \pi(\lambda, \gamma | \mathbf{y})$ . For we discard the first  $N_{\text{burn-in}}$  samples.

We approximate  $f(\lambda)$  and  $g(\lambda)$  around the mode  $(\lambda_0, \gamma_0)$  of  $\pi(\lambda, \gamma | \mathbf{y})$  provided by the `scipy.optimize.fmin` function, with a limit of 25 function evaluations. Then we approximate  $f(\lambda)$  with a 1-st order Taylor series and  $g(\lambda)$  with a linear approximation in the log-space, where for the approximation in  $g(\lambda)$  we set  $\lambda_{\max}$  to the maximum value of  $\lambda$  on the TT grid (see next section). We initialised the MWG algorithm at the mode  $(\lambda^{(0)}, \gamma^{(0)}) = (\lambda_0, \gamma_0)$  and take  $N = 10100$ , which includes burn in period of  $N_{\text{burn-in}} = 100$  steps. The standard deviation of the normal proposal distribution  $\lambda' \sim \mathcal{N}(\lambda^{(k)}, \sigma_\lambda^2)$  is empirically set to  $\sigma_\lambda = 0.8\lambda_0$ , so that the acceptance rate is  $\approx 0.5$  as suggested in [40]. This takes  $\approx 0.5$ s and we plot in Fig. 4.4 as well as the trace of the MWG to show ergodicity. We calculate the IACT with the Python implementation of [62] provided by [26], which gives us  $\tau_{\text{int},\gamma} \approx 4.4 \pm 0.2$  and  $\tau_{\text{int},\lambda} = 10.4 \pm 1.0$  (see Fig. 4.3 and Fig. B.2).



**Figure 4.3:** Provided by [26], the IACT  $\tau_{\text{int},\lambda}$  at summation windows  $W$  as well as the estimated autocorrelation function  $\Gamma_\lambda$  at lag  $t$  of the samples  $\lambda \sim \pi(\cdot | \mathbf{y})$ .

### TT approximation of marginal posterior

Alternatively, we can utilise a TT approximation of the square root of the marginal posterior over a predefined grid and calculate the marginals  $\pi(\gamma | \mathbf{y})$  and  $\pi(\lambda | \mathbf{y})$  (see Sec. 2.3.1).

We predefined the univariate grid with  $n = 20$  grid points (see Fig. 4.7) over  $\gamma = [0.8 \times 10^{15}, 1.2 \times 10^{16}]$  and  $\lambda = [10^{-5}, 8 \times 10^{-4}]$ . We introduce a ‘‘normalisation constant’’  $c = -300/2$  to avoid underflow so that the values of  $\sqrt{\pi(\lambda, \gamma|\mathbf{y})} = \exp\{0.5 \log \pi(\lambda, \gamma|\mathbf{y}) + c\}$  are within computer precision. Then we initialise the `tt.cross.rectcross.rect_cross.cross` function based on the TT cross algorithm in [37, 12] from the Python package `ttipy` [36] with a random tensor. The number of ranks is set to constant  $r = 4$ , we do 1 sweep with  $2n_{\text{sweeps}}2nr = 400$  function evaluations and obtain a TT approximation of  $\pi(\lambda, \gamma|\mathbf{y})$  in about 0.02s. Ironically, this the same number of functions evaluations to approximate a  $20 \times 20$  point grid. The TT format is especially advantageous for larger grid sizes and higher dimensional parameter spaces. To compute the marginals  $\pi(\lambda|\mathbf{y})$  and  $\pi(\gamma|\mathbf{y})$  we set the TT error  $\xi = 1/\lambda(\mathcal{X})$  and  $\lambda(x) = 1$ , so that for Cartesian basis  $\mathbf{M}_k = \text{diag}(\lambda_k(\mathcal{X}_k))$  (see Eq. 2.28). We calculate the coefficient tensor  $\mathbf{B}$  and  $\mathbf{R}_{\text{pre}}$  as in Prop. 1 and Prop. 2 (see Sec. 2.3.1).

We plot the TT approximation as a colour map on top of the obtained samples in Fig. 4.4. The relative RMS TT approximation error over the whole grid is  $\approx 7\%$  and similar to the propagation error in  $\pi(\lambda, \gamma|\mathbf{y})$  due to the approximations of  $f(\lambda)$  and  $g(\lambda)$  (see further up).

#### 4.2.2 Full Conditional Posterior – Ozone

Finally, we evaluate the normally distributed full conditional posterior distribution

$$\mathbf{x}|\delta, \gamma, \mathbf{y} \sim \mathcal{N}\left(\underbrace{(\mathbf{A}^T \mathbf{A} + \delta/\gamma \mathbf{L})^{-1} \mathbf{A}^T \mathbf{y}}_{\mathbf{x}_\lambda}, \underbrace{(\gamma \mathbf{A}^T \mathbf{A} + \delta \mathbf{L})^{-1}}_{\gamma \mathbf{B}_\lambda}\right), \quad (4.16)$$

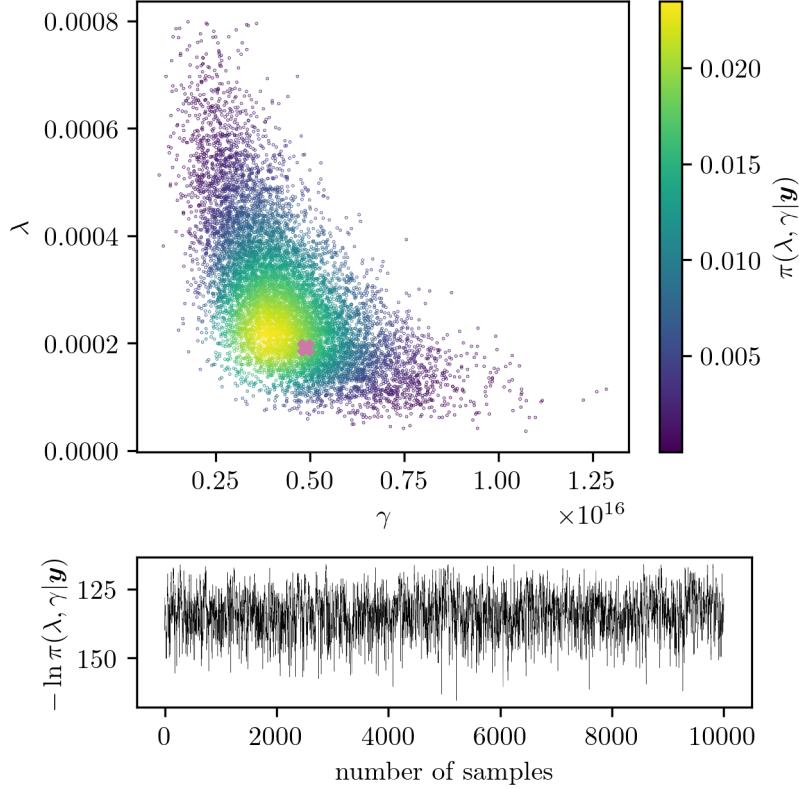
as in Eq. 2.15, with  $\lambda = \delta/\gamma$ . In this thesis, we compute the mean

$$\mu_{\mathbf{x}|\mathbf{y}} = \int \mathbf{x}_\lambda \pi(\lambda|\mathbf{y}) d\lambda \approx \sum \mathbf{x}_{\lambda_i} \pi(\lambda_i|\mathbf{y}), \quad (4.17)$$

and covariance

$$\Sigma_{\mathbf{x}|\mathbf{y}} = \int \gamma^{-1} \pi(\gamma|\mathbf{y}) d\gamma \int \mathbf{B}_\lambda^{-1} \pi(\lambda|\mathbf{y}) d\lambda \approx \sum \gamma_i^{-1} \pi(\gamma_i|\mathbf{y}) \sum \mathbf{B}_{\lambda_i}^{-1} \pi(\lambda_i|\mathbf{y}) \quad (4.18)$$

of  $\pi(\mathbf{x}|\mathbf{y})$  as weighted expectations over the marginal posterior  $\pi(\lambda, \gamma|\mathbf{y})$  by quadrature [10, Sec. 2.1] with  $\sum \pi(\lambda_i|\mathbf{y}) = \sum \pi(\gamma_i|\mathbf{y}) = 1$ . The weights  $\pi(\lambda_i|\mathbf{y})$  and  $\pi(\gamma_i|\mathbf{y})$  are either given by the TT approximation or by the bars of the sample-based histograms. More precisely, the heights of the sample-based histogram bars act as quadrature weights, where  $\lambda_i$  is defined at the centre of each bar. We use Cholesky decomposition of  $\mathbf{B}_\lambda = \mathbf{A}^T \mathbf{A} + \lambda \mathbf{L}$  to invert  $\mathbf{B}_\lambda$  and to calculate  $\mathbf{x}_\lambda = (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{L})^{-1} \mathbf{A}^T \mathbf{y}$  both via `scipy.linalg.cho_solve`. It is sufficient to evaluate  $\mathbf{x}_\lambda$  and invert  $\mathbf{B}_\lambda$  20 times to obtain mean and covariance values of  $\pi(\mathbf{x}|\mathbf{y})$  within a reasonable error (see Fig. 4.7). We



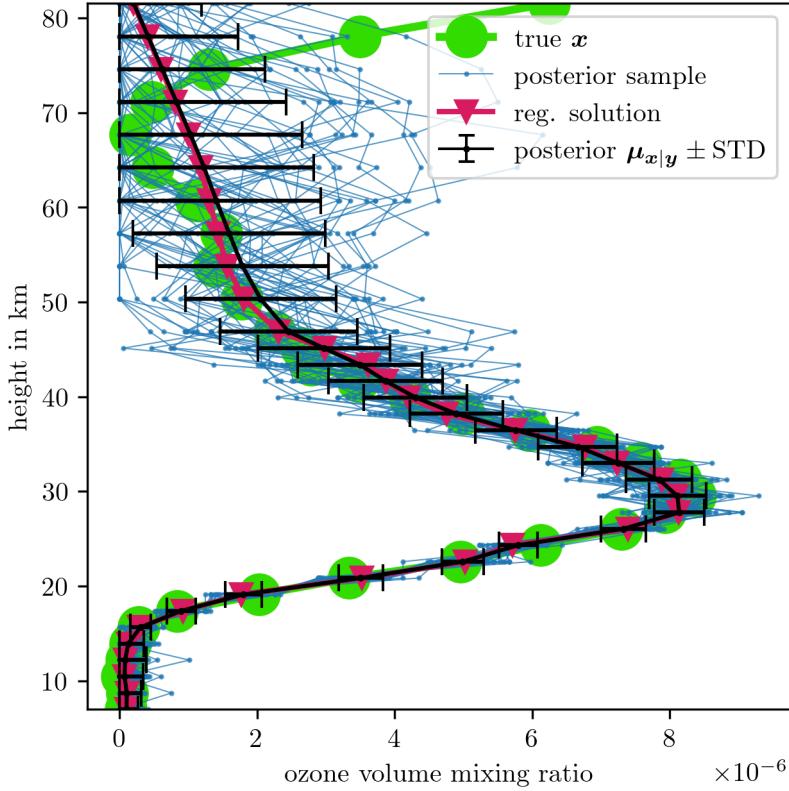
**Figure 4.4:** Samples from the marginal posterior colour-coded using the TT approximation of  $\pi(\lambda, \gamma|\mathbf{y})$ . The mode of  $(\lambda_0, \gamma_0)$  of  $\pi(\lambda, \gamma|\mathbf{y})$  is marked with the pink cross. To show ergodicity, we plot the trace of the samples of the MWG algorithm.

need roughly 0.025s to calculate the full posterior mean and variance including finding the mode of  $\pi(\lambda, \gamma|\mathbf{y})$ , running the TT cross and calculating the marginals. If we use the MWG sampler we need  $\approx 0.5$ s for the same results, so most computational effort lays within the sampling procedure and the time to calculate full posterior mean and variance is negligible. We plot posterior samples of  $\pi(\mathbf{x}|\mathbf{y})$  in Fig. 4.5 and set negative ozone values to zero, which we observe in almost every sample. The fact that we have to deal with negative ozone values is due to the poor prior choice in  $\pi(\mathbf{x}|\delta)$ . This indicates that we should use a different, more physically based prior or model a parametrised ozone profile. Note that the posterior samples do not capture the second ozone peak at around 80km.

If calculating the variance is too costly, the RTO method (see Sec. 6.2.2) may be a feasible alternative to draw a sample from Eq. 4.16.

### Eigenvalues full conditional posterior covariance

We can visualise how the posterior samples are affected through the prior  $\delta \mathbf{L}$  and the forward model  $\gamma \mathbf{A}^T \mathbf{A}$  by comparing their eigenvalues to the eigenvalues of the precision matrix  $\mathbf{Q}_{\mathbf{x}|\delta, \gamma, \mathbf{y}} = \gamma \mathbf{A}^T \mathbf{A} + \delta \mathbf{L}$  for a random  $\delta, \gamma \sim \pi(\delta, \gamma|\mathbf{y})$ . We order the eigenvalues in

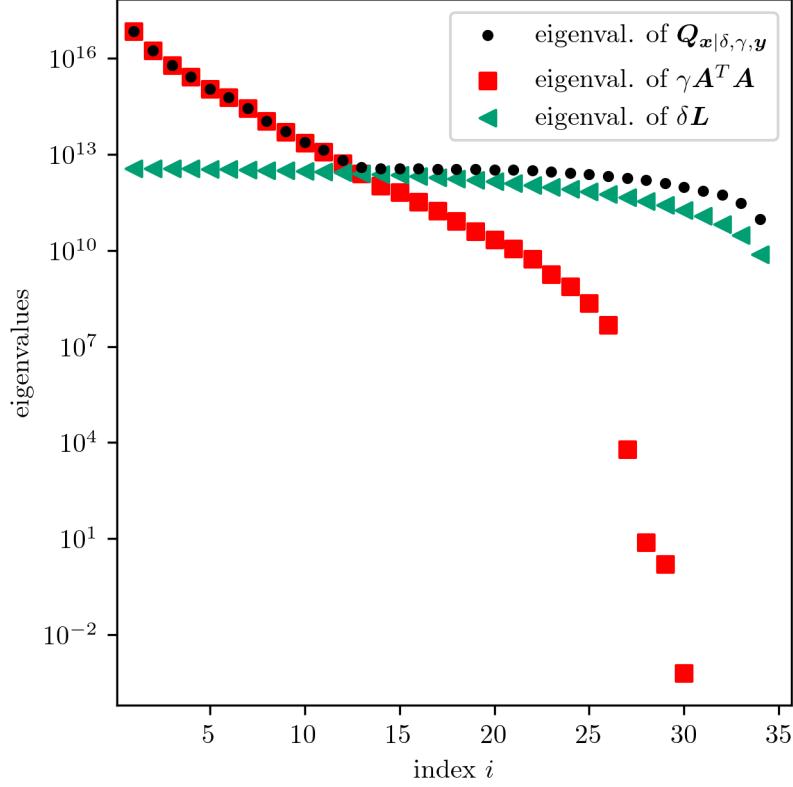


**Figure 4.5:** Ozone samples from the full posterior distribution  $\pi(\mathbf{x}|\mathbf{y})$  after characterising full posterior mean and covariance by weighted expectations over the marginal posterior  $\pi(\lambda, \gamma|\mathbf{y})$  based on the linear forward map  $\mathbf{A}_L$ . We set negative ozone VMR values to zero.

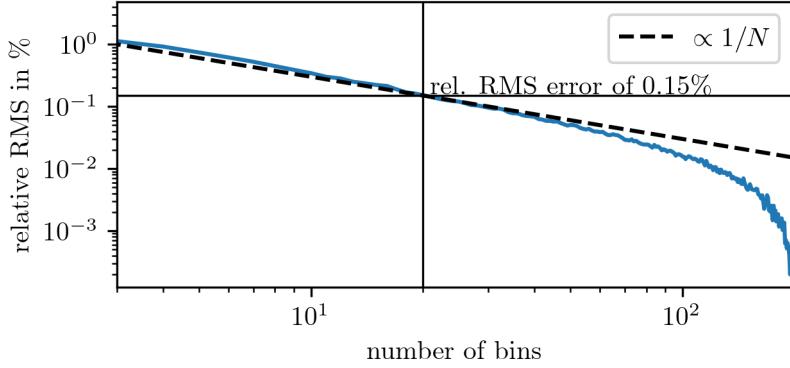
size and plot those in Fig. 4.6. We observe that the larger eigenvalues of  $\mathbf{Q}_{\mathbf{x}|\delta,\gamma,\mathbf{y}}$  are very much the same as the larger eigenvalues of  $\gamma \mathbf{A}^T \mathbf{A}$ . Once the eigenvalues of  $\gamma \mathbf{A}^T \mathbf{A}$  are significantly smaller than the eigenvalues of  $\mathbf{Q}_{\mathbf{x}|\delta,\gamma,\mathbf{y}}$  the structure of the eigenvalues is dominated by the eigenvalues of  $\delta \mathbf{L}$ . The largest 10 eigenvalues of  $\mathbf{Q}_{\mathbf{x}|\delta,\gamma,\mathbf{y}}$  include ozone profile structure at lower altitudes, where the other eigenvector mainly represent structures at higher altitudes (see Fig. B.5 and Fig. B.5). Note that the eigenvalues of each matrix may correspond to different eigenvectors even if the eigenvalues of two matrices are the same.

### Errors of Full Posterior Mean and Covariance

In Fig. 4.7, we plot the relative RMS error for the mean  $\mu_{\mathbf{x}|\mathbf{y}}$  and covariance  $\Sigma_{\mathbf{x}|\mathbf{y}}$  of  $\pi(\mathbf{x}|\mathbf{y})$ . We obtain those results by calculating the weighted expectation over normalised histograms of  $\pi(\lambda, \gamma|\mathbf{y})$ , where we vary the number of bins and compare to a solution calculated from a histogram with 200 bins. The relative error behaves roughly proportional to  $1/N$ , and we consider a relative error less than 0.5% good enough, which we easily meet at 20 bins. This sets the TT grid size and the number of evaluations of  $\mathbf{x}_\lambda$  in Eq. 4.17 and  $(\gamma \mathbf{B}_\lambda)^{-1}$  in Eq. 4.18.



**Figure 4.6:** Eigenvalues of the precision matrix of  $\mathbf{Q}_{\mathbf{x}|\delta,\gamma,\mathbf{y}} = \gamma \mathbf{A}^T \mathbf{A} + \delta \mathbf{L}$  of the full posterior distribution  $\pi(\mathbf{x}|\delta, \gamma, \mathbf{y})$  for ozone. We see that large eigenvalues of  $\gamma \mathbf{A}^T \mathbf{A}$  and  $\delta \mathbf{L}$  are rather unaffected by the prior compared to small eigenvalues. The eigenbasis may differ.



**Figure 4.7:** Relative RMS error of  $\mu_{\mathbf{x}|\mathbf{y}}$  and covariance  $\Sigma_{\mathbf{x}|\mathbf{y}}$  calculated by the weighted expectations and compared to a “ground truth” given by weighted expectations over 200 bins.

### 4.3 Solution by Regularisation

Since we claim that the Bayesian approach is superior to regularisation methods, we compare the MTC method to a regularisation approach, which is most similar to our chosen linear-Gaussian Bayesian framework [17].

The regularised solution is defined as in [23, 17]

$$\mathbf{x}_\lambda = \arg \min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_{L^2}^2 + \lambda \mathbf{x}^T \mathbf{L} \mathbf{x}, \quad (4.19)$$

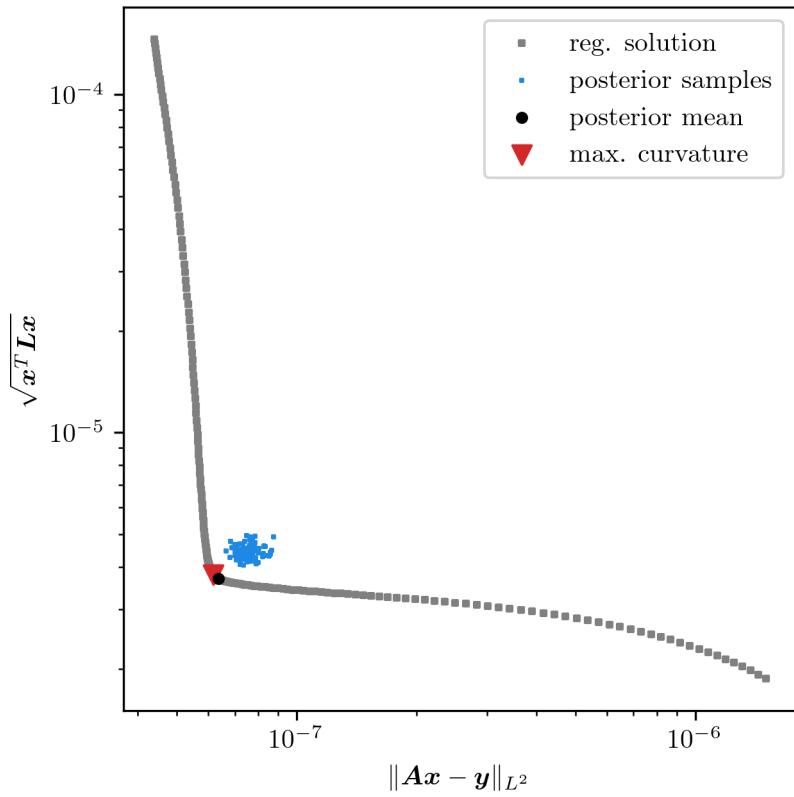
with the regularisation parameter  $\lambda$ , linear forward model matrix  $\mathbf{A}$  and data  $\mathbf{y}$ . A regularised solution

$$\mathbf{x}_\lambda = (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{L})^{-1} \mathbf{A}^T \mathbf{y} \quad (4.20)$$

is calculated as in Sec. 2.4.

To find the regularised solution, we use the L-curve method, and follow [25]. Within this method we compute  $\mathbf{x}_\lambda$ , for 200 different  $\lambda$  values in between  $10^0$  to  $10^6$  and plot the regularisation norm  $\sqrt{\mathbf{x}_\lambda^T \mathbf{L} \mathbf{x}_\lambda}$  against the data misfit norm  $\|\mathbf{A}\mathbf{x}_\lambda - \mathbf{y}\|_{L^2}$  (see Figure 4.8). The regularised solution corresponds to the “corner” of the L-curve at the point of maximum curvature provided by the kneedle algorithm [48] using the function `kneed.KneeLocator` in  $\approx 0.015$ s, which is slightly faster than the TT approach to obtain full posterior mean and covariance. We plot the corresponding regularisation parameter in Fig. 5.3, which can vary significantly compared to the  $\lambda$ -samples of the marginal posterior.

The regularised solution in Fig. 5.4 is very similar to the posterior mean. It is pretty clear that the regularised solution accounts for only one possible solution and does not provide uncertainties. The regularised solution is not similar to the samples drawn from the posterior  $\pi(\mathbf{x}|\mathbf{y})$  (see Fig. 4.5). The samples of  $\pi(\mathbf{x}|\mathbf{y})$  plotted in Fig. 4.8 lie above the L-Curve, whereas the posterior mean and the regularised solution are on the L-Curve. This does make sense, if one thinks about the mean as the (smooth) average over less-smooth samples and the regularised solution as an extremely smooth ozone profile (see Lagrangian in Sec. 2.4). In contrast, the samples are less regularised and hence lie above the L-Curve, but have a similar data misfit norm, and as already mentioned, are all feasible solutions to the data. Neither the regularisation solution nor the posterior ozone profiles capture the second ozone peak of the ground truth at high altitudes.



**Figure 4.8:** L-Curve of regularised semi norm  $\sqrt{\mathbf{x}^T \mathbf{L} \mathbf{x}}$  against the data misfit norm  $\|\mathbf{A}\mathbf{x}_\lambda - \mathbf{x}\|_{L^2}$  for different  $\lambda$  values, where  $\mathbf{x}_\lambda$  is calculated as in Eq. 4.20. The best regularised solution is at the point of maximum curvature (pink triangle). Additionally, we calculate the data misfit norm and the regularised norm for the mean (black circle) and samples (blue squares) of the full posterior of ozone.

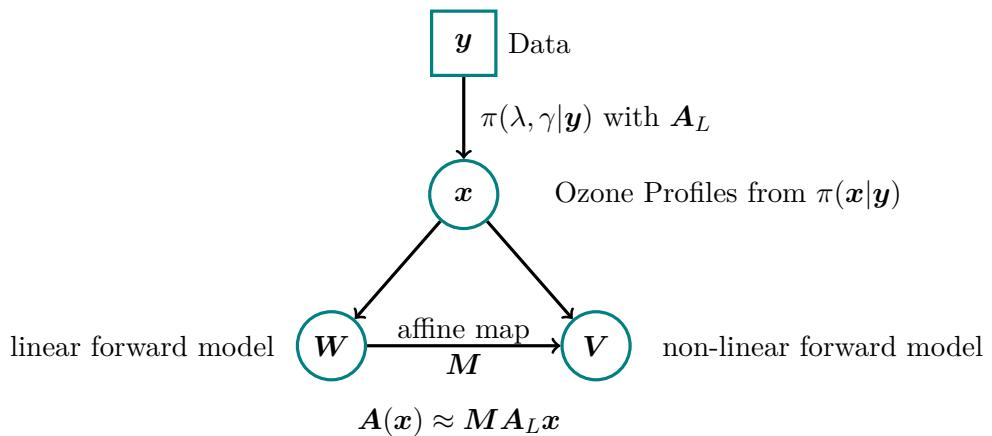


# 5

## Affine Approximation of the Non-Linear Model

The forward map, introduced in Chapter 3, poses a weakly non-linear forward problem. One could tackle the non-linearity by treating the inverse problem as a linear inverse problem and then iteratively updating the non-linear part after each parameter sample. Instead, as in Fig. 5.1 illustrated, we approximate the non-linear model using an affine map, which is a linear map with a translation, e.g.  $\mathbf{A}\mathbf{x} + \mathbf{b}$ , and maps Gaussians onto Gaussians. For parameters  $\mathbf{x}$  around the posterior mean  $\mu_{\mathbf{x}|\mathbf{y}}$  the affine map  $\mathbf{M}$  provides an approximation to the non-linear model  $\mathbf{A}(\mathbf{x})$  based on the linear model  $\mathbf{A}_L$ .

### 5.1 Finding an Affine Approximation



**Figure 5.1:** The strategy to find the affine map consists of first evaluating the marginal posterior for ozone  $\pi(\lambda, \gamma|\mathbf{y})$  based on the linear forward model. Based on ozone samples from the full posterior, we find an affine map  $\mathbf{M}$  which approximates between noise-free linear data  $\mathbf{A}_L$  and noise-free non-linear data  $\mathbf{A}(\mathbf{x})$ .

We find an affine map by creating the vector spaces  $\mathbf{W}$  based on the linear forward model and  $\mathbf{V}$  based on the non-linear forward model with ground truth pressure and temperature. More specifically  $m - 1$  samples  $\mathbf{x}^{(j)} \sim \pi(\mathbf{x}|\mathbf{y})$  for  $j = 2, \dots, m$  from the posterior and the posterior mean  $\mu_{\mathbf{x}|\mathbf{y}}$  generate,

$$\mathbf{W} = \begin{bmatrix} | & | & | & | \\ \mathbf{A}_L \mu_{\mathbf{x}|\mathbf{y}} & \mathbf{A}_L \mathbf{x}^{(2)} & \cdots & \mathbf{A}_L \mathbf{x}^{(j)} & \cdots & \mathbf{A}_L \mathbf{x}^{(m)} \\ | & | & | & | & & | \end{bmatrix} \in \mathbb{R}^{m \times m}$$

and

$$\mathbf{V} = \begin{bmatrix} | & | & | & | \\ \mathbf{A}(\mu_{\mathbf{x}|\mathbf{y}}) & \mathbf{A}(\mathbf{x}^{(2)}) & \cdots & \mathbf{A}(\mathbf{x}^{(j)}) & \cdots & \mathbf{A}(\mathbf{x}^{(m)}) \\ | & | & | & | & & | \end{bmatrix} = \begin{bmatrix} \_ & v_1 & \_ \\ \vdots & \vdots & \vdots \\ \_ & v_j & \_ \\ \vdots & \vdots & \vdots \\ \_ & v_m & \_ \end{bmatrix} \in \mathbb{R}^{m \times m}.$$

Then the non-linear forward model is approximated as

$$\mathbf{A}(\mathbf{x}) \approx \mathbf{M} \mathbf{A}_L \mathbf{x}, \quad (5.1)$$

where we solve  $v_j = r_j \mathbf{W}$  for each row  $r_j$  in

$$\mathbf{V} \mathbf{W}^{-1} = \mathbf{M} = \begin{bmatrix} \_ & r_1 & \_ \\ \vdots & \vdots & \vdots \\ \_ & r_j & \_ \\ \vdots & \vdots & \vdots \\ \_ & r_m & \_ \end{bmatrix} \in \mathbb{R}^{m \times m}.$$

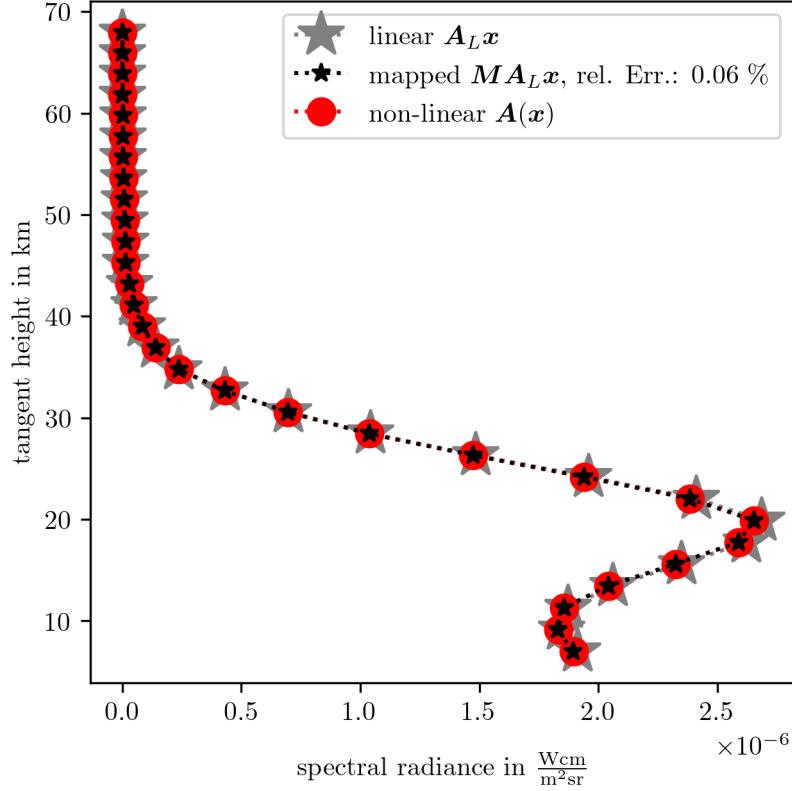
using the Python function `numpy.linalg.solve`. This is feasible since every noise-free measurement is independent of each other, and then every row  $v_j$  of  $\mathbf{V} \in \mathbb{R}^{m \times m}$  is independent of each other as well. For an  $\mathbf{x} = \mu_{\mathbf{x}|\mathbf{y}} + \Delta \mathbf{x}$  we rewrite Eq. 5.1 to

$$\mathbf{A}(\mathbf{x}) \approx \underbrace{\mathbf{M} \mathbf{A}_L \mu_{\mathbf{x}|\mathbf{y}}}_{= \mathbf{A}(\mu_{\mathbf{x}|\mathbf{y}})} + \underbrace{\mathbf{M} \mathbf{A}_L \Delta \mathbf{x}}_{= \mathbf{A}'(\mu_{\mathbf{x}|\mathbf{y}}) \Delta \mathbf{x}} \quad (5.2)$$

$$= \underbrace{\mathbf{A}'(\mu_{\mathbf{x}|\mathbf{y}}) \mathbf{x}}_{\mathbf{Ax}} + \underbrace{\mathbf{A}(\mu_{\mathbf{x}|\mathbf{y}}) - \mathbf{A}'(\mu_{\mathbf{x}|\mathbf{y}}) \mu_{\mathbf{x}|\mathbf{y}}}_{\mathbf{b}} \quad (5.3)$$

to show that  $\mathbf{M} : \mathbf{A}_L \mathbf{x} \rightarrow \mathbf{A}(\mathbf{x})$  is an affine map.

The relative RMS difference  $\|\text{vec}(\mathbf{MW}) - \text{vec}(\mathbf{V})\|_{L^2} / \|\text{vec}(\mathbf{MW})\|_{L^2}$  between the mapped linear noise-free data and the non-linear noise-free data is approximately 0.001%. This is much smaller than the relative RMS difference between  $\mathbf{W}$  and  $\mathbf{V}$  of about 1%. Fig. 5.2 shows the mapping for one posterior ozone sample with a relative RMS error  $\approx 0.06\%$ . This posterior ozone sample has not been used to create this mapping; in other words, this is an unseen event not occurring in the training data. Consequently, from here onwards, we use the approximated forward map.

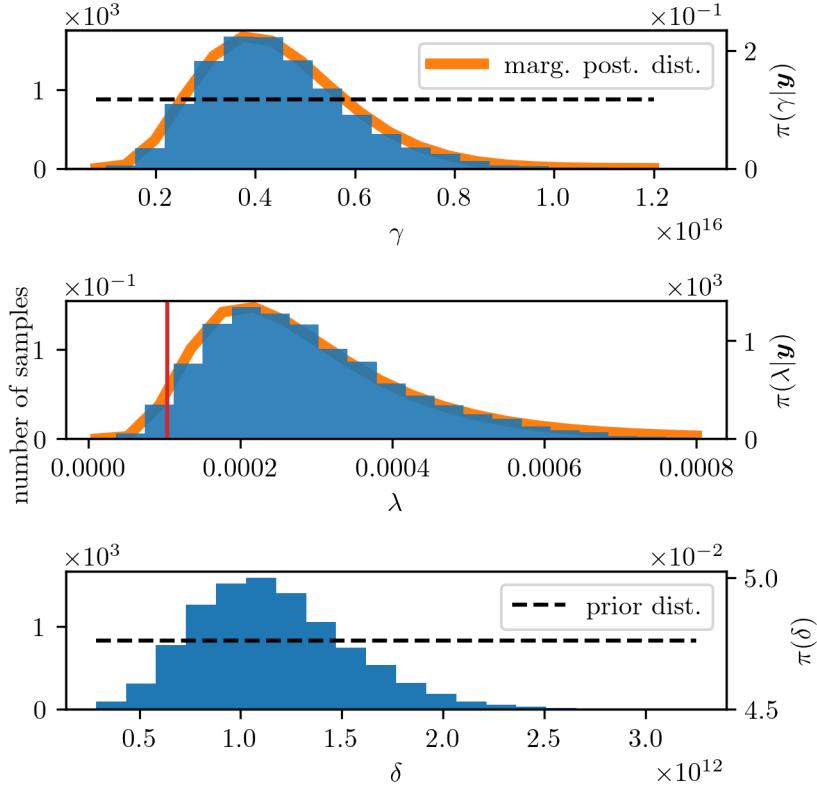


**Figure 5.2:** Assessment of how well we can approximate noise-free non-linear data  $\mathbf{A}(\mathbf{x})$  (red circles) with noise-free linear data  $\mathbf{A}_L\mathbf{x}$  (grey stars) and the previously calculated affine map  $\mathbf{M}$ . The approximated noise-free data (black stars) has a relative RMS error of  $\approx 0.06\%$  compared to the true non-linear noise-free data. The ozone sample to generate this noise-free data has not been used to create the affine map.

## 5.2 Marginal and Then Conditional Posterior – Ozone

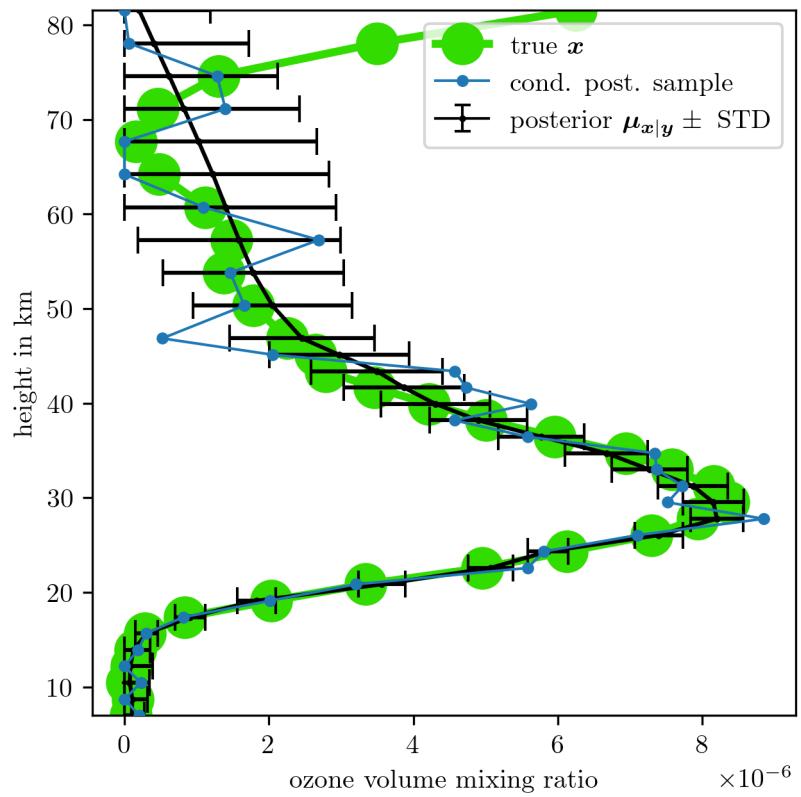
Again, we use the MTC scheme and the exact same setup and procedure as in Sec. 4.2 to evaluate the marginal posterior and then the full conditional posterior of ozone with similar computational time.

The marginal posterior is defined as in Eq. 4.5, but with the approximated forward model. The MWG is initialised at the mode of  $\pi(\lambda, \gamma|\mathbf{y})$  and  $f(\lambda)$  and  $g(\lambda)$  are approximated around the mode as in Sec. 4.2.1 (see Eq. 4.9 and Eq. 4.10). We take  $N = 10000$  plus  $N_{\text{burn-in}} = 100$  steps in  $\approx 0.5\text{s}$ . The IACTs provided by [26] are  $\tau_{\text{int},\gamma} \approx 5.2 \pm 0.3$  and  $\tau_{\text{int},\lambda} = 11 \pm 1$  (see Fig. B.4 and Fig. B.3) and similar to the previously calculated values. We plot the samples in Fig. 5.3 as well as the TT approximation of the marginal posterior using 400 function evaluations (same grid; same number of ranks; see Sec. 4.2.1). The relative RMS error of the TT approximation over the whole grid and the relative RMS approximation error of  $\pi(\lambda, \gamma|\mathbf{y})$  due to the approximations of  $f(\lambda)$  and  $g(\lambda)$  are both  $\approx 8\%$ .



**Figure 5.3:** The TT approximation of the marginal posterior in orange and the samples as a histogram, as well as the prior distribution with a dotted line. We sample  $\lambda$  and  $\gamma$  using the MWG algorithm and then calculate  $\delta$  for every sample of the marginal posterior. The regularised parameter corresponding to the best regularised solution (see Fig. 5.4 and Fig. 4.8) is marked with the red vertical line. We mark the ground truth noise precision with the back vertical line.

Again, we calculate the full posterior mean  $\mu_{x|y}$ , see Eq. 4.17, and covariance matrix  $\Sigma_{x|y}$  4.18 as weighted expectation. We plot the results and one sample of  $\pi(x|y)$ , which represents a feasible solution to this inverse problem, in Fig. 5.4, as well as the regularised solution (see next section), and one sample from the posterior. We can see that the ground truth lies within the STD around the mean, except for the peak at around 80km. Compared to the previously calculated mean and variance based on the linear forward model  $A_L$  (see 4.5), the posterior distribution based on  $MA_L$  does not differ significantly. This is expected since the difference between the linear and non-linear forward map of  $\approx 1\%$  is small.



**Figure 5.4:** Full posterior mean and variance and one ozone sample from the full posterior. We plot the regularised solution on top of the ground truth ozone profile in green. The results are based on the approximated forward model  $MA_L$ .



# 6

## Joint Retrieval of Ozone, Pressure and Temperature

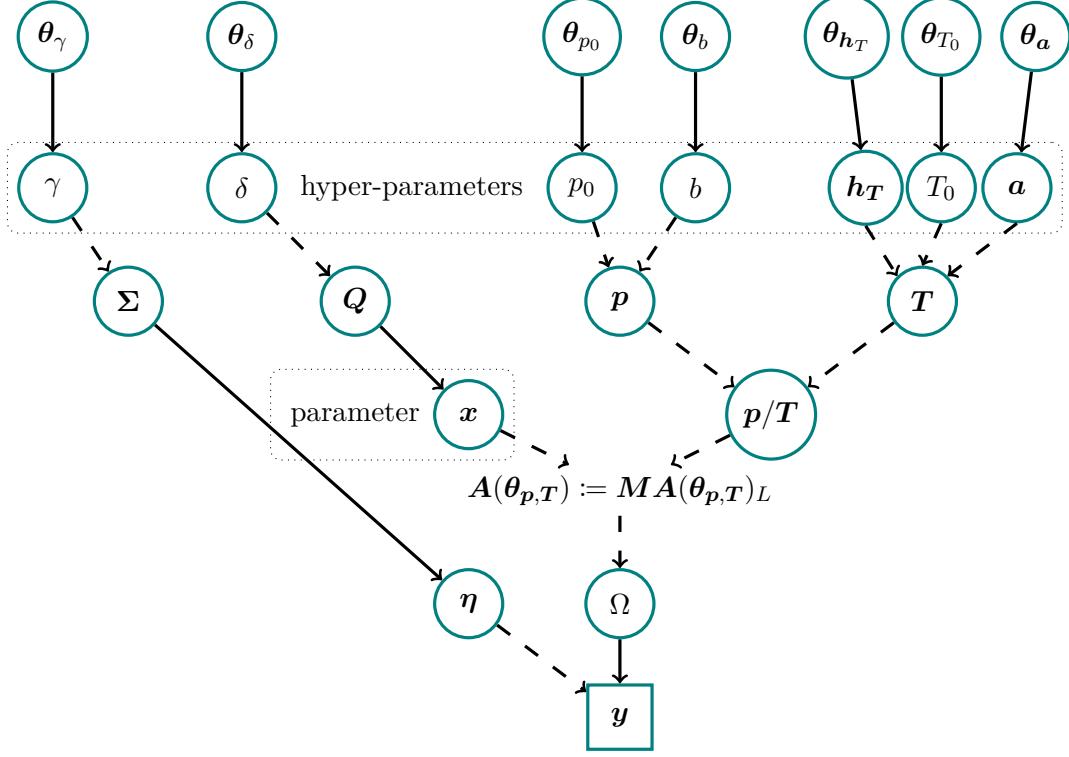
Here, the hierarchical Bayesian model in Sec. 4.1 and the MTC scheme is extended to jointly provide posterior distributions of ozone ,pressure and temperature. We elaborate on some aspects of prior modelling. Additionally we guide the reader through the process of approximating the higher-dimensional marginal posterior more efficiently using a TT.

We use the affine approximation  $\mathbf{M}$  from the previous Chapter, and define a linear forward model matrix as

$$\mathbf{A} := \mathbf{M}\mathbf{A}_L. \quad (6.1)$$

### 6.1 Hierarchical Bayesian Model – Ozone, Pressure, Temperature

As in Sec. 4.1, we use a DAG as in Fig. 6.1 to visualise the measurement process and correlations between pressure  $\mathbf{p}$ , temperature  $\mathbf{T}$  and ozone  $\mathbf{x}$ , which progress deterministically (dashed line) into the forward model, via  $\mathbf{x} \times \mathbf{p}/\mathbf{T}$ . Note that other variables such as absorption constant, internal partition function and the black body radiation are also dependent on temperature. Through their respective prior distributions, they generate a space of all possible noise-free data  $\Omega$ , from which we observe some data, including some added normally distributed noise  $\boldsymbol{\eta}$ . This hierarchical Bayesian framework includes the hyper-parameters  $p_0, b$  for pressure (see Eq. 6.3),  $\mathbf{a}, \mathbf{h}_T, T_0$  for temperature (see Eq. 3.9),  $\delta$  for ozone smoothness and  $\gamma$  for noise precision. Each of those hyper-parameters is described by the hyper-prior distribution  $\pi(p_0, b, T_0, \mathbf{h}_T, \mathbf{a}, \delta, \gamma)$  defined by us. Here  $\boldsymbol{\theta}_\gamma, \boldsymbol{\theta}_\delta$  determine gamma distributions, e.g.  $\gamma \sim \Gamma(\alpha_\gamma, \beta_\gamma)$  with  $\boldsymbol{\theta}_\gamma = \{\alpha_\gamma, \beta_\gamma\}$ , and  $\boldsymbol{\theta}_{p_0}, \boldsymbol{\theta}_b, \boldsymbol{\theta}_h, \boldsymbol{\theta}_{T_0}, \boldsymbol{\theta}_a$  determine a normal distribution  $p_0, b, \mathbf{h}_T, T_0, \mathbf{a} \sim \pi(\boldsymbol{\theta}_{p_0}, \boldsymbol{\theta}_b, \boldsymbol{\theta}_h, \boldsymbol{\theta}_{T_0}, \boldsymbol{\theta}_a)$ ,



**Figure 6.1:** DAG of Bayesian model for ozone, pressure and temperature. The hyper-parameters  $\mathbf{h}_T = \{h_{T,1}, h_{T,2}, h_{T,3}, h_{T,4}, h_{T,5}, h_{T,6}\}$ ,  $\mathbf{a} = \{a_0, a_1, a_2, a_3, a_4, a_5, a_6\}$ ,  $T_0$ ,  $b$  and  $p_0$  deterministically (dotted line) describe pressure parameter  $\mathbf{p}$  through the function in Eq. 6.3, and temperature parameter  $\mathbf{T}$  through the function in Eq. 3.9. In this case, we choose the hyper-priors  $\pi(p_0, b, T_0, \mathbf{h}_T, \mathbf{a})$  to be a normally distributed apriori, determined by  $\theta_{h_T}, \theta_a, \theta_{T_0}, \theta_b, \theta_{p_0}$ , which represent mean and variances. The ozone parameter  $\mathbf{x}$  is statistically (solid line) described by the prior distribution  $\mathbf{x}|\delta \sim \mathcal{N}(0, (\delta \mathbf{L})^{-1})$ . Here, the hyper-parameter  $\delta$  accounts for smoothness in the ozone profile and defines the precision matrix  $\mathbf{Q} = \delta \mathbf{L}$ , where  $\mathbf{L}$  is the graph Laplacian as in Eq. 4.2. The noise covariance  $\Sigma = \gamma^{-1} \mathbf{I}$  of the random noise vector  $\boldsymbol{\eta} \sim \mathcal{N}(0, \gamma^{-1} \mathbf{I})$  is defined by the noise hyper-parameter  $\gamma$ . As in Sec. 4.1 described,  $\theta_\delta$  and  $\theta_\gamma$  define the hyper-priors  $\pi(\delta, \gamma)$ . Then, we randomly observe a data set  $\mathbf{y}$  from the space of all measurables  $\Omega$  through the approximated forward model  $\mathbf{A}(\theta_{p,T}) := \mathbf{M}\mathbf{A}(\theta_{p,T})_L$ , depending on the hyper-parameter  $\theta_{p,T} := \{p_0, b, \mathbf{h}_T, T_0, \mathbf{a}\}$ , including some added noise. Given the data we like to determine the posterior distribution over the hyper-parameters  $\pi(p_0, b, T_0, \mathbf{h}_T, \mathbf{a}, \delta, \gamma | \mathbf{y})$  first and then  $\pi(\mathbf{x}|p_0, b, T_0, \mathbf{h}_T, \mathbf{a}, \delta, \gamma, \mathbf{y})$ , utilising the MTC scheme.

e.g.  $b \sim \mathcal{N}(\mu_b, \sigma_b^2)$  and  $\theta_b = \{\mu_b, \sigma_b\}$ . We use the approximated forward model

$\mathbf{A}(\theta_{p,T}) := \mathbf{M}\mathbf{A}(\theta_{p,T})_L$  and denote  $\theta_{p,T} := \{p_0, b, T_0, \mathbf{h}_T, \mathbf{a}\}$ , which includes all hyper-parameter related to pressure and temperature.

model parameters	priors	TT bounds		t-walk	
		lower	upper	$\tau_{\text{int}}$	Context
$\mathbf{x}$	$\mathcal{N}(0, (\delta \mathbf{L})^{-1})$	-	-	-	$\mathbf{x}$
$\delta$	$\mathcal{T}(1, 10^{-35})$	-	-		$\mathbf{x}$
$\gamma$	$\mathcal{T}(1, 10^{-35})$	$8 \times 10^{14}$	$1.2 \times 10^{16}$	$507 \pm 29$	$\mathbf{y}$
$\lambda = \delta/\gamma$	-	$1 \times 10^{-5}$	$2.5 \times 10^{-3}$	$979 \pm 75$	$\mathbf{x}$
$b$	$\mathcal{N}(0.174, (0.01)^2)$	0.129	0.214	$830 \pm 60$	$\mathbf{p}$
$h_{T,1}$	$\mathcal{N}(11, (1.5)^2)$	5.4	16.3	$286 \pm 13$	$\mathbf{T}$
$T_0$	$\mathcal{N}(288.15, (10)^2)$	247	326	$279 \pm 12$	$\mathbf{T}$
$p_0$	$\mathcal{N}(1311, (20)^2)$	1237	1387	$279 \pm 12$	$\mathbf{p}$
$h_{T,3}$	$\mathcal{N}(32.3, (2.5)^2)$	22.9	41.7	$254 \pm 11$	$\mathbf{T}$
$a_1$	$\mathcal{N}(0, (0.1)^2)$	-0.38	0.38	$295 \pm 13$	$\mathbf{T}$
$h_{T,2}$	$\mathcal{N}(20.1, (0.7)^2)$	17.2	22.7	$296 \pm 13$	$\mathbf{T}$
$a_0$	$\mathcal{N}(-6.5, (0.01)^2)$	-6.54	-6.47	$252 \pm 10$	$\mathbf{T}$
$a_2$	$\mathcal{N}(1, (0.01)^2)$	0.97	1.03	$267 \pm 11$	$\mathbf{T}$
$a_3$	$\mathcal{N}(2.8, (0.1)^2)$	2.5	3.1	$267 \pm 11$	$\mathbf{T}$
$h_{T,4}$	$\mathcal{N}(47.4, (0.5)^2)$	45.5	49.3	$270 \pm 12$	$\mathbf{T}$
$a_4$	$\mathcal{N}(0, (0.1)^2)$	-0.38	0.38	$254 \pm 11$	$\mathbf{T}$
$h_{T,5}$	$\mathcal{N}(51.4, (0.5)^2)$	49.5	53.3	$280 \pm 12$	$\mathbf{T}$
$a_5$	$\mathcal{N}(-2.8, (0.1)^2)$	-3.18	-2.43	$278 \pm 12$	$\mathbf{T}$
$h_{T,6}$	$\mathcal{N}(71.8, (3)^2)$	60.5	83.1	$250 \pm 10$	$\mathbf{T}$
$a_6$	$\mathcal{N}(-2, (0.01)^2)$	-2.04	-1.96	$272 \pm 12$	$\mathbf{T}$

**Table 6.1:** Summary of relevant parameter characteristics, bounds and sampling statistics. We denote  $\mathcal{N}(\mu, \sigma^2)$  as the Gaussian and  $\mathcal{T}(\alpha = \text{scale}, \beta = \text{rate})$  as the gamma distribution. The IACT  $\tau_{\text{int}}$  is estimated as in [63] from posterior samples based on the approximated forward map.

Then, we set up the hierarchical Bayesian framework

$$\mathbf{y} | \mathbf{x}, p_0, b, T_0, \mathbf{h}_T, \mathbf{a}, \delta, \gamma \sim \mathcal{N}(\mathbf{A}(p_0, b, T_0, \mathbf{h}_T, \mathbf{a}) \mathbf{x}, \gamma^{-1} \mathbf{I}) \quad (6.2a)$$

$$\mathbf{x} | \delta \sim \mathcal{N}(\mathbf{0}, (\delta \mathbf{L})^{-1}) \quad (6.2b)$$

$$\delta \sim \Gamma(\alpha_\delta, \beta_\delta) \quad (6.2c)$$

$$\gamma \sim \Gamma(\alpha_\gamma, \beta_\gamma) \quad (6.2d)$$

$$\mathbf{a} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{a}}, \boldsymbol{\Sigma}_{\mathbf{a}}) \quad (6.2e)$$

$$\mathbf{h}_T \sim \mathcal{N}(\boldsymbol{\mu}_T, \boldsymbol{\Sigma}_{\mathbf{h}_T}) \quad (6.2f)$$

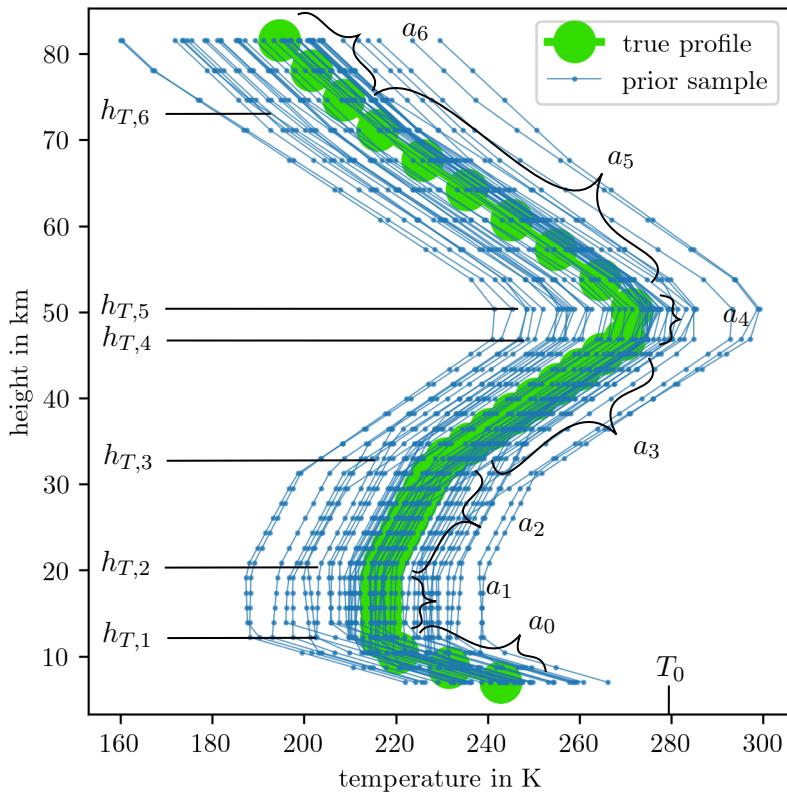
$$T_0 \sim \mathcal{N}(\mu_{T_0}, \sigma_{T_0}) \quad (6.2g)$$

$$p_0 \sim \mathcal{N}(\mu_{p_0}, \sigma_{p_0}) \quad (6.2h)$$

$$b \sim \mathcal{N}(\mu_b, \sigma_b) \quad (6.2i)$$

and define a normally distributed likelihood (due to Gaussian noise) and normally distributed priors. Before we formulate the posterior distribution, we carefully define  $\theta_\gamma, \theta_\delta, \theta_{p_0}, \theta_b, \theta_h, \theta_{T_0}, \theta_a$ , the hyper-prior scales, shapes, means and variances, which are explicitly given in Tab. 6.1.

### 6.1.1 Prior Modelling

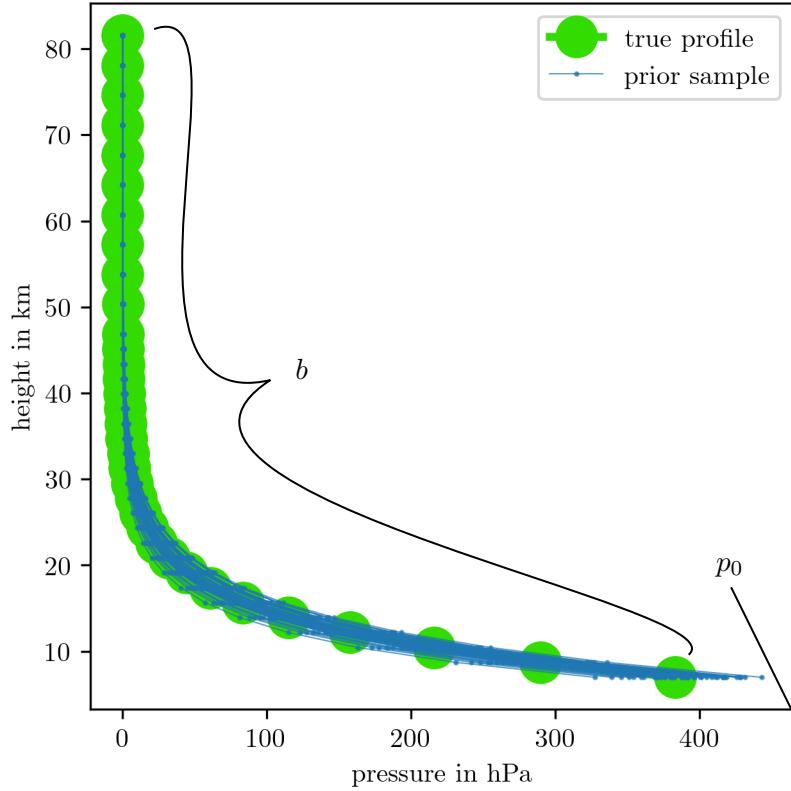


**Figure 6.2:** Prior samples from the hyper-prior distribution of  $\mathbf{h}_T$ ,  $\mathbf{a}$  and  $T_0$ , as defined in Tab. 6.1, where we calculate  $\mathbf{T}$  according to the function in Eq. 3.9.

We start by describing the pressure  $\mathbf{p}$  in between  $h_{L,0} \approx 7\text{km}$  and  $h_{L,n} \approx 82\text{km}$  with an exponential function

$$p(h) = \exp(-b h) p_0 \quad , h_{L,0} \leq h \leq h_{L,n} \quad (6.3)$$

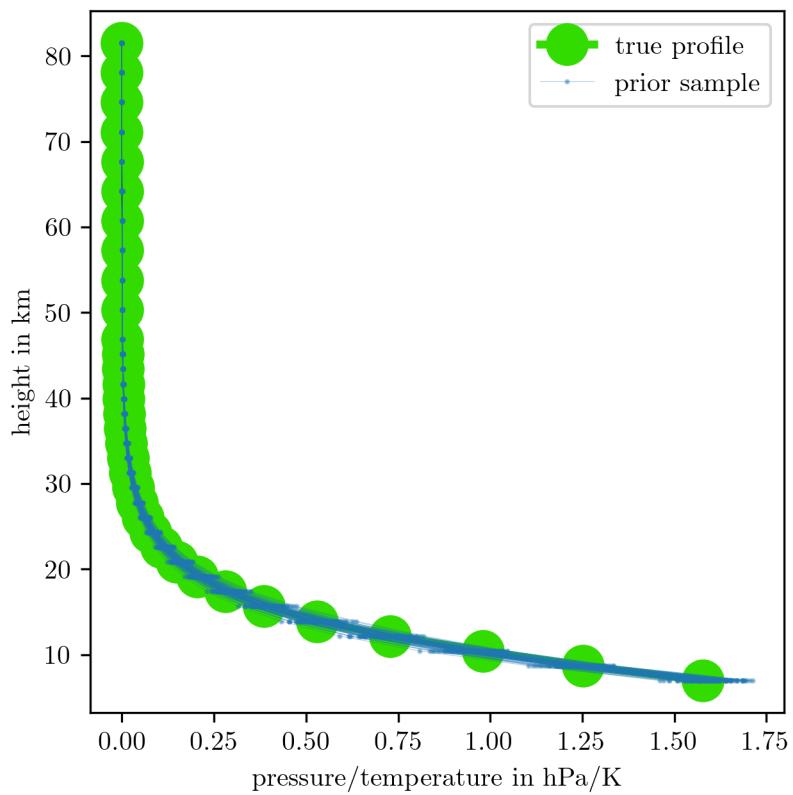
depending on two hyper-parameters  $p_0, b$ , see Fig. 6.3. Similarly, the temperature as described in Eq. 3.9 can be parametrised with 14 hyper-parameters  $\mathbf{h}_T = \{h_{T,1}, h_{T,2}, h_{T,3}, h_{T,4}, h_{T,5}, h_{T,6}\}$ ,  $\mathbf{a} = \{a_0, a_1, a_2, a_3, a_4, a_5, a_6\}$  and  $T_0$  (see Fig. 6.2 and Eq. 3.9). To complete the model, we have to define sensible hyper-prior variances and means of the normally distributed hyper-prior distribution for pressure and temperature-related hyper-parameters, where  $\pi(\delta, \gamma)$  are the same gamma distributions as previously



**Figure 6.3:** Prior samples from the hyper-prior distribution of  $b$  and  $p_0$  as defined in Tab. 6.1, where we calculate  $\mathbf{p}$  according to the function in Eq. 6.3.

defined in Sec. 4.1.1. We tune the normal distribution  $\pi(\mathbf{h}_T)$ , so that the temperature profile maintains its structure,  $h_{T,i} < h_{T,i+1}$  for  $i = 1, \dots, 5$  (see Fig. B.7) and set  $\pi(\mathbf{a})$  to a normal distribution as well. Similarly, we set  $\pi(T_0)$  to a normal distribution, so that it mimics a daily temperature variability of roughly 30K. We choose those rather informative hyper-prior distributions, because we find (see Fig. 6.4) that the data is uninformative about the temperature profile. The hyper-prior distribution  $\pi(p_0, b)$  for pressure-related hyper-parameters is also normally distributed, but with rather large variance  $\sigma_b^2$ , where  $p_0$  has a variability of around 80hPa, close to what we can observe when looking at weather data. We set means of the normal distribution  $\pi(p_0, b, T_0, \mathbf{h}_T, \mathbf{a})$  to the ground truth values of  $\mathbf{T}$  and  $\mathbf{p}$  which for  $\mathbf{p}$  we find with the python function `scipy.optimize.curve_fit` (see Tab. 6.1).

We plot prior samples of the pressure  $\mathbf{p}$  in Fig. 6.3, the temperature  $\mathbf{T}$  in Fig. 6.2 and the ratio  $\mathbf{p}/\mathbf{T}$  in Fig. 6.4 against the ground truth profiles. Additionally, we plot prior samples of  $1/\mathbf{T}$  in Fig. B.8. Here we already observe that  $\mathbf{p}/\mathbf{T}$  inherits the structure of the pressure function and hence the model is uninformative about the temperature.



**Figure 6.4:** Prior samples from the hyper-prior distribution of  $\mathbf{h}_T$ ,  $\mathbf{a}$  and  $T_0$  for temperature as in Eq. 3.9 and  $\mathbf{b}$  and  $p_0$  for pressure as in Eq. 6.3. We plot  $\mathbf{p}/\mathbf{T}$ . The hyper-priors are defined in Tab. 6.1.

## 6.2 Posterior Distribution

### 6.2.1 Marginal Posterior Distribution – Pressure and Temperature

Here, we define the marginal and then the full conditional posterior distribution for the described Bayesian model. We either use the t-walk algorithm [6] to draw samples from  $\pi(p_0, b, T_0, \mathbf{h}_T, \mathbf{a}, \lambda, \gamma | \mathbf{y})$  or we utilise a TT approximation on a predefined grid to generate samples via the SIRT method with an MH correction step. In doing so, we guide the reader through the procedure and point out some key aspects of how we obtain an efficient TT approximation. Lastly, we use the RTO method to draw samples from the full conditional posterior  $\pi(\mathbf{x} | p_0, b, T_0, \mathbf{h}_T, \mathbf{a}, \lambda, \gamma, \mathbf{y})$ .

The marginal posterior is given as

$$\pi(\boldsymbol{\theta}_{p,T}, \lambda, \gamma | \mathbf{y}) \propto \lambda^{n/2} \gamma^{m/2} \exp \left\{ -\frac{1}{2} g(\boldsymbol{\theta}_{p,T}, \lambda) - \frac{\gamma}{2} f(\boldsymbol{\theta}_{p,T}, \lambda) \right\} \pi(\boldsymbol{\theta}_{p,T}, \lambda, \gamma), \quad (6.4)$$

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

$$f(\boldsymbol{\theta}_{p,T}, \lambda) = \mathbf{y}^T \mathbf{y} - (\mathbf{A}(\boldsymbol{\theta}_{p,T})^T \mathbf{y})^T (\mathbf{A}(\boldsymbol{\theta}_{p,T})^T \mathbf{A}(\boldsymbol{\theta}_{p,T}) + \lambda \mathbf{L})^{-1} (\mathbf{A}(\boldsymbol{\theta}_{p,T})^T \mathbf{y}), \quad (6.5a)$$

$$\text{and } g(\boldsymbol{\theta}_{p,T}, \lambda) = \log \det (\mathbf{A}(\boldsymbol{\theta}_{p,T})^T \mathbf{A}(\boldsymbol{\theta}_{p,T}) + \lambda \mathbf{L}). \quad (6.5b)$$

For each evaluation of  $\pi(\boldsymbol{\theta}_{p,T}, \lambda, \gamma | \mathbf{y})$  we compose  $\mathbf{A}_L(\boldsymbol{\theta}_{p,T})$  as in Chapter 3, and calculate  $f$  and  $g$  directly using the Cholesky decomposition.

#### Sampling from marginal posterior

T-walk Sampler as a Black Box If the number of hyper-parameters is large, we sample from the marginal posterior  $\pi(\boldsymbol{\theta} | \mathbf{y})$  using the t-walk sampler as by Christen and Fox [6], because it is easy-to-use and a quick-to-implement sampler. The t-walk chooses between four different types of steps on the target distribution and is employed as a black-box algorithm in default settings, requiring the specification of the number of samples, burn-in period, support region, and the target distribution. Convergence to the target distribution is guaranteed by the construction of this algorithm (see [6]). **Oh, are you defining some kind of decision tree about what kind of sampler you run. Did you introduce that ? At the moment this reads like a detective novel – I have no idea what is going on - - I have to guess. Not a good thing to have the reader guessing.**

For a ground truth, we run the t-walk [6] algorithm on  $\pi(\boldsymbol{\theta}_{p,T}, \lambda, \gamma | \mathbf{y})$ , where we set our objective to generate 1000 independent samples from the marginal posterior. We bound the maximum IACT (see Tab. 6.1 and Fig. B.10 to Fig. B.27) by 1100, so we run the t-walk for  $N = 1000 \times 1100$  steps with a burn-in period of  $N_{\text{burn-in}} = 100 \times 1100$ . We initialise the Python implementation of the t-walk [7] around the hyper-prior mean values and the mode of  $\pi(\lambda, \gamma | \mathbf{y})$ . We take a total number of  $N' = N + N_{\text{burn-in}} = 1210000$  steps in  $\approx 10$  mins within bounds given by the iteratively defined TT grid (see Tab. 6.1). We plot the resulting histograms in Fig. 6.7 to Fig. 6.11 and the trace of the samples in Fig. B.9.

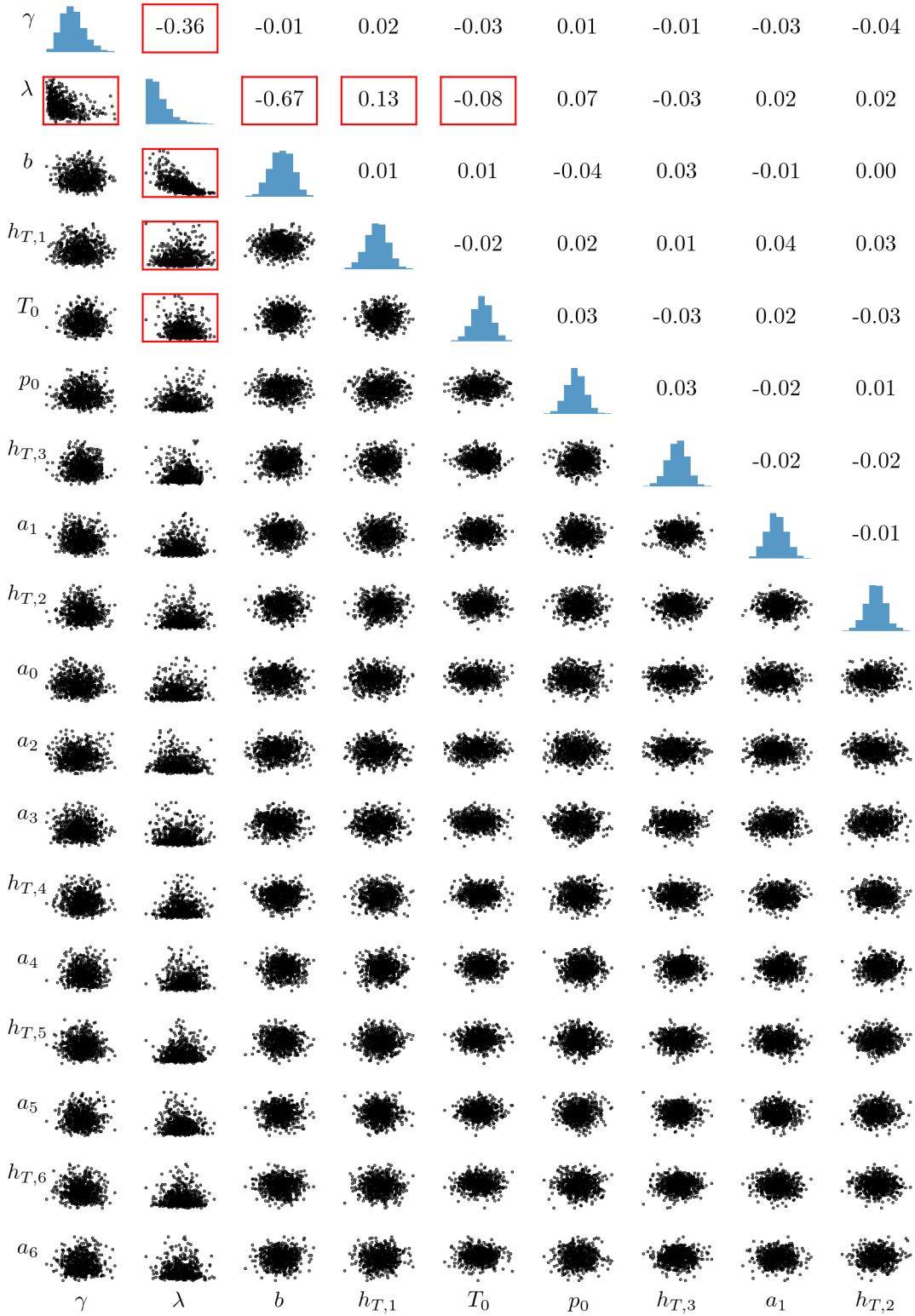
### TT Approximation of marginal posterior

The aim now is to approximate the square root of the marginal posterior

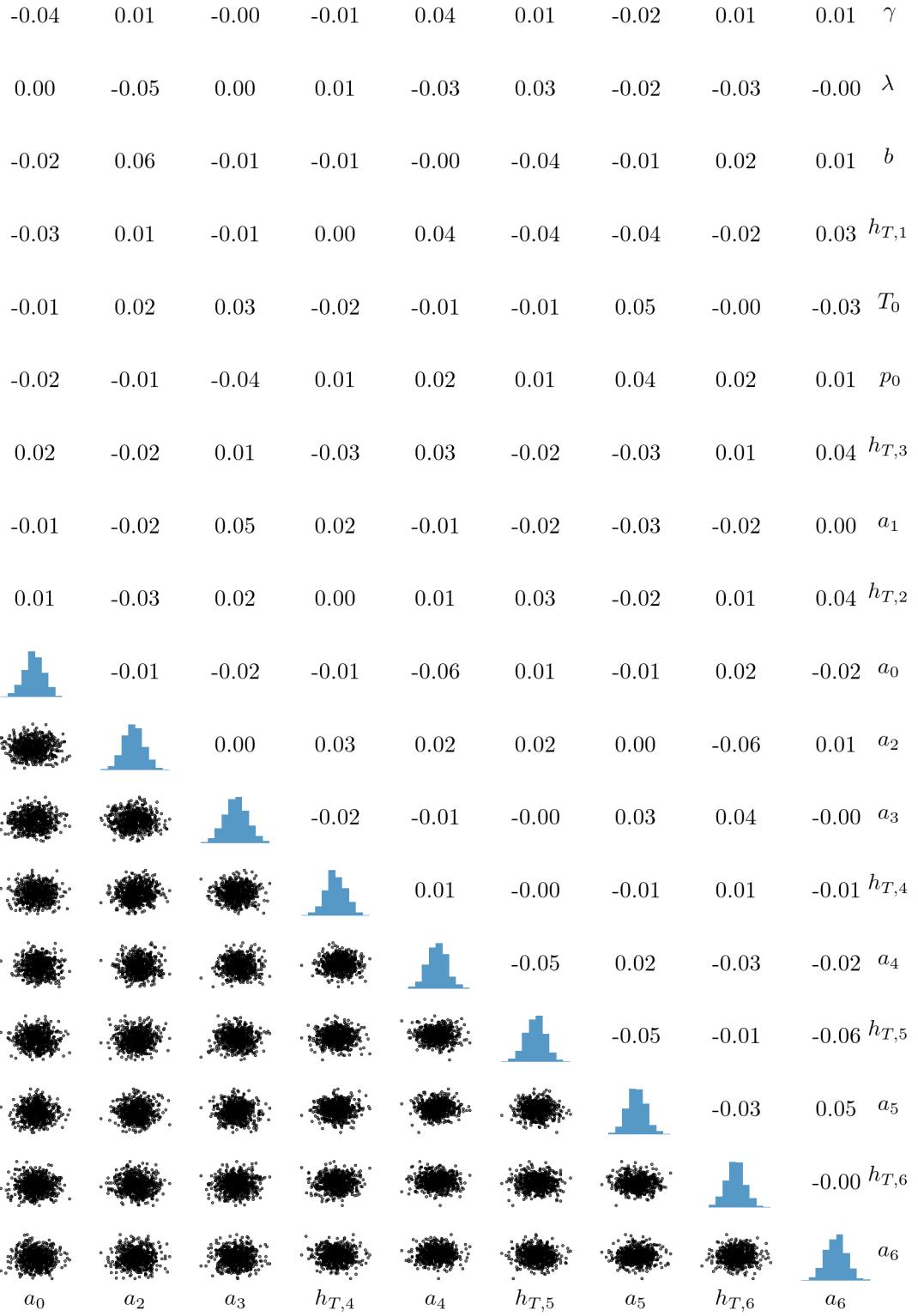
$$\sqrt{\pi(\lambda, \boldsymbol{\theta}_{\mathbf{p}, \mathbf{T}}, \gamma | \mathbf{y})} \propto \exp\left\{0.5 \log \pi(\lambda, \boldsymbol{\theta}_{\mathbf{p}, \mathbf{T}}, \gamma | \mathbf{y}) + c\right\}, \quad (6.6)$$

where we introduce a “normalisation constant”  $c = -50$  to avoid under or overflow and stay within computer precision. In doing so we run the `tt.cross.rectcross.rect_cross.cross` function from the `ttypy` python package [36]. We set the grid according to the results of the t-walk. If we compute the marginal as in Sec. 2.3, we set  $\xi = 1/\lambda(\mathcal{X})$  and  $\lambda(x) = 1$  so that for Cartesian basis  $\mathbf{M}_k = \text{diag}(\lambda_k(\mathcal{X}_k))$ . To draw samples from this TT approximation we use the SIRT-MH scheme as in Sec. 2.3.2.

**Correlation structure** First, we order the hyper-parameters according to their correlation structure to improve the efficiency of the TT approximation. Specifically, we arrange the hyper-parameter space  $\mathcal{X}_\gamma \times \mathcal{X}_\lambda \times \mathcal{X}_b \times \dots$  in such a way that highly correlated hyper-parameter pairs are adjacent and directly linked through their shared TT rank. In Fig. 6.5 we plot 1000 independent samples drawn via the SIRT-MH scheme from the TT approximation of  $\sqrt{\pi(\boldsymbol{\theta}_{\mathbf{p}, \mathbf{T}}, \lambda, \gamma | \mathbf{y})}$  and the Pearson correlation coefficient between hyper-parameter pairs. A coefficient close to 1 or  $-1$  indicates strong correlation, while values near zero suggest weak or no correlation. We observe that the hyper-parameters  $\lambda$  and  $b$ , and  $\lambda$  and  $\gamma$  are highly correlated. Additionally,  $h_{T,1}$  and  $T_0$  describing the temperature at low altitudes (strong signal) are mildly correlated to  $b$  as well. This is because  $h_{T,1}$  and  $T_0$  influence “the smoothness” of  $\mathbf{p}/\mathbf{T}$ , which is hard to see in Fig. 6.4. Interestingly,  $p_0$  appears largely uncorrelated with other hyper-parameters, while  $b$  is the key parameter linking pressure to ozone and temperature. Hyper-parameters describing temperature at higher altitudes are very much uncorrelated and the IACTs in Tab. 6.1 agree with those results.

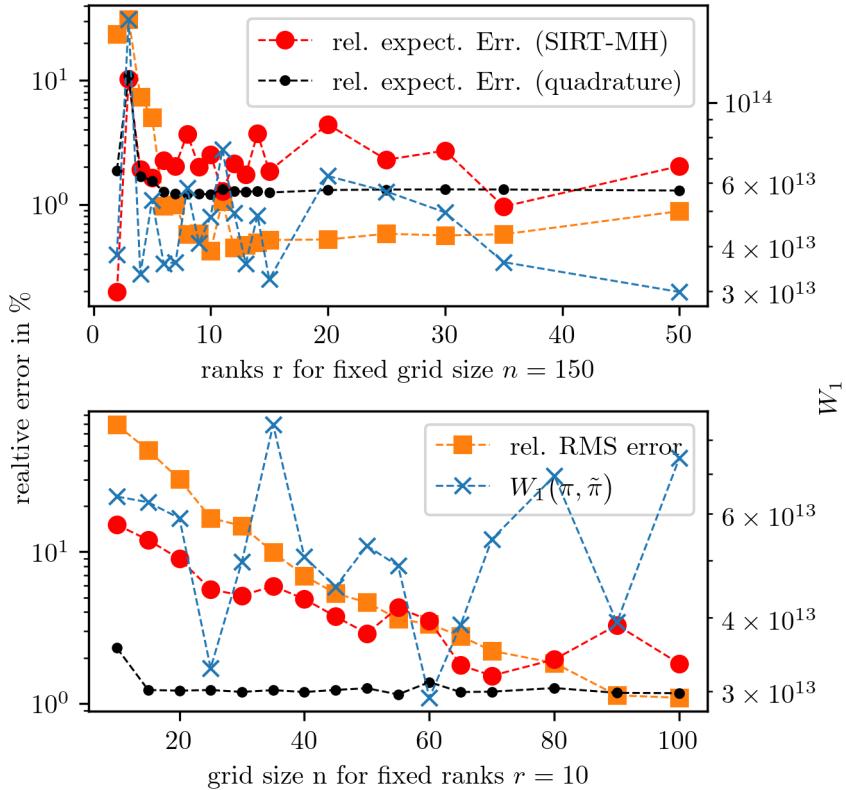


**Figure 6.5:** Plot of 1000 independent samples from TT approximation of  $\sqrt{\pi(\theta_{p,T}, \lambda, \gamma | \mathbf{y})}$  via SIRT-MH scheme. We plot the Pearson correlation coefficient ranging from  $-1$  to  $1$  for each hyper-parameter pair.



Correlation plot of samples from TT-approximation of  $\sqrt{\pi(p_0, b, T_0, \mathbf{h}_T, \mathbf{a}_T | \mathbf{y}, \gamma, \mathbf{x})}$  via SIRT scheme.

**Find optimal rank and grid size** Next we aim to determine the optimal rank and grid size to accurately approximate the marginal posterior, while decreasing the number of function evaluation. We set the number of grid points to  $n = 150$  and calculate different error measures for decreasing number of ranks to find the optimal number of ranks, where we compare to true marginal posterior function values and 1000 independent t-walk samples from that distribution. Then we fix a small but tolerable rank and decrease the number of grid points until sufficient accuracy.



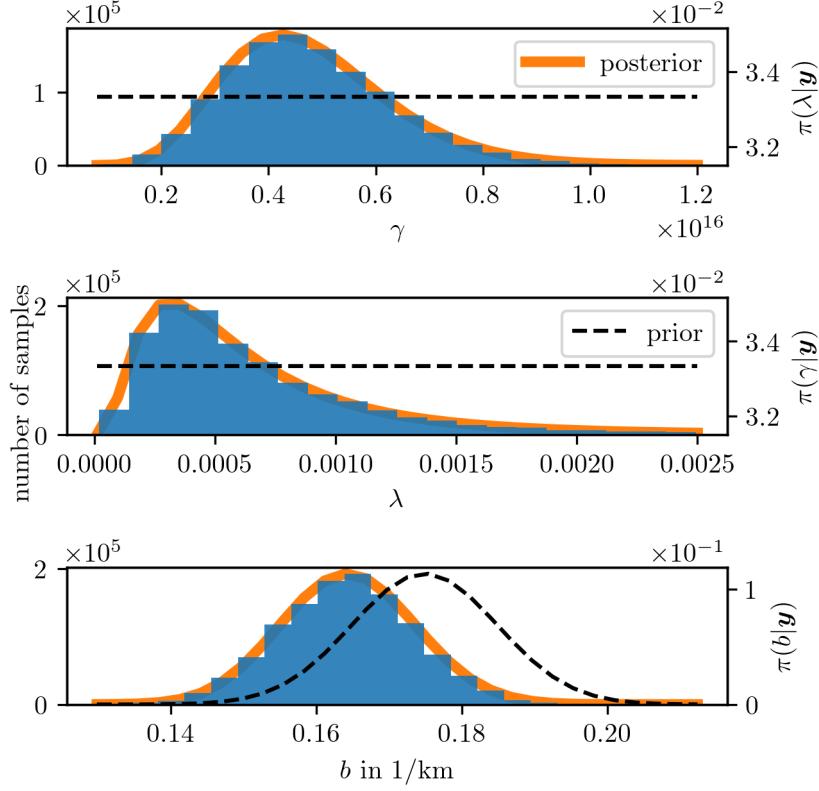
**Figure 6.6:** Given a TT approximation of  $\sqrt{\pi(\lambda, \theta_{p,T}, \gamma | \mathbf{y})}$ , we calculate the relative RMS error (orange squares) and the 1-Wasserstein distance (blue cross) between approximated values at sample points provided by the SIRT-MH and the true function values. We calculate the relative RMS error between the sample mean provided by the t-walk and mean values for the hyper-parameters calculated by quadrature (black dots), where we use the marginal function from the TT approximation as weights. Additionally, we plot the relative RMS error between the sample-based mean from the SIRT-MH and the t-walk (red circles).

For stable and comparable results, we do five sweeps in the `tt.cross.rectcross.rect_cross.cross` python function initialised with a random TT. Then we draw 1000 independent samples from the TT approximation of the marginal posterior via the SIRT-MH scheme. To calculate the 1-Wasserstein distance, as in Eq. 2.48, between SIRT-MH samples weighted with the TT approximation of marginal posterior and the t-walk samples, weighted by the true marginal posterior values, we use

the `SamplesLoss("sinkhorn", p=1, blur=0.05, scaling=0.8)` function with default settings from the Python package `geomloss` [16]. This provides the unbiased Sinkhorn divergence, which converges towards the Wasserstein distance and can be understood as the generalised Quicksort algorithm [15]. Here,  $p = 1$  defines the distance measure  $\|\mathbf{x} - \tilde{\mathbf{x}}\|_{L^2}$ , the blur parameter is an entropic penalty and the scaling parameter specifies the trade-off between speed ( $\text{scaling} < 0.4$ ) and accuracy ( $\text{scaling} > 0.9$ ) [16]. Additionally, we use the marginal functions of each TT approximation to calculate the means  $\boldsymbol{\mu}_{\text{TT}} \in \mathbb{R}^{18}$  of each hyper-parameter by weighted expectations. Then we obtain the relative RMS difference  $\|\boldsymbol{\mu}_{\text{TT}} - \boldsymbol{\mu}_{\text{t-walk}}\|_{L^2}/\|\boldsymbol{\mu}_{\text{t-walk}}\|_{L^2}$ . Here  $\boldsymbol{\mu}_{\text{t-walk}}$  denotes the “true sample-based means” from the t-walk. Further, we calculate the relative RMS between the SIRT-MH sample-based mean  $\boldsymbol{\mu}_{\text{SIRT-MH}}$  and  $\boldsymbol{\mu}_{\text{t-walk}}$ , and the relative RMS error at SIRT-MH samples compared to ground truth function values (Eq. 2.44).

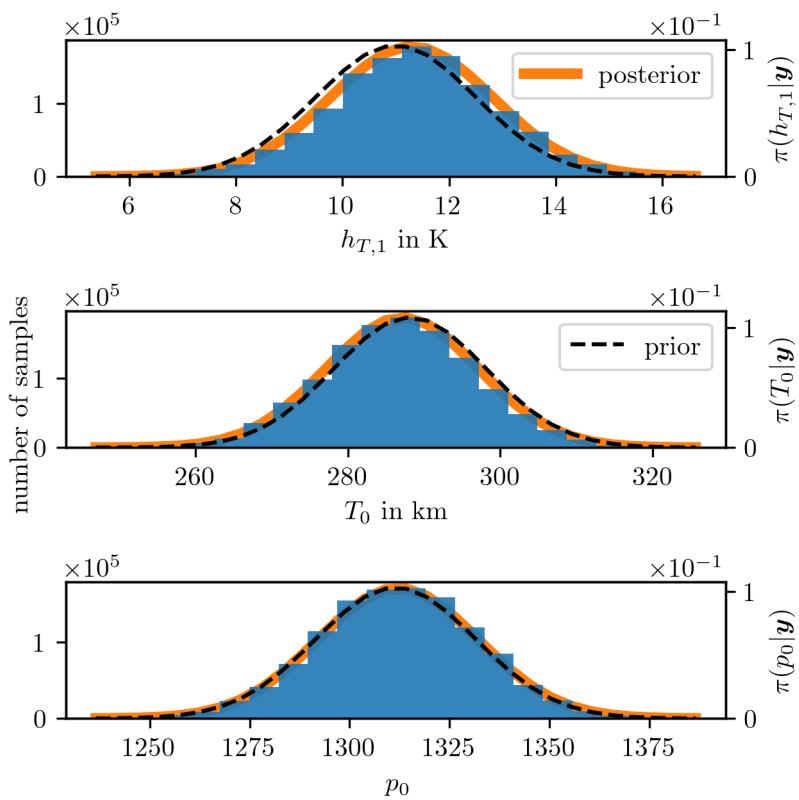
We plot all of these measures in Fig. 6.6, where we observe that a rank  $r = 10$  is sufficient because the error measures are relatively stable for  $r \geq 10$ . Then we decrease the grid size and decide that for  $n \geq 20$  the relative differences of sample-based means to  $\boldsymbol{\mu}_{\text{t-walk}}$  (red circles in Fig. 6.6) and the RMS error at the SIRT-MH samples (orange squares in Fig. 6.6) around 20% is good enough. We relate a more accurate interpolation in between grid points to decreasing sample-based relative errors for an increasing number of grid points, since the chosen linear interpolation (see Eq. 2.41) is a rather rudimentary choice. The quadrature-based relative expectation error (black dots in Fig. 6.6) is almost constant for ranks  $\geq 7$  and grid sizes  $\geq 15$ . Since the hyper-parameters have different length scales, we are only interested in the trend of the 1-Wasserstein distance (blue crosses in Fig. 6.6). The 1-Wasserstein distance is quite fluctuant but decreases with increasing ranks and stays within a similar range for decreasing grid size.

Further, we decrease the number of functions evaluations and define ranks  $r = [1, 10, 10, 10, 10, 10, 5, 5, 5, 5, 3, 2, 2, 2, 2, 2, 2, 1]$  harvesting the correlation structure of  $\pi(\boldsymbol{\theta}_{\mathbf{p}, \mathbf{T}}, \lambda, \gamma | \mathbf{y})$  even more. We do one sweep in the `tt.cross.rectcross.rect_cross.cross`, reducing the computation time to  $\approx 7$ s and the number of function evaluations to 24120, where we initialise at a previously calculated approximation. We report an average IACT (provided by [62, 26]) of  $\approx 1.2 \pm 0.2$  for the samples drawn via the SIRT-MH scheme, which means that we need two function evaluations per independent sample, once the TT approximation is available. **To draw 1000 independent samples, including generating a TT approximation, takes  $\approx 30$ s, and we report a relative RMS error of  $\approx 25\%$  evaluated over those 1000 independent samples. Additionally, we report a relative RMS error of  $\leq 1\%$  on 1000 randomly chosen grid points, indicating that the linear interpolation causes most of the approximation error.** We plot the marginals for each hyper-parameter in Fig. 6.7 to Fig. 6.12 and samples in

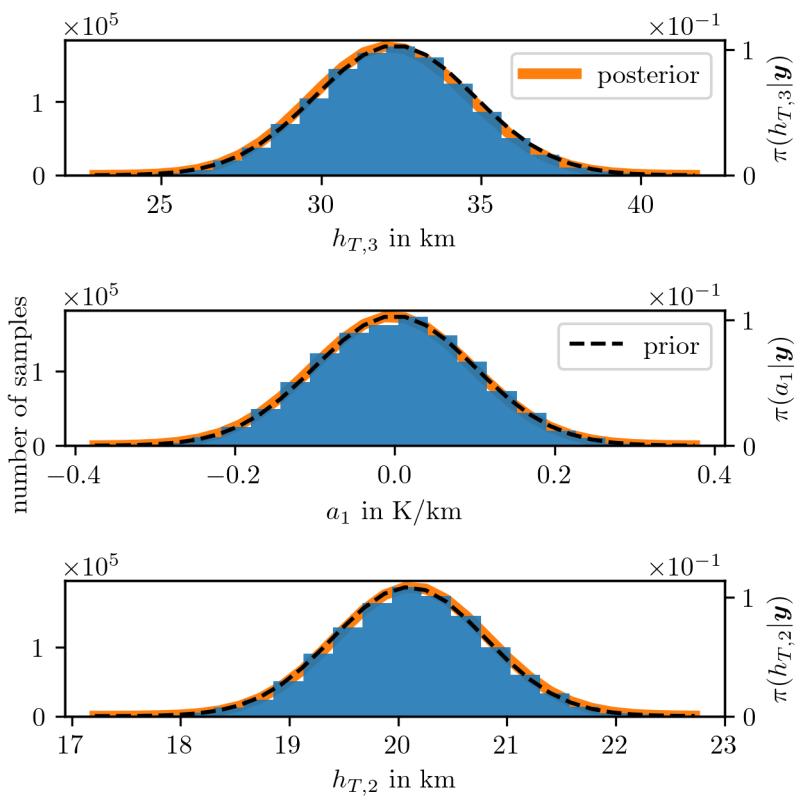


**Figure 6.7:** TT approximation of the marginal posterior in orange and the samples as a histogram as well as the prior distribution with a dotted line.

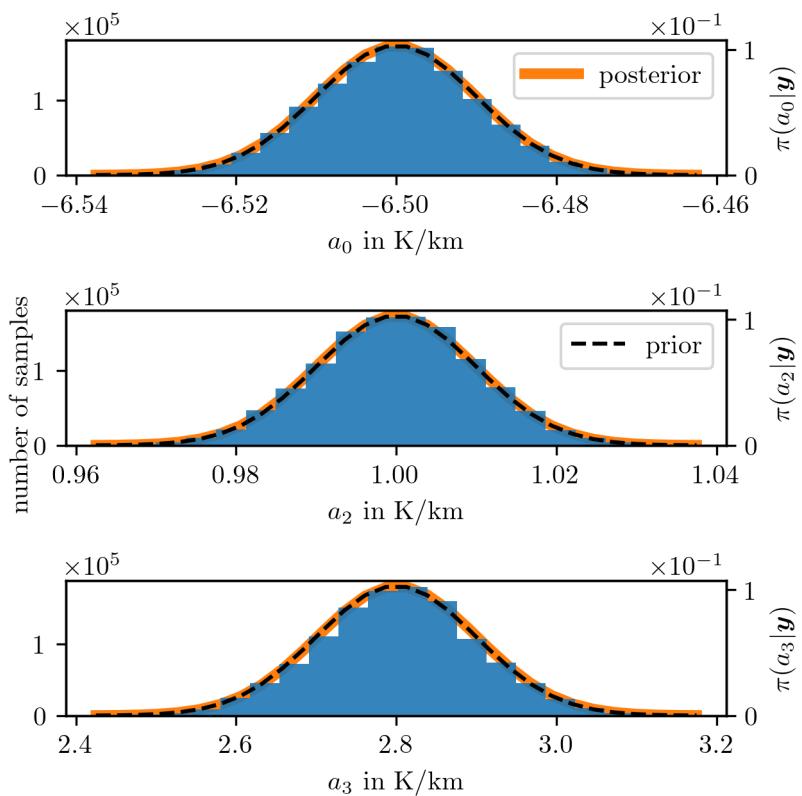
Fig. 6.5. We observe that, besides  $\lambda$  and  $\gamma$ , only the marginal posterior of the  $b$  hyper-parameter is seriously affected by the data and has significantly changed compared to the hyper-prior distribution.



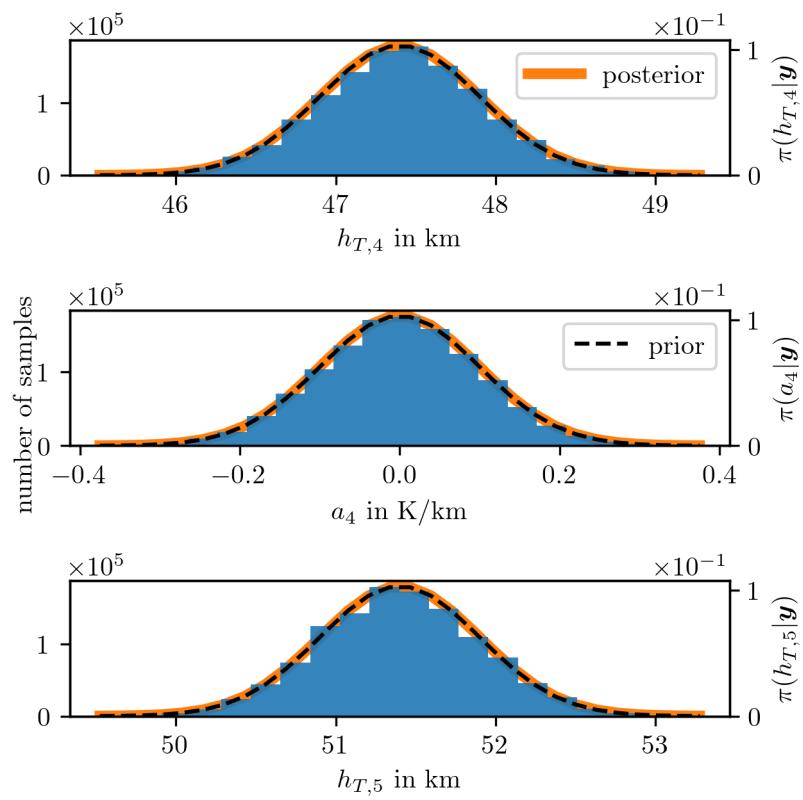
**Figure 6.8:** TT approximation of the marginal posterior in orange and the samples as a histogram as well as the prior distribution with a dotted line.



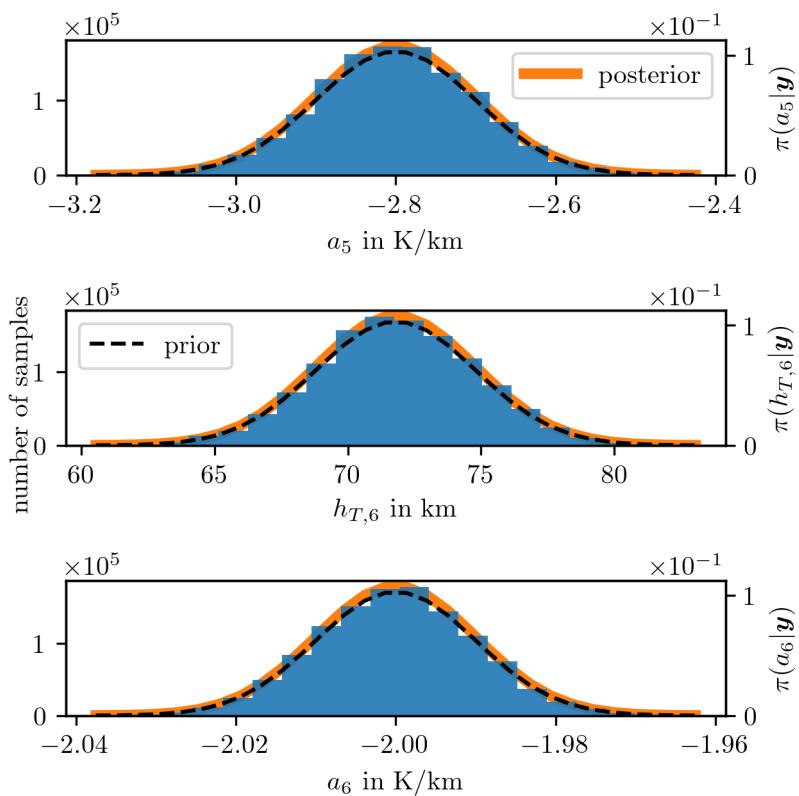
**Figure 6.9:** TT approximation of the marginal posterior in orange and the samples as a histogram as well as the prior distribution with a dotted line.



**Figure 6.10:** TT approximation of the marginal posterior in orange and the samples as a histogram as well as the prior distribution with a dotted line.

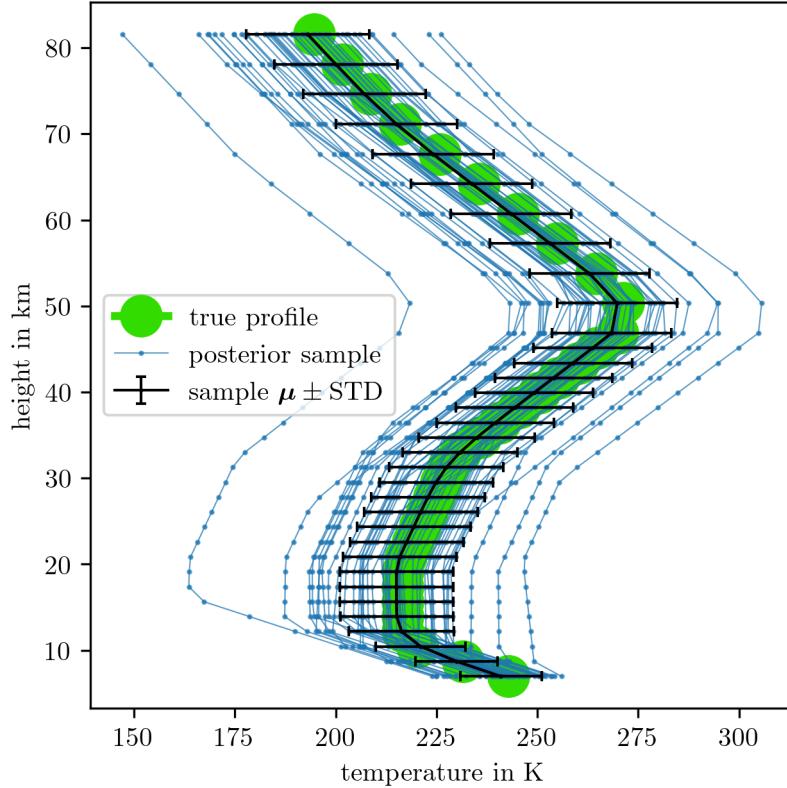


**Figure 6.11:** TT approximation of the marginal posterior in orange and the samples as a histogram as well as the prior distribution with a dotted line.



**Figure 6.12:** TT approximation of the marginal posterior in orange and the samples as a histogram as well as the prior distribution with a dotted line.

### Posterior pressure and temperature



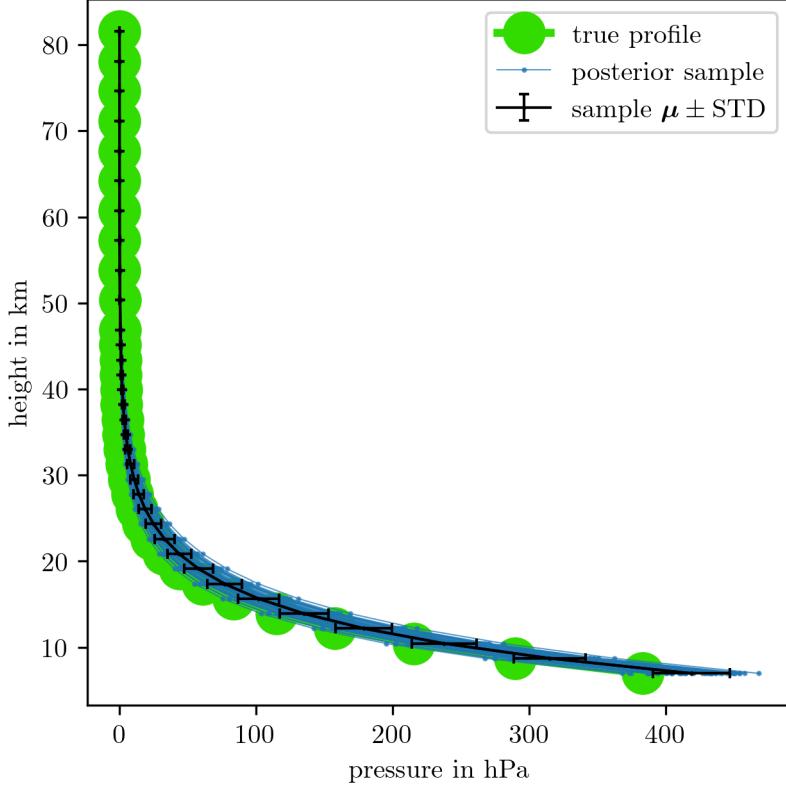
**Figure 6.13:** According to the hyper-parameter samples from the marginal posterior distribution (see Fig. 6.8 to Fig. 6.12) we plot the corresponding posterior temperature profile as given by Eq. 3.9.

#### 6.2.2 Full Conditional Posterior Distribution – Ozone

We use the RTO method (see Sec. 6.2.2) to obtain ozone samples from the full conditional posterior and sample hyper-parameter samples directly from the marginal posterior to compute posterior temperature and pressure profiles according to their respective function (see Eq. 3.9 and Eq. 6.3). We plot posterior profiles of ozone in Fig. 6.15, temperature in Fig. 6.13 and pressure in Fig. 6.14.

#### Randomise then optimise

If it is computationally not feasible to calculate the mean and the covariance matrix of the full posterior (see Eq. 2.17 and 2.18) via quadrature due to a large number of hyper-parameters, e.g.  $\boldsymbol{\theta} \in \mathbb{R}^{n_\theta}$  and  $n_\theta \geq 4$ , we need an alternative way to draw samples from  $\pi(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$ . For the linear-Gaussian Bayesian model, we condition on  $\boldsymbol{\theta}$  and draw samples from the full conditional normal posterior distribution  $\pi(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$  using the RTO method [2].



**Figure 6.14:** According to the hyper-parameter samples from the marginal posterior distribution (see Fig. 6.7 and Fig. 6.8) we plot the corresponding posterior pressure profile as given by Eq. 6.3.

OK, another condition and another method. You definitely need some kind of explanatory introduction to this. So these are going to be specific algorithms, but this section is 'general', so these end up being waffly and of no clear purpose. You have some work to do.

The full conditional posterior, as in Eq. 2.15, can be rewritten as

$$\pi(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y}) \propto \pi(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})\pi(\mathbf{x}|\boldsymbol{\theta}) \quad (6.7)$$

$$\propto \exp\left(-\frac{1}{2}(\mathbf{A}_{\boldsymbol{\theta}}\mathbf{x} - \mathbf{y})^T \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}(\mathbf{A}_{\boldsymbol{\theta}}\mathbf{x} - \mathbf{y})\right) \exp\left(-\frac{1}{2}(\boldsymbol{\mu} - \mathbf{x})^T \mathbf{Q}_{\boldsymbol{\theta}}(\boldsymbol{\mu} - \mathbf{x})\right), \quad (6.8)$$

$$= \exp\left(-\frac{1}{2} \|\hat{\mathbf{A}}\mathbf{x} - \hat{\mathbf{y}}\|_{L^2}^2\right), \quad (6.9)$$

where we define

$$\hat{\mathbf{A}} := \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1/2} \mathbf{A}_{\boldsymbol{\theta}} \\ \mathbf{Q}_{\boldsymbol{\theta}}^{1/2} \end{bmatrix}, \quad \hat{\mathbf{y}} := \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1/2} \mathbf{y} \\ \mathbf{Q}_{\boldsymbol{\theta}}^{1/2} \boldsymbol{\mu} \end{bmatrix} \quad [3, 4], \quad (6.10)$$

with  $\mathbf{A}(\boldsymbol{\theta}) := \mathbf{A}_{\boldsymbol{\theta}}$ ,  $\mathbf{Q}(\boldsymbol{\theta}) := \mathbf{Q}_{\boldsymbol{\theta}}$  and  $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}) := \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}$ , which are all dependent on the hyper-parameters  $\boldsymbol{\theta}$ . A sample  $\mathbf{x}^{(k)} \sim \pi(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$  from the full conditional posterior is

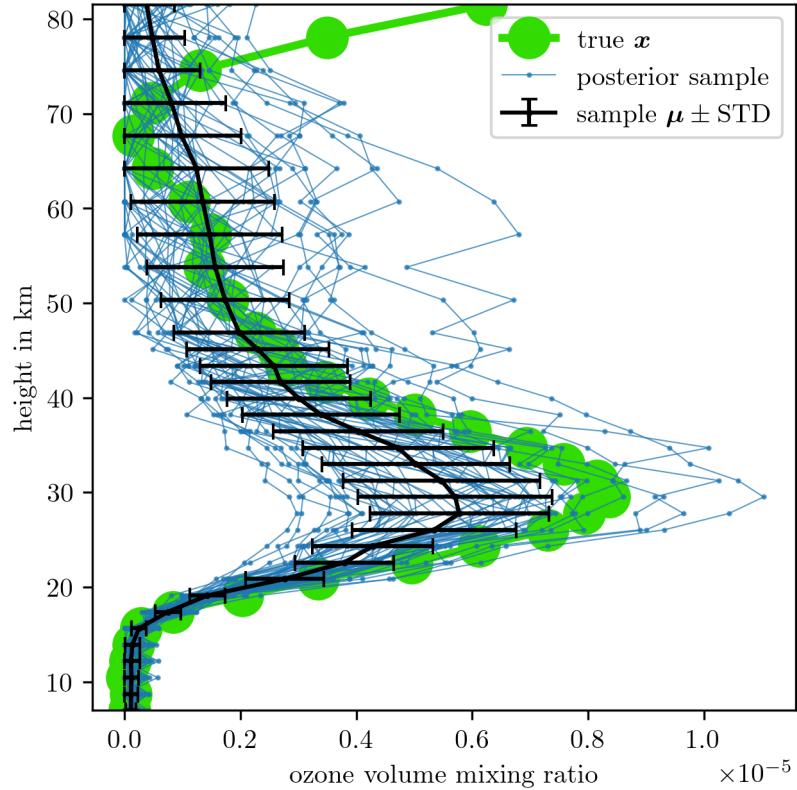
obtained by minimising the following equation with respect to  $\mathbf{x}$  :

$$\mathbf{x}^{(k)} = \arg \min_{\mathbf{x}} \|\hat{\mathbf{A}}\mathbf{x} - (\hat{\mathbf{y}} + \mathbf{b})\|_{L^2}^2, \quad \mathbf{b} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad (6.11)$$

where we add a random perturbation  $\mathbf{b}$ . Similar to Section ??, this expression becomes

$$(\mathbf{A}_\theta^T \Sigma_\theta^{-1} \mathbf{A}_\theta + \mathbf{Q}_\theta) \mathbf{x}^{(k)} = \mathbf{A}_\theta^T \Sigma_\theta^{-1} \mathbf{y} + \mathbf{Q}_\theta \mu + \mathbf{v}_1 + \mathbf{v}_2, \quad (6.12)$$

with  $\mathbf{v}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{A}_\theta^T \Sigma_\theta^{-1} \mathbf{A}_\theta)$  and  $\mathbf{v}_2 \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_\theta)$ , representing independent Gaussian random variables [2, 17].



**Figure 6.15:** According to the hyper-parameter samples from the marginal posterior distribution (see Fig. 6.7) we plot the corresponding ozone sample from the full conditional posterior using the RTO method.

We observe that pressure and ozone are highly correlated. Since the hyper-parameter  $b$  is smaller than its ground truth value, the posterior pressure profile does not exponentially decrease as strongly as the ground truth pressure (see Fig. 6.7). This results in posterior pressure values which are slightly larger than the ground truth, and in an average posterior ozone profile with much smaller peak values compared to the ground truth. Additionally, the individual posterior samples are more prior-dominated through larger  $\lambda$  values (see Fig. 6.7) and hence slightly smoother. Again, we are not able to recover the second peak at high altitudes. The posterior temperature profiles look (as expected) similar to the prior temperature profiles.



# 7

## Summary and Outlook

In this chapter, we summarise the key results and conclusions of our work and provide an outlook for future research. We compare the Bayesian approach to a regularisation approach and elaborate on the differences between sampling-based methods and the TT approximation. Lastly, we situate our results within the broader context of atmospheric modelling and discuss the implications for the future development of an atmospheric limb sounder.

### 7.1 Regularisation Solution vs. Hierarchical Bayesian Approach

Using a regularisation approach, we need 200 solves of  $\mathbf{x}_\lambda$  to obtain one solution of this inverse problem. In contrast, the hierarchical Bayesian approach involves 25 function evaluations of the marginal posterior (to find the mode) and then 10100 samples from the approximated marginal posterior followed by 20 evaluations of  $\mathbf{x}_\lambda$  and  $\mathbf{B}_\lambda^{-1}$  to characterise the full posterior in  $\approx 0.5$ s. Utilising a TT approximation (including finding the mode) to compute the full posterior mean and covariance takes  $\approx 0.025$ s, requiring only 400 function evaluations to approximate  $\pi(\lambda, \gamma | \mathbf{x})$  and is almost as fast as the regularisation approach ( $\approx 0.015$ ). Regardless, either method has a runtime of much less than a second on a basic laptop.

While regularisation yields a single optimal solution (point estimate), a Bayesian framework provides a distribution of ozone profiles, which are all feasible solutions to the inverse problem, and hence true errors. Moreover, within the hierarchical Bayesian approach, we can include prior knowledge about the noise, ozone profile and many more physical processes through hyper-parameters, offering an arbitrarily flexible and informative inference framework.

## 7.2 Sampling Methods vs. TT Approximation

Using the TT approximation involves far fewer function evaluations of the target distribution compared to sample-based methods, but requires a predefined grid and a normalisation constant, which, for now, we have to find iteratively. Relying solely on TT approximations may lead to a substantial amount of trial and error and dealing with numerical issues. Nevertheless, once properly configured, we have shown the potential and advantages of TT methods.

More specifically, the TT approximation of the 2-dimensional marginal posterior ( $\approx 0.02s$ ) is more than 20 times faster than the MWG sampler ( $\approx 0.5s$ ). Excluding the function evaluations for finding the mode of the marginal posterior, the MWG sampler takes 10100 steps while the TT approximation only needs 400 function evaluations; this is a factor of  $\approx 25$ . Alternatively, for low-dimensional distributions, it may be preferable to approximate integrals directly using existing freely available quadrature libraries and packages such as `quadpy`.

In higher dimensions, such as the 18-dimensional marginal posterior considered in this thesis, TT methods ( $\approx 0.5\text{min}$ ) outperform samplers like the t-walk ( $\approx 10\text{min}$ ), once a grid and normalisation constant have been defined. Although the t-walk may not be the best sampler for this specific marginal posterior and the underlying correlation structure, it is robust and easy to implement. To illustrate the efficiency of TT approximations, we compare the number of function evaluations per 1000 independent samples. For 1000 independent samples with a maximum IACT of 550 and a burn-in period of 100 independent samples, the t-walk needs 1210000 function evaluations. In contrast, 24120 function evaluations are enough to approximate the marginal posterior in the TT format. Then drawing 1000 independent samples via the SIRT-MH scheme requires another 2000 function evaluations with an IACT of  $\approx 1.2$ . So the cost per independent sample for the t-walk is 1210 and for the TT approach is  $\approx 27$  function evaluations, including the burn-in period and the TT approximation via the `rect_cross.cross` Python function. Or, after the burn-in phase, the t-walk requires around 1100 function evaluations per independent sample, while with an approximation of a probability density in the TT format available, only two function evaluations per independent sample are needed.

For future application, we suggest improving the efficiency of the TT approximation by, e.g., reducing the correlation structure through a coordinate system rotation or using better interpolators in between grid points to reduce the approximation error. This may be particularly important when the CDF in the SIRT scheme is not smooth due to poor approximations of the target density at previous samples. Moreover, using a different reference measure for integration as in [8], such as a Gaussian measure instead

of the current Lebesgue measure, may increase numerical stability. Currently, we have to predefine a normalisation constant and lower ranks manually, bounding the ranks automatically would be helpful (see e.g. [44]).

### 7.3 Atmospheric Physics

Here we summarise results within the context of our simplified physical atmospheric limb-sounding model. We demonstrated that the underlying non-linear forward model can be approximated with an affine map and the linear model, making this a linear inverse problem. For future application, we wish to include more measurement device-specific hyper-parameters in the forward model. This could include e.g. uncertainty in pointing accuracy or an antenna response function.

In Sec. 3.2.1, we showed that we do not gain more information if we measure more frequently or collect more data in noise-dominated regions and that we need an SNR of  $\approx 10^4$  to produce data, which is informative about ozone at higher altitudes.

Fig. 6.15 and Fig. 6.14 illustrate that pressure and ozone are highly correlated. One has to consider that when conditioning on pressure estimates from other measurements, as a slight change in pressure does skew the ozone VMR significantly. We could fix that by choosing a more restrictive prior for the pressure-related hyper-parameter  $b$ , but that would not be objective. By explanatory analysis, we found that data with an SNR of  $\approx 1000$  recovers an ozone (without a peak in higher altitudes) and pressure profile close to the ground truth. As previously mentioned in the prior analysis (see Fig. 6.4), the model as well as the data are uninformative about temperature and dominated by the exponentially decreasing pressure.

All the samples plotted in Fig. 5.4, Fig. 4.5, and Fig. 6.15 present valid solutions to the inverse problem, but consistently fail to capture the ozone peak at higher altitudes. This is due to noise-dominated data (see Fig. 3.7) and low signal strength in upper atmospheric regions, where the variability of the posterior ozone is large and primarily determined by the prior. We conclude that the main objective for future research is to develop a more accurate, potentially parametrised (prior) model, which captures physical properties and chemical processes of ozone in the atmosphere.



## References

- [1] Ambrosio, L., Brué, E., and Semola, D. *Lectures on Optimal Transport*. Cham: Springer Nature Switzerland, 2024.
- [2] Bardsley, J. “MCMC-based image reconstruction with uncertainty quantification”. In: *SIAM Journal on Scientific Computing* 34.3 (2012), A1316–A1332.
- [3] Bardsley, J., Solonen, A., Haario, H., and Laine, M. “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.
- [4] Bardsley, J. and Cui, T. “A Metropolis-Hastings-Within-Gibbs Sampler for Nonlinear Hierarchical-Bayesian Inverse Problems”. In: *2017 MATRIX Annals*. Vol. 2. MATRIX Book Series. Switzerland: Springer, 2019, pp. 2–12.
- [5] Champ, C. W. and Sills, A. V. “The Generalized Law of Total Covariance”. In: *preprint* (2022).
- [6] Christen, J. A. and Fox, C. “A general purpose sampling algorithm for continuous distributions (the t-walk)”. In: *Bayesian Analysis* 5.2 (2010), pp. 263 –281.
- [7] Christen, J. A. and Fox, C. *The t-walk software*. <https://www.cimat.mx/~jac/twalk/>. [Online; accessed 25/11/24]. CIMAT, Mexico, and University of Otago, New Zealand.
- [8] Cui, T. and Dolgov, S. “Deep composition of tensor-trains using squared inverse rosenblatt transports”. In: *Foundations of Computational Mathematics* 22.6 (2022), pp. 1863–1922.
- [9] Davis, P. J. and Rabinowitz, P. *Methods of numerical integration*. San Diego, CA: Academic Press, Inc., 1984.
- [10] Dick, J., Kuo, F. Y., and Sloan, I. H. “High-dimensional integration: The quasi-Monte Carlo way”. In: *Acta Numerica* 22 (2013), 133–288.
- [11] Dolgov, S., Anaya-Izquierdo, K., Fox, C., and Scheichl, R. “Approximation and sampling of multivariate probability distributions in the tensor train decomposition”. In: *Statistics and Computing* 30 (2020), pp. 603–625.
- [12] Dolgov, S. and Scheichl, R. “A Hybrid Alternating Least Squares–TT-Cross Algorithm for Parametric PDEs”. In: *SIAM/ASA Journal on Uncertainty Quantification* 7.1 (2019), pp. 260–291.
- [13] Duncan, B. *Aura at 20 Years*.  
<https://science.nasa.gov/science-research/earth-science/aura-at-20-years/>. [Online; accessed 31/08/25]. NASA’s Goddard Space Flight Center (GSFC), 2024.
- [14] Facility, A. N. C. D. *CubeSat Microwave Radiometer Mission to Support Global Ozone Layer Monitoring. Concept Study - Summary Report*. unpublished, internal report. Canberra BC: UNSW Canberra Space, 2023.
- [15] Feydy, J. “Analyse de données géométriques, au delà des convolutions”. Ph.D. Thesis. Université Paris-Saclay, July 2020.
- [16] Feydy, J. *GeomLoss – Geometric Loss functions between sampled measures, images and volumes*. <https://www.kernel-operations.io/geomloss/api/pytorch-api.html>. [Online; accessed 12/09/25].
- [17] Fox, C. and Norton, R. A. “Fast sampling in a linear-Gaussian inverse problem”. In: *SIAM/ASA Journal on Uncertainty Quantification* 4.1 (2016), pp. 1191–1218.

- [18] Fox, C. *Blokkurs on computing MCMC for inverse problems*. unpublished. Physics Department, University of Otago, 2025.
- [19] Fox, C., Dolgov, S., Morrison, M. E., and Molteno, T. C. “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.
- [20] Froidevaux, L. et al. “Validation of Aura Microwave Limb Sounder stratospheric ozone measurements”. In: *Journal of Geophysical Research: Atmospheres* 113.D15 (2008).
- [21] Geyer, C. J. “Practical markov chain monte carlo”. In: *Statistical science* (1992), pp. 473–483.
- [22] Gordon, I. E et al. “The HITRAN2020 molecular spectroscopic database”. In: *Journal of Quantitative Spectroscopy and Radiative Transfer* 277 (2022), p. 107949.
- [23] Hansen, P. C. *Discrete Inverse Problems: Insight and Algorithms*. Philadelphia: SIAM, 2010.
- [24] Hansen, P. C. “Regularization, GSVD and truncated GSVD”. In: *BIT numerical mathematics* 29.3 (1989), pp. 491–504.
- [25] Hansen, P. C. and O’Leary, D. P. “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.
- [26] Hesse, D. *py-uwerr; Python implementation of Monte Carlo error analysis a la Wolff*. <https://github.com/dhesse/py-uwerr>. [Online; accessed 09/09/25].
- [27] Kaipio, J. P. and Somersalo, E. *Statistical and Computational Inverse Problems*. New York: Springer-Verlag New York, 2005.
- [28] Lee, J. N. and Wu, D. L. “Solar Cycle Modulation of Nighttime Ozone Near the Mesopause as Observed by MLS”. In: *Earth and Space Science* 7.4 (2020).
- [29] Li, B., Miao, L., Zhang, C., and Yang, W. “A Lagrange Multiplier-based Regularization Algorithm for Image Super-resolution”. In: *2018 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*. 2018, pp. 422–426.
- [30] Livesey, N. J., Van Snyder, W., Read, W. G., and Wagner, P. A. “Retrieval algorithms for the EOS Microwave limb sounder (MLS)”. In: *IEEE Transactions on Geoscience and Remote Sensing* 44.5 (2006), pp. 1144–1155.
- [31] Livesey, N. J. et al. *Earth Observing System (EOS) Microwave Limb Sounder (MLS) Version 5.0x Level 2 and 3 data quality and description document*. Version 5.0-1.1a. NASA Goddard Earth Sciences Data and Information Services Center, 2022.
- [32] Livesey, N. J. et al. “Validation of Aura Microwave Limb Sounder O3 and CO observations in the upper troposphere and lower stratosphere”. In: *Journal of Geophysical Research: Atmospheres* 113.D15 (2008).
- [33] Meyn, S. P. and Tweedie, R. *Markov Chains and Stochastic Stability. 2nd Edition*. New York: Cambridge University Press, 2009.
- [34] Oseledets, I. “DMRG Approach to Fast Linear Algebra in the TT-Format”. In: *Computational Methods in Applied Mathematics* 11.3 (2011), pp. 382–393.
- [35] Oseledets, I. “Tensor-train decomposition”. In: *SIAM Journal on Scientific Computing* 33.5 (2011), pp. 2295–2317.
- [36] Oseledets, I., Bershtsky, D., and Saluev, T. *tpty - a Python implementation of the TT-toolbox*. <https://github.com/oseledets/tpty>. [Online; accessed 23/06/25]. 2018.
- [37] Oseledets, I. and Tyrtyshnikov, E. “TT-cross approximation for multidimensional arrays”. In: *Linear Algebra and its Applications* 432.1 (2010), pp. 70–88.
- [38] Read, W., Shippony, Z., Schwartz, M., Livesey, N. J., and Van Snyder, W. “The clear-sky unpolarized forward model for the EOS aura microwave limb sounder (MLS)”. In: *IEEE Transactions on Geoscience and Remote Sensing* 44.5 (2006), pp. 1367–1379.

- [39] Readings, C. *Envisat MIPAS An Instrument for Atmospheric Chemistry and Climate Research*. Noordwijk: ESA Publications Division, 2000.
- [40] Roberts, G. *ST911 Fundamentals of Statistical Inference - Part III*. Department of Statistics, University of Warwick, 2015.
- [41] Roberts, G. O. and Rosenthal, J. S. "General state space Markov chains and MCMC algorithms". In: *Probability Surveys* 1 (2004), pp. 20–71.
- [42] Roberts, G. O. and Rosenthal, J. S. "Harris recurrence of Metropolis-within-Gibbs and trans-dimensional Markov chains". In: *The Annals of Applied Probability* 16.4 (2006), 2123–2139.
- [43] Rodgers, C. D. "Retrieval of atmospheric temperature and composition from remote measurements of thermal radiation". In: *Reviews of Geophysics* 14.4 (1976), pp. 609–624.
- [44] Rohrbach, P. B., Dolgov, S., Grasedyck, L., and Scheichl, R. "Rank Bounds for Approximating Gaussian Densities in the Tensor-Train Format". In: *SIAM/ASA Journal on Uncertainty Quantification* 10.3 (2022), pp. 1191–1224.
- [45] Rue, H. and Held, L. *Gaussian Markov random fields: theory and applications*. London: CRC press, 2005.
- [46] Rybicki, G. B. and Lightman, A. P. *Radiative processes in Astrophysics*. Weinheim: Wiley-VCH, 2004.
- [47] Santosh, K., Das, N., and Ghosh, S. "Chapter 3 - Deep learning models". In: *Deep Learning Models for Medical Imaging*. Primers in Biomedical Imaging Devices and Systems. Academic Press, 2022, pp. 65–97.
- [48] Satopää, V., Albrecht, J., Irwin, D., and Raghavan, B. "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.
- [49] Schwartz, M., Froidevaux, L., Livesey, N., and Read, W. *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]. NASA Goddard Earth Sciences Data and Information Services Center, 2020.
- [50] Sedlmeir, F. et al. "Detecting THz in the telecom range: All resonant THz up-conversion in a whispering gallery mode resonator". In: *2014 39th International Conference on Infrared, Millimeter, and Terahertz waves (IRMMW-THz)*. 2014, pp. 1–2.
- [51] Šimečková, M., Jacquemart, D., Rothman, L. S., Gamache, R. R., and Goldman, A. "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.
- [52] Simpson, D., Lindgren, F., and Rue, H. "Think continuous: Markovian Gaussian models in spatial statistics". In: *Spatial Statistics* 1 (2012), pp. 16–29.
- [53] Sokal, A. "Monte Carlo Methods in Statistical Mechanics: Foundations and New Algorithms". In: *Functional Integration: Basics and Applications*. Boston, MA: Springer US, 1997, pp. 131–192.
- [54] Suresh, M. I. et al. "Multichannel upconversion of terahertz radiation in an optical disk resonator". In: *Opt. Express* 33.5 (Mar. 2025), pp. 10302–10311.
- [55] Tan, S. M., Fox, C., and Nicholls, G. K. *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]. Physics Department, University of Otago, 2016.
- [56] Thickstun, J. *Kantorovich-rubinstein duality*. [https://courses.cs.washington.edu/courses/cse599i/20au/resources/L12\\_duality.pdf](https://courses.cs.washington.edu/courses/cse599i/20au/resources/L12_duality.pdf). [Online; accessed 31/08/25]. University of Washington, 2019.

- [57] 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.
- [58] Ustin, S. and Middleton, E. M. “Current and near-term Earth-observing environmental satellites, their missions, characteristics, instruments, and applications”. In: *Sensors* 24.11 (2024), p. 3488.
- [59] Vats, D., Acosta, F., Huber, M. L., and Jones, G. L. “Understanding Linchpin Variables in Markov Chain Monte Carlo”. In: *preprint* (2022).
- [60] Wang, Y.-X., Sharpenack, J., Smola, A. J., and Tibshirani, R. J. “Trend Filtering on Graphs”. In: *Journal of Machine Learning Research* 17.105 (2016), pp. 1–41.
- [61] Waters, J. et al. “The earth observing system microwave limb sounder (EOS MLS) on the Aura satellite”. In: *IEEE Transactions on Geoscience and Remote Sensing* 44.5 (2006), pp. 1075–1092.
- [62] Wolff, U. “Monte Carlo errors with less errors”. In: *Computer Physics Communications* 156.2 (2004), pp. 143–153.
- [63] Wolff, U. *UWerr.m Version6*. <https://www.physik.hu-berlin.de/de/com/ALPHAssoft>. [Online; accessed 5/11/23]. Humboldt-Universität to Berlin, 2004.
- [64] Wolff, U., Bunk, B. hard, Korzec, T., Knechtli, F., and Bär, O. *Lecture Notes on Computational Physics II [in german]*. [www-com.physik.hu-berlin.de/comphys/comphys.htm](http://www-com.physik.hu-berlin.de/comphys/comphys.htm). [Online; accessed 29/08/25]. Humboldt University, Berlin, 2016.

# Appendices



# A

## Theoretical and Technical Background

### A.1 Correlation Structure

In the book Gaussian Markov Random Fields [45], Rue and Held demonstrate that a strong correlation between the hyper-parameter  $\mu$  and the latent field  $\mathbf{x}$  can significantly slow down convergence particularly when using Gibbs samplers. They consider the hierarchical model

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

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

and apply a Gibbs sampler based on the full conditional distributions

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

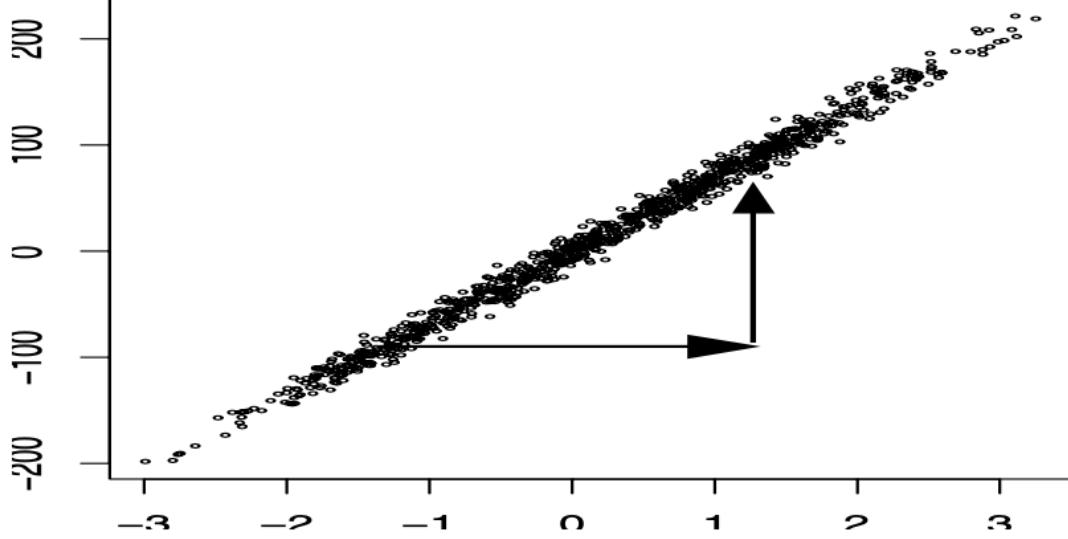
$$\mathbf{x}^{(k)} | \mu^{(k)} \sim \mathcal{N}(\mu^{(k)} \mathbf{1}, \mathbf{Q}^{-1}). \quad (\text{A.3})$$

As illustrated in Figure A.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 A.1 as a “squeeze” of the distribution.

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

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

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



**Figure A.1:** The figure taken from [45, Figure 4.1 (b)] shows samples from the chain of  $\mu$  (x-axis) and  $\mathbf{1}^T \mathbf{Q} \mathbf{x}^{(k)}$  (y-axis) for over 1000 iterations, based on the hierarchical model in Eq. A.1, with an autoregressive process encoded in  $\mathbf{Q}$ . The algorithm updates  $\mu$  and  $\mathbf{x}$  successively from their conditional distributions (see Eq. A.2 and Eq. A.3). 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.

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.

## A.2 Monte-Carlo Error and Integrated Autocorrelation Time

To assess the error  $(\sigma^{(i)})^2$  of a samples-based estimate

$$\bar{h}_N^{(i)} := \text{E}_{\mathbf{x}|\mathbf{y}}[h(\mathbf{x})] = \frac{1}{N} \sum_{k=1}^N h(\mathbf{x}^{(k)}), \quad (\text{A.6})$$

from the chain  $\mathcal{M}^{(i)} = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(k)}, \dots, \mathbf{x}^{(s)}, \dots, \mathbf{x}^{(N)}\} \sim \pi(\mathbf{x}|\mathbf{y})$ , we ignore systematic error due to initialisation bias (burn-in period), but we have to take into account that samples produced by any system or algorithm are correlated. To derive the IACT, we follow Ulli Wolff's lecture notes [64] (or alternatively [62]).

In general, the error of a Monte-Carlo estimate is:

$$(\sigma^{(i)})^2 = \text{Var}(\bar{h}_N^{(i)}) = \text{Var}(\text{E}_{\mathbf{x}|\mathbf{y}}[h(\mathbf{x})]) = \left( \frac{1}{N} \sum_{k=1}^N h(\mathbf{x}^{(k)}) - \bar{h}_N^{(i)} \right)^2. \quad (\text{A.7})$$

Expanding this summation, we see that

$$(\sigma^{(i)})^2 = \frac{1}{N^2} \sum_{k,s=1}^N \Gamma(k-s) \quad (\text{A.8})$$

with the autocorrelation coefficient  $\Gamma(k-s) = (h(\mathbf{x}^{(k)}) - \bar{h}_N^{(i)})(h(\mathbf{x}^{(s)}) - \bar{h}_N^{(i)})$ . Next we rewrite

$$\sum_{k,s=1}^N \Gamma(k-s) = \text{Var}(h(\mathbf{x})) \sum_{k,s=1}^N \frac{\Gamma(k-s)}{\Gamma(0)} = \text{Var}(h(\mathbf{x})) \sum_{k,s=1}^N \rho(k-s), \quad (\text{A.9})$$

with the normalised autocorrelation coefficient  $\rho(k-s) = \Gamma(k-s)/\Gamma(0)$  at lag  $k-s$  and  $\Gamma(0) = \text{Var}(h(\mathbf{x}))$  for  $k=s$ . Typically  $\Gamma(t)$  decays exponentially so that, for  $N \gg \tau$ ,  $\Gamma(t) \xrightarrow{t \rightarrow \infty} \exp\{-|t|/\tau\}$  and we can approximate

$$\sum_{k,s=1}^N \rho(k-s) = N \sum_{t=-(N-1)}^{N-1} \left(1 - \frac{t}{N}\right) \rho(t) \approx N \sum_{t=-\infty}^{\infty} \rho(t), \quad (\text{A.10})$$

see [53, p. 137]. If  $\tau \gg 1$

$$\sum_{t=-\infty}^{\infty} \rho(t) = 1 + 2 \sum_{t=1}^{\infty} (e^{-1/\tau})^t = 1 + 2 \frac{e^{-1/\tau}}{1 - e^{-1/\tau}} \approx 1 + 2 \frac{1 - 1/\tau}{1/\tau} = 2\tau - 1. \quad (\text{A.11})$$

Here we use the geometric power series  $\sum_{n=0}^{\infty} x^n = 1/(1-x)$  and the Taylor series  $e^x \approx 1+x$  for small  $x$ . In practice, the estimate for the Monte-Carlo error is:

$$(\sigma^{(i)})^2 \approx \frac{\text{Var}(h(\mathbf{x}))}{N} \sum_{t=-\infty}^{\infty} \rho(t) \approx \frac{\text{Var}(h(\mathbf{x}))}{N} \underbrace{\left(1 + 2 \sum_{t=1}^W \rho(t)\right)}_{:=\tau_{\text{int}}} = \text{Var}(h(\mathbf{x})) \frac{\tau_{\text{int}}}{N}, \quad (\text{A.12})$$

where  $W$  is the summation window and we define the IACT as twice the value as in [64, pp. 103-105]. The IACT provides a good estimate of how many steps the sampling algorithm needs to take to produce one independent sample. More specifically, the effective sample size  $\frac{\tau_{\text{int}}}{N}$  gives an estimate of how efficient a sampler is.

### A.3 Python Code

```

1 def MargBack(TTCore, univarGrid):
2     ''' Backward marginalisation (see Prop. 1) as in SIRT from Cui et al. [8] '''
3
4     dim = len(univarGrid)
5     B = dim * [None]    # coeffTensor
6     B[-1] = TTCore[-1]
7     R = [None] * dim
8     C = [None] * dim
9
10    for k in range(dim - 1, 0, -1):
11        r_kmin1, n, r_k = np.shape(TTCore[k])
12        # Eq. 2.28, [8, Eq. 22] !! we set Lebesgue Measure to const = one
13        M = np.identity(n) * (univarGrid[k][-1] - univarGrid[k][0])  # Mass matrix
14        L = scy.linalg.cholesky(M)
15
16        # construct Tensor C Eq. 2.33, [8, Eq. 27]
17        C[k] = np.zeros((r_kmin1, n, r_k))
18        for alpha in range(0, r_kmin1):
19            for l in range(0, r_k):
20                C[k][alpha, :, l] = B[k][alpha, :, l] @ L[:, :]
21
22        # unfold along first coordinate and compute thin QR decomposition of C^T
23        # Eq. 2.34, [8, Eq. 28]
24        Q, R[k] = np.linalg.qr(C[k].reshape((r_kmin1, n * r_k)), order='C').transpose(), mode='reduced')
25
26        # compute next coefficient tensor Eq. 2.35, [8, Eq. 29]
27        r_kmin2, n, r_kmin1 = np.shape(TTCore[k - 1])
28        B[k - 1] = np.zeros(np.shape(TTCore[k - 1]))
29        for alpha_2 in range(0, r_kmin2):
30            for l_1 in range(0, r_kmin1):
31                B[k - 1][alpha_2, :, l_1] = TTCore[k - 1][alpha_2, :, :] @ R[k][l_1, :]
32
33    return B

```

**Listing A.1:** Python code to calculate Backward marginals, as in Prop. 1 and [8].

```

1 def MargForw(TTCore, univarGrid):
2     ''' Forward marginalisation (see Prop. 2)
3         similar to backward marginalisation as in Cui et al. [8] '''
4
5     # compute pre marginal coefficients starting at dim = 1, k = 0
6     BPre = dim * [None] # coeffTensor
7     LebLam = 1 # !! Lebesgue Measure
8     BPre[0] = TTCore[0]
9     RPre = [None] * dim
10    CPre = [None] * dim
11
12    for k in range(0, dim-1):
13        r_kmin1, n, r_k = np.shape(TTCore[k])
14        # Eq. 2.28, [8, Eq. 22] !! we set Lebesgue Measure to const = one
15        M = np.identity(n) * (univarGrid[k][-1] - univarGrid[k][0]) # Mass matrix
16        L = scy.linalg.cholesky(M)
17
18        # construct Tensor C Eq. 2.36
19        CPre[k] = np.zeros((r_kmin1, n, r_k))
20        for alpha in range(0, r_kmin1):
21            for l in range(0, r_k):
22                CPre[k][alpha, :, l] = BPre[k][alpha, :, l] @ L[:, :]
23
24        # unfold along first coordinate and compute thin QR decomposition of C
25        # Eq. 2.37
26        Q, RPre[k] = np.linalg.qr(CPre[k].reshape((r_kmin1 * n, r_k)), order='C'), mode='reduced')
27
28        # compute next coefficient tensor Eq. 2.38
29        r_k, n, r_kpls1 = np.shape(TTCore[k + 1])
30        BPre[k + 1] = np.zeros(np.shape(TTCore[k + 1]))
31        for alpha_1 in range(0, r_kpls1):
32            for l_1 in range(0, r_k):
33                BPre[k + 1][l_1, :, alpha_1] = RPre[k][l_1, :] @ TTCore[k + 1][:, :, alpha_1]
34
35
36    return BPre
37

```

**Listing A.2:** Python code to calculate forward marginals, as in Prop. 2.

```

1 def SIRT(seeds, SQTT, univarGrid, BackMarg, absError):
2     ''' do squared inverse rosenblatt transform (SIRT) as in Cui et al. [8] '''
3
4     dim, numbSampl = seeds.shape
5     sampls = np.zeros(seeds.shape) # samples from approximated PDF
6     probVal = np.zeros(seeds.shape) # PDF values, for MH-correction step
7     Approx = np.zeros(seeds.shape[1]) # TT-Approx., to compare to true function
8
9     # Lebesgue measure for quadrature Eq. 2.20
10    WholeLebLam = np.zeros(dim)
11    for k in range(0, dim):
12        WholeLebLam[k] = (univarGrid[k][-1] - univarGrid[k][0])
13    lamX = np.ones(dim)
14    for k in range(1, dim):
15        lamX[k - 1] = np.prod(WholeLebLam[k:])
16
17
18    gamError = absError / np.prod(WholeLebLam) # error as in Eq. 2.25 [8, Eq. (35)]
19
20    # sample from first dimension [8, Eq. (30)]
21    firstMarg = gamError * lamX[0] + np.sum(BackMarg[0][0, :, :] ** 2, 1)
22    # cumulative distribution function, normalised numerically Eq. 2.39 [8, Eq. (17)]
23    firstCDF = np.cumsum(firstMarg / np.sum(firstMarg))
24    # draw samples as 'inverse transform'
25    sampls[0] = np.interp(seeds[0], firstCDF, univarGrid[0])
26    probVal[0] = np.interp(sampls[0], univarGrid[0], firstMarg / np.sum(firstMarg))
27
28    # sample from other dimensions
29    for n in range(0, numbSampl):
30        # piecew. poly. interpol in first dimension Eq. 2.41 [11]
31        CurrApprCore = LinInterPolTT(SQTT[0], univarGrid[0], sampls[0][n])
32        for d in range(1, dim):
33            # marginal function, conditioned on previous samples
34            rank_min, gridSize, rank_pls = BackMarg[d].shape
35            MargDep = np.zeros((BackMarg[d].shape))
36            for r in range(0, rank_min):
37                # condition on previous samples
38                MargDep[r, :, :] = CurrApprCore[0, r] * BackMarg[d][r, :, :]
39
40            currMarg = gamError * lamX[d] + np.sum(np.sum(np.copy(MargDep), axis=0)** 2,
41                axis=1) # Eq. 2.40 [8, Eq. (31)]
42
43            currCDF = np.cumsum(currMarg / np.sum(currMarg)) # Eq. 2.39 [8, Eq. (17)]
44
45            # draw sample as 'inverse transform'
46            sampls[d][n] = np.interp(seeds[d][n], currCDF, univarGrid[d])
47            probVal[d][n] = np.interp(sampls[d][n], univarGrid[d],
48                currMarg / np.sum(currMarg))
49            # piecew. poly. interpol., Eq. 2.41 [11], cond. on samp. for next PDF
50            CurrApprCore = np.copy(CurrApprCore) @ LinInterPolTT(SQTT[d], univarGrid[d],
51                sampls[d][n])
52
53            Approx[n] = gamError + CurrApprCore ** 2
54
55    return sampls, probVal, Approx

```

**Listing A.3:** Python code to draw samples via SIRT, as in Alg. Box 1.

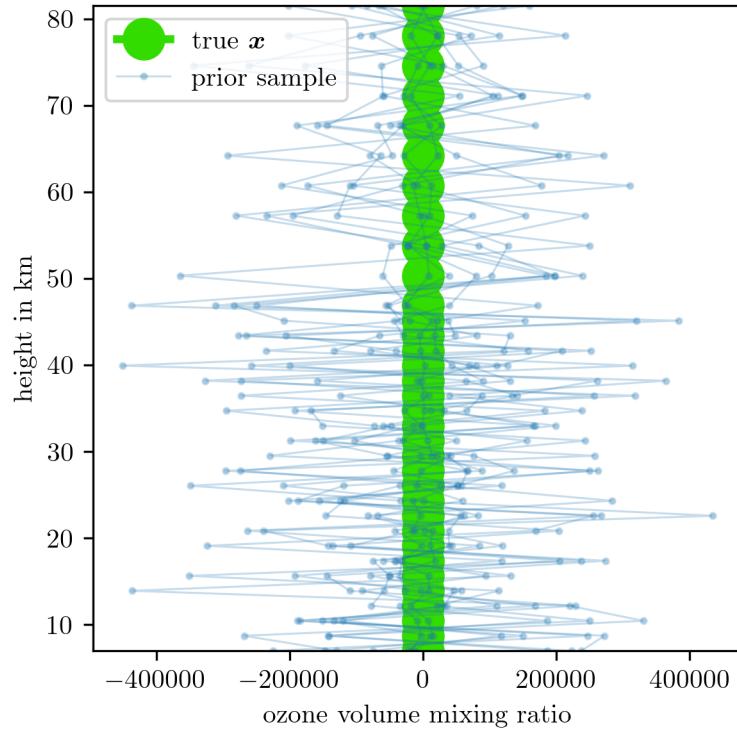
# B

## Additional Figures

Some of the additional figures shown are repetitive and omitted from the main document; others provide details that are not necessary for understanding the main results but may be interesting and offer a more visual understanding for the curious reader.

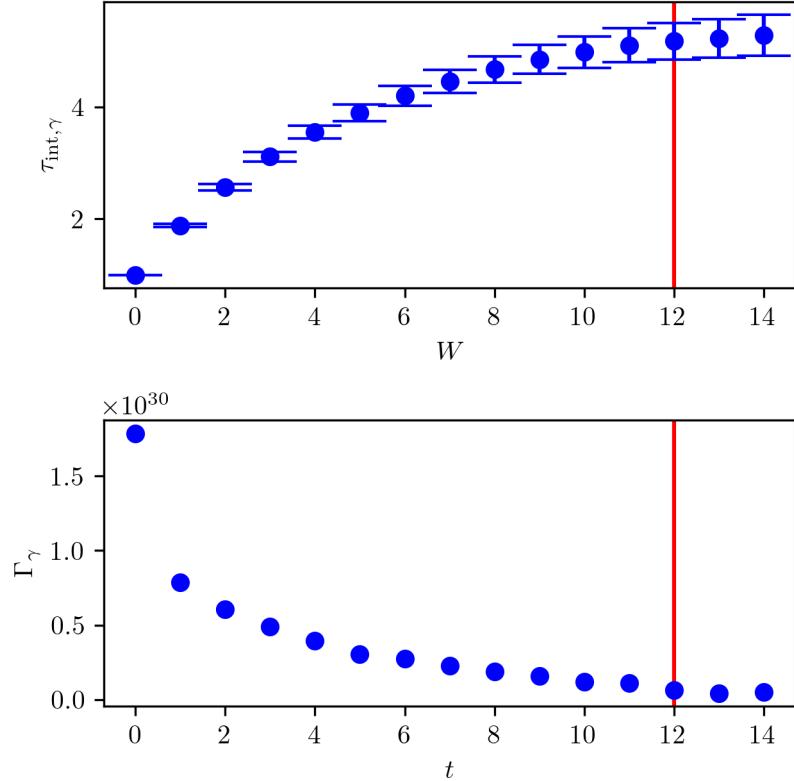
## B.1 Ozone

### B.1.1 Ozone Prior

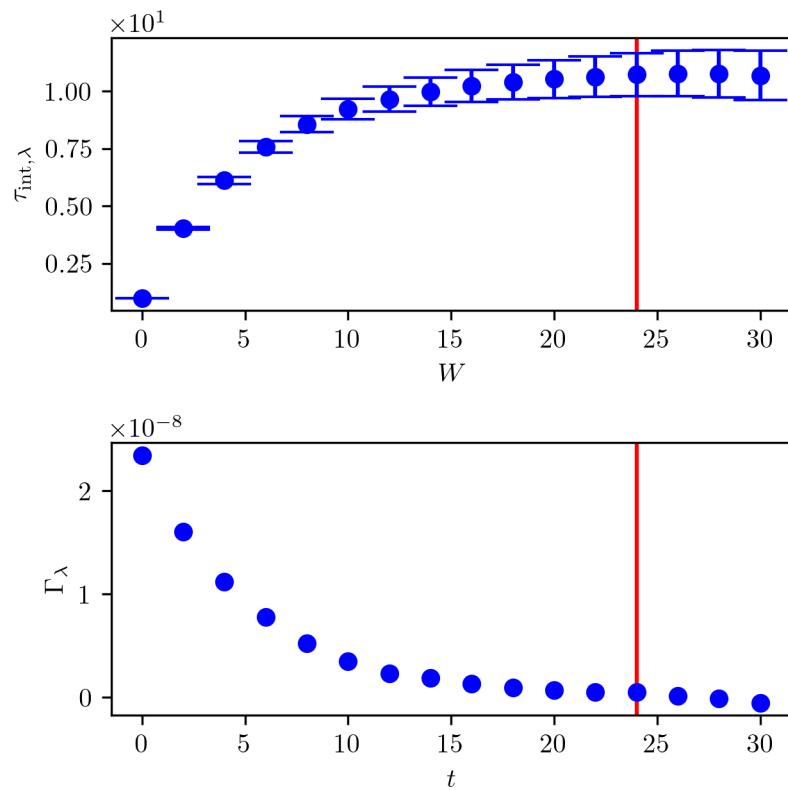


**Figure B.1:** We draw samples from ozone prior distribution  $\mathbf{x} \sim \mathcal{N}(0, \delta \mathbf{L})$  after generating a sample from the hyper-prior distribution  $\delta \sim \mathcal{T}(1, 10^{-10})$ . Note that since the variance of prior samples is very large compared to the ozone volume mixing ratios, the ozone profile appears to be constant, which is not the case, see e.g. Fig. 4.5.

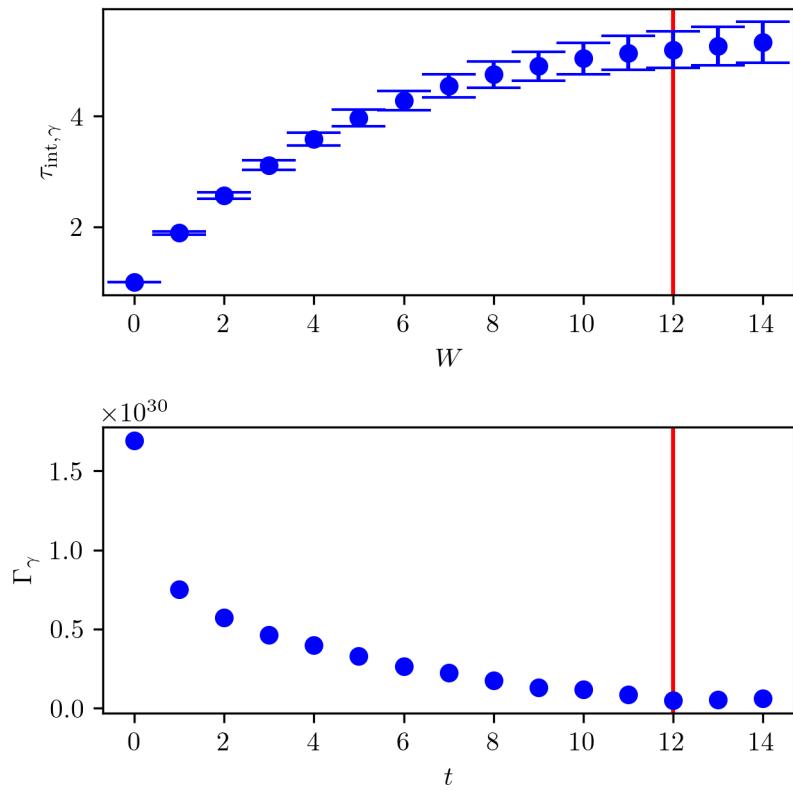
### B.1.2 Integrated Autocorrelation Time



**Figure B.2:** Provided by [26], the IACT  $\tau_{\text{int},\gamma}$  at summation windows  $W$  as well as the estimated autocorrelation function  $\Gamma_\gamma$  at lag  $t$  of the samples  $\gamma \sim \pi(\cdot|\mathbf{y})$  based on the linear forward model.

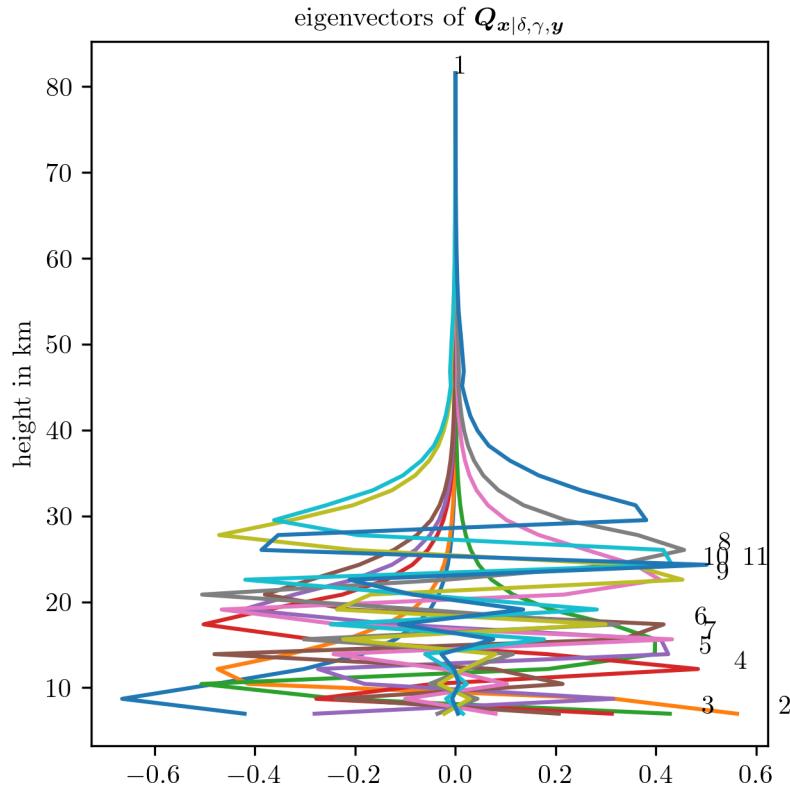


**Figure B.3:** Provided by [26], the IACT  $\tau_{\text{int},\lambda}$  at summation windows  $W$  as well as the estimated autocorrelation function  $\Gamma_\lambda$  at lag  $t$  of the samples  $\lambda \sim \pi(\cdot | \mathbf{y})$  based on the approximated forward model.

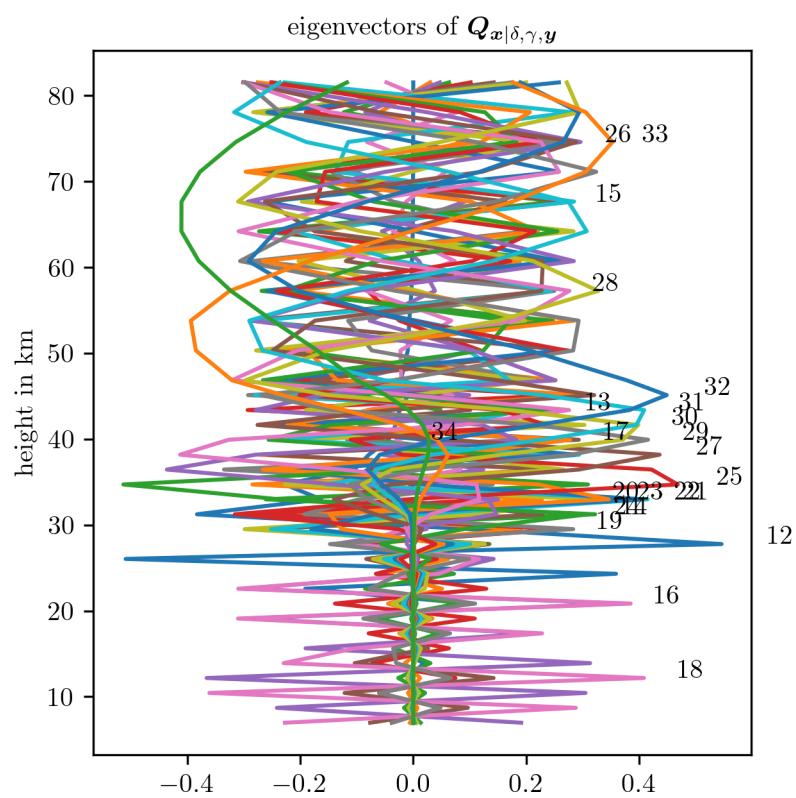


**Figure B.4:** Provided by [26], the IACT  $\tau_{\text{int},\gamma}$  at summation windows  $W$  as well as the estimated autocorrelation function  $\Gamma_\gamma$  at lag  $t$  of the samples  $\gamma \sim \pi(\cdot|\mathbf{y})$  based on the approximated forward model.

### B.1.3 Eigenvectors of Full Conditional Posterior Precision Matrix



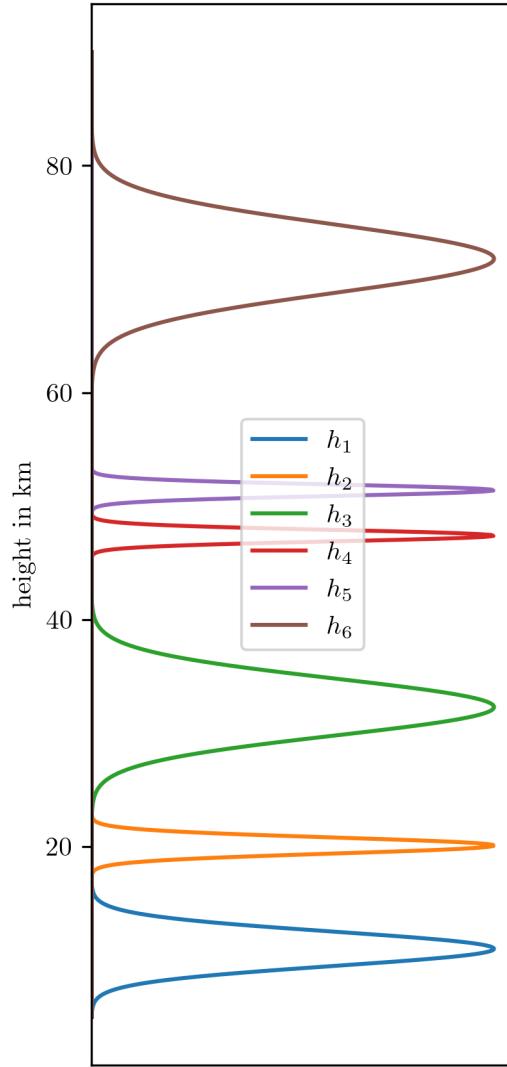
**Figure B.5:** First 11 eigenvectors corresponding to in size ordered eigenvalues of conditional precision matrix  $\mathbf{Q}_{\mathbf{x}|\delta,\gamma,\mathbf{y}}$ . We see that the eigenvectors span structures for heights  $\leq 40$ .



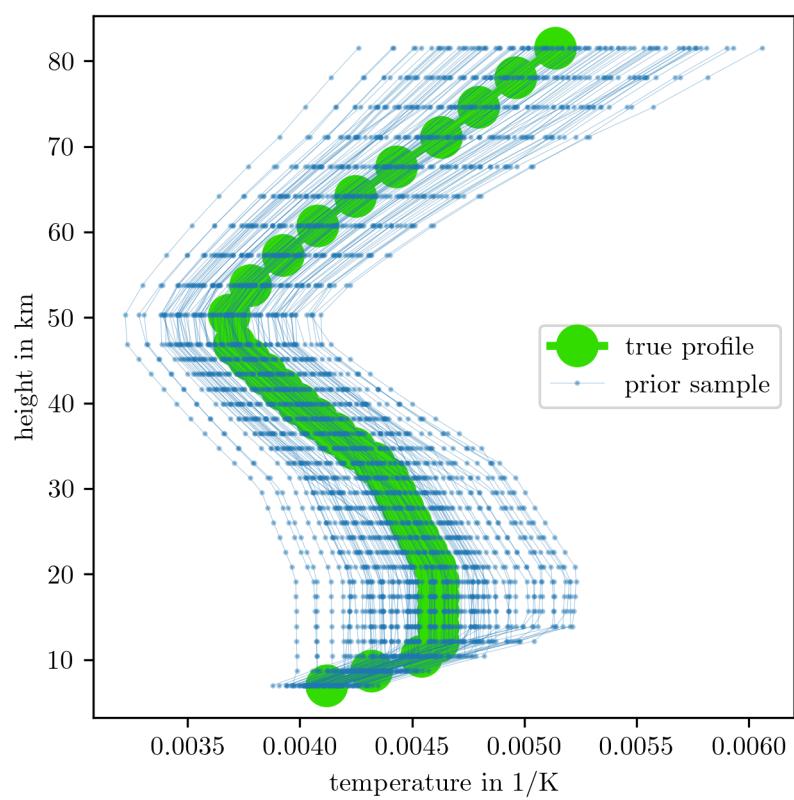
**Figure B.6:** Last 23 eigenvectors corresponding to in size ordered eigenvalues of conditional precision matrix  $\mathbf{Q}_{\mathbf{x}|\delta,\gamma,\mathbf{y}}$ . The eigenvectors represent structures according to the prior.

## B.2 Pressure and Temperature

### B.2.1 Priors

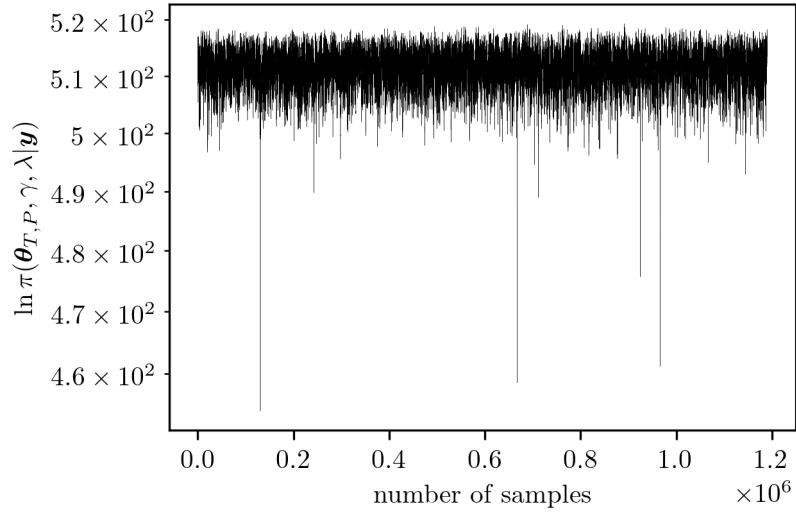


**Figure B.7:** Prior distributions  $\pi(\mathbf{h}_T)$ , which we choose so that they do not overlap and not conflict with the temperature function in Eq. 3.9.



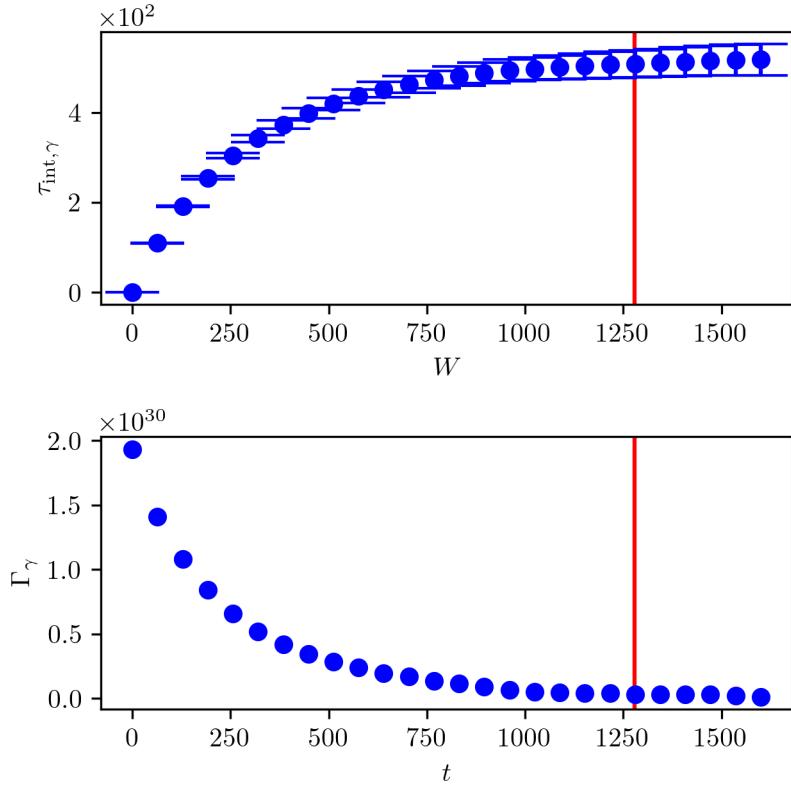
**Figure B.8:** Prior samples of the inverted temperature profile.

### B.2.2 T-walk Trace

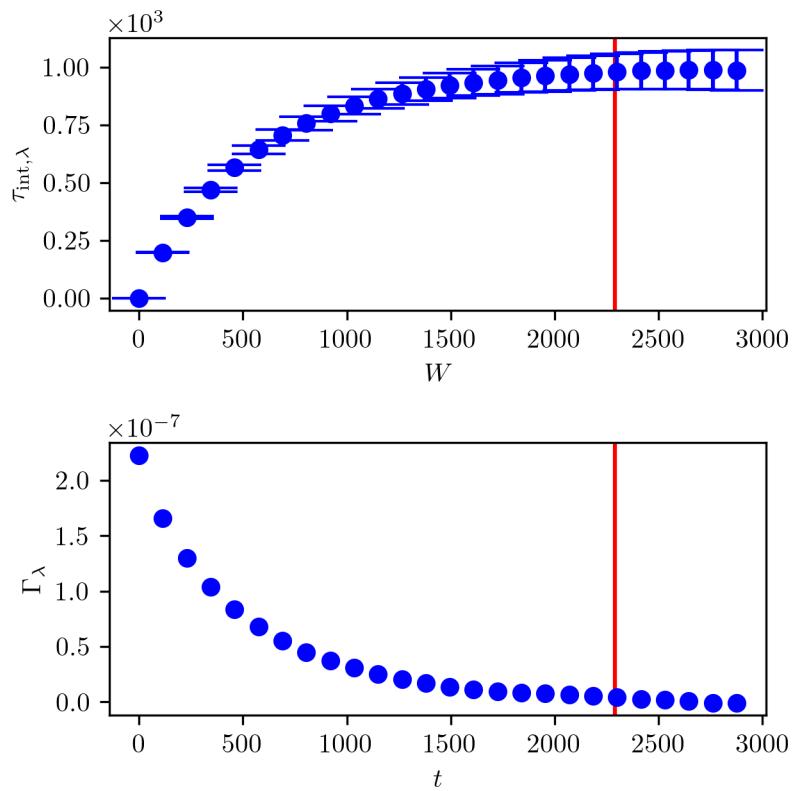


**Figure B.9:** Output trace of the t-walk on the posterior distribution  $\pi(p_0, b, \mathbf{h}_T, \mathbf{a} | \gamma, \mathbf{y})$ .

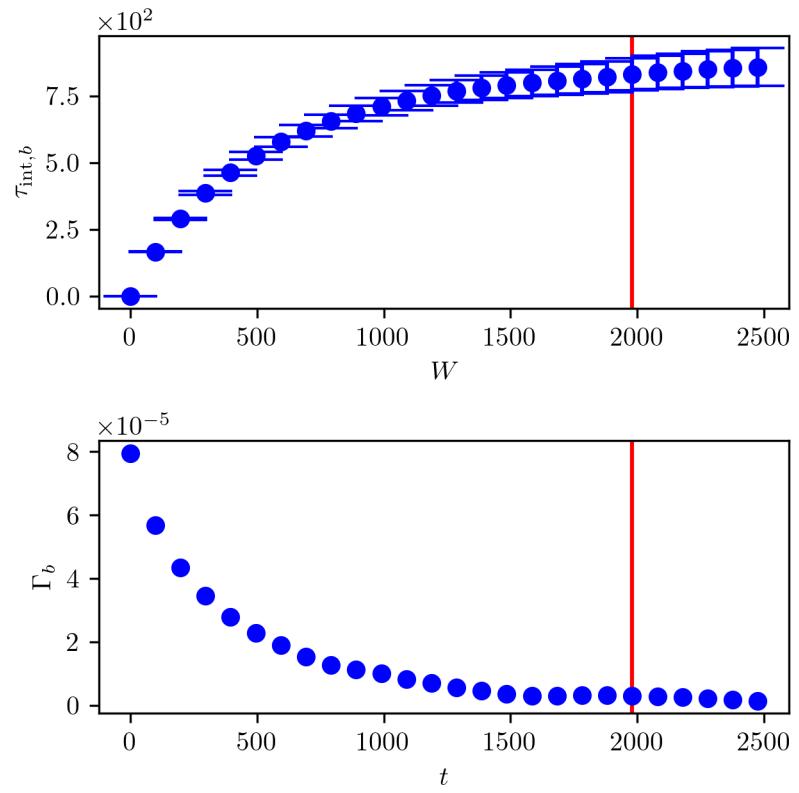
### B.2.3 Integrated Autocorrelation Time



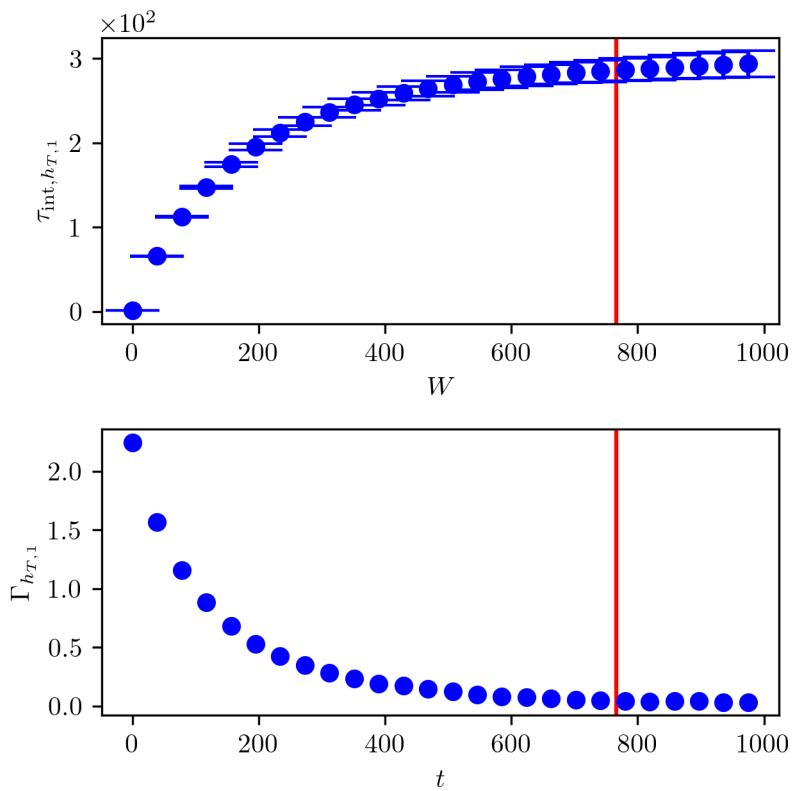
**Figure B.10:** Provided by [26], the IACT  $\tau_{\text{int},\gamma}$  at summation windows  $W$  and the estimated autocorrelation function  $\Gamma_\gamma$  at lag  $t$  of samples  $\gamma \sim \pi(\cdot|\mathbf{y})$  from the t-walk for the approximated forward model.



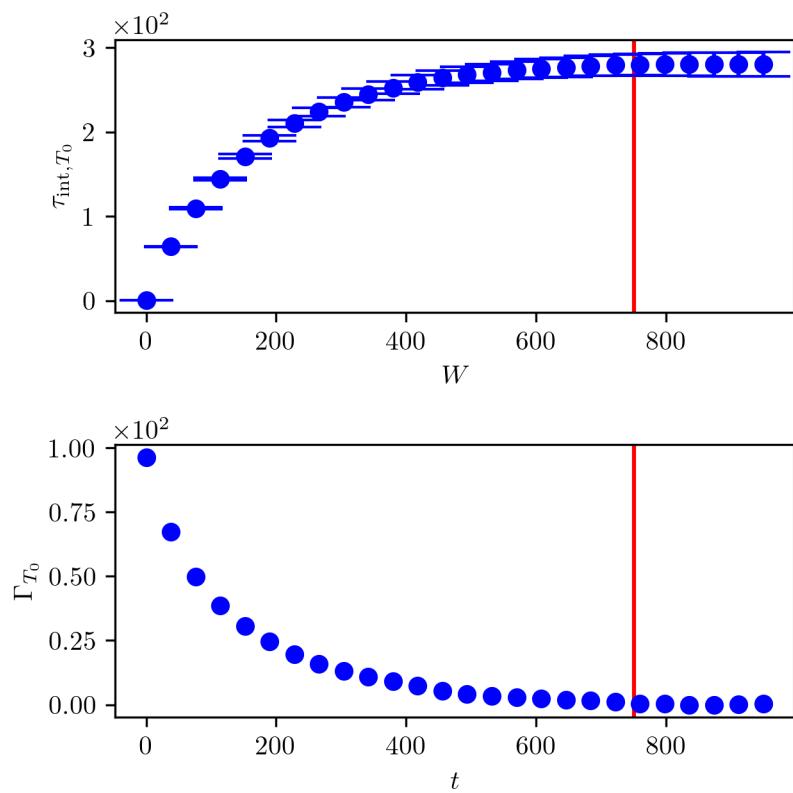
**Figure B.11:** Provided by [26], the IACT  $\tau_{\text{int},\lambda}$  at summation windows  $W$  and the estimated autocorrelation function  $\Gamma_\lambda$  at lag  $t$  of samples  $\lambda \sim \pi(\cdots | \mathbf{y})$  from the t-walkfor the approximated forward model.



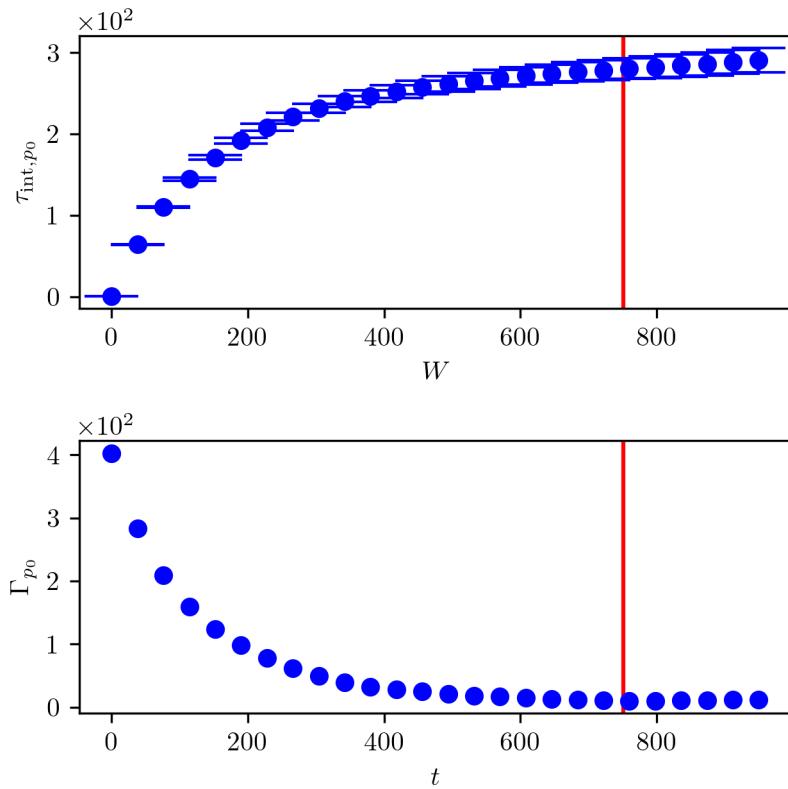
**Figure B.12:** Provided by [26], the IACT  $\tau_{\text{int},b}$  at summation windows  $W$  and the estimated autocorrelation function  $\Gamma_b$  at lag  $t$  of samples  $b \sim \pi(\cdot | \mathbf{y})$  from the t-walk for the approximated forward model.



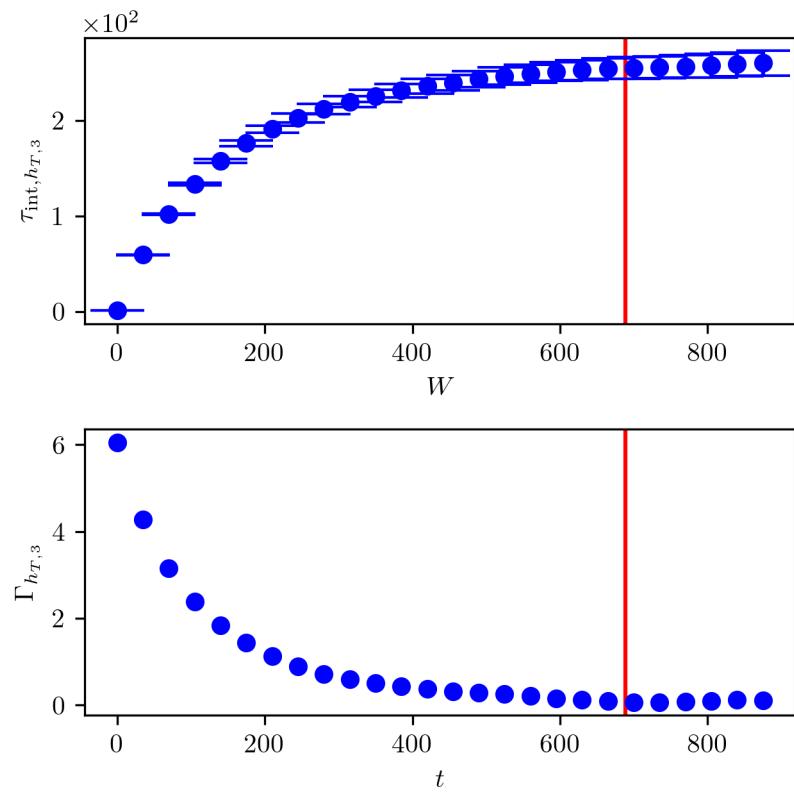
**Figure B.13:** Provided by [26], the IACT  $\tau_{\text{int},h_{T,1}}$  at summation windows  $W$  and the estimated autocorrelation function  $\Gamma_{h_{T,1}}$  at lag  $t$  of samples  $h_{T,1} \sim \pi(\cdot | \mathbf{y})$  from the t-walk for the approximated forward model.



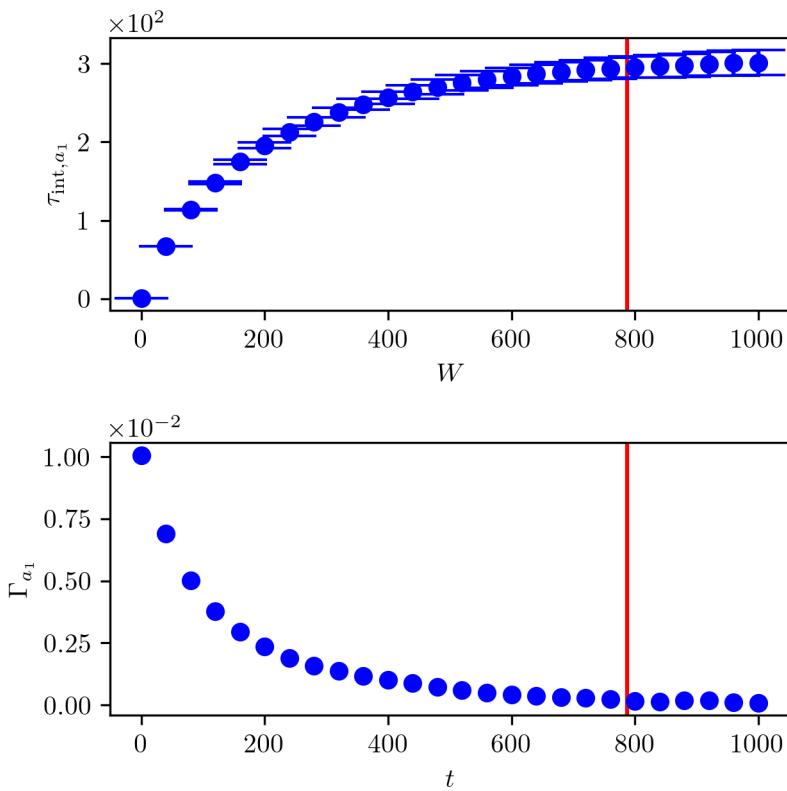
**Figure B.14:** Provided by [26], the IACT  $\tau_{\text{int}, T_0}$  at summation windows  $W$  and the estimated autocorrelation function  $\Gamma_{T_0}$  at lag  $t$  of samples  $T_0 \sim \pi(\cdot | \mathbf{y})$  from the t-walk for the approximated forward model.



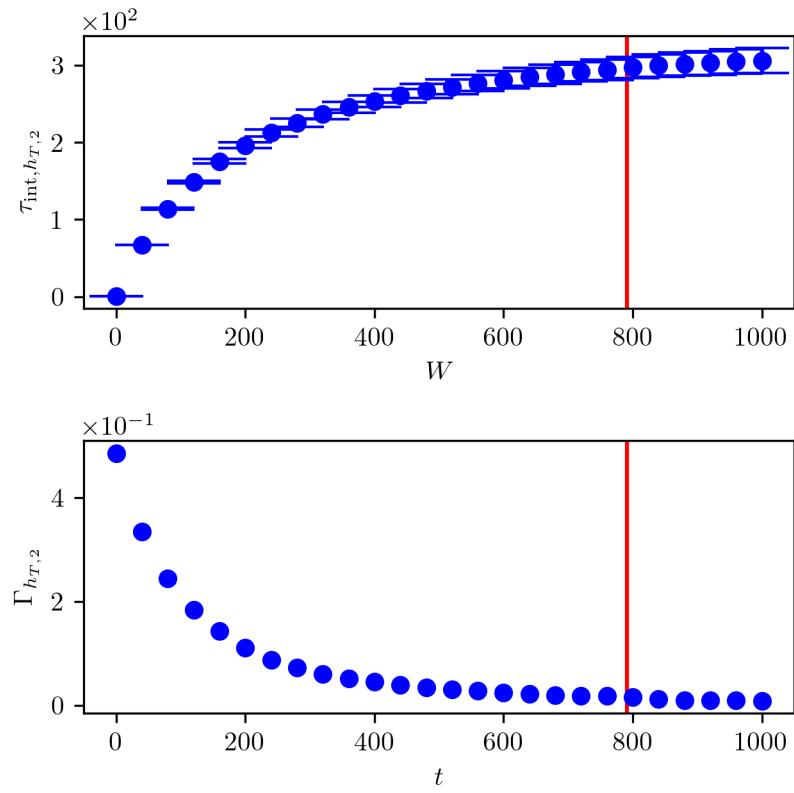
**Figure B.15:** Provided by [26], the IACT  $\tau_{\text{int},p_0}$  at summation windows  $W$  and the estimated autocorrelation function  $\Gamma_{p_0}$  at lag  $t$  of samples  $p_0 \sim \pi(\cdot|\mathbf{y})$  from the t-walk for the approximated forward model.



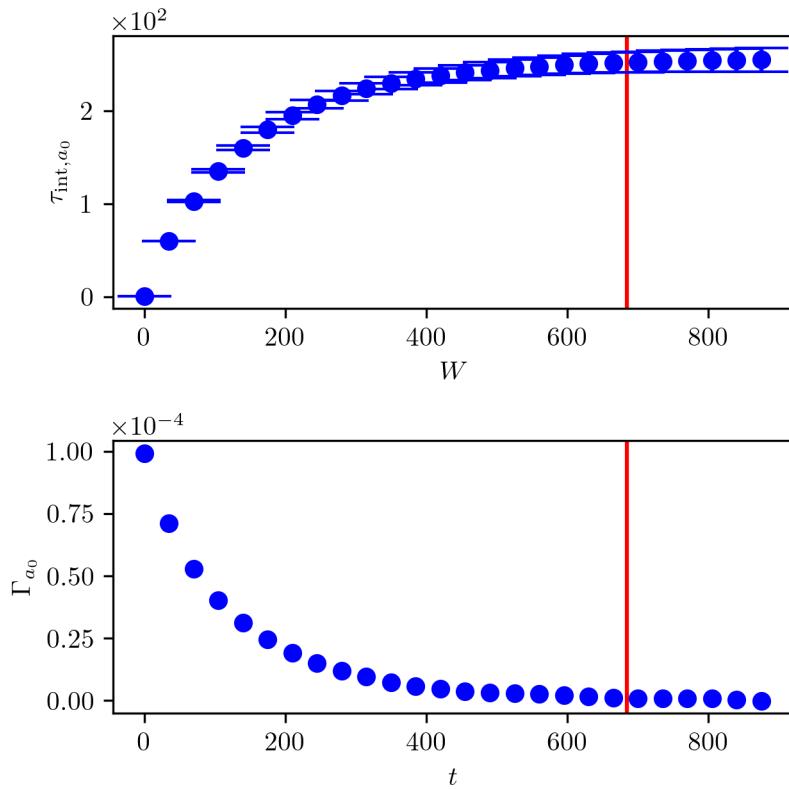
**Figure B.16:** Provided by [26], the IACT  $\tau_{\text{int}, h_{T,3}}$  at summation windows  $W$  and the estimated autocorrelation function  $\Gamma_{h_{T,3}}$  at lag  $t$  of samples  $h_{T,3} \sim \pi(\cdot | \mathbf{y})$  from the t-walk for the approximated forward model.



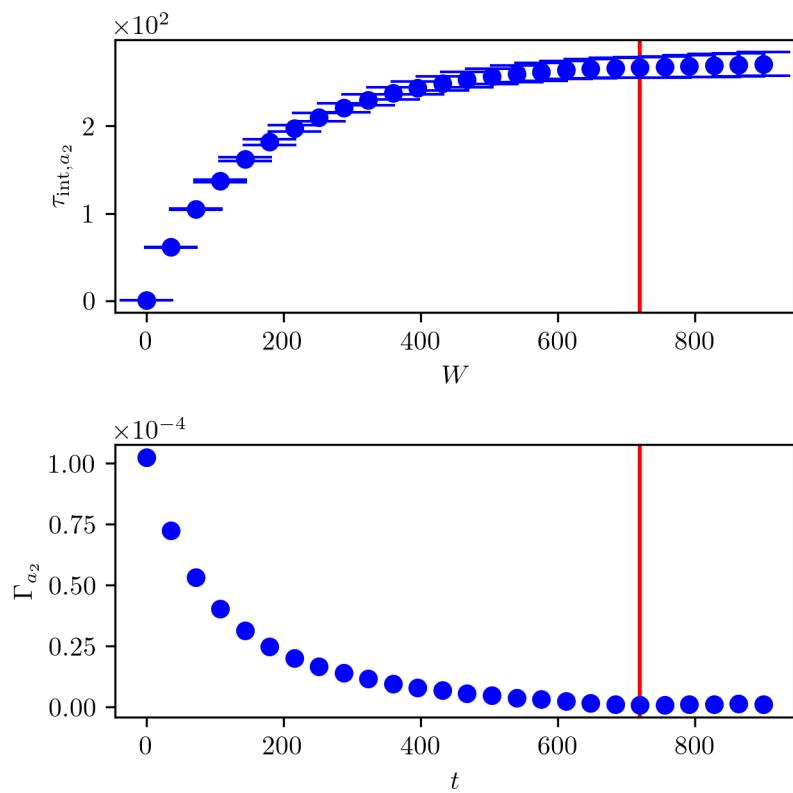
**Figure B.17:** Provided by [26], the IACT  $\tau_{\text{int},a_1}$  at summation windows  $W$  and the estimated autocorrelation function  $\Gamma_{a_1}$  at lag  $t$  of samples  $a_1 \sim \pi(\cdot|\mathbf{y})$  from the t-walk for the approximated forward model.



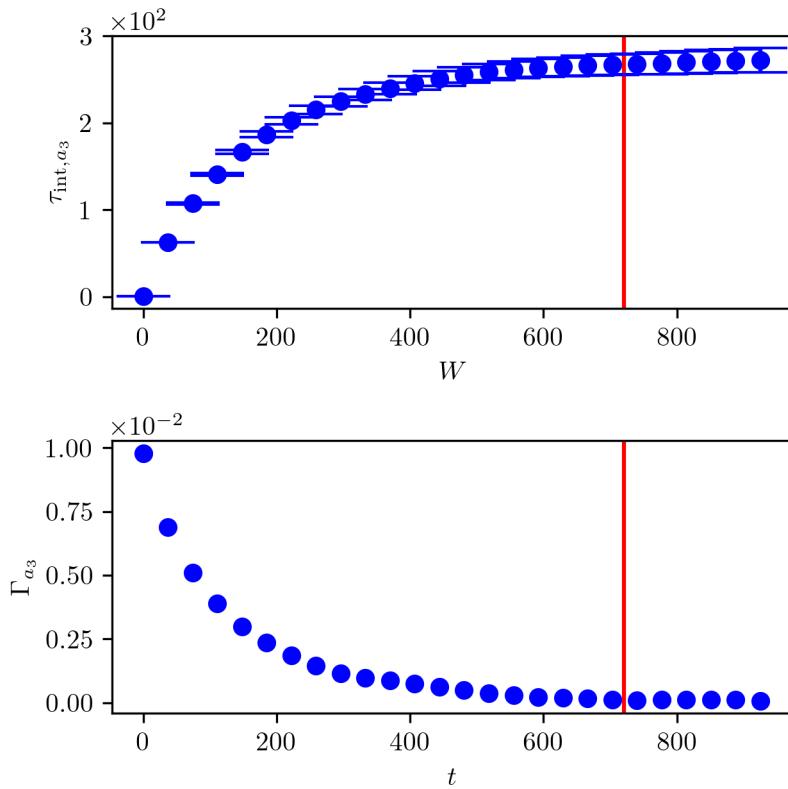
**Figure B.18:** Provided by [26], the IACT  $\tau_{\text{int}, h_{T,2}}$  at summation windows  $W$  and the estimated autocorrelation function  $\Gamma_{h_{T,2}}$  at lag  $t$  of samples  $h_{T,2} \sim \pi(\cdot \cdot \cdot | \mathbf{y})$  from the t-walk for the approximated forward model.



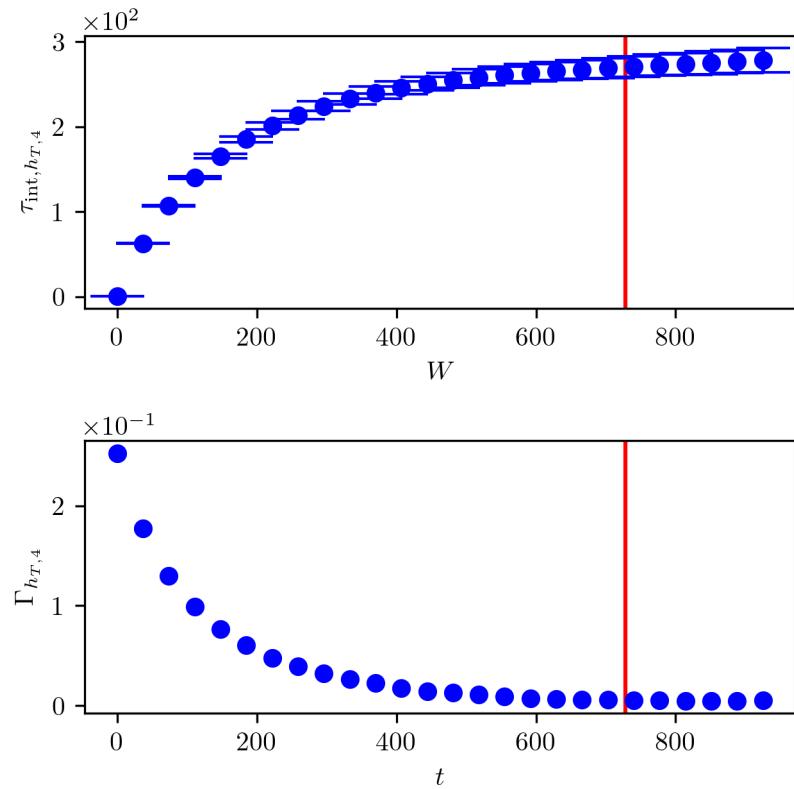
**Figure B.19:** Provided by [26], the IACT  $\tau_{\text{int},a_0}$  at summation windows  $W$  and the estimated autocorrelation function  $\Gamma_{a_0}$  at lag  $t$  of samples  $a_0 \sim \pi(\cdot|\mathbf{y})$  from the t-walk for the approximated forward model.



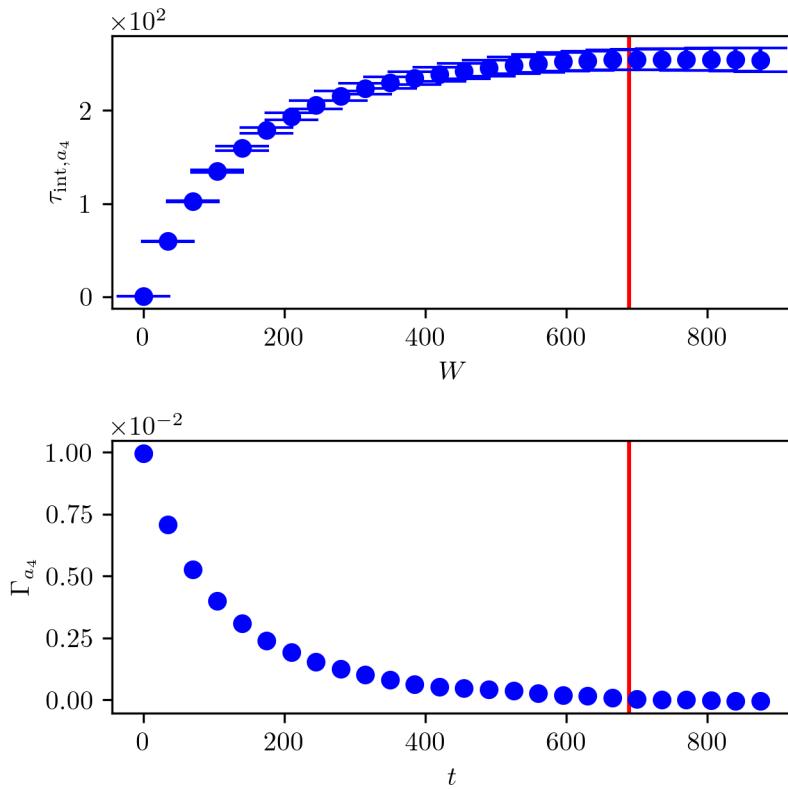
**Figure B.20:** Provided by [26], the IACT  $\tau_{\text{int},a_2}$  at summation windows  $W$  and the estimated autocorrelation function  $\Gamma_{a_2}$  at lag  $t$  of samples  $a_2 \sim \pi(\cdot | \mathbf{y})$  from the t-walk for the approximated forward model.



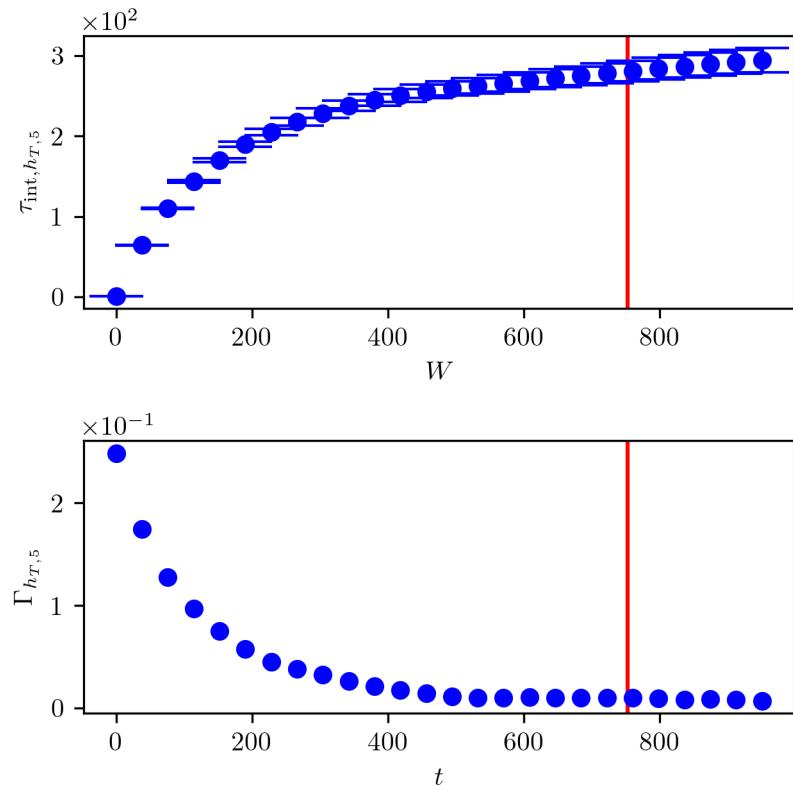
**Figure B.21:** Provided by [26], the IACT  $\tau_{\text{int},a_3}$  at summation windows  $W$  and the estimated autocorrelation function  $\Gamma_{a_3}$  at lag  $t$  of samples  $a_3 \sim \pi(\cdot|\mathbf{y})$  from the t-walk for the approximated forward model.



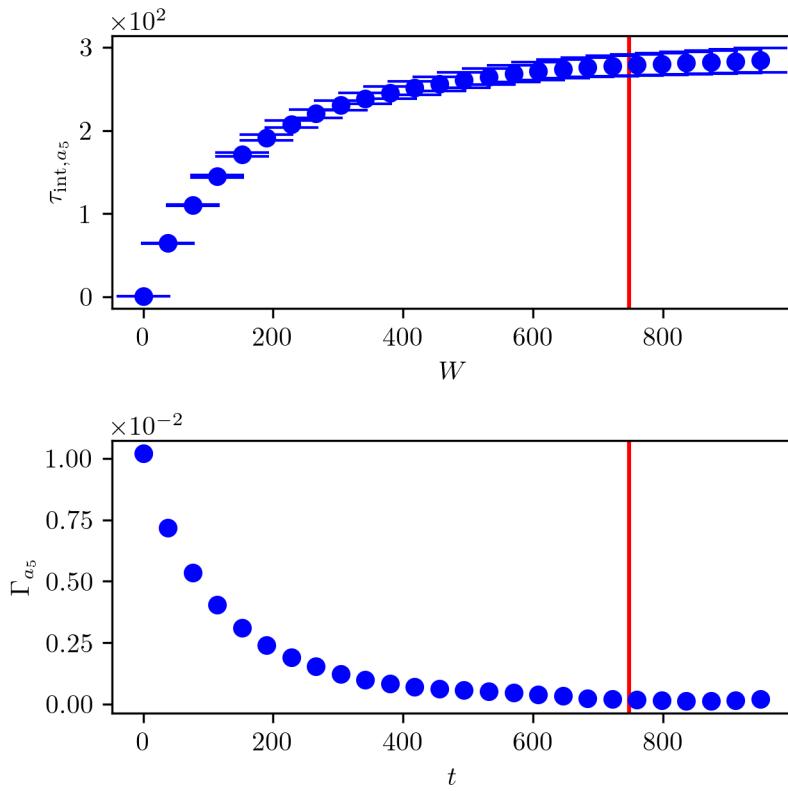
**Figure B.22:** Provided by [26], the IACT  $\tau_{\text{int}, h_{T,4}}$  at summation windows  $W$  and the estimated autocorrelation function  $\Gamma_{h_{T,4}}$  at lag  $t$  of samples  $h_{T,4} \sim \pi(\cdot | \mathbf{y})$  from the t-walk for the approximated forward model.



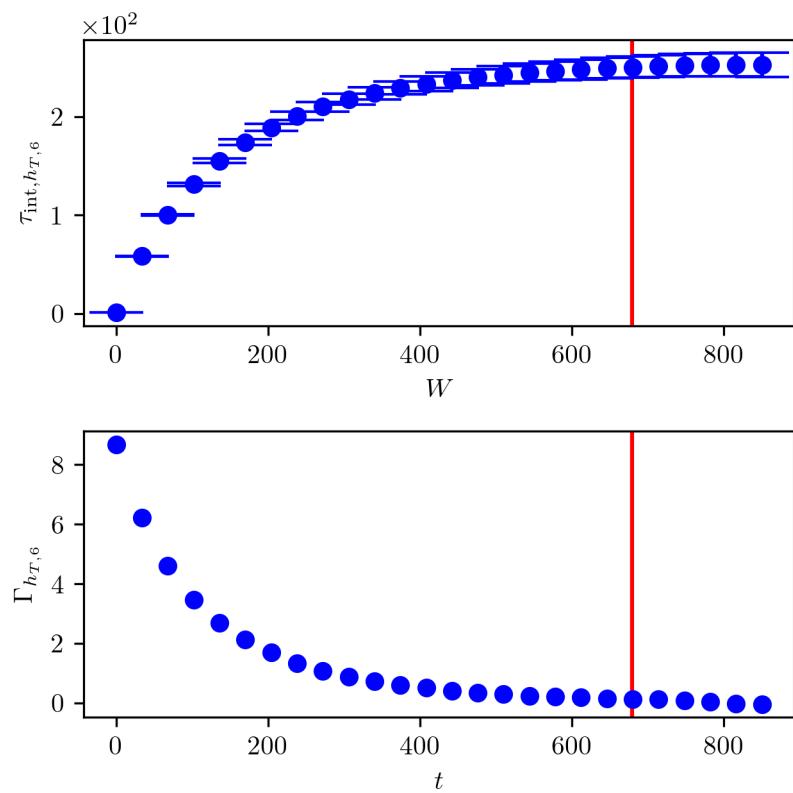
**Figure B.23:** Provided by [26], the IACT  $\tau_{\text{int}, a_4}$  at summation windows  $W$  and the estimated autocorrelation function  $\Gamma_{a_4}$  at lag  $t$  of samples  $a_4 \sim \pi(\cdot | \mathbf{y})$  from the t-walk for the approximated forward model.



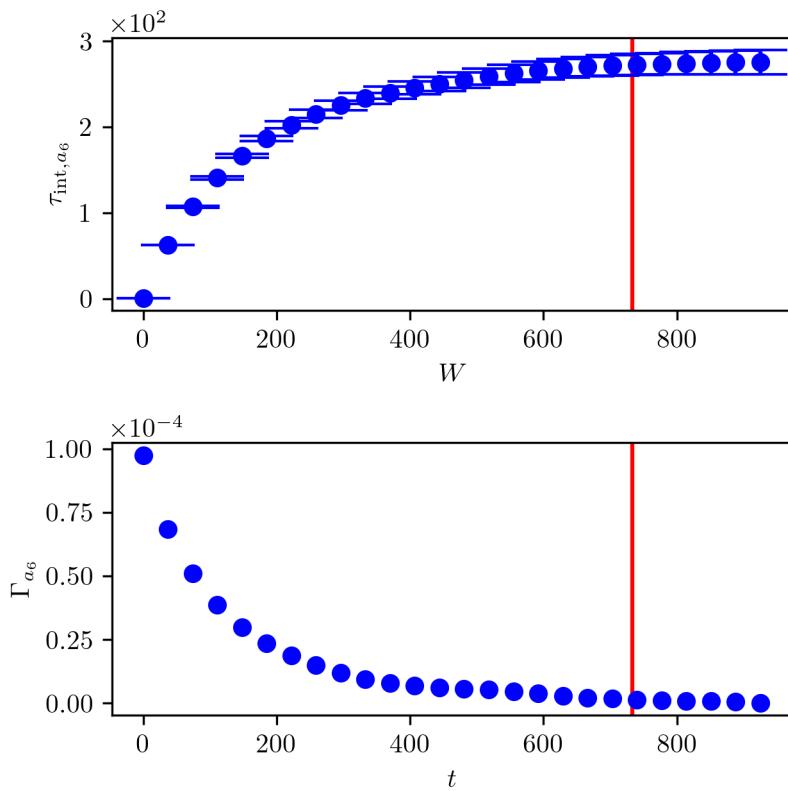
**Figure B.24:** Provided by [26], the IACT  $\tau_{\text{int}, h_{T,5}}$  at summation windows  $W$  and the estimated autocorrelation function  $\Gamma_{h_{T,5}}$  at lag  $t$  of samples  $h_{T,5} \sim \pi(\cdot | \mathbf{y})$  from the t-walk for the approximated forward model.



**Figure B.25:** Provided by [26], the IACT  $\tau_{\text{int},a_5}$  at summation windows  $W$  and the estimated autocorrelation function  $\Gamma_{a_5}$  at lag  $t$  of samples  $a_5 \sim \pi(\cdot|\mathbf{y})$  from the t-walk for the approximated forward model.



**Figure B.26:** Provided by [26], the IACT  $\tau_{\text{int}, h_{T,6}}$  at summation windows  $W$  and the estimated autocorrelation function  $\Gamma_{h_{T,6}}$  at lag  $t$  of samples  $h_{T,6} \sim \pi(\cdot | \mathbf{y})$  from the t-walk for the approximated forward model.



**Figure B.27:** Provided by [26], the IACT  $\tau_{\text{int},a_6}$  at summation windows  $W$  and the estimated autocorrelation function  $\Gamma_{a_6}$  at lag  $t$  of samples  $a_6 \sim \pi(\cdot|\mathbf{y})$  from the t-walk for the approximated forward model.