Contents

List of Figures

columnwidth 421.10046pt

1 Introduction

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

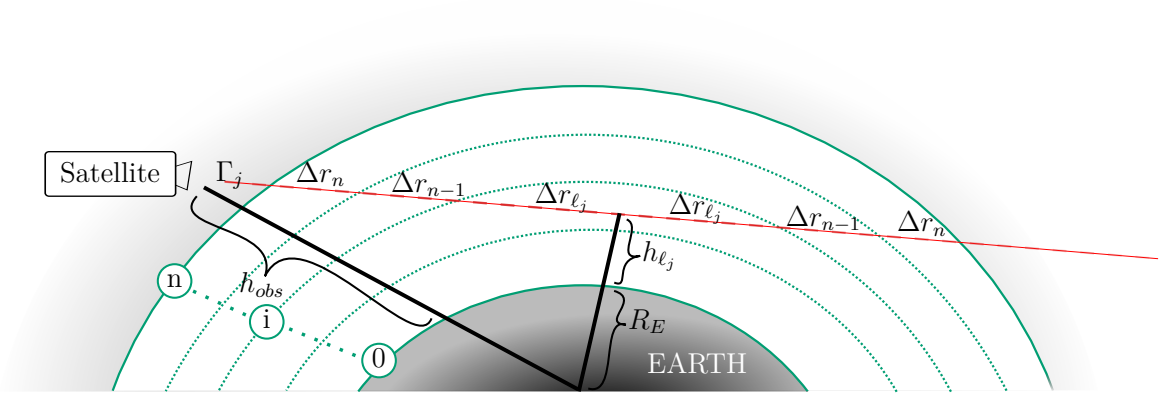
- hierachical Bayesian model, sampling to TT approx
- RTE as an example
- nonLinear to Linear Affine funciton (affine RTO)

1.2 Thesis Outline

2

Theoretical Background

2.1 Forward Model



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

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

The j^{th} measurement, taken on line of sight Γ_j is modelled by the the radiative

transfer equation (RTE) [1]

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

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

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

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

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

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

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

2.2 Affine Map

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

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

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

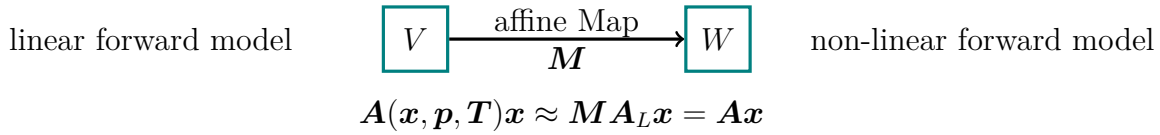


Figure 2.1: Schematics of Affine Map, which approximates the linear forward model to the non-linear forward model.

2.3 Bayesian Inference

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

We can visualize a measuremnt process through a hierarchiallly ordered directed acyclic graph (DAG), see Figure 2.2. As an observatory process naturally includes some random noise we classify the noise precision for example as hyper-parameters $\boldsymbol{\theta}$ as well as other hyperparameters influence the parmeters $\boldsymbol{x}$ detemernistaclly. Then through the forward model the parameters $\boldsymbol{x}$ are mapped onto the space of all measurables $\boldsymbol{u}$, from which we observe some data including noise as previoously described. This help us to inderstand correlation wiotin in the measiuremnt and modelling process. To inferer the underling parameters and hyperparmeters we follow the arrows backwards and set up a Bayesain hierachlly orded model.

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

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

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

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

with the noise covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, so that $\boldsymbol{\eta} \sim \mathcal{N}(0, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$ as in Eq. **??**, the prior precision matrix $\boldsymbol{Q}(\boldsymbol{\theta})$, prior mean $\boldsymbol{\mu}$ and some prior distribution over the hyper-parameters $\pi(\boldsymbol{\theta})$. Trhough the prior distributions $\pi(\boldsymbol{x}|\boldsymbol{y})$ and $\pi(\boldsymbol{\theta})$., we can incorpate functional depende ncies as well physical propetires. The likelhood function $\pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})$ includes information about how the parameters and hyper-parameters fit to the data according to our forward model and including the measurement process.

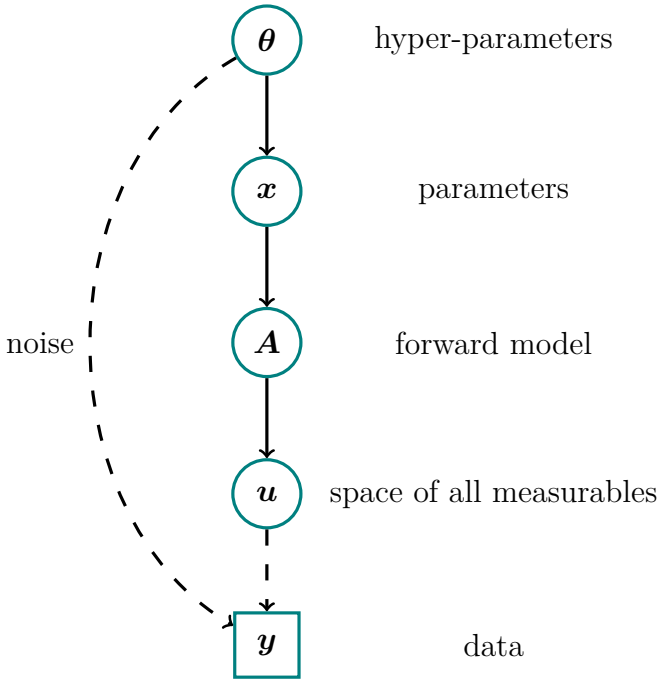


Figure 2.2: The directed acyclic graph (DAG) for a typical linear inverse problem visualises forward dependencies as solid line arrows for deterministic dependencies and dotted arrows for statistical dependencies. Naturally the data $\boldsymbol{y}$ has some noise described through included in some hyper-parameters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

With a normally distributed prior and likelihood function this becomes a linear-Gaussian Bayesian hierarchical model. For more detailed byesiasn analys we recoomend [**<empty citation>**].

According to Bayes' theorem we focus on the posterior distribution

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

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

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

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

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

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

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

2.3.1 Marginal and then Conditional

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

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

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

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

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

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

with

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

See lemma [**<empty citation>**].

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

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

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

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

with weights given by $\pi(\boldsymbol{\theta}|\boldsymbol{y})$. [**<empty citation>**] Note that the noise covariance $\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})$ and the prior precision $\boldsymbol{Q} = \boldsymbol{Q}(\boldsymbol{\theta})$ are dependent of hyper-parameters $\boldsymbol{\theta}$.

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

2.4 Sampling Methods

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

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

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

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

2.4.1 Metropolis

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

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

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

or reject and set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)}$. To assure convergence independent of the intial sample we discard samples after the so-called burn-in period and effectively generate a Markov-Chain of length $N - N_{\text{burn-in}}$.

Algorithm 1: Metropolis

1: Initialize $\boldsymbol{\theta}^{(0)}$
2: **for** $k = 1, \ldots, N$ **do**
3: Propose $\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta}^{(t-1)}, \boldsymbol{\Sigma})$
4: Compute
$$\alpha(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t-1)}) = \min\left\{1, \frac{\pi(\boldsymbol{\theta}|\boldsymbol{y})}{\pi(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{y})}\right\}$$
5: Draw $u \sim \mathcal{U}(0, 1)$
6: **if** $\alpha \geq u$ **then**
7: Accept and set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}$
8: **else**
9: Reject and set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)}$
10: **end if**
11: **end for**
12: Output: $\boldsymbol{\theta}^{(0)}, \ldots, \boldsymbol{\theta}^{(k)}, \ldots, \boldsymbol{\theta}^{(N)} \sim \pi(\boldsymbol{\theta}|\boldsymbol{y})$

Since using a symmetric proposal distribution, Metropolis chains are reversible and as you can see in $\alpha(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t-1)})$ converge towards $\pi(\boldsymbol{\theta}|\boldsymbol{y})$. A more mathematical prove of ergodicity can be found in [**<empty citation>**].

2.4.2 Gibbs

Each transition follows a specific rule: selecting one variable and sampling it from its conditional distribution while keeping the others fixed.

Algorithm 2: Gibbs

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

$\boldsymbol{\theta}_{<j} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{j-1}\}$ $\boldsymbol{\theta}_{>j} = \{\boldsymbol{\theta}_{j+1}, \ldots, \boldsymbol{\theta}_d\}$which denotes all dimesnions except $\boldsymbol{\theta}_j$

2.4.3 t-walk

We use the t-walk sampler as a black box sampling algorithm [**<empty citation>**], so it prodices a ergodic markov chain

2.4.4 Draw a sample from the Conditional posterior distribution – RTO

In the case of marginalising out the latent field we use a smaple of choice to genrate a amrkov chain. Then we can draw a sample from that markov chain and condition on it so that we can draw a sample from the As the full conditional distribution for $\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}$ is a normal distribution we can rewrite to:

$$\pi(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}) \propto \pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta}) \tag{2.16}$$

$$= \exp\|\hat{\boldsymbol{A}}\boldsymbol{x} - \hat{\boldsymbol{y}}\|^2 , \tag{2.17}$$

where

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

One sample from the posterior can be computed by minimizing the following equation with respect to $\hat{\boldsymbol{x}}$:

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

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

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

where $\boldsymbol{v}_1 \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{A})$ and $\boldsymbol{v}_2 \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q})$ are independent random variables. Finally, we can draw an independent sample from the posterior $(\boldsymbol{x}_i, \boldsymbol{\theta}_i) \sim \pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})$.

2.5 Numerical Approxiamtion Methods

To approximate multi dimesnioan functions we can use the tensor train format to approximate marginal funcitons

2.5.1 Tensor Train

One of the main contributions of this paper is to show that the conditional distribution method is feasible, and efficient, once a PDF has been put into TT format. This section presents those calculations.

First, we describe the computation of the marginal PDFs p_k, defined in (2), given π in a TT format (3). Note that integrals over the variable x_k appear in all conditionals (2) with $k < d$. The TT format allows computing the $r_{k-1} \times 1$ vector P_k required for evaluating the marginal PDF p_{k-1} by the following algorithm:

Algorithm 3: Computation of marginal PDFs

1: Initialize $P_{d+1} = 1$.
2: **for** $k = d, d-1, \ldots, 2$ **do**
3: Compute

$$(P_k)_{\alpha_{k-1}} = \sum_{\alpha_k=1}^{r_k} \left(\int_{\mathbb{R}} \pi^{(k)}_{\alpha_{k-1},\alpha_k}(x_k)\, dx_k \right) (P_{k+1})_{\alpha_k}$$

4: **end for**

Since $\pi^{(k)}(x_k) \in \mathbb{R}^{r_{k-1} \times r_k}$ for each fixed x_k, the integral $\int \pi^{(k)}(x_k)dx_k$ is a $r_{k-1} \times r_k$ matrix, where α_{k-1} is the row index and α_k is the column index. Hence, we can write Line 3 as the matrix–vector product:

$$P_k = \left(\int_{\mathbb{R}} \pi^{(k)}(x_k) dx_k \right) P_{k+1}.$$

Assuming n quadrature points for each x_k, and the uniform rank bound $r_k \leq r$, the asymptotic complexity of this algorithm is $O(dnr^2)$.

The first marginal PDF is approximated by $p_1^*(x_1) = |\pi^{(1)}(x_1)P_2|$. We take the absolute value because the TT approximation π^* (and hence, $\pi^{(1)}(x_1)P_2$) may be negative at some locations. In the k-th step of the sampling procedure, the marginal PDF also requires the first $k-1$ TT blocks, restricted to the components of the sample that are already determined:

$$p_k^*(x_1, \ldots, x_{k-1}|x_k) = \left| \pi^{(1)}(x_1) \cdots \pi^{(k-1)}(x_{k-1}) \pi^{(k)}(x_k) P_{k+1} \right|.$$

However, since the loop goes sequentially from $k = 1$ to $k = d$, the sampled TT blocks can be accumulated in the same fashion as the integrals P_k. Again, we take the absolute value to ensure positivity.

The **full** marginals are then defined as:

$$p_k^*(x_k) = \left| \left(\int_{\mathbb{R}} \pi^{(1)}(x_1) dx_1 \right) \cdots \left(\int_{\mathbb{R}} \pi^{(k-1)}(x_{k-1}) dx_{k-1} \right) \pi^{(k)}(x_k) \left(\int_{\mathbb{R}} \pi^{(k+1)}(x_{k+1}) dx_{k+1} \right) \cdots \left(\int_{\mathbb{R}} \pi^{(d)}(\right.\right.$$

2.6 SIRT - Marginal Functions and Conditional PDFs

We represent each TT core of the decomposition in (18) as

$$G_k^{(\alpha_{k-1}, \alpha_k)}(x_k) = \sum_{i=1}^{n_k} \phi_k^{(i)}(x_k) \boldsymbol{A}_k[\alpha_{k-1}, i, \alpha_k], \quad k = 1, ..., d, \tag{2.21}$$

where $\{\phi_k^{(i)}(x_k)\}_{i=1}^{n_k}$ is the set of basis functions for the k-th coordinate and $\boldsymbol{A}_k \in \mathbb{R}^{r_{k-1} \times n_k \times r_k}$ is the associated k-th coefficient tensor. For the k-th set of basis functions, we define the mass matrix $\boldsymbol{M}_k \in \mathbb{R}^{n_k \times n_k}$ by

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

2.7 right to left

Proposition 1 *Starting with the last coordinate $k = d$, we set $\boldsymbol{B}_d = \boldsymbol{A}_d$. Suppose for the first k dimensions ($d > k \geq 1$), we have a coefficient tensor $\boldsymbol{B}_k \in \mathbb{R}^{r_{k-1} \times n_k \times r_k}$ that defines a marginal function $\pi_{\leq k}(x_{\leq k})$ as in (??). The following procedure **can** be used to obtain the coefficient tensor $\boldsymbol{B}_{k-1} \in \mathbb{R}^{r_{k-2} \times n_{k-1} \times r_{k-1}}$ for defining the next marginal function $\pi_{<k}(x_{<k})$:*

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

$$\boldsymbol{C}_k[\alpha_{k-1}, \tau, l_k] = \sum_{i=1}^{n_k} \boldsymbol{B}_k[\alpha_{k-1}, i, l_k] \boldsymbol{L}_k[i, \tau]. \tag{2.23}$$

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

$$\boldsymbol{Q}_k \boldsymbol{R}_k = (\boldsymbol{C}_k^{(R)})^\top. \tag{2.24}$$

3. *Compute the new coefficient tensor:*

$$\boldsymbol{B}_{k-1}[\alpha_{k-2}, i, l_{k-1}] = \sum_{\alpha_{k-1}=1}^{r_{k-1}} \boldsymbol{A}_{k-1}[\alpha_{k-2}, i, \alpha_{k-1}] \boldsymbol{R}_k[l_{k-1}, \alpha_{k-1}]. \tag{2.25}$$

2.8 left to right

Proposition 2 *Starting with the first coordinate $k = 1$, we set $\boldsymbol{B}_{pre,1} = \boldsymbol{A}_1$. Suppose for the last k dimensions ($k < d$), we have a coefficient tensor $\boldsymbol{B}_{pre,k} \in \mathbb{R}^{r_{k-1} \times n_k \times r_k}$ that defines a marginal function $\pi_{\leq k}(x_{\leq k})$ as in (??). The following procedure **can** be used to obtain the coefficient tensor $\boldsymbol{B}_{pre,k+1} \in \mathbb{R}^{r_k \times n_{k+1} \times r_{k+1}}$ for defining the next marginal function $\pi_{>k}(x_{>k})$:*

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

$$\boldsymbol{C}_{pre,k}[\alpha_{k-1}, \tau, l_k] = \sum_{i=1}^{n_k} \boldsymbol{L}_k[i, \tau] \boldsymbol{B}_{pre,k}[\alpha_{k-1}, i, l_k]. \tag{2.26}$$

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

$$\boldsymbol{Q}_{pre,k}\boldsymbol{R}_{pre,k} = (\boldsymbol{C}^{(R)}_{pre,k}). \tag{2.27}$$

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

$$\boldsymbol{B}_{pre,k+1}[l_{k+1}, i, \alpha_{k+1}] = \sum_{\alpha_k=1}^{r_k} \boldsymbol{R}_{pre,k}[l_{k+1}, \alpha_k]\boldsymbol{A}_{k+1}[\alpha_k, i, \alpha_{k+1}]. \tag{2.28}$$

2.9 Calc Marginals

Proposition 3 *The marginal PDF of* X_1 *can be expressed as*

$$f_{X_1}(x_1) = \frac{1}{z}\left(\gamma \prod_{i=2}^{d} \lambda_i(X_i) + \sum_{l_1=1}^{r_1}\left(\sum_{i=1}^{n_1} \phi_1^{(i)}(x_1)\boldsymbol{D}_1[i, l_1]\right)^2\right)\lambda_1(x_1), \tag{2.29}$$

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

The marginal PDF of X_n *can be expressed as*

$$f_{X_n}(x_n) = \frac{1}{z}\left(\gamma \prod_{i=1}^{d-1} \lambda_i(X_i) + \sum_{l_{n-1}=1}^{r_{n-1}}\left(\sum_{i=1}^{n_1} \phi_1^{(i)}(x_1)\boldsymbol{D}_n[l_{n-1}, i]\right)^2\right)\lambda_n(x_n), \tag{2.30}$$

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

The marginal PDF of X_k *can be expressed as*

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

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

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

Appendices

References

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

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