Contents

List of Figures

columnwidth 421.10046pt

1 Introduction

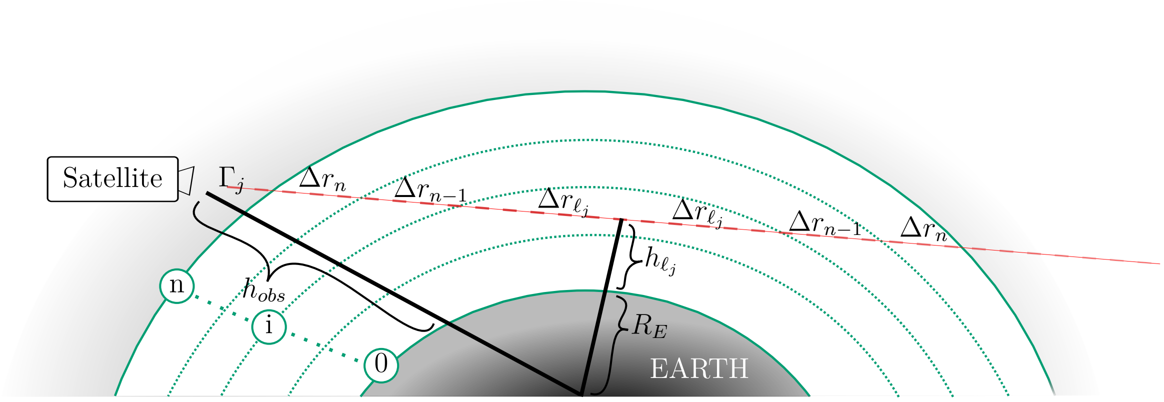
1.1 What is going on?, 3 facts, What is new in this thesis?

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

1.2 Thesis Outline

2
Theoretical and Technical Background

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 $h_{i-1} < h_i$, $i = 1, \dots, n$, 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 $\boldsymbol{x} = \{x_1, x_2, \dots, x_n\} \in \mathbb{R}^n$ or the fraction of pressure and temperature $\boldsymbol{p}/\boldsymbol{T} = \{p_1/T_1, p_2/T_2, \dots, p_n/T_n\} \in \mathbb{R}^n$, 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 $\boldsymbol{\eta} \in \mathbb{R}^{m}$.

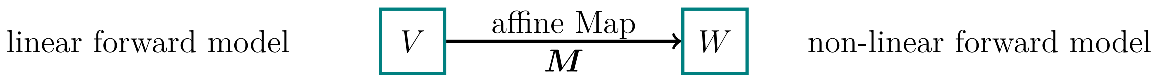
Since the absorption $\tau(r)$ reduces measurements by of order 1%, or less, making the inverse problem only weakly non-linear. We use that to approximate the non-linear forward model $\boldsymbol{A}(\boldsymbol{x}, \boldsymbol{p}, \boldsymbol{T})$ with a map $\boldsymbol{M}$ so that $\boldsymbol{A}(\boldsymbol{x}, \boldsymbol{p}, \boldsymbol{T}) \approx \boldsymbol{M}\boldsymbol{A}_L$ 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[]. 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
$V = \left\{ \boldsymbol{A}(\boldsymbol{x}^{(1)}, \boldsymbol{p}, \boldsymbol{T}), \dots, \boldsymbol{A}(\boldsymbol{x}^{(m)}, \boldsymbol{p}, \boldsymbol{T}) \right\}$ and $W = \left\{ \boldsymbol{A}\boldsymbol{x}^{(1)}, \dots, \boldsymbol{A}\boldsymbol{x}^{(m)} \right\}$ over the same field, with fixed $\boldsymbol{p}, \boldsymbol{T}$. The parameter $\boldsymbol{x}$ is distributed as the so-called posterior distribution $\left\{ \boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(m)} \right\} \sim \pi(\boldsymbol{x} | \boldsymbol{\theta}, \boldsymbol{y})$, with hyper-parameters $\boldsymbol{\theta}$, according to a Bayesian hierarchical model.



$$\boldsymbol{A}(\boldsymbol{x},\boldsymbol{p},\boldsymbol{T})\boldsymbol{x} \approx \boldsymbol{M}\boldsymbol{A}_L\boldsymbol{x} = \boldsymbol{A}\boldsymbol{x}$$

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

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

based on a linear forward model $\boldsymbol{A}$ and some noise $\boldsymbol{\eta}$. Later in section **??** we set up a more sophisticated Bayesian framework applied to the forward model in section **??**.

We can visualise the correlation structure of a measurement process through a hierarchiallly ordered directed acyclic graph (DAG), see Figure 2.2. As an observatory process naturally includes some random noise we include that in our DAG and classify the noise as a hyper-parameter in $\boldsymbol{\theta}$. Other hyper-parameters influence the parmeters $\boldsymbol{x}$ detemernistaclly, which are then mapped through the forward model onto the space of all measurables $\boldsymbol{u}$, from which we observe some data $\boldsymbol{y}$ including noise as previously mentioned. Drawing a DAG can help us to dependences within the measurement and modelling process. Given some data we inferer the distribution of the underling parameters and hyper-parameters by following the arrows in Figure **??** backwards and set up a Bayesain hierachlly ordered 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 $\boldsymbol{\theta}$.

$$\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{A}\boldsymbol{x}, \boldsymbol{\Sigma}(\boldsymbol{\theta})) \tag{2.5a}$$

$$\boldsymbol{x}|\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{Q}^{-1}(\boldsymbol{\theta})) \tag{2.5b}$$

$$\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})\,, \tag{2.5c}$$

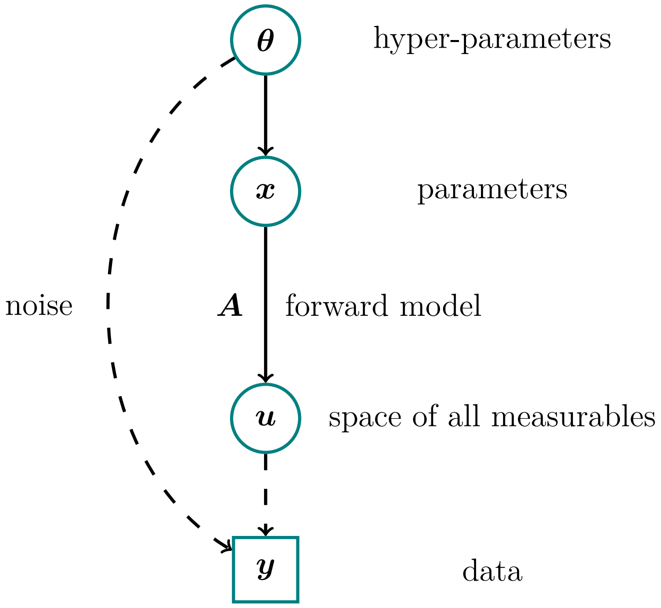


Figure 2.2: The directed acyclic graph (DAG) for a typical linear inverse problem visualises forward dependencies as solid line arrows for deterministic dependencies and dotted arrows for statistical dependencies. Naturally the data $\boldsymbol{y}$ has some noise described through included in some hyper-parameters $\boldsymbol{\theta}$. 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 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 $\pi(\boldsymbol{\theta})$. Through sensibly choosing the prior distributions $\pi(\boldsymbol{x}|\boldsymbol{y})$ as well as the hyper-parameters $\boldsymbol{\theta}$ and their prior distribution $\pi(\boldsymbol{\theta})$, we can incorporate functional dependencies as well physical properties of the parameters. The likelhood function $\pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})$ is a measure of how the parameters and hyper-parameters fit to the data according to our forward model, including information of the measurement process.

With a normally distributed prior and likelihood function this becomes a linear-Gaussian Bayesian hierarchical model. For more detailed Bayesian analysis we recommend [].

The posterior distribution, the function of interest,

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

is given according to Bayes' theorem [], 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

$$\pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y}) \propto \pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})\pi(\boldsymbol{x}, \boldsymbol{\theta}) . \tag{2.7}$$

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

$$\mathrm{E}_{\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y}}[h(\boldsymbol{x})] = \int \int h(\boldsymbol{x}, \boldsymbol{\theta}) \, \pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y}) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{\theta} , \tag{2.8}$$

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

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

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

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

$$\propto \sqrt{\frac{\det(\boldsymbol{\Sigma}^{-1}) \, \det(\boldsymbol{Q})}{\det(\boldsymbol{Q} + \boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{A})}} \times \exp\Big[-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\mu})^T \boldsymbol{Q}_{\boldsymbol{\theta}|\boldsymbol{y}}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\mu})\Big]\pi(\boldsymbol{\theta}) , \tag{2.11}$$

with

$$\boldsymbol{Q}_{\boldsymbol{\theta}|\boldsymbol{y}} = \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{A} + \boldsymbol{Q})^{-1}\boldsymbol{A}^T\boldsymbol{\Sigma}^{-1} \,. \tag{2.12}$$

See lemma [].

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

$$\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta} \sim \mathcal{N}\Big(\boldsymbol{\mu} + (\boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{A} + \boldsymbol{Q})^{-1}\boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\mu}), (\boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{A} + \boldsymbol{Q})^{-1}\Big)\,, \tag{2.13}$$

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

$$\mathrm{E}_{\boldsymbol{x}|\boldsymbol{y}}[h(\boldsymbol{x})] = \int \mathrm{E}_{\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y}}[h(\boldsymbol{x})]\, \pi(\boldsymbol{\theta}|\boldsymbol{y})\, \mathrm{d}\boldsymbol{\theta} \,, \tag{2.14}$$

with weights given by $\pi(\boldsymbol{\theta}|\boldsymbol{y})$. [] Note that the noise covariance $\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})$ and the prior precision $\boldsymbol{Q} = \boldsymbol{Q}(\boldsymbol{\theta})$ are depending on 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 Regularisation

Another method to find a solution to a linear inverse problem as in Eq. **??** is to find a solution $\boldsymbol{x}_\lambda$ accdoring to a data misfit norm and a regularisation semi-norm []. We will discuss the case of Tikhonov regularisation.

For a parameter $\boldsymbol{x}$ a linear forward model and some data $\boldsymbol{y}$ the data misfit norm is defined as

$$\left\| \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x} \right\| \,. \tag{2.15}$$

The regularisation semi norm

$$\lambda \left\| \boldsymbol{T}\boldsymbol{x} \right\| \tag{2.16}$$

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

$$\boldsymbol{x}_\lambda = \arg\min_{\boldsymbol{x}} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|^2 + \lambda \|\boldsymbol{T}\boldsymbol{x}\|^2 \tag{2.17}$$

can be found by the derivativ

$$\begin{aligned} & \nabla_{\boldsymbol{x}} \Big\{ (\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}_\lambda)^T (\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}_\lambda) + \lambda \boldsymbol{x}_\lambda^T \boldsymbol{T}^T \boldsymbol{T} \boldsymbol{x}_\lambda \Big\} = 0 & (2.18) \\ \iff \quad & \nabla_{\boldsymbol{x}} \Big\{ \boldsymbol{y}^T \boldsymbol{y} + \boldsymbol{x}_\lambda^T \boldsymbol{A}^T \boldsymbol{A} \boldsymbol{x}_\lambda - \boldsymbol{y}^T \boldsymbol{A} \boldsymbol{x}_\lambda - \boldsymbol{x}_\lambda^T \boldsymbol{A}^T \boldsymbol{y} + \lambda \boldsymbol{x}_\lambda^T \boldsymbol{T}^T \boldsymbol{T} \boldsymbol{x}_\lambda \Big\} = 0 & (2.19) \\ \iff \quad & 2\boldsymbol{A}^T \boldsymbol{A} \boldsymbol{x}_\lambda - 2\boldsymbol{A}^T \boldsymbol{y} + 2\lambda \boldsymbol{T}^T \boldsymbol{T} \boldsymbol{x}_\lambda = 0 & (2.20) \end{aligned}$$

Then a regularised solution is given as:

$$\boldsymbol{A}^T \boldsymbol{y} (\boldsymbol{A}^T \boldsymbol{A} + \lambda \boldsymbol{L})^{-1} = \boldsymbol{x}_\lambda, \tag{2.21}$$

where we can set $\boldsymbol{T}^T \boldsymbol{T} = \boldsymbol{L}$, which is typically, banded matrix approximation to the nth derivative[]. To find the best regularised solution $\boldsymbol{x}_\lambda$ is calculated for a range of λ so thate the regularised solution minimizes the data-misfit as well as the regularised semi-norm, which can be done via a L-Curve. We will further specify $\boldsymbol{L}$ in section **??** as well as plot an L-Curve.

2.5 Sampling Methods

In this chapter we present the sampling methods used in this thesis and also argue why we use sampling methods to calculate the expectation

$$\mathrm{E}_{\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y}}[h(\boldsymbol{x},\boldsymbol{\theta})] = \underbrace{\int\int h(\boldsymbol{x},\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} \, \mathrm{d}\boldsymbol{x}}_{\boldsymbol{\mu}_{\text{int}}} \tag{2.22}$$

of a function $h(\boldsymbol{x},\boldsymbol{\theta})$ with respect to a probability density $\pi(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})$. In the case where the calculation of the integral is not feasible we can approximate Eq. **??** with a sample based estimate

$$\mathrm{E}_{\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y}}[h(\boldsymbol{x},\boldsymbol{\theta})] \approx \underbrace{\frac{1}{N} \sum_{k=1}^{N} h(\boldsymbol{x}^{(k)},\boldsymbol{\theta}^{(k)})}_{\boldsymbol{\mu}_{\text{samp}}}, \tag{2.23}$$

for large enough samples size N of a sample set $\mathcal{M} = \{(\boldsymbol{x}, \boldsymbol{\theta})^{(1)}, \dots, (\boldsymbol{x}, \boldsymbol{\theta})^{(k)}, \dots, (\boldsymbol{x}, \boldsymbol{\theta})^{(N)}\} \sim \pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})$. The variance of this unbiased estimator is $\mathcal{O}(1/N)$ so large samples sizes can reduce the uncertainty of $\boldsymbol{\mu}_{samp}$ []. see chapter 5 in Ridley and Roberts 2004 We can do this as the central limit theorem states that the samples means $\boldsymbol{\mu}_{samp}^{(i)}$, of samples sets $\mathcal{M}_i$ for $i = 1, \dots, n$ of any distribution , converge in distribution to a normal distribution so that

$$\sqrt{n}(\boldsymbol{\mu}_{\text{samp}}^{(i)} - \boldsymbol{\mu}_{\text{int}}) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2), \tag{2.24}$$

if $\sigma^2 < \infty$ with a bound error $\boldsymbol{\mu}_{\text{samp}}^{(i)} - \boldsymbol{\mu}_{\text{int}}$.

For us it is now important to show that we the methods we use draw samples from a target distribution so that $\mathcal{M} = \{(\boldsymbol{x}, \boldsymbol{\theta})^{(1)}, \dots, (\boldsymbol{x}, \boldsymbol{\theta})^{(k)}, \dots, (\boldsymbol{x}, \boldsymbol{\theta})^{(N)}\} \sim \pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})$ and that we then use sample based estimates as in Eq. **??**. In doing so we will use Markov-Chain Monte Carlo (MCMC) methods, where we hava to show that the produces Markov-Chain $\mathcal{M}$ is ergodic. The ergodicy theorem states that, if an aperiodic and irreducible Markov chain $\mathcal{M}$ is reversible **??** then it converges towards a stationnary distrbution unique equilibirum distribution $\pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})$. 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 the chain converges and we can use sample based estimates. In practise one can look at the trace $\pi(\boldsymbol{x}^{(k)}, \boldsymbol{\theta}^{(k)}|\boldsymbol{y})$ for $k = 1, \dots, N$ of the samples and eyeball ergodicity.

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

2.5.1 Metropolis- within Gibbs sampling

As introduced in the MTC sektion we will sample from $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ seperatly from $\pi(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$. To sample from $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ we use a Metropolis-within-Gibbs sampler and introduce this for the 2 dimesnional case, as this is what we deal with in this tehsis, where $\boldsymbol{\theta} = (\theta_1, \theta_2)$, we do a metroplis step in θ_1 deirection and a gibbs stepp in θ_2 direction.

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

$$\alpha(\theta_1|\theta_1^{(t-1)}) = \min\left\{1, \frac{\pi(\theta_1|\theta_2^{(t-1)}, \boldsymbol{y})q(\theta_1^{(t-1)}|\theta_1)}{\pi(\theta_1^{(t-1)}|\theta_2^{(t-1)}, \boldsymbol{y})q(\theta_1|\theta_1^{(t-1)})}\right\} \tag{2.25}$$

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

The we do a gibbs step in θ_2 direction, where Gibbs sampling is a special case of the metropolis hastings algorihtm with the acceptance probaility of 1. The next samples $\theta_2^{(t)} \sim \pi(\cdot|\theta_1^{(t)}, \boldsymbol{y})$ conditioned on $\theta_1^{(t)}$ at step t.

We reapet this N times where and assure convergence independent of the intial sample (irreducibility) we discard samples after the so-called burn-in period so that we produce a Markov-Chain of length $N - N_{\text{burn-in}}$.

A more mathematical prove of ergodicity for a more general Metropolis-within-Gibbs algorithm can be found in []. ARRIS RECURRENCE OF METROPOLIS-WITHIN-GIBBS AND TRANS-DIMENSIONAL MARKOV CHAINS By Gareth O. Roberts and Jeffrey S. Rosentha 2006

2.5.2 Draw a sample from a multivariate normal distribution

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

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

$$\pi(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}) \propto \pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta}) \tag{2.26}$$

$$= \exp\|\hat{\boldsymbol{A}}\boldsymbol{x} - \hat{\boldsymbol{y}}\|^2 , \tag{2.27}$$

Algorithm 1: Metropolis within Gibbs

```
Initialize and suppose two dimensional vector $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)})$
for $k = 1, \ldots, N$ do
    Propose $\theta_1 \sim q(\cdot|\theta_1^{(t-1)}) = q(\theta_1^{(t-1)}|\cdot)$
    Compute
        $\alpha(\theta_1|\theta_1^{(t-1)}) = \min\left\{1, \frac{\pi(\theta_1|\theta_2^{(t-1)}, \boldsymbol{y})\cancel{q(\theta_1^{(t-1)}|\theta_1)}}{\pi(\theta_1^{(t-1)}|\theta_2^{(t-1)}, \boldsymbol{y})\cancel{q(\theta_1|\theta_1^{(t-1)})}}\right\}$
    Draw $u \sim \mathcal{U}(0, 1)$
    if $\alpha \geq u$ then
        Accept and set $\theta_1^{(t)} = \theta_1$
    else
        Reject and keep $\theta_1^{(t)} = \theta_1^{(t-1)}$
    end if
    Draw $\theta_2^{(t)} \sim \pi(\cdot|\theta_1^{(t)}, \boldsymbol{y})$
end for
Output: $\boldsymbol{\theta}^{(0)}, \ldots, \boldsymbol{\theta}^{(k)}, \ldots, \boldsymbol{\theta}^{(N)} \sim \pi(\boldsymbol{\theta}|\boldsymbol{y})$
```

where

$$\hat{\boldsymbol{A}} = \begin{bmatrix} \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\theta})\boldsymbol{A} \\ \boldsymbol{Q}^{1/2}(\boldsymbol{\theta}) \end{bmatrix}, \quad \hat{\boldsymbol{y}} = \begin{bmatrix} \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\theta})\boldsymbol{y} \\ \boldsymbol{Q}^{1/2}(\boldsymbol{\theta})\boldsymbol{\mu} \end{bmatrix} []. \tag{2.28}$$

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

$$\boldsymbol{x}_i = \arg\min_{\hat{\boldsymbol{x}}} \|\hat{\boldsymbol{A}}\hat{\boldsymbol{x}} - (\hat{\boldsymbol{y}} + \boldsymbol{b})\|^2, \quad \boldsymbol{b} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{2.29}$$

where we add a randomised perturbation $\boldsymbol{b}$. Similarly as in section **??** we can rewrite the argument of Eq. 2.28 to

$$(\boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})\boldsymbol{A} + \boldsymbol{Q}(\boldsymbol{\theta}))\boldsymbol{x}_i = \boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})\boldsymbol{y} + \boldsymbol{Q}(\boldsymbol{\theta})\boldsymbol{\mu} + \boldsymbol{v}_1 + \boldsymbol{v}_2, \tag{2.30}$$

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

Within the MTC scheme ergodicity is implied if the Metropolis-within-Gibbs produces an ergodic chain , $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(k)}, \ldots, \boldsymbol{\theta}^{(N)}\} \sim \pi(\boldsymbol{\theta}|\boldsymbol{y})$ then $\mathcal{M} = \{(\boldsymbol{x}, \boldsymbol{\theta})^{(1)}, \ldots, (\boldsymbol{x}, \boldsymbol{\theta})^{(k)}, \ldots, (\boldsymbol{x}$ $\pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})$ is ergodic as well with independent samples $\boldsymbol{x}^{(k)}$ from the full conditional $\pi(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$ [].

2.5.3 t-walk

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

2.6 Numerical Approxiamtion Methods - Tensor Train

Using the tensor train format to approximate a d-dimensional function $\pi(x)$ enables us to compute marginal posterior probability distribution cheaply. As the name suggest the tensor train format is a train of tensors, more specifically two and three dimensional tensors which we call cores $\pi_k \in \mathbb{R}^{r_{k-1}\times n \times r_k}$ and which are connected through a ranks r_k and r_{k-1} for the kth dimension and defined by the number of gridpoints n, as in Figure **??** displayed. For the first and last dimensional core the outer ranks are $r_0 = r_d = 1$.

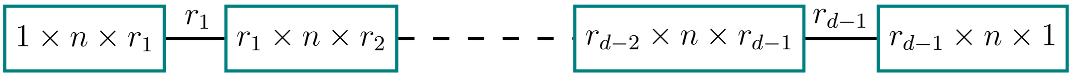

$1 \times n \times r_1$
r_1
$r_1 \times n \times r_2$
$r_{d-2} \times n \times r_{d-1}$
r_{d-1}
$r_{d-1} \times n \times 1$

Figure 2.3: text

Figure 2.4: nice matries picture

The approximated marginal target function

$$\begin{aligned} f_{X_k}(x_k) = \frac{1}{z}\Bigg| &\left(\int_{\mathbb{R}} \pi_1(x_1)\boldsymbol{\lambda}_1(x_1)\mathrm{d}x_1\right)\cdots\left(\int_{\mathbb{R}} \pi_{k-1}(x_{k-1})\boldsymbol{\lambda}_{k-1}(x_{k-1})\mathrm{d}x_{k-1}\right) \\ &\pi_k(x_k)\boldsymbol{\lambda}_k(x_k) \\ &\left(\int_{\mathbb{R}} \pi_{k+1}(x_{k+1})\boldsymbol{\lambda}_{k+1}(x_{k+1})\mathrm{d}x_{k+1}\right)\cdots\left(\int_{\mathbb{R}} \pi_d(x_d)\boldsymbol{\lambda}_d(x_d)\mathrm{d}x_d\right)\Bigg|, \end{aligned} \tag{2.31}$$

is given by integration over each core, where kth is in $\mathbb{R}^{r_{k-1}\times r_k}$ and z is some normalising constant. Here we introduce some Lebesgue measurable weight function $\boldsymbol{\lambda}(x) = \prod_{i=1}^{d} \boldsymbol{\lambda}_i(x_i)$. Why? [].

From here the notation and procedure is taken mostly from []. For numerical stability we can approximate the sqaure root of

$$\sqrt{\pi(x)} \approx \tilde{g}(x) = \boldsymbol{G}_1(x_1), \dots, \boldsymbol{G}_k(x_k), \dots, \boldsymbol{G}_d(x_d) \tag{2.32}$$

where the TT-core

$$G_k^{(\alpha_{k-1},\alpha_k)}(x_k) = \sum_{i=1}^{n_k} \phi_k^{(i)}(x_k) \boldsymbol{A}_k[\alpha_{k-1}, i, \alpha_k], \quad k = 1, ..., d, \tag{2.33}$$

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

We assume the function

$$\pi(x) \approx \gamma' + g^2(x)\,, \tag{2.34}$$

where $g(x)$ is defined through the tensor train decomposition plus an error γ' according to the l2 norm. Then the normalised target function is

$$f_X(x) = \frac{1}{z}\pi(x)\boldsymbol{\lambda}(x) = \frac{1}{z}(\gamma'\boldsymbol{\lambda}(x) + g^2(x)\boldsymbol{\lambda}(x)) \tag{2.35}$$

with a normalisation constant z. Consquently the approximated marginal functions can be expressed as

$$\begin{aligned} f_{X_k}(x_k) = \frac{1}{z}\Bigg(\gamma' \prod_{i=1}^{k-1} \boldsymbol{\lambda}_i(\mathcal{X}_i) \prod_{i=k+1}^{d} \boldsymbol{\lambda}_i(\mathcal{X}_i) \\ + \left(\int_{\mathbb{R}} \boldsymbol{G}_1^2(x_1)\boldsymbol{\lambda}_1(x_1)\mathrm{d}x_1\right) \cdots \left(\int_{\mathbb{R}} \boldsymbol{G}_{k-1}^2(x_{k-1})\boldsymbol{\lambda}_{k-1}(x_{k-1})\mathrm{d}x_{k-1}\right) \\ \boldsymbol{G}_k^2(x_k)\boldsymbol{\lambda}_k(x_k) \\ \left(\int_{\mathbb{R}} \boldsymbol{G}_{k+1}^2(x_{k+1})\boldsymbol{\lambda}_{k+1}(x_{k+1})\mathrm{d}x_{k+1}\right) \cdots \left(\int_{\mathbb{R}} \boldsymbol{G}_d^2(x_d)\boldsymbol{\lambda}_d(x_d)\mathrm{d}x_d\right)\Bigg)\,, \end{aligned} \tag{2.36}$$

where $\boldsymbol{\lambda}_k(\mathcal{X}_k) = \int_{\mathcal{X}_k} \boldsymbol{\lambda}_k(x_k)\mathrm{d}x_k$.

To effeciently calculate these marginals on can us a procedure which is called left and right orthogonalization of cores [] [32]. To do so we define the mass matrix $\boldsymbol{M}_k \in \mathbb{R}^{n_k \times n_k}$ by

$$\boldsymbol{M}_k[i,j] = \int_{X_k} \phi_k^{(i)}(x_k)\phi_k^{(j)}(x_k)\boldsymbol{\lambda}(x_k)\, dx_k, \quad i = 1, ..., n_k, \quad j = 1, ..., n_k, \tag{2.37}$$

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

2.6.1 Marginal Functions

We calculate the marginal functions through procedures, which we call backward marginalisation [] and forward marginalisation. We gain the coefficient matrices $\boldsymbol{B}_k$ through backward marginalisation and the coefficient matrices $\boldsymbol{B}_{pre,n}$ through forward marginalisation, which enables us to calculate marginal fuunction similar to [].

The proposition 1 to caculte $\boldsymbol{B}_k$ is taken from [].

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

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

$$\boldsymbol{C}_k[\alpha_{k-1}, \tau, l_k] = \sum_{i=1}^{n_k} \boldsymbol{B}_k[\alpha_{k-1}, i, l_k] \boldsymbol{L}_k[i, \tau]. \qquad (2.38)$$

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

$$\boldsymbol{Q}_k \boldsymbol{R}_k = \left(\boldsymbol{C}_k^{(R)}\right)^\top. \qquad (2.39)$$

3. Compute the new coefficient tensor:

$$\boldsymbol{B}_{k-1}[\alpha_{k-2}, i, l_{k-1}] = \sum_{\alpha_{k-1}=1}^{r_{k-1}} \boldsymbol{A}_{k-1}[\alpha_{k-2}, i, \alpha_{k-1}] \boldsymbol{R}_k[l_{k-1}, \alpha_{k-1}]. \qquad (2.40)$$

Then we need to do this the other way as well.

In addiotn we also have to do forward marginalistion starig with the first dimension

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

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

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

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

$$\boldsymbol{Q}_{pre,k} \boldsymbol{R}_{pre,k} = (\boldsymbol{C}_{pre,k}^{(R)}). \tag{2.42}$$

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

$$\boldsymbol{B}_{pre,k+1}[l_{k+1}, i, \alpha_{k+1}] = \sum_{\alpha_k=1}^{r_k} \boldsymbol{R}_{pre,k}[l_{k+1}, \alpha_k] \boldsymbol{A}_{k+1}[\alpha_k, i, \alpha_{k+1}]. \tag{2.43}$$

The marginal PDF of X_k can be expressed as

$$f_{X_k}(x_k) = \frac{1}{z} \left(\gamma' \prod_{i=1}^{k-1} \lambda_i(X_i) \prod_{i=k+1}^{d} \lambda_i(X_i) + \sum_{l_{k-1}=1}^{r_{k-1}} \sum_{l_k=1}^{r_k} \left(\sum_{i=1}^{n_k} \phi_k^{(i)}(x_k) \boldsymbol{D}_k[l_{k-1}, i, l_k] \right)^2 \right) \lambda_k(x_k), \tag{2.44}$$

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

$$\boldsymbol{D}_k[l_{k-1}, i, l_k] = \sum_{\alpha_{k-1}=1}^{r_{k-1}} \boldsymbol{R}_{pre,k-1}[l_{k-1}, \alpha_{k-1}] \boldsymbol{B}_k[\alpha_{k-1}, i, l_k]. \tag{2.45}$$

Special Cases The marginal PDF of X_1 can be expressed as

$$f_{X_1}(x_1) = \frac{1}{z} \left(\gamma' \prod_{i=2}^{d} \boldsymbol{\lambda}_i(\boldsymbol{\mathcal{X}}_i) + \sum_{l_1=1}^{r_1} \left(\sum_{i=1}^{n_1} \phi_1^{(i)}(x_1) \boldsymbol{D}_1[i, l_1] \right)^2 \right) \boldsymbol{\lambda}_1(x_1), \tag{2.46}$$

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

$$f_{X_n}(x_n) = \frac{1}{z} \left(\gamma' \prod_{i=1}^{d-1} \boldsymbol{\lambda}_i(\boldsymbol{\mathcal{X}}_i) + \sum_{l_{n-1}=1}^{r_{n-1}} \left(\sum_{i=1}^{n_1} \phi_1^{(i)}(x_1) \boldsymbol{D}_n[l_{n-1}, i] \right)^2 \right) \boldsymbol{\lambda}_n(x_n), \tag{2.47}$$

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

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.