

WESTFÄLISCHE WILHELMS-UNIVERSITÄT MÜNSTER
INSTITUTE OF THEORETICAL PHYSICS

MARKOV CHAIN MONTE CARLO STUDIES OF THE PROTON STRUCTURE

MASTERS THESIS

Peter Risse

Primary Supervisor

P.D. Dr. Karol Kovařík

Secondary Supervisor

Prof. Dr. Michael Klasen

14.09.2020

Contents

1. Introduction	1
2. Parton Distribution Functions and Deep Inelastic Scattering	3
2.1. The QCD Lagrangian	3
2.2. Deep inelastic scattering	4
2.3. Parton Model	7
2.4. QCD improved Parton Model	10
2.5. F_2 measurements and PDF parametrisation	16
3. Bayesian inference and the χ^2 function	18
3.1. The principle of maximum entropy	18
3.2. Location and scale parameters	19
3.3. Mean and variance	20
3.4. Likelihood for several uncorrelated measurements	23
3.5. Likelihood for several correlated measurements	24
3.6. Final Bayesian model	26
4. Theory of Markov Chains	28
4.1. Properties of Markov Chains	28
4.2. Detailed balance	31
4.3. Parameter mean and the statistical error	31
4.4. Numerical estimators and the error of the error	34
4.5. Automatic summation window W	37
4.6. Test case: synthetic data	38
5. Monte Carlo algorithms	44
5.1. The Metropolis-Hastings Algorithm	44
5.2. Optimal acceptance rate	48
5.3. Adaptive Metropolis-Hastings	50
6. Error estimation	53
6.1. The statistical error of secondary observables	53
6.2. The Hessian method for error estimation	54
6.3. Markov Chain Monte Carlo inference	57
7. Proton valence PDFs from DIS experiments	61
7.1. Generating samples	61
7.2. Comparison with experimental measurements	65
7.3. Comparison with Levenberg-Marquardt minimisation	67
8. Conclusion	69
A. Calculations regarding the theory of Markov Chains	71
A.1. Leading bias of the autocorrelation function	71
A.2. The variance of the autocorrelation function	72
A.3. Error of the projected normalized autocorrelation function	73

A.4. The variance of the full and the naive covariance matrix	75
A.5. Analytic results for the test case	76
B. Relations and identities	78
B.1. Lab frame kinematics	78
B.2. Gaussian integrals	79
B.3. Matrix manipulation	79
C. Additional figures	80
References	83

1. Introduction

Parton distribution functions are an essential part of predicting and understanding high energy experiments. With the search for new physics and the rising accuracy of standard model parameters, it is becoming increasingly important to create reliable data sets describing the density functions and to have a good understanding of their uncertainties. Constraining the different densities for the proton or for hadrons is an ongoing research topic. The current procedure (e.g. [22], [11]) for obtaining PDFs is to propose a universal function and determine its parameters from experiments such as deep inelastic scattering or Drell-Yan processes. The determination of parameters is done via minimising a figure of merit, namely the correlated χ^2 -function and the uncertainty intervals are obtained by making use of quadratic approximations around the best fit value. The accuracy hereby does not only depend on the viable data, but also on the theoretical predicts carried out in perturbation theory and the evolution of the PDFs from an initial scale to the required scale of the interaction.

In this masters thesis the focus lies within a different way of investigating the figure of merit. In the current state-of-the-art procedure the handling of error estimation is very difficult, when the region around the best value is not well estimated via a quadratic expansion. This leads to less well constrained predictions. Furthermore it is not straightforward to judge if the algorithm used did indeed find the global minimum. It is therefore advantageous to have a second independent method to investigate the χ^2 function. The framework of Markov Chain Monte Carlo might provide such an independent procedure of constraining parton distribution functions.

In fact there has recently been a proof of concept by Y. Gbedo and M. Mangin-Brinet [13], where PDF and their uncertainties were extracted with Markov Chain Monte Carlo methods. The goal of this thesis is to understand the relevant concepts, expand the codebase (written in **C++** and **Fortran**) of the nCTEQ collaboration and try to recreate the successful determination of PDFs. Although the paper addresses the exact same topic as this thesis, we will not use the paper itself as a guideline and recreate it step by step, because it is only a proof of concept paper and therefore does not go into great detail. We will concentrate more on the general features of Markov Chain Monte Carlos. The thesis starts by introducing the Parton model in the framework of deep inelastic scattering. Here the general concepts of the model are stated and its combination with QCD leading to the DGLAP equations is discussed. Also the experimental measurements and the nCTEQ codebase are introduced. The third section aims at validating the χ^2 function as the figure of merit in the context of Bayesian probability theory. By this stage it has become clear, why solving this inverse problem for the parameters from experimental data is highly non-trivial. However the next section introduces Markov Chains, which allow to compute expectation values from arbitrarily complicated distributions in a simple manner. Here it is also discussed how to analyse such a chain and extract its key properties. In section five certain Monte Carlo algorithms for generating a Markov Chain are investigated and then the focus lies with the adaptive Metropolis-Hastings algorithm, which optimises the properties of the chain with small amounts of χ^2 evaluations. The last preparatory section deals with the error estimation of functions, which take the

samples of a chain as parameters. In the final analysis the functions will be set to the PDFs or theoretical predictions like F_2 -values for the proton.

At this point in the thesis all preparations are done. The nCTEQ codebase has been extended and is able to fit parameters with Markov Chain Monte Carlo methods. The analysis of a Markov Chain can be done with a self written `Python`-package. This has been excluded from the nCTEQ code, since the analysis requires a lot of plotting due to visual keys, which is very fast and convenient with the `Pyplot`-package.

In the final section of this thesis all of the earlier sections are combined and a proton valence distribution fit from deep inelastic scattering data is carried out. This fit is explained in detail and its physical predictions are not only compared to the data set used for fitting, but also to a complementary fit, which is the core analysis of P. Duwentäster's masters thesis [12]. Here well-known minimisation techniques were used to solve the inverse problem.

2. Parton Distribution Functions and Deep Inelastic Scattering

In this introductory section the aim is to lay the theoretical foundation for the Markov Chain Monte Carlo techniques to be applied. It starts by a short overview of the Quantum Chromodynamics Lagrangian. This can also be found in every standard text book, e.g. from Schwartz [39], Peskin and Schroeder [30], etc. Then the concepts of Deep Inelastic proton scattering are introduced, which is used to motivate a naive Parton model. Here the notation is kept similar to Schwartz's. The Parton model is, after stating the core ideas and results, combined with QCD calculations, leading to the DGLAP evolution equations for the parton distributions functions. Here the paper by Martin [25] provided an introduction and further information can also be found in [6]. Additional papers used are cited within the text. Lastly experimental measurements used for the actual fit are discussed and the nCTEQ codebase is introduced.

2.1. The QCD Lagrangian

The Lagrangian of Quantum chromodynamics is separated from the electroweak section of the Standard Model. Furthermore it can be split into three different parts:

$$\mathcal{L}_{\text{QCD}} = \mathcal{L}_{\text{classical}} + \mathcal{L}_{\text{gauge-fixing}} + \mathcal{L}_{\text{ghost}}.$$

It exhibits a $\text{SU}(3)$ colour symmetry which is thought to be exact, since a single colour state has not been observed up until now. From Baryons such as Δ^{++} , which is made up of three up-quarks, we know that we need the colour quantum number in order to comply with the Pauli exclusion principle. Starting with the latter two terms, the gauge-fixing and ghost parts are given by

$$\mathcal{L}_{\text{gauge-fixing}} = -\frac{1}{2\xi}(\partial_\mu A_\mu^a)(\partial^\mu A^{a\mu}) \quad \mathcal{L}_{\text{ghost}} = (\partial_\mu \bar{c}^a)(\delta^{ac}\partial^\mu + gf^{abc}A^{b\mu})c^c,$$

where ξ is the gauge-fixing parameter and c^a and \bar{c}^a are the Faddeev-Popov ghosts and anti-ghosts. The ghosts cancel unphysical degrees of freedom of the gluons in loop calculations. The reason for their existence is a consequence of the Lagrangian formulation of field theory (see for example [39, chapter 25.4]).

The classical part consists of the gauge fields A_μ^a and fermion fields ψ_j . It takes on the form of

$$\begin{aligned} \mathcal{L}_{\text{classical}} &= \bar{\psi}_i[(i\not{\partial} - m_i)\delta_{ij} + g\not{A}_{ij}^a T_{ij}^a]\psi_j - \frac{1}{4}F_{\mu\nu}^a F^{a\mu\nu} \\ \text{with } F_{\mu\nu}^a &= \partial_\mu A_\nu^a - \partial_\nu A_\mu^a + gf^{abc}A_\mu^b A_\nu^c. \end{aligned} \tag{2.1}$$

In contrast to Quantum electrodynamics the field strength tensor has a third non-abelian term, giving rise to self-interactions between three or even four gluons. In fact by investigating higher order loop calculations one finds, that the coupling shrinks at higher

energies due to this third term. The coupling is strong at large distances and weak for short distances (asymptotic freedom). For the detection in 1973 Gross, Wilczek [16] and Politzer [31] were awarded the Nobel Prize in 2004. It is for this reason that no free quarks are found in experiments.

The T^a in (2.1) are the generators of $SU(3)_c$ in the fundamental representation obeying

$$T_{ij}^a = \frac{\lambda_{ij}^a}{2} \quad [T^a, T^b] = if^{abc}T^c,$$

with the Gell-Mann matrices λ^a and the fully antisymmetric structure constants f^{abc} . The normalisation is given by

$$\text{Tr}(T^a T^b) = T_F \delta^{ab} \quad T_F = \frac{1}{2}$$

and the Casimir operator is

$$\sum_{a,k} T_{ik}^a T_{kj}^a = C_F \delta_{ij} \quad C_F = \frac{N^2 - 1}{2N} = \frac{4}{3}.$$

The adjoint representation, acting on the vector space spanned by the generators in the fundamental representation, are constructed from the structure constants

$$(T_A^a)_{bc} = -if^{abc}$$

and their normalisation is given by

$$\text{Tr}(T_A^a T_A^b) = T_A \delta^{ab} \quad T_A = 3.$$

The Casimir operator leads to an identical equation implying $C_A = T_A$.

2.2. Deep inelastic scattering

This section introduces the concepts of deep inelastic scattering, a framework, where the parton model is explained best. In fact we will later use experimental measurements from these kind of processes to fit the parton distribution functions. We consider an electron scattering of a proton with such a high momentum transfer that the proton splits up into its constituents: $e^- P \rightarrow e^- X$. Figure 2.1 illustrates the situation schematically, with the greyed circle representing our ignorance about scattering off a bound state. The bound state itself can not be calculated via perturbative methods due to the running coupling of QCD. We will therefore make a general ansatz with so-called structure functions. The next section introducing the parton model will make predictions for these functions.

The differential cross section for the process above can be written in terms of a leptonic

2.2. Deep inelastic scattering

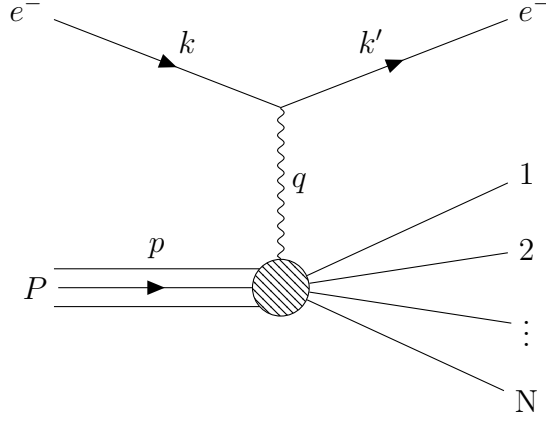


Figure 2.1: Schematic diagram for deep inelastic scattering. The electron with momentum k scatters of a proton with momentum p . The momentum transfer, given by q , is large enough for the proton splitting up into an undefined final state.

tensor $L_{\mu\nu}$ and a hadronic tensor $W_{\mu\nu}$:

$$\frac{d\sigma}{dE'd\Omega} = \frac{\alpha^2 E'}{q^4 E} \sum_i \eta_i L_{\mu\nu}^i W^{\mu\nu}. \quad (2.2)$$

E and E' are the initial and final state energies respectively. We will compute the unpolarised cross section. The hadron tensor has to be kept as general as possible due to our ignorance regarding the final state. The most general form, only using $g^{\mu\nu}$, q and p , is given by

$$W^{\mu\nu} = -g^{\mu\nu}W_1 + \frac{p^\mu p^\nu}{M^2}W_2 + \frac{i}{2M^2}\epsilon^{\mu\nu\rho\lambda}p_\rho q_\lambda W_3 + \frac{q^\mu q^\nu}{M^2}W_4 + \frac{p^\mu q^\nu + p^\nu q^\mu}{M^2}W_5.$$

M is the proton mass. By making use of current conservation $q_\mu W^{\mu\nu} = 0$ we can simplify the expression above and rewrite W_4 and W_5 in terms of W_1 and W_2 . The reduced tensor equals

$$W^{\mu\nu} = \left(\frac{q^\mu q^\nu}{q^2} - g^{\mu\nu} \right) W_1 + \left(p^\mu - \frac{q^\mu p \cdot q}{q^2} \right) \left(p^\nu - \frac{q^\nu p \cdot q}{q^2} \right) \frac{W_2}{M^2} + \frac{i}{2M^2}\epsilon^{\mu\nu\rho\lambda}p_\rho q_\lambda W_3. \quad (2.3)$$

The third term W_3 only plays a role, if the leptonic tensor is antisymmetric. As the sum over i in (2.2) runs over all contributions from the possible interactions, we include photon and Z_0 exchange, giving i the 'values' $\gamma\gamma$, ZZ and the interference term γZ . The η_i include the couplings and factors from the propagators. If we neglect the electron

mass, the lepton tensors yield:

$$\begin{aligned}
 L_{\gamma\gamma}^{\mu\nu} &= 2(k^\mu k'^\nu + k^\nu k'^\mu - (k \cdot k')q^{\mu\nu}) \\
 L_{\gamma Z}^{\mu\nu} + L_{Z\gamma}^{\mu\nu} &= 4c_V(k^\mu k'^\nu + k^\nu k'^\mu - (k \cdot k')q^{\mu\nu}) + 4ic_A\epsilon^{\mu\nu\rho\lambda}k_\rho k'_\lambda \\
 L_{ZZ}^{\mu\nu} &= 2(c_V^2 + c_A^2)(k^\mu k'^\nu + k^\nu k'^\mu - (k \cdot k')q^{\mu\nu}) + 4ic_Vc_A\epsilon^{\mu\nu\rho\lambda}k_\rho k'_\lambda,
 \end{aligned} \tag{2.4}$$

with the vector and axial couplings

$$c_V = t_3 - 2e \sin \theta_W \quad \text{and} \quad c_A = t_3,$$

which entail the third component of the weak isospin t_3 , the electric charge e and the Weinberg angle θ_W .

We choose the proton to be at rest, resulting in the kinematics given in (B.1). Now contracting the leptonic tensor to the hadronic one yields

$$\frac{d\sigma}{dE'd\Omega} = 4\frac{\alpha^2 E'^2}{q^4} \left(2W_1 \sin^2 \frac{\theta}{2} + W_2 \cos^2 \frac{\theta}{2} + \frac{E+E'}{M} W_3 \sin^2 \frac{\theta}{2} \right)$$

after a lot of algebra. Here it is useful to define dimensionless structure functions:

$$F_1 = MW_1 \quad F_2 = \frac{p \cdot q}{M} W_2 = (E - E')W_2 \quad F_3 = \frac{p \cdot q}{M} W_3 = (E - E')W_3.$$

Furthermore the cross section is usually written in terms of the energy scale of the collision $Q = \sqrt{-q^2}$ and the so called Bjorken variable $x = -q^2/(2p \cdot q)$. The change of variables is then simply given by

$$\frac{d\sigma}{dx dQ^2} = \frac{\pi y}{xE'} \frac{d\sigma}{dE' d\Omega} \quad \text{with} \quad y = 1 - \frac{E'}{E}.$$

Finally collecting the definitions above allows us to write the final differential cross section as¹

$$\frac{d\sigma}{dx dQ^2} = \frac{4\pi\alpha^2}{xQ^2} \left[xy^2 F_1 + \left(1 - y - \frac{Mxy}{2E} \right) F_2 + xy \left(1 - \frac{y}{2} \right) F_3 \right] \tag{2.5}$$

In this way writing down the differential cross section the kinematics (x, y and Q) decoupled from the proton structure information. By measuring the cross section it will be possible to extract these functions. On the other hand the parton model, which will be introduced in the following section, makes predictions for F_1, F_2 and F_3 , making it possible to verify the model and fix missing parameters.

¹If the incoming particle were an e^+ instead of an e^- , the sign in front of the third structure function would have to be negative.

2.3. Parton Model

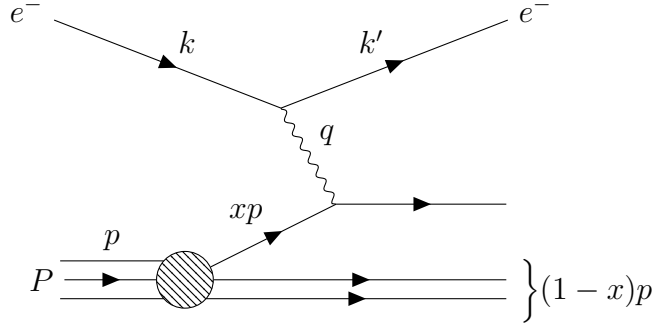


Figure 2.2: Deep inelastic scattering in the framework of the parton model, where the electron scatters off one of the constituents of the proton elastically. This quark carries a fraction of x of the total momentum of the proton.

2.3. Parton Model

The parton model was originally suggested by Feynman and its key assumption is that there are point-like particles within the proton, which are essentially free. These particles are quarks, gluons and antiquarks. So when an electron scatters off a proton, it interacts with one of its constituents elastically. This situation is depicted in fig. 2.2. We say, that the parton has a momentum p_i collinear to the proton momentum before the interaction with the electron and then has a momentum p_f afterwards. From $p_i^\mu + q^\mu = p_f^\mu$ we get

$$m_q^2 + 2p_i \cdot q + q^2 = m_q^2 \quad \Rightarrow \quad \frac{Q^2}{2p_i \cdot q} = 1$$

by squaring. Thus if we say the quark has a fraction x_q of the momentum of the proton ($p_i = x_q p$), we see by evaluating

$$x_q = x_q \frac{Q^2}{2p_i \cdot q} = \frac{Q^2}{2p \cdot q} = x$$

that the fraction x_q is in fact the Bjorken x so the subscript will be dropped from now on.

With this assumption at hand it is still unclear how to compute cross sections with this model, since it is unclear how many partons are present in the proton at a given moment in time. The time scale of the interaction is of order $\mathcal{O}(Q^{-1})$, whilst the dynamics inside the proton are much larger, being of order $\mathcal{O}(M^{-1})$ (again: $M \ll Q$). The quarks and gluons cannot interact at the time scale of the interaction and are approximately free. The electron therefore only sees a 'snap shot' of the frozen partons and since we cannot calculate the internal dynamics, we model these 'snap shots' as distributions $f_i(x)$. So the quantity $f_i(x)dx$ gives the total number of partons of type i inside the proton with a momentum fraction in the range $[x, x + dx]$. The distributions $f_i(x)$ are therefore

similar to classical probability distributions, but with a slightly different normalisation as we will see below. In the literature the quantity $f_i(x)dx$ is therefore (simplified) called 'probability of finding a quark of type i with momentum fraction x ', where it is actually the number.

The proton has certain quantum numbers which need to be fulfilled. For example the proton has an up-quark number of 2, so the distributions of up-type and anti-up-type have to be normalised in this way

$$\int_0^1 dx (f_u(x) - f_{\bar{u}}(x)) = 2 \quad (2.6)$$

and furthermore for the other types²

$$\int_0^1 dx (f_d(x) - f_{\bar{d}}(x)) = 1 \quad \text{and} \quad \int_0^1 dx (f_c(x) - f_{\bar{c}}(x)) = 0. \quad (2.7)$$

Only with this combination it is possible to get the charge to be 1, the baryon number 1 and the strangeness 0. It is important to note that only the difference is normalised and not the distributions themselves. The gluon number is not conserved, so f_g has no associated sum rule, except for the momentum conservation, which takes the form of

$$\sum_i \int_0^1 dx (x f_i(x)) = 1 \quad (2.8)$$

in this context. So the expectation value of the momentum fraction for each parton type added up has to be equal to 1. In other words: The proton is fully made up of partons. The proton distributions from the nCTEQ15 global analysis are shown in fig. 2.3. As one can see the up-density peaks at a much higher value than the down-density due to the different normalisations.

The neutron can be built up in the same manner and since it forms an isospin doublet one can relate the up-quark distribution from the neutron with the down-quark distribution from the proton. The opposite case is also possible:

$$u^p(x) = d^n(x) \quad d^p(x) = u^n(x) \quad s^p(x) = s^n(x).$$

Here we added the obvious relation for the strange-quarks and used the conventional naming for the distributions.

It is also often useful to look at the valence-quark distributions, by removing the portion of the quark distribution, which is due to the creation of quark and antiquark pairs. The

²If one also included charm- and bottom-quarks, their normalisation would be analogous to the strange densities.

2.3. Parton Model

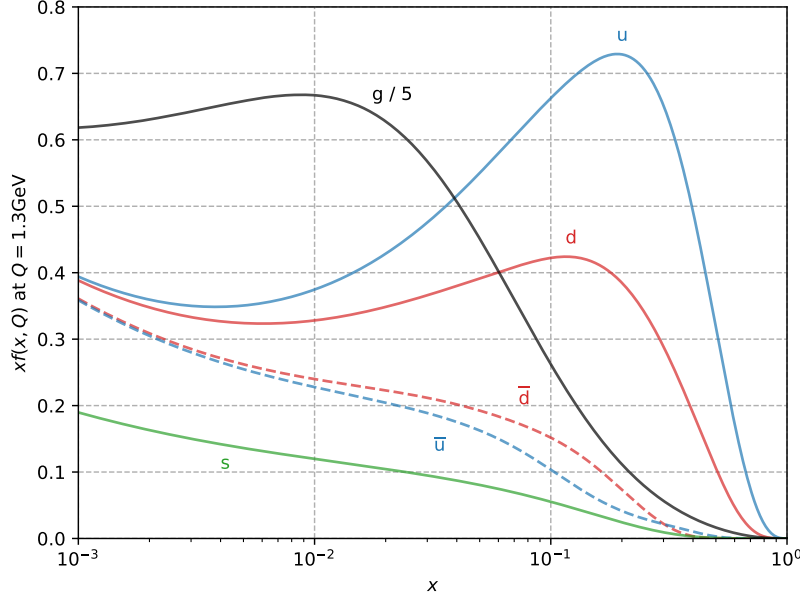


Figure 2.3: The nCTEQ15 [22] proton PDFs at $Q = 1.3\text{GeV}$ for $u, \bar{u}, d, \bar{d}, s = \bar{s}$ and g . Here the numeric values were obtained by making use of the LHAPDF [8] package.

latter ones are called sea-quark distributions. The valence distributions are given by

$$u_v(x) = u(x) - \bar{u}(x) \quad \text{and} \quad d_v(x) = d(x) - \bar{d}(x).$$

These are the only two valence distributions, since every other type of quark has to be created as a quark antiquark pair.

In the parton model the cross section for electron proton scattering is written as the sum over the expectation values for the electron quark cross sections for every quark type. Obviously the gluons do not participate due to their lack of electric or weak charge. In equations:

$$\frac{d\sigma}{dx dQ^2} = \sum_q \int_0^1 d\xi f_q(\xi) \left(\frac{d\hat{\sigma}}{dx dQ^2} \right). \quad (2.9)$$

Therefore by comparing this equation with (2.5) one is able to read off predictions made by the parton model. The cross section for point-like photon interaction is for example:

$$\frac{d\hat{\sigma}_\gamma}{dx dQ^2} = \frac{4\pi\alpha^2}{xQ^2} e_q^2 \left(1 + (1-y)^2 \right) \delta(x - \xi).$$

If one combines all three interaction terms one arrives at

$$\begin{aligned}
 2xF_1^\gamma &= F_2^\gamma = x \sum_q e_q^2 f_q(x) & F_3^\gamma &= 0 \\
 2xF_1^{\gamma Z} &= F_2^{\gamma Z} = x \sum_q e_q c_{V,q} f_q(x) & F_3^{\gamma Z} &= x \sum_q \pm e_q c_{V,q} f_q(x) \\
 2xF_1^Z &= F_2^Z = x \sum_q (c_{V,q}^2 + c_{A,q}^2) f_q(x) & F_3^Z &= x \sum_q \pm (c_{V,q}^2 + c_{A,q}^2) f_q(x),
 \end{aligned} \tag{2.10}$$

where the positive sign in F_3 is for quarks and the negative sign for antiquarks. There are two important observations from these predictions. First F_1 and F_2 are always connected via a factor of $2x$. This is known as the Callan-Gross relation. This can be traced back to the fact that the quarks are spin- $\frac{1}{2}$ particles (see for example [39, chapter 32.1]). The second important prediction is the non-existent dependence on Q^2 . This is called Bjorken scaling. The experimental verification is plotted in fig. 2.4 and is best seen in the region $[0.1, 0.6]$ for x . Bjorken scaling was historically confirmed but later experiments showed a logarithmic Q^2 dependence, especially in the low x -region ($x < 0.01$). This feature can only be seen if one includes corrections from QCD and will lead to the well-known DGLAP-equations.

2.4. QCD improved Parton Model

The combination of perturbative QCD with the Parton Model shows the violation of the Bjorken scaling discussed above. In the process we will need to redefine the parton distribution functions beyond leading order of perturbation theory. But we will assume that the basic statements of the Parton Model hold insofar, that the distribution function $f_i(\xi)d\xi$ still describes the number of quarks of type i at the momentum fraction ξ . We will first need to write down $\hat{W}^{\mu\nu}$ as the parton version of $W^{\mu\nu}$. By introducing the partonic version of the Bjorken variable

$$z = \frac{Q^2}{2p_i \cdot q},$$

which connects x and ξ via $x = z\xi$, we can write the hadronic tensor as

$$\begin{aligned}
 W^{\mu\nu} &= \sum_i \int_0^1 dz \int_0^1 d\xi f_i(\xi) \hat{W}^{\mu\nu}(z, Q) \delta(x - z\xi) \\
 &= \sum_i \int_x^1 \frac{d\xi}{\xi} f_i(\xi) \hat{W}^{\mu\nu}\left(\frac{x}{\xi}, Q\right).
 \end{aligned} \tag{2.11}$$

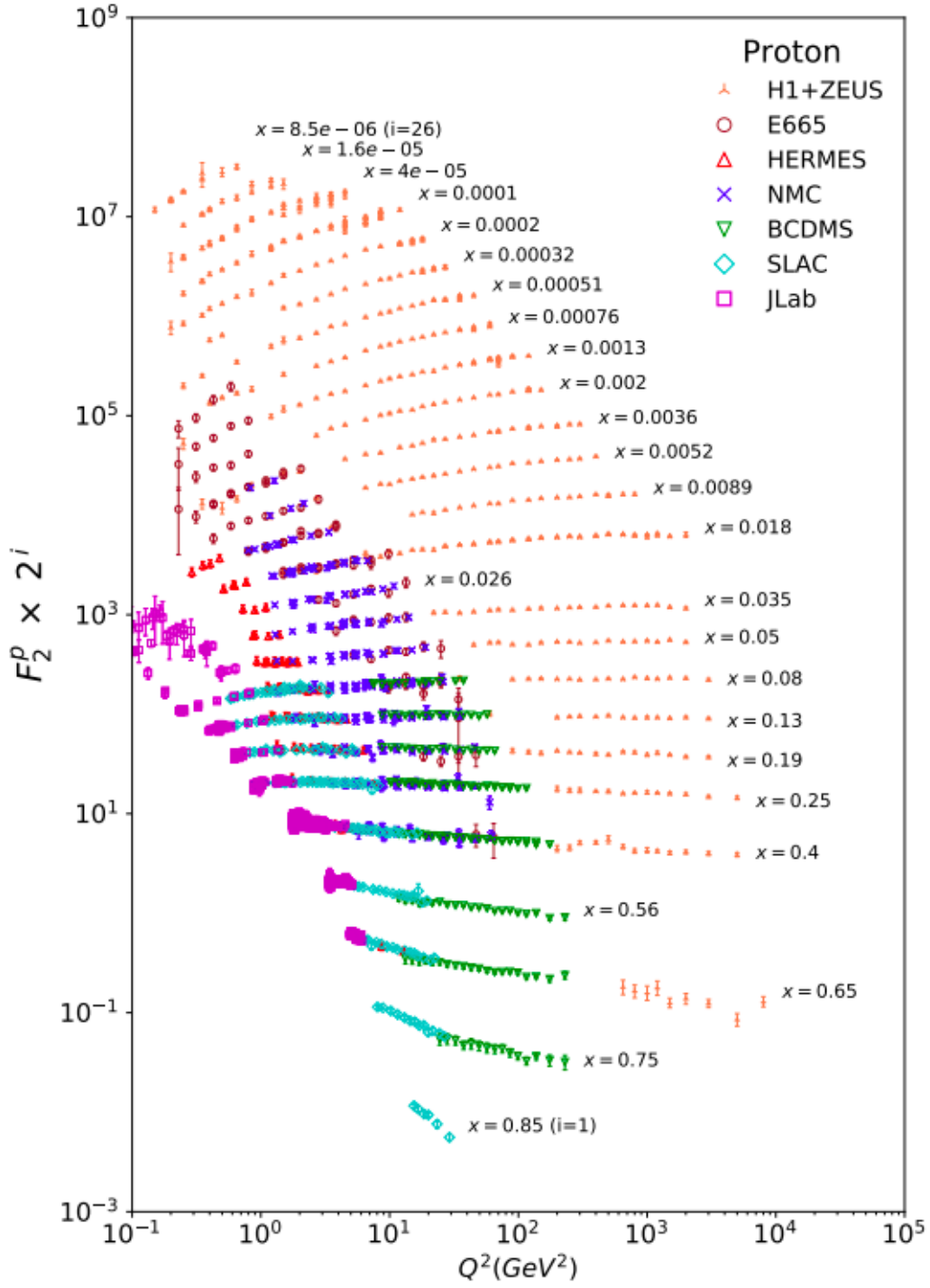


Figure 2.4: F_2 measurements from different experiments, where the value of F_2 gets multiplied by 2^i for each batch of fixed x . The data shows logarithmic Q^2 dependence for low x values, but is nearly independent of Q^2 for higher values. The figure is from the particle data group [17].

2. Parton Distribution Functions and Deep Inelastic Scattering

For simplicity we will only consider photonic interactions here. At leading order the only contribution is given by $\gamma^* q \rightarrow q$ and with $p_f = p_i + q$ we find

$$\begin{aligned}\hat{W}^{\mu\nu}(z, Q) &= \frac{e_i^2}{2} \int \frac{d^3 p_f}{(2\pi)^3} \frac{1}{2E_f} \text{Tr} \left[\gamma^\mu \not{p}_i \gamma^\nu \not{p}_f \right] (2\pi)^4 \delta^4(p_i + q - p_f) \\ &= 2\pi e_i^2 \left[\left(-g^{\mu\nu} + \frac{q^\mu q^\nu}{q^2} \right) + \frac{4z}{Q^2} \left(p_i^\mu - \frac{p_i \cdot q}{q^2} q^\mu \right) \left(p_i^\nu - \frac{p_i \cdot q}{q^2} q^\nu \right) \right] \delta(1-z)\end{aligned}$$

By comparing the expression above to the hadronic tensor from (2.3), we get the partonic form of the Callan-Gross relation

$$\hat{W}_1 = \frac{Q^2}{4z} \hat{W}_2 = 2\pi e_i^2 \delta(1-z) \quad (2.12)$$

which still holds at leading order. We will now introduce a new structure function W_0 in order to simplify the following discussion and provide a basis for calculating the Q dependence of the PDFs. We will define $W_0 \equiv -q^{\mu\nu} W_{\mu\nu}$, which is then closely related to the unpolarised cross section. We get

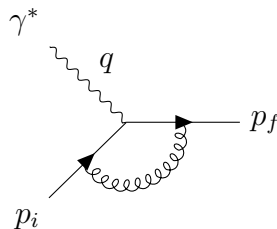
$$W_0(z, Q) = 3W_1(z, Q) - W_2(z, Q) \left(M^2 + \frac{Q^2}{4z^2} \right) \stackrel{Q \gg M}{\approx} 2W_1(z, Q)$$

and in particular

$$W_0(z, Q) = 4\pi \sum_i e_i^2 f_i(z). \quad (2.13)$$

By turning this equation around we can use it to define Parton distribution functions beyond leading order. We will proceed in presenting the solution of diagrams at next to leading order contributing to W_0 and use these results to redefine the PDFs thereby arriving at their Q dependence.

Since this calculation will be up to NLO the following amplitudes will be multiplied by the leading-order contribution. The first diagram to be considered is


(2.14)

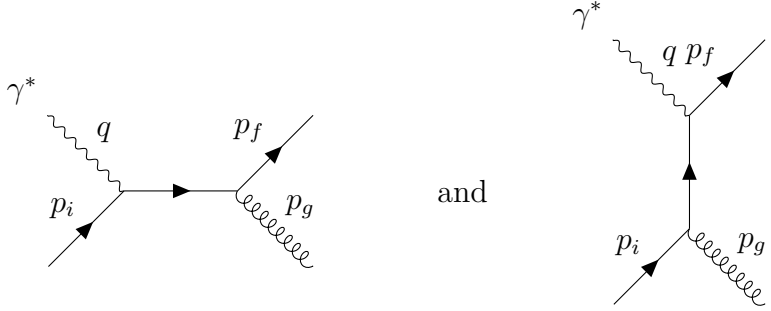
The interference between the leading-order graph and this virtual contribution can be calculated with the divergences by dimensional regularisation. The divergences are then

2.4. QCD improved Parton Model

explicitly shown as poles. The result in $D = 4 - \epsilon$ dimension is³

$$\hat{W}_0^V = 4\pi e_i^2 \frac{\alpha_s}{2\pi} C_F \left(\frac{4\pi\mu^2}{Q^2} \right)^{\frac{\epsilon}{2}} \frac{\Gamma(1 - \frac{\epsilon}{2})}{\Gamma(1 - \epsilon)} \left(\frac{\pi^2}{3} - 8 - \frac{8}{\epsilon^2} - \frac{6}{\epsilon} \right) \delta(1 - z) + \mathcal{O}(\epsilon). \quad (2.15)$$

All terms proportional to ϵ are omitted for visual clarity. We see two divergences for $\epsilon \rightarrow 0$. Here all UV divergences have been removed via the counterterm, so all ϵ 's are of IR kind and can be partially cancelled by the real contributions. As we will see the double pole will vanish, but one needs to pay close attention to the other pole. The real contributions are given by two diagrams, where either the initial or the final quark radiates a gluon:



(2.16)

The calculation of these diagrams is more tedious and in order to make the poles explicit one needs to use the plus distribution, which is defined as

$$\int_0^1 dx f(x) [g(x)]_+ = \int_0^1 dx (f(x) - f(1)) g(x) \quad \text{and} \quad [g(x)]_+ = g(x) \quad \text{for} \quad x \neq 1.$$

The solution is then given by⁴

$$\begin{aligned} \hat{W}_0^R = 4\pi e_i^2 \frac{\alpha_s}{2\pi} C_F \left(\frac{4\pi\mu^2}{Q^2} \right)^{\frac{\epsilon}{2}} \frac{\Gamma(1 - \frac{\epsilon}{2})}{\Gamma(1 - \epsilon)} & \left(3 + 2z - \frac{1 + z^2}{1 - z} \ln z + (1 + z^2) \left[\frac{\ln(1 - z)}{1 - z} \right]_+ \right. \\ & \left. + \left(\frac{8}{\epsilon^2} + \frac{3}{\epsilon} + \frac{7}{2} + \right) \delta(1 - z) - \left(2 \frac{1 + z^2}{\epsilon} + \frac{3}{2} \right) \left[\frac{1}{1 - z} \right]_+ \right) + \mathcal{O}(\epsilon) \end{aligned} \quad (2.17)$$

with three different poles. As mentioned above, one can see that the double pole proportional to the delta distribution does indeed appear with a different sign, so that it gets cancelled if one adds the virtual and real contributions. On the other hand, the single pole lacks a factor of two in order to get removed. Additionally there is the term $2 \frac{1 + z^2}{\epsilon} \left[\frac{1}{1 - z} \right]_+$, which has no counter part in (2.15). So we are still left with the divergent

³The calculation can be found in [39, chapter 20.A].

⁴This has been calculated in [2].

terms

$$\frac{2C_F}{\epsilon} \left[(1+z^2) \left[\frac{1}{1-z} \right]_+ + \frac{3}{2} \delta(1-z) \right] \equiv \frac{2}{\epsilon} P_{qq}(z), \quad (2.18)$$

where we identified the so-called splitting function $P_{qq}(z)$. This function will be of importance below. The rest of (2.17) is well behaved and we are able to add the NLO result to the leading order:

$$\begin{aligned} \hat{W}_0 &= \hat{W}_0^{\text{LO}} + \hat{W}_0^{\text{V}} + \hat{W}_0^{\text{R}} \\ &= 4\pi e_i^2 \left\{ \left[\delta(1-z) - \frac{1}{\epsilon} \frac{\alpha_s}{\pi} P_{qq}(z) \left(\frac{4\pi\mu^2}{Q^2} \right)^{\frac{\epsilon}{2}} \frac{\Gamma(1-\frac{\epsilon}{2})}{\Gamma(1-\epsilon)} \right] + \frac{\alpha_s}{2\pi} C_F \left[-\frac{3}{2} \left[\frac{1}{1-z} \right]_+ \right. \right. \\ &\quad \left. \left. + (1+z^2) \left[\frac{\ln(1-z)}{1-z} \right]_+ - \frac{1+z^2}{1-z} \ln z + 3 + 2z - \left(\frac{9}{2} + \frac{1}{3}\pi^2 \right) \delta(1-z) \right] \right\}. \end{aligned} \quad (2.19)$$

Having the pole proportional to $P_{qq}(z)$ is not a problem, since this is a parton-level calculation and for the physical prediction one needs to integrate (2.11) additionally over x . If one then uses the first line of the equation, the integral

$$\int_0^1 dz P_{qq}(z) = 0$$

appears. Due to the properties of this splitting function the integral vanishes making the physical prediction finite. Furthermore this calculation was done disregarding the mass of the quark, which cuts the infrared divergence off similar to Compton scattering in QED. However it turns out that working with differences is a more practical quantity as we will see below.

Before we insert (2.19) into (2.11), we need to expand

$$\frac{1}{\epsilon} \left(\frac{4\pi\mu^2}{Q^2} \right)^{\frac{\epsilon}{2}} = \frac{1}{\epsilon} + \frac{1}{2} \left(\ln \frac{\mu^2}{Q^2} + \ln 4\pi \right) + \mathcal{O}(\epsilon),$$

where we will drop the terms proportional to ϵ . Then concentrating on the pole and the scale dependence we get

$$W_0(x, Q) = 4\pi \sum_i e_i^2 \int_x^1 \frac{d\xi}{\xi} f_i(\xi) \left[\delta \left(1 - \frac{x}{\xi} \right) - \frac{\alpha_s}{2\pi} P_{qq} \left(\frac{x}{\xi} \right) \left(\frac{2}{\epsilon} + \ln \frac{\mu^2}{Q^2} \right) + \text{finite} \right].$$

Here one can see that if one were to choose the quark mass as the cutoff, the pole would vanish, but the logarithm of m_q/Q gets very large since $Q \gg m_q$. By taking the difference at Q and Q_0 we arrive at the simpler expression

$$W_0(x, Q) - W_0(x, Q_0) = 4\pi \sum_i e_i^2 \int_x^1 \frac{d\xi}{\xi} f_i(\xi) \left[\frac{\alpha_s}{2\pi} P_{qq} \left(\frac{x}{\xi} \right) \ln \frac{Q^2}{Q_0^2} \right], \quad (2.20)$$

2.4. QCD improved Parton Model

which finishes this 'tour de force' through the NLO calculation.

Instead of calculating differences one can use renormalisation and work in terms of renormalised quantities. Here⁵ we define

$$W_0(x, Q) \equiv 4\pi \sum_i e_q^2 f_i(x, Q)$$

for every scale Q . As hinted at the beginning of this section, the equation defines renormalised Parton distribution functions. In order to comply with (2.20) the distributions need to fulfil

$$f_i(x, \mu_1) = f_i(x, \mu) + \frac{\alpha_s}{2\pi} \int_x^1 \frac{d\xi}{\xi} f_i(x, \mu) P_{qq} \left(\frac{x}{\xi} \right) \ln \frac{\mu_1^2}{\mu^2}$$

which implies after differentiating with respect to μ :

$$\mu \frac{d}{d\mu} f_i(x, \mu) = \frac{\alpha_s}{\pi} \int_x^1 \frac{d\xi}{\xi} f_i(x, \mu) P_{qq} \left(\frac{x}{\xi} \right).$$

This is however only one part of the full answer, since we only investigated $\gamma^* q \rightarrow q$. Other contributions to $e^- P \rightarrow e^- X$ are $g \rightarrow q\bar{q}$ or $g \rightarrow gg$, because there exists the possibility to find antiquarks and gluons as well. This leads to a mixing of these PDFs under the evolution. The full form can then be given by

$$\mu \frac{d}{d\mu} \begin{pmatrix} f_i(x, \mu) \\ f_g(x, \mu) \end{pmatrix} = \sum_j \frac{\alpha_s}{\pi} \int_x^1 \frac{d\xi}{\xi} \begin{pmatrix} P_{q_i q_j} \left(\frac{x}{\xi} \right) & P_{q_i g} \left(\frac{x}{\xi} \right) \\ P_{g q_j} \left(\frac{x}{\xi} \right) & P_{gg} \left(\frac{x}{\xi} \right) \end{pmatrix} \begin{pmatrix} f_j(\xi, \mu) \\ f_g(\xi, \mu) \end{pmatrix}. \quad (2.21)$$

This is the well-known DGLAP evolution equation after Dokshitzer [10], Gribov, Lipatov [15], Altarelli and Parisi [1]. The splitting functions⁶ are defined by

$$\begin{aligned} P_{qq}(z) &= C_F \left(\frac{1+z^2}{[1-z]_+} + \frac{3}{2} \delta(1-z) \right) \\ P_{qg}(z) &= T_F (z^2 + (1-z)^2) \\ P_{gq}(z) &= C_F \left(\frac{1+(1-z)^2}{z} \right) \\ P_{gg}(z) &= 2C_A \left(\frac{z}{[1-z]_+} + \frac{1-z}{z} + z(1-z) \right) + \frac{\beta_0}{2} \delta(1-z) \end{aligned}$$

with $\beta_0 = \frac{11}{3}C_A - \frac{4}{3}T_F n_f$ as the first term of the renormalisation group equation of the

⁵This is known as the DIS-scheme, where everything gets absorbed into the PDFs. An alternative scheme, the $\overline{\text{MS}}$ scheme, incorporates only the pole into the distributions. Further information can be found in [6, chapter IV.B].

⁶These are the splitting functions at leading order, whilst the processes calculated were of next to leading order. Currently the functions are known at NNLO (two levels deeper than the functions given here); see [27] and [28]).

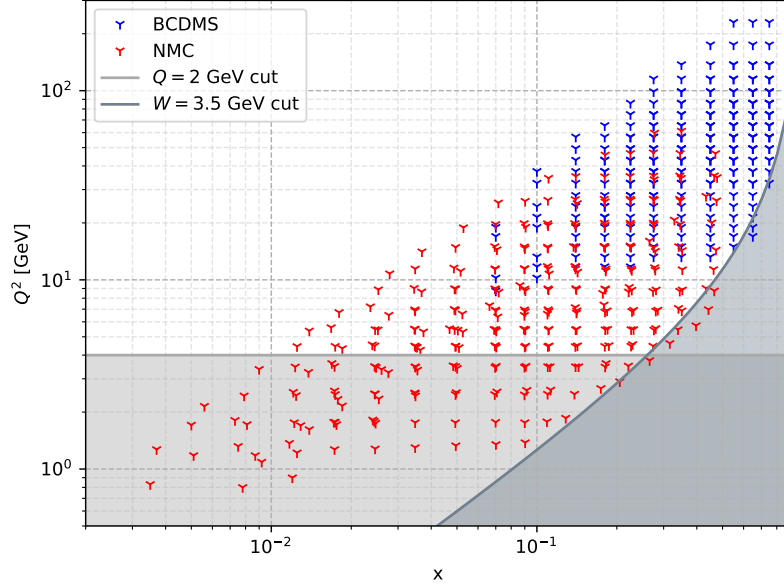


Figure 2.5: The kinematic coverage of the two data sets (BCDMS: [5] and [4]; NMC: [3]). The grey lines indicate the Q_{\min} and W_{\min} cuts.

strong coupling. A derivation can be found in [30, chapter 17]. With this equation at hand it is possible to calculate the PDFs at every scale, if the PDFs' x dependence is fixed at an initial scale via experimental measurements.

2.5. F_2 measurements and PDF parametrisation

As experimental measurements results from the New Muon Collaboration (NMC [3]) and the BCDMS Collaboration ([5] and [4]) are taken. These are agreeing measurements for the F_2 structure function of the proton and deuteron. They were measured at the Super Proton Synchrotron (SPS) at CERN and are shown in fig. 2.5. We approximate the deuteron as the sum of a proton and neutron, where we neglect the nuclear effects due to the weak binding.

In the sections above the calculations were given within approximations. Therefore we will restrict the available data to regions, where the approximations definitely hold, because we do not want to fit the PDFs to effects, which they should not model in the first place. We neglect data points below $Q \leq 2\text{GeV}$ in order to avoid higher twist corrections ($M \ll Q$), shown as the light grey region. A secondary cut is applied for the invariant mass $W^2 = (p + q)^2$. The line of Q^2 as a function of x is given by

$$Q_{\text{cut}}^2(x) = (W^2 - M^2) \frac{x}{1 - x}.$$

We want to make sure we are in the region of deep inelastic scattering instead of elastic scattering or resonances. Therefore we remove data points with $W \leq 3.5\text{GeV}$ (dark grey

2.5. F_2 measurements and PDF parametrisation

line). In total 992 data points are taken into account.

In order to predict F_2 values the Parton densities are needed as explained in the previous section. The distributions can be calculated as a function of Q , if the x dependence is fixed at some initial scale Q_0 . The x dependence however can not be calculated with perturbative QCD, therefore a function is proposed at Q_0 with open parameters, which are then to be determined by comparison to experimental measurements. The nCTEQ collaboration used the following parametrisation

$$xf_i(x, Q_0) = p_0 x^{p_1} (1 - x)^{p_2} e^{p_3 x} (1 + e^{p_4 x})^{p_5} \quad (2.22)$$

at $Q_0 = 1.3\text{GeV}$ for the global nuclear analysis [22]. Since this thesis uses the nCTEQ codebase (written in **C++** and **Fortran**), we will use it also. The predictions for data at a different scale Q is then done via the DGLAP evolution at NLO. Since the evolution (2.21) connects a derivative with an integration the prediction is numerically very expensive. In fact every evaluation of the correlated χ^2 -function, with the data set from above, takes about 50 seconds. Due to the nature of the parameters \mathbf{p} and the high costs of every χ^2 evaluation, fitting turns into a non-trivial task.

The data set is well understood and serves as a testing ground for fitting algorithms. In fact it is possible to fit the data well by only using u-valence and d-valence distributions by making use of the isospin symmetry between the proton and neutron. The parameters $p_0^{u_v, d_v}$ are not part of the fit and determined via the normalisation equations (see (2.6) to (2.8)), resulting in ten parameters to be determined.

3. Bayesian inference and the χ^2 function

The last section dealt with theoretical predictions for experimental measurements and ended with parameters of the theory which are to be determined. The overall goal is to solve for the parameter values starting from the measurements. Therefore we need a figure of merit to display the goodness of a set of parameter values. The theory of Bayesian inference yields a framework where one can calculate how probable a set of parameter values is given the measurements. The starting point is the Bayesian theorem: The components of \mathbf{D} represent the individual data points with their uncertainty σ , respectively. Then the Bayesian Theorem for the model parameters \mathbf{X} states:

$$p(\mathbf{X}|\mathbf{D}, I) = \frac{p(\mathbf{D}|\mathbf{X}, I)p(\mathbf{X}|I)}{p(\mathbf{D}|I)}, \quad (3.1)$$

where I encodes the information that we have on the experimental measurements (their variance σ) and on the model parameters, namely the function $F(\{X_i\})$ which predicts the experimental values for a given set of parameters. Usually one is not interested in the absolute probability for the parameters \mathbf{X} given the data \mathbf{D} , but in the best estimate, which translates to the maximum of $p(\mathbf{X}|\mathbf{D}, I)$. Therefore the absolute probability for the given Data is not relevant⁷ and we only denote the proportionality:

$$p(\mathbf{X}|\mathbf{D}, I) \propto p(\mathbf{D}|\mathbf{X}, I)p(\mathbf{X}|I) \quad (3.2)$$

Now we have to find the likelihood $p(\mathbf{D}|\mathbf{X}, I)$ and the prior $p(\mathbf{X}|I)$ with the knowledge encoded in I , without biasing ourselves. To put this idea of maximum ignorance into mathematical terms, we use the technique of maximum entropy, which are explained in [42] and [23], to generate these probabilities.

One has to note that this (and the following) section is in the literature often explained in terms of measure theory in order to incorporate discontinuous and less well behaved distributions. Due to the fact that every distribution of interest in this thesis is of continuous nature and in principle well behaved, we simplified the notation.

3.1. The principle of maximum entropy

We use the continuous definition of the Shannon entropy S . It is in fact the negative of the Kullback-Leibler divergence, which is a measure for information gained by learning that variables are distributed as $p(\mathbf{X}|I)$ instead of a probability distribution $\tilde{p}(\mathbf{X})$, which does not have the information I incorporated. We make use of this fact by maximising the entropy with respect to $p(\mathbf{X}|I)$ under constraints given by I . This determines the probability distribution (pdf) by minimising the information gained meaning that no

⁷If one is interested in the absolute value, one can get the normalisation factor from the integral over the whole domain $\int p(\mathbf{X}|\mathbf{D}, I)d^M\mathbf{X} = 1$, albeit this is close to impossible in most cases.

3.2. Location and scale parameters

additional bias is added. S is given by

$$S = - \int p(\mathbf{X}) \log \left(\frac{p(\mathbf{X})}{m(\mathbf{X})} \right) d^M \mathbf{X}. \quad (3.3)$$

Following the reasoning above the Shannon entropy measures the difference in information between $p(\mathbf{X})$ and a so-called measure $m(\mathbf{X})$, which is equal to the probability distribution function expressing complete ignorance of \mathbf{X} .

In order to explain the general concept we prove the statement above for a one dimensional case, where the distributions are defined in a range from 0 to R . The only information we have about $p(X)$ is its normalisation:

$$\int_0^R p(X) dX = 1$$

To compute $p(X)$ without biasing ourselves, we maximise the entropy under the normalisation constraint using Lagrange multipliers and functional derivatives.

$$\begin{aligned} Q &= - \int_0^R p(X) \log \left(\frac{p(X)}{m(X)} \right) dX + \lambda \left(1 - \int_0^R p(X) dX \right) \\ \frac{\delta Q}{\delta p(X)} &= - \log \left(\frac{p(X)}{m(X)} \right) - 1 - \lambda = 0 \\ \Rightarrow p(X) &= m(X) \exp(-1 - \lambda) \end{aligned}$$

The Lagrange parameter λ can now be determined from the constraint on $p(X)$, by using the normalisation of $m(X)$.

$$\begin{aligned} 1 &= \int_0^R p(X) dX \\ &= \exp(-1 - \lambda) \int_0^R m(X) dX \\ &= \exp(-1 - \lambda) \quad \Rightarrow \lambda = -1 \end{aligned}$$

Finally we get $p(X|\text{normalisation}) = m(X)$ as stated above. Although we have an (intuitive) understanding of $m(X)$, it is yet unclear how to arrive at an expression of it.

3.2. Location and scale parameters

In certain cases it is possible to construct a prior probability from general arguments. Suppose the parameter X_i from \mathbf{X} is a location parameter. This implicitly means that the pdf should be invariant under translational transformations in this direction, since

the total probability should not be dependent on the choice of the coordinate system. Mathematically written:

$$p(X_i|I)dX_i = p(X_i + x_0|I)d(X_i + x_0)$$

which is solved by $p(X_i|I) = \text{const}$ and can be determined via normalisation. Thus complete ignorance of a location parameter is represented by the assignment of a uniform pdf.

For a scale parameter however the pdf should be invariant with respect to stretching or shrinking of the X_i axis:

$$p(X_i|I)dX_i = p(\beta X_i|I)d(\beta X_i)$$

This equation is solved via $p(X_i|I) = \frac{\text{const}}{X_i}$, which is called *Jeffreys' prior* since it was first suggested by him in 1939. This pdf is uniform for the logarithm of X_i , since $p(X_i)dX_i \rightarrow p(\exp(U_i))\exp(U_i)dU_i = \text{const } dU_i$.

3.3. Mean and variance

Often the case occurs where we have testable information about a parameter, such as the mean μ and variance σ . Normally experimental measurements are given that way. The probabilities for the parameters are then given as a Gaussian, which we are now capable of proving. Here we focus on only one measurement, such that we do not have to care about correlated measurements. The latter case will be discussed in section 3.5. The constraints are

$$1 = \int_a^b p(X)dX \tag{3.4}$$

$$\sigma^2 = \int_a^b (x - \mu)^2 p(X)dX, \tag{3.5}$$

where we left the integrals within a range $[a, b]$ to be more general (obviously $\mu \in [a, b]$). We say the parameter is a location parameter, such that we can set the measure uniform ($m(X) = m = \text{const}$). We do not attempt to normalise m , since it is not possible for every potential value of a or b . However, we will see that $p(X)$ does not depend on m justifying our ignorance.

The functional Q to be maximised yields

$$Q = - \int_a^b p(X) \ln \left(\frac{p(X)}{m} \right) dX + \lambda_1 \left(1 - \int_a^b p(X)dX \right) + \lambda_2 \left(\sigma^2 - \int_a^b (X - \mu)^2 p(X)dX \right)$$

3.3. Mean and variance

$$\begin{aligned} \Rightarrow \frac{\delta Q}{\delta p(X)} &= -\ln\left(\frac{p(X)}{m}\right) - 1 - \lambda_1 - \lambda_2(X - \mu)^2 \\ \Leftrightarrow p(X) &= m \exp\left(-1 - \lambda_1 - \lambda_2(X - \mu)^2\right). \end{aligned} \quad (3.6)$$

Now we can evaluate the constraints and solve the equation for $\lambda_{1,2}$. The resulting integrals are solvable in terms of the error function $\text{Erf}(z) = \frac{1}{\sqrt{\pi}} \int_{-z}^z \exp(-t^2) dt$.

$$\begin{aligned} \Rightarrow 1 &= \frac{m}{2} e^{-1-\lambda_1} \sqrt{\frac{\pi}{\lambda_2}} \left[\text{Erf}(\sqrt{\lambda_2}(\mu - a)) - \text{Erf}(\sqrt{\lambda_2}(\mu - b)) \right] \\ &=: \frac{m}{2} e^{-1-\lambda_1} \sqrt{\frac{\pi}{\lambda_2}} K1(\lambda_2, \mu, a, b) \\ \Rightarrow \sigma^2 &= \frac{m}{4\sqrt{\lambda_2}^3} e^{-1-\lambda_1} \left[2\sqrt{\lambda_2}(a - \mu)e^{-\lambda_2(a-\mu)^2} + 2\sqrt{\lambda_2}(b - \mu)e^{-\lambda_2(b-\mu)^2} + \sqrt{\pi}K1 \right] \\ &=: \frac{m}{4\sqrt{\lambda_2}^3} e^{-1-\lambda_1} \left[K2(\lambda_2, \mu, a, b) + \sqrt{\pi}K1 \right] \end{aligned}$$

We have defined $K1(\lambda_2, \mu, a, b)$ and $K2(\lambda_2, \mu, a, b)$ to abbreviate the terms containing the bounds a and b . Eliminating λ_1 results in a transcendental equation for λ_2 due to the boundaries.

$$\lambda_2 = \frac{1}{2\sigma^2} \left(1 + \frac{1}{\sqrt{\pi}} \frac{K2(\lambda_2, \mu, a, b)}{K1(\lambda_2, \mu, a, b)} \right) \quad (3.7)$$

As one can see from the graphical approach in fig. 3.1, there is a unique solution to this equation. We now consider only the ratio $\left| \frac{K2}{\sqrt{\pi}K1} \right|$ with $\mu = 2$ and $\sigma = 1$ at the value of λ_2 which solves (3.7) as a function of the interval length d , where we set the interval symmetrical around μ ($a = \mu - d/2$ and $b = \mu + d/2$). The values of λ_2 are obtained from a numerical algorithm. As one can see in fig. 3.2, the ratio goes strongly to zero. The blue line follows the linear axis (left) and the green line the logarithmic axis. If one sets d too low no solution is possible besides $\lambda_2 = 0$, which is nonsensical since it represents a solution, where there is no second constraint. In conclusion we may take the boundaries of the parameter towards infinity, since it is in very good agreement with the exact result. Even if the parameter cannot take negative values by its definition. In fig. 3.1 it is also possible to verify that this argumentation holds for the experimental measurements.

Applying $a \rightarrow -\infty, b \rightarrow \infty$ to $K1$ and $K2$ yields:

$$\begin{aligned} K1(\lambda_2, \mu, a \rightarrow -\infty, b \rightarrow \infty) &\rightarrow 2 \\ K2(\lambda_2, \mu, a \rightarrow -\infty, b \rightarrow \infty) &\rightarrow 0 \\ \lambda_2 &= \frac{1}{2\sigma^2}. \end{aligned}$$

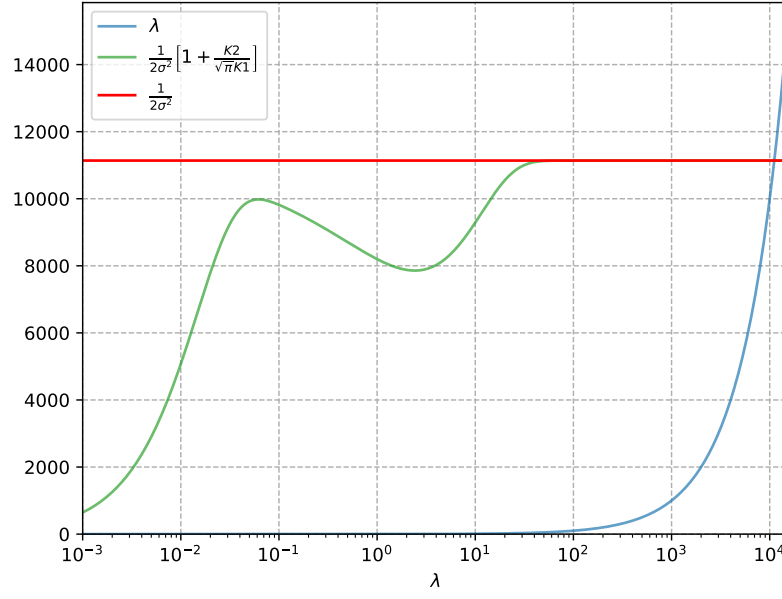


Figure 3.1: The graphic shows the unique solution to (3.7). We have set $\mu = 0.38165$, $\sigma = 0.0067$, $a = 0$ and $b = 10$. This is one of the measurements from [5].

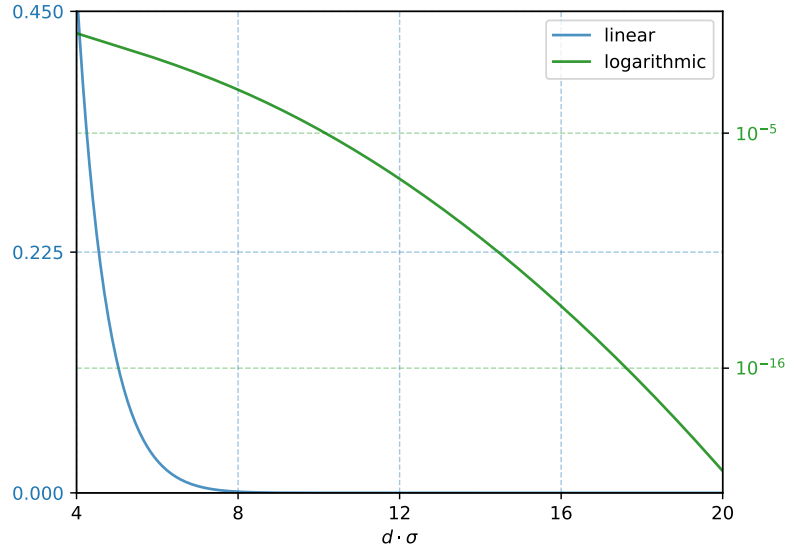


Figure 3.2: The ratio $\left| \frac{K2}{\sqrt{\pi}K1} \right|$ with $\mu = 2$ and $\sigma = 1$ at the value of λ_2 which solves (3.7) as a function of the interval length d . The integration interval has been set to $[\mu - d/2, \mu + d/2]$. This ratio drops stronger than exponentially as seen from the logarithmic depiction.

3.4. Likelihood for several uncorrelated measurements

It is now also possible to solve (3.6) for λ_1 resulting in

$$m \exp(-1 - \lambda_1) = \sqrt{\frac{\lambda_2}{\pi}} = \frac{1}{\sqrt{2\pi}\sigma}.$$

and finally

$$p(X) = m \exp(-1 - \lambda_1 - \lambda_2(X - \mu)^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(X - \mu)^2}{2\sigma^2}\right) \quad (3.8)$$

which does not depend on m and is a normalized Gaussian with mean μ and variance σ as stated above.

3.4. Likelihood for several uncorrelated measurements

As stated above the probability for the parameters \mathbf{X} given the data \mathbf{D} can be denoted as (see (3.2))

$$p(\mathbf{X}|\mathbf{D}, I) \propto p(\mathbf{D}|\mathbf{X}, I)p(\mathbf{X}|I).$$

If we now assume no further information about our parameters, we might indicate this by a flat prior $p(\mathbf{X}|I) = \text{const.}$ Thus the prior can be absorbed into the normalisation and we end up with

$$p(\mathbf{X}|\mathbf{D}, I) \propto p(\mathbf{D}|\mathbf{X}, I).$$

The probability to find a single data point given the parameters is independent for uncorrelated measurements. The total probability to find every data point is then simply given by the product over the probabilities for the individual measurements:

$$p(\mathbf{D}|\mathbf{X}, I) = \prod_{k=1}^N p(D_k|\mathbf{X}, I).$$

This fact can also be easily derived as a maximal entropy solution.

Experimental measurements are usually given by the best estimate and a corresponding error ($\hat{=}$ variance). With the discussion from the last section in mind, we therefore conclude that the probability for a single measurement is given by a Gaussian with the mean as the best estimate and the variance as the corresponding error:

$$p(D_k|\mathbf{X}, I) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{(F_k - D_k)^2}{2\sigma_k^2}\right).$$

Instead of X_k we denoted F_k , which should be the theory prediction evaluated at the corresponding measurement $F_k = F(\mathbf{X}_k)$, since it is obviously not the set of parameters which should be equal to the data point but the theory prediction.

Finally the probability for the parameters is now given by

$$\begin{aligned}
 p(\mathbf{X}|\mathbf{D}, I) &\propto \prod_{k=1}^N \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{(F_k - D_k)^2}{2\sigma_k^2}\right) \\
 &= \left[\prod_{k=1}^N \frac{1}{\sqrt{2\pi}\sigma_k} \right] \exp\left(-\sum_{l=1}^N \frac{(F_l - D_l)^2}{2\sigma_l^2}\right) \\
 &= \left[\prod_{k=1}^N \frac{1}{\sqrt{2\pi}\sigma_k} \right] \exp\left(-\frac{1}{2}\chi_u^2\right). \tag{3.9}
 \end{aligned}$$

In the last step we inserted the χ_u^2 -function for uncorrelated data. And for a optimal parameter estimation one wants to find the point where the probability for the parameters is at its maximum, which is the minimum of the χ_u^2 -function. This is how Bayesian probability theory validates a χ_u^2 -fit (least squares) as a maximum likelihood estimation.

3.5. Likelihood for several correlated measurements

In the case above we stated that experimental measurements were usually given by the best estimate and a corresponding error. This fact is not entirely true since one often has to deal with correlated measurements. This breaks the factorisation of the probabilities for the individual data points. Instead one has to deal with a matrix C which can be defined via

$$C_{ij} = \int (x_i - \mu_i)(x_j - \mu_j)p(\mathbf{X})d^N\mathbf{X} =: \int \delta x_i \delta x_j p(\mathbf{X})d^N\mathbf{X} \tag{3.10}$$

making use of the expected value μ_i for a parameter and its probability distribution $p(\mathbf{X})$. Explicitly the diagonal parts of the matrix are the squares of the variances σ_i and the non-diagonal parts are the covariance contributions σ_{ij}^2 , giving C the name *covariance matrix*. As one can see from the definition, C is symmetric and connects different measurements. As above we can use (3.10) as a constraint to construct a probability distribution via the maximum entropy ansatz. The functional then has the form

$$\begin{aligned}
 Q = & - \int_{-\infty}^{\infty} p(\mathbf{X}) \ln\left(\frac{p(\mathbf{X})}{m}\right) d^N\mathbf{X} + \lambda_0 \left(1 - \int_{-\infty}^{\infty} p(\mathbf{X}) d^N\mathbf{X}\right) \\
 & + \sum_{i=1}^N \lambda_i \left(\mu_i - \int_{-\infty}^{\infty} x_i p(\mathbf{X}) d^N\mathbf{X}\right) \\
 & + \sum_{i,j=1}^N \Lambda_{ij} \left(C_{ij} - \int_{-\infty}^{\infty} \delta x_i \delta x_j p(\mathbf{X}) d^N\mathbf{X}\right), \tag{3.11}
 \end{aligned}$$

3.5. Likelihood for several correlated measurements

which results in

$$p(\mathbf{X}) = m \exp \left(-\lambda_0 - 1 - \sum_{i=1}^N \lambda_i x_i - \sum_{i,j=1}^N \Lambda_{ij} \delta x_i \delta x_j \right). \quad (3.12)$$

In order to solve the equation for the different λ 's, we begin with the normalisation constraint. This is however manageable analytically since we only have to solve a multidimensional integral. We first perform the substitution $\delta x_i = y_i$ to solve the Gaussian integral with the identity (B.2):

$$\begin{aligned} 1 &= \int_{-\infty}^{\infty} p(\mathbf{X}) d^N \mathbf{X} = \int_{-\infty}^{\infty} m \exp \left(-\lambda_0 - 1 - \sum_{i=1}^N \lambda_i (y_i + \mu_i) - \sum_{i,j=1}^N \Lambda_{ij} y_i y_j \right) d^N \mathbf{Y} \\ &= m e^{-\lambda_0 - 1 - \sum_{i=1}^N \lambda_i \mu_i} \int_{-\infty}^{\infty} \exp \left(\sum_{i=1}^N \lambda_i y_i - \sum_{i,j=1}^N \Lambda_{ij} y_i y_j \right) d^N \mathbf{Y} \\ &= m \sqrt{\frac{(2\pi)^N}{\det(2\Lambda)}} e^{-\lambda_0 - 1 - \sum_{i=1}^N \lambda_i \mu_i + \frac{1}{4} \sum_{i=1}^N \Lambda_{ij}^{-1} \lambda_i \lambda_j} \\ &\Leftrightarrow e^{\lambda_0 + 1} = m \sqrt{\frac{(2\pi)^N}{\det(2\Lambda)}} e^{-\sum_{i=1}^N \lambda_i \mu_i + \frac{1}{4} \sum_{i=1}^N \Lambda_{ij}^{-1} \lambda_i \lambda_j} \end{aligned} \quad (3.13)$$

This result will abbreviate the following constraints, not only because the right hand side collapses to $e^{\lambda_0 + 1}$, but also since we can reuse the way of solving the integral.

The constraint given by the mean values is easily calculated by making use of

$$\int_{-\infty}^{\infty} x_i p(\mathbf{X}) d^N \mathbf{X} = -\frac{\partial}{\partial \lambda_i} \int_{-\infty}^{\infty} p(\mathbf{X}) d^N \mathbf{X}$$

so we only have to calculate the derivative of the third line of (3.13) with respect to λ_i and insert our result for λ_0 :

$$\begin{aligned} \mu_i &= -m \sqrt{\frac{(2\pi)^N}{\det(2\Lambda)}} e^{-\lambda_0 - 1} \frac{\partial}{\partial \lambda_i} \exp \left(-\sum_{i=1}^N \lambda_i \mu_i + \frac{1}{4} \sum_{i=1}^N \Lambda_{ij}^{-1} \lambda_i \lambda_j \right) \\ &= m \sqrt{\frac{(2\pi)^N}{\det(2\Lambda)}} e^{-\lambda_0 - 1 - \sum_{i=1}^N \lambda_i \mu_i + \frac{1}{4} \sum_{i=1}^N \Lambda_{ij}^{-1} \lambda_i \lambda_j} \left(\mu_i - \frac{1}{4} (\Lambda_{ij}^{-1} + \Lambda_{ji}^{-1}) \lambda_j \right) \\ &= \mu_i - \frac{1}{4} (\Lambda_{ij}^{-1} + \Lambda_{ji}^{-1}) \lambda_j \end{aligned} \quad (3.14)$$

This yields $\lambda_i = 0 \forall i$.⁸ The last constraint however is not that easily dealt with. First

⁸Otherwise we would have to set a constraint on Λ , but this matrix is to be set by the last constraint.

we use the ideas from above.

$$\begin{aligned}
 C_{ij} &= \sqrt{\frac{\det(2\Lambda)}{(2\pi)^N}} \int_{-\infty}^{\infty} x_i x_j \exp\left(-\sum_{i,j=1}^N \Lambda_{ij} \delta x_i \delta x_j\right) d^N \mathbf{X} \\
 &= \sqrt{\frac{\det(2\Lambda)}{(2\pi)^N}} \int_{-\infty}^{\infty} y_i y_j \exp\left(-\sum_{i,j=1}^N \Lambda_{ij} y_i y_j\right) d^N \mathbf{Y} \\
 &= -\sqrt{\frac{\det(2\Lambda)}{(2\pi)^N}} \frac{\partial}{\partial \Lambda_{ij}} \int_{-\infty}^{\infty} \exp\left(-\sum_{i,j=1}^N \Lambda_{ij} y_i y_j\right) d^N \mathbf{Y} \\
 &= -\sqrt{\frac{\det(2\Lambda)}{(2\pi)^N}} \frac{\partial}{\partial \Lambda_{ij}} \sqrt{\frac{(2\pi)^N}{\det(2\Lambda)}} = -\sqrt{\det(2\Lambda)} \frac{\partial}{\partial \Lambda_{ij}} \frac{1}{\sqrt{\det(2\Lambda)}}
 \end{aligned}$$

The resulting derivative can be computed via *Jacobi's formula* (see (B.3) and (B.4)).

$$\begin{aligned}
 \Rightarrow C_{ij} &= \frac{1}{2\det(2\Lambda)} \frac{\partial}{\partial \Lambda_{ij}} \det(2\Lambda) \\
 &= \frac{1}{2\det(2\Lambda)} 2(\text{adj}^T 2\Lambda)_{ij} \\
 &= (2\Lambda)_{ij}^{-1} \\
 \Leftrightarrow \Lambda_{ij} &= \frac{1}{2} C_{ij}^{-1}
 \end{aligned} \tag{3.15}$$

Combining all results we get

$$e^{\lambda_0+1} = m \sqrt{(2\pi)^N \det(C)}$$

and finally

$$p(\mathbf{X}) = \frac{1}{\sqrt{(2\pi)^N \det(C)}} \exp\left(-\frac{1}{2} \sum_{i,j=1}^N \delta x_i C_{ij}^{-1} \delta x_j\right). \tag{3.16}$$

The term in the exponent is (one half) of the generalised χ^2 function for correlated measurements if we replace the expectation values with the results of the measurements and replace the variable \mathbf{X} with the theory prediction. The covariance matrix is therefore to be identified with the correlated errors given by the experiment.

3.6. Final Bayesian model

In the sections above we showed how one can use the Entropy definition of (3.3) to arrive at probability density functions if we have testable information. The Bayesian Theorem allows us to explicitly write a formula which assigns a probability density to a specific parameter set \mathbf{X} for our model. Hence we can tune our model parameters to the point where this density is maximised. For a list of correlated measurements we showed

3.6. Final Bayesian model

the connection between the generalised χ^2 function and the likelihood. If we now have no preference for our parameters we might assign a uniform prior to them yielding the overall probability density function

$$p(\mathbf{X}|\mathbf{D}, C, I) = \frac{1}{Z} \exp\left(-\frac{1}{2}\chi^2(\mathbf{X}, \mathbf{D}, C)\right) \quad (3.17)$$

with a normalisation constant Z . This formula is maximised if one minimises the χ^2 -function. This gives rise to the whole lot of fitting algorithms concerning χ^2 -minimisation. This framework however is more powerful since we now have a systematic way of including additional information about the parameters. In the latter work we will use the full probability density function to build up a fitting algorithm. Therefore the algorithm can easily be changed if new information is available.

Additionally one has to note that this way of assigning probability densities does not rely on the experiments returning Gaussian error estimates. The solution displayed above is the most conservative model one can use if the only information given is the mean and variance. If there were more information given, the entropy would return a different density function. Further comments on this topic can be found in [20, chapter 5].

There are several problems one faces when trying to derive the best parameter set from this distribution. Evaluating the distribution on a grid is not an applicable approach, since the theoretical predictions are very costly (see section 2.5). Furthermore we want to construct an algorithm which could also be applied to more experimental measurements and more theory parameters. In a full analysis the parameter space can easily grow to $\mathcal{O}(100)$ dimensions making grids extremely inefficient.

The next idea would be direct sampling from the distribution via Monte Carlo techniques, but in order to do so one needs the normalisation constant Z . This in return can only be calculated on a grid, so this idea must also be discarded. It is however possible to write Monte Carlo algorithms, which can sample from a distribution by only considering differences between two points in the parameter space. For this method the normalisation is not needed. They have also the benefit of not relying on any approximation, so that the investigated distribution can in principle be as complicated as possible.

There is however a disadvantage: A new sample will depend on the current sample, which makes the statistical analysis complicated. Usually this is circumvented by generating a large amount of samples, but as discussed this is not helpful in our case. Thus we will investigate these series of samples, which are called Markov Chains, in detail to decrease the amount of samples needed to its minimum.

4. Theory of Markov Chains

As explained above the desired probability distribution is too complicated to allow for direct sampling or evaluating the density on a grid. We will therefore use a simpler ansatz, where new samples are taken in the vicinity of the last sample following a certain rule. This however introduces dependencies between the points, which are to be treated with care. A series of said samples is a so-called Markov Chain. In this section we want to introduce Markov Chains with their theoretical background. In particular we will rephrase the task of finding the best parameter set as computing expectation values from the samples. In order to extract the properties of the given chains we will also investigate their time series analysis.

In the definition of Markov Chains and the argumentation why these chains are useful for solving the above problem we mostly follow [29, chapter 3] and [7, chapter 1]. We also want to introduce a tool to check whether a given chain has the desired properties, in particular whether it has reached the invariant distribution. Therefore we follow the argumentation from (again) [7, chapter 1] and [21, chapter 5] and we will use the integrated autocorrelation time τ . It is most convenient that it can also be used to rate the efficiency of the algorithm generating the Markov Chain. The implementation and discussion on how to compute τ is based on [44] and is called the 'Gamma-Method'.

4.1. Properties of Markov Chains

A Markov Chain is a series of random variables $\mathbf{X}^0, \mathbf{X}^1, \mathbf{X}^2, \dots$ up to a total number of n , where the distribution of the $(n+1)$ th element \mathbf{X}^n only depends on the last element \mathbf{X}^{n-1} . Formally:

$$p(\mathbf{X}^n | \mathbf{X}^{n-1}, \{\mathbf{X}^t : t \in \Omega\}) = p(\mathbf{X}^n | \mathbf{X}^{n-1}), \quad (4.1)$$

with Ω being any subset of $\{0, \dots, n-2\}$. So it is only natural to view the indices n as a time (later Monte Carlo time). The range of all elements in \mathbf{X} is called state space and is always expected to be continuous in this context. The full Markov Chain can be defined by giving a distribution for the initial state \mathbf{X}^0 and a transition probability density $T_n(\mathbf{X}, \mathbf{X}')$, which denotes the probability at a step n to jump from the state \mathbf{X} to \mathbf{X}' . If the transition probability does not depend on the time n we say the Markov chain is homogeneous (or stationary/at equilibrium). Given a initial distribution and a transition kernel T_t it is possible to calculate the probabilities for a state at a point in time t :

$$p_n(\mathbf{X}) = \int p_{n-1}(\mathbf{X}') T_n(\mathbf{X}', \mathbf{X}) d\mathbf{X}'. \quad (4.2)$$

This equation gives rise to the so called invariant (or stationary) distribution $\pi(\mathbf{X})$ over the state space. Once reached the invariant distribution persists for ever:

$$\pi(\mathbf{X}) = \int \pi(\mathbf{X}') T_n(\mathbf{X}', \mathbf{X}) d\mathbf{X}'. \quad (4.3)$$

We are especially interested in homogenous Markov Chains, which are ergodic meaning they reach the invariant distribution if one lets the time go to infinity regardless of the

4.1. Properties of Markov Chains

choice of initial probabilities $p_0(\mathbf{X})$. This is the connection to the probability distribution for the parameters: We want to construct a Markov Chain which converges to this very probability distribution, so we can sample from it. This is useful since there exists a fundamental theorem stating that every expectation value computed with respect to a probability distribution can be approximated by $p_n(\mathbf{X})$ if $\pi(\mathbf{X})$ is the desired probability distribution. Furthermore if $n \rightarrow \infty$ the expectation value taken with respect to $p_n(\mathbf{X})$ converges to the aimed expectation value. These statements are made precise by proving the above mentioned theorem (slightly modified from [29] chapter 3.3):

Fundamental Theorem 1 *If a homogeneous Markov Chain on a continuous state space with transition probabilities $T(\mathbf{X}, \mathbf{X}')$ has π as an invariant distribution and*

$$\nu = \min_{\mathbf{X}} \min_{\mathbf{X}': \pi(\mathbf{X}') > 0} \frac{T(\mathbf{X}, \mathbf{X}')}{\pi(\mathbf{X}')} > 0 \quad (4.4)$$

then the Markov Chain is geometrically ergodic, i.e. regardless of the initial probabilities $p_0(\mathbf{X})$

$$\lim_{n \rightarrow \infty} p_n(\mathbf{X}) = \pi(\mathbf{X}) \quad (4.5)$$

for all \mathbf{X} . Furthermore if $a(\mathbf{X})$ is any (bounded) real valued function of the state, then the expectation value of $a(\mathbf{X})$ with respect to the distribution $p_n(\mathbf{X})$, written $E_n[a]$, converges to its expectation value with respect to $\pi(\mathbf{X})$, written $\langle a \rangle$, with

$$|\langle a \rangle - E_n[a]| \leq (1 - \nu)^n \max_{\mathbf{X}, \mathbf{X}'} |a(\mathbf{X}) - a(\mathbf{X}')| \quad (4.6)$$

The proof first demonstrates that the distribution $p_n(\mathbf{X})$ can be written as a combination of the invariant distribution and another arbitrary distribution, which we will call r_n here. As n goes to infinity the proportion of $\pi(\mathbf{X})$ in $p_n(\mathbf{X})$ approaches one. In particular we will show

$$p_n(\mathbf{X}) = (1 - (1 - \nu)^n) \pi(\mathbf{X}) + (1 - \nu)^n r_n(\mathbf{X}) \quad (4.7)$$

by induction. Note here that $\nu \leq 1$ since $T(\mathbf{X}, \mathbf{X}')$ cannot be greater than $\pi(\mathbf{X}')$ for all \mathbf{X}' , because they must obey (4.3). To start the induction we choose $r_0 = p_0$. The verification for $n \rightarrow n + 1$ begins with the definition for p_{n+1} (see (4.2)):

$$\begin{aligned} p_{n+1}(\mathbf{X}) &= \int p_n(\mathbf{X}') T(\mathbf{X}', \mathbf{X}) d\mathbf{X}' \\ &= \int [(1 - (1 - \nu)^n) \pi(\mathbf{X}') + (1 - \nu)^n r_n(\mathbf{X}')] T(\mathbf{X}', \mathbf{X}) d\mathbf{X}' \\ &= (1 - (1 - \nu)^n) \pi(\mathbf{X}) + (1 - \nu)^n \int r_n(\mathbf{X}') T(\mathbf{X}', \mathbf{X}) d\mathbf{X}' \end{aligned}$$

At this point we perform a clever insertion of $\nu\pi(\mathbf{X}) - \nu\pi(\mathbf{X})$ and use the normalisation of r_n over the state space.

$$\begin{aligned}
 p_{n+1}(\mathbf{X}) &= (1 - (1 - \nu)^n)\pi(\mathbf{X}) \\
 &\quad + (1 - \nu)^n \int r_n(\mathbf{X}') [T(\mathbf{X}', \mathbf{X}) + \nu\pi(\mathbf{X}) - \nu\pi(\mathbf{X})] d\mathbf{X}' \\
 &= (1 - (1 - \nu)^n)\pi(\mathbf{X}) + (1 - \nu)^n \nu\pi(\mathbf{X}) \\
 &\quad + (1 - \nu)^n \int r_n(\mathbf{X}') [T(\mathbf{X}', \mathbf{X}) - \nu\pi(\mathbf{X})] d\mathbf{X}' \\
 &= (1 - (1 - \nu)^{n+1})\pi(\mathbf{X}) + (1 - \nu)^{n+1} \int r_n(\mathbf{X}') \frac{T(\mathbf{X}', \mathbf{X}) - \nu\pi(\mathbf{X})}{1 - \nu} d\mathbf{X}' \\
 &= (1 - (1 - \nu)^{n+1})\pi(\mathbf{X}) + (1 - \nu)^{n+1} r_{n+1}(\mathbf{X})
 \end{aligned} \tag{4.8}$$

The last line defines the computation of r_{n+1} at each time step:

$$r_{n+1}(\mathbf{X}) = \int r_n(\mathbf{X}') \frac{T(\mathbf{X}', \mathbf{X}) - \nu\pi(\mathbf{X})}{1 - \nu} d\mathbf{X}' \tag{4.9}$$

for which the normalisation can be easily verified and from (4.4) we conclude $r_{n+1}(\mathbf{X}) > 0 \forall \mathbf{X}$ making r_{n+1} a proper probability distribution. This establishes (4.5) and the Markov Chain is ergodic.

The second part of the Theorem also makes use of the rewriting of $p_n(\mathbf{X})$:

$$\begin{aligned}
 |\langle a \rangle - E_n[a]| &= \left| \int a(\mathbf{X})\pi(\mathbf{X})d\mathbf{X} - \int a(\mathbf{X})p_n(\mathbf{X})d\mathbf{X} \right| \\
 &= (1 - \nu)^n \left| \int a(\mathbf{X})[\pi(\mathbf{X}) - r_n(\mathbf{X})]d\mathbf{X} \right| \\
 &\leq (1 - \nu)^n (\sup(a) - \inf(a)) \\
 &= (1 - \nu)^n \max_{\mathbf{X}, \mathbf{X}'} |a(\mathbf{X}) - a(\mathbf{X}')|
 \end{aligned} \tag{4.10}$$

For the third line we used Hölder's inequality combined with the normalisation of π and r_n yielding:

$$\begin{aligned}
 \left| \int a(\mathbf{X})[\pi(\mathbf{X}) - r_n(\mathbf{X})]d\mathbf{X} \right| &\leq \int |a(\mathbf{X})\pi(\mathbf{X})| + |(-a(\mathbf{X}))r_n(\mathbf{X})|d\mathbf{X} \\
 &= \|a\pi\|_1 + \|(-a)r_n\|_1 \\
 &\leq \|a\|_\infty \|\pi\|_1 + \|(-a)\|_\infty \|r_n\|_1 \\
 &= \sup(a) - \inf(a)
 \end{aligned}$$

The last line is the reason we need $a(\mathbf{X})$ to be bounded. The proof is completed.

Having established this theorem, it becomes clear how to construct a Markov Chain Monte Carlo algorithm. We have to design the algorithm in a way such that it converges to the probability distribution as its invariant distribution π . Then we draw samples from this distribution to compute expectation values of any function a we are interested

in. In the section 5 we show how to do this efficiently.

4.2. Detailed balance

A restrictive, but easy to verify condition to show that a transition kernel $T(\mathbf{X}, \mathbf{X}')$ has an invariant distribution $\pi(\mathbf{X})$ is to verify that it fulfils the following so-called 'detailed balance condition':

$$\pi(\mathbf{X})T(\mathbf{X}, \mathbf{X}') = \pi(\mathbf{X}')T(\mathbf{X}', \mathbf{X}), \quad (4.11)$$

since inserting this equation into the definition of an invariant distribution yields

$$\begin{aligned} \pi(\mathbf{X}) &= \int \pi(\mathbf{X}')T(\mathbf{X}', \mathbf{X})d\mathbf{X}' \\ &= \int \pi(\mathbf{X})T(\mathbf{X}, \mathbf{X}')d\mathbf{X}' \\ &= \pi(\mathbf{X}). \end{aligned}$$

It is however possible for a transition kernel to have an invariant distribution, without detailed balance holding.

When discussing Monte Carlo algorithms we will use detailed balance to show that the invariant distribution is precisely the distribution we want to sample from. If one is also able to show the condition in (4.4), we know the algorithm will generate an ergodic chain from the desired distribution. It is thus a proper Markov Chain Monte Carlo algorithm.

4.3. Parameter mean and the statistical error

As pointed out in the previous sections, the Markov Chain returns a time series of parameter estimates. We will now assume that the chain is at equilibrium and the time needed to get close enough to the invariant distribution has been removed. How to identify these different stages will be discussed in the upcoming sections. We now want to focus on parameter estimation and reliability.

Formally we assume that there is a set of true parameters \mathbf{X} and the estimates are closely scattered around this value. In other words we sample from the region, where $\pi(\mathbf{X})$ has most of its probability mass. The natural estimator for the mean is given by

$$\bar{x}_\alpha = \sum_{i=1}^N x_\alpha^i$$

or alternatively

$$\tilde{x}_\alpha^r = \frac{1}{N_r} \sum_{i=1}^{N_r} x_\alpha^{i,r} \quad \text{and} \quad \bar{x}_\alpha = \frac{1}{N} \sum_{r=1}^R N_r \tilde{x}_\alpha^r \quad (4.12)$$

if we have R statistically independent time series (called replica) with a length of N_r respectively. The total number of samples is given by $\sum_{r=1}^R N_r = N$. Obviously all chains have to have the same invariant distribution. These estimators are unbiased, meaning the expectation value of their deviations from the true values $\tilde{\delta}_\alpha^r = \tilde{x}_\alpha^r - X_\alpha$ and $\bar{\delta}_\alpha = \bar{x}_\alpha - X_\alpha$

is given by

$$\langle \tilde{\delta}_\alpha^r \rangle = 0 \quad \text{and} \quad \langle \bar{\delta}_\alpha \rangle = 0. \quad (4.13)$$

Because of the central limit theorem we can assume \tilde{x}_α^r and \bar{x}_α to be Gaussian distributed for large N_r , although the $x_\alpha^{i,r}$ are not necessarily Gaussian. Naively one would expect the corresponding density distribution to be fully determined by the covariance matrix given by

$$C_{\alpha\beta}^0 \delta_{r,s} \equiv \langle (x_\alpha^{i,r} - X_\alpha)(x_\beta^{i,s} - X_\beta) \rangle. \quad (4.14)$$

The delta function arises from the statistical independence between the different chains. $\hat{C}_{\alpha\beta}^0$ however is not the true covariance matrix since we are dealing with Markov Chains and thus the consecutive estimates for \mathbf{X} depend on each other. This results in an underestimation of the variance because the points tend to be closer to each other. To incorporate these effects we will try to estimate the full autocorrelation function

$$\Gamma_{\alpha\beta}(t) \delta_{r,s} \equiv \langle (x_\alpha^{i,r} - X_\alpha)(x_\beta^{i+t,s} - X_\beta) \rangle \quad \text{with} \quad \Gamma_{\alpha\beta}(0) = C_{\alpha\beta}^0 \quad (4.15)$$

This function calculates the dependency between the estimates for x_α and x_β at separation times. Since we are dealing with homogenous and converged chains, we imply a symmetry in time ($\Gamma_{\alpha\beta}(t) = \Gamma_{\alpha\beta}(-t)$) and also in the indices α, β . The formal definition of the full covariance matrix is then

$$C_{\alpha\beta} \equiv \sum_{t=-\infty}^{\infty} \Gamma_{\alpha\beta}(t) = \Gamma_{\alpha\beta}(0) + 2 \sum_{t=1}^{\infty} \Gamma_{\alpha\beta}(t) \quad (4.16)$$

which is obviously not computable, since we are not dealing with infinite chains. This will be resolved in the upcoming section. But this definition allows us to define the error of the sample mean. In order to do this we have to investigate the expectation value of $\bar{\delta}_\alpha$ multiplied by $\bar{\delta}_\beta$:

$$\begin{aligned} \langle \bar{\delta}_\alpha \bar{\delta}_\beta \rangle &= \frac{1}{N^2} \langle \sum_{r=1}^R \sum_{i=1}^{N_r} (x_\alpha^{i,r} - X_\alpha) \sum_{s=1}^R \sum_{j=1}^{N_s} (x_\beta^{j,s} - X_\beta) \rangle \\ &= \frac{1}{N^2} \sum_{r,s=1}^R \sum_{i,j=1}^{N_r, N_s} \langle (x_\alpha^{i,r} - X_\alpha)(x_\beta^{j,s} - X_\beta) \rangle \\ &= \frac{1}{N^2} \sum_{r,s=1}^R \sum_{i,j=1}^{N_r, N_s} \Gamma_{\alpha\beta}(i-j) \delta_{r,s} = \frac{1}{N^2} \sum_{r=1}^R \sum_{i,j=1}^{N_r} \Gamma_{\alpha\beta}(i-j) \\ &\approx \frac{1}{N^2} \sum_{r=1}^R N_r \sum_{t=-\infty}^{\infty} \Gamma_{\alpha\beta}(t) \\ &= \frac{1}{N} C_{\alpha\beta} \end{aligned} \quad (4.17)$$

4.3. Parameter mean and the statistical error

Here we used in the second to last line an approximation

$$\sum_{i,j=1}^M g(i-j) \approx M \sum_{k=-\infty}^{\infty} g(k)$$

which is justified for rapidly decaying functions. The autocorrelation function is rapidly decaying in this case, since every new estimate for \mathbf{X} is drawn randomly from a distribution which only explicitly depends on the last estimate. Points which are separated by large times should therefore be approximately independent, resulting in values of $\Gamma_{\alpha\beta}(t)$ close to zero. It turns out to be close to an exponential function with time scale $\tau_{\alpha\beta}$, which is named the integrated autocorrelation time.

(4.17) shows the (squared) statistical error from a sample mean. This defines the one sigma confidence interval in which we can expect the true mean value of this time series to be found in. This will not be the confidence interval which one can use for fitting parameters to experimental measurements. We will discuss this matter in section 6. If this were the confidence interval usable for fitting, one could easily enlarge statistics (meaning a larger total number of samples N) and shrink the error to zero thus obtaining a perfect fit. Therefore we will use this confidence interval only as a measure of how well we can be certain that there is a true value \mathbf{X} about which the estimates are scattered.

It is useful to redefine this error by separating the autocorrelation effects from the naive error:

$$\begin{aligned} \frac{1}{N}C_{\alpha\beta} &= \frac{1}{N} \left(\Gamma_{\alpha\beta}(0) + 2 \sum_{t=1}^{\infty} \Gamma_{\alpha\beta}(t) \right) = \frac{2}{N} \Gamma_{\alpha\beta}(0) \left(\frac{1}{2} + \sum_{t=1}^{\infty} \frac{\Gamma_{\alpha\beta}(t)}{\Gamma_{\alpha\beta}(0)} \right) \\ &= \frac{2}{N} \Gamma_{\alpha\beta}(0) \tau_{\alpha\beta} = \frac{2\tau_{\alpha\beta}}{N} C_{\alpha\beta}^0, \end{aligned} \quad (4.18)$$

where we have inserted the integrated autocorrelation time $\tau_{\alpha\beta}$, which includes all autocorrelation effects. This equation means that $2\tau_{\alpha\beta}$ effectively reduces the total number of samples. Therefore two times the integrated autocorrelation time counts the number of steps one needs starting at one sample to arrive at a new sample, which is then independent of it. The minimal possible value is 0.5. If we assume $\Gamma_{\alpha\beta}$ to be proportional to an exponential decay with time scale $\tau_{\alpha\beta}$, we can verify the step from the first to the second line:

$$\begin{aligned} \frac{1}{2} + \sum_{t=1}^{\infty} \frac{\Gamma_{\alpha\beta}(t)}{\Gamma_{\alpha\beta}(0)} &= \frac{1}{2} + \sum_{t=1}^{\infty} \exp\left(\frac{-t}{\tau_{\alpha\beta}}\right) \\ &= \frac{1}{2} + \left(\frac{1}{1 - \exp(-1/\tau_{\alpha\beta})} - 1 \right) \\ &\approx \frac{1}{2} + \left(\tau - \frac{1}{2} + \mathcal{O}\left(\frac{1}{\tau_{\alpha\beta}}\right) \right) \\ &= \tau_{\alpha\beta} + \mathcal{O}\left(\frac{1}{\tau_{\alpha\beta}}\right). \end{aligned} \quad (4.19)$$

If we were to integrate over the separation time instead of summing, the equation would be exact. We will assume from here on that the autocorrelation function decays exponentially as stated above, which is justified for sample sizes much larger than $\tau_{\alpha\beta}$. In reality however there are many contributions making a reliable approximation of autocorrelation effects extremely difficult.

The interpretation of the autocorrelation time to effectively reducing the number of samples until only independent samples are left gives rise to the idea of the autocorrelation time being a good measure of efficiency for the Markov Chain generating algorithm. Since an algorithm which can produce many samples in a short amount of time is a less well working algorithm if the chains suffer from strong autocorrelations, it may be advantageous to switch to a slower algorithm which produces less but better quality samples in the same time.

4.4. Numerical estimators and the error of the error

Up to this point we only defined the relevant objects in an unusable form, since the autocorrelation time in this form depends on the true values \mathbf{X} which are not known. Also the summation is carried out up to infinity, which is also not of practical use. These issues are to be resolved in this section.

In general we will construct estimators for the uncertainty estimation of the parameter values and of $\tau_{\alpha\beta}$. Then we try to estimate their reliability respectively giving us the error of the error. Finally we will choose the set of estimators which minimise the error of the computation of the full covariance matrix.

The autocorrelation function $\Gamma_{\alpha\beta}(t)$, as seen in the last section, is the figure of main interest. From there we can derive all other quantities. With this in mind we want to construct an estimator easy to handle numerically specking. The first step is to replace the exact values \mathbf{X} by the ensemble mean. This yields:

$$\bar{\Gamma}_{\alpha\beta}(t) = \frac{1}{N - Rt} \sum_{r=1}^R \sum_{i=1}^{N_r-t} (x_{\alpha}^{i,r} - \bar{x}_{\alpha})(x_{\beta}^{i+t,r} - \bar{x}_{\beta}) \quad (4.20)$$

One has to note that by inserting the ensemble mean, we introduced a leading bias given by

$$\langle \bar{\Gamma}_{\alpha\beta}(t) - \Gamma_{\alpha\beta}(t) \rangle \approx -\frac{C_{\alpha\beta}}{N}. \quad (4.21)$$

The calculation is quite tedious and can therefore be found in appendix A.1. Normally this systematic error is removed by changing the normalisation from N to $N - 1$. We will not do this here since we are interested in minimising the total uncertainty of the covariance matrix, which then ensures a reliable confidence interval for the parameter estimates. We will find error contributions of the order of $1/\sqrt{N}$ there, making contributions from (4.21) negligible.

Next the naive covariance matrix is estimated by

$$\bar{C}_{\alpha\beta}^0 = \bar{\Gamma}_{\alpha\beta}(0) \quad (4.22)$$

4.4. Numerical estimators and the error of the error

and the full covariance matrix via

$$\bar{C}_{\alpha\beta}(W) = \bar{\Gamma}_{\alpha\beta}(0) + 2 \sum_{t=1}^W \bar{\Gamma}_{\alpha\beta}(t) \quad (4.23)$$

where we cut off the summation at W . This value has to be chosen with care, since if we cut off the summation too early, we underestimate the autocorrelation effects. Otherwise if we sum over a too large window, we include more and more noise, since it gets less and less suppressed in (4.20), whilst the signal becomes negligible. It is possible to find an optimal summation window W , which we show in this section and explain in detail in section 4.5.

Finally the autocorrelation time will be computed by

$$\bar{\tau}_{\alpha\beta} = \frac{\bar{C}_{\alpha\beta}(W)}{2\bar{C}_{\alpha\beta}^0}. \quad (4.24)$$

Similar to the autocorrelation function we introduce a systematic error in the computation of $\bar{C}_{\alpha\beta}(W)$. This time this results from the truncation of the autocorrelation sum at W . For a purely exponential decay of the autocorrelation function one estimates (by using (4.21))

$$\begin{aligned} \langle \bar{\Gamma}_{\alpha\beta}(0) + 2 \sum_{t=1}^W \bar{\Gamma}_{\alpha\beta}(t) \rangle &\approx \Gamma_{\alpha\beta}(0) - \frac{C_{\alpha\beta}}{N} + 2 \sum_{t=1}^W \left(\Gamma_{\alpha\beta}(t) - \frac{C_{\alpha\beta}}{N} \right) \\ &= -\frac{2W+1}{N} C_{\alpha\beta} + \Gamma_{\alpha\beta}(0) + 2 \sum_{t=1}^{\infty} \Gamma_{\alpha\beta}(t) - 2 \sum_{t=W+1}^{\infty} \Gamma_{\alpha\beta}(t) \\ &= C_{\alpha\beta} \left(1 - \frac{2W+1}{N} \right) - 2\Gamma_{\alpha\beta}(0) \sum_{t=W+1}^{\infty} \exp(-t/\tau_{\alpha\beta}) \\ \Leftrightarrow \langle \bar{C}_{\alpha\beta}(W) - C_{\alpha\beta} \rangle &\approx -\frac{2W+1}{N} C_{\alpha\beta} - 2C_{\alpha\beta}^0 \frac{\exp(-W/\tau_{\alpha\beta})}{\exp(1/\tau_{\alpha\beta}) - 1} \end{aligned}$$

The first contribution is $\mathcal{O}(1/N)$ so we neglect it. The second term however has to be handled with care. In order to get a concise and fast computable formula we will reformulate it with the use of $2\tau_{\alpha\beta}C_{\alpha\beta}^0 = C_{\alpha\beta}$:

$$\begin{aligned} 2C_{\alpha\beta}^0 \frac{\exp(-W/\tau_{\alpha\beta})}{\exp(1/\tau_{\alpha\beta}) - 1} &= \frac{C_{\alpha\beta}}{\tau_{\alpha\beta}} \frac{\exp(-W/\tau_{\alpha\beta})}{\exp(1/\tau_{\alpha\beta}) - 1} \\ &\approx \frac{C_{\alpha\beta}}{\tau_{\alpha\beta}} \exp(-W/\tau_{\alpha\beta}) \left(\tau_{\alpha\beta} - \frac{1}{2} + \mathcal{O}(1/\tau_{\alpha\beta}) \right) \\ &= C_{\alpha\beta} \exp(-W/\tau_{\alpha\beta}) (1 + \mathcal{O}(1/\tau_{\alpha\beta})). \end{aligned}$$

Reinserting leads to

$$\frac{\langle \bar{C}_{\alpha\beta}(W) - C_{\alpha\beta} \rangle}{C_{\alpha\beta}} \approx -\exp(-W/\tau_{\alpha\beta}) \quad (4.25)$$

Following the argumentation from [44, chapter 3] this can only be interpreted as an order of magnitude statement if the autocorrelation function is not purely exponential. A treatment including other contributions is discussed in [38, chapter 2].

The second contribution to the error of the covariance matrix is computed via the variance of $\bar{C}_{\alpha\beta}(W)$. It involves approximating terms like

$$\langle \bar{\Gamma}_{\alpha\beta}(s) \bar{\Gamma}_{\gamma\delta}(t) \rangle = \frac{1}{(N-s)(N-t)} \sum_{i=1}^{N-s} \sum_{j=1}^{N-t} \langle (x_{\alpha}^i - \bar{x}_{\alpha})(x_{\beta}^j - \bar{x}_{\beta})(x_{\gamma}^i - \bar{x}_{\gamma})(x_{\delta}^j - \bar{x}_{\delta}) \rangle.$$

which is a four-point correlator (here for one replica to avoid even more indices). Its approximation was first done in [24] and a full derivation can be found in [41, appendix A]. A calculation for the variance of $\bar{C}_{\alpha\beta}(W)$ starting from this approximation is given in appendix A.4. The result is (see (A.4))

$$\langle (\bar{C}_{\alpha\beta}(W) - C_{\alpha\beta})(\bar{C}_{\gamma\delta}(W) - C_{\gamma\delta}) \rangle \approx \frac{2W+1}{N} (C_{\alpha\gamma} C_{\beta\delta} + C_{\alpha\delta} C_{\gamma\beta}). \quad (4.26)$$

Thus combining the absolute value of (4.25) and the square root of the above formula we get the sum of all error contributions for the covariance matrix:

$$\text{err}(\bar{C}_{\alpha\beta}(W)) \approx C_{\alpha\beta} \exp(-W/\tau_{\alpha\beta}) + \sqrt{\frac{2W+1}{N} (C_{\alpha\alpha} C_{\beta\beta} + C_{\alpha\beta}^2)} \quad (4.27)$$

This is the error we wish to minimise with respect to W . Therefore we could proceed and calculate the optimal value for W , which leads to solving a transcendental equation. In principle this has to be done for every possible combination of α and β and then one could compute every $\bar{C}_{\alpha\beta}$, $\bar{C}_{\alpha\beta}^0$ and $\bar{\tau}_{\alpha\beta}$ characterising the full set of samples. This is however an enormous numerical task especially for a large number of parameters and makes it difficult to compare chains generated from different algorithms. To solve this issue we will consider only a 'projected' version of this calculation⁹. Instead of calculating each component of $\bar{C}_{\alpha\beta}$ individually we will only consider the sum over all components. This leads to:

$$C \equiv \sum_{\alpha\beta} C_{\alpha\beta} \quad C^0 \equiv \sum_{\alpha\beta} C_{\alpha\beta}^0 \quad \tau \equiv \sum_{\alpha\beta} \tau_{\alpha\beta} \quad (4.28)$$

$$\bar{C}(W) \equiv \sum_{\alpha\beta} \bar{C}_{\alpha\beta}(W) \quad \bar{C}^0 \equiv \sum_{\alpha\beta} \bar{C}_{\alpha\beta}^0 \quad \bar{\tau} \equiv \sum_{\alpha\beta} \bar{\tau}_{\alpha\beta} \quad (4.29)$$

Thus we only get a single autocorrelation time for the whole set of samples. As a check for consistency one could also compute the autocorrelation times for each parameter individually by pretending the set of samples consists only of samples for the one parameter of interest.

⁹This can be seen as taking the sum of the parameters instead of the individual parameters as the figure of interest. How to handle these so-called secondary observables is explained in section 6.

4.5. Automatic summation window W

The total error of the projected covariance simplifies to

$$\text{err}(\bar{C}(W)) \approx C \exp(-W/\tau) + \sqrt{\frac{2W+1}{N}} 2C^2 \approx C \exp(-W/\tau) + 2\sqrt{\frac{W}{N}} C. \quad (4.30)$$

This is the function we will minimise to get the optimal value for W . But since this explicitly depends on τ , this task is not trivial and will be explained in section 4.5.

With the help of

$$\begin{aligned} \langle (\bar{C}_{\alpha\beta}^0 - C_{\alpha\beta}^0)(\bar{C}_{\gamma\delta}(W) - C_{\gamma\delta}) \rangle &\approx \frac{1}{N} (C_{\alpha\gamma} C_{\beta\delta} + C_{\alpha\delta} C_{\gamma\beta}) \\ \Rightarrow \langle (\bar{C}^0 - C^0)(\bar{C}(W) - C) \rangle &\approx \frac{2}{N} C^2 \end{aligned} \quad (4.31)$$

and

$$\begin{aligned} \langle (\bar{C}_{\alpha\beta}^0 - C_{\alpha\beta}^0)(\bar{C}_{\gamma\delta}^0 - C_{\gamma\delta}^0) \rangle &\leq \frac{1}{2N} (C_{\alpha\gamma}^0 C_{\beta\delta}^0 + C_{\alpha\gamma} C_{\beta\delta}^0 + C_{\alpha\delta}^0 C_{\gamma\beta}^0 + C_{\alpha\delta} C_{\gamma\beta}^0) \\ \Rightarrow \langle (\bar{C}^0 - C^0)^2 \rangle &\leq \frac{2}{N} C^0 C \end{aligned} \quad (4.32)$$

which are derived in appendix A.4 ((A.5) and (A.6)), we can compute the error for the integrated autocorrelation time τ via error propagation:

$$\begin{aligned} \langle (\bar{\tau} - \tau)^2 \rangle &\approx \left(\frac{\partial \bar{\tau}}{\partial \bar{C}} \right)^2 \text{var}(\bar{C})^2 + 2 \frac{\partial^2 \bar{\tau}}{\partial \bar{C} \partial \bar{C}^0} \text{var}(C \bar{C}^0) + \left(\frac{\partial \bar{\tau}}{\partial \bar{C}^0} \right)^2 \text{var}(\bar{C}^0)^2 \\ &= \frac{1}{N} \left((2W+1) \frac{C^2}{2(C^0)^2} - \frac{C^3}{(C^0)^3} + \frac{C^3}{2(C^0)^3} \right) \\ &= \frac{4}{N} \left(W + \frac{1}{2} - \tau \right) \tau^2. \end{aligned} \quad (4.33)$$

In order to compute these values we replace the analytic exact values with the estimators defined above, which technically introduces new errors, but since we are already at the level of the error of the error we stop here.

4.5. Automatic summation window W

In the previous section we derived the function we want to minimise with respect to W to find the optimal estimation of the projected covariance matrix. In this section we explain the procedure.

As previously mentioned the decay of the autocorrelation function is not always purely exponential. And by revisiting the total error of the projected covariance (4.30):

$$\text{err}(\bar{C}(W)) \approx C \exp(-W/\tau) + 2\sqrt{\frac{W}{N}} C.$$

we see the explicit dependence on the true (or most dominant) autocorrelation time τ_{true} . If there are more contributions to the decay of the autocorrelation function, the integrated autocorrelation time will not coincide with τ_{true} (see (4.19)). In order to compensate for this we start with a hypothesis of τ_{true} being proportional to the integrated autocorrelation time. More explicit we use

$$2\bar{\tau}(W) \approx \sum_{t=-\infty}^{\infty} \exp\left(-\frac{|t|}{\bar{\tau}(W)}\right) \stackrel{(\text{hyp})}{=} \sum_{t=-\infty}^{\infty} \exp\left(-\frac{S|t|}{\tau_{\text{true}}}\right) \quad (4.34)$$

with the proportionality given by $\tau_{\text{true}} = S\bar{\tau}(W)^{10}$. We then solve this equation for τ_{true} to get an estimation:

$$\begin{aligned} 2\bar{\tau}(W) &= \sum_{t=-\infty}^{\infty} \exp\left(-\frac{S|t|}{\tau_{\text{true}}}\right) = 1 + 2 \frac{\exp\left(-\frac{S}{\tau_{\text{true}}}\right)}{1 - \exp\left(-\frac{S}{\tau_{\text{true}}}\right)} \\ \Leftrightarrow 2\bar{\tau}(W) - 1 &= \exp\left(-\frac{S}{\tau_{\text{true}}}\right) (2 + 2\bar{\tau}(W) - 1) \\ \Rightarrow \bar{\tau}_{\text{true}} &= \frac{S}{\ln\left(\frac{2\bar{\tau}(W)+1}{2\bar{\tau}(W)-1}\right)} \end{aligned} \quad (4.35)$$

In order to obtain the minimal error of the projected covariance we compute the sign change in

$$g(W) = \exp\left(-\frac{W}{\bar{\tau}_{\text{true}}}\right) - \frac{\bar{\tau}_{\text{true}}}{\sqrt{WN}} \quad (4.36)$$

which is (up to the factor $-\frac{C}{\tau}$) the W derivative of (4.30).

In conclusion the algorithm starts with $W = 1$, computes $\bar{\tau}(W)$ and from there $\bar{\tau}_{\text{true}}$. This is done for increasing values of W until the first value of W is reached, where $g(W)$ is negative for the first time, indicating a minimum in the error estimation of \bar{C} . This is then defined as the optimal value of W called W_{opt} .

4.6. Test case: synthetic data

Having established the general concept and relevant formulas, we will now turn to a test case, where we control the auto- and crosscorrelation effects. In particular we will generate a Markov chain for two parameters X_1 and X_2 . The individual estimates will consist of the signal, being the actual values we want to estimate, and noise, which we will prepare for autocorrelation. Every sample will be generated from

$$x_1^i = X_1 + q(\nu_c^i + \nu_1^i) \quad x_2^i = X_2 + q(\nu_c^i + \nu_2^i) \quad (4.37)$$

¹⁰Usually S is chosen to be in the range of 1...5. We will later see the optimal value of W being at a plateau of $\bar{\tau}(W)$ and the values of S shift W_{opt} by such small amounts, that $\bar{\tau}(W)$ does not change too much.

4.6. Test case: synthetic data

where ν_1 and ν_2 are independent sources of autocorrelated noise for X_1 and X_2 respectively. ν_c on the other hand is autocorrelated noise added to both parameters in the same way thus introducing a crosscorrelation between the samples (hence the index). q controls the strength of the noise compared to the signal.

The autocorrelated noise will be generated from a recursive sequence:

$$\nu^1 = \eta^1 \quad \nu^{i+1} = \sqrt{1 - a^2} \eta^{i+1} + a \nu^i \quad (4.38)$$

where η^i is a random and independent sample generated from a Gaussian distribution with mean zero and unit variance. The factor a controls the autocorrelation effects as we will see shortly. Since every new sample in this sequence only depends on the current sample we are dealing with a proper Markov Chain. A simple induction¹¹ yields a vanishing mean for any ν sequence (see (A.7)). The autocorrelation function can be calculated leading to

$$\Gamma(t) = \langle \nu^i \nu^{i+t} \rangle = a^{|t|}, \quad (4.39)$$

so for $a < 1$ it results in an exponential decay. Subsequently we can calculate the integrated autocorrelation time τ , which we would like to set as an input. Therefore we will set the sum over $\Gamma(t)$ equal to 2τ and solve this equation for a yielding

$$\begin{aligned} \sum_{t=-\infty}^{\infty} \frac{\Gamma(t)}{\Gamma(0)} &= \frac{2}{1-a} - 1 \stackrel{!}{=} 2\tau \\ \Leftrightarrow a &= \frac{2\tau - 1}{2\tau + 1}. \end{aligned} \quad (4.40)$$

Hence we are able to control every feature of this chain. We imagine generating three independent chains with an integrated autocorrelation time given by τ_1 , τ_2 and τ_c respectively. From there on we calculate the samples.

Obviously the expectation value of $x_{1,2}^i$ is given by $X_{1,2}$ and using the outcome above, we are easily able to derive the other analytic characteristics (see (A.9) to (A.11)):

$$\begin{aligned} C_{11}^0 &= 2q^2 & C_{12}^0 &= q^2 & C_{22}^0 &= 2q^2 \\ \Gamma_{11}(t) &= q^2(a_c^{|t|} + a_1^{|t|}) & \Gamma_{12}(t) &= q^2 a_c^{|t|} & \Gamma_{22}(t) &= q^2(a_c^{|t|} + a_2^{|t|}) \\ C_{11} &= 2q^2(\tau_c + \tau_1) & C_{12} &= 2q^2\tau_c & C_{22} &= 2q^2(\tau_c + \tau_2) \end{aligned}$$

As already stated above we see q controlling the magnitude of the errors and see τ_c introducing an autocorrelation for each component individually and a crosscorrelation between them.

The projected integrated autocorrelation time (here called τ_{int} to avoid confusion) turns out to be

$$\tau_{int} = \frac{\sum_{\alpha\beta} C_{\alpha\beta}}{2 \sum_{\alpha\beta} C_{\alpha\beta}^0} = \frac{1}{6}(4\tau_c + \tau_1 + \tau_2).$$

We now run a simulation with $X_1 = 1$ and $X_2 = 0.5$. By choosing $\tau_1 = \tau_2 = 8, \tau_c = 4$

¹¹Every calculation not explicitly shown in this section can be found in appendix A.5

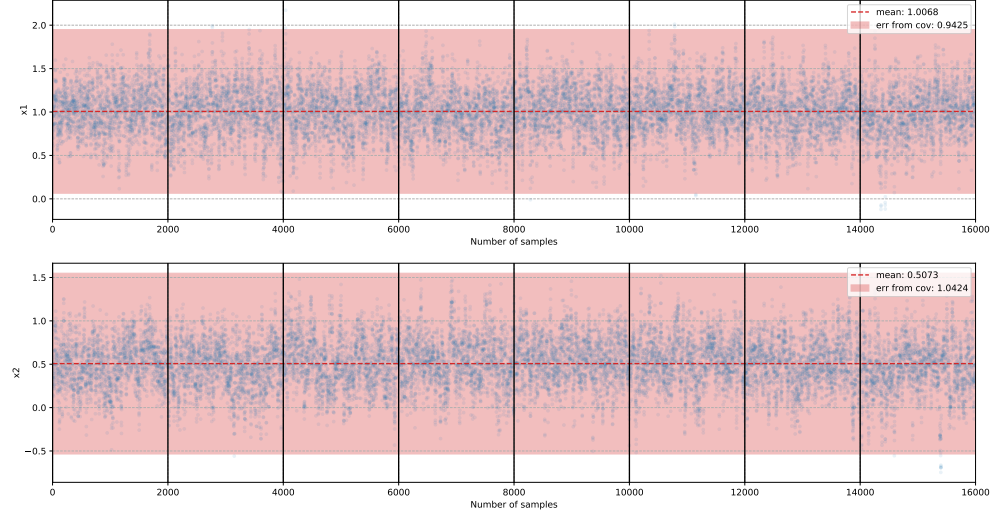


Figure 4.1: The time series of a synthetic data set, which consists of 16000 samples for two parameters divided into 8 replicas. The replica are separated via a solid black line. The red dashed line denotes the statistical mean value, whilst the red belt shows the marginalized 1σ error band from the covariance matrix.

and $q = 0.2$ the estimation of \mathbf{X} turns into a highly non-trivial task, since the spread given by the covariance is of comparable size for the first parameter and nearly twice as large as the second parameter. The exact numerical values are

$$C_{11} = \frac{24}{25} \quad C_{12} = \frac{8}{25} \quad C_{22} = \frac{24}{25} \\ \tau_{int} = \frac{16}{3}.$$

The data is generated in eight independent replicas with a chain length of 2000 samples providing 16000 samples overall. By making use of $N_{\text{indep}} = \frac{N}{2\tau_{int}}$ we conclude that there are a total of 2500 independent data points. The full data set is displayed in fig. 4.1. Here we see the time series of each replicum (divided by the solid black lines). The red dashed line represents the mean value obtained via the statistical estimator above defined. It is possible to see a definite substructure in the time series indicating strong autocorrelations. The mean values with the statistical errors are equal to

$$\bar{x}_1 = 0.9972 \pm 0.0053 \quad \text{and} \quad \bar{x}_2 = 0.5019 \pm 0.0055.$$

We can conclude that the generated time series has indeed a mean value at the exact $\mathbf{X} = (1, 0.5)$ and the Γ -method is able to find this value within the estimated errors. The red belt on the other hand is the marginalized error band from the covariance matrix. If we were to fit parameters from a distribution, the covariance matrix characterizes the

4.6. Test case: synthetic data

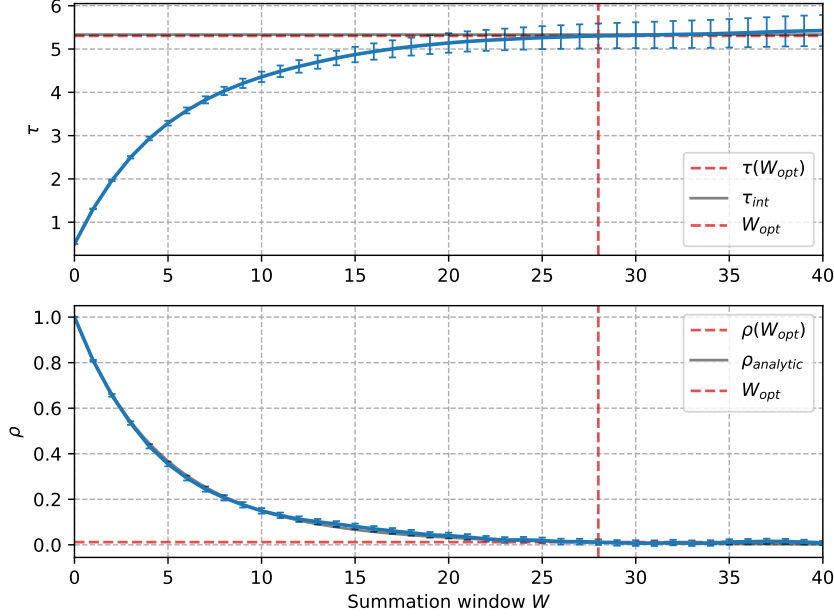


Figure 4.2: The upper plot shows the estimation of τ as a function of W and the lower plot the estimation of $\rho(W)$. The optimal region for W_{opt} is where $\rho(W)$ is sufficiently close to zero and does not yet fluctuate too strongly. Consequently $\tau(W)$ shows a plateau in this region. The red dashed lines show the position and value of τ and ρ at the optimal value for W . The analytic curves have been added in black.

local maximum of a Gaussian approximation. In particular it defines how well we can estimate a parameter from this Markov Chain. This procedure is given in section 6.1. The full covariance matrix is calculated to be

$$\bar{C}_{11} = 0.91 \pm 0.01 \quad \bar{C}_{12} = 0.322 \pm 0.006 \quad \bar{C}_{22} = 0.96 \pm 0.01$$

Whilst $C_{22} = 0.96$ and $C_{12} = 0.32$ are estimated correctly, we see a slight underestimation for the analytic $C_{11} = 0.96$. This could either be explained by the Markov Chain having these properties by chance, since the actual points are generated via random numbers from a distribution. Or the second possibility is the chain having properties which do not go with the approximations made when estimating the variance of the covariance matrix. A second run with twice as many samples has an estimation of the analytic covariance matrix within the 1σ error band. So the supposed disagreement is solved by better statistics.

The next feature of this Markov Chain to analyse is the autocorrelation, where we can see the impact of the summation W more clearly. A good consistency check is given by a plot of the calculated integrated autocorrelation time as a function of the summation window and also the normalised autocorrelation function $\rho(t)$ which are connected via

$$\tau = \frac{C}{2C^0} = \frac{1}{2} + \sum_{t=1}^{W_{opt}} \frac{\Gamma(t)}{\Gamma(0)} \equiv \frac{1}{2} + \sum_{t=1}^{W_{opt}} \rho(t). \quad (4.41)$$

As $\rho(t)$ displays the normalised decay of the autocorrelation function it is always (aside from statistical fluctuations) between 0 and 1. This serves well for comparing different Markov Chains. An error estimate of $\rho(t)$ is worked out in appendix A.3. In fig. 4.2 we can see $\bar{\tau}(W)$ and $\bar{\rho}(W)$ as the blue lines in each plot respectively. The red dashed lines indicate the point, where the optimal value of W has been chosen by the automatic summation window procedure. It is chosen at a point, where $\rho(W)$ is already sufficiently close to zero, but the noise has not taken over yet. At the right hand side of the depiction, we can see the beginning of small fluctuations. Consequently τ has to exhibit a plateau in the vicinity of W_{opt} . For this analysis $S = 1$ has been chosen. Smaller values for S shift W_{opt} more to the left and greater values for S shift W_{opt} to higher values. If τ is still in the plateau region, the value of S does not matter. Therefore one can see this depiction as an indication for a suitable analysis.

$\bar{\tau}(W_{opt})$ is estimated to be 5.30 ± 0.29 which is very close to the analytic value of $\tau_{int} = 5.3$. Figure 4.2 also shows the agreement of the analytic results with the numerical estimations. This and the other results validate the Γ -method to be able to solve such a non trivial task.

At last it is very helpful to have a look at the data set in terms of marginalised 1D and 2D density representations. Here it is possible to see whether the parameters are indeed Gaussian distributed and if there are, it might be sufficient to use a Gaussian approximation, which is fully characterised by the covariance matrix, for further applications of the parameter estimates. If the samples are not Gaussian distributed it might be beneficial to use the samples themselves.

The marginalised one-dimensional distributions are obtained via a simple histogram for each parameter. This is depicted in fig. 4.3 in the plots diagonally positioned (from left to right). The histograms clearly show a Gaussian distribution, which should come not as a surprise since the noise was generated via Gauss functions. The red curves show a fit for comparison.

The lower left plot displays a scatter plot of the individual samples. We see a positive correlation between them, which results from the positive value of C_{12} . The density estimation however is more complicated: One could try to generate a two-dimensional Gaussian distribution from the covariance matrix obtained priorly. But this turns out to be numerically unstable and hard to visualise especially for non Gaussian distributed samples. Instead we do a kernel density estimation. One tries to estimate the distribution (for now) denoted as $f(x_1, x_2)$ via a sum of kernel functions over all samples:

$$f(x_1, x_2) = \frac{1}{Nh} \sum_{i=1}^N K\left(\frac{x_1 - x_1^i}{h}, \frac{x_2 - x_2^i}{h}\right),$$

with a certain bandwidth h , which can be chosen automatically. We use a simple Gaussian as the kernel function $K(z_1, z_2)$. A detailed introduction can be found in [40, chapter 6]. Once obtained it is possible to draw a line around the region where a certain percentage of the probability mass is located. This has been done in fig. 4.3 via the red line for 1σ ($f(x_1, x_2) = 1/\sqrt{e}$) and the black line for 2σ ($f(x_1, x_2) = 1/e^2$) respectively. With this procedure it is possible to estimate an arbitrary complicated function $f(x_1, x_2)$.

4.6. Test case: synthetic data

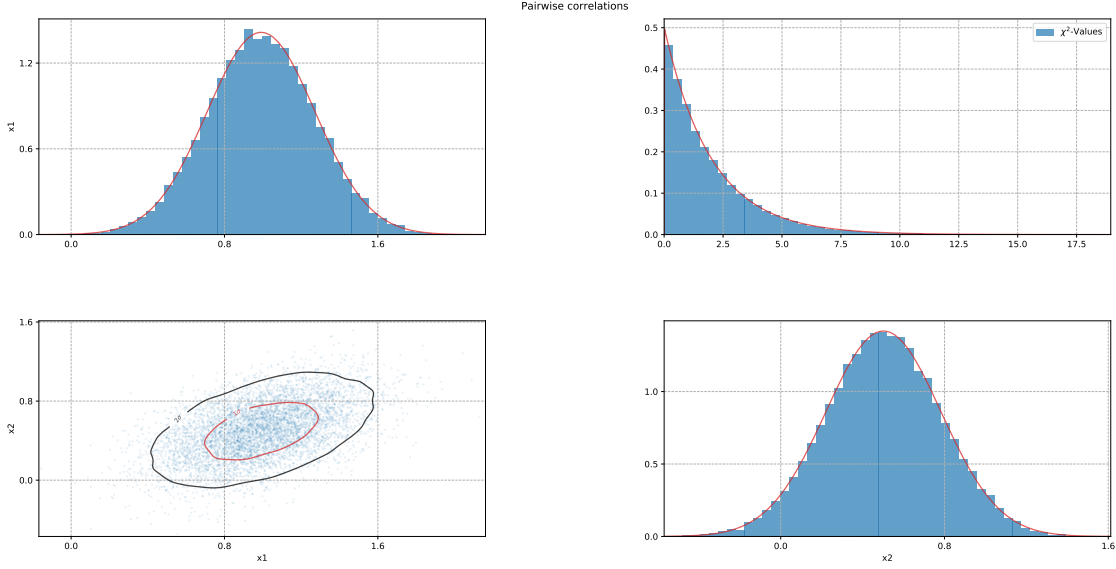


Figure 4.3: The 1D (diagonal plots) and 2D (lower left plot) distributions of the samples. As we can see, the 1D distribution is well described via a Gaussian, which has been fitted and is displayed as the red curve. The 2D distribution is shown as a scatter plot of the samples and the red (black) line indicates a 1σ (2σ) distance to the maximum of the kernel density estimation. The upper right plot shows the χ^2 values generated from the samples and has been fitted via a χ^2 -distribution.

The last plot in the upper right corner shows a histogram of χ^2 values for the samples. These values were obtained by using the exact values of \mathbf{X} and the naive covariance matrix. The red curve is a fit with a χ^2 -distribution since the probability of obtaining a certain $\chi^2 = z$ value is given by this function

$$f(z, k) = \frac{z^{k/2-1}}{2^{k/2}\Gamma(k/2)} \exp(-z/2).$$

k is the degree of freedom which is given by two in this case. This was done for the synthetic data since in the later analysis we will try to estimate parameters from a χ^2 -figure of merit. So each estimate has a specific χ^2 value and investigating the distribution of these values yields another consistency check. By artificially generating these values, we introduce this concept.

By carefully comparing the plots from fig. 4.3 to the obtained covariance matrix, one notices an apparent disagreement between the raw samples and the numerical estimators: By the looks of the Gaussian 1D density distributions one gets a smaller variance estimation. This disagreement is not real, since we have no temporal resolution and are therefore fully ignorant regarding autocorrelation effects. The variances obtained from these depictions are obtainable via the naive covariance matrix. This is also the reason we used the naive covariance matrix when generating the χ^2 -values.

5. Monte Carlo algorithms

One of the most popular and widely used algorithms was proposed by Metropolis et al. [26] in 1953. It deals with the efficient computation of expectation values in statistical mechanics. Instead of choosing configurations randomly, then weighting them with the corresponding Boltzmann factor, it chooses configurations with a probability of the Boltzmann factor and weights them evenly. This formed the basis of Monte Carlo simulations. Hastings [19] generalised the update mechanism in 1970, naming it Metropolis-Hastings Algorithm.

We will use this scheme as an introduction to Markov Chain generating rules. Whilst discussing its properties we will find a much needed improvement in order to compute large enough statistics in a reasonable amount of time. This will lead us to the adaptive Metropolis-Hastings algorithm, which we will use in a slightly adjusted form as proposed by Haario et al. [18] in 2001.

A general overview and introduction to these topics can be found in the *Handbook of Markov Chain Monte Carlo* [7, Chapter 1 and 4], which summarizes the scientific literature and presents it in an educational format. A survey of various results concerning the convergence of these algorithms was presented by Roberts and Rosenthal [35]. The convergence of the adaptive Metropolis Hastings algorithm was proven in the paper mentioned above by Haario et al. and by Roberts and Rosenthal [34] in 2007 in a more general manner concerning various adaptive algorithms.

5.1. The Metropolis-Hastings Algorithm

We are interested in an update mechanism that preserves a specific invariant distribution $\pi(\mathbf{X})$. More generally we suppose we are only able to compute the desired distribution up to a constant: $h(\mathbf{X}) = c\pi(\mathbf{X})$. The Metropolis-Hastings update mechanism starts by proposing a state \mathbf{Y} dependent on the current state \mathbf{X} . The proposal is defined by a conditional probability distribution denoted by $q(\mathbf{X}, \cdot)$. Then the so-called Hastings ratio¹²

$$r(\mathbf{X}, \mathbf{Y}) = \frac{h(\mathbf{Y})q(\mathbf{Y}, \mathbf{X})}{h(\mathbf{X})q(\mathbf{X}, \mathbf{Y})} = \frac{\pi(\mathbf{Y})q(\mathbf{Y}, \mathbf{X})}{\pi(\mathbf{X})q(\mathbf{X}, \mathbf{Y})} \quad (5.1)$$

is computed to evaluate the acceptance probability of the proposed move \mathbf{Y} given by

$$a(\mathbf{X}, \mathbf{Y}) = \min(1, r(\mathbf{X}, \mathbf{Y})). \quad (5.2)$$

That is, the state after the update is \mathbf{Y} with a transition probability density

$$T(\mathbf{X}, \mathbf{Y}) = q(\mathbf{X}, \mathbf{Y})a(\mathbf{X}, \mathbf{Y})$$

¹²The special case of a symmetric proposal distribution $q(\mathbf{X}, \mathbf{Y}) = q(\mathbf{Y}, \mathbf{X})$ simplifies the ratio to the original proposal from Metropolis [26].

5.1. The Metropolis-Hastings Algorithm

and the algorithm remains at \mathbf{X} with probability

$$\int q(\mathbf{X}, \mathbf{Y})(1 - a(\mathbf{X}, \mathbf{Y}))d\mathbf{Y}.$$

It is important to notice the possibility of a rejection. If one were to propose new states until eventually one gets accepted, one would destroy the update mechanisms property of preserving the distribution $\pi(\mathbf{X})$.

As it can be easily seen from the Hastings ratio there is no need to find the normalisation constant c . Furthermore the ratio is well defined, since a state with $q(\mathbf{X}, \mathbf{Y}) = 0$ has zero probability to get proposed in the first place. And $\pi(\mathbf{X}) = 0$ is also impossible since the Hastings ratio would have been zero, when \mathbf{X} was proposed and thus could not get accepted. In fact only the initial state has to be chosen with care.

It is simple to show that the invariant distribution is preserved by verifying the detailed balance condition (4.11). Trivially the condition is fulfilled for remaining at the current state. First¹³ considering the case where $r(\mathbf{X}, \mathbf{Y}) \geq 1$ results in $a(\mathbf{X}, \mathbf{Y}) = 1$ and $a(\mathbf{Y}, \mathbf{X}) = r(\mathbf{Y}, \mathbf{X})$. This yields:

$$\begin{aligned} \pi(\mathbf{X})T(\mathbf{X}, \mathbf{Y}) &= \pi(\mathbf{X})q(\mathbf{X}, \mathbf{Y}) = \pi(\mathbf{X})q(\mathbf{X}, \mathbf{Y})\frac{\pi(\mathbf{Y})q(\mathbf{Y}, \mathbf{X})}{\pi(\mathbf{Y})q(\mathbf{Y}, \mathbf{X})} \\ &= r(\mathbf{Y}, \mathbf{X})\pi(\mathbf{Y})q(\mathbf{Y}, \mathbf{X}) \\ &= \pi(\mathbf{Y})q(\mathbf{Y}, \mathbf{X})a(\mathbf{Y}, \mathbf{X}) \\ &= \pi(\mathbf{Y})T(\mathbf{Y}, \mathbf{X}) \end{aligned}$$

The second case supposes $r(\mathbf{X}, \mathbf{Y}) < 1$, with $a(\mathbf{X}, \mathbf{Y}) = r(\mathbf{X}, \mathbf{Y})$ and $a(\mathbf{Y}, \mathbf{X}) = 1$. Then

$$\begin{aligned} \pi(\mathbf{X})T(\mathbf{X}, \mathbf{Y}) &= \pi(\mathbf{X})q(\mathbf{X}, \mathbf{Y})r(\mathbf{X}, \mathbf{Y}) \\ &= \pi(\mathbf{X})q(\mathbf{X}, \mathbf{Y})\frac{\pi(\mathbf{Y})q(\mathbf{Y}, \mathbf{X})}{\pi(\mathbf{X})q(\mathbf{X}, \mathbf{Y})} \\ &= \pi(\mathbf{Y})q(\mathbf{Y}, \mathbf{X}) = \pi(\mathbf{Y})q(\mathbf{Y}, \mathbf{X})a(\mathbf{Y}, \mathbf{X}) \\ &= \pi(\mathbf{Y})T(\mathbf{Y}, \mathbf{X}). \end{aligned}$$

In either case we can exchange \mathbf{X} and \mathbf{Y} in the detailed balance condition, showing that $\pi(\mathbf{X})$ is indeed the invariant distribution and it is sufficient to know the function up to an constant in order to sample from it.

In contrast to the former case, showing ergodicity is not as simple. We have to investigate the ratio $\frac{T(\mathbf{X}, \mathbf{Y})}{\pi(\mathbf{Y})}$ from the fundamental theorem (see esp. (4.4)). In order to find the

¹³Please note throughout this section that switching the arguments of the functions changes their meaning drastically, e.g. $r(\mathbf{X}, \mathbf{Y}) = 1/r(\mathbf{Y}, \mathbf{X})$.

minimum we assume $\mathbf{X} \neq \mathbf{Y}$ and furthermore $r(\mathbf{X}, \mathbf{Y}) < 1$, yielding

$$\begin{aligned} \frac{T(\mathbf{X}, \mathbf{Y})}{\pi(\mathbf{Y})} &= \frac{q(\mathbf{X}, \mathbf{Y})r(\mathbf{X}, \mathbf{Y})}{\pi(\mathbf{Y})} \\ &= \frac{q(\mathbf{Y}, \mathbf{X})}{\pi(\mathbf{X})} \end{aligned}$$

The denominator is controllable if we require the invariant distribution to be bounded from above (as required in the theorem). The numerator however will vanish, since the proposal distribution is a probability density over an infinite state space and therefore has to vanish at infinity. A simple way of resolving the issue is to confine the state space. For infinite state spaces it is still possible to find convergence [35, theorem 4], but without any statement about the rate. The only condition relevant in this context is, if one is able to argue that there is a probability to reach starting from any region any other region in a finite amount of jumps. This has to be kept in mind when choosing a proposal distribution. Throughout the whole thesis we will use a symmetric proposal distribution, which is given by a simple Gaussian around the current state in every component, yielding

$$q(\mathbf{X}, \mathbf{Y}) = \prod_i \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(x_i - y_i)^2}{2\sigma_i^2}\right). \quad (5.3)$$

The values of σ_i will be carefully tuned later. This class of proposal distributions has the required properties and is simple (and fast) to sample from. This setup is called a random walk Metropolis Hastings algorithm, since the proposal is given by a function fulfilling $q(\mathbf{X}, \mathbf{Y}) = q(\mathbf{Y}, \mathbf{X})$.

To prove any kind of convergence rates one has to show additional properties of the invariant distribution. These are listed in [35, Chapter 3]. It is for example possible to show geometric ergodicity for distributions with a polynomial in the exponent (see [37]). The distribution we care about is highly non-linear. It also strongly depends on the experimental measurements used and confidence intervals. If we consider the general case, we might also want to include data from different experiments where the relevant interaction is for example not governed by DIS. This will change the shape of the distribution. Therefore we will not attempt to prove one of the convergence rate theorems and restrain ourselves with simple convergence, leaving the question aside of how one can be sure that the chain actually has the desired properties. And the answer is we can never be definitely be sure:

When running a Markov Chain Monte Carlo simulation the beginning of a chain will enter a so called 'burn-in' phase. At this point of Monte Carlo time the chain is strongly influenced by the initial state, since this is often a simple guess and will in general not be close to the maximum. Here we experience a drift in parameter space. So usually this part of the chain is dropped and will not be further analysed. If the chain is sufficiently stationary one hopes the chain converged properly. A perk of using the Gamma Method to analyse the chain, is the fact that even a small drift in parameter space (maybe over

5.1. The Metropolis-Hastings Algorithm

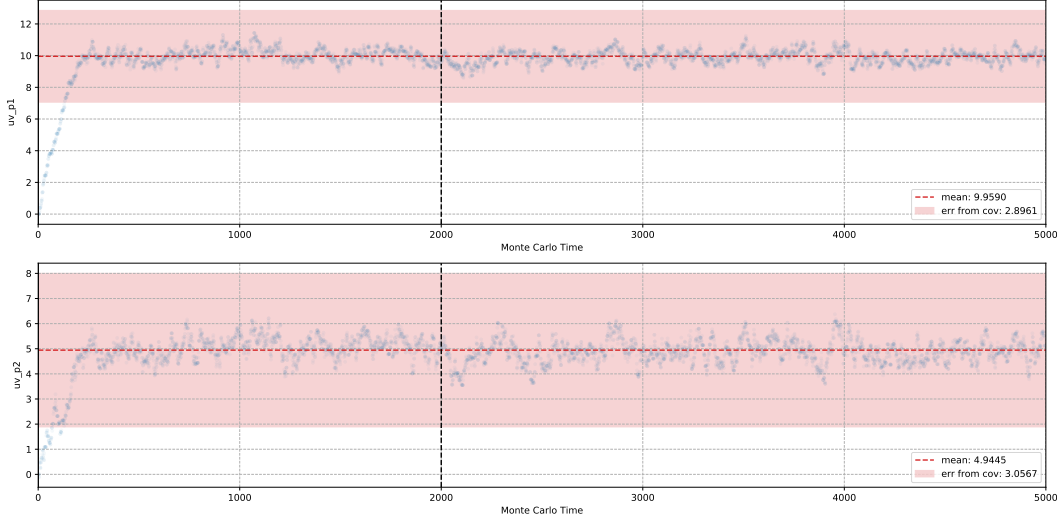


Figure 5.1: The first 5000 samples generated by a random walk Metropolis Hastings algorithm. The invariant distribution is a multivariate Gaussian centred around $x_1 = 10$ and $x_2 = 5$. The first 200 samples characterise the 'burn-in' phase. The vertical dashed line marks the beginning of the analysed batch of samples. The red dashed lines indicate the mean values, which are very close to the true value.

several thousand samples) will still be detected, since the integrated autocorrelation time is sensitive towards this behaviour. So in conclusion one can not be sure the invariant distribution was reached, but by identifying the 'burn-in'-phase and performing a autocorrelation time estimation, one can be reasonably sure that there is no further drift. In the case of a multimodal distribution it may still be possible that the chain converged to a local maximum and is 'stuck' there, meaning the probability to jump from the local to the global maximum is very low. This can happen if there is a valley of low probability dividing the two regions.

As a simple test case to visualise the concepts introduced above we imagine having a simple 2D Gaussian distribution given by

$$\pi(\mathbf{X}) \propto \exp\left(-\frac{1}{2}(\mathbf{X} - \mathbf{X}_0)C^{-1}(\mathbf{X} - \mathbf{X}_0)\right) \text{ with } \mathbf{X}_0 = \begin{pmatrix} 10 \\ 5 \end{pmatrix} \text{ and } C = \begin{pmatrix} 0.2 & 0.1 \\ 0.1 & 0.2 \end{pmatrix}$$

as the invariant distribution. Although the normalisation constant is analytically known, we do not have to care about it. For the proposal distribution we choose $\sigma_1 = \sigma_2 = 0.2$ and record 10000 samples. The first 5000 samples can be seen in fig. 5.1. The starting point was set to the origin and within the first 200 samples the algorithm seems to have converged. The analysis started at the 2000th sample indicated by the black dashed line. The mean values are at $\bar{x}_1 = 9.959$ and $\bar{x}_2 = 4.9445$ therefore very close to the exact centre of the Gaussian. The confidence estimate from the covariance matrix is however very large. This has to do with the way the chain oscillates around its mean value. Every point seems to be very dependent on the last state. If we investigate the integrated

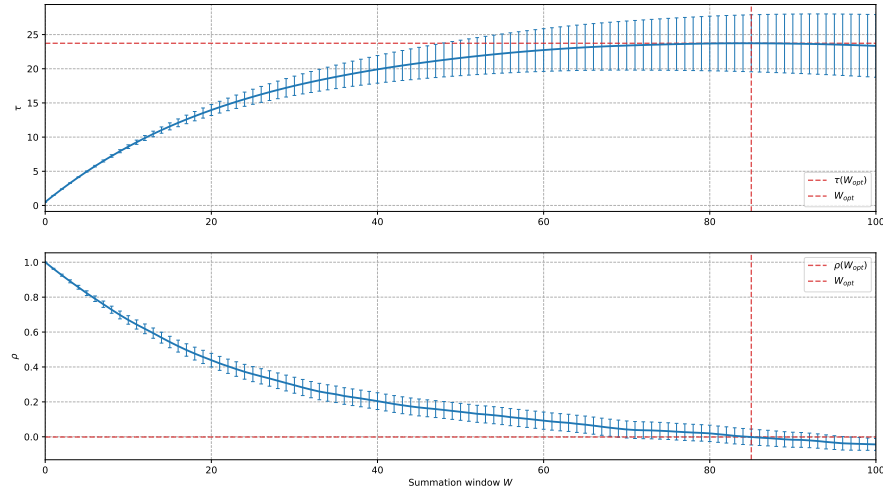


Figure 5.2: The integrated autocorrelation time τ and the normalised autocorrelation function $\rho(t)$. Due to the sharp proposal distribution the samples are strongly autocorrelated and only about every 50th point is an independent sample. S was taken to be 1.

autocorrelation time (see fig. 5.2 upper plot), we find our expectations fulfilled. The normalised autocorrelation function (lower plot) decays very slowly, leading to the fact that only about every 50th point is an independent sample. So we see that the algorithm does not sample very efficiently. This has to do with the values for σ_1 and σ_2 as we will see in the next section.

5.2. Optimal acceptance rate

The example from the last section shows convergence towards the target distribution, but suffers from strong autocorrelations. If we investigate the acceptance rate (the ratio of accepted proposals over all proposals) we will find a rate of 74.5%. So most of the proposals are accepted, which at first sight seems like a good property. But in fact it is not, since this happens because the proposed states are relatively close to each other with respect to the spread of the maximum of the invariant distribution. This means that the difference in the density between proposed states and the current state is small, making the acceptance probability large.

Figure 5.3 shows 100 consecutive jumps made by the algorithm in red, additionally the blue ellipses draw the contour levels of the target distribution. With respect to Monte Carlo time, the algorithm starts close to the centre of the maximum and ends in the lower left part of the figure. Here we can clearly see, that the jumps made are close to each other, resulting in only probing the lower left part of the ellipses. So a batch mean of this passage is strongly biased towards lower values for both x_1 and x_2 , showing the

5.2. Optimal acceptance rate

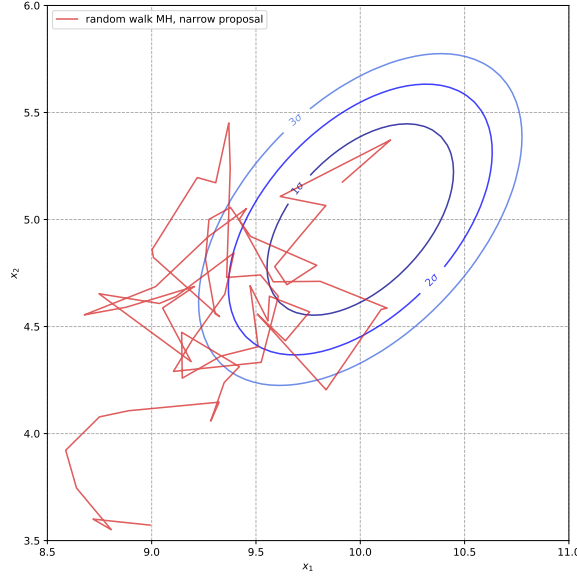


Figure 5.3: The blueish ellipses show the contour levels of the invariant distribution. The red line shows 100 steps of the random walk MH from the last section. The algorithm starts in the centre of the distribution and ends further outside. Due to the jumps being close to each other, the algorithm probes only the lower left part of the maximum.

effects of strong autocorrelation. Furthermore the time series ends far outside of the region of the maximum, where the total density is very low. This is a secondary effect which occurs when the proposals are too close to each other. In fact this can only happen because the relative difference between two nearby points does not reflect the factor of the total probability of points in this region being extremely light.

On the other hand, a too wide proposal distribution results in a very low acceptance rate, with the chain hardly ever moving. This is obviously just as unfortunate. For the very special (and unrealistic) class of target densities

$$\pi(\mathbf{X}) = \prod_i f(x_i)$$

it is possible to show¹⁴, that the optimal acceptance rate is precisely 0.234. Now since this class of target densities is almost never a realistic case, because if it would be, one could find the maximum with more simple methods, one usually relaxes the statement. The algorithms efficiency remains high whenever the acceptance rate is between about 0.1 and 0.6.

Later by Roberts and Rosenthal (2001) [36] it was shown that, if the invariant distribution is of the form $\mathcal{N}(\cdot, C)$, then a proposal which draws samples from $\mathcal{N}(\cdot, kC)$, with k being a constant, is optimal and the result for the optimal acceptance rating still holds. In order to achieve an acceptance rate of 0.234 the constant can be fixed to $k = \frac{(2.38)^2}{d}$,

¹⁴See [33] for the original paper and [7, chapter 4.2] for an introductory text.

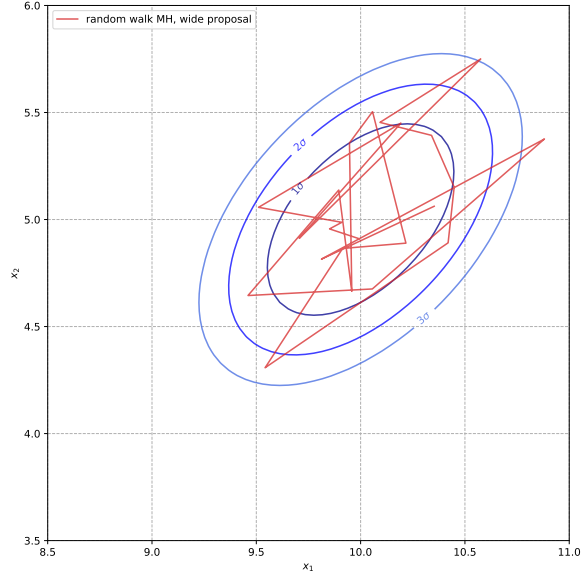


Figure 5.4: The same plot as above, but now the red line is generated with a random walk Metropolis with optimal settings. The proposal is done via $\mathcal{N}(\cdot, kC)$ with $k = (2.38)^2$, resulting in an acceptance rate of 24%. The step size is much larger and the shape of the maximum is covered entirely.

where d is the dimension of the state space.

The theorems described above only hold as the number of parameters d goes to infinity, but finite dimensional tests show the applicability of these results for small dimensions. If we now redo the analysis with a proposal distribution as described above¹⁵ we get an acceptance rate of 24%. The integrated autocorrelation time drops to $\tau = 5.1 \pm 0.5$ which should mostly be due to the fact that only about every fourth point gets accepted. Again 100 consecutive jumps are shown in fig. 5.4. This time the jumps cover a much greater distance and the centre of the ellipses is covered entirely. Also points outside the 99% region are much less likely. The overall performance of the algorithm has been improved.

This form of optimising the proposal distribution can also be used if the target distribution is not a Gaussian, because often it is the case that the maximum of a distribution is locally shaped as a Gaussian. However the covariance matrix is seldom known, which makes it quite tedious to find a good proposal distribution especially in high dimensions. There are algorithms which try to learn a optimal proposal distribution automatically.

5.3. Adaptive Metropolis-Hastings

In the above example the covariance matrix was exactly known. In more realistic cases it is seldom known or the target distribution is in fact not Gaussian. However it is often

¹⁵We set $k = (2.38)^2$ due to the small amount of dimensions. In fact dividing this k -value by two yields a too high acceptance rate.

5.3. Adaptive Metropolis-Hastings

the case that the maximum of a distribution can be locally approximated quadratically. Suppose we have a target distribution $\pi(\mathbf{X})$ with a maximum occurring at \mathbf{X}_0 . If we consider the natural logarithm $L = \log(\pi(\mathbf{X}))$ we may use a Taylor expansion around the maximum:

$$L \approx L(\mathbf{X}_0) + \frac{1}{2} \sum_{\alpha, \beta} (x_\alpha - x_\alpha^0)(x_\beta - x_\beta^0) \left. \frac{\partial^2 L}{\partial x_\alpha \partial x_\beta} \right|_{\mathbf{X}_0}$$

The linear terms vanish due to expanding around the maximum and we dropped the terms beyond the quadratic one. By exponentiating again we arrive at

$$\pi(\mathbf{X}) \propto \exp \left(\frac{1}{2} (\mathbf{X} - \mathbf{X}_0)^T \nabla \nabla L|_{\mathbf{X}_0} (\mathbf{X} - \mathbf{X}_0) \right) \quad (5.4)$$

where $\nabla \nabla L|_{\mathbf{X}_0}$ can be viewed as the Hessian matrix of the logarithmic probability density. Furthermore it is the inverse of the (negative) covariance matrix. In fact every distribution is connected to a normal distribution and it boils down to how well one can approximate its maximum by a quadratic expansion. The algorithm here to discuss is making use of this fact. It tries to learn the optimal proposal distribution from the samples at runtime. It is then possible to propose wider jumps, since the algorithm knows where most of the probability mass lies.

The adaptive Metropolis-Hastings algorithm starts like regular random walk with a fixed covariance matrix C_0 until the N_{pre} th iteration is reached. Then the proposal distribution is changed to

$$\mathbf{Y}_{n+1} \sim (1 - \beta) \mathcal{N} \left(\mathbf{X}_n, \frac{(2.38)^2}{d} \bar{C}_n \right) + \beta \mathcal{N}(\mathbf{X}_n, C_0) \quad (5.5)$$

where \bar{C}_n is the n th step empirical covariance matrix. The factor $0 < \beta < 1$ ensures that even if \bar{C}_n collapses to zero, the proposal is still valid and the algorithm converges. Computationally the empirical covariance matrix can be computed in constant time, since it fulfils a recursion relation. In addition to the usual conditions for a Metropolis-Hastings algorithm there are two requirements to meet: First every proposal distribution at each time step respectively has to converge to the invariant distribution if it were used singularly. This criterion is realised since every step can be seen as a simple random walk proposal. Second the adaptation between two consecutive proposal probabilities has to vanish. More formally:

$$\lim_{n \rightarrow \infty} \sup_{\mathbf{X} \in \mathbb{R}^d} \|P_{n+1}(\mathbf{X}, \cdot) - P_n(\mathbf{X}, \cdot)\| = 0 \quad (\text{in probability}) \quad (5.6)$$

so the probability of two consecutive proposal probabilities being different has to vanish as the number of iterations goes to infinity. The algorithm uses the Metropolis-Hastings acceptance probability at every step, reducing the condition that only the adaptation of the proposal itself has to diminish. This is given for the algorithm above, since the covariance changes at the n th iteration only by $\mathcal{O}(1/n)$.

We simulate the problem from above again with the adaptive Metropolis-Hastings

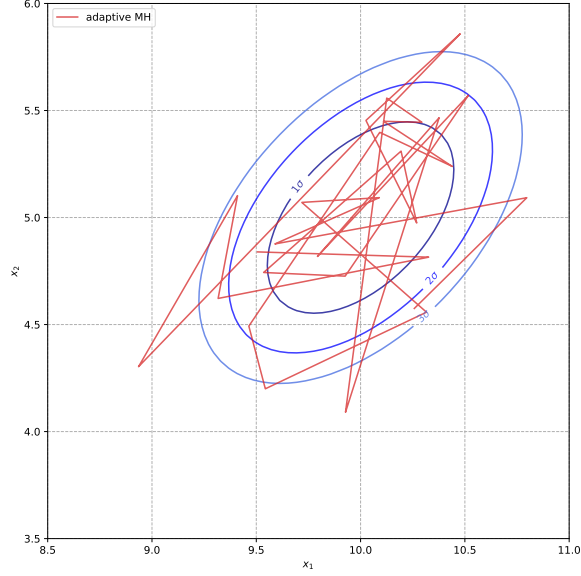


Figure 5.5: 100 consecutive iterations with the adaptive Metropolis-Hastings algorithm (after $N_{\text{pre}} = 2000$ iterations). The maximum of the target distribution is well probed and the algorithm converged towards the optimal solution.

algorithm. We choose the parameters:

$$N_{\text{pre}} = 2000 \quad \beta = 0.9 \quad C_0 = \begin{pmatrix} 0.2 & 0 \\ 0 & 0.2 \end{pmatrix} \quad S = 1$$

The n th step empirical covariance is computed without taking autocorrelation into account. With these settings we reach an acceptance rate of 23.5% and an integrated autocorrelation of 5.5 ± 0.7 . The estimates for the parameters are $\bar{x}_1 = 9.989$ and $\bar{x}_2 = 5.015$, which are again very close to the exact values. From these results and fig. 5.5 we conclude that the algorithm converged towards the optimal solution. In fact the covariance matrix estimated at the final step is

$$C_{10000} = \begin{pmatrix} 0.22 & 0.11 \\ 0.11 & 0.22 \end{pmatrix}$$

which is 1.1 times the predefined covariance matrix.

6. Error estimation

Usually we are not interested in the parameters themselves, but in a function $f(\mathbf{X})$ which takes the parameters as an argument. This is called a secondary observable. Given a valid sampling algorithm and parameter estimates in the form of a Markov Chain, we would like to construct a confidence interval. This should translate how well we can estimate parameters to how well we can estimate a function of the estimators.

The Gamma-Method [44] explained in section 4 also provides a framework for analysing secondary observables. It will however only produce the statistical error for the best estimate. Again this error testifies that there is a statistical mean and we are 68% sure that the true expectation value lies within the given confidence interval. An infinite amount of samples will shrink this confidence interval to zero. So this error estimate can only be used to state that the Markov Chain actually converges towards a stationary expectation value.

There are other methods, which are specifically tuned for χ^2 -fitting of parton distribution functions. One method uses the Hessian matrix of the χ^2 function as a quadratic approximation of the maximum. This was introduced in [32] and a short overview can be found in [43]. From there on it is possible to construct confidence intervals by allowing for a well defined deviation from the minimal χ^2 value. The method is easily applicable to Markov Chains since the Hessian matrix can be estimated from the samples. Alternatively one could use finite differences.

Lastly we will consider the natural way of error estimation in the framework of Markov Chain Monte Carlos: Using the samples directly to get a statistical representation of the desired secondary observable. This method does not approximate anything as long as the invariant distribution is sampled correctly. In a paper the samples themselves should be stated, giving the readers the opportunity to decide how to define an accurate representation of the secondary observable. Additional comments are also to be found in [21, chapter 8].

6.1. The statistical error of secondary observables

As it is possible to study the statistical properties of the samples, we can analyse the secondary observable in the same manner. Since we already developed the relevant equations and estimators we can simply translate the notation. The best estimate for the function value is obviously defined as

$$\bar{f} = f(\bar{\mathbf{X}}) \tag{6.1}$$

We can now use a Taylor expansion around the true parameter values to leading order. Then we make use of the covariance matrix of the parameter estimates to construct a

confidence interval for the secondary function¹⁶:

$$\begin{aligned}\bar{f} &\approx f(\mathbf{X}) + \sum_{\alpha} (\bar{x}_{\alpha} - x_{\alpha}) \left. \frac{\partial f}{\partial x_{\alpha}} \right|_{\mathbf{X}} \\ \Rightarrow \langle (\bar{f} - f(\mathbf{X}))^2 \rangle &\approx \frac{1}{N} \sum_{\alpha\beta} \left. \frac{\partial f}{\partial x_{\alpha}} \right|_{\mathbf{X}} \left. \frac{\partial f}{\partial x_{\beta}} \right|_{\mathbf{X}} \bar{C}_{\alpha\beta} \equiv \frac{1}{N} \bar{C}_f\end{aligned}$$

In other words the second line 'projects' the covariance matrix onto the function $f(\mathbf{X})$ via its partial derivatives. Naturally one has to evaluate the derivatives at the best estimate $\bar{\mathbf{X}}$ if an analytic expression is at hand. Otherwise a numerical estimation is needed. Either way it introduces (small) additional contributions, which we do not attempt to cancel.

It is possible to rewrite the above expression in the context of the integrated autocorrelation time. The derivatives are merely factors, so there is no need to repeat the whole analysis on how to construct good estimators. We can simply recycle the results, yielding

$$\begin{aligned}\bar{C}_f &= 2\bar{\tau}_f \bar{C}_{0,f} \\ \bar{C}_{0,f} &= \sum_{\alpha\beta} \left. \frac{\partial f}{\partial x_{\alpha}} \right|_{\bar{\mathbf{X}}} \left. \frac{\partial f}{\partial x_{\beta}} \right|_{\bar{\mathbf{X}}} \bar{\Gamma}_{\alpha\beta}(0) \quad \text{and} \quad \bar{\tau}_f = \frac{1}{2} + \sum_{t=1}^W \sum_{\alpha\beta} \left. \frac{\partial f}{\partial x_{\alpha}} \right|_{\bar{\mathbf{X}}} \left. \frac{\partial f}{\partial x_{\beta}} \right|_{\bar{\mathbf{X}}} \bar{\rho}_{\alpha\beta}(t)\end{aligned}$$

where we implied the automatic summation window procedure from section 4.5. The integrated autocorrelation time $\bar{\tau}_f$ can be used to analyse the efficiency of the sampling algorithm.

This method is a very convenient way of constructing confidence intervals since it only uses the partial derivatives of the secondary observable in addition to the covariance and autocorrelation time estimates, which are already present, when the Markov Chain has been analysed. Since it is a statistical confidence interval the error is of order $\mathcal{O}(1/\sqrt{N})$. Thus generating an infinite amount of samples shrinks the error to zero, which is therefore not a suitable way of extracting bounds from experimental measurements. For this application other methods are needed.

6.2. The Hessian method for error estimation

As explained in section 5.3 the maximum of a distribution is directly connected to a Gaussian. It turned out, that the inverse of the (negative) covariance matrix is the Hessian of the logarithmic probability density. The Bayesian analysis revealed the likelihood of a set of parameters with respect to the experimental measurements to be (see section 3.6)

$$p(\mathbf{X}) \propto \exp\left(-\frac{1}{2}\chi^2\right).$$

¹⁶Neglecting the quadratic term introduces a bias unless the function $f(\mathbf{X})$ is linear. If one produces several replicas it is possible to cancel the bias if needed (see [44, chapter 2.2]). Usually by producing large enough statistics this contribution is negligible since it scales quadratically in the fluctuations between the parameter estimates and the true values.

6.2. The Hessian method for error estimation

Combining these two facts yields

$$C_{\alpha\beta}^{-1} = \frac{1}{2} \frac{\partial^2}{\partial x_\alpha \partial x_\beta} \chi^2 \Big|_{\mathbf{x}_0} \equiv \frac{1}{2} H_{\alpha\beta}. \quad (6.2)$$

It is therefore possible to estimate the Hessian matrix directly from the samples if the quadratic expansion is justified. By making use of orthonormal eigenvectors \mathbf{v}_α with a corresponding set of eigenvalues $\{\epsilon_\alpha\}$ defined by¹⁷

$$\sum_\beta C_{\gamma\beta} v_{\beta\alpha} = \epsilon_\alpha v_{\gamma\alpha} \quad \text{and} \quad \sum_\beta v_{\beta\alpha} v_{\beta\gamma} = \delta_{\alpha\gamma}$$

we constructed a hyperellipsoid in parameter space. The χ^2 function is constant along this surface. We will now use rescaled coordinates z_α to transform the hyperellipsoid into a hypersphere. If the maximum is approximated perfectly by a normal distribution the scaling factors s_α are given by $\sqrt{\epsilon_\alpha}$. If one used the Hessian matrix to access the eigenvectors the scaling s_α would be given by $\sqrt{2/\epsilon_\alpha^H}$ with the eigenvalues of the Hessian matrix ϵ_α^H respectively. Since the maximum is seldom of true quadratic form, we will discuss a rescaling procedure later. Either way the difference in the parameters can be expressed as

$$X_\alpha - X_\alpha^0 = \sum_\beta v_{\alpha\beta} s_\beta z_\beta.$$

This yields a easy to calculate difference in χ^2

$$\Delta\chi^2 = \chi^2(\mathbf{X}) - \chi^2(\mathbf{X}^0) = \sum_\alpha z_\alpha^2 \quad (6.3)$$

and the surface of constant χ^2 is a hypersphere.

Now turning to the secondary function $f(\mathbf{X})$, we want to know how much the function varies, when the difference in χ^2 is given by $\Delta\chi^2 = t^2$. Explicitly we deflect the parameters in a way such that $\sum_\alpha z_\alpha^2 = t^2$. The value of t is to be defined later and kept general for now. The change in the secondary function can then be calculated via a directional derivative. To be most conservative we will choose the direction \mathbf{d} along the line, where the secondary function varies the most. That is the gradient direction. More explicitly we choose

$$\mathbf{d} = T \frac{\nabla f}{|\nabla f|}$$

where the magnitude of the direction is given by T . It will be possible to relate t and T later on. Summing up the discussion above we arrive at

$$\Delta f = (\nabla f) \cdot T \frac{\nabla f}{|\nabla f|} = T |\nabla f|$$

¹⁷In the original literature (see [32] or [43]) the eigenvectors and -values are defined with respect to the Hessian matrix. We will define the vectors with respect to the covariance matrix, since this is the accessible quantity in the framework of Markov Chain Monte Carlos.

for the confidence interval for the secondary function given by $f(\mathbf{X}_0) \pm \Delta f$. The gradient (which has to be taken with respect to z_α) will be approximated by

$$\frac{\partial f}{\partial z_\alpha} \approx \frac{f(\mathbf{X}_0 + t\mathbf{e}_\alpha) - f(\mathbf{X}_0 - t\mathbf{e}_\alpha)}{2t}$$

where \mathbf{e}_α is a rescaled vector in the direction of the component α , namely the column vectors \mathbf{v}_α times the rescaling factor s_α . This leaves us with the final expression

$$\Delta f = \frac{T}{2t} \sqrt{\sum_\alpha [f(\mathbf{X}_0 + t\mathbf{e}_\alpha) - f(\mathbf{X}_0 - t\mathbf{e}_\alpha)]^2}. \quad (6.4)$$

The only thing left to do is to find specific values for T and t . Luckily it is possible to relate these two by pretending the secondary function to be the χ^2 function itself. By making use of $\chi^2(\mathbf{X}_0 \pm t\mathbf{e}_\alpha) \stackrel{!}{=} \chi^2(\mathbf{X}_0) + T^2$ and the formula above one arrives at

$$\begin{aligned} T^2 = \Delta\chi^2 &= \frac{T}{2t} \sqrt{\sum_\alpha (2t^2)^2} \\ \Leftrightarrow T &= \sqrt{N_{\text{param}}} t. \end{aligned} \quad (6.5)$$

The value for T can be defined by making use of the χ^2 -distribution. The probability density function

$$f(\chi^2, k) = \frac{(\chi^2)^{k/2-1}}{2^{k/2}\Gamma(k/2)} \exp(-\chi^2/2).$$

has the mean $\langle\chi^2\rangle = k$ and a variance $\langle(\chi^2 - k)^2\rangle = 2k$ for a large degree of freedom ($k \gg 1$). If we had perfect statistics (the quadratic approximation holds perfectly) we could set $T^2 = \sqrt{2k}$, but usually one enlarges the value of T by a factor $\frac{\chi_{\text{fit}}^2}{k}$ resulting in

$$T = \sqrt{\chi_{\text{fit}}^2 \sqrt{\frac{2}{k}}}. \quad (6.6)$$

This enlargement of T serves as a more conservative procedure, since usually several experiments are considered within the χ^2 function. The data sets often do not agree fully and combining these sets turns into a non-trivial task. This is circumvented by defining the confidence intervals more conservative.

The scaling factors s_α need to be adjusted, since the maximum is usually not perfectly approximated by a Gaussian. Therefore we use

$$\min(\Delta\chi^2(\mathbf{X}_0 + ts_\alpha\mathbf{v}_\alpha), \Delta\chi^2(\mathbf{X}_0 - ts_\alpha\mathbf{v}_\alpha)) = t^2 \quad (6.7)$$

to define the optimal s_α . With this equation the uncertainties are always overestimated on the one side and exactly estimated on the other side¹⁸. The equation can be solved

¹⁸Another benefit of using this equation is the fact that numerical algorithms for solving an eigenvector

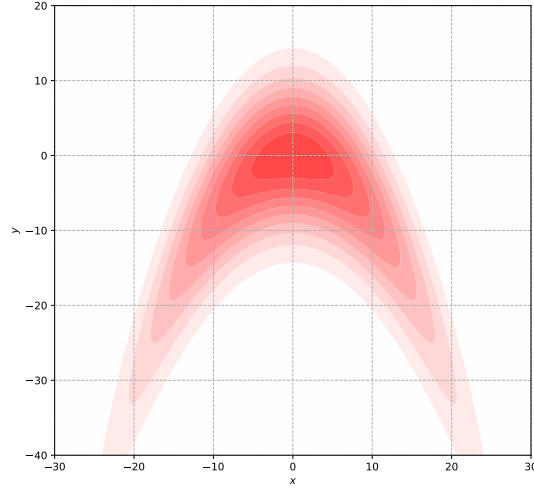


Figure 6.1: The contours of the distribution described in (6.8). The maximum can not be well described by a Gaussian approximation.

numerically.

This method of assigning confidence intervals is the standard procedure when fitting parton distribution functions with minimising techniques. The most important disadvantage is the reliance on the quadratic expansion of the maximum of the likelihood. After performing an analysis it is possible to scan the minimum of the χ^2 function along the eigendirections. In the ideal case the outcome should be (when plotted against the z_α values) a unit parabola. Every deviation shows the discrepancy between the approximation and the true shape of the χ^2 function. It is often the case (see for example [22, chapter 4 or appendix B]) that the maximum does not follow a perfect unit parabola along every eigendirection. This way systematic errors are made when estimating the confidence intervals.

6.3. Markov Chain Monte Carlo inference

The best way to illustrate the usefulness of publishing the actual samples instead of mean values and confidence intervals, is to consider a probability density function with a not Gaussian maximum. An example are two dimensional densities with a single 'banana-shaped' maximum. An unnormalised distribution of this kind is given by

$$h(\mathbf{X}) = \exp\left(-q\mathbf{X}^T C^{-1} \mathbf{X}\right) \quad \text{with} \quad \mathbf{X} = \begin{pmatrix} x \\ y + sx^2 - as \end{pmatrix} \quad C = \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix} \quad (6.8)$$

problem are usually only approximate solutions. By rescaling the factors in this way one gets a better estimation of the hypersphere and therefore a better uncertainty estimation.

The additional quadratic term in the second component of \mathbf{X} generates this special shape, which is controlled by the parameter s . It bends the shape of the outer parts from a normal Gaussian ellipse towards negative (for positive s -values) y -values, giving it the banana-like shape.

The contours of the distribution are shown in fig. 6.1 with the special choice of

$$a = 0.025 \quad b = 0.075 \quad s = 0.075 \quad q = 0.015.$$

Obviously the maximum can not be described by a Gaussian approximation. The error estimation from the last section will fail for two reasons: First the lines along the eigendirections will not be of parabola shape. Therefore the confidence intervals (even for the parameters themselves) will not reflect the true behaviour. Second the estimation of the covariance matrix itself will not be accurate, since the Gamma-Method implies the existence of a quadratic maximum. A Gaussian approximation with a well defined covariance matrix only exists very close to the maximum. The true shape of the distribution cannot be described in this manner.

The problem becomes even more pronounced when looking at the 1D marginal distributions for x and y . In fig. 6.2 the samples from the distribution generated via the adaptive Metropolis Hastings algorithm are displayed. From the lower left plot we see that the distribution gets accurately sampled. If we look at the marginal distribution for x (upper left plot), we see that this parameter can be accurately described via a normal distribution. So all relevant information about this parameter can be communicated via the mean and variance. On the other hand the distribution for y (lower right plot) is skewed. The red line indicates the 'best fit' with a normal distribution. In fact if one were to describe the distribution with the mean and some confidence interval, one would have to use asymmetric ranges. This is the reason why the methods explained above for error estimation will fail. The error estimation for the covariance matrix from the Gamma-Method is $\mathcal{O}(10^4)$ times larger than the actual value indicating not convergence. Naturally the question arises how to generate confidence intervals in such a case. The answer is to use the samples directly to investigate the distribution of the secondary observable. More explicit compute the desired quantity for every sample and use the outcome as the proper probability distribution of the secondary observable, since this is what the samples actually represent. This way no information gets lost and the user can still decide if a Gaussian approximation is sufficient or if asymmetric error bands defined via the distribution are necessary. If the latter is the case one usually computes the 1σ band by including the same amount of probability mass until 68% of the total mass is reached on the left- and right-hand side of the mean value. The length from the mean to the borders on both sides are the confidence intervals respectively.

Usually there are several thousands of samples and some observables may take a long time to compute. Therefore it would be advantageous to reduce the amount of samples, whilst retaining their properties. This matter has been investigated in [14] with the result that any subsampling decreases the amount of information stored. It should only be done when compromising the (time) costs, see especially chapter 3.6. In the framework of parton distribution functions there is a method developed in [9], but it discusses the

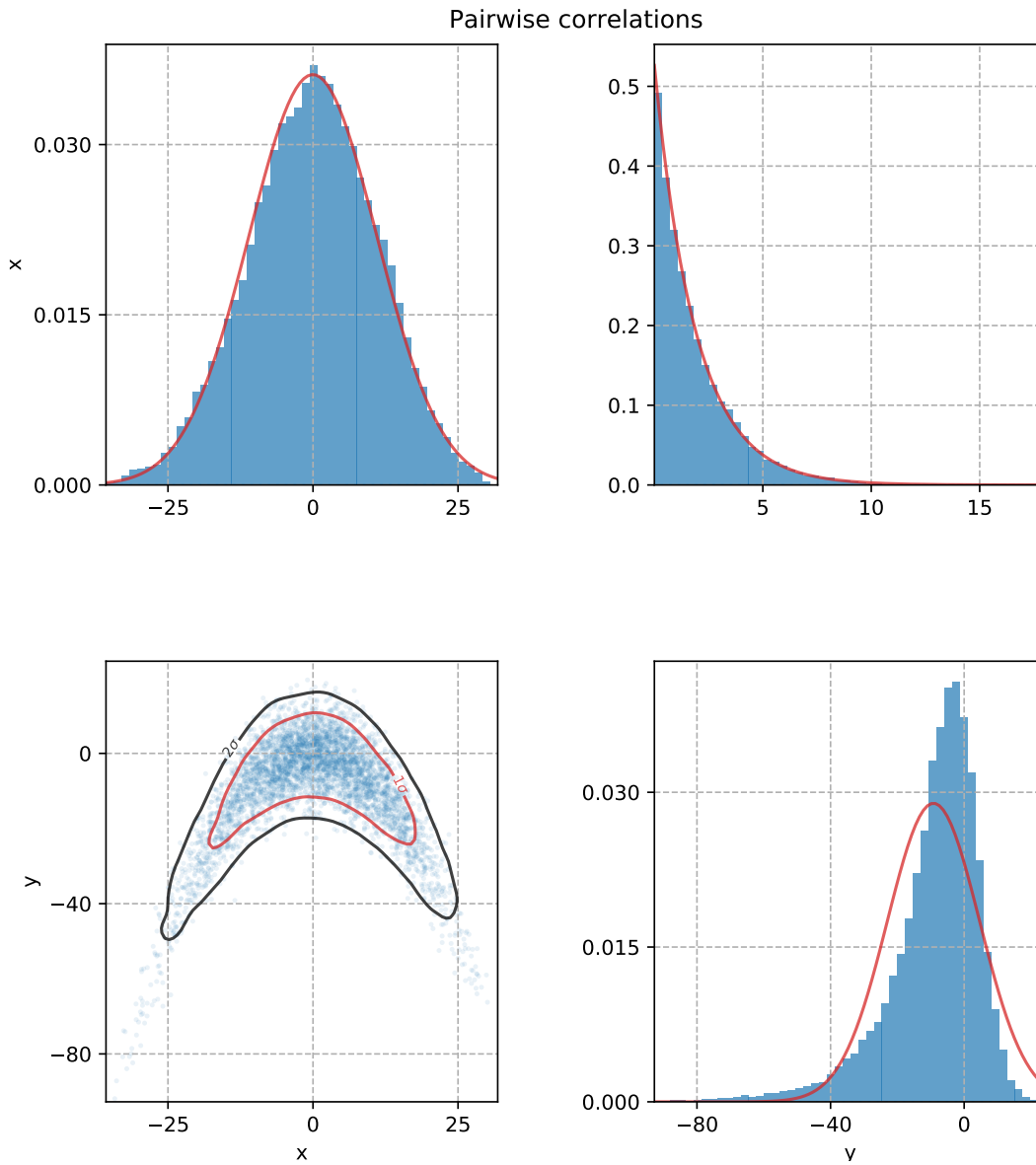


Figure 6.2: A pairwise plot of samples generated from the distribution described above via the adaptive Metropolis Hastings algorithm. The 2D density representation (lower left plot) displays the distribution accurately. Whilst the samples for the x -parameter (upper left plot) are well described via a normal distribution, we find the samples for the y -parameter skewed towards higher values. The upper right plot shows a histogram of the absolute values in the exponent of the distribution.

special setting, when the distributions are well described by a normal distribution and not in a general setting. The latter case is open for research.

7. Proton valence PDFs from DIS experiments

This section finally combines the results of every section above. In particular we will consider a ten parameter fit to the experimental measurements of the proton and deuterium F_2 values. The ten parameters are divided into the five up-valence quark distributions and the latter five are from the down valence PDFs, fitting the full valence contribution of the proton. The underlying parametrisation has been presented in section 2.5. Since these are measurements with correlated uncertainties, the invariant distribution is given by exponentiating the correlated χ^2 -function (see section 3.6). We will use the adaptive Metropolis Hastings algorithm to generate a Markov Chain, which is then to be analysed. From there on the error PDFs are obtained. To ensure reliability of the data, the fit is investigated by the Hessian method (see section 6.2) and compared to the experimental measurements. Furthermore we compare the results to a complementary analysis with well understood minimisation techniques.

7.1. Generating samples

The main problem in generating the samples is the numeric complexity of the problem. Every χ^2 -evaluation takes about 50 seconds for the 992 experimental data points. Therefore we tried to reduce the number of evaluations as much as possible. We generated a starting point via a normal distribution around the best fit values of the nCTEQ15 fit. Starting at a point close to the maximum allows for a reduction of the burn-in length. Second in order to be more certain that the Markov Chain converges, we decided not to use many but shorter runs and rather concentrate on a single but very long run.

In total 30.000 samples have been generated in six and a half days of computing time. The time series is shown in fig. 7.1. As one can see from the plot the algorithm has been restarted three times at its current mean value in the burn-in phase. This resetting helps the convergence, since the proposal distribution is then only estimated with the help of the newer samples. The old samples, which are strongly biased towards the starting point, therefore do not influence the proposal distribution. Here we set the number of random walk proposals¹⁹ to 2000 and then the adaptive proposal distribution is used. After every reset one can clearly see that the chain mixes better, benefiting convergence and therefore less samples are needed.

After 20.000 samples have been taken, the chain seems to have converged²⁰. A zoom-in on the last 10.000 samples is shown in fig. C.2. Here one can see that the chain does not drift any more and furthermore shows good mixing, indicating an integrated autocorrelation time close to 0.5. In fact the acceptance rate in this area is at 89.5%,

¹⁹The random walk proposal uses normal distributions around the nCTEQ15 best fit value before resetting and afterwards the mean value at the point of resetting. The widths are first given by 1% of the parameter value and when reset to the square root of the current diagonal entries of the covariance matrix.

²⁰Since we only created one long chain, it is not possible to say whether this is the global maximum of the invariant distribution. Several other, but shorter runs also converged towards this region giving at least some confidence.

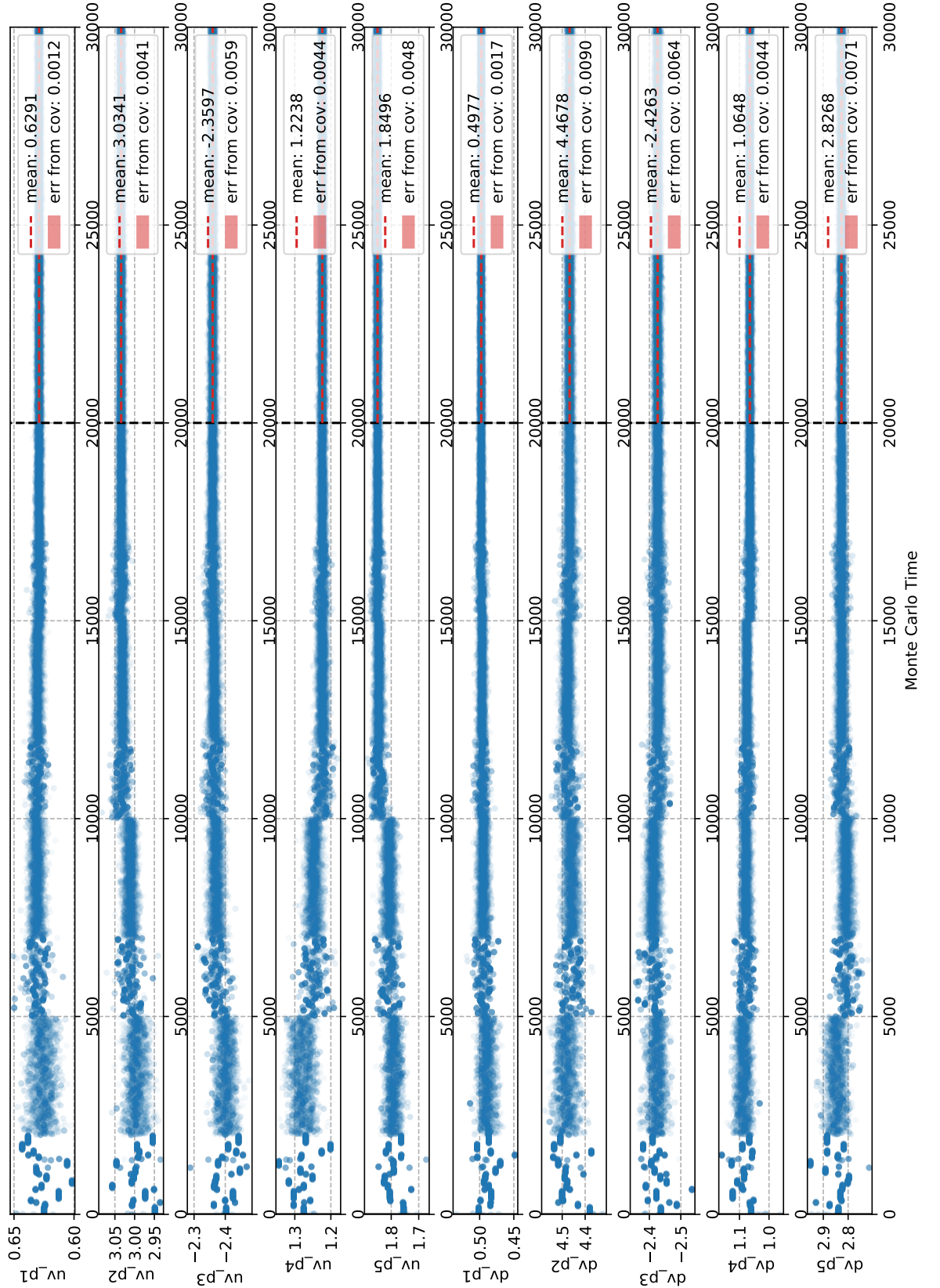


Figure 7.1: The full time series (30.000 samples) of the ten parameter fit generated via the adaptive Metropolis Hastings algorithm. The algorithm was reset to its current mean value at three stages. The dashed black line indicates the starting point of the analysis.

7.1. Generating samples

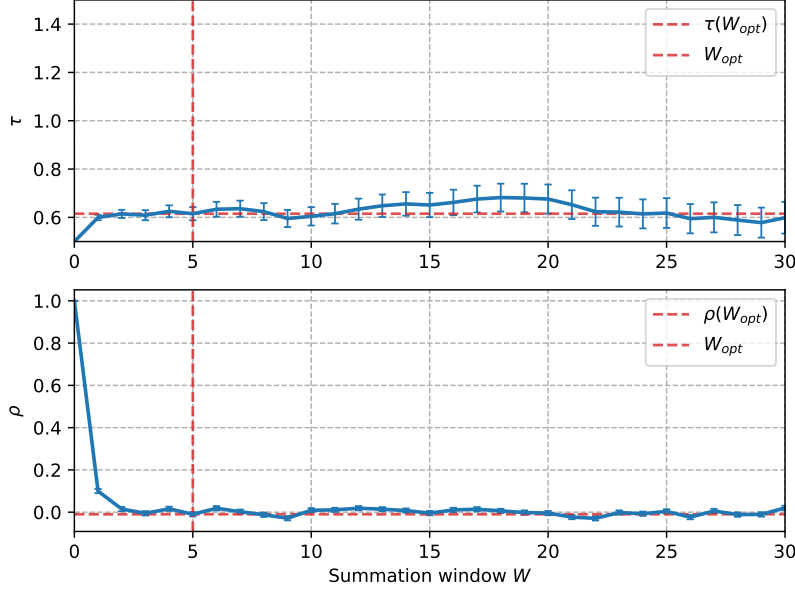


Figure 7.2: The projected integrated autocorrelation time τ and normalised autocorrelation function $\rho(t)$. The chain shows very good statistics, which results in a very strong decrease of $\rho(t)$ and τ to be very close to 0.5. S has been taken to be 2.

which is surprisingly high, because an acceptance rate close to 100% is often connected to strong autocorrelations. Given the good properties of this part of the chain, the rest of the analysis will be based on these samples.

The first analysis to carry out is to estimate the covariance matrix and the integrated autocorrelation time with the Gamma method. By making use of the automatic summation windowing procedure, the projected integrated autocorrelation time, which is depicted as a function of the summation window in fig. 7.2, is estimated as $\tau = 0.62 \pm 0.03$. In fact the autocorrelation time for each parameter individually lies in $[0.58, 0.66]$. The chain indeed shows very good statistics confirming the presumptions above.

The next indicator for a well behaved Markov Chain are the marginal one and two dimensional distributions. They have been created in fig. 7.3, with the 1D distributions fitted via a normal distribution on the diagonal and the 2D distributions on the lower left part. There is no sign of skewness on the diagonal plots. The kernel density estimations show that the pairwise distributions are well described by 2D Gaussians. Putting these two facts together along with low integrated autocorrelation time one can conclude that the data set is well described via the full covariance matrix. The marginalised confidence intervals from the covariance matrix are already given in fig. 7.1 (and also in fig. C.2). Consequentially this leads to the assumption that the minimum in the χ^2 function is of quadratic nature. The only mismatch is the χ^2 distribution itself, which has been plotted and fitted in the upper right corner of fig. 7.3. The larger χ^2 -values deviate more strongly compared to their overall distribution fit. An explanation would be that the chain in fact did not yet converge, because then the χ^2 -values would be governed by

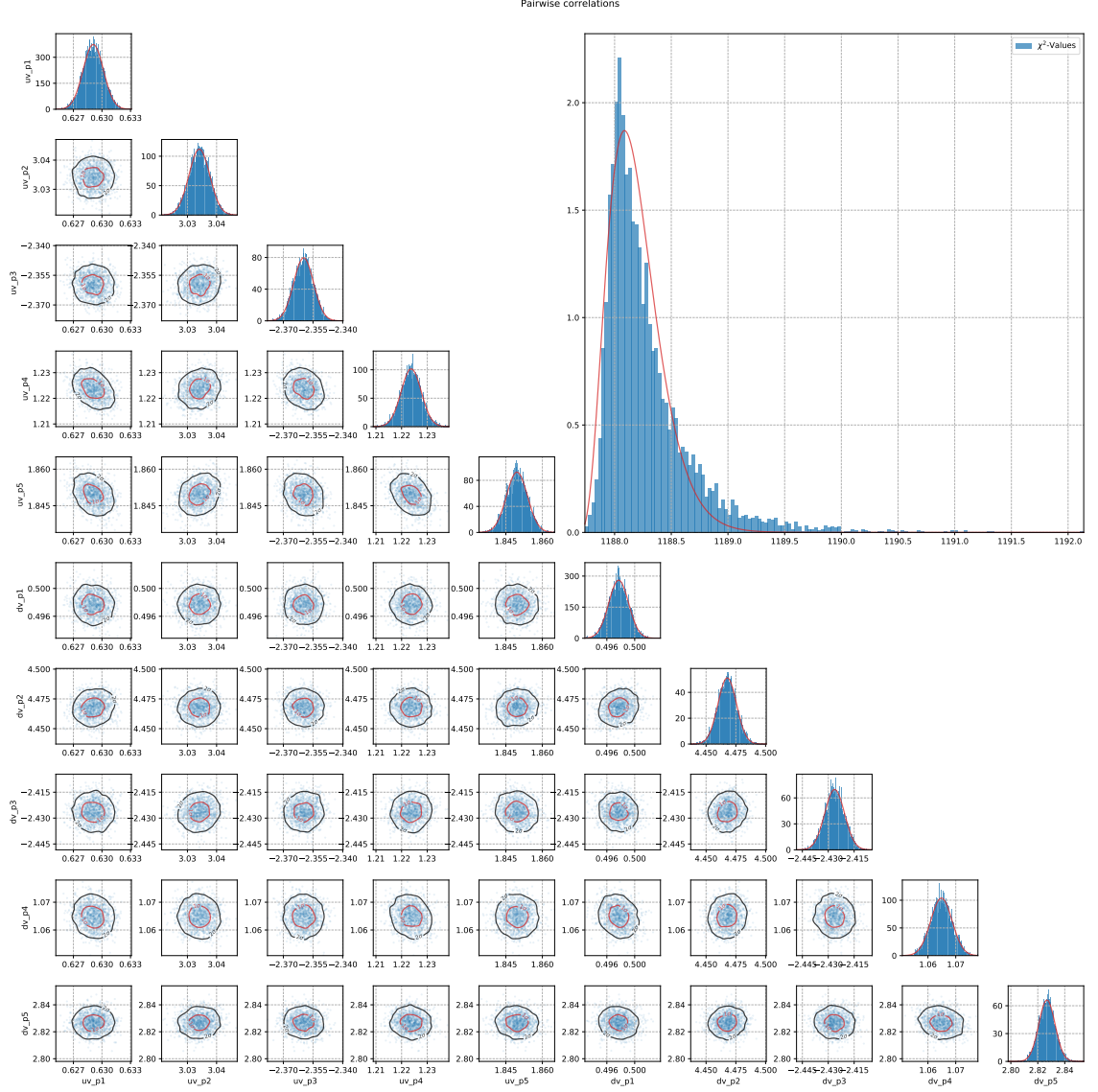


Figure 7.3: The one and two dimensional marginal distributions for each parameter, along with the χ^2 distribution. The 1D distributions are all well estimated by a normal distribution and 2D representation are all of elliptic type. The χ^2 values are slightly too heavy weighted on the higher side.

7.2. Comparison with experimental measurements

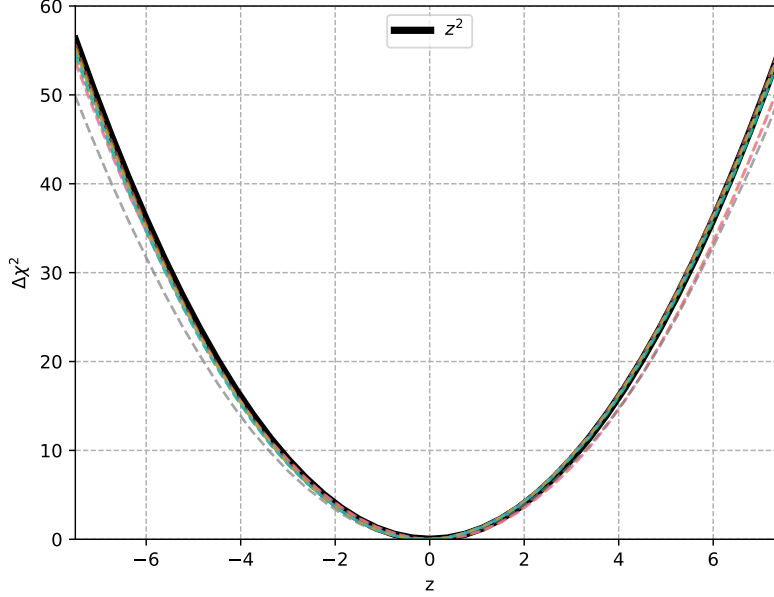


Figure 7.4: Scans of $\Delta\chi^2$ along all eigenvectors with a unit parabola in black as comparison. The curves follow the black line closely, indicating a quadratic minimum.

higher values in the beginning of the time series and would drift in time towards lower values. By looking at fig. C.1, which displays the χ^2 values as a function of the Monte Carlo time, we can rule out this idea or more precisely if this were true, the drift would be at a very slow pace, since it is not visible. After all the question is left open.

In conclusion, we have generated a Markov chain of 10.000 samples with excellent properties. Furthermore we are confident that the chain converged towards a stationary distribution. The samples are close to being independent from each other and the marginal distributions are well described by Gaussians. The data set is characterised by the covariance matrix and an investigation of secondary observables with the Hessian method is justified and as we will see yields similar results.

7.2. Comparison with experimental measurements

After validating the process of sample generation one has to check whether the data set can recreate the experimental measurements. As discussed above the Hessian approach is applicable. The mean value of the fit has $\chi^2_{\text{mean}} = 1188.3$ resulting in $T = 7.32$ which we will round up slightly to $T = 7.5$ giving $t = 2.4$. With these values at hand we will first investigate the χ^2 -function along the eigendirections of the covariance matrix. As explained in section 6.2 these scans along z_α should follow a unit parabola after appropriate rescaling of the eigendirections. The scans along the eigendirections are displayed in fig. 7.4 and are close to the unit parabola indicating good agreement with the discussion from the last section.

The comparison with the experimental measurements is given in fig. 7.5. For this depiction

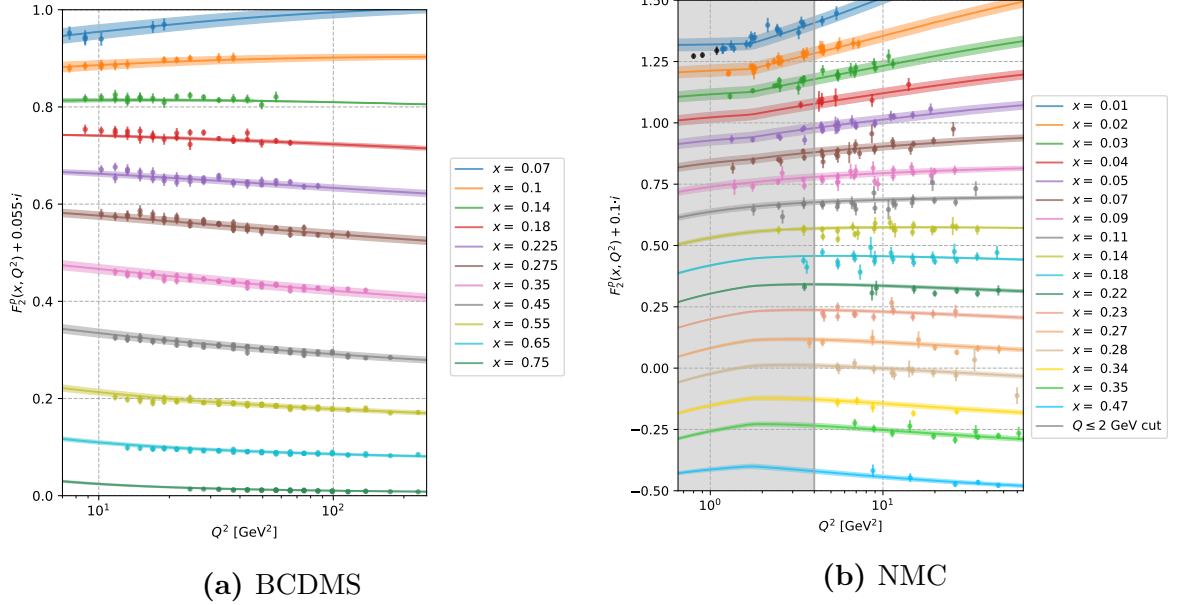


Figure 7.5: The F_2^p proton predictions with confidence intervals compared to the experimental measurements ([5],[4] and [3]). The $Q \leq 2$ GeV excluded region has been marked with a grey layer and the black points are due to the $W \leq 3.5$ GeV cut.

the F_2 values for the proton have been calculated along Q^2 values whilst x has been kept fixed. The measurements from the BCDMS collaboration are performed on a regular grid regarding the x direction. The NMC collaboration however is neither regular in the x nor the Q^2 direction (see fig. 2.5). In order to compare the data with the fit in a single plot, the x values have been rounded to two decimal places. The prediction is then made with the rounded x values. By making use of the eigendirections and (6.4) the confidence intervals are obtained.

The predictions are in very good agreement with the data from the BCDMS group, almost every point is met within the 1σ confidence interval. The NMC data however shows discrepancies (e.g. $x = \{0.07, 0.11, 0.28\}$), but they are single measurements and might rather be statistical outliers than physically meaningful results. In the excluded region the data is predicted consistently except for points within $x = 0.01$ and $Q^2 < 1$. These disagreements are likely due to higher twist effects, which was the reason for excluding them in the first place. The deuteron measurements are equally well fit.

One has to keep in mind that this fit is only done for the valence distributions and only from DIS data. It is due to these shortcomings that the uncertainties have to be discussed with care. Gluon and sea quark distributions play an important role, especially in the small x region. Because of the nature of the Hessian method, there is no straightforward way of including uncertainties for parameters which are kept fixed. This leads to an overconfidence of the predictions. But after all this analysis is just to show that the methods discussed are useful for application and have not (yet) been applied for a full analysis.

7.3. Comparison with Levenberg-Marquardt minimisation

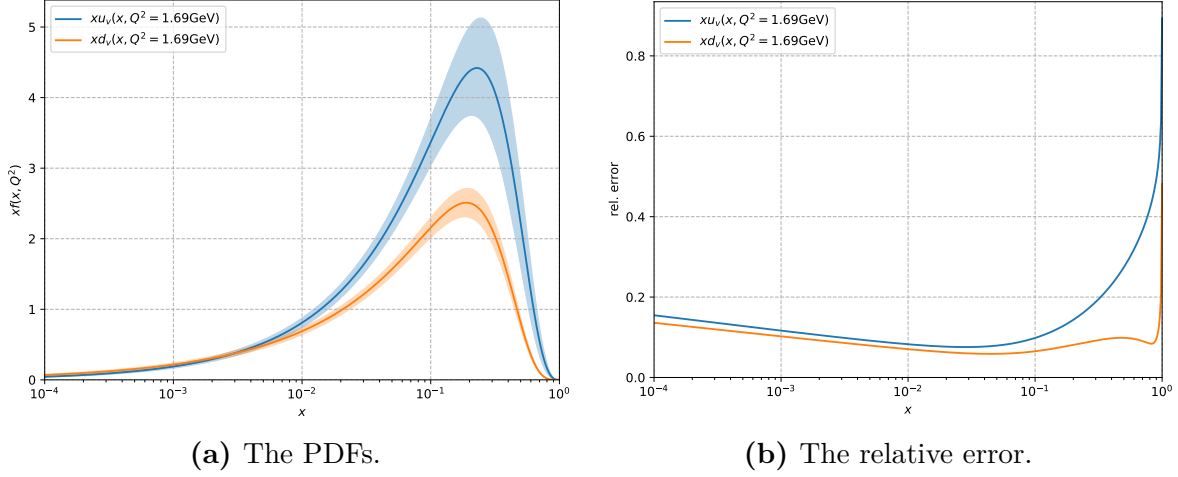


Figure 7.6: The u_v and d_v parton distribution functions with the confidence interval on the left and the relative uncertainty on the right. The confidence intervals have been obtained with the Hessian approach.

7.3. Comparison with Levenberg-Marquardt minimisation

Before we can compare the results it is useful to look at the error PDFs themselves. Again the confidence intervals are obtained by making use of the Hessian approach to ensure better comparability later. On the left side of fig. 7.6 the u_v and d_v distributions with their uncertainty intervals are shown. The plot on the right shows the relative uncertainty given by $\Delta f_i / f_i$. The absolute uncertainty is the largest at the maxima of the distributions, whereas the relative plot shows that it is actually slightly less than the uncertainty in the low x regime. Here one has to neglect the right most part ($x \approx 1$), because the distributions tend towards zero at this point, giving large contributions of numerical inaccuracy in the ratio. This result is not surprising, since almost all of the data lies in $x \in [0.01, 0.75]$. The actual parameter values are given in table 7.1.

The complementary analysis, which minimised the χ^2 function with the Levenberg-Marquardt algorithm, has been performed in [12]. The same experimental measurements and the nCTEQ codebase were used for calculating the minimisation. For the analysis the Hessian approach was also used to verify the result and get confidence intervals. The comparison of the parameter values and the χ^2 value are also given in table 7.1.

Table 7.1: The parameter values (upper row u_v , lower row d_v) from the MCMC approach in comparison to the Levenberg-Marquardt algorithm. The parameters differ significantly and the obtained χ^2 values indicates a better fit obtained via LM.

method	p_1	p_2	p_3	p_4	p_5	χ^2
MCMC	0.6291	3.0341	-2.3598	1.2238	1.8496	1188.3
	0.4977	4.4678	-2.4263	1.0648	2.8268	
LM	0.6242	3.0439	-1.7325	1.5296	1.3055	1184.2
	0.5331	7.7468	4.4962	0.2900	2.1484	

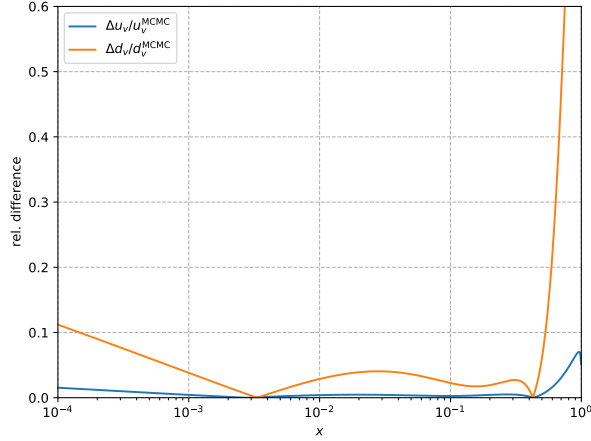


Figure 7.7: The relative difference between the u_v and d_v distribution obtained with the Levenberg-Marquardt minimisation and the MCMC method. The relative difference is below the confidence interval (see fig. 7.6). This means that both parameter sets yield the same prediction.

As one can see a different minimum was obtained. The χ^2 value of the minimiser is lower by 4.1 points and if one compares the individual parameters, a significant difference is apparent, especially for the d_v parameters p_{2-5} . With this information at hand it is clear that the Monte Carlo algorithm did not converge towards the global maximum of the distribution. Instead it found a pronounced local maximum as one can see from the scans along the eigendirections (see fig. 7.4).

Naturally the question arises how both sets of parameters are able to describe the experimental measurements, since the predictions met the data well as seen above. Here the nCTEQ parametrisation comes into plan. It turns out that different parameter values can describe the same overall behaviour of the distributions. In fact if one considers the relative difference of the best fit functions given by

$$\frac{|f_i^{\text{MCMC}} - f_i^{\text{LM}}|}{f_i^{\text{MCMC}}},$$

one can see that the parameter set from the LM-algorithm lies within the confidence interval from the MCMC fit. The ratio has been depicted in fig. 7.7.

In conclusion both fits yield the same physical prediction, which explains why the difference in χ^2 is so low ($\Delta\chi^2/\text{d.o.f.} \approx 0.004$). This test further confirms the applicability of Markov Chain Monte Carlos as a valid fitting method, albeit one caveat has to be made concerning the guarantee of finding the global maximum: There is none. But this is also the case for trying to find the minimum with a minimiser.

8. Conclusion

This thesis started by introducing the Parton model in the framework of Deep Inelastic scattering. The key statements and results have been explained before connecting the naive Parton model with QCD, which lead towards the DGLAP equations. With these facts at hand the difficulties of numerical predictions and the necessity of fitting theory parameters from experimental measurements became clear.

In order to explain how to solve such an inverse problem Bayesian probability theory was introduced, which allows to assign probabilities for parameter values when additional information is given. The final conclusion was to use a probability distribution, which has the correlated χ^2 -function as its key ingredient. Maximising this distribution with respect to the parameters yields the best fit.

Since direct sampling from the distribution is not possible the concepts of Markov Chains were explained giving the opportunity to compute expectation values from educatedly chosen samples. This rephrases the goal 'to find the best fit' in 'computing expectation values from probability distributions'. In this context also the analysis of chains was discussed. In particular, the Gamma-Method was used to find out certain properties and the reliability of given chains. A further section investigated the construction of uncertainty intervals for functions, which take the parameters as arguments. Here statistical inference, the Hessian approach and Markov Chain Monte Carlo inference was discussed. The Hessian approach is a widely used and well understood technique, but has shortcomings if the best fit is not well approximated via normal distributions. Here the MCMC inference demonstrates its strengths, because it does not rely on any approximations.

Markov Chain Monte Carlo algorithms are a numeric way of constructing chains with the desired property of computing expectation values from it. First the omnipresent Metropolis-Hastings algorithm was considered, until some shortcomings lead to the adaptive Metropolis-Hastings algorithm. This algorithm tries to estimate the correlations between the parameters at runtime making it extremely efficient, which is needed due to the difficult theoretical predictions.

Finally the concepts above were combined in a test case: The proton valence distributions were fitted from DIS experiments. In total 30.000 parameter samples were created, from which, after detailed investigation of their properties, 10.000 were used in the final analysis. Then the fit predictions were compared to the experimental measurements and the fit itself was collated with a complementary fit obtained with a minimisation algorithm. The Markov Chain Monte Carlo analysis was able to meet the experimental measurement consistently and also both fitting techniques coincided in the physical predictions.

In conclusion the Markov Chain Monte Carlo approach demonstrates being equally applicable as standard minimisation techniques, when estimating Parton Distribution functions. The nCTEQ codebase has been successfully extended and it is in theory possible to carry out a global analysis. The code is structured in a way, which makes it easy to expand it further with other Monte Carlo algorithms.

In this thesis the MCMC approach was only used in the well known regime of valence

distributions from DIS experiments. It still has to verify its strength at a full analysis with parton distributions, where the quadratic approximations do not hold very well. More explicitly: In situations where the standard approaches do not guarantee reliable results.

Furthermore a procedure to compress Markov Chains will be needed if one wants to publish the samples directly. It would be advantageous if the chain was reduced to hundreds samples without losing its properties. This would shorten the time a reader would need for further analysis of the data.

With the current pace a full analysis is not within reach, because the code needs a speed-up. Therefore the codebase could be modified to be able to run on a cluster. Parallelisation of certain parts of the code is also a possibility. This would also allow for different Monte Carlo algorithms, because most of them are optimised to sample from easy to calculate distributions (or distributions where the difference between two points can be computed fast). Here the 'Hamiltonian Monte Carlo' algorithm is the state-of-the-art, which needs to calculate gradients of the probability distribution, but shows very good convergence properties.

A. Calculations regarding the theory of Markov Chains

Some derivations were excluded from the main text. Here the calculations are carried out.

A.1. Leading bias of the autocorrelation function

First, we want to calculate the bias of the estimator for the autocorrelation function $\Gamma_{\alpha\beta}$. The main idea is to replace the ensemble mean by $\bar{x}_\alpha = \bar{\delta}_\alpha + X_\alpha$. This way we can identify analytic expressions for the autocorrelation function and the covariance matrix.

$$\begin{aligned}
\langle \bar{\Gamma}_{\alpha\beta}(t) \rangle &= \left\langle \frac{1}{N - Rt} \sum_{r=1}^R \sum_{i=1}^{N_r-t} (x_\alpha^{i,r} - \bar{x}_\alpha)(x_\beta^{i+t,r} - \bar{x}_\beta) \right\rangle \\
&= \frac{1}{N - Rt} \sum_{r=1}^R \sum_{i=1}^{N_r-t} \langle (x_\alpha^{i,r} - \bar{\delta}_\alpha - X_\alpha)(x_\beta^{i+t,r} - \bar{\delta}_\beta - X_\beta) \rangle \\
&= \frac{1}{N - Rt} \sum_{r=1}^R \sum_{i=1}^{N_r-t} \left[\langle (x_\alpha^{i,r} - X_\alpha)(x_\beta^{i+t,r} - X_\beta) \rangle - \langle \bar{\delta}_\alpha(x_\beta^{i+t,r} - X_\beta) \rangle \right. \\
&\quad \left. - \langle (x_\alpha^{i,r} - X_\alpha)\bar{\delta}_\beta \rangle + \langle \bar{\delta}_\alpha\bar{\delta}_\beta \rangle \right] \\
&\approx \Gamma_{\alpha\beta}(t) - \frac{1}{N - Rt} \sum_{r=1}^R \sum_{i=1}^{N_r-t} \left[\langle \bar{\delta}_\alpha(x_\beta^{i+t,r} - X_\beta) \rangle + \langle (x_\alpha^{i,r} - X_\alpha)\bar{\delta}_\beta \rangle \right] + \frac{C_{\alpha\beta}}{N}
\end{aligned}$$

The last line uses the definition of the autocorrelation function and the result from (4.17). We see that the estimator indeed produces values centred around the exact autocorrelation function $\Gamma_{\alpha\beta}(t)$, but there are also other contributions. We will now complete the summation in the centred part in order to obtain the ensemble mean again:

$$\begin{aligned}
\frac{1}{N - Rt} \sum_{r=1}^R \sum_{i=1}^{N_r-t} \langle \bar{\delta}_\alpha(x_\beta^{i+t,r} - X_\beta) \rangle &= \frac{1}{N - Rt} \sum_{r=1}^R \sum_{i=1}^{N_r} \langle \bar{\delta}_\alpha(x_\beta^{i,r} - X_\beta) \rangle \\
&\quad - \frac{1}{N - Rt} \sum_{r=1}^R \sum_{i=1}^t \langle \bar{\delta}_\alpha(x_\beta^{i,r} - X_\beta) \rangle \\
&= \frac{N}{N - Rt} \langle \bar{\delta}_\alpha\bar{\delta}_\beta \rangle - \frac{1}{N - Rt} \sum_{r=1}^R \sum_{i=1}^t \langle \bar{\delta}_\alpha(x_\beta^{i,r} - X_\beta) \rangle \\
&\approx \frac{1}{N - Rt} C_{\alpha\beta} - \frac{1}{N - Rt} \sum_{r=1}^R \sum_{i=1}^t \langle \bar{\delta}_\alpha(x_\beta^{i,r} - X_\beta) \rangle
\end{aligned}$$

Similarly we obtain

$$\frac{1}{N - Rt} \sum_{r=1}^R \sum_{i=1}^{N_r-t} \langle (x_\alpha^{i,r} - X_\alpha)\bar{\delta}_\beta \rangle \approx \frac{1}{N - Rt} C_{\alpha\beta} - \frac{1}{N - Rt} \sum_{r=1}^R \sum_{i=1}^t \langle (x_\alpha^{N_r-t+i,r} - X_\alpha)\bar{\delta}_\beta \rangle$$

where we added the final part of the sum instead of the beginning. Collecting all contributions including the correlation matrix we get

$$C_{\alpha\beta} \left(\frac{1}{N} - \frac{2}{N - Rt} \right) = -\frac{C_{\alpha\beta}}{N} \left(1 + 2 \sum_{k=1}^{\infty} \left(\frac{Rt}{N} \right)^k \right).$$

Finally we arrive at

$$\begin{aligned} \langle \bar{\Gamma}_{\alpha\beta}(t) - \Gamma_{\alpha\beta}(t) \rangle &= -\frac{C_{\alpha\beta}}{N} \left(1 + 2 \sum_{k=1}^{\infty} \left(\frac{Rt}{N} \right)^k \right) \\ &\quad - \frac{1}{N - Rt} \sum_{r=1}^R \sum_{i=1}^t \left[\langle \bar{\delta}_{\alpha}(x_{\beta}^{i,r} - X_{\beta}) \rangle + \langle (x_{\alpha}^{Nr-t+i,r} - X_{\alpha}) \bar{\delta}_{\beta} \rangle \right] \\ &\approx -\frac{C_{\alpha\beta}}{N}. \end{aligned} \tag{A.1}$$

Thus the leading bias is given by $-\frac{C_{\alpha\beta}}{N}$. One could remove this bias by normalising the summation with $N - 1$ instead of N . This matter is explained in section 4.4.

A.2. The variance of the autocorrelation function

Here we compute the variance of the autocorrelation function, where we keep the indices as general as possible, since we will use the result for various other calculations. Specifically we will deal with

$$\begin{aligned} &\langle (\bar{\Gamma}_{\alpha\beta}(s) - \Gamma_{\alpha\beta}(s))(\bar{\Gamma}_{\gamma\delta}(t) - \Gamma_{\gamma\delta}(t)) \rangle \\ &= \langle \bar{\Gamma}_{\alpha\beta}(s) \bar{\Gamma}_{\gamma\delta}(t) \rangle + \Gamma_{\alpha\beta}(s) \Gamma_{\gamma\delta}(t) - \Gamma_{\alpha\beta}(s) \langle \bar{\Gamma}_{\gamma\delta}(t) \rangle - \Gamma_{\gamma\delta}(t) \langle \bar{\Gamma}_{\alpha\beta}(s) \rangle \\ &\approx \langle \bar{\Gamma}_{\alpha\beta}(s) \bar{\Gamma}_{\gamma\delta}(t) \rangle + \Gamma_{\alpha\beta}(s) \Gamma_{\gamma\delta}(t) + \Gamma_{\alpha\beta}(s) \frac{C_{\gamma\delta}}{N} + \Gamma_{\gamma\delta}(t) \frac{C_{\alpha\beta}}{N} \\ &\approx \langle \bar{\Gamma}_{\alpha\beta}(s) \bar{\Gamma}_{\gamma\delta}(t) \rangle - \left(\Gamma_{\alpha\beta}(s) - \frac{C_{\alpha\beta}}{N} \right) \left(\Gamma_{\gamma\delta}(t) - \frac{C_{\gamma\delta}}{N} \right). \end{aligned}$$

The last line neglects a term $\propto 1/N^2$. We have brought the equation to this form, because the four-point correlator given by

$$\langle \bar{\Gamma}_{\alpha\beta}(s) \bar{\Gamma}_{\gamma\delta}(t) \rangle = \frac{1}{(N-s)(N-t)} \sum_{i=1}^{N-s} \sum_{j=1}^{N-t} \langle (x_{\alpha}^i - \bar{x}_{\alpha})(x_{\beta}^j - \bar{x}_{\beta})(x_{\gamma}^i - \bar{x}_{\gamma})(x_{\delta}^j - \bar{x}_{\delta}) \rangle,$$

can be approximated by its disconnected part (see [44, appendix A], [24, appendix C] or [41, appendix A]). We also assume t, s to be small compared to N , since this is fulfilled

A.3. Error of the projected normalized autocorrelation function

in every application we would like to carry out. This simplifies the equation further to

$$\begin{aligned} \langle \bar{\Gamma}_{\alpha\beta}(s) \bar{\Gamma}_{\gamma\delta}(t) \rangle &\approx \left(\Gamma_{\alpha\beta}(s) - \frac{C_{\alpha\beta}}{N} \right) \left(\Gamma_{\gamma\delta}(t) - \frac{C_{\gamma\delta}}{N} \right) \\ &+ \frac{1}{N^2} \sum_{i,j=1}^N [\Gamma_{\alpha\gamma}(j-i) \Gamma_{\beta\delta}(j-i+t-s) + \Gamma_{\alpha\delta}(j-i+t) \Gamma_{\beta\gamma}(j-i-s)]. \end{aligned}$$

Lastly there is one further simplification possible by making use of $\sum_{i,j=1}^M g(i \pm j) \approx M \sum_{k=-\infty}^{\infty} g(k)$ for rapidly decaying functions. This yields the final result:

$$\begin{aligned} &\langle (\bar{\Gamma}_{\alpha\beta}(s) - \Gamma_{\alpha\beta}(s)) (\bar{\Gamma}_{\gamma\delta}(t) - \Gamma_{\gamma\delta}(t)) \rangle \\ &\approx \frac{1}{N} \sum_{k=-\infty}^{\infty} [\Gamma_{\alpha\gamma}(k) \Gamma_{\beta\delta}(k+t-s) + \Gamma_{\alpha\delta}(k+t) \Gamma_{\beta\gamma}(k-s)] \\ &= \frac{1}{N} \sum_{k=-\infty}^{\infty} [\Gamma_{\alpha\gamma}(k) \Gamma_{\beta\delta}(k+t-s) + \Gamma_{\alpha\delta}(k) \Gamma_{\beta\gamma}(k-t-s)]. \end{aligned} \quad (\text{A.2})$$

This formula is needed for every further application.

A.3. Error of the projected normalized autocorrelation function

We wish to derive an error estimation for the projected normalized autocorrelation function, which is given by

$$\bar{\rho}(t) = \frac{\bar{\Gamma}(t)}{\bar{\Gamma}(0)} = \frac{\sum_{\alpha\beta} \bar{\Gamma}_{\alpha\beta}(t)}{\sum_{\alpha'\beta'} \bar{\Gamma}_{\alpha'\beta'}(0)}.$$

The first thing to do is calculating the individual variances for $\bar{\Gamma}(t)$ and $\bar{\Gamma}(0)$ and also the crosscorrelation between them. This allows us to use error propagation to get a final approximation.

Starting with $\bar{\Gamma}(t)$ we arrive at

$$\begin{aligned} \langle (\bar{\Gamma}(t) - \Gamma(t))^2 \rangle &= \sum_{\alpha\beta} \sum_{\gamma\delta} \langle (\bar{\Gamma}_{\alpha\beta}(t) - \Gamma_{\alpha\beta}(t)) (\bar{\Gamma}_{\gamma\delta}(t) - \Gamma_{\gamma\delta}(t)) \rangle \\ &\stackrel{(\text{A.2})}{\approx} \sum_{\alpha\beta} \sum_{\gamma\delta} \frac{1}{N} \sum_{k=-\infty}^{\infty} [\Gamma_{\alpha\gamma}(k) \Gamma_{\beta\delta}(k) + \Gamma_{\alpha\delta}(k+t) \Gamma_{\beta\gamma}(k-t)] \\ &= \frac{1}{N} \sum_{k=-\infty}^{\infty} [\Gamma^2(k) + \Gamma(k+t) \Gamma(k-t)]. \end{aligned}$$

Calculating the crosscorrelation term analogously yields

$$\begin{aligned}\langle(\bar{\Gamma}(t) - \Gamma(t))(\bar{\Gamma}(0) - \Gamma(0))\rangle &\approx \sum_{\alpha\beta} \sum_{\gamma\delta} \frac{1}{N} \sum_{k=-\infty}^{\infty} [\Gamma_{\alpha\gamma}(k)\Gamma_{\beta\delta}(k+t) + \Gamma_{\alpha\delta}(k+t)\Gamma_{\beta\gamma}(k)] \\ &= \frac{2}{N} \sum_{k=-\infty}^{\infty} \Gamma(k)\Gamma(k+t)\end{aligned}$$

and finally the last term is given by

$$\begin{aligned}\langle(\bar{\Gamma}(0) - \Gamma(0))^2\rangle &\approx \sum_{\alpha\beta} \sum_{\gamma\delta} \frac{1}{N} \sum_{k=-\infty}^{\infty} [\Gamma_{\alpha\gamma}(k)\Gamma_{\beta\delta}(k) + \Gamma_{\alpha\delta}(k)\Gamma_{\beta\gamma}(k)] \\ &= \frac{2}{N} \sum_{k=-\infty}^{\infty} \Gamma^2(k).\end{aligned}$$

Combining all results from above via error propagation we get a quite detailed formula for the variance of $\bar{\rho}(t)$:

$$\begin{aligned}\langle(\bar{\rho}(t) - \rho(t))^2\rangle &\approx \frac{1}{N} \sum_{k=-\infty}^{\infty} \left[\frac{\Gamma^2(k) + \Gamma(k+t)\Gamma(k-t)}{\Gamma^2(0)} + 2\frac{\Gamma^2(k)\Gamma^2(t)}{\Gamma^4(0)} - 4\frac{\Gamma(k)\Gamma(t)\Gamma(k+t)}{\Gamma^3(0)} \right] \\ &= \frac{1}{N} \sum_{k=-\infty}^{\infty} \left[\rho^2(k) + \rho(k+t)\rho(k-t) + 2\rho^2(k)\rho^2(t) - 4\rho(k)\rho(t)\rho(k+t) \right].\end{aligned}$$

In order to simplify this result we will alter the terms of the sum until we get a perfect square. First we notice the sum over k is carried out over the whole domain of \mathbb{Z} . This can be used to our advantage since the $\rho^2(k)$ can be written as $\frac{1}{2}(\rho^2(k+t) + \rho^2(k-t))$ and $4\rho(k)\rho(t)\rho(k+t)$ can be modified to $2\rho(k)\rho(t)\rho(k+t) + 2\rho(k)\rho(t)\rho(k-t)$. Also we change the summation over the whole domain to a term starting at $k=0$ and taking twice the sum over the positive integers. This is permitted since $\Gamma(t)$ and thus $\rho(t)$ is only dependent on the absolute value of its argument. Also one has to note that the terms add up to 0 when $k=0$. Collecting everything results in a perfect square:

$$\begin{aligned}\langle(\bar{\rho}(t) - \rho(t))^2\rangle &\approx \frac{1}{N} \sum_{k=1}^{\infty} \left[\rho^2(k+t) + \rho^2(k-t) + 2\rho(k+t)\rho(k-t) + 4\rho^2(k)\rho^2(t) \right. \\ &\quad \left. - 4\rho(k)\rho(t)\rho(k+t) - 4\rho(k)\rho(t)\rho(k-t) \right] \\ &= \frac{1}{N} \sum_{k=1}^{\infty} [\rho(k+t) + \rho(k-t) - 2\rho(k)\rho(t)]^2\end{aligned}\tag{A.3}$$

As for usual the sum over k is only carried out until W_{opt} in a numerical implementation.

A.4. The variance of the full and the naive covariance matrix

The most important result from appendix A.2 is the variance of the full covariance matrix. Luckily this is a straight forward calculation. We begin with inserting the variance of the autocorrelation function:

$$\begin{aligned}
 & \langle (\bar{C}_{\alpha\beta}(W) - C_{\alpha\beta})(\bar{C}_{\gamma\delta}(W) - C_{\gamma\delta}) \rangle \\
 &= \sum_{s,t=-W}^W \langle (\bar{\Gamma}_{\alpha\beta}(s) - \Gamma_{\alpha\beta}(s))(\bar{\Gamma}_{\gamma\delta}(t) - \Gamma_{\gamma\delta}(t)) \rangle \\
 &\stackrel{(A.2)}{\approx} \sum_{s,t=-W}^W \frac{1}{N} \sum_{k=-\infty}^{\infty} [\Gamma_{\alpha\gamma}(k)\Gamma_{\beta\delta}(k+t-s) + \Gamma_{\alpha\delta}(k)\Gamma_{\beta\gamma}(k-t-s)]
 \end{aligned}$$

Now we exchange the order of summation and make use of $\sum_{i,j=-M}^M g(i \pm j) \approx (1 + 2M) \sum_{k=-\infty}^{\infty} g(k)$ again. This introduces a second summation over the whole domain of \mathbb{Z} , which we will use to identify the exact covariance matrix.

$$\begin{aligned}
 & \frac{1}{N} \sum_{k=-\infty}^{\infty} \sum_{s,t=-W}^W [\Gamma_{\alpha\gamma}(k)\Gamma_{\beta\delta}(k+t-s) + \Gamma_{\alpha\delta}(k)\Gamma_{\beta\gamma}(k-t-s)] \\
 &\approx \frac{2W+1}{N} \sum_{k,u=-\infty}^{\infty} [\Gamma_{\alpha\gamma}(k)\Gamma_{\beta\delta}(k+u) + \Gamma_{\alpha\delta}(k)\Gamma_{\beta\gamma}(k-u)] \\
 &= \frac{2W+1}{N} \sum_{k,u=-\infty}^{\infty} [\Gamma_{\alpha\gamma}(k)\Gamma_{\beta\delta}(u) + \Gamma_{\alpha\delta}(k)\Gamma_{\beta\gamma}(u)]
 \end{aligned}$$

After separating the sums in the last line, we are able to use the exact covariance matrix, leaving us with a formula easily calculated.

$$\Rightarrow \langle (\bar{C}_{\alpha\beta}(W) - C_{\alpha\beta})(\bar{C}_{\gamma\delta}(W) - C_{\gamma\delta}) \rangle = \frac{2W+1}{N} (C_{\alpha\gamma}C_{\beta\delta} + C_{\alpha\delta}C_{\beta\gamma}) \quad (A.4)$$

The crossvariance between the naive and the full covariance matrix is calculated in a similar manner, but we only have to deal with one summation:

$$\langle (\bar{C}_{\alpha\beta}^0 - C_{\alpha\beta}^0)(\bar{C}_{\gamma\delta}(W) - C_{\gamma\delta}) \rangle \approx \frac{1}{N} \sum_{k=-\infty}^{\infty} \sum_{t=-W}^W [\Gamma_{\alpha\gamma}(k)\Gamma_{\beta\delta}(k+t) + \Gamma_{\alpha\delta}(k)\Gamma_{\beta\gamma}(k-t)]$$

At this point we assume every contribution of the autocorrelation function to be negligible unless the argument is (much) smaller than W . This is consistent with the way W_{opt} is chosen. In other words: We let the summation over t go to infinity again ending up with

the exact covariance matrix:

$$\begin{aligned} \langle (\bar{C}_{\alpha\beta}^0 - C_{\alpha\beta}^0)(\bar{C}_{\gamma\delta}(W) - C_{\gamma\delta}) \rangle &\approx \frac{1}{N} \sum_{k,t=-\infty}^{\infty} [\Gamma_{\alpha\gamma}(k)\Gamma_{\beta\delta}(t) + \Gamma_{\alpha\delta}(k)\Gamma_{\beta\gamma}(t)] \\ &= \frac{1}{N} (C_{\alpha\gamma}C_{\beta\delta} + C_{\alpha\delta}C_{\gamma\beta}). \end{aligned} \quad (\text{A.5})$$

The variance of the naive covariance matrix is more complicated:

$$\langle (\bar{C}_{\alpha\beta}^0 - C_{\alpha\beta}^0)(\bar{C}_{\gamma\delta}^0 - C_{\gamma\delta}^0) \rangle \approx \frac{1}{N} \sum_{k=-\infty}^{\infty} [\Gamma_{\alpha\gamma}(k)\Gamma_{\beta\delta}(k) + \Gamma_{\alpha\delta}(k)\Gamma_{\beta\gamma}(k)]$$

Here we only see one sum. In order to get a formula easily computed, we now try to find an upper limit. This can be achieved by replacing one term in the product of two autocorrelation functions by its biggest possible value. More specifically we replace $\Gamma(k)\Gamma(k)$ by $\Gamma(0)\Gamma(k)$ (omitting all indices at this point), since the first term in these decaying functions is the biggest. In order to treat all indices equally, we split $\Gamma_{\alpha\gamma}(k)\Gamma_{\beta\delta}(k)$ up into $\frac{1}{2}(\Gamma_{\alpha\gamma}(k)\Gamma_{\beta\delta}(k) + \Gamma_{\alpha\gamma}(k)\Gamma_{\beta\delta}(k))$ and replace $\Gamma_{\alpha\gamma}(k)$ in the first term and $\Gamma_{\beta\delta}(k)$ in the second one. Overall we get

$$\begin{aligned} \langle (\bar{C}_{\alpha\beta}^0 - C_{\alpha\beta}^0)(\bar{C}_{\gamma\delta}^0 - C_{\gamma\delta}^0) \rangle &\leq \frac{1}{2N} \sum_{k=-\infty}^{\infty} [\Gamma_{\alpha\gamma}(0)\Gamma_{\beta\delta}(k) + \Gamma_{\alpha\gamma}(k)\Gamma_{\beta\delta}(0) \\ &\quad + \Gamma_{\alpha\delta}(0)\Gamma_{\beta\gamma}(k) + \Gamma_{\alpha\delta}(k)\Gamma_{\beta\gamma}(0)] \\ &= \frac{1}{2N} (C_{\alpha\gamma}^0 C_{\beta\delta} + C_{\alpha\gamma} C_{\beta\delta}^0 + C_{\alpha\delta}^0 C_{\gamma\beta} + C_{\alpha\delta} C_{\gamma\beta}^0), \end{aligned} \quad (\text{A.6})$$

where we have identified the full and naive covariance matrix respectively.

A.5. Analytic results for the test case

We want to analyse the properties of the following Markov Chain

$$\nu^1 = \eta^1 \quad \nu^{i+1} = \sqrt{1 - a^2} \eta^{i+1} + a \nu^i.$$

The random numbers η^i are independently generated from $\mathcal{N}(0, 1)$. By induction it is possible to show that the mean value is zero. The starting point $\langle \nu^1 \rangle$ as the first step of induction is obviously fulfilled due to the properties of the random numbers. The induction step yields:

$$\langle \nu^{i+1} \rangle = \sqrt{1 - a^2} \langle \eta^{i+1} \rangle + a \langle \nu^i \rangle = 0. \quad (\text{A.7})$$

The first term vanishes due to η^{i+1} begin sampled from $\mathcal{N}(0, 1)$ and the second term also yields zero due to the proposition.

In order to compute the autocorrelation function we first have to investigate the variance.

A.5. Analytic results for the test case

After inserting the definition we arrive at

$$\begin{aligned}\langle \nu^i \nu^i \rangle &= (1 - a^2) \langle \eta^i \eta^i \rangle + 2\sqrt{1 - a^2} \langle \eta^i \nu^{i-1} \rangle + a^2 \langle \nu^{i-1} \nu^{i-1} \rangle \\ &= (1 - a^2) + a^2 \langle \nu^{i-1} \nu^{i-1} \rangle.\end{aligned}$$

The term in the middle vanishes since η^i and ν^{i-1} are independent (ν^{i-1} only depends on η^{i-1}) and we therefore get the product of the individual mean values, which are both zero. The last term however points out that the variance is to be proven by induction. With $\langle \nu^1 \nu^1 \rangle = \langle \eta^1 \eta^1 \rangle = 1$ we can use the proposition $\langle \nu^{i-1} \nu^{i-1} \rangle = 1$, yielding

$$\langle \nu^i \nu^i \rangle = (1 - a^2) + a^2 \langle \nu^{i-1} \nu^{i-1} \rangle = 1,$$

which completes the proof.

Finally it is possible to turn to the autocorrelation function. By making use of

$$\nu^{i+k} = \sqrt{1 - a^2} \sum_{j=0}^{k-1} a^j \eta^{i+k-j} + a^k \nu^i$$

it is possible to simplify the autocorrelation function. Starting from the definition one gets

$$\begin{aligned}\Gamma(t) = \langle \nu^i \nu^{i+t} \rangle &= \sqrt{1 - a^2} \sum_{j=0}^{t-1} a^j \langle \eta^{i+t-j} \nu^i \rangle + a^t \langle \nu^i \nu^i \rangle \\ &= a^t.\end{aligned}$$

To get to the second line, one has to notice that the sum over j only introduces η 's which are not of the same index as ν^i . Therefore the expectation splits into the products of the individual mean values, which vanish. Analogously, one could repeat the calculation for $\Gamma(-t)$ and this time lower the index of ν^i to get to the same result. All in all we conclude

$$\Gamma(t) = a^{|t|}. \tag{A.8}$$

With the choice of $a = \frac{2\tau-1}{2\tau+1}$ the integrated autocorrelation time is controllable (see section 4.6):

$$\sum_{t=-\infty}^{\infty} \frac{\Gamma(t)}{\Gamma(0)} = 2\tau.$$

We now imagine generating a Markov Chain with two parameters via

$$x_1^i = X_1 + q(\nu_c^i + \nu_1^i) \quad \text{and} \quad x_2^i = X_2 + q(\nu_c^i + \nu_2^i).$$

In particular, we have three independent (noise) chains with the properties above. Obviously the mean value is given by X_1 and X_2 respectively. A less obvious case is the

autocorrelation function. We begin with x_1 :

$$\begin{aligned}
 \Gamma_{11}(t) &= \langle (x_1^i - X_1)(x_1^{i+t} - X_1) \rangle \\
 &= \langle x_1^i x_1^{i+t} \rangle - X_1 \langle x_1^i \rangle - X_1 \langle x_1^{i+t} \rangle + X_1^2 \\
 &= \langle x_1^i x_1^{i+t} \rangle - X_1^2 \\
 &= X_1^2 + X_1 q \left(\langle \nu_c^i \rangle + \langle \nu_1^i \rangle + \langle \nu_c^{i+t} \rangle + \langle \nu_1^{i+t} \rangle \right) \\
 &\quad + q^2 \left(\langle \nu_c^i \nu_c^{i+t} \rangle + \langle \nu_c^{i+t} \nu_1^i \rangle + \langle \nu_c^i \nu_1^{i+t} \rangle + \langle \nu_1^i \nu_1^{i+t} \rangle \right) - X_1^2 \\
 &= q^2 \left(\langle \nu_c^i \nu_c^{i+t} \rangle + \langle \nu_1^i \nu_1^{i+t} \rangle \right) \\
 &= q^2 \left(a_c^{|t|} + a_1^{|t|} \right)
 \end{aligned}$$

From the fourth to the fifth line, we used the independence between ν_c and ν_1 to split the expectation value of the product into the product of the expectation values. The computation for $\Gamma_{22}(t)$ is analogous one simply has to substitute '2' for each '1', therefore: $\Gamma_{22}(t) = q^2 (a_c^{|t|} + a_2^{|t|})$. For $\Gamma_{12}(t)$ we will get a very similar equation, but $\langle \nu_c^i \nu_c^{i+t} \rangle$ will be the only non-vanishing term, since every other term will be either an expectation value of a single noise chain or the expectation value of the product between two independent noise chains. This yields $\Gamma_{12}(t) = q^2 a_c^{|t|}$, which should not be surprising, since ν_c is the only term correlating both chains. In conclusion we get

$$\Gamma_{11}(t) = q^2 (a_c^{|t|} + a_1^{|t|}) \quad \Gamma_{12}(t) = q^2 a_c^{|t|} \quad \Gamma_{22}(t) = q^2 (a_c^{|t|} + a_2^{|t|}). \quad (\text{A.9})$$

From here it is simple to calculate the naive and full covariance matrix, since we only have to plug in zero for the naive matrix or have to sum over all values for t for the full matrix. This yields

$$C_{11}^0 = 2q^2 \quad C_{12}^0 = q^2 \quad C_{22}^0 = 2q^2 \quad (\text{A.10})$$

$$C_{11} = 2q^2(\tau_c + \tau_1) \quad C_{12} = 2q^2\tau_c \quad C_{22} = 2q^2(\tau_c + \tau_2) \quad (\text{A.11})$$

where we have replaced the sum by the respective autocorrelation time, due to the choice of a_c, a_1 and a_2 as explained above.

B. Relations and identities

B.1. Lab frame kinematics

We use the momentum definitions from fig. 2.1:

$$\begin{array}{ll}
 k : \text{incoming } e^- & p : \text{proton} \\
 k' : \text{outgoing } e^- & q : \text{momentum transfer}
 \end{array}$$

B.2. Gaussian integrals

Furthermore we neglect the electron mass and take the proton at rest, yielding

$$\begin{aligned}k \cdot k' &= 2EE' \sin^2 \frac{\theta}{2} \\k \cdot p &= EM \\k' \cdot p &= E'M \\q &= k - k' \\q^2 &= -4EE' \sin^2 \frac{\theta}{2}.\end{aligned}\tag{B.1}$$

B.2. Gaussian integrals

$$\int_{-\infty}^{\infty} \exp \left[-\frac{1}{2} \sum_{i,j} A_{ij} x_i x_j + \sum_i B_i x_i \right] d^m x = \sqrt{\frac{(2\pi)^N}{\det(A)}} \exp \left[\frac{1}{2} \sum_{ij} A_{ij}^{-1} B_i B_j \right] \tag{B.2}$$

B.3. Matrix manipulation

Jacobi's formula:

$$\frac{\partial}{\partial t} \det A(t) = \text{tr} \left(\text{adj}(A(t)) \frac{\partial A(t)}{\partial t} \right) \tag{B.3}$$

with the special case

$$\frac{\partial}{\partial A_{ij}} \det A = (\text{adj}^T A)_{ij} \tag{B.4}$$

where $\text{adj}(A)$ is the adjugate matrix of A . For invertible matrices it can be computed as $\text{adj} A = \det(A) A^{-1}$.

C. Additional figures

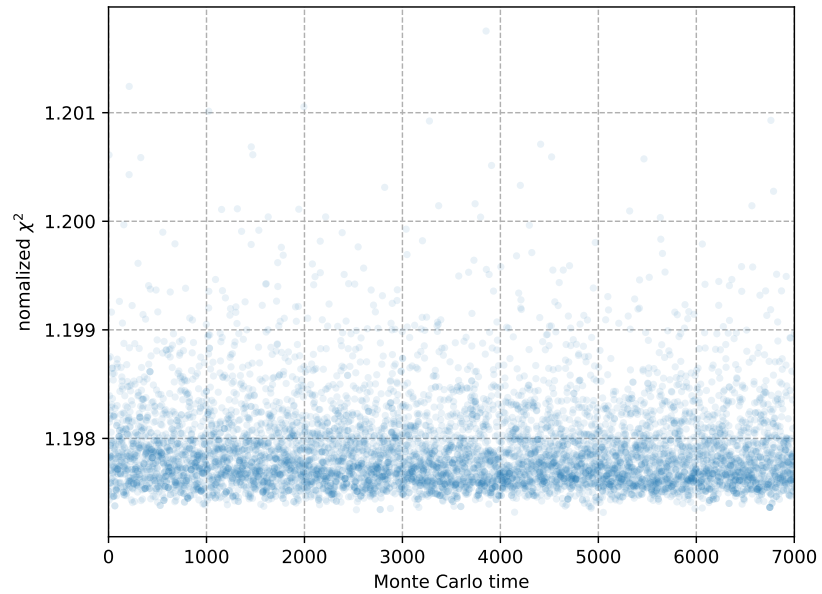


Figure C.1: The time series of the normalised χ^2 values. A drift in this depiction would indicate that the chain did not converge. Please note that some of these χ^2 -values have a multiplicity, meaning the algorithm did not jump directly and stayed at this point several iterations. It is for this reason, that here only ~ 7000 points are drawn.

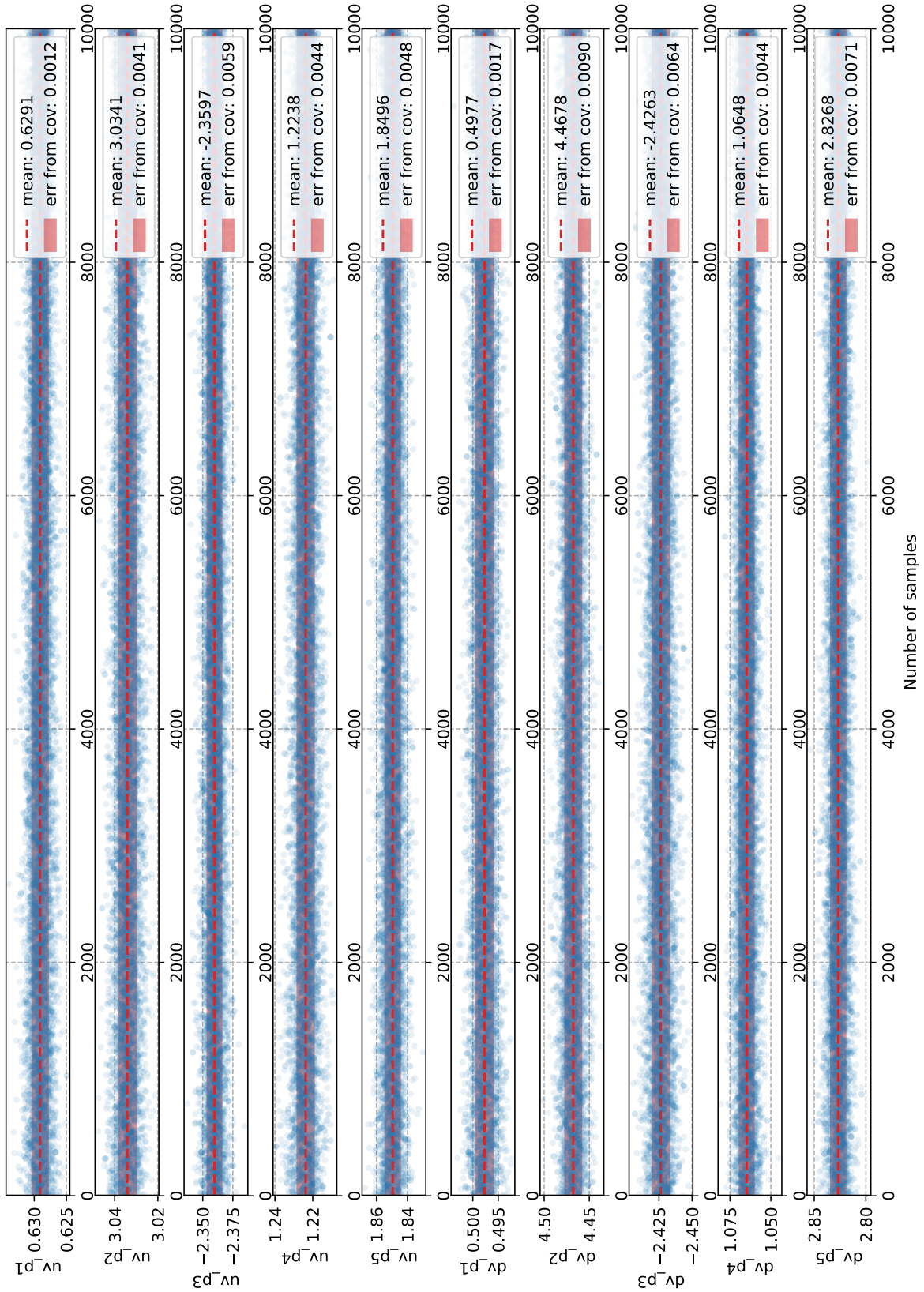


Figure C.2: The analysed 10.000 samples. The chain shows a good mixing, indicating a integrated correlation time close to 0.5. The chain seems to have converged since there is no drift visible.

References

- [1] G. Altarelli and G. Parisi. “Asymptotic freedom in parton language”. In: *Nucl. Phys. B* 126 (1977), pp. 298–318. DOI: 10.1016/0550-3213(77)90384-4.
- [2] Guido Altarelli, R.Keith Ellis, and G. Martinelli. “Large Perturbative Corrections to the Drell-Yan Process in QCD”. In: *Nucl. Phys. B* 157 (1979), pp. 461–497. DOI: 10.1016/0550-3213(79)90116-0.
- [3] M. Arneodo, A. Arvidson, and B. Badeek et al. “Measurement of the proton and the deuteron structure functions, F_2^p and F_2^d ”. In: *Physics Letters B* 364.2 (1995), pp. 107–115. DOI: [https://doi.org/10.1016/0370-2693\(95\)01318-9](https://doi.org/10.1016/0370-2693(95)01318-9).
- [4] A.C. Benvenuti, D. Bollini, and G. Bruni et al. “A high statistics measurement of the deuteron structure functions $F_2(x, Q^2)$ and R from deep inelastic muon scattering at high Q^2 ”. In: *Physics Letters B* 237.3 (1990), pp. 592–598. DOI: [https://doi.org/10.1016/0370-2693\(90\)91231-Y](https://doi.org/10.1016/0370-2693(90)91231-Y).
- [5] A.C. Benvenuti, D. Bollini, and G. Bruni et al. “A high statistics measurement of the proton structure functions $F_2(x, Q^2)$ and R from deep inelastic muon scattering at high Q^2 ”. In: *Physics Letters B* 223.3 (1989), pp. 485–489. DOI: [https://doi.org/10.1016/0370-2693\(89\)91637-7](https://doi.org/10.1016/0370-2693(89)91637-7).
- [6] Raymond Brock et al. “Handbook of perturbative QCD: Version 1.0”. In: *Rev. Mod. Phys.* 67 (1995), pp. 157–248. DOI: 10.1103/RevModPhys.67.157.
- [7] Steve Brooks et al. *Handbook of Markov Chain Monte Carlo*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, 2011. ISBN: 9781420079425.
- [8] Andy Buckley et al. “LHAPDF6: parton density access in the LHC precision era”. In: *The European Physical Journal C* 75.3 (Mar. 2015). ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-015-3318-8.
- [9] Stefano Carrazza et al. “An unbiased Hessian representation for Monte Carlo PDFs”. In: *The European Physical Journal C* 75.8 (Aug. 2015). DOI: 10.1140/epjc/s10052-015-3590-7.
- [10] Yuri L. Dokshitzer. “Calculation of the Structure Functions for Deep Inelastic Scattering and $e^+ e^-$ Annihilation by Perturbation Theory in Quantum Chromodynamics.” In: *Sov. Phys. JETP* 46 (1977), pp. 641–653.
- [11] Sayipjamal Dulat et al. “New parton distribution functions from a global analysis of quantum chromodynamics”. In: *Physical Review D* 93.3 (Feb. 2016). DOI: 10.1103/physrevd.93.033006.
- [12] Pit Duwentäster. “Parton Distribution Functions from Deep Inelastic Scattering”. MA thesis. University of Münster, 2019.
- [13] Yémalin Gabin Gbedo and Mariane Mangin-Brinet. “Markov chain Monte Carlo techniques applied to parton distribution functions determination: Proof of concept”. In: *Physical Review D* 96.1 (July 2017). DOI: 10.1103/physrevd.96.014015.

- [14] Charles J. Geyer. “Practical Markov Chain Monte Carlo”. In: *Statistical Science* 7.4 (Nov. 1992), pp. 473–483. DOI: 10.1214/ss/1177011137.
- [15] V.N. Gribov and L.N. Lipatov. “Deep inelastic electron scattering in perturbation theory”. In: *Physics Letters B* 37.1 (1971), pp. 78–80. DOI: [https://doi.org/10.1016/0370-2693\(71\)90576-4](https://doi.org/10.1016/0370-2693(71)90576-4).
- [16] David J. Gross and Frank Wilczek. “Ultraviolet Behavior of Non-Abelian Gauge Theories”. In: *Physical Review Letters* 30.26 (June 1973), pp. 1343–1346. DOI: 10.1103/PhysRevLett.30.1343.
- [17] Particle Data Group and P A et al. Zyla. “Review of Particle Physics”. In: *Progress of Theoretical and Experimental Physics* 2020.8 (Aug. 2020). DOI: 10.1093/ptep/ptaa104.
- [18] Heikki Haario, Eero Saksman, and Johanna Tamminen. “An adaptive Metropolis algorithm”. In: *Bernoulli* 7.2 (Apr. 2001), pp. 223–242.
- [19] W. K. Hastings. “Monte Carlo Sampling Methods Using Markov Chains and Their Applications”. In: *Biometrika* 57.1 (1970), pp. 97–109.
- [20] David W. Hogg, Jo Bovy, and Dustin Lang. *Data analysis recipes: Fitting a model to data*. 2010.
- [21] David W. Hogg and Daniel Foreman-Mackey. “Data Analysis Recipes: Using Markov Chain Monte Carlo”. In: *The Astrophysical Journal Supplement Series* 236.1 (2018). DOI: 10.3847/1538-4365/aab76e.
- [22] K. Kovarik et al. “nCTEQ15: Global analysis of nuclear parton distributions with uncertainties in the CTEQ framework”. In: *Physical Review D* 93.8 (Apr. 2016). DOI: 10.1103/physrevd.93.085037.
- [23] Wolfgang von der Linden, Volker Dose, and U. Toussaint. *Bayesian probability theory, applications in physical sciences*. Cambridge University Press, 2014. DOI: 10.1017/CB09781139565608.
- [24] Neal Madras and Alan D. Sokal. “The pivot algorithm: A highly efficient Monte Carlo method for the self-avoiding walk”. In: *Journal of Statistical Physics* 50.1-2 (Jan. 1988), pp. 109–186. DOI: 10.1007/BF01022990.
- [25] A. D. Martin. “Proton Structure, Partons, QCD, DGLAP and Beyond”. In: *Acta Physica Polonica B* 39.9 (Sept. 2008).
- [26] Nicholas Metropolis et al. “Equation of State Calculations by Fast Computing Machines”. In: *Journal of Chemical Physics* 21.6 (June 1953), pp. 1087–1092. DOI: 10.1063/1.1699114.
- [27] S. Moch, J.A.M. Vermaseren, and A. Vogt. “The three-loop splitting functions in QCD: the non-singlet case”. In: *Nuclear Physics B* 688.1 (2004), pp. 101–134. DOI: <https://doi.org/10.1016/j.nuclphysb.2004.03.030>.
- [28] S. Moch, J.A.M. Vermaseren, and A. Vogt. “The three-loop splitting functions in QCD: the singlet case”. In: *Nuclear Physics B* 691.1 (2004), pp. 129–181. DOI: <https://doi.org/10.1016/j.nuclphysb.2004.04.024>.

- [29] Radford M. Neal. *Probabilistic Inference Using Markov Chain Monte Carlo Methods*. Tech. rep. 1998, p. 140.
- [30] Michael E. Peskin and Daniel V. Schroeder. *An Introduction to Quantum Field Theory*. Westview Press, 2015. ISBN: 9780813350196.
- [31] H. David Politzer. “Reliable Perturbative Results for Strong Interactions?” In: *Physical Review Letters* 30.26 (June 1973), pp. 1346–1349. DOI: 10.1103/PhysRevLett.30.1346.
- [32] J. Pumplin et al. “Uncertainties of predictions from parton distribution functions. II. The Hessian method”. In: *Physical Review D* 65.1 (Dec. 2001). DOI: 10.1103/physrevd.65.014013.
- [33] Gareth O. Roberts, A. Gelman, and W. R. Gilks. “Weak convergence and optimal scaling of random walk Metropolis algorithms”. In: *Ann. Appl. Probab.* 7.1 (Feb. 1997), pp. 110–120. DOI: 10.1214/aoap/1034625254.
- [34] Gareth O. Roberts and Jeffrey S. Rosenthal. “Coupling and Ergodicity of Adaptive Markov Chain Monte Carlo Algorithms”. In: *Journal of Applied Probability* 44.2 (2007), pp. 458–475. DOI: 10.1239/jap/1183667414.
- [35] Gareth O. Roberts and Jeffrey S. Rosenthal. “General state space Markov chains and MCMC algorithms”. In: *Probab. Surveys* 1 (2004), pp. 20–71. DOI: 10.1214/1549578041000000024.
- [36] Gareth O. Roberts and Jeffrey S. Rosenthal. “Optimal scaling for various Metropolis-Hastings algorithms”. In: *Statistical Science* 16.4 (Nov. 2001), pp. 351–367. DOI: 10.1214/ss/1015346320.
- [37] Gareth O. Roberts and R. L. Tweedie. “Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms”. In: *Biometrika* 83.1 (Mar. 1996), pp. 95–110. DOI: 10.1093/biomet/83.1.95.
- [38] Stefan Schaefer, Rainer Sommer, and Francesco Virotta. “Critical slowing down and error analysis in lattice QCD simulations”. In: *Nuclear Physics B* 845.1 (Apr. 2011), pp. 93–119. DOI: 10.1016/j.nuclphysb.2010.11.020.
- [39] Matthew D. Schwartz. *Quantum Field Theory and the Standard Model*. Cambridge University Press, 2013. ISBN: 9781107034730.
- [40] David W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley and Sons, Inc., 1992. DOI: 10.1002/9780470316849.
- [41] Robert Shumway and David Stoffer. *Time Series Analysis and Its Applications*. 2nd ed. Springer-Verlag New York, 2006. ISBN: 978-0-387-36276-2.
- [42] D.S. Sivia. *Data Analysis: A Bayesian Tutorial*. Clarendon Press, 1996. ISBN: 9780198518891.
- [43] G. Watt and R. S. Thorne. “Study of Monte Carlo approach to experimental uncertainty propagation with MSTW 2008 PDFs”. In: *Journal of High Energy Physics* 2012.8 (Aug. 2012). DOI: 10.1007/jhep08(2012)052.

- [44] Ulli Wolff. “Monte Carlo errors with less errors”. In: *Computer Physics Communications* 156.2 (2004), pp. 143–153. DOI: 10.1016/S0010-4655(03)00467-3.

Declaration of Academic Integrity

I hereby confirm that this thesis on _____
_____ is solely my own work and that I have used no sources or aids other than the ones stated. All passages in my thesis for which other sources, including electronic media, have been used, be it direct quotes or content references, have been acknowledged as such and the sources cited.

(date and signature of student)

I agree to have my thesis checked in order to rule out potential similarities with other works and to have my thesis stored in a database for this purpose.

(date and signature of student)