

GAUSSIAN PROCESSES FOR SUPERSYMMETRY

by

Ingrid Holm

THESIS

for the degree of

MASTER OF SCIENCE



Faculty of Mathematics and Natural Sciences
University of Oslo

May 2018

Abstract

This is an abstract text.

To someone

This is a dedication to my cat.

Acknowledgements

I acknowledge my acknowledgements.

Contents

1	Introduction	1
2	Physics Background	3
2.1	The Standard Model	3
2.2	The Hierarchy Problem	4
2.3	Supersymmetry	6
2.3.1	Superfields	8
2.3.2	Supersymmetric Lagrangian	9
2.3.3	Soft Supersymmetry Breaking	11
2.4	The Minimal Supersymmetric Standard Model	12
2.4.1	Field Content	12
2.4.2	MSSM Lagrangian	14
2.4.3	R-parity	15
2.4.4	Soft Breaking terms	16
2.4.5	Radiative Electroweak Symmetry Breaking	16
2.4.6	Sparticles	18
2.4.7	MSSM-24	20
2.4.8	Constrained MSSM	20
3	Supersymmetry at Hadron Colliders	23
3.1	Hadron Colliders	23
3.1.1	Parton Distribution Functions	24
3.1.2	Luminosity	25
3.2	Phenomenology	25
3.2.1	Searches For Supersymmetry	25
3.2.2	Current Bounds on Sparticles	26
3.3	Squark-Squark Cross Section	29
3.3.1	Leading Order Cross Section	29
3.4	Next-to-leading Order Corrections	34
3.4.1	State-of-the-art Tools	37
4	Gaussian Processes	41
4.1	Introduction to Bayesian Statistics	41

Contents

4.1.1	Bayes' Theorem	42
4.1.2	Priors and Likelihood	43
4.1.3	Best Estimate and Reliability	43
4.1.4	Covariance	46
4.2	Covariance Functions	48
4.2.1	The Squared Exponential Covariance Function	49
4.2.2	The Matérn Class of Covariance Functions	50
4.2.3	Noise	50
4.2.4	Hyperparameters	50
4.3	Gaussian Process Regression	51
4.3.1	The Prior and Posterior Distribution	54
4.4	Model Selection	59
4.4.1	Log Marginal Likelihood	59
4.4.2	Cross Validation	60
4.4.3	Relative Deviance	62
5	Evaluating Cross Sections using Gaussian Processes	65
5.1	Data Generation	65
5.2	Dataset Transformations	71
5.3	Learning the Gaussian Process	72
5.3.1	The Benchmark	74
5.3.2	Outliers	76
5.3.3	Cuts on Cross Sections	76
5.3.4	Features	77
5.3.5	Kernel	77
5.3.6	Cummulative Settings	79
5.4	Distributed Gaussian Processes	81
5.4.1	Product-of-Experts	81
5.4.2	Bayesian Committee Machine	83
5.4.3	Robust Bayesian Comittee Machine	83
5.4.4	Evaluating Cross Sections using Distributed Gaussian Processes	86
5.4.5	Cross Validation	86
6	Results	89
6.1	Learning Curves	89
6.2	Comparison with Prospino and NLL-fast	91
6.2.1	Relative Deviance	91
6.2.2	Cross Sections	94
6.3	Optimizing the Model	95
Appendices		103

Contents

A Notation	105
A.1 Quantum Field Theory	105
A.2 Statistics	106
B Data Quality	107
C Benchmark for Distributed Gaussian Processes	109
D Relative Deviance Distributions	111

Chapter 1

Introduction

- why nlo?
 - why gp?
 - why dgp?

Chapter 2

Physics Background

In this chapter supersymmetry and some of the motivations for an extension of the Standard Model are introduced, assuming familiarity with quantum field theory, the Standard Model of particle physics and some group theory. The Higgs mechanism and the hierarchy problem are reviewed, before supersymmetry is outlined. The Minimal Supersymmetric Standard Model is introduced, with its corresponding field content. Finally, the versions of the Minimal Supersymmetric Standard Model used in this project are briefly outlined.

2.1 The Standard Model

The *Standard Model of particle physics* (SM) has successfully explained almost all experimental results in particle physics and predicted several phenomena. One of the main attributes of the Standard Model is that particles with different values of the *spin* quantum number behave differently. Particles with half-integer and integer spin values are called *fermions* and *bosons*, respectively. Fermions are particles such as quarks and leptons, which interact through the exchange of force carrying bosons. The Standard Model bosons are the *photon* (electromagnetic interaction), the *gluon* (strong interaction that holds atoms together), the W and Z bosons (the weak interaction) and the famously elusive *Higgs boson*, that provides masses for the Standard Model particles. The equations of motion and allowed interactions can all be derived from the *Lagrangian* of the Standard Model. The Lagrangian, of which the time integral is the action S , is invariant to transformations under the Lorentz group — known as a change of reference frame.

The Higgs Mechanism

The Standard Model is a gauge theory based on the symmetry group $SU(3)_C \times SU(2)_L \times U(1)_Y$. The $SU(3)_C$ group is the symmetry group for strong inter-

actions, or quantum chromodynamics, and $SU(2)_L \times U(1)_Y$ is the electroweak symmetry group. In order for the particles to obtain masses the electroweak symmetry must be spontaneously broken down to $U(1)_{\text{em}}$. The symmetry is broken when the Higgs field obtains a non-zero *vacuum expectation value* (vev) — meaning that it has some field value when the governing potential is at its minimum. The Higgs field Φ is a self-interacting complex $SU(2)_L$ doublet whose Lagrangian is given by

$$\mathcal{L}_\Phi = \partial_\mu \Phi^\dagger \partial^\mu \Phi + V(\Phi), \quad (2.1)$$

where the first term is the kinetic term, and the scalar potential describing the Higgs $V(\Phi)$ is the famous Mexican hat potential

$$V(\Phi) = \mu^2 \Phi^\dagger \Phi + \lambda (\Phi^\dagger \Phi)^2. \quad (2.2)$$

For $\mu^2 < 0$ and $\lambda > 0$ this potential acquires a non-trivial minimum given by

$$|\Phi_0| = \sqrt{\frac{-\mu^2}{2\lambda}} \equiv \frac{v}{\sqrt{2}}, \quad (2.3)$$

where v is the vacuum expectation value. A special parameterization of the Higgs $SU(2)_L$ doublet, $\Phi^T(x) = \frac{1}{\sqrt{2}}(0, v + h(x))$, leads to the Lagrangian developing mass terms for fermions and the gauge bosons Z and W^\pm . The mass terms are proportional to v , *e.g.*

$$M_W = \frac{1}{2}vg,$$

where M_W is the mass of the W and g is the $SU(2)_L$ -coupling. The Higgs' own mass provides one of the strongest arguments for introducing supersymmetry, namely the *hierarchy problem*, which is discussed in the following section.

Another argument for the introduction of supersymmetry is gauge coupling unification. Gauge coupling unification is the assumption that the Standard Model symmetry group is a unified gauge group, *e.g.* $SU(5)$ or $SO(10)$, broken down to $SU(3)_C \times SU(2)_L \times U(1)_Y$ at some high energy scale. However, gauge unification is not discussed in this project.

2.2 The Hierarchy Problem

The Higgs boson was discovered at the Large Hadron Collider (LHC) in 2012, and its mass was measured at $m_H \sim 125$ GeV [?]. The expression for the Higgs mass in the Standard Model includes loop corrections, which provide a large discrepancy between theory and experiment. The Higgs mass receives fermionic or scalar loop-contributions to its mass, such as those shown in Fig. 2.1. The

expression for the mass can then be written in terms of the bare parameter m_{H0} and the corrections Δm_H

$$m_H^2 = m_{H0}^2 + \Delta m_H^2.$$

Loop diagrams contain divergences, because of integrals over all possible momenta for the virtual particles in the loops. A way to get rid of these divergences is to regularize the expressions. Regularization is a neat trick that introduces a *cut-off scale*, which sets an upper limit on the momentum that is integrated over. A common choice for the cut-off scale Λ is the Planck scale, as this is where new physics is needed to explain gravity in the Standard Model. The Planck scale is of the order of $\Lambda_P \sim 10^{18}$ GeV. After regularization, the mass correction terms are

$$\Delta m_H^2 = -\frac{|\lambda_f|^2}{8\pi^2} \Lambda_P^2 + \frac{\lambda_s}{16\pi^2} \Lambda_P^2 + \dots, \quad (2.4)$$

where λ_s is the coupling of the Higgs to the scalar, and λ_f is the Higgs coupling to the fermion. The problem now becomes apparent: the correction to the mass is proportional to the Planck scale, placing it at the order of 10^{18} GeV, yet the mass has been experimentally measured around 125 GeV. There must be some colossal cancellation of terms with a tremendous tuning of the SM parameters in λ_s and λ_f^2 . Tuning of parameters is undesirable — the model should be as natural as possible.

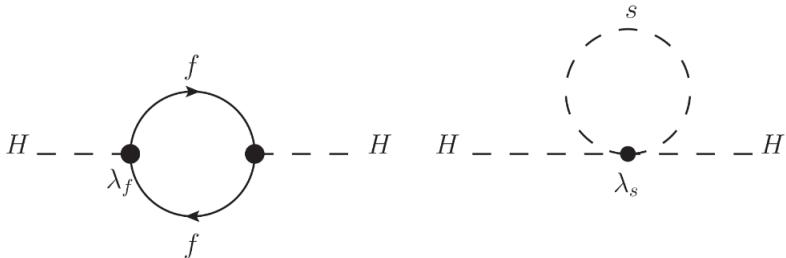


Figure 2.1: Fermion and scalar one-loop corrections to the Higgs mass. Figure from [1].

Supersymmetry provides an elegant solution to the hierarchy problem. In simple terms, supersymmetry introduces fermionic superpartners for the bosons, and vice versa. These are called sparticles, and in unbroken supersymmetry the particles and sparticles have identical mass, and their couplings to the Higgs are the same $\lambda_s = |\lambda_f|^2$. In addition, there are twice as many scalars as fermions, which gives a perfect cancellation of these enormous corrections. Unbroken supersymmetry therefore solves the hierarchy problem. As will be discussed, supersymmetry must in fact be a broken symmetry, and broken supersymmetry is revisited in Sec. 2.3.3.

2.3 Supersymmetry

Supersymmetry is an extension of symmetries in relativistic quantum field theory. Relativistic field theories are invariant under boosts, rotations and translations in spacetime, called Poincaré transformations. A Poincaré transformation of the position four-vector of a particle, x^μ , is given as

$$x^\mu \rightarrow x'^\mu = \Lambda^\mu_\nu x^\nu + a^\mu, \quad (2.5)$$

where Λ^μ_ν is a Lorentz transformation and a^μ is a translation. The assumption behind supersymmetry is that Nature obeys a non-trivial extension of the related Poincaré algebra, namely the *superalgebra*.

Superalgebra

The super-Poincaré algebra is given by the following commutation and anticommutation relations [2]

$$\{Q_A, Q_B\} = \{\bar{Q}_A, \bar{Q}_B\} = 0, \quad (2.6)$$

$$\{Q_A, \bar{Q}_{\dot{A}}\} = 2(\sigma^\mu)_{A\dot{A}} P_\mu, \quad (2.7)$$

$$[Q_A, P^\mu] = [\bar{Q}_A, P^\mu] = 0, \quad (2.8)$$

$$[Q_A, M^{\mu\nu}] = \frac{1}{2}(\sigma^{\mu\nu})_A^B Q_B, \quad (2.9)$$

$$[\bar{Q}_{\dot{A}}, M^{\mu\nu}] = \frac{1}{2}(\bar{\sigma}^{\mu\nu})_{\dot{a}}^{\dot{b}} Q_{\dot{b}}^\dagger, \quad (2.10)$$

where Q_A and $\bar{Q}_{\dot{A}}$ are the superalgebra generators, $A = 1, 2$ and $\dot{A} = 1, 2$ are the indices of two Weyl spinors, P^μ are the generators of translation, and $M^{\mu\nu}$ are the generators of the Lorentz group.

The supersymmetry generators turn fermions into bosons and vice versa. More specifically, these operators have the following commutation relations with the rotation generator J^3

$$[Q_A, J^3] = \frac{1}{2}(\sigma^3)_A^B Q_B, \quad (2.11)$$

which for the Q_1 generator becomes

$$[Q_1, J^3] = \frac{1}{2}Q_1. \quad (2.12)$$

Using this operator on a state in an irreducible representation of the Poincaré algebra with mass m and spin j_3 gives

$$J^3 Q_1 |m, j_3\rangle = (j_3 - \frac{1}{2}) Q_1 |m, j_3\rangle, \quad (2.13)$$

thus lowering the spin of the state by $1/2$. Similarly, Q_2 would increase the spin. They do not, however, change the mass. This can be seen from Eq. (2.8)

$$P^\mu P_\mu Q_A |m, j_3\rangle = Q_A P^\mu P_\mu |m, j_3\rangle = m^2 Q_A |m, j_3\rangle. \quad (2.14)$$

Superpartners

States that transform into each other via Q_A and $\bar{Q}_{\dot{A}}$ are called *superpartners*. From Eq. (2.14) it can be seen that, in unbroken supersymmetry, the partnering fermions and bosons have the same mass. Superpartners fall into irreducible representations of the superalgebra called *supermultiplets*. The superpartners are the fermionic and bosonic states of the supermultiplets. The supersymmetry generators commute with the gauge generators as well, so superpartners that are part of the same supermultiplet must also have the same electric charges, weak isospin and color degrees of freedom [3]. The number of bosonic and fermionic degrees of freedom in each supermultiplet is equal,

$$n_B = n_F. \quad (2.15)$$

The simplest possibility for a supermultiplet that obeys Eq. (2.15) has a single Weyl fermion, with $n_F = 2$, and two real scalars, each with $n_B = 1$, assembled to a complex scalar field. This is called a *chiral multiplet*. Chiral multiplets are the only supermultiplets that can contain fermions whose left-handed partners transform differently under the gauge-group than their right-handed partners [3]. Since this is the case for the SM fermions, these must be members of chiral supermultiplets. The superpartners of the quarks and leptons must therefore be spin-0 bosons. The scalar partners of the fermions are denoted by the prefix ‘s’ for scalar, such as the *squarks* and the *sleptons*. The left- and right-handed pieces of the quarks and leptons are separate two-component Weyl spinors, so each has its own complex scalar partner. For example, the up-type quark u has two scalar partners, \tilde{u}_R and \tilde{u}_L , where superpartners are denoted by a tilde \sim .

The next-to-simplest supermultiplet has a spin-1 vector boson, with $n_B = 2$, and a massless spin-1/2 Weyl spinor, also with $n_F = 2$. The vector bosons are the gauge bosons, and their fermionic superpartners are called *gauginos*. These multiplets are called *vector*, or *gauge* supermultiplets.

Superspace

The elements of the superalgebra and their representations can be described using *superspace*. Coordinates in superspace are given by $z^\pi = (x^\mu, \theta^A, \bar{\theta}_{\dot{A}})$, where x^μ are the well-known Minkowski coordinates, and $\theta^A, \bar{\theta}_{\dot{A}}$ are four *Grassmann numbers* contained in Weyl spinors with indices A and \dot{A} . Grassmann numbers are numbers that *anti-commute*, and for $\theta_A, \bar{\theta}^{\dot{B}}$ the following is therefore true

$$\{\theta^A, \theta^B\} = \{\theta^A, \bar{\theta}^{\dot{B}}\} = \{\bar{\theta}^{\dot{A}}, \theta^B\} = \{\bar{\theta}^{\dot{A}}, \bar{\theta}^{\dot{B}}\} = 0, \quad (2.16)$$

which gives the relationships

$$\theta_A^2 \equiv \theta_A \theta_A = -\theta_A \theta_A = 0, \quad (2.17)$$

$$\theta^2 \equiv \theta \theta = \theta_A \theta^A = -2\theta_1 \theta_2, \quad (2.18)$$

$$\bar{\theta}^2 \equiv \bar{\theta} \bar{\theta} = \bar{\theta}_A \bar{\theta}^A = -2\bar{\theta}_1 \bar{\theta}_2. \quad (2.19)$$

Any power series of functions of Grassmann numbers therefore terminates as a function of θ_1

$$f(\theta_1) = a + b\theta_1, \quad \frac{df}{d\theta_1} = a. \quad (2.20)$$

Integrals over Grassmann numbers are

$$\int \theta \theta \, d^2\theta = 1, \quad \int \bar{\theta} \bar{\theta} \, d^2\bar{\theta} = 1, \quad \int (\theta \theta)(\bar{\theta} \bar{\theta}) = 1 \, d^4\theta, \quad (2.21)$$

where $d^2\theta = -1/4d\theta^A d\theta_A$, $d^2\bar{\theta} = -1/4d\bar{\theta}^A d\bar{\theta}_A$ and $d^4\theta = d^2\theta d^2\bar{\theta}$.

2.3.1 Superfields

In superspace, any *superfield* Φ — a function on superspace $\Phi(x^\mu, \theta^A, \bar{\theta}_{\dot{A}})$ — can be expanded as a power series in the anti-commuting variables, with components that are functions of the four-vector x^μ . A superfield can therefore be written as

$$\begin{aligned} \Phi(x, \theta, \bar{\theta}) = & a(x) + \theta \xi(x) + \theta^\dagger \chi^\dagger(x) + \theta \theta b(x) + \theta^\dagger \theta^\dagger c(x) \\ & + \theta^\dagger \bar{\sigma}^\mu \theta v_\mu(x) + \theta^\dagger \theta^\dagger \theta \eta(x) + \theta \theta \theta^\dagger \zeta(x) + \theta \theta \theta^\dagger \theta^\dagger d(x), \end{aligned} \quad (2.22)$$

where $\bar{\sigma}^\mu$ is defined in Appendix A, and $a(x)$, $b(x)$, $c(x)$, $v_\mu(x)$ and $d(x)$ are bosonic fields, and $\xi(x)$, $\chi(x)$, $\eta(x)$ and $\zeta(x)$ are two-component fermionic fields.

The general covariant derivatives working on superfields, that are invariant under supersymmetry transformations, are defined as

$$D_A \equiv \partial_A + i(\sigma^\mu \bar{\theta})_A \partial_\mu, \quad (2.23)$$

$$\bar{D}^{\dot{A}} \equiv -\partial^{\dot{A}} - i(\sigma^\mu \theta)^{\dot{A}} \partial_\mu. \quad (2.24)$$

The covariant derivatives work on the aforementioned superfields, Φ . A chiral superfield must obey the following constraints

$$\bar{D}_{\dot{A}} \Phi(x, \theta, \bar{\theta}) = 0 \quad (\text{left-handed chiral superfield}), \quad (2.25)$$

$$D^A \Phi^\dagger(x, \theta, \bar{\theta}) = 0 \quad (\text{right-handed chiral superfield}). \quad (2.26)$$

The fields Φ are required to be Lorentz scalars or pseudoscalars, which restricts the properties of their component fields. Using the general form of a superfield,

Eq. (2.22), and the constraints from the covariant derivatives, it can be shown that the left- and right-handed chiral fields can be written in terms of their component fields as [1]

$$\begin{aligned}\Phi(x, \theta, \bar{\theta}) = & A(x) + i(\theta\sigma^\mu\bar{\theta})\partial_\mu A(x) - \frac{1}{4}\theta\theta\bar{\theta}\bar{\theta}\square A(x) + \sqrt{2}\theta\psi(x) \\ & - \frac{i}{\sqrt{2}}\theta\theta\partial_\mu\psi(x)\sigma^\mu\bar{\theta} + \theta\theta F(x),\end{aligned}\quad (2.27)$$

$$\begin{aligned}\Phi^\dagger(x, \theta, \bar{\theta}) = & A^*(x) - i(\theta\sigma^\mu\bar{\theta})\partial_\mu A^*(x) - \frac{1}{4}\theta\theta\bar{\theta}\bar{\theta}\square A^*(x) + \sqrt{2}\bar{\theta}\bar{\psi}(x) \\ & - \frac{i}{\sqrt{2}}\bar{\theta}\bar{\theta}\theta\sigma^\mu\partial_\mu\bar{\psi}(x) + \bar{\theta}\bar{\theta}F^*(x),\end{aligned}\quad (2.28)$$

where $A(x)$ and $F(x)$ are complex scalars, $A^*(x)$ and $F^*(x)$ are their complex conjugates, and $\psi_A(x)$ and $\bar{\psi}^A(x)$ are left-handed and right-handed Weyl spinors, respectively.

A vector superfield $V(x, \theta, \bar{\theta})$ is defined by the constraint

$$V(x, \theta, \bar{\theta}) = V^\dagger(x, \theta, \bar{\theta}). \quad (2.29)$$

From Eq. (2.29) the structure of a general vector field in terms of component fields is [3]

$$\begin{aligned}V(x, \theta, \bar{\theta}) = & f(x) + \theta\varphi(x) + \bar{\theta}\bar{\varphi}(x) + \theta\theta m(x) + \bar{\theta}\bar{\theta}m^*(x) \\ & + \theta\sigma^\mu\bar{\theta}V_\mu(x) + \theta\theta\bar{\theta}\bar{\lambda}(x) + \bar{\theta}\bar{\theta}\theta\lambda(x) + \theta\theta\bar{\theta}\bar{\theta}d(x),\end{aligned}\quad (2.30)$$

where $f(x)$, $d(x)$ are real scalar fields, $\varphi_A(x)$, $\lambda_A(x)$ are Weyl spinors, $m(x)$ is a complex scalar field and $V_\mu(x)$ is a real Lorentz four-vector. Given a chiral superfield Φ , the combinations $\Phi + \Phi^\dagger$, $i(\Phi - \Phi^\dagger)$ and $\Phi^\dagger\Phi$ are all real, vector superfields. In the $j = \frac{1}{2}$ representation of the superalgebra, the superfield $V = \Phi^\dagger\Phi$ does not correspond to the promised number of degrees of freedom. This problem is fixed by the *super-gauge* introduced in Sec. 2.3.2.

2.3.2 Supersymmetric Lagrangian

Symmetry transformations of the Lagrangian should leave the action

$$S \equiv \int d^4x \mathcal{L}, \quad (2.31)$$

invariant. This is automatically fulfilled if the Lagrangian only changes by a total derivative. It can be shown that the highest order component fields in θ and $\bar{\theta}$ of the scalar and vector superfields have this property. To ensure that the action is invariant under supersymmetry transformations, the Lagrangian is redefined

such that

$$S = \int d^4x \int d^4\theta \mathcal{L}, \quad (2.32)$$

where there is now an integral over Grassmann numbers and $d^4\theta = d^2\theta d^2\bar{\theta}$.

Restrictions on the supersymmetric Lagrangian, such as invariance under supersymmetric transformations and renormalizability, mean that the most general Lagrangian as a function of the scalar superfields Φ_i is

$$\mathcal{L} = \Phi_i^\dagger \Phi_i + \bar{\theta}\bar{\theta}W[\Phi] + \theta\theta W[\Phi^\dagger], \quad (2.33)$$

where $\Phi_i^\dagger \Phi_i$ is the kinetic term, and $W[\Phi]$ is the *superpotential*

$$W[\Phi] = g_i \Phi_i + m_{ij} \Phi_i \Phi_j + \lambda_{ijk} \Phi_i \Phi_j \Phi_k, \quad (2.34)$$

where m_{ij} and λ_{ijk} are symmetric. Specifying a supersymmetric Lagrangian then only requires specifying the superpotential.

Supergauge

The Lagrangian should also be gauge invariant. Consider a general, Abelian or non-Abelian, group G with the Lie algebra of group generators t_a that fulfill

$$[t_a, t_b] = i f_{ab}{}^c t_c, \quad (2.35)$$

where f_{ab}^c are the structure constants. An element g in the group G can be written down in the unitary representation

$$U(g) = e^{i\Lambda_a t^a}. \quad (2.36)$$

The supergauge transformation (global or local) on left-handed chiral superfields Φ_i is then defined as [1]

$$\Phi \rightarrow \Phi' = e^{-i\Lambda_a T^a q} \Phi, \quad (2.37)$$

where q is the charge of the superfield Φ under G , Λ_a are the parameters of the transformation, and T^a are the generators of the gauge group in a chosen representation. For a left-handed superfield Φ_i the Λ^a must also be left-handed superfields, and correspondingly a right-handed superfield Φ^\dagger must have right-handed superfields $\Lambda^{\dagger a}$.

For the Lagrangian to be gauge invariant the potential W must be gauge invariant as well. From the requirement that $W[\Phi] = W[\Phi']$, some restrictions on the superpotential follow

$$g_i = 0 \text{ if } g_i U_{ir} \neq g_r, \quad (2.38)$$

$$m_{ij} = 0 \text{ if } m_{ij} U_{ir} U_{js} \neq m_{rs}, \quad (2.39)$$

$$\lambda_{ijk} = 0 \text{ if } \lambda_{ijk} U_{ir} U_{js} U_{kt} \neq \lambda_{rst}, \quad (2.40)$$

where the indices on U are matrix indices.

The kinetic term must also be invariant under gauge transformations. For this term to be invariant, a gauge compensating vector superfield V^a for each Lie algebra generator T_a with the appropriate gauge transformation is introduced. The kinetic term can then be written as $\Phi^\dagger e^{qV^a T_a} \Phi$, and it transforms as

$$\Phi^\dagger e^{qV^a T_a} \Phi \rightarrow \Phi'^\dagger e^{qV'^a T_a} \Phi' = \Phi^\dagger e^{iq\Lambda^{a\dagger} T_a} e^{qV'^a T_a} e^{-iq\Lambda^a T_a} \Phi, \quad (2.41)$$

meaning that the vector superfield V^a must transform as

$$e^{qV'^a T_a} = e^{-iq\Lambda^{a\dagger} T_a} e^{qV^a T_a} e^{iq\Lambda^a T_a}. \quad (2.42)$$

Supersymmetric Field Strength

The supersymmetric Lagrangian also requires field strengths, analogous to the electromagnetic field strength $F_{\mu\nu}$. The supersymmetric field strengths are

$$W_A \equiv -\frac{1}{4} \bar{D}\bar{D} e^{-V} D_A e^V, \quad (2.43)$$

$$\bar{W}_{\dot{A}} \equiv -\frac{1}{4} D D e^{-V} \bar{D}_{\dot{A}} e^V, \quad (2.44)$$

where $V = V^a T_a$. W_A ($\bar{W}_{\dot{A}}$) is a left-handed (right-handed) superfield, and it can be shown that the trace $\text{Tr}[W_A W^A]$ is supergauge invariant [1].

The Lagrangian for a supersymmetric theory with (possibly) non-Abelian gauge groups is then

$$\mathcal{L} = \Phi^\dagger e^V \Phi + \delta^2(\bar{\theta}) W[\Phi] + \delta^2(\theta) W[\Phi^\dagger] + \frac{1}{2T(R)} \delta^2(\bar{\theta}) \text{Tr}[W_A W^A], \quad (2.45)$$

where $T(R)$ is the Dynkin index for normalization, $\delta^2(\bar{\theta}) = \bar{\theta}\bar{\theta}$ and $\delta^2(\theta) = \theta\theta$. The Dynkin index of the representation R in terms of matrices T_a is given by $\text{Tr}[T_a, T_b] = T(R)\delta_{ab}$.

2.3.3 Soft Supersymmetry Breaking

As previously mentioned, particles and their corresponding sparticles have identical masses in unbroken supersymmetry. Since sparticles have not yet been observed, supersymmetry must be a broken symmetry. In this section soft supersymmetry breaking is considered as a way of providing masses to particles, without compromising the solution to the hierarchy problem.

The Standard Model particles obtain mass through spontaneous symmetry breaking of the electroweak symmetry, as described in Sec. 2.1. Supersymmetry cannot be broken in the same way, as it would require that the sum of scalar particles squared be equal to the sum of fermion masses squared. Consequently,

not all scalar partners could be heavier than the known particles, which is known not to be the case [1].

Instead, supersymmetry can be spontaneously broken via *soft breaking*. Soft breaking entails adding terms to the Lagrangian that break supersymmetry explicitly, while preserving the cancellations of divergences that fixes the hierarchy problem. These are called *soft terms*, and there are several restrictions on them. That a term is *soft* means that the coupling should have mass dimension one or higher, to avoid divergences from loop contributions to scalar masses. The possible soft terms can be written in terms of their component fields

$$\begin{aligned} \mathcal{L}_{\text{soft}} = & -\frac{1}{2}M\lambda^A\lambda_A - \left(\frac{1}{6}a_{ijk}A_iA_jA_k + \frac{1}{2}b_{ij}A_iA_j + t_iA_i + \text{c.c.} \right) \\ & - m_{ij}^2A_i^*A_j, \end{aligned} \quad (2.46)$$

where λ_A are Weyl spinor fields and A_i are chiral fields. The couplings consist of gaugino masses M for each gauge group, scalar squared-mass terms m_{ij}^2 and b_{ij} , scalar couplings a_{ijk} and tadpole terms t_i . The soft breaking terms thereby give masses to both the scalar and fermionic superpartners of the SM particles.

Restrictions on the new parameters are necessary to avoid reintroducing the hierarchy problem. If the breaking terms are soft, the correcting mass terms are at most

$$\Delta m_h^2 = -\frac{\lambda_s}{16\pi^2}m_s^2 \ln \frac{\Lambda_{UV}}{m_s^2} + \dots,$$

at leading order in the breaking scale Λ_{UV} , where m_s is the soft breaking scale. In this scheme m_s is restricted to $m_s \sim \mathcal{O}(1 \text{ TeV})$.

2.4 The Minimal Supersymmetric Standard Model

The *Minimal Supersymmetric Standard Model* (MSSM) is minimal in the sense that it requires the least amount of new fields introduced in order to have all the SM fields and supersymmetry. The MSSM is based on the minimal extension of the Poincaré algebra. In this section the field content of the MSSM and the introduction of R -parity is discussed, before the MSSM-24 and CMSSM and their parameters are introduced.

2.4.1 Field Content

As discussed in Sec. 2.3, the SM fermions and the Higgs boson are contained in chiral supermultiplets, and the SM gauge bosons are contained in vector supermultiplets. A chiral supermultiplet contains one Weyl spinor with two fermionic degrees of freedom, and two real scalars, with one bosonic degree of freedom each.

To form a Dirac fermion, both a left-handed and a right-handed Weyl spinor are needed. These are obtained from a left-handed chiral supermultiplet, and a *different* right-handed chiral supermultiplet. The four fermionic degrees of freedom become two fermions — a particle and an antiparticle — and the four scalar degrees of freedom become four scalar particles, a pair of left- and right-handed scalars, and their antiparticles.

Leptons

For leptons the left-handed chiral superfields are contained in the $SU(2)_L$ doublets L_i , and the right-handed chiral superfields are in $SU(2)_L$ singlets \bar{E}_i , given by

$$L_i = \begin{pmatrix} \nu_i \\ l_i \end{pmatrix} \quad \text{and} \quad \bar{E}_i, \quad (2.47)$$

where $i = 1, 2, 3$ is the generation index. The superfields l_i and \bar{E}_i combine to give charged leptons and sleptons, and ν_i gives the (left-handed) neutrinos and sneutrinos. Note that there is no right-handed \bar{N}_i . This is a convention, as MSSM is older than the discovery of massive neutrinos.

Quarks

Similarly, for up-flavour and down-flavour quarks the left-handed chiral superfields contained in the $SU(2)_L$ doublet Q_i , and the right-handed chiral superfields \bar{U}_i and \bar{D}_i are given by

$$Q_i = \begin{pmatrix} u_i \\ d_i \end{pmatrix}, \quad \bar{U}_i \quad \text{and} \quad \bar{D}_i. \quad (2.48)$$

The superfields u_i and \bar{U}_i give the up-type SM quarks and squarks, and the superfields d_i and \bar{D}_i give the down-type SM quarks and squarks. Color indices are omitted for simplicity.

Gauge Bosons

To construct the gauge bosons, vector supermultiplets are used, as discussed in Sec. 2.3. A vector supermultiplet contains two fermionic and two bosonic degrees of freedom, in the form of a massless vector boson and one Weyl spinor of each handedness. As noted in Sec. 2.3.2, one superfield V^a is needed per generator of the algebra T_a for each of the gauge groups in $SU(3)_C$, $SU(2)_L$ and $U(1)_Y$. These vector supermultiplets are

$$C^a, \quad W^a, \quad \text{and} \quad B^0. \quad (2.49)$$

The spin-1 bosons constructed from these superfields are the SM gauge bosons g , B^0 , W^0 , W^+ and W^- . The spin-1/2 superpartners of the gluons g are the *gluinos* \tilde{g} . For the $SU(2)_L \times U(1)_Y$ gauge bosons they are \tilde{B}^0 , \tilde{W}^0 , \tilde{W}^+ and \tilde{W}^- , named the *bino* and the *winos*. After electroweak symmetry breaking, B^0 and W^0 mix to give the mass eigenstates Z^0 and γ , and the corresponding mixtures of \tilde{B}^0 and \tilde{W}^0 are called *zino* \tilde{Z}^0 and *photino* $\tilde{\gamma}$.

Higgs Boson

Finally, supermultiplets are needed for the Higgs. The Higgs is a scalar particle, so it must reside in a chiral supermultiplet, as discussed in Sec. 2.3. The supersymmetry version of the SM Higgs $SU(2)_L$ doublet would mix left- and right-handed superfields, and therefore cannot appear in the superpotential. The minimal allowed Higgs content are two $SU(2)_L$ Higgs doublets H_u and H_d , indexed according to the quarks they give mass to. The doublets are

$$H_u = \begin{pmatrix} H_u^+ \\ H_u^0 \end{pmatrix}, \quad H_d = \begin{pmatrix} H_d^0 \\ H_d^- \end{pmatrix}. \quad (2.50)$$

These left-handed chiral supermultiplets contain in total four Weyl spinors and eight bosonic degrees of freedom. Three degrees of freedom are used to give masses to the W^\pm and Z^0 bosons through the Higgs mechanism. The remaining five are manifest through the mass eigenstates h^0 , H^0 , A^0 and H^\pm . The Weyl spinors combine into the *higgsinos*. The entire field content of the MSSM is listed in Table 2.1.

2.4.2 MSSM Lagrangian

The Lagrangian for the MSSM may now be constructed from the supermultiplets, and consists of kinetic terms \mathcal{L}_{kin} , supersymmetric field strength terms \mathcal{L}_V , the superpotential terms \mathcal{L}_W and the soft breaking terms $\mathcal{L}_{\text{soft}}$,

$$\mathcal{L}_{\text{MSSM}} = \mathcal{L}_{\text{kin}} + \mathcal{L}_V + \mathcal{L}_W + \mathcal{L}_{\text{soft}}. \quad (2.51)$$

The kinetic terms are constructed from the supermultiplets introduced above

$$\begin{aligned} \mathcal{L}_{\text{kin}} = & L_i^\dagger e^{\frac{1}{2}g\sigma W - \frac{1}{2}g'B} L_i + Q_i^\dagger e^{\frac{1}{2}g_s\lambda C + \frac{1}{2}g\sigma W + \frac{1}{3}\cdot\frac{1}{2}g'B} Q_i \\ & + \bar{U}_i^\dagger e^{\frac{1}{2}g_s\lambda C - \frac{4}{3}\cdot\frac{1}{2}g'B} \bar{U}_i + \bar{D}_i^\dagger e^{\frac{1}{2}g_s\lambda C - \frac{2}{3}\cdot\frac{1}{2}g'B} \bar{D}_i \\ & + \bar{E}_i^\dagger e^{2\frac{1}{2}g'B} \bar{E}_i + H_u^\dagger e^{\frac{1}{2}g\sigma W + \frac{1}{2}g'B} H_u + H_d^\dagger e^{\frac{1}{2}g\sigma W - \frac{1}{2}g'B} H_d, \end{aligned} \quad (2.52)$$

where g' , g and g_s are the couplings of the $U(1)_Y$, $SU(2)_L$ and the $SU(3)_C$.

The supersymmetric field strength contributions with pure gauge terms are

$$\mathcal{L}_V = \frac{1}{2}\text{Tr}\{W^A W_A\} \bar{\theta} \bar{\theta} + \frac{1}{2}\text{Tr}\{C^A C_A\} \bar{\theta} \bar{\theta} + \frac{1}{4}B^A B_A \bar{\theta} \bar{\theta} + \text{c.c.}, \quad (2.53)$$

Supermultiplet	Scalars	Fermions	Vectors	$SU(3)_c$	$SU(2)_L$	$U(1)_Y$
Q_i	$(\tilde{u}_{iL}, \tilde{d}_{iL})$	(u_{iL}, d_{iL})		3	2	$\frac{1}{6}$
\bar{u}_i	\tilde{u}_{iR}^*	u_{iR}^\dagger		$\bar{3}$	1	$-\frac{2}{3}$
\bar{d}_i	\tilde{d}_{iR}^*	d_{iR}^\dagger		$\bar{3}$	1	$\frac{1}{3}$
L_i	$(\tilde{\nu}_{iL}, \tilde{e}_{iL})$	(ν_{iL}, e_{iL})		1	2	$-\frac{1}{2}$
\bar{e}_i	\tilde{e}_{iR}^*	e_{iR}^\dagger		1	1	1
H_u	(H_u^+, H_u^0)	$(\tilde{H}_u^+, \tilde{H}_u^0)$		1	2	$\frac{1}{2}$
H_d	(H_d^0, H_d^-)	$(\tilde{H}_d^0, \tilde{H}_d^-)$		1	2	$-\frac{1}{2}$
g		\tilde{g}	g	8	1	0
W		$\tilde{W}^{1,2,3}$	$W^{1,2,3}$	1	3	0
B		\tilde{B}	B	1	1	0

Table 2.1: Gauge and chiral supermultiplets in the Minimal Supersymmetric Standard Model with SM gauge group representations. The index $i = 1, 2, 3$ runs over the three generations of quarks and lepton. Table from [2].

with the field strengths W_A , C_A and B_A given by

$$W_A = -\frac{1}{4}\bar{D}\bar{D}e^{-W}D_AeW, \quad W = \frac{1}{2}g\sigma^aW^a, \quad (2.54)$$

$$C_A = -\frac{1}{4}\bar{D}\bar{D}e^{-C}D_Ae^C, \quad C = \frac{1}{2}g_s\lambda^aC^a, \quad (2.55)$$

$$B_A = -\frac{1}{4}\bar{D}\bar{D}D_AB, \quad B = \frac{1}{2}g'B^0. \quad (2.56)$$

The gauge invariant terms in the superpotential are

$$\begin{aligned} W = & \mu H_u H_d + \mu' L_i H_u + y_{ij}^e L_i H_d E_j + y_{ij}^u Q_i H_u \bar{U}_j + y_{ij}^d Q_i H_d \bar{D}_j \\ & + \lambda_{ijk} L_i L_j \bar{E}_k + \lambda'_{ijk} L_i Q_j \bar{D}_k + \lambda''_{ijk} \bar{U}_i \bar{D}_j \bar{D}_k, \end{aligned} \quad (2.57)$$

where $H_u H_d$ is shorthand for $H_u^T i\sigma_2 H_d$ — and similarly for the other doublet pairs — which is a construction invariant under $SU(2)_L$. The parameter μ is the Lagrangian mass parameter and μ' is some other mass parameter in the superpotential.

2.4.3 R-parity

The most general supersymmetric Lagrangian with the fields in Sec. 2.4.2 results in couplings that violate lepton and baryon numbers, such as $\mu' L_i H_u$ and $\lambda_{ijk} \bar{U}_i \bar{D}_j \bar{D}_k$. However, these violations are under strict restrictions from experiment, such as the search for proton decay $p \rightarrow e^+ \pi^0$. This decay would violate

both baryon and lepton number by 1 unit, but has not been observed. A new, multiplicative conserved quantity is therefore introduced, namely *R-parity*

$$P_R \equiv (-1)^{3(B-L)+2s}, \quad (2.58)$$

where s is spin, B is baryon number and L is lepton number. This quantity is +1 for SM particles, and -1 for sparticles. If *R*-parity is to be conserved sparticles must therefore always be produced and annihilated in pairs. A further consequence is that there must exist a stable, *lightest supersymmetric particle* (LSP), to which all other supersymmetric particles eventually decay. For this particle to have gone undetected it should have zero electric and color charge. These properties make the LSP a good candidate for dark matter [4].

2.4.4 Soft Breaking terms

The allowed soft breaking terms that conserve *R*-parity and gauge invariance are, in component fields, as follows

$$\begin{aligned} \mathcal{L}_{\text{soft}} = & -\frac{1}{2}M_1\tilde{B}\tilde{B} + M_2\widetilde{W}^a\widetilde{W}^a + M_3\widetilde{g}^a\widetilde{g}^a + c.c. \\ & - a_{ij}^u\tilde{Q}_iH_u\tilde{u}_{jR}^* - a_{ij}^d\tilde{Q}_iH_d\tilde{d}_{jR}^* - a_{ij}^e\tilde{L}_iH_d\tilde{e}_{jR}^* + c.c. \\ & - (m_u^2)_{ij}\tilde{u}_{iR}^*\tilde{u}_{jR} - (m_d^2)_{ij}\tilde{d}_{iR}^*\tilde{d}_{jR} - (m_e^2)_{ij}\tilde{e}_{iR}^*\tilde{e}_{jR} \\ & - (m_Q^2)_{ij}\tilde{Q}_i^\dagger\tilde{Q}_j - (m_L^2)_{ij}\tilde{L}_i^\dagger\tilde{L}_j \\ & - m_{H_u}^2H_u^*H_u - m_{H_d}^2H_d^*H_d - (bH_uH_d + c.c.), \end{aligned} \quad (2.59)$$

where the M_i are potentially complex valued, introducing six new parameters; the a_{ij} are potentially complex values, introducing 54 new parameters, b is potentially complex valued, introducing two new parameters; the m_{ij}^2 are complex valued and hermitian, introducing 47 new parameters. After removing excessive degrees of freedom, the MSSM Lagrangian has introduced a total of 105 new parameters, where 104 come from the soft terms and μ comes from the superpotential.

2.4.5 Radiative Electroweak Symmetry Breaking

As discussed in Sec. 2.1, the SM particles obtain their mass when the Higgs has a field value at the minimum of its governing potential. In supersymmetry, the

scalar potential for the Higgs component fields is

$$\begin{aligned}
V(H_u, H_d) = & |\mu|^2(|H_u^0|^2 + |H_u^+|^2 + |H_d^0|^2 + |H_d^-|^2) \\
& + \frac{1}{8}(g^2 + g'^2)(|H_u^0|^2 + |H_u^+|^2 - |H_d^0|^2 - |H_d^-|^2)^2 \\
& + \frac{1}{2}g^2|H_u^+ H_d^{0*} + H_u^0 H_d^{-*}|^2 \\
& + m_{H_u}^2(|H_u^0|^2 + |H_u^+|^2) + m_{H_d}^2(|H_d^0|^2 + |H_d^-|^2) \\
& + [b(H_u^+ H_d^- - H_u^0 H_d^0) + \text{c.c.}]
\end{aligned} \tag{2.60}$$

Using gauge freedom, this potential can be simplified to

$$\begin{aligned}
V(H_u^0, H_d^0) = & (|\mu|^2 + m_{H_u}^2)|H_u^0|^2 + (|\mu|^2 + m_{H_d}^2)|H_d^0|^2 \\
& + \frac{1}{8}(g^2 + g'^2)(|H_u^0|^2 - |H_d^0|^2)^2 - (bH_u^0 H_d^0 + \text{c.c.})
\end{aligned} \tag{2.61}$$

Analogous to the SM, $SU(2)_L \times U(1)_Y$ should be broken down to $U(1)_{\text{em}}$ in order to give masses to gauge bosons and SM fermions. It can be shown that this potential has a minimum for finite field values, that this minimum has a remaining $U(1)_{\text{em}}$ symmetry, and that the potential is bounded from below. For the potential to have a negative mass term the following is required

$$b^2 > (|\mu|^2 + m_{H_u}^2)(|\mu|^2 + m_{H_d}^2), \tag{2.62}$$

and the potential is bounded from below if

$$2b < 2|\mu|^2 + m_{H_u}^2 + m_{H_d}^2. \tag{2.63}$$

Assumed that $m_{H_u} = m_{H_d}$ at some high scale, the requirements Eqs. (2.62) – (2.62) cannot be simultaneously satisfied at that scale. However, to one-loop the Renormalization Group Equation (RGE)¹ for $m_{H_u}^2$ and $m_{H_d}^2$ are

$$16\pi^2\beta_{m_{H_u}^2} \equiv 16\pi^2 \frac{dm_{H_u}^2}{dt} = 6|y_t|^2(m_{H_u}^2 + m_{Q_3}^2 + m_{u3}^2) + \dots \tag{2.64}$$

$$16\pi^2\beta_{m_{H_d}^2} \equiv 16\pi^2 \frac{dm_{H_d}^2}{dt} = 6|y_b|^2(m_{H_d}^2 + m_{Q_3}^2 + m_{d3}^2) + \dots, \tag{2.65}$$

where y_t and y_b are the top and bottom quark Yukawa couplings, respectively, and $m_{Q_3} = m_{33}^Q$, $m_{u3} = m_{33}^u$ and $m_{d3} = m_{33}^d$. Since $y_t \gg y_b$, $m_{H_u}^2$ runs much faster than $m_{H_d}^2$ as they approach the electroweak scale, and can become negative. This is called the *radiative electroweak symmetry breaking*.

The vector boson masses are known from experiment, and provide constraints on Higgs vevs $v_u = \langle H_u^0 \rangle$ and $v_d = \langle H_d^0 \rangle$

$$v_u^2 + v_d^2 \equiv v^2 = \frac{2m_Z^2}{g^2 + g'^2} \approx (174 \text{ GeV})^2. \tag{2.66}$$

¹The RGE describes how parameters change as a function of energy scale.

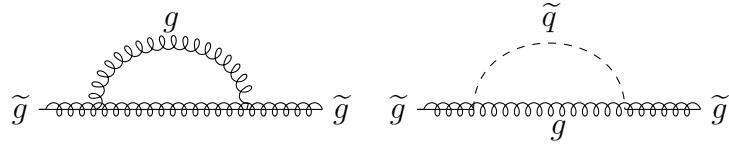


Figure 2.2: One loop contributions to the gluino mass.

The vevs therefore provide a single free parameter, which can be expressed as

$$\tan \beta \equiv \frac{v_u}{v_d}. \quad (2.67)$$

The parameters b and $|\mu|$ can be eliminated as free parameters of the model, but the sign of μ , $\text{sgn } \mu$, cannot.

2.4.6 Sparticles

The most important sparticles in this project are the gluinos and the squarks. In this section the masses of the gluinos and the squarks, as well as the neutralinos, are quickly overviewed.

Gluinos

The gluino is the superpartner of the gluon, which is the boson responsible for the strong interaction. At tree level the gluino does not mix with anything in the MSSM and the mass is the soft term M_3 , but with loop contributions such as those in Fig. 2.2 the mass runs quickly with energy μ . The gluino mass with one loop contributions in the \overline{DR} scheme is

$$m_{\tilde{g}} = M_3(\mu) \left[1 + \frac{\alpha_s}{4\pi} \left(15 + 6 \ln \frac{\mu}{M_3} + \sum_{\text{all } \tilde{q}} A_{\tilde{q}} \right) \right], \quad (2.68)$$

where the squark contributions are

$$A_{\tilde{q}} = \int_0^1 dx \ x \ln \left(x \frac{m_{\tilde{q}}^2}{M_3^2} + (1-x) \frac{m_q^2}{M_3^2} x(1-x) - i\epsilon \right). \quad (2.69)$$

Squarks

In supersymmetry there are two squarks \tilde{q}_L, \tilde{q}_R per quark q , and in the MSSM several terms contribute to their masses. For the first two generations, which are the ones relevant to this project, the main contributions come from the soft terms and the scalar potential. The contributions from soft terms assume that the soft

masses are close to diagonal, and provide contributions $-m_Q^2 \tilde{Q}_i^\dagger \tilde{Q}_i$ and $-m_q^2 \tilde{q}_{iR}^* \tilde{q}_{iR}$ for an $SU(2)_L$ doublet \tilde{Q}_i and singlet \tilde{q}_i with index i , respectively. Note that same generation left-handed squarks share the same soft mass parameter, $m_{Q_i}^2$, which makes the splitting between $m_{\tilde{d}_L}^2$ and $m_{\tilde{u}_L}^2$ much smaller than the splitting for $m_{\tilde{d}_R}^2$ and $m_{\tilde{u}_R}^2$.

The scalar potential contributes with hyperfine terms $\frac{1}{2} \sum g_a^2 (A^* T^a A)^2$. This becomes of the form $(\text{sfermion})^2 (\text{Higgs})^2$ when one of the scalar fields A is a Higgs field. When the Higgs develops vacuum expectation values v , these terms become mass terms for the squarks

$$\Delta_Q = (T_{3F}g^2 - Y_Fg'^2)(v_d^2 - v_u^2) = (T_{3F} - Q_F \sin^2 \theta_W) \cos 2\beta m_Z^2, \quad (2.70)$$

where the isospin T_3 , hypercharge Y , and electric charge Q are the charges of the left-handed supermultiplet to which the squark belongs. These terms are the same for other sfermions.

The mass terms of the squarks are then *e.g.*

$$m_{\tilde{u}_L}^2 = m_{Q_1}^2 + \Delta \tilde{u}_L, \quad (2.71)$$

$$m_{\tilde{d}_L}^2 = m_{Q_1}^2 + \Delta \tilde{d}_L, \quad (2.72)$$

$$m_{\tilde{u}_R}^2 = m_{u_1}^2 + \Delta \tilde{u}_R, \quad (2.73)$$

with the mass splittings between the same generation left-handed squarks given by

$$m_{\tilde{d}_L}^2 - m_{\tilde{u}_L}^2 = -\frac{1}{2}g^2(v_d - v_u) = -\cos 2\beta m_W^2. \quad (2.74)$$

Neutralinos and Charginos

Electroweak symmetry breaking allows the gauge fields to mix. The only requirement is that fields of the same $U(1)_{\text{em}}$ charge mix. This gives fields like the photino and zino, which are supersymmetric partners to the photon and Z -boson. The photino and zino are mixes of the neutral fields \tilde{B}^0 and \tilde{W}^0 . However, the gauge fields are also free to mix with the Higgsinos, the fermions in the Higgs superfields, giving particles known as *neutralinos*. There are four neutralinos,

$$\tilde{\chi}_i^0 = N_{i1} \tilde{B}^0 + N_{i2} \tilde{W}^0 + N_{i3} \tilde{H}_d^0 + N_{i4} \tilde{H}_u^0, \quad (2.75)$$

for $i = 1, 2, 3, 4$, where N_{ij} indicates how much of each component field is mixed in the neutralino. There are also charged particles, known as *charginos*, which are similar to the neutralinos but mixes \tilde{W}^+ , \tilde{H}_u^+ , \tilde{W}^- and \tilde{H}_d^- .

2.4.7 MSSM-24

As noted in Sec. 1.4.3, the MSSM introduces 105 new parameters, where 104 come from soft terms and one comes from the scalar superpotential. However, experimental results put restrictions on some parameters at high energy scales. One of the models with a restricted number of parameters is MSSM-24, where the number of parameters is reduced to 24. This is the model used to generate data in this project.

Off-diagonal terms in the slepton and squark mass matrices $(m_f^2)_{ij}$ could induce flavour-changing processes, such as $\mu \rightarrow e\gamma$. Since these processes have not yet been observed experimentally, the squark and lepton mass matrices are assumed to be diagonal,

$$(m_f^2)_{ij} = \text{diag}(m_{f1}^2, m_{f2}^2, m_{f3}^2), \quad f = u, d, e, Q, L. \quad (2.76)$$

Another restriction comes from CP-violation. To avoid inducing large CP-violating phases, the gaugino masses and three-scalar couplings are assumed to be real

$$\text{Im}(M_1) = \text{Im}(M_2) = \text{Im}(M_3) = \text{Im}(A_0^u) = \text{Im}(A_0^d) = \text{Im}(A_0^e) = 0. \quad (2.77)$$

Finally, as there is a one-to-one correspondence between the three-scalar terms in the SM y_{ij}^f and in the MSSM a_{ij}^f , these are taken to be related through proportionality constants

$$a_{ij}^u = A_0^u y_{ij}^u, \quad a_{ij}^d = A_0^d y_{ij}^d, \quad a_{ij}^e = A_0^e y_{ij}^e. \quad (2.78)$$

The parameters of the MSSM-24 are then the following

$M_1, M_2, M_3,$	Gaugino mass parameters,
A_0^u, A_0^d, A_0^e	Trilinear couplings,
$\tan \beta, \mu, b,$	Higgs parameters,
$m_{\tilde{Q}_1}^2, m_{\tilde{Q}_2}^2, m_{\tilde{Q}_3}^2$	Squark mass parameters,
$m_{\tilde{u}_1}^2, m_{\tilde{u}_2}^2, m_{\tilde{u}_3}^2$	
$m_{\tilde{d}_1}^2, m_{\tilde{d}_2}^2, m_{\tilde{d}_3}^2,$	
$m_{\tilde{L}_1}^2, m_{\tilde{L}_2}^2, m_{\tilde{L}_3}^2,$	Slepton mass parameters
$m_{\tilde{e}_1}^2, m_{\tilde{e}_2}^2, m_{\tilde{e}_3}^2.$	

2.4.8 Constrained MSSM

Another version of the MSSM is the *Constrained MSSM* (CMSSM), also called *minimal supergravity* (mSUGRA). The CMSSM assumes that soft supersymmetry breaking happens in a *hidden sector*. A hidden sector is some non-accessible

high energy scale where the fields of the sector have very small or no direct couplings to the fields in the visible sector. These fields acquire a non-zero vacuum expectation value that is mediated down to the visible sector, via an interaction that is common for both sectors.

In the CMSSM the hidden sector used is the *Planck-Mediated Symmetry Breaking* (PMSB), with some gravity mechanism at the Planck scale $\Lambda_P \sim 10^{18}$ GeV, which explains the name minimal supergravity. In this scheme it is assumed that soft terms only provide the following four (and a half) parameters

$$m_{1/2}, \quad m_0^2, \quad A_0, \quad B_0, \quad \text{sgn } \mu, \quad (2.79)$$

which give, at the renormalization scale $m_s = M_P$, for the soft breaking terms

$$M_3 = M_2 = M_1 = m_{1/2}, \quad (2.80)$$

$$(m_Q^2)_{ij} = (m_u^2)_{ij} = (m_d^2)_{ij} = (m_e^2)_{ij} = (m_L)_{ij}^2 = \text{diag } (m_0), \quad (2.81)$$

$$a_{ij}^u = A_0 y_{ij}^u, \quad a_{ij}^d = A_0 y_{ij}^d, \quad a_{ij}^e = A_0 y_{ij}^e, \quad (2.82)$$

$$b = B_0 \mu. \quad (2.83)$$

The mass parameters are then run down to the electroweak scale, where *e.g.* the hyperfine mass splitting for $m_{\tilde{u}_L}$ and $m_{\tilde{d}_L}$ is included.

Chapter 3

Supersymmetry at Hadron Colliders

The Large Hadron Collider (LHC) at CERN is one of the largest and most important particle physics experiments in the world. In this chapter, some of the advantages and challenges of using hadron colliders are discussed, along with techniques for moving from theory to observable signals. A short description of supersymmetric phenomenology follows, and current bounds on the squark and gluino masses are given. Finally, the cross section for squark pair production is calculated to leading order, and next-to-leading order terms are investigated.

3.1 Hadron Colliders

Colliding hadrons makes it possible to reach very high center-of-mass energies, as they allow for the use of circular accelerators. While most linear colliders collide leptons, such as e^-e^+ , leptons are not advantageous for circular accelerators because of *synchrotron radiation*. Synchrotron radiation is the radiation of energy from a particle being accelerated. The power radiated by a relativistic charged particle forced to move in circular motion with radius R is given by the Schwinger's formula [5]

$$P_e = \frac{2}{3} \frac{e^2 c}{R^2} \left(\frac{E}{mc^2} \right)^4, \quad (3.1)$$

where E is the particle energy, m is the mass of the particle and c is the speed of light in vacuum. For light particles, such as leptons, a lot of energy is therefore wasted in circular accelerators. Because the proton mass is much larger than the electron mass, this effect is relatively small, allowing the LHC to operate at energies currently as high as 13 TeV. In this project the data is generated at 8 TeV. The circular form means the accelerating structures can be reused

as many times as one desires, thus putting ‘no limits’ on the energies obtained. There are, of course, limits. At around energies of around $5 - 7$ TeV per particle, synchrotron radiation becomes an important effect also for protons. The photons emitted from synchrotron radiation hit the walls of the vacuum chamber walls, where they can interact with electrons and cause *electron clouds*. Electron clouds occur when electrons are ejected from the walls of the vacuum chamber and accelerated towards a passing beam bunch. When the electrons reach the center of the chamber, however, the beam bunch has passed, and the now-energetic electrons hit the other side of the chamber, producing more free electrons. This becomes a ‘cloud’ of electrons that can in turn affect the particle beam.

A challenge more specific to hadron collisions is the internal distribution of momentum in a hadron. Hadrons are made up of valence and sea quarks, which make the kinematics of collisions very difficult to calculate. Valence quarks are the quarks used to classify a hadron, such as *uud* for the proton and *udd* for the neutron. In addition to these, hadrons contain a sea of virtual quarks and gluons. The quarks and gluons, called *partons*, distribute the hadron momentum somewhat randomly amongst themselves. Since the distribution of energy and momentum is unknown, so is the momenta of the ingoing parton in the collision. The *transverse* momentum and energy, not to mention the *missing* transverse energy, are therefore central when analysing data from hadron colliders. There is, however, some experimental knowledge of the distribution of longitudinal momentum contained in *parton distribution functions*.

3.1.1 Parton Distribution Functions

Partonic cross sections are calculated for colliding partons in ingoing protons, such as two quarks $q_1 q_2$. The cross section is a function of the center-of-mass energy, s , which can be written as a fraction of the center of mass energy of the colliding protons

$$s = Sx_1x_2, \quad (3.2)$$

where S is the center of mass energy of the colliding protons, and x_i is the momentum fraction of the quark q_i . The fractions of momenta are then integrated over, using *parton distribution functions* (PDFs) $f(x_i)$, which are specified for the different partons in different hadrons. For example, the fraction x_u for an up-flavour quark in a proton would be much larger than that for a top-flavour quark. Integrating over parton distribution functions yields the *total cross section*

$$\sigma_{q_1 q_2} = \int f(x_1)f(x_2)\hat{\sigma}_{q_1 q_2}(s)dx_1 dx_2, \quad (3.3)$$

where $\hat{\sigma}_{q_1 q_2}$ is the partonic cross section. In this project the CTEQ6 parton distribution functions from the LHAPDF Fortran library [6] are used. The total

number of events can be found by multiplying the total cross section by the *luminosity* of the collider.

3.1.2 Luminosity

The *instantaneous luminosity* \mathcal{L} is a measure of the number of collisions in a collider per unit time. The integrated luminosity, \mathcal{L}_{int} , is the luminosity integrated with respect to time, and is given in inverse femto-barn $1 \text{ fb}^{-1} = 10^{43} \text{ m}^{-2}$. The luminosity of the data can be used to set limits on the size of cross sections, by considering the following relation for a process where a particle A is produced

$$n_A = \mathcal{L}_{\text{int}} \sigma_A, \quad (3.4)$$

where \mathcal{L}_{int} is the integrated luminosity, σ_A is the cross section for the production of A , and n_A is the number of produced particles. Setting $n = 1$ for a single produced particle, and using the integrated luminosity $\mathcal{L}_{\text{int}} = 20.3 \text{ fb}^{-1}$ for the 8 TeV data set considered in this project, the lower limit on cross sections is

$$\sigma = \frac{1}{20.3 \text{ fb}^{-1}} \approx 0.05 \text{ fb}. \quad (3.5)$$

Cross sections below $\sigma \sim \mathcal{O}(10^{-3} \text{ fb}^{-1})$ correspond to 0.02 produced particles in the 8 TeV data set, and will therefore be considered less important in this project.

3.2 Phenomenology

Phenomenology is the application of theory to high energy experiments such as the Large Hadron Collider described in Sec. 3.1. This section deals with searches for supersymmetry at hadron colliders using missing transverse energy and quark jets. Some current bounds on sparticles are discussed as well.

3.2.1 Searches For Supersymmetry

Supersymmetry at hadron colliders will likely be in the form of QCD processes, as the colliding particles are quarks and gluons. The traces left in detectors by such processes will be closely related to the conservation of R -parity, in that all sparticles are produced in pairs and eventually decay to the LSP, as mentioned in Sec. 2.4.3. If the LSP is indeed only weakly interacting, this makes it very difficult to detect. An indirect manner of detection involves looking for *missing transverse energy* \cancel{E}_T . The energy is considered solely in the transverse plane, as the longitudinal momentum is difficult to predict in the hadronic case, see the discussion of hadron colliders in Sec. 3.1. To look for \cancel{E}_T the *effective mass* is

defined as

$$M_{eff} = \sum p_T^{jet} + \cancel{E}_T, \quad (3.6)$$

and used to search for deviations from SM expectations. The momentum p_T^{jet} is the transverse momentum of *jets*. Jets are collimated¹ bunches of final-state partons and hadrons that appear in hard interactions². They can be thought of as energetic partons that undergo showering and then hadronization. Jets are common traces in hadron collision detectors. Because of asymptotic freedom³ partons do not remain unbound for long and form the jet hadrons. In supersymmetric processes from hadron collisions the production of jets should also be common, as the LSP is assumed color neutral and the color charge of gluinos and squarks must end up in colored SM particles. An example of squark pair production with the subsequent decay into jets is shown in Fig. 3.1. The hadronic jets produced from supersymmetric processes can be complicated and difficult to identify. The significant background from SM processes offers another difficulty in searching for supersymmetry.

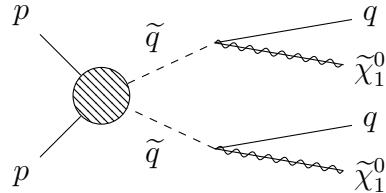


Figure 3.1: Possible signature of a supersymmetric QCD process, with two quark jets and large missing transverse energy in the final state.

Allowing for *R*-parity violation opens up for more final-state possibilities. The LSP can then decay, and sparticles can be produced one at a time. The violation of *R*-parity allows for *massive metastable charged particles* (MMCPs), typical for scenarios with a gravitino LSP [1].

3.2.2 Current Bounds on Sparticles

Searches and limits are available from the data recorded in 2015 by the ATLAS experiment in $\sqrt{s} = 13$ TeV proton-proton collisions at the LHC, with an integrated luminosity of 2.3 fb^{-1} . The analysis included searches for jets and missing transverse energy. Simplified models were assumed, with *R*-parity conservation

¹A collimated beam is parallel and does not disperse with distance.

²Hard interactions have very energetic ingoing particles

³The strong coupling constant becomes larger as the distance between partons increases, meaning that partons are *confined* in hadrons.

and the lightest neutralino as the lightest supersymmetric particle. At 95% confidence level, exclusion limits of the gluino and squark masses are set at 1.51 TeV and 1.03 TeV, respectively [7], assuming a *massless lightest neutralino*. A plot of exclusion limits is shown in Fig. 3.2. Note that for a massive lightest neutralino the bounds on the gluino and squark masses are significantly reduced, down to 400 GeV for the squark mass and around 650 GeV for the gluino mass. Values below these mean that the lightest neutralino is no longer the lightest particle $m_{\tilde{q}} < m_{\tilde{\chi}_1^0}$, $m_{\tilde{g}} < m_{\tilde{\chi}_1^0}$, and sparticles must decay to another particle. In this project squark and gluino masses in the interval [0 – 4000] GeV are explored, as the neutralino is assumed to be massive.

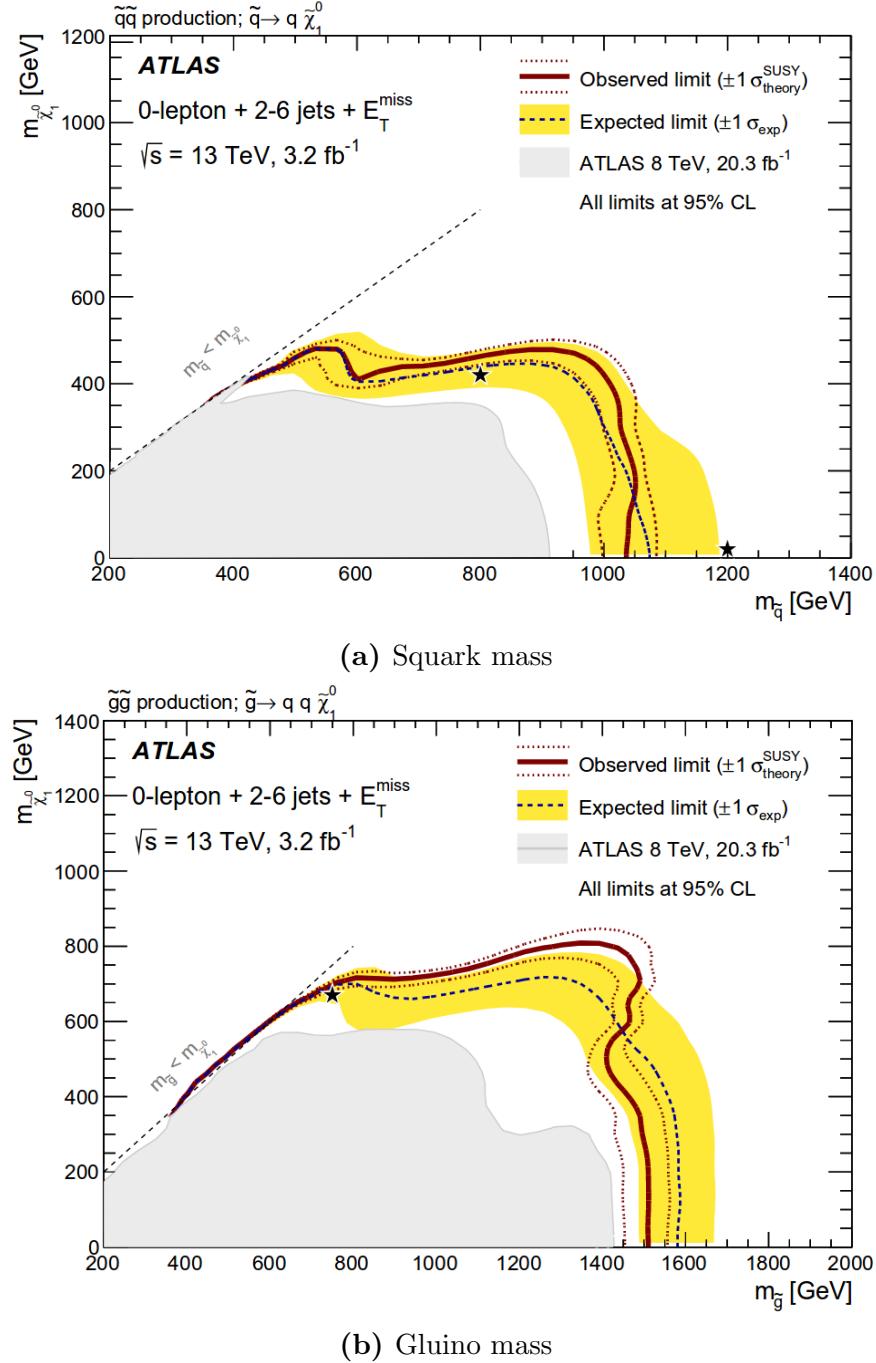


Figure 3.2: Exclusion limits from the ATLAS experiment in $\sqrt{s} = 13 \text{ TeV}$ proton-proton collisions. The analysis assumes conservation of R -parity and a lightest neutralino LSP. The exclusion limits also assume a massless lightest neutralino. The dashed lines indicate the limit where $m_{\tilde{q}/\tilde{g}} < m_{\tilde{\chi}_1^0}$, so the squarks and gluinos cannot decay to the lightest neutralino above this limit. Figures from [7].

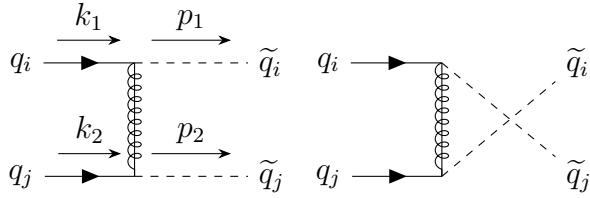


Figure 3.3: Feynman diagrams for squark pair production in quark-quark collisions, both t - and u -channel diagrams. Note that the u -channel (right diagram) is only possible for $i = j$.

3.3 Squark-Squark Cross Section

In this project the relevant QCD process will be the production of squark pairs in quark-quark collisions,

$$q_i q_j \rightarrow \tilde{q}_i \tilde{q}_j, \quad (3.7)$$

where i, j are the first and second generation quark flavours u, d, s, c . Feynman diagrams for tree-level contributions are found in Fig. 3.3. For equal flavour quarks the t - and u -channel contribute, while only the t -channel contributes for different flavours.

3.3.1 Leading Order Cross Section

In this section the partonic cross section for squark pair-production is calculated to leading order. The exchanged gluino momentum is denoted p in the calculations, and defined as $p = k_2 - p_2$ for the t -channel and $p = k_2 - p_1$ for the u -channel. The following set of kinematical invariants are used

$$\begin{aligned} s &= (k_1 + k_2)^2 = 2k_1 \cdot k_2, & t_1 &= (k_2 - p_2)^2 - m_{\tilde{q}}^2, & t_g &= (k_2 - p_2)^2 - m_{\tilde{g}}^2, \\ t &= (k_2 - p_2)^2 = m_{\tilde{q}}^2 - 2(k_2 \cdot p_2), & u_1 &= (k_1 - p_2)^2 - m_{\tilde{q}}^2, & u_g &= (k_1 - p_2)^2 - m_{\tilde{g}}^2, \\ u &= (k_1 - p_2)^2 = m_{\tilde{q}}^2 - 2(k_1 \cdot p_2), \end{aligned}$$

where the Mandelstam variables are related by $t+u+s=p_1^2+p_2^2$. The expressions for the different chiralities are considered later, until then the chiral projection operators in the matrix element are denoted as P and P' . The Feynman gauge is used, and the $n_f = 5$ light flavour quarks are treated as massless. The total matrix element is

$$i\mathcal{M} = i\mathcal{M}_t + i\mathcal{M}_u, \quad (3.8)$$

which gives four terms for the matrix element squared

$$|\mathcal{M}|^2 = |\mathcal{M}_t|^2 + |\mathcal{M}_t \mathcal{M}_u| + |\mathcal{M}_u \mathcal{M}_t| + |\mathcal{M}_u|^2. \quad (3.9)$$

Pure t -channel

The matrix element for the pure t -channel from Eq. (3.8), $i\mathcal{M}_t$, becomes (reading direction is from q_j to q_i)

$$\begin{aligned} i\mathcal{M}_t &= \bar{v}(k_1) \left(-i\sqrt{2}g P(t_a)^{ij} \right) \times \delta^{ab} \frac{i}{\not{p} - m_{\tilde{g}}} \times \left(-i\sqrt{2}g P'(t_b)^{lk} \right) \times u(k_2) \\ &= -(t_a)^{ij} (t^a)^{lk} \times \frac{i2g^2}{t_g^2} \times \bar{v}(k_1) P(\not{p} + m_{\tilde{g}}) P' u(k_2), \end{aligned}$$

where the color factor has been factored out. The matrix element squared is then

$$\begin{aligned} |\mathcal{M}_t|^2 &= (t_a)^{ij} (t^a)^{lk} (t_b)_{ij} (t^b)_{lk} \times \frac{4g^4}{t_g^2} (\bar{v}(k_1) P(\not{p} + m_{\tilde{g}}) P' u(k_2)) \\ &\quad \times (\bar{u}(k_2) P(\not{p} + m_{\tilde{g}}) P' v(k_1)). \end{aligned}$$

To sum over colors, the following relation is used [8]

$$\sum_a (t^a)_{ij} (t_a)_{lk} = \frac{1}{2} (\delta_{ik} \delta_{lj} - \frac{1}{N} \delta_{ij} \delta_{lk}), \quad (3.10)$$

which gives for the color factor

$$\begin{aligned} \sum (t^a)^{ij} (t_a)^{kl} (t^b)_{ij} (t_b)_{kl} &= \frac{1}{4} (\delta_{ik} \delta_{lj} - \frac{1}{N} \delta_{ij} \delta_{lk}) (\delta^{ik} \delta^{lj} - \frac{1}{N} \delta^{ij} \delta^{lk}) \\ &= \frac{1}{4} (N^2 - 1) \\ &= \frac{1}{2} NC_F, \end{aligned}$$

where $C_F = (N^2 - 1)/(2N)$. Averaging over spin gives

$$\sum |\mathcal{M}_t|^2 = \frac{1}{2} NC_F \times \frac{4g^4}{t_g^2} \text{tr} [\not{k}_1 P(\not{p} + m_{\tilde{g}}) P' \not{k}_2 P(\not{p} + m_{\tilde{g}}) P'].$$

As previously mentioned, the quarks are considered massless. The final state squarks can have equal or different chiralities, and both combinations contribute to the total cross section.

Different chiralities $P = P_{R/L}$, $P' = P_{L/R}$

For squarks of different chiralities the trace becomes

$$\begin{aligned} \text{tr} [\not{k}_1 P_{R/L} (\not{p} + m_{\tilde{g}}) P_{L/R} \not{k}_2 P_{R/L} (\not{p} + m_{\tilde{g}}) P_{L/R}] &= 2(2(p \cdot k_2)(k_1 \cdot p) - p^2(k_1 \cdot k_2)) \\ &= (t - m_{\tilde{q}}^2)s + (t - m_{\tilde{q}}^2)(u - m_{\tilde{q}}^2) - ts \\ &= t_1 u_1 - m_{\tilde{q}}^2 s, \end{aligned}$$

where $P_{R/L} P_{R/L} = P_{R/L}$, $(\gamma^5)^2 = 1$ and $\text{tr}[\gamma^5 \gamma^\mu \gamma^\nu] = 0$. For same flavour ingoing quarks, qq , the only possibility for different chiralities is $\tilde{q}_R \tilde{q}_L$. For different flavour quarks, $q'q$, there are two possibilities, namely $\tilde{q}_R \tilde{q}'_L$ and $\tilde{q}'_L \tilde{q}_R$.

Equal chiralities $P = P_{R/L}$, $P' = P_{R/L}$

For squarks of equal chiralities the trace becomes

$$\text{tr}[\not{k}_1 P_{R/L}(\not{p} + m_{\tilde{g}}) P_{R/L} \not{k}_2 P_{R/L}(\not{p} + m_{\tilde{g}}) P_{R/L}] = 2m_{\tilde{g}}^2(k_1 \cdot k_2) = m_{\tilde{g}}^2 s.$$

For equal flavour quarks qq there are now two possibilities for equal chirality squarks, namely $\tilde{q}_R \tilde{q}_R$ and $\tilde{q}_L \tilde{q}_L$. For different flavour quarks qq' the two possibilities for equal chiralities are $\tilde{q}_L \tilde{q}'_L$ and $\tilde{q}_R \tilde{q}'_R$.

Note that the contribution from $\tilde{q}_R \tilde{q}_R$ is identical to $\tilde{q}_L \tilde{q}_L$ in QCD processes. If no electroweak corrections are included in the higher order term, the NLO cross section should also be identical as a function of $m_{\tilde{q}}$, because only the electroweak interaction couples to right- and left-handed states differently.

Sum over Chiralities

For incoming quarks $q_i q_j$ the sum over chiralities yields

$$\sum |\mathcal{M}_t|^2 = NC_F \times \frac{4g^4}{t_g^2} \left[\delta_{ij} \left(\frac{1}{2}(t_1 u_1 - sm_q^2) + sm_{\tilde{g}}^2 \right) + (1 - \delta_{ij}) \left(t_1 u_1 - s(m_q^2 - m_{\tilde{g}}^2) \right) \right],$$

where the terms proportional to δ_{ij} only contribute for equal flavour quarks, and the terms proportional to $1 - \delta_{ij}$ contribute for different flavour quarks.

Pure u -channel

The expression for the u -channel matrix element in Eq. (3.8) is identical to the t -channel matrix element, but for the exchange $t_g^2 \rightarrow u_g^2$. The u -channel only contributes for equal flavour quarks, $i = j$, so the matrix element squared is

$$\sum |\mathcal{M}_u|^2 = \delta_{ij} NC_F \frac{4g^4}{u_g^2} \left[\frac{1}{2}(t_1 u_1 - sm_q^2) + sm_{\tilde{g}}^2 \right].$$

Cross term

The cross term between the t - and u -channel diagrams in Eq. (3.8), consists of two terms, where $\mathcal{M}_t \mathcal{M}_u$ is given by

$$\begin{aligned} \mathcal{M}_t \mathcal{M}_u &= (t^a)^{ij} (t_a)^{kl} (t^b)_{ik} (t_b)_{jl} \frac{-i2g^2}{t_g} (\bar{v}(k_1) P(\not{p} + m_{\tilde{g}}) P' u(k_2)) \\ &\quad \times \frac{i2g^2}{u_g} (\bar{u}(k_2) P(\not{p} + m_{\tilde{g}}) P' v(k_1)) \end{aligned}$$

The color factor is again found using Eq. (3.10), which gives

$$\sum_{a,b} (t^a)^{ij} (t_a)^{kl} (t^b)_{ik} (t_b)_{jl} = -\frac{1}{2} C_F.$$

Averaging over spins gives

$$\sum \mathcal{M}_t \mathcal{M}_u = -\frac{1}{2} C_F \times \text{tr} \left[\frac{4g^4}{u_g t_g} (\not{k}_1 P(\not{k}_2 - \not{p}_2 + m_{\tilde{q}}) P' \not{k}_2 P(\not{k}_2 - \not{p}_1 + m_{\tilde{q}}) P') \right].$$

Summing over equal and different chiralities gives

$$\sum \mathcal{M}_t \mathcal{M}_u = -\frac{1}{2} C_F \frac{4g^4}{u_g t_g} (-t_1 u_1 + m_{\tilde{q}}^2 s + m_{\tilde{g}}^2 s)$$

The other cross term in Eq. (3.8), $\mathcal{M}_u \mathcal{M}_t$, is similar, but yields a slightly different expression

$$\sum \mathcal{M}_u \mathcal{M}_t = -\frac{1}{2} C_F \frac{4g^4}{u_g t_g} (u_1 t_1 - m_{\tilde{q}}^2 s + m_{\tilde{g}}^2 s).$$

Adding the cross terms then gives

$$\sum |\mathcal{M}_t \mathcal{M}_u + \mathcal{M}_u \mathcal{M}_t| = -\frac{1}{2} C_F \delta_{ij} \frac{4g^4}{u_g t_g} (2m_{\tilde{g}}^2 s).$$

Matrix Elements

The matrix elements can be divided into contributions from $\tilde{q}_{iR} \tilde{q}_{iR}$, $\tilde{q}_{iR} \tilde{q}_{iL}$, $\tilde{q}_{iR} \tilde{q}_{jR}$ and $\tilde{q}_{iR} \tilde{q}_{jL}$, and equivalently for $R \leftrightarrow L$. The matrix elements of the different contributions are

$$\tilde{q}_{iR} \tilde{q}_{iR}, \tilde{q}_{iL} \tilde{q}_{iL} : \quad \sum |\mathcal{M}|_{iRiR} = 4g^4 s m_{\tilde{g}}^2 \left[\frac{1}{2} N C_F \left(\frac{1}{t_g^2} + \frac{1}{u_g^2} \right) - C_F \frac{1}{t_g u_g} \right], \quad (3.11)$$

$$\tilde{q}_{iR} \tilde{q}_{iL} : \quad \sum |\mathcal{M}|_{iRiL} = 4g^4 (u_1 t_1 - s m_{\tilde{q}}^2) \left[\frac{1}{2} N C_F \left(\frac{1}{t_g^2} + \frac{1}{u_g^2} \right) \right], \quad (3.12)$$

$$\tilde{q}_{iR} \tilde{q}_{jR}, \tilde{q}_{iL} \tilde{q}_{jL} : \quad \sum |\mathcal{M}|_{iRjR} = 4g^4 s m_{\tilde{g}}^2 \left[\frac{1}{2} N C_F \frac{1}{t_g^2} \right], \quad (3.13)$$

$$\tilde{q}_{iR} \tilde{q}_{jL}, \tilde{q}_{iL} \tilde{q}_{jR} : \quad \sum |\mathcal{M}|_{iRjL} = 4g^4 (t_1 u_1 - s m_{\tilde{q}}^2) \left[\frac{1}{2} N C_F \frac{1}{t_g^2} \right]. \quad (3.14)$$

Note that, as mentioned, the contribution from $\tilde{q}_R \tilde{q}_R$ and $\tilde{q}_L \tilde{q}_L$ are identical as functions of $m_{\tilde{q}}$.

Summing over the terms from different chirality combinations gives for the total sum over matrix element squared

$$\begin{aligned} \sum |\mathcal{M}|^2 = & \delta_{ij} \left[2g^4 NC_F (u_1 t_1 - s m_{\tilde{q}}^2) \left(\frac{1}{t_g^2} + \frac{1}{u_g^2} \right) \right. \\ & + 4g^4 s m_{\tilde{q}}^2 \left(NC_F \left(\frac{1}{t_g^2} + \frac{1}{u_g^2} \right) - 2C_F \frac{1}{u_g t_g} \right) \Big] \\ & + (1 - \delta_{ij}) \left[4g^4 NC_F \frac{u_1 t_1 - s(m_g^2 - m_{\tilde{q}}^2)}{t_g^2} \right], \end{aligned} \quad (3.15)$$

where, again, the terms proportional to δ_{ij} contribute for equal flavour quarks, and terms proportional to $1 - \delta_{ij}$ contribute for different flavour quarks.

Partonic Cross Section

The $SU(3)$ color factors are given by $N = 3$, which gives $C_F = 4/3$. To find the total, lowest order partonic cross section an n -dimensional phase space integral is performed, and spin and color averaging are taken into account [9]. The lowest order double-differential distributions are then given by

$$\begin{aligned} s^2 \frac{d^2 \sigma^B}{dt du} = & K_{ij} \frac{\pi S_\varepsilon}{\Gamma(1-\varepsilon)} \left[\frac{(t-p_2^2)(u-p_2^2) - p_2^2 s}{\mu^2 s} \right]^{-\varepsilon} \Theta([t-p_2^2][u-p_2^2] - p_2^2 s) \\ & \times \Theta(s-4m^2) \delta(s+t+u-p_1^2-p_2^2) \sum |\mathcal{M}_B|^2, \end{aligned} \quad (3.16)$$

where the averaging over initial state colours and spins is given by the factor K_{ij} , which for squark and antisquark production is [9]

$$K_{qq} = K_{\bar{q}\bar{q}} = \frac{1}{4N^2}. \quad (3.17)$$

Integration over the remaining parameters gives the Born level⁴ partonic cross section for squark pair production $q_i q_j \rightarrow \tilde{q}_i \tilde{q}_j$ [9]

$$\begin{aligned} \hat{\sigma}^B = & \frac{\pi \alpha_s^2}{s} \left[\beta_{\tilde{q}} \left(-\frac{4}{9} - \frac{4m_-^4}{9(m_{\tilde{g}}^2 s + m_-^4)} \right) + \left(-\frac{4}{9} - \frac{8m_-^2}{9s} \right) L_1 \right] \\ & + \delta_{ij} \frac{\pi \alpha_s^2}{s} \left[\frac{8m_{\tilde{g}}^2}{27(s+2m_-^2)} L_1 \right], \end{aligned} \quad (3.18)$$

where

$$L_1 = \ln \left(\frac{s+2m_-^2 - s\beta_{\tilde{q}}}{s+2m_-^2 + s\beta_{\tilde{q}}} \right), \quad \beta_{\tilde{q}} = \sqrt{1 - \frac{4m_{\tilde{q}}^2}{s}}, \quad m_-^2 = m_g^2 - m_{\tilde{q}}^2, \quad \alpha_s = \frac{g_s^2}{4\pi}.$$

⁴Tree-level

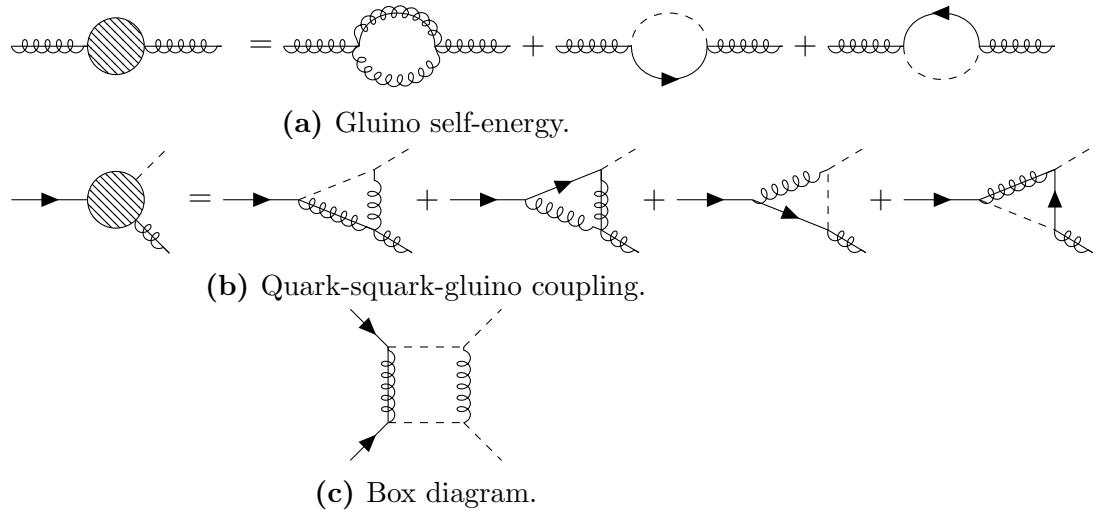


Figure 3.4: A selected set of Feynman diagrams for the virtual corrections to $q_i q_j \rightarrow \tilde{q}_i \tilde{q}_j$ [9].

3.4 Next-to-leading Order Corrections

The next-to-leading order (NLO) terms contain virtual corrections to the Born level Feynman diagrams. By allowing more complicated Feynman diagrams, virtual particles can appear in the process. These never make it to the final state and become real particles, but are instead absorbed by the process. Since they are never real, or on-shell, they can have any momentum and mass. Thus, it is necessary to integrate over all possible momenta, which leads to divergences and infinities.

There are three main categories of divergences, namely ultraviolet, infrared and collinear. Ultraviolet divergences appear when the integral is over an energy without upper bounds. When the opposite problem occurs, that low momentum gives infinities, this is called an infrared divergence. Collinear divergences are also called mass singularities, and stem from masses going to zero, where infinities can appear from splitting at zero angle.

In order to include the corrections and still get physical cross sections, it is necessary to integrate out the divergences using dimensional regularization and renormalise parameters. Renormalising a parameter means giving it a scale dependence, and baking the infinities into this expression. Physically, this can be interpreted as giving the coupling constant a length scale (energy scale $^{-1}$) dependence. For example, the electromagnetic coupling becomes stronger the closer to a charged particle you get. These calulations are often very complicated and tedious. Some of the diagrams for virtual corrections to $q_i q_j \rightarrow \tilde{q}_i \tilde{q}_j$ are shown in Fig. 3.4.

There are several reasons for calculating these cumbersome NLO terms. Firstly,

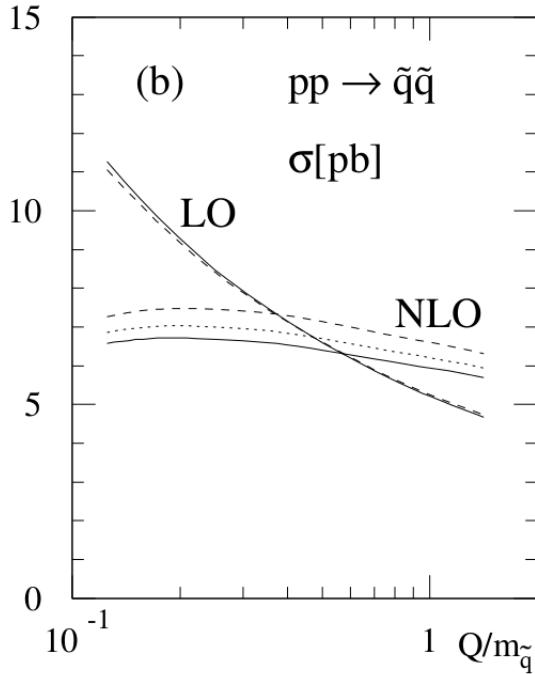


Figure 3.5: The dependence on the renormalization/factorization scale Q for the LO and NLO cross sections for squark-squark production at the LHC ($\sqrt{s} = 14$ TeV). Parton densities are GRV94 (solid), CTEQ3 (dashed) and MRS(A') (dotted). Mass parameters are $m_{\tilde{q}} = 600$ GeV, $m_{\tilde{g}} = 500$ GeV and $m_t = 175$ GeV. Figure from [9].

the leading order terms are highly dependent on the *a priori* unknown renormalization scale, leading to a large uncertainty in theoretical predictions (up to a factor of two). Adding higher order terms reduces this scale dependency. An example is shown in Fig. 3.5, where the LO and NLO cross sections for $qq \rightarrow \tilde{q}\tilde{q}$ are plotted as a function of the renormalization/factorization scale [9]. The NLO terms are clearly less dependent on the scale. Secondly, the NLO contributions are expected to be large and positive, thereby raising the cross sections significantly and allowing for stronger bounds on the gluino and squark masses [9].

The calculations from [9] assume degenerate squark masses $m_{\tilde{q}}$, set the 5 lightest quark masses to zero as they are much lighter than the squarks, and the top quark mass to $m_t = 175$ GeV. The two free parameters left are then $m_{\tilde{g}}$ and $m_{\tilde{q}}$. This indicates that these might be good features for the learning. The renormalization scheme used is the \bar{MS} scheme. Masses and couplings are renormalized, and the resulting parameters can be found in [9].

Next-to-leading Order Partonic Cross-Section

As discussed in Sec. 3.1, cross sections for hadron colliders require first finding the partonic cross sections. In order to analyze these, scaling functions are introduced [9] giving the LO+NLO result as

$$\hat{\sigma}_{ij} = \frac{\alpha_s^2(Q^2)}{m^2} \left\{ f_{ij}^B(\eta, r) + 4\pi\alpha_s(Q^2) \left[f_{ij}^{V+S}(\eta, r, r_t) \right. \right. \\ \left. \left. + f_{ij}^H(\eta, r) + \bar{f}_{ij}(\eta, r) \log\left(\frac{Q^2}{m^2}\right) \right] \right\}, \quad (3.19)$$

where Q^2 is the renormalization scale, often set to $Q^2 = m^2$, and $m = (\sqrt{p_1^2} + \sqrt{p_2^2})/2$ is the average mass of the produced particles. The scaling functions f are as follows: the Born term f^B from Eq. (3.18), the sum of virtual and soft-gluon corrections f^{V+S} , the hard gluon corrections f^H , and the scale-dependent contributions \bar{f} . The partonic cross section is written as a function of the parameters

$$\eta = \frac{s}{4m^2} - 1, \quad r = \frac{m_{\tilde{g}}^2}{m_q^2}, \quad r_t = \frac{m_t^2}{m^2}. \quad (3.20)$$

The energy near the threshold is the base for an important part of the contributions to the cross section [9]. In this region the scaling functions can be expanded in the low velocity of produced particles β , leading to the following expressions [9]

$$f_{qq}^B = \frac{8\pi\beta m_{\tilde{q}}^2 m_{\tilde{g}}^2}{27(m_q^2 + m_{\tilde{g}}^2)^2}, \quad f_{q'q}^B = \frac{8\pi\beta m_{\tilde{q}}^2 m_{\tilde{g}}^2}{9(m_q^2 + m_{\tilde{g}}^2)^2} \\ f_{qq}^{V+S} = f_{qq}^B \frac{1}{24\beta}, \quad f_{q'q}^{V+S} = f_{q'q}^B \frac{1}{24\beta} \\ f_{qq}^H = f_{qq}^B \left[\frac{2}{3\pi^2} \log^2(8\beta^2) - \frac{7}{2\pi^2} \log(8\beta^2) \right], \quad f_{q'q}^H = f_{q'q}^B \left[\frac{2}{3\pi^2} \log^2(8\beta^2) - \frac{19}{6\pi^2} \log(8\beta^2) \right] \\ \bar{f}_{qq} = -f_{qq}^B \frac{2}{3\pi^2} \log(8\beta^2), \quad \bar{f}_{q'q} = -f_{q'q}^B \frac{2}{3\pi^2} \log(8\beta^2). \quad (3.21)$$

The main contributions to the partonic cross section come from this region.

Hadronic Cross-Section

As per the above discussion, the partonic cross-sections must be integrated over in order to obtain the total hadronic cross section. A convolution integral over the parton distribution functions yields the expression

$$\sigma(S, Q^2) = \sum_{i,j=q,\bar{q}} \int_\tau^1 dx_1 \int_{\tau/x_1}^1 dx_2 f_i^{h_1}(x_1, Q^2) f_j^{h_2}(x_2, Q^2) \hat{\sigma}_{ij}(x_1 x_2 S, Q^2) \Big|_{\tau=4m^2/S}. \quad (3.22)$$

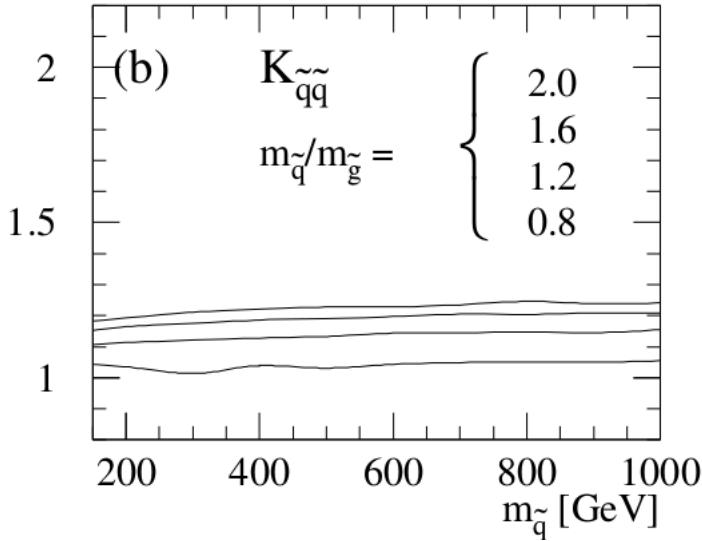


Figure 3.6: The K -factors for the LHC ($\sqrt{s} = 14$ TeV). Parton densities are GRV94, with scale $Q = m$, and the top squark mass is $m_t = 175$ GeV. Figure from [9].

The integrals are calculated numerically using the VEGAS integration routine [10] in [9]. The uncertainty due to different parameterizations of the parton densities for the LHC calculations of the NLO terms amount to $\lesssim 13\%$ at the central scale.

K-Factor

To quantify the change in the cross section from adding NLO terms, the K -factor is introduced. The K -factor is the ratio between the cross sections

$$K = \sigma_{NLO}/\sigma_{LO}. \quad (3.23)$$

The K -factor for squark production for varying mass ratios $m_{\tilde{q}}/m_{\tilde{g}} = 2.0, 1.6, 1.2, 0.8$ at the LHC is shown in Fig. 3.6. As seen from the figure, the K -factor is larger than one and quite stable as a function of the squark mass.

3.4.1 State-of-the-art Tools

There are two numerical tools for the calculation of NLO SUSY cross sections, namely **Prospino** [11] and **NLL-fast** [12].

Prospino

Prospino 2.1 [11] is a numerical tool that calculates supersymmetric cross sections to next-to-leading order for non-degenerate squark masses. It uses the

K -factor to calculate NLO cross sections, by first computing LO cross sections for the five or four lightest squark flavors. The LO and NLO rates are then calculated for a mean value of the squark mass, and the corresponding K -factor is calculated. The LO cross sections for different squarks are then multiplied by the K -factor, giving an approximation to the NLO terms for non-degenerate masses. The calculations are outlined in the section above, and are described in more detail in [11].

Prospino 2.1 uses a renormalization scale to calculate cross sections, which leads to uncertainties. The uncertainty from the scale dependence is estimated by calculating cross sections for twice and half the chosen renormalization scale, and seeing how the cross sections change. As demonstrated in Fig. 3.5 the scale dependence of cross sections is reduced by adding higher order terms. As a consequence, the scale dependence is also a way of estimating the order of magnitude of higher order terms (in this case, next-to-next-to-leading order).

In addition, as was discovered during the project, **Prospino 2.1** has a weakness in the heavy squark mass space. Consider the cross section for the production of $\tilde{c}_L \tilde{c}_L$. When all squark masses are heavy, but $m_{\tilde{c}_L}$ is slightly lighter than the others, the K -factor can be set to zero if the average mass is above threshold ($m^2 > s$). This means that $LO \neq 0$ and $NLO = 0$, which gives problematic outlier points that cause trouble during learning.

Evaluating cross sections with **Prospino 2.1** is also highly time-consuming. Evaluating the strong process cross sections for a single CMSSM benchmark point takes about 15 minutes of CPU time on a modern processor [13].

NLL-fast

NLL-fast 2.1 [9, 14, 15, 16, 17] computes the hadronic cross sections of gluino and squark pair production including NLO supersymmetric QCD corrections, and the resummation of soft gluon emission at next-to-leading-logarithmic (NLL) accuracy. It also provides an error estimation, based on the errors from renormalization scale dependence and the parton distribution functions. The calculation is done by reading in tables of LO and NLO+NLL results, the scale uncertainty and the PDF and α_s error, and uses fast interpolation to calculate cross sections for any squark and gluino mass within the range [200, 2500] GeV for $\sqrt{s} = 8$ TeV. **NLL-fast 2.1** is limited, however, in that it only calculates cross sections for mass-degenerate squarks. Since the true total cross section for squark pair production consists of 36 different cross sections for combinations of the four lightest quarks, the estimate from **NLL-fast 2.1** is very simplified and unfit for *e.g.* MSSM-24, as will be demonstrated in Sec. 6.2.2.

Accuracy

Including the NLO+NLL cross sections has reduced the theoretical error to 10% for a wide range of processes and masses, see the discussion in [13]. There are other uncertainties, from the parton distribution functions (PDFs) and α_s . These must also be included in the total error estimate. PDFs are based on experimental data, so the uncertainty increases with the sparticle masses because the PDFs are most poorly constrained at large scales and at large partons x .

For illustration, consider both the gluino and (degenerate) squark mass to be 1.5 TeV, which is at the edge of what the LHC is able to produce at 8 TeV. **NLL-fast 2.1** gives errors of (+24.3%, -22.2%) for the PDFs and (+8.3%, -7.3%) for α_s , when using the MSTW2008 NLO PDF set [18]. In this case the total error will not be much lower than 25%. However, with new data from the LHC errors from PDFs and α_s will be reduced over time.

In order to keep the regression errors subdominant the relative regression errors obtained in this project should not exceed 10%.

Chapter 4

Gaussian Processes

In this chapter Gaussian process regression is introduced. First, some concepts and terminology in Bayesian statistics are reviewed. The concept of covariance, and how it can be determined by covariance functions is then explored. The following section introduces the mathematical framework of Gaussian processes, and the Gaussian noise model is used as an example. Bayesian model selection and cross validation are introduced as tools to quantify and improve the quality of predictions.

4.1 Introduction to Bayesian Statistics

There are two general philosophies in statistics, namely *Bayesian* and *frequentist* statistics. Statisticians from both branches would probably consider the following statement to be true

Statisticians use probability to describe uncertainty.

The difference between Bayesian and frequentist statistics lies in the definition of the *uncertain*. Since uncertainty is described by probability, the definition of probability must also be different. A distinction is made between *objective* and *subjective* probability, where only the former is used by frequentists, and Bayesians use a combination of the two. Consider an example in which a statistician throws a die. Before throwing, he is uncertain about the outcome of the toss. This uncertainty related to the outcome is *objective*: no one can know if he will throw a 1 or a 4. On the other hand, he might also be uncertain about the underlying probability distribution of the toss. Is the die loaded? Is one of the edges sharper than the others? This uncertainty is *subjective*, as it may vary depending on how much information is available about the system, and how that information is used.

One of the main criticisms of subjective probability posed by frequentists is that the final probability depends on who you ask.

4.1.1 Bayes' Theorem

The difference between frequentist and Bayesian statistics also lies in the application of *Bayes' theorem* [19]. Bayes' theorem can be derived from the familiar rules of probability for random variables X and Y given the information I ,

$$P(X|I) + P(\bar{X}|I) = 1, \quad (4.1)$$

$$P(X, Y|I) = P(X|Y, I)P(Y|I), \quad (4.2)$$

commonly known as the *sum rule* and *product rule*, respectively. The probability $P(X|I)$ is the probability of outcome X given the information I , and $P(X|Y, I)$ is the probability of outcome X given the information I *and* outcome Y . The bar over \bar{X} means that the outcome X does *not* happen. The sum rule states that the total probability of the outcomes X and \bar{X} is equal to 1. This is rather intuitive, considering that an event either takes place or not. The product rule concerns the probability of both outcomes X and Y . This is equal to the probability of Y times the probability of X given that Y has already occurred. The product rule can be used to give Bayes' theorem, first formulated by the reverend Thomas Bayes in 1763,

$$P(Y|X, I) = \frac{P(X|Y, I)P(Y|I)}{P(X|I)}. \quad (4.3)$$

Bayes' theorem states that the probability of Y given X is proportional to the probability of X given Y , times the probability of Y . The proportionality factor is one over the probability of X . Surprisingly, there is nothing Bayesian — in the modern statistical sense — about Bayes' theorem. It is merely a reformulation of the rules of logical consistent reasoning by Richard Cox in 1946 [20]. Laplace was the one to make Bayes' theorem Bayesian, when he used the theorem to perform inference about probability distributions [21]. Probability distributions are functions that describe how the probability of a random variable X is distributed, and are denoted $P(X|\Theta, I)$. The *parameters*, Θ , of these functions determine the shape of the distribution, and are *not* random variables. An example of probability distribution parameters are the mean and variance of a Gaussian distribution. Bayes' theorem can be used to find the probability of the *parameters* Θ given a set of random variables X drawn from the distribution, and the resulting expression is the *posterior probability distribution*

$$P(\Theta|X, I) = \frac{P(X|\Theta, I)P(\Theta|I)}{P(X|I)}, \quad (4.4)$$

where $P(X|\Theta, I)$ and $P(\Theta|I)$ are the *likelihood* and *prior*, respectively, and $P(X|I)$ is a normalization constant called the *marginal likelihood* or evidence.

The marginal likelihood is independent of the parameters

$$P(X|I) = \int P(X|\Theta, I)P(\Theta|I) d\Theta. \quad (4.5)$$

In other words, Eq. (4.4) states the probability of the parameters Θ given the knowledge of outcomes X . The parameters Θ are sometimes called the *hypothesis*, and the variables X are called the *evidence*. Bayes' theorem is then used to modify the hypothesis using the evidence.

A crucial parting of Bayesian statistics from frequentist statistics is at the introduction of the *prior*, which expresses a probability distribution on the *parameters* of the probability distribution before data. The prior and likelihood are discussed in the next section, while the marginal likelihood is revisited in Sec. 1.4.

4.1.2 Priors and Likelihood

The likelihood $P(X|\Theta, I)$ is simply the probability of the observations X given the parameters of the probability distribution Θ , and is revisited in Sec. 4.4.1. Conversely, the prior expresses a prior belief or assumption of the parameters, and has to be determined beforehand. As mentioned previously, the measure $P(\Theta|X, I)$ from Eq. (4.4) is called the posterior distribution. This can be thought of as the prior belief, modified by how well this belief fits the data,

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{marginal likelihood}}.$$

Consider an example. The statistician mentioned before now sets about tossing a coin. Before tossing he assumes the probability of heads is uniformly distributed, and so adopts a flat, or uniform, prior probability distribution. The uniform distribution is illustrated in the first panel in Fig. 4.1. After one toss he gets heads, and the posterior changes to a function with high probability for heads, and low for tails, illustrated in the second panel. After four tosses, of which two gave heads and two gave tails, the posterior in the third panel shows an equal probability for heads and tails, with a wide distribution centered at 0.5. After several tosses the distribution converges to a narrow peak around 0.25, illustrated in the fourth panel. This indicates an unfair coin that is biased towards tails.

4.1.3 Best Estimate and Reliability

The quantity of interest, such as the parameters of a distribution, will now be denoted X . Given a posterior distribution $P(X|\mathcal{D}, I)$ over X , where the prior has been modified by some data \mathcal{D} , it is important to decide how well the posterior fits the data. As will be shown, the posterior can be approximated by a Gaussian

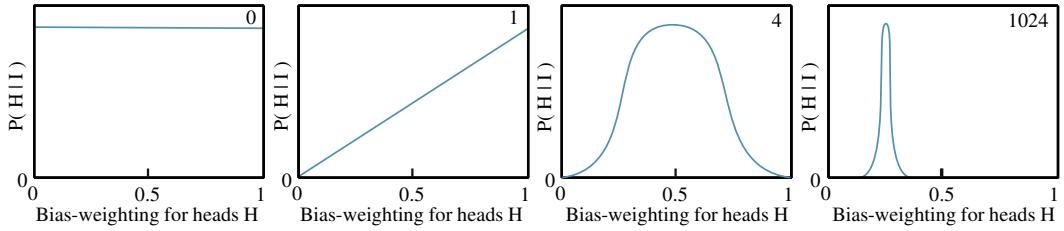


Figure 4.1: The posterior probability distribution of the bias-weighting of a coin for the uniform prior, $P(H|I)$. The first panel from the left is before the coin is tossed, the second panel is after 1 toss, the third is after 4 tosses, and the fourth is after 1024 tosses. The posterior converges towards a narrow peak at 0.25, so the coin is biased.

distribution with a mean and variance, where the mean and variance are given by the *best estimate* and the *reliability* of the posterior, respectively. The best estimate X_0 is the outcome with the highest probability. In other words, it maximizes the posterior distribution

$$\frac{dP}{dX} \Big|_{X_0} = 0, \quad \frac{d^2P}{dX^2} \Big|_{X_0} < 0, \quad (4.6)$$

where P is the posterior $P(X|\mathcal{D}, I)$. The second derivative must be negative to ensure that X_0 is, in fact, a maximum.

Once a best estimate is found, it is important to know how reliable it is. Reliability, or uncertainty, is connected to the width of the distribution. The width of the distribution indicates how much the posterior distribution is smeared out around the best estimate X_0 . A narrow distribution has low uncertainty, while a wide distribution has large uncertainty. As an example, the third panel in Fig. 4.1 shows a distribution with a mean value of 0.5 with large uncertainty, while the fourth panel shows a distribution with mean 0.25 with small uncertainty.

The width is found by taking the logarithm¹ and Taylor expanding the posterior distribution $P(X|\mathcal{D}, I)$

$$L = L(X_0) + \frac{1}{2} \frac{d^2L}{dX^2} \Big|_{X_0} (X - X_0)^2 + \dots, \quad L = \log_e [P(X|\mathcal{D}, I)] \quad (4.7)$$

The first term, $L(X_0)$, is just a constant. From Eq. (4.6) the condition of the best estimate is that $dL/dX|_{X_0} = 0$. The dominant term in determining the width is therefore the quadratic term in Eq. (4.7).

¹ L is a monotonic function of P , so the maximum of L is at the maximum of P .

The Gaussian Distribution

Taking the exponential of Eq. (4.7) and ignoring higher order terms, the posterior can then be approximated as

$$P(X|\mathcal{D}, I) \approx A \exp \left[\frac{1}{2} \frac{d^2 L}{dX^2} \Big|_{X_0} (X - X_0)^2 \right], \quad (4.8)$$

where $A = \exp [L(X_0)]$ is a constant.

Equation (4.8) is now in the shape of a *Gaussian distribution*, given by

$$P(X|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp \left[-\frac{(X - \mu)^2}{2\sigma^2} \right], \quad (4.9)$$

where μ and σ^2 are the two parameters of the Gaussian distribution mentioned above. An example of a Gaussian distribution is shown in Fig. 4.2. The parameter μ is the *mean value* of X , which will be denoted $m(X)$, and σ^2 is the *variance* of the distribution around the mean $m(X)$, denoted $\mathbb{V}(X)$. The variance is discussed further below. The mean and variance for the approximated Gaussian distribution of $P(X|\mathcal{D}, I)$ are then given by

$$m(X) = X_0, \quad \mathbb{V}(X) = \left(-\frac{d^2 L}{dX^2} \right)^{-1/2}. \quad (4.10)$$

The Gaussian distribution is also referred to as the *normal distribution*, and a Gaussian distribution with mean $m(X) = \mu$ and variance $\mathbb{V}(X) = \sigma^2$ is therefore denoted $\mathcal{N}(\mu, \sigma^2)$. The notation $Y \sim \mathcal{N}(\mu, \sigma^2)$ means a *random variable* Y drawn from a Gaussian distribution with mean μ and variance σ^2 . The Gaussian distribution is symmetric with respect to the maximum at the mean μ , and has a full width at half maximum (FWHM) at around 2.35σ , as seen in Fig. 4.2 where $\sigma = \sqrt{\sigma^2}$ is the standard deviation.

Since the basic idea in Gaussian process regression is to predict a Gaussian distribution over function values $f(\mathbf{x})$ for each input vector \mathbf{x} , the Gaussian distribution is central to Gaussian processes. Gaussian processes are discussed in Sec. 4.3, so for now it will suffice to sum up that the quality of a posterior distribution, $P(X|\mathcal{D}, I)$, can be described by two measures: the best estimate and the reliability. These can be seen as the mean and variance of a Gaussian distribution $\mathcal{N}(m(X), \mathbb{V}(X))$, where

$$m(X) : \text{Mean of } X, \quad (4.11)$$

$$\mathbb{V}(X) : \text{Variance of } X. \quad (4.12)$$

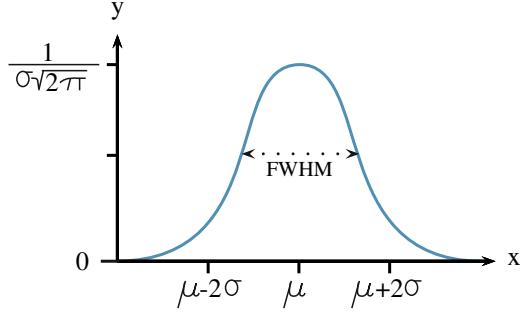


Figure 4.2: A Gaussian probability distribution. The maximum is at the mean value μ , with a full width at half maximum (FWHM) at around 2.35σ . Figure from [20].

Variance

Now that some basics of Bayesian statistics have been covered, a more formal definition of the variance of a distribution will be given. The variance $\mathbb{V}(X)$ is defined as the expectation value of the square of deviations from the mean. The expectation value of X following a probability distribution $P(X|\mathcal{D}, I)$ is denoted $\mathbb{E}[X]$, and defined as

$$\mathbb{E}[X] = \int X P(X|\mathcal{D}, I) dX. \quad (4.13)$$

For the posterior probability distribution $P(X|\mathcal{D}, I)$ the variance of X is then given by [20]

$$\mathbb{V}(X) = \mathbb{E}[(X - X_0)^2] = \int (X - X_0)^2 P(X|\mathcal{D}, I) dX. \quad (4.14)$$

The variance in X is often denoted $\sigma_X^2 = \mathbb{V}(X)$, and its square root is the *standard deviation* $\sqrt{\sigma_X^2} = \sigma_X$.

4.1.4 Covariance

In distributions over several quantities of interest X_i , denoted $P(X_i|\mathcal{D}, I)$, varying one quantity X_i can affect the variance of another quantity X_j . This is called *covariance*. For distributions over several quantities, called *multivariate distributions*, the equations are not as simple to solve as in Eq. (4.7). A set of *simultaneous equations* must then be solved to get the best estimate

$$\frac{dP}{dX_i} \Big|_{X_{0j}} = 0, \quad \frac{d^2P}{dX_i^2} \Big|_{X_{0j}} < 0 \quad (4.15)$$

For simplicity the problem is considered in two dimensions, $\{X_i\} = (X, Y)$. Analogously to Eq. (4.7), the Taylor expansion of $L = \log_e [P(X, Y | \mathcal{D}, I)]$ is found

$$\begin{aligned} L = & L(X_0, Y_0) + \frac{1}{2} \left[\frac{d^2 L}{dX^2} \Big|_{X_0, Y_0} (X - X_0)^2 \right. \\ & \left. + \frac{d^2 L}{dY^2} \Big|_{X_0, Y_0} (Y - Y_0)^2 + 2 \frac{d^2 L}{dXdY} \Big|_{X_0, Y_0} (X - X_0)(Y - Y_0) \right] + \dots \quad (4.16) \end{aligned}$$

There are now four partial derivatives, reduced to three using the rules for mixed partial derivatives $\frac{\partial^2}{\partial X \partial Y} = \frac{\partial^2}{\partial Y \partial X}$. Writing the quadratic terms of Eq. (4.16) in matrix form gives

$$Q = (X - X_0 \quad Y - Y_0) \begin{pmatrix} A & C \\ C & B \end{pmatrix} \begin{pmatrix} X - X_0 \\ Y - Y_0 \end{pmatrix}, \quad (4.17)$$

where the matrix elements are

$$A = \frac{\partial^2 L}{\partial X^2} \Big|_{X_0, Y_0}, \quad B = \frac{\partial^2 L}{\partial Y^2} \Big|_{X_0, Y_0}, \quad C = \frac{\partial^2 L}{\partial X \partial Y} \Big|_{X_0, Y_0}. \quad (4.18)$$

The expression for the variance of X given the distribution $P(X, Y | \mathcal{D}, I)$ is very similar to Eq. (4.14), except for an additional integral over Y

$$\mathbb{V}(X) = \sigma_X^2 = \mathbb{E}[(X - X_0)^2] = \int \int (X - X_0)^2 P(X, Y | \mathcal{D}, I) dX dY. \quad (4.19)$$

A similar expression, σ_Y^2 , can be found for Y by substituting X and Y .

The simultaneous deviations of X and Y constitute the aforementioned covariance, often denoted σ_{XY}^2 . In two dimensions the covariance is given by

$$\sigma_{XY}^2 = \mathbb{E}[(X - X_0)(Y - Y_0)] = \int \int (X - X_0)(Y - Y_0) P(X, Y | \mathcal{D}, I) dX dY. \quad (4.20)$$

The covariance indicates how an over- or underestimation of one quantity affects another. If, for example, an overestimation of X leads to an overestimation of Y , the covariance is positive. An example of positive covariance is shown in the third panel of Fig. 4.3. If the overestimation of X has little or no effect on the estimation of Y , the covariance is negligible or zero $|\sigma_{XY}| \ll \sqrt{\sigma_X^2 \sigma_Y^2}$, as seen in the first panel of Fig. 4.3. The second panel shows negative covariance.

Covariance Matrix

The variances and covariances are the elements of the *covariance matrix*. For N quantities of interest X_1, \dots, X_N the covariance matrix is an $N \times N$ -matrix. The

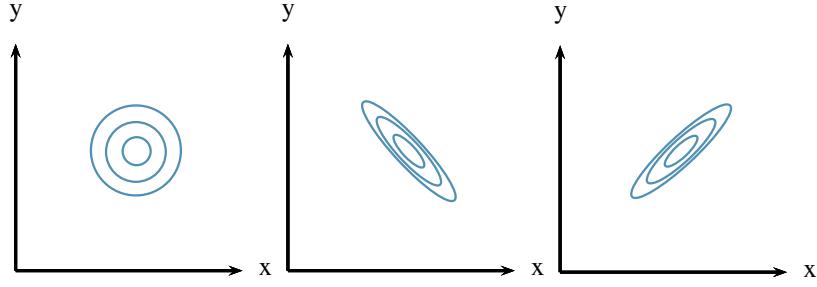


Figure 4.3: A schematic illustration of covariance and correlation. (a) The contours of a posterior pdf with zero covariance, where the inferred values of X and Y are uncorrelated. (b) The corresponding plot when the covariance is large and negative; (c) The case of positive correlation.

covariance matrix for X and Y is denoted $\text{cov}(X, Y)$, and it can be shown that [20]

$$\text{cov}(X, Y) = \begin{pmatrix} \sigma_X^2 & \sigma_{XY}^2 \\ \sigma_{XY}^2 & \sigma_Y^2 \end{pmatrix} = - \begin{pmatrix} A & C \\ C & B \end{pmatrix}^{-1}, \quad (4.21)$$

where A , B and C are the covariances defined in Eq. (4.18).

In short, the posterior probability distribution over X , $P(X|\mathcal{D}, I)$, can be approximated as a Gaussian distribution $\mathcal{N}(m(X), \mathbb{V}(X))$, where the mean is the best estimate and the variance is the associated reliability. The Gaussian distribution is defined by the mean value $m(X)$ and the variance $\mathbb{V}(X)$. For multivariate distributions, the covariance $\sigma_{X_i X_j}^2$ can also be calculated for the variables X_i , and all variances and covariances are contained in the covariance matrix $\text{cov}(X_i)$. The elements of a covariance matrix can also be calculated using *covariance functions*.

4.2 Covariance Functions

The elements of a covariance matrix can be determined by *covariance functions*, or *kernels*. A function that maps two arguments $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ into \mathbb{R} , where \mathcal{X} is the input space, is generally called a kernel k . Covariance functions are symmetric kernels, meaning that $k(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}', \mathbf{x})$. In this project both terms, kernel and covariance function, refer to the covariance function.

A kernel function that only depends on the difference between two points, $\mathbf{x} - \mathbf{x}'$, is called *stationary*, i.e. invariant to translations in input space. If, in addition, it only depends on the length $r = |\mathbf{x} - \mathbf{x}'|$, the function is *isotropic*².

²Invariant to rigid rotations in input space.

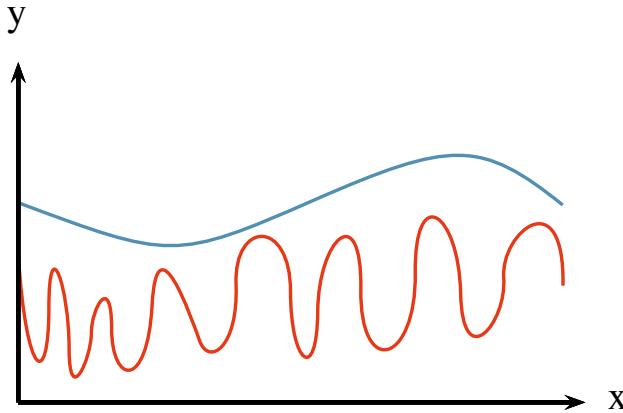


Figure 4.4: The effect of varying the length scale ℓ . A long length scale (blue) gives a smooth, slowly varying function, while a short length scale (red) gives a more staccato, quickly varying function.

Isotropic functions are commonly referred to as *radial basis functions*, as they are only functions of r .

The covariance function can also depend on the dot product, $\mathbf{x} \cdot \mathbf{x}'$, and is then called a *dot product* covariance function. The most important covariance functions for this project are the *squared exponential covariance function* and the *Matérn class of covariance functions*.

4.2.1 The Squared Exponential Covariance Function

The *squared exponential covariance function* (SE) has the form

$$k_{\text{SE}}(r) = \exp\left(-\frac{r^2}{2\ell^2}\right), \quad (4.22)$$

where ℓ is the *characteristic length scale*. Equation (4.22) is sometimes also called the *radial basis function* (RBF). The length scale, ℓ , determines the smoothness of the function. It can be loosely interpreted as how far you need to move (along a particular axis) in input space for the function values to become uncorrelated. For a large length scale one should expect a very slowly varying function, while a shorter length scale means a more rapidly varying function, see the illustration in Fig. 4.4. The SE is infinitely differentiable and therefore very smooth. Stein [22] argues that such strong smoothness assumptions are unrealistic for most physical problems, and recommends another class of kernels, namely the *Matérn class of covariance functions*.

4.2.2 The Matérn Class of Covariance Functions

The *Matérn class of covariance functions* is given by

$$k_{\text{Matérn}}(r) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}r}{\ell} \right)^\nu K_\nu \left(\frac{\sqrt{2\nu}r}{\ell} \right), \quad (4.23)$$

where $\nu, \ell > 0$, and K_ν is a modified Bessel function [23]. The hyperparameter ν controls the smoothness of the function. For $\nu \rightarrow \infty$ this becomes the SE kernel, and for $\nu = 1/2$ it becomes the very rough absolute exponential kernel $k(r) = \exp(-r/\ell)$. In the case of half integer ν , $\nu = p + \frac{1}{2}$ for $p \in \mathbb{N}$, the covariance function is simply the product of an exponential and a polynomial

$$k_{\nu=p+\frac{1}{2}} = \exp \left(-\frac{\sqrt{2\nu}r}{\ell} \right) \frac{\Gamma(p+1)}{\Gamma(2p+1)} \sum_{i=0}^p \frac{(p+i)!}{i!(p-i)!} \left(\frac{\sqrt{8\nu}r}{\ell} \right)^{p-i}. \quad (4.24)$$

In machine learning the two most common cases are for $\nu = 3/2$ and $\nu = 5/2$

$$k_{\nu=3/2}(r) = \left(1 + \frac{\sqrt{3}r}{\ell} \right) \exp \left(-\frac{\sqrt{3}r}{\ell} \right), \quad (4.25)$$

$$k_{\nu=5/2}(r) = \left(1 + \frac{\sqrt{5}r}{\ell} + \frac{5r^2}{3\ell^2} \right) \exp \left(-\frac{\sqrt{5}r}{\ell} \right). \quad (4.26)$$

4.2.3 Noise

The covariance function can also contain information about noise in the data. If the noise follows a Gaussian distribution $\varepsilon \sim \mathcal{N}(0, \sigma_n^2)$, the noise is represented by adding the variance, σ_n^2 , to the diagonal of the covariance matrix

$$k(\mathbf{x}_i, \mathbf{x}_j)_{\text{noise}} = \sigma_n^2 \delta_{ij}, \quad (4.27)$$

where δ_{ij} is the Kronecker delta. The model where the noise is assumed to follow a Gaussian distribution is called the *Gaussian noise model*, and is discussed in Sec. 4.3.1.

4.2.4 Hyperparameters

The RBF kernel in Eq. (4.22) and the Matérn kernel in Eq. (4.23) are both isotropic, *i.e.* they are functions only of the distance between input points, r . Isotropic kernels can, however, be generalized to an anisotropic form by setting

$$r^2(\mathbf{x}, \mathbf{x}') = (\mathbf{x} - \mathbf{x}')^T M (\mathbf{x} - \mathbf{x}'), \quad (4.28)$$

where M is some positive definite matrix. The matrix M can take multiple forms, such as

$$M_1 = \ell^{-2} \mathbb{I}, \quad M_2 = \text{diag}(\boldsymbol{\ell})^{-2}, \quad (4.29)$$

where ℓ^2 is a scalar, and ℓ is a vector of the same dimension as the input vectors \mathbf{x} , \mathbf{x}' . Choosing ℓ to be a vector instead of a scalar is in many cases useful, especially if the input vector \mathbf{x} contains values of different scales.

The matrix M is now part of the *hyperparameters* of the kernel. Each kernel has a vector of hyperparameters, denoted $\boldsymbol{\theta}$. An example is the following generalisation of the RBF kernel,

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sigma_f^2 \exp\left(-\frac{1}{2}(\mathbf{x}_i - \mathbf{x}_j)^T M (\mathbf{x}_i - \mathbf{x}_j)\right) + \sigma_n^2 \delta_{ij}, \quad (4.30)$$

where σ_f is a scaling factor and σ_n^2 is the Gaussian noise parameter, discussed in Sec. 4.2.3. The hyperparameters of this kernel are $\boldsymbol{\theta} = (\{M\}, \sigma_f^2, \sigma_n^2)^T$, where $\{M\}$ denotes the parameters in the symmetric matrix M .

Other Covariance Functions

There are several other types of covariance functions that are not discussed here. In addition, kernels can be multiplied and summed to form new kernels, making the space of possible kernels infinite. For further details see Chapter 4 in [24].

4.3 Gaussian Process Regression

Gaussian processes (GP) is a supervised machine learning method, designed to solve regression and probabilistic classification problems. Only regression is explored in this project. This section begins by introducing Gaussian processes and important notation, first in a general sense, and then in the function space view. Then distributions over functions are considered, followed by a short discussion on how functions can be drawn from these distributions. Finally, a quick overview of the noise-free model and the Gaussian noise model, and their covariance matrices, are given.

Gaussian Processes

Consider a set of points $\mathcal{D} = \{\mathbf{x}_i, y_i\}$, where y is some (possibly noisy) function of \mathbf{x} , $y(\mathbf{x}) = f(\mathbf{x}) + \varepsilon$. Here, ε is the noise and $f(\mathbf{x})$ is the true value of the function. These points are illustrated by the black dots in Fig. 4.5. In machine learning \mathcal{D} is the *training data*, as it is used to train the model. It consists of *features*, which are the components of the input vectors \mathbf{x}_i , and *targets*, which are the function values y_i . The set of points is discrete, so there is some \mathbf{x}^* for which the target y^* is unknown. The test point \mathbf{x}^* is marked on the x -axis in Fig. 4.5.

Gaussian processes (GP) predict a Gaussian distribution *over function values* at this point \mathbf{x}^* . The distribution for a single test point \mathbf{x}^* has a corresponding mean $m(\mathbf{x}^*)$ and variance $\mathbb{V}(\mathbf{x}^*)$. Note that the mean $m(\mathbf{x}^*)$ is not the mean of

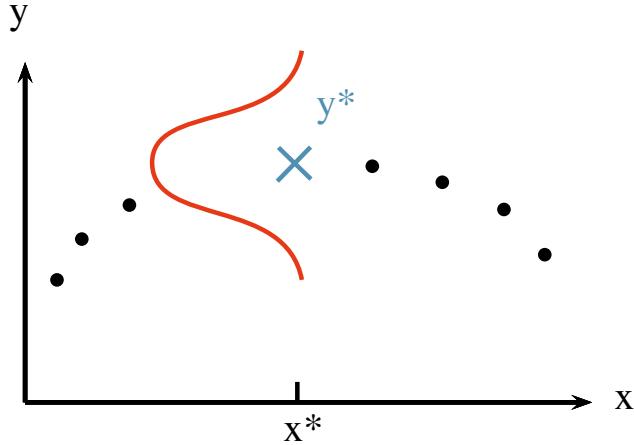


Figure 4.5: An illustration of a GP prediction of the target value y^* (blue cross), given the known set of points $\{x_i, y_i\}$ (black dots). The prediction is a Gaussian distribution in y with mean y^* and variance σ^2 . The Gaussian distribution is drawn in red with y on the vertical axis, with uncertainty in the y -direction.

the input vector \mathbf{x}^* , but rather *the mean of function values* $f(\mathbf{x})$ evaluated at \mathbf{x}^* . The GP prediction for the target value $y^* = f(\mathbf{x}^*)$ is the mean $m(\mathbf{x}^*)$. Similarly, the variance, $V(\mathbf{x}^*)$, is in fact the variance in function values $f(\mathbf{x}^*)$, or the width of the Gaussian distribution (red line) in the y -direction in Fig. 4.5. The mean value y^* is drawn as a blue cross in Fig. 4.5. As will be shown, the predicted target y^* is a linear combination of the known targets y_i , where the weights are controlled by the covariances between training points \mathbf{x}_i and the test point \mathbf{x}^* .

Some Notation

As Gaussian process regression is notoriously confusing, it is helpful to begin with some notation.

As discussed in Sec. 4.3, the Gaussian distribution with a mean μ and variance σ^2 is denoted $\mathcal{N}(\mu, \sigma^2)$. The distribution can be over a single random variable, *e.g.* f , or a finite set of random variables, f_i . A Gaussian distribution over several random variables is called a *multivariate Gaussian distribution*. The n random variables f_i , $i = 1, \dots, n$ drawn from a multivariate Gaussian distribution make up the n -dimensional vector \mathbf{f} . The mean values, μ_i , are then contained in the n -dimensional *mean vector* $m(\mathbf{f})$. The variance, σ^2 , is replaced by an $n \times n$ -dimensional covariance matrix, $\text{cov}(\mathbf{f})$. The multivariate Gaussian distribution over \mathbf{f} is written as

$$\mathbf{f} \sim \mathcal{N}(m(\mathbf{f}), \text{cov}(\mathbf{f})). \quad (4.31)$$

For n points $\{\mathbf{x}_i, y_i\}$, where \mathbf{x}_i is an m -dimensional feature vector and y_i is the target, the features comprise the $n \times m$ -matrix X . The targets make up the corresponding n -dimensional vector \mathbf{y} . A central assumption in Gaussian processes is that the covariance between targets y_i, y_j is given by the covariance of their features $\mathbf{x}_i, \mathbf{x}_j$

$$\text{cov}(y_i, y_j) = k(\mathbf{x}_i, \mathbf{x}_j), \quad (4.32)$$

where $k(\mathbf{x}_i, \mathbf{x}_j)$ is a covariance function, as described in Sec. 4.2. In Gaussian processes the distribution over target values \mathbf{y} with feature matrix X will therefore be written

$$\mathbf{y} \sim \mathcal{N}(m(\mathbf{y}), K(X, X)), \quad (4.33)$$

where $K(X, X)$ is the covariance matrix containing the covariances of the features in X , calculated using the covariance function $k(\mathbf{x}, \mathbf{x}')$.

Finally, a Gaussian process will be denoted \mathcal{GP} . It may be difficult to distinguish between a Gaussian *distribution*, \mathcal{N} , and a Gaussian *process*, \mathcal{GP} . The difference can be thought of as the difference between a finite collection of function values $f_i = f(\mathbf{x}_i)$, and the continuous function, $f(\mathbf{x})$. The former can be viewed as a vector \mathbf{f} , and can be drawn from a distribution such as the one in Eq. (4.33), $\mathbf{f} \sim \mathcal{N}(m(\mathbf{f}), K(X, X))$. The latter is a *function*, drawn from a distribution *over functions*, where the mean $m(\mathbf{x})$ and covariances $k(\mathbf{x}, \mathbf{x}')$ are functions as well. A function $f(\mathbf{x})$ drawn from a Gaussian process is written as

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')). \quad (4.34)$$

Function Space View

Gaussian processes are distributions over functions. It is therefore useful to consider the problem in the function space view introduced in [24]. For a function $f(\mathbf{x})$ the Gaussian process *mean function*, $m(\mathbf{x})$, and *covariance function*, $k(\mathbf{x}, \mathbf{x}')$, are defined as

$$m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})], \quad (4.35)$$

$$k(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))]. \quad (4.36)$$

In other words, the mean $m(\mathbf{x})$ is the expected value of the function $f(\mathbf{x})$ at \mathbf{x} , and the covariance function is the expected value of the simultaneous deviations of $f(\mathbf{x})$ from the mean $m(\mathbf{x})$ at \mathbf{x} , and $f(\mathbf{x}')$ from the mean $m(\mathbf{x}')$ at \mathbf{x}' . As mentioned, the mean and covariance are now *functions* of the input vector \mathbf{x} . This means that for every input vector \mathbf{x} , there is a Gaussian distribution over function values with a mean $m(\mathbf{x})$ and covariance with the function value at \mathbf{x} given by $k(\mathbf{x}, \mathbf{x}')$. This is a generalization of the single test point in Fig. 4.5, where every point \mathbf{x}^* gets a similar distribution.

The collection of Gaussian distributions over functions that are now a function of the input vector \mathbf{x} are the Gaussian processes, \mathcal{GP} . In the same way that a random variable is drawn from a distribution, random functions can be drawn from the \mathcal{GP} ,

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')). \quad (4.37)$$

How functions are drawn from distributions is discussed in more detail in a later section.

The covariance between two points $f(\mathbf{x})$ and $f(\mathbf{x}')$ is determined by the covariance function $k(\mathbf{x}, \mathbf{x}')$, as discussed in Sec. 4.2. As mentioned, the covariance function calculates the covariance between the *function values*, $f(\mathbf{x}_i)$ and $f(\mathbf{x}_j)$, and *not* the input vectors, \mathbf{x}_i and \mathbf{x}_j . Rather, the covariance between two function values is a function of the input vectors. Consider again the illustration in Fig. 4.4. The variation in function values $f(\mathbf{x}_i)$ and $f(\mathbf{x}_j)$ depends on the distance between the points \mathbf{x}_i and \mathbf{x}_j , and the characteristic length scale of the process. If the length scale is large, the two input vectors \mathbf{x}_i and \mathbf{x}_j can be far away, and still have similar function values $f(\mathbf{x}_i)$ and $f(\mathbf{x}_j)$. For short length scales, however, nearby points can have very different function values, because the function varies rapidly.

4.3.1 The Prior and Posterior Distribution

Gaussian processes are Bayesian in that there is a prior and a posterior distribution over functions, where the posterior is obtained by conditioning the prior on the training data. Consider the $n^* \times m$ -matrix of test points X^* , containing n^* test points \mathbf{x}_i^* , with unknown function values $f(\mathbf{x}^*)$. Using the kernel function, $k(\mathbf{x}_i^*, \mathbf{x}_j^*)$, on the matrix X^* gives the covariance matrix, $K(X^*, X^*)$, as discussed in Sec. 4.2. The covariance matrix now contains the covariance of all test points \mathbf{x}_i^* . Combined with an initial mean of zero³ one obtains the *prior* distribution of predicted target values \mathbf{f}^*

$$\mathbf{f}^* \sim \mathcal{N}(\mathbf{0}, K(X^*, X^*)). \quad (4.38)$$

This distribution contains the prior assumptions about the function values $f(\mathbf{x})$, in that the smoothness of the function and the correlation between function values are encoded in the covariance matrix. This is the prior probability distribution discussed in Sec. 4.1.2, that will be modified by the data to provide the posterior probability distribution. The choice of kernel is therefore one of the most important steps in learning with GPs.

³The mean does not have to be zero, it could for example be the mean of the training data.

Noise-Free Model

Now consider the addition of training data to the prior distribution in Eq. 4.38, in the form of a noise-free set of n training points $\{\mathbf{x}_i, y_i\}$, so that $y = f(\mathbf{x})$. The input vectors \mathbf{x}_i form the $n \times m$ -matrix X , where the rows are the input vectors. The training targets y_i form the corresponding n -dimensional vector \mathbf{f} . The test points are still contained in the $n^* \times m$ matrix X^* . The goal is to predict an n^* -dimensional vector \mathbf{f}^* containing the predictions of the function values at the points \mathbf{x}_i^* , which is conditioned on the known function values \mathbf{f} .

The joint distribution of training outputs, \mathbf{f} , and test outputs, \mathbf{f}^* , according to the prior in Eq. 4.38 is

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}^* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(X, X) & K(X, X^*) \\ K(X, X^*) & K(X^*, X^*) \end{bmatrix}\right), \quad (4.39)$$

where, as before, $K(X_i, X_j')$ is the covariance matrix between the sets of points $\{\mathbf{x}_i\}$ and $\{\mathbf{x}'_j\}$ calculated using the covariance function $k(\mathbf{x}_i, \mathbf{x}'_j)$. By conditioning the distribution of \mathbf{f}^* on the observations \mathbf{f} , the posterior distribution over \mathbf{f}^* is obtained⁴ [24]

$$\mathbf{f}^* | X^*, X, \mathbf{f} \sim \mathcal{N}(K(X^*, X)K(X, X)^{-1}\mathbf{f}, \quad (4.40)$$

$$K(X^*, X^*) - K(X^*, X)K(X, X)^{-1}K(X, X^*)), \quad (4.41)$$

where $\mathbf{f}^* | X^*, X, \mathbf{f}$ means the target values \mathbf{f}^* for the features X^* , given the known features X and targets \mathbf{f} . The prior in Eq. 4.38 has been modified by the data, X and \mathbf{f} , to give the posterior in Eq. (4.40), leaving two distributions from which function values can be drawn.

Drawing Samples

All functions $f(\mathbf{x})$ that are drawn from the Gaussian process \mathcal{GP} must have a covariance between function values $f(\mathbf{x}), f(\mathbf{x}')$ determined by the covariance function $k(\mathbf{x}, \mathbf{x}')$. If for example the covariance function has a large length scale, quickly varying functions are not allowed. The mean of the distribution, $m(\mathbf{x})$, is *not* the mean value of each function $f(\mathbf{x})$ drawn from \mathcal{GP} , but rather the mean function value you would get in \mathbf{x} if you drew enough functions from the \mathcal{GP} . Drawing functions from a distribution in this way will be referred to as *drawing samples*.

To generate samples $\mathbf{f} \sim \mathcal{N}(\mathbf{m}, K)$ with mean \mathbf{m} and covariance matrix K using a scalar Gaussian generator⁵, one proceeds as follows: first the Cholesky

⁴For more details, see Appendix A.2 in [24].

⁵A scalar Gaussian generator generates random numbers from a Gaussian distributions, and can be found in most programming environments, such as the `random` environment in Python.

decomposition L — also known as the matrix square root — of the covariance matrix is found using $K = LL^T$, where L is a lower triangular matrix. A vector \mathbf{u} is then generated by multiple calls to the scalar Gaussian generator $\mathbf{u} \sim \mathcal{N}(0, I)$. Then $\mathbf{f} = \mathbf{m} + L\mathbf{u}$ has the desired distribution with mean \mathbf{m} and covariance $L\mathbb{E}[\mathbf{u}\mathbf{u}^T]L^T = LL^T = K$, as described in [24].

Using the method described above, random functions are drawn from the prior and posterior in Eq. (4.38) and Eq. (4.40), respectively, and shown in Fig. 4.6. As discussed, the samples drawn from the prior have mean equal to zero $m(\mathbf{x}) = \mathbf{0}$, and constant covariance, $K(X^*, X^*)$, meaning that if you drew enough functions the mean of all function values at every \mathbf{x} would be zero. The prior is shown in the upper panel of Fig. 4.6, where the mean of the distribution is represented by the thick, black line, and the covariance is the light blue band around the mean. In the posterior distribution the mean values and covariances have been modified by the training data, represented by red dots in the lower panel of Fig. 4.6. In a point where there is training data the uncertainty is zero⁶, and so all samples drawn from the posterior distribution must pass through this point. Far away from training points the covariance is large. The mean has also been modified to pass through the training points, as seen by the thick black line in the lower panel.

Gaussian Noise Model

Noise-free observations are rare. In most cases targets will contain some noise $y = f(\mathbf{x}) + \varepsilon$, where the noise ε is assumed to follow a Gaussian distribution $\varepsilon \sim \mathcal{N}(0, \sigma_n^2)$. This is the *Gaussian noise model*. As discussed in Sec. 4.2 the covariance of a function with Gaussian noise can be expressed as

$$\text{cov}(y_i, y_j) = k(\mathbf{x}_i, \mathbf{x}_j) + \sigma_n^2 \delta_{ij}. \quad (4.42)$$

Training targets are now contained in the n -dimensional vector \mathbf{y} , while training features are contained in the $n \times m$ -matrix X , test features in the $n^* \times m$ -matrix X^* and predicted targets in the n^* -dimensional vector \mathbf{f}^* , as before. With the addition of the noise the prior distribution becomes

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}^* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(X, X) + \sigma_n^2 \mathbb{I} & K(X, X^*) \\ K(X^*, X) & K(X^*, X^*) \end{bmatrix}\right). \quad (4.43)$$

The conditioned distribution is then

$$\mathbf{f}^* | X^*, X, \mathbf{y} \sim \mathcal{N}(m(\mathbf{f}^*), \text{cov}(\mathbf{f}^*)) \quad (4.44)$$

⁶Assuming there is no noise in the data.

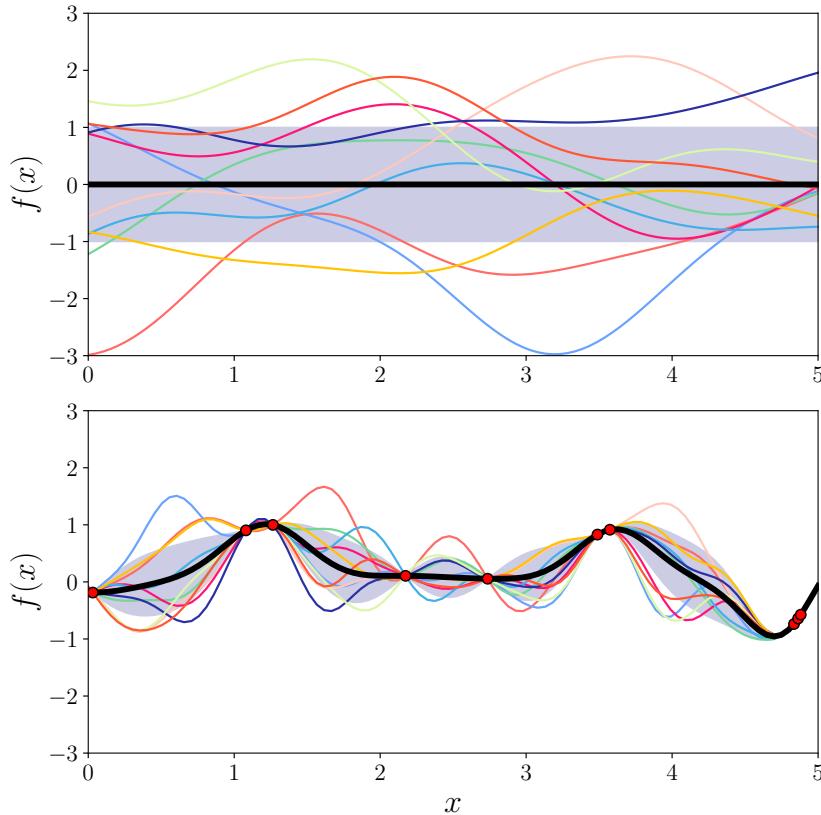


Figure 4.6: Drawing functions from the prior (top) and posterior (bottom) distributions. The thick, black line represents the mean of the distribution, while the shaded, blue area is the variance. The multiple colored lines are functions drawn randomly from the prior and posterior distributions, whose correlation are dictated by the covariance function. The prior has mean 0 and covariance given by the squared exponential function. The posterior has been modified by training points (red dots), giving rise to zero uncertainty at the points where training data exists, and an altered mean value for the distribution. The kernel of the prior distribution has hyperparameters $\sigma_f = 1$ and $\ell = 1$, while for the posterior they are $\sigma_f = 0.594$ and $\ell = 0.279$. Figure generated using scikit-learn.

where

$$m(\mathbf{f}^*) = K(X^*, X)[K(X, X) + \sigma_n^2 \mathbb{I}]^{-1} \mathbf{y}, \quad (4.45)$$

$$\text{cov}(\mathbf{f}^*) = K(X^*, X^*) - K(X^*, X)[K(X, X) + \sigma_n^2 \mathbb{I}]^{-1} K(X, X^*). \quad (4.46)$$

Equations (4.45) – (4.46) are the key predictive equations for Gaussian process regression. The calculation requires the inverting the $n \times n$ -matrix $[K(X, X) + \sigma_n^2 \mathbb{I}]$, which for large n becomes computationally unviable. This is discussed further in Sec. 5.4, where distributed Gaussian processes are proposed as a way of scaling Gaussian processes to larger training sets.

To tidy up the expression in Eqs. (4.45) – (4.46) the matrices $K \equiv K(X, X)$ and $K^* \equiv K(X, X^*)$ are defined. For a single test point \mathbf{x}^* the matrix K^* is written as a vector $\mathbf{k}(X, \mathbf{x}^*) = \mathbf{k}^*$ to denote the covariances between the n training points and the test point \mathbf{x}^* . Using this compact notation the GP prediction of $f^* = f(\mathbf{x}^*)$ for a single test point \mathbf{x}^* is

$$m(f^*) = \mathbf{k}^{*T} (K + \sigma_n^2 \mathbb{I})^{-1} \mathbf{y}, \quad (4.47)$$

$$\mathbb{V}[f^*] = k(\mathbf{x}^*, \mathbf{x}^*) - \mathbf{k}^{*T} (K + \sigma_n^2 \mathbb{I})^{-1} \mathbf{k}^*. \quad (4.48)$$

Note that, as mentioned in the beginning of Sec. 4.3, the predicted mean value $m(f^*)$ can be viewed as a linear combination of y_i of the form $\alpha \mathbf{y}$, where $\alpha = \mathbf{k}^{*T} (K + \sigma_n^2 \mathbb{I})^{-1}$. The vector α then only contains the covariance between features.

Equations (4.47) – (4.48) form the basis for GP prediction in `scikit-learn` [25]. The algorithm for Gaussian process regression on a single test point \mathbf{x}^* , with training data X , \mathbf{y} is outlined in Algorithm 1. The algorithm uses the Cholesky decomposition, L , of the covariance matrix to find the weights $\boldsymbol{\alpha}$ used to calculate the predictive mean f^* . The variance $\mathbb{V}[\mathbf{x}^*]$ is calculated using L and the covariance of the test point with the training points, $\mathbf{k}^* = k(X, \mathbf{x}^*)$.

Data: X (inputs), \mathbf{y} (targets), k (covariance function/kernel), σ_n^2 (noise level), \mathbf{x}_* (test input).

$L = \text{Cholesky decomposition } (K + \sigma_n^2 I) ;$

$\boldsymbol{\alpha} = (L^T)^{-1} (L^{-1} \mathbf{y}) ;$

$\bar{f}_* = \mathbf{k}_*^T \boldsymbol{\alpha} ;$

$\mathbf{v} = L^{-1} \mathbf{k}_* ;$

$\mathbb{V}[f_*] = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{v}^T \mathbf{v} ;$

$\log p(\mathbf{y}|X) = -\frac{1}{2} \mathbf{y}^T \boldsymbol{\alpha} - \sum_i \log L_{ii} - \frac{n}{2} \log 2\pi ;$

Result: f_* (mean), $\mathbb{V}[f_*]$ (variance), $\log p(\mathbf{y}|X)$ (log marginal likelihood).

Algorithm 1: Algorithm 2.1 from [24].

4.4 Model Selection

Choosing the right kernel and hyperparameters, introduced in Sec. 4.2, is an important part of Gaussian process regression. Finding the kernel and corresponding hyperparameters that best fit the data is often called *model selection*. Model selection is also referred to as *training* in machine learning. In this section Bayesian model selection for the marginal likelihood is quickly overviewed, and the log marginal likelihood and cross validation are introduced as tools to optimize Gaussian process estimators.

4.4.1 Log Marginal Likelihood

The *marginal likelihood* can be used to find the optimal hyperparameters of covariance functions. Gaussian process regression models with Gaussian noise have the wonderful trait of analytically tractable integrals for the marginal likelihood, $P(\mathbf{y}|X, \boldsymbol{\theta})$. Note that the marginal likelihood defined here looks different than in Eq. (4.4) — $\boldsymbol{\theta}$ are the hyperparameters of the covariance functions introduced in Sec. 4.2.4, \mathbf{y} are the training outputs and X are the training inputs. The parameters in Eq. (4.4), denoted Θ , are now the true — or *latent* — function values, \mathbf{f} where $\mathbf{y} = \mathbf{f} + \varepsilon$. The marginal likelihood is the integral of the likelihood times the prior over the latent function values \mathbf{f} , given by

$$P(\mathbf{y}|X, \boldsymbol{\theta}) = \int P(\mathbf{y}|\mathbf{f}, X, \boldsymbol{\theta})P(\mathbf{f}|X, \boldsymbol{\theta})d\mathbf{f}, \quad (4.49)$$

where $P(\mathbf{f}|X, \boldsymbol{\theta})$ is the prior and $P(\mathbf{y}|\mathbf{f}, X, \boldsymbol{\theta})$ is the likelihood. Under the Gaussian process model the prior is a Gaussian, $\mathbf{y} \sim \mathcal{N}(\mathbf{0}, K + \sigma_n^2 \mathbb{I})$, so the logarithm of the marginal likelihood in Eq. (4.49) gives for the *log marginal likelihood* (LML) [24]

$$\log P(\mathbf{y}|X, \boldsymbol{\theta}) = -\frac{1}{2}\mathbf{y}^T(K + \sigma_n^2 \mathbb{I})^{-1}\mathbf{y} - \frac{1}{2}\log|K + \sigma_n^2 \mathbb{I}| - \frac{n}{2}\log 2\pi, \quad (4.50)$$

where n is the number of training points. Each term in Eq. (4.50) has an interpretation: $-\frac{1}{2}\mathbf{y}^T(K + \sigma_n^2 \mathbb{I})^{-1}\mathbf{y}$ is the only term involving the data, and is therefore the data-fit; $-\frac{1}{2}\log|K + \sigma_n^2 \mathbb{I}|$ is the complexity penalty depending only on the covariance function and the inputs; and $-\frac{n}{2}\log 2\pi$ is a normalization term.

As mentioned, the goal is to use the marginal likelihood to determine the optimal hyperparameters of the covariance function, $\boldsymbol{\theta}$. Note that the marginal likelihood in Eq. (4.49) is the probability of the training outputs, \mathbf{y} , given the training inputs X and the hyperparameters $\boldsymbol{\theta}$. Maximizing the log marginal likelihood for the hyperparameters will therefore give the best estimate for $\boldsymbol{\theta}$, as per the discussion in Sec. 4.1.3. Maximizing the LML requires finding the partial

derivatives

$$\frac{\partial}{\partial \theta_j} \log p(\mathbf{y}|X, \boldsymbol{\theta}) = \frac{1}{2} \mathbf{y}^T K^{-1} \frac{\partial K}{\partial \theta_j} K^{-1} \mathbf{y} - \frac{1}{2} \text{tr}(K^{-1} \frac{\partial K}{\partial \theta_j}). \quad (4.51)$$

Using partial derivatives, or the gradients, to find the optimal hyperparameters is called a *gradient based optimizer*. Computing the inverse of a matrix, K^{-1} , is computationally complex, and for n training points goes as $\mathcal{O}(n^3)$. Once this is done, however, finding the partial derivatives only requires complexity $\mathcal{O}(n^2)$, and so gradient based optimizers are advantageous.

The LML can have several local optima, as seen in the contour plot of the log marginal likelihood in Fig. 4.7. These correspond to different interpretations of the data. The leftmost optimum in Fig. 4.7, for example, favors a small length scale and smaller noise level. This means that little of the data is considered to be noise. The rightmost optimum has a higher noise level, and allows for a range of length scales, as most of the data is considered to be noise. Features with very large length scales are considered superfluous, as the function value depends little on them. To avoid ending up in a local optima, it can be wise to restart the optimizer a few times during learning.

4.4.2 Cross Validation

Cross validation is a means of monitoring the performance of a model. In k -fold validation this is done by dividing the data into k subsets and using $k-1$ folds to train the model, and a single fold to validate it. This is repeated k times, one for each possible validation set. Cross-validation requires a scoring function, such as the R^2 score. The R^2 -score is given by

$$R^2 = 1 - \frac{\sum_{i=0}^{N-1} (y_i - \hat{y}_i)^2}{\sum_{i=0}^{N-1} (y_i - \bar{y})^2}, \quad (4.52)$$

where \hat{y}_i is the predicted value of the i th sample, y_i is the true value and $\bar{y} = \frac{1}{N} \sum_{i=0}^{N-1} y_i$ for N samples. This is the score used for cross validation in this project.

Cross-validation can be used to plot *learning curves*, which are used to find out whether the estimator benefits from adding more data. The learning curve plots the *training score* and *validation score*. The training score is the mean R^2 -score, $m(R^2)$, for the k predictions of the training points, *i.e.* where the $k-1$ folds used for training are also used as test points. The validation score is the mean R^2 score, $m(R^2)$, for the k predictions of validation points, *i.e.* the one fold that is not a part of the training data. The variances of the R^2 -scores, $\mathbb{V}(R^2)$, are plotted as error bands. The test- and validation scores are used to find out if the model is *overfitting* or *underfitting*. *Overfitting* means that the model is a perfect fit to the training data, but predicts poorly for test data because it

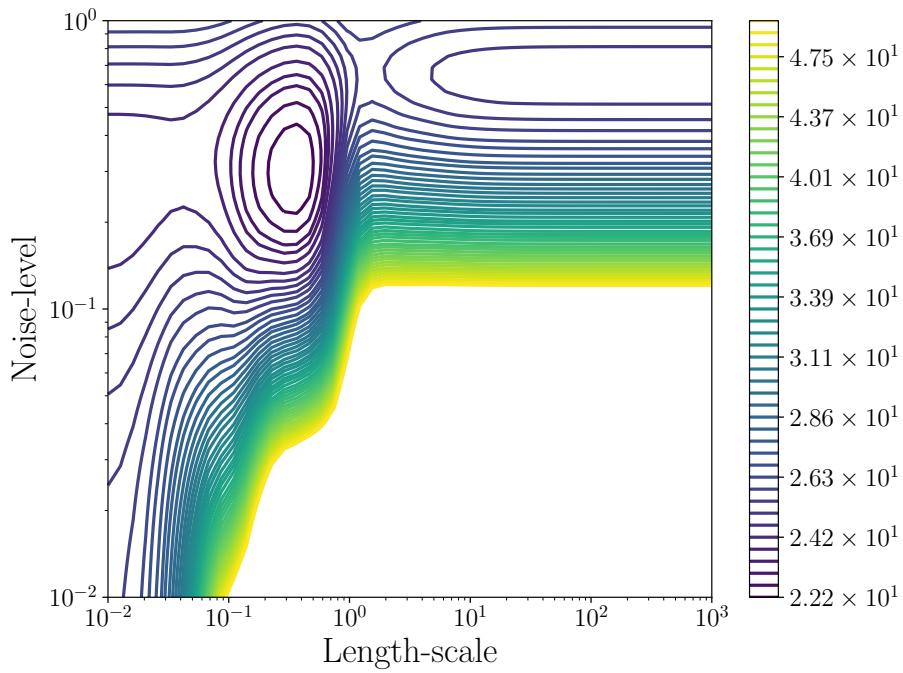


Figure 4.7: A contour plot of the log marginal likelihood with two local optima for a Gaussian process with kernel $k(\mathbf{x}_i, \mathbf{x}_j) = \sigma_f^2 \exp(-(\mathbf{x}_i - \mathbf{x}_j)^2 / \ell^2) + \sigma_n^2 \delta_{ij}$. The rightmost optima favours a short length scale and low noise, with $\sigma_f = 0.64$, $\ell = 0.365$ and $\sigma_n^2 = 0.29$, while the leftmost favors a high noise level and therefore several large length scales, with $\sigma_f = 0.00316$, $\ell = 109$ and $\sigma_n^2 = 0.6$. The optimum to the right has LML -21.8 and the optimum to the right has LML -23.87 . Plots were generated using scikit-learn.

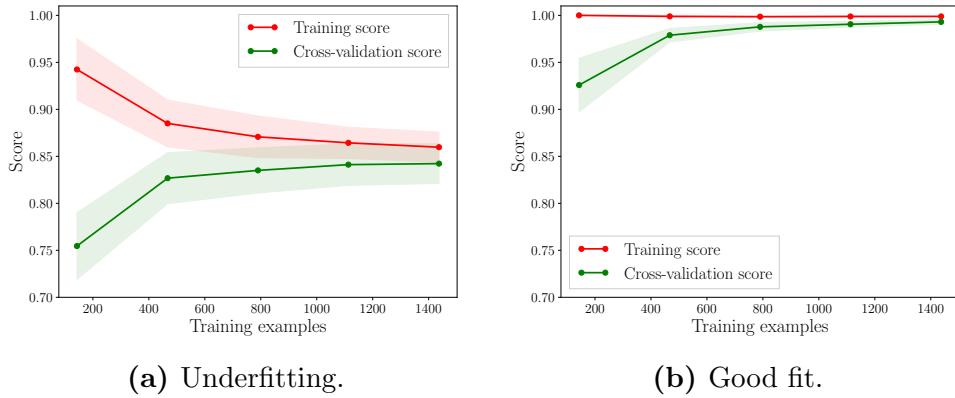


Figure 4.8: Learning curves for two different estimators. The training scores are shown as red lines, with uncertainty bands in light red. The validation scores are shown as green lines, with uncertainty bands in light green. The estimator in (a) is underfitting, as both the training and validation tend to a value less than one. The estimator in (b) is a good fit, and could benefit from more data.

is not general. *Underfitting* occurs when the model is not able to capture the underlying structure of the data. Ideally, the training score should be 1, and the validation score should approach 1 with the addition of data.

Examples of learning curves are shown in Fig. 4.8 as a function of training points, for two machine learning techniques called Naive Bayes and Support Vector Machine on an example problem. In (a) both the training score and cross-validation score tend to a value below 1, which indicates underfitting. This model will not benefit from more data. The example in (b) shows a training score of approximately 1, and a cross validation score that converges towards 1. This model could benefit from more data.

4.4.3 Relative Deviance

In this project predictions are compared using the *relative deviance*. For true values y_i and values predicted by the estimator \hat{y}_i the relative deviance is given by

$$\varepsilon_i = \frac{y_i - \hat{y}_i}{y_i}. \quad (4.53)$$

In this project the target values have a very wide span, ranging from about 10^{-30} fb to 10^9 fb. The data is therefore divided into decades, meaning one set contains $\sigma \in [10^i, 10^{i+1}]$, where σ is the cross section. Then a distribution over the relative deviances within each decade is found, with a mean value, $m(\varepsilon_i)$, and standard

deviation, $\text{std}(\varepsilon_i)$. These are plotted as a function of i , and denoted

$$m(\varepsilon_i) = m\left(\frac{y_i - \hat{y}_i}{y_i}\right), \quad (4.54)$$

$$\sigma^2(\varepsilon_i) = \mathbb{V}\left(\frac{y_i - \hat{y}_i}{y_i}\right), \quad (4.55)$$

$$\text{std}(\varepsilon_i) = \sqrt{\sigma^2(\varepsilon)}. \quad (4.56)$$

Chapter 5

Evaluating Cross Sections using Gaussian Processes

This chapter is dedicated to improving the evaluation of cross sections for squark pair production with Gaussian processes. A benchmark Gaussian process estimator is considered, and compared to estimators with possible improvements in the data set, in the kernel and in the features. Then the possibility of scaling Gaussian processes to larger datasets using distributed Gaussian processes is explored.

5.1 Data Generation

In this section the generation of MSSM-24 training and test data is discussed, following closely the discussion in [26].

Sampling of Data

An MPI parallelized Python script generates a sample point in the MSSM-24 parameter space by drawing random values from the distributions in Table 5.1. When a parameter point has been sampled, it is run through the program `softpoint.x` which calculates its supersymmetric spectrum using the `Softsusy 3.6.2`-package [27]. The spectrum is then written to a SLHA-file [28] that is given as input to `Prospino 2.1`, which subsequently calculates the LO and NLO cross sections according to the method outlined in Section 3.4.1, and the results are written to the SLHA-file. The relevant masses and NLO cross sections are later harvested to input files, which are used by the Gaussian processes.

The weak scale MSSM model used in this project, MSSM-24, requires a soft breaking scale m_s for its definition, as discussed in Sec. 2.3.3. This scale is set to $m_s = 1$ TeV. It is worth noting that the parameter space for the squark cross sections is significantly reduced from that of the MSSM-24. The cross

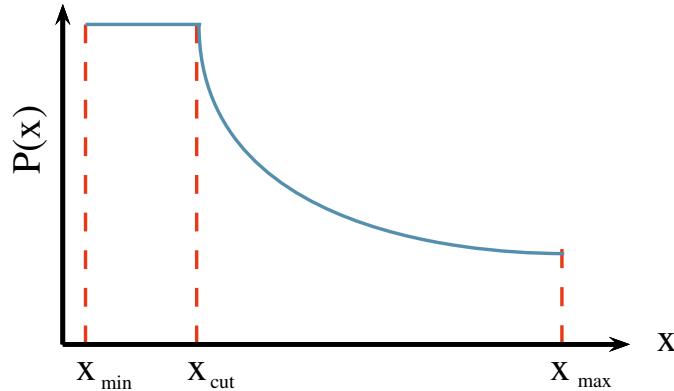


Figure 5.1: Illustration of the log prior distribution $P(x)$. Around $x = 0$ the prior would blow up, so for $x < x_{cut}$ a flat prior is used.

sections depend on the values of the masses $m_{\tilde{q}}$ and $m_{\tilde{g}}$. Since only first and second generation squarks are considered, $m_{\tilde{q}}$ contributes with eight masses¹ which is combined with $m_{\tilde{g}}$ to reduce the parameter space to nine dimensions. The parameter space is therefore better sampled than it would appear from the 24 parameters in MSSM-24.

Priors

To get a reasonable distribution in parameter space a flat and a log prior distribution is used in parameter space. The flat prior covers the edges of parameter space well, while a log prior covers the innermost points². The log prior is flat in log space, so the number of points sampled decrease exponentially as one moves away from zero. FORKLARE MER HER Therefore, a combination of the log and flat priors is used in order to properly cover parameter space. To avoid divergence of the log prior close to zero this region is covered by a flat prior. The limits on the priors are `[start, flat_start, flat_end, end]` for priors that include negative values, and `[flat_start, start, end]` for priors with only positive values. An illustration of a prior with positive values is shown in Fig. 5.1, where a flat distribution is used close to $x = 0$ to avoid divergences.

Data Quality

Data quality plots are used to ensure that the sampled data is properly distributed in parameter space. In Fig. 5.3 scatter plot examples for $m_{\tilde{g}}$, $m_{\tilde{d}_L}$ and $m_{\tilde{u}_L}$ are shown. Mass distributions for the other squark masses are similar, and

¹Four flavours with a pair of lefthanded and righthanded squarks each.

²The points close to zero.

Parameter	Log prior range	Flat prior range
M_1	[0,100,4000]	[0,4000]
M_2	[0,100,4000]	[0,4000]
M_3	[0,100,4000]	[0,4000]
A_t	[-4000, -100, 100, 4000]	[-4000, 4000]
A_b	[-4000, -100, 100, 4000]	[-4000, 4000]
A_τ	[-4000, -100, 100, 4000]	[-4000, 4000]
μ	[-4000, -100, 100, 4000]	[-4000, 4000]
m_A^{pole}	[0,100,4000]	[0,4000]
$\tan \beta$	[2, 60]	[2, 60]
m_{L_1}	[0, 100, 4000]	[0, 4000]
m_{L_2}	[0, 100, 4000]	[0, 4000]
m_{L_3}	[0, 100, 4000]	[0, 4000]
m_{e_1}	[0, 100, 4000]	[0, 4000]
m_{e_2}	[0, 100, 4000]	[0, 4000]
m_{e_3}	[0, 100, 4000]	[0, 4000]
m_{Q_1}	[0, 100, 4000]	[0, 4000]
m_{Q_2}	[0, 100, 4000]	[0, 4000]
m_{Q_3}	[0, 100, 4000]	[0, 4000]
m_{u_1}	[0, 100, 4000]	[0, 4000]
m_{u_2}	[0, 100, 4000]	[0, 4000]
m_{u_3}	[0, 100, 4000]	[0, 4000]
m_{d_1}	[0, 100, 4000]	[0, 4000]
m_{d_2}	[0, 100, 4000]	[0, 4000]
m_{d_3}	[0, 100, 4000]	[0, 4000]

Table 5.1: Table showing the sampling intervals used for the parameters when sampling the MSSM-24, where the soft breaking scale is set to $Q = 1$ TeV. The log priors have three and four limit values, which are of the form `[flat_start, start, end]` and `[start, flat_start, flat_end, end]`. All values in GeV except $\tan \beta$ which is unitless. Table from [26].

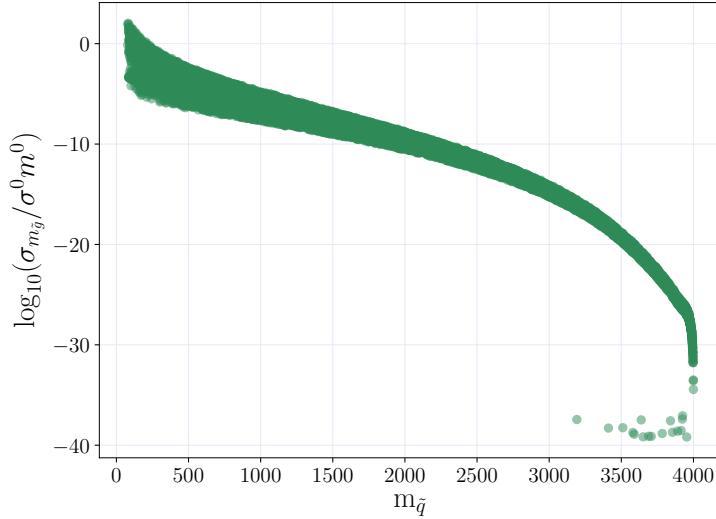


Figure 5.2: Plot of $\log(\sigma_{m_{\tilde{g}}})$ for $\tilde{d}_L \tilde{d}_L$ as a function of $m_{\tilde{d}_L}$, where outliers are included. The outliers are originally $\sigma = 0$ fb, but are set to $\sigma = 10^{-32}$ fb so as to avoid infinities.

can be found in Appendix B. The scatter plots use 40 000 points. All parts of the parameter space seem to be covered, although the density of points is higher for small masses, as can be expected from the log prior. However, in Fig. ?? the number of points with high cross sections is small. The density of points is lower for small squark masses where the gluino mass is large, WHY?.

As discussed in Sec. 2.4.6, the masses of same-generation left-handed squarks are connected. BLABLABLA FYLL UT

Calculations in `Prospino 2.1` set some NLO terms to zero as discussed in Sec. 3.4.1. These points become outliers in the dataset. The outliers can be seen as a cluster of points well below the others for large masses in Fig. 5.2, where $\sigma_{m_{\tilde{g}}}$ is plotted as a function of $m_{\tilde{q}}$. These points have zero cross sections, which are set to 10^{-32} fb in the calculation to avoid divergences. In Fig. 5.2 they have different values because the cross sections have been divided by the gluino mass.

The panel in Fig. ?? shows a scatter plot of $m_{\tilde{d}_L}$ versus $m_{\tilde{u}_L}$. The plot is almost linear, which comes from the mass splittings of the MSSM. Same-generation left-handed squarks come in $SU(2)$ -doublets, and their masses are predominantly determined by *one* mass parameter because they must form a complete $SU(2)$ -representation. The mass parameter responsible is the soft mass $m_{Q_i}^2$ for generation i . Right-handed squarks, on the other hand, get their masses from different parameters $m_{\tilde{d}_i}^2$ and $m_{\tilde{u}_i}^2$, and the masses are therefore independent of each other in the MSSM-24. The mass splitting between same-generation left-

handed squarks, *e.g.* \tilde{d}_L and \tilde{u}_L , was given in Sec. 1.4.5,

$$m_{\tilde{d}_L}^2 - m_{\tilde{u}_L}^2 = -\cos 2\beta \ m_W^2,$$

where $m_W^2 \approx 80$ GeV. The mass splitting is relatively small for large masses, but becomes significant for small $m_{\tilde{d}_L}$ and $m_{\tilde{u}_L}$. It is therefore a possibility that training on processes with same-generation left-handed squarks, $\tilde{u}_L \tilde{d}_L$, $\tilde{s}_L \tilde{c}_L$, will effectively be training on a single squark mass.

Noise

The data contains some noise that originates in the numerical **Prospino 2.1** calculation, as discussed in Sec. 3.4.1. In a parameter point chosen at random the relative error ϵ_i has a typical numerical variance of $\sigma_\epsilon^2 \simeq 4.0 \cdot 10^{-6}$. This error fluctuates little between parameter points because there is a convergence criterium in **Prospino 2.1**, so it is considered a good approximation for the order of magnitude of errors in all points. The goal is now to incorporate this information in the Gaussian process. For that purpose the following relation is considered,

$$Y_i = y_i^{true} + \Delta y_i = y_i^{true}(1 + \epsilon_i), \quad (5.1)$$

where Y_i are cross sections provided by Prospino, y_i^{true} are true cross sections and Δy_i is the error in cross sections. The relative error from the calculation is assumed to follow a Gaussian distribution $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$.

The target values used in this project will in fact be the logarithm of the cross sections,

$$\begin{aligned} X_i &= \log_{10} Y_i \\ &= \log_{10}(y_i^{true}) + \log_{10}(1 + \epsilon_i). \end{aligned} \quad (5.2)$$

For small ϵ_i the second term in Eq. (5.2) can be expanded in a Taylor series

$$\log_{10}(1 + \epsilon_i) = \frac{\epsilon_i}{\log 10} - \frac{\epsilon_i^2}{2 \log 10} + \mathcal{O}(\epsilon_i^3). \quad (5.3)$$

The first term in Eq. (5.3) is dominant, and for small enough ϵ_i the following approximation can be made

$$\log_{10}(1 + \epsilon_i) \approx \frac{1}{\log 10} \ \epsilon_i. \quad (5.4)$$

The targets in Eq. (5.2) can now be written as

$$X_i \approx \log_{10}(y_i^{true}) + \frac{1}{\log 10} \ \epsilon_i = \log_{10}(y_i^{true}) + \varepsilon_i \quad (5.5)$$

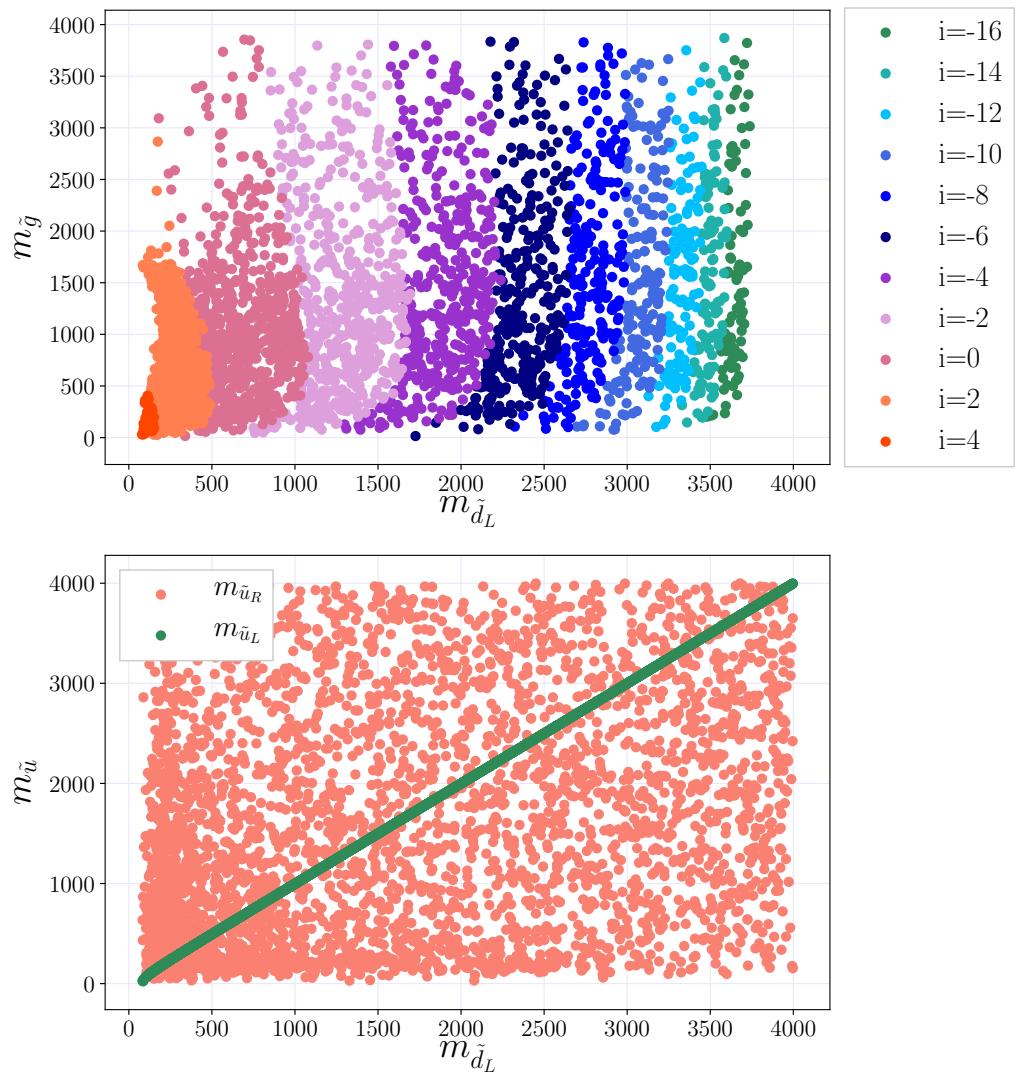


Figure 5.3: Data quality plots of the distribution of mass parameters $m_{\tilde{d}_L}$, $m_{\tilde{u}_L}$ and $m_{\tilde{g}}$ for 40 000 points. In (c) different orders of magnitude of the cross sections for $\tilde{d}_L \tilde{d}_L$ are shown in different colors, indicating that there are few points in the very high cross sections $\sigma \in [10^4 - 10^6]$, and that lower cross sections are more spread across the gluino mass spectrum.

where the total error, ε_i , now follows a Gaussian distribution $\mathcal{N}(0, \sigma_\varepsilon^2)$. For a random variable X multiplied by a number c the variance is given by

$$\mathbb{V}[cX] = c^2 \cdot \mathbb{V}[X], \quad (5.6)$$

so the variance σ_ε^2 is given by

$$\sigma_n^2 = \frac{\sigma_\epsilon^2}{(\log 10)^2}. \quad (5.7)$$

Equation (5.5) is now on the form of the Gaussian noise model discussed in Sec. 4.3.1. The distribution of the relative error ϵ_i has variance $\sigma_\epsilon^2 = 4.0 \cdot 10^{-6}$, so the variance of the Gaussian distributed noise should be

$$\sigma_\varepsilon^2 = \frac{4.0 \cdot 10^{-6}}{(\log 10)^2} \simeq 7.5 \cdot 10^{-7}.$$

The noise term predicted by the Gaussian process estimator should therefore have a variance of the order $\sigma_\varepsilon^2 \sim \mathcal{O}(10^{-7} - 10^{-6})$.

5.2 Dataset Transformations

The plot in Fig. 5.3 indicates that cross sections, in particular of the order $\mathcal{O}(1 \text{ fb})$ and lower, are very spread as a function of $m_{\tilde{q}}$. The spread is also evident in the upper left panel of Fig. 5.4, where the logarithm of the cross section for the production of $d_L d_{\tilde{L}}$ is plotted as a function of the gluino mass. Spread function values are problematic because they make it difficult for the Gaussian process regressor to recognize the shape of the function. The upper left panel shows σ as a function of the squark mass $m_{\tilde{q}}$, which is also quite spread. In addition, for small cross sections the cross section is a very steep function of $m_{\tilde{q}}$, while it is relatively flat for large σ . High derivatives in the function means that training points need to be close to estimate the actual shape of the function, which presents another difficulty. Some of the steepness and spread is remedied by using $\log_{10} \sigma$ instead of σ . This section will be devoted to reducing the spread caused by the gluino mass dependency.

Scaling Functions

As discussed in Sec. 2.4, the partonic cross sections can be written in terms of scaling functions f , see Eq. (3.19). The scaling functions are the different contributions to the cross section, as explained in Sec. 2.4. The total cross section only differs from the partonic cross section by an integral over parton distribution functions, so the mass dependencies in Eq. (3.19) are relevant for the total cross sections as well.

The main contributions to the cross section come from the scaling functions at the threshold energy given in Eq. (3.21), which makes it possible to remove some of the complexity of the function. Note that all terms are proportional to f_{qq}^B ($f_{q'q}^B$), which is again proportional to $m_q^2 m_{\tilde{g}}^2$,

$$\sigma \propto m_q^2 m_{\tilde{g}}^2. \quad (5.8)$$

Are partonic cross section are proportional to $\sigma \propto 1/m^2$, where m is the average mass of the final state particles. In squark pair production $m^2 = m_{\tilde{q}}^2$, so the squark mass dependency in Eq. (5.8) is automatically cancelled. However, the following transformation can be made

$$\sigma \rightarrow \sigma_{m_{\tilde{g}}} = \frac{\sigma}{m_{\tilde{g}}^2}, \quad (5.9)$$

reducing the gluino mass dependency. Another possibility is to further reduce the mass dependency by defining σ_{fac} as

$$\sigma_{fac} = \sigma \frac{(m_{\tilde{g}}^2 + m_{\tilde{q}}^2)^2}{m_{\tilde{g}}^2}. \quad (5.10)$$

The middle and bottom panels in Fig. 5.4 show $\sigma_{m_{\tilde{g}}}$ and σ_{fac} as functions of $m_{\tilde{g}}$ and $m_{\tilde{q}}$. The spread as a function of the squark mass is reduced for both expressions, but notably more for σ_{fac} than for $\sigma_{m_{\tilde{g}}}$. However, for large cross sections the shape of $\sigma_{m_{\tilde{g}}}$ as a function of squark and gluino mass are very similar, while both σ and σ_{fac} are very flat functions of $m_{\tilde{g}}$. The similarity in $\sigma_{m_{\tilde{g}}}$ may contribute to finding a kernel that fits well in both dimensions. Both $\sigma_{m_{\tilde{g}}}$ and σ_{fac} were attempted as target values, and $\sigma_{m_{\tilde{g}}}$ gave the best results. The target value in this project is therefore the logarithm of $\sigma_{m_{\tilde{g}}}$,

$$\log_{10} \sigma_{m_{\tilde{g}}}. \quad (5.11)$$

In the threshold region the partonic version of $\sigma_{m_{\tilde{g}}}$ is given by

$$\hat{\sigma}_{ij, m_{\tilde{g}}} = \frac{8\pi\beta\alpha_s^2(Q^2)}{27(m_q^2 + m_{\tilde{g}}^2)^2} \left\{ 1 + 4\pi\alpha_s(Q^2) \left[\frac{1}{24\beta} + \frac{2}{3\pi^2} \log^2(8\beta^2) - \frac{7}{2\pi^2} \log(8\beta^2) - \frac{2}{3\pi^2} \log(8\beta^2) \log\left(\frac{Q^2}{m^2}\right) \right] \right\}. \quad (5.12)$$

5.3 Learning the Gaussian Process

In this section a Gaussian process estimator is trained with benchmark settings, and a selected set of modifications to the benchmark are introduced. The quality

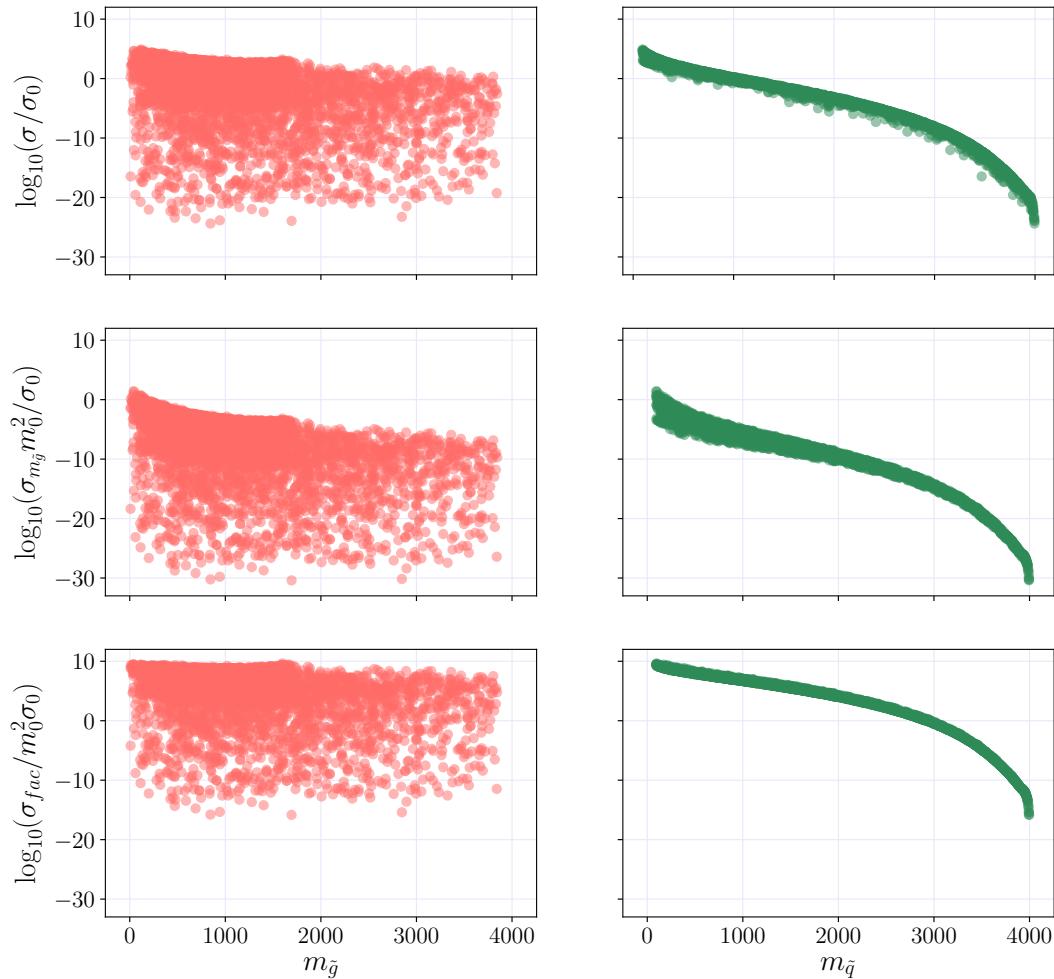


Figure 5.4: The cross section, σ , and the modified cross sections, $\sigma_{m_{\tilde{g}}}$ and σ_{fac} , as functions of the gluino mass $m_{\tilde{g}}$ and squark mass $m_{\tilde{q}} = m_{\tilde{d}_L}$ for the production of $\tilde{d}_L \tilde{d}_L$. The distributions are less spread when some of the mass dependency is removed.

of estimators are quantified in plots of the relative deviance distributions defined in Sec. 4.4.3. Note that in many cases the parameter σ_f^2 is at its upper limit, $\sigma_{f_{\max}}^2 = 1000$. Increasing the upper limit on σ_f^2 alters the optimized length scale parameters, but does not change the prediction. The upper limit of σ_f^2 has therefore been kept at $\sigma_{f_{\max}}^2 = 1000$.

5.3.1 The Benchmark

The example processes are the pair production of $\tilde{d}_L \tilde{d}_L$ and $\tilde{d}_L \tilde{u}_R$. The benchmark settings (BM) are

- A GP estimator with 2000 training points.
- 20 000 test points.
- Features $m_{\tilde{g}}, m_{\tilde{d}_L}$ ($m_{\tilde{g}}, m_{\tilde{d}_L}, m_{\tilde{u}_R}$) for $\tilde{d}_L \tilde{d}_L$ ($\tilde{d}_L \tilde{u}_R$) production.
- The exponential squared kernel (RBF) with a white noise term,

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sigma_f^2 \exp\left(-\frac{1}{2}(\mathbf{x}_i - \mathbf{x}_j)^T M (\mathbf{x} - \mathbf{x}')\right) + \sigma_n^2 \delta_{ij}, \quad (5.13)$$

where $M = \text{diag}(\boldsymbol{\ell})^{-2}$, and σ_n^2 is the variance of the distribution of white noise $\varepsilon_{WK} \sim \mathcal{N}(0, \sigma_n^2)$.

The kernel is implemented in **scikit-learn** in the following way

```
kernel_BM = ConstantKernel(constant_value=10,
    constant_value_bounds=(1e-3, 1e4)) * RBF(length_scale =
        np.array([1000, 1000]), length_scale_bounds=(1, 1e6)) +
    WhiteKernel(noise_level=1, noise_level_bounds=(2e-10, 1e2))
```

Note that the length scale of the RBF is given as a vector of the same dimension as the feature vector $\mathbf{x}, \boldsymbol{\ell} \in \mathbb{R}^D$. This is to allow different features to have different characteristic length scales, as they appear to have from the lower panels in Fig. 5.4. The target values are $\log_{10} \sigma_{m_{\tilde{g}}}$, where $\sigma_{m_{\tilde{g}}}$ was defined in Sec. 5.2.

The means and standard deviations $m(\varepsilon_i)$ and $\text{std}(\varepsilon_i)$, as defined in Eqs. (4.54) – (4.55), are calculated and plotted for the BM settings in Fig. 5.5 for $\tilde{d}_L \tilde{d}_L$. For $\tilde{d}_L \tilde{u}_R$ the results are very similar, and therefore not shown here. The optimized kernel parameters found by the GP are listed in Tables 5.2 – 5.3, while computation times on a standard laptop with 4 cores are found in Table 5.4. The predicted noise level variances, σ_n^2 , are very high for both processes, at $\sigma_n^2 = 0.47$ for $\tilde{d}_L \tilde{d}_L$ and $\sigma_n^2 = 0.65$ for $\tilde{d}_L \tilde{u}_R$, which is far from the value $\sigma_\varepsilon^2 \sim 7.5 \cdot 10^{-7}$ predicted in Sec. 5.1. In addition the estimator is very unstable, predicting values that are too small for small cross sections, as can be seen in Fig. 5.5.

	σ_f	$\ell_{m_{\bar{g}}}$	$\ell_{m_{\tilde{d}_L}}$	$\ell_{\bar{m}}$	σ_n^2
BM	59.6	5470	2190		0.47
No outliers	98.5	5740	215		0.00372
$\sigma > 10^{-16}$ fb	22.7	1170	998		0.0036
\bar{m}	33.1	1190	200	846	0.0000119
Matern	31.7	30200	8600		0.462

Table 5.2: Optimized kernel parameters for different settings for $\tilde{d}_L \tilde{d}_L$.

	σ_f	$\ell_{m_{\bar{g}}}$	$\ell_{m_{\tilde{d}_L}}$	$\ell_{m_{\tilde{u}_R}}$	$\ell_{\bar{m}}$	σ_n^2
BM	22.7	3070	2700	3770		0.65
No outliers	31.6	1010	3030	3240		0.00316
$\sigma > 10^{-16}$ fb	31.6	1150	3230	3310		0.00276
\bar{m}	166	795	5010	3910	742	0.0000697
Matérn	10.9	3220	1700	6950		0.402

Table 5.3: Optimized kernel parameters for different settings for $\tilde{d}_L \tilde{u}_R$.

	Time $\tilde{d}_L \tilde{d}_L$	Time $\tilde{d}_L \tilde{u}_R$
BM	00:07:48	00:08:40
No outliers	00:08:42	00:11:20
$\sigma > 10^{-16}$ fb	00:07:24	00:11:07
\bar{m}	00:11:20	00:16:37
Matérn	00:07:28	00:13:05

Table 5.4: Computation times with 2000 training points and 20 000 test points on a laptop with four cores.

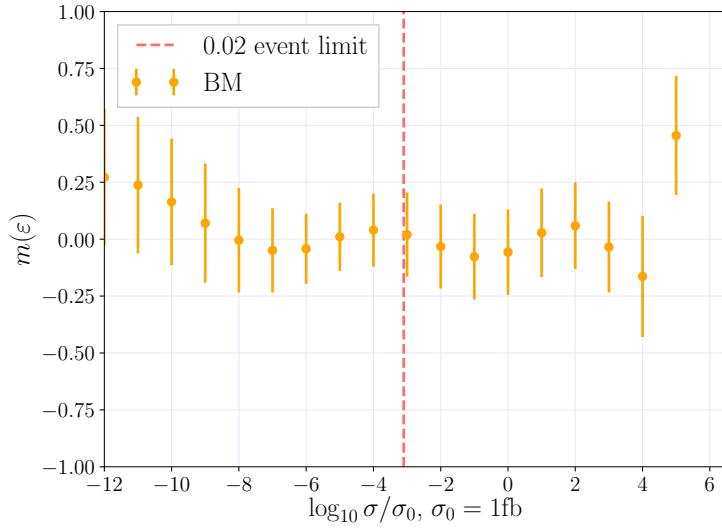


Figure 5.5: The mean and standard deviation of the relative deviance distributions, $m(\varepsilon_i)$ and $\text{std}(\varepsilon_i)$, as a function of $i = \log_{10} \sigma/\sigma_0$, for the process $\tilde{d}_L \tilde{d}_L$ with benchmark settings. The optimized kernel parameters are listed in Table 5.2.

5.3.2 Outliers

As discussed in Sec. 5.1 the data contains outliers. To see the effect of the outlier points in the GP prediction, a GP estimator with the BM settings is trained on a dataset where outliers have been removed. Removing outliers improves the prediction for small cross sections, but also stabilizes the prediction for larger values. This can be seen in Fig. 5.6a where BM estimators with and without outliers are compared. The predicted noise levels are significantly reduced with the removal of outliers for both processes, from $\sigma_n^2 = 0.47$ to $\sigma_n^2 = 0.00372$ for $\tilde{d}_L \tilde{d}_L$ and from $\sigma_n^2 = 0.65$ to $\alpha = 0.00316$ for $\tilde{d}_L \tilde{u}_R$. This implies that including the outliers leads the GP to underfit, which means that much of the signal is considered noise.

5.3.3 Cuts on Cross Sections

Smooth functions are easier to fit with Gaussian processes, and the target values are very steep functions of large squark masses. This can be seen from the righthand panels in Fig. 5.4. In addition, small target values comprise the regions with the most spread as a function of the gluino mass. Since the limit for 0.02 events at the LHC with 20 fb^{-1} of integrated luminosity is at $\sigma = 10^{-3} \text{ fb}^3$, a

³Cross sections with lower values than this predict less than 0.02 events, and are therefore less interesting in this project.

lower cut is set at $\sigma_{cut} = 10^{-16}$ fb. The cut excludes all cross sections lower than σ_{cut} from both training and testing.

The resulting distributions of the relative deviance for $\tilde{d}_L \tilde{d}_L$ are shown in Fig. 5.6b, and the optimized kernel parameters are found in Tables 5.2 – 5.3. Noise levels are further reduced from the case where outliers are removed, with the variance going from $\sigma_n^2 = 0.00372$ to $\sigma_n^2 = 0.00336$ for $\tilde{d}_L \tilde{d}_L$ and from $\sigma_n^2 = 0.00316$ to $\sigma_n^2 = 0.00276$ for $\tilde{d}_L \tilde{u}_R$. The estimated noise level seems to be approaching the value estimated in Sec. 5.1, indicating that the exclusion of low cross sections improves the prediction.

The relative deviances of the largest target value predictions are significantly improved with respect to the BM with and without outliers. Training and testing exclusively on large cross sections renders a very stable prediction for all (included) orders of magnitude.

5.3.4 Features

`Prospino 2.1` calculates the NLO cross section for the mean of the squark masses, \bar{m} , and uses this to find the K -factor. The K -factor is then multiplied with the LO cross sections for non-degenerate squark masses to find the individual NLO terms, as discussed in Sec. 3.4.1. Adding the mean squark mass as a feature could therefore improve the prediction.

The optimized kernel values are found in Tables 5.2 – 5.3. Adding the mean mass as a feature reduces the estimated noise level considerably, at $\sigma_n^2 = 1.19 \cdot 10^{-5}$ for $\tilde{d}_L \tilde{d}_L$, and $\sigma_n^2 = 6.97 \cdot 10^{-5}$ for $\tilde{d}_L \tilde{u}_R$. This is an indication that the estimator is improved, considering the two processes should have a noise level of the same order of magnitude, as discussed in Sec. 5.1.

Further, the estimator predicts a shorter length scale for \bar{m} than for $m_{\tilde{g}}$. This implies that there is a higher dependence on the mean mass than the gluino mass, since GPs attribute large length scales to irrelevant features. The resulting distributions for $d_L d_L$ are shown in Fig. 5.6c and compared to the BM. With some exceptions at $\log_{10} \sigma / \sigma_0 \in [2, 4]$, where the variances are very large, adding \bar{m} as a feature gives a mean of almost zero and very small variances of the relative deviances for cross sections over the 0.02 event limit.

5.3.5 Kernel

The RBF kernel is intrinsically smooth, while the Matérn kernel has a hyperparameter ν to control its smoothness. It is sometimes argued that this makes Matérn a better kernel for physical processes, as mentioned in Sec. 4.2.2, and the Matérn kernel is therefore tested. In `scikit-learn` the hyperparameter ν is not optimized during training, and must be determined beforehand. In this section the Matérn kernel with $\nu = \frac{3}{2}$ is compared to the BM RBF kernel, and

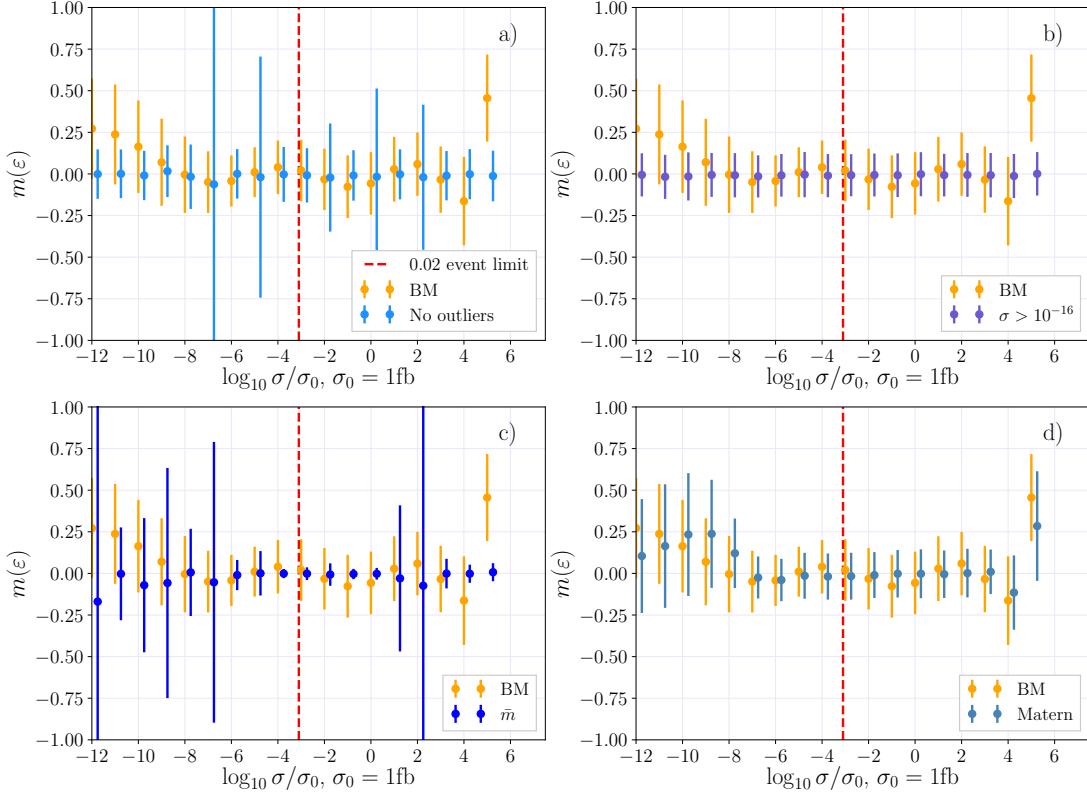


Figure 5.6: The mean and standard deviation of the relative deviance distributions, $m(\varepsilon_i)$ and $\text{std}(\varepsilon_i)$ per decade $i = \log_{10} \sigma/\sigma_0$, for the process $\tilde{d}_L \tilde{d}_L$ with **a)** benchmark settings (orange) and removed outliers (blue); **b)** benchmark settings (orange) and a lower cut on cross sections (blue); **c)** benchmark settings (orange) and the added feature \bar{m} (blue); **d)** the benchmark settings (orange) and the Matérn kernel (blue).

the resulting error distributions for $\tilde{d}_L \tilde{d}_L$ are found in Fig. 5.6d.

The hyperparameter ν is set to $\frac{3}{2}$ as this is one of the values for which covariances are quick to calculate. For values not in $\nu = [\frac{1}{2}, \frac{3}{2}, \frac{5}{2}, \infty]$ scikit-learn evaluates Bessel functions explicitly, which increases the computational cost by up to a factor 10.

The predictions are somewhat more stable for high cross sections than with the RBF kernel. For low cross sections the error distributions have larger variances, but as this is not the region of interest the Matérn kernel is considered a better fit than the RBF. As can be seen from the optimized kernel values in Tables 5.2 – 5.3, the Matérn kernel predicts a slightly lower noise level than the RBF. The noise variances go from $\sigma_n^2 = 0.47$ with RBF to $\sigma_n^2 = 0.462$ with Matérn for $\tilde{d}_L \tilde{d}_L$, and from $\sigma_n^2 = 0.65$ with RBF to $\sigma = 0.402$ with Matérn for $\tilde{d}_L \tilde{u}_R$.

5.3.6 Cummulative Settings

A combination of the improved settings from the foregoing sections is used in this section. To sum up, the cummulative settings are as follows

- A GP estimator with 2000 training points.
- 20 000 test points.
- Features $[m_{\tilde{g}}, m_{\tilde{d}_L}, \bar{m}]$ ($[m_{\tilde{g}}, m_{\tilde{d}_L}, m_{\tilde{u}_R}, \bar{m}]$) for $\tilde{d}_L \tilde{d}_L$ ($\tilde{d}_L \tilde{u}_R$) production.
- The Matérn kernel with $\nu = 1.5$ and a white noise term

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sigma_f^2 \exp \left(1 + \sqrt{3} [(\mathbf{x}_i - \mathbf{x}_j)^T M (\mathbf{x}_i - \mathbf{x}_j)]^{1/2} \right) \quad (5.14)$$

$$\times \exp \left(\sqrt{3} [(\mathbf{x}_i - \mathbf{x}_j)^T M (\mathbf{x}_i - \mathbf{x}_j)]^{1/2} \right) + \sigma_n^2 \delta_{ij}, \quad (5.15)$$

where $M = \text{diag}(\boldsymbol{\ell})^{-2}$.

- A cut on the cross sections $\sigma > \sigma_{cut} = 10^{-16}$ fb.⁴

The resulting relative deviance distributions are found in Fig. 5.7. Note that the limits on the y -axis are now $[-0.2, 0.2]$, as opposed to $[-1, 1]$ in Fig. 5.6. With the cummulative settings all standard deviations, $\text{std}(\varepsilon_i)$, are less than 5% and all means, $m(\varepsilon_i)$, are close to zero for $\tilde{d}_L \tilde{u}_R$. The estimator for the process $\tilde{d}_L \tilde{d}_L$ performs considerably worse than $\tilde{d}_L \tilde{d}_L$, but most standard deviations are within 10%. The computation with the cummulative settings for 2000 training points takes 00:09:30 for $\tilde{d}_L \tilde{d}_L$ and 00:10:28 for $\tilde{d}_L \tilde{u}_R$ under the same conditions as in Table 5.4.

Noise

The noise level is known to some degree, as discussed in Sec. 5.1, so a more direct approach of explicitly adding noise to the covariance is here tested with the cummulative settings. In `scikit-learn` an option to letting the `WhiteKernel` estimate the level of noise is to specify the noise by passing it as to the Gaussian process regressor function.

For the cummulative settings the `WhiteKernel` predicts values relatively close to $\sigma_\varepsilon^2 = 7.5 \cdot 10^{-7}$, with $\sigma_n^2 = 10^{-5}$ for $\tilde{d}_L \tilde{d}_L$ and $5.1 \cdot 10^{-5}$ for $\tilde{d}_L \tilde{u}_R$. Except for the length scale for the mean mass, $\ell_{\bar{m}}$, which changes from 9500 to 5900 for $\tilde{d}_L \tilde{u}_R$, the remaining kernel parameters therefore hardly change when α is fixed, as seen in Table 5.5. The prediction changes very little with the addition of α_{fix} . A marginal improvement can be seen in the standard deviations of relative deviances when $\alpha = \alpha_{fix}$ in Fig. 5.8, but the mean values are almost identical.

⁴Effectively also removing the outliers $\sigma = 0 < 10^{-16}$ fb.

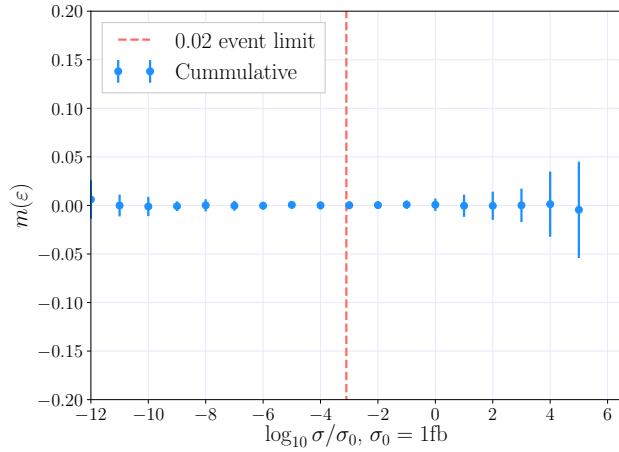
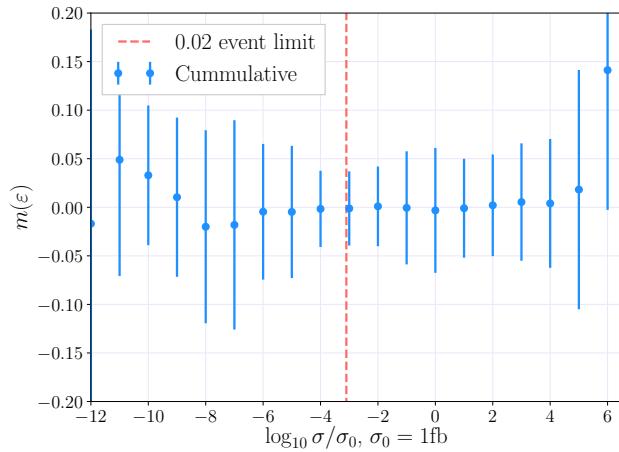
(a) The process $\tilde{d}_L \tilde{d}_L$ (b) The process $\tilde{d}_L \tilde{u}_R$

Figure 5.7: The mean and standard deviation of the relative deviance distributions, $m(\varepsilon_i)$ and $\text{std}(\varepsilon_i)$ per decade $i = \log_{10} \sigma / \sigma_0$ for $\tilde{d}_L \tilde{d}_L$ with the cummulative settings from Sec. 5.3.6. All orders of magnitude have standard deviations smaller than 5%.

	σ_f	$\ell_{m_{\tilde{g}}}$	$\ell_{m_{\tilde{d}_L}}$	$\ell_{m_{\tilde{u}_R}}$	$\ell_{\bar{m}}$	σ_n^2
$\tilde{d}_L \tilde{d}_L$	27.4	30500	17400		74800	$1 \cdot 10^{-5}$
$\tilde{d}_L \tilde{d}_L$	31.6	28300	18300		69900	$7.5 \cdot 10^{-7}$ (fixed)
$\tilde{d}_L \tilde{u}_R$	31.6	1890	4100	4140	9500	$5.1 \cdot 10^{-5}$
$\tilde{d}_L \tilde{u}_R$	31.6	1780	4020	4030	5900	$7.5 \cdot 10^{-7}$ (fixed)

Table 5.5: Optimized kernel parameters for the cumulative settings with the noise level estimated by the GP, and given explicitly to the estimator $\sigma_\varepsilon^2 = 7.5 \cdot 10^{-7}$.

For calculations with few training points the computation time is not affected in any great way, but for larger datasets the removal of `WhiteKernel` from the total kernel may reduce the computation time, as it leaves one less hyperparameter to optimize. For the remainder of the project the `WhiteKernel` with a free σ_n^2 is used.

5.4 Distributed Gaussian Processes

A significant disadvantage of Gaussian processes is that they scale poorly with the size of the data set, as discussed in Sec. 4.3. For n training points and n^* test points, training and predicting scale as $\mathcal{O}(n^3)$ and $\mathcal{O}(n^{*2})$, respectively, giving GP training data sets a practical limit of $\mathcal{O}(10^4)$.

In [29] a method of scaling GPs to large data sets is proposed, in the form of a robust Bayesian Committee Machine (rBCM). This method is based on the concepts of product-of-experts and Bayesian Committee Machine introduced below, and has the advantage of providing an uncertainty for the prediction.

5.4.1 Product-of-Experts

Product-of-expert (PoE) models are a way of parallelising large computations. They combine several independent computations on subsets of the total data, called ‘experts’. In the case of distributed Gaussian processes each expert performs GP on a subset of the training data, and the predictions on a common test set are combined.

Consider the training data set $\mathcal{D} = \{\mathbf{X}, \mathbf{y}\}$, which is partitioned into M subsets $\mathcal{D}^k = \{X^k, \mathbf{y}^k\}$, $k = 1, \dots, M$. The feature vectors are D -dimensional, $\mathbf{x} \in \mathbb{R}^D$. Each GP expert does learning on its training data set \mathcal{D}^k , then predictions are combined at the parent node. This node could also be considered an expert for a PoE with several layers, see Fig. 5.9.

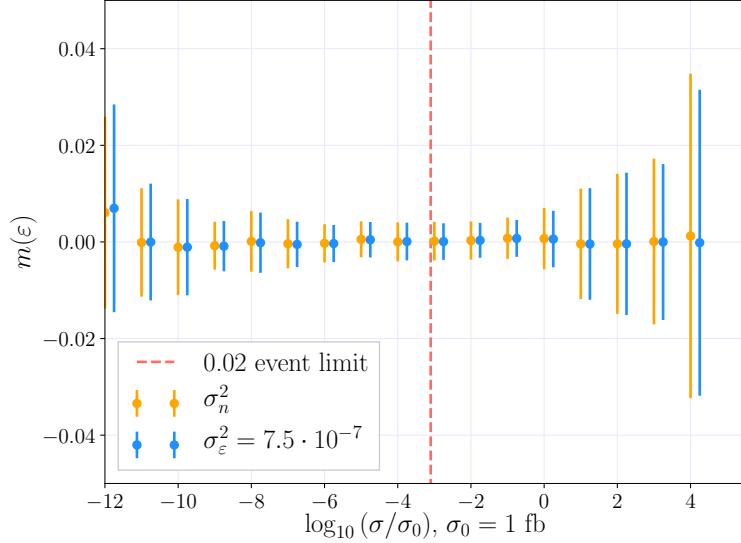


Figure 5.8: The mean and standard deviation of the relative deviation distributions, $m(\varepsilon_i)$ and $\text{std}(\varepsilon_i)$, per decade $i = \log_{10} \sigma / \sigma_0$ for $\tilde{d}_L \tilde{d}_L$ with the cummulative settings from Sec. 5.3.6. The noise level σ_n^2 is estimated by the GP (orange), and given explicitly as $\sigma_\varepsilon^2 = 7.5 \cdot 10^{-7}$ (blue).

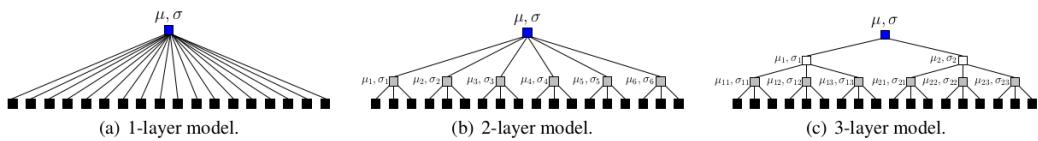


Figure 5.9: Computational graphs of hierarchical product-of-expert models. Main computations are at the leaf nodes (GP experts in black). All other nodes recombine computations from their children nodes. The blue node at the top represents the final prediction. Figure from [29].

5.4.2 Bayesian Committee Machine

The Bayesian Committee Machine [30] is a type of PoE where a new weighting scheme is introduced. Consider the training data set from the previous section partitioned into M sets, $\mathcal{D}^k = \{X^k, \mathbf{y}^k\}$, $k = 1, \dots, M$. For simplicity a single test point, \mathbf{x}^* , is considered, where f^* is the unknown target value.

M estimators are trained separately, one for each training data set. Now let $\mathbf{D}^i = \{\mathcal{D}^1, \dots, \mathcal{D}^i\}$ denote the data sets with indices smaller than or equal to i for $i = 1, \dots, M$. For the i th estimator the posterior probability distribution for the test point \mathbf{x}^* is $P(f^*|\mathcal{D}^i)$. Then in general

$$P(f^*|\mathbf{D}^{i-1}, \mathcal{D}^i) \propto P(f^*)P(\mathbf{D}^{i-1}|f^*)P(\mathcal{D}^i|\mathbf{D}^{i-1}, f^*), \quad (5.16)$$

meaning that the posterior distribution for the datasets $\{\mathcal{D}^1, \dots, \mathcal{D}^i\}$ is proportional to the posterior for the dataset \mathcal{D}^i given the datasets $\{\mathcal{D}^1, \dots, \mathcal{D}^{i-1}\}$, times the posterior of the datasets $\{\mathcal{D}^1, \dots, \mathcal{D}^{i-1}\}$, multiplied by the prior $P(f^*)$. Note that this is simply a generalization of the product rule in Eq. (4.2).

Assuming the datasets \mathbf{D}^{i-1} and \mathcal{D}^i are independent, or at least have a very small correlation, given f^* , the following approximation is made

$$P(\mathcal{D}^i|\mathbf{D}^{i-1}, f^*) \approx P(\mathcal{D}^i|f^*). \quad (5.17)$$

Using this approximation and applying Bayes' theorem gives

$$P(f^*|\mathbf{D}^{i-1}, \mathcal{D}^i) \propto \frac{P(f^*|\mathbf{D}^{i-1})P(f^*|\mathcal{D}^i)}{P(f^*)}. \quad (5.18)$$

The approximate posterior distribution using the entire dataset, \mathcal{D} , is then

$$\hat{P}(f^*|\mathcal{D}) \propto \frac{\prod_{i=1}^M P(f^*|\mathcal{D}^i)}{P(f^*)^{M-1}}, \quad (5.19)$$

where the hat over \hat{P} means that the distribution is approximate. The posterior distributions of each expert are simply multiplied, and divided by the prior $M-1$ times.

5.4.3 Robust Bayesian Committee Machine

The Bayesian Committee Machine (BCM) can be modified to a *robust Bayesian Committee Machine* (rBCM), which is the scheme that will be used in this project for distributed Gaussian processes (DGP). The M experts are assumed to be independent [29] as in Eq. (5.17), effectively block-diagonalizing the covariance matrix. The marginal likelihood from Eq. 4.50 now factorizes into the product of M individual terms because of the independence assumption. For the training

data of one data set $\mathcal{D}^k = \{X^k, \mathbf{y}^k\}$ the log marginal likelihood is given by Eq. (4.50)

$$\log P(\mathbf{y}^k | X^k, \boldsymbol{\theta}) = -\frac{1}{2} \mathbf{y}^{kT} (K_\psi^k + \sigma_\varepsilon^2 \mathbb{I})^{-1} \mathbf{y}^k - \frac{1}{2} \log |K_\psi^k + \sigma_\varepsilon^2 \mathbb{I}|, \quad (5.20)$$

where K_ψ^k is the covariance matrix of the training features X^k , with the kernel hyperparameters given by $\boldsymbol{\theta} = \{\psi, \sigma_\varepsilon^2\}$. Computing the LML now entails inverting a $n_k \times n_k$ matrix, $K_\psi^k + \sigma_\varepsilon^2 \mathbb{I}$, which requires time $\mathcal{O}(n_k^3)$ and memory consumption $\mathcal{O}(n_k^2 + n_k D)$ for $\mathbf{x} \in \mathbb{R}^D$. For $n_k \ll N$, this reduces the computation time and memory use considerably, and allows for parallel computing.

The rBCM predicts a function value f^* at a corresponding test input \mathbf{x}^* according to a predictive distribution that is similar to the BCM in Eq. (5.19), with the addition of powers β_k ,

$$P(f^* | \mathbf{x}^*, \mathcal{D}) = \frac{\prod_{k=1}^M P_k^{\beta_k}(f^* | \mathbf{x}^*, \mathcal{D}^{(k)})}{P^{-1+\sum_k \beta_k}(f^* | \mathbf{x}^*)}, \quad (5.21)$$

where the powers β_k control the importance of the individual experts, but also how strong the influence of the prior is. In [29], these are chosen according to the predictive power of each expert at \mathbf{x}^* . More specifically, β_k is the change in differential entropy between the prior $P(f^* | \mathbf{x}^*)$ and the posterior $P(f^* | \mathbf{x}^*, \mathcal{D}^{(k)})$, which can be calculated as

$$\beta_k = \frac{1}{2} (\log \sigma_{**}^2 - \log \sigma_k^2(\mathbf{x}^*)), \quad (5.22)$$

where $\sigma_{**}^2 = k(\mathbf{x}^*, \mathbf{x}^*)$ is the prior variance of the test point, and $\sigma_k^2(\mathbf{x}^*)$ is the predictive variance of the k th expert given by Eq. 4.48.

The combined predictive mean and variance are denoted μ_{rbcm}^* and σ_{rbcm}^{*2} , and are then given by

$$\mu_{rbcm}^* = (\sigma_{rbcm}^*)^2 \sum_k \beta_k \sigma_k^{-2}(\mathbf{x}^*) \mu_k(\mathbf{x}^*), \quad (5.23)$$

$$(\sigma_{rbcm}^*)^{-2} = \sum_{k=1}^M \beta_k \sigma_k^{-2}(\mathbf{x}^*) + \left(1 - \sum_{k=1}^M \beta_k\right) \sigma_{**}^{-2}. \quad (5.24)$$

Implementation

There is currently no functionality for an rBCM in the **scikit-learn** library, so the combined predictive mean and variance in Eqs. (5.23) – (5.24) were implemented in a **Python** routine. The **scikit-learn** library's existing framework for regular Gaussian processes were used for the individual experts. The algorithm was parallelised, so that each expert can learn and predict in parallel, before being combined to the final prediction. Pseudocode for the implementation is

found in Algorithm 2, which takes the training data X , \mathbf{y} , the initial kernel k , the noise level variance σ_n^2 , the number of experts $N_{experts}$ and the test features \mathbf{x}^* as input, and computes the combined predictive mean and variance, μ_{rbcm}^* , σ_{rbcm}^{*2} .

For parallelisation the `scikit-learn` function `Parallel` from `joblib` was used, which runs Python functions as pipeline jobs. It uses the Python function `multiprocessing` as a backend. An example of usage with 3 parallel jobs is as follows

```
>>> from joblib import Parallel, delayed
>>> from math import sqrt
>>> Parallel(n_jobs=3)(delayed(sqrt)(i**2) for i in range(10))
[0.0, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0]
```

where `delayed` is a simple trick to be able to create a tuple with a function-call syntax.

Data: N (number of experts), X (inputs), \mathbf{y} (targets), k (initial kernel), σ_n^2 (noise level), \mathbf{x}^* (test input)

Split training data into N subsets: X_i, \mathbf{y}_i ;

for expert $i = 1, \dots, N$ **do**

- Fit GP to training data X_i, \mathbf{y}_i ;
- Predict μ_{i*}, σ_{i*}^2 for \mathbf{x}^* using GP ;
- $\sigma_{i**}^2 = k(\mathbf{x}^*, \mathbf{x}^*)$;

end

for expert $i = 1, \dots, N$ **do**

- $\beta_i = \frac{1}{2}(\log(\sigma_{i**}^2) - \log(\sigma_{i*}^2))$;
- $(\sigma_*^{rbcm})^{-2} += \beta_i \sigma^{-2} + (\frac{1}{N} - \beta_i) \sigma_{i**}^{-2}$

end

for expert $i = 1, \dots, N$ **do**

- $\mu_*^{rbcm} += (\sigma_{i*}^{rbcm})^2 \beta_i \sigma_{i*}^{-2} \mu_{i*}$

end

Result: Approximative distribution of $f_* = f(\mathbf{x}_*)$ with mean μ_*^{rbcm} and variance $(\sigma_*^{rbcm})^2$.

Algorithm 2: Pseudocode for distributed Gaussian processes on a single test point \mathbf{x}_* . For the fit and prediction of each GP expert Algorithm (1) is used.

Number of experts	Points per expert	Time
1	2000	00:03:32
4	2000	00:05:46
1	8000	01:35:21

Table 5.6: Table of computation times for GP fit and prediction on the Abel HPC cluster.

5.4.4 Evaluating Cross Sections using Distributed Gaussian Processes

In this section the distributed Gaussian processes described in the previous section are applied to estimators with the benchmark settings from Sec. 5.3.1 to scale the problem to larger training sets. The prediction for a single expert with 8000 training points is compared to the combined prediction of 4 experts with 2000 training points each, both in terms of computation time and quality of prediction. Larger training sets per expert could be used, such as four experts with 8000 training points each, but comparing to a single expert with the same amount of data would be unfeasible because of the scale of the computation, as it would require an expert with $4 \times 8000 = 32000$ training points.

Adding Experts

The addition of experts with 2000 training points improves the prediction from Sec. 5.3.1 significantly, with very little increase in computational cost. In Fig. 5.10 a comparison of error distributions between one expert and four experts with 2000 training points each, and one expert with 8000 training points is shown. For comparison the settings are the benchmark settings, and outliers are included in the dataset. The improvement and stability of the prediction with the addition of data is large. From the relative deviation distributions there is little difference between using four experts with 2000 training points each and using a single expert with 8000 training points. The difference in computation times, however, is very large. Four experts with 2000 points take a little under six minutes to compute, while the single expert with 8000 points takes over an hour and a half. Computation times are shown in Table 5.6.

5.4.5 Cross Validation

Since there is no `scikit-learn` functionality for distributed Gaussian processes, an algorithm for k -fold cross validation as a function of experts was implemented, according to the outline in Sec. 4.4.2. The algorithm uses the `scikit-learn` function `KFold` to find training and test indices for k splits of the data. The

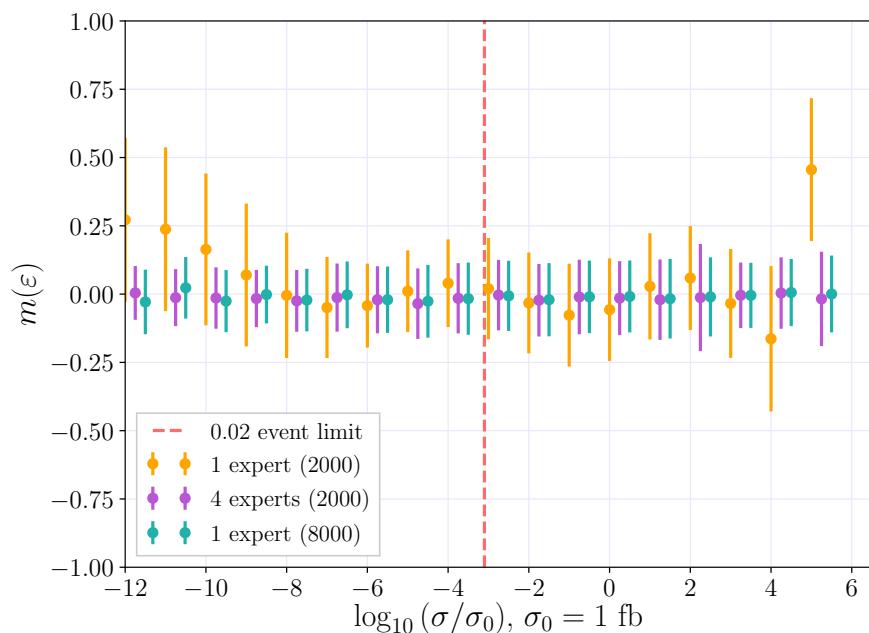


Figure 5.10: The mean and standard deviation of the relative deviance distributions, $m(\varepsilon_i)$ and $\text{std}(\varepsilon_i)$, per decade $i = \log_{10} \sigma/\sigma_0$ for $\tilde{d}_L \tilde{d}_L$, using GP with the BM settings, for one (orange) and four (violet) experts with 2000 training points each, and one expert with 8000 training points (green).

scoring function used in the CV is the R^2 -score introduced in Sec. 4.4.2, and pseudocode for the algorithm is found in Algorithm 3.

Data: N (max number of experts), n (training points per expert),
 X (inputs), \mathbf{y} (targets), k (number of folds for cross validation)
number of experts $n = 1, \dots, N$;
for $n = 1, \dots, N$ **do**

- Training size = $n \cdot (i + 1)$;
- Total size = training size $\cdot \frac{k}{k-1}$;
- Split training data into subsets;
- Use KFold to create k -fold cross validation instance kf ;
- for** *training indices, test indices in kf* **do**

 - Fit GP to $k - 1$ folds of training data;
 - Use 1 fold as test data;
 - Predict values \hat{y}_{train} for training data;
 - Predict values \hat{y}_{test} for test data;
 - $$R_{train}^2(\hat{y}, y) = 1 - \frac{\sum_{i=0}^{\text{Training size}-1} (y_i - \hat{y}_i)^2}{\sum_{i=0}^{\text{Training size}-1} (y_i - \bar{y})^2};$$
 - $$R_{test}^2(\hat{y}, y) = 1 - \frac{\sum_{i=0}^{n-1} (y_i - \hat{y}_i)^2}{\sum_{i=0}^{n-1} (y_i - \bar{y})^2};$$

- end**
- Find mean m and std of R_{test}^2 -values and R_{train}^2 -values

end

Result: $\mathbf{n}, \mathbf{m}(R_{train}^2), \sigma_{std}(R_{train}^2), \mathbf{m}(R_{test}^2), \sigma_{std}(R_{test}^2)$.

Algorithm 3: Pseudocode for k -fold cross validation of distributed Gaussian processes, which calculates the R^2 -scores for training and test data as a function of the number of experts, *e.g.* to be used for learning curves. In the R^2 -score calculation, y_i are true values, \hat{y}_i are GP predicted values, and \bar{y} is the mean of all y_i .

Chapter 6

Results

In this chapter distributed Gaussian processes trained on the MSSM-24 dataset are used to predict cross sections for MSSM-24 and CMSSM. The DGP estimators use the cumulative settings from the previous chapter. Learning curves as a function of number of experts are plotted for example processes for experts of different sizes. Furter, plots of the relative deviances from the Prospino data are shown, and the resulting predictions are compared to cross sections from Prospino and NLL-fast. Finally, optimizing the model with respect to predictive capabilities, model size and computation times is discussed.

6.1 Learning Curves

Learning curves as a function of number of experts for the cumulative settings from Sec. 5.3.6 are shown in Fig. 6.1 for $\tilde{d}_L \tilde{d}_L$ and $\tilde{d}_L \tilde{u}_R$. The experts are trained with 500, 1000 and 2000 points each, and learning curves are calculated according to the cross validation method in Sec. 4.4.2.

The training scores for the estimators of $\tilde{d}_L \tilde{d}_L$ and $\tilde{d}_L \tilde{u}_R$ are 1, indicating that neither model is underfitting. The validation curves for both processes converge towards 1, albeit faster and for less training points per expert for $\tilde{d}_L \tilde{d}_L$ than for $\tilde{d}_L \tilde{u}_R$. In both cases the training and validation scores are very high, even for few experts. Adding more data, both in the form of more experts and more points per expert, gives higher validation scores. Including a ‘bad’¹ expert affects the validation score negatively and increases the uncertainty in the score, *e.g.* from nine to ten experts with 1000 training points for $\tilde{d}_L \tilde{d}_L$. Nevertheless, the addition of data generally improves the score. Predictions in this chapter therefore use the largest reasonable² models, of ten experts with 5000 and 8000 training points each, depending on whether or not the models need to be stored.

¹Bad in the sense that the R^2 -score is low

²Taking into account computation times, matrix sizes and model sizes

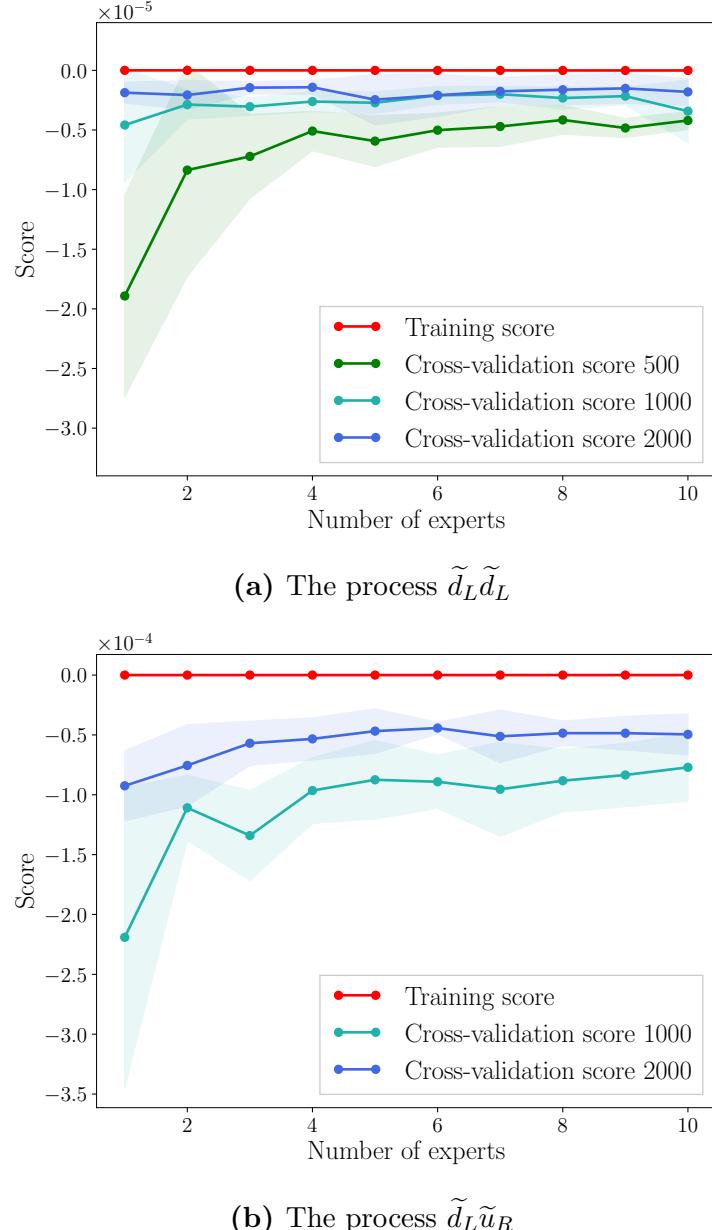


Figure 6.1: Learning curves as a function of number of experts, with 500, 1000 and 2000 training points per expert for the processes (a) $\tilde{d}_L \tilde{d}_L$ and (b) $\tilde{d}_L \tilde{u}_R$. The validation curve for $\tilde{d}_L \tilde{u}_R$ with 500 training points per expert is omitted because the uncertainties in the validation scores are very large. The k -fold cross validation uses R^2 -score as described in Sec. 4.4.2, and here $R^2 - 1$ is plotted. Note that the units on the y -axes are different for (a) and (b).

6.2 Comparison with Prospino and NLL-fast

In this section plots of the relative deviance distributions defined in Sec. 4.4.3 are shown for the squark pair-production cross sections from the MSSM-24 and CMSSM datasets. Cross sections predicted by the DGP are compared to cross sections calculated using **Prospino**, and **NLL-fast** where this is possible.

All first and second generation squarks are considered; $m_{\tilde{u}_L}$, $m_{\tilde{d}_L}$, $m_{\tilde{s}_L}$, $m_{\tilde{c}_L}$, $m_{\tilde{u}_R}$, $m_{\tilde{d}_R}$, $m_{\tilde{s}_R}$ and $m_{\tilde{c}_R}$. These make up 36 different processes for squark pair production. A separate distributed Gaussian processes estimator (DGP) is trained for each squark process, resulting in $36 \times 10 = 360$ trained experts.

6.2.1 Relative Deviance

The relative deviance distributions for the MSSM-24 datasets are from DGPs consisting of ten experts with 8000 trainings points per expert. For CMSSM each expert uses 5000 training points, as the models had to be stored between training and prediction. Model size is discussed in Sec. 6.3.

MSSM-24

Relative deviance distributions of DGP predictions from the true values for 20 000 test points from the MSSM-24 data set are shown in Fig. 6.2 for selected squark processes. I

The relative deviance distributions for the DGP prediction of left-handed squarks of equal chirality are shown in Fig. 6.2a. All distributions have a mean of approximately zero, and a standard deviation well within the desired value of 10%. The largest cross sections have been excluded from the plots, as the small number of training and test points leave distributions with large uncertainties. These squark processes have only three features, as opposed to four, which appear to make them easier to learn for the DGP. The corresponding right-handed processes show similar results, and plots of relative deviance distributions are shown in Appendix D.

In Fig. 6.2b relative deviance distributions are shown for the process $\tilde{u}\tilde{d}$ for different chirality combinations, which yield different expressions for the cross section as shown in Sec. 3.3.1. The DGPs for $\tilde{d}_L\tilde{u}_L$ are superior to the other chirality combinations, presumably because the strong correlation between $m_{\tilde{d}_L}$ and $m_{\tilde{u}_L}$ — discussed in Sec. 5.1 — means the DGPs effectively train on three features. The DGPs for $\tilde{d}_R\tilde{u}_R$ perform better than for different chiralities, but worse than for $\tilde{u}_L\tilde{d}_L$.

All squark processes in Fig. 6.2b have four features, which appears to make prediction more difficult for the DGPs. Nevertheless, mean relative deviances are close to zero, particularly for cross sections above the 0.02 event limit. Standard deviations are larger than for left-handed equal-flavour chirality processes, but

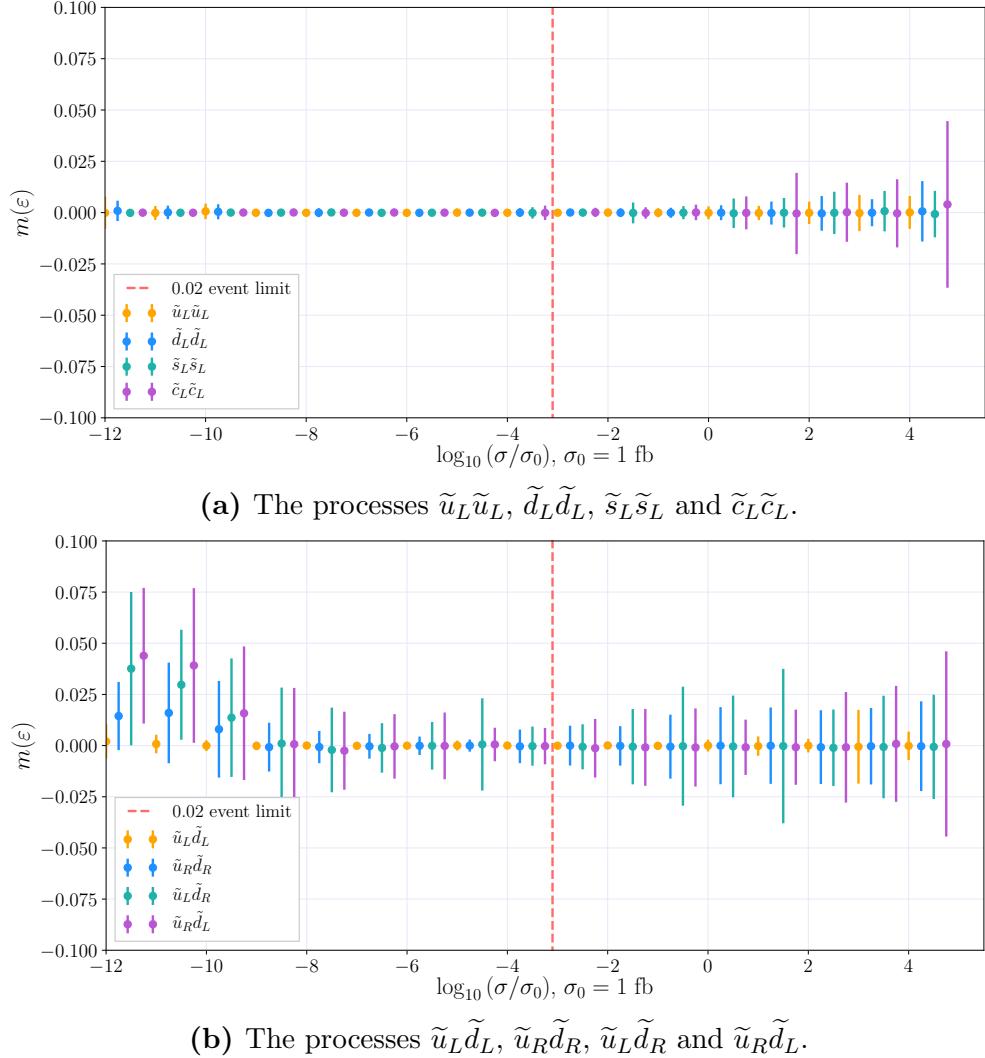


Figure 6.2: Relative deviance distributions as a function of the logarithm of the normalized cross sections $\log_{10} \sigma / \sigma_0$. Ten experts with 8000 training points from the MSSM-24 dataset each were combined on 20 000 test points from the MSSM-24 dataset for processes **(a)** $\tilde{u}_L\tilde{u}_L$, $\tilde{d}_L\tilde{d}_L$, $\tilde{s}_L\tilde{s}_L$ and $\tilde{c}_L\tilde{c}_L$; and **(b)** $\tilde{u}_L\tilde{d}_L$, $\tilde{u}_R\tilde{d}_R$, $\tilde{u}_L\tilde{d}_R$ and $\tilde{u}_R\tilde{d}_L$.

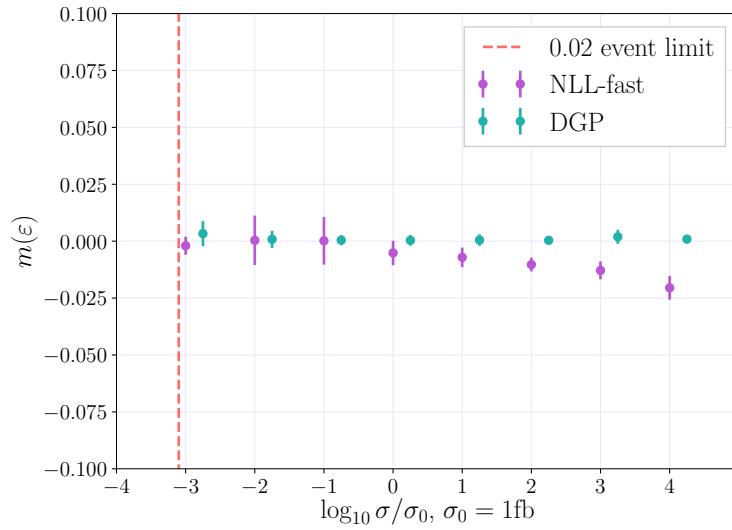


Figure 6.3: Distribution of the relative deviance ε as a function of the logarithm of the total cross section for all squark pair production processes for NLL-fast and Gaussian processes for mSUGRA data. The ‘true’ values are the values generated by Prospino. 10 experts with 5000 training points each were trained on MSSM-24 data for each process $\tilde{q}_i \tilde{q}_j$.

within 2.5% for most cross sections larger than $> 10^{-8}$ fb.

CMSSM

The DGPs trained on MSSM-24 data are also tested on an CMSSM dataset and compared to cross sections calculated by **NLL-fast** 2.1 for the same parameter points. The true values are the sums of cross sections calculated by **Prospino** 2.1 for all 36 processes for each parameter point. These are compared to sums over the corresponding 36 cross sections estimated using the DGPs. **NLL-fast** calculates the total cross section for each parameter point, which is also compared to the **Prospino** 2.1 calculations.

The resulting relative deviance distributions are shown in Fig. 6.3. The cross sections calculated by **NLL-fast** are close to the true values for the CMSSM data. This is because the squark masses in CMSSM have much smaller splittings than in MSSM-24, as discussed in Sec. 2.4.8, and **NLL-fast** assumes degenerate squark masses. For large cross sections, however, **NLL-fast** predicts values that are too large, while the DGPs predict cross sections very close to the true value.

6.2.2 Cross Sections

In this section the distributed Gaussian processes trained on the MSSM-24 dataset are used to predict cross sections as a function of a single mass, either squark or gluino mass, with the remaining masses held fixed. Cross sections for individual processes, such as $\tilde{d}_L \tilde{d}_L$, are plotted as a function of the squark mass. The total cross section for squark pair production is plotted both as a function of the squark mass and the gluino mass. The predictions are then compared to cross sections calculated by **Prospino 2.1** and **NLL-fast** for the same squark and gluino masses.

Cross sections from the distributed Gaussian processes and **Prospino 2.1**, and from **NLL-fast** where these are included, are plotted with uncertainties: **Prospino 2.1** gets uncertainties from the scale dependence described in Sec. 3.4.1, Gaussian process uncertainties are the predictive variances in Eq. 5.24, and **NLL-fast** gets uncertainties from scale dependence, PDFs and α_s .

Cross Sections for Two Processes

Figure 6.4 shows the cross sections for $\tilde{d}_L \tilde{d}_L$ and $\tilde{d}_L \tilde{u}_R$ plotted as functions of the squark mass $m_{\tilde{d}_L}$ for $m_{\tilde{d}_L} \in [200, 2500]$ GeV. The mass splitting with \tilde{u}_L is given by

$$m_{\tilde{d}_L}^2 - m_{\tilde{u}_L}^2 \approx m_W^2, \quad (6.1)$$

for $m_W = 80$ GeV. All other squark masses are held at 1000 GeV, and the gluino mass is held at $m_{\tilde{g}} = 500$ GeV. The cross sections predicted by the distributed Gaussian processes have small variances, as can be seen from the $50\sigma_{\text{std}}$ uncertainty bands plotted in Fig. 6.4. The predictions overlap with **Prospino 2.1** calculations for small $m_{\tilde{d}_L}$, and are slightly — but systematically — underestimated for large $m_{\tilde{d}_L}$.

Total Cross Section Varying the Squark Mass

Figure 6.5 shows the total cross section for squark pair production as a function of $m_{\tilde{d}_L}$, for the same mass values as in Fig. 6.4. The cross sections predicted by the Gaussian processes again coincide with the true values for small $m_{\tilde{d}_L}$, and the underestimation for large masses from Fig. 6.4 is repeated here. In comparison, the **NLL-fast 2.1** cross sections are far from the values calculated by **Prospino 2.1** for all values of $m_{\tilde{d}_L}$, save one. For $m_{\tilde{d}_L} = 1000$ GeV, where the squark masses are in fact degenerate, the calculations from **NLL-fast 2.1** and **Prospino 2.1** coincide.

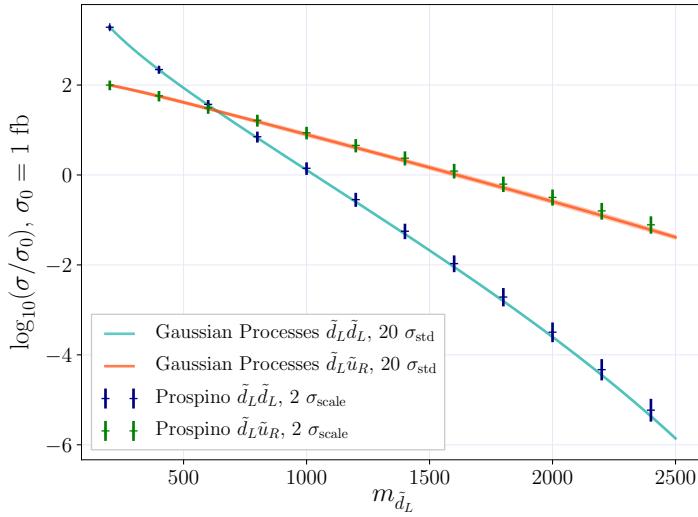


Figure 6.4: Cross sections for $\tilde{d}_L \tilde{d}_L$ and $\tilde{u}_L \tilde{d}_L$, using $m_{\tilde{d}_L} = [200, 2400]$, GeV and all other masses set to $m_i = 1000$ GeV generated by prospino (crosses) and predicted by the GP (lines with errors). The GP models used are for $\tilde{d}_L \tilde{d}_L$ and $\tilde{d}_L \tilde{u}_L$.

Total Cross Section Varying the Gluino Mass

Figure 6.6 shows the total cross section for squark pair production as a function of the gluino mass, for $m_{\tilde{g}} \in [200, 2400]$ GeV. All squark masses are held at 1000 GeV. The uncertainty from Prospino 2.1 is not shown here. Cross sections predicted by the DGPs and calculated using Prospino 2.1 and NLL-fast 2.1 coincide for all values of the gluino mass, which can be expected as the squark masses are degenerate in all points. The uncertainties in the predictions from the distributed Gaussian processes are fairly small, but increase with the gluino mass.

6.3 Optimizing the Model

The distributed Gaussian process predictions with 10 experts using 8000 training points each are very accurate. Unfortunately, the Gaussian process models take up a lot of space when stored. In addition, larger models take longer to predict values. In this section the model sizes and computation times are discussed, to find the optimal model as a compromise between prediction quality, size and computation time.

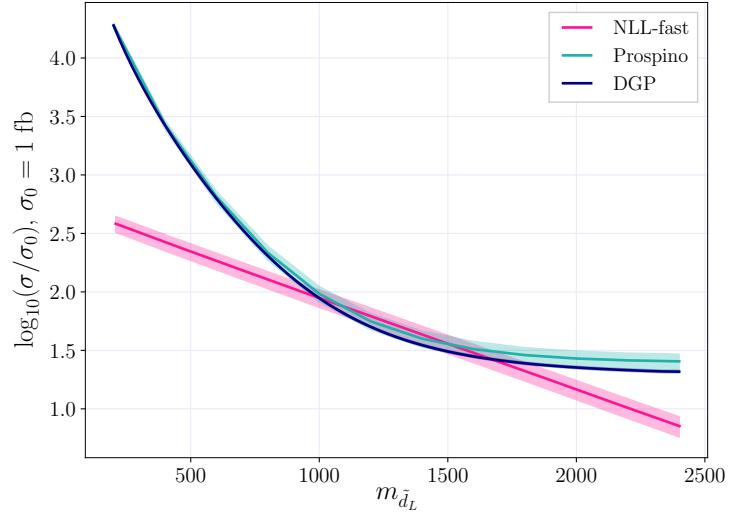


Figure 6.5: Total cross sections for all 36 processes as a function of $m_{\tilde{d}_L}$ calculated by Prospino and NLL-fast, and estimated by DGP. The masses $m_{\tilde{d}_L}$ and $m_{\tilde{u}_L}$ are varied from [200, 2400] GeV while all other squark masses are held at 1000 GeV, and $m_{\tilde{g}} = 500$ GeV.

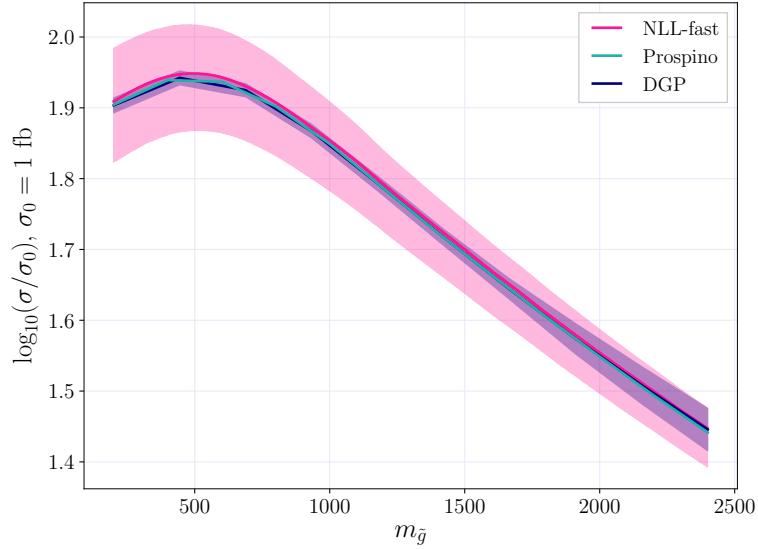


Figure 6.6: Total cross sections for all 36 processes as a function of $m_{\tilde{g}}$ calculated by Prospino and NLL-fast, and estimated by Gaussian processes. The mass m_g is varied [200, 2400] GeV while all squark masses are held at 1000 GeV.

Training data size	Model size [MB]
2000	31
3000	69
5000	191

Table 6.1: Size of saved GP models with 3 or 4 features and the Matérn kernel.

Model Size

Storing the distributed Gaussian processes models requires a lot of memory. A single Gaussian process model trained with 2000 training points takes up 31 MB. Since one model is needed per process, this means that a relatively small model with a single expert with 2000 training points will require

$$36 \times 31 \text{ MB} \approx 1.09 \text{ GB}. \quad (6.2)$$

As seen from Table 6.1 the model size scales approximately as $\mathcal{O}(n^2)$ for a model with n training points. For M experts with n training points each, the model size then scales as $\mathcal{O}(M \cdot n^2)$. Where it is possible it is therefore advisable to add *more* experts as opposed to *larger* experts.

The model size puts limits on the number of training points an optimal model should have. There is very little difference in model sizes for models with 3 and 4 features, and a difference of 29 bytes for all training sizes between using the RBF kernel and the Matérn kernel. A model with 36 processes with 10 experts each would require

$$10 \times 36 \times 191 \text{ MB} = 67.15 \text{ GB}.$$

By comparison, reducing the experts to 3000 training points each would require

$$10 \times 36 \times 69 \text{ MB} = 24.26 \text{ GB}.$$

Comparisons of the relative deviance distributions with 3000, 5000 and 8000 training points per expert for the processes $\tilde{d}_L \tilde{d}_L$ and $\tilde{d}_L \tilde{u}_R$ are shown in Fig. 6.7. The improvement in the prediction for $\tilde{d}_L \tilde{u}_R$ from 3000 to 8000 training points per expert is large for all σ . Processes with different squarks and chiralities should therefore use experts with as many training points as possible, as the main contribution to the prediction error will come from these processes. In contrast, the process $\tilde{d}_L \tilde{d}_L$ is modelled very accurately even for 3000 training points per expert. For cross sections larger than 10^2 fb there is a visible improvement from 3000 to 8000 training points per expert, likely rooted in the addition of

data³ in the form of larger experts. For 3000 training points per expert this could be remedied simply by adding more experts trained on large cross sections. In conclusion, processes with equal flavour and equal chirality can use smaller experts, while processes with different flavour or different chirality benefit from having as large experts as possible.

Number of Models

In this project 36 Gaussian process models were trained with ten experts each, resulting in 360 experts. The effective number of models may be reduced, however, as many of the models are almost equivalent.

Equal-flavour equal-chirality (EFEC) processes are, to leading order, identical as functions of the squark mass, as shown in Sec. 3.3.1. This is because **Prospino** calculates cross sections for strong interactions, and the NLO terms only contain QCD corrections, as discussed in Sec. 3.4. Since there is no electroweak correction, there is no distinction between left- and right-handed squarks in the calculation. They also share the same underlying pdf from the quark, *e.g.* the pdf's used for calculating $\tilde{d}_R \tilde{d}_R$ and $\tilde{d}_L \tilde{d}_L$ is the one for the d -flavour squark. The number of models can therefore be reduced from 36 to 32.

Further, the masses of different generation left-handed squarks are also independent, meaning that the process $\tilde{u}_L \tilde{s}_L$ is, to leading order, identical to $\tilde{u}_R \tilde{s}_R$ as a function of $m_{\tilde{u}_L}$ and $m_{\tilde{s}_L}$, and $m_{\tilde{u}_R}$ and $m_{\tilde{s}_R}$, respectively. The number of models is therefore further reduced from 32 to 28.

Processes with different flavours and different chiralities are also identical to leading order, such as $\tilde{d}_L \tilde{u}_R$ and $\tilde{u}_L \tilde{d}_R$ as functions of $m_{\tilde{d}_L}$ and $m_{\tilde{d}_R}$, respectively. In Fig. 6.8 the models trained on $\tilde{d}_L \tilde{u}_R$ and $\tilde{u}_L \tilde{d}_R$ are used to predict cross sections as a function of $m_{\tilde{d}_L}$ and $m_{\tilde{d}_R}$, respectively. The relevant d -mass was varied in the interval [200, 2400] GeV, while the remaining seven squark masses were held at 1000 GeV. The cross sections overlap almost completely, with differences in the variances seen first when $100\sigma_{\text{std}}$ is plotted. As there are 12 such processes, six can be removed. The number of models is then reduced from 28 to 22.

It is worth stressing that the processes are *not identical* — the leading order cross sections are indeed identical, but the next-to-leading order cross sections depend on the mean mass. The mean mass is different for *e.g.* $\tilde{d}_L \tilde{d}_L$ and $\tilde{d}_R \tilde{d}_R$ if the relevant mass is very different from the other squark masses, as $\tilde{d}_L \tilde{d}_L$ is trained on points where $m_{\tilde{d}_L}$ and $m_{\tilde{u}_L}$ are close in value, and $\tilde{d}_R \tilde{d}_R$ is trained on points where $m_{\tilde{d}_R}$ is independent of $m_{\tilde{u}_R}$. This is, however, a relatively small effect, so the effective number of models described above can be used to a good approximation.

In addition, the models for EFEC processes, and $\tilde{d}_L \tilde{u}_L$ and $\tilde{c}_L \tilde{s}_L$ perform better than the other processes for fewer training points. It may therefore not be

³The endpoints are usually not well covered

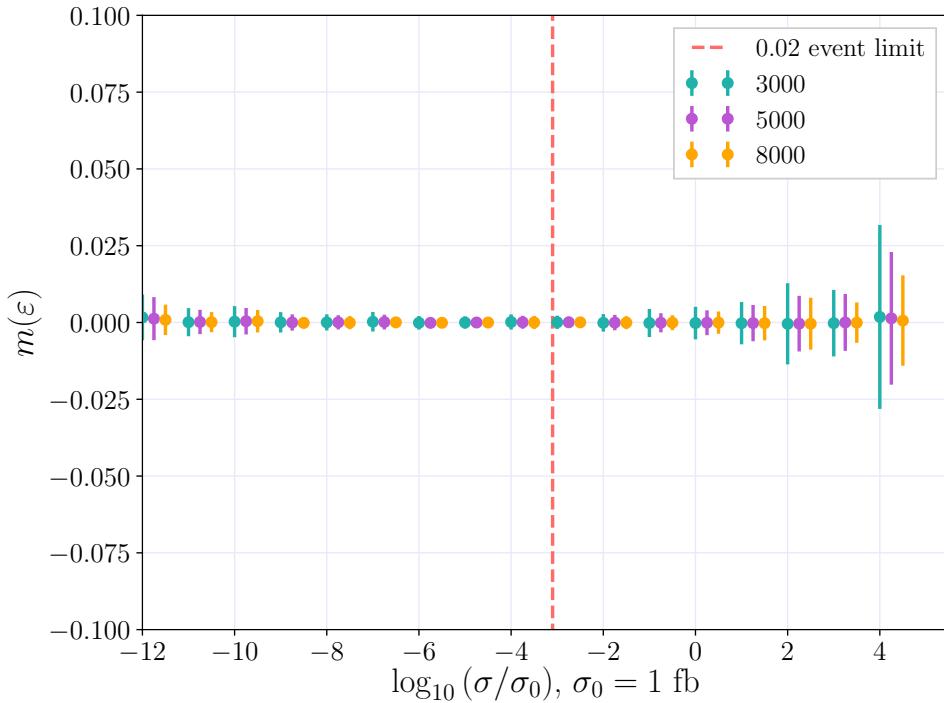
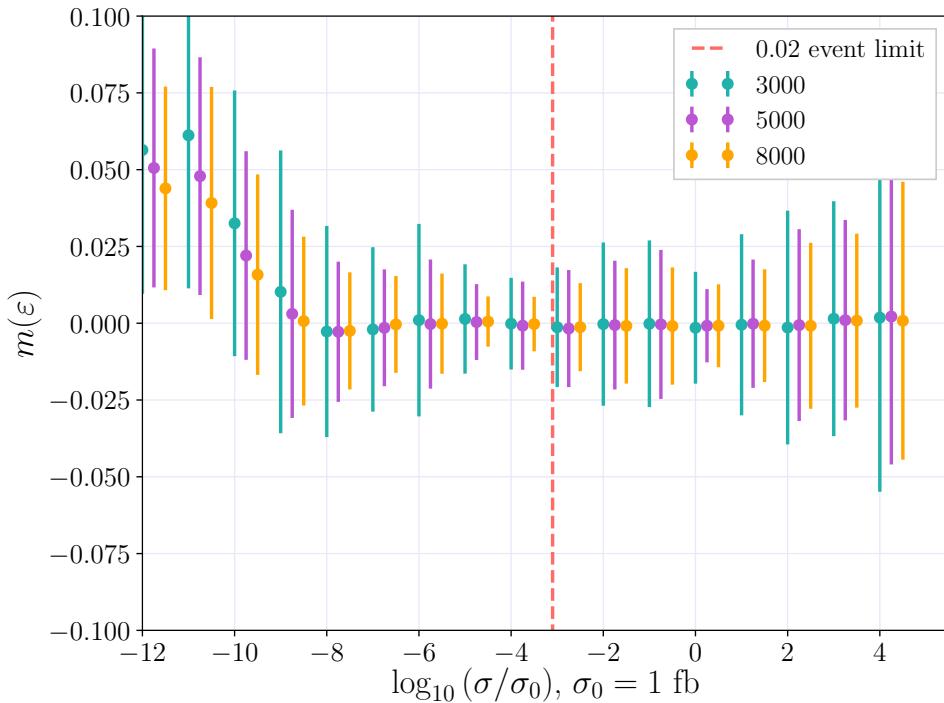
(a) The process $\tilde{d}_L \tilde{d}_L$.(b) The process $\tilde{d}_L \tilde{u}_R$.

Figure 6.7: Relative deviance distributions as a function of the logarithm of the normalized cross sections $\log_{10} \sigma / \sigma_0$. Ten experts with 3000 (green), 5000 (violet) and 8000 (orange) training points from the MSSM-24 dataset each were combined on 20 000 test points from the MSSM-24 dataset for processes (a) $\tilde{d}_L \tilde{d}_L$ and (b) $\tilde{d}_L \tilde{u}_R$.

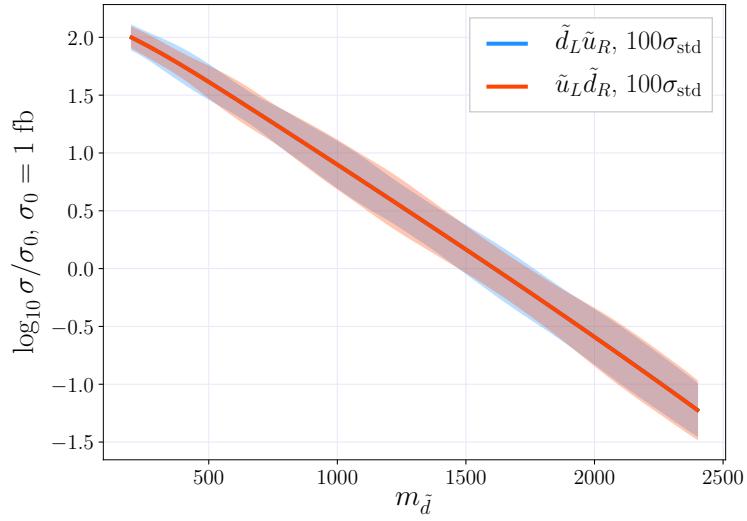


Figure 6.8: The predicted cross sections for $\tilde{d}_L \tilde{u}_R$ as a function of $m_{\tilde{d}_L}$ (blue) and $\tilde{u}_L \tilde{d}_R$ as a function of $m_{\tilde{d}_R}$ (red), with $100\sigma_{\text{std}}$ uncertainty bands. The cross sections were estimated for $m_{\tilde{d}_L} = [200, 2400]$ GeV and $m_{\tilde{d}_R} = [200, 2400]$ GeV, respectively, and all other squark masses were held at 1000 GeV. The gluino mass is $m_{\tilde{g}} = 500$ GeV. The cross sections completely overlap.

necessary to use 10 experts for each of these processes. In Fig. 6.1a the learning curve for $\tilde{d}_L \tilde{d}_L$ with 2000 training points per expert shows little difference in validation score for more than seven experts. For larger experts it could therefore be sufficient with *e.g.* 4 experts. Reducing the number of experts for all EFEC processes, and $\tilde{d}_L \tilde{u}_L$ and $\tilde{c}_L \tilde{s}_L$, and using the equal-flavour left-handed models on equal-flavour right-handed processes reduces the number of experts to

$$4 \times 4 \text{ experts} + 2 \times 4 \text{ experts} + 26 \times 10 \text{ experts} = 280 \text{ experts.}$$

The minimal set of models needed is therefore

$$\begin{array}{ccccc}
 \tilde{u}_L \tilde{u}_L & \tilde{u}_L \tilde{u}_R & \tilde{u}_L \tilde{d}_L & \tilde{u}_L \tilde{s}_L & \tilde{u}_L \tilde{d}_R \\
 \tilde{d}_L \tilde{d}_L & \tilde{d}_L \tilde{d}_R & \tilde{u}_R \tilde{d}_R & \tilde{u}_L \tilde{c}_L & \tilde{u}_L \tilde{s}_R \\
 \tilde{s}_L \tilde{s}_L & \tilde{s}_L \tilde{s}_R & \tilde{s}_L \tilde{c}_L & \tilde{d}_L \tilde{s}_L & \tilde{u}_L \tilde{c}_R \\
 \tilde{c}_L \tilde{c}_L & \tilde{c}_L \tilde{c}_R & \tilde{s}_R \tilde{c}_R & \tilde{d}_L \tilde{c}_L & \tilde{d}_L \tilde{s}_R \\
 & & & & \tilde{d}_L \tilde{c}_R \\
 & & & & \tilde{s}_L \tilde{c}_R.
 \end{array}$$

Computation Times

Training 10 experts with 5000 training points each, in parallel, takes approximately 40 minutes for a single process. The prediction time for the DGP was calculated by letting the model predict values for 2000 test points for a single process, and dividing the total computation time by 2000. The average computation time for predicting a single point for a single process is

$$0.46925 \text{ s.} \quad (6.3)$$

The computation time for predicting all 36 cross sections for a single parameter point is therefore

$$36 \times 0.46925 \text{ s} = 16.893 \text{ s.} \quad (6.4)$$

In Tab. 6.2 the prediction times for all 36 process cross sections for a single parameter point are shown for **Prospino** and distributed Gaussian processes. For **Prospino** three cross sections were calculated for each process, with scale factors 0.5, 1.0 and 2.0, to include the uncertainty from scale dependence. The time for **NLL-fast** is also shown for one parameter point. **NLL-fast** is considerably faster than both DGP and **Prospino**, but has a very large error relative to the **Prospino** computation, as seen in Fig. 6.5. Although it is much slower than **NLL-fast**, the DGP is faster than **Prospino** by a factor of approximately 61.

The prediction of each DGP expert is *not* done in parallel. Parallelising the prediction algorithm could reduce the computational time in Tab. 6.2 by a factor of approximately 10

$$16.893 \text{ s} : 10 = 1.689 \text{ s.}$$

The prediction of each of the 36 squark production processes was also done in sequence. Predicting for each process in parallel could further reduce the computation time by a factor of 36

$$1.689 \text{ s} : 36 = 0.0469 \text{ s.}$$

Note that these are idealized times, meant to illustrate how the current algorithm can be improved. The prediction time of Gaussian processes goes as $\mathcal{O}(n^2)$ for n training points, so using smaller experts would also reduce the computation time.

Tool	Computation time [s]
Prospino	1711.96
NLL-fast	0.00739
Distributed Gaussian Processes	16.893

Table 6.2: Computation times for 1 parameter point for all 36 squark pair production processes.

Appendices

Appendix A

Notation

A.1 Quantum Field Theory

Indices running over four values are represented by letters from the Greek alphabet, $\mu, \nu, \gamma, \rho, \dots = 0, 1, 2, 3$. The four-vector for position and momentum of a particle are given by

$$x^\mu = (t, \mathbf{x}), \quad p^\mu = (E, \mathbf{p}), \quad (\text{A.1})$$

and the four-vector derivative is

$$\partial_\mu = (\partial/\partial t, \boldsymbol{\nabla}). \quad (\text{A.2})$$

The spacetime metric is

$$\eta_{\mu\nu} = \text{diag}(+1, -1, -1, -1), \quad (\text{A.3})$$

so that $m^2 = p^2$ for an on-shell particle of mass m .

The right- and left handed projection operators are

$$P_R = \frac{1 + \gamma^5}{2}, \quad P_L = \frac{1 - \gamma^5}{2}, \quad (\text{A.4})$$

where γ^5 is defined as

$$\gamma^5 = i\gamma^0\gamma^1\gamma^2\gamma^3, \quad (\text{A.5})$$

where γ^μ are the Dirac matrices

$$\gamma^\mu = \begin{pmatrix} 0 & \sigma^\mu \\ \bar{\sigma}^\mu & 0 \end{pmatrix}, \quad (\text{A.6})$$

where

$$\sigma^0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \sigma^1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma^2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma^3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad (\text{A.7})$$

and $\bar{\sigma}^0 = \sigma^0$ and $\bar{\sigma}^i = -\sigma^i$, $i = 1, 2, 3$.

Dirac Spinor

The Dirac spinor is written in terms of 2 two-component, anti-commuting, complex objects

$$\Phi_D = \begin{pmatrix} \xi_\alpha \\ \chi^{\dagger\dot{\alpha}} \end{pmatrix}, \quad \bar{\Phi}_D = \Phi_D^\dagger \sigma^1 = (\chi^\alpha, \xi_{\dot{\alpha}}^\dagger), \quad (\text{A.8})$$

where undotted indices, α , are used for the first two components of a Dirac spinor, and dotted indices, $\dot{\alpha}$, are used for the last two. The field ξ is a *left-handed Weyl spinor*, and χ^\dagger is a *right-handed Weyl spinor*, and are projected out from the Dirac spinor using the projection operators in Eq. (A.4).

A.2 Statistics

The expectation value of a quantity X following a probability distribution $P(X)$ is denoted $\mathbb{E}[X]$ and defined as

$$\mathbb{E}[X] = \int P(X)X \, dX \quad (\text{A.9})$$

Appendix B

Data Quality

Data quality plots for all squark masses are found here. SETT INN!

Appendix C

Benchmark for Distributed Gaussian Processes

The benchmark function for parallelised distributed Gaussian processes is

$$f(x_1, x_2) = 4x_1x_2,$$

where the vectors $\mathbf{x} = (x_1, x_2)$ were drawn from a random normal distribution using the `numpy` function `random.randn`. Gaussian processes implemented by `scikit-learn` in the function `GaussianProcessRegressor` were compared to distributed Gaussian processes with 4 experts. 2000 training points and 1000 test points were used, and the resulting times for the GP and DGP were

$$\text{Gaussian processes time: } 154.12 \text{ s} \quad (\text{C.1})$$

$$\text{Distributed Gaussian processes time: } 5.61 \text{ s} \quad (\text{C.2})$$

Histograms of the relative deviances for Gaussian processes (GP) and Distributed Gaussian processes (DGP) are found in Fig. (C.1).

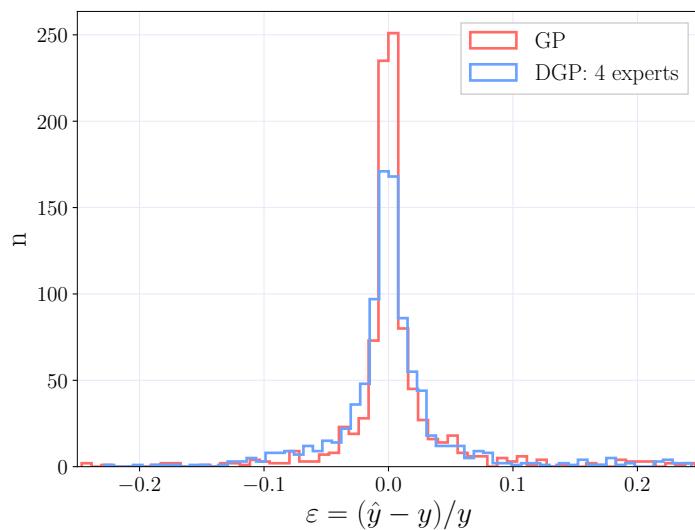


Figure C.1: Histogram of the relative deviance between true value y and predicted value \hat{y} for Gaussian process regression (GP) and Distributed gaussian process regression (DGP) for the function $f(x_1, x_2) = 4x_1x_2$.

Appendix D

Relative Deviance Distributions

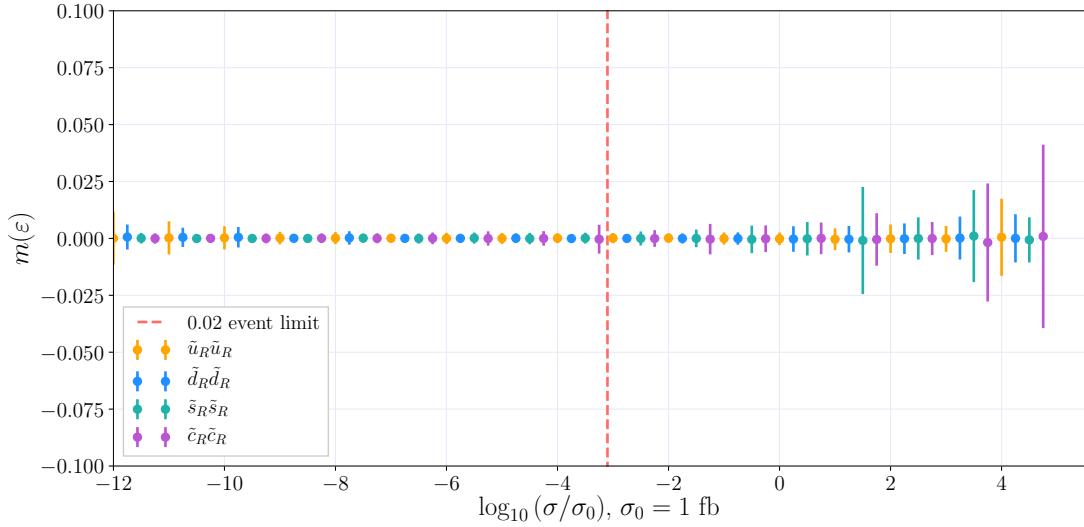


Figure D.1: Relative deviance distributions for $\tilde{u}_R \tilde{u}_R$, $\tilde{d}_R \tilde{d}_R$, $\tilde{s}_R \tilde{s}_R$ and $\tilde{c}_R \tilde{c}_R$.

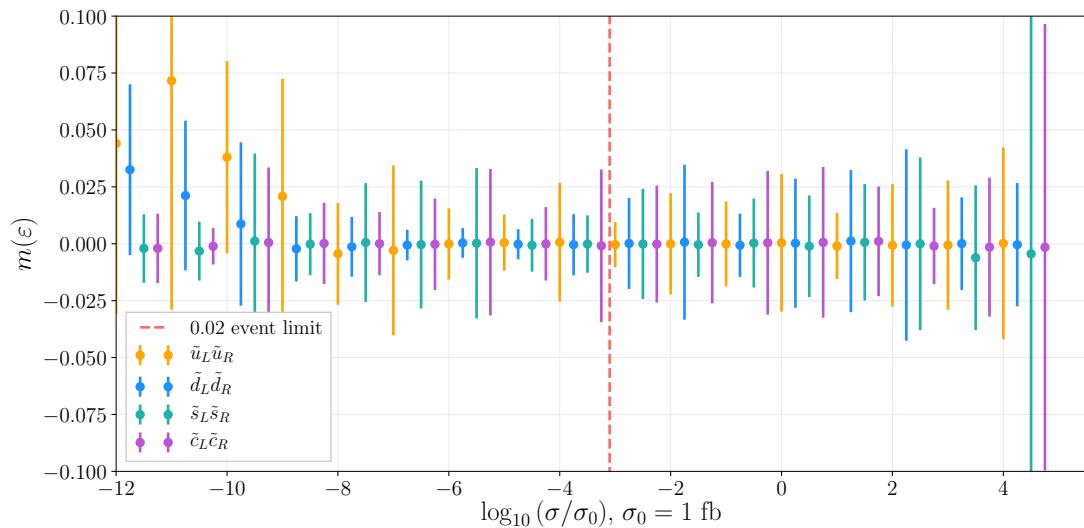


Figure D.2: Relative deviance distributions for $\tilde{u}_L \tilde{u}_R$, $\tilde{d}_L \tilde{d}_R$, $\tilde{s}_L \tilde{s}_R$ and $\tilde{c}_L \tilde{c}_R$.

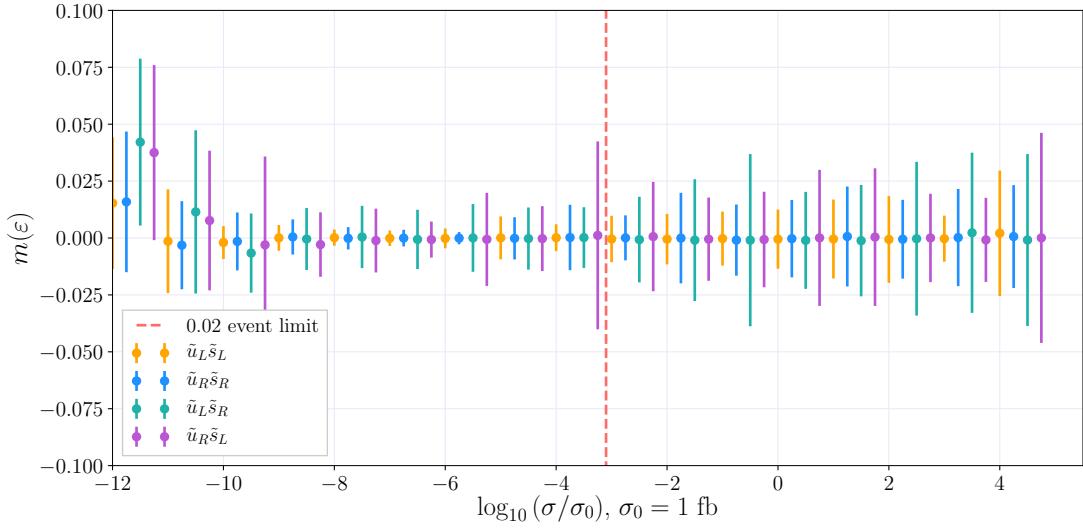


Figure D.3: Relative deviance distributions for $\tilde{u}_L \tilde{s}_L$, $\tilde{u}_R \tilde{s}_R$, $\tilde{u}_L \tilde{s}_R$ and $\tilde{u}_R \tilde{s}_L$.

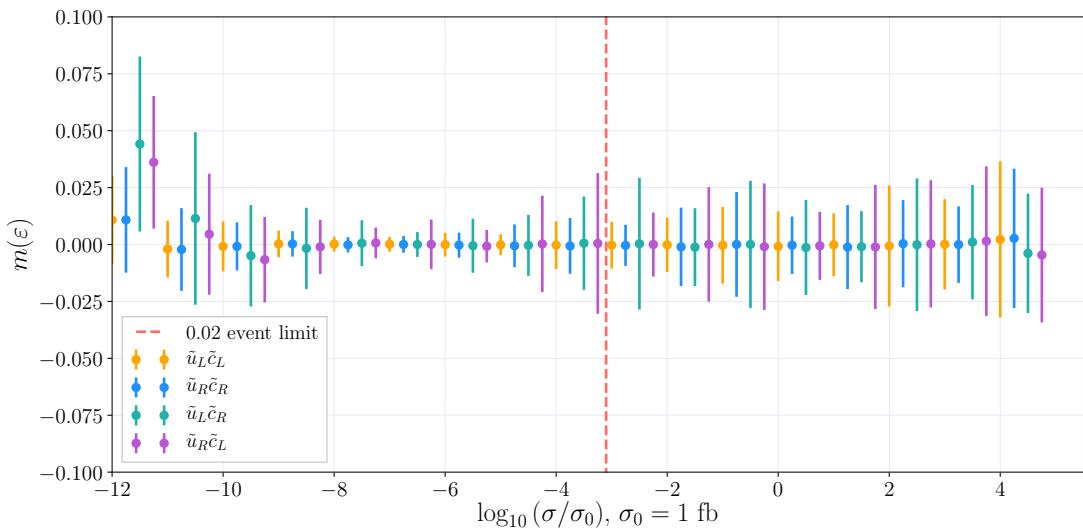


Figure D.4: Relative deviance distributions for $\tilde{u}_L \tilde{c}_L$, $\tilde{u}_R \tilde{c}_R$, $\tilde{u}_L \tilde{c}_R$ and $\tilde{u}_R \tilde{c}_L$.

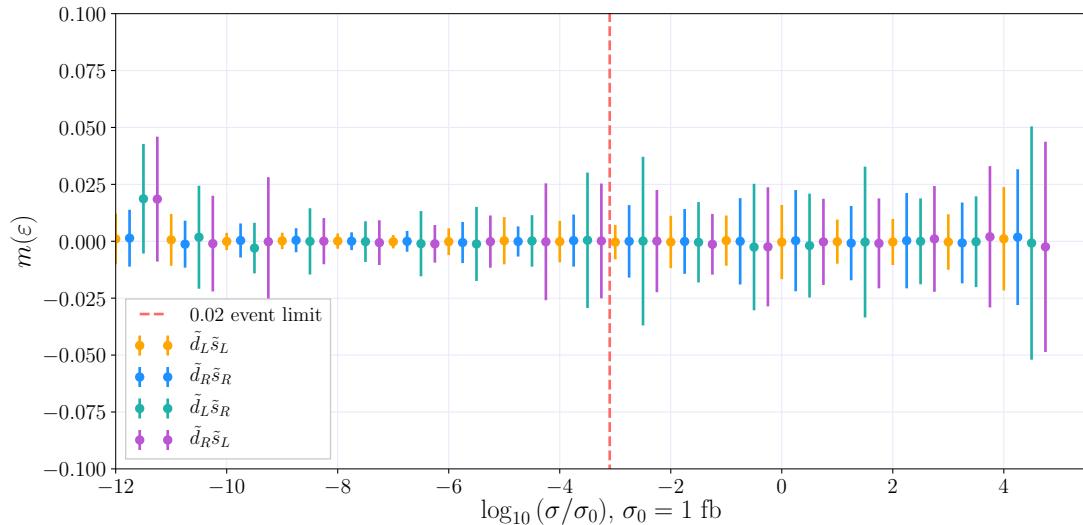


Figure D.5: Relative deviance distributions for $\tilde{d}_L \tilde{s}_L$, $\tilde{d}_R \tilde{s}_R$, $\tilde{d}_L \tilde{s}_R$ and $\tilde{d}_R \tilde{s}_L$.

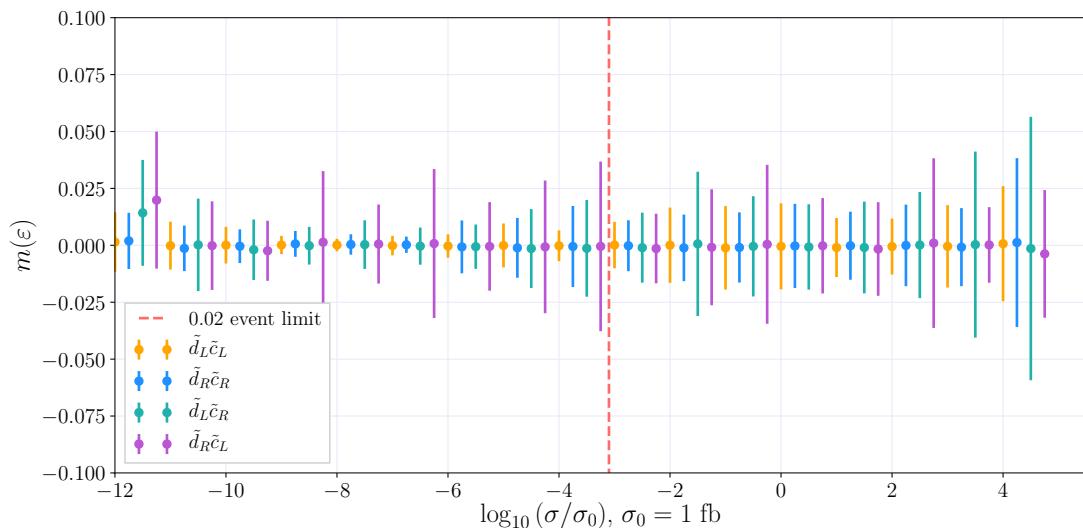


Figure D.6: Relative deviance distributions for $\tilde{d}_L \tilde{c}_L$, $\tilde{d}_R \tilde{c}_R$, $\tilde{d}_L \tilde{c}_R$ and $\tilde{d}_R \tilde{c}_L$.

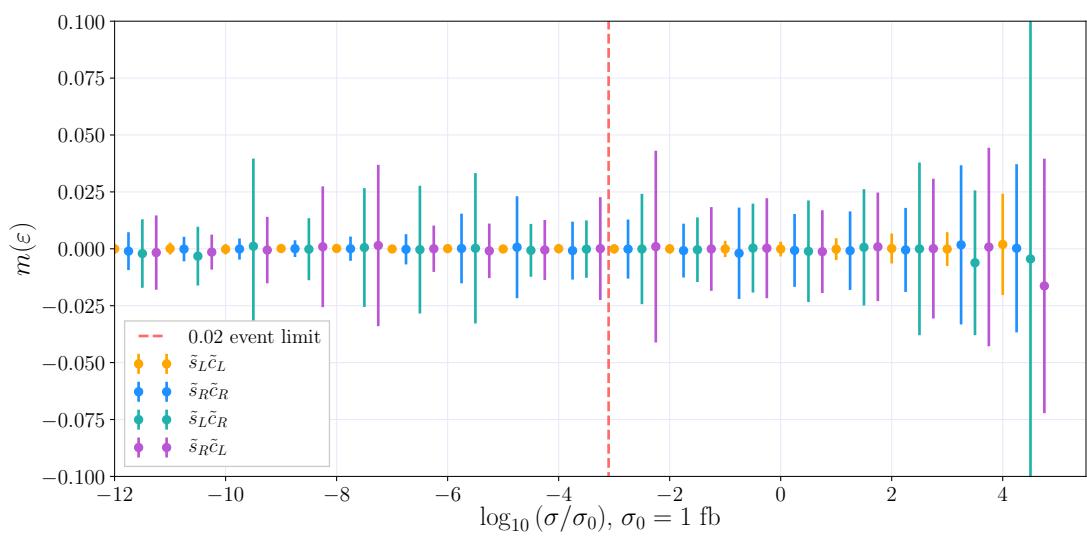


Figure D.7: Relative deviance distributions for $\tilde{s}_L \tilde{c}_L$, $\tilde{s}_R \tilde{c}_R$, $\tilde{s}_L \tilde{c}_R$ and $\tilde{d}_R \tilde{c}_L$.

Bibliography

- [1] P. Batzing and A. Raklev, *Lecture notes for fys5190/fys9190*, .
- [2] A. Kvellestad, *Chasing susy through parameter space*, .
- [3] S. P. Martin, *A supersymmetry primer*, in *Perspectives on supersymmetry II*, pp. 1–153. World Scientific, 2010.
- [4] S. Weinberg, *The Quantum Theory of Fields*, vol. 1. Cambridge University Press, 1995, [10.1017/CBO9781139644167](https://doi.org/10.1017/CBO9781139644167).
- [5] A. Balerna and S. Mobilio, *Introduction to synchrotron radiation*, in *Synchrotron radiation*, pp. 3–28. Springer, 2015.
- [6] P. M. Nadolsky, H.-L. Lai, Q.-H. Cao, J. Huston, J. Pumplin, D. Stump et al., *Implications of cteq global analysis for collider observables*, *Phys. Rev. D* **78** (Jul, 2008) 013004.
- [7] M. Aaboud, G. Aad, B. Abbott, J. Abdallah, O. Abdinov, B. Abeloos et al., *Search for squarks and gluinos in final states with jets and missing transverse momentum at $\sqrt{s} = 13\text{ TeV}$ with the atlas detector*, *The European Physical Journal C* **76** (2016) 392.
- [8] R. K. Ellis, W. J. Stirling and B. R. Webber, *QCD and collider physics*. Cambridge university press, 2003.
- [9] W. Beenakker, R. Höpker, M. Spira and P. Zerwas, *Squark and gluino production at hadron colliders*, *Nuclear Physics B* **492** (1997) 51–103.
- [10] G. P. Lepage, *A new algorithm for adaptive multidimensional integration*, *Journal of Computational Physics* **27** (1978) 192 – 203.
- [11] W. Beenakker, R. Höpker and M. Spira, *Prospino: a program for the production of supersymmetric particles in next-to-leading order qcd*, *arXiv preprint hep-ph/9611232* (1996) .
- [12] W. Beenakker, C. Borschensky, M. Krämer, A. Kulesza, E. Laenen,

- S. Marzani et al., *Nlo+ nll squark and gluino production cross sections with threshold-improved parton distributions*, *The European Physical Journal C* **76** (2016) 53.
- [13] C. Balázs, A. Buckley, L. A. Dal, B. Farmer, P. Jackson, A. Krislock et al., *Colliderbit: a gambit module for the calculation of high-energy collider observables and likelihoods*, *The European Physical Journal C* **77** (2017) 795.
 - [14] A. Kulesza and L. Motyka, *Threshold resummation for squark-antisquark and gluino-pair production at the lhc*, *Physical review letters* **102** (2009) 111802.
 - [15] A. Kulesza and L. Motyka, *Soft gluon resummation for the production of gluino-gluino and squark-antisquark pairs at the lhc*, *Physical Review D* **80** (2009) 095004.
 - [16] W. Beenakker, S. Brensing, M. Krämer, A. Kulesza, E. Laenen and I. Niessen, *Soft-gluon resummation for squark and gluino hadroproduction*, *Journal of High Energy Physics* **2009** (2009) 041.
 - [17] W. Beenakker, S. Brensing, M. Krämer, A. Kulesza, E. Laenen, L. Motyka et al., *Squark and gluino hadroproduction*, *International Journal of Modern Physics A* **26** (2011) 2637–2664.
 - [18] A. D. Martin, W. J. Stirling, R. S. Thorne and G. Watt, *Parton distributions for the LHC*, *Eur. Phys. J.* **C63** (2009) 189–285, [[0901.0002](#)].
 - [19] M. Bayes and M. Price, *An essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, frs communicated by mr. price, in a letter to john canton, amfrs, Philosophical Transactions (1683-1775)* (1763) 370–418.
 - [20] D. Sivia and J. Skilling, *Data analysis: a Bayesian tutorial*. OUP Oxford, 2006.
 - [21] P. S. Laplace, *Théorie analytique des probabilités*. Courcier, 1820.
 - [22] M. Stein, “Interpolation of spatial data: some theory for kriging. 1999.”
 - [23] M. Abramowitz and I. A. Stegun, *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*, vol. 55. Courier Corporation, 1964.
 - [24] C. E. Rasmussen and C. K. Williams, *Gaussian processes for machine learning*, vol. 1. MIT press Cambridge, 2006.

- [25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel et al., *Scikit-learn: Machine learning in Python*, *Journal of Machine Learning Research* **12** (2011) 2825–2830.
- [26] J. V. Sparre, *Fast evaluation of supersymmetric cross sections*, .
- [27] B. Allanach, *Softsusy: A program for calculating supersymmetric spectra*, *Computer Physics Communications* **143** (2002) 305 – 331.
- [28] P. Skands, B. C. Allanach, H. Baer, C. Balázs, G. Bélanger, F. Boudjema et al., *Susy les houches accord: Interfacing susy spectrum calculators, decay packages, and event generators*, *Journal of High Energy Physics* **2004** (2004) 036.
- [29] M. P. Deisenroth and J. W. Ng, *Distributed gaussian processes*, *arXiv preprint arXiv:1502.02843* (2015) .
- [30] V. Tresp, *A bayesian committee machine*, *Neural computation* **12** (2000) 2719–2741.