Applications of Higher-Order Quadrature Methods to Econometric Models and Estimators

Alexandros Philipp Gilch

Born December 8, 1995 in Bremen, Deutschland

30th April 2020

Master's Thesis Mathematics

Advisor: Prof. Dr. Michael Griebel

Second Advisor: Dr. Bastian Bohn

Institut für Numerische Simulation

Mathematisch-Naturwissenschaftliche Fakultät der

Rheinischen Friedrich-Wilhelms-Universität Bonn

Contents

1 Introduction

Empirical research in economics is concerned with quantifying economic relationships and empirically evaluating economic hypotheses and models with statistical (or *econometric*) methods. Ranging from data analysis to the creation and assessment of complex dynamic models, econometricians use a wide spectrum of mathematical, statistical and computational techniques to combine observations with theory. Therefore, econometrics constitutes a basic pillar for the utilization of economic theory by policy makers like central banks and national governments.
The standard approach for evaluating data of an economic phenomenon can be separated in two steps: Firstly, a model has to be formalized which includes important variables and characterizes a mathematical relation between them. The structural approach in economic modeling tries to identify causal relationships or mathematically describable characteristics in the examined phenomenon and assumes that the resulting model correctly specifies this phenomenon. As this assumption can usually not be justified rigorously, quasi- or semi-structural models are used which view structural assumptions as approximation or abstraction of reality or leave some features of the phenomenon unspecified ([7], [42]). Non-structural models rely even less on structural assumptions and only analyze correlation between the observed variables [77].
In contemporary econometrics, neither structural nor non-structural models are considered to be deterministic. Therefore, the observable data is treated as realizations of random variables. Furthermore, unobservable variables are introduced which account for errors and noise in measurement or for unspecified structural components. Together this implies that most currently used econometric models are statistical models.
Three types of variables can be encountered in an econometric model, which can be explained exemplary by considering a prototypical linear model,

$$Y = X\,\beta + \varepsilon\,. \tag{1.1}$$

It contains a *dependent variable* or *regressand* $Y \in \mathbb{R}$ (usually the outcome of an economic action) and the *explanatory variables* or *regressors* $X \in \mathbb{R}^{1\times q}$, $\varepsilon \in \mathbb{R}$ (usually some features or inputs to the acting individual). Here, X and Y are observable to the researcher, i.e. he can observe realizations $\{x_n, y_n \mid n = 1, ..., N\}$ of the random variables X and Y, while ε is unobservable with known or unknown distribution F.
Another design decision regards the parametrization of the proposed model: The linear model (1.1) is *parametric* with respect to the finite-dimensional parameter $\beta \in \mathbb{R}^q$. Leaving F unspecified yields a *non-parametric* model as the search space for the parameter F is infinite-dimensional. Often, this is avoided by assuming that F constitutes a Gaussian (or some other parameterized) distribution which is fully characterized by two finite-dimensional

parameters, the mean and the variance.

Once the structure and relevant variables for the model have been determined, the second step is to *estimate* the proposed model, i.e. evaluate it with the observed data. Within the given structure, an *estimator* is optimal if it gives the "best" specification of the model for the observed data. Common metrics for the quality of an estimator are unbiasedness, consistency, asymptotic normality, efficiency and robustness.

Various estimators have been designed for finding the optimal parameters. The most popular estimators, in particular Least Squares, Maximum Likelihood and Generalized Method of Moments estimators, are tailored to parametric models and can be summarized in the class of *extremum estimators* [45]. For collected data points $z_n = (x_n, y_n)$, these estimators maximize an objective function Q_N which is constructed from the data $z_1, ..., z_N$.

Different estimators have different requirements regarding the distribution of observable and unobservable variables, especially whether the samples/ observations are independent or correlated. In fact, often model assumptions are made in order to allow the estimation using a particularly well-behaved estimator. For example, the unobservable variable ε in (1.1) is usually assumed to have mean 0 and constant variance for all individuals and to be uncorrelated between individuals in order to allow estimation with the *best linear unbiased estimator*, the Least Squares estimator.

On the other hand, with growing and less expensive computational power, more difficult and comprehensive models became feasible in the past decades. This encouraged the development of ever more complex designs where both, models and estimators, involve complex mathematical expressions like differential equations, dynamic problems or multivariate integrals. In this context, the increasing demand for fast converging and precise approximation techniques led to a variety of numerical methods enabling the evaluation of multidimensional problems for many individuals, countries, firms and time-periods.

For example, the objective function Q_N of Maximum Likelihood and Generalized Method of Moments estimators can be interpreted as an approximation of an integral over the data space. Similarly, the unobservable variables in a model are usually integrated out to obtain a mean prediction which leads to multidimensional integrals. These integrals are typically only tractable if strong assumptions on the model are made, otherwise they need to be approximated.

As most intractable integrals in econometrics emerge from unobserved stochastic components, a classic remedy is to simulate those stochastic components using Monte Carlo techniques ([39], [84]). Then, the integral is approximated with a quadrature rule using the simulated values (x_n) as basis, i.e.

$$Q_N(f) := \sum_{n=1}^{N} w_n f(x_n) \approx \mathcal{I}(f) := \int_\Omega f(x)dx \tag{1.2}$$

for some integrand f on the domain $\Omega \subset \mathbb{R}^d$. With uniform weights $w_n = 1/N$ for all $n = 1, ..., N$, this leads to a robust but only rather slowly converging approximation with probabilistic error rate $O(N^{-1/2})$. However, the great advantage of this rate is its independence of the dimension d. The canonical extension of one-dimensional rules, the product rule, succumbs to the curse of dimensionality [3]: For an r-times differentiable integrand, its convergence rate $O(N^{-r/d})$ decreases exponentially in d and is therefore not competitive for high-dimensional integrals.
Higher-order algorithms like Sparse Grid (SG) [24] or Quasi Monte Carlo (QMC) quadrature ([12], [53]) may replace classic simulation techniques and help to break the curse of dimensionality. Depending on the regularity r of f, they achieve deterministic error rates $O(N^{-r}\log(N)^{t(r,d)})$ which are only impaired by the secondary rate $\log(N)^{t(r,d)}$ with some scalar $t(r,d) > 0$ and can therefore outperform Monte Carlo simulations. While Quasi Monte Carlo quadrature aims at finding seemingly random but more evenly distributed nodes in Ω, Sparse Grid quadrature strategically omits most of the nodes of the product rule to reduce computational effort.
The Sparse Grid technique has also been used in other fields of numerical simulation, e.g. for partial differential equations, molecular dynamics or machine learning ([9], [32], [88]). In general, it could be applied whenever a high-dimensional net or tensor product causes computational problems due to the exponential increase in the number of nodes or cross-products.

Given the available numerical quadrature algorithms and the need for accurate and fast integral approximations in econometrics, this thesis seeks to provide a systematic overview of occurrences of multidimensional integrals in econometrics. Once the necessary background is introduced, an analysis of the integrands for regularity clarifies whether and when higher-order quadrature rules can improve computability of econometric models.
Developing such a comprehensive survey is difficult as econometric theory consists of a wide variety of custom-made tools and techniques. We begin this endeavor in Part I of this thesis by generalizing existing works on SG quadrature in econometrics. As result, parameterized sets of integrals emerge as the correct framework for econometric quadrature, where the optimal formula has to be identified depending on the parameter specification. Based on the works of Heiss and Wintschel ([47], [48]), Judd and Skrainka [50] and Griebel and Oettershagen ([33], [34]) we derive parameterized sets

of integrals for Generalized Linear Mixed models (GLMM) of the form

$$\mathcal{I}(y|\,\beta_0, \Sigma, \phi) = \int_{\mathbb{R}^q} f(y|\,\beta_0, u, \phi) h(u|\Sigma) du \tag{1.3}$$

and analyze them in terms of regularity. Here, f constitutes a distribution from an exponential family and h some distribution which specifies correlation between the unobservable variables u.

We find that regularity assumptions of higher-order quadrature rules are usually fulfilled as f and h comprise smooth probability density functions (p.d.f.) for almost all model specifications. Yet, the constant in the O-notation of the convergence rate is highly dependent on the parameters $\theta = (\beta_0, \phi, \Sigma)$ and may cause a suboptimal pre-asymptotic convergence performance for SG and QMC quadrature. Only for very high $N > 10^7$ the higher main rate $O(N^{-r})$ surpasses the secondary log-rate and the O-constant and enters an asymptotic convergence phase.

We obtain similar results for integrals from two exemplary Dynamic Economic models: For the presented Dynamic Discrete Choice model ([1], [55]), the choice of a smoothing parameter has major impact on the performance of higher-quadrature rules but also affects the approximation result. Furthermore, we consider a simple Neo-classical Stochastic Growth model [49] with many parameters which we independently vary. While similar behavior as before could be observed for some of those parameters, particularly for changes of the variance Σ, others seem to have a smaller effect on the O-constant.

Moreover, econometricians are often interested in rough but fast approximations (often about 2 or 3 accurate digits) rather than high-precision results as the models are approximations or not completely correctly specified themselves. Choosing a fixed quadrature rule for an entire model class is therefore not practical. Even for one model specification (e.g. Mixed Logit or Probit), it is appropriate to choose an optimal quadrature rule depending on the tested parameter vector. In Chapter 4, we present examples where SG or QMC quadrature are superior and where Monte Carlo simulation is sufficient, and give heuristics for determining the proper quadrature rule.

In the course of investigating integrals in econometric models, we realized that estimating model parameters inherently comprises the simulation of an integral. Part II of this thesis deals with a new approach to computing estimators by using the SG method on nested integrals with an intermediate function.

Two important classes of extremum estimators, M-estimators and Generalized Method of Moments (GMM) estimators, are defined as maximizers of objective functions Q_N. In their exact form they fulfill beneficial properties like consistency and asymptotic normality [69]. Yet, Q_N often contains an intractable integral and has to be approximated, e.g. if one of the models in part I is estimated. Continuing the work of Griebel et al. [34], we prove consistency and asymptotic normality for approximated GMM-estimators. This justifies the use of deterministic approximation algorithms like SG quadrature for the GMM objective function whereas this was previously only the case for simulated Q_N.

Subsequently, we observe that Q_N is already the Monte Carlo simulation of an expected value over the observable variables, so Q_N acts as approximator for a multidimensional integral. Together with the approximation of Q_N itself we obtain the approximation of two nested integrals

$$\mathcal{I}(\theta) = \int_{\Omega_1} F\left(z, \theta, \int_{\Omega_2} \varphi(x, z|\theta) dx\right) dz \tag{1.4}$$

for some parameter θ, functions F, φ and integration domains Ω_1, Ω_2 which are each defined by the estimation method and the model.

The expression (1.4) denotes an integral on the tensor product domain $\Omega_1 \times \Omega_2$. Continuing the works by Griebel, Harbrecht and Multerer [30], [31] we adapt the sparse tensor product space method for integration on $\Omega_1 \times \Omega_2$: Given two quadrature rules Q^1 and Q^2 with N_1 and N_2 quadrature nodes on Ω_1 and Ω_2, respectively, the classic tensor product approach evaluates Q^2 for every node x^1 of Q^1, thus requiring $N_1 \cdot N_2$ evaluations of φ in total. Similar to the SG method, sparse tensor product (STP) or *Multilevel* quadrature omits many of the quadrature nodes (x^1, x^2) strategically such that the number of nodes is only of order $O(\max\{N_1, N_2\} \log(\max\{N_1, N_2\}))$ instead of order $O(N_1 \cdot N_2)$. In particular, we prove that with STP quadrature the lower convergence rate of Q^1 and Q^2 in each separate domain can almost be preserved for the product domain and is only impaired by a log-factor. This implies that STP quadrature can achieve a similar convergence rate with considerably fewer quadrature nodes. To account for the intermediate function F, we require F to be Hölder continuous.

Finally, we test sparse tensor product quadrature for the Maximum Likelihood estimator of a Mixed Logit model [83], the GMM estimator of a Probit model [68] and for one nested integral arising directly from a Mixed Probit model. Our results clearly display that STP quadrature outperforms quadrature on the full tensor product and can therefore reduce the computational effort for the computation of estimators.

Contributions

The main contributions of this thesis are:

- Classification and extension of existing applications of Sparse Grids and Frolov quadrature in econometrics to the class of Generalized Linear Mixed models. Exemplary applications of higher-order quadrature to integrals from Dynamic Economic models.
- Identification and evaluation of parameterized sets of integrals as the correct framework for the application of higher-order quadrature to econometric integrals.

- Proof of consistency and asymptotic normality for (deterministically) approximated Generalized Method of Moments estimators.

- Adaption of the Sparse Tensor Product method to quadrature on product domains and extension of the concept to nested integrals with intermediate function.

Structure

The following overview summarizes the preceding paragraphs and illustrates the structure of this thesis:

- In Chapter 2, we present the class of Generalized Linear Mixed models (GLMM) and two examples for Dynamic Economic models (DEM) all of which incorporate intractable multidimensional integrals. We derive several specifications of GLMM and observe that all examples yield parameterized sets of integrals.

- Chapter 3 introduces modern methods for the approximation of multidimensional integrals, in particular SG quadrature and its generalizations and Frolov cubature.

- In Chapter 4, these methods are applied to the previously obtained integrals and we examine how they perform for several combinations of parameters.

- In Chapter 5, we analyze two important classes of estimators, M- and GMM-estimators, and extend the definitions to include their approximated counterparts. Furthermore, we state conditions under which the approximated estimators preserve consistency and asymptotic normality.

- Lastly, we see in Chapter 6 how estimation can be linked to integration on a tensor product space. We prove error bounds for sparse tensor product quadrature and underscore them with numerical results.

The chapters are grouped in two parts devoted to the respective general topic. References and literature reviews are given at the beginning of each chapter.

As this thesis merges topics from econometrics and numerical mathematics, different customs for the notation of mathematical expressions meet. Since this is a mathematical thesis, we rely on the more consistent mathematical notation and try to fit the econometric terms in this framework.

Acknowledgments

I would like to thank Prof. Dr. Michael Griebel for providing me with this interesting topic at the intersection of numerical mathematics and econometrics and the opportunity to work at the Institute for Numerical Simulation.

Without his guidance and thoughtful suggestions this thesis would not have been possible.
I also want to thank Dr. Bastian Bohn for agreeing to be the second advisor of my thesis and Dr. Jens Oettershagen who, together with Prof. Griebel, laid the groundwork for this thesis and provided me with some basic code elements to build my framework on.
Furthermore, I am thankful to Prof. Dr. Philipp Eisenhauer and Rafael Suchy who offered me the opportunity to utilize my achieved knowledge directly to an econometric use case and to present intermediate results at the Open Source Economics Hackathon of the Department for Economics in Bonn.
Additionally, I want to thankfully mention Christina Brinkmann and Shayan Hundrieser who kindly proof-read a draft of this thesis and made valuable comments regarding style and econometric theory.
Lastly, I wish to thank my family and Anna Jauch for the motivational support and patience throughout the past year of ups and downs.

Part I
Quadrature in Economic Models

2 Econometric Models

2.1 Objective

Econometric modeling is concerned with describing a statistical relation between economic quantities based on the qualitative analysis of an economic phenomenon. These quantities are divided into three types: Observable dependent variables (also regressands or outcome variables) and observable and unobservable explanatory variables (also regressors or control variables for the observable and residuals for the unobservable variables). The goal is to design a statistical model which reproduces the original data generating process and quantifies the dependence of the outcome on the regressors. Usually the researcher attempts to account as much variance as possible to observable variables and only little variance to the unobservable residuals or "errors".

Often parametric models are used to reduce model complexity: Instead of considering a distinct variance parameter for every observation a shared probability distribution is assumed from which all observations are (independently or dependently) drawn and which is parameterized by finitely many parameters. This approach also simplifies the *estimation* of the model where the optimal parameter or parameter vector in terms of a given error criterion is determined.

In this chapter, we introduce one general class of econometric models, Generalized Linear Mixed models (GLMM), and two examples from the broad spectrum of Dynamic Economic models (DEM). With the unobservable residuals they include a stochastic component which is integrated out by taking the expected value over the unobservables. Hence, the described models include the computation of high-dimensional, analytically intractable integrals which require numerical approximation.

GLMM are used for the evaluation of a wide variety of economic problems, e.g. from educational and transportation but also medical research. We develop them step by step starting with the simple linear regression model for continuous and unbounded data. Following the comprehensive discussion by McCullagh and Nelder [65] we extend it to *Generalized Linear models* which already encompass specifications for bounded and discrete data. Finally, GLMM further extend this notion and incorporate correlation between subgroups of data by allowing for random effects in the model. Based on [21], [67] and [83] we present several specifications of GLMM, mostly from Discrete Choice modeling, and identify why and which multidimensional integrals arise in GLMM.

In the third section of this chapter, we consider two examples for DEM and observe that they also require multidimensional integration. Based on the extensive presentation by Aguirregabiria et al. [1] and the seminal works by Keane and Wolpin ([57], [56]) and Eisenhauer [16] we shortly derive a general *Dynamic Discrete Choice model* (DDCM) and point out how integration comes to play in its evaluation. Afterwards we study a simple *Neo-classical Stochastic Growth model* which already served as prototypical model for the application of higher-order interpolation methods ([8], [49], [61]).

2.2 Linear and Generalized Linear Models

We assume that a researcher has collected data on some economic phenomenon, such that the set of observations consists of outcomes $y_n \in \Omega \subseteq \mathbb{R}$ and explanatory factors $x_n \in \mathbb{R}^{1\times q}$ for $n = 1, ..., N$.

The basic linear regression model is widely used in non-structural modeling but is often also justified by structural assumptions. It proposes a linear relation between explanatory and outcome variables with error term ε_n,

$$y_n = x_n \,\beta + \varepsilon_n \quad \text{for } n = 1, ..., N\,. \tag{2.1}$$

The unknown parameter vector $\beta \in \mathbb{R}^q$ determines the exact relationship and needs to be estimated. Estimation is based on collected data of the observable variables and further explained in Chapter 5. We treat X, Y and ε as random variables with N independently distributed realizations $x_1, ..., x_N$, $y_1, ..., y_N$ and $\varepsilon_1, ..., \varepsilon_N$ and obtain

$$Y = X\,\beta + \varepsilon$$

as relation of stochastic components. We call $\eta := X\,\beta$ the *(linear) predictor*. One can think of the error term as defective observations, unobserved heterogeneity in the examined agents (e.g. individuals, groups of individuals, companies, countries, ...) or missing factors in the proposed model.

Models with multiple outcomes are equivalently derived. In general, we assume in this thesis that J outcomes are observed, $y_n \in \Omega \subset \mathbb{R}^J$, and each observed regressor is a matrix $x_n \in \mathbb{R}^{J\times q}$, i.e. q factors are considered for each outcome. To complete the regression (2.1), also the residual variables are vector-valued, $\varepsilon_n \in \mathbb{R}^J$.

Several assumptions regarding distribution and correlation of ε are normally made to enable easier estimation and interpretability. These include $\mathrm{E}[\varepsilon] = 0$ and constant, finite variance for all observations and that realizations of ε may be uncorrelated between individuals.

However, while linear models with Gaussian errors are frequently and successfully applied in natural and social science, they are not suitable for any circumstances. For instance, count data can only incorporate non-negative errors and bounded outcome variables cannot be represented properly by an unbounded linear predictor.

Generalized Linear models (GLM) address this issue and relax the conditions on ε and the relation between X and Y. Popular examples are the

Logit and the Probit model for discrete data e.g. from Discrete Choice modeling and the Poisson model for count data e.g. from damage modeling or the evaluation of drug treatments.
The GLM consists of three components:

- A *linear predictor* $\eta := X\beta$,
- an *exponential family of distributions* f for Y with mean $\mathrm{E}(Y) = \mu$, and
- a *link function* g s.t. $g(\mu) = \eta$.

Generalized Additive Models use non-linear predictors η of the form $\eta = \beta_0 + \sum_{j=1}^{q} f_j(X_j)$ where the f_j are components of the vector-valued function f. Since for most econometric applications a linear setting is sufficient we focus on the more common GLM which introduces flexibility into the standard linear model via the link function g and the distribution f.
Via $g(\mathrm{E}(Y)) = \eta = X\,\beta$ the outcome Y is again modeled by the explanatory variables X and the parameter vector β. Instead of adding an unobserved variable ε with mean 0, this variability is directly incorporated in Y by specifying a distribution f for Y. The mean of μ of Y is used to parameterize f.
The choice of a distribution f is determined by its domain (finite, countably infinite, bounded continuous or unbounded continuous) and the assumptions of the researcher. A GLM requires that its probability density function f has the general form of an overdispersed exponential family,

$$f(y|\zeta,\phi) := \exp\left(\frac{T(y)\cdot\theta(\zeta) + \tilde{b}(\zeta)}{a(\phi)} + c(y,\varphi)\right),$$

where $T(y)$ is a sufficient statistic and $\cdot$ denotes the dot-product. If $\zeta = \theta(\zeta)$, then f is in natural form. This can simply be achieved by considering a transformation of the parameter ζ, i.e. $\theta = \theta(\zeta)$. The replacement $b(\theta) = \tilde{b}(\zeta)$ is well defined for exponential families and this transformation. Additionally, we usually assume that y is a sufficient statistic by the choice of the link function and hence obtain a distribution from a natural exponential family,

$$f(y|\theta,\phi) := \exp\left(\frac{y\cdot\theta + b(\theta)}{a(\phi)} + c(y,\varphi)\right).$$

The functions b, c fix the distribution type (Gaussian, Poisson, Multinomial, ...) while the parameters θ and ϕ and the function a serve to parameterize the exact shape. Here, θ is called the *natural parameter* and can be written w.r.t. the mean $\mu = \mathrm{E}(Y)$ of Y as

$$\tilde{g}(\mu) = \theta\,.$$

The function $\tilde{g} : M \to \mathbb{R}^J$ is defined on the set $M \subset \mathbb{R}$ where M is the domain of the mean μ. It is uniquely determined by the choice of b and c and is bijective and differentiable for exponential families. In fact, MuCullagh and Nelder [65] show

$$\mu = \nabla_\theta b(\theta)\,,$$

so $\tilde{g} = \nabla_\theta b^{-1}$.

Furthermore, ϕ is called the *dispersion parameter* and determines the variance of Y. Assuming the information inequality holds we get

$$\mathrm{Var}(Y) = a(\phi)\nabla_{\theta\theta} b(\theta)\,.$$

We write $\nabla_{\theta\theta} b(\theta)$ as the *variance function* $V(\mu)$ and consider it as function of the mean via the canonical link $\tilde{g}$.

Most models assume $a(\phi) = \frac{\phi}{w}$ for a known prior weight w and even assume a fixed value $\phi = \phi_0$ if the main concern of estimation is the natural parameter θ. The simultaneous modeling of more than one shape determining parameter (e.g. for a negative binomial distribution) is realized in the context of *vector GLMs* yet this thesis focuses on the given GLM representation as it already covers the most common models in applied econometrics. One exception are the later reviewed Linear Quantile Mixed models.

The link function $g : M \to \mathbb{R}^J$ relates the mean $\mu = \mathrm{E}(Y)$ with the predictor η,

$$g^{-1}(\eta) = \mu\,.$$

Similar to $\tilde{g}$, g is defined on the domain M of the mean μ and assumed to be bijective and differentiable. Indeed, $\tilde{g}$ is called the *canonical link function* as it arises naturally from the distribution f.

The link function transforms the unbounded range of the linear predictor to the allowed range for μ. This concerns primarily models for count data, i.e. $M = [0, \infty)$ or $M = \mathbb{N}_0$, and models for categorical data where $\mu = (\mu_1, ..., \mu_J)$ denotes the probabilities for each of the J possible outcomes.

We can then rewrite f in terms of the parameter β of the predictor

$$f(y|\,\beta, X, \phi_0) = \exp\left(\frac{y \cdot \tilde{g}(g^{-1}(X\,\beta)) + b(\tilde{g}(g^{-1}(X\,\beta)))}{a(\phi_0)} + c(y, \phi_0)\right) \quad (2.2)$$

and reduce the estimation to the original parameter β. In particular, we see that only the inverse of the link function g is required. The representation (2.2) implies the convenient choice $g = \tilde{g}$, yet other options have also proven valuable. Link functions are typically derived in the context of a particular model and can then be examined theoretically.

Usual link functions for binomial data (i.e. $Y \in \{0, 1\}$ and $\mu \in [0, 1]$) are

- the *Logit* function

$$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right) , \tag{2.3}$$

- the *Probit* function

$$g(\mu) = \Phi^{-1}(\mu) \tag{2.4}$$

  where Φ is the c.d.f. of the Gaussian distribution,

- and the *complementary log-log* function

$$g(\mu) = \operatorname{cloglog}(\mu) := \log(-\log(1-\mu)) . \tag{2.5}$$

The Gaussian c.d.f. in the Probit function can theoretically be replaced by the c.d.f. of any probability distribution on $\mathbb{R}$. The Probit function can be adapted for general categorical data with $\mu \in [0,1]^J$ by using a multivariate Gaussian c.d.f.. For the Logit function, the typical extension to the multivariate case is

$$g_i^{-1}(\eta) = \frac{\exp(\eta_i)}{\sum_{j=1}^J \exp(\eta_j)} \tag{2.6}$$

for $i = 1, .., J$. Categorical observations often arise from Discrete Choice models where an individual or agent has the choice between $J \geq 2$ alternatives. Then we call $\mathrm{E}(Y_i) = \mu_i = P(Y_i = 1,\ Y_j = 0 \text{ for } j \neq i)$ the *choice probability* for choice i.

For data with positive mean the power family of link functions is available,

$$g_\lambda(\mu) = \begin{cases} \mu^\lambda , \text{ for} & \lambda \neq 0 , \\ \log(\mu) , \text{ for} & \lambda = 0 . \end{cases}$$

In the multivariate case, these links are applied componentwise. The associated exponential families with positive mean are Poisson and Negative Binomial distribution for count data and Gamma and inverse Gaussian distribution for scale data (i.e. continuous positive data).

Finally, we state the associated canonical link functions:

- Binomial: $\tilde{g}(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$,
- Multinomial: $\tilde{g}_i(\mu) = \log(\mu)$ for $i = 1, ..., J$,
- Poisson: $\tilde{g}(\mu) = \log(\mu)$,
- Negative Binomial with number of failures r: $\tilde{g}(\mu) = \log\left(\frac{\mu}{\mu+r}\right)$,

Distribution f	canonical link $\tilde{g}(\mu) =$	other possible links $g(\mu) =$	log-partition $b(\tilde{g}(\mu)) =$
Categorial	$\log(\mu)$	Φ^{-1}, other inverse CDFs	0
Bernoulli	$\text{Logit}(\mu) := \log(\frac{\mu}{1-\mu})$	Φ^{-1}, cloglog, other inverse CDFs	0
Poisson	$\log(\mu)$	$\frac{1}{\mu},\ \mu^{\lambda}$	μ
Negative binomial (for fixed r)	$\log\left(\frac{\mu}{\mu+r}\right)$	$\log(\mu),\ \mu^{\lambda}$	$r \log\left(\frac{r}{\mu+r}\right)$
Gamma (for fixed ν)	$-1/\mu$	$\log(\mu),\ \mu^{\lambda}$	$-\log(\mu)$
Inverse Gaussian (for fixed ν)	$1/\mu^2$	$\log(\mu),\ \mu^{\lambda}$	$\frac{2}{\mu}$
Gaussian (for fixed σ)	μ (Identity link)	-	$\mu^2/2$
Laplace	μ (Identity link)	-	(not an exponential family)

Table 1: Specifications of Generalized Linear Models (bracketed variables are fixed dispersion parameters).

- Gamma with scale parameter ν: $\tilde{g}(\mu) = -\frac{1}{\mu}$,
- Inverse Gaussian with shape parameter ν: $\tilde{g}(\mu) = \frac{1}{\mu^2}$,
- Gaussian: $\tilde{g}(\mu) = \mu$.

All combinations are also collected in Table 1. Here, f is defined for one-dimensional outcomes y and can be extended to a vector of J independently distributed outcomes y_i by taking the J-fold product of f. Otherwise, each outcome can be considered separately by means of a J-dimensional vector of J independent distributions f from Table 1. Correlation between $y_1, ..., y_J$ is introduced by an additional mixing distribution in the following section. In principal, many more exponential families and link functions are conceivable but their treatment is rather theoretical. Actual applications rely, as far as our research went, solely on the mentioned distributions and functions.

2.3 Generalized Linear Mixed Models (GLMM)

The setting of the previous section with observations $z_n = (x_n, y_n)$, $n = 1, ..., N$, and the linear predictor $\eta = X\,\beta$ is well suited for cross-sectional data where the observations are assumed to be independent and identically distributed for all N observed agents. This changes if clustered data is used: Here, the individuals are sorted in groups so that correlation between the observations in one group is allowed.

A classic example for such data comes from educational research: Test results y_{nk} and explanatory variables x_{nk} like sex or education of the parents are observed for students $n = 1, ..., N$ across several colleges $k = 1, ..., K$. While still considering a linear relationship between y_{nk} and x_{nk} it is reasonable to assume an unobservable college-specific random effect u_k additional to the student-specific random effect ε_{nk},

$$y_{nk} = x_{nk}\,\beta + u_k + \varepsilon_{nk}\,. \tag{2.7}$$

In this so called *random intercept model*, the random variable U with realizations u_k, $k = 1, ..., K$, is assumed to follow some random, typically a Gaussian, distribution with mean 0. The unobservable residual $(u_k + \varepsilon_{nk})$ is not independent for each separate individual n but correlated for individuals in group k.

The model (2.7) can be extended to a *random effects* or *mixed model* by considering another set $w_{nk} \in \mathbb{R}^{J\times p}$ of observable explanatory variables and a vector-valued random effect $u_k \in \mathbb{R}^p$ with some multivariate distribution and $\mathrm{E}[U] = 0$ so that

$$y_{nk} = x_{nk}\,\beta + w_{nk}u_k + \varepsilon_{nk}\,. \tag{2.8}$$

A simpler form (for $p = q$ and $x_{nk} = w_{nk}$),

$$y_{nk} = x_{nk}(\beta + u_k) + \varepsilon_{nk}\,, \tag{2.9}$$

can be obtained when we assume that u_k is not some feature of the group k but in fact the variation of the parameter vector β in every group, i.e. $\beta_k := \beta + u_k$ is drawn independently from a probability distribution for all groups $k = 1, ..., K$. A prominent example is the cluster interpretation of panel data: Here, every "group" of observations is actually the set of observations for one person at different points in time. Then, the parameter vector β_k might be different for every individual k but constant over time while the residual ε_{nk} is independent for every observation at every time.

As for the standard linear model in the previous section, we can generalize the mixed model (2.9) by introducing a link function g and assuming an exponential family distribution f for y_{nk}. We again use the linear predictor

$$\eta = X\,\beta = X(\beta_0 + U)$$

but assume that the previously fixed parameter vector β is now composed of a fixed component β_0 and a random effect U with realizations u_k, $k = 1, ..., K$. Together with a distribution h with mean 0 and variance Σ for U, the components f, g, h and η constitute a *Generalized Linear Mixed model* (GLMM).

While the random intercept model (2.7) features the easily calculable random effects estimator (an extension of Least Squares) this estimator cannot be used for GLMM. Instead, a likelihood approach is applied for which the likelihood of the outcome $Y = y_{nk}$, given the data x_{nk} and a guess for the estimated parameter β_0, is calculated. As we do not want to model u_k separately for every k we integrate it out and obtain

$$P(y|\,\beta_0, \Sigma, \phi) = \int_{\mathbb{R}^q} f(y|\,\beta_0, u, \phi)h(u|\Sigma)du \tag{2.10}$$

where the values of Σ, ϕ are design choices or separately estimated.
For most of the specifications in Table 1 and common choices for h, this integral has no analytic solution and therefore must be computed numerically. Instead, every combination of such specifications for f, g and h leads to a different parameterized set of integrals which has to be analyzed numerically.
GLMM extent the GLM from the previous section so naturally the specifications in Table 1 for f and g hold also for GLMM. As for GLM, the established notation allows for much more variation in the choice of f, g and h than actually used. The additional distribution h is often assumed to be multivariate Gaussian for similar reasons as for the derivation of the linear model. Alternatively, the multivariate Laplace distribution can be used. For most choices of f, g and h, the integral (2.10) has no analytic solution. Instead, every combination (f, g, h) leads to a different set of integrals which are parameterized by (β_0, Σ, ϕ) and have to be computed numerically.

In the following, we derive four model specifications with different integrals (2.10) and see how they fit in the setting of Generalized Linear Mixed models. They are widely used in econometrics and related statistical fields like psychology, medicine and social science.
In the context of Discrete Choice models (DCM) two popular variants, the *Mixed Logit* and the *Multinomial Probit* model, are derived: Here, we are interested in understanding how an individual chooses between finitely many alternatives. For each alternative we want to find a *choice probability* as a function of the observed attributes for each individual. In the finite case we can compute a probability for each outcome separately, hence the outcome is described by a multinomial distribution. In terms of GLMM, the mean μ becomes a vector of means so that

$$\mu_i = P(Y = i) = \mathrm{E}[\mathbf{1}_{\{Y=i\}}]$$

for every choice i where the expectation is taken over the distribution of Y.
Discrete Choice models have been used for many years in different branches of econometrics: research applications include the analysis of market equilibria [4], transportation ([5], both for Mixed Logit) or debt crisis in developing countries ([35], for Multinomial Probit). Train [83] gives a comprehensive overview of various models, applications, estimation techniques and respective numerical methods.

Normally, a latent variable model is deployed where the choice is made according to some unobserved utility measure and the alternative with maximal utility is chosen. Let J be the number of alternatives and $\beta \in \mathbb{R}^q$ a vector of parameters for the utility measure $U(x, \beta) \in \mathbb{R}^J$. For each individual we observe a vector (or matrix) of exogenous variables $X \in \mathcal{X} \subset \mathbb{R}^{J \times r}$ and a choice $Y \in \{1, ..., J\}$. Here $\mathcal{X}$ is the domain for the random variable X and we require $r \geq q$ so that the model is identified. Furthermore, we suppose there are unobservable factors $\varepsilon \in \mathbb{R}^J$ which affect the utility and are distributed according to a known (or assumed) distribution. The common utility function $U : \mathcal{X} \times \mathbb{R}^q \to \mathbb{R}^J$ is linear in X and β, assumes $r = q$ and is additive in ε,

$$U(X, \beta) = X\,\beta + \varepsilon\,.$$

Then $Y = i$ exactly if $U_i(X, \beta) > U_j(X, \beta)$ for all $j \neq i$ where U_i are the components of the vector U, i.e. the utility of the particular choice i. In order to find the choice probabilities

$$P(Y = i|X, \beta) = \Pr\left((X\,\beta)_i - (X\,\beta)_j > \varepsilon_j - \varepsilon_i \;:\; \forall j \neq i\right) \tag{2.11}$$

we need to propose a distribution for ε. The Mixed Logit model assumes an extreme value distribution with p.d.f.

$$f(\varepsilon_j) = e^{-\varepsilon_j} e^{-e^{-\varepsilon_j}}$$

for $j = 1, ..., J$, while the Multinomial Probit model uses a multivariate Gaussian distribution $\mathcal{N}(0, \Sigma)$ with covariance matrix $\Sigma \in \mathbb{R}^{J \times J}$.
The choice probabilities for Mixed Logit are based on the choice probabilities of the more basic Logit model which assumes a fixed parameter vector β: Given the error ε_i, we have

$$\begin{aligned} P(Y = i|X, \beta, \varepsilon_i) &= \Pr\left((X\,\beta)_i - (X\,\beta)_j + \varepsilon_i > \varepsilon_j \;:\; \forall j \neq i\right) \\ &= \prod_{j \neq i} e^{-\varepsilon_j} e^{-e^{-(\varepsilon_i + (X\,\beta)_i - (X\,\beta)_j)}} \end{aligned}$$

and then integrate over the distribution of ε_i to obtain

$$\begin{aligned} P(Y = i|X, \beta) &= \int_{\mathbb{R}} P(Y = i|X, \beta, \varepsilon_i) e^{-\varepsilon_i} e^{-e^{-\varepsilon_i}} d\varepsilon_i \\ &= \frac{e^{(X\,\beta)_i}}{\sum_{j=1}^{J} e^{(X\,\beta)_j}}\,. \end{aligned}$$

Thus we have derived and justified the link function described in (2.6). If the available data is in panel form or clustered, we need to define a mixing distribution h for β in order to specify correlation within a series or a cluster of choices. As mentioned for the general GLMM, we then want to estimate mean β_0 and variance Σ of h instead of every single parameter vector β_k for every individual (”cluster”) $k = 1, ..., K$. Hence, we integrate β out

$$P(y = i|X, \beta_0, \Sigma) = \int_{\mathbb{R}^q} \frac{e^{(X\,\beta)_i}}{\sum_{j=1}^{J} e^{(X\,\beta)_j}} h(\beta\,|\,\beta_0, \Sigma) d\,\beta \;. \tag{2.12}$$

This constitutes the Mixed Logit choice probabilities for $i = 1, ..., J$. The commonly used mixing distribution is the Gaussian, leading to an intractable integral. Other mixing distributions are possible but McCulloch and Searle [66] point out that the choice of mixture seems to have only marginal effect on the model performance.

The Multinomial Probit model is even easier derived from (2.11). Since only the differences in utility affect the choice we can define $\tilde{U}_{ij} = U_i - U_j$, $\tilde{V}_{ij} = (X\,\beta)_i - (X\,\beta)_j$ and $\tilde{\varepsilon}_{ij} = \varepsilon_i - \varepsilon_j$ for $i \neq j$ and rewrite (2.11) as

$$P(Y = i|X, \beta) = \Pr\left(\tilde{U}_{ij} > 0 \;:\; \forall j \neq i\right) \;. \tag{2.13}$$

The vector $\tilde{\varepsilon}_i = (\tilde{\varepsilon}_{i1}, ..., \tilde{\varepsilon}_{i(i-1)}, \tilde{\varepsilon}_{i(i+1)}, ..., \tilde{\varepsilon}_{iJ})^T \in \mathbb{R}^{J-1}$ is again normally distributed with covariance matrix $\tilde{\Sigma}_i$ derived from Σ. Then (2.13) evaluates to

$$P(Y = i|X, \beta) = \int_{\mathbb{R}^{J-1}} \mathbf{1}_{\{\tilde{V}_{ij} + \tilde{\varepsilon}_{ij} > 0 \;\forall j \neq i\}} \phi(\tilde{\varepsilon}_i) d\varepsilon_i = \Phi(\tilde{V}_i) \,, \tag{2.14}$$

where ϕ is the p.d.f. and Φ is the c.d.f. of $\tilde{\varepsilon}_i$. This resembles the link function in (2.4). The multivariate c.d.f. Φ cannot be computed analytically for non-trivial Σ and also has to be approximated numerically. In contrast to the Mixed Logit model, this means that the integral does not stem from a mixing distribution h but directly from the assumed probability distribution in the utility function.

Within-cluster or -series correlation is usually expressed already by the freely chosen covariance matrix Σ. Hence, an additional mixing distribution is rarely used, but mentioned e.g. in [66] and [83] as possibly beneficial but even less feasible numerically. We examine an approach to reduce quadrature costs for such a double or *nested* integral in Section 6.3 and present results in Section 6.4.

For count data we cannot estimate a probability for every natural number $k \in \mathbb{N}_0$. Instead, Poisson and Negative Binomial distribution define probabilities for every k in terms of one or two parameters and thus also fix mean and variance of the outcome. Hence, estimation is concerned with finding those values. Count data arises in different contexts, e.g. in health economics or research on labor mobility (for further references see [87]) and modeling approaches for it are a recurrent topic in textbooks ([29] or, in the

context of GLM, [65]). *Mixed Poisson models* are mentioned e.g. in [62] and [66] but infrequently used due to intractable integrals of the form (2.10). Zhao et al. [89] and Zhu and Lee [90] consider data from medical research and provide first approaches to approximate the integrals by Monte Carlo simulation. An alternative to the fully mixed model is the Poisson-Gamma model which is equivalent to Negative Binomial regression.
Similar to the DCM above, we let $\beta \in \mathbb{R}^q$ be a parameter vector and $X \in \mathbb{R}^q$ be the vector of q observed attributes for each count $y \in \mathbb{N}_0$ for a single individual (or firm, country,...). We use again the linear predictor $\eta = X\,\beta$. The Poisson distribution has proven to be suitable for many instances of count data and was originally deployed by Bortkiewicz (1898) to the infamous Prussian horse-kicking data.
The Poisson distribution comes from an exponential family and is given by

$$\Pr(y = k|\mu) = \frac{\mu^k}{k!}e^{-\mu}$$

for mean $\mu > 0$. Then the link function g s.t. $g(\eta) = \mu$ must take positive real values. Usually the canonical link $g = \log$ is applied leading to

$$\Pr(y = k|X, \beta) = \frac{1}{k!}e^{kX\,\beta}e^{-e^{X\,\beta}}\,.$$

Now, a general mixing distribution h for β can account for correlations between counts, e.g. for groups of individuals or for a series of counts for one individual. We parameterize h again by β_0 and Σ and get

$$\Pr(y = k|X, \beta_0, \Sigma) = \int_{\mathbb{R}^q} \frac{1}{k!}e^{kX\,\beta}e^{-e^{X\,\beta}}h(\beta\,|\,\beta_0, \Sigma)d\beta\,.$$

Analogous to the Mixed Logit model this integral cannot be computed analytically, e.g. if h denotes the Gaussian p.d.f.

Finally, we return to continuous data on the real line and the original model

$$Y = X\,\beta + \varepsilon\,.$$

In Section 2.2, we mentioned that often i.i.d. $\varepsilon \sim \mathcal{N}(0, \sigma)$ is assumed, among other reasons due to the simple least squares estimator which arises from this assumption. However, the Gaussian distribution is a light-tailed distribution which can be problematic in two cases: The true distribution might deviate strongly from the Gaussian, so the narrow dispersion of mass around the mean might be too restrictive. The second, related issue are outliers in the data which highly affect and distort the least squares estimator.
One solution for this difficulty is the choice of a more robust distribution. The Laplace distribution with p.d.f.

$$f(z|m, b) = \frac{1}{2b}\exp\left(-\frac{|z-m|}{b}\right) \tag{2.15}$$

has heavier tails as the Gaussian and leads in the linear regression model to the *Least Absolute Deviations* estimator. Kotz et al. [58] propose a multi-variate extension with mean $\mu \in \mathbb{R}^d$ and covariance $\Sigma \in \mathbb{R}^{d\times d}$,

$$f(z|\mu,\Sigma) = \frac{2\exp(z^T\Sigma^{-1}\mu)}{(2\pi)^{d/2}\det(\Sigma)^{1/2}}\left(\frac{z^t\Sigma^{-1}z}{2+\mu^T\Sigma^{-1}\mu}\right)^{\nu/2} K_\nu\left(\sqrt{(2+\mu^T\Sigma^{-1}\mu)\,(z^t\Sigma^{-1}z)}\right)$$

for $z \in \mathbb{R}^d$ and $\nu = \frac{2-d}{2}$ and the modified Bessel function of second kind K_ν. In the context of GLM and GLMM, models with Gaussian or Laplacian f do not require a link function g as the mean μ can be any real number (or vector in $\mathbb{R}^q$). Yet, we can still introduce a mixing distribution for β. The resulting model is simply known as *Linear Mixed model* for Gaussian errors and has been referred to as *Linear Quantile Mixed model* by Geraci and Bottai [22] for Laplacian errors. It does not directly fit in the presented GLM framework as the Laplace distribution is no exponential family for variable mean μ. We still describe it in this context as it poses similar computational questions and can at least be characterized in terms of a vector-GLM, which omits the assumption that f is from an exponential family.

The survey [21] reviews four combinations of Gaussian and Laplace distribution for errors and random effects. It demonstrates that once more numerical approximation of an integral is necessary if the Laplacian is involved: Let the error ε be i.i.d. according to the p.d.f. from (2.15) with $m = 0$ for some scale $b > 0$. Given a sample β from the mixing distribution $h(\beta\,|\,\beta_0, \Sigma)$, the outcome Y is Laplace distributed with mean $\mu = \mathrm{E}[Y] = X\,\beta$. We integrate over the random effect and get the p.d.f. of Y,

$$f(y|X,\beta_0,\Sigma) = \int_{\mathbb{R}^q} \frac{1}{2b}\exp\left(-\frac{|y - X\,\beta\,|}{b}\right) h(\beta\,|\,\beta_0,\Sigma)d\,\beta\ .$$

Concluding this section, we see that the GLMM framework offers plenty of options to the researcher to specify and adapt the traditional linear model to his economic problem and data. Yet, many of these options include an intractable multidimensional integral and can only be computed at high cost.

2.4 Dynamic Economic Models (DEM)

Researchers are often concerned with describing economic processes where utility is not only defined at one particular moment in time but over the course of a longer time period. In such cases, a decision for a job or the allocation of resources in a factory not only determines the immediate return but also influences future decisions and expected utilities. The static GLMM is unable to incorporate such inter-temporal correlations for highly-involved multidimensional models. Hence, a dynamic approach is used, where the utility over all time periods is maximized by applying Bellman's principle of optimality to compute the optimal choice at each time step. Depending on the model formulation and solution method this strategy might create a large state space for each time period, where every state has to be evaluated to determine the optimal choice. In other cases the characterization of stochastic PDEs for the model parameters allows for more direct iterative techniques.

In the following, we present two dynamic economic models from quite different branches of economics: The *Dynamic Discrete Choice model* (DDCM) is the natural extension of the static Discrete Choice model encountered in the previous chapter. We now assume that the decision maker takes his future decisions into account and therefore might make choices whose benefit will only be rewarded in the future.

The second example originates in macroeconomics. In contrast to the previously developed models from microeconomics the following *Neo-classical Stochastic Growth model* (NSGM) is concerned with the simulation of national economies. Based on quantities such as capital, consumption, investments and productivity it solves the Social planner's problem of maximizing the long-term overall welfare (from today's point of view for the current and all future periods).

As before, both models require the computation of a multidimensional integral which stems from the definition as stochastic models via unobservable variables. However, the derivation of these integrals is less canonical as for GLMM and often depends on the maximization algorithm. Hence, both cases can only exemplary illustrate the advantages of higher-order quadrature for the solution of Dynamic Economic models.

Dynamic Discrete Choice models

The GLMM with a multinomial distribution is designed to model static DCM. We allowed correlation between individuals in clusters but evaluated choices only at one particular moment for those individuals. However, economists are also interested in investigating how agents make choices over a long period of time e.g. in labor economics [56] or industrial organization [80] (further research areas are referenced in [55]). This changes the setting from single independent choices to series of choices which are correlated temporally.

If the choices made by an individual are short-sighted (i.e. not taking future decisions into account), then we can form a cluster out of a series of (independent) choices and model the clustered data with a GLMM. This is not possible anymore if the individual uses his knowledge about the existence of future decisions and can make assumptions regarding the expected utility of a certain decision. Allowing for foresighted planning introduces a dynamic component into the previously static model.

We will base our definition of *Dynamic Discrete Choice Models (DDCM)* on the review paper by Aguirregabiria [1] and the review chapter by Keane

[55]. They define DDCMs in the context of utility and demonstrate why this new model class again requires the computation of a high-dimensional integral.

We assume a setting with finitely many time periods $t = 0, ..., T$ where an individual n makes one decision d_{nt} per time period. In each time period he has the choice between M alternatives $d_{nt} \in \mathcal{D} := \{1, ..., M\}$. For each decision we have observable variables x_{nt} and unobservable variables ε_{nt}. Here "observable" and "unobservable" refer to the researcher whereas the decision maker perceives all variables and incorporates them into his decision process. We summarize them into the *state vector* $s_{nt} := (x_{nt}, \varepsilon_{nt})$. The researcher can only collect data for x_{nt} hence he cannot know the entire state vector. We denote the state space of all state vectors by $\mathcal{S}$ and the state space of observable variables by $\mathcal{X}$.
In static decision models, for a given time step this utility of a choice only depends on the current data (exogenous variables) and parameters:

$$U_{nt}^{(m)} = U(m, s_{nt}) = x_{nt}^{(m)}\,\beta + \varepsilon_{nt}$$

for a utility function $U = U(d, s)$, a decision d and a state s. As mentioned in 2.2 most models rely on linear predictors/utility functions as they are easier to interpret. The residual ε_{nt} can be interpreted as the error in measurement as well as factors the researcher does not know. The decision is then modeled as

$$d_{it}^{(m)} = \begin{cases} 1, & \text{if } U_{nt}^{(m)} \geq U_{nt}^{(m')} \text{ for all } m' \neq m\,. \\ 0, & \text{otherwise.} \end{cases}$$

for all $m \in \mathcal{D}$.
In the dynamic case, we assume that the agent takes into account how his decision influences future states and utilities. E.g. an individual choosing to go to school in period t will be better educated in period $(t + 1)$ than if he did not and can factor this in with regard to his later choice of occupation. Hence, he can include this knowledge to make his decision not only based on the immediate utility but on the discounted utility for all future time periods.
Let $\delta \in (0, 1)$ be the discount factor which discounts future utilities according to how distant in the future they are. Then, the expected utility for every time period t and a series of choices $\{d_{n(t+\tau)}\}_{\tau=0,...,T-t}$ is defined as

$$\mathrm{E}\left[\sum_{\tau=0}^{T-t} \delta^{\tau} U\left(d_{n(t+\tau)}, s_{n(t+\tau)}\right) | d_{nt}, s_{nt}\right] \tag{2.16}$$

and its maximization can be rephrased as a *Dynamic Programming (DP)* problem. Using the *Bellman principle of optimality* we obtain the *Value* *function* (of the state s_{nt}),

$$\begin{aligned} V(s_{it}) &= \max_{d\in\mathcal{D}} \left\{ U(d, s_{nt}) + \mathrm{E}\left[\sum_{\tau=t+1}^{T} \delta^{\tau-1} U(d_{n\tau}, s_{n\tau}) \middle| d, s_{nt} \right] \right\} \\ &= \max_{d\in\mathcal{D}} \left\{ U(d, s_{nt}) + \delta \int V(s_{n(t+1)}) dF(s_{n(t+1)}|d, s_{nt}) \right\} . \end{aligned}$$

The probability measure $F(s_{n(t+1)}|d, s_{nt})$ represents the individual's beliefs about the future depending on the choice he makes in time step t. It includes the distribution G of the unobserved variables $\varepsilon_{n(t+1)}$ as well as the Markov transition probabilities for the states x_n and therefore is a product of continuous and discrete probability measure. The expectation in the upper form is defined over all future states $s_{n(t+1)}, ..., s_{nT}$.

Finally, the value function $V(s_{nt})$ cannot be computed since s_{nt} still includes unobservable components. An observable solution of the DP-problem is obtained by integrating over the remaining unobservable factor ε_{nt}

$$\bar{V}_{nt} = \bar{V}(x_{nt}) := \int_{\mathbb{R}^q} V(s_{nt}) dG(\varepsilon_{nt}) . \tag{2.17}$$

For finite horizons $T < \infty$, a standard approach to solving this DP-problem is backwards induction. Starting with $\bar{V}_{nT}$ we iteratively compute $\bar{V}_{n(T-1)}$, $\bar{V}_{n(T-2)}, ...$ and so on. The individual has multiple alternatives in each time period which influence the state vectors in later periods. Hence, we have to compute $\bar{V}_{nt}$ for all $x_{nt} \in \mathcal{X}$. If $\mathcal{X}$ is discrete, we can solve the problem exactly, otherwise we have to rely on discretization or interpolation.

Nevertheless, even in the discrete case, finding the exact maximum of (2.16) is usually not feasible. We need to integrate over the ε_{nt} for each state x_{nt} rendering the evaluation of the value function numerically challenging already for one state. Unfortunately the number of states is growing exponentially with T since every additional time period introduces another layer of alternatives for the agent. Therefore, it is important to make the integration step numerically as cheap as possible in order to be able to compute as many states $x \in \mathcal{X}$ as possible and approximate the exact solution of the DP-problem well. Another method to reduce computational costs is to decrease the number of states for which the utility is calculated. Keane and Wolpin present in [57] an interpolation method which interpolates between a small sample of states. Since this thesis is focused on investigating the approximation of integrals this method is left for further research.

The optimal policy function can be found by computing the value function in (2.17). For precomputed $\bar{V}_{n(t+1)}$ the integral writes

$$\bar{V}_{nt} = \int_{\mathbb{R}^q} \max_{d\in\mathcal{D}} U(d, s_{nt}) + \bar{V}_{n(t+1)} dG(\varepsilon_{nt}) .$$

So far, we have not discussed any assumptions on the exact form of U and how the unobservables ε_{nt} are distributed. Neither did we specify the transition process $x_{nt} \to x_{n(t+1)}$ and how it depends on the decision d_{nt}. Basic specifications of DDCM have been established by Rust ([79], [80]) and by Keane and Wolpin [56].
Rust makes comparatively strong assumptions on U and ε_{nt}: He assumes *additive separability* for U, meaning $U(d,s) = \tilde{U}(d,x) + \varepsilon$, and that the unobservables are i.i.d.. Then, ε is assumed to follow an extreme value distribution so the integral again results in the Logit term from Section 2.2. The term $\tilde{U}$ is typically a (piecewise) linear function in x but in any case does not affect the integration problem. Under these conditions analytic solutions for the choice probabilities of the DDCM are available, so numerical approximation is not necessary.
Keane and Wolpin relaxed some of these restrictions to allow for a more realistic reasoning, e.g. in modeling of the career choices of young men ([16], [56]). They drop the additive separability of U and assume a multivariate Gaussian distribution with mean 0 and covariance structure Σ for the ε_{nt}. In particular, Σ permits correlation between the outcome of different choices. A generalized version of the utility function used by Keane and Wolpin is

$$U(d,s) = \begin{cases} \tilde{U}(d,x) + \varepsilon & \text{for } d \in \mathcal{D}', \\ e^{\tilde{U}(d,x)+\varepsilon} & \text{for } d \in \mathcal{D} \setminus \mathcal{D}' \end{cases}$$

where $\mathcal{D}'$ is a subset of $\mathcal{D}$. The assumptions for $\tilde{U}$ remain similar to Rust's model.

Neo-classical Stochastic Growth models

Lastly, we consider a Neo-classical Stochastic Growth model which can be used to model the allocation of capital k_t and consumption c_t of nations or companies. Basic forms of such models are stated with infinite horizon, and a numerical solution is only possible under additional constraints on the evolution of capital and productivity rates. The standard solution approach then involves reformulation of the maximization problem as Euler formula, time iteration and polynomial interpolation of policy functions for k_t and c_t. The model is quite simple to estimate if only one nation or company is considered but becomes numerically challenging for the multi-country case. Several research works on the interpolation of multidimensional functions in these models examined advanced numerical methods which can circumvent the *curse of dimensionality*. Sparse grid interpolation was first used by Krueger and Kubler [61] in a International Real Business Cycle model and further investigated in [8] and [64]. Judd, Maliar and Maliar applied sparse grid interpolation to a multi-country NSGM ([49], [51]).
However, not only interpolation suffers from numerical infeasibility for high dimensions but also multidimensional intractable integrals appear in these

models. As research on numerical properties has, to date, focused on interpolation, this leaves us the opportunity to further improve the numerical performance by employing higher-order quadrature methods to these integrals. As prototypical example we present the model also used by Judd et al..

The so called social planner is concerned with maximizing the welfare

$$U = \max_{\{k_{t+1}^i, c_t^i\}_{t\in\mathbb{N}_0}, i=1,\dots,J} \mathrm{E}\left[\sum_{i=1}^{J} \lambda^i \sum_{t=0}^{\infty} \beta^t\, u^i(c_t^i)\right] \tag{2.18}$$

of nations $i = 1, ..., J$ for an infinite time span where capital k_t^i and consumption c_t^i of each nation i fulfill the budget constraint

$$\sum_{i=1}^{J} c_t^i + k_{t+1}^i = \sum_{i=1}^{J} (1-\delta) k_t^i + a_t^i f^i(k_t^i)\,. \tag{2.19}$$

The associated productivity parameters a_t^i are subject to external shocks $\varepsilon_{t+1}^i \sim \mathcal{N}(0, \Sigma)$ which are i.i.d. over time and develops with

$$\log a_{t+1}^i = \rho \log a_t^i + \varepsilon_{t+1}^i \tag{2.20}$$

for $i = 1, ..., J$. The parameter vector $\theta = (\beta, \delta, \rho, \Sigma)$ comprises the discount factor $\beta \in (0, 1]$, the capital depreciation rate $\delta \in (0, 1]$, the autocorrelation coefficient $\rho \in (-1, 1)$ and covariance matrix $\Sigma \in \mathbb{R}^{J\times J}$.

The solution of (2.18) under the condition that (2.19) and (2.20) hold is given by stochastic processes $\{k_{t+1}^i, c_t^i\}_{t\in\mathbb{N}_0, i=1,\dots,J}$ which are measurable w.r.t. $\{a_t^i\}_{t\in\mathbb{N}_0}$. Judd et al. assume $f^i = f$ and $u^i = u$ for all $i = 1, ..., J$ and assign equal weights $\lambda^i = 1$ to each country so that we have $c_t^i = c_t$ for all $t \in \mathbb{N}_0$. For strictly increasing, continuously differentiable and concave utility and production functions u and f the Euler equation

$$k_{t+1}^i = \mathrm{E}\left[\beta\, \frac{u'(c_{t+1})}{u'(c_t)} \left(1 - \delta + a_{t+1}^i f'(k_{t+1}^i)\right) k_{t+1}^i\right] \tag{2.21}$$

for $i = 1, ..., J$ provides a solution $\{k_{t+1}^i, c_t\}_{i=1,\dots,J}$ for every time period t. The expectation is taken over the unobserved shocks ε_{t+1}. The state space in this model formulation is formed by the known state variables $(k_t^i, a_t^i)_{i=1,\dots,J}$ and it is our goal to find policy functions $k_{t+1}^i = K(k_t^i, a_t^i)$ for capital and $c_t = C(k_t, a_t)$ for consumption.

The expected value on the right hand side of (2.21) can usually not be evaluated analytically, hence numerical quadrature is necessary here.

3 Multidimensional Quadrature Rules

3.1 Objective

High dimensional integrals appear not only in econometrics but also in various other fields like physics, engineering or finance. As they are often not analytically solvable there have been many attempts to find fast and precise quadrature rules which are applicable in as many situations as possible.
We consider quadrature rules of the following kind:

$$Q_N(f) := \sum_{n=1}^{N} w_n f(x_n) \approx \mathcal{I}(f) := \int_\Omega f(x)dx \tag{3.1}$$

for a function $f : \Omega \to \mathbb{R}$ on the domain $\Omega \subset \mathbb{R}^d$. We assume that Ω is closed and bounded and w.l.o.g. let $\Omega = [0,1]^d$. For $n = 1, ..., N$, the points $x_n \in \Omega$ are called *(quadrature) nodes* and $w_n \in \mathbb{R}$ *weights*.
There are two measures to determine the power of a quadrature rule: *Polynomial exactness* specifies the maximal degree of polynomials which are integrated exactly by a formula. It can be extended to exactness regarding any Chebychev system of functions instead of polynomials (see [33] for definition and applications of this generalization).
The error of a quadrature rule w.r.t. the true value,

$$E_N(f) := |\,\mathcal{I}(f) - Q_N(f)|\,, \tag{3.2}$$

typically depends on the dimension d and regularity conditions of f. Often, the error $E_N(f)$ can be bounded asymptotically for $N \to \infty$ by a function $g(N)$ which we call *convergence rate*. It is commonly given in Landau notation $E_N(f) = O(g(N))$ so that constant factors are incorporated in the O since they are only relevant for small N.
Numerical analysis is usually concerned with one of the following two issues in understanding $E_N(f)$: The first case arises if f is fixed. Then, a quadrature rule can be tailored directly to the properties of f leading to a minimal error $E_N(f)$. Yet, such a rule often does not approximate other functions very well or only with unnecessary high computational costs.
The second setting asks for an optimal N-point quadrature rule for a space $\mathcal{F}$ of considered functions. The *worst case error* of a rule Q_N

$$e(Q_N) = e(Q_N, \mathcal{F}) := \sup_{f\in\mathcal{F}, ||f||\leq 1} E_N(f)$$

depends only on $\mathcal{F}$ and denotes the operator norm of $||\,\mathcal{I} - Q_N||$ for $\mathcal{F}$. Then, the optimal error bound among all N-point rules is defined as

$$e(N) = e(N, \mathcal{F}) := \inf_{Q_N} e(Q_N, \mathcal{F})$$

for the space $\mathcal{F}$ of considered functions. For some function spaces, e.g. periodic Sobolev spaces on $[0,1]$, optimal bounds $e(N)$ have been proved although the corresponding rules are not always known.
The *worst case error* of Q_N is frequently bounded asymptotically by a term of the form $O(N^{-\gamma}\log(N)^{t(r,d)})$ where $\gamma > 0$ and d is the dimension of the domain. If $\mathcal{F}$ is defined via some kind of regularity, then r stands for the minimal regularity of functions in $\mathcal{F}$. In this case, the pure polynomial rate $O(N^{-\gamma})$ is rarely achieved and the number $t(r,d) > 0$ in the factor $\log(N)^{t(r,d)}$ indicates how large this deviation is for given r and d. Furthermore, the bound for $e(Q_N)$ differs depending on the chosen $\mathcal{F}$-norm $||\cdot||$ and can have different constants in the O-notation up to possibly exponentially increasing constants in r or d.
However, the general rate $e(Q_N)$ does not necessarily describe the approximation behavior of a particular function well for small N: Since

$$E_N(f) \leq e(Q_N)||f||\,,$$

also the norm of f and the constant in the O-estimate of $e(Q_N)$ enter in the error of Q_N. Hence, the rate $e(Q_N)$ is not sharp for small N if high constants are involved. Instead, a suboptimal convergence behavior is observed at first, with the rate $N^{-\gamma}$ only kicking in for large N. This can be prohibitive of using a rule with optimal error bound if available computing capacities are limited or temporal restrictions hold so that the optimal main rate cannot be exploited for large N.
In order to observe the convergence behavior for integrals from real world applications we consider series $Q_l = Q_{N_l}$ of quadrature rules such that N_l is dependent on the *level* l, and $N_l < N_{l+1}$ with $\lim_{l\to\infty} Q_l(f) = \mathcal{I}(f)$. Consequently, we then write E_l for the error.
For *nested* rules we obtain the helpful consequence, that the set of nodes for the l-th level is a subset of the nodes for the $(l+1)$-th level, so we can use previous function evaluations for higher levels. This strategy often requires that N_l may grow exponentially with l, e.g. $N_l = O(2^l)$ for the *Trapezoid* and the *Clenshaw-Curtis* rule. Alternatives, which are maximally nested in the sense that $N_{l+1} - N_l = 1$, are randomly drawn nodes, nodes from *Halton*- or *Sobol*-sequences or *Leja*-points. *Gaussian* quadrature rules provide non-nested nodes although there are extensions by Kronrod [60] and Patterson [74] which introduce nestedness for Gauss-Legendre points.

This chapter is structured as follows: Firstly, we shortly recall the well-known quadrature on one-dimensional domains and present several formulas and their properties.
We then move to multidimensional quadrature which is also called *cubature*. Herein, we see that using the d-fold product of a one-dimensional rule in d dimensions entails the *curse of dimension* indicating the exponential growth

of computational costs for the same convergence rate. We describe a practical alternative to the product rule, the so called *Sparse Grid method* which is also known as *Smolyak rule* [82]. Sparse Grids (SG) have originally been developed in order to define grids and according spaces of basis functions on which partial differential equations can be solved [88]. Further applications include sparse grids in machine learning and data mining (see [9] and [18] for more references).

Our presentation of the sparse grid construction and the underlying one-dimensional quadrature formulas is based on the seminal paper by Gerstner and Griebel [24] and the review article by Bungartz and Griebel [9]. Finally, our examination of Sparse Grids concludes with the depiction of extensions of the basic SG approach to functions with boundary singularities and adaptive quadrature rules ([23], [33]).

Subsequently, we move to the *Monte Carlo* (MC) method which chooses the quadrature nodes randomly from a uniform distribution on Ω. It offers the appealing property of a convergence rate which is independent of d. MC integration is covered in many text books on numerical integration in multiple dimensions, e.g. in [70].

Quasi Monte Carlo (QMC) methods seek to mimic the uniform distribution of MC integration but choose the nodes deterministically, leading to more evenly distributed nodes which avoid large "gaps". This property is also called *low discrepancy.* We describe families of (digital) nets and series and of lattice rules which have been most popular in the past. In particular, we include Frolov cubature which almost achieves the optimal bound in the considered Sobolev spaces of mixed regularity. We base our review on the presentation in [12], [13], [70] and [52].

Lastly, we consider alternatives to the uniform weight quadrature (i.e. MC and QMC rules) and introduce *optimal weights* cubature ([72]). Depending on the selected function space optimal weights cubature enables fast convergence rates even for random point sets.

3.2 One-dimensional Quadrature and the Product Rule

The approximation of integrals in one dimension, i.e. $\Omega = [0, 1]$, is one of the fundamental fields of interests in numerical mathematics. Over the past decades and centuries, a versatile toolbox containing various different quadrature formulas with different properties has evolved and is consequently also the starting point for discussions of multidimensional integration ([11], [41]). We start by presenting three important groups of one-dimensional quadrature formulas.

Newton-Cotes formulas rely on the polynomial interpolation of the integrand and use equidistant points on an open or closed interval. Depending on the number N quadrature nodes, different weights can be computed yielding formulas like the Trapezoid-, Simpson- or Midpoint rule. The associated weights are obtained by the integration of Lagrange polynomials and lead to a polynomial degree of exactness N. Basic Newton-Cotes formulas can be composed such that for the space of r-times differentiable functions $\mathcal{C}^r$, $r \in \mathbb{N}$ the optimal error bound

$$e(Q^{\text{Newton-Cotes}}, \mathcal{C}^r) = e(N, \mathcal{C}^r) = O(N^{-r})$$

is achieved.
However, these rules suffer from instability for large N due to *Runge's phenomenon* [78], meaning that large deviations occur close to the boundary. This can be mitigated by using a composite rule: Ω is iteratively subdivided into smaller intervals where on each of them a basic formula is applied. For example, the iterated Trapezoid rule leads to a nested rule with an almost duplication $N_l = 2^{l-1} + 1$ of the number of nodes in each level and achieves quadratic convergence

$$e(Q_l^{\text{IterTrapez}}, \mathcal{C}^2) = O(N_l^{-2}) \,.$$

Using basic formulas with smaller step size (i.e. more nodes) generates iterated quadrature rules of higher order.

Clenshaw-Curtis quadrature is motivated by the change of variables $x = \cos(\theta)$ and relaxes the limitation to equidistant points. It instead uses zeros or extrema of Chebychev polynomials as nodes. The latter choice produces nested points if we set $N_l = 2^{l-1} + 1$. Clenshaw-Curtis formulas yield $N_l - 1$ as degree of polynomial exactness and are adaptive to the regularity of the integrand in the sense that

$$e(Q_l^{\text{CC}}, \mathcal{C}^r) = O(N_l^{-r})$$

for functions $f \in \mathcal{C}^r(\Omega)$ and any $r \in \mathbb{N}$. There are further variations of the Clenshaw-Curtis rule like Filippi or Féjer formulas.

Finally, *Gaussian* quadrature is designed to realize the maximally possible polynomial exactness $2N - 1$ for N nodes. We consider a weighted integral with weight function $\omega : \Omega \to \mathbb{R}$

$$\mathcal{I}^\omega(f) := \int_\Omega f(x)\omega(x)dx \,.$$

Different formulas have been found for several weight functions: Relevant cases for us are

- The Gauss-Legendre formula for $\omega \equiv 1$,
- the Gauss-Laguerre formula for $\omega(x) = e^{-x}$ on $\Omega = [0, \infty)$, and

- the Gauss-Hermite formula for $\omega(x) = e^{-x^2}$ on $\Omega = \mathbb{R}$.

The weights are again computed by integrating Lagrange basis polynomials, but the nodes are chosen as the zeros of a polynomial p_N of degree N. Here, p_N is the unique polynomial (up to scalars) which is orthogonal to the space $\mathcal{P}_{N-1}$ of polynomials of degree $\leq N-1$ w.r.t. the scalar product

$$\langle f, g\rangle_\omega := \int_\Omega f(x)g(x)\omega(x)dx\,.$$

The same error bounds as for Clenshaw-Curtis formulas hold, i.e.

$$|\,\mathcal{I}^\omega(f) - Q^\omega(f)| = O(N^{-r})$$

for $f \in \mathcal{C}^r$ and any $r \in \mathbb{N}$.
However, most Gauss quadrature formulas (and in particular the ones mentioned above) are not nested. Kronrod [60] developed an extension of the N-point Gauss-Legendre formula by adding $N+1$ points which are the zeros of the $N+1$-th Stieltjes polynomial w.r.t. the N-th Legendre polynomial. Patterson [74] iterated this construction to get a nested *Gauss-Patterson* formula for $N_l = 2^l - 1$ based on the 3-point Gauss-Legendre formula. It has polynomial degree of exactness $3^{l-1} - 1$ for $l \geq 2$ and preserves the error bounds $E_l(f) = O(N_l^{-r})$ for $f \in \mathcal{C}^r$.
However, this method cannot be used for all N-point Gauss-Legendre formulas, e.g. the 2-point rule can only be extended 4 times. In particular, Patterson's as well as Kronrod's techniques cannot necessarily be applied for other weight functions, e.g. Gauss-Hermite and Gauss-Laguerre quadrature cannot be expanded in this way [54].

Moving from a one-dimensional to a d-dimensional domain $\Omega = [0,1]^d$ the first, straight-forward idea is to use the available formulas in each dimension separately. For d formulas Q_{l_i} on $[0,1]$, $l_1, ..., l_d \in \mathbb{N}$,

$$Q^d_{(l_1,\dots,l_d)} = Q_{l_1} \otimes \cdots \otimes Q_{l_d}(f) := \sum_{n_1=1}^{N_{l_1}} \cdots \sum_{n_d=1}^{N_{l_d}} w_{l_1 n_1} \cdots w_{l_d n_d} f(x_{l_1 n_1}, ..., x_{l_d n_d}) \tag{3.3}$$

yields the so called *product rule*. Letting $l = l_i$ for $i = 1, ..., d$, the error bound $O(N_l^{-r})$ for level l and N_l nodes in one direction persists for $f \in \mathcal{C}^r$. But the total number of nodes for $Q^d_{(l,\dots,l)}$ is $N = N_l^d = \prod_{i=1}^d N_{l_i}$, so this bound actually deteriorates to

$$|\,\mathcal{I}(f) - Q^d_{(l,\dots,l)}(f)| = O(N^{-r}) = O(N_l^{-\frac{r}{d}})\,.$$

This means that the product rule has an error bound which decreases exponentially fast with the dimension d for a rule with N_l nodes. This phenomenon is called *curse of dimensionality* and has been described first by Bellman [3].

3.3 Sparse Grid Quadrature

The *Smolyak* or *Sparse Grid* (SG) method presents an option to circumvent this problem. So far, we discussed convergence rates of quadrature errors for spaces such as $\mathcal{C}^r$. These results can be generalized to *Sobolev spaces*

$$H^r(\Omega) := \{f : \Omega \to \mathbb{R} \ : \ D^\alpha f \in L_2(\Omega) \ \forall |\alpha|_1 \leq r\} \ ,$$

where the partial derivatives D^α are defined in the weak sense. For the multiindex $\alpha \in \mathbb{N}_0^d$, we introduce the norms $|\alpha|_\infty := \max_{i=1,\dots,d} |\alpha_i|$ and $|\alpha|_1 := |\alpha_1| + \cdots + |\alpha_d|$. This leads to the notion of a *Sobolev space of mixed regularity*

$$H^r_{\text{mix}}(\Omega) := \{f : \Omega \to \mathbb{R} \ : \ D^\alpha f \in L_2(\Omega) \ \forall |\alpha|_\infty \leq r\} \ ,$$

which demands that more mixed derivatives are bounded than in the usual case. This is the right setting for the investigation of SG quadrature.
We again consider one of the one-dimensional formulas Q_l from Section 3.2 and compute the difference formulas

$$\Delta_l(f) := Q_l(f) - Q_{l-1}(f) \tag{3.4}$$

for $l \geq 1$ and $Q_0(f) := 0$. Here, $\Gamma_l = \{x_{li} \ , \ i = 1, ..., N_l\}$ defines the set of nodes for Q_l and $\Gamma_l \cup \Gamma_{l-1}$ is the node set of Δ_l. With this telescopic decomposition of Q_l we can reproduce the product rule as

$$Q_{l_1} \otimes \cdots \otimes Q_{l_d}(f) = \sum_{|k|_\infty \leq l} \Delta_{k_1} \otimes \cdots \otimes \Delta_{k_d}(f)$$

for indices $k \in \mathbb{N}^d$.
The origin of the SG method lies in the observation that for large $|k|_1$ the terms $\Delta_{k_1} \otimes \cdots \otimes \Delta_{k_d}$ only contribute little to the overall sum. Their contribution becomes even more negligible when compared to the cost for their computation, i.e. the number of function evaluations. This motivates the truncation of the sum to

$$Q_l^d(f) = \sum_{|k|_1 \leq l+d-1} \Delta_{k_1} \otimes \cdots \otimes \Delta_{k_d}(f) \, . \tag{3.5}$$

In some cases the function evaluation $f(x)$ is exceptionally costly, e.g. if it includes solving an ODE or approximating an integral itself. The formula stated in (3.5) might include multiple computations of f at some nodes. We can rewrite it such that f is only computed once at every node and the weights are precomputed. We define difference grids $\Xi_l := \Gamma_l \setminus \Gamma_{l-1}$ where $\Gamma_0 := \emptyset$ and write the node grid as

$$\Gamma_l^d := \bigcup_{|k|_1 \leq l+d-1} \Xi_{k_1} \times \cdots \times \Xi_{k_d} \, .$$

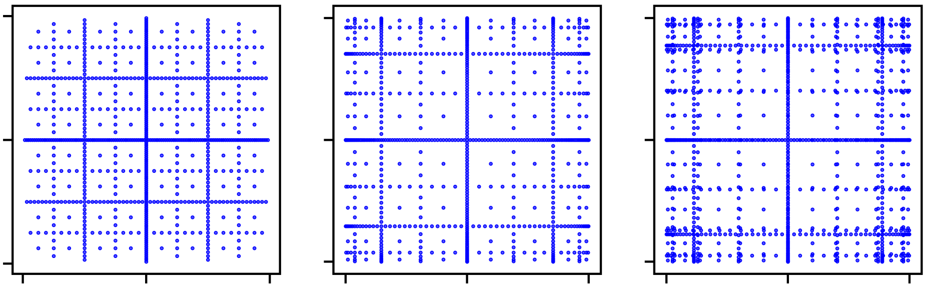

Figure 1: Sparse grid points for Trapezoid, Clenshaw-Curtis and Gauss-Legendre formulas.

Examples of sparse grid points can be seen in figure 1 for underlying Trapezoid, Clenshaw-Curtis and Gauss-Legendre formulas. The weights for each unique node in Γ_l^d can be computed in advance based on the underlying one-dimensional formula. Novak and Ritter [71] calculated the size of the sparse grid

$$N_l^{(d)} := \#\Gamma_l^d = \sum_{|k|_1 \leq l+d-1} m_{k_1} \cdot \ldots \cdot m_{k_d} = O(N_l \log(N_l)^{d-1}) \,,$$

for $m_{k_i} = \#\Xi_{k_i}$. Although this estimate holds for nested and non-nested rules alike, the constant in $O(N_l \log(N_l)^{d-1})$ is smaller for nested rules since nestedness is passed on from one to multiple dimensions, i.e.

$$\Gamma_l^d \subset \Gamma_{l+1}^d \,.$$

In both cases the size of the SG rule is substantially smaller than the number of nodes N_l^d for the product rule. It is reduced to an almost linear term with an additional log-factor instead of an exponential dependence on d.
If the product rule is already implemented, we can conveniently expand the form (3.5) to

$$Q_l^d(f) = \sum_{l \leq |k|_1 \leq l+d-1} (-1)^{l+d-|k|_1-1} \binom{d-1}{|k|_1 - l} Q_{k_1} \otimes \cdots \otimes Q_{k_d}(f)$$

to utilize the existing implementation. SG quadrature in this form is also called *combination technique.*
Although we reduced the number of nodes considerably compared to the product rule the error bound for sparse grid quadrature remains almost unchanged. For the SG formula based on Clenshaw-Curtis or Gauss formulas we have the bound [86]

$$e(Q_l^{\mathrm{SG}}, H_{\mathrm{mix}}^r) = O(N_l^{-r} \log(N_l)^{\frac{(d-1)(r+1)}{2}}) \,. \tag{3.6}$$

The algebraic convergence of order r is preserved and only impaired by a log-factor. However, the pre-asymptotic behavior might not reflect this rate as the constant can be quite high and exponentially dependent on r and d. Polynomial exactness is also transferred from one to multiple dimensions. Here, the corresponding d-dimensional space of polynomial is constructed in the same fashion as the SG rule by

$$\mathcal{P}_l^d := \{\mathcal{P}_{k_1} \otimes \cdots \otimes \mathcal{P}_{k_d} \mid |k|_1 = l + d - 1\}$$

from the spaces $\mathcal{P}_{k_i}$ of one-dimensional polynomials of degree $\leq k_i$.

Extensions of the sparse grid method

There have been several approaches to further improve sparse grids quadrature for specific regularity conditions and integrands. These include anisotropic and dimension-adaptive sparse grids [23] and SG quadrature for integrands with boundary singularities [33].

One way to generalize sparse grids is to change the index set $\mathcal{A} \subset \mathbb{N}^d$ in the summation. Originally we had

$$\mathcal{A}_1 = \left\{(k_1, ..., k_d) \in \mathbb{N}^d \ : \ |k|_1 \leq l + d - 1\right\} .$$

Changing the norm $|\cdot|_1$ to $|\cdot|_v$ for any $v \in \mathbb{R}_+^d$ and defining

$$|x|_v := \sum_{k=1}^{d} v_k x_k$$

leads to a sparse grid that emphasizes certain dimensions over others. This way pre-existing knowledge about the regularity of the integrand in multiple dimensions can be incorporated and more specific refinements of the grid are possible where necessary.

If such knowledge is not available, we can build up the index set iteratively. In [23], a corresponding algorithm is explained which is not only based on the estimated error of the terms $\Delta_{k_1} \otimes \cdots \otimes \Delta_{k_d}$ but also on the cost of each additional step.

Finally, Oettershagen and Griebel [33] investigate a technique to account for boundary singularities of the integrand. Such singularities might arise when transforming an integral on $[0, \infty)^d$ or $\mathbb{R}^d$ to the bounded domain Ω. They apply *Generalized Gaussian* formulas which are based on a Chebychev system of basis functions instead of polynomials.

Two useful Chebychev systems correspond directly to weight functions for Gaussian quadrature: Instead of polynomials in x, polynomials in the transformed argument $\psi(x)$ are used to build the space $\mathcal{P}_{N-1}$ for which an orthonormal polynomial p_N is sought and whose zeros serve as quadrature nodes (cf. Section 3.2). The transform $\psi = \log$ is motivated by the Gauss-Laguerre formula and adds weakly differentiable functions with one boundary singularity in Ω to the space of fast approximated integrands. Similarly,

$\psi = \operatorname{erf}^{-1}$ where erf is the Error function (motivated by Gauss-Hermite) accounts for weakly differentiable functions with boundary singularities on both sides of the interval $[0, 1]$.

The error bounds of Gaussian quadrature are preserved but now extended to functions with algebraic boundary singularities. Additionally, they appear to be transferred to the multidimensional case for (adaptive) sparse grids [33].

This concludes our short review of SG quadrature. Due to our focus on understanding econometric integrals as parameterized sets of integrals, we only apply isotropic SG rules based on Clenshaw-Curtis or (generalized) Gauss formulas in this thesis and leave adaptive approaches for further research. Furthermore, the idea of reducing the number of nodes in a d-fold tensor product gives rise to sparse tensor product quadrature which we investigate in Chapter 6.

3.4 Monte Carlo Quadrature

In contrast to product and SG rules which are inherently based on one-dimensional rules we can also define nodes directly in $\Omega = [0, 1]^d$. We distinguish probabilistic *Monte Carlo* and deterministic *Quasi Monte Carlo* methods. They both solely rely on the proper choice of nodes and use the uniform weight $\frac{1}{N}$ for all nodes. Hence, all nodes contribute equally to the total result.

Monte Carlo (MC) quadrature can be derived from a probabilistic point of view. We understand $\mathcal{I}(f)$ as the expected value over the random variable $f(X)$ where X is uniformly distributed on Ω. The nodes $x_1, ..., x_N$ are drawn independently and randomly from a uniform distribution and yield a quadrature formula which is exact for all $f \in L_2(\Omega)$ in expectation,

$$\mathrm{E}\left[\frac{1}{N}\sum_{n=1}^{N} f(x_n)\right] = \frac{1}{N}\sum_{n=1}^{N} \mathrm{E}[f(x_n)] = \mathrm{E}[f(x)] = \mathcal{I}(f)\,.$$

Consequently, we do not consider the error $E_N(f)$ but instead the root mean square error

$$\mathrm{MSE}[f] := \mathrm{E}[|\,\mathcal{I}(f) - Q_N(f)|^2] = \frac{\mathrm{Var}(f)}{N}$$

which implies a “probabilistic” convergence rate of $O(N^{-\frac{1}{2}})$, independent of the dimension d. This means that even for integrands f with very low regularity in a high-dimensional domain at least a slow convergence can be assured. In particular, MC integration attains a faster convergence rate than the product rule if $d > \frac{r}{2}$ and $f \in H^r_{\mathrm{mix}}$ and is often faster than SG quadrature for small N.

A further advantage is the sample variance of $Q_N(f)$ which provides a directly available error estimate for any N. Additionally, simple implementations of Q_N only require a sufficiently good pseudo random number generator. Thus, for both mathematical and implementational reasons, MC quadrature is very popular among econometricians and often used without further analysis of other alternatives.

Yet, the probabilistic nature of the MSE does not allow a direct comparison with the deterministic error bounds from previous sections. The mean squared error does not comprise a norm, thus, common proof techniques, e.g. for error bounds of tensor product quadrature rules (see Chapter 6), have to be adapted for this different error measure.

Furthermore, the actual performance of MC integration also depends on the possibly large factor $\mathrm{Var}(f)$. *Variance reduction* techniques try to mitigate this dependence in the constant of the rate. E.g. we can lower $\mathrm{Var}(f)$ by drawing the samples x_N from a distribution Φ with p.d.f. φ and use the identity

$$\int_\Omega f(x)dx = \int_\Omega \frac{f(x)}{\varphi(x)} d\Phi(x)\,.$$

If f and φ are shaped similarly (i.e. $\mathrm{dist}(f, \varphi)$ is minimized by φ for some distance measure dist), then the variance of the new integrand $\frac{f}{\varphi}$ can be considerably smaller than $\mathrm{Var}(f)$. This method is called *importance sampling.* Other techniques include *stratified sampling* and *correlated sampling.*

Importance sampling promoted the interest in sampling from distributions other than the uniform one. Since most econometric models assume Gaussian errors the question arose how to draw samples from multivariate Gaussian distributions with arbitrary covariance structure. In the 1980s and 1990s importance sampling techniques like the GHK-Simulator [25], [39] and Markov Chain Monte Carlo methods like Gibbs Sampling [19] were developed to tackle this problem. Hajivassiliou compares several methods simulating Gaussian distributions in [38]. Genz' approach [20] relies on transformations of the covariance matrix and is also applicable to t-distributions.

3.5 Quasi Monte Carlo Quadrature

Even though the convergence rate of MC quadrature is independent of the dimension, it is also distinctly slower than one-dimensional and SG quadrature. Quasi Monte Carlo (QMC) quadrature improves the random approach by choosing its points deterministically. They are designed to cover Ω evenly which is often achieved by minimizing the *discrepancy* of the node set. With *lattice rules* and *(digital) nets and sequences* we describe two main families to define such node sets and also give examples for each family.

Random points might leave large portions of the integration domain empty and without any quadrature node. We see in figure 2 how QMC rules can

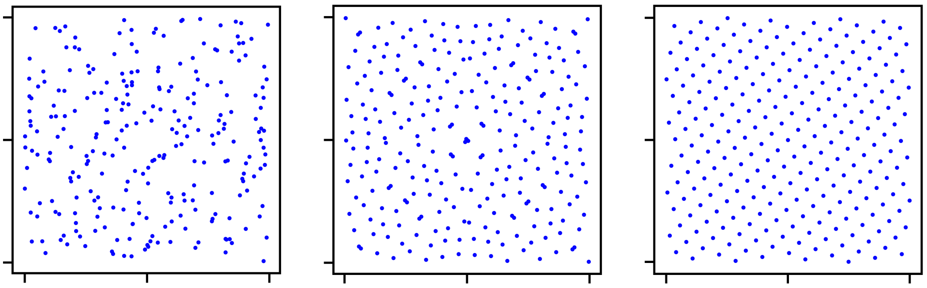

Figure 2: (Quasi) Monte Carlo points in two dimensions: Randomly drawn, from the Sobolev sequence and from the Chebychev-Frolov lattice

provide a more uniform distribution. Beside the classification into lattice rules and nets and sequences (and some others) we can differentiate between *closed* and *open* rules. The former recompute the nodes $x_1, ..., x_N$ for every N from scratch while the latter reuse already computed nodes and hence values $f(x_n)$.

As mentioned in Section 3.1 optimal error bounds $e(N, \mathcal{F})$ have been computed for several function spaces, e.g. Besov and Triebel-Lizorkin spaces (which are generalizations of Sobolev spaces) [85]. Their bound also hold for Sobolev spaces which are our primary subject of interest and reads

$$e(N, H^r_{\text{mix}}) = O(N^{-r} \log(N)^{\frac{d-1}{2}}) \, .$$

Thus, the optimal rate from one-dimensional quadrature can be transferred to the multidimensional case at the cost of a log-factor with exponent independent of regularity r. In fact, this is even possible for Sobolev spaces with differing regularities in each dimension $\mathbf{r} = (r_1, ..., r_d)$ if we set $r = \min\{r_1, ..., r_d\}$ in the above formula and replace d by the multiplicity of r in $\{r_1, ..., r_d\}$.

Unfortunately, the error for the SG quadrature (3.6) does not fulfill this rate as its log-exponent does depend on r and therefore prolongs the suboptimal pre-asymptotic phase caused by the higher secondary rate. Kacwin et al. [53] prove that Frolov cubature attains the optimal bound for Sobolev spaces, although the pre-asymptotic behavior seems to be problematic.

A comprehensive overview of digital nets and their construction is given in [13]. Dick, Kuo and Sloan explain both families in their review article [12] and give an introduction into error analysis and proper function spaces for QMC quadrature. We will follow their presentation in this and the next section. Furthermore, we consider lattices based on *Frolov cubature* which have been developed by Kacwin et al. ([52], [53]) after their original proposition by Frolov [17].

(Digital) nets and sequences

With nodes from *nets* or *sequences* we attempt to equip certain cubic subsections of $\Omega = [0,1]^d$ with the same number of points. The resulting node set Γ_N is well-distributed in Ω. The following definition from [12] is also known as (t,m,s)-net with the replacement s for d.

Definition 3.1. *((t,m,d)-net)*
Let $t \geq 0,\ m \geq 1,\ d \geq 1$ and $b \geq 2$ be integers with $t \leq m$. A (t,m,d)-net in base b is a point set Γ consisting of b^m points in $[0,1)^d$ such that all elementary boxes of the form

$$\prod_{j=1}^{d} \left[\frac{a_j}{b^{k_j}}, \frac{a_j+1}{b^{k_j}} \right) \tag{3.7}$$

for $k_j \in \mathbb{N}_0,\ 0 \leq a_j \leq b^{k_j}-1$ where $k_1 + \dots + k_d = m-t$ contain exactly b^t points.

This gives a node set of size $N = b^{m-t}b^t = b^m$ since there are $\prod_{j=1}^d b^{k_j} = b^{m-t}$ elementary boxes for every choice $(k_1, ..., k_d) \in \mathbb{N}^d$. The (t,m,d)-net becomes finer for smaller t or larger $(m-t)$ since this implies that more and smaller boxes have to be covered. The net is a closed rule since we cannot simply increase $N = b^m$ and retain the same nodes. A similar definition for an open rule is the following.

Definition 3.2. *((t,d)-sequence)*
Let $t \geq 0$ and $d \geq 1$ be integers. A (t,d)-sequence in base b is a sequence of points $\Gamma = (x_0, x_1, ...)$ in $[0,1)^d$ such that for any integers $m > t$ and $l \geq 0$, every block of b^m points

$$x_{lb^m}, ..., x_{(l+1)b^m-1}$$

in the sequence Γ forms a (t,m,d)-net in base b.

The *digital* construction scheme gives a method to build such nets and sequences. It is based on the solution of linear equations in finite fields $\mathbb{Z}_p$. Here as well as for lattice rules, number theoretic techniques come to play when developing and analyzing new quadrature formulas.
The most common example for a digital (t,d)-sequence is the *Sobol* sequence which was first constructed by Sobol in 1967. Niederreiter (1987) introduced the general (t,d)-sequences and gave a generalized version (the *Niederreiter* sequence) of the Sobol sequence. We abstain from discussing the construction here since we use the Sobol sequence only as a benchmark for SG and Frolov quadrature. This allows for a direct comparison with results presented in finance and economics which to date mainly rely on MC integration and sometimes QMC quadrature.

The same holds for the popular *Halton* sequences which can neither be put into the concept of (t,d)-sequences nor the following lattice rules.
We already mentioned that the proposed QMC rules shall cover Ω evenly. How well a set of points is distributed in Ω is normally measured in terms of the *discrepancy.* There are multiple definitions of discrepancy for different purposes but we only consider the *star discrepancy*

$$D_\Gamma^* := \sup_{z\in[0,1)^d} \left| \frac{1}{N} \sum_{x\in\Gamma} \mathbf{1}_{[0,z]}(x) - \prod_{j=1}^{d} z_j \right|$$

for a set Γ consisting of N points. Here $[0,z]$ is the box spanned by the intervals $[0,z_i]$ for $i=1,...,d$ and $\mathbf{1}_{[0,z]}(x)=1$ only if $x_i \in [0,z_i]$ for $i=1,...,d$.
A classical estimate for the error term is the *Koksma-Hlawka inequality*

$$E_N(f) \leq D_\Gamma^* V(f)$$

for a function f of bounded variation $V(f)$ in the sense of Hardy and Krause. The sequences presented above are constructed s.t. D_Γ^* is as small as possible, hence the name *low-discrepancy sequences.* Bounds of the form

$$D_\Gamma^* = O\left(\frac{\log(N)^{d-1}}{N}\right)$$

have been shown for many of these constructions. This is a significant improvement compared to MC integration and comes only at the cost of the demand for bounded variation of the integrand. On the other hand, if more knowledge concerning the regularity of the integrand is available, the bound for $e(N, H_{\text{mix}}^r)$ implies that even better rates are possible for N-point rules.

Lattice rules
Lattice rules originated from the desire to utilize the maximal regularity of the integrand and achieve higher-order convergence rates. A *lattice* Λ is a discrete subset of $\mathbb{R}^d$ which is closed under addition and subtraction. If $\mathbb{Z}^d \subset \Lambda$ and $\Gamma_N := \Lambda \cap \Omega$ denotes the node set of a quadrature rule with uniform weight $\frac{1}{N}$, then this rule is called a *lattice rule.* It can be seen as a generalization of the trapezoid rule since the projection of Γ_N on each axis produces a set of equi-distant points.
One of the first lattice rules were the *good lattice points* (Korobov, 1959). For a generating vector $z=(z_1,...,z_d) \in \mathbb{Z}^d$ we define the quadrature points as

$$x_n = \left\{\frac{nz}{N}\right\} \text{ for } n = 1,...,N$$

where $\{a\}$ is the fractional part of $a \in \mathbb{R}$ and z has no common factor with N. The additional condition $\gcd(z_i, N) = 1$ for all $i=1,...,d$ assures that

all one-dimensional projections of the lattice rule each contain N distinct values.
A simple example is the *Fibonacci* rule for $d = 2$ based on the sequence $(F_k)_{k\in\mathbb{N}_0}$ of the same name which sets $z = (1, F_k)$ and $N_k = F_{k+1}$. Although there is no apparent generalization to higher dimensions, the Fibonacci lattice seems to fulfill certain optimality conditions in $\mathbb{R}^2$.
A different approach is followed by Frolov cubature rules,

$$Q_N(f) = \frac{1}{N} \sum_{x\in\mathbb{Z}^d\cap\Omega} f(A_N x)\,,$$

where $A_N = N^{-\frac{1}{d}}A$ is defined for $A \in \mathbb{R}^{d\times d}$ with $\det(A) = 1$. Since we assumed $\Omega = [0,1]^d$ (or, more generally, that Ω is bounded and closed) at the beginning of this chapter, $N = \#\mathbb{Z} \cap \Omega$ is finite. If the lattice $A(\mathbb{Z}^d) = \{Az \mid z \in \mathbb{Z}^d\}$ is *admissible* in the sense that

$$\inf_{x\in\mathbb{Z}^d\setminus\{0\}} \left|\prod_{i=1}^{d}(Ax)_i\right| > 0\,,$$

then Q_N attains *universal optimality* for many useful function spaces (e.g. the aforementioned mixed Sobolev and Besov spaces with zero boundary conditions). Here, "universal" indicates that the optimal convergence rate is obtained for any regularity $r \in \mathbb{N}$, i.e. the same rule can be used for arbitrary smooth integrands and will always provide the maximal convergence.
This distinguishes Frolov quadrature from other formulas which all have either some maximal regularity which they can utilize or have a worse than optimal log-exponent. However, the constant in the O-notation depends on the choice of A and can be hindering for the actual applicability of Frolov cubature. In particular, the search for matrices A which yield a constant-minimizing or (in the case of high dimensions) any admissible lattice has not yet terminated.
One way for finding such a matrix was already given by Frolov: Consider a polynomial $p(x)$ that fulfills the following conditions

- p has integer coefficients,
- p has leading coefficient 1,
- p is irreducible over $\mathbb{Q}$ and
- p has distinct roots $\xi_1, ..., \xi_d$.

Then the corresponding Vandermonde matrix

$$\begin{pmatrix} 1 & \xi_1 & \xi_1^2 & \cdots & \xi_1^{d-1} \\ 1 & \xi_2 & \xi_2^2 & \cdots & \xi_2^{d-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \xi_d & \xi_d^2 & \cdots & \xi_d^{d-1} \end{pmatrix} \in \mathbb{R}^{d\times d}$$

constitutes an admissible lattice.
For $d = 2^m$, Chebychev polynomials meet these assumptions [52]. While there are other techniques for most of the low dimensions $d = 2, 4, 8, ...$ for some cases like $d = 7, 13$ satisfactory polynomials were only found by brute force [53]. The corresponding nodes for the Frolov formula in up to 10 dimensions can be found on the website of the Institute for Numerical Simulation Bonn (`https://ins.uni-bonn.de/content/software-frolov`).

3.6 Optimal Weights Cubature

While the aforementioned MC and QMC methods all use the uniform weight $\frac{1}{N}$ a recent work by Oettershagen [72] evaluates the utilization of so called *optimal weights*. Based on the notion of *Reproducing Kernel Hilbert spaces* and a given set of points in Ω we can obtain optimal weights for these points by solving a system of linear equations.
We start the derivation of this technique with a definition from [12].

Definition 3.3. *(Reproducing Kernel Hilbert space)*
A Hilbert space $\mathcal{H}_K$ of functions $f : \Omega \to \mathbb{R}$ on a set Ω with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_K}$ is a reproducing Kernel Hilbert space (RKHS) with kernel $K : \Omega \times \Omega \to \mathbb{R}$ if

(i) $K(\cdot, x) \in \mathcal{H}_K$ for all $x \in \Omega$ and

(ii) $f(x) = \langle f, K(\cdot, x) \rangle_{\mathcal{H}_K}$ for all $x \in \Omega$ and all $f \in \mathcal{H}_K$.

The kernel K defined by the above conditions is unique for $\mathcal{H}_K$ and also satisfies

(iii) $K(x, y) = K(y, x)$ for all $x, y \in \Omega$, and

(iv) that the kernel matrix $G(\Gamma_N)$ for any set $\Gamma_N = \{x_1, ..., x_N\}$ of pairwise distinct points $x_i \in \Omega$ defined by

$$G_{ij} := K(x_i, x_j)$$

is positive definite.

In particular, the *Riesz-representer* of the point evaluation functional $\delta_x : f \mapsto f(x)$ is $K(\cdot, x)$ for any $x \in \Omega$. By the Riesz-representation theorem, a representer can be found for any bounded linear functional in the dual space $\mathcal{H}_K^*$ of $\mathcal{H}_K$, thus also for the error functional $E_N : f \mapsto \mathcal{I}(f) - Q_N(f)$ of a quadrature rule Q_N. Using this representer, a closed form expression for the worst case error $e(Q_N, \mathcal{H}_K)$ can be derived [72]

$$e(Q_N, \mathcal{H}_K)^2 = ||E_N||^2_{\mathcal{H}_K^*} = \int_\Omega \int_\Omega K(x, y) dx dy - 2 \sum_{n=1}^{N} w_n \int_\Omega K(x_n, x) dx + \sum_{n=1}^{N} \sum_{m=1}^{N} w_n w_m K(x_n, x_m)\,. \tag{3.8}$$

The transformed expression greatly simplifies the identification of sharp bounds. Hence, many research works focused on calculating such bounds for RKHS like periodic and non-periodic Sobolev spaces and the space of absolutely continuous functions.

For fixed nodes $x_1, ..., x_N$, the worst case error is minimized w.r.t. the weights $w_1, ..., w_N$ by computing the partial derivatives and setting them to 0. This leads to the system

$$b(\Gamma_N)_m := \int_\Omega K(x_m, x)dx + \sum_{n=1}^{N} w_n K(x_m, x_n) \text{ for } m = 1, ..., N\,.$$

Thus, the optimal weights vector $w^*(\Gamma_N) = (w_1^*, ..., w_N^*)$ is given by

$$w^*(\Gamma_N) = G(\Gamma_N)^{-1} b(\Gamma_N)\,. \tag{3.9}$$

It is noteworthy that these weights can be computed for any set of pairwise distinct points including the random points from MC integration. We only need to identify the correct RKHS and corresponding kernel $K(\cdot, \cdot)$ in order to solve (3.9). Then optimal weights improve the MC-rate from $O(N^{-\frac{1}{2}})$ to

$$E_N(f) = O(N^{-r+\frac{1}{2}} \log(N)^{rd-\frac{1}{2}})$$

for $f \in \tilde{H}^r_{\text{mix}}([0,1]^d)$, the periodic Sobolev space.

For the non-periodic case and for the optimally weighted Halton sequence similar results could be obtained. For other RKHS it remains unclear whether this rate can actually be achieved. Furthermore, the solution of the system (3.9) requires $O(N^3)$ operations which is expensive for large N.

Concluding, the optimal weights approach gives us the chance to exploit regularity even if only random points are available. In particular, we can almost reach the optimal rate with this approach and are only hindered by the dependence of the log-factor on r.

This is of major importance for the computation of likelihoods as defined in Section 5.2. There, the quadrature points are measurements or observations which are inherently random, so we only treat them as basis for Monte Carlo integration. Most notably, if we can determine the domain of the data samples and their (approximate) distribution, optimal weights might allow us to boost convergence rates drastically.

4 Numerical Results

4.1 Objective

In Chapter 2 we encountered and described various econometric models which all included the computation of a multidimensional integral. These integrals often cannot be evaluated analytically, thus numerical approximation is necessary. This chapter is concerned with the application of the quadrature methods we developed in the previous sections to these econometric integrals.

Our simulations show that SG and Frolov cubature provide viable alternatives to currently dominating methods and perform considerably better for Probit and Mixed Logit models. Similar results could be obtained for integrals arising from the Neo-classical Stochastic Growth model in Section 2.4. However, they are less useful for Linear Quantile Mixed and Dynamic Discrete Choice models where they only match the performance of QMC rules.

We can retrace these differences directly to kinks or even singularities of the integrands. Kinks are a frequent issue in econometric integrals as they are often defined via an indicator function $\mathbf{1}_S$ for some subset $S \subset \Omega$. A smoothing operator can mitigate this problem but leads to biased results.

In contrast to common numerical analysis which focuses on the regularity of the integrands we examine parameterized sets of integrals where regularity is not enough to achieve fast and precise approximations: Even though all integrands lie in the same function space, different specifications lead to better or worse convergence plots. This issue is clearly visible for Mixed Poisson and NSG models. Already a small change in the assumed covariance matrix, mean or other parameters strongly influences the performance of SG and Frolov quadrature.

This leads us to the conclusion that a direct relation between model and function space does usually not exist. A general preference for one quadrature rule for every specification of a particular model cannot be justified and a separate choice of quadrature has to be made for any newly considered model specification. Still, our survey offers some heuristics and first results about where the application of higher-order rules is beneficial for the estimation process.

In the following, we present approximation results for all of the described models and specifications from Chapter 2, i.e. various GLMM (Section 2.3) and two exemplary DEM (Section 2.4). As the integrals are not analytically solvable the approximated values cannot be compared with the exact result. Instead for all integrals a "true" result $\tilde{\mathcal{I}}(f)$ is computed with the respective best performing quadrature rule and a high number of nodes. For better comparability we then plot the relative error

$$R_l(f) = \frac{|Q_l(f) - \tilde{\mathcal{I}}(f)|}{|Q_l(f)}$$

against the number of nodes (i.e. function evaluations) N_l in each level l. Depending on the performance we present results for a varying set of quadrature rules. In general, MC and QMC quadrature based on Sobol and Halton sequences serve as benchmark for high-order rules. For Frolov cubature, we use the precomputed points by Kacwin et al. [53], available on `https://ins.uni-bonn.de/content/software-frolov` for integrals of dimension 10 or less. Finally, the applied SG quadrature rules are based either on the one-dimensional Gauss-Hermite or Clenshaw-Curtis formula. In some cases it proved to be sufficient to display convergence for one of the two formulas whereas both rules performed reasonable in other cases.

Bhat [6] was one of the first to investigate Halton-points for quadrature in econometrics and used them to compute a Mixed Logit model. The application of SG quadrature on Mixed Logit and Probit models was first investigated by Winschel and Heiss ([47], [48]), Judd and Skrainka [50] and Oettershagen ([33], [72]). Our simulations support their results and illustrate them in a broader context by including Mixed Poisson and Linear Quantile Mixed models and adding Frolov cubature as another alternative to sequence based QMC.
As yet, approximation of integrals in DDCM is based solely on Monte Carlo integration ([1], [15], [55]). Similarly, SG and Frolov quadrature apparently has not yet been considered for NSGM although the product rule seems to be sufficient in many cases ([49], [63]).
In the following two sections, we will denote quadrature based on Sobol- or Halton-sequences jointly as QMC quadrature and state results for Frolov cubature separately although it is also a QMC rule by nature. This notation is reasonable since Frolov cubature, in contrast to Sobol and Halton rules, is able to achieve optimal error bounds and has not yet been examined for the application to econometric integrals.

4.2 Quadrature for GLMM Integrals

In Chapter 3, we established two important criteria for the integrand to achieve high convergence rates, namely regularity and boundary conditions. We recall that for GLMM the integrand s is specified by the distribution of the outcome f, the link function g and the mixing distribution h and parameterized by mean and covariance of the latter

$$\begin{aligned} s(u|y, \beta_0, \Sigma, \phi_0) &= f(y|\,\beta_0, u, \phi_0)h(u|\Sigma) \\ &= \exp\left(\frac{y^T\tilde{g}(g^{-1}(X(\beta_0+u))) + b(\tilde{g}(g^{-1}(X(\beta_0+u))))}{a(\phi_0)} + c(y,\phi_0)\right) h(u|\Sigma)\,. \end{aligned}$$

The most popular combinations are presented in Table 1 and Section 2.3. In general, the subject of interest is the natural parameter θ, so $a(\phi_0)$ and $c(y, \phi_0)$ constitute constants in terms of u and the functions a and c can be neglected in the examination of regularity of s. The functions b and $\tilde{g}$ presented in Table 1 are all smooth in their domains except for possible singularities at the boundaries. Nevertheless, in the setting of an exponential family they yield smooth functions in terms of the mean μ. The only exception is the Laplace distribution with p.d.f. (2.15) which is only continuous in μ.

The dependence on μ is converted to a dependence on $\beta_0 + u$ via the link function g, more specifically via g^{-1}. Any (inverse of a) link function in Table 1 including the canonical links is again smooth in its domain, hence we expect the successful applicability of high-order quadrature methods.

This is also backed by the usual choice of a multivariate Gaussian for the mixing distribution h which is also smooth. In contrast, the multivariate Laplacian exhibits a singularity at 0 and, indeed, all formulas based on regularity failed immediately for it in our simulations so they are not included for further examination.

However, both Gaussian and Laplacian are defined on the whole space $\mathbb{R}^d$, so the (Q)MC rules characterized in sections 3.4 and 3.5 cannot directly be applied as they are designed for $\Omega = [0,1]^d$ (or any bounded domain). Hence, the integrand has to be transformed to fit Ω. This transformation naturally induces singularities at the boundaries. Yet, the term e^{-x^2} of the Gaussian can mitigate them if the singularities are of lower order. Simple trial and error showed that the *tangens*-transformation has sufficiently flat singularities, so the transformed integrand has zero boundary in Ω.

Analogous to the structure of Section 2.3, we start with integrals arising from the Mixed Logit model: For dimensions $q = 2, 4, 8, 12$ of the parameter vector $\beta = \beta_0 + u$ and $J = 8$ alternatives we draw a random vector $x \in [0,1]^{J \times q}$ of observed exogenous factors in the Mixed Logit model. The number of alternatives can be arbitrary in this setting as the probability is computed separately for every choice. We let $\beta_0 = (0.2, ..., 0.2)^T$ be a guess for the mean of the mixing distribution and let its covariance $\Sigma = \Sigma_\rho$ be parameterized by $\rho = 0.2$ where $\Sigma_{ii} = 1$ and $\Sigma_{ij} = \rho$.

In accordance with our knowledge about the regularity of the integrand, figure 3 displays the clear superiority of SG and Frolov quadrature compared to (Q)MC rules for low dimensions 2 and 4. The proposed rates from Section 3.3 even lead to exponential convergence. For $q = 8$ and $q = 12$ this behavior cannot be preserved as the log-factor dominates the performance before the main rate kicks in. This pre-asymptotic phase is problematic in real-world applications as econometrics only rarely require high precision approximations and are rather interested in rough but easily obtainable estimates.

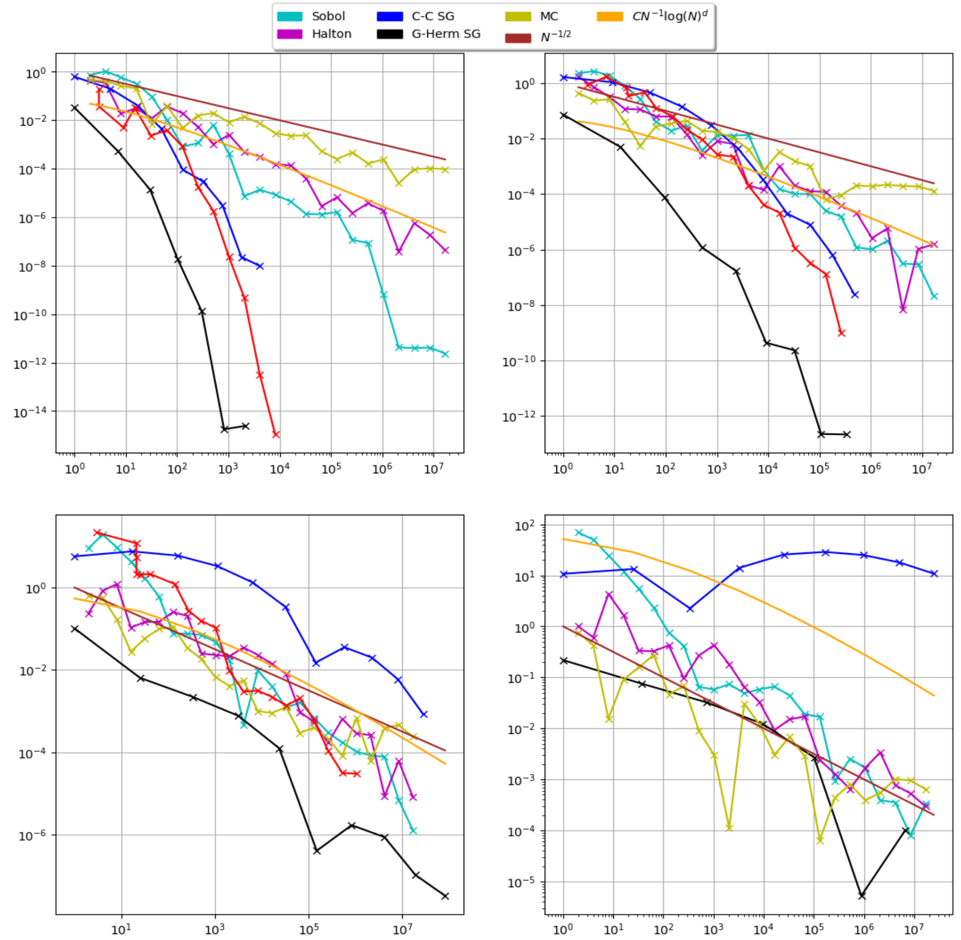

Figure 3: Multinomial Mixed Logit for dimensions 2, 4, 8, 12 (left to right, top to bottom) with Gaussian mixing distribution.

The general advantage of the Gauss-Hermite formula over Clenshaw-Curtis follows from its definition for the weight function e^{-x^2}. Thus, the Gauss-Hermite points have the whole real line as domain and the tangens-transformation can be omitted. The form e^{-x^2}, or e^{-x^Tx} in the multivariate case, can be recovered for arbitrary positive definite covariance matrix by linearly transforming the integration variable with the matrix C from the Cholesky decomposition $\Sigma = CC^T$.

Yet, the performance of SG quadrature not only depends on the dimension but is also affected by the chosen parameters. In figure 4, we investigated quadrature for fixed dimension $q = 4$ and scaled covariance matrices $\kappa\Sigma_{0.3}$ with $\kappa = 0.1,\ 2,\ 100,\ 1000$. In the latter two cases, SG quadrature does not exhibit any converging behavior at all, while Frolov cubature maintains a better rate.

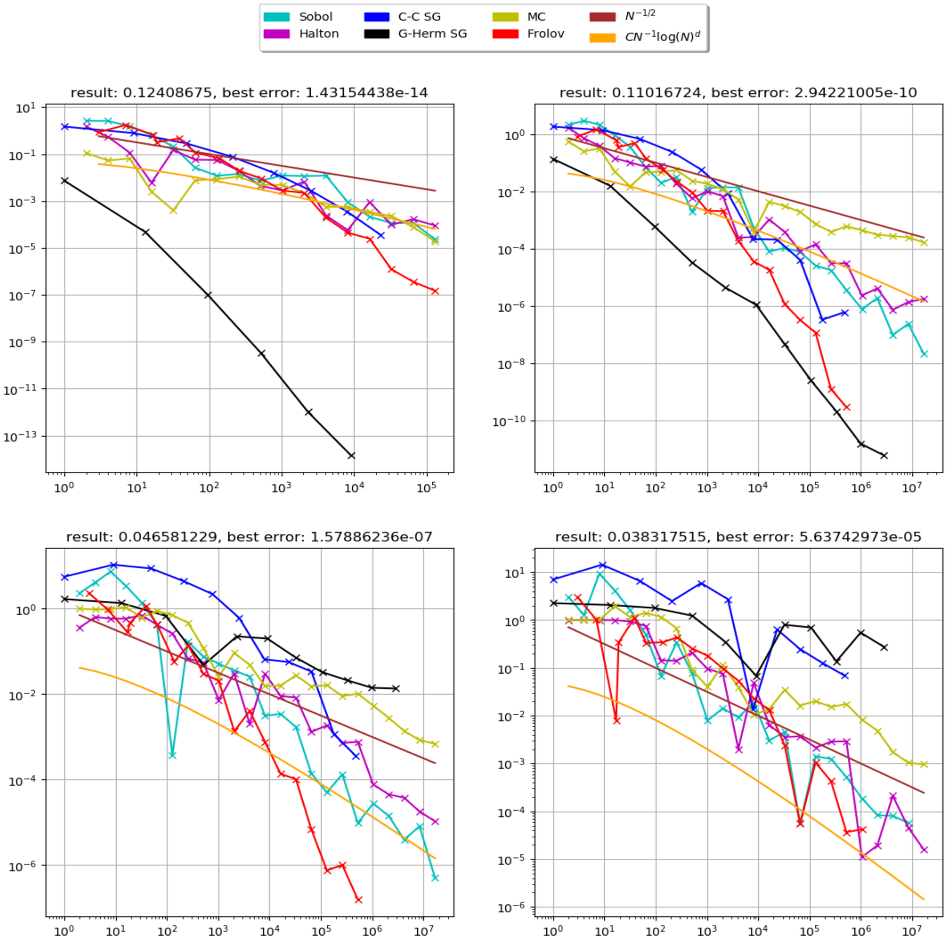

Figure 4: Multinomial Mixed Logit for dimensions 4 with Gaussian mixing distribution and covariance $\kappa\Sigma_{0.3}$ with $\rho = 0.3$ and $\kappa = 0.1,\ 2,\ 100,\ 1000$ (left to right, top to bottom).

We can explain the erratic, diverging nature of the SG plots in figure 4 despite the expected higher-order convergence rate with constants in the O-notation: In the previous chapter we estimated the approximation error by

$$\begin{aligned} E(f) = |\mathcal{I}(f) - Q(f)| \leq ||\mathcal{I} - Q|| \cdot ||f|| \leq C(r,d)N^{-r}\log(N)^{t(r,d)}||f|| \\ = O(N^{-r}\log(N)^{t(r,d)}) \end{aligned} \tag{4.1}$$

for regularity r and some exponent $t(r,d)$. The constant $C(r,d)$ can be exponentially dependent on r or d and therefore extend the pre-asymptotic phase notably before the main rate N^{-r} kicks in. The additional factor $||f||$ in the O-constant depends not only on f but also on the choice of a norm $||\cdot||$. Hence, predictions from the error bound (4.1) should be treated

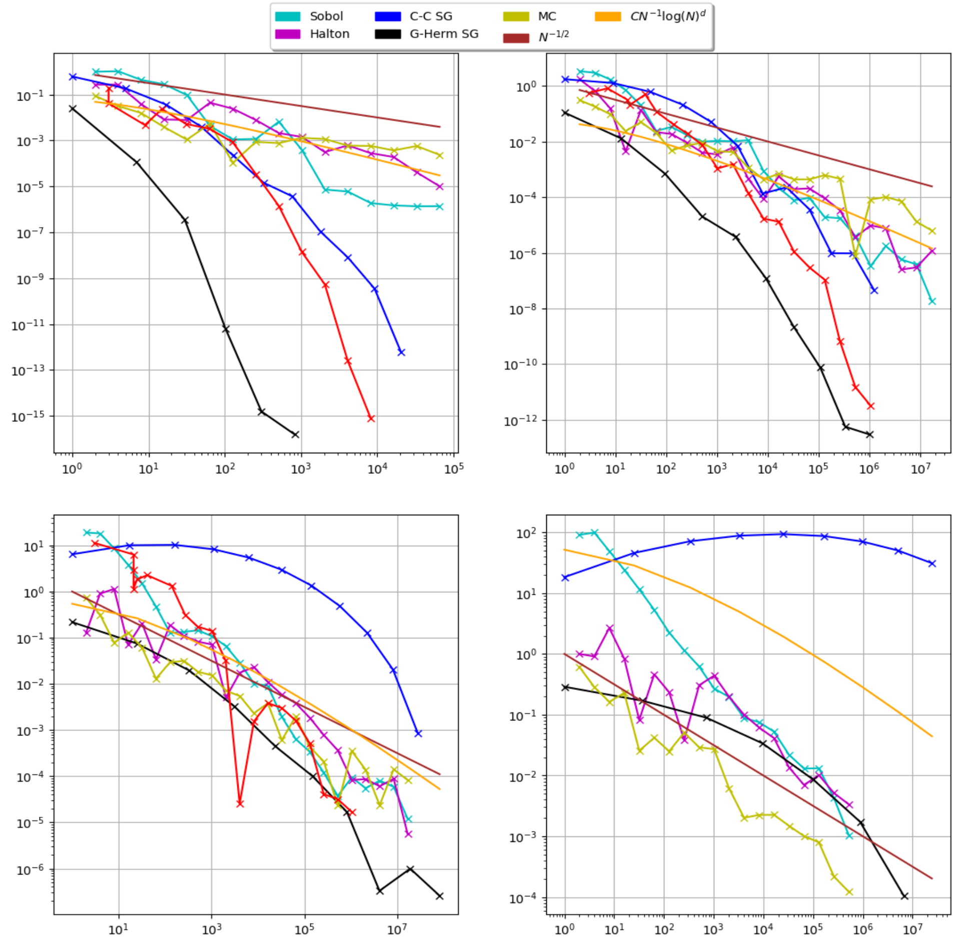

Figure 5: Mixed Poisson model for dimensions 2, 4, 8, 12 (left to right, top to bottom) with Gaussian mixing distribution $\mathcal{N}(0, 0.1\Sigma_{0.2})$.

sensibly for low N as the bound might strongly depend on "invisible" constants in the O-notation. While the log-factor only flattens the curve the O-constant leads to curves as in figure 4 or later in figures 6 and 12.

It particular, we realize that the analysis of quadrature rules for econometric parameterized integrals cannot be executed along the lines of classical numerical mathematics: For a set of integrals, e.g. all GLMM integrals, one would normally try to find a function space which includes all integrands and then determine a quadrature rule which achieves the optimal convergence rate in this function space. This approach is not practical here although the described integrands have sufficient regularity to be in a Sobolev space of mixed regularity, as the constant in the error bound is problematic for small N. On the other hand, it is not reasonable to consider adaptive quadrature methods, since it would require to much computational effort to generate

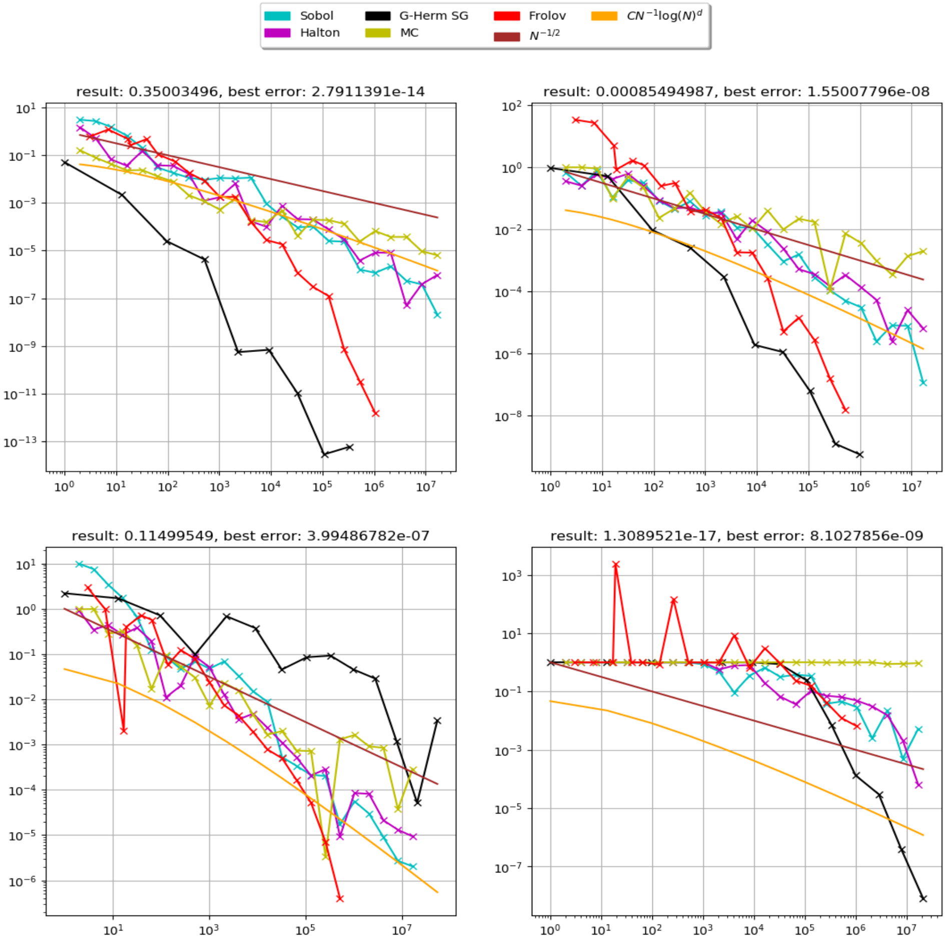

Figure 6: Mixed Poisson model for dimensions 4 with Gaussian mixing distribution $\mathcal{N}(\mu, \kappa\Sigma_{0.3})$ and combinations $(\mu, \kappa) = (0, 0.1)$, $(4, 0.1)$, $(0, 10)$, $(8, 0.1)$ (left to right, top to bottom).

an individual rule for each integral from the parameterized set of integrals. Instead, it is possible to identify regions of parameters where SG quadrature outperforms currently used methods and others where (Q)MC quadrature is optimal. In figure 4, we observe that low covariance parameters justify the use of SG quadrature whereas Frolov cubature is the better choice for higher covariance.

For the Mixed Poisson model in figure 5, we used the same specifications for x, J and β_0 but scaled $\kappa\Sigma_{0.2}$ with $\kappa = 0.1$. Then similar results as for Mixed Logit and low covariance could be achieved: Considerable improvements result from SG quadrature for low to moderate dimensions but they quickly vanish for higher q due to the unfavorable log-factor in the error bound and the resulting pre-asymptotic behavior. Still, figure 5 shows that

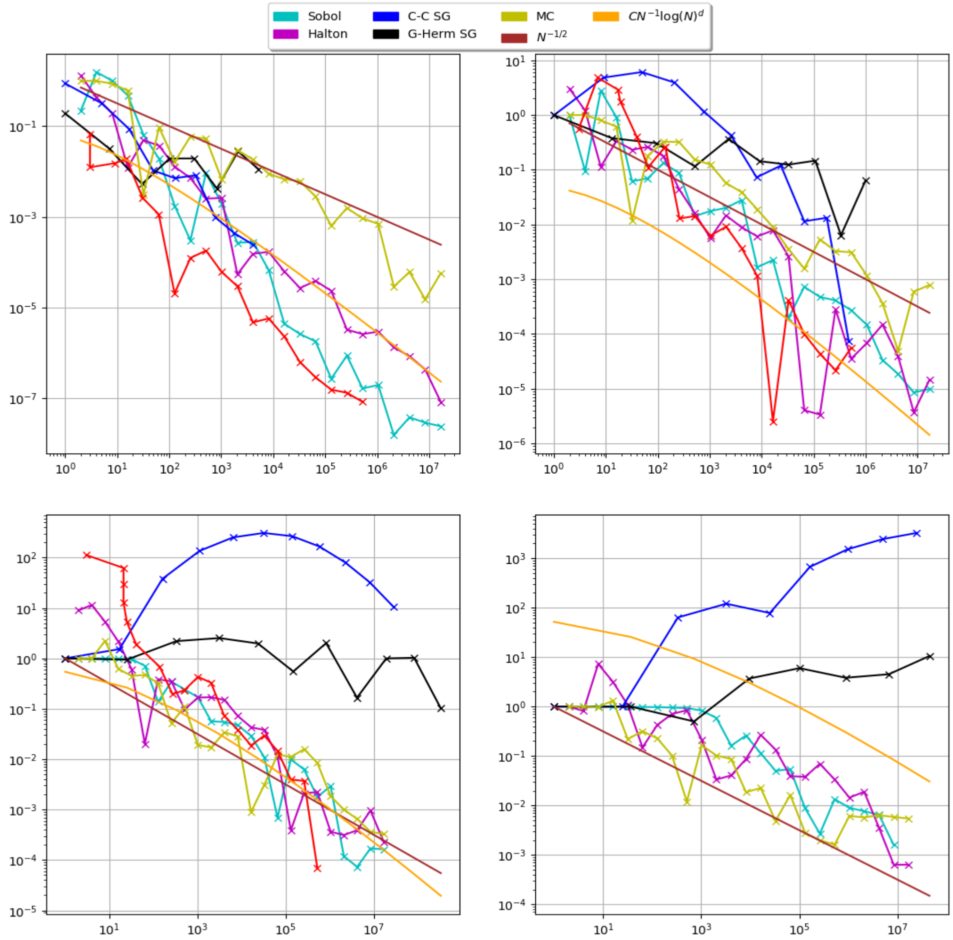

Figure 7: Linear Quantile Mixed model with Gaussian mixing distribution for dimensions 2, 4, 8, 12 (left to right, top to bottom).

a sparse grid based on the Gauss-Hermite rule can keep up with (Q)MC rules even in higher dimensions.

As for Mixed Logit these results depend on the choice of parameters. Figure 6 displays convergence plots for four combinations of mean μ and scale parameter κ of the Gaussian mixing distribution $\mathcal{N}(\mu, \kappa\Sigma)$: Again a high value for κ extends the pre-asymptotic phase of the SG rule significantly so that Frolov cubature and even Sobol- and Halton rules are better choices for moderately accurate approximations.

This is different for increased mean: Here SG quadrature is the best choice even for higher μ since the pre-asymptotic behavior is poor not only for SG but also for QMC rules. In particular, we note that SG quadrature indeed achieves an exponential convergence rate once the pre-asymptotic phase is left.

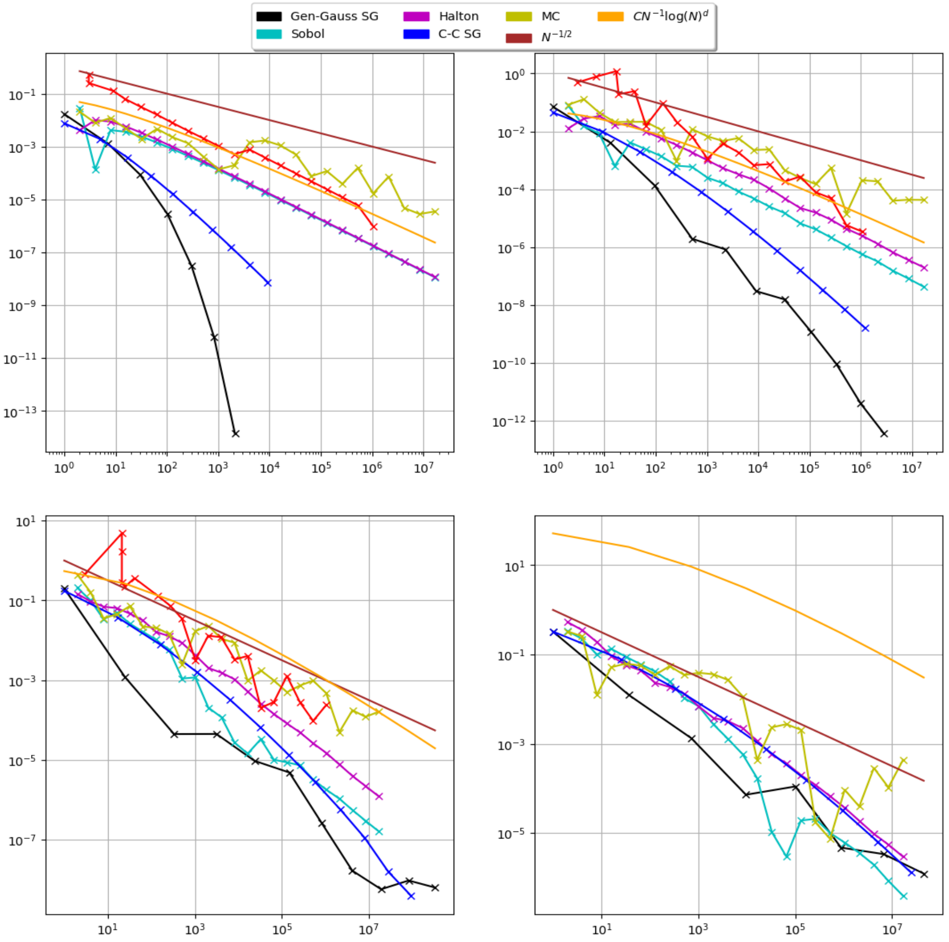

Figure 8: Multinomial Probit for dimensions 2, 4, 8, 12 (left to right, top to bottom).

For the Linear Quantile Mixed model the failure of SG quadrature is caused by the lacking regularity of the Laplace p.d.f. due to the absolute value in the exponent (see equation (2.15)). For non-differentiable functions the main term of any error bound is limited to $O(N^{-1})$, so the pre-asymptotic phase due to the log-factor extends even further for SG quadrature. Similarly, QMC rules follow the MC-rate $O(N^{-1/2})$ in their pre-asymptotic phase until the better main rate dominates. Hence QMC rules are superior to MC integration for low dimensions but match it for only higher dimensions and moderate N.
The Multinomial Probit choice probability defined by (2.14) cannot directly be approximated by the given quadrature rules. The Genz-algorithm [20] which is equivalent to the GHK-simulator ([39], [57]) transforms the integral

to the unit cube

$$\mathcal{I}(\mathbf{v}) = \int_{(0,1)^{d-1}} \prod_{i=1}^{d} \hat{v}_i(w_1, ..., w_{i-1}) dw \tag{4.2}$$

where $d = J - 1$ for the number of choices J and

$$\mathbf{v} = (\tilde{V}_{ki})_{i=1, i \neq k}^{J} = ((X\,\beta)_k - (X\,\beta)_i)_{i=1, i \neq k}^{J} \tag{4.3}$$

for a fixed choice $k \in \{1, ..., J\}$. The $\hat{v}_i$ are recursively defined by

$$\hat{v}_i(w_1, ..., w_{i-1}) = \Phi\left(C_{ii}^{-1} \cdot \left(\mathbf{v}_i - \sum_{j=1}^{i-1} C_{ij} \Phi^{-1}(w_j \hat{v}_j(w_1, ..., w_{j-1}))\right)\right).$$

Here, Φ is the c.d.f. of the standard univariate Gaussian and C is a factor from the Cholesky decomposition of Σ. The inverse c.d.f. Φ^{-1} induces a boundary singularity for the integrand in (4.2) but the product over the $\hat{v}_i$ is still analytic for $w \in (0,1)^{d-1}$. Due to the discontinuity in the integrand in (2.14) it is impractical to apply SG quadrature to the untransformed integral.

Instead, Oettershagen and Griebel [33] propose to use a Sparse Grid which is based on a generalized Gauss formula. In our case, this formula is generated similarly to conventional Gauss-formulas only that polynomials in $\log(x)$ are used instead of polynomials in x. This way, r-times differentiable functions with boundary singularities are included in the space of functions for which a main rate of $O(N^{-r})$ is achieved. This property is preserved for multidimensional integrands and SG quadrature.

In figure 8, we compute the Probit integral for covariance matrix $\Sigma_{0.2}$ and $\mathbf{v} = (0.5, ..., 0.5)$: For J =3, 5, 9, 13 (hence dimension 2, 4, 8, 12), the plots display that SG quadrature again surpasses (Q)MC for low dimensions while the secondary rate again reduces this advantage for higher dimensions. Frolov cubature does not perform well for Probit as the transformed integrand does not meet the zero boundary condition.

Oettershagen and Griebel explored further variants of generalized Gauss-SG quadrature with other basis functions instead of common polynomials and showed that the above results also sustain for other parameter combinations.

We summarize the results for GLMM integrals in Table 2 where each cell indicates whether the respective quadrature rule performs well in terms of its optimal convergence rate for the respective model and specification. The table illustrates that MC and QMC quadrature provide viable methods for a wide variety of integrals but are often surpassed by at least one of the higher-order rules.

In particular, it makes sense to switch between quadrature methods depending on the choice of parameters and specification of the model. Often,

Model	MC: $O(N^{-1/2})$, Sobolev/Halton-QMC: $O(N^{-1}\log(N)^{d-1})$	SG (Gauss-Hermite or CC): $O(N^{-r}\log(N)^{(d-1)(r+1)/2})$	Frolov: $O(N^{-r}\log(N)^{(d-1)/2})$
Mixed Logit (small κ)	Yes.	Yes (for small to moderate d).	Yes (for small d).
Mixed Logit (large κ)	Yes.	No.	Yes (but deteriorating for larger κ).
Mixed Poisson (small κ, μ)	Yes.	Yes.	Yes.
Mixed Poisson (large μ)	Yes.	Yes (but prolonged pre-asymptotic).	Yes (but prolonged pre-asymptotic).
Mixed Poisson (large κ)	Yes.	No (possibly prolonged pre-asymptotic).	Yes
LQMM	Yes.	No.	Yes, but only rate $O(N^{-1}\log(N)^{d-1})$.
Probit	Yes.	Yes (for small to moderate d).	No (only MC rate).

Table 2: Overview of whether quadrature rules achieve their optimal convergence rates for given models and specifications.

it is possible to utilize polynomial or even exponential convergence rates of higher-order rules if small variance is assumed. On the other hand, QMC and Frolov quadrature achieved better results for high variance, undermining the importance of the the careful choice of quadrature for each newly considered parameter.

4.3 Quadrature for DEM Integrals

Similar to the previous section, we now present some numerical results for the integrals derived in Section 2.4, first for DDCM and then for NSGM. For DDCM, we replace the non-differentiable integrand by a smoothed maximum function and investigate how quadrature is affected by this change. For NSGM, we encounter the same dependence of the approximation performance on the choice of parameters as for some GLMM integrals.

Researchers have used different functions to model the utility (2.17) in DDCM. Rust [79] uses linear functions whereas Keane and Wolpin and Eisenhauer [16] use functions involving exponentiation. We can generalize both settings to get

$$g(x) = \max\{c_1x_1, ..., c_Kx_K, e^{c_{K+1}x_{K+1}}, ..., e^{c_dx_d}\} \tag{4.4}$$

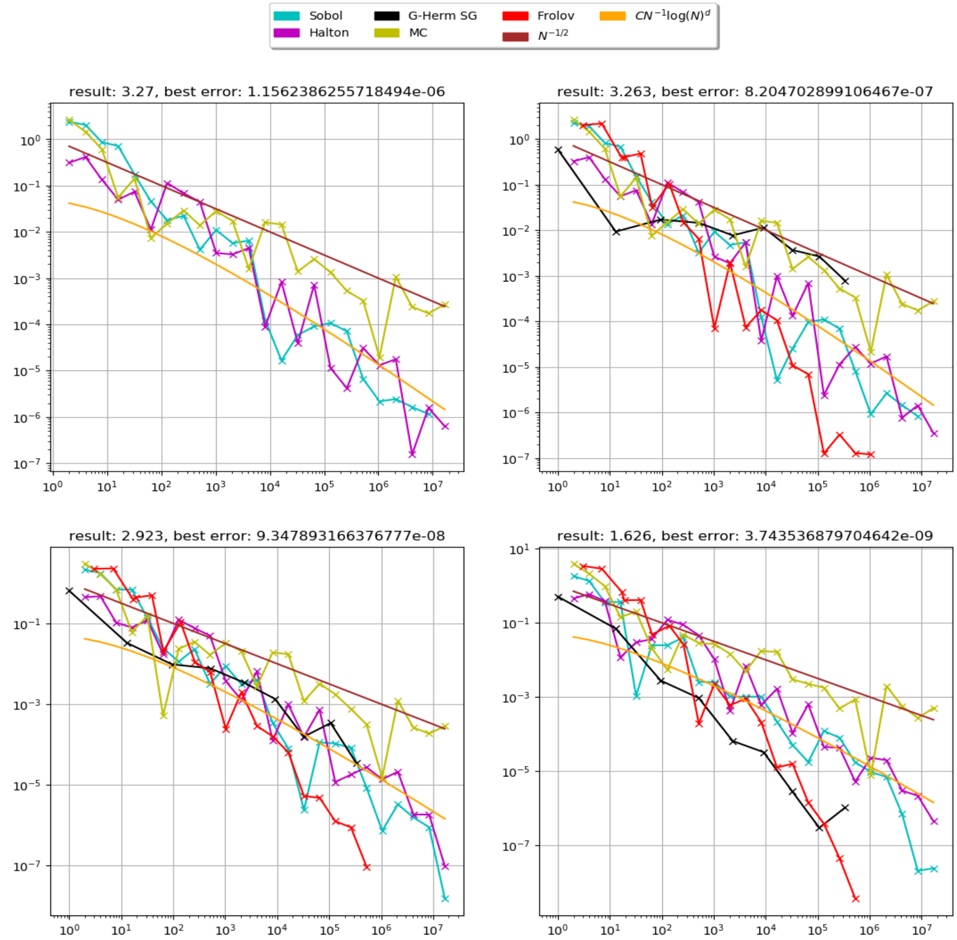

Figure 9: Dynamic Discrete Choice model for dimension 4 with real, smoothed ($\alpha = 10$), smoothed ($\alpha = 1$), smoothed ($\alpha = 0.1$) maximum.

and

$$g(x) = \max\{c_1 + x_1, ..., c_K + x_K, c_{K+1}x_{K+1}, ..., c_d x_d\} \tag{4.5}$$

respectively. The number $0 \leq K \leq d$ depends on the modeling choices made by the researcher.

The maximum function introduces a kink into the integrand resulting in $g \in H^{3/2}_{\text{mix}}$. Although the arguments of the maximum are smooth we can therefore not expect the same exponential or polynomial convergence as in the previous chapter. Instead of the true kinked maximum we use the smoothed maximum function

$$S(u_1, ..., u_d|\alpha) = \frac{\sum_{i=1}^d u_i e^{\alpha u_i}}{\sum_{i=1}^d e^{\alpha u_i}}$$

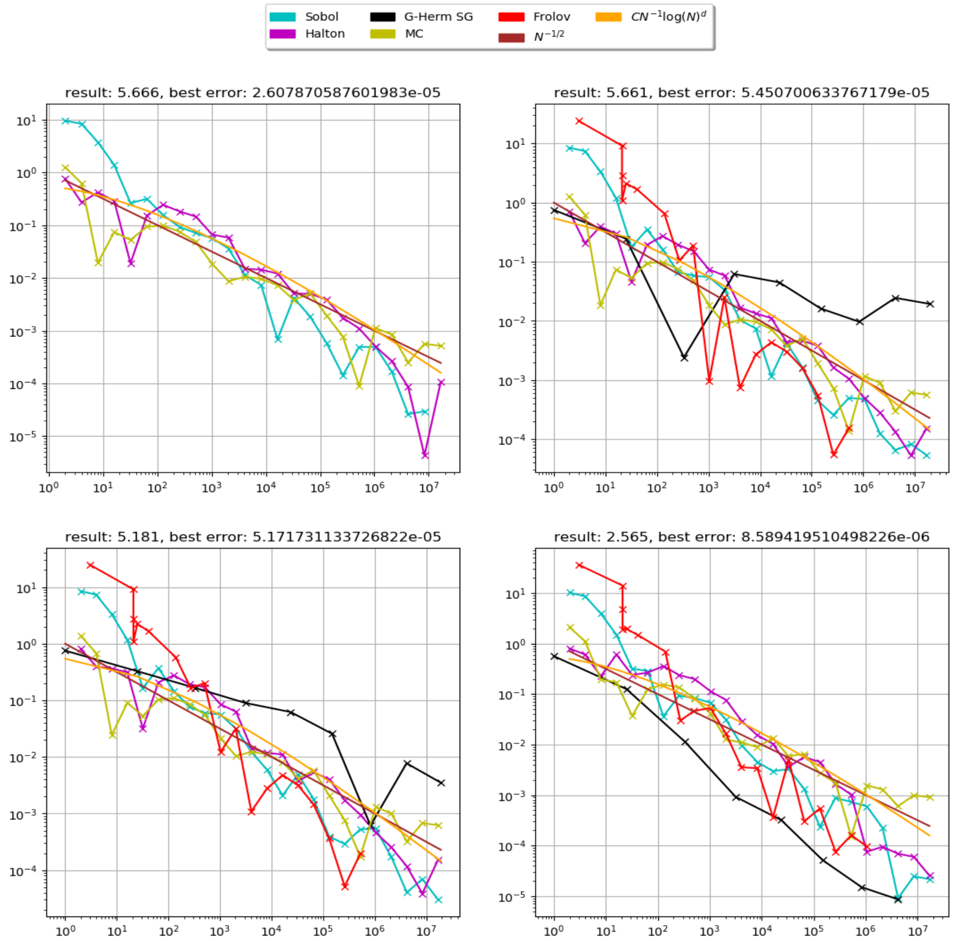

Figure 10: Dynamic Discrete Choice model for dimension 8 with real, smoothed ($\alpha = 10$), smoothed ($\alpha = 1$), smoothed ($\alpha = 0.1$) maximum (left to right, top to bottom).

where the smoothing gets stronger for smaller $\alpha > 0$.

Figure 9 displays convergence plots for true and smoothed maxima with $\alpha = 10, 1, 0.1$ and $d = 4$. We used g from (4.4) with $K = 2$, $c_i = 1$ for all $i = 1, .., d$ and covariance matrix $\Sigma = \Sigma_{0.2}$ with similar results arising for g from (4.5) and other choices of c and Σ.

For strong smoothing, SG and Frolov quadrature indeed outrun (Q)MC rules which is also visible for higher dimension $d = 8$ (see figure 10). In particular, SG and Frolov also perform better for the smoothed maximum function than Sobol and Halton do for the true maximum function, therefore offering a faster converging substitute for QMC rules.

However, for all integrands also the "true" results (i.e. best approximations) are displayed. We observe that the results for strongly smoothed integrands

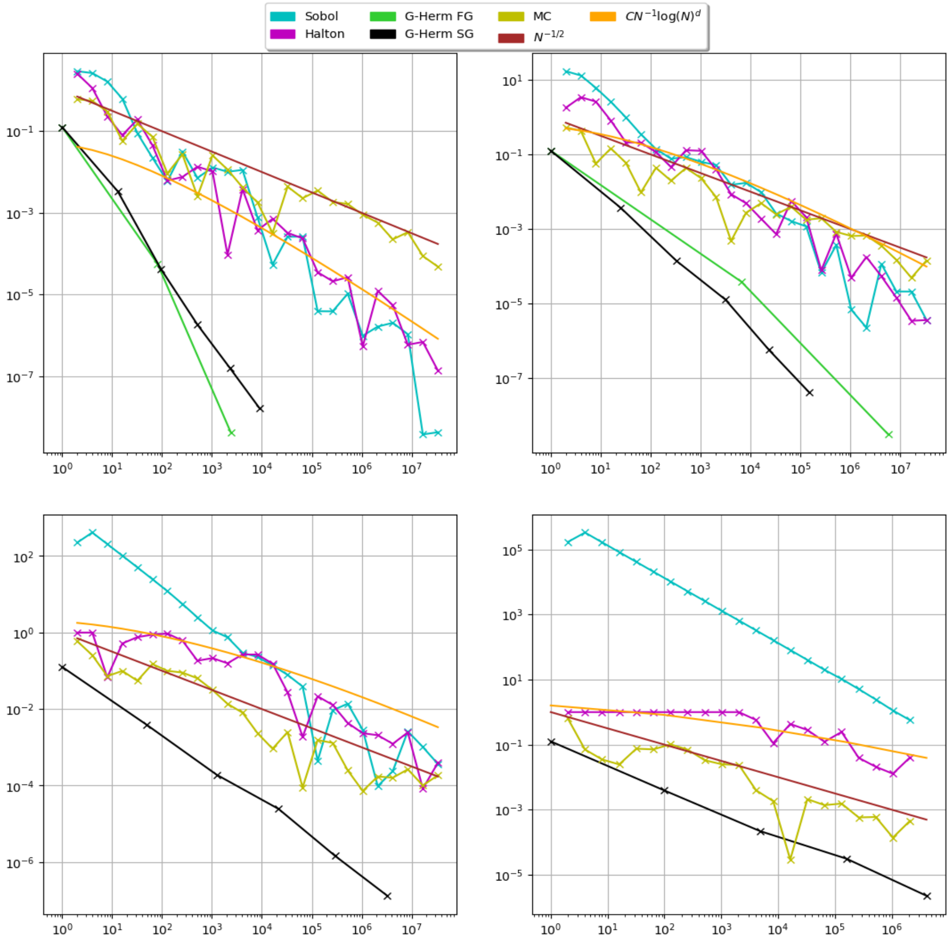

Figure 11: Neo-classical stochastic growth model for dimensions 4, 8, 16, 32 (left to right, top to bottom).

highly diverge from the real maximum result. Hence, applying fast converging SG or Frolov quadrature to a smoothed integrand is no practicable alternative to approximating the true integral (2.17). Only if the use of the smoothed maximum function can be justified econometrically, e.g. by a trembling hand approach, higher-order formulas could be applied successfully.

Finally, we consider the integral (2.21) which arises from a prototypical NSGM. As yet, the utility function u and the production function f remained unspecified. We rely on the specifications in [49],

$$u(c) = \begin{cases} \frac{c^{1-\gamma}-1}{1-y} & \text{for } \gamma > 0, \gamma \neq 1 \\ \log(c) & \text{for } \gamma = 1 \end{cases}$$

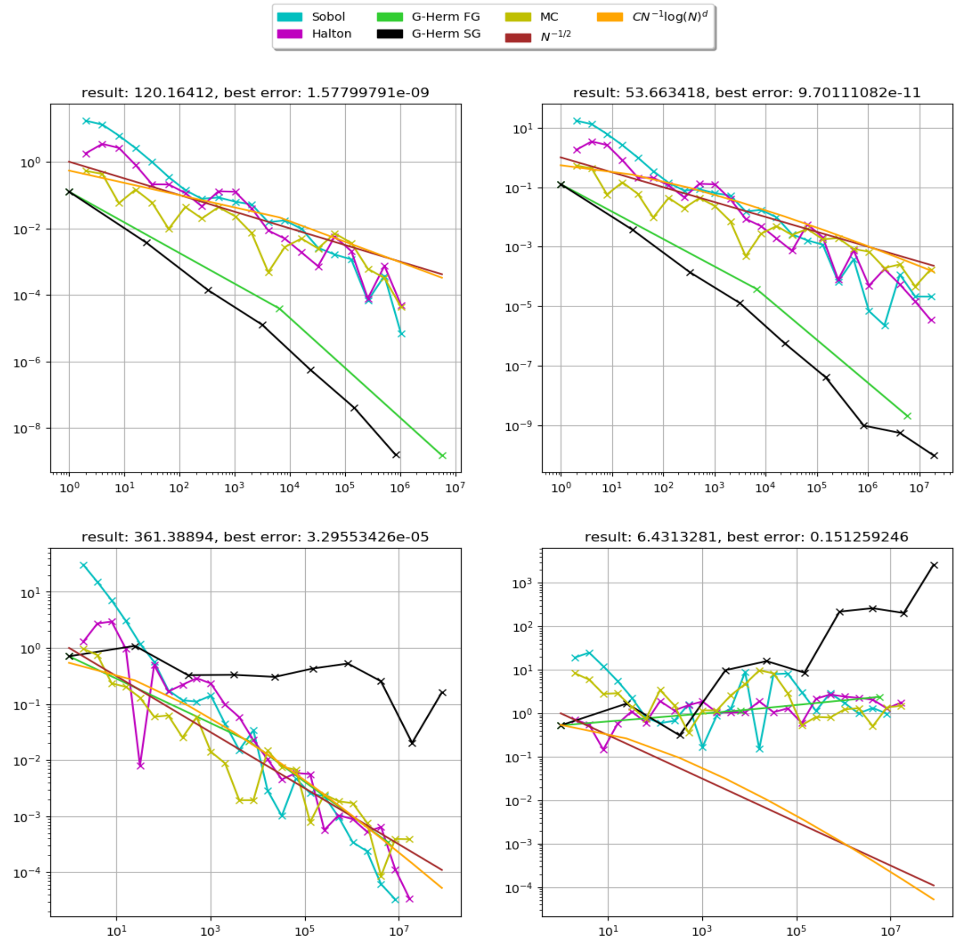

Figure 12: Neo-classical stochastic growth model for dimension 8 and base case $\gamma = 1$, $\sigma = .5$, $\delta = 0.02$ (upper left corner) and with modifications $\gamma = 0.1$ (upper right corner), $\sigma = 10$ (lower left corner), $\delta = 1$ (lower right corner).

and

$$f(k) = k^{\alpha} \text{ for some } \alpha > 0\,.$$

Furthermore, we let $k_{t+1} = 0.95k_t + 0.05\bar{k}a_t$ be the iteration method for capital and $\bar{k}$ the steady-state capital and get an iterative process for c_t by rewriting the budget constraint (2.19). Together with the original parameters $\beta, \delta, \rho, \Sigma$, we now have the vector $\theta = (\alpha, \beta, \gamma, \delta, \rho, \Sigma, \bar{k})$ which fully parameterizes the model. As we do not want to estimate the complete model we need to fix values for θ and provide start values for the iteration of k_t. We set $\alpha = 0.36,\ \beta = 0.99,\ \rho = 0.95,\ \bar{k} = 10$, assume $k_t = \bar{k}$ and $a_t = 1.2$ and let Σ be of the form $\Sigma_{ii} = 2\sigma^2$, $\Sigma_{ij} = \sigma$ for $i \neq j$, $i, j = 1, ..., J$ and

$\sigma > 0$.
In figure 11, we further set $\sigma = 0.5$, $\gamma = 1$ and $\delta = 0.02$ and varied the number of countries which are modeled. We considered $J = 4, 8, 16, 32$ and see that SG quadrature based on the Gauss-Hermite formula outperforms (Q)MC rules even for high dimensions. In particular, we observe exponential convergence rates for $J = 4$ and $J = 8$ and less clearly for $J = 16$ and $J = 32$. For low dimensions, the product rule is an alternative as it achieves similar convergence rates but it becomes infeasible for higher dimensions since the number of nodes per level increases exponentially in the dimension.
Additionally, the secondary log-rate is visible for both QMC and SG quadrature throughout all plots: The slopes decrease with rising dimensions so that Halton and Sobol are entirely impractical for high dimension. Equivalently the dominance of SG over MC quadrature diminishes.
This behavior is strongly dependent on the chosen parameter vector θ. In figure 12, we demonstrate how a slight perturbation of γ ,δ and σ affects the integrand and has major impact on the performance of SG quadrature. As base case (in the upper left corner) we use the configuration from figure 11 for dimension 8. Then we separately set $\gamma = 0.1$ (upper right corner), $\sigma = 10$ (lower, left corner) and $\delta = 1$ (lower right corner) to get plots for different parameter configuration.
While the change in γ only affected the result of the approximation (53.663 instead of 120.164) but left the convergence behavior for all rules untouched, this is considerably different for σ and δ: Similar to Mixed Logit and Mixed Poisson, SG quadrature performs notably worse for increased covariance $\sigma = 10$ and does not leave the pre-asymptotic phase for small to moderate N. The product rule does not exhibit this unfavorable pre-asymptotic behavior and achieves a similar rate as the (Q)MC rules. However, it does not surpass them due to the relatively high dimension of the integral and the curse of dimensionality. Hence, for higher dimensions and if moderately accurate approximations are sufficient, (Q)MC rules are again the superior choice.
For $\delta = 1$, approximation was unsuccessful for all applied quadrature methods. Since regularity of the integrand is not affected by the choice of δ we assume that this is again caused by high constants in the O-notation. For MC integration it means that the integrand exhibits a high variance for $\delta = 1$ while $||f||$ influences convergence for deterministic rules.

We again collect our findings for different models and specifications in Table 3. While MC and QMC quadrature still perform reasonably for almost all cases, the dependence of higher-order rules on the chosen parameters is even more visible than in the previous section. For SG quadrature, we assume that the slow or non-existing convergence only holds for small to moderate N and the main rate kicks in for larger N. However, this is often

Model	MC: $O(N^{-1/2})$, QMC: $O(N^{-1}\log(N)^{d-1})$	SG (Gauss-Hermite or CC): $O(N^{-r}\log(N)^{(d-1)(r+1)/2})$	Frolov: $O(N^{-r}\log(N)^{(d-1)/2})$
DDCM (real maximum)	Yes.	No.	No.
DDCM (mild smoothing)	Yes.	No (possibly prolonged pre-asymptotic).	Yes, but only rate $O(N^{-1}\log(N)^{d-1})$.
DDCM (strong smoothing)	Yes.	Yes (for small d).	Yes (for small d).

Model	MC: $O(N^{-1/2})$, QMC: $O(N^{-1}\log(N)^{d-1})$	SG (Gauss-Hermite or CC): $O(N^{-r}\log(N)^{(d-1)(r+1)/2})$	FG: $O(N^{-r/d})$
NSGM (small σ, δ, medium-sized γ)	Yes.	Yes.	Yes.
NSGM (small γ)	Yes.	Yes.	Yes.
NSGM (large σ)	Yes.	No (possibly prolonged pre-asymptotic).	(Yes, but too few points to evaluate).
NSGM (large δ)	No.	No (possibly prolonged pre-asymptotic).	No.

Table 3: Overview of whether quadrature rules achieve their optimal convergence rates for given models and specifications.

no option for econometricians as they require fast and rough approximations rather than highly accurate results.

Part II
Quadrature in Estimation

5 Econometric Estimation

5.1 Objective

Given a theoretical economic model the next step is to evaluate this model with respect to real world data. There are various ways to collect such data depending on the object of investigation: Economists who study the behavior of individuals in microeconomic situations can design studies where they observe the behavior of test persons in a controlled environment. Other models might require data from real economic actions of and interactions between individuals, companies or countries. Public agencies, research institutions, banks, insurance and software companies have collected large amounts of data during the last decades which are suitable for this task.

In this chapter we examine methods for the evaluation of a parametric model, i.e. the structural form of the model is fixed but flexible in terms of a set of parameters. Finding the best parameter of a model based on the available data is called *estimating a model.* The optimal parameter for a given N-point data set $\mathcal{Z}_N = \{z_1, ..., z_N\}$ in the parameter set Θ is called the *estimator* θ_N. The *true* or *population parameter* of the model is denoted by θ_0.

If the model under consideration is parameterized by q parameters, then $\Theta \subset \mathbb{R}^q$ and all parameters θ are in fact parameter vectors. All following statements regarding θ can be extended to this case with the necessary alterations for each statement (i.e. vectors or matrices instead of scalars whenever necessary).

One data point z_n incorporates all observable variables, including exogenous factors and outcomes (i.e. x_n and y_n in the setting of Chapter 2), so it might be a vector or a matrix. W.l.o.g. we let $z \in \mathcal{Z} \subset \mathbb{R}^r$ and combine, where needed, exogenous variable x and outcome variable y into the tuple $z = (x, y)$. Furthermore, we omit the index n if we do not talk about a specific data point. We assume that the z_n are independent identically distributed (i.i.d.) samples from a known or unknown distribution. If the data is structured as panel data or correlated for other reasons, we can build independent sets of data points and denote them as $z_1, ..., z_N$. E.g. panel data contains $N \cdot T$ data points where N is the number of individuals and T the number of time periods, so we can typically merge T data points for one individual into one variable z. This way we again get independent data points $z_1, ..., z_N$.

In this chapter we describe desirable properties of estimators and introduce the concept of extremum estimators which provide a general class for many popular estimation methods. Following the presentation by Hayashi [45] we investigate two major sub-classes, namely M-estimators and Generalized Method of Moments estimators, both of which also include Maximum

Likelihood estimation. Newey and McFadden [69] and the introductory paper by Hansen [43] describe under which conditions they are good estimators in terms of the beneficial properties in Definition 5.1.
Subsequently, we observe that in many use cases the computation of an estimator requires the numerical integration of an integral (see e.g. [27]). In fact, M- and GMM-estimators are by definition simulators of an expected value. Motivated by similar investigations for Maximum Likelihood estimation [34] we develop a framework for approximated M- and GMM-estimators and show under which conditions they retain the following properties.

Definition 5.1. *(Desirable properties for estimators)*
An estimator θ_N is called

(I) ***consistent*** *if* $\lim_{N\to\infty} P\left(|\theta_N - \theta_0| > \varepsilon\right) = 0$ *for all* $\varepsilon > 0$*, i.e.* $\theta_N \xrightarrow{P} \theta_0$ *converges in probability for* $N \to \infty$*,*

(II) ***asymptotically normal*** *if:* $\sqrt{N}(\theta_N - \theta_0) \xrightarrow{D} \mathcal{N}(0, V)$ *for some scalar- or matrix-valued variance* V*,*

(III) ***efficient*** *if:* $\mathrm{MSE}[\theta_N] \leq \min_{\theta\in\Theta_N} \mathrm{MSE}[\theta_N]$, *where* $\mathrm{MSE}[\theta] = \mathrm{E}[|\theta - \theta_0|^2]$ *is the mean squared error and* Θ_N *is the set of admissible estimators given the data* $\mathcal{Z}_N$*. If the estimator is unbiased, then the mean squared error equals the variance. If it is additionally vector-valued,* MSE *is replaced by the covariance matrix of* θ_N *and for two matrices* A, B*, "*$A \leq B$*" means* $B - A$ *is positive semi-definite.*

Consistency can be regarded as the most basic property as it just demands that θ_N converges in probability to the population parameter for large N. In particular, a consistent estimator is also *unbiased* as we have $\mathrm{E}\left(\theta_N\right) = \theta_0$.
Asymptotic Normality (stemming from the Central Limit Theorem) implies that we can approximate the distribution of the estimator asymptotically by a (multivariate) Gaussian distribution. An asymptotically normal estimator is also consistent.
An *efficient* estimator outperforms other estimators in a certain class of estimators (e.g. all consistent or asymptotically normal estimators) based on a loss function which measures the error w.r.t. the population parameter θ_0. The most common loss function is the *mean squared error* MSE which is equal to $\mathrm{Var}(\theta)$ for unbiased θ. Efficiency is often achieved asymptotically, e.g. for the Maximum Likelihood estimator. The Cramér-Rao inequality defines a lower bound for the variance for unbiased estimators, provided the distribution fulfills two weak regularity conditions.
Many estimators only perform well for certain model assumptions, like specific distributions for the unobservable variables. For example, an estimator

might be efficient if the data is normally or almost normally distributed but may fail for heavy-tailed distributions. This is also the case if low-tails are assumed but the data contains many or extreme outliers. We call θ_N *robust* if it is un- or only lightly affected by such issues.

In the following we investigate estimators from the class of *extremum estimators*. Hayashi [45] defines them as

$$\theta_N = \operatorname*{argmax}_{\theta \in \Theta} Q_N(\theta)\,,$$

where we assume that Q_N is continuous in the data $z_1, ..., z_N$. Gouriéroux and Monfort [28] show that the estimator θ_N exists and is measurable in the data $z_1, ..., z_N$ if $\Theta \subset \mathbb{R}^q$ is compact and Q_N is measurable in $z_1, ..., z_N$ and continuous in θ. Although these assumptions are not always fulfilled in econometrics we do not further discuss them here, since we focus on computing the objective function Q_N rather than investigating the maximization process.
Newey and McFadden state two theorems regarding consistency (Theorem 2.1) and asymptotic normality (Theorem 7.1) of extremum estimators. We revisit them in order to provide a basis for further considerations.

Lemma 5.2. *(Consistency of the extremum estimator)*
Let $Q_0(\theta)$ be a function s.t.

(i) *$Q_0(\theta)$ is uniquely maximized at θ_0,*

(ii) *Θ is compact,*

(iii) *$Q_0(\theta)$ is continuous for all θ, and*

(iv) *$Q_N(\theta)$ converges uniformly in probability to $Q_0(\theta)$, i.e.*

$$\sup_{\theta \in \Theta} |Q_N(\theta) - Q_0(\theta)| \xrightarrow{P} 0$$

as $N \to \infty$.

Then $\theta_N \xrightarrow{P} \theta_0$ for $N \to \infty$.

Newey and McFadden provide two theorems concerning the asymptotic normality of extremum estimators. The first demands that Q_N is twice continuously differentiable in a neighborhood of the population parameter θ_0. This condition is too strong for our purposes since we investigate the estimation of approximated objective functions. Approximation does not necessarily preserve differentiability as was also noticed by Newey and McFadden (e.g. for simulated Probit choice probabilities). Therefore, they specify a second, more general theorem which is tailored to this issue and

only requires differentiability of Q_0 at θ_0.
Let D_N be the derivative or approximate derivative of Q_N at θ_0 and define the error term w.r.t. D_N as

$$A_N(\theta) = \sqrt{N}\frac{Q_N(\theta) - Q_N(\theta_0) - D_N(\theta - \theta_0) - (Q_0(\theta) - Q_0(\theta_0)}{||\theta - \theta_0||}.$$

Here and later on, if not stated otherwise, $||\cdot||$ denotes the Euclidean norm for vectors and the operator norm for matrices, functions and other operators. Then we have the following theorem:

Lemma 5.3. *(Asymptotic Normality of the extremum estimator)*
Assume that

(i) $Q_N(\theta_N) \geq \sup_{\theta\in\Theta} Q_N(\theta) - o_p(N^{-1})$, *i.e.* Q_N *is almost maximized at* θ_N,

(ii) $\theta_N \xrightarrow{P} \theta_0$ *as* $N \to \infty$, *i.e.* θ_N *is consistent,*

(iii) $\theta_0 \in \overset{\circ}{\Theta}$, *i.e.* θ_0 *is an interior point of* Θ,

(iv) Q_0 *is twice differentiable at* θ_0 *with invertible second derivative* $H = \nabla_{\theta\theta} Q_0(\theta_0)$,

(v) $\sqrt{N} D_N \xrightarrow{d} \mathcal{N}(0, \Psi)$, *and*

(vi) for any sequence $\delta_N \to 0$, *we have*

$$\sup_{||\theta-\theta_0||\leq\delta_N} \frac{|A_N(\theta)|}{1 + \sqrt{N}||\theta - \theta_0||} |A_N(\theta)| \xrightarrow{P} 0$$

Then $\sqrt{N}(\theta_N - \theta_0) \xrightarrow{d} \mathcal{N}(0, H^{-1}\Psi H^{-1})$.

Here, $X_N = o_p(c_N)$ denotes the Landau-notation in probability for a series of random variables X_N and a set of constants c_N. It implies that X_N/c_N converges to 0 in probability.
Lemmas 5.2 and 5.3 now provide a framework in which more specific conditions for exact and approximated M- and GMM-estimators can be defined.

5.2 M-Estimators

The *M-estimator* is defined for the objective function

$$Q_N(\theta) := \frac{1}{N}\sum_{i=1}^{N} m(\theta|z_i)$$

where $m : \Theta \times \mathcal{Z} \to \mathbb{R}$. The function m is usually also denoted by ρ and yields a consistent estimator if it is continuously differentiable and an asymptotically normal estimator if it is twice continuously differentiable. As we will focus on the prevalent Maximum Likelihood estimator, we refer to Amemiya [2] and Gouriéroux and Monfort ([28], chapter 8) for further discussions on properties of M-estimators and more examples.

The objective function $Q_N(\theta)$ can be seen as the sample mean of the random variable $m(\theta|Z)$. If the observations $z_1, ..., z_N$ are drawn independently, Q_N is equivalent to the Monte Carlo simulation (cf. Section 3.4) of the integral

$$\mathrm{E}(m(\theta|z)) = \int_{\mathcal{Z}} m(\theta|z) dF(z) =: Q_0(\theta)\,. \tag{5.1}$$

We call Q_0 the *integrated M-estimator.*

The most commonly used M-estimator is the *Maximum Likelihood estimator.* In Discrete Choice models (cf. Section 2.3) we characterize a choice probability $P(y_n = j|x_n, \theta)$ for each individual $n = 1, ..., N$ and alternative $j = 1, ..., J$ (recall the notation $z = (x, y)$). In other contexts and models, y_n might reprise an outcome with continuous support. Then, $P(y_n|x_n, \theta)$ denotes the value of the p.d.f. of the assumed distribution for the outcome at the argument y_n. In this case, P is usually replaced by f but for the sake of consistency we keep the present notation with P.

The goal of Maximum Likelihood estimation (MLE) is to find the parameter that maximizes the joint probability of the independent data points $x_1, ..., x_N$,

$$\mathcal{L}(\theta|z_1, ..., z_N) := \prod_{n=1}^{N} P(y_n|x_n, \theta)\,.$$

Then $\mathcal{L}(\theta|z_1, ..., z_N)$ denotes the *likelihood* of any parameter θ for given data $\mathcal{Z}_N$.

The maximization can be approached in two ways: We can either maximize the objective function directly or differentiate and equate with 0. The latter leads again to a minimization problem (which is numerically the same as maximization) if the resulting equation or system of equations is not analytically solvable. Train [83] describes multiple common maximization algorithms including Newton-Raphson, BHHH-maximization and Steepest ascent, which can also be applied to the analogous problem in Generalized Method of Moments estimation.

Maximization algorithms are more likely to converge to the true solution if the objective function is concave w.r.t. the maximization variable θ since concavity yields a unique maximum. Since many common probability distributions are only logarithmically concave we apply the logarithm to $\mathcal{L}$ and get the more convenient *Loglikelihood function*

$$\ell(\theta|\mathcal{Z}_N) := \log\left(\mathcal{L}(\theta|\mathcal{Z}_N)\right) = \sum_{n=1}^{N} \log(P(z_n|\theta)).$$

This obviously fits in the M-estimator scheme with

$$m(\theta|z) = \log(P(z|\theta)) .$$

Under certain conditions [69] the ML-estimator is consistent and asymptotically normal but there are more robust alternatives like *Maximum Spacing Estimation* which also provides an M-estimator [75]. The ML-estimator is efficient among asymptotically normal GMM estimators (see next section) but not among all asymptotically normal estimators. Therefore, given a reasonably difficult model and modest model assumptions, the use of Maximum Likelihood is wide-spread throughout econometrics and other data-driven disciplines like biology, medicine or psychology.

5.3 Generalized Method of Moments (GMM) Estimators

The computation of the likelihood (i.e. the choice probability in DCMs) requires knowledge or well justified assumptions on the distribution of the data $\mathcal{Z}_N$. If this requirement cannot be satisfied, the *Generalized Method of Moments (GMM) estimator* provides a less restrictive alternative. It was originally formalized by Hansen [43] and extensive discussions and lists of applications in econometric research can be found in [40] and [45]. It roots in solving the estimating equations

$$G_0(\theta) := \mathrm{E}_z[m(\theta|z)] = 0 \tag{5.2}$$

where the expected value is taken over the complete data domain $\mathcal{Z}$. The *moment function* $m : \Theta \times \mathcal{Z} \rightarrow \mathbb{R}^q$ is defined such that (5.2) is uniquely solved by the true parameter vector θ_0, and such that $G_0(\theta) \neq 0$ for $\theta \neq \theta_0$. The moment function should be at least q-dimensional so that the parameter can be identified. Furthermore, we assume that $m(\theta|z)$ is measurable in z for each $\theta \in \Theta$ and continuous in θ for every $z \in \mathbb{R}^r$. Often m is chosen to incorporate orthogonality conditions which are imposed by the model. We point out that the ML-estimator arises from (5.2) if we set

$$m(\theta|z) = \nabla_\theta \log\left(P(z|\theta)\right) \tag{5.3}$$

and compute the maximum by differentiating the Loglikelihood and setting it to 0.

In the classical *Method of Moments* we have as moment functions

$$m^{(k)}(\theta|z) = \theta \cdot z^k$$

for $k = 1, ..., q$ which defines $G_N^{(k)}$ as the k-th sample moment (with componentwise defined z^k).

Naturally, G_0 cannot be computed analytically, thus the expected value is replaced by the sample average $G_N(\theta) := \frac{1}{N}\sum_{n=1}^N m(\theta|z_n)$. Furthermore, the task of solving equation (5.2) is replaced by the minimization of the objective function

$$Q_N(\theta) := -||G_N(\theta)||^2_{W_N}$$

where $W \in \mathbb{R}^{q\times q}$ is a symmetric positive definite *weight matrix.* Asymptotically, $W_N \to W_0$ and the objective function becomes

$$Q_0(\theta) := -||G_0(\theta)||^2_{W_0} .$$

The choice of the weight matrix affects the asymptotic efficiency of the estimator. The exact form of the optimal weight matrix is known to be

$$W_0 = \mathrm{E}_z \left[m(\theta_0|z)m(\theta_0|z)^T\right]^{-1} .$$

Obviously W_0 needs also to be approximated, namely the expectation over z and the parameter θ_0. This can be done in parallel to the approximation of $Q_0(\theta_0)$ by $Q_N(\theta)$.

In general, many moment conditions arise if the data generating process for the outcome variable y is simulated by a model. For Discrete Choice models we derive the following condition: Given the observed data x_n, the choice probability $P(y = j|x_n, \theta)$ should exactly equal the expected value of the observed choice y_n:

$$G_0^{(j)}(\theta_0) = \mathrm{E}\left[y_n - P(y = j|x, \theta_0)\right] = 0 \tag{5.4}$$

for $j = 1, ..., J$. This is also the original example considered by McFadden [68] when he introduced the Method of Simulated Moments. Other moment functions are derived similarly by considering the expected value of the difference between an observed outcome and the simulation of the data generating process. If (5.4) is fulfilled, then for any $h(x)$ the function

$$m(\theta|z_n) = (y_n - P(y = j|x_n)) \otimes h(x_n)$$

gives a valid orthogonality condition and thereby defines a suitable moment condition. In particular, the GMM-estimator is efficient for the choice

$$h(x_n) = \nabla_\theta \log\left(P(z_n|\theta)\right) \cdot \nabla_\theta \log\left(P(z_n|\theta)\right)^T , \tag{5.5}$$

since this rebuilds the ML-estimator in the same fashion as in (5.3).

In [69], we find conditions for the GMM-estimator to be consistent (Theorem 2.6) and asymptotically normal (Theorem 7.2).

Lemma 5.4. *(Consistency of the GMM-estimator)*
Assume that

(i) $z_1, ..., z_N$ *are i.i.d.,*

(ii) the weight matrix converges to the optimal weight matrix: $W_N \xrightarrow{P} W_0$ *as* $N \to \infty$*,*

(iii) W_0 *is positive semi-definite and* $W_0 \mathrm{E}_z[m(\theta|z)] = 0$ *only if* $\theta = \theta_0$*,*

(iv) $\theta_0 \in \Theta$ *and* Θ *is compact,*

(v) $m(\theta|z)$ *is continuous for all* $\theta \in \Theta$ *with probability 1, and*

(vi) $\mathrm{E}_z[\sup_{\theta\in\Theta} ||m(\theta|z)||] < \infty$.

Then $\theta_N \xrightarrow{P} \theta_0$.

As for extremum estimators there are two theorems on asymptotic normality of the GMM-estimator possible, one requiring differentiability of G_0 in a neighborhood of θ_0, the other one only at θ_0. Based on Lemma 5.3 and the choice

$$D_N = (\nabla_\theta Q_0(\theta_0))^T W G_N(\theta_0)$$

for the approximate derivative of G_N at θ_0 we get:

Lemma 5.5. *(Asymptotic Normality of the GMM-estimator)*
Assume that $||G_N(\theta_N)||_W \leq \inf_{\theta\in\Theta} ||G_N(\theta)||_W + o_p(N^{-1})$, $\theta_N \xrightarrow{P} \theta_0$ *and* $W_N \to W_0$*, where* W_0 *is positive semi-definite and* $G_0(\theta)$ *satisfies the following conditions*

(i) $G_0(\theta_0) = 0$*,*

(ii) $\theta_0 \in \mathring{\Theta}$*, i.e.* θ_0 *is an interior point of* Θ*,*

(iii) G_0 *is differentiable in* θ_0 *with derivative* $D = \nabla_\theta G(\theta_0)$ *s.t.* $D^T W_0 D$ *is invertible,*

(iv) $\sqrt{N} G_N(\theta_0) \xrightarrow{d} \mathcal{N}(0, \Sigma)$*, and*

(v) for any sequence $\delta_N \to 0$*, we have*

$$\sup_{||\theta-\theta_0||\leq\delta_N} \frac{\sqrt{N}}{1+\sqrt{N}||\theta-\theta_0||} |G_N(\theta) - G_N(\theta_0) - G_0(\theta)| \xrightarrow{P} 0\,.$$

Then $\sqrt{N}(\theta_N - \theta_0) \xrightarrow{d} \mathcal{N}(0, (D^T W_0 D)^{-1} D^T W_0 \Sigma W_0 D (D^T W_0 D)^{-1})$.

5.4 Approximated Estimators

Both, M-estimators and GMM estimators, rely on functions m which either represent outcome probabilities or moments of the exogenous variables. They are defined to include all available knowledge and assumptions from the data and the model. In sections 2.3 and 2.4 we saw that the computation of the choice probability is not always analytically possible and requires numerical approximation. As the desired properties (consistency, asymptotic normality, efficiency) where defined and proven for exact functions this raises the question how approximation changes the behavior of the estimators.

An early example in this subject is the application of Gauss-Hermite quadrature to a Probit model by Butler and Moffitt [10]. However, most econometricians in the past have used simulation methods, i.e. Monte Carlo integration, for their models. Starting with *Maximum Simulated Likelihood*, numerous papers ([27], [39], [36]) examined whether the simulated ML-estimator remains consistent and asymptotically normal. McFadden [68] and Pakes and Pollard [73] introduced the *Method of Simulated Moments* in order to evaluate Multinomial Probit models. Schennach [81] provided an updated version of this approach where the specification of a distribution for the unobservables is avoided. Hajivassiliou [37] added the *Method of Simulated Scores* and compared all three simulated estimators [39]. Train [83] gave an overview of the application of simulated estimators to Discrete Choice models.

A later development was the introduction of Quasi Monte Carlo Methods to econometrics e.g. by Bhat [6] which make the expensive computation of samples obsolete and replace it with the utilization of a deterministic series of nodes.

Kristensen and Salanié [59] discussed approximated estimators in a more general context covering M-estimators as well as GMM. They consider stochastic and deterministic approximation of the objective function alike and study how bias and variance are affected and how this effect can be mitigated in the maximization step.

Finally, Griebel et al. [34] investigate the usage of more advanced quadrature methods for the approximation of the objective functions for extremum estimators. They determine ratios between sample number N and approximation accuracy $R(N)$ such that consistency and asymptotic normality of the *Maximum Approximated Likelihood* estimator are preserved.

We keep the notation general in this chapter in order to cover various approximation techniques. Later, in Chapter 6, we focus on multidimensional integrals as they appeared in the previous chapters and for moment functions in GMM-estimation. Other approximation algorithms are used if the moment function requires the solution of a dynamic programming problem (originally in [14] or later e.g. in [49] or [61]).

We assume that for both, M- and GMM-estimators, m cannot be evaluated analytically, i.e. it includes the computation of an intractable integral, differential equation or solution to a dynamic programming problem. We approximate m by m_R which applies an R-point quadrature formula or discretization of the intractable part of m.
Let $\widehat{Q}_N$ denote the approximated objective function,

$$\widehat{Q}_N(\theta) = Q_{NR} := \begin{cases} -\left|\left|\frac{1}{N}\sum_{n=1}^N m_R(\theta|z_n)\right|\right|^2_{\widehat{W}_N}, & \text{for GMM estimation,} \\ \frac{1}{N}\sum_{n=1}^N m_R(\theta|z_n), & \text{for M-estimation.} \end{cases}$$

In general, we assume that $R = R(N)$ is monotonically increasing such that $Q_{NR} \to Q_N$ for $N \to \infty$ and hence $Q_{NR} \to Q_0$. We omit the dependence from R and only write $\widehat{Q}_N$. Similarly, $\hat{\theta}_N$ denotes the maximizer of $\widehat{Q}_N$. For GMM-estimation, we further have

$$\widehat{G}_N = G_{NR} := \frac{1}{N}\sum_{n=1}^N m_R(\theta|z_n) \tag{5.6}$$

for the approximated moment function and $\widehat{W}_N$ for the approximated weight matrix, in case it also depends on m.
This raises the question under which conditions consistency and asymptotic normality can be assured for $\hat{\theta}_N$. This issue is investigated in [34] for the case of Maximum Approximated Likelihood and we adapt this approach for the Generalized Method of Approximated Moments (GMAM).
McFadden [68] and Pakes and Pollard [73] showed consistency and asymptotic normality for simulated moments already in 1989. Theorems 5.6 and 5.7 formalize those results for any approximated moment function by proving that the assumptions in Lemmas 5.2 and 5.5 are fulfilled by $\hat{\theta}_N$ and $\widehat{Q}_N$ respectively.
With Lemma 5.4 we obtained a statement about the consistency of the GMM-estimator θ_N. The following theorem extends this property to the GMAM-estimator under two additional conditions on the convergence of $\widehat{Q}_N$ to Q_N.

Theorem 5.6. *(Consistency of the GMAM-estimator)*
Assume that

(i) *Conditions (i)-(vi) from Lemma 5.4 hold,*

(ii) $\lim_{R\to\infty} \sup_{z\in\mathcal{Z},\theta\in\Theta} |(m_R(\theta|z) - m(\theta|z))||m(\theta|z)| = 0$,

(iii) $\lim_{R\to\infty} ||\widehat{W}_N - W_N|| \to 0$ *and* $\sup_{z\in\mathcal{Z},\theta\in\Theta} ||m(\theta|z)||^2 < \infty$ *or* $\widehat{W}_N = W_N$, *and*

(iv) $R(N) \to \infty$ *as* $N \to \infty$.

Then $\hat{\theta}_N$ is a consistent estimator of θ_0, i.e.

$$\hat{\theta}_N \xrightarrow{P} \theta_0 \ \text{as}\ N \to \infty\,.$$

Proof. In Lemma 5.2, conditions are given under which $\hat{\theta}_N$ is consistent. Conditions 5.2(i)-(iii) for $\hat{\theta}_N$ and $\widehat{Q}_N$ are assured by assumption 5.6(i). Hence, it only remains to prove 5.2(iv) for $\widehat{Q}_N$ instead of Q_N,

$$\sup_{\theta\in\Theta} |\widehat{Q}_N(\theta) - Q_0(\theta)| \xrightarrow{P} 0\,. \tag{5.7}$$

Using the triangle-inequality and that 5.2(iv) is fulfilled for Q_N by Lemma 5.4 it suffices to show $\sup_{\theta\in\Theta} |\widehat{Q}_N(\theta) - Q_N(\theta)| \xrightarrow{P} 0$:

$$\begin{aligned}\sup_{\theta\in\Theta}|\widehat{Q}_N(\theta) - Q_N(\theta)| &= \sup_{\theta\in\Theta}\left| \left|\left|\widehat{G}_N(\theta)\right|\right|^2_{\widehat{W}_N} - ||G_N(\theta)||^2_{W_N}\right| \\ &\le \sup_{\theta\in\Theta}\left|\left\langle \frac{1}{N}\sum_{n=1}^{N} m_R(\theta|z_n) - m(\theta|z_n),\ \frac{1}{N}\sum_{n=1}^{N} m_R(\theta|z_n) + m(\theta|z_n)\right\rangle_{\widehat{W}_N}\right| \\ &\quad + \left|G_N(\theta)(\widehat{W}_N - W_N)G_N(\theta)\right| \\ &\le \sup_{\theta\in\Theta, z\in\mathcal{Z}} ||\widehat{W}_N||\,|m_R(\theta|z) - m(\theta|z)|\,|m_R(\theta|z) + m(\theta|z)| \\ &\quad + \sup_{\theta\in\Theta, z\in\mathcal{Z}} |m(\theta|z)|^2 ||\widehat{W}_N - W_N||\,.\end{aligned}$$

Assumption (iv) leads to $m_R(\theta|z) \to m(\theta|z)$ for any $z \in \mathcal{Z}$ and $N \to \infty$ (and implicitly $R \to \infty$). Hence, $|m_R(\theta|z) + m(\theta|z)|$ is asymptotically bounded by $C\,|m(\theta|z)|$ and assumption (ii) for the first and (iii) for the second term finish the proof of (5.7) for $N \to \infty$. □

Next, we show asymptotic normality of $\hat{\theta}_N$ by showing that the conditions of Lemma 5.5 hold for the approximated moment function $\widehat{G}_N$.

Theorem 5.7. *(Asymptotic Normality of the GMAM-estimator)*
Assume that

(i) the assumptions of Lemma 5.5 hold for the infeasible estimator θ_N,

(ii) the assumptions of Theorem 5.6 hold,

(iii) $||\widehat{G}_N(\hat{\theta}_N)||_{\widehat{W}_N} \le \inf_{\theta\in\Theta} ||\widehat{G}_N(\theta)||_W + o_p(N^{-1})$,

(iv) $\sup_{z\in\mathcal{Z}} \sqrt{N}|m_R(\theta_0|z) - m(\theta_0|z)| \to 0$,

(v) for any sequence $\delta_N \to 0$, we have

$$\sup_{z\in\mathcal{Z}, ||\theta-\theta_0||\le\delta_N} \frac{\sqrt{N}}{1+\sqrt{N}||\theta-\theta_0||}|m_R(\theta|z) - m(\theta|z)| \to 0.$$

Then $\sqrt{N}(\hat{\theta}_N - \theta_0) \xrightarrow{d} \mathcal{N}(0, (D^T W_0 D)^{-1} D^T W_0 \Sigma W_0 D (D^T W_0 D)^{-1})$.

Proof. By assumptions (i) and (ii) the conditions 5.5(i)-(iii) as well as consistency of $\hat{\theta}_N$ and $\widehat{W}_N \to W_0$ are fulfilled. Similarly, (iii) states that $\hat{\theta}_N$ maximizes $\widehat{Q}_N$ approximately. Thus, it remains to show 5.5(iv) and (v). Firstly, we have

$$\begin{aligned}\sqrt{N}\widehat{G}_N(\theta_0) &= \sqrt{N}G_N(\theta_0) + \sqrt{N}(\widehat{G}_N(\theta_0) - G_N(\theta_0)) \\ &\leq \sqrt{N}G_N(\theta_0) + \sup_{z\in\mathcal{Z}} \sqrt{N}|m_R(\theta_0|z) - m(\theta_0|z)| \xrightarrow{d} \mathcal{N}(0,\Sigma)\end{aligned}$$

by (i) (the same statement for G_N) and (iv).
Again by the triangle inequality, we get for any sequence $\delta_N \to 0$ and $C_N = \frac{\sqrt{N}}{1+\sqrt{N}||\theta-\theta_0||}$:

$$\begin{aligned}&\sup_{||\theta-\theta_0||\leq\delta_N} C_N||\widehat{G}_N(\theta) - \widehat{G}_N(\theta_0) - G_0(\theta)|| \\ &\qquad\leq \sup_{||\theta-\theta_0||\leq\delta_N} C_N||G_N(\theta) - G_N(\theta_0) - G_0(\theta)|| \\ &\qquad\quad + C_N||\widehat{G}_N(\theta) - G_N(\theta)|| + C_N||\widehat{G}_N(\theta_0) - G_N(\theta_0)|| \\ &\qquad\leq \sup_{||\theta-\theta_0||\leq\delta_N} C_N||G_N(\theta) - G_N(\theta_0) - G_0(\theta)|| \\ &\qquad\quad + 2 \sup_{z\in\mathcal{Z}, ||\theta-\theta_0||\leq\delta_N} C_N|m_R(\theta|z) - m(\theta|z)|\,.\end{aligned}$$

Then, the first term converges in probability to 0 due to (i) and so does the second due to (v). □

With consistency and asymptotic normality proven for $\hat{\theta}_N$ we can safely estimate approximated choice probabilities and moment functions. In particular, we can apply deterministic quadrature as analyzed in part I.
Now, the notion of integrated M- and GMM-estimators raises the question whether advanced quadrature can be used for the objective function itself. Since the integration domain would be the real world data in this case, deterministic quadrature nodes cannot be chosen and we have to settle for MC (or possibly optimal weights) quadrature. However, understanding the objective function as an integral and having an additional approximated integral as moment function allows us to utilize an SG approach for approximated estimators.

6 Multilevel Estimation

6.1 Objective

In Chapter 4, we extended the range of applicable quadrature rules in econometrics for Generalized Linear Mixed and Dynamic Economic models. In both cases, choice probabilities were an essential part of the model and were defined as multidimensional integrals.
In the previous chapter, we observed that common objective functions of extremum estimators theoretically denote multidimensional integrals which are approximated with real-world or population samples as quadrature nodes. Furthermore, we found that approximation of choice probabilities and moment functions does not impair desirable properties of extremum estimators like consistency and asymptotic normality.
Combining simulation of the objective function and integration of a choice probability we obtain an approximation problem on a product domain $\Omega_1 \times \Omega_2$ where Ω_1 is the population space and Ω_2 is the integration domain for the choice probability. Harbrecht and Griebel [30] investigated interpolation in such tensor product spaces and constructed a *sparse tensor product space* based on the sparse grid method. Sparse tensor product spaces are used e.g. in [31] where elliptic PDEs are solved with quadrature on Ω_1 and interpolation on Ω_2. Heinrich [46] and Giles [26] proposed a similar approach called *Multilevel Monte Carlo* method solely for Monte Carlo simulations of stochastic and deterministic PDEs and intractable integrals.
Applied to the estimation of an econometric model, we consider numerical quadrature on both Ω_1 and Ω_2. However, most objective functions have an intermediate function "between" the two integration steps, e.g. for integrated Loglikelihood

$$Q_0(\theta) = \int_{\mathcal{Z}} \log(P(z|\theta)) d\nu(z) \tag{6.1}$$

and $P(z|\theta)$ is given by another integral. We show that the theorems on convergence rates from [30] hold similarly for quadrature and provide conditions on coupling functions F such that these rates are preserved.

6.2 General Setup

The asymptotic GMM-estimator is given by $Q_0(\theta) = ||G_0(\theta)||^2_{W_0}$ and similar to (6.1) we assume that G_0 has the general form

$$G_0(\theta) = \int_{\mathcal{Z}} F\left(z, \theta, \int_{\mathcal{U}} \varphi(u, z|\theta) d\mu(u)\right) d\nu(z) \,. \tag{6.2}$$

The function $F : \mathcal{Z} \times \Theta \times \mathbb{R} \to \mathbb{R}$ is defined by the chosen moments m and the variable u often represents unobservable variables or errors in measurement.

Since GMM-estimators are fairly general and also include Maximum Likelihood estimation, our subject of interest is the objective function described in (6.2). Results for (6.1) can simply be derived with $F(z,\theta,t) = \log(t)$. In the following, we assume that

(I) φ is μ-integrable in u for all $\theta \in \Theta$ and ν-almost all z and

(II) F is ν-integrable in z for all $\theta \in \Theta$.

We omit the dependence on θ as the integral is computed separately for each θ and write (6.2) more generally as

$$\mathcal{I}_1(F_\varphi) := \int_{\Omega_1} F_\varphi(z) d\nu(z) \tag{6.3}$$

$$\mathcal{I}_2(\varphi(z)) := \int_{\Omega_2} \varphi(z)(u) d\mu(u) \tag{6.4}$$

for functions $F_\varphi \in \mathcal{H}_1(\Omega_1;\nu)$ and $\varphi \in \mathcal{H}_1(\Omega_1, \mathcal{H}_2(\Omega_2;\mu);\nu)$ and domains $\Omega_i \subset \mathbb{R}^{d_i}$. We write F_φ to indicate that we consider functions F which always include the computation of the integral $\mathcal{I}_2$ but might also depend on z in a direct way. We express this dependence via

$$F_\varphi(z) = F(z, \mathcal{I}_2(\varphi(z)))$$

for a function $F \in \mathcal{H}_1(\Omega_1 \times \mathbb{R}, \nu)$.

The spaces $\mathcal{H}_1$ and $\mathcal{H}_2$ are Banach spaces of functions. In order to apply quadrature rules like SG or QMC which require some regularity of the integrand we assume them to be Sobolev spaces of mixed regularity $\mathcal{H}_i = H^{r_i}_{\text{mix}}$. Here, $\mathcal{H}_1$ constitutes Bochner space as its target space is another Banach space. As integration and quadrature are executed separately due to the intermediate function F and φ is assumed to have sufficient regularity in both arguments, this notion is unproblematic for the following results.

We can view (6.3) and (6.4) for fixed F as single integration problem of an integrand in

$$\mathcal{H}^F := \left\{ \varphi \in \mathcal{H}_1(\Omega_1, \mathcal{H}_2(\Omega_2;\mu);\nu) \text{ s.t. } F\left(z, \int_{\Omega_2} \varphi(z)(u) d\mu(u)\right) \in \mathcal{H}_1(\Omega_1,\nu) \right\}.$$

For $F_\varphi(z) = F(|\mathcal{I}_2|)$ this definition resembles the so called *Orlicz-Bochner space* $L^F(\Omega, X)$. We want to shortly mention its definition [76] in order to compare properties:

Definition 6.1. *(Orlicz-Bochner space)*
Let (Ω,Σ,ν) be a measure space, $(X, ||\cdot||)$ a Banach space and F a Young function, i.e. $F : \mathbb{R} \rightarrow [0,\infty]$ is convex and lower semi-continuous and

$$\lim_{s\to\infty} \frac{F(s)}{s} = \infty\,,$$

$$\lim_{s\to 0} \frac{F(s)}{s} = 0\,.$$

Then the Orlicz-Bochner space is defined as

$$L^F(\Omega, X) := \left\{ \varphi : \Omega \to X \text{ measurable and } \exists \alpha > 0 : \int_\Omega F\big(\alpha ||u(z)||_X\big) d\mu(z) < \infty \right\} .$$

Furthermore, the set

$$M^F(\Omega, X) := \left\{ \varphi : \Omega \to X \text{ measurable and } \forall \alpha > 0 : \int_\Omega F\big(\alpha ||u(z)||_X\big) d\mu(z) < \infty \right\}$$

is a closed subspace of L^F and sometimes called small Orlicz-Bochner space.

This notion can be further extended to include weakly differentiable functions φ [44]. Setting $\Omega = \Omega_1$ and $X = \mathcal{H}_2(\Omega_2, \mu)$ we get $H^F \subset L^F$. However, the assumption that F is a Young function and hence convex is too strong for the general form (6.2). For example, the case $F(z,t) = \log(z)$ does not satisfy these conditions rendering any further investigation of Orlicz-Bochner spaces inadequate as Maximum Likelihood estimation is the most relevant application of our theoretical results. We therefore proceed by examining F_φ and φ in separate spaces.

6.3 Sparse Tensor Product Quadrature

The goal of this section is to approximate the integrals $\mathcal{I}_1$ and $\mathcal{I}_2$ from (6.3) and (6.4) and make use of the nested structure of the integration problem. Harbrecht and Griebel consider the case $F(z,t) = t$, i.e. $\varphi \in \mathcal{H}(\Omega_1 \times \Omega_2)$ without any intermediate function F, and prove results on the convergence of interpolation methods on the tensor product space $\Omega_1 \times \Omega_2$.

There, as well as in our problem, approximation is initially considered on each domain Ω_i , $i = 1, 2$, separately. In particular, different regularity assumptions might hold on each Ω_i so that the quadrature formulas Q^i have different rates of convergence. Then, the classical approach is to balance Q^1 and Q^2 such that equal convergence on both domains is achieved. If one converged asymptotically faster than the other, the smaller error would be outweighed by the other asymptotically, so additional cost for the quicker convergence would be wasted. A product rule similar to the one in Section 3.2 balances both quadratures and leads to a cost-efficient and optimal convergence, as both quadrature formulas contribute an error of the same order.

Likewise, the SG method can also be transferred to the product domain $\Omega_1 \times \Omega_2$: Instead of using the full node set of Q^2 for each node of Q^1, we build a sparse grid of nodes on $\Omega_1 \times \Omega_2$ in the same fashion as we did for nodes on $\mathbb{R}^d = \mathbb{R} \times \cdots \times \mathbb{R}$, the d-fold tensor product of $\mathbb{R}$. Harbrecht et al. ([30], [31]) call this construction the sparse tensor product (STP) in contrast to the (full) tensor product (FTP). Alternatively, they speak of *Multilevel quadrature* which underscores the use of different levels of accuracy for the

quadrature nodes in the outer loop.
They prove that STP/Multilevel quadrature preserves the rates from Q^1 and Q^2 except for a log-factor in a similar fashion as the original sparse grid method does. Unfortunately, we cannot directly utilize their results due to the intermediate function F but generalize them by imposing an additional condition on F.

In the following, we assume that the Q^i , $i = 1, 2$, have error bounds of the form

$$e(Q^i, H^{r_i}_{\text{mix}}) = O\left(N^{-s_i} \log(N)^{t_i}\right) \tag{6.5}$$

for some $0 \leq s_i \leq r_i$. The log-exponent takes the general form $t_i = t_i(s_i, d_i)$ so to include (Q)MC, SG, Frolov and product rule quadrature. We assume that Q^i_l is a rule with

$$N_{il} = O(2^l)$$

nodes if it is an MC or QMC rule or with

$$N_{il} = O(2^l l^{(d-1)(r+1)/2})$$

nodes if it is an SG rule. Thus, N_{il} behaves asymptotically like $O(2^{l(1+\varepsilon)})$ where $\varepsilon = 0$ for MC and QMC rules and $\varepsilon > 0$ is arbitrarily small for SG rules. We will usually just write N_l whenever the number of nodes of Q^1 or Q^2 is meant.
Now difference quadrature formulas are defined similar to (3.4) as

$$\Delta^1_l(F_\varphi) := \begin{cases} Q^1_l(F_\varphi) - Q^1_{l-1}(F_\varphi)\,, & \text{for } l \geq 2\,, \\ Q^1_1(F_\varphi)\,, & \text{for } l = 1\,. \end{cases}$$

$$\Delta^2_l(z, \varphi(z)) := \begin{cases} F(z, Q^2_l\, \varphi(z)) - F(z, Q^2_{l-1}\, \varphi(z))\,, & \text{for } l \geq 2\,, \\ F(z, Q^2_1\, \varphi(z))\,, & \text{for } l = 1\,. \end{cases}$$

This allows for telescopic expansions of Q^1 and $F(z, Q^2(z))$ for any $z \in \mathcal{Z}$. In particular, we can sum up Δ^i over $l \in \mathbb{N}$ and get representations of $\mathcal{I}_1$ and $F_\varphi(z)$ resulting in

$$\begin{aligned} \mathcal{I}_1(F_\varphi) &= \sum_{j_1=1}^{\infty} \Delta^1_{j_1}(F_\varphi) = \sum_{j_1=1}^{\infty} \Delta^1_{j_1} \left(\sum_{j_2=1}^{\infty} \Delta^2_{j_2}(\cdot, \varphi(\cdot)) \right) \\ &= \sum_{(j_1, j_2) \in \mathbb{N}^2} \Delta^1_{j_1} \otimes \Delta^2_{j_2}(\cdot, \varphi(\cdot))\,. \end{aligned} \tag{6.6}$$

For a general level set $\mathcal{A} \subset \mathbb{N}^2$, we obtain the general sparse grid quadrature rule on $\Omega_1 \times \Omega_2$ by truncating the above sum

$$Q_{\mathcal{A}}(F_\varphi) := \sum_{(j_1, j_2) \in \mathcal{A}} \Delta^1_{j_1} \otimes \Delta^2_{j_2}(\cdot, \varphi(\cdot))\,.$$

For our considerations, we use the basic anisotropic SG index set $\mathcal{A}_1^\sigma(J) := \{(j_1, j_2) \ : \ \sigma j_1 + \frac{j_2}{\sigma} \leq J\}$ and compare it to the full grid set $\mathcal{A}_\infty^\sigma(J) := \{(j_1, j_2) \ : \ \max\{\sigma j_1, \frac{j_2}{\sigma}\} \leq J\}$. The parameter $\sigma > 0$ accounts for different convergence rates of inner and outer quadrature and "balances" them.
We write $Q_{\infty J}^\sigma$ for the level-J-FTP rule and Q_{1J}^σ for the level-J-STP rule and define the corresponding errors as

$$E_{\infty J}^\sigma(F, \varphi) := |\mathcal{I}_1(F_\varphi) - Q_{\infty J}^\sigma(F_\varphi)|\,, \tag{6.7}$$

$$E_{1J}^\sigma(F, \varphi) := |\mathcal{I}_1(F_\varphi) - Q_{1J}^\sigma(F_\varphi)|\,. \tag{6.8}$$

Then, $e_{\infty J}^\sigma$ and e_{1J}^σ are the corresponding worst case errors and we omit σ in most cases later on.
First, we count the number of nodes in $Q_{\infty J}$ and Q_{1J}. Since F does not affect the number of nodes $N_{\infty J}$ and N_{1J} we can adapt Theorem 4.1 from [30] and adjust it for the additional case where Q^1 might be an SG rule.

Theorem 6.2. *(Size of full and sparse tensor product quadrature)*
For quadrature formulas Q^i, $i = 1, 2$, as above the full tensor product rule has $N_{\infty J}$ nodes and

$$N_{\infty J} = O\left(2^{J(\sigma + 1/\sigma)(1+\varepsilon)}\right)\,.$$

The sparse tensor product rule has N_{1J} nodes and

$$N_{1J} = \begin{cases} O\left(2^{J\max\{1/\sigma, \sigma\}(1+\varepsilon)}\right) & \text{for } \sigma \neq 1\,, \\ O\left(J 2^{J(1+\varepsilon)}\right) & \text{for } \sigma = 1\,. \end{cases}$$

Here, $\varepsilon = 0$ if neither Q^1 nor Q^2 are SG rules and $\varepsilon > 0$ arbitrarily small otherwise.

Proof. The estimate for $N_{\infty J}$ follows directly from expanding $Q_{\infty J}$:

$$Q_{\infty J}(F_\varphi) = \sum_{\sigma j_1, \frac{j_2}{\sigma} \leq J} \Delta_{j_1}^1 \otimes \Delta_{j_2}^2\,(\cdot, \varphi(\cdot)) = Q_{J/\sigma}^1 \otimes F\left(\cdot, Q_{\sigma J}^2(\varphi(\cdot))\right)\,. \tag{6.9}$$

Then we apply $N_{il} = O(2^{l(1+\varepsilon)})$ for each Q^i and get

$$N_{\infty J} = O\left(2^{\frac{J}{\sigma}(1+\varepsilon)} 2^{\sigma J(1+\varepsilon)}\right) = O\left(2^{J(\sigma + \frac{1}{\sigma})(1+\varepsilon)}\right)\,.$$

For Q_{1J}, we consider each term $\Delta_{j_1}^1 \otimes \Delta_{j_2}^2$ and see that it has $2^{(j_1+j_2)(1+\varepsilon)}$ nodes. We sum over the index set $\mathcal{A}_1^\sigma$

$$N_{1J} = \sum_{\sigma j_1 + \frac{j_2}{\sigma} \leq J} O\left(2^{(j_1+j_2)(1+\varepsilon)}\right) = \sum_{j_1=0}^{J/\sigma} \sum_{j_2=0}^{M} O\left(2^{(j_1+j_2)(1+\varepsilon)}\right)$$

$$= O\left(2^{\sigma J(1+\varepsilon)} \sum_{j_1=0}^{J/\sigma} 2^{j_1(1-\sigma^2)(1+\varepsilon)}\right)$$

where $M = J\sigma - j_1\sigma^2$. For $\sigma > 1$, this implies $N_{1J} = O(2^{\sigma J(1+\varepsilon)})$ and for $\sigma < 1$ we get $N_{1J} = O(2^{(1+\varepsilon)J/\sigma})$. Only for $\sigma = 1$ we have $O(J2^{J(1+\varepsilon)})$. □

In the case $\sigma = 1$ this is exactly the size of the basic SG quadrature formula in two dimensions and we can substitute the log-term J by an additional factor $(1 + \varepsilon)$ in the exponent. We observe that the reduction from FTP to STP is most substantial if σ is close to 1, since the factor $\sigma + 1/\sigma$ is reduced to $\max\{\sigma, 1/\sigma\}$.
To prove bounds for $e_{\infty J}$ and e_{1J} similar to those in [30] but generalized w.r.t. the intermediate function F we need the notion of *Hölder continuity.*

Definition 6.3. *(Hölder continuity)*
A function $f : \Omega \subset \mathbb{R}^d \to \mathbb{R}$ is called Hölder continuous if there exist $\alpha, C > 0$ s.t.

$$|f(x) - f(y)| \leq C||x - y||^\alpha$$

for all $x, y \in \Omega$.

Theorem 6.4. *(Error bound for full tensor product quadrature)*
Let $0 < s_i \leq r_i$ for $i = 1, 2$ and Q^i be quadrature formulas on Ω_i with error bounds as in (6.5). Suppose $F \in H^{r_1}_{mix}(\Omega \times \mathbb{R}; \nu)$ and $F(z, \cdot)$ is Hölder continuous with exponent α for any $z \in \mathcal{Z}$ and $\varphi \in H^{r_1}_{mix}(\Omega_1, H^{r_2}_{mix}(\Omega_2; \mu); \nu)$. Then the error of the tensor product quadrature rule is given by

$$e_{\infty J} = O\left((J\sigma)^{\alpha t_2} \left(\frac{J}{\sigma}\right)^{t_1} 2^{-J\min(s_1/\sigma, \sigma\alpha s_2)(1+\varepsilon)} \right).$$

Proof. We reuse the expansion (6.9) and omit for simplicity the dependence on z in the following. With the triangle inequality we get

$$\begin{aligned}
e_{\infty J}(F_\varphi) &= |\mathcal{I}_1(F_\varphi) - Q_{\infty J}(F_\varphi)| \\
&= \left|\mathcal{I}_1 \otimes F\left(\mathcal{I}_2(\varphi)\right) - Q^1_{J/\sigma} \otimes F\left(Q^2_{\sigma J}(\varphi)\right)\right| \\
&\leq \left|\left(\mathcal{I}_1 - Q^1_{J/\sigma}\right) \otimes F\left(\mathcal{I}_2(\varphi)\right)\right| + \left|Q^1_{J/\sigma} \otimes \left(F\left(\mathcal{I}_2(\varphi)\right) - F\left(Q^2_{\sigma J}(\varphi)\right)\right)\right| \\
&\leq O(N^{-s_1}_{J/\sigma} \log(N_{J/\sigma})^{t_1}) + ||Q^1_{J/\sigma}|| \left|F\left(\mathcal{I}_2(\varphi)\right) - F\left(Q^2_{\sigma J}(\varphi)\right)\right| \\
&\leq O(N^{-s_1}_{J/\sigma} \log(N_{J/\sigma})^{t_1}) + C\left|\mathcal{I}_2(\varphi) - Q^2_{\sigma J}(\varphi)\right|^\alpha \\
&= O(N^{-s_1}_{J/\sigma} \log(N_{J/\sigma})^{t_1}) + O(N^{-\alpha s_2}_{\sigma J} \log(N_{\sigma J})^{\alpha t_2}) \\
&= O\left((J\sigma)^{\alpha t_2} \left(\frac{J}{\sigma}\right)^{t_1} 2^{-J\min(s_1/\sigma, \alpha s_2 \sigma)(1+\varepsilon)} \right).
\end{aligned}$$

The first term measures the approximation accuracy of Q^1. Since $F, \varphi \in H^{r_1}_{\text{mix}}$ and $s_2 \leq r_2$ it obtains the rate (6.5). For the second term, we use Hölder continuity, boundedness of the linear operator $Q^1_{J/\sigma}$ and that $\varphi(z) \in H^{r_2}_{\text{mix}}$ for all z, so Q^2 also obtains its rate (6.5). □

The following extension of Theorem 4.3 in [30] shows that STP quadrature gives a similar result.

Theorem 6.5. *(Error bound for sparse tensor product quadrature)*
Let $0 < s_i \leq r_i$, Q^i for $i = 1, 2$, F and φ be as in Theorem 6.4. Then the error of the sparse tensor product quadrature is given by

$$e_{1J} = \begin{cases} O\left((J\sigma)^{\alpha t_2} \left(\frac{J}{\sigma}\right)^{t_1} 2^{-J\min\{s_1/\sigma, \alpha s_2\sigma\}(1+\varepsilon)}\right) & \textit{for } \frac{s_1}{\sigma} \neq \alpha s_2 \sigma\,, \\ \\ O\left((J\sigma)^{\alpha t_2} \left(\frac{J}{\sigma}\right)^{t_1+1} 2^{-J\alpha s_2\sigma(1+\varepsilon)}\right) & \textit{for } \frac{s_1}{\sigma} = \alpha s_2 \sigma\,. \end{cases}$$

Proof. Using the triangle inequality we get a bound for Δ^1 for functions $f \in H^{r_1}_{\text{mix}}$,

$$\begin{aligned} ||\Delta^1_l|| &= \max_{f\in H^{r_1}_{\text{mix}}, ||f||\leq 1} \left|\left|\Delta^1_l(f)\right|\right| \\ &\leq \max_{f\in H^{r_1}_{\text{mix}}, ||f||\leq 1} \left|\left|Q^1_l(f) - \mathcal{I}_1(f)\right|\right| + \left|\left|\mathcal{I}_1(f) - Q^1_{l-1}(f)\right|\right| \\ &= O(N_l^{-s_1} \log(N_l)^{t_1})\,. \end{aligned}$$

Expanding $\mathcal{I}_1(F_\varphi)$ according to (6.6) we get

$$\begin{aligned} e_{1J}(F_\varphi) &= \left| \mathcal{I}_1(F_\varphi) - \sum_{\sigma j_1 + \frac{j_2}{\sigma} \leq J} \Delta^1_{j_1} \otimes \Delta^2_{j_2}(\varphi) \right| \\ &\leq \sum_{\sigma j_1 + \frac{j_2}{\sigma} > J} \left|\Delta^1_{j_1} \otimes \Delta^2_{j_2}(\varphi)\right| \\ &\leq \sum_{\sigma j_1 + \frac{j_2}{\sigma} > J} ||\Delta^1_{j_1}||\, ||\Delta^2_{j_2}(\varphi)|| \\ &\leq O\left(\sum_{\sigma j_1 + \frac{j_2}{\sigma} > J} 2^{-(s_1 j_1 + \alpha s_2 j_2)(1+\varepsilon)} j_1^{t_1} j_2^{\alpha t_2} \right). \end{aligned}$$

We split the index set $\left\{(j_1, j_2) : \sigma j_1 + \frac{j_2}{\sigma} > J\right\}$ into two disjoint sets

$$\begin{aligned} I_1 &:= \left\{(j_1, j_2) : 0 \leq j_1 \leq \frac{J}{\sigma}, M < j_2\right\}, \\ I_2 &:= \left\{(j_1, j_2) : \frac{J}{\sigma} < j_1, 0 \leq j_2\right\} \end{aligned}$$

where again $M = J\sigma - j_1\sigma^2$ and sum over each index set separately:

$$\begin{aligned}
\sum_{(j_1,j_2)\in I_1} & 2^{-(s_1j_1+\alpha s_2 j_2)(1+\varepsilon)} j_1^{t_1} j_2^{\alpha t_2} \\
&= \sum_{j_1=0}^{J/\sigma} j_1^{t_1} 2^{-s_1j_1(1+\varepsilon)} \sum_{j_2=M+1}^{\infty} j_2^{\alpha t_2} 2^{-\alpha s_2 j_2(1+\varepsilon)} \\
&\le \sum_{j_1=0}^{J/\sigma} j_1^{t_1} (M+1)^{\alpha t_2} 2^{-(s_1j_1+\alpha s_2 M)(1+\varepsilon)} \sum_{j_2=1}^{\infty} j_2^{\alpha t_2} 2^{-\alpha s_2 j_2(1+\varepsilon)} \\
&\le (J\sigma)^{\alpha t_2} \left(\frac{J}{\sigma}\right)^{t_1} \mathrm{Li}_{-\alpha t_2}(2^{-\alpha s_2(1+\varepsilon)}) \sum_{j_1=0}^{J/\sigma} 2^{-(s_1j_1+\alpha s_2 M)(1+\varepsilon)} \\
&= C\ (J\sigma)^{\alpha t_2} \left(\frac{J}{\sigma}\right)^{t_1} 2^{-\alpha s_2 J\sigma(1+\varepsilon)} \sum_{j_1=0}^{J/\sigma} 2^{-j_1\sigma(s_1/\sigma-\alpha s_2\sigma)(1+\varepsilon)}
\end{aligned}$$

where Li denotes the *Polylogarithm*. The constant C depends on s_2 and d_2 but this was already the case for the constants in the error estimates for Q^1 and Q^2. In the same fashion we get

$$\begin{aligned}
\sum_{(j_1,j_2)\in I_2} 2^{-(s_1j_1+\alpha s_2 j_2)(1+\varepsilon)} &= C \left(\frac{J}{\sigma}\right)^{t_1} 2^{-s_1 J(1+\varepsilon)/\sigma} \\
&= C \left(\frac{J}{\sigma}\right)^{t_1} 2^{-\alpha s_2 J\sigma(1+\varepsilon)}\ 2^{-J(s_1/\sigma-\alpha s_2\sigma)(1+\varepsilon)}
\end{aligned}$$

with C including two Polylogarithms (one for each summation over j_1 or j_2). Joining both sums we distinguish three cases. For $\frac{s_1}{\sigma} < \alpha s_2\sigma$,

$$\begin{aligned}
\sum_{\sigma j_1+\frac{j_2}{\sigma}>J} & 2^{-(s_1j_1+\alpha s_2 j_2)(1+\varepsilon)} j_1^{t_1} j_2^{\alpha t_2} \\
&= O\left(2^{-\alpha s_2 J\sigma(1+\varepsilon)} \left(\frac{J}{\sigma}\right)^{t_1} \left((J\sigma)^{\alpha t_2} \sum_{j_1=0}^{J/\sigma} 2^{-j_1\sigma(s_1/\sigma-\alpha s_2\sigma)(1+\varepsilon)} + 2^{-J(s_1/\sigma-\alpha s_2\sigma)(1+\varepsilon)}\right)\right) \\
&= O\left(2^{-\alpha s_2 J\sigma(1+\varepsilon)} \left(\frac{J}{\sigma}\right)^{t_1} \left((J\sigma)^{\alpha t_2} 2^{-J(s_1/\sigma-\alpha s_2\sigma)(1+\varepsilon)} + 2^{-J(s_1/\sigma-\alpha s_2\sigma)(1+\varepsilon)}\right)\right) \\
&= O\left((J\sigma)^{\alpha t_2} \left(\frac{J}{\sigma}\right)^{t_1} 2^{-\frac{s_1}{\sigma}J(1+\varepsilon)}\right).
\end{aligned}$$

For $\frac{s_1}{\sigma} > \alpha s_2 \sigma$ we can bound the expression by

$$\begin{aligned}\sum_{\sigma j_1 + \frac{j_2}{\sigma} > J} 2^{-(s_1 j_1 + \alpha s_2 j_2)(1+\varepsilon)} j_1^{t_1} j_2^{\alpha t_2} &= O\left(2^{-\alpha s_2 J \sigma (1+\varepsilon)} \left(\frac{J}{\sigma}\right)^{t_1} \left((J\sigma)^{\alpha t_2} + 1\right)\right) \\ &= O\left((J\sigma)^{\alpha t_2} \left(\frac{J}{\sigma}\right)^{t_1} 2^{-\alpha s_2 J \sigma (1+\varepsilon)}\right) .\end{aligned}$$

Finally, for $\frac{s_1}{\sigma} = \alpha s_2 \sigma$,

$$\begin{aligned}\sum_{\sigma j_1 + \frac{j_2}{\sigma} > J} 2^{-(s_1 j_1 + \alpha s_2 j_2)(1+\varepsilon)} j_1^{t_1} j_2^{\alpha t_2 (1+\varepsilon)} &= O\left(2^{-\alpha s_2 J \sigma (1+\varepsilon)} \left(\frac{J}{\sigma}\right)^{t_1} \left((J\sigma)^{\alpha t_2} \sum_{j_1=0}^{J/\sigma} 1 + 1\right)\right) \\ &= O\left((J\sigma)^{\alpha t_2} \left(\frac{J}{\sigma}\right)^{t_1+1} 2^{-J\alpha s_2 \sigma (1+\varepsilon)}\right) .\end{aligned}$$

This concludes the proof. □

Theorems 6.4 and 6.5 can also be stated for probabilistic error rates of the outer quadrature, e.g. from MC integration. Then, the mean squared error is used instead of a norm and the proof proceeds similar to the derivation of the mean squared error of MC integration.
For both, $Q_{\infty J}$ and Q_{1J}, we can combine Theorems 6.2 and 6.4 or 6.5 respectively to obtain an error bound in terms of the size of the corresponding quadrature formula.

Corollary 6.6. *Let $0 < s_i \leq r_i$, Q^i for $i = 1, 2$, F and φ as in Theorem 6.4 and set*

$$\begin{aligned}\gamma_\infty &:= \frac{\min\{s_1/\sigma, \alpha s_2 \sigma\}}{\sigma + 1/\sigma} , \\ \gamma_1 &:= \frac{\min\{s_1/\sigma, \alpha s_2 \sigma\}}{\max\{\sigma, 1/\sigma\}} .\end{aligned}$$

For $N = N_{\infty J}$ the FTP error is

$$e_{\infty J} = O\left(\log(N)^{t_1 + \alpha t_2} N^{-\gamma_\infty}\right) .$$

For STP quadrature we separate four cases for $N = N_{1J}$ according to Theorem 6.2. If $\sigma \neq 1$ we have

$$e_{1J} = \begin{cases} O\left((\log N)^{t_1 + \alpha t_2} N^{-\gamma_1}\right) & \textit{for } \frac{s_1}{\sigma} \neq \alpha s_2 \sigma , \\ \\ O\left((\log N)^{t_1 + \alpha t_2 + 1} N^{-\gamma_1}\right) & \textit{for } \frac{s_1}{\sigma} = \alpha s_2 \sigma , \end{cases}$$

and if $\sigma = 1$

$$e_{1J} = \begin{cases} O\left((\log N)^{t_1+\alpha t_2+\gamma_1} N^{-\gamma_1}\right) & \textit{for } \frac{s_1}{\sigma} \neq \alpha s_2 \sigma\,, \\ \\ O\left((\log N)^{t_1+\alpha t_2+\gamma_1+1} N^{-\gamma_1}\right) & \textit{for } \frac{s_1}{\sigma} = \alpha s_2 \sigma\,. \end{cases}$$

Proof. All results can be obtained by plugging in the corresponding results from Theorems 6.2, 6.4 and 6.5. □

Finally, we identify the optimal σ to balance error bounds of Q^1 and Q^2 and get an optimal joint convergence rate.

Theorem 6.7. *(Optimal* σ *for full and sparse tensor product quadrature) Both,* $Q_{\infty J}$ *and* Q_{1J}*, achieve their best error bound for*

$$\sigma^* = \sqrt{\frac{s_1}{\alpha s_2}}\,.$$

If $\kappa = \frac{s_1}{\alpha s_2} \neq 1$*, then any* σ *with* $\sigma^2 \in [1, \kappa]$ *or* $\sigma^2 \in [\kappa, 1]$ *respectively is optimal for* Q_{1J}*. The optimal exponents are then*

$$\gamma_\infty^* = \frac{\alpha s_1 s_2}{s_1 + \alpha s_2}\,, \tag{6.10}$$

$$\gamma_1^* = \min\{s_1, \alpha s_2\}\,. \tag{6.11}$$

Proof. In order to achieve optimal bounds, we have to maximize γ_∞ and γ_1. The former is maximized if $s_1/\sigma = \alpha s_2 \sigma$ and the same holds for γ_1 if $s_1 = \alpha s_2$. If $s_1 \neq \alpha s_2$, we have

$$\gamma_1^* := \max_{\sigma>0} \gamma_1 = \max_{\sigma>0} \left(\alpha s_2 \min\{\kappa, \sigma^2\} \min\{1, \frac{1}{\sigma^2}\}\right)\,. \tag{6.12}$$

For $\kappa < 1$, i.e. $s_1 < \alpha s_2$, we distinguish the cases (I) $\sigma^2 < \kappa$, (II) $\kappa \leq \sigma^2 \leq 1$ and (III) $\sigma^2 > 1$ and have $\gamma_1^* = s_1$ for (II) and $\gamma_1^* < s_1$ for (I) and (III). Similar cases result from $\kappa > 1$ with γ_1^* maximal for $1 \leq \sigma^2 \leq \kappa$. □

6.4 Numerical Results

This section is devoted to the validation of the previously obtained results on the convergence order of FTP and STP quadrature. We present numerical results for a synthetic test function and for two exemplary integrands from Maximum Likelihood and GMM estimation. Then, the latter can be called *Multilevel estimators* as the approximation of the inner integral is executed with different levels of accuracy for each data sample. While the true value of the integral is available for the test function we use the same estimate for the error for the other two integrals as in Chapter 4.
The Tables 4 and 5 give an overview of the expected convergence rates for

		$Q^1 =$		
		MC	QMC	SG/Frolov
$Q^2 =$	MC: $O(N^{-1/2})$	$O(N^{-1/4})$	$O(N^{-1/3}\log(N)^{d_1})$	$O(N^{-\frac{r}{2r+1}}\log(N)^{t_1})$
	QMC: $O(N^{-1}\log(N)^{d})$	$O(N^{-1/3}\log(N)^{d_2})$	$O(N^{-1/2}\log(N)^{d_1+d_2})$	$O(N^{-\frac{r}{r+1}}\log(N)^{d_2+t_1})$
	SG/Frolov: $O(N^{-r}\log(N)^{t(r,d)})$	$O(N^{-\frac{r}{2r+1}}\log(N)^{t_2})$	$O(N^{-\frac{r}{r+1}}\log(N)^{d_1+t_2})$	$O(N^{-\frac{r_1 r_2}{r_1+r_2}}\log(N)^{t_1+t_2})$

Table 4: Expected convergence rates for FTP quadrature with optimal σ^* and $\alpha = 1$.

various combinations of Q^1 and Q^2. In general, we see that the overall convergence rate is always bounded by the lesser rate of Q^1 and Q^2. If one of the formulas or convergence rates are fixed either by the availability of quadrature nodes or regularity of the integrand, it can only be the goal to approximate this rate as close as possible. Naturally, one can always try to do this by using higher order rules for the other integral.
Table 5 shows that the optimal (main) rate can be achieved with any complimentary rule of at least equal order via STP quadrature. This performance is especially impressive if both formulas have the same order: For the same number of nodes the main rate of STP is squared compared to FTP. This is also exactly the behavior which can be observed for traditional sparse grid quadrature, justifying the treatment of STP as a generalized version of SG quadrature.

We start with a synthetic test function in order to demonstrate the general applicability of the suggested methods. Let

$$\varphi_{\text{test}}(u, z|\theta) := u^{z+\theta-1}e^{-u} \tag{6.13}$$

and $\Omega_1 = [0, \infty),\ \Omega_2 = [0, 1]$. Then

$$\begin{aligned} Q_0(\theta) &= \int_0^1 \log\left(\int_0^\infty \varphi_{\text{test}}(u, z|\theta)dx\right) dz = \int_0^1 \log\left(\Gamma(z+\theta)\right) dz \\ &= -\log(\Gamma(\theta)) - \theta + (\theta - 1)\log(\theta) + \log(\Gamma(\theta+1)) + \frac{1}{2}\log(2\pi)\,, \end{aligned}$$

where Γ denotes the Gamma function.
Although the general setup is intended for multidimensional integration this example with one-dimensional Ω_1 and Ω_2 already illustrates the improvements resulting from the Multilevel approach. In particular, φ is smooth, so $\varphi_{\text{test}} \in H^r_{\text{mix}}$ for any $r > 0$, implying that every presented quadrature

		$Q^1 =$		
		MC	QMC	SG/Frolov
$Q^2 =$	MC: $O(N^{-1/2})$	$O(N^{-1/2}\log(N)^{3/2})$	$O(N^{-1/2}\log(N)^{d_1})$	$O(N^{-1/2}\log(N)^{t_1})$
	QMC: $O(N^{-1}\log(N)^{d})$	$O(N^{-1/2}\log(N)^{d_2})$	$O(N^{-1}\log(N)^{d_1+d_2+2})$	$O(N^{-1}\log(N)^{d_2+t_1})$
	SG/Frolov: $O(N^{-r}\log(N)^{t(r,d)})$	$O(N^{-1/2}\log(N)^{t_2})$	$O(N^{-1}\log(N)^{d_1+t_2})$	$O(N^{-\min\{r_1,r_2\}}\log(N)^{t_1+t_2})$

Table 5: Expected convergence rates for STP quadrature with optimal σ^* and $\alpha = 1$.

method achieves its maximal order of convergence and allowing the direct comparison of computed and observed rates.

The choice $F(z, \theta, t) = \log(t)$ resembles the Maximum Likelihood setup (6.1). Written in the general framework (6.2) the estimator based on the Loglikelihood is constructed with some function φ of an unobservable variable which integrates to the choice probability $P(z|\theta) = \int_U \varphi(u, z|\theta)du$. Theorems 6.4 and 6.5 required that F be Hölder continuous in t for all $z \in \mathcal{Z}$ and $\theta \in \Theta$. The logarithm is in fact *Lipschitz* continuous, i.e. Hölder continuous with constant $\alpha = 1$ but only for $P(z|\theta) \geq \delta > 0$ for some small constant δ. In an econometric context, it makes sense to assume that $P(z|\theta) > 0$ but the additional bound from 0 is less easy to justify. As ML-estimation also requires the integration over z we will assume, that such a bound can be prescribed by the choice of the search region θ. Additionally, the function log is smooth in $(0, \infty)$ hence $\log \in H^r_{\text{mix}}$ for any $r > 0$.

Figure 13 presents six combinations of quadrature formulas for Q^1 and Q^2: All plots on the left hand side display the expected better rate of STP versus FTP quadrature. The generally higher convergence of SG quadrature demonstrates the shift from $r/2$ to r best. A similar result is obtained for the combination of optimal weight cubature with Sobol quadrature, since both also achieve comparatively high rates on their own. In contrast, the combinations of Monte Carlo integration with SG and Sobol rules support the claim that the slow convergence of the MC rule can barely be ameliorated by the use of higher order formulas for the other integral.

We shortly want to explain the sometimes erratic behavior of the curves: MC integration naturally exhibits some variance in the error due to the probabilistic nature of the choice of the quadrature nodes. The structure of the STP rule strengthens this behavior if Q^1 and Q^2 are not equally balanced, i.e. $\sigma \neq 1$. Then $\frac{J}{\sigma}$ and σJ are not necessarily integers, so the conditions $j_1 \leq \frac{J}{\sigma}$ and $j_2 \leq \sigma J$ lead to an uneven increase of j_1 and j_2, where e.g. j_2 is only increased every $1/\sigma^2$-th increase of j_1.

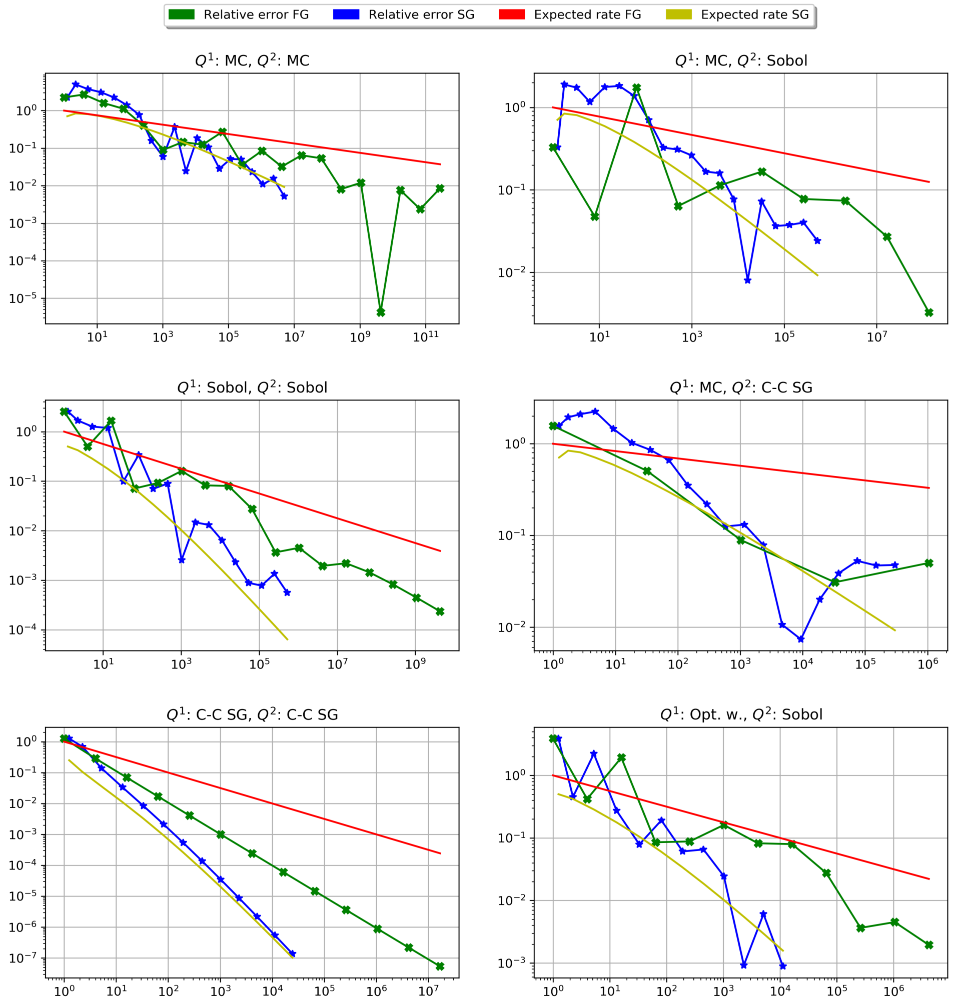

Figure 13: Full (FG) and sparse (SG) tensor product quadrature for $F(z,t) = \log(t)$ and φ_{test} with $\theta = 1$.
"MC" = Monte Carlo quadrature, "Sobol" = Quasi Monte Carlo quadrature with Sobol points, "C-C SG" = Sparse grid quadrature for the Clenshaw-Curtis rule, "G-Leg SG" = Sparse grid quadrature for the Gauss-Legendre rule, "Opt. w." = Optimal weights quadrature.

The next two examples are based on Generalized Linear Mixed models from Sections 2.3 and 4.2 and can both be estimated either with ML- or GMM-estimation. We already encountered the Mixed Logit model (see Section 2.3 and [83]) where the choice probability is calculated by integrating the plain Logit probability over a parameter distribution,

$$P(z|\theta) = \int_U \frac{e^{u \cdot z_i}}{\sum_{j=1}^{J} e^{u \cdot z_j}} d\mu(u) \,.$$

Here, the observable variable $z \in \mathbb{R}^{q \times J}$ contains q values for each of the J choices and the parameter $u \in \mathbb{R}^q$ is distributed according to the measure μ in its domain $U = \mathbb{R}^q$. We consider a multivariate Gaussian distribution for μ with mean $(0, ..., 0)$ and covariance matrix $\Sigma = \sigma_{0.1}$ as in Section 4.2.

For the Multinomial Probit model, the variables z, x and J have the same shapes and we consider the choice probability (2.14) in its transformed version (4.2) with covariance matrix $\Sigma_{0.1}$. Instead of a fixed utility $\mathbf{v} = (0.5, ..., 0.5)$ as in Section 4.2 we now compute the utility for each data point z_n, i.e. every quadrature node of the outer integral, and for one fixed choice $k = 1$ by definition (4.3). In our notation, the random variable X in (4.3) has the realizations z_n and the parameter vector β is denoted by the estimate θ. For our exemplary computations, we used $\theta = (1, ..., 1) \in \mathbb{R}^q$.

Since both models define a proper choice probability, Maximum Likelihood would be the natural choice for estimation. However, the fact that we cannot compute the choice probability exactly calls for the use of approximated estimators, i.e. MAL- and GMAM-estimation. The already mentioned research works [34], [39] and [68] showed that under certain circumstances GMAM might be the better option.

For both estimators, the outer integral is defined over the full data space from the real world. While it might be easy to quantify the range of the data ($\mathcal{Z} \subset \mathbb{R}^d$) it is much harder to determine their distribution ν in $\mathcal{Z}$. In particular, we cannot choose the quadrature nodes deterministically. Hence, the sampling of data points is inherently random and limits the choice of Q^1 to quadrature methods which are based on random nodes. Therefore, we only consider Monte Carlo and Optimal weights cubature for Q^1 and combine them with low- (MC), medium- (Sobol) and high-order (SG or Frolov) rules for Q^2.

In terms of the established notations in Section 6.3, the Mixed Logit model with ML-estimation yields the integrand

$$\varphi_{\mathrm{MixL}}(u, z|\theta) = \frac{e^{u \cdot z_i}}{\sum_{j=1}^{J} e^{u \cdot z_j}}$$

and again the intermediate function $F(z, \theta, t) = \log(t)$. We let $\mathcal{Z} = [0, 1]^{q \times J}$ and ν be the uniform distribution and set $J = 3$ and $q = 4$, so $\mathcal{I}_1$ denotes a 12- and $\mathcal{I}_2$ denotes a 4-dimensional integral. Similar to the synthetic case, φ_{MixL} is smooth, so theoretically any order of convergence could be obtained asymptotically. The possibly problematic cases of $P(z|\theta)$ being very close to 0 however remains.

Figure 14 now supports the claims made in Section 6.3: If MC integration is used for Q^1, then MC integration also gives the highest improvement of the usage of STP compared with FTP. In particular, a sparse grid rule for Q^2 improves FTP and STP simultaneously. Similar results were obtained for optimal weights cubature where the difference between STP and FTP

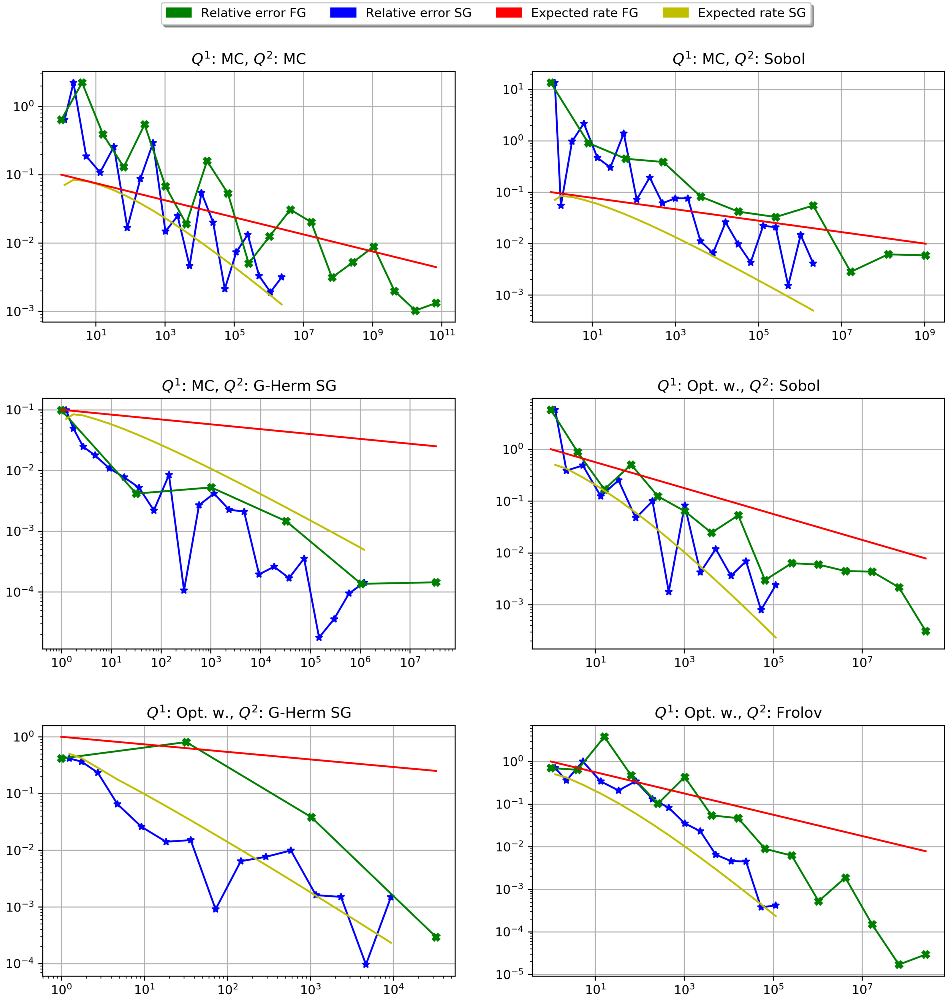

Figure 14: Full (FG) and sparse (SG) tensor product quadrature for the Mixed Logit model estimated with Maximum Likelihood. (Legend in figure 13)

quadrature can be observed clearly for all combinations.

Finally, figure 15 displays results for the Multinomial Probit model which is estimated by a GMM estimator based on the seminal paper by McFadden [68]. Then $G_0(\theta)$ is exactly of the form (5.4) and we can use orthogonality conditions defined by (5.5). We consider the approximation G_N of G_0 with nodes z_n and weights w_n for $n = 1, ..., N$. In Section 2.3, we defined the Multinomial Probit choice probability $P(y_n = j|x_n, \theta) = P_j^{(n)}(\theta)$ for every data point $z_n = (x_n, y_n)$ and let $y_j^{(n)} = 1$ if j is the observed choice of individual n and 0 otherwise. Now, the GMM-estimator is the maximizer

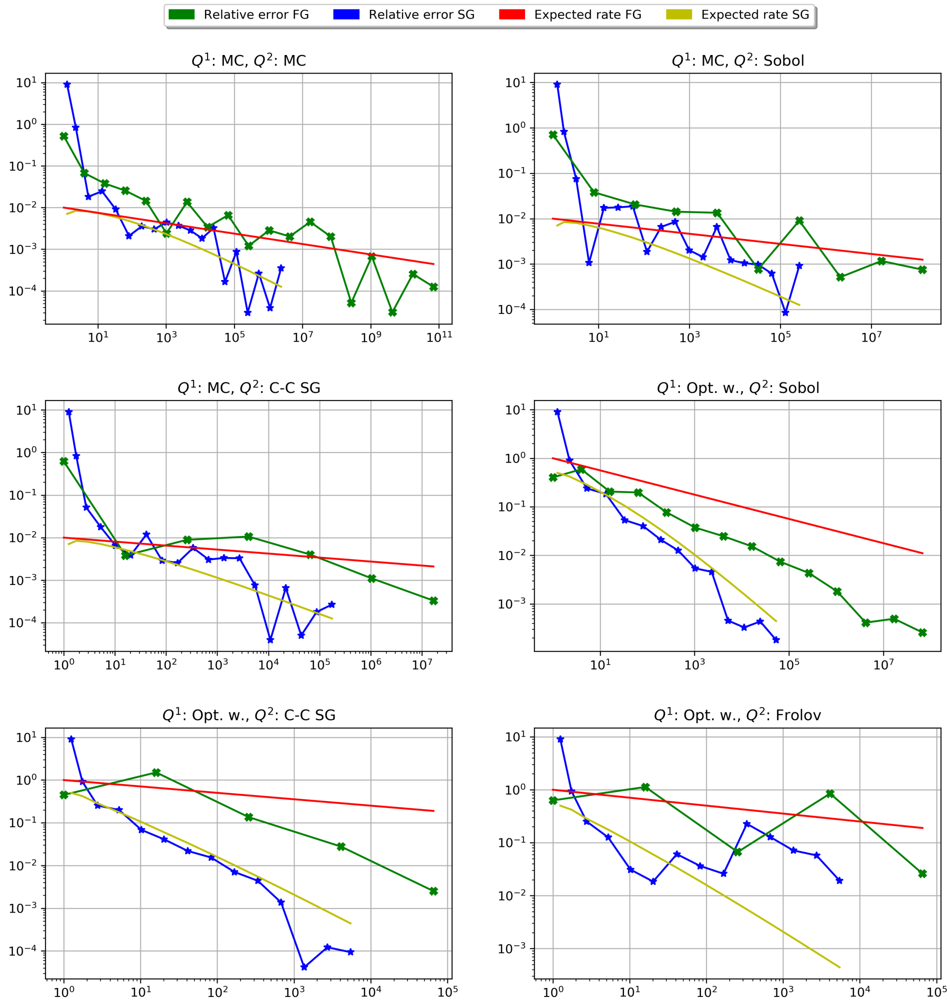

Figure 15: Full (FG) and sparse (SG) tensor product quadrature for the Multinomial Probit model estimated with GMM. (Legend in figure 13)

of the euclidean norm of

$$G_N(\theta) = \nabla_\theta \left(\log P^{(n)}(\theta)\right)^T \sum_{n=1}^{N} w_n \begin{pmatrix} y_1^{(n)} - P_1^{(n)}(\theta) \\ \vdots \\ y_J^{(n)} - P_J^{(n)}(\theta) \end{pmatrix}$$
$$= \sum_{n=1}^{N} w_n \sum_{j=1}^{J} \nabla_\theta P_j^{(n)}(\theta) \left(\frac{y_j^{(n)}}{P_j^{(n)}(\theta)} - 1 \right) .$$

According to the definition (2.14) of $P_j^{(n)}(\theta)$, the choice probability is given as the c.d.f. of a multivariate Gaussian distribution, so the derivative exists

and is given by the corresponding p.d.f.. Furthermore $P_j^{(n)}(\theta)$ needs to be computed only for the case $y_j^{(n)} = 1$ so the approximation problem for this estimator boils down to the computation of one Multinomial Probit integral for each node/data point z_n.
As already discussed in Section 4.2, the Genz-transformed Probit integral is better suited for numerical quadrature. In the setting of tensor product integration, its definition (4.2) yields the inner integrand

$$\varphi_{\mathrm{MNP}}(u, z|\theta) = \prod_{i=1}^{J-1} \hat{v}_i(u_1, ..., u_{i-1})$$

As intermediate function, we obtain $F(z,t) = \frac{1}{t}$.
We let again $\mathcal{Z} = [0,1]^{q \times J}$ and ν be the uniform distribution and set $J = 5$ and $q = 3$, so $\mathcal{I}_1$ denotes a 15- and $\mathcal{I}_2$ a 4-dimensional integral. Again, φ_{MNP} is smooth, so any quadrature formula should achieve its best rate. F is Lipschitz if t is bounded away from 0, so the conditions of Theorems 6.4 and 6.5 are met.
We see in figure 15 that STP clearly outperforms FTP for all combinations of MC or Optimal weight cubature with low and high order formulas for Q^2. In particular, for Monte Carlo integration STP and FTP follow the expected rates closely, similar to the above case of Mixed Logit/Maximum likelihood. Only the combination of Optimal Weights with Frolov cubature fails due to the original bad performance of the latter (see Section 4.2).

As final example, we recall the Mixed Probit model from Section 2.3: There we noticed that although the multivariate Probit model already allows for correlation between choices, a mixture distribution might be superior in some cases. Yet, a Mixed Probit model is computationally even more challenging since it involves not only the approximation of a multivariate Gaussian distribution but also of the integral over the parameter mixture. In particular, the multivariate Gaussian has to be calculated at every quadrature node for the mixture integral.
Hence, we have again two nested integrals for which we can compare FTP and STP quadrature. The inner integrand remains φ_{MNP} with covariance matrix $\Sigma = \Sigma_{0.2}$ but in the computation of the utility $\mathbf{v}$ the roles of integration and fixed variable are interchanged: We now integrate over β and fix a set of observed variables X, so in the established notation we have "z" = "β" and X is a value in the parameter vector θ. This also changes the dimensionality of Ω_1 from $q \cdot J$ to q. We draw $q \cdot J$ values randomly from a uniform distribution to assemble $X \in \mathbb{R}^{q \times J}$ and set $q = 4$ and $J = 5$.
Furthermore, the intermediate function F becomes

$$F(z, \theta, t) = t \cdot \varphi(z|\mu, \Psi)$$

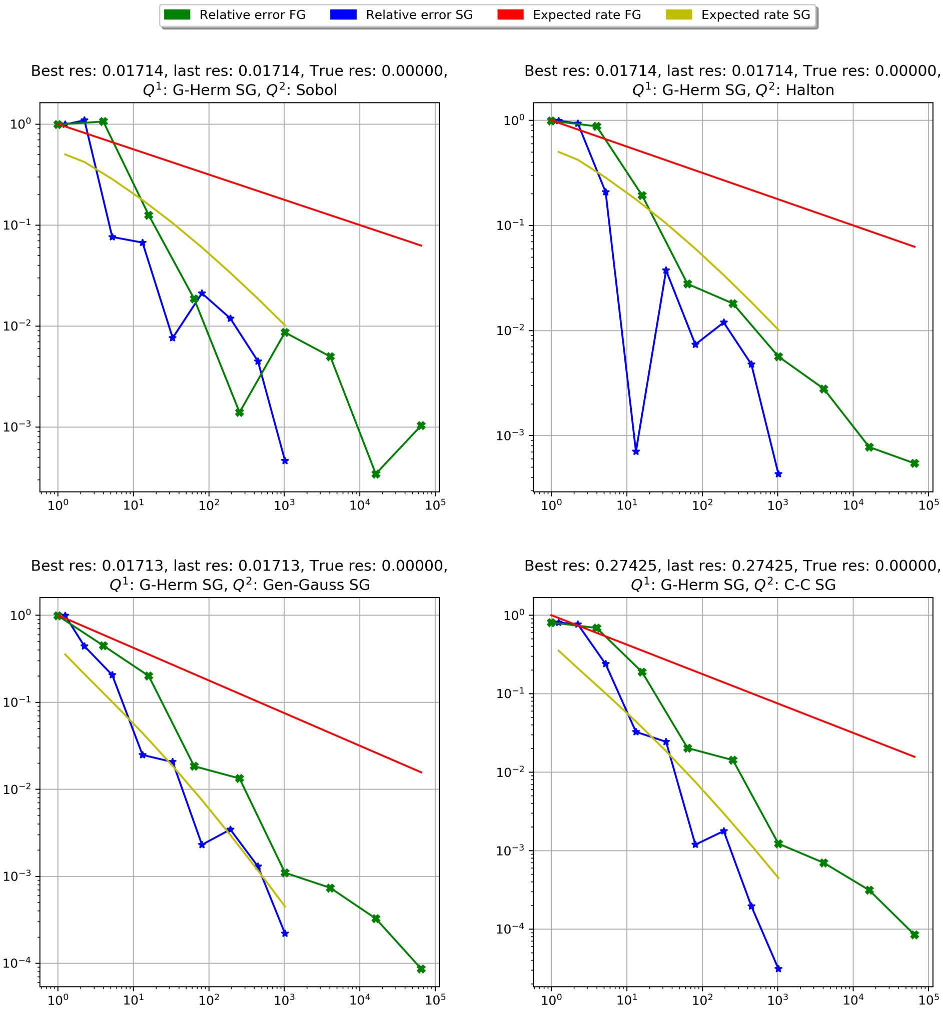

Figure 16: Full (FG) and sparse (SG) tensor product quadrature for the Mixed Multinomial Probit model. (Legend in figure 13)

for a mixing distribution $\varphi(\cdot|\mu, \Psi)$ with mean μ and covariance matrix Ψ. As in previous chapters, we let φ be a multivariate Gaussian distribution, set $\mu = (0.2, ..., 0.2)$ and $\Psi = \Sigma_{0.1}$ and have the complete parameter vector $\theta = (X, \mu, \Psi, \Sigma)$.

Lastly, we are not restricted to MC and Optimal Weights cubature for the outer integral as the integration is completely model induced. Therefore, we can test STP quadrature for higher-order rules and display the resulting improvements in figure 16. Once more, it underscores the predictions we made in tables 4 and 5 as STP clearly outperforms FTP quadrature. Furthermore, we see how the benefits are more visible if same-order rules are used for Q^1 and Q^2 and how the high order of SG quadrature is sustained.

Summarizing this and the previous findings of this section we conclude that econometric estimation and nested integrals arising from econometric models can significantly benefit from using sparse tensor product quadrature. In estimation, it enables us to reach the best possible main rate $N^{1/2}$ for any rule Q^2 and hence increases the accuracy of an ML- or GMM-estimator for a fixed set of observations. For nested integrals as in the Mixed Probit model, we preserve polynomial (and possibly even exponential) convergence rates and make intractable models feasible.

7 Conclusion

Overview

The objective of this thesis was to systematize the use of numerical quadrature in econometrics and investigate properties of recurring econometric integrals in order to promote higher-order quadrature. Our second goal was to understand how econometric estimators are affected by approximation. In this context, Multilevel quadrature offers another approach to improve the joint convergence behavior of approximated estimators.

Part I was devoted to defining and evaluating integrals from econometric models. We began with the derivation of Generalized Linear Mixed models which are generalizations of both, Generalized Linear models and Linear Mixed models. GLMM provide a wide variety of possible specifications which allow the evaluation of many different kinds of data. Two prominent examples are categorical data arising from Discrete Choice modeling and count data. However, econometricians usually rely on only a few heavily used models and specifications, namely Mixed Logit, (Mixed) Multinomial Probit (categorical data) and Mixed Poisson (count data) models.

For all three models and similarly for the fourth example, a Linear Quantile Mixed model, the introduction of correlation between individuals or over time added another unobservable random variable to the model. This variable is integrated out to get a deterministic expression, leading to a multidimensional integral without analytical solution. Additionally, the plain Multinomial Probit requires the computation of a multivariate Gaussian c.d.f. which is also not analytically achievable.

Afterwards, we introduced integrals from two exemplary Dynamic Economic models. For both DEM, the objective is to maximize a temporally aggregated utility function using Bellman's principle of optimality. Hence, in every step (and possibly for many states) the next-period unobservable variables have to be integrated out.

Having defined a series of econometric integrals, the next chapter was dedicated to the presentation of numerical quadrature formulas. We described one-dimensional Newton-Cotes, Clenshaw-Curtis and Gauss formulas and how they are extended to multidimensional integrals by the product rule. Next, we encountered Sparse Grid quadrature and portrayed its construction and error bounds. SG quadrature circumvents the curse of dimensionality exhibited by the product rule by strategically omitting nodes. It achieves high convergence rates in mixed Sobolev spaces with further extensions, like SG quadrature based on Generalized Gauss formulas, adapting it to functions with boundary singularities.

We continued our discussion of quadrature formulas with Monte Carlo and Quasi Monte Carlo methods: In addition to the classic MC integration and Sobol and Halton sequences, Frolov cubature was presented as an example for a quadrature rule which almost achieves the optimal bound on mixed

Sobolev spaces. Finally, the relatively new Optimal Weights formula offered a way to improve randomly drawn nodes by computing weights based on the notion of the Reproducing Kernel Hilbert space.

In the last chapter of part I, we applied the established methods to the previously defined integrals. We found that most integrands are highly regular or even smooth, so higher-order quadrature should lead to polynomial or even exponential convergence rates. However, the pre-asymptotic performance of these rules was often less promising: The exponential d-dependence of the secondary rate prevented fast convergence for high-dimensional integrals. Furthermore, all specifications of GLMM and DEM led to parameterized sets of integrals. Depending on the chosen parameter we supposed that the constant in the O-notation of the error bounds grows excessively and outweighs the high convergence rate for small and medium N. Yet, we found reasonable parameter regions where higher-order quadrature clearly outperformed (Q)MC rules and is therefore a viable alternative.

In the second part of the thesis, we considered econometric estimators. After defining extremum estimators and the popular subclasses of M- and GMM-estimators, we stated conditions for them to be consistent and asymptotically normal. As the respective objective functions often need to be approximated, these conditions have to be proven again for the approximated estimators. We gave sufficient conditions for consistency and asymptotic normality of the Generalized Methods of Approximated Moments (GMAM) estimator.

The thesis concluded with the derivation of Multilevel quadrature and Multilevel estimation: Firstly, we discovered that M- and GMM-estimators can be considered as Monte Carlo simulations of integrals over the domain of the observable variables, i.e. the “real world data” space. Together with the integrals posed by the respective models and objective functions they comprise nested integrals separated by an intermediate function.

Next, we adapted the sparse tensor product space technique for integration problems and proved the corresponding theorems on error bounds and the optimal balancing factor. In particular, we proposed a Hölder continuity condition for the intermediate function to preserve error bounds. It turned out that the improvements of Multilevel quadrature compared with classic quadrature are most significant if the rules used for inner and outer integral achieve similar convergence rates. Then, the total rate is almost squared for Multilevel quadrature and almost equals the rate of each separate formula.

We combined both notions to obtain significantly improved approximations of Maximum Approximated Likelihood- and GMAM-estimators. Here, again Mixed Logit and Multinomial Probit served as examples for models with intractable multidimensional integrals. Lastly, the Mixed Multinomial Probit model directly includes a nested integral where Multilevel quadrature was similarly effective.

Outlook

With rising complexity of econometric models, advanced numerical approximation methods remain a crucial instrument to ensure computability, and even increase their importance for future research. Based on the presented topics and results, three paths seem promising for further investigations:
Our survey of multidimensional econometric integrals only covered one larger class of models and only two examples of Dynamic Economic models. It could be extended by including even more models and could be deepened by making the parameter regions more precise for which SG or Frolov quadrature are more effective than (Q)MC rules.
The second possibility was mentioned for Dynamic Discrete Choice models: Here, each choice probability, i.e. integral, has to be computed for each state in the state space to determine the optimal utility for an individual. Although first approaches were made to interpolate the values for each state [57], a Multilevel/Sparse Grid method could possibly enhance these efforts. Similarly, Sparse Grid interpolation for multidimensional PDEs in DEM has already been tested [8] but has not yet been surveyed comprehensively.
The original Sparse Tensor Product and Multilevel methods where developed for joint interpolation problems while we extended it for nested integrals. It is also possible to use it on mixed interpolation-integration problems [31]. Besides multidimensional integrals and PDEs, econometric models and estimators also often require the numerical maximization of some objective function over a multidimensional parameter space. Hence, it could be possible to reduce computational efforts with a Multilevel approach for different approximated quantities for the estimation of an econometric model.

From the presented results, we conclude that advanced numerical methods offer the opportunity to further enhance current approximation schemes in econometrics. As yet, most research into the application of Sparse Grid methods to econometrics has focused on interpolation of high-dimensional functions. This thesis has shown that higher-order quadrature is also a valuable tool for econometrics as multidimensional integrals are a frequent issue in complex econometric models. We believe that a further cooperation between numerical mathematics and econometrics could identify and utilize even more potential than the few suggestions above already illustrate.

References

[1] V. Aguirregabiria and P. Mira. “Dynamic discrete choice structural models: A survey”. In: *Journal of Econometrics* 156.1 (2010), pp. 38–67.

[2] T. Amemiya. *Advanced Econometrics.* Harvard University Press, 1985.

[3] R. Bellmann. *Adaptive Control Processes: A Guided Tour.* Princeton University Press, 1961.

[4] S. Berry, J. Levinsohn and A. Pakes. “Automobile Prices in Market Equilibrium”. In: *Econometrica* 63.4 (1995), pp. 841–890.

[5] C. R. Bhat. “Accommodating variations in responsiveness to level-of-service measures in travel mode choice modeling”. In: *Transportation Research Part A: Policy and Practice* 32.7 (1998), pp. 495 –507.

[6] C. R. Bhat. “Quasi–random maximum simulated likelihood estimation of the mixed multinomial logit model”. In: *Transportation Research Part B: Methodological* 35.7 (2001), pp. 677–693.

[7] R. Blundell. “What Have We Learned from Structural Models?” In: *American Economic Review* 107.5 (2017), pp. 287–292.

[8] J. Brumm and S. Scheidegger. “Using Adaptive Sparse Grids to Solve High–Dimensional Dynamic Models”. In: *Econometrica* 85.5 (2017), pp. 1575–1612.

[9] H.-J. Bungartz and M. Griebel. “Sparse grids”. In: *Acta Numerica* 13 (2004), pp. 147–269.

[10] J. S. Butler and R. Moffitt. “A computationally efficient quadrature procedure for the one–factor multinomial probit model”. In: *Econometrica* 50.3 (1982), pp. 761–764.

[11] P. J. Davis and P. Rabinowitz. *Methods of Numerical Integration.* Computer science and applied mathematics. Academic Press, 1975.

[12] J. Dick, F. Y. Kuo and I.H. Sloan. “High–dimensional integration: The quasi–Monte Carlo way”. In: *Acta Numerica* 22 (2013), pp. 133–288.

[13] J. Dick and F. Pillichshammer. *Digital Nets and Sequences: Discrepancy Theory and Quasi-Monte Carlo Integration.* Cambridge University Press, 2010.

[14] D. Duffie and K. Singleton. “Simulated Moments Estimator of Markov Models of Asset Prices”. In: *Econometrica* 61.4 (1993), pp. 929–952.

[15] P. Eisenhauer. “The approximate solution of finite–horizon discrete-choice dynamic programming models”. In: *Journal of Applied Econometrics* 34.1 (2019), pp. 149–154.

[16] P. Eisenhauer, J. J. Heckman and S. Mosso. "Estimation of dynamic discrete choice models by maximum likelihood and the simulated method of moments". In: *International economic review* 56.2 (2015), pp. 331–357.

[17] K. Frolov. "Upper bounds for the errors of quadrature formulae on classes of functions." In: *Dokl. Akad. Nauk SSSR* 231 (1976), 818–821.

[18] J. Garcke. "Sparse Grids in a Nutshell". In: *Sparse grids and applications.* Ed. by J. Garcke and M. Griebel. Vol. 88. Lecture Notes in Computational Science and Engineering. Springer, 2013, pp. 57–80.

[19] S. Geman and D. Geman. "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6.6 (1984), pp. 721–741.

[20] A. Genz. "Numerical Computation Of Multivariate Normal Probabilities". In: *Journal of Computational and Graphical Statistics* 1 (2000).

[21] M. Geraci. "Mixed-effects models using the normal and the Laplace distributions: A $\mathbf{2} \times \mathbf{2}$ convolution scheme for applied research". Available as arXiv preprint 1712.07216. 2017.

[22] M. Geraci and M. Bottai. "Linear quantile mixed models". In: *Statistics and computing* 24.3 (2014), pp. 461–479.

[23] T. Gerstner and M. Griebel. "Dimension–Adaptive Tensor–Product Quadrature". In: *Computing* 71.1 (2003), pp. 65–87.

[24] T. Gerstner and M. Griebel. "Numerical Integration using Sparse Grids". In: *Numerical Algorithms* 18.3 (1998), pp. 209–232.

[25] J. Geweke. "Efficient Simulation from the Multivariate Normal and Student-t Distributions Subject to Linear Constraints and the Evaluation of Constraint Probabilities". In: *Computing Science and Statistics: Proceedings of the Twenty-Third Symposium on the Interface.* Vol. 23. 1998.

[26] M. B. Giles. "Multilevel monte carlo methods". In: *Acta Numerica* 24 (2015), pp. 259–328.

[27] C. Gourieroux and A. Monfort. *Simulation-based Econometric Methods.* Oxford University Press, 1997.

[28] C. Gourieroux and A. Monfort. *Statistics and Econometric Models.* Vol. 2. Themes in Modern Econometrics. Cambridge University Press, 1995.

[29] W. H. Greene. *Econometric analysis.* Pearson Education India, 2003.

[30] M. Griebel and H. Harbrecht. "On The Construction Of Sparse Tensor Product Spaces". In: *Mathematics of Computations* 82.282 (2013), pp. 975–994.

[31] M. Griebel, H. Harbrecht and M. Multerer. “Multilevel quadrature for elliptic parametric partial differential equations in case of polygonal approximations of curved domains”. In: *SIAM Journal on Numerical Analysis* 58.1 (2020), pp. 684–705.

[32] M. Griebel, S. Knapek and G. Zumbusch. *Numerical Simulation in Molecular Dynamics: Numerics, Algorithms, Parallelization, Applications*. Springer, 2007.

[33] M. Griebel and J. Oettershagen. “Dimension-adaptive sparse grid quadrature for integrals with boundary singularities”. In: *Sparse grids and Applications*. Vol. 97. Lecture Notes in Computational Science and Engineering. Springer, 2014, pp. 109–136.

[34] M. Griebel et al. “Maximum approximated likelihood estimation”. Submitted to Econometrica. Available as INS Preprint No. 1905. 2019.

[35] V. A. Hajivassiliou. “A simulation estimation analysis of the external debt crises of developing countries”. In: *Journal of Applied Econometrics* 9.2 (1994), pp. 109–131.

[36] V. A. Hajivassiliou. “Some practical issues in maximum simulated likelihood”. In: *Simulation-Based inference in econometrics: Methods and Applications*. Ed. by R. Mariano, T. Schuermann and M. J. Weeks. Cambridge University Press Cambridge, 2000, pp. 71–99.

[37] V. A. Hajivassiliou and D. L. McFadden. “The method of simulated scores for the estimation of LDV models”. In: *Econometrica* 66.4 (1998), pp. 863–896.

[38] V. A. Hajivassiliou, D. L. McFadden and P. A. Ruud. “Simulation of multivariate normal rectangle probabilities and their derivatives theoretical and computational results”. In: *Journal of Econometrics* 72.1 (1996), pp. 85 –134.

[39] V. A. Hajivassiliou and P. A. Ruud. “Chapter 40 Classical estimation methods for LDV models using simulation”. In: *Handbook of Econometrics*. Vol. 4. Elsevier, 1994, pp. 2383 –2441.

[40] A. R. Hall. *Generalized method of moments*. Oxford University Press, 2005.

[41] M. Hanke-Bourgeois. *Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens*. Vieweg+Teubner Verlag, 2008.

[42] B. Hansen. *Econometrics*. 2020.

[43] L. Hansen. “Large Sample Properties Generalized Method of Moments Estimators”. In: *Econometrica* 50.4 (1982), pp. 1029–1054.

[44] P. Harjulehto and P. Hästö. *Orlicz spaces and generalized Orlicz spaces*. Lecture Notes in Mathematics. Springer, 2019.

[45] F. Hayashi. *Econometrics.* Princeton University Press, 2000.

[46] S. Heinrich. “Multilevel Monte Carlo Methods”. In: *Large-Scale Scientific Computing.* Ed. by S. Margenov, J. Waśniewski and P. Yalamov. Springer, 2001, pp. 58–67.

[47] F. Heiss. “The panel probit model: Adaptive integration on sparse grids”. In: *Advances in Econometrics* 26 (2010), pp. 41–64.

[48] F. Heiss and V. Winschel. “Likelihood approximation by numerical integration on sparse grids”. In: *Journal of Econometrics* 144.1 (2008), pp. 62–80.

[49] K. Judd, L. Maliar and S. Maliar. “Numerically stable and accurate stochastic simulation approaches for solving dynamic economic models”. In: *Quantitative Economics* 2.2 (2011), pp. 173–210.

[50] K. Judd and B. Skrainka. *High performance quadrature rules: how numerical integration affects a popular model of product differentiation.* CeMMAP working papers. 2011.

[51] K. Judd et al. “Smolyak method for solving dynamic economic models: Lagrange interpolation, anisotropic grid and adaptive domain”. In: *Journal of Economic Dynamics and Control* 44 (2014), pp. 92–123.

[52] C. Kacwin. “Realization of the Frolov cubature formula via orthogonal Chebyshev-Frolov lattices”. Master Thesis. Institut für Numerische Simulation, Universität Bonn, 2016.

[53] C. Kacwin et al. “Numerical performance of optimized Frolov lattices in tensor product reproducing kernel Sobolev spaces”. In: (2018). INS Preprint No. 1801, pp. 1–40.

[54] D. K. Kahaner and G. Monegato. “Nonexistence of extended Gauss-Laguerre and Gauss-Hermite quadrature rules with positive weights”. In: *Zeitschrift für angewandte Mathematik und Physik ZAMP* 29.6 (1978), pp. 983–986.

[55] M. P. Keane, P. E. Todd and K. I. Wolpin. “Chapter 4 - The Structural Estimation of Behavioral Models: Discrete Choice Dynamic Programming Methods and Applications”. In: *Handbook of Labor Economics.* Ed. by O. Ashenfelter and D. Card. Vol. 4. Elsevier, 2011, pp. 331–461.

[56] M. P. Keane and K. I. Wolpin. “The career decisions of young men”. In: *Journal of Political Economy* 105.3 (1997), pp. 473–522.

[57] M. P. Keane and K. I. Wolpin. “The solution and estimation of discrete choice dynamic programming models by simulation and interpolation: Monte Carlo evidence”. In: *The Review of Economics and Statistics* 76.4 (1994), pp. 648–672.

[58] S. Kotz, T. Kozubowski and K. Podgorski. *The Laplace Distribution and Generalizations*. Springer, 2001.

[59] D. Kristensen and B. Salanié. "Higher-order properties of approximate estimators". In: *Journal of Econometrics* 198.2 (2017), pp. 189–208.

[60] A. S. Kronrod. *Nodes and weights of quadrature formulas. Sixteen-place tables*. Consultants Bureau, New York, 1965.

[61] D. Krueger and F. Kubler. "Computing OLG Models with Stochastic Production". In: *Journal of Economic Dynamics and Control* 28.7 (2004), pp. 1411–1436.

[62] N. T. Longford. "Random Coefficient Models". In: *Handbook of Statistical Modeling for the Social and Behavioral Sciences*. Ed. by G. Arminger, C. C. Clogg and M. E. Sobel. Springer, 1995, pp. 519–570.

[63] L. Maliar and S. Maliar. "Chapter 7 - Numerical Methods for Large-Scale Dynamic Economic Models". In: *Handbook of Computational Economics*. Ed. by Karl Schmedders and Kenneth L. Judd. Vol. 3. Elsevier, 2014, pp. 325 –477.

[64] B. A. Malin, D. Krueger and F. Kubler. "Solving the multi-country real business cycle model using a Smolyak-collocation method". In: *Journal of Economic Dynamics and Control* 35.2 (2011). Computational Suite of Models with Heterogeneous Agents II: Multi-Country Real Business Cycle Models, pp. 229 –239.

[65] P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Monographs on statistics and applied probability. 1989.

[66] C. McCulloch, S. Searle and J. Neuhaus. *Generalized, Linear, and Mixed Models*. Wiley, 2001.

[67] C. E. McCulloch. "Generalized Linear Mixed Models". In: *NSF-CBMS Regional Conference Series in Probability and Statistics* 7 (2003), pp. 1–84.

[68] D. McFadden. "A Method of Simulated Moments for Estimation of Discrete Response Models without Numerical Integration". In: *Econometrica* 57.5 (1989), pp. 995–1026.

[69] W. K. Newey and D. McFadden. "Chapter 36 Large sample estimation and hypothesis testing". In: *Handbook of Econometrics*. Vol. 4. Elsevier, 1994, pp. 2111 –2245.

[70] H. Niederreiter. *Random Number Generation and Quasi-Monte Carlo Methods*. CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics, 1992.

[71] E. Novak and K. Ritter. “The Curse of Dimension and a Universal Method For Numerical Integration”. In: *Multivariate Approximation and Splines.* Ed. by G. Nürnberger, J. W. Schmidt and G. Walz. Birkhäuser Basel, 1997, pp. 177–187.

[72] J. Oettershagen. “Construction of Optimal Cubature Algorithms with Applications to Econometrics and Uncertainty Quantification”. Dissertation. Institut für Numerische Simulation, Universität Bonn, 2017.

[73] A. Pakes and D. Pollard. “Simulation and the Asymptotics of Optimization Estimators”. In: *Econometrica* 57.5 (1989), pp. 1027–1057.

[74] T. N. L. Patterson. “The Optimum Addition of Points to Quadrature Formulae”. In: *Mathematics of Computation* 22.104 (1968), pp. 847–856.

[75] B. Ranneby. “The Maximum Spacing Method. An Estimation Method Related to the Maximum Likelihood Method”. In: *Scandinavian Journal of Statistics* 11.2 (1984), pp. 93–112.

[76] M. M. Rao and Z. D. Ren. *Theory of Orlicz Spaces.* Chapman & Hall Pure and Applied Mathematics. Taylor & Francis, 1991.

[77] P. C. Reiss and F. A. Wolak. “Chapter 64 Structural Econometric Modeling: Rationales and Examples from Industrial Organization”. In: *Handbook of Econometrics.* Ed. by James J. Heckman and Edward E. Leamer. Vol. 6. Elsevier, 2007, pp. 4277 –4415.

[78] C. Runge. “Über empirische Funktionen und die Interpolation zwischen äquidistanten Ordinaten”. In: *Zeitschrift für Mathematik und Physik,* 46 (1901), pp. 224–243.

[79] J. Rust. “Chapter 51 Structural estimation of markov decision processes”. In: *Handbook of Econometrics.* Vol. 4. Elsevier, 1994, pp. 3081 –3143.

[80] J. Rust. “Optimal Replacement of GMC Bus Engines: An Empirical Model of Harold Zurcher”. In: *Econometrica* 55.5 (1987), pp. 999–1033.

[81] S. M. Schennach. “Entropic Latent Variable Integration via Simulation”. In: *Econometrica* 82.1 (2014), pp. 345–385.

[82] S. A. Smolyak. “Quadrature and interpolation formulas for tensor products of certain classes of functions”. In: Soviet Mathematics Doklady 4, 1963, pp. 240–243.

[83] K. Train. *Discrete Choice Methods With Simulation.* Cambridge University Press, 2009.

[84] F. Tuerlinckx et al. “Statistical inference in generalized linear mixed models: A review”. In: *British Journal of Mathematical and Statistical Psychology* 59.2 (2006), pp. 225–255.

[85] M. Ullrich and T. Ullrich. “The Role of Frolov’s Cubature Formula for Functions with Bounded Mixed Derivative”. In: *SIAM Journal on Numerical Analysis* 54.2 (2016), pp. 969–993.

[86] G.W. Wasilkowski and H. Wozniakowski. “Explicit Cost Bounds of Algorithms for Multivariate Tensor Product Problems”. In: *Journal of Complexity* 11.1 (1995), pp. 1 –56.

[87] R. Winkelmann. *Econometric Analysis of Count Data*. Springer, 2008.

[88] C. Zenger. “Sparse grids”. In: *Parallel Algorithms for Partial Differential Equations*. Ed. by W. Hackbush. 1990.

[89] Y. Zhao et al. “General design Bayesian generalized linear mixed models”. In: *Statistical Science* 21.1 (2006), pp. 35–51.

[90] H.-T. Zhu and S.-Y. Lee. “Analysis of generalized linear mixed models via a stochastic approximation algorithm with Markov chain Monte-Carlo method”. In: *Statistics and Computing* 12.2 (2002), pp. 175–183.