Institute of Parallel and Distributed Systems

University of Stuttgart
Universitätsstraße 38
D–70569 Stuttgart

Bachelorarbeit Nr. 2489

The Sparse Grid Combination Technique for Quantities of Interest

Felix Huber

Course of Study:	Simulation Technology
Examiner:	Jun.-Prof. Dr. rer. nat. Dirk Pflüger
Supervisor:	Prof. Markus Hegland (ANU), Jun.-Prof. Dr. rer. nat. Dirk Pflüger
Commenced:	May 09, 2016
Completed:	November 08, 2016
CR-Classification:	G.1.0

Abstract

The curse of dimensionality is a major problem for large scale simulations. One way to tackle this problem is the sparse grid combination technique. While a full grid requires $\mathcal{O}\left(h_n^{-d}\right)$ grid points the sparse grid combination technique needs significantly less points.

In contrast to the traditional combination technique, which combines solution functions themselves, this work puts its focus on the combination technique with quantities of interest and their surpluses. After introducing the concept of surpluses that describe how much the solution changes if the grids are refined, we defined the combination technique as a sum of these surpluses. We show how the concept of surpluses can be utilized to deduce error bounds for the quantity of interest and helps to adapt the combination technique to problems with different error models. To improve the error bound we introduce a new extrapolated version of the combination technique and see how the surpluses are affected.

To evaluate our theoretical results we perform numerical experiments where we consider integration problems and the gyrokinetic plasma turbulence simulation GENE. The experimental results for the integration problems nicely confirm our derived theoretical results.

Contents

1 Intro

Energy consumption and the number of unknowns become major challenges when computing large scale simulations. However, large simulations are often needed to give high accuracy solutions for the problem at hand. In physics, large simulations are required to resolve fine grained effects that might have an impact on the overall solution. This causes high accuracy simulations to become quite expensive. Other simulations run on a large simulated area, which naturally leads to a large number of unknowns if one does not aggressively simplify the discretization. For simulations that are integrated over the time, accurate solutions for each time step are even more important, as the error accumulates over time. The more accurate the solutions are computed for each time step, the more steps one can simulate until the errors become to large.

So how do we get high accurate solutions for our problem? The most straight forward answer is to simply increase the grid resolution for the numerical approximation. However, increasing the grid resolution quickly becomes unfeasible, especially for high dimensional problems. This is due to the fast increasing computational power and memory requirements for solving the problem on a fine grid. As an example, assume that the number of grid points in each dimension is proportional to the parameter n, which controls the accuracy of the solution. The total number of grid points for a full grid solution in d dimensions is then of order $\mathcal{O}\left(n^d\right)$ for increasing n. As the memory requirement is typically at least of the same order as the number of grid points, this soon becomes a major limitation for further refinements. A similar problem arises with the computational complexity, which usually increases even faster than the number of grid points. This problem is often referred to as the *curse of dimensionality*.

These issues illustrate the need for other methods that reduce the complexity. Examples are extrapolation methods or the combination technique [GSZ92].

Throughout this work we focus on the combination technique. The rough idea is to split the full grid solution into a sum of so called *surpluses*. These surpluses describe the difference between consecutive full grid solutions. Instead of using all surpluses of the full grid solution, we reduce the computational effort by using only those surpluses that contribute the most to the solution.

In this work we present an alternative view on the combination technique. We take the focus off the solutions for individual component grids and define the combination

technique in terms of surpluses instead. Furthermore we do not consider solutions functions in different functions spaces, which are combined in case of the traditional combination technique. Instead we deal with certain properties of these solutions, which we refer to as *quantities of interest*. Now we first compute the quantities of interest on different component grids and combine the results in a second step. This is due to the fact that one often is not interested in the actual solution itself, but in certain properties of this solution. While the solution is usually a function over a grid, like an electric potential or distribution function, quantities of interest are scalar values derived from this solution. Examples for quantities of interest are integrals over the function, the average and minimal value, moments like the variance and skewness or even eigenvalues [HGC07].

We will first consider integration problems, where the quantity of interest is the integral over a fixed integration area. Finally we focus on quantities of interest computed by the plasma turbulence simulation GENE [GM13], which are deduced by a particle distribution function.

Additionally we present a new method, which aims to combine the properties of extrapolation with the combination technique. This method performed particularly well in experiments with different integration methods.

Overview

In chapter 2 we first discuss extrapolation methods. As we consider integration problems as a sample problem within this work we first do some ground by revisiting the Euler-Maclaurin formula, which gives us an error expansion for the trapezoidal rule. This error expansion will not only allow us to extrapolate the trapezoidal rule, but will also help us to understand the combination technique later throughout this work. The error expansion plays an important role to understand the behavior of the surpluses that are crucial for the combination technique in the next chapter. Based on the error expansion we then define the Romberg extrapolation and deduce a generalization for two dimensions. These extrapolation methods are then used in chapter 4 to deduce a combination technique working on extrapolated values. This chapter also includes the first part of my propaedeutic, which is section 2.1 about the Euler-Maclaurin formula.

In chapter 3 we present a novel, alternative view on the combination technique for quantities of interest. We first define the combination technique in terms of surpluses, based on their decay. In case of integration problems, this decay depends on the Euler-Maclaurin formula from the previous chapter. We then deduce error bounds for different types of surpluses. Unlike existing approaches that combine the solution functions, we

do not assume anything about the underlying function spaces on which the solution was computed. Instead we only assume a certain type of decay for the surpluses. We end this section with a short overview of variations of the combination technique in section 3.5, which forms the second part of my propaedeutic.

In chapter 4 we introduce a new method, which tries to combine the best of combination technique and Romberg extrapolation, based on our results of the two previous chapters. By mixing both methods we get a very accurate result. Additionally we reduce the computational effort compared to the two dimensional extrapolation due to the combination technique.

In chapter 5 we present results of numerical experiments to verify our theoretical results of all previous chapters. We consider both, integration problems and the plasma turbulence simulation GENE. For them we discuss the behavior of surpluses and problems that arise when applying the combination technique to quantities of interest with GENE.

2 Extrapolation

We start with a chapter on extrapolation methods. In this chapter we first deduce the Euler-Maclaurin formula, which leads to an error expansion for the trapezoidal rule. This error expansion is not only for the Romberg method but is also important to understand the combination technique and its error in chapter 3. The Romberg method will then be used in chapter 4 to introduce a combination technique using extrapolated results.

But first we consider an easy example to explain the fundamental idea of extrapolation methods.

Assume a numerical method to compute a quantity of interest $q(h)$ with a parameter h, which controls the accuracy of the method. Often this parameter is the grid resolution and controls how fine the grid used by the method is. Further, assume that the error $e(h) = q(h) - q$ is known to be of the form

$$e(h) \coloneqq c_2 h^2 + c_4 h^4 + c_6 h^6$$

with unknown coefficients c_2,c_4 and c_6, which are all independent of h. We clearly see that the method is of second order as the lowest order terms is h^2. For a known solution $q(h_1)$ the error is given by $e(h_1)$. However, assume that this error is known to be too large and we are interested in a better approximation. Choosing $h_2 < h_1$ clearly gives a better approximation. Reducing h_1 to one half should decrease the error by one fourth. But we can do better than this if we use two solutions for different $h_1 \neq h_2$. For each h_i we get a solution with a different error

$$\begin{aligned} q(h_1) &= q + e(h_1) = q + c_2 h_1^2 + c_4 h_1^4 + c_6 h_1^6 \\ q(h_2) &= q + e(h_2) = q + c_2 h_2^2 + c_4 h_2^4 + c_6 h_2^6 \end{aligned}$$

with the same three unknown constants for each solution. This actually looks like a system of equations for c_i. As the system is undetermined we cannot solve it for the coefficients, but we know enough to cancel out one of the terms. Canceling out the second order term should be the best choice, as for small values of h we can assume that $h^2 \ll h^4$. This increases the order of the method from second order to fourth

order. Multiplying the the second equation by $\alpha^{-2} = {}^{h_1^2}/_{h_2^2}$, subtracting the results and substituting α gives

$$\begin{aligned} q(h_1) - \frac{h_1^2}{h_2^2} q(h_2) = &\left(1 - \frac{h_1^2}{h_2^2}\right) q + c_4 \left(1 - \frac{h_1^2}{h_2^2}\right) \left(h_1^4 - h_2^4\right) \\ &+ c_4 \left(1 - \frac{h_1^2}{h_2^2}\right) \left(h_1^6 - h_2^6\right) \\ = &\left(1 - \alpha^{-2}\right) q + \left(1 - \alpha^{-2}\right) \left(1 - \alpha^4\right) c_4 h_1^4 \\ &+ \left(1 - \alpha^{-2}\right) \left(1 - \alpha^6\right) c_6 h_1^6 \end{aligned}$$

in which only the fourth and sixth order terms remain. This already looks promising, however, as we can see by the factor in front of q, we no longer approximate the exact solution. This is easily fixed by multiplying the equation by $(1 - \alpha^2)^{-1}$ and we have

$$\frac{q(h_1) - \alpha^{-2} q(\alpha h_1)}{1 - \alpha^{-2}} = q + \frac{1 - \alpha^4}{1 - \alpha^{-2}} c_4 h_1^4 + \frac{1 - \alpha^6}{1 - \alpha^{-2}} c_6 h_1^6,$$

which now approximates q correctly. This means that we are able to get a higher (fourth) order approximation for q by using a linear combination of two lower (second) order approximations. We can now compare this new approximation to our initial second order approximation. For small values of h_1 and $\alpha = {}^1/_2$, which is $h_2 = {}^1/_2 h_1$, we can approximate the error to be $e(h_1) = -{}^5/_{16} c_4 h_1^4 + \mathcal{O}(h_1^6)$. The error without the extrapolation would be $e(h) = c_2 h^2 + \mathcal{O}(h^4)$. This shows that we have to choose $h = \sqrt{{}^{c_4}/_{c_2}} h_1^2$ to get an error similar to the extrapolation method. A comparison of grids needed for the extrapolation and the non-extrapolated method shows that the non-extrapolated method needs a much finer grid: For the extrapolation method we need two solutions with $h = h_1$ and $h = {}^1/_2 h_1$, while we need only one solution with $h = \sqrt{{}^{c_4}/_{c_2}} h_1^2$ for the non-extrapolated method to obtain a similar error. It should become clear that for small errors and therefore for small values of h_1 the non-extrapolated method needs a much finer grid to get an error of the same order as the extrapolated method. As the computational complexity usually increases with at least h^{-1} this implies that using the extrapolated method for small values of h yields better results for the same computational effort.

Quantities of interest like spacial, temporal or phase space averages and moments like the variance often contain some sort of integral. In the next chapters we therefore focus on calculating the integral $q = \int_0^1 f(x)\,\mathrm{d}x$, using the trapezoidal rule. In order to perform the extrapolation we first discuss how to get an error expansion for the trapezoidal rule, which leads to the Euler-Maclaurin formula. We then discuss how to use the error expansion to further increase the order of the method. We do this by using the Romberg method to approximate the integral using multiple trapezoidal rules to eliminate more error terms.

Eventually, we relate this method to the surpluses of the combination technique to increase its order and discuss how the extrapolation affects the surpluses.

2.1 The Euler-Maclaurin Formula*

In the introduction we showed how an error expansion can be used to cancel error terms. However, to apply extrapolation methods the error expansion has to be known. As we consider integration problems as sample problem, we first have to find an error expansion for the trapezoidal rule. We now deduce the Euler-Maclaurin formula, which gives us an error expansion for the trapezoidal rule.

We first assume that our quantity of interest is the one dimensional integral $q = I = \int_0^1 f(x)\,\mathrm{d}x$ for a given function f. To approximate the integral we use the trapezoidal rule

$$q(h) = T(h) = \frac{f(0)}{2}h + \sum_{i=1}^{n-1} f(ih)h + \frac{f(1)}{2}h$$

with n grid points $x_i = ih$ for $i = 0, \ldots, n$. As we have n intervals, the grid resolution h is given by $h = 1/n$.

Next we derive a one dimensional error expansion for smooth functions f and later discuss a formula for the two dimensional case.

2.1.1 The One Dimensional Case

We now derive the Euler-Maclaurin formula in one dimension. The following proof is based on [Lyn65] where the function f is assumed to be analytic. As this is rarely the case, we generalize the result for $p+1$ times differentiable functions $f \in C^{p+1}[-h, h]$.

As the trapezoidal rule is composed of many small trapezoids we first focus on the error of each trapezoid on it's interval. Afterwards we sum up all the errors to get an error for the trapezoidal rule on the entire interval.

The trapezoidal rule for a single trapezoid over the interval $[-h, h]$ is given by

$$T(h) = 2h\frac{f(-h) + f(h)}{2}.$$

*This chapter is part of the propaedeutic.

We can now use the Taylor expansion of f at $x_0 = 0$ to rewrite $f(-h)$ and $f(h)$ as

$$f(h) = f(0) + \sum_{k=1}^{p} \frac{f^{(k)}(0)}{k!} h^k + r(h)$$
$$f(-h) = f(0) + \sum_{k=1}^{p} (-1)^k \frac{f^{(k)}(0)}{k!} h^k + r(-h)$$

with the remainder term $r \in \mathcal{O}(h^{p+1})$ for $h \to 0$. We now substitute the trapezoidal rule and have

$$T(h) = 2h \left(f(0) + \sum_{k=1}^{p/2} \frac{f^{(2k)}(0)}{(2k)!} h^{2k} + \tilde{r}(h) \right) \tag{2.1}$$

with the new remainder term $\tilde{r}(h) = \frac{1}{2}(r(h) + r(-h)) \in \mathcal{O}(h^{p+1})$. Note that the sum only contains even exponents of h as the odd terms canceled out.

To calculate the error for the single trapezoid approximation, we now rewrite the exact integral in terms of a Taylor expansion. After substituting the expansion of f in in the integral, we can integrate the sum term term-wise:

$$\begin{aligned} I &= \int_{-h}^{h} f(x)\,\mathrm{d}x = \int_{-h}^{h} f(0) + \sum_{k=1}^{p} \frac{f^{(k)}(0)}{k!} x^k + r(x)\,\mathrm{d}x \\ &= 2h \left(f(0) + \sum_{k=1}^{p/2} \frac{f^{(2k)}(0)}{(2k+1)!} h^{2k} \right) + \int_{-h}^{h} r(x)\,\mathrm{d}x. \end{aligned} \tag{2.2}$$

Note that only the even exponents contribute to the integral, while the odd exponents vanish as we integrate over a symmetric interval. As the remainder term $r(x)$ is $\mathcal{O}(x^{p+1})$, the integral over the remainder from $-h$ to h is of order $\mathcal{O}(h^{p+2})$.

Substituting the the results for the trapezoid (2.1) and the integral (2.2) we can calculate the error as

$$\begin{aligned} e(h) &= T(h) - I \\ &= 2h \sum_{k=1}^{p/2} \left(\left(\frac{1}{(2k)!} - \frac{1}{(2k+1)!} \right) f^{(2k)}(0) h^{2k} \right) + \mathcal{O}\left(h^{p+2}\right) \end{aligned} \tag{2.3}$$

and see that the error only contains even powers of h, too.

Before we calculate the error for the composed trapezoidal rule, we first rewrite the derivatives in the equation above in terms of integrals over derivatives. This is important, as we don not only consider single trapezoids but the trapezoidal rule consisting of many trapezoids. In its current form the error for the trapezoidal rule would contain derivative terms in the center of each single sub-interval. This means that the number

of evaluations of derivative terms changes if we change h and therefore the number of sub-intervals, too. By rewriting the derivatives in terms of integrals over the sub-intervals, we can later combine all the integrals over the smaller intervals into bigger integrals over the entire interval. Changing h, or respectively the number of intervals used by the trapezoidal rule, still leads to more and smaller sub-intervals. However, the small integrals can always be combined into integrals over the entire interval and are therefore explicitly independent of h.

Similar to (2.2) we can rewrite the derivatives $f^{(k)}(0)$ using the Taylor expansion of f and it holds

$$\int_{-h}^{h} f^{(k)}(x)\,\mathrm{d}x = 2h\left(f^{(k)}(0) + \sum_{\substack{i=1\\ \text{even}}}^{p-k} \frac{f^{(i+k)}(0)}{(i+1)!}h^i\right) + \int_{-h}^{h} r_k(x)\,\mathrm{d}x$$

with a remainder term $r_k(x) \in \mathcal{O}\left(x^{p-k+1}\right)$. For $k = 2$ we can substitute $f^{(2)}(0)$ in (2.3) and have

$$\begin{aligned} T(h) - I = {} & h^2\left(\frac{1}{2!} - \frac{1}{3!}\right)\int_{-h}^{h} f^{(2)}(x)\,\mathrm{d}x \\ & + 2h\sum_{k=2}^{p/2}\left(\left(\frac{1}{(2k+1)!} - \frac{1}{(2k)!} - \frac{1}{(2k+1)!}\left(\frac{1}{2!} - \frac{1}{3!}\right)\right) f^{(2k)}(0)h^{2k}\right) + \mathcal{O}\left(h^{p+2}\right). \end{aligned}$$

We now repeat this step for all higher derivatives, where each derivative only adds terms to higher order terms. This process leads to the one dimensional error expansion for the trapezoid in integral form

$$I(f) - T(f,h) = \sum_{k=1}^{p/2} c_{2k}h^{2k}\int_{-h}^{h} f^{(2k)}(x)\,\mathrm{d}x + \mathcal{O}\left(h^{p+2}\right) \tag{2.4}$$

for some coefficients $c_{2k} \in \mathbb{R}$. These coefficients do depend on neither f or h. Further information on how to calculate these coefficients can be found in [Lyn65]. For our purpose the exact value is of no further importance.

2.1.2 The Two Dimensional Case

The steps for higher dimensions are similar to the steps for the one dimensional error expansion.

Similar to our previous procedure we assume $f \in C^{p+1}$ and use the multidimensional Taylor expansion for f

$$f(x,y) = f(0,0) + \sum_{i=0}^{1} \frac{\partial f}{\partial x_i} x_i + \frac{1}{2!} \sum_{i=0}^{1} \sum_{j=0}^{1} \frac{\partial^2 f}{\partial x_i \partial x_j} x_i x_j + \frac{1}{3!} \sum_{i=0}^{1} \sum_{j=0}^{1} \sum_{k=0}^{1} \frac{\partial^3 f}{\partial x_i \partial x_j \partial x_k} x_i x_j x_k + \dots$$

and for its integral

$$\begin{aligned} I(f) &= \int_{-h_y}^{h_y} \int_{-h_x}^{h_x} f(x,y)\,\mathrm{d}x\mathrm{d}y \\ &= 4h_x h_y \Bigg(f(0,0) + \frac{1}{3!}\frac{\partial^2 f}{\partial x^2} h_x^2 + \frac{1}{3!}\frac{\partial^2 f}{\partial y^2} h_y^2 \\ &\quad + \frac{1}{5!}\frac{\partial^4 f}{\partial x^4} h_x^4 + \frac{1}{9}\frac{4}{4!}\frac{\partial^4 f}{\partial x^2 \partial y^2} h_x^2 h_y^2 + \frac{1}{5!}\frac{\partial^4 f}{\partial y^4} h_y^4 + \dots \Bigg) \end{aligned}$$

where all the derivatives are evaluated at $(x,y)^T = (0,0)^T$ to make the formula more readable. As f is smooth and therefore its derivatives are continuous, the Schwarz' theorem can be applied to swap the order of differentiation for the mixed derivatives.

Again, note that all Taylor terms containing an odd power of either x or y did not contribute to the integral.

Rewriting the trapezoid rule in terms of the Taylor expansion gives

$$\begin{aligned} T(h_x, h_y, f) &= 4h_x h_y \frac{f(-h_x,-h_y) + f(-h_x,h_y) + f(h_x,-h_y) + f(h_x,h_y)}{4} \\ &= 4h_x h_y \Bigg(f(0,0) + \frac{1}{2!}\frac{\partial^2 f}{\partial x^2} h_x^2 + \frac{1}{2!}\frac{\partial^2 f}{\partial y^2} h_y^2 \\ &\quad + \frac{1}{4!}\frac{\partial^4 f}{\partial x^4} h_x^4 + \frac{6}{4!}\frac{\partial^4 f}{\partial x^2 \partial y^2} h_x^2 h_y^2 + \frac{1}{4!}\frac{\partial^4 f}{\partial y^4} h_y^4 + \dots \Bigg) \end{aligned}$$

in which all the Taylor terms with an odd exponent in either h_x or h_y cancel out.

The two dimensional error expansion for an individual trapezoid can then be calculated as

$$\begin{aligned} &I(f) - T(h_x, h_y, f) = \\ &4h_x h_y \Bigg(\left(\frac{1}{3!} - \frac{1}{2!}\right)\frac{\partial^2 f}{\partial x^2} h_x^2 + \left(\frac{1}{3!} - \frac{1}{2!}\right)\frac{\partial^2 f}{\partial y^2} h_y^2 + \left(\frac{1}{5!} - \frac{1}{4!}\right)\frac{\partial^4 f}{\partial x^4} h_x^4 \\ &\qquad + \left(\frac{1}{9}\frac{4}{4!} - \frac{6}{4!}\right)\frac{\partial^4 f}{\partial x^2 \partial y^2} h_x^2 h_y^2 + \left(\frac{1}{5!} - \frac{1}{4!}\right)\frac{\partial^4 f}{\partial y^4} h_y^4 + \dots \Bigg). \end{aligned}$$

which is very similar to the one dimensional case.

Again, similar to the one dimensional case, we substitute the derivatives for the Taylor expansion of the integrated derivative to get the two dimensional error expansion

$$I(f) - T(h_x, h_y, f) = \sum_{\substack{i=2 \\ \text{even}}}^{p} \sum_{\substack{k=0 \\ \text{even}}}^{i} c_{k,i-k} h_x^k h_y^{i-k} \int_{-h_y}^{h_y} \int_{-h_x}^{h_x} \frac{\partial^k \partial^{i-k}}{\partial x^k \partial y^{i-k}} f(x,y) \,\mathrm{d}x\mathrm{d}y + R \quad (2.5)$$

with a high order remainder term R.

2.1.3 Error Expansion for the Trapezoidal Rule

We now use the error expansions we deduced for single trapezoid approximations to get an error expansion for the trapezoidal rule. As the trapezoidal rule only consists of subdividing the interval and applying the single trapezoid approximation on each of the subintervals, all we have to do is to sum up the errors done on each subinterval:

As before we consider a function $f \in C^{p+1}[0,1]$. Let I_i and T_i be the integral and the trapezoid approximation over the i^{th} interval $[2ih, 2(i+1)h]$. The error for the trapezoid rule is then given by by the sum of the errors on each interval

$$I - T(h) = \sum_{i=0}^{n-1} (I_i - T_i(h)).$$

Furthermore, let f_i be the transformed function $f_i(x) \coloneqq f(x + 2ih + h)$, which maps the center of the i^{th} interval to the origin. Now, we rewrite the error on each subinterval in terms of f_i and it holds

$$\begin{aligned} I_i - T_i(h) &= \int_{2ih}^{2(i+1)h} f(x) \,\mathrm{d}x - 2h \frac{f(2ih) + f(2(i+1)h)}{2} \\ &= \int_{-h}^{h} f_i(x) \,\mathrm{d}x - 2h \frac{f_i(-h) + f_i(h)}{2}. \end{aligned}$$

Substituting our error expansion (2.4) we get

$$\begin{aligned} \ldots &= \sum_{k=1}^{p/2} c_{2k} h^{2k} \int_{-h}^{h} f_i^{(2k)}(x) \,\mathrm{d}x + \mathcal{O}\Big(h^{p+2}\Big) \\ &= \sum_{k=1}^{p/2} c_{2k} h^{2k} \int_{2hi}^{2(i+1)h} f^{(2k)}(x) \,\mathrm{d}x + \mathcal{O}\Big(h^{p+2}\Big). \end{aligned}$$

Note that we found the coefficients c_k to be independent of the integrand and therefore independent of the interval i. We now substitute the error for the trapezoidal rule and obtain the *Euler-Maclaurin formula* for the trapezoidal rule

$$\begin{aligned} I - T(h) &= \sum_{i=0}^{n-1} \sum_{k=1}^{p/2} c_{2k} h^{2k} \int_{2hi}^{2(i+1)h} f^{(2k)}(x)\,\mathrm{d}x + \mathcal{O}\Big(h^{p+1}\Big) \\ &= \sum_{k=1}^{p/2} c_{2k} h^{2k} \int_{0}^{1} f^{(2k)}(x)\,\mathrm{d}x + \mathcal{O}\Big(h^{p+1}\Big)\,. \end{aligned} \tag{2.6}$$

We see, that the error expansion for the trapezoidal rule only contains even powers of h, too.

The same argument can be applied to the two dimensional case for a function $f \in C^{p+1,p+1}\left[0,1\right]^2$, which is $p+1$ times differentiable in each dimension. For a grid with the resolution h_x in the x-direction and resolution h_y in the y-direction we get the *Euler-Maclaurin formula*

$$I(f) - T(h_x, h_y, f) = \sum_{\substack{i=2 \\ \text{even}}}^{p} \sum_{\substack{k=0 \\ \text{even}}}^{i} c_{k,i-k} h_x^k h_y^{i-k} \int_{-b}^{b} \int_{-a}^{a} \frac{\partial^k \partial^{i-k}}{\partial x^k \partial y^{i-k}} f(x,y)\,\mathrm{d}x\mathrm{d}y + \mathcal{O}\left(h_x^{p+1} + h_y^{p+1}\right)$$

for the two dimensional trapezoidal rule.

2.2 Romberg Quadrature

In the previous section we deduced an error expansion for the one and two dimensional trapezoidal rule $T(h)$, which showed that the expansions only consists of even powers of h. We will now use this knowledge to define the Romberg quadrature [DR06]. The idea of the Romberg quadrature is to repeatedly apply the extrapolation scheme we discussed in the last chapter, to cancel more and more error terms of the error expansion for the trapezoidal rule. This leads to an sequence of quadrature rules of increasing order.

2.2.1 How it Works

As shown in (2.6) the error expansion for the trapezoidal rule $T(h)$ for a function $f \in C^{2p+2}$ only contains even powers of h. With the Euler-Maclaurin formula (2.6) we found

$$T(h) - I = \sum_{k=1}^{p} c_{2k} h^{2k} + \mathcal{O}\Big(h^{2p+2}\Big)$$

for some coefficients c_k, which are independent of h. We now follow the same steps as in the introduction, but repeat them multiple times in order to cancel more than just one error term.

Suppose two grids with a different grid widths $h_0 = h$ and $h_1 = \alpha h$. The order of the quadrature can then be increased by combining the two error expansions

$$\begin{aligned} T(h_0) - I(f) &= c_2 h_0^2 + c_4 h_0^4 + \cdots + c_{2p} h_0^{2p} + \mathcal{O}\left(h_0^{2p+2}\right) \\ T(h_1) - I(f) &= c_2 h_1^2 + c_4 h_1^4 + \cdots + c_{2p} h_1^{2p} + \mathcal{O}\left(h_1^{2p+2}\right) \\ &= c_2 \alpha^2 h_0^2 + c_4 \alpha^4 h_0^4 + \cdots + c_{2p} \alpha^{2p} h_0^{2p} + \mathcal{O}\left(h_0^{2p+2}\right) \end{aligned}$$

such that the h^2 term gets eliminated. As the coefficient c_2 is the same for both quadratures (they are independent of h), the exact value is no special importance. Eliminating the h_0^2 term as before gives

$$\frac{T(h_0) - \alpha^{-2} T(h_1)}{1 - \alpha^{-2}} - I(f) = \frac{1 - \alpha^2}{1 - \alpha^{-2}} c_4 h_0^4 + \cdots + \frac{1 - \alpha^{2p-2}}{1 - \alpha^{-2}} c_{2p} h_0^{2p} + \mathcal{O}\left(h_0^{2p+2}\right).$$

Note the increased order of the new quadrature rule $T_0^{(1)} = T^{(1)}(h_0)$. For $h_1 = {}^1\!/_2 h_0$ the new quadrature rule is the Simpson rule written as the linear combination of two trapezoidal rules.

Further error terms can be eliminated by adding more trapezoidal rules. Using $h_2 = \alpha h_1 = \alpha^2 h_0$ and the corresponding trapezoidal rule, the same steps can be repeated with h_1 and h_2 and we get two quadrature rules of fourth order:

$$\begin{aligned} T^{(1)}(h_0) &= \frac{T(h_0) - \alpha^{-2} T(h_1)}{1 - \alpha^{-2}} - I(f) \\ &= \frac{1 - \alpha^2}{1 - \alpha^{-2}} c_4 h_0^4 + \cdots + \frac{1 - \alpha^{2p-2}}{1 - \alpha^{-2}} c_{2p} h_0^{2p} + \mathcal{O}\left(h_0^{2p+2}\right) \\ T^{(1)}(h_1) &= \frac{T(h_1) - \alpha^{-2} T(h_2)}{1 - \alpha^{-2}} - I(f) \\ &= \frac{1 - \alpha^2}{1 - \alpha^{-2}} c_4 h_1^4 + \cdots + \frac{1 - \alpha^{2p-2}}{1 - \alpha^{-2}} c_{2p} h_1^{2p} + \mathcal{O}\left(h_1^{2p+2}\right) \end{aligned}$$

We see that for both quadrature rules the coefficients are the same. We now substitute $h_1 = \alpha h_0$ and eliminating the h_0^4 term as before and get a new quadrature rule

$$T_0^{(2)} = \frac{T^{(1)}(h_0) - \alpha^{-2} T^{(1)}(h_1)}{1 - \alpha^{-2}},$$

which is of sixth order. Note that this new quadrature rule with $\alpha = {}^1\!/_2$ is equivalent to Boole's rule.

This procedure can be continued for every new grid resolution where each new grid resolution cancels one more error term in the Euler-Maclaurin formula. For each new grid $h_{i+1} = \alpha h_i$ we calculate

$$T^{(1)}(h_i) = \frac{T(h_{i+1}) - \alpha^{-2} T(h_i)}{1 - \alpha^{-2}} \tag{2.7}$$

and then repeat this procedure for each pair $T^{(k)}(h_i), T^{(k)}(h_{i+1})$, which cancels one more term of the error expansion. Each new grid therefore allows to cancel one further term. Note that this can only done until all terms are eliminated. This means that for smoother functions f more error terms can be eliminated.

A nice property of the Romberg quadrature is that it can be applied on the (scalar) results of different trapezoidal rules as a post-processing step. This can be done without changing the way the trapezoidal rules are evaluated, as long as the integrand is smooth enough.

If hierarchical trapezoidal rules with $h_i = 2^{-i}h$ are used, the computation of the trapezoidal rules can be optimized further. In this case the grid points for the different grid widths h_i are nested, which means that function evaluations can be reduced by reusing potentially expensive evaluations using evaluations form previous grids. Each new trapezoidal rule can be written using the result from the previous trapezoidal rule

$$T(h_i) = h_i \left(\frac{1}{2} f(0) + \sum_{i=1}^{n_i - 2} f(i h_i) + \frac{1}{2} f(0) \right)$$

such that the new trapezoidal rule only has to take those grid point into account that are not contained in the previous grid:

$$\begin{aligned} T(\frac{h_i}{2}) &= \frac{h_i}{2} \left(\frac{1}{2} f(0) + \sum_{j=1}^{2n_i - 3} f(i \frac{h_i}{2}) + \frac{1}{2} f(0) \right) \\ &= \frac{1}{2} T(h_i) + \sum_{j=0}^{n_i - 1} f((1 + 2i) \frac{h_i}{2}) \\ &= \frac{1}{2} T(h_i) + \sum_{j=0}^{n_i - 1} f(\frac{h_i}{2} + j h_j). \end{aligned}$$

This reduces the computational costs for each trapezoidal rule to about one half.

To do the Romberg steps on N trapezoidal rules $T_i^{(0)} = T(h_i)$ with $i = 0, \ldots, N-1$ one can then use the iteration formula

$$T_i^{(n+1)}(h) = \frac{T_i^{(n)}(h) - 4^{n+1} T_{i+1}^{(n)}(h)}{1 - 4^{n+1}} \tag{2.8}$$

for $n = 0, \ldots, N-1$ and $i = 0, \ldots, N-1-n$. This iteration leads to the following pattern:

$$
\begin{array}{ccccccc}
T_0 & & & & & & \\
 & \searrow & & & & & \\
T_1 & \rightarrow & T_0^{(1)} & & & & \\
 & \searrow & & \searrow & & & \\
T_2 & \rightarrow & T_1^{(1)} & \rightarrow & T_0^{(2)} & & \\
 & \searrow & & \searrow & & \searrow & \\
T_3 & \rightarrow & T_2^{(1)} & \rightarrow & T_1^{(2)} & \rightarrow & T_0^{(3)}
\end{array}
$$

This extrapolation scheme is obviously not tied to quadrature problems, but can be used with any quantity of interest with as similar error expansion.

2.2.2 Tensor Romberg Method in Two Dimensions

In two dimensions we use a tensor product version of the Romberg method introduced in the previous section. This means that we will first apply the one dimensional Romberg method along the first axis and then apply the Romberg method a second time on the result along the second axis.

To be more specific, suppose the solution for the two dimensional trapezoidal rule $T(h_i, h_j)$ for different grid resolutions in each dimension.

First, we apply the Romberg method along the first dimension. This means that for each h_j in the y dimension we consider $T_j(h_i) \coloneqq T(h_i, h_j)$ with a fixed value of j. On $T_j(h_i)$ we now apply the Romberg method (2.8) using different values for h_i. The grid resolution h_j therefore acts like a constant. This leads to the extrapolated results $T_j^{(i)}(h_i)$ of order i in the x dimension, which is the extrapolation using the grid resolutions $h_0, \ldots, h_i$. For each h_j in y direction we now have a sequence of extrapolation results $T_j(h_0, \ldots, h_i)$ where larger values of i cancel more and more error terms in the x dimension.

As a second step we now repeat this procedure along the other dimension. For each h_i in x dimension we now fix i and consider $T^{(i)}(h_j) \coloneqq T_j(h_0, \ldots, h_i)$. Now i acts as a constant. Similar to before we apply the one dimensional Romberg method on $T^{(i)}(h_j)$ and get $T^{(i)}(h_0, \ldots, h_j)$ using the grid resolutions $h_0, \ldots, h_j$. Larger values of j now cancel more error terms in the y dimension.

This new extrapolated result $T^{(i,j)} = T(h_0, \ldots, h_i, h_0, \ldots, h_j) \coloneqq T^{(i)}(h_0, \ldots, h_j)$ depends on all results with grid resolutions (h_k, h_l) with $k = 0, \ldots, i$ and $l = 0, \ldots, j$ and was extrapolated i times in the x dimension and j times in the y dimension.

The reader should be sufficiently confused by now. To illustrate the method we consider the following example step by step: Assume the trapezoidal rue with the error expansion

$$T(h_i, h_j) = I + c_{2,0}h_i^2 + c_{0,2}h_j^2 + c_{4,0}h_i^4 + c_{2,2}h_i^2h_j^2 + c_{0,4}h_j^4$$

where we omitted a remainder terms, for the sake of clarity. To apply the Romberg method in the first dimension we fix j and consider $T_j(h_i) \coloneqq T(h_i, h_j)$. We now apply the Romberg method (2.8) on $T_j(h_i)$, which cancels error terms in the first dimension: For the zeroth Romberg step in the first dimension we do nothing at all and have

$$T_j(h_0) = I + c_{2,0}h_0^2 + c_{0,2}h_j^2 + c_{4,0}h_0^4 + c_{2,2}h_0^2h_j^2 + c_{0,4}h_j^4.$$

The next Romberg step in the first dimension cancels all the h_i^2 terms:

$$T_j(h_0, h_1) = I + c_{0,2}h_j^2 + \frac{1-\alpha^2}{1-\alpha^{-2}}c_{4,0}h_0^4 + c_{0,4}h_j^4.$$

Remember that h_j can be considered as a constant and thus the mixed term $h_i^2h_j^2$ cancels out, too. The next Romberg step in the first dimension cancels all the h_i^4 terms and we have

$$T_j(h_0, h_1, h_2) = I + c_{0,2}h_j^2 + c_{0,4}h_j^4.$$

For each j we now have different $T_j(h_0, \dots, h_i)$, which differ in the terms in the first dimension but still contain all terms h_j in the second dimension. We now aim to cancel some of the remaining h_j terms in the second dimension. Therefore we now consider $T^{(i)}(h_j) \coloneqq T_j(h_0, \dots, h_i)$. Analogous to the steps before we now apply the Romberg method for different values of i. For $i = 0$ we get the following series

$$\begin{aligned}
T^{(0,0)} = T^{(0)}(h_0) &= I + c_{2,0}h_0^2 + c_{0,2}h_0^2 + c_{4,0}h_0^4 + c_{2,2}h_0^2h_0^2 + c_{0,4}h_0^4 \\
T^{(0,1)} = T^{(0)}(h_0, h_1) &= I + c_{2,0}h_0^2 + c_{4,0}h_0^4 + \frac{1-\alpha^2}{1-\alpha^{-2}}c_{0,4}h_0^4 \\
T^{(0,2)} = T^{(0)}(h_0, h_1, h_2) &= I + c_{2,0}h_0^2 + c_{4,0}h_0^4
\end{aligned}$$

where we cancel an increasing number of error terms in the second dimension. For $i = 1$ we have

$$\begin{aligned}
T^{(1,0)} = T^{(1)}(h_0) &= I + c_{0,2}h_0^2 + \frac{1-\alpha^2}{1-\alpha^{-2}}c_{4,0}h_0^4 + c_{0,4}h_0^4 \\
T^{(1,1)} = T^{(1)}(h_0, h_1) &= I + \frac{1-\alpha^2}{1-\alpha^{-2}}c_{4,0}h_0^4 + \frac{1-\alpha^2}{1-\alpha^{-2}}c_{0,4}h_0^4 \\
T^{(1,2)} = T^{(1)}(h_0, h_1, h_2) &= I + \frac{1-\alpha^2}{1-\alpha^{-2}}c_{4,0}h_0^4.
\end{aligned}$$

Note that due to the error cancellation in the first dimension, this result does not contain any h_i^2 terms. And for $i = 2$ we have

$$\begin{aligned} T^{(2,0)} &= T^{(2)}(h_0) = I + c_{0,2}h_0^2 + c_{0,4}h_0^4 \\ T^{(2,1)} &= T^{(2)}(h_0, h_1) = I + \frac{1-\alpha^2}{1-\alpha^{-2}} c_{0,4}h_0^4 \\ T^{(2,2)} &= T^{(2)}(h_0, h_1, h_2) = I. \end{aligned}$$

The important point is that in general $T^{(i,j)}$ will only contain error terms that are of higher order than $2i$ in the first dimension and of higher order than $2j$ in the second dimension. The highest order extrapolation in this example is $T^{(2,2)}$, which was extrapolated twice in each dimension. Hence, $T^{(2,2)}$ is the most accurate result.

A big advantage of this method is that it goes along very nicely with the combination technique discussed in the next section. Later we discuss how we can combine this method with the combination technique to improve the error and approximate a full tensor Romberg solution $T^{(n,n)}$.

3 Combination Technique

In this chapter we discuss the combination technique [GSZ92] based on surpluses as introduced in [HHK+16]. But first we consider a basic, one dimensional example to to explain the idea behind the combination technique. Even though the one dimensional case does not have any real applications it is well suited to explain the general idea of the method.

Assume a quantity of interest q_i depending on i, which controls the accuracy of the numerical method used to calculate the value. Furthermore, assume that the sequence q_i converges to the exact solution q for $i \to \infty$. The error for each i is then given by $e_i = q_i - q$ or $q_i = q + e_i$. As q_i converges to q the error converges to zero for $i \to \infty$.

Consider the series $q_i = q + c2^{-2i}$. In this case the error is given by $e_i = c2^{-2i}$ with an unknown coefficient c. As the exact solution q is unknown we want to choose i such that q_i is a good approximation of q. Choosing i such that $|e_i| \leq \varepsilon$ clearly gives a good approximation, but as c is not known this cannot be solved for i. Instead of $e_i = q_i - q$, the difference

$$w_{i+1} \coloneqq q_{i+1} - q_i$$

might be a good surrogate. For our example it holds $w_i = c2^{-2i}\left(\frac{1}{4} - 1\right) = -\frac{3}{4}c2^{-2i} = -\frac{3}{4}e_i$. Choosing i to satisfy $|w_i| \leq \varepsilon$ implies $|e_i| \leq \frac{4}{3}w_i$. This means even without knowing the exact error we are able give a bound on the error that is based on the difference $w_{i+1} = q_{i+1} - q_i$.

This motivates to focus more on the differences w_i, which we will refer to as *surpluses* from now on. Instead of looking at the sequence q_i we now look at the series $q_i^c = \sum_{k=0}^{i} w_k$ and truncate the sum depending on the value of the surpluses. If all truncated surpluses w_k with $k > i$ are small and decay fast enough, their contribution to the solution q_i^c is rather small. The error e_i is then given by the truncated part of the series: $e_i = \sum_{k=i+1}^{\infty} w_k$. In the case of our example this series becomes a geometric series and we can calculate the error as

$$e_i = \frac{3}{4}c\left(\frac{(1/4)^{i+1}}{1 - 1/4}\right).$$

In one dimension the truncated series $q_i^c = \sum_{k=0}^{i} w_i$ with $w_0 = q_0$ is equivalent to q_i itself and the combination technique does not add any new advantages or seems to be particularly interesting. However, this should give an idea about what the combination technique is all about: truncating a series of surpluses $\sum_{k=0}^{i} w_k$ such that the error is small enough. If the surpluses decay fast enough this leads to a good approximation for the full grid solution.

We will now generalize this idea to higher dimensions, where the advantages of the combination technique become more apparent. After a more detailed definition of surpluses and an introduction to the combination technique, we discuss the relation between error and surpluses more deeply in section 3.2.2. We then analyze the error of the combination technique for specific error models in section 3.2.3.

3.1 Defining the Surpluses for Higher Dimensions

For higher dimensional problems the quantities of interest depend on multiple parameters $\vec{\mathbf{l}} \in \mathbb{N}_0^d$, which control the accuracy of each dimension. To simplify the following explanations we now restrict ourself to the two dimensional case where the quantity of interest $q_{i,j}$ only depends on two parameters $\vec{\mathbf{l}} = (i, j)$. Higher dimensions can be dealt with in a similar way.

In more than one dimension it is not immediately apparent how the surpluses should be defined. In one dimension we could just take the difference of a solution and its predecessor. However, in multiple dimensions each solution $q_{\vec{\mathbf{l}}}$ has multiple predecessors. Namely both $q_{i-1,j}$ and $q_{i,j-1}$ are predecessors of $q_{i,j}$ in the two dimensional case. Using the constraint that each solution $q_{\vec{\mathbf{l}}}$ contains all surpluses with a smaller multi-index, that is

$$q_{\vec{\mathbf{l}}} = \sum_{\vec{\mathbf{k}} \leq \vec{\mathbf{l}}} w_{\vec{\mathbf{k}}}, \tag{3.1}$$

one can calculate the surpluses. One possibility to calculate the surpluses is to solve the linear system of equations $\mathbf{q} = \underline{\mathbf{M}}\mathbf{w}$ for $\mathbf{w}$ where $\mathbf{q}$ and $\mathbf{w}$ are vectors containing all the quantities of interest $q_{\vec{\mathbf{l}}}$ and surpluses $w_{\vec{\mathbf{l}}}$. Each row of $\underline{\mathbf{M}}$ represents (3.1) for a different level $\vec{\mathbf{l}}$. The surpluses are then defined by

$$\mathbf{w} \coloneqq \underline{\mathbf{M}}^{-1}\mathbf{q}. \tag{3.2}$$

In two dimensions for the first 3 solutions in each dimensions this gives this nice looking system of equations

$$\begin{pmatrix} q_{0,0} \\ q_{0,1} \\ q_{0,2} \\ q_{1,0} \\ q_{1,1} \\ q_{1,2} \\ q_{2,0} \\ q_{2,1} \\ q_{2,2} \end{pmatrix} = \begin{pmatrix} 1 & & & & & & & & \\ 1 & 1 & & & & & & & \\ 1 & 1 & 1 & & & & & & \\ 1 & & & 1 & & & & & \\ 1 & 1 & & 1 & 1 & & & & \\ 1 & 1 & 1 & 1 & 1 & 1 & & & \\ 1 & & & 1 & & & 1 & & \\ 1 & 1 & & 1 & 1 & & 1 & 1 & \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} w_{0,0} \\ w_{0,1} \\ w_{0,2} \\ w_{1,0} \\ w_{1,1} \\ w_{1,2} \\ w_{2,0} \\ w_{2,1} \\ w_{2,2} \end{pmatrix}$$

and the solution

$$\begin{pmatrix} w_{0,0} \\ w_{0,1} \\ w_{0,2} \\ w_{1,0} \\ w_{1,1} \\ w_{1,2} \\ w_{2,0} \\ w_{2,1} \\ w_{2,2} \end{pmatrix} = \begin{pmatrix} +1 & & & & & & & & \\ -1 & +1 & & & & & & & \\ & -1 & +1 & & & & & & \\ -1 & & & +1 & & & & & \\ +1 & -1 & & -1 & +1 & & & & \\ & +1 & -1 & & -1 & +1 & & & \\ & & & -1 & & & +1 & & \\ & & & +1 & -1 & & -1 & +1 & \\ & & & & +1 & -1 & & -1 & +1 \end{pmatrix} \begin{pmatrix} q_{0,0} \\ q_{0,1} \\ q_{0,2} \\ q_{1,0} \\ q_{1,1} \\ q_{1,2} \\ q_{2,0} \\ q_{2,1} \\ q_{2,2} \end{pmatrix}$$

where it becomes apparent how the surpluses have to be calculated [HHK+16]. We see that we have

$$w_{i,j} = \begin{cases} q_{i,j} - q_{i-1,j} - q_{i,j-1} + q_{i-1,j-1} & i,j \neq 0 \\ q_{0,j} - q_{0,j-1} & i = 0, j \neq 0 \\ q_{i,0} - q_{i-1,0} & i \neq 0, j = 0 \\ q_{0,0} & i,j = 0 \end{cases} \tag{3.3}$$

for the surpluses in two dimensions. This also holds if one considers more than just the first nine surpluses. The procedure to find the surpluses for higher dimensions is very similar. First, one defines a new matrix $\underline{\mathbf{M}}$ that describes which surpluses each solution $q_{i,j}$ contains. This leads to a similar result for the surpluses by inverting the matrix.

3.2 How does the Combination Technique Work?

Now as we defined the surpluses for higher dimensions, we can explain the combination technique for higher dimensions. But first let us recapitulate what we got so far: We have a problem, which can be solved at different levels of accuracy and we defined surpluses, which describe how much the solution changes between different solutions. Furthermore we can write each quantity of interest as a sum of surpluses. In the one dimensional case we choose an index i based on the decay of the surpluses to truncate the series such that we skipped surpluses, which did not contribute enough to the final solution.

We now try to do something similar for higher dimensional problems: Based on the decay of the surpluses we only select those surpluses that contribute most to the final solution and skip surpluses that are smaller to reduce unnecessary computations.

First consider the two dimensional case with the surpluses

$$w_{i,j} = c_{i,j} 2^{-(i+j)}$$

where $|c_{i,j}| \leq M$ are unknown constants, which are bounded by $M \geq 0$. According to (3.1) the quantity of interest $q_{n,n}$ is then computed by $q_{n,n} = \sum_{k,l \leq n} w_{k,l} = \sum_{k=0}^{n} \sum_{l=0}^{n} w_{k,l}$. Looking at the decay of the surpluses we see that they decrease in in each dimension, and that all $w_{k,l}$ with the same value $n = k + l$ should have a similar value $|w_{k,l}| \leq M 2^{-n}$. To reduce unnecessary computations we skip small surpluses $|w_{k,l}| \leq M 2^{-n} < \varepsilon$. Hence we truncate all surpluses from our sum for which $n = k + l > -\log(\varepsilon / M)$ holds. This gives $q_n^c = \sum_{k+l \leq n} w_{k,l}$, which is the classical combination technique. As we only included a subset of the surpluses that make up $q_{n,n}$ we introduced an additional error, which we discuss in greater detail in section 3.2.3. The difference of the included surpluses is shown in figure 3.1.

In general, assuming that the surpluses decay fast for increasing $|\vec{\mathbf{l}}|$ and satisfy

$$\left| w_{\vec{\mathbf{l}}} \right| \leq M 2^{-\beta f(|\vec{\mathbf{l}}|)}$$

for a monotonically increasing function f and a constant M we truncate the multidimensional sum and define the *classical combination technique* as

$$q_n^c := \sum_{|\vec{\mathbf{l}}| \leq n} w_{\vec{\mathbf{l}}}, \tag{3.4}$$

which basically truncated the sum along a hyperplane. Here we assumed that $w_{\vec{\mathbf{l}}}$ decays at the same rate in each dimension (hence $f(|\vec{\mathbf{l}}|)$ in the exponent). If this is not the case and $w_{\vec{\mathbf{l}}}$ decreases at a different speed in different dimension, e.g. $w_{i,j} = c_{i,j} 2^{-(i+2j)}$, the

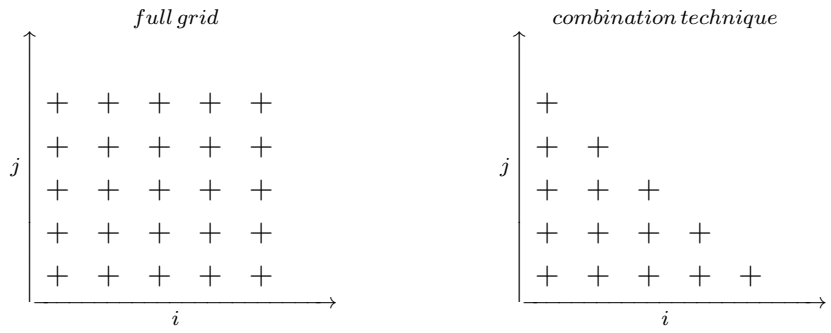

Figure 3.1: Surpluses for the full grid quantity of interest $q_{4,4}$ (left) and for the combined quantity of interest q_4^c. Plus marks surpluses included in the sum.

sum can be adjusted accordingly. For this example $\sum_{2i+j\leq n} w_{i,j}$ might be a better choice to balance the error in each dimension. We see that depending on the decay of surpluses other variations of the combination technique are feasible. In section 3.5 we give a short overview over other common variations.

3.2.1 Convergence of the Combination Technique

Now as we defined the classical combination technique it is important to show under which conditions this technique still converges. In order to make sure everything is well defined we assume that the series over the surpluses is absolutely convergent, which is

$$\sum_{\vec{\mathbf{l}}\in\mathbb{N}_0^d} |w_{\vec{\mathbf{l}}}| < \infty.$$

We now discuss, how fast the surpluses must decay such that the sum converges, and how it depends on the number of dimensions of the considered problem.

Let $q_n \coloneqq q_{\vec{\mathbf{n}}}$ with $\vec{\mathbf{n}} = n\vec{\mathbf{1}}$. The error e_n for the full grid solution

$$q_n = \sum_{\vec{\mathbf{l}}\leq\vec{\mathbf{n}}} w_{\vec{\mathbf{l}}}$$

is then given by all the surpluses that are not included in the sum:

$$e_n = q_n - q = \sum_{\vec{\mathbf{l}}>\vec{\mathbf{n}}} w_{\vec{\mathbf{l}}}.$$

Similarly, the error for the combination technique

$$q_n^c = \sum_{|\vec{\mathbf{l}}|\leq n} w_{\vec{\mathbf{l}}}$$

is then given by

$$e_n^c = q_n^c - q = \sum_{|\vec{\mathbf{l}}|>n} w_{\vec{\mathbf{l}}}.$$

Under the assumption of absolute convergence it becomes clear that if q_n converges to q, the combined solution q_n^c does so too and vice versa: One can just think of q_n^c as a reordering of the summed surpluses of q_n. The rearrangement theorem states that rearranging the order of the terms in an absolute converging series does not change the value of the series [AE07]. However, this does not give any information on how fast the error decreases for $n \to \infty$. We will discuss more explicit error bound in the next section.

Further assume that the surpluses are bound by a function only depending on on the level $|\vec{\mathbf{l}}|$: $w_{\vec{\mathbf{l}}} \leq Mf(|\vec{\mathbf{l}}|)$. In the next section 3.2.2 we deduce this function f for the case that there exists an error expansion for the quantity of interest.

We can now estimate the error for the combination technique as

$$e_n^c = \sum_{|\vec{\mathbf{l}}|>n} w_{\vec{\mathbf{l}}} \leq M \sum_{|\vec{\mathbf{l}}|>n} f(|\vec{\mathbf{l}}|).$$

On each level of surpluses with a constant value of $|\vec{\mathbf{l}}|$ different multi-indices contribute the same value $f(|\vec{\mathbf{l}}|)$ multiple times to the sum. Therefore we now count the number of indices on each level

$$N_l = \left|\left\{\vec{\mathbf{l}} \in \mathbb{N}_0^d \mid |\vec{\mathbf{l}}| = l\right\}\right|$$

and sum up all values level wise:

$$e_n^c \leq M \sum_{l=n+1}^{\infty} N_l f(l).$$

As the number of indices on each level in d dimensions has the order $N_l \in \mathcal{O}\left(l^{d-1}\right)$ for $n \to \infty$ the error converges to zero if and only if

$$\sum_{l=0}^{\infty} l^{d-1} f(l)$$

converges to a finite value. This directly implies that f has to converge to zero to satisfy this condition. Note that this explicitly depends on the dimensionality of the problem.

The higher dimensional the problem is, the faster the surpluses have to decay. This means that for a specific function f like $f(l) = 1/l^4$ the series converges for $d \leq 2$ but diverges for higher dimensional problems.

We now discuss different kinds of f for which the this sum converges, which implies the convergence of q_n^c.

Decay of Surpluses

Assuming a general function $f(k)$ there is no general rule to tell whether the series

$$\sum_{k=0}^{\infty} k^{d-1} f(k)$$

converges or not. This is due to the fact that there is no universal function one could compare the series against. However, we can provide some examples of functions f for which the convergence of $\sum_{k=1}^{\infty} k^{d-1} f(k)$ can be determined. For the sake of simplicity in the following discussion we will start the series from $k = 1$ to avoid cases with logarithms or fractions. There one has to add an offset to the indices in order to avoid problems for $k = 0$, which would only obfuscate the formulas.

We first consider the previous example for $f(k) = 1/k^p$. It follows that for $p > d$ the series converges and for $p \leq d$ the series diverges. Substituting f gives $\sum_{k=1}^{\infty} 1/k^{p-d+1}$. Comparing this to the well known series $\sum_{k=1}^{\infty} 1/n^{1+\varepsilon}$that converges for all $\varepsilon > 0$ and diverges for $\varepsilon = 0$ gives our proposition.

Now consider $f(k) = \mathrm{e}^{-\beta g(k)}$ for different functions g. It is clear that if $g(k)$ is a polynomial the series converges. Furthermore we have following proposition:

Theorem 3.1. *Consider the series $\sum_{k=1}^{\infty} k^{d-1}\mathrm{e}^{-\beta g(k)}$ for $d \geq 1$ and $\beta > 0$. For $g(k) = \sqrt[p]{k}$ the series converges for all $p > 0$ and $\beta > 0$. For $g(k) = \ln \sqrt[p]{k}$ the series converges for all $d < \beta/p$ and it diverges for $d \geq \beta/p$. For $g(k) = \ln \ln k$ the series diverges.*

Proof. The first property is shown by comparing $\mathrm{e}^{-\beta \sqrt[p]{k}} < \mathrm{e}^{-\gamma \ln k} = k^{-\gamma}$ for all $\gamma > 0$ and for large values of k. Choosing $\gamma > d$ and substituting the series gives $\sum_{k=1}^{\infty} k^{d-1}\mathrm{e}^{-\beta \sqrt[d]{k}} < \sum_{k=1}^{\infty} k^{d-1} 1/k^{\gamma} < \infty$, which means that the series converges for all $\beta > 0$.

We show the second property by comparing the series to $\sum_{k=1}^{\infty} 1/n^{1+\varepsilon}$ once again. Rewriting the series as $\sum_{k=1}^{\infty} k^{d-1}\mathrm{e}^{-\beta \ln \sqrt[p]{k}} = \sum_{k=1}^{\infty} k^{d-1} k^{-\beta/p} = \sum_{k=1}^{\infty} k^{d-1-\beta/p}$ we see that for $d < \beta/p$ the series converges and diverges otherwise.

The last property is shown by using $(\ln k)^{\beta} < k$ for all $\beta > 0$ and large values of k: $\sum_{k=1}^{\infty} k^{d-1}\mathrm{e}^{-\beta \ln \ln k} = \sum_{k=1}^{\infty} k^{d-1} (\ln k)^{-\beta} > \sum_{k=1}^{\infty} k^{d-2} > \sum_{k=1}^{\infty} k^{-1}$ as $d \geq 1$. □

These functions can be seen as some examples for functions whose convergence or divergence can be used to determine the behavior of other functions by comparing them: As the series diverges for $g(k) = \ln \ln k$ it also diverges for any slower growing g. As the series converges with $g(k) = \sqrt[p]{k}$ for all $\beta > 0$, it also converges for any faster growing g like polynomials.

3.2.2 Effect of an Error Expansion for q on the Surpluses

In the previous section 3.2.1 we assumed that the decay for the surpluses is known. Usually this is not the case. More often error bounds for the quantities of interest are known instead that depend on the method used to calculate the solutions.

We now consider the two dimensional combination technique for quantities of interest. Assuming that an error expansion for the quantity of interest

$$e_{i,j} = q_{i,j} - q = \sum_{k=0}^{s} \sum_{l=0}^{t} c_{k,l} h_{x,i}^{k} h_{y,j}^{l} + R_{i,j}$$

is known we now investigate how this error expansion relates to the surpluses $w_{i,j}$. Here we consider an error expansion containing all powers of the grid resolutions. In case of more specific expansions like the Euler-Maclaurin formula for the trapezoidal rule, the error expansion only contains even powers of the grid resolutions. In this case one can just set some of the coefficients to zero.

For $i \neq 0$ and $j \neq 0$ the $w_{i,j}$ is given by

$$w_{i,j} = q_{i,j} - q_{i-1,j} - q_{i,j-1} + q_{i-1,j-1}.$$

Substituting the error expansion all the q terms cancel out and we get

$$\begin{aligned}
w_{i,j} &= e_{i,j} - e_{i-1,j} - e_{i,j-1} + e_{i-1,j-1} \\
&= \sum_{k=0}^{s} \sum_{l=0}^{t} c_{k,l} h_{x,i}^{k} h_{y,j}^{l} + R_{i,j} - \sum_{k=0}^{s} \sum_{l=0}^{t} c_{k,l} h_{x,i-1}^{k} h_{y,j}^{l} - R_{i-1,j} \\
&\quad - \sum_{k=0}^{s} \sum_{l=0}^{t} c_{k,l} h_{x,i}^{k} h_{y,j-1}^{l} - R_{i,j-1} + \sum_{k=0}^{s} \sum_{l=0}^{t} c_{k,l} h_{x,i-1}^{k} h_{y,j-1}^{l} + R_{i-1,j-1} \\
&= \sum_{k=0}^{s} \sum_{l=0}^{t} c_{k,l} \left(h_{x,i}^{k} h_{y,j}^{l} - h_{x,i-1}^{k} h_{y,j}^{l} - h_{x,i}^{k} h_{y,j-1}^{l} + h_{x,i-1}^{k} h_{y,j-1}^{l} \right) + R'_{i,j}
\end{aligned}$$

where $R'_{i,j} = R_{i,j} - R_{i-1,j} - R_{i,j-1} + R_{i-1,j-1}$. By splitting the error expansion in mixed and non-mixed terms we can further simplify this expression into

$$\begin{aligned} w_{i,j} = &\sum_{k=0}^{s} c_{k,0} \left(h_{x,i}^k - h_{x,i-1}^k - h_{x,i}^k + h_{x,i-1}^k \right) \\ &+ \sum_{l=0}^{t} c_{0,l} \left(h_{y,j}^l - h_{y,j}^l - h_{y,j-1}^l + h_{y,j-1}^l \right) \\ &+ \sum_{k=1}^{s} \sum_{l=1}^{t} c_{k,l} \left(h_{x,i}^k h_{y,j}^l - h_{x,i-1}^k h_{y,j}^l - h_{x,i}^k h_{y,j-1}^l + h_{x,i-1}^k h_{y,j-1}^l \right) \\ &+ R'_{i,j} \\ = &\sum_{k=1}^{s} \sum_{l=1}^{t} c_{k,l} \left(h_{x,i}^k h_{y,j}^l - h_{x,i-1}^k h_{y,j}^l - h_{x,i}^k h_{y,j-1}^l + h_{x,i-1}^k h_{y,j-1}^l \right) \\ &+ R_{i,j} - R_{i-1,j} - R_{i,j-1} + R_{i-1,j-1} \end{aligned}$$

and see that only the mixed terms contribute to the expansion for the surpluses. Again, this only holds if $i \neq 0$ and $j \neq 0$. If either i or j is zero, non-mixed terms are still present in the expansion for the surplus $w_{i,j}$. For $i = 0$ and $j > 0$ it holds

$$\begin{aligned} w_{0,j} &= e_{0,j} - e_{0,j-1} \\ &= \sum_{k=0}^{s} \sum_{l=0}^{t} c_{k,l} h_{x,0}^k \left(h_{y,j}^l - h_{y,j-1}^l \right) + R_{0,j} - R_{0,j-1} \end{aligned}$$

and

$$w_{i,0} = \sum_{k=0}^{s} \sum_{l=0}^{t} c_{k,l} \left(h_{x,i}^k - h_{x,i-1}^k \right) h_{y,0}^l + R_{i,0} - R_{i-1,0}$$

and for $i = j = 0$ it holds

$$w_{0,0} = q_{0,0} = q + \sum_{k=0}^{s} \sum_{l=0}^{t} c_{k,l} h_{x,0}^k h_{y,0}^l + R_{0,0}.$$

However, these expansions result in a telescope sum when summing over i or j, which means that only the first and the last terms remain in $\sum_{j=0}^{n} w_{0,j}$ and $\sum_{i=0}^{n} w_{i,0}$ respectively.

It is important to note that the cancellation of non-mixed error terms for the surpluses cannot be directly applied to the combination technique. Surpluses with $i = 0$ or $j = 0$ still contain non-mixed terms, which do not cancel out when we combination technique is applied.

Furthermore, the cancellation of non-mixed terms also supports our previous assumption that the value of the surpluses are similar for all $w_{i,j}$ with the same value of $i + j$.

Assuming an error expansion of the form

$$e_{i,j} = q_{i,j} - q = \sum_{\substack{k=2 \\ \text{even}}}^{s} \sum_{\substack{l=0 \\ \text{even}}}^{k} c_{k,k-l} h_i^k h_j^{k-l} + R_{i,j}$$

with $h_i = 2^{-i}$ as we have for the trapezoidal rule the lowest order error term is

$$c_{2,2} h_i^2 h_j^2 = c_{2,2} 2^{-2i} 2^{-2j} = c_{2,2} 4^{-(i+j)}$$

and we see that the error only depends on the level $i + j$ in the lowest order.

3.2.3 Error of the Combination Technique

Now that we have a general combination technique q_n^c for a quantity of interest it is of major interest how large the error is and how it compares to q_n.

Based on the results from the previous chapter we now assume that with $h_i = 2^{-i}$ the surpluses satisfy

$$w_{i,j} = \sum_{k=2}^{s} \sum_{l=2}^{s} c_{k,l} 2^{-(ik+jl)} = c_{2,2} 4^{-(i+j)} + o\Big(4^{-(i+j)}\Big)$$

for $i \to \infty$ and $j \to \infty$. As we showed before only mixed terms remain in the expansion. Hence, we can estimate the surpluses with

$$|w_{i,j}| \leq M 4^{-(i+j)}$$

for a nonnegative constant M.

Theorem 3.2. *Consider surpluses $|w_{i,j}| \leq M 4^{-(i+j)}$ and the combination technique $q_n^c = \sum_{i+j \leq n} w_{i,j}$. Then it holds for the difference between the full grid solution $q_{n,n}$ and the combination technique q_n^c*

$$|q_{n,n} - q_n^c| \leq \frac{M}{3} 4^{-n} n. \tag{3.5}$$

Proof. We can calculate the additional error for the combination technique over $q_{n,n}$ by summing up all surpluses that are included in $q_{n,n}$ but not q_n^c:

$$|q_{n,n} - q_n^c| = \sum_{\substack{k+l>n \\ k,l \leq n}} w_{k,l} \leq M \sum_{\substack{k+l>n \\ k,l \leq n}} 4^{-(k+l)}.$$

We now sum up all the surpluses level-wise

$$\cdots = M \sum_{l=n+1}^{2n} (2n - l + 1)\, 4^{-l}$$

using that on each level $l = k + l > n$ there are $2n - l + 1$ surpluses with a similar value. Evaluating the geometric sum leads to

$$\ldots = \frac{M}{9} 4^{-n} \left(3n - 1 + 4^{-n}\right) \leq \frac{M}{3} 4^{-n} n.$$

Substituting $h_n = 2^{-n}$ then gives

$$\begin{aligned} |q_{n,n} - q_n^c| &\leq \frac{M}{9} h_n^2 \left(-3 \log_2 h_n - 1 + h_n\right) \\ &\leq -\frac{M}{3} h_n^2 \log_2 h_n \\ &= \frac{M}{3} h_n^2 \log_2 h_n^{-1}. \end{aligned}$$

□

We see that the error decreases quickly for increasing n. Note that this result is very similar to the result based on the traditional pointwise error splitting assumption for a solution in [GSZ92], which also leads to a pointwise convergence order of $\mathcal{O}\left(h_n^2 \log_2 h_n^{-1}\right)$. The traditional proof does not only work for the combination of functions, but can also be applied to quantities of interest with a similar error splitting assumption. However, we only focused on properties of surpluses to deduce similar error bounds. This surplus based approach can also be applied to the combination of solution functions: For each point x one defines the quantity of interest $q \coloneqq f(x)$ with the numerical approximations $q_{i,j} = f_{i,j}(x) \approx f(x)$ for which the pointwise error splitting holds

$$|q - q_{i,j}| = C_1(h_i) h_i^2 + C_2(h_j) h_j^2 + D(h_i, h_j) h_i^2 h_j^2$$

with the bounded functions C_1, C_2 and D. As the coefficient functions are bounded, we estimate them by M and get

$$|q - q_{i,j}| \leq M h_i^2 + M h_j^2 + M h_i^2 h_j^2.$$

Due to our previous result that surpluses only contain mixed error terms, we estimate the surpluses as

$$|w_{i,j}| \leq M h_i^2 h_j^2 = M 4^{-(i+j)}$$

for $h_i = 2^{-i}$ and $h_j = 2^{-j}$. We see that we can now apply our error estimation for the combination technique based on the surpluses. In contrast to the traditional approach this formula gives an estimation for the difference between the full grid solution and the combined solution $|q_{n,n} - q_n^c|$ in stead of the total error $|q - q_n^c|$. However, as the error $|q - q_{i,j}|$ is of order $\mathcal{O}\left(h_i^2 + h_j^2\right)$, which vanishes faster than $|q_{n,n} - q_n^c| \in \mathcal{O}(h_n^2 \log_2 h_n^{-1})$ for increasing grid resolutions the overall order for the error of the combined solution does not change.

Note that this error bound holds without any further assumptions. This is unlike existing convergence proofs for the combination of the solution (not the quantity of interest), which assume hierarchical function spaces for the sparse grid approximation [Gar12].

Up to now we assumed that surpluses decrease in both dimensions, which generally might not be the case. Therefore we now consider surpluses

$$|w_{i,j}| \leq M 2^{-(\alpha i + \beta j)}$$

which decrease at a non-uniform rate. With $\alpha = \beta = 2$ we essentially have our previous assumption. However, if $\alpha \neq \beta$ the level on which all surpluses have a similar value is no longer given by $i + j = const$. Instead the two coefficients move the plane, such that surpluses of similar value now satisfy $\alpha i + \beta j \approx const$.

3.2.4 Error of the Combination Technique with Spectral Expansions

So far we assumed that the error for both dimensions is of the form

$$e_i = \sum_{k=1}^{r} c_k h_i^k + R.$$

An example for which this is true is the trapezoidal rule to approximate an integral. In this case the error is given by the Euler-Maclaurin formula. However, there also exist other expansions like a Taylor series or a Fourier series, where the length of the series is varied instead of h. In case of a Taylor series this leads to an point-wise error of the form

$$e_i = -\sum_{k=i+1}^{r} c_k h^k - R$$

for a point $h = x - x_0$ with the coefficients $c_k = {}^{f^{(k)}(x_0)}/_{k!}$. Note hat even though the two expansions look very similar they behave very differently. In contrast to the previous sum, the index i now appears in the lower bound for the sum, while in the former sum, i controlled the size of h_i.

To investigate how the error of the combination technique behaves with these kind of expansions, we first calculate what the surpluses look like. In one dimensions we defined the surpluses w_i as

$$w_i = q_i - q_{i-1}$$

for $0 < i < r$. Substituting $q_i = q + e_i$ gives

$$\begin{aligned} w_i &= q - \sum_{k=i+1}^{r} c_k h^k - R - q + \sum_{k=i}^{r} c_k h^k + R \\ &= c_i h^i = \frac{f^{(i)}(x_0)}{i!} h^i \end{aligned}$$

which shows that the i^{th} surplus is just given by the i^{th} element of the Taylor expansion.

Now consider the case that one dimensions has an error expansion as before, but the second dimension has a Taylor-like expansion. Assume that the surpluses are given by

$$w_{i,j} = 4^{-i} \frac{h^j}{j!} c_{i,j}$$

for bounded coefficients $c_{i,j} \leq M$. An example for surpluses like this is the quantity of interest $q = \int_{-1}^{1} f(h, y)\,\mathrm{d}y$ for a fixed value of h, where the integral is approximated by a trapezoidal rule and $f(h, \cdot)$ is approximated by using a Taylor expansion. It should be clear that a finer grid for the trapezoidal rule results in a better approximation for the integral. This is the 4^{-i} term. But the further away from the expansion point we evaluate the integral, the worse the results become. This is the ${}^{h^j}/_{j!}$ term.

With these surpluses we get the following result.

Theorem 3.3. *Consider surpluses $|w_{i,j}| \leq M 4^{-i} \frac{h^j}{j!}$ and the combination technique $q_n^c = \sum_{i+j \leq n} w_{i,j}$. Then it holds for the difference between the full grid solution $q_{n,n}$ and the combination technique q_n^c*

$$|q_{n,n} - q_n^c| \leq \frac{M}{3} \mathrm{e}^{4h} 4^{-n} \tag{3.6}$$

and with $h_n = 2^{-n}$ it holds

$$|q_{n,n} - q_n^c| \leq \frac{M}{3} \mathrm{e}^{4h} h_n^2.$$

Proof. To calculate the additional error of the combination technique we sum up the surpluses which were truncated from q_n:

$$\begin{aligned} |q_{n,n} - q_n^c| &\leq \sum_{j=1}^{n} \sum_{i=n-j+1}^{n} w_{i,j} \\ &\leq M \sum_{j=1}^{n} \sum_{i=n-j+1}^{n} 4^{-i} \frac{h^j}{j!}. \end{aligned}$$

Using the geometric sum to calculate the sum over i we get

$$\begin{aligned}\ldots &= M \sum_{j=1}^{n} \frac{h^j}{j!} \frac{4^{-(n-j+1)} - 4^{-(n+1)}}{1 - 4^{-1}} = M \sum_{j=1}^{n} \frac{h^j}{j!} 4^{-n} \frac{4^j - 1}{4 - 1} \\ &= \frac{M}{3} 4^{-n} \sum_{j=1}^{n} \left(\frac{(4h)^j}{j!} - \frac{h^j}{j!} \right).\end{aligned}$$

We can estimate the first term in the sum by e^{4h}, as it is the first part of the power expansion $\mathrm{e}^x = 1 + \sum_{k=1}^{\infty} {}^{x^k}/_{k!}$. For the second part we just skip all but the last term of the sum with $j = n$. This leads to

$$\ldots \leq \frac{M}{3} 4^{-n} \left(\mathrm{e}^{4h} - 1 - \frac{h^n}{n!} \right) \leq \frac{M}{3} \mathrm{e}^{4h} 4^{-n}$$

and

$$\ldots = \frac{M}{3} \mathrm{e}^{4h} h_n^2$$

with $h_n = 2^{-n}$. □

Comparing this error bound to the error bound for exponential decay in both dimensions 3.5 we see that the additional error decreases faster: We now have the factor 4^{-n} instead of $n4^{-n}$. This is exactly the same order we assumed for the j-axis, which became the dominant error. This shows that the error in the i-dimension must decrease faster than the error along the j-axis.

We now adjust the combination technique to balance the error in both dimensions, such that no error dominates the error in the other dimension. In the previous chapter we assumed that the error decays with 4^{-i} and 4^{-j} in each dimension respectively. To balance the order of the errors we demanded

$$4^{-i} 4^{-j} = 4^{-n} = const$$

which leads to $i = n - j$ and therefore the classical combination technique

$$q_n^c = \sum_{i+j \leq n} w_{i,j} = \sum_{j=0}^{n} \sum_{i=0}^{n-j} w_{i,j}.$$

We now try to do something similar for $w_{i,j} = 4^{-i} {}^{h^j}/_{j!}$. To balance the error orders we demand

$$4^{-i} \frac{h^j}{j!} = 4^{-c} = const,$$

which leads to

$$i = j \log_4 h - \log_4 j! + c.$$

Approximating the logarithm of the factorial with Stirling's formula we have

$$i \approx j \log_4 h - j \log_4 j + c = j \log_4 \frac{h}{j} + c,$$

which describes a border for the surpluses. Surpluses below this line are included in the sum. Surpluses above this line are considered too small. The constant can be determined with the additional constraint that $i = n$ for $j = 0$, which leads to $c = n$. The adapted combination technique is then defined by

$$q_n^{c'} = \sum_{i=0}^{n} \sum_{j=0}^{\max\left\{0, j \log_4 \frac{h}{j} + n\right\}} w_{i,j}.$$

3.3 Evaluating the Combination Technique

The combination technique, as we defined in (3.4), is a sum of surpluses. This formula is important to understand the combination technique and methods like thresholding [HHK+16]. However, the combined solution can be computed differently, especially when using the classical combination technique. In this case the initial calculation of all surpluses can be avoided and the solution is evaluated directly as a linear combination on the solutions $q_{\vec{\mathbf{l}}}$.

Using the definitions from section 3.1 we defined the combination technique (3.4) as

$$q_n^c = \sum_{|\vec{\mathbf{l}}| \leq n} w_{\vec{\mathbf{l}}}.$$

We now introduce weights

$$c_{\vec{\mathbf{l}}}^{(w)} = \begin{cases} 1 & |\vec{\mathbf{l}}| \leq n \\ 0 & \text{otherwise} \end{cases}$$

masking elements that are included in the combination technique. The sum above is then equivalent to

$$q_n^c = \sum_{\vec{\mathbf{l}} \leq n \vec{\mathbf{1}}} c_{\vec{\mathbf{l}}}^{(w)} w_{\vec{\mathbf{l}}} = \mathbf{c}_w^T \mathbf{w}$$

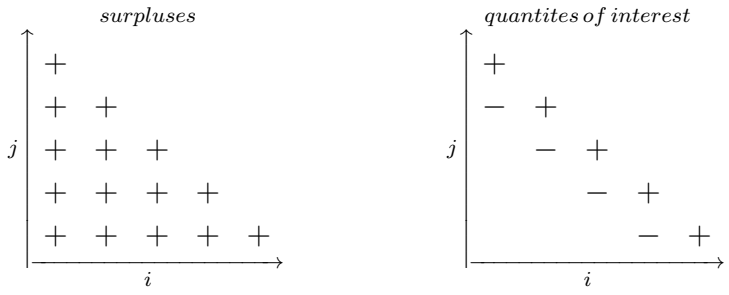

Figure 3.2: The coefficients for the combination technique for q_4^c. Left: Coefficients for the surpluses $w_{i,j}$. Right: Coefficients for the quantities of interest $q_{i,j}$. Plus stands for positive one, minus stands for minus one.

where $\mathbf{c}_w$ is the vector of $c_{\vec{\mathbf{l}}}^{(w)}$ and $\mathbf{w}$ the vector of $w_{\vec{\mathbf{l}}}$. Substituting the definition of the surpluses (3.2) we get

$$q_n^c = \mathbf{c}_w^T \mathbf{w} = \mathbf{c}_w^T \left(\underline{\mathbf{M}}^{-1} \mathbf{q} \right) = \left(\mathbf{c}_w^T \underline{\mathbf{M}}^{-1} \right) \mathbf{q}.$$

Thus the coefficients for the solutions $\mathbf{c}_q$ are therefore given by

$$\mathbf{c}_q^T = \mathbf{c}_w^T \underline{\mathbf{M}}^{-1}.$$

In two dimensions this leads to the weights

$$c_{i,j}^{(q)} = \begin{cases} 1 & i+j = n \\ -1 & i+j = n-1 \\ 0 & \text{otherwise} \end{cases}$$

for the classical combination technique, which results in the more common definition

$$q_n^c = \sum_{i,j \leq n} c_{i,j}^{(q)} q_{i,j} = \sum_{i+j=n} q_{i,j} - \sum_{i+j=n-1} q_{i,j} \tag{3.7}$$

as it is seen in literature [Gar12; GSZ92; Har16]. The emerging pattern for the weights is illustrated in figure 3.2. Summing up all the unknowns on each grids shows that the combination technique uses much less grid points than the full grid solution. While the full grid solution uses $\mathcal{O}(n^2)$ grid points, the combination technique only uses $\mathcal{O}(2n \log n)$ grid points spread over multiple grids [BG04; Pfl10].

For higher dimensional problems the combination technique is given by

$$q_{\vec{\mathbf{n}}}^c = \sum_{k=0}^{d-1} (-1)^k \binom{d-1}{k} \sum_{|\vec{\mathbf{l}}|=n-k} f_{\vec{\mathbf{l}}}$$

as given in [HGC07; Pfl10].

To compute the combination technique it is therefore not necessary to calculate all the surpluses first, as each solution is a sum of surpluses itself. It is sufficient to calculate only the solutions for the highest levels and combine them with the right coefficients. This avoids computations on smaller grids.

The same approach can also be used for more general combination techniques. Given a finite set of indices $I \subseteq \mathbb{N}_0^d$ we can define the combination technique as

$$q_I^c = \sum_{\vec{\mathbf{k}} \in I} w_{\vec{\mathbf{k}}}.$$

Defining the weights $c_{\vec{\mathbf{k}}}^{(w)} = \chi_i(\vec{\mathbf{k}})$ for each surplus by the characteristic function of the set of indices we can apply the same steps as before and get weights $\mathbf{c}_q^T = \mathbf{c}_w^T \underline{\mathbf{M}}^{-1}$ for the solutions such that

$$q_I^c = \sum_{\vec{\mathbf{k}} \in I} c_{\vec{\mathbf{k}}}^{(q)} q_{\vec{\mathbf{k}}}.$$

3.4 Discussion

With the combination technique we presented a way to reduce the costs of a full grid solution at the cost of an additional error.

We showed how an error expansion influences the decay of the surpluses. Depending on the decay of the surpluses we then defined the combination technique. With an double-exponential and a exponential-spectral decay we gave two new alternative proofs for the error bounds of the combination technique for quantities of interest. Moreover we used the exponential-spectral decay to show the importance of balancing the errors in each dimension, as the error along the spectral axis decays at a faster rate than the exponential error axis. This information can then be used to define an alternative combination technique, which suits the decay of the surpluses.

Using the combination technique one has to compute many coarser component grids instead of one fine grid. For example, in two dimensions this reduces the total number of grid points from $\mathcal{O}(n^2)$ on the full grid to $\mathcal{O}(n \log n)$ [BG04]. This makes larger computations possible even if the full grid becomes too large due to computational complexity or even memory constraints. Furthermore this also adds a new layer of parallelism. As all computations on the individual component grids are independent they can easily be run in parallel. This can be used to further reduce the wall time needed for the combined solution.

3.5 Variations*

In the introduction we defined the combination technique as

$$q_n^c \coloneqq \sum_{|\vec{\mathbf{k}}|\leq n} w_{\vec{\mathbf{k}}}$$

assuming a specific error model in which the surpluses decay at the same rate in both dimensions. However in the following chapters we gave different examples where it is a good idea to use an alternative version of the combination technique. Examples for this are cases in which surpluses decay at a different rate in each dimension. In this case the constraint to balance the error in each dimension directly leads us to alternative combination techniques. In a more abstract way we can therefore see the combination technique just as a sum of a set of surpluses. The selection of surpluses, which are summed up then depend on a specific combination technique and can be adjusted to match the problem at hand. Hence we introduce the *index set* $I \subset \mathbb{N}_0^d$, which contains all multi-indices for the surpluses in the sum [HHK+16]. The combination technique in terms of this index set is then described by

$$q_I^c \coloneqq \sum_{\vec{\mathbf{k}}\in I} w_{\vec{\mathbf{k}}}.$$

The index set I_n corresponding to the classical combination technique is given by the

$$I_n = \left\{\vec{\mathbf{k}} \in \mathbb{N}_0^d \mid |\vec{\mathbf{k}}| \leq n\right\}.$$

We basically just moved the bounds from the sum into the curly braces of the index set. An equivalent definition would be

$$I_n = \downarrow \left\{\vec{\mathbf{k}} \in \mathbb{N}_0^d \mid |\vec{\mathbf{k}}| = n\right\}$$

with the *downset*

$$\downarrow A = \left\{\vec{\mathbf{k}} \in \mathbb{N}_0^d \mid \exists_{\vec{\mathbf{a}}\in A}\, \vec{\mathbf{k}} \leq \vec{\mathbf{a}}\right\}.$$

We give a short overview of some alternative combination techniques for different kind of problems. For all methods an example is given in figure 3.3 to illustrate the differences visually.

*This chapter is part of the propaedeutic.

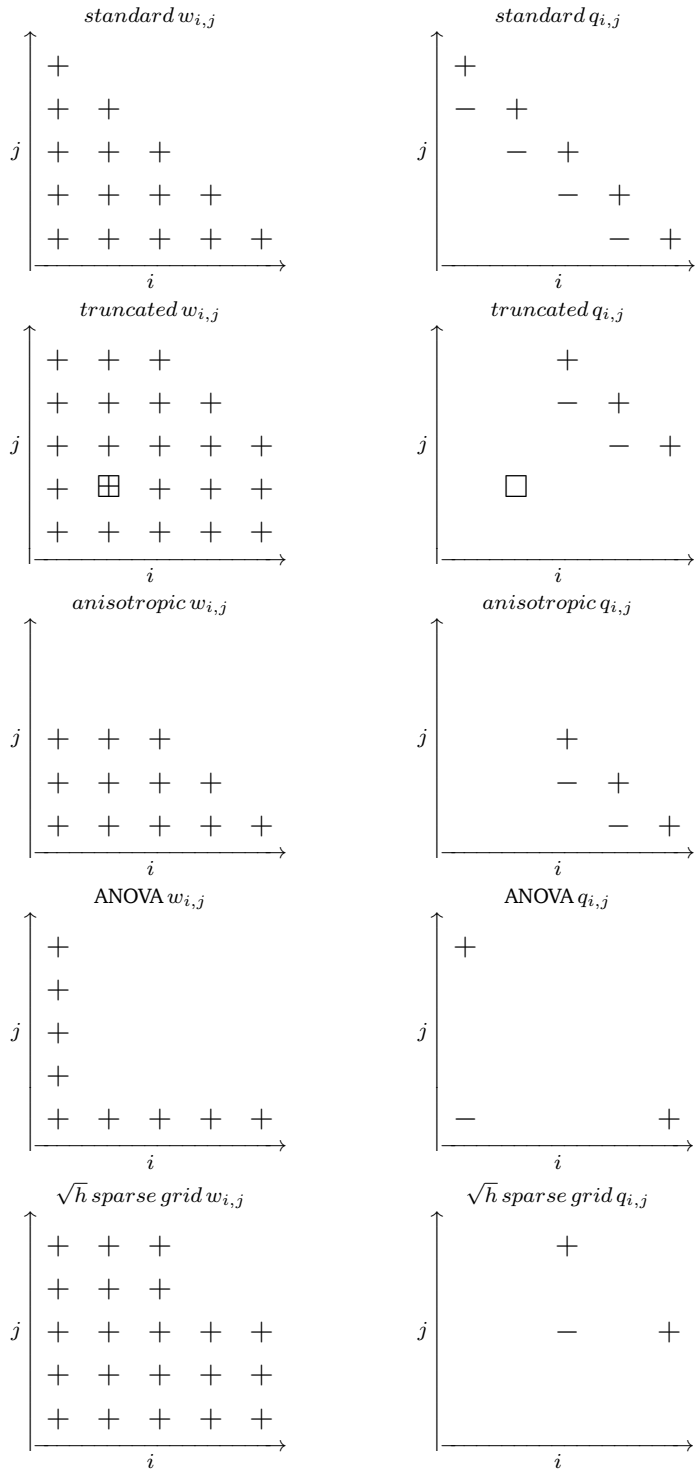

Figure 3.3: Comparison of different combination techniques with $n = 4$. The Plus stands for a factor of plus one in the sum, minus stands for minus one. The truncated version uses $\vec{\mathbf{m}} = (2, 2)$, which is marked with □. The anisotropic grids use $\vec{\mathbf{n}} = (4, 2)$ and ANOVA uses $m = 1$.

3.5.1 Truncated Combination Technique

For many applications including GENE a minimal grid resolution is needed for some dimensions, as otherwise the solver cannot calculate a solutions for grids that are very coarse in one dimension. The solution for this problem becomes apparent when looking at the combination pattern for quantities of interests in figure 3.2. As computations with one index close to zero cannot be computed we take the combination technique corresponding to a larger full grid and skip all computations that are too small to compute. This looks like cutting the corners of the triangular area. We can compute the surpluses belonging to this method by using the results from section 3.3. This gives the index set for the *truncated combination technique* [BBNS12; BP12]

$$I_n^{(tr)} =\downarrow \left\{\vec{\mathbf{k}} \in \mathbb{N}_0^d \mid |\vec{\mathbf{k}}| \leq n + \max(\vec{\mathbf{m}}) \wedge \vec{\mathbf{m}} \leq \vec{\mathbf{k}} \leq \vec{\mathbf{n}}\right\}$$

with $\vec{\mathbf{n}} = n \cdot \vec{\mathbf{1}}$ and a multi-index $\vec{\mathbf{m}}$ describing the minimal resolution needed in each dimension. As the Gene simulations in chapter 5 need a minimal resolution in each dimension we used the truncated combination technique to combine the results of the component grids.

The classical index set is therefore equivalent to the index set of the truncated combination technique with $\vec{\mathbf{m}} = \vec{\mathbf{0}}$. With $\vec{\mathbf{m}} = n \cdot \vec{\mathbf{1}}$ the truncated combination technique is equivalent to the full grid solution.

Using a truncated version with $\vec{\mathbf{m}} \neq \vec{\mathbf{0}}$ increases the computational effort, as the sizes for the needed grids increase. For $\vec{\mathbf{m}} = (d-1)n \cdot \vec{\mathbf{1}}$ the combination technique becomes identical to solving the $\vec{\mathbf{n}}$-grid directly. However, while the error decreases as we use more surpluses, the convergence of the solution function remains unchanged [Har16]. A conclusion from proposition 4.5 in [Har16] shows that the truncated combination technique converges at the same rate to the exact solution as the classical combination technique for increasing n.

3.5.2 Anisotropic Grids

For problems, where different resolutions are required in each dimension, these definitions can be further generalized. As one is interested in a solution using a higher resolution in one dimension it makes sense to take more surpluses in this direction into account. An example for an index set with these properties is

$$I_{\vec{\mathbf{n}}} = \left\{\vec{\mathbf{k}} \in \mathbb{N}_0^d \mid |\vec{\mathbf{k}}| \leq \max(\vec{\mathbf{n}}) \wedge \vec{\mathbf{k}} \leq \vec{\mathbf{n}}\right\}.$$

Combining these different index sets is also an possible. A truncated version with a minimum level of $\vec{\mathbf{m}}$ can then be defined as

$$I_{\vec{\mathbf{n}}}^{(tr)} = \downarrow \left\{ \vec{\mathbf{k}} \in \mathbb{N}_0^d \mid |\vec{\mathbf{k}}| \leq \max_{i,j \leq d} (\vec{\mathbf{n}}_i + \vec{\mathbf{m}}_j) \wedge \vec{\mathbf{m}} \leq \vec{\mathbf{k}} \leq \vec{\mathbf{n}} \right\}.$$

These two definitions are generalizations of the two former definitions. For $\vec{\mathbf{n}} = n \cdot \vec{\mathbf{1}}$ these index sets are equivalent to I_n and $I_n^{(tr)}$ respectively.

However, this uses quite a lot surpluses in one dimension. An other alternative is to define a hyperplane, such that the index set contains all surpluses below this plane:

$$I_{\vec{\mathbf{n}}}^{\vec{\mathbf{a}}} = \left\{ \vec{\mathbf{k}} \in \mathbb{N}_0^d \mid \vec{\mathbf{k}} \leq |\vec{\mathbf{a}} \cdot \vec{\mathbf{n}}| \right\}.$$

The skewness of the plane is thereby controlled by the muti-index $\vec{\mathbf{a}}$, which can be seen as a vector of weights applied to $\vec{\mathbf{n}}$. For $\vec{\mathbf{a}} = \vec{\mathbf{1}}$ this is equivalent to I_n.

3.5.3 Analysis of Variance (ANOVA)

The combination technique is also very similar to a so called ANOVA decomposition of a function, which can be used to estimate the variance of a function as shown in [ES81]. A function depending on random variables $x_1, \ldots, x_n$ is thereby decomposed into a sum of functions, each depending only on $m \leq n$ variables while holding all other variables constant. These functions are uncorrelated and have mean zero [ES81]. Usually only the low-order interactions with $m \ll n$ are used. This simplifies the analysis as the remaining functions have much less arguments than the initial function. For example, the decomposition for a function with 3 arguments looks as follows

$$\begin{aligned} f(x,y,z) = \mu &+ f_1(x) + f_2(y) + f_3(z) \\ &+ f_{12}(x,y) + f_{13}(x,z) + f_{23}(y,z) \\ &+ f_{123}(x,y,z) \end{aligned}$$

which is then approximated with $m = 1$ as

$$f(x,y,z) \approx \mu + f_1(x) + f_2(y) + f_3(z).$$

This method corresponds to a combination technique using only the surpluses $w_{i,0,0}$, $w_{0,j,0}$ and $w_{0,0,k}$. The grid resolutions then control how good the solution functions $f_i(x_i)$ are approximated.

The index can therefore be described as [HHK+16]

$$I_{\vec{\mathbf{n}},m}^{anova} = \left\{ \vec{\mathbf{k}} \in \mathbb{N}_0^d \mid |\operatorname{supp} \vec{\mathbf{k}}| \leq m \wedge \vec{\mathbf{k}} \leq \vec{\mathbf{n}} \right\}$$

where $|\operatorname{supp} \vec{\mathbf{k}}|$ is the number of nonzero entries in $\vec{\mathbf{k}}$. For $m = d$ the ANOVA method becomes equivalent to the full grid solution.

The ANOVA has also applications outside the field of statistics. For example, this decomposition can be used to analyze error bounds for quadrature rules as shown in [Hic96]. In case the magnitude of the individual ANOVA effects is known, one can draw conclusions on the performance of the quadrature rules and reduce the error by choosing an appropriate quadrature rule.

3.5.4 $\sqrt{h}$ Sparse Grids/Two Scale Combination

While sparse grids reduce the number of unknowns for a problem, they usually converge slower than full grid solutions. An other combination technique uses $\sqrt{h}$ *sparse grids* [LZ06]. The total number of unknowns is with $\mathcal{O}\left(h_n^{-d/2}\right)$ higher than usual sparse grids, but still significantly lower than a full grid with $\mathcal{O}\left(h_n^{-d}\right)$. However for a specific class of PDEs the solution functions converge with $\mathcal{O}(h_n^2)$, which is the same order as full grids [Pfl10]. The corresponding index set is given by

$$I_n^{(\sqrt{h})} = \left\{ \vec{\mathbf{l}} \in \mathbb{N}_0^d \mid \exists_k \, \vec{\mathbf{l}}_k \leq n \wedge \forall_{r \neq k} \, \vec{\mathbf{l}}_r \leq \left\lfloor \frac{n}{2} \right\rfloor \right\}$$

which means that all but one entry of the multi-indices in the set must be smaller than $\lfloor n/2 \rfloor$.

4 A Combination Technique for the Romberg Method

In section 2.2.2 we showed a way to use standard extrapolation rules to improve higher dimensional solutions. While this leads to higher accurate approximations this method is quite expensive as all solutions $q_{\vec{\mathbf{k}}}$ with $|\vec{\mathbf{k}}| \leq n$ are needed to compute an extrapolated $q_{n,n}$. With the combination technique in chapter 3 we introduced a way to reduce the costs of normal solutions by just using a subset of the surpluses, which only leads to a slightly worse error. We now try to combine the best of both ideas in a new "extrapolated" combination technique, which has similar asymptotic error bound as the extrapolation method. At the same time it uses significantly less effort to calculate than the tensor Romberg method and thus is closer to the computational effort needed for the traditional combination technique.

This method is based on the tensor extrapolation scheme introduced in section 2.2.2. Instead of the quantities of interest we considered before, we now consider the extrapolated quantities of interest. In the context of integration problems, these quantities of interest are the extrapolated trapezoidal rules $q_{i,j} \coloneqq T^{(i,j)}$. Each solution $q_{i,j}$ is therefore the result of an extrapolation in both dimensions. Based on the extrapolated $q_{i,j}$ we then calculate surpluses and apply the combination technique as we did before.

In the following we first investigate how the new surpluses decay. We then use this knowledge to deduce error bounds for the difference between the full extrapolation $q_{n,n} = T^{(n,n)}$ and the combined result q_n^c using the extrapolated quantities of interest.

4.1 Effect of the Extrapolation on the Surpluses

In order to estimate the error for this technique, we first investigate how the tensor Romberg method affects the surpluses. Starting with the one dimensional case we denote the extrapolated quantity of interest as $q_i^{(e)}$ with the surpluses $w_i^{(e)}$. For $i = 0$ we just have $w_0^{(e)} = q_0^{(e)} = q_0$ as before. For $i = 1$ we have

$$w_1^{(e)} = q_1^{(e)} - q_0^{(e)} = \frac{4}{3}q_1 - \frac{1}{3}q_0 - q_0 = \frac{4}{3}\left(q_1 - q_0\right) = \frac{4}{3}w_1.$$

Otherwise, using the definition of the surpluses (3.2) and substituting the one dimensional Romberg method (2.7) we have

$$\begin{aligned} w_{i+1}^{(e)} &= q_{k+1}^{(e)} - q_k^{(e)} \\ &= \frac{4}{3}q_{k+1} - \frac{1}{3}q_k - \frac{4}{3}q_k + \frac{1}{3}q_{k-1} \\ &= \frac{4}{3}\left(q_{k+1} - q_k\right) - \frac{1}{3}\left(q_k - q_{k-1}\right) \\ &= \frac{4}{3}w_{k+1} - \frac{1}{3}w_k \end{aligned}$$

and we see that the extrapolation on the quantity of interest corresponds directly to the extrapolation of the surpluses. This also holds for higher extrapolation orders, as for each new order a similar pattern to (2.8) can be applied to the previous extrapolation level. This means that we can eliminate further terms from the expanded form of the surpluses. However, as each new extrapolation level contains one less extrapolated quantity of interest, we shift the indices by one such that the first surpluses remain unchanged.

In two dimensions we apply the one dimensional Romberg method on each dimension separately and get the same result by rearranging the terms

$$\begin{aligned} w_{i,j}^{(e)} &= q_{i,j}^{(e)} - q_{i-1,j}^{(e)} - q_{i,j-1}^{(e)} + q_{i-1,j-1}^{(e)} \\ &= \frac{16}{9}\left(q_{i,j} - q_{i-1,j} - q_{i,j-1} + q_{i-1,j-1}\right) \\ &\quad - \frac{4}{9}\left(q_{i-1,j} - q_{i-2,j} - q_{i-1,j-1} + q_{i-2,j-1}\right) \\ &\quad - \frac{4}{9}\left(q_{i,j-1} - q_{i-1,j-1} - q_{i,j-2} + q_{i-1,j-2}\right) \\ &\quad + \frac{1}{9}\left(q_{i-1,j-1} - q_{i-2,j-1} - q_{i-1,j-2} + q_{i-2,j-2}\right) \\ &= \frac{16}{9}w_{i,j} - \frac{4}{9}w_{i-1,j} - \frac{4}{9}w_{i,j-1} + \frac{1}{9}w_{i-1,j-1} \end{aligned}$$

which corresponds to the first extrapolation step in two dimensions. The coefficients come from the tensor product of the weights $(4/3, -1/3)$ with themselves. Again, the fact that the extrapolation applied to the quantities directly applies to the surpluses, still holds in higher dimensions due to the tensor product properties of our method.

To perform the one dimensional extrapolations we can use a similar scheme as for the quantities of interest and apply the scheme directly to the old surpluses:

$$\begin{array}{ccccccc} w_0 & = & w_0^{(1)} & = & w_0^{(2)} & = & w_0^{(3)} \\ & \searrow & & & & & \\ w_1 & \rightarrow & w_1^{(1)} & = & w_1^{(2)} & = & w_1^{(3)} \\ & \searrow & & \searrow & & & \\ w_2 & \rightarrow & w_2^{(1)} & \rightarrow & w_2^{(2)} & = & w_2^{(3)} \\ & \searrow & & \searrow & & \searrow & \\ w_3 & \rightarrow & w_3^{(1)} & \rightarrow & w_3^{(2)} & \rightarrow & w_3^{(3)} \end{array}$$

where $w_i^{(k)}$ are the k^{th} extrapolated surpluses and w_i the non extrapolated surpluses. After the extrapolation we have the new extrapolated surpluses $w_i^e := w_i^{(i)}$, which we denote by w_i or $w_{i,j}$ in two dimensions in this chapter. Hence we can directly deduce the extrapolated surpluses $w_{i,j}^{(e)}$ from the original surpluses $w_{i,j}$ by applying the extrapolation scheme.

4.2 Error for the Combined Romberg Method

In the previous chapter we showed that the extrapolation for the quantities of interest directly trickles down to the surpluses. This means that after the extrapolation the surpluses are of increasing order in each dimension.

The extrapolation as presented earlier only makes sense if the error expansion for the quantity of interest only contains even exponents. By adjusting the steps in section 2.2.1, one could easily adjust the method to work with an error expansion containing odd coefficients as well. However for surpluses that decay in a spectral way like for a Taylor expansion, the Romberg method is less effective.

Therefore, we now consider a quantity of interest with an error expansion of the form

$$q_{x,y} - q = \sum_{\substack{i=2 \\ \text{even}}}^{\infty} \sum_{\substack{k=0 \\ \text{even}}}^{i} c_{k,i-k} h_x^k h_y^{i-k}$$

which is similar to the two dimensional Euler-Maclaurin formula we discussed in section 2.1.2. As shown in 2.2.1 each one dimensional extrapolation step is given by

$$T^{(k)}(h) = \frac{T^{(k-1)}(h) - \alpha^{-2k} T^{(k-1)}(\alpha h)}{1 - \alpha^{-2k}}$$

if T satisfies the error expansion. Each extrapolation step does not only cancel one additional error term from the error expansion but also modifies the coefficients for

all higher order error terms. We now investigate how the coefficients for the highest remaining error terms change over multiple extrapolation steps. This helps us to calculate how fast the extrapolated quantities of interest converge. As shown in the previous section, the extrapolation on the quantity of interest corresponds to extrapolating the surpluses. Once we know the decay for the surpluses, we can estimate the error for our method similar to 3.2.3 by summing up the surpluses.

4.2.1 Decay of the Extrapolation Coefficients

To calculate the decay of the coefficients we first focus on the one dimensional case and a quantity of interest q with an error expansion

$$q(h) - q = \sum_{\substack{i=2 \\ \text{even}}}^{p} c_i h^i.$$

We now denote the k^{th} extrapolation with a fixed α as $q^{(k)}(h)$. The extrapolation does not only cancel error terms but also changes the coefficients of the remaining error terms. In order to estimate the error of the extrapolated results we now investigate how the coefficients change for increasing extrapolation orders. Therefore we write the error expansion of $q^{(k)}$ as

$$q^{(k)}(h) - q = \sum_{\substack{i=2(k+1) \\ \text{even}}}^{p} d_i^{(k)} c_i h^i$$

where the coefficients are changed by an additional factor $d_i^{(e)}$. In particular $d_{2k+2}^{(k)}$ is of special interest, as it is the additional factor for the lowest order term of the k^{th} extrapolation. The following lemma gives a bound for these additional factors:

Lemma 4.1. *Consider a quantity of interest with an error expansion of the form*

$$q(h) - q = \sum_{\substack{i=2 \\ \textit{even}}}^{p} c_i h^i$$

and the extrapolated quantities of interest $q^{(k)}$ with an error expansion

$$q^{(k)}(h) - q = \sum_{\substack{i=2(k+1) \\ \textit{even}}}^{p} d_i^{(k)} c_i h^i. \tag{4.1}$$

The factors $d_{2k+2}^{(k)}$ for the extrapolated quantity of interest with $\alpha = 1/2$ satisfy

$$\left| d_{2k+2}^{(k)} \right| \leq 2^{-\left(k^2+k\right)} \in \mathcal{O}\left(2^{-k^2}\right).$$

Proof. As the 0^{th} extrapolation corresponds to the initial quantity of interest we have $d_i^{(0)} = 1$ for all i. After the first extrapolation step the h^2 error term is eliminated, which means that $d_2^{(1)} = 0$. For all other terms in the error expansion we have

$$d_i^{(1)} = \frac{\alpha^{i-2} - 1}{\alpha^{-2} - 1}$$

for $i = 4, 6, \dots$ according to the extrapolation formula. For the second extrapolation step it holds

$$d_i^{(2)} = \begin{cases} 0 & i \in \{2, 4\} \\ \frac{\alpha^{i-4}-1}{\alpha^{-4}-1} d_i^{(1)} & i = 6, 8, \dots \end{cases}$$

where we eliminate the h^4 error term. Continuing this pattern gives

$$d_i^{(k)} = \begin{cases} 0 & i = 2, 4, \dots, 2k \\ \frac{\alpha^{i-2k}-1}{\alpha^{-2k}-1} d_i^{(k-1)} & i = 2k+2, 2k+4, \dots \end{cases}$$

which gives us a recursive formula for the non-zero values. We can write the non-zero values for $i \geq 2k + 2$ as

$$d_i^{(k)} = \prod_{e=1}^{n} d_i^{(e)} = \prod_{e=1}^{n} \frac{\alpha^{i-2e} - 1}{\alpha^{-2e} - 1}.$$

Note that all factors $d_i^{(k)}$ have an absolute value smaller one. We know that the k^{th} extrapolation eliminates all terms h^i with $i \leq 2k$ and that the the lowest order term remaining is $d_{2k+2}^{(k)} c_{2k+2} h^{2k+2}$. We are therefore interested in the factor $d_{2k+2}^{(k)}$ to estimate the error. Using the previous formula we see

$$d_{2k+2}^{(k)} = \prod_{i=1}^{k} d_{2i+2}^{(i)} = \prod_{i=1}^{k} \frac{\alpha^{2k+2-2i} - 1}{\alpha^{-2i} - 1}.$$

As the absolute value of each numerator is smaller than one we can estimate the coefficients with $\alpha = 1/2$ as

$$\left| d_{2k+2}^{(k)} \right| \leq \prod_{i=1}^{k} \frac{1}{4^k - 1} \leq \prod_{i=1}^{k} \frac{1}{4^k} = 4^{-\sum_{i=1}^{k} i} \in \mathcal{O}\left(2^{-k^2}\right)$$

for $k \to \infty$. □

While the error for the original quantity of interest $q(h)$ behaves like

$$\left| q(2^{-k}h) - q \right| \leq C 2^{-2k} h^2$$

for a constant C, it holds

$$\left|q^{(k)}(h, 2^{-1}h, \ldots, 2^{-k}h) - q\right| \leq C^{(e)}2^{-k^2}h^{2k+2}$$

for the extrapolated quantity of interest and an new constant $C^{(e)}$. We see that the extrapolated quantities decay much faster for increasing k. As shown before the extrapolation for the quantity of interest this is equivalent to applying the extrapolation to the surpluses for which a similar formula holds.

4.2.2 Calculating the Error

We now apply the combination technique to the extrapolated surpluses and calculate some error bounds. Based on the result of the previous section and the tensor product properties of applying the extrapolation on each dimension individually we now assume that our surpluses satisfy

$$|w_{i,j}| \leq M2^{-\left(i^2+j^2\right)}$$

for a two dimensional problem and a nonnegative constant M. Note that we assumed $|w_{i,j}| \leq M2^{-(i+j)}$ in section 3.2.3 where we discussed the error for the normal combination technique without extrapolation. In figure 4.1 is shown how levels of similar absolute value behave for each kind of decay.

Once again we are interested in the difference between the full grid solution $q_{n,n}$ and the combined solution q_n^c, which now uses the extrapolated results. Here the full grid solution corresponds to the full extrapolated solution using the tensor Romberg method. The following theorem gives an error bound for the additional error.

Theorem 4.2. *Consider the extrapolated surpluses $|w_{i,j}| \leq M2^{-\left(i^2+j^2\right)}$ and the combination technique $q_n^c = \sum_{i+j\leq n} w_{i,j}$. Then it holds for the difference between the extrapolated full grid solution $q_{n,n}$ and the combination technique q_n^c*

$$|q_{n,n} - q_n^c| \leq \frac{M\pi}{4\ln 2}2^{-\frac{1}{2}n^2}.$$

Proof. The difference between the solutions can be estimated by

$$|q_{n,n} - q_n^c| = \sum_{\substack{k+l>n \\ k,l\leq n}} |w_{k,l}| \leq M \sum_{\substack{k+l>n \\ k,l\leq n}} 2^{-\left(i^2+j^2\right)}.$$

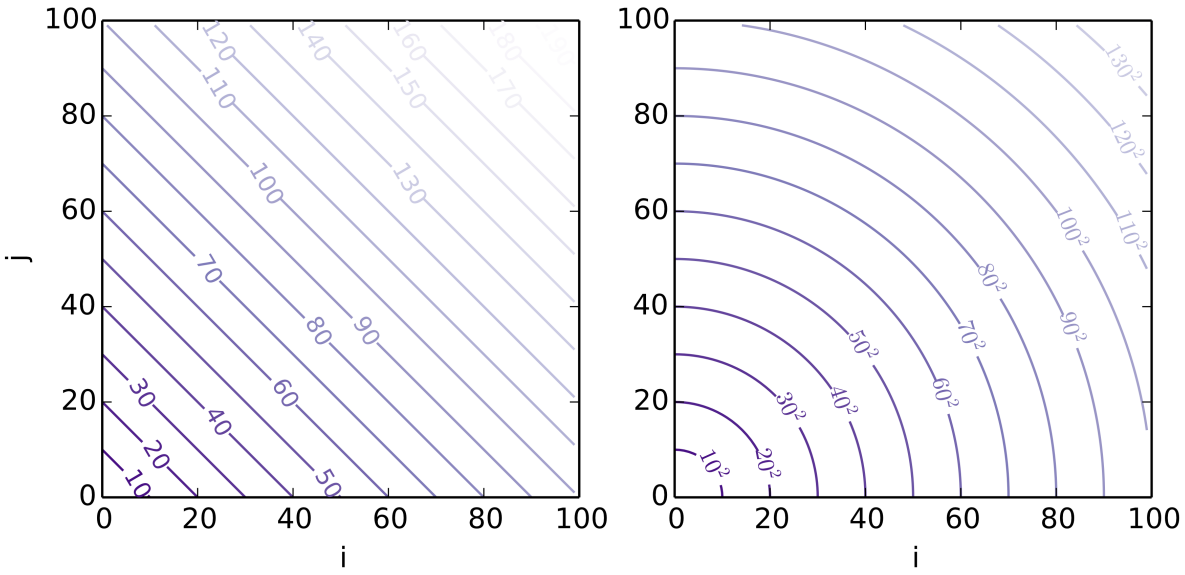

Figure 4.1: Levels, on which surpluses have a similar value. Left: levels with $i+j = const$ form lines. Right: levels with $i^2 + j^2 = const$ form circles.

However, we cannot sum up the surpluses level-wise as before as the solution for $i^2 + j^2 = const$ are circles, which does not map nicely to our discrete values of i and j. We therefore approximate the sum by an integral, which is easier to calculate:

$$\dots \leq M \int_0^n \int_{n-x}^n 2^{-\left(x^2+y^2\right)} \,\mathrm{d}y\mathrm{d}x.$$

We now switch to polar coordinates and generously estimate the triangular integration region by a torus part covering the triangle and we have

$$\begin{aligned}
\dots &\leq M \int_{\frac{1}{2}\sqrt{2}n}^{\sqrt{2}n} \int_0^{\frac{\pi}{2}} 2^{-r^2} r \,\mathrm{d}\varphi\mathrm{d}r \\
&= M\frac{\pi}{2} \int_{\frac{1}{2}\sqrt{2}n}^{\sqrt{2}n} r 2^{-r^2} \,\mathrm{d}r \\
&= \frac{M\pi}{4\ln 2} \left(2^{-\frac{1}{2}n^2} - 2^{-2n^2}\right) \\
&\leq \frac{M\pi}{4\ln 2} 2^{-\frac{1}{2}n^2}
\end{aligned}$$

where we use the fact that all the surpluses along a circle around the origin have similar absolute values. □

In contrast to the combination technique without the extrapolation, which decays proportional to $n2^{-n}$, we now observe a decay proportional of $2^{-\frac{1}{2}n^2}$, which should be significantly smaller for reasonably big values of n.

Based on the observation that surpluses on a circle around the origin are similarly sized, we can define an alternative combination technique $q_n^{c'}$, which gets rid of the $1/2$ factor in the exponent at the cost of using some more surpluses:

$$q_n^{c'} = \sum_{k^2+l^2 \leq n^2} w_{k,l}.$$

Note that we now use $k^2 + l^2 \leq n^2$ as an upper bound for the indices instead of $k + l \leq n$. This implies that all surpluses on this boundary have a similar absolute value. We calculate the additional error as before and find

$$\begin{aligned}\left|q_{n,n} - q_n^{c'}\right| &\leq M \int_n^{\sqrt{2}n} \int_0^{\frac{\pi}{2}} 2^{-r^2} r \,\mathrm{d}\varphi \mathrm{d}r \\ &= \frac{M\pi}{4\ln 2}\left(2^{-n^2} - 2^{-2n^2}\right) \\ &\leq \frac{M\pi}{4\ln 2} 2^{-n^2}.\end{aligned}$$

However this alternative combination technique $q_n^{c'}$ uses about $\frac{1/4\pi n^2 - 1/2n^2}{n^2/2} \approx 57\%$ more surpluses than q_n^c.

Instead of adding more surpluses to improve the error bounds, we can also remove surpluses while maintaining the error bounds of q_n^c. In the error estimation for q_n^c we loosely estimated the triangular integration region by an torus segment. We can now define another alternative combination technique

$$q_n^{c''} = \sum_{k^2+l^2 \leq \frac{n^2}{2}} w_{k,l}$$

where we adapt the set of used surpluses such that they resemble the torus estimation much better. We now have an error bound of

$$\left|q_{n,n,} - q_n^{c''}\right| \leq \frac{M\pi}{4\ln 2} 2^{-\frac{1}{2}n^2}$$

while reducing the amount of used surpluses by approximately $\frac{1/4\pi(n/\sqrt{2})^2 - 1/2n^2}{n^2/2} \approx -21\%$.

It is important to note that these percentage numbers do not relate to the computational complexity or similar things as the cost for calculating the individual surpluses can vary a lot. Many different factors like the used grid sizes and the complexity of the solver have an effect on the total cost. The numbers are only given for a very rough idea of how much the amount of surpluses differ for each version and to give an example of things to consider when deciding which surpluses one wants to use.

4.3 Discussion

While this method achieves similar results to the full tensor Romberg method with much less surpluses (see chapter 5), both have some shortcomings. Similar to the normal one dimensional Romberg extrapolation for integration problems these methods strongly depend on the smoothness of the function. If the error expansion does not hold or is too short for the degree of extrapolation, these methods can lead to large errors or totally wrong results. For high extrapolation degrees these methods additionally become very prone to small inaccuracies such as rounding errors. This can make high extrapolation errors unfeasible. However, due to the good approximation order of the extrapolation, lower order extrapolations might give results that are good enough.

An alternative similar method described in [Har16] and [Rei04] uses the calculated solutions to generate many first or second order extrapolations, which are then combined with the combination technique. While this method does not have such a high approximation order it is much more robust with respect to rounding errors and improves the accuracy of the final result. Thanks to our previous work we can even estimate the error quite easily. As a first order extrapolation cancels all h_i^2 terms in the expansion for both, quantity of interest and the surpluses, we estimate the new surpluses as $w_{i,j} \leq M2^{-4(i+j)}$ and repeat the steps from section 3.2.3.

5 Numerical Results

After we discussed many different theoretical aspects of the combination technique and the surpluses, we now present some numerical results for different test cases. Namely we test our results with different artificial integration problems, where the quantity of interest is the integral over a two dimensional function. Also we test our methods for various quantities of interests calculated by the plasma turbulence simulation GENE.

5.1 Experimental Setup

5.1.1 Integration Problems

For some easy to verify experiments we used various two dimensional integration problems. A big advantage of this is that the exact solution is often known analytically and therefore the error and convergence of the methods can easily be verified. As quantities of interest we calculated

$$q_1(f) = \int_0^1 \int_0^1 f(x,y)\,\mathrm{d}x\mathrm{d}y$$
$$q_2(f) = \int_0^1 \int_0^1 \left(f(x,y)\right)^2\,\mathrm{d}x\mathrm{d}y$$

for different functions and calculated the surpluses and analyzed the convergence behavior. To calculate numerical results for the quantities of interest different integration methods were used. Besides a hierarchical and non-hierarchical trapezoidal rule with $2^k + 1$ and 2^k grid points respectively, other methods like Gauss-Legendre, Clenchaw-Curtis, truncated Taylor series and a least-squares polynomial approximation were used. Additionally noise was added on the grid points for some of the methods to investigate the effect. In this work we will only discuss the results of the normal trapezoidal rules, as they nicely show our theoretical results due to their error expansion. The evaluation of other methods is beyond the scope of this work.

To use the trapezoidal rule we first have to discretize the integration area $\Omega = [0,1]^2$. For the trapezoidal rule we choose different equidistant grids in each dimension:

$$G_{i,j} = \left\{ (x,y) \in \Omega \mid x = \frac{k}{N_i}, y = \frac{l}{N_j}, k = 0,\dots,N_i, l = 0,\dots,N_j \right\}$$

with $N_k = 2^k$ for the hierarchical and $N_k = 2^k - 1$ for the non-hierarchical trapezoidal rule. On $G_{i,j}$ the trapezoidal rule is then defined as

$$T_{i,j}(f) = \frac{1}{N_i}\frac{1}{N_j}\sum_{k=0}^{N_i}\sum_{l=0}^{N_j} c_{k,l} f\left(\frac{k}{N_i}, \frac{l}{N_j}\right)$$

with the weights

$$c_{k,l} = \begin{cases} 4 & k = 1,\dots,N_i - 1 \wedge l = 1,\dots,N_j - 1 \\ 1 & k = 0, N_i \wedge l = 0, N_j \\ 2 & \text{otherwise.} \end{cases}$$

The combination technique is then applied over the results of the trapezoidal rules $q_{i,j} = T_{i,j}(f)$ as quantity of interest with different numbers of grid points in each dimension.

We use these integration problems to verify our theoretical results on the behavior of the surpluses and the error in the next chapter.

5.1.2 GENE

To apply the combination technique to a real world example we used the gyrokinetic plasma turbulence simulation GENE [GM13]. GENE was initially developed by F. Jenko *et al.* [JDKR00] and successively improved by others. It solves the nonlinear gyrokinetic equations, but also contains a switch to only use the linear part of the equation.

In this work we consider GENE as a black box, which runs a simulation for a given set of grid parameters and outputs a set of quantities of interests. For a more in depth explanation of GENE and the underlying physics we refer to the manual [GM13] and

the dissertations of Dannert [Dan05] and Grler [Gör10]. The quantities of interest we consider are all integrals over the velocity space d^3v:

$$
\begin{aligned}
n_1 &= \int v f_1 \,\mathrm{d}^3v & \mathbf{Q}_{es} &= \int \frac{1}{2} m v^2 f_1 \mathbf{v}_D \,\mathrm{d}^3v \\
u_{1\parallel} &= \frac{1}{n_0} \int v_\parallel f_1 \,\mathrm{d}^3v & \mathbf{Q}_{em} &= 0 \\
T_{1\parallel} &= \frac{m}{n_0} \int \left(v_\parallel - u_{1\parallel}\right)^2 f_1 \,\mathrm{d}^3v - T_0 \frac{n_1}{n_0} & \mathbf{\Gamma}_{es} &= \int f_1 \mathbf{v}_D \,\mathrm{d}^3v \\
T_{1\perp} &= \int v\, v_\perp^2 f_1 \,\mathrm{d}^3v - T_0 \frac{n_1}{n_0} & \mathbf{\Gamma}_{em} &= 0
\end{aligned}
$$

with the particle distribution function f_1, particle density n_1, parallel velocity $v_\parallel$, drift velocity $\mathbf{v}_D$, parallel and perpendicular temperature $T_\parallel$ and $T_\perp$, the electrostatic and electromagnetic particle flux Γ_{es} and Γ_{em} and the electrostatic and electromagnetic heat flux Q_{es} and Q_{em}. Each property is then averaged over the simulation volume and normalized to a reference value [GM13]. The symbols for the different physical properties were chosen to match the symbols in the GENE manual.

This also shows the general importance of considering integration problems as all quantities of interest contain some sort of integral.

For our experiments we considered linear runs with an equidistant μ-grid, nonlinear runs with an equidistant μ-grid and nonlinear runs with an Guass-Laguerre μ-grid. Multiple simulations were run with different grid resolutions. The resolutions in the spacial dimensions x, y and z were kept constant while the resolutions in the dimensions v (parallel velocity $v_\parallel$) and w (magnetic moment μ) were varied. The number of grid points in the velocity and magnetic dimension was chosen to be 2^k for the equidistant grids and k for the Gauss-Laguerre grid. As GENE fails to run for to small grids, k has to be at least 3 and at least 4 for the w dimension as the solutions for the quantities of interest collapse to zero otherwise. Therefore, all GENE plots contain an offset and start at $k = 3$. The parameter files with further settings for the simulations can be found in the Appendix.

We then applied the combination technique on the different solutions and analyzed the result.

5.2 Results for Integration Problems

We now present our results with different integration problems and check if they confirm our theoretical results of the previous chapters.

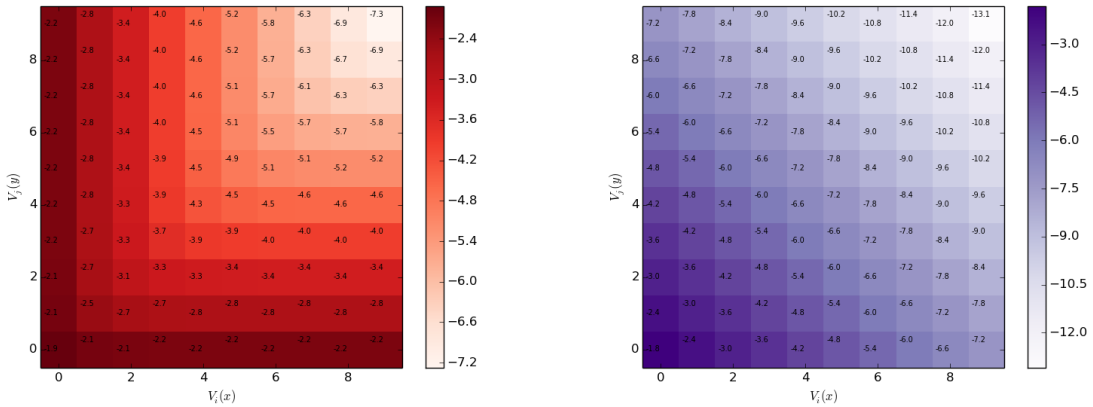

Figure 5.1: Logarithmic absolute value for the error of the quantity of interest $q_1(f)$ (left) and the surpluses (right) for the two dimensional integral with the trapezoidal rule for $f(x, y) = x\,(x-1)\,y\,(y-1)$. The parameters i and j effect the grid resolution in each dimension. Increasing i or j doubles the amount of grid points in x or y dimension respectively.

5.2.1 Effect of the Error Expansion on the Surpluses

In chapter section 3.2.2 we assumed a general error expansion for a quantity of interest and concluded that the expansion for the surpluses only contains terms of the form $h_x^k h_y^l$ with $k, l \geq 1$. That means non-mixed terms like h_x^k or h_y^l are not present any more, even though the error expansion for the quantity of interest can contain these terms.

This effect can clearly be seen in our experiments. In figure 5.1 this effect can be seen for a two dimensional integration problem. Lines on which the calculated quantity of interest (the value of the integral approximated by a trapezoidal rule) has similar values are L-shaped. The lines on which the surpluses have similar values form parallel lines going from the upper left to the lower right. This is a direct consequence of the error expansions by examining the most significant error terms:

As the quantity of interest is calculated using the trapezoidal rule the lowest order error terms are h_x^2 and h_y^2. The value of the error mostly depends on $\max\left\{h_x^2, h_y^2\right\}$. As $h_x \propto 2^{-i}$ and $h_y \propto 2^{-j}$ are monotonically decreasing the levels of similar values are described by $\min\{i, j\} = const$, which leads to the L-shaped levels. Note that the highest mixed term is $h_x^2 h_y^2$ and is therefore of higher order.

As shown in section 3.2.2 calculating the surpluses cancels out all non-mixed terms like from the error expansion. The h_x^2 and h_y^2 terms are therefore no longer present in the surpluses and the lowest order term $h_x^2 h_y^2$ is proportional to $2^{-i}2^{-j} = 2^{-(i+j)}$. For the surpluses, levels of similar values are therefore lines with $i + j = const$, which leads to

the parallel lines from the upper left to the lower right. Summing up only the surpluses with the larges values therefore leads to the combination technique $q_n^c := \sum_{|\vec{\mathbf{l}}|\leq n} w_{\vec{\mathbf{l}}}$ as defined in 3.4.

5.2.2 Error of the Combination Technique

We now revisit the error estimations for the combination technique. As the expansion for the surpluses for the trapezoidal rule only contains mixed terms of at least second order in each grid resolution it holds

$$|w_{i,j}| \leq M4^{-(i+j)}.$$

In (3.2.3) we discussed the error for the combination technique and showed that

$$|q_{n,n} - q_n^c| \leq \frac{M}{3} n4^{-n}$$

holds according to equation (3.5).

The results for one of the integration problems is shown in figure 5.2. The plots show the measured additional error $|q_{n,n} - q_n^c|$ and an approximation for the upper bound $\frac{M}{3}n4^{-n}$. The upper bound was calculated using $M_{max} = \max_{i,j}\left(4^{i+j}\,|w_{i,j}|\right)$ for all known surpluses. As the surpluses decay quickly, this should be a good approximation for M. The results shown in figure 5.2 are particularly interesting, as the measured error $|q_{n,n} - q_n^c|$ is just slightly smaller then upper bound $\frac{M}{3}n4^{-n}$, which shows that the given formula for the upper bound is a good approximation.

5.2.3 Effects of Extrapolation on the Surpluses

In section 4.1 we discussed the effect of the extrapolation on the surpluses and found that the extrapolating the quantity of interest is equivalent to extrapolating the surpluses that therefore decay much faster. Furthermore we found that surpluses on a circular line around the origin have similar absolute values. The results for the integration problems in figure 5.3 show this effect: While the surpluses for the extrapolated combination technique are slightly larger close to the origin they decay significantly faster than the standard combination technique due the performed extrapolation.

The actual rate at which the surpluses decrease are shown in figure 5.4. The surpluses for the combination technique clearly show that the absolute value decay by an factor of 2^{-4} every time the number of grid pints doubles in both dimensions. This directly relates to the results from section 3.2.2 saying that the surpluses only contain mixed terms. In

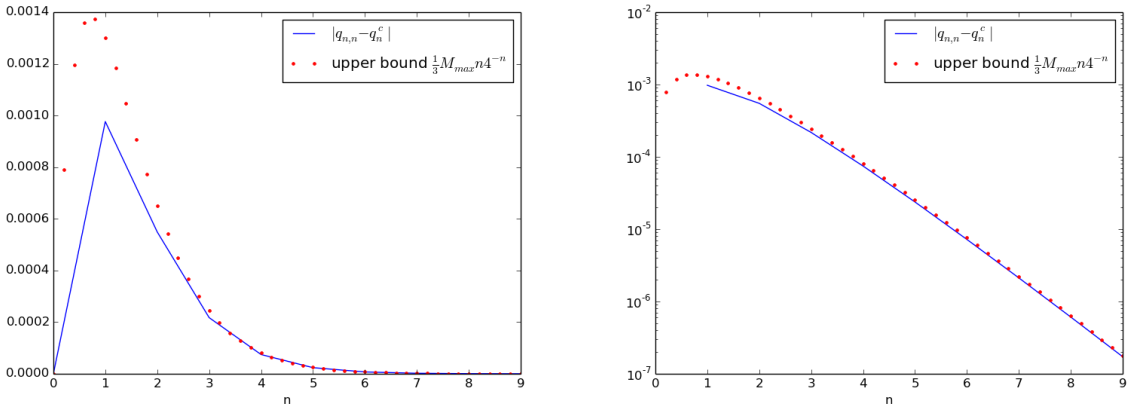

Figure 5.2: Additional error $|q_{n,n} - q_n^c|$ for the two dimensional trapezoidal rule with $f(x, y) = x\,(x-1)\,y\,(y-1)$ as quantity of interest: Measured error (blue line) and approximated upper bound (red dots). The upper bound was approximated, by using the maximum M for all known surpluses. Note that both values are discrete values drawn continuously to show the effect more clearly.

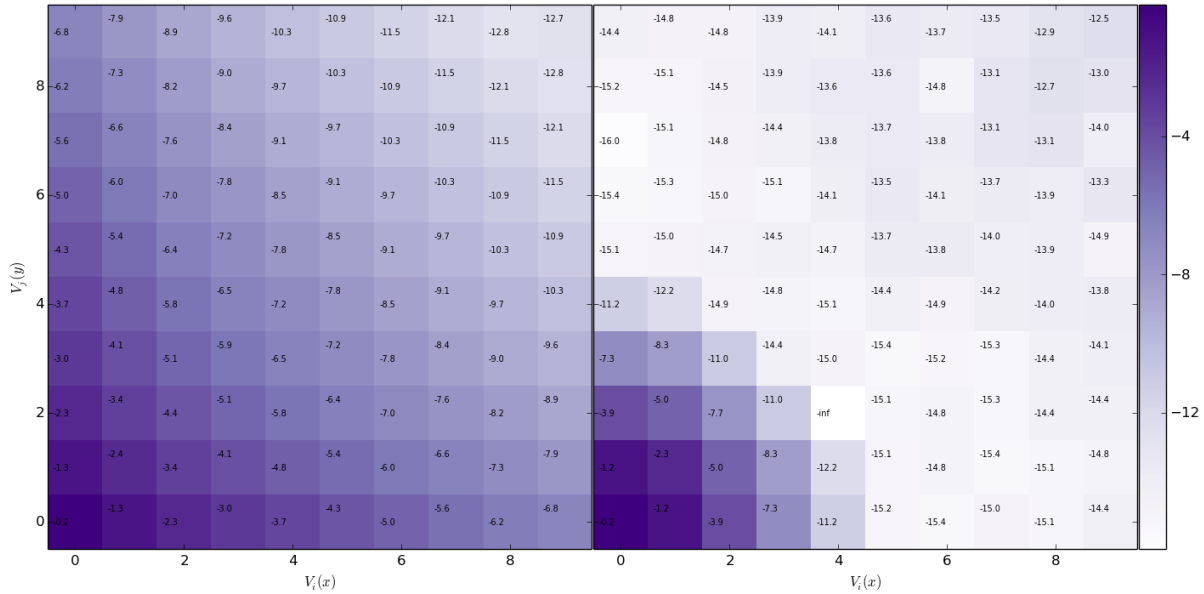

Figure 5.3: The logarithmic absolute value of the surpluses for the combination technique (left) and the combination technique with extrapolation (right) using the non-hierarchical trapezoidal rule for $\sin(x + y)$.

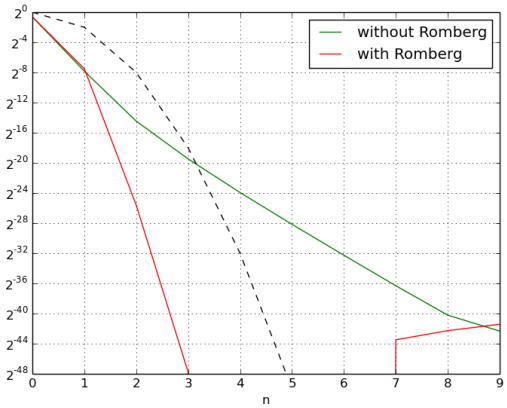

Figure 5.4: Decay of the absolute value of the surpluses $w_{n,n}$ for the combination technique (green) and the combination technique with extrapolation (red) using the non-hierarchical trapezoidal rule for $\sin(x+y)$. The dashed black line is the expected decay $c2^{-(n^2+n^2)}$ as a reference.

case of the error expansion for the trapezoidal rule the lowest order term remaining is $h_x^2 h_y^2$. Doubling the number of grid points then gives $(2^{-1}h_x)^2(2^{-1}h_y)^2 = 2^{-4}h_x h_y$. The extrapolated surpluses are slightly larger at first but start decreasing much faster than the original surpluses. This is due to the fact that the extrapolated surpluses decay like $2^{-(i^2+j^2)}$ for $i, j \to \infty$ while the original surpluses decay like $4^{-(i+j)}$ as shown in section 4.1 and section 3.2.2. The larger surpluses for the extrapolated version, which can be seen at higher levels, are probably artifacts of rounding errors and the high extrapolation order.

5.2.4 Error for the Combination Technique with Extrapolation

In chapter 4 we introduced a combination technique built on top of an extrapolation method to improve the error bounds of the method. We found that

$$\left|q_{n,n} - q_n^{c'}\right| \leq \frac{M\pi}{4\ln 2} 2^{-\frac{1}{2}n^2}$$

holds for the additional error for a constant M, which should lead to significantly better results than the normal combination technique. In our experiments with the extrapolated combination technique we used the adjusted version

$$q_n^{c'} = \sum_{k^2+l^2 \leq n^2} w_{k,l},$$

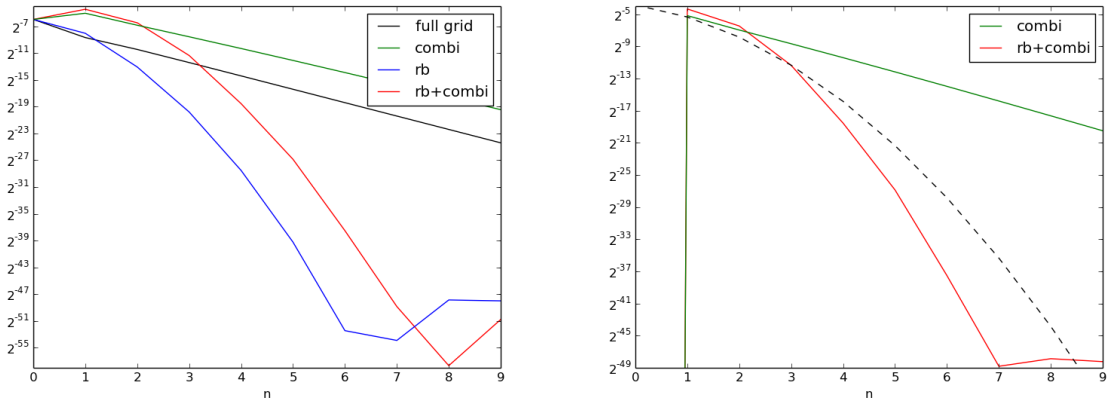

Figure 5.5: Comparison of or different methods for the hierarchical trapezoidal rule for $f(x, y) = \sin(3x + 5y)$: Left: Error of the full grid solution (black), combination technique (green), full tensor Romberg method (blue), Combined Romberg method (red). Right: Difference between the full grid solution and the combined solution for the combination technique (green) and the Combined Romberg method (red). The dashed black line is an example for the expected decay $c2^{-1/2n^2}$ as a reference. Note that the corresponding full grid solution for the Combined Romberg Method is the tensor Romberg method.

which matches the decay of the surpluses. The resulting errors for the function $f(x, y) = \sin(3x + 5y)$ can be seen in figure 5.5. On can see that the error for the Romberg versions decreases much faster than the error for the standard combination technique. For each versions of the normal and the extrapolated techniques achieve both, the full grid solution and the combined solution, similar convergence order. Also the difference between the full grid solution and the combined solution look as out theory suggested. While the difference for the normal combination technique decreases exponentially, the Romberg version decreases even faster than $2^{-\left(i^2+j^2\right)}$. Besides the fact that we only gave an upper bound for the error, one further reason for this might be that the coefficients $c_{i,j}$ in front of the error terms in (4.1) on page 50 decay, too. As mentioned before the increasing error for $n > 7$ might be an error due to rounding errors.

5.3 Plasma Turbulence Simulation GENE

In the previous section it was quite easy to apply our results to different integration problems. However, for GENE this turned out to be more challenging.

As mentioned before, GENE has two simulation modes: a linear mode and a non-linear mode. However for the linear mode the quantities of interest diverges exponentially to infinity for large simulation times as shown in figure 5.6. The rate at which the quantities increase can still give interesting physical insight [GM13]. The fact that the computed quantity of interest drops towards zero might be due to overflow effects in the solver.

In case of the nonlinear run the growth of the quantity of interest is dominated by the linear part for the first few second where one can see the exponential growth once again. Afterwards the nonlinear part shows effect and the solution fluctuates around a constant level. These fluctuations make it difficult to apply the combination technique over different solutions as the extend is quite large even on smaller time scales: If one takes one point in time and measures the quantity of interest the value might differ a lot to the value of the quantity of interest a few seconds later. Changes in the scale of 2500 in about 20 seconds can be observed, which is a lot compared to an average value of about 4500. Also the standard deviation is quite large with a value of 720, which is equivalent to a variance of 5200. The trend for a nonlinear run with average and standard deviation is shown in figure 5.6. For both the mean

$$\mu = \frac{1}{\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} w_i x_i$$

and the variance [Pri72]

$$\sigma^2 = \frac{\sum_{i=1}^{n} w_i}{\left(\sum_{i=1}^{n} w_i\right)^2 - \sum_{i=1}^{n} w_i^2} \sum_{i=1}^{n} w_i \left(x_i - \mu\right)^2$$

we used a weighted version of the formulas where the weights for each data point is the length of each time step simulated.

Due to the high fluctuations the surpluses for the quantities of interest are very noisy and the quality of the result of the combination technique is not clear. One possibility to tackle this problem is to say that one is not interested in the exact values of the simulations but rather the temporal average of the nonlinear part of the solution. The different surpluses for the non averaged and averaged methods can be seen in figure 5.7.

This approach was expected to be more robust under the effect of fluctuations. However, while the surpluses for the averaged method seem to be much smaller than for the non averaged method both surpluses do not seem to have a specific trend and look rather noisy. The high values for the surpluses on the left are because the GENE cannot simulate the quantities of interests properly for only 8 grid points in the v dimension, which causes in a huge initial increase of the values up to a magnitude of 10^8, followed by a decay of similar size and large fluctuations.

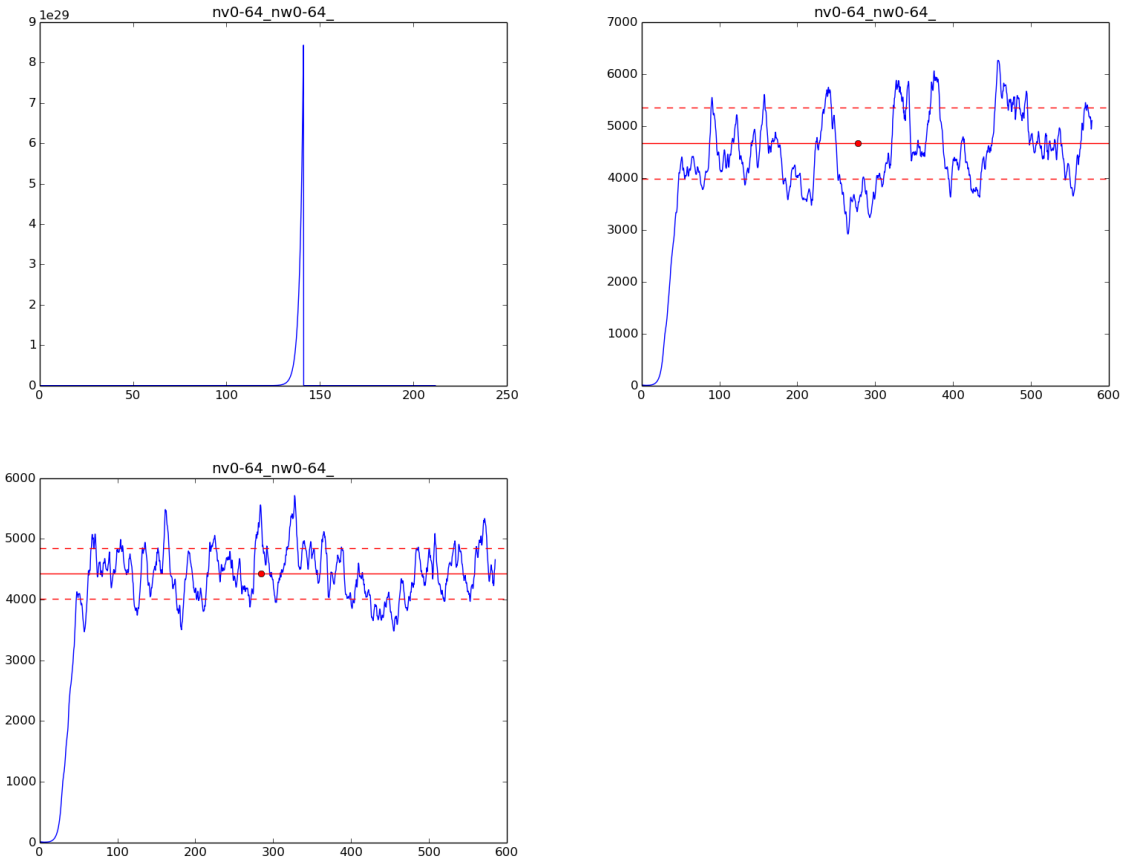

Figure 5.6: Top: Particle density for a linear (left) and nonlinear (right) GENE run with an equidistant μ-grid. Bottom: Particle density for a nonlinear Gene run with a Gauss-Laguerre grid. Red lines mark the average value for the time after the red dot. The red dashed line is the average plus minus the standard deviation for the same time interval.

While the surpluses in our experiments do not seem to converge due to very different results for different grid resolutions, ongoing experiments by Mario Heene suggest that for finer grids the solutions seem to converge. Here, further work is needed to verify this.

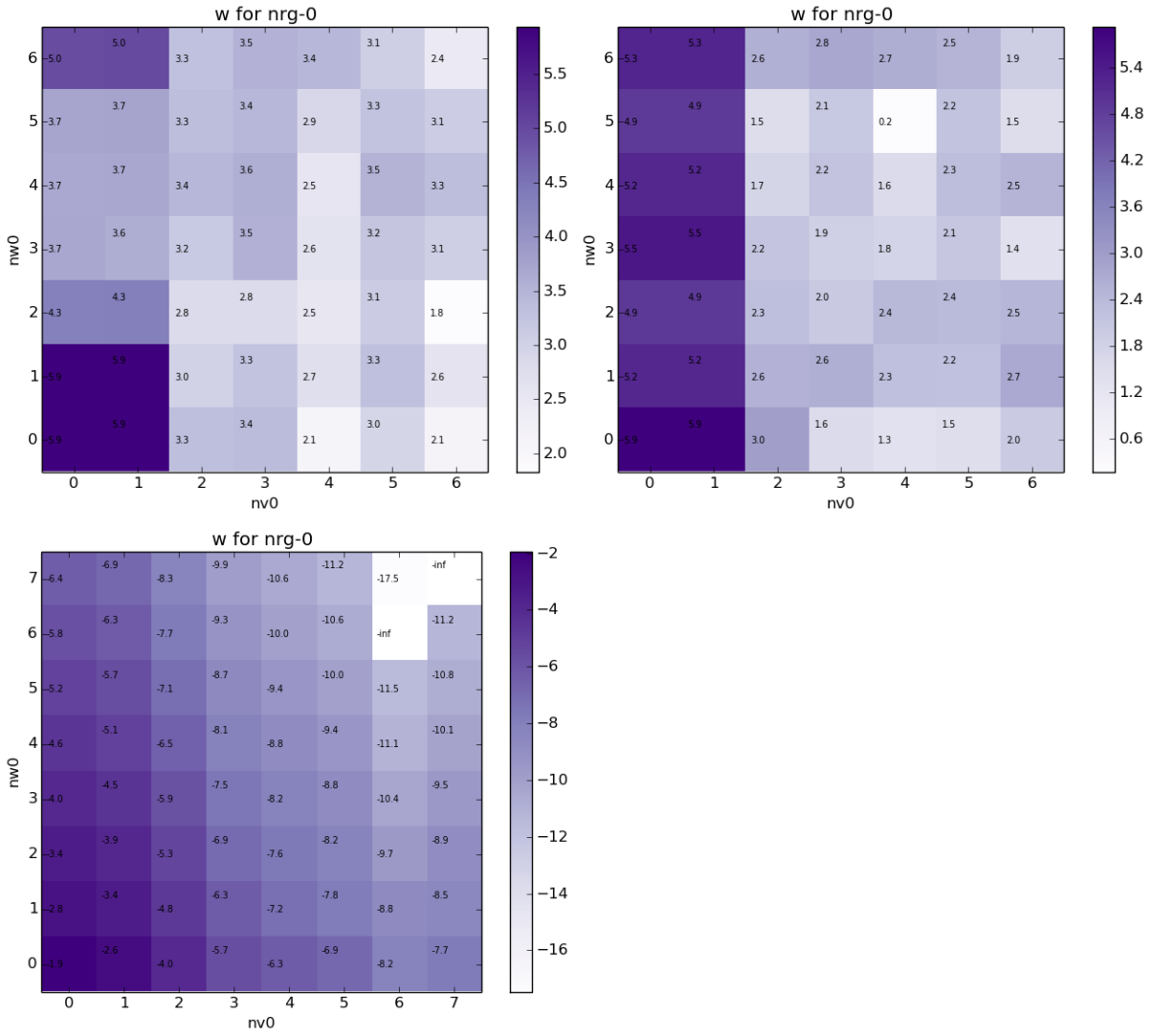

Figure 5.7: Top: Logarithmic surpluses for particle density for the nonlinear runs using a Gauss-Laguerre grid in the w dimension. Left: Sampling the last simulated value. Right: Using the average over the last 300s of simulated time.
Bottom: Logarithmic surpluses for particle density for the linear runs. One can see, that the surpluses are much more regular than the nonlinear runs.

6 Conclusion

In this work we presented an alternative view on the combination technique. While most existing literature defines the combination technique as linear combination of different solution functions, we took a slightly different approach. We skipped the constraint on the hierarchical properties and shifted the focus away from the solutions themselves towards the surpluses. Based on an error model we defined the combination technique to only include the largest surpluses. In this process we skipped small surpluses that do not contribute much to the solution. This reduced the number of grid points from $\mathcal{O}(h_n^{-2})$ to $\mathcal{O}(h_n^{-1} \log h_n^{-1})$ for two dimensional problems [BG04], but also introduced an additional error. To quantify this error we used the surpluses with different error models. In case of the trapezoidal rule we found that the additional error is of order $\mathcal{O}(h_n^2 \log h_n^{-1})$, which is not much worse than the full grid solution with an error of $\mathcal{O}(h_n^2)$. Our focus on quantities of interest also allowed to use extrapolation methods to improve the accuracy of our method. By combining the Romberg method with the combination technique we were able to improve the performance of our methods.

GENE proved itself to be a bigger problem for the combination technique. While we avoided high fluctuations by averaging the quantity of interest over time, large differences between the component grids remained. Hence, the quantity of interest did not show any convergence behavior. Further work is required to apply the combination technique to GENE and averaged quantities of interest.

In the theory part we focused on two dimensional integration problems an related error estimations. These insights are an example where profit from the combination technique. The error bounds and, more importantly, the general idea of the methods should be easily expansible to higher dimensions.

Using the concept of surpluses makes it more natural to define the combination technique for different decay models. It is intuitive to take the largest surpluses to get a good approximation for the full grid solution and simultaneously reduce the computational complexity. Furthermore, reasoning about the additional error simply comes down to summing up the surpluses excluded by the combination technique. This demonstrates that surpluses describe a key principle of the combination technique.

A Appendix

A.1 Parameter Files for GENE

A.1.1 Linear Run with Equidistant Grids

```
&parallelization
!n_procs_s =   1
!n_procs_v =   1
!n_procs_w =   8
!n_procs_x =   1
!n_procs_y =   1
!n_procs_z =   1
!n_procs_sim =      8
/

&box
n_spec =    1
nx0    =    24
nky0   =    1
nz0    =    24
nv0    =    96
nw0    =    16

kymin =   0.2000
lv    =     3.00
lw    =     9.00
adapt_lx = T
ky0_ind =   1
mu_grid_type = 'equidist'
/

&in_out
diagdir = './'
chptdir = './'

read_checkpoint  = F
write_checkpoint = T

istep_field =     100
istep_mom   =     100
istep_nrg   =      10
istep_omega =      20
istep_vsp   =     500
istep_schpt =    5000
istep_energy =     500

write_std = T
write_h5 = F
chpt_h5 = F
momentum_flux = F
/

&general
nonlinear =   F
comp_type = 'IV'
!perf_vec  =  2 2 1 1 1 2 1 1 2
!nblocks   =      16
arakawa_zv =   T
arakawa_zv_order =   4
hypz_opt=F
```

```
timescheme = 'RK4'
dt_max     =   1.0E-03
calc_dt = T

timelim    =      64500
ntimesteps =      100000
omega_prec =   0.1000E-03

beta       =    0.0000000
debye2     =    0.0000000
collision_op = 'none'

init_cond = 'alm'

hyp_z =   0.2500
hyp_v =   0.2000

perf_tsteps =  20

/

&geometry
magn_geometry = 's_alpha'
q0       =    1.4000000
shat     =   0.7960
trpeps   =   0.18000000
major_R  =    1.0000000
norm_flux_projection  = F
/

&species
name   = 'ions'
omn    =    2.2200000
omt    =    6.8900000
mass   =    1.0000000
temp   =    1.0000000
dens   =    1.0000000
charge =  1
/

&info
step_time  =     0.0009
number of computed time steps =     699
time for initial value solver =      0.637
!calc_dt = T  ! original
calc_dt = F
init_time =     2.7914
n_fields = 1
n_moms   =  6
lx =   4.18760
ly =   20.9440
PRECISION  = DOUBLE
ENDIANNESS = LITTLE
OMP_NUM_THREADS =   1
SVN_REV = 3943M
RELEASE = 1.6 - alpha 0
/

&units
/


&scan
scan_dims = 18
par_in_dir = './out/scanfiles0000/in_par'
/
```

A.1.2 Nonlinear Run with Equidistant Grids

```
&parallelization
n_procs_s =   1
n_procs_v =   1
n_procs_w =   8
n_procs_x =   1
n_procs_y =   1
```

```
n_procs_z =   1
n_procs_sim =      8
/

&box
n_spec =    1
nx0    =    8
nky0   =   16
nz0    =   24
nv0    =   64
nw0    =   64

kymin =   0.5000E-01
lv    =     3.00
lw    =     9.00
lx    =   125.628
nexc  =   5
adapt_lx = F
mu_grid_type = 'equidist'
/

&in_out
diagdir = './nv0-64_nw0-64_/'

read_checkpoint  = F
write_checkpoint = T

istep_field  =       0
istep_mom    =       0
istep_nrg    =      10
istep_vsp    =       0
istep_schpt  =       0

write_std = T
/

&general
nonlinear =   T
comp_type = 'IV'
perf_vec  =  2 2 1 1 1 1 1 1
nblocks   =   12288
arakawa_zv_order =   2

timescheme = 'RK4'
dt_max     =   0.4331E-01
courant    =     1.25

timelim    =      86000
ntimesteps =       4000
simtimelim =   0.1000E+05

beta       =    0.0000000
debye2     =    0.0000000
collision_op = 'none'

init_cond = 'ppj'

hyp_z =   -1.000
hyp_v =   0.2000

/

&geometry
magn_geometry = 's_alpha'
q0       =    1.4000000
shat     =   0.7960
trpeps   =   0.18000000
major_R  =    1.0000000
dpdx_term= 'gradB_eq_curv'
dpdx_pm  =    0.0000000
norm_flux_projection  = F
/

&species
name   = 'ions'
omn    =    2.2200000
omt    =    6.9600000

mass   =    1.0000000
temp   =    1.0000000
dens   =    1.0000000
charge =  1
```

```
/

&info
step_time  =     6.4009
number of computed time steps =    4000
time for initial value solver =  25603.373
calc_dt = T
hypz compensation = T
init_time =   189.0467
n_fields = 1
n_moms   =  0
nrgcols  = 10
```

```
lx =   125.628
ly =   125.664
PRECISION  = DOUBLE
ENDIANNESS = LITTLE
OMP_NUM_THREADS =   1
SVN_REV = Unversioned director
RELEASE = 1.8 - alpha 0
/

&units
/
```

A.1.3 Nonlinear Run with Gauss-Laguerre Grid in the μ-Dimension

```
&parallelization
n_procs_s =   1
n_procs_v =   1
n_procs_w =   9
n_procs_x =   1
n_procs_y =   1
n_procs_z =   2
n_procs_sim =     18
/

&box
n_spec =    1
nx0    =    8
nky0   =   16
nz0    =   24
nv0    =  512
nw0    =    9

kymin =   0.5000E-01
lv    =     3.00
lw    =     9.00
lx    =   125.628
nexc  =   5
adapt_lx = F
mu_grid_type = 'gau_lag'
```

```
/

&in_out
diagdir = './nv0-512_nw0-9_/'

read_checkpoint  = F
write_checkpoint = T

istep_field  =       0
istep_mom    =       0
istep_nrg    =      10
istep_vsp    =       0
istep_schpt  =       0

write_std = T
/

&general
nonlinear =   T
comp_type = 'IV'
perf_vec  =  2 1 2 1 1 1 2 1 2
nblocks   =     256
arakawa_zv_order =   2

timescheme = 'RK4'
```

```
dt_max     =   0.1851E-01
courant    =     1.25

timelim    =      86000
ntimesteps =      20000
simtimelim =   0.1000E+05

beta       =    0.0000000
debye2     =    0.0000000
collision_op = 'none'

init_cond = 'ppj'

hyp_z =   -1.000
hyp_v =   0.2000

/

&geometry
magn_geometry = 's_alpha'
q0       =    1.4000000
shat     =   0.7960
trpeps   =   0.18000000
major_R  =    1.0000000
dpdx_term= 'gradB_eq_curv'
dpdx_pm  =    0.0000000
norm_flux_projection  = F
/

&species
name   = 'ions'
omn    =    2.2200000
omt    =    6.9600000

mass   =    1.0000000
temp   =    1.0000000
dens   =    1.0000000
charge =  1
/

&info
step_time  =     0.9569
number of computed time steps =   20000
time for initial value solver =  19138.734
calc_dt = T
hypz compensation = T
init_time =    17.0150
n_fields = 1
n_moms   =  0
nrgcols  = 10
lx =   125.628
ly =   125.664
PRECISION  = DOUBLE
ENDIANNESS = LITTLE
OMP_NUM_THREADS =   1
SVN_REV = Unversioned director
RELEASE = 1.8 - alpha 0
/

&units
/
```

Bibliography

[AE07] H. Amann, J. Escher. *Analysis 1, dritte Auflage*. Birkhäuser Verlag: Basel-Boston-Berlin, 2007 (cit. on p. 30).

[BBNS12] J. Benk, H.-J. Bungartz, A.-E. Nagy, S. Schraufstetter. "Variants of the combination technique for multi-dimensional option pricing." In: *Progress in Industrial Mathematics at ECMI 2010*. Springer, 2012, pp. 231–237 (cit. on p. 44).

[BG04] H.-J. Bungartz, M. Griebel. "Sparse grids." In: *Acta numerica* 13 (2004), pp. 147–269 (cit. on pp. 40, 41, 69).

[BP12] J. Benk, D. Pflüger. "Hybrid parallel solutions of the Black-Scholes PDE with the truncated combination technique." In: *High Performance Computing and Simulation (HPCS), 2012 International Conference on*. IEEE. 2012, pp. 678–683 (cit. on p. 44).

[Dan05] T. Dannert. "Gyrokinetische Simulation von Plasmaturbulenz mit gefangenen Teilchen und elektromagnetischen Effekten." PhD thesis. Universität München, 2005 (cit. on p. 59).

[DR06] W. Dahmen, A. Reusken. *Numerik für Ingenieure und Naturwissenschaftler*. Springer-Verlag, 2006 (cit. on p. 18).

[ES81] B. Efron, C. Stein. "The jackknife estimate of variance." In: *The Annals of Statistics* (1981), pp. 586–596 (cit. on p. 45).

[Gar12] J. Garcke. "Sparse grids in a nutshell." In: *Sparse grids and applications*. Springer, 2012, pp. 57–80 (cit. on pp. 36, 40).

[GM13] *The Gyrokinetic Plasma Turbulence Code Gene: User Manual*. Gene Development Team. Aug. 2013 (cit. on pp. 8, 58, 59, 65).

[Gör10] T. Görler. "Multiscale effects in plasma microturbulence." PhD thesis. Universität Ulm, 2010 (cit. on p. 59).

[GSZ92] M. Griebel, M. Schneider, C. Zenger. "A combination technique for the solution of sparse grid problems." In: *Iterative Methods in Linear Algebra*. Ed. by P. deGroen, R. Buwens. IMACS. Elsevier, North Holland, 1992, pp. 263–281 (cit. on pp. 7, 25, 35, 40).

[Har16] B. Hardig. “Fault tolerant computation of hyperbolic PDEs with the sparse grid combination technique.” PhD thesis. The Australian National Univerity, 2016 (cit. on pp. 40, 44, 55).

[HGC07] M. Hegland, J. Garcke, V. Challis. “The combination technique and some generalisations.” In: *Linear Algebra and its Applications* 420.2 (2007), pp. 249–275 (cit. on pp. 8, 41).

[HHK+16] M. Hegland, B. Harding, C. Kowitz, D. Pflüger, P. Strazdins. “Recent developments in the theory and application of the sparse grid combination technique.” In: *Software for Exascale Computing-SPPEXA 2013-2015*. Springer, 2016, pp. 143–163 (cit. on pp. 25, 27, 39, 42, 45).

[Hic96] F. J. Hickernell. “Quadrature error bounds with applications to lattice rules.” In: *SIAM Journal on Numerical Analysis* 33.5 (1996), pp. 1995–2016 (cit. on p. 46).

[JDKR00] F. Jenko, W. Dorland, M. Kotschenreuther, B. Rogers. “Electron temperature gradient driven turbulence.” In: *Physics of Plasmas (1994-present)* 7.5 (2000), pp. 1904–1910 (cit. on p. 58).

[Lyn65] J. N. Lyness. “Symmetric integration rules for hypercubes. I. Error coefficients.” In: *Mathematics of Computation* 19.90 (1965), pp. 260–276 (cit. on pp. 13, 15).

[LZ06] F. Liu, A. Zhou. “Two-scale finite element discretizations for partial differential equations.” In: *Journal of Computational Mathematics* (2006), pp. 373–392 (cit. on p. 46).

[Pfl10] D. Pflüger. “Spatially adaptive sparse grids for high-dimensional problems.” PhD thesis. Verlag Dr. Hut, 2010 (cit. on pp. 40, 41, 46).

[Pri72] G. R. Price. “Extension of covariance selection mathematics.” In: *Annals of human genetics* 35.4 (1972), pp. 485–490 (cit. on p. 65).

[Rei04] C. Reisinger. “Numerische Methoden für hochdimensionale parabolische Gleichungen am Beispiel von Optionspreisaufgaben.” In: (2004) (cit. on p. 55).

All links were last followed on November 07, 2016.

Declaration

I hereby declare that the work presented in this thesis is entirely my own and that I did not use any other sources and references than the listed ones. I have marked all direct or indirect statements from other sources contained therein as quotations. Neither this work nor significant parts of it were part of another examination procedure. I have not published this work in whole or in part before. The electronic copy is consistent with all submitted copies.

---

place, date, signature