Christian Schitter

Distribution-Free Portfolio Strategies and Algorithms for Sequential Investment

MASTERARBEIT

zur Erlangung des akademischen Grades eines Diplom-Ingenieur

Masterstudium Finanz- und Versicherungsmathematik

Graz University of Technology

Technische Universität Graz

Betreuer:
Univ.-Prof. Dipl.-Ing. Dr.techn. Ernst Stadlober

Institut für Statistik

Graz, im Mai 2011

Christian Schitter

Distribution-Free Portfolio Strategies and Algorithms for Sequential Investment

MASTER THESIS

written to obtain the academic degree of a Master of Science (MSc)

Master programme financial and actuarial mathematics

Graz University of Technology

Advisor:
Univ.-Prof. Dipl.-Ing. Dr.techn. Ernst Stadlober

Institute of Statistics

Graz, May 2011

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Graz, am ....

(Unterschrift)

Statutory Declaration

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitely marked all material which has been quotes either literally or by content from the used sources.

....

date (signature)

Abstract

This master thesis is concerned with the growth-optimal portfolio framework for sequential investment. In this setting, one is interested in finding portfolio strategies that asymptotically achieve the best possible expected average growth rate in a set of reference a-priori portfolio strategies. After laying out important preliminaries, we present optimal strategies which work under rather general assumptions for the underlying processes of returns. Next, distribution-free algorithms for empirical growth-optimal portfolio selection are presented, using methods from nonparametric regression and prediction by expert advice. An application on real world commodity data (implemented in C++ code) shows the applicability and effectivity of each of these algorithms. The results are promising, especially for kernel and nearest neighbour algorithms.

Zusammenfassung

Diese Arbeit beschäftigt sich mit dem Konzept wachstumsoptimaler Portfoliostrategien für sequentielle Investments. Ziel in diesem Modell ist das Finden von Portfolios, die asymptotisch die bestmögliche erwartete durchschnittliche Wachstumsrate in einer festgelegten Menge von a-priori Portfoliostrategien erreichen. Nach der Präsentation wichtiger wahrscheinlichkeitstheoretischer Konzepte werden optimale Portfoliostrategien unter sehr allgemeinen Anforderungen an die zugrundeliegenden Returns präsentiert. Danach werden verteilungsfreie empirische Portfolioalgorithmen in diesem Modell erarbeitet, die unter anderem Ergebnisse aus der nichtparametrischen Regression und der Vorhersage mit Experten verwenden. Eine Anwendung auf Rohstoffkurse (implementiert in C++) zeigt die Vor- und Nachteile eines jeden dieser Algorithmen. Die Ergebnisse dieser Backtests sind insgesamt sehr vielversprechend, besonders für die Kernel und Nearest-Neighbour Algorithmen.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

The problem of choosing an optimal combination of investments has a long history closely related to the theory of gambling (like most of todays financial models). [Markowitz, 1970] developed a risk/return framework which found a lot of response by the academic and professional community. Another portfolio concept, the idea of growth-optimal investment for sequential investment decisions, appealed less to academics and has almost been neglected by practitioners, to some extent because calculating this portfolio is difficult. Advances in computer technology now allow to derive powerful algorithms to calculate growth-optimal portfolios, making this framework interesting again. These new algorithms have the special property that they do not assume a certain probability distribution on the price processes, but try to establish optimality conditions with as little assumptions as possible. The main focus of this thesis is on such distribution-free portfolio strategies and algorithms. A second aspect covered is a test of these methods on a set of commodity data to assess their applicability.

This leads to the following structure of the thesis: At the beginning, relevant general definitions and results from stochastics are presented, as well as the basics from nonparametric regression and prediction by expert advice, which are among the core distribution-free methods in other areas. In chapter 3, the concept of growth-optimality is introduced and important theoretical results are summarized. Special attention is drawn to the difference of strategies for dependent and independent returns. The main focus, again, lies on results that are independent of specific distributions. A critical discussion of the properties of growth-optimal portfolios finishes this part. Chapter 4 presents several selection algorithms utilizing previous ideas and results. Concluding, several algorithms are applied onto real world data, leading to a comparison of their benefits and drawbacks in chapter 5.

At this point, I am very grateful to my advisor, Prof. Dr. techn. Ernst Stadlober, who

patiently led me through the process of writing this thesis. I am thankful to Julia, who supported me so much in the last months.
Finally, I am especially grateful to my parents, Peter and Annemarie Schitter, without whom I would have never been able to undertake my studies. To them I dedicate this thesis.

Chapter 2

Preliminaries

2.1 Stochastic Definitions and Results

2.1.1 Stochastic Processes and Convergence

One of the main interests in stochastics today is to model the behaviour of a random variable over time. This is done by the theory of stochastic processes. The selection of definitions and results presented here is limited to time-discrete stochastic processes, as these are the ones of main interest for this thesis. A comprehensive and general introduction to stochastic processes can be found in [Kallenberg, 2002], on which this section also mainly relies.

Definition 2.1 (Stochastic Process)**.** *Define a probability space* $(\Omega, \mathcal{A}, P)$*, a* σ*-algebra* $\mathcal{Z}$ *on a space* Ξ *and a set of indices* $\mathcal{T}$*. A stochastic process is defined as a sequence of random variables* $X_t : \Omega \to \Xi$ *with* $t \in \mathcal{T}$ *such that for every* $t \in \mathcal{T}$ *the variable* X_t *is* $\mathcal{A}$*-*$\mathcal{Z}$*-measurable.*

The whole sequence and every of its sub-sequences have a (marginal) distribution. In a time-discrete setting, one usually uses $\mathcal{T} \subseteq \mathbb{Z}$, which will be assumed throughout the thesis. It should be noted, that the X_t can still be continuously distributed, which will usually be assumed in the following. To compare the equalities of two random variables, we use the concept of equality in distribution:

Definition 2.2 (Equality in distribution)**.** *Two random variables* X *and* Y *are said to be equal in distribution,* $X \stackrel{d}{=} Y$*, if*

$$P(X^{-1}(Z)) = P(Y^{-1}(Z)) \ \ \forall Z \in \mathcal{Z}.$$

We will also be interested in the asymptotic behaviour of functions of the stochastic process over time. In a stochastic setting, special types of convergence need to be defined in contrast to a deterministic setting, as the limit itself is usually stochastic:

Definition 2.3 (Convergence in distribution)**.** *A sequence of random variables X_n, $n = 1, 2, \ldots$, is said to be converging in distribution to X, $X_n \xrightarrow{d} X$, if*

$$\lim_{n\to\infty} P(X_n^{-1}(Z)) = P(X^{-1}(Z)) \ \ \forall Z \in \mathcal{Z}.$$

Definition 2.4 (Almost sure convergence)**.** *A sequence of random variables X_n, $n = 1, 2, \ldots$, is said to be converging almost surely to X, $X_n \xrightarrow{a.s.} X$, if*

$$P(\lim_{n\to\infty} X_n = X) = 1.$$

Remark 2.1. *Almost sure convergence is stronger than convergence in distribution in the sense that the former implies the latter, which is usually not true the other way round.*

Remark 2.2. *There are other types of stochastic convergence that are not mentioned here, as they are not relevant in this paper.*

2.1.2 Stationarity and Ergodicity

It is common to group classes of stochastic processes according to certain characteristics (like Markov processes or the Brownian motion). Among the most important classes (as there are a lot of fundamental results about them) is the rather general class of stationary and ergodic processes, that is briefly described in the following. Note that general properties of these processes do not depend on a specific underlying distribution. The following results are mainly collected from [Györfi et al., 2002] and [Kallenberg, 2002].

Definition 2.5 (Shift operator)**.** *The n-th shift operator θ^n on a vector x is defined as*

$$\theta^n(x_i, x_{i+1}, \ldots, x_{i+k}) = (x_{i+n}, x_{i+1+n}, \ldots, x_{i+k+n})$$

for arbitrary $i, k, n \in \mathbb{N}$.

Definition 2.6 (Stationary process)**.** *A stochastic process X_t is called stationary if for every selection of integers n, k and t*

$$\theta^n(X_t, X_{t+1}, \ldots, X_{t+k}) \stackrel{d}{=} (X_t, X_{t+1}, \ldots, X_{t+k}).$$

Remark 2.3. *This definition implies that sequences of equal length in this stochastic process always have the same distribution, no matter from which position of the process they are taken.*

Definition 2.7 (Measure-preserving operator)**.** *On a probability space $(\Omega, \mathcal{A}, P)$, an operator $T : \Omega \to \Omega$ is called measure-preserving, if for all $A \in \mathcal{A}$*

$$P(A) = P(T^{-1}A).$$

The following result shows a connection between discrete stationary random processes and measure-preserving operators.

Lemma 2.1. *For every stationary sequence X_t, $t \in \mathbb{Z}$, there is a random variable X and a measure-preserving operator T, such that*

$$X_t = X(T^t\omega).$$

Proof. Defining $Z_t = X_t(\omega)$, identify $\omega_t = X_t^{-1}(Z_t)$. Because of stationarity, $\theta^k X_t(\omega) = X_{t+k}(\omega) = Z_{t+k} \stackrel{d}{=} Z_t = X_t(\omega)$ for all $k \in \mathbb{Z}$ implies that $P(\omega_t) = P(\omega_{t+k}) = P(\theta^{-k}\omega_n)$. Therefore, the shift operator is measure-preserving for stationary processes. By choosing $Z = X_0$ and $T = \theta$, one immediately gets the result. □

Definition 2.8 (Ergodic process)**.** *A stationary process is called ergodic, if its inherent measure-preserving transformation T has the property that*

$$T^{-1}A = A \Rightarrow P(A) = 0 \vee 1 \ \ \forall A \in \mathcal{A}.$$

Definition 2.9 (Stationary and ergodic process)**.** *A stochastic process is called stationary and ergodic (s-a-e), if it has both the properties of stationarity and ergodicity together.*

There is a central result for the class of ergodic processes, Birkhoff's ergodic theorem, which can be seen as a law of large numbers for this class of processes and highlights the power and significance of s-a-e processes. A special case of this theorem for functions of s-a-e processes will be presented here as well:

Theorem 2.1 (Birkhoff's ergodic theorem for s-a-e processes)**.** *Let X_t be a stationary and ergodic process with $E(|X_1|) < \infty$. Then*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i = E(X_1).$$

Proof. See for example [Györfi et al., 2002]. □

Theorem 2.2 (Breimann's generalized ergodic theorem for s-a-e processes)**.** *Let* $X = \{X_t\}_{t=1,2,\dots}$ *be a s-a-e process and* θ^n *denotes the shift operator. Let* $f_1(.), f_2(.), \dots$ *be a sequence of real valued functions such that* $\lim_{n\to\infty} f_n(X) = f(X)$ *almost surely for some function* $f(.)$*. Assume* $E(\sup_n |f_n(X)|) < \infty$*. Then*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} f_i(\theta^i X) = E(f(X)).$$

Proof. See [Breiman, 1957]. □

Remark 2.4. *Stationarity and ergodicity indicates a strong dependence of a process over time. Due to the previously stated laws of large numbers, very powerful results can be achieved for models based on s-a-e processes by transforming similar results for the standard law of large numbers.*

Remark 2.5. *In practice, a lot of dependent time series that do not show a specific trend and/or pass certain statistical tests for stationarity are treated as stationary and ergodic for this reason.*

2.2 Prediction by Expert Advice

We will see that portfolio selection can be closely related to predicting the outcomes of a sequence. In machine learning, a similar problem regularly occurs (see for example [Cesa-Bianchi and Lugosi, 2006]). From a given set of observations, one wants to deduce - as general and accurate as possible - the next outcome. There is a whole range of theories and special adaptations of this task for certain problems. Prediction by expert advice is a generalization of many of these concepts and will be closer looked at here, serving as a resource of ideas for portfolio algorithms later on.

2.2.1 The Basic Model

Think of the following framework of predicting: A forecaster has to make a decision for an action according to a certain input. Each time the forecaster acts, it can evaluate the success of its action and thereby "learn" which reactions on which input where good and which were not. Optimally, this should lead to a forecaster with a good reaction to the given inputs after a certain learning period. If possible, the forecaster should automatically adapt when the required reactions for certain inputs change.

More formally, this can be described as follows: We want to sequentially predict a series of outcomes $y_1, y_2, \ldots$ of a finite outcome space $\mathcal{Y}$. The forecaster's predictions $\hat{p_1}, \hat{p_2}, \ldots$ belong to the convex decision space $\mathcal{D}$, that is often - but not necessarily - a subspace of $\mathcal{Y}$. To make these predictions, the forecaster has access to a set of experts $\{f_{i,t} \in \mathcal{D} : i \in \mathcal{I}\}$ for a given set of indices $\mathcal{I}$. He does not know how reliable the single experts are, but he calculates his own prediction according to those of the experts. After the real outcome is revealed, the forecaster can see how reliable the experts' and his own predictions were for this specific case. He weighs this reliability by a nonnegative loss function

$$l : \mathcal{D} \times \mathcal{Y} \to \mathbb{R}$$

usually used with either the prediction ($l(\hat{p}_t, y_t)$) or the experts ($l(f_{i,t}, y_t)$). According to the performance, he can choose to use the information given by the experts differently in the next prediction, thereby adopting to what he just learned.
The forecaster's goal is usually set to keep the cumulative regret for each forecaster i,

$$R_{i,n} = \hat{L}_n - L_{i,n} \tag{2.1}$$

with $\hat{L}_n = \sum_{t=1}^{n} l(\hat{p}_t, y_t)$ and $L_{i,n} = \sum_{t=1}^{n} l(f_{i,t}, y_t)$, as small as possible. Usually, this means that the loss of the prediction should converge in the mean towards the loss of the best expert

$$R_n = \frac{1}{n}\left(\hat{L}_n - \inf_{i \in \mathcal{I}} L_{i,n}\right) \overset{n\to\infty}{\longrightarrow} 0. \tag{2.2}$$

There are surprisingly many possibilities to find a mixing of the experts that fulfills (2.2), each with benefits and drawbacks.
Probably the simplest form of such a calculation is the weighted average forecaster, a simple convex combination of the experts:

$$\hat{p}_t = \frac{\sum_{i \in \mathcal{I}} w_{i,t} f_{i,t}}{\sum_{i \in \mathcal{I}} w_{i,t}}. \tag{2.3}$$

As $\mathcal{D}$ was assumed to be convex, the forecast is again in this space. Assume for now that $\mathcal{I} = \{1, 2, \ldots, N\}$ is finite. As the forecaster has access to a certain history of the expert's prediction, it makes sense to calculate the weights as a function of the past performance of the experts, $w_{i,t}(R_{t-1})$ with $R_{t-1} = (R_{1,t-1}, \ldots, R_{N,t-1})$. A very powerful choice for this function is the so called exponential weight, given by

$$w_{i,t}(R_{t-1}) = \frac{\exp(\eta R_{i,t-1})}{\sum_{j=1}^{N} \exp(\eta R_{j,t-1})}$$

for an arbitrary but positive parameter η. Using these weights in (2.3) together with substitution by (2.1) delivers

$$\hat{p}_t = \frac{\sum_{i=1}^{N} \exp(-\eta \sum_{j=1}^{t-1} l(f_{i,j}, y_i)) f_{i,t}}{\sum_{i=1}^{N} \exp(-\eta \sum_{j=1}^{t-1} l(f_{i,j}, y_i))} = \frac{\sum_{i=1}^{N} \exp(-\eta l(f_{i,t}, y_i)) w_{i,t-1} f_{i,t}}{\sum_{i=1}^{N} \exp(-\eta l(f_{i,j}, y_i)) w_{i,t-1}}. \tag{2.4}$$

For this forecaster, the following upper bound of regret can be derived:

Theorem 2.3 (Upper bound for exponentially weighted average forecasters)**.** *Assume that the loss function is convex in its first argument and that it takes values in* $[0, 1]$*. For any* n *and* $\eta > 0$ *and for all* $y_1, \ldots, y_n \in \mathcal{Y}$*, the regret of the exponentially weighted average forecaster satisfies*

$$R_n \leq \frac{\ln N}{\eta n} + \frac{\eta}{2}.$$

Proof. See [Cesa-Bianchi and Lugosi, 2006]. □

The upper bound is optimal for $\eta = \sqrt{\frac{2 \ln N}{n}}$, leading to

$$R_n \leq \sqrt{\frac{2 \ln N}{n}}$$

which is obviously converging to 0 and therefore fulfilling (2.2). Still, every other choice for η that leads to convergence is also feasible, like $\eta = 1$.

Generalizing this idea, one is interested in finding a forecaster in the class $\mathcal{P}$ of possible sequences of forecasters, with the smallest possible regret for a fixed loss function l and experts $f_{i,n}$ for $i \in \mathcal{I}$, the so called minimax regret $V_n(\mathcal{I})$,

$$V_n(\mathcal{I}) = \inf_{\mathcal{P}} \sup_{y_1^n \in \mathcal{Y}^n} \left(\sum_{t=1}^{n} l(\hat{p}_t(y_1^{t-1}), y_t) - \inf_{i \in \mathcal{I}} \sum_{t=1}^{n} l(f_{i,t}(y_1^{t-1}), y_t) \right).$$

An upper bound for the minimax regret shows the worst case regret of the forecasting strategy, while a lower bound shows the best case regret. In the following section, an explicit minimax regret of the logarithmic loss function in combination with the exponentially weighted average forecaster will be derived.

2.2.2 Logarithmic Loss and Mixture Forecaster

Let the outcome space be $\mathcal{Y} = \{1, \ldots, m\}$ and let the decision space be the so called probability simplex

$$\mathcal{D} = \left\{ p = (p(1), \ldots, p(m)) : \sum_{j=1}^{m} p(j) = 1, p(j) \geq 0, j = 1, \ldots, m \right\}.$$

We will also assume that m experts predict according to the known past of outputs at time t, y_1^{t-1}, assigning each outcome a probability:

$$f_t = (f_t(1|y_1^{t-1}), \ldots, f_t(m|y_1^{t-1})).$$

$f_t(i|y_1^{t-1})$ can be interpreted as the conditional probability of the occurrence of i at time t. The forecaster therefore chooses at each time instance t a probability vector

$$\hat{p}_t = (\hat{p}_t(1|y_1^{t-1}), \ldots, \hat{p}_t(m|y_1^{t-1})).$$

Continuing to interpret the experts and the forecaster as conditional probabilities, one can write

$$f_n(y_1^n) = \prod_{t=1}^{n} f_t(y_t|y_1^{t-1}) \text{ and } \hat{p}_n(y_1^n) = \prod_{t=1}^{n} \hat{p}_t(y_t|y_1^{t-1}).$$

Considering that the main operation here is multiplication, the choice of a loss function based on the logarithm is rather natural:

$$l(p, y) = \ln \frac{1}{p(y)} \quad y \in \mathcal{Y}, p \in \mathcal{D}. \tag{2.5}$$

This loss function is obviously aiming at assigning a high probability to the outcomes in the sequence. Replacing the loss function in (2.2) by (2.5) gives

$$R_n = \frac{1}{n}\left(\sum_{t=1}^{n} \ln \frac{1}{\hat{p}_t(y_t|y_1^{t-1})} - \inf_{i\in\mathcal{I}} \sum_{t=1}^{n} \ln \frac{1}{f_{i,t}(y_t|y_1^{t-1})}\right) = \sup_{i\in\mathcal{I}} \ln \frac{f_{i,n}(y_1^n)}{\hat{p}_n(y_1^n)}.$$

The forecaster that follows is the so called mixture forecaster:

$$\hat{p}_n(y_1^n) = \prod_{t=1}^{n} \frac{\sum_{i\in\mathcal{I}} f_{i,t}(y_1^t)}{\sum_{i\in\mathcal{I}} f_{i,t-1}(y_1^{t-1})}.$$

Using the obvious property that $\sum_{y_1^n \in \mathcal{Y}^n} f_{i,n}(y_1^n) = 1$ the former formula reduces to

$$\hat{p}_n(y_1^n) = \frac{1}{N} \sum_{i\in\mathcal{I}} f_{i,n}(y_1^n).$$

Here $f_{i,0}(y_0^0) = 1$. The mixture forecaster in this case is just the uniform mixture of the N experts. There is the following extension if $\mathcal{I}$, and therefore the number of experts, is countable. Choose π_i for each $i \in \mathcal{I}$ such that $\sum_{i\in\mathcal{I}} \pi_i = 1$. Then we can get a mixture forecaster for countable many experts by

$$\hat{p}_n(y_1^n) = \sum_{i \in \mathcal{I}} \pi_i f_{i,n}(y_1^n).$$

In the special case of the logarithmic loss function, the minimax regret can be found explicitly. Observe here, that

$$V_n(\mathcal{I}) = \inf_{\hat{p}} \sup_{y_1^n \in \mathcal{Y}^n} \ln \frac{\sup_{i \in \mathcal{I}} f_{i,n}(y_1^n)}{\hat{p}_n(y_1^n)}.$$

The following theorem delivers the desired forecaster.

Theorem 2.4 (Minimax regret for logarithmic loss functions)**.** *For any class of experts $f_{i,n}$, $i \in \mathcal{I}$, and integer $n > 0$, the so called likelihood forecaster*

$$\hat{p}_n^*(y_1^n) = \frac{\sup_{i \in \mathcal{I}} f_{i,n}(y_1^n)}{\sum_{x_1^n \in \mathcal{Y}^n} \sup_{i \in \mathcal{I}} f_{i,n}(x_1^n)}$$

is the unique forecaster such that

$$V_n(\mathcal{I}) = \sup_{y_1^n \in \mathcal{Y}^n} \ln \frac{\sup_{i \in \mathcal{I}} f_{i,n}(y_1^n)}{\hat{p}_n^*(y_1^n)}.$$

Moreover, $\hat{p}_n^*$ is an equalizer; that is, for all $y_1^n \in \mathcal{Y}^n$,

$$\ln \frac{\sup_{i \in \mathcal{I}} f_{i,n}(y_1^n)}{\hat{p}_n^*(y_1^n)} = \ln \sum_{x_1^n \in \mathcal{Y}^n} \sup_{i \in \mathcal{I}} f_{i,n}(x_1^n) = V_n(\mathcal{I}).$$

Proof. See [Cesa-Bianchi and Lugosi, 2006]. □

Knowing the minimax regret and the best forecaster in a minimax sense facilitates the analysis of a forecaster significantly, for example in the case of the Laplace mixture forecaster in the following example.

Example 2.1 (Laplace mixture forecaster)**.** *The Laplace mixture forecaster makes predictions on $\mathcal{Y} = \{1, 2\}$, and is defined by*

$$\hat{p}_n(y_1^n) = \int_0^1 q^{n_1}(1-q)^{n_2} dq$$

where n_1 and n_2 are the number of occurrences of 1 and 2 in y_1^n respectively. This means, we got a countable number of constant experts q and $1-q$ and mix them by taking the average according to the uniform distribution. For this forecaster, the following can be stated:

Theorem 2.5 (Minimax regret of Laplace mixture forecaster)**.** *The Laplace mixture forecaster satisfies*

$$\sup_{y_1^n \in \{1,2\}^n} \left(\hat{L}(y_1^n) - \inf_{i \in \mathcal{I}} L_{i,n}(y_1^n) \right) = \ln(n+1).$$

Proof. See [Cesa-Bianchi and Lugosi, 2006]. □

Remark 2.6. *Note that the methods described here are purely deterministic (even though the experts can be interpreted as probabilities in the logarithmic loss case). Yet, these methods can be used in a stochastic setting. Assume for example that the experts depend on stochastic observations and methods. This idea will be of use in the next section, but also later on.*

2.3 Nonparametric Regression

2.3.1 Nonparametric Regression for Independent and Identically Distributed Observations

In regression estimation, one is usually interested in defining a functional dependence between an observed vector X and a response variable Y, where the variables are all assumed to be independent and identically distributed (iid). This means finding a function depending on X that is a "good" approximation of Y. In a nonparametric setting, this requires a measurable estimator $m^*(X) : \mathbb{R}^d \to \mathbb{R}$ such that the so called L_2-risk is minimized (compare for example [Härdle, 1992]):

$$E(m^*(X) - Y) = \min_{f:\mathbb{R}^d \to \mathbb{R} \; measurable} E(|f(X) - Y|^2).$$

The well known solution to this problem is the regression function

$$m(x) = E(Y|X = x). \tag{2.6}$$

The difference to parametric regression estimation now lies in the fact, that the estimator for $m(x)$ is not dependent on a finite number of estimated parameters (like the moments of the underlying distribution), but directly on the whole history of observations. This means that for a given set of n observations $D_n = ((X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n))$, the regression function estimate $m_n(x)$ of Y_{n+1} depends on the current observation X_{n+1} and the whole history D_n, therefore $m_n(x) = m_n(X_{n+1}, D_n)$. This setting leads to more flexibility and a better integration of the available data compared to parametric methods. At this point, an estimator for $m_n(x)$ is needed, that either fulfills

- weak consistency: $\lim_{n\to\infty} E(\int (m_n(x) - m(x))^2 \mu(dx)) = 0$, or

- strong consistency: $\lim_{n\to\infty} \int (m_n(x) - m(x))^2 \mu(dx) = 0 \;\; a.s.$

for a given distribution, as this means that the estimator converges toward the regression function. If the estimator is weakly/strongly consistent for all distributions, it is called weakly/strongly universally consistent.
One of the standard methods to serve as an estimator is the kernel-based prediction method (which is more or less local averaging). Here, $m_n(x)$ can be written as

$$m_n(x) = \sum_{i=1}^{n} W_{n,i}(x) Y_i \tag{2.7}$$

for some weights $W_{n,i}(x) = W_{n,i}(X_{n+1}, D_n)$. The weight for Y_i is smaller, the "further" X_i is away from X_{n+1}. A simple idea for choosing the weights is a naive kernel estimation as follows:

$$m_n(D_n, X_{n+1}) = \frac{\sum_{i=1}^{n} \mathbb{1}_{\{\|X_i - X_{n+1}\| \leq r\}} Y_i}{\sum_{i=1}^{n} \mathbb{1}_{\{\|X_i - X_{n+1}\| \leq r\}}}.$$

In this formula, r is the so called bandwidth. Its value decides which observations are "similar" or "near" enough to the current one to be considered in the local averaging. A generalization of this idea is the so called Nadaraya-Watson kernel

$$m_n(D_n, X_{n+1}) = \frac{\sum_{i=1}^{n} K_r(X_{n+1} - X_i) Y_i}{\sum_{i=1}^{n} K_r(X_{n+1} - X_i)}$$

where $K_r(X - X_i)$ is an arbitrary function (the kernel function) for which the method is weakly or strongly (universally) consistent and converges - preferably as fast as possible - towards the regression function. Finding a useful kernel function is not an easy task, nor is deciding the size of the bandwidth $r > 0$. r is a kind of smoothing parameter, which is typical for many nonparametric regression estimates. Therefore, the regression estimate itself in such a case also depends on r: $m_{n,r}(x)$.
Choosing smoothing parameters is a rather difficult task. The possible, undesired effects of choosing r too large is over-smoothing, as many observations are considered in the local average that are actually too far away. On the other hand, choosing r too small results in under-smoothing, which means that there are often nearly no points over which to take the average. Compare figure 2.1 and 2.2 for a visual representation of this problem.
Clearly, there is a need for a sound procedure to select the smoothing parameter. Several ones have been proposed, mainly such which rely on estimating r from a "training sample" in a way that the fit of the estimate on the training sample is good (for extensive coverage of these aspects, see [Györfi et al., 2002]). A more recent, alternative approach, that does

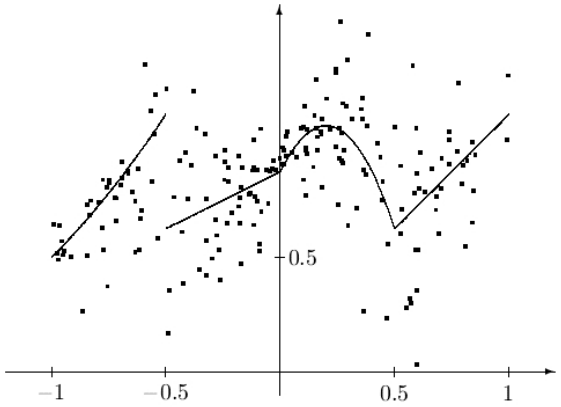
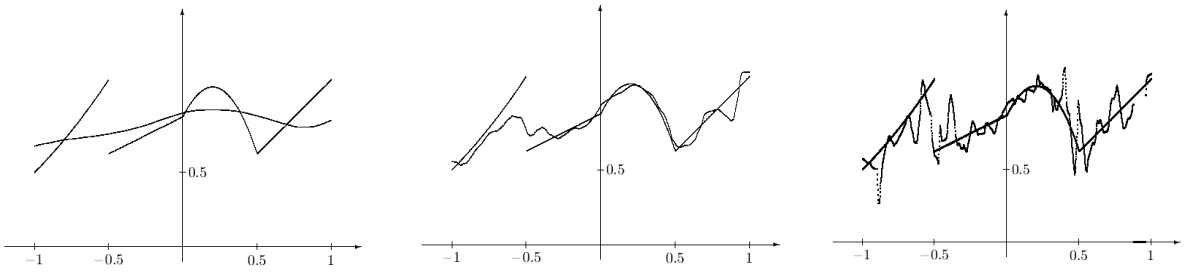


Figure 2.1: Simulated data with original regression function. *Source: [Györfi et al., 2002].*



(a) Over-smoothing (b) Good smoothing (c) Under-smoothing

Figure 2.2: Kernel regression estimate for figure 2.1 with increasing values for r. *Source: [Györfi et al., 2002].*

not need an a-priori estimation for r or a training sample, will be introduced later in the case of dependent observations.
However, even with the challenges at hand it can be shown that the following rather general result holds, which assures the user of the usefulness of this method:

Theorem 2.6 (Weak consistency of kernel estimator for iid processes)**.** *Assume that the observations are iid and that there are balls $S_{0,r}$ of radius r and balls $S_{0,R}$ of radius R centered at the origin ($0 < r \leq R$), and constant $b > 0$ such that*

$$\mathbb{1}_{\{x \in S_{0,R}\}} \geq K(x) \geq b\mathbb{1}_{\{x \in S_{0,r}\}}$$

for the Kernel $K(x)$, and consider the kernel estimate $m_n(x)$ introduced before with a bandwidth function r_n depending on the number of observations n. If $r_n \to 0$ and $nr_n^d \to \infty$, then the kernel estimate is weakly consistent.

Proof. See [Györfi et al., 2002]. □

Another paradigm is the so called k-nearest neighbour (k-NN) estimator. Here one applies again local averaging, but only on the k "nearest" points in the known history. More specifically, reorder the history D_n to $D_{(n)}(x) = ((X_{(1)}, Y_{(1)}), \ldots, (X_{(n)}, Y_{(n)}))$ according to increasing values of $\|X_i - x\|$. With this, define the k_n-nearest neighbour estimate by

$$m_n(x) = \frac{1}{k_n} \sum_{i=1}^{k_n} Y_{(i)}(x).$$

In our previous setup, this means taking again $x = (X_{n+1}, D_n)$.
If the kth ordered point has the same distance from x as the $(k+1)$th, one has a tie and should therefore declare an applicable rule for choosing the appropriate point (like using the one whose original index is the highest for example). In a theoretical setting, it is usually sufficient to assume that ties occur with probability 0. This is done in the following theorem, which proves the universal consistency of this concept:

Theorem 2.7 (Weak consistency of k-NN estimator)**.** *If $k_n \to \infty$ and $\frac{k_n}{n} \to 0$ then the k_n-NN estimator is weakly consistent for all distributions of (X, Y) where ties occur with probability 0 and $E(Y^2) < \infty$.*

Proof. See [Györfi et al., 2002]. □

Remark 2.7. *k-NN and Kernel estimators follow a similar concept, but have different implications. It usually depends on the problem and data at hand to decide which of the two is the better choice. The advantage of the kernel-based estimator lies in the fact, that one can assign a different importance to historic observations according to their distance from the current observation by choosing an appropriate kernel function. Furthermore, points that are "too far" away can be excluded, which should result in a better fit. However, for little data, it can often happen that few or even no historic observations are close enough to be considered, thus making the estimation obsolete. The k-NN estimator on the other hand does exactly the opposite: While it always produces a valid estimation (as it uses exactly k observations for the local average) it has the disadvantage of probably using points too far away to serve as a good prediction. Precise analysis of the data and a comparison of both methods therefore usually proves to be useful before either one is finally applied.*

2.3.2 Nonparametric Sequential Prediction for Dependent Observations

While in the classical regression case one is confronted with iid observations, in several applications (especially in finance and economics) this assumption is too restrictive. In

these areas, one usually looks at time series and wants to predict the next outcome from the known past, taking into account possible dependencies over time (see for example [Biau et al., 2010]). A combination of kernel estimation and expert advice will prove to deliver a very robust method for this kind of time series prediction problems.
Consider the time series $x_1^n = (x_1, x_2, \ldots, x_n)$. One is interested to predict x_{n+1} from the known past with the predictor $g_n(x_1^n)$. Ideally, one wants to minimize the cumulative squared prediction error

$$L_n(g) = \frac{1}{n}\sum_{t=1}^{n}(g_t(x_1^n) - x_{t+1})^2.$$

[Algoet, 1994] shows that the best bound L^* for any sequential prediction strategy g of s-a-e processes is given by

$$\liminf_{n\to\infty} L_n(\hat{g}) \geq L^* = E((X_1 - E(X_1|X_{-\infty}^0))^2) \quad a.s. \tag{2.8}$$

Usually, the lower bound cannot be reached for finite n, but one can find strategies, such that

$$\lim_{n\to\infty} L_n(\hat{g}) = L^* \quad a.s.$$

with respect to the class of all s-a-e processes. This means that if the underlying time series is generated by a s-a-e process, the strategy asymptotically achieves the best lower bound. Combining the basic ideas of the last sections, one can define the following sequential prediction strategies:
Define an array of experts $f_n^{(k,l)}$, where k and l are positive integers, and the function

$$T_a(z) = \begin{cases} a & \text{if } z > a \\ z & \text{if } |z| \leq a \\ -a & \text{if } z < -a \end{cases}$$

Define further a radius $r_{k,l}$ for each pair (k,l) such that for fixed k

$$\lim_{l\to\infty} r_{k,l} = 0. \tag{2.9}$$

Then the kernel-based expert $f_n^{(k,l)}$ at time n can be defined as

$$f_n^{(k,l)}(x_1^n) = T_{\min(n^\delta,l)}\left(\frac{\sum_{\{k<t<n\}} K_{r_{k,l}}(\|x_{t-k+1}^t - x_{n-k+1}^n\|)x_t}{\sum_{\{k<t<n\}} K_{r_{k,l}}(\|x_{t-k+1}^t - x_{n-k+1}^n\|)}\right)$$

for a kernel function $K_r(x)$.

Further define an arbitrary probability distribution $\{q_{k,l}\}$, arbitrary $\eta_n > 0$ and the weights

$$w_{k,l,n} = q_{k,l} \exp\left(-\eta_n (n-1) L_{n-1}(f_{n-1}^{(k,l)})\right).$$

Then the following holds:

Theorem 2.8 (Asymptotic optimality of the sequential kernel prediction for s-a-e processes)**.** *Choose*

$$\eta_n = \frac{1}{\sqrt{n}}.$$

Assuming that (2.9) is verified, the prediction scheme for the kernel weights defined above running over all integers (k, l),

$$g_n(x_1^n) = \frac{\sum_{k,l=1}^{\infty} w_{k,l,n} f_n^{(k,l)}(x_1^n)}{\sum_{i,j=1}^{\infty} w_{i,j,n}}$$

asymptotically achieves the lower bound from (2.8) for the class of all s-a-e processes where $E(X_0^4) < \infty$.

Proof. See [Biau et al., 2010]. □

Remark 2.8. *The previous result holds if the x_i are vectors by just using the Frobenius norm in the Kernel function.*

In quite the same way, we can define a nearest-neighbour based estimator with similar properties. Consider again experts $f_n^{(k,l)}$ for integers k and l. This time, choose $p_l \in (0, 1)$ such that

$$\lim_{l \to \infty} p_l = 0 \tag{2.10}$$

and set $\bar{l} = \lfloor p_l n \rfloor$ where $\lfloor . \rfloor$ is the floor function. Introduce the set of the $\bar{l}$ nearest neighbours

$$\begin{aligned} \mathcal{J}_{k,l,n} &= \{k < t < n : x_{t-k+1}^t \text{ is among the } \bar{l} \text{ observations in} \\ &\quad (x_1^k, \ldots, x_{n-k}^{n-1}) \text{ with smallest distance to } x_{n-k+1}^n\}. \end{aligned}$$

Then the experts are defined for $n > k + \bar{l} + 1$ by

$$f_n^{(k,l)}(x_1^n) = T_{min(n^\delta,l)} \left(\frac{1}{\bar{l}} \sum_{j \in \mathcal{J}_{k,l,n}} x_{j+1} \right)$$

for $\delta > 0$ if the sum is nonvoid and 0 otherwise. As before, define an arbitrary probability distribution $\{q^{(k,l)}\}$, $\eta_n > 0$ and weights

$$w_{k,l,n} = q_{k,l} \exp\left(-\eta_n (n-1) L_{n-1}(f_{n-1}^{(k,l)})\right).$$

With these definitions, the following theorem holds:

Theorem 2.9 (Asymptotic optimality of the sequential NN prediction for s-a-e processes)**.** *Choose*

$$\eta_n = \frac{1}{\sqrt{n}}.$$

Assuming that (2.10) is verified, the prediction scheme for the nearest neighbour weights defined above running over all integers (k, l),

$$g_n(x_1^n) = \frac{\sum_{k,l=1}^{\infty} w_{k,l,n} f_n^{(k,l)}(x_1^n)}{\sum_{i,j=1}^{\infty} w_{i,j,n}}$$

asymptotically achieves the lower bound from (2.8) for the class of all s-a-e processes where $E(X_0^4) < \infty$.

Proof. See [Biau et al., 2010] □

Remark 2.9. *These sequential prediction strategies for time series can of course be extended by looking at s-a-e time series that have observation and response variables as in the general regression case. This means defining predictors $g_n(x_1^n, y_1^{n-1})$ in the same way as before to predict y_n. All of the above results hold for this case (see [Biau et al., 2010] and [Györfi and Schäfer, 2003]).*

Remark 2.10. *In an applied setting one uses a finite set of experts, restricting (k, l) by $k = 1, \ldots, K$, $l = 1, \ldots, L$.*

Chapter 3

Growth-Optimality

After laying out basic ideas and important results from other areas, we are now turning our attention towards portfolio theory. Here, one basically wants to "optimally" distribute money among a given choice of assets under different constraints. The fractions invested in every single asset need to be predicted, which was the reason for investigating this topic earlier. We will now lay out what is considered "optimal" investment in the growth-optimal framework and present core results of this theory. But first, we will look at the game of coin tossing to understand the idea of growth-optimal strategies in an easy setting.

3.1 The Kelly Strategy

It is well known that stochastics evolved mainly out of the interest in analyzing gambling situations and trying to find optimal strategies for certain games. [Kelly, 1956] is supposed to be the first to have introduced the growth-optimal strategies to sequential gambling by adopting the concept of information rate from signal transforming. The question he had in mind was the following: If a player gambles several times in a row in a favorite game (that is, a game in which the expected return rate is positive), without drawing any money away from the game, what is the optimal gambling strategy to maximize his expected wealth in the long run? The objective was therefore to find a "portfolio" of two "assets", where $0 \leq b^{(1)} \leq 1$ is the fraction of the gamblers money he should bet at each try and $b^{(2)} = 1 - b^{(1)}$ is the money he should keep back.
To grasp the concept, assume that the game is tossing coins with an unfair coin (thus making it a favourable game), where the probability of head is given by $0.5 < p < 1$ and that of number is given by $q = 1 - p$. The gambler receives double the amount of money he bets if he gets head and looses the money he bets otherwise. He is allowed to play

as many games as he likes with his initial capital S_0. Now, how much money should the gambler bet in each try on head? If he bets all his money ($b^{(1)} = 1$) all the time, he will be broke almost surely at some point, namely when the coin shows number for the first time. We first write down his fortune S_n after n tries

$$S_n = S_0(2b^{(1)} + 1 - b^{(1)})^H(1 - b^{(1)})^L = S_0(1 + b^{(1)})^H(1 - b^{(1)})^L$$

where H and L are the numbers of wins and losses in the n bets. Looking at this representation of wealth after n tries, one sees that wealth will grow exponentially fast, that is

$$\frac{S_n}{S_0} = \exp(nW_n)$$

where W_n is called the average growth rate. It is prudent to try to maximize $E(W_n)$ instead of $E(S_n)$ in the long run, which in this sense means asymptotically. A simple argument is that maximizing $E(S_n)$ means betting everything on head all the time, leading to bankruptcy with probability one. Therefore, we calculate

$$W = \lim_{n\to\infty} \frac{1}{n} \ln\left(\frac{S_n}{S_0}\right) = \lim_{n\to\infty} \left(\frac{H}{n} \ln(1 + b^{(1)}) + \frac{L}{n} \ln(1 - b^{(1)})\right)$$

which satisfies

$$W = p \ln(1 + b^{(1)}) + q \ln(1 - b^{(1)})$$

with probability one by the law of large numbers. The maximum with respect to $b^{(1)}$ is obtained straightforward by differentiating

$$\frac{d}{db^{(1)}} p \ln(1 + b^{(1)}) + q \ln(1 - b^{(1)}) = \frac{p}{1 + b^{(1)}} - \frac{q}{1 - b^{(1)}} \stackrel{!}{=} 0.$$

Solving for $b^{(1)}$ with respect to the fact that $1 + b^{(1)} + 1 - b^{(1)} = 2$ delivers

$$1 + b^{(1)} = 2p$$
$$1 - b^{(1)} = 2q$$

which is equivalent to $b^{(1)} = 2p - 1$ or $b^{(1)} = 1 - 2q$ respectively. This strategy is thus asymptotically the best strategy for this game if one can play infinitely many games as it is generating the best average growth rate

$$W^* = p\ln(2p) + q\ln(2q)$$

in the long run. This partition of wealth is called the Kelly strategy and achieves in this game the highest expected average rate of growth. Such a strategy is therefore also called growth-optimal. A nice consequence of this strategy is that the gambler will never bet all of his money, thus not risking ruin.

3.2 The Growth-Optimal Portfolio Model

The main idea of [Kelly, 1956] was that of considering a model with exponential growth for sequential gambling situations, as all winnings are reinvested every time. [Breiman, 1960] expanded this idea to sequential investments in a portfolio selection context. In this setting, one is concerned with the problem of optimally distributing at time n a given amount of money S_n among d assets, whose prices at time n are denoted by

$$s_n = (s_n^{(1)}, \cdots, s_n^{(d)}).$$

One is especially interested in the vector of growth factors

$$x_n = (x_n^{(1)}, \cdots, x_n^{(d)})$$

which represents the growth of the assets over one period, given by

$$x_n^{(j)} = \frac{s_n^{(j)}}{s_{n-1}^{(j)}}.$$

Note that growth factors and returns of an investment are not the same. The return $r_n^{(i)}$ of an asset is the percentage increase, therefore $r_n^{(i)} = x_n^{(i)} - 1$. If we now denote the vector of the fractions of the money invested in the assets at time n (with no short selling allowed) by

$$b_n = (b_n^{(1)}, \cdots, b_n^{(d)}) \in \mathcal{B} = \left\{ (b^{(1)}, \ldots, b^{(d)}) : 0 \leq b^{(i)} \leq 1, \sum_{i=1}^{d} b^{(i)} = 1 \right\}$$

the given amount of money that results from a trading strategy $B = \{b_i\}_{i=1,2,\ldots}$ after one period can be represented as

$$S_1(B) = \sum_{j=1}^{d} b_1^{(j)} s_1^{(j)}. \tag{3.1}$$

By using the vector of growth factors and the fact that

$$S_{k+1}(B) = S_k(B) \sum_{j=1}^{d} b_{k+1}^{(j)} x_{k+1}^{(j)}$$

the wealth after n periods can be represented by

$$S_n(B) = S_0 \prod_{i=1}^{n} \sum_{j=1}^{d} b_i^{(j)} x_i^{(j)} = S_0 \prod_{i=1}^{n} \langle b_i, x_i \rangle \tag{3.2}$$

where $\langle ., . \rangle$ denotes the standard inner product and S_0 is the given initial wealth. Note that we simply write $S_n(b)$ if the strategy is constant over time with $B = \{b\}_{i=1,2,\ldots}$.
The objective of a portfolio strategy in general is to find a series of a-priori portfolio vectors that should usually maximize the return of the portfolio under given constraints. Different setups have been proposed, some taking into account constraints to limit risk, some taking into account time constraints. The proposed setups also differ in assumptions over the underlying distribution of the process that drives s_n. The objective of the portfolio strategies in this thesis is to achieve so called growth-optimality, in a sense as derived by [Kelly, 1956] for gambling. In the case of investment, the exponential growth of the sequential strategy is implied by the behaviour of risk-free assets that grow by sequential compounding (compare also [Luenberger, 1998]). No one would invest money without having a high chance to gain wealth on average, that is $E(\langle b_i, x_i \rangle) > 1$. This makes investing a "favourable game". Therefore the price process grows exponentially on average,

$$s_n^{(j)} = e^{nW_n^{(j)}} \tag{3.3}$$

which gives the so called average growth rate $W_n^{(j)}$,

$$W_n^{(j)} = \frac{1}{n} \ln s_n^{(j)} \tag{3.4}$$

for asset j and

$$W_n = \frac{1}{n} \ln S_n$$

for the portfolio. Note that a favourable game is equivalent to having a positive asymptotic average growth rate. Using representation (3.2), the asymptotic average growth rate W of the portfolio can be derived as

$$W = \lim_{n\to_\infty} W_n = \lim_{n\to\infty} \frac{1}{n} \ln S_n = \lim_{n\to\infty} \left(\frac{1}{n} \ln S_0 + \frac{1}{n} \sum_{i=1}^{n} \ln \langle b_i, x_i \rangle \right) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \ln \langle b_i, x_i \rangle .$$

Assume now that we can recalibrate the portfolio vector each day based on the information of the known past, $b_n = b(x_1^{n-1})$ at time n. Then, we are looking for a strategy $B^* = \{b_i^*\}_{i=2,3,\ldots}$ such that

$$\limsup_{n\to\infty} \frac{1}{n} \ln \frac{S_n(B)}{S_n(B^*)} = \limsup_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \ln \frac{\langle b_i, x_i \rangle}{\langle b_i^*, x_i \rangle} \leq 0 \quad a.s.$$

for any strategy B that is different from B^*. This means that in the long run, there is no admissible strategy B that can beat the strategy B^* almost surely. An equivalent definition is given by

$$\limsup_{n\to\infty} \frac{1}{n} \ln S_n(B) \leq \limsup_{n\to\infty} \frac{1}{n} \ln S_n(B^*) \ \ a.s.$$

and also by

$$\limsup_{n\to\infty} \left(\frac{1}{n} \ln(S_n(B)) - \frac{1}{n} \ln(S_n(B^*)) \right) \leq 0 \ \ a.s.$$

Such a choice B^* of portfolio vectors is called the growth-optimal portfolio (sometimes also log-optimum or Kelly portfolio). As this definition usually is too general and we do not necessarily want to choose specific distributions of the returns to investigate, it is useful to restrict results to a certain choice of the underlying processes of X_t, thus giving rise to the following definition:

Definition 3.1 (Universal consistency)**.** *A portfolio strategy B^* is called universally consistent with respect to a class $\mathcal{C}$ of stochastic processes $\{X_n\}_{n=-\infty}^{\infty}$ if for each process in the class and every strategy B that is different from B^**

$$\limsup_{n\to\infty} \frac{1}{n} \ln \frac{S_n(B)}{S_n(B^*)} \leq 0 \ \ a.s.$$

The more general the class can be kept, the better. From a similar point of view, one can also be interested in the behaviour of a strategy B with respect to a set of reference strategies $\{\bar{B}_i\}_{i\in\mathcal{I}}$ with $\mathcal{I}$ being the set of indices of the strategies. This implies the following definitions:

Definition 3.2 (Deterministic superiority)**.** *We call a strategy B to be deterministically superior with respect to a reference set of strategies $\{\bar{B}_i\}_{i\in\mathcal{I}}$ and a class of processes $\mathcal{C}$, if*

$$\limsup_{n\to\infty} \sup_{\bar{B}\in\{\bar{B}_i\}_{i\in\mathcal{I}}} \left(\frac{1}{n}\ln(S_n(\bar{B})) - \frac{1}{n}\ln(S_n(B))\right) \leq 0$$

for all processes in $\mathcal{C}$*.*

Definition 3.3 (Stochastic superiority)**.** *We call a strategy* B *to be stochastically superior with respect to a reference set of strategies* $\{\bar{B}_i\}_{i\in\mathcal{I}}$ *and a class of processes* $\mathcal{C}$ *if*

$$\limsup_{n\to\infty} \sup_{\bar{B}\in\{\bar{B}_i\}_{i\in\mathcal{I}}} \left(\frac{1}{n}\ln(S_n(\bar{B})) - \frac{1}{n}\ln(S_n(B))\right) \leq 0 \ \ a.s.$$

for all processes in $\mathcal{C}$*.*

Remark 3.1. *It is obvious that if a strategy is stochastically superior with respect to all possible strategies and a class* $\mathcal{C}$*, it is also universally consistent with respect to all processes in* $\mathcal{C}$*.*

3.3 Universal Consistency for Independent and Identically Distributed Returns

[Breiman, 1961] further investigates the application of growth-optimality to gambling with independent and identically distributed (iid) outcomes, coming to conclusions that can be transferred to growth-optimal investments for iid returns. Note that if the returns are iid, the growth factors are as well and vice versa. He arrives at a universally consistent strategy for iid processes by stating the following result:

Theorem 3.1 (Universally consistent portfolio strategy for iid returns)**.** *Let the vectors of growth factors* X_i *be iid and* $E((\ln\langle b, X_i\rangle)^2) < \infty$*. Then the growth-optimal constant portfolio vector* b^* *is given by*

$$b^* = \arg\max_{b\in\mathcal{B}} E(\ln\langle b, X_i\rangle)$$

and the resulting strategy is universally consistent with respect to all iid processes (or equivalently: with respect to all processes that are memoryless). The resulting maximal asymptotic average growth rate is

$$W^* = E(\ln(\langle b^*, X_i\rangle)) \ \ a.s.$$

for arbitrary i*.*

Proof. ([Györfi et al., 2007]) Look at the constantly rebalanced portfolio strategy $B = \{b_i = b\}_{i=1,2,...}$ and its average growth rate

$$\frac{1}{n} \ln S_n(b) = \frac{1}{n} \sum_{i=1}^{n} \ln \langle b, X_i \rangle = \frac{1}{n} \sum_{i=1}^{n} E(\ln \langle b, X_i \rangle) + \frac{1}{n} \sum_{i=1}^{n} (\ln \langle b, X_i \rangle - E(\ln \langle b, X_i \rangle)).$$

By the strong law of large numbers, the last part tends to zero almost surely,

$$\sum_{i=1}^{n} (\ln \langle b, X_i \rangle - E(\ln \langle b, X_i \rangle)) \to 0 \ \ a.s.$$

as the returns are iid and $E((\ln \langle b, X_i \rangle)^2) < \infty$. This implies that maximizing the average growth rate

$$\lim_{n \to \infty} \frac{1}{n} \ln S_n(B) = E(\ln(\langle b, X_i \rangle) \ \ a.s.$$

is equivalent to maximizing $E(\ln \langle b, X_i \rangle)$ for any i. Consequently,

$$b^* = \arg \max_{b \in \mathcal{B}} E(\ln \langle b, X_i \rangle)$$

and the maximal average growth rate is

$$\lim_{n \to \infty} \frac{1}{n} \ln S_n(b^*) = E(\ln \langle b, X_i \rangle) \ \ a.s.$$

□

Remark 3.2. *This result has the consequence that even if the portfolio vector can be chosen dynamically at each time step, the growth-optimal portfolio vector is constant over time for iid returns.*

In an applied setting for iid returns, the theorem implies that the problem of calculating a growth-optimal portfolio can be adressed by estimating the expected logarithm of growth factors of the strategy, which will be useful later on.

3.4 The Importance of Using the Logarithm

In the last sections, growth-optimality was defined by finding a portfolio strategy that delivers asymptotically the best expected average growth rate of all possible strategies. But why do we want to maximize the growth rate, that is more or less the logarithm of wealth, and not wealth itself? The reason for this lies in the fact, that $S_n(b)$ ($B = b$

constant here) is not close to $E(S_n(b))$ if S_n grows exponentially, that is $E(S_n) = E(e^{nW_n})$, as shown in [Györfi et al., 2007].
The last proof indicates that together with the iid property

$$\begin{aligned} 1 &= \lim_{n\to\infty} P\left(-\delta < \frac{1}{n}\ln S_n(b) - E(\ln\langle b, X_i\rangle) < \delta\right) \\ &= \lim_{n\to\infty} P\left(e^{n(-\delta+E(\ln\langle b,X_i\rangle))} < S_n(b) < e^{n(\delta+E(\ln\langle b,X_i\rangle))}\right) \end{aligned}$$

for any $\delta > 0$. This means that $S_n(b)$ is close to $\exp(nE(\ln\langle b, X_i\rangle))$. Looking at

$$E(S_n(b)) = E\left(\prod_{i=1}^{n}\langle b, X_i\rangle\right) = \exp(n\ln\langle b, E(X_i)\rangle)$$

shows that $E(S_n(b))$ is close to $\exp(n\ln\langle b, E(X_i)\rangle)$ on the other hand. Applying Jensen's inequality,

$$\exp(n\ln\langle b, E(X_i)\rangle) > \exp(nE(\ln\langle b, X_i\rangle))$$

implies that $S_n(b)$ is less than $E(S_n(b))$. This shows that maximizing the average growth rate is the right thing to do in the iid case. The same holds for the general case, though the proof is much more difficult (compare again [Györfi et al., 2007]). In the setting of sequential investment with exponential growth, this means that maximizing the average growth rate leads to a much higher (in fact infinitely higher) performance than any strategy that maximizes wealth itself in the long run.

3.5 Universal Consistency for Stationary and Ergodic Returns

Things become more complicated if returns are not iid any more. Here it is usually not possible to achieve universal consistency. However, for a special kind of highly dependent returns, namely s-a-e returns, we can indeed find a strategy that is universal with respect to this class of processes. That means that the optimal portfolio strategy given the known past performs at least equally as well as the portfolio strategy given the full (even unknown) past for s-a-e processes in the long run. This is shown in [Algoet and Cover, 1988] and [Algoet, 1994].
If we assume dependence among the returns, it makes sense to look at conditional expectations. Transforming the strategy for iid returns, this results in the following theorem:

Theorem 3.2 (Universally consistent strategy for s-a-e returns)**.** *Choose the portfolio strategy* $B^* = \{b_i^*\}_{i=1,2,\dots}$ *that sequentially invests at time* i *according to*

$$b_i^* = \arg\sup_{b\in\mathcal{B}} E(\ln\langle b, X_i\rangle \,|X_1, \cdots, X_{i-1}) \tag{3.5}$$

where $(X_1, \cdots, X_{i-1})$ *is the known past. Then this strategy is universally consistent with respect to the class of all s-a-e processes. Furthermore, for s-a-e returns, this strategy achieves the maximal expected growth rate*

$$\lim_{n\to\infty} \frac{1}{n} S_n(B^*) = W^*$$

where

$$W^* = E(\max_{b\in\mathcal{B}} E(\ln(\langle b(X_{-\infty}^{-1}), X_0\rangle)|X_{-\infty}^{-1}))$$

given the information about the full past.

Proof. [Algoet and Cover, 1988] or [Györfi et al., 2007]. □

Corollary 3.1. *From the previous theorem, it follows immediately that the universally consistent strategy with respect to stochastic processes with iid realisations is constant (as these processes are memoryless) and this fact reinforces theorem 3.1.*

With the previous results, the problem of finding a universally consistent growth-optimal portfolio for the class of s-a-e processes can be adressed by finding an algorithm that evaluates (3.5) for any s-a-e process.

3.6 Stochastic Superiority for General Returns

In the general case, when no restrictions on the processes generating the returns are assumed, universal consistency cannot be established by any strategy any more. But there is still a possibility to receive a rather strong result: The last section showed that sequentially maximizing the conditional expected logarithm of growth factors given the known past leads to achieving the maximal expected return knowing the infinite past in the s-a-e case. Now we state a result due to [Algoet and Cover, 1988] that the same strategy beats any other strategy that acts on the same information.

Theorem 3.3 (Stochastic superior strategy for any process)**.** *Let* $\mathcal{F}_t = \sigma(X_1, \dots, X_t)$ *be the sigma-algebra of the known past. Then the strategy* $B^* = \{b_i^*(\mathcal{F}_i)\}_{i=1,2,\dots}$ *with*

$$b_i^*(\mathcal{F}_i) = \arg\sup_{b\in\mathcal{B}} E(\ln\langle b, X_{i-1}\rangle \,|\mathcal{F}_i)$$

is stochastically superior with respect to every other strategy $b_i(\mathcal{F}_i)$ *and every process, therefore*

$$\limsup_{n\to\infty} \frac{1}{n} \ln \frac{S_n(B)}{S_n(B^*)} \leq 0 \;\; a.s.$$

Here,

$$\left\{ \bar{S}_n(B, B^*) = \frac{S_n(B)}{S_n(B^*)} \right\}_{0\leq n<\infty}$$

is a non-negative submartingale, that is

$$E(\bar{S}_t(B, B^*)|\bar{S}_s(B, B^*) = \bar{s}) \geq \bar{s} \quad \forall s < t.$$

$\bar{S}_n(B, B^*)$ *is converging almost surely to a random variable* Y *with* $E(Y) \leq 1$.

Proof. See [Algoet and Cover, 1988]. □

Remark 3.3. *This theorem shows that even though in the general case we cannot guarantee to achieve the highest possible growth rate of every strategy, we can at least count on the fact that it is impossible to find a better strategy that relies on the same information about the past.*

3.7 Critical Discussion

The growth-optimal portfolio theory is purely concerned with asymptotically maximizing wealth in sequential investment. This means that one needs to have the possibility to "play" as many games as one wants. In this setting, the theory guarantees that there is no better strategy for this "game" if the strategy is universal or that a proposed strategy is better than a set of strategies if it is superior. Even though this seems to be a reasonable objective for a portfolio strategy, this theory was controversially discussed among economists and widely ignored by practitioners for a long time. This chapter critically adresses these discussions in short, thereby providing valuable insights into the properties and also shortfalls of the growth-optimal portfolio. A good overview of these points and the academic work done in the field of growth-optimality can be found for example in [Christensen, 2005].

3.7.1 Growth-Optimal versus Efficient Portfolios

In the 1970s, Markowitz's mean-variance portfolio was the dominating portfolio theory by far. This framework assumed that only expected return and variance (or equivalently the standard deviation) of the returns are important in making a portfolio choice. A strategy b'_t is said to dominate another strategy b''_t in this framework if

$$E(S_t(b'_t)) > E(S_t(b''_t)) \text{ and } Var(S_t(b'_t)) \leq Var(S_t(b''_t))$$

or

$$E(S_t(b'_t)) \geq E(S_t(b''_t)) \text{ and } Var(S_t(b'_t)) < Var(S_t(b''_t)).$$

If a portfolio $\hat{b}_t$ is not dominated by any other portfolio, it is said to be efficient. Usually, this means that there are infinitely many possible portfolio choices that are efficient, and each one is as good as the others. The results from section 3.4 already imply, that this portfolio choice can usually not coincide with the growth-optimal portfolio, as it tries to maximize the expected wealth instead of the expected growth rate, which we saw to be two completely different objectives.

Now, it needs to be said that Markowitz was purely concerned with investments over a single period, which is obviously a different objective than the sequential investment framework of the growth-optimal portfolio. But it is usually true, that larger investors recalibrate their investments very often. Therefore, it is interesting to look at a comparison of the efficient and the growth-optimal portfolio for sequential investment. This was investigated in a simple example by [Hakansson, 1971], which is recapitulated in short in the following.

Example 3.1 ([Hakansson, 1971])**.** *Assume that there are two risky assets with*

$$x_t^{(1)} = \begin{cases} 0 & \textit{with probability 0.1} \\ 1.5 & \textit{with probability 0.9} \end{cases}$$

and

$$x_t^{(2)} = \begin{cases} 1.15 & \textit{with probability 0.9} \\ 2.65 & \textit{with probability 0.1} \end{cases}$$

Furthermore,

$$P(x_t^{(1)} = 0, x_t^{(2)} = 1.15) = 0.1,$$

$$P(x_t^{(1)} = 0, x_t^{(2)} = 2.65) = 0,$$

$$P(x_t^{(1)} = 1.5, x_t^{(2)} = 1.15) = 0.8,$$

$$P(x_t^{(1)} = 1.5, x_t^{(2)} = 2.65) = 0.1.$$

The relevant variances and covariances are given by

$$Var(x_t^{(1)}) = Var(x_t^{(2)}) = 0.2025$$

and

$$Cov(x_t^{(1)}, x_t^{(2)}) = 0.0225.$$

Therefore, the expected growth and standard deviation of the portfolio can easily be calculated by

$$E(S_t(b_t)) = E(b_t^{(1)} x_t^{(1)} + b_t^{(2)} x_t^{(2)}) = 1.35 b_t^{(1)} + 1.30 b_t^{(2)}$$

and

$$\sqrt{Var(S_t(b_t))} = \sqrt{0.2025(b_t^{(1)^2} + b_t^{(2)^2}) + 0.045 b_t^{(1)} b_t^{(2)}}.$$

Figure 3.1 shows the expected growth versus standard deviation of all admissible portfolios, starting with $b_t = (0, 1)$ on the outer left going to $b_t = (1, 0)$ on the outer right of the curve. The efficient portfolios are marked in blue.
The growth optimal portfolio is obtained by maximizing

$$\begin{aligned} E(\ln S_t(b_t)) &= E(\ln(b_t^{(1)} x_t^{(1)} + b_t^{(2)} x_t^{(2)})) \\ &= E(\ln(b_t^{(1)}(x_t^{(1)} - x_t^{(2)}) + x_t^{(2)})) \\ &= 0.1 \ln(1.15 - 1.15 b_t^{(1)}) + 0.8 \ln(1.15 + 0.35 b_t^{(1)}) + 0.1 \ln(2.65 - 1.15 b_t^{(1)}). \end{aligned}$$

This is simply done by differentiating with respect to $b_t^{(1)}$ and solving for the root. The resulting growth optimal portfolio is given by the unique solution of

$$0.35075 - 1.07237 b_t^{(1)} + 0.462875 b_t^{(1)^2} = 0$$

that lies between 0 and 1. This is the vector $b_t^* = (0.3941, 0.6058)$, indicated by a black dot in figure 3.1.

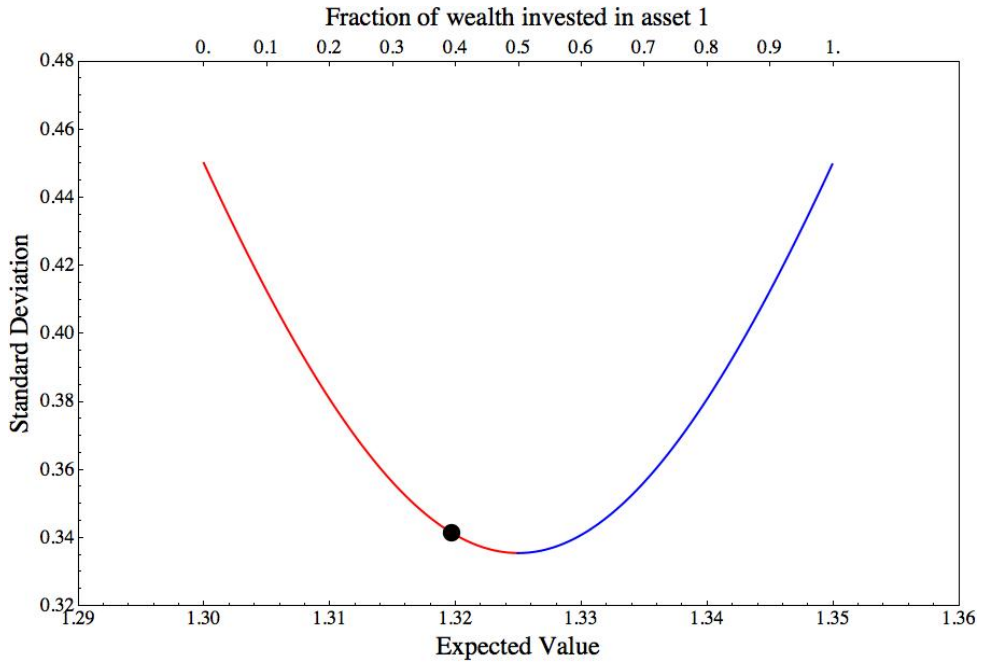


Figure 3.1: Expected growth versus standard deviation of all admissible portfolios in example 3.1. *After: [Hakansson, 1971].*

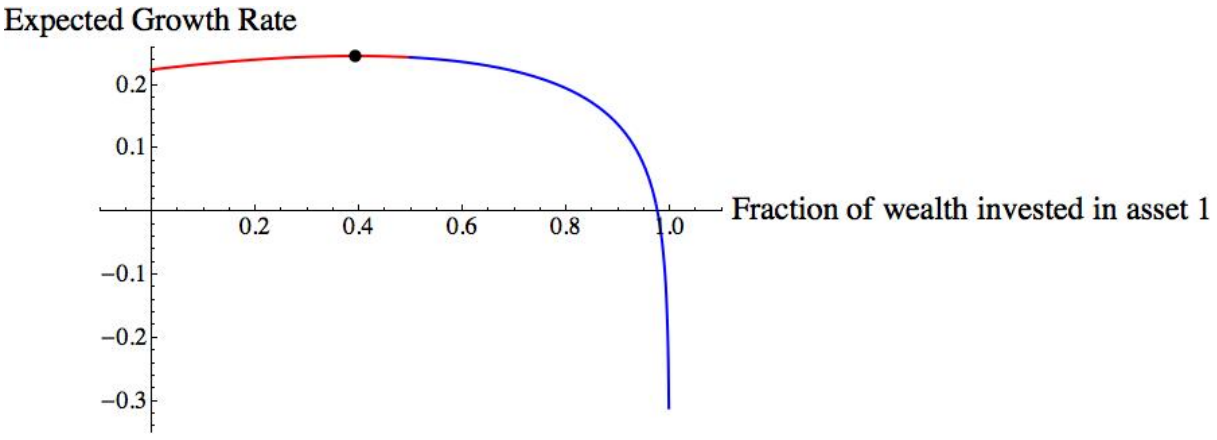


Figure 3.2: Expected average growth rate of all admissible portfolios in example 3.1.

Analyzing this example, the following conclusions on the relationship between efficient and growth-optimal portfolios can be obtained:

- The growth-optimal portfolio does not necessarily need to be efficient, and vice versa. In this example, the growth-optimal portfolio is in fact not efficient. What is even more, the "worst" portfolio in the mean variance sense, namely the one with all money in asset 2, has still a higher expected growth rate (about 22.3%) than most of the efficient portfolios (portfolios with more than 68.9% in asset 1 have a lower and sometimes even negative growth rate than that of the worst not efficient portfolio). This can easily be seen from figure 3.2.
- Efficient portfolios can lead to ruin almost surely as they can have negative expected growth rates. In this example, this involves all portfolios where more than 97.6% of wealth is invested in asset 1, as can be seen in figure 3.2. For those, ruin in the long run is almost sure, as

  $$P(S_n = 0) \to 1 \text{ for } n \to \infty$$

  by the law of large numbers.

It should once again be noted, that the mean-variance approach is explicitly derived for a single period investment decision. Still, this requirement is very rarely discussed and this framework is often used when being clearly in a situation of sequential investment. The previous analysis clearly points out that a growth-optimal approach is more desirable in a sequential investment framework for the long run than that of the mean-variance theory.

3.7.2 Growth-Optimality in the Context of Utility Theory

Utility theory is probably one of the most important principles in economics today. We quickly summarize important concepts here, for an extensive coverage of utility theory we refer to common microeconomics textbooks like [Pindyck and Rubinfeld, 2009].
In short, utility theory states that a consumer assigns each possible basket X containing an arbitrary amount of L commodities a certain value, expressed by a personal utility function $U(X)$. X can therefore be represented by a vector $(x_1, \ldots, x_L)$ where the i-th element represents the amount of commodity i in the basket. A basket X_1 is said to be preferred to X_2, if $U(X_1) > U(X_2)$, here denoted by the relation $X_1 \succ X_2$. A consumer is said to be indifferent between two baskets X_1 and X_2 if both have the same utility $U(X_1) = U(X_2)$ which we denote by $X_1 = X_2$. A utility function always implies a system

of preferences. A set of preferences represented by the relation $\succeq$ has the property that it is

- **complete**: Every two possible baskets can be compared, that is we have either $X_1 \succeq X_2$ or $X_2 \succeq X_1$ (or both which is $X_1 = X_2$)
- and **transitive**: If $X_1 \succeq X_2$ and $X_2 \succeq X_3$ then $X_1 \succeq X_3$.

Furthermore, there are the following requirements for standard utility functions:

- $U(X)$ is assumed to be monotonically increasing in every commodity, as the consumer prefers more to less.
- $U(X)$ is assumed to be concave, as this reflects the fact that the utility added decreases with the amount of goods possessed. (Getting one apple when having one apple increases my utility enormously. But getting one additional apple when already possessing 10,000 apples is no big deal and therefore increases my utility only a little bit.)

The utility function in financial sciences usually refers to the utility to be gained from a portfolio choice with payout $\langle b, X_i \rangle$. Here, the properties of the utility function reflect the risk awareness of the investor:

- The investor prefers more money to less, therefore the monotonicity of the utility function.
- The average investor is assumed to regret loosing one unit of money more than he enjoys winning one unit. Therefore, the utility function is concave.

These seem to be agreeable assumptions, as - depending on the individual preferences - an investor might not enter an investment with positive expected return if the probability to loose a lot of money is in his view too high. The most important utility functions used are the logarithm and the square root, but obviously there is an infinite choice of possibilities.
Connecting this with the concept of growth-optimality, we implicitly state that we can ignore the individual preferences of the investor in this framework. This is based on the fact that on average we will make more money with this than with any other investment strategy and the law of large numbers assures us to reach this goal.
Contrary to this argumentation, [Samuelson, 1963] and [Samuelson, 1971] show in a simple model that arguing with the law of large numbers is not necessarily consistent with the

axioms of utility theory presented earlier. More precisely, he considers the game of coin tossing, where one looses one Euro if the guess is wrong and receives two Euros if the guess is right. It seems reasonable that there are people who do not accept this bet as they fear more loosing one Euro than they enjoy winning two Euros. But in a long row of bets, the fortune of the gambler would obviously explode by the law of large numbers as he wins one Euro on average per game. So, the more chances the gambler gets to bet, the more likely he will enter the bet. That is, he should prefer more tries to less.
This can be translated into the following preference system:

$$X_1 \succ X_2 \text{ if } P(\text{Gain of } X_1 > \text{Gain of } X_2) > \frac{1}{2}.$$

If strategy X_i represents betting i times, this should lead to $X_i \succ X_{i+1}$ for all $i \geq 1$. Still, depending on the gamblers risk awareness and how the game is designed, he could still refrain from betting two times, as this still bears to much risk for him, making $X_1 = X_2$ - a contradiction to the previous statement! In this spirit, one can always think of a way to construct a game which a given risk averse investor will not enter for a certain number of tries. This shows that even though the law of large numbers "guarantees" to win more than with another strategy, it still depends on the preferences of the gambler if this strategy is his personal favourite choice.
In a similar argument [Samuelson, 1971] states that a strategy that dominates in the long run, does not necessarily dominate every strategy in a shorter time horizon.
[Markowitz, 1976] argued against Samuelson by stating - in yet another simple model - that no long-term investor should have any other than the log-utility function, which is the only one consistent with growth-optimality (as the average growth rate is exactly the log-utility of the portfolio choice). It should be stressed here that both arguments are valid in their respective frameworks and no final conclusion was reached by either one on this topic. However, the theory of growth-optimality is specifically proposed to be independent of utility theory. It simply states that a rational investor wants to achieve the highest growth rate in the long run, not caring about individual preferences. This is probably a bit exaggerated for private investors, but big institutional investors like hedge and pension funds find this objective probably reasonable enough.

3.7.3 How Long is the "Long Horizon"?

A question that arises naturally when talking about asymptotic strategies is: "How long is a long period of time here?". Do I need to have a horizon of 20, 30 or 100 years to be rather sure to attain more wealth by using this strategy than the others?

To investigate this matter, [Thorp and Beach, 2006] propose to look at a market with one asset and a risk free interest rate and to approximate the behaviour of the discrete returns by a continuous process. The money invested in the savings account grows by an interest rate r with probability 1, while the asset returns $X = \mu \pm \sigma$ with equal probability 0.5 (that is, $E(X) = \mu$ and $Var(X) = \sigma^2$). In this case, the expected average growth rate is given by

$$E(\ln(1 + b^{(1)}r + b^{(2)}X)) = 0.5\ln(1 + b^{(1)}r + b^{(2)}(\mu + \sigma)) + 0.5\ln(1 + b^{(1)}r + b^{(2)}(\mu - \sigma)).$$

Subdividing each time step into n independent time steps while keeping expectation and variance proportional leads to an asset growth of $X_i = \frac{\mu}{n} \pm \frac{\sigma}{\sqrt{n}}$, again with probability 0.5 each.
The accumulated wealth after one "original" time step is then given by

$$S_1(b) = S_0 \prod_{i=1}^{n} \left(1 + b_t^{(1)} \frac{r}{n} + b_t^{(2)} X_i\right).$$

This is exactly the well known case of the binomial model in option pricing (as for example presented in [Hull, 2006]) that converges toward a lognormal diffusion process for $n \to \infty$ with drift μ and variance σ^2 for the asset part and growth r for the savings account part. The expected growth rate of this portfolio is

$$W = E\left(\lim_{n\to\infty} \ln\left(\frac{S_1(b)}{S_0}\right)\right) = r + b^{(2)}(\mu - r) - \frac{\sigma^2 b^{(2)^2}}{2}$$

per time step for which μ and σ where originally defined. Note that by this approximation rebalancing would need to be done "instantaneously" which is impossible in reality. Therefore, it can only be seen as an approximation for the problem - but with valuable insights. Maximizing this growth rate leads to

$$b^* = \left(1 - \frac{\mu - r}{\sigma^2}, \frac{\mu - r}{\sigma^2}\right)$$

giving

$$W^* = r + \frac{1}{2}\frac{(\mu - r)^2}{\sigma^2}.$$

Note that in this setup short selling (that is $b^{(i)} \in \mathbb{R}$) is allowed. In this case, the return on the growth-optimal portfolio (the logarithm of S_n^*) is normally distributed and therefore nW^* has expectation

$$n\mu^* = n\left(r + \frac{1}{2}\frac{(\mu - r)^2}{\sigma^2}\right)$$

and standard deviation

$$\sqrt{n}\sigma^* = \sqrt{n}\frac{\mu - r}{\sigma}.$$

If one is interested in the probability that the growth optimal portfolio reaches a certain growth k in a given time N, we can simply standardize NW^* and use the standard normal distribution function $\Phi(.)$

$$P\left(\frac{NW^* - N\mu^*}{\sqrt{N}\sigma^*} > \frac{k - N\mu^*}{\sqrt{N}\sigma^*}\right) = 1 - \Phi\left(\frac{k - N\mu^*}{\sqrt{N}\sigma^*}\right). \tag{3.6}$$

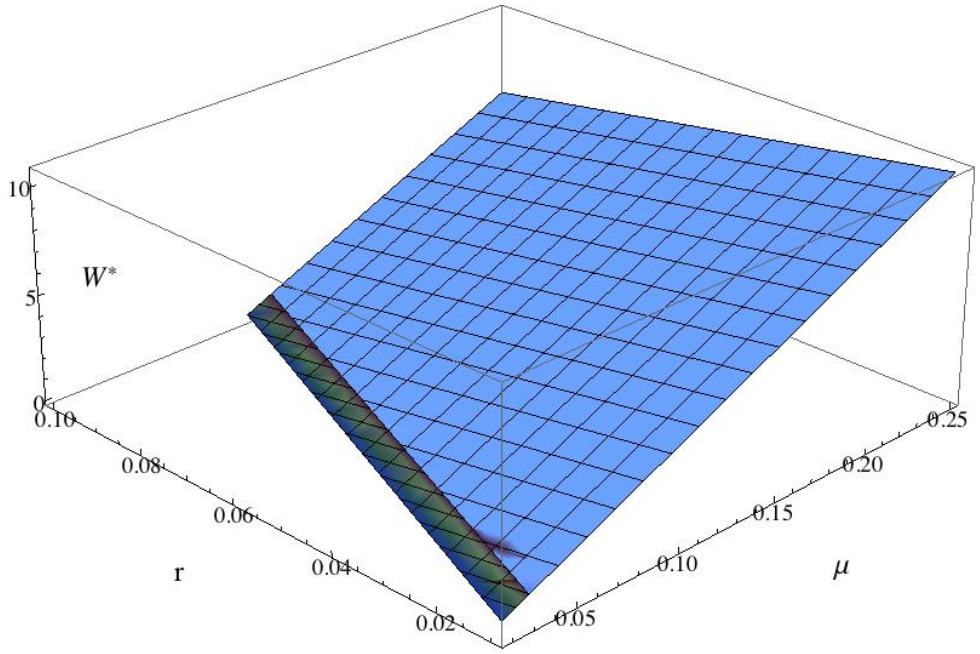


Figure 3.3: Optimal growth rate depending on r and μ.

Example 3.2. *Assume a yearly interest rate $r = 0.02$, an expected yearly growth rate of the asset of $\mu = 0.1$ and yearly standard deviation $\sigma = 0.15$.*
Then, in the given model, the probability that we achieve double wealth $e^k = 2$ within the next $N = 10$ years (which corresponds to around 2000 trading days) is almost 71%.
If we want to determine the time needed to achieve $e^k = 2$ with a given probability, we simply need to solve for N in (3.6). To have a 95% chance of reaching $k = \ln(2)$ the investor needs 37.3 years in this setting.
Figures 3.3, 3.4 and 3.5 give further insight into the behaviour of this example. The dark areas in the figures represent the optimal portfolios that do not require short selling.

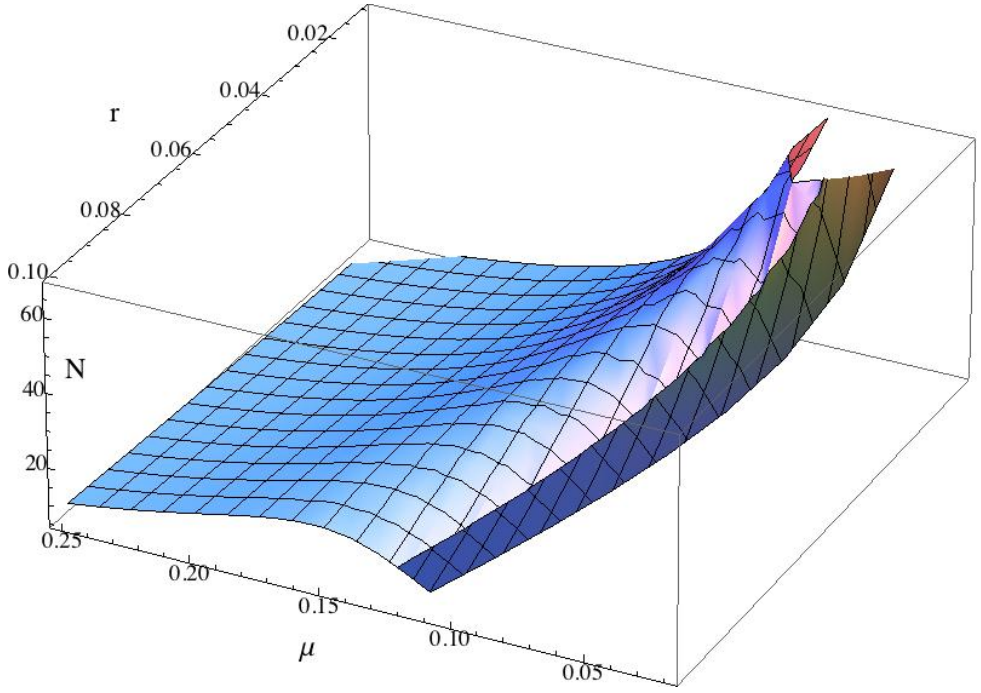


Figure 3.4: Time to have a 95% chance to achieve $e^k = 2$ depending on r and μ.

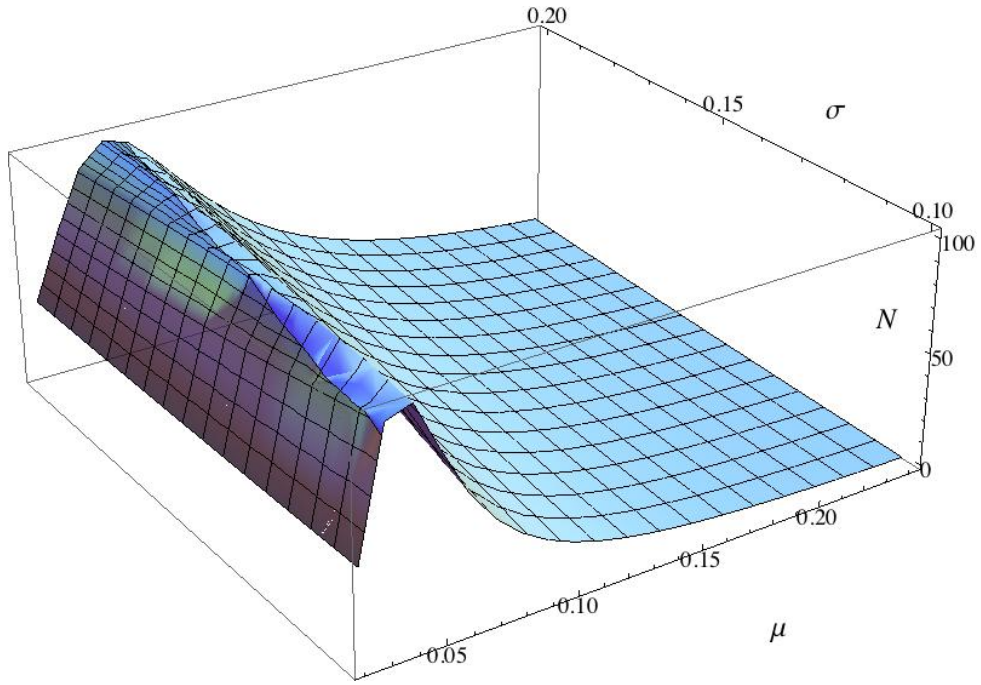


Figure 3.5: Time to have a 95% chance to achieve $e^k = 2$ depending on μ and σ.

[Rubinstein, 1998] also dealt with this topic and answered the question in the context of the famous Capital Asset Pricing Model (CAPM). For a reasonable choice of parameters (again in the case of one risky and one riskless asset) he arrives at the conclusion, that it needs over 105 years to have a 90% chance to outperform an account that grows with the risk-less interest rate.
Empirical experiments however contradict these disappointing results for several strategies in the growth-optimal framework, as is explained in section 5.2. Furthermore, [Thorp and Beach, 2006] state that Thorp's investment firm invested over twenty years according to the principles of growth-optimality, thereby outperforming most benchmarks by far. The analysis given so far in the literature on this topic therefore seems to be insufficient to capture the true time-related behaviour of growth-optimal strategies in investing and could need further investigations.

3.7.4 Practical Calculation

Another point why Markowitz' model found more acceptance among practitioners lies in the fact, that the mean-variance portfolio problem is easy to solve under the assumption of normally distributed returns and its calculation is easy (see [Markowitz, 1970]). Although the assumption of normally distributed returns seems to be invalid in reality, it is considered good enough for practical risk management purposes.
Calculating a growth-optimal portfolio on the other hand that delivers reasonable results in reality is a much more complex task, thus it seemed not useful in practice for a long time (see for example [Cover, 1991] and [Cover, 2002]). Together with the computational complexity of possible algorithms as presented in section 4.6, these restrictions could be among the reasons why the growth-optimal portfolio theory found less application in practice than the mean-variance framework.

Chapter 4

Portfolio Algorithms

In the last chapter, the concept of growth-optimality was investigated as a reasonable framework for sequential investment from a theoretical point of view. As usual, these results cannot be applied immediately onto real world data. Algorithms are needed that can then be measured by results and definitions (universal consistency, deterministic and stochastic superiority) from the previous chapter. In this chapter several distribution-free algorithms from the literature are presented that either rely on the findings of the last chapter (approximating the expected logarithm of growth factors) or the bounds of prediction strategies by expert advice.

4.1 Best Constantly Rebalanced Portfolio

One of the basic results of the last chapter states that it is more useful to maximize the log of the portfolio growth factors than the growth factors themselves. This gives rise to the very simple concept of the best constantly rebalanced portfolio. For this, one determines the portfolio vector prediction at time n by

$$\hat{b}_n = \arg\max_{b \in \mathcal{B}} \prod_{i=1}^{n-1} \langle b, x_i \rangle \,.$$

This is equivalent to

$$\hat{b}_n = \arg\max_{b \in \mathcal{B}} \sum_{i=1}^{n-1} \ln \langle b, x_i \rangle \tag{4.1}$$

because the transformation with the logarithm does not change the maximum, as the log-function is monotonically increasing. One can easily see a very simple, deterministic property of this strategy, as shown in [Cover, 1991]:

Lemma 4.1 (Basic properties of the best constantly rebalanced portfolio)**.** *Denote by e_j the j-th basis vector, that is the vector that has 1 at the j-th place and 0 on all others. Define the empirical portfolio strategy $\hat{B} = \{\hat{b}_i\}_{i=1,2,\dots}$ by (4.1). Then $S_n(\hat{B})$ exceeds the arithmetic mean (for arbitrary weights $\alpha_j \geq 0, \sum_{j=1}^{d} \alpha_j = 1$)*

$$S_n(\hat{B}) \geq \sum_{j=1}^{d} \alpha_j S_n(e_j)$$

the geometric mean

$$S_n(\hat{B}) \geq \left(\prod_{j=1}^{d} S_n(e_j) \right)^{\frac{1}{d}}$$

and the maximum of the individual stocks

$$S_n(\hat{B}) \geq \max_j S_n(e_j).$$

Proof. These properties are simple consequences of the fact, that $S_n(\hat{B}) \geq S_n(e_j)$ for each $j = 1, \dots, d$. □

Thus, from a pure backward-looking perspective this strategy is probably the best we can come up with at first. But it is also an estimation of the growth-optimal strategy for iid returns, when considering the theoretical results from the last chapter. Remember that the best sequential investment strategy for iid returns was given by

$$\arg\max_{b \in \mathcal{B}} \mu(b) = \arg\max_{b \in \mathcal{B}} E(\ln(\langle b, x_i \rangle)).$$

The usual estimate $\hat{\mu}$ for the expected value is nothing else than the arithmetic mean of the given observations, which converges to the expected value as the number of observations tends to infinity by the law of large numbers. Therefore,

$$b^* \approx \arg\max_{b \in \mathcal{B}} \hat{\mu}(b) = \arg\max_{b \in \mathcal{B}} \frac{1}{n} \sum_{i=1}^{n} \ln(\langle b, x_i \rangle)$$

which is exactly the idea from the beginning of this section, because the additional fraction here is not relevant for the maximization.

Under the assumption of iid returns this portfolio vector would need to be calculated only once, as the growth-optimal portfolio in this case is constant over time. Still, it can be useful to recalculate this vector at each time step, as the estimate gets better by each added observation. The algorithm $BCR - PREDICT_i$ that follows is summarized in algorithm 1.

Algorithm 1: $BCR - PREDICT_i$

Input: Vectors of growth factors x_1^i.
Calculate

$$\hat{b}_{i+1} = \arg\max_{b \in \mathcal{B}} \sum_{j=1}^{i} \ln(\langle b, x_j \rangle).$$

Output: Portfolio vector $\hat{b}_{i+1}$.

Remark 4.1. *This is a very simple method and tends to be not very effective. The estimation of the expected value can be improved if further assumptions about the underlying distribution of the return vector and its log-sum are imposed. As this task requires a lot of assumptions and calculations, it can be assumed to be usually rather difficult. Furthermore, the results would tend to be very specific compared to the rather general results we derive in this thesis and are therefore not investigated any more.*

4.2 The EG Investment Strategy

Improving on the last algorithm and following [Cesa-Bianchi and Lugosi, 2006], we want to directly have a closer look at the ratio

$$\max_{x_1^n} \max_{b \in \mathcal{B}} \ln \frac{\prod_{i=1}^{n} \langle b, x_i \rangle}{\prod_{i=1}^{n} \left\langle \hat{b}_i, x_i \right\rangle}$$

where the numerator represents the best constantly rebalanced portfolio strategy. We call this ratio the worst-case logarithmic wealth ratio. If we find a strategy $\hat{B} = \{\hat{b}_i\}_{i=1,2,\ldots}$ for which this fraction converges towards 0, it would be deterministically superior to the set of constantly rebalanced portfolios and all processes. We know that

$$1 + u \leq \exp(u) \text{ for } u \geq 0$$

as $1+u$ represent the first two summands of the taylor series expansion of $\exp(u)$, followed by only positive summands when $u \geq 0$. We can conclude that

$$\ln(1 + u) \leq u.$$

This leads to the following inequality:

$$\begin{aligned}
\ln \frac{\prod_{i=1}^{n} \langle b, x_i \rangle}{\prod_{i=1}^{n} \left\langle \hat{b}_i, x_i \right\rangle} &= \sum_{i=1}^{n} \ln \left(1 + \frac{\left\langle b - \hat{b}_i, x_i \right\rangle}{\left\langle \hat{b}_i, x_i \right\rangle} \right) \\
&\leq \sum_{i=1}^{n} \sum_{j=1}^{d} \frac{(b^{(j)} - \hat{b}_i^{(j)}) x_i^{(j)}}{\left\langle \hat{b}_i, x_i \right\rangle} \\
&= \sum_{i=1}^{n} \left(\sum_{j=1}^{d} b^{(j)} \frac{x_i^{(j)}}{\left\langle \hat{b}_i, x_i \right\rangle} - \sum_{j=1}^{d} \hat{b}_i^{(j)} \frac{x_i^{(j)}}{\left\langle \hat{b}_i, x_i \right\rangle} \right). \qquad (4.2)
\end{aligned}$$

We want to maximize this equation with respect to b, which would be much easier if the right hand side of the inequality would be the negative value of the current. This can be done by bounding the growth factors between two constants,

$$0 < c \leq x_i^{(j)} \leq C$$

for each $i = 1, 2, \ldots$ and $j = 1, 2, \ldots, d$. An immediate consequence of this requirement on the growth factors are the bounds

$$c \leq \min_{j=1,\ldots,d} x_i^{(j)} = \sum_{k=1}^{d} b^{(k)} \min_{j=1,\ldots,d} x_i^{(j)} \leq \langle b, x_i \rangle$$

for each $i = 1, 2, \ldots$ and arbitrary $b \in \mathcal{B}$. With this we get

$$0 \leq \frac{x_i^{(j)}}{\left\langle \hat{b}_i, x_i \right\rangle} \leq \frac{C}{c}. \qquad (4.3)$$

By defining a function $l_i^{(j)}$ such that

$$l_i^{(j)} - \frac{C}{c} = -\frac{x_i^{(j)}}{\left\langle \hat{b}_i, x_i \right\rangle}$$

and replacing it in (4.2) we get the desired result (as $l_i^{(j)}$ is positive by (4.3)):

$$
\begin{aligned}
& \sum_{i=1}^{n} \left(\sum_{j=1}^{d} b^{(j)} \frac{x_i^{(j)}}{\left\langle \hat{b}_i, x_i \right\rangle} - \sum_{j=1}^{d} \hat{b}_i^{(j)} \frac{x_i^{(j)}}{\left\langle \hat{b}_i, x_i \right\rangle} \right) \\
= & \sum_{i=1}^{n} \left(\sum_{j=1}^{d} \hat{b}_i^{(j)} \left(l_i^{(j)} - \frac{C}{c} \right) - \sum_{j=1}^{d} b^{(j)} \left(l_i^{(j)} - \frac{C}{c} \right) \right) \\
= & \sum_{i=1}^{n} \left(\sum_{j=1}^{d} \hat{b}_i^{(j)} l_i^{(j)} - \sum_{j=1}^{d} b^{(j)} l_i^{(j)} \right) - \sum_{i=1}^{n} \left(\frac{C}{c} \underbrace{\sum_{j=1}^{d} \hat{b}_i^{(j)}}_{=1} - \frac{C}{c} \underbrace{\sum_{j=1}^{d} b^{(j)}}_{=1} \right) \\
= & \sum_{j=1}^{d} \sum_{i=1}^{n} \hat{b}_i^{(j)} l_i^{(j)} - \sum_{j=1}^{d} b^{(j)} \sum_{i=1}^{n} l_i^{(j)}.
\end{aligned}
$$

Maximizing with respect to b means minimizing the second sum of the right side of this equality. This is equivalent to choosing the vector b that has 1 on the place where $\sum_{i=1}^{n} l_i^{(j)}$ is minimal:

$$
\max_{b \in \mathcal{B}} \ln \frac{\prod_{i=1}^{n} \langle b, x_i \rangle}{\prod_{i=1}^{n} \left\langle \hat{b}_i, x_i \right\rangle} \leq \sum_{j=1}^{d} \sum_{i=1}^{n} \hat{b}_i^{(j)} l_i^{(j)} - \min_{j=1,\ldots,d} \sum_{i=1}^{n} l_i^{(j)}. \tag{4.4}
$$

The right hand side now has exactly the form of a regret function in the context of prediction by expert advice. Interpreting $l_i^{(j)} = \frac{C}{c} - \frac{x_i^{(j)}}{\left\langle \hat{b}_i, x_i \right\rangle}$ as the loss of an exponential forecaster, we immediately get a portfolio algorithm (summarized in algorithm 2) by using theorem 2.4:

$$
\hat{b}_1 = \left(\frac{1}{d}, \ldots, \frac{1}{d} \right),
$$

$$
\hat{b}_n^{(j)} = \frac{\hat{b}_{n-1}^{(j)} \exp\left(\eta \frac{x_{n-1}^{(j)}}{\left\langle \hat{b}_{n-1}, x_{n-1} \right\rangle} \right)}{\sum_{k=1}^{d} \hat{b}_{n-1}^{(k)} \exp\left(\eta \frac{x_{n-1}^{(k)}}{\left\langle \hat{b}_{n-1}, x_{n-1} \right\rangle} \right)}.
$$

As the formula in the exp-function has the form of a gradient, this method is called exponential gradient (EG) method (for an interpretation along these lines, see for example [Helmbold et al., 1998]). For this strategy, we can use theorem 2.3 to establish an upper bound for the worst-case logarithmic wealth ratio as it is bounded by the regret in equation (4.4). By directly bounding the regret of this special forecaster, this result can even be improved:

Algorithm 2: $EG - PREDICT_i$

Input: Vectors of growth factors x_1^i, last portfolio vector $\hat{b}_i$, bounds c and C.
Calculate

$$\eta = \frac{c}{C}\sqrt{\frac{8\ln d}{i}}.$$

for $j \in \{0, \ldots, d\}$ **do**
Calculate

$$\hat{b}_{i+1}^{(j)} = \frac{\hat{b}_i^{(j)} \exp\left(\eta \frac{x_i^{(j)}}{\langle \hat{b}_i, x_i \rangle}\right)}{\sum_{k=1}^{d} \hat{b}_i^{(k)} \exp\left(\eta \frac{x_i^{(k)}}{\langle \hat{b}_i, x_i \rangle}\right)}.$$

Output: Portfolio vector $\hat{b}_{i+1}$.

Theorem 4.1 (Deterministic superiority of the EG-Investment algorithm)**.** *Assume the growth factors are bounded by constants*

$$0 < c \leq x_i^{(j)} \leq C.$$

For the EG investment algorithm $\hat{B} = \{\hat{b}_i\}_{i=1,2,\ldots}$ *with*

$$\eta = \frac{c}{C}\sqrt{\frac{8\ln d}{i}}$$

the worst case logarithmic wealth ratio fulfills

$$\max_{x_1^n} \max_{b \in \mathcal{B}} \ln\left(\frac{\prod_{i=1}^{n} \langle b, x_i \rangle}{\prod_{i=1}^{n} \left\langle \hat{b}_i, x_i \right\rangle}\right) \leq \frac{C}{c}\sqrt{\frac{n}{2}\ln d}.$$

This is equivalent to

$$\lim_{n\to\infty} \max_{x_1^n} \max_{b \in \mathcal{B}} \frac{1}{n} \ln\left(\frac{S_n(b)}{S_n(\hat{B})}\right) \leq \lim_{n\to\infty} \frac{C}{c}\sqrt{\frac{1}{2n}\ln d} = 0.$$

This shows that the EG investment algorithm is deterministically superior with respect to the set of constantly rebalanced portfolios and all processes.

Proof. See [Cesa-Bianchi and Lugosi, 2006]. □

The previous result is rather strong as it shows that this portfolio should perform asymptotically always as good as the best constantly rebalanced portfolio from the last section

without adding much computational complexity. Still, we have no guarantee that it performs well, as the constantly rebalanced portfolio itself does not need to perform well. The next chapter will establish a first algorithm that is universally consistent.

Remark 4.2. *As the EG-strategy is designed to compete with the best constant portfolio, it is not surprising that it converges towards a constant portfolio itself as n gets large (Note that η depends on n and converges to 0 as $n \to \infty$).*

4.3 Universal Portfolios

In this section we investigate an algorithm that is inspired by the Laplace mixture forecaster from example 2.1. For this, consider the portfolio prediction

$$\hat{b}_1 = \left(\frac{1}{d}, \dots, \frac{1}{d}\right),$$

$$\hat{b}_n = \frac{\int_{\mathcal{B}} \cdots \int bS_{n-1}(b)db}{\int_{\mathcal{B}} \cdots \int S_{n-1}(b)db}.$$

This is more or less weighting each portfolio vector with its historic performance - the better a constant portfolio did in the past, the more the strategy tends to use this vector. One can interpret this also as estimating an empirical distribution over the possible outcomes of all portfolio strategies. That is, this strategy "learns" in the long run the real distribution of all possible constantly rebalanced portfolios and acts accordingly. The following result is an immediate consequence of this observation:

Theorem 4.2 (Universal consistency of the universal portfolio with respect to iid processes)**.** *Let the vectors of growth factors $X_i \overset{\text{iid}}{\sim} F(x)$. Then the average growth rate induced by the previous strategy $\hat{B} = \{\hat{b}_i\}_{i=1,2,\dots}$ achieves asymptotically that of the best constantly rebalanced portfolio a.s.,*

$$\frac{1}{n} \ln(S_n(\hat{B})) \to \max_{b \in \mathcal{B}} E(\langle b, X_i \rangle) \quad a.s.$$

It is therefore universally consistent with respect to the class of all iid processes.

Proof. See [Cover, 1991]. □

With this result, the proposed algorithm seems to be a powerful tool if the underlying returns are assumed to be iid.
To understand the algorithm even better, look at the resulting wealth after n investment periods (as did [Cesa-Bianchi and Lugosi, 2006]):

$$\begin{aligned} S_n(\hat{B}) &= \prod_{i=1}^{n} \left\langle \hat{b}_i, x_i \right\rangle = \prod_{i=1}^{n} \left\langle \frac{\int_{\mathcal{B}} bS_{i-1}(b)db}{\int_{\mathcal{B}} S_{i-1}(b)db}, x_i \right\rangle \\ &= \prod_{i=1}^{n} \frac{\left\langle \int_{\mathcal{B}} bS_{i-1}(b)db, x_i \right\rangle}{\int_{\mathcal{B}} S_{i-1}(b)db} = \prod_{i=1}^{n} \frac{\int_{\mathcal{B}} \langle bS_{i-1}(b), x_i \rangle \, db}{\int_{\mathcal{B}} S_{i-1}(b)db} \\ &= \prod_{i=1}^{n} \frac{\int_{\mathcal{B}} \langle b, x_i \rangle \, S_{i-1}(b)db}{\int_{\mathcal{B}} S_{i-1}(b)db} = \prod_{i=1}^{n} \frac{\int_{\mathcal{B}} S_i(b)db}{\int_{\mathcal{B}} S_{i-1}(b)db} \\ &= \int_{\mathcal{B}} S_n(b)db \end{aligned}$$

as the telescope products cancel out. This simple derivation shows that the universal portfolio strategy delivers the average of all constantly rebalanced portfolios - a reasonable choice if we want to beat the constantly rebalanced portfolios.
Following this line of thought, we can look at how well this portfolio performs in comparison to the class of all constantly rebalanced portfolios. We will see that this can be done by defining the forecaster $\hat{p}_n$ induced by a fixed constantly rebalanced portfolio $\bar{b}$,

$$\hat{p}_n(y_1^n, \bar{b}) = \bar{b}^{(1)^{n_1}} \cdots \bar{b}^{(d)^{n_d}}$$

where $y_1^n \in \mathcal{Y}^n = \{1, \ldots, d\}^n$ and n_i is the number of occurrences of i in y_1^n.
The wealth resulting from such a constantly rebalanced portfolio can then be rewritten as

$$\begin{aligned} S_n(b) &= \prod_{i=1}^{n} \langle \bar{b}, x_i \rangle \\ &= \sum_{y_1^n \in \mathcal{Y}^n} \prod_{i=1}^{n} x_i^{y_i} \hat{p}_n(y_1^n, \bar{b}) \\ &= \sum_{y_1^n \in \mathcal{Y}^n} \hat{p}_n(y_1^n, \bar{b}) \prod_{i=1}^{n} x_{y_i}^i. \end{aligned}$$

Substituting this part in the wealth of the universal portfolio (which is consisting of constantly rebalanced portfolios) results in

$$S_n(\hat{B}) = \sum_{y_1^n \in \mathcal{Y}^n} \int_{\mathcal{B}} \hat{p}_n(y_1^n, b)db \prod_{i=1}^{n} x_{y_i}^i.$$

This shows that the universal portfolio strategy essentially is a mixture of Laplace mixture

strategies as in example 2.1. It can therefore be treated along those lines, leading to the following result:

Theorem 4.3 (Deterministic superiority of the universal portfolio algorithm)**.** *The ratio of wealth of the universal portfolio strategy and the best constantly rebalanced strategy is bounded by*

$$\sup_{x^n} \sup_{b \in \mathcal{B}} \ln \frac{S_n(b)}{S_n(\hat{B})} \leq (d-1) \ln(n+1).$$

Considering the asymptotic average growth rate, this gives

$$\sup_{x^n} \sup_{b \in \mathcal{B}} \frac{1}{n} \ln \frac{S_n(b)}{S_n(\hat{B})} \leq (d-1) \frac{\ln(n+1)}{n}.$$

Therefore,

$$\lim_{n \to \infty} \sup_{x^n} \sup_{b \in \mathcal{B}} \frac{1}{n} \ln \frac{S_n(b)}{S_n(\hat{B})} \leq 0.$$

This means that the universal portfolio strategy is deterministically superior with respect to the set of all constantly rebalanced strategies and all processes.

Proof. See [Cesa-Bianchi and Lugosi, 2006]. □

4.3.1 Approximation by the Trapezoidal Rule

For practical purposes, the integrals in the universal portfolio strategy need to be solved numerically. [Cover, 1991] does this by using the trapezoidal rule for the case of two assets. It is rather obvious that this leads to

$$\hat{b}_n = \frac{\sum_{i=0}^{T} (\frac{i}{T}, 1 - \frac{i}{T}) S_{n-1}(\frac{i}{T}, 1 - \frac{i}{T})}{\sum_{i=0}^{T} S_{n-1}(\frac{i}{T}, 1 - \frac{i}{T})}$$

for $d = 2$ and $T \in \mathbb{N}$ nodes for the interpolation. Note that the equidistant weights induced by using the trapezoidal rule cancel out.
Adding a third asset, it is important not to violate the property that the elements of b must add up to 1. Therefore, we limit the second sum and get

$$\hat{b}_n = \frac{\sum_{i_1=0}^{T} \sum_{i_2=0}^{T-i_1} (\frac{i_1}{T}, \frac{i_2}{T}, 1 - \frac{i_1+i_2}{T}) S_{n-1}(\frac{i_1}{T}, \frac{i_2}{T}, 1 - \frac{i_1+i_2}{T})}{\sum_{i_1=0}^{T} \sum_{i_2=0}^{T-i_1} S_{n-1}(\frac{i_1}{T}, \frac{i_2}{T}, 1 - \frac{i_1+i_2}{T})}.$$

We generalize this to the d-asset case. For this, define the running index for sum k by i_k and the variable $I_k = \sum_{j=1}^{k-1} i_j$. With this, an approximation for the integral with equidistant nodes in the trapezoidal rule is given by

$$\hat{b}_n = \frac{\sum_{i_1=0}^{T}\sum_{i_2=0}^{T-I_2}\sum_{i_3=0}^{T-I_3}\cdots\sum_{i_{d-1}=0}^{T-I_{d-1}}(\frac{i_1}{T},\ldots,\frac{i_{d-1}}{T},1-\frac{I_{d-1}}{T})S_{n-1}((\frac{i_1}{T},\ldots,\frac{i_{d-1}}{T},1-\frac{I_{d-1}}{T}))}{\sum_{i_1=0}^{T}\sum_{i_2=0}^{T-I_2}\sum_{i_3=0}^{T-I_3}\cdots\sum_{i_{d-1}=0}^{T-I_{d-1}}S_{n-1}((\frac{i_1}{T},\ldots,\frac{i_{d-1}}{T},1-\frac{I_{d-1}}{T}))}.$$

The resulting prediction algorithm $UP-PREDICT_i$ is given in algorithm 4. Note that we need to use a matrix B of all portfolio vectors that occur in the application of the trapezoidal rule. This matrix results from calling the function $TRAP-MATRIX$ with $v=1$, $I_k=0$, $k=1$, the number of nodes T, $b=(0,\ldots,0)$ and the number of assets d as provided in algorithm 3.

Algorithm 3: $TRAP-MATRIX$

Input: I_k, k, T, a vector b and an integer d that stops the recursion as well as access to the global variable v and global matrix B.

if $k<d$ **then**
- $s=0$
- **for** $j\in\{0,\ldots,T-I_k\}$ **do**
  - $b_k=\frac{j}{T}$
  - $TRAP-MATRIX(I_k+j,k+1,T,b,d)$

else
- **for** $j\in\{0,\ldots,d-1\}$ **do**
  - Set $B_{v,j}=b_j$.
- $B_{v,d}=1-\frac{I_k}{T}$
- $v=v+1$

Output: No output.

Algorithm 4: $UP-PREDICT_i$

Input: Vectors of growth factors x_1^i, integer T, matrix B and $S_{i-1}(b)$ for each row b in B.

Let v be the number of rows in B and B_j the j-th row of B.

for $j\in\{1,\ldots,v\}$ **do**
- Calculate
$$S_i(B_j)=S_{i-1}(B_j)\left\langle B_j,x_i\right\rangle.$$

Calculate
$$\hat{b}_{i+1}=\frac{\sum_{j=1}^{v}B_jS_i(B_j)}{\sum_{j=1}^{v}S_i(B_j)}.$$

Output: Portfolio vector $\hat{b}_{i+1}$ and $S_i(b)$ for each row b in B.

4.3.2 Approximation by a Monte-Carlo Method

Even though rather simple to implement, the trapezoidal approximation of an integral is usually not a reasonable choice, especially considering the computational costs in higher dimensions (compare section 4.6). For high dimensional numerical integration, Monte-Carlo methods are usually more appropriate. Recapitulating the basics (as in [Hammersley and Handscomb, 1964], for example), in a Monte-Carlo setting one interprets a standard multidimensional integral over a d-dimensional region $\mathcal{R}$

$$I_{\mathcal{R}} = \int_{\mathcal{R}} \cdots \int f(x)dx$$

as the expected value of $f(X)$ of the uniformly on $\mathcal{R}$ distributed d-dimensional random vector X. By the law of large numbers, the expected value can always be approximated by taking observations of X, averaging them and multiplying with the volume of $\mathcal{R}$. Therefore, approximating $I_{\mathcal{R}}$ can be done by simulating uniformly distributed random vectors $X_1, \ldots, X_n$ on $\mathcal{R}$ and calculating

$$I_{\mathcal{R}} \approx V(\mathcal{R})\frac{1}{n}\sum_{i=1}^{n} f(X_i)$$

where $V(.)$ is the volume function. This method delivers better approximation error bounds than the trapezoidal rule. In fact, this bound decreases by the factor $\frac{1}{\sqrt{n}}$ for the simple Monte-Carlo method as presented here.

For approximating the universal portfolio, we now need to simulate uniformly distributed vectors in the unity simplex $\mathcal{B}$. Fortunately, there is a well-known method to transform uniformly distributed random vectors from the unit cube $[0,1]^d$ into the unit simplex (see for example [Devroye, 1986]). To do so, simulate d uniformly distributed variables $U_1, \ldots, U_d$ by using a Halton or Sobol sequence or similar methods. Order them increasingly into $U_{(1)}, \ldots, U_{(d)}$ and add $U_0 = U_{(0)} = 0$ and $U_{d+1} = U_{(d+1)} = 1$. The vector

$$\tilde{b} = (U_{(1)} - U_{(0)}, U_{(2)} - U_{(1)}, \ldots, U_{(d+1)} - U_{(d)})$$

is then uniformly distributed on the unit simplex. With this knowledge, the universal portfolio prediction can be approximated by

$$\hat{b}_{n+1} = \frac{\sum_{i=1}^{T} \tilde{b}_i S_n(\tilde{b}_i)}{\sum_{i=1}^{T} S_n(\tilde{b}_i)}$$

with T d-dimensional vectors $\tilde{b}_i$. Note that the weights again cancel out as in the trape-

zoidal rule. This concept leads to the function $MC - PREDICT_i$, summarized in algorithm 6 after getting T simulated portfolio vectors by using algorithm 5.

Algorithm 5: SIM

Input: Integer T and dimension d.
Produce Td uniformly distributed random variables $U_1, \ldots, U_{Td}$.
Set $U_{(0)} = 0$ and $U_{(d+1)} = 1$.
for $i \in \{1, \ldots, T\}$ **do**
  Order $U_{(i-1)d+1}, \ldots, U_{id}$ increasingly into $U_{(1)}, \ldots, U_{(d)}$.
  Calculate $\tilde{b}_i = (U_{(1)} - U_{(0)}, U_{(2)} - U_{(1)}, \ldots, U_{(d+1)} - U_{(d)})$.
Output: T portfolio vectors $\tilde{b}_1, \ldots, \tilde{b}_T$.

Algorithm 6: $MC - PREDICT_i$

Input: Vectors of growth factors x_1^i and $\tilde{b}_j$, $S_{i-1}(\tilde{b}_j)$ for $j = 1, \ldots, T$.
for $j \in \{1, \ldots, T\}$ **do**
  Calculate

$$S_i(\tilde{b}_j) = S_{i-1}(\tilde{b}_j) \left\langle \tilde{b}_j, x_i \right\rangle .$$

Calculate

$$\hat{b}_{i+1} = \frac{\sum_{j=1}^{T} \tilde{b}_j S_i(\tilde{b}_j)}{\sum_{j=1}^{T} S_i(\tilde{b}_j)}.$$

Output: Portfolio vector $\hat{b}_{i+1}$ and $S_i(\tilde{b}_j)$ for $j = 1, \ldots, T$.

4.4 A Kernel Based Algorithm with Expert Advice

Thinking back to the result in theorem 3.2, we saw that solving the problem of finding a sequential growth-optimal portfolio selection procedure reduces to solving

$$\max_{b(.) \in \mathcal{B}} E(\ln \left\langle b(x_1^{n-1}), x_n \right\rangle | x_1^{n-1})$$

for each investment period. With the knowledge of the kernel regression estimate for s-a-e processes, we have an idea how to estimate $E(x_n | x_1^{n-1})$. By letting the radius r depend on a "window"-length k and a "precision"-parameter l as in section 2.3.2 and using the Frobenius norm, we define

$$\mathcal{J}_{k,l,n} = \{k < i < n : \|x_{i-k+1}^{i} - x_{n-k+1}^{n}\| \leq r_{k,l}\}$$

and accordingly the experts

$$\hat{b}_n^{(k,l)}(x_1^{n-1}) = \arg\max_{b\in\mathcal{B}} \sum_{j\in\mathcal{J}_{k,l,n-1}} \ln \langle b, x_{j+1}\rangle \,. \tag{4.5}$$

Let

$$\hat{B}_n^{(k,l)} = (\hat{b}_1^{(k,l)}, \ldots, \hat{b}_n^{(k,l)})$$

and

$$\hat{B}^{(k,l)} = \{\hat{B}_i^{(k,l)}\}_{i=1,2,\ldots}.$$

Then, the resulting portfolio vector prediction $\hat{b}_n(x_1^{n-1})$ is the simple weighing of the individual experts with their past performance and an arbitrary probability distribution $\{q_{k,l}\}$:

$$\hat{b}_n(x_1^{n-1}) = \frac{\sum_{k,l} q_{k,l} S_{n-1}(\hat{B}^{(k,l)}) \hat{b}_n^{(k,l)}(x_1^{n-1})}{\sum_{k,l} q_{k,l} S_{n-1}(\hat{B}^{(k,l)})}. \tag{4.6}$$

For this method, the following result can be shown:

Theorem 4.4 (Universal consistency of kernel algorithm with respect to s-a-e processes)**.** *Let k, l run over all positive integers and let $r_{k,l}$ be strictly decreasing in l for fixed k, that is*

$$\lim_{l\to\infty} r_{k,l} = 0.$$

Then the portfolio strategy (4.9) is universally consistent with respect to the class of all s-a-e processes with $E(|\ln X_j|) < \infty$ for $j = 1, 2, \ldots, d$.

Proof. See [Györfi et al., 2006]. □

With this theorem, we have found a portfolio selection algorithm for the growth-optimal portfolio, that works for the general class of s-a-e processes without further adaptation. The algorithm $KERNEL - PREDICT_i$ for a one-step prediction at time i is given in pseudo code in algorithm 7, assuming finite k and l, as well as "sensibly" chosen $r_{k,l}$. Computational issues arising in this context will be adressed in section 4.6.

Algorithm 7: $KERNEL - PREDICT_i$

Input: Vectors of growth factors x_1^i as well as $r_{k,l}$, $q_{k,l}$, $\hat{b}_i^{(k,l)}(x_1^{i-1})$ and $S_{i-1}(\hat{B}^{(k,l)})$ for each (k,l), $k = 1, \ldots, K$, $l = 1, \ldots, L$.

for $l \in \{1, \ldots, L\}$ **do**

  for $k \in \{1, \ldots, K\}$ **do**

    Update and store

$$S_i(\hat{B}^{(k,l)}) = S_{i-1}(\hat{B}^{(k,l)}) \left\langle \hat{b}_i^{(k,l)}(x_1^{i-1}), x_i \right\rangle .$$

    Collect all data points $k < j < i$ in the set $\mathcal{J}_{k,l,i}$, where

$$\|x_{j-k+1}^j - x_{i-k+1}^i\| \leq r_{k,l}.$$

    if $\mathcal{J}_{k,l,i} = \varnothing$ **then**

$$b_{i+1}^{(k,l)}(x_1^i) = \left(\frac{1}{d}, \ldots, \frac{1}{d}\right)$$

    else

      Calculate and store

$$\hat{b}_{i+1}^{(k,l)}(x_1^i) = \arg\max_{b \in \mathcal{B}} \sum_{j \in \mathcal{J}_{k,l,i}} \langle b, x_{j+1} \rangle .$$

Calculate

$$\hat{b}_{i+1}(x_1^i) = \frac{\sum_{k,l} q_{k,l} S_i(\hat{B}^{(k,l)}) \hat{b}_{i+1}^{(k,l)}(x_1^i)}{\sum_{k,l} q_{k,l} S_i(\hat{B}^{(k,l)})}.$$

Output: Portfolio vector $\hat{b}_{i+1}$.

4.5 A Nearest Neighbour Based Algorithm with Expert Advice

Similarly to the way we came up with $KERNEL - PREDICT_i$, it is rather natural to apply the nearest neighbour idea from nonparametric regression to portfolio selection. Therefore, we again define experts for each k and l. Choose for each l a $p_l \in (0,1)$ such that

$$\lim_{l\to\infty} p_l = 0. \tag{4.7}$$

Define further

$$\bar{l} = \lfloor p_l n \rfloor$$

and the set of $\bar{l}$ nearest neighbours

$$\mathcal{J}_{k,l,n} = \{k < i < n : \|x_{i-k+1}^i - x_{n-k+1}^n\| \text{ is among the } \bar{l} \text{ smallest values}\}$$

where $\|.\|$ is again the well known Frobenius norm. With this new set we can define the experts

$$\hat{b}_n^{(k,l)}(x_1^{n-1}) = \arg\max_{b\in\mathcal{B}} \prod_{j\in\mathcal{J}_{k,l,n-1}} \langle b, x_{j+1}\rangle . \tag{4.8}$$

Once more, let

$$\hat{B}_n^{(k,l)} = (\hat{b}_1^{(k,l)}, \ldots, \hat{b}_n^{(k,l)})$$

and

$$\hat{B}^{(k,l)} = \{\hat{B}_i^{(k,l)}\}_{i=1,2,\ldots}.$$

Combine the experts using an arbitrary probability distribution $\{q_{k,l}\}$ by

$$\hat{b}_n(x_1^{n-1}) = \frac{\sum_{k,l} q_{k,l} S_{n-1}(\hat{B}^{(k,l)}) \hat{b}_n^{(k,l)}(x_1^{n-1})}{\sum_{k,l} q_{k,l} S_{n-1}(\hat{B}^{(k,l)})}. \tag{4.9}$$

Then, we can state the following:

Theorem 4.5 (Universal consistency of nearest neighbour portfolio algorithm with respect to s-a-e processes)**.** *Assume (4.7) and assume that ties occur with probability 0. Then the nearest neighbour portfolio scheme as defined in (4.9) asymptotically achieves*

the highest possible expected growth rate almost surely with respect to the class of s-a-e processes such that $E(|\ln(X_i^{(j)})|) < \infty$ *for all assets* $j = 1, \ldots, d$.

Proof. See [Györfi et al., 2008b]. □

The resulting algorithm $NN - PREDICT_i$ for a one-step prediction at time i is given in algorithm 8, assuming finite k and l and "sensibly" chosen p_l.

Algorithm 8: $NN - PREDICT_i$

Input: Vectors of growth factors x_1^i as well as $p_l \in (0,1)$ for each l, $q_{k,l}$, $\hat{b}_i^{(k,l)}(x_1^{i-1})$ and $S_{i-1}(\hat{B}^{(k,l)})$ for each (k,l), $k = 1, \ldots, K$, $l = 1, \ldots, L$.

for $l \in \{1, \ldots, L\}$ **do**

  for $k \in \{1, \ldots, K\}$ **do**

    Update and store

$$S_i(\hat{B}^{(k,l)}) = S_{i-1}(\hat{B}^{(k,l)}) \left\langle \hat{b}_i^{(k,l)}(x_1^{i-1}), x_i \right\rangle .$$

    for $j \in \{1, \ldots, i\}$ **do**

      Calculate the Frobenius Norm $F_j = \|x_{j-k+1}^j - x_{i-k+1}^i\|$.

    Order the F_j increasingly from $F_{(1)}$ being the smallest to $F_{(i)}$ being the biggest value. Calculate $\bar{l} = \lfloor p_l i \rfloor$. Collect the set $\mathcal{J}_{k,l,i} = (F_{(1)}, \ldots, F_{(\bar{l})})$.

    if $\mathcal{J}_{k,l,i} = \varnothing$ **then**

$$\hat{b}_{i+1}^{(k,l)}(x_1^i) = \left(\frac{1}{d}, \ldots, \frac{1}{d}\right)$$

    else

      Calculate and store

$$\hat{b}_{i+1}^{(k,l)}(x_1^i) = \arg\max_{b \in \mathcal{B}} \prod_{j \in \mathcal{J}_{k,l,i}} \langle b, x_{j+1} \rangle .$$

Calculate

$$\hat{b}_{i+1}(x_1^i) = \frac{\sum_{k,l} q_{k,l} S_i(\hat{B}^{(k,l)}) \hat{b}_{i+1}^{(k,l)}(x_1^i)}{\sum_{k,l} q_{k,l} S_i(\hat{B}^{(k,l)})}.$$

Output: Portfolio vector $\hat{b}_{i+1}$.

4.6 Computational Complexity

As indicated in the critical discussion of growth-optimal models in section 3.7.4, calculating a growth-optimal portfolio is not an easy task. This was further confirmed when

looking at the final algorithms as defined above. While the constantly rebalanced portfolios are rather easy to obtain (the necessary optimizations can be done numerically, for example by using Spelucci's donlp2[1]), the more advanced methods, namely the kernel and nearest neighbour based algorithms, are much more complex. As discussed in [Györfi et al., 2008a] and [Györfi et al., 2007], an efficient implementation is essential for the applicability of an empirical portfolio selection method. We will examine each algorithm individually:

- **Algorithm 1:** The best constantly rebalanced portfolio is the easiest algorithm in this thesis and straightforward. At each prediction, one optimization is needed. Depending on the efficiency of the optimization algorithm, this should be done fast.

- **Algorithm 2:** Even though theoretically better than the best rebalanced portfolio, the EG-Investment Strategy is probably even faster than algorithm 1, as no optimization is needed at all. It is linear in its arguments for each prediction and therefore extremely fast.

- **Algorithm 4 (using algorithm 3):** Again, we do not need to optimize for the universal portfolio. But the computational difficulty here lies in the numerical evaluation of an integral. Using the trapezoidal rule is not very efficient (especially as it requires the use of a recursive function), but if the number of nodes is small, the computational costs are bearable. Unfortunately the number of calculations - which depends clearly on the number of summands in the approximation - explodes fast as shown in figure 4.1 (even though it seems to stay sub-exponential). This method is therefore not feasible for a bigger number of assets. This stays true even if the wealth associated with each possible vector is stored at each time step, which saves valuable time that would be needed for the recalculation of wealth.

- **Algorithm 6 (using algorithm 5):** With a Monte-Carlo approximation, the integrals can be solved in however little time one wants. Obviously, the accuracy of the approximation strongly depends on T - but so does the computational cost. Valuable time can be saved again by simulating the portfolio vectors only once and updating the associated wealth at each time step instead of recalculating it. A big advantage here is that the time needed for calculation can be estimated easily as it depends linearly on T, that is for double the amount of portfolio vectors one needs twice the time to evaluate the integral.

[1] See http://www.mathematik.tu-darmstadt.de/fbereiche/numerik/staff/spellucci/DONLP2/.

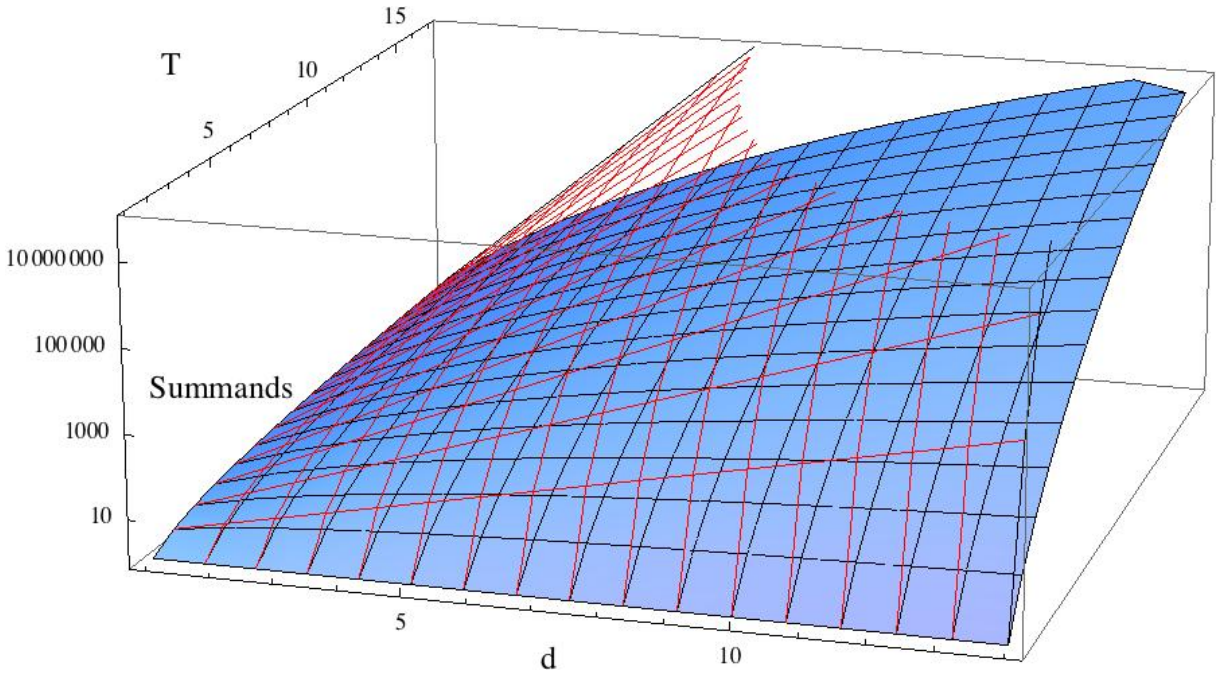


Figure 4.1: Number of summands for the trapezoidal rule over the unit simplex. Red mesh represents number of summands for the trapezoidal rule over the unit cube.

- **Algorithm 7:** There are two crucial factors that determine the complexity of this algorithm: First, the search for similarities in the local averaging is of quadratic order as the comparison of n points with each other needs $\binom{n}{2} = \frac{n(n-1)}{2}$ calculations, which is a well known result from the handshake lemma. Secondly, as the number of similarities grows the numerical optimization gets more complex too, which is the other time consuming step. This complexity depends strongly on the choice of the radius function: The smaller the radius, the fewer similarities, the faster the optimization. Unfortunately, there is no way to solve these two points efficiently - one has always a tradeoff between runtime and precision. But one can at least reduce the number of these calculations to a minimum by storing the historic growth of the individual experts, thus calculating the history always only once. This analysis indicates that radius function as well as K and L need to be chosen carefully. Ideally, one tries several configurations before deciding which one to use.

- **Algorithm 8:** The nearest neighbour based algorithm poses the same challenges as the kernel-based method. Additionally, one needs to sort the historic differences of the norms at each step. This at least can be implemented sufficiently by using an efficient sorting algorithm (of which many are available) and is thus a minor problem.

Besides considering these points before choosing an algorithm, it is also important to choose a low-level programming language for the whole implementation, especially for the complex algorithms 4, 7 and 8. The importance of this point is underlined by the following: An implementation of the kernel based algorithm in the statistics software R with an external optimization package needed over three weeks to run a backtest on the commodities data for the strategy S2 that will be presented in chapter 5, while an implementation in $C++$ needed around 10 hours only, i.e. a reduction of the runtime by a factor of 50.

Chapter 5

Empirical Results

5.1 Evaluation of an Investment Strategy

In the last chapter, several algorithms were provided which can be applied in reality to create an investment strategy. For practitioners, the practicability and effectiveness of these algorithms on real-world data is important. Therefore, it is common to test such strategies with a backtest on historic data sets. Given a time series of n real-world data points and an a-priori algorithm $PREDICT_i$ that predicts at time i the portfolio vector for $i+1$, this so called method $BACKTEST$ is summarized in Algorithm 9.

Algorithm 9: $BACKTEST$

Input: Vectors of growth factors x_1^n.
for $i \in \{1, \ldots, n\}$ **do**
  $b_{i+1} = PREDICT_i$.
Output: Portfolio vectors $b_2, \ldots, b_{n+1}$.

After running the backtest, the portfolio returns resulting from the algorithm can easily be calculated by setting

$$r_i = \langle b_i, x_i \rangle \quad \forall i \in \{2 \ldots, n\}.$$

These returns can then be analysed. It should of course be mentioned that the results of such a backtest are highly hypothetical, as the following assumptions are made:

- **There are no transaction costs.** In reality, rebalancing a portfolio can be very costly.
- **Arbitrary numbers of assets can be bought for the price of their last quote on each trading day** (the price given by the historic time series). In reality, one

can only buy for a price close to the observed one. Furthermore, buying a lot of stocks at once changes prices considerably, which again means that stocks cannot be bought or sold at the price of the last quote. Huge transactions can even lead to charges for manipulating the stock market.

- **Assets are arbitrarily divisible.** This means that one can invest every desired fraction of wealth into any asset, however small the fraction is. Although this is nearly true with currencies, it can normally not be done with stocks. These have to be bought in full pieces, meaning that one can only invest a multiple of the given stock prices. If wealth is large, e.g. ten thousand times the price of the most expensive stock, the assumption becomes more or less valid again.

Despite those unrealistic assumptions, a backtest is a good indicator of the effectiveness of a strategy (especially in liquid markets like the Dow Jones), even though one cannot expect to achieve the same gains in a real-world environment (or needs to adjust the model to take into account the restrictions mentioned above).
There are a lot of performance measures to evaluate a portfolio strategy. We will concentrate on the following ones as provided in the R package *PerformanceAnalytics* ([Carl and Peterson, 2010]):

- **Daily standard deviation (volatility):** This is a standard measurement of risk given by the usual formula for the standard deviation

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(r_i - \bar{r})^2}$$

where r_i, $i = 1, \ldots, n$, are the daily returns and $\bar{r}$ is their arithmetic average. One would like to have as little variation as possible in the strategy, as this means a rather constant return rate with little insecurity.

- **Daily Sharpe ratio:** While it is desirable to have little variation in the returns, this is most likely not the only objective. Usually one would allow more variance in return for higher returns. Therefore, it is useful (and common) to compare investment strategies by the Sharpe ratio, that is the ratio of average return and standard deviation

$$R_{\text{Sharpe}} = \frac{\bar{r}}{s}. \tag{5.1}$$

A higher Sharpe ratio is naturally preferred, as this means high returns with low volatility.

- **Daily Sortino ratio:** To define risk by standard deviation means that we desire constant returns. But actually we do not care if there are high peaks in the positive returns - only negative returns are an issue. Therefore, these positive returns should probably be excluded when talking about risk, which gives rise to the definition of the downside deviation: Collect the $n_\delta \leq n$ returns that are smaller than 0 in Δ and define

$$\delta = \sqrt{\frac{1}{n}\sum_{r\in\Delta} r^2}$$

where $\bar{r}_\delta$ is the arithmetic average of all returns in Δ. The Sortino ratio now picks up the idea of the Sharpe ratio, only by using the downside deviation instead of the standard deviation:

$$R_{\text{Sortino}} = \frac{\bar{r}}{\delta}. \tag{5.2}$$

Again, a high Sortino ratio is preferred. Taking into account only negative returns for the downside deviation (the arithmetic mean is again taken over all returns), the Sortino ratio is more risk sensitive than the Sharpe ratio.

- **Maximum drawdown:** To describe the maximum drawdown, we first define the peak in the cumulated wealth of a strategy B before a time T as the maximum wealth up to T

$$P(T) = \max_{j=1,\ldots,T} S_j(B).$$

A drawdown at time T is the percentage loss incurred since the last peak before T,

$$D(T) = \max\left(0, \frac{P(T) - S_T(B)}{S_T(B)}\right).$$

The drawdown is zero if $S_T(B)$ is the peak. Obviously, drawdowns are undesirable but inevitable. Still, we would like them to be at least small. It is therefore useful to look at the largest (the maximum) drawdown in the backtest until time T:

$$MD(T) = \max_{t=1,\ldots,T} D(t).$$

A small maximum drawdown again indicates a rather constant growth from peak to peak.

- **Annualized return:** The annual geometric average return per year in the backtest is given as AR, which is nothing else than the yearly interest on the investment strategy. Making this even more precise: If the backtest produced cumulated wealth of $S_N(B)$ over N given years, then the annualized return is given by

$$AR = S_N(B)^{\frac{1}{N}} - 1.$$

Apart from those measures, we will also give the daily average growth rate W, the percentage of months with positive returns PR_{mon} and the percentage of years with positive returns PR_{ann}. Obviously, one desires high values for all those three. Additionally, we also provide information about the approximate runtime RT of the backtest on a Mac with a 2.26 GHz Intel Core 2 Duo processor, as the time needed to calculate the portfolio prediction could also be relevant when rebalancing (and therefore recalculation) is done often.
For visual interpretation, we provide figures on the behaviour of the portfolio vector over time, the time series of returns and cumulated growth (how would an original investment of 1 currency unit have turned out?) on log-scale as well as the time series of the drawdowns $D(t)$ for each strategy. In those last three figures, we will also draw a benchmark for the strategy by adding the performance of the equally weighted portfolio, which is the constantly rebalanced portfolio with $b = \left(\frac{1}{d}, \ldots, \frac{1}{d}\right)$.

5.2 Backtests in the Literature

Several of the algorithms presented above have already been applied onto real world data. The best constantly rebalanced portfolio over the whole period of historic returns (that means the best constantly rebalanced portfolio in hindsight) usually serves as a benchmark for the performance. Those backtests usually confine themselves to presenting the average growth rate of the method.
[Cover, 1991] tested the universal portfolio strategy on two stocks from the New York Stock Exchange which were picked for their volatility, yielding a wealth of almost 40 times the invested capital after 5651 trading days corresponding to an observed period of approximately 22 years. It still falls short from the best constantly rebalanced portfolio when calculated in hindsight, but performs much better than the individual stocks on their own.

[Helmbold et al., 1998] used the same data with the EG strategy, doing considerably better with the same stocks than the universal portfolio (for a constant $\eta = 0.05$ in the formula). In fact, they achieved 70 times the initial wealth with this strategy. For other combinations of stocks they even achieved a wealth up to 110.2 times S_0.
For testing the kernel and nearest neighbour based methods, [Györfi et al., 2008a] used data from the stock and currency markets. The results are surprisingly good: For 36 stocks from the New York Stock Exchange (from the same data set as used by [Cover, 1991] and [Helmbold et al., 1998]), investing 1 currency unit by the different empirical portfolio selection rules results in a wealth of 1.12×10^9 to 3.31×10^{11} currency units after 5651 trading days (approximately 22 years). This corresponds to an annualized return of 159% to 234%, while the best constantly rebalanced portfolio in hindsight delivered an annualized return of 29%. Similar results were obtained for 18 pairs of currencies. Investing one currency unit into different exchange rates resulted in an accumulated wealth of 22.22 to 393.00 currency units after 3429 trading days, or an annualized return of 26.9% to 58.3% percent. Even though this is much less than the results for the stocks, it has to be taken into account that while stocks are investment products that usually gain in prices over time, currency exchange rates are more or less swinging around a constant mean which do not generate wealth on their own. This is also reflected by the result of the constantly rebalanced portfolio, which only achieves an annualized return of 1.9% for the currency data. Restricting the possible stocks to only the two most volatile ones shows here as well to further increase profitability and also save computational costs.

5.3 A Test with a Selection of Commodities

Similar to the ones mentioned in the last section, we conduct a backtest on a selection of commodities under the assumptions of arbitrary divisibility and no transaction costs. Commodities are traded on a wide variety of markets and are usually difficult to use for investors as they need to be stored and transported. To invest in commodities, one usually uses financial derivatives like futures, which can be bought and sold without really possessing the products. Futures allow investors to buy products in the future for a price that is set today. They are especially convenient, as there are only small transaction costs on big markets. A future's price is closely linked to the price of its underlying asset, which should allow the same conclusions for futures like those that we draw from the backtest of the prices of the commodities (see for example [Hull, 2006]).

5.3.1 Description of Data

Asset#	Name	Quoted by	Unit
1	LME-Aluminium 99.7%	London Metal Exchange	US$/Metric Tonne
2	Coffee-Colombian (NY)	Wall Street Journal	US-Cents/Pound
3	LME-Copper, Grade A	London Metal Exchange	US$/Metric Tonne
4	Corn No.2 Yellow	US Department of Agriculture	US-Cents/Bushel
5	Cotton,1 1/16 Str Low	US Department of Agriculture	US-Cents/Pound
6	Crude Oil-Brent	ICIS Pricing	US$/Barrel
7	Gold Bullion LBM	London Bullion Market	US$/Troy Ounce
8	Natural Gas-Henry Hub	Dow Jones Energy Service	US$/Million British Thermal Units
9	LME-Nickel	London Metal Exchange	US$/Metric Tonne
10	Silver Fix LBM	London Bullion Market	US-Cents/
11	Soya beans, No.1 Yellow	US Department of Agriculture	US-Cents/Bushel
12	Soya Oil, Crude Decatur	US Department of Agriculture	US-Cents/Pound
13	Raw Sugar-ISO	Public Ledger	US-Cents/Pound
14	Gasoline	Dow Jones Energy Service	US-Cents/Gallon
15	Wheat No.2,Soft Red	US Department of Agriculture	US-Cents/Bushel
16	LME-SHG Zinc 99.995%	London Metal Exchange	US$/Metric Tonne
17	Cash	-	US$

Table 5.1: Overview over the commodities. *Source: Reuters Datastream.*

Table 5.1 gives a description of the 16 commodities that are used for the backtest. Additionally, we use cash as the 17th asset with zero growth. The time series comprises a range of 3900 trading days, starting with April 14th, 1995 and ending on March 23rd, 2010. The cumulated growth of the individual assets (except cash) can be found in figure 5.1.

Observe how the growth and volatility of the individual assets differ from each other. Also observe that most of the assets have no notable growth over time, except for oil (asset 6), natural gas (asset 8), nickel (asset 9), gasoline (asset 14) and zinc (asset 16). At the end of the time series we see a distinct downturn in most of the assets before they partly recover again.

To get a reasonable function for the radius in the kernel-function, it is necessary to examine the given data further. As the analysed data points cannot be considered in the analysis of the backtest performance (in reality, they need to be "known" in advance to be used), we limit this analysis on the first 500 data points (or 12.8% of the sample). We call these points the training sample and will exclude them from the performance review. Still, they can be used to train the algorithms.

Calculating the norm of the differences for window lengths $k = 1, 2, 3$,

$$\|x_i^{i+k} - x_j^{j+k}\| \quad \forall\, i \neq j \in \{1, \ldots, 500 - k\}$$

where x_i is the vector of the growth factors of the assets at time i and $\|.\|$ is the Frobenius norm, results in $\sum_{k=1}^{3} \binom{501-k}{2} \approx \binom{500}{2} 3 = 374250$ data points for the differences. The fact

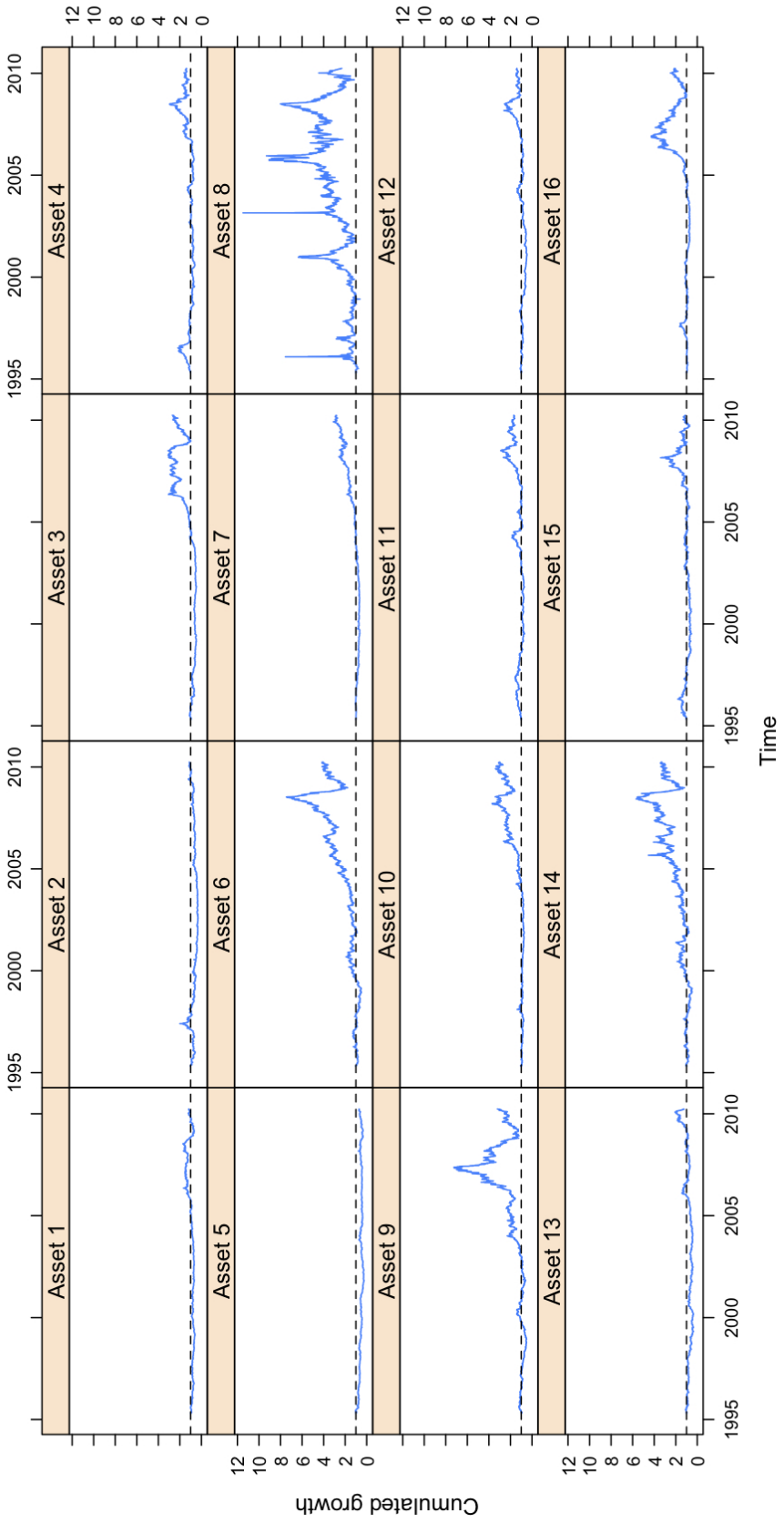
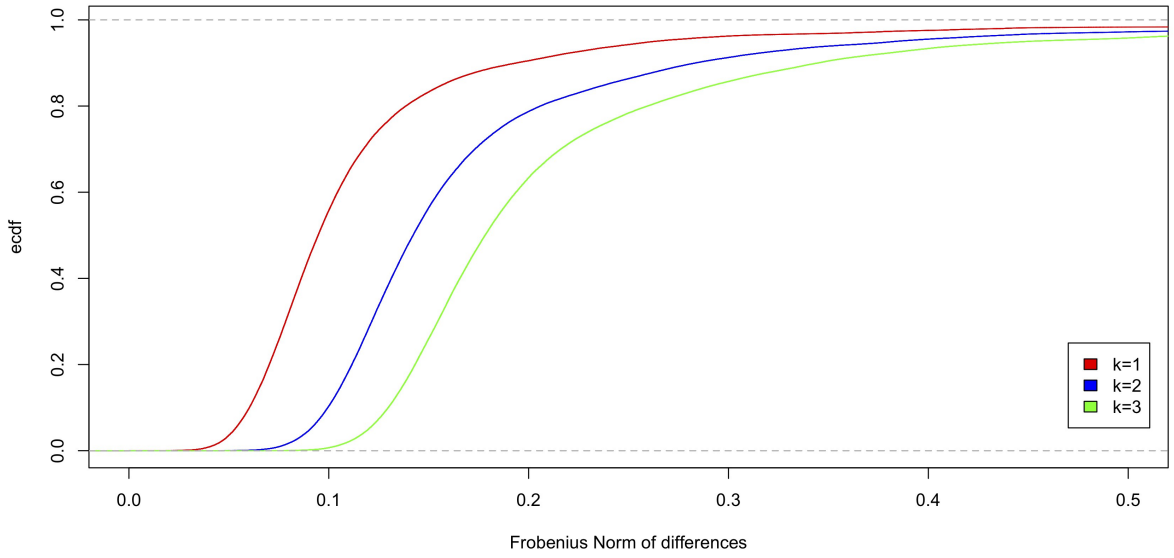
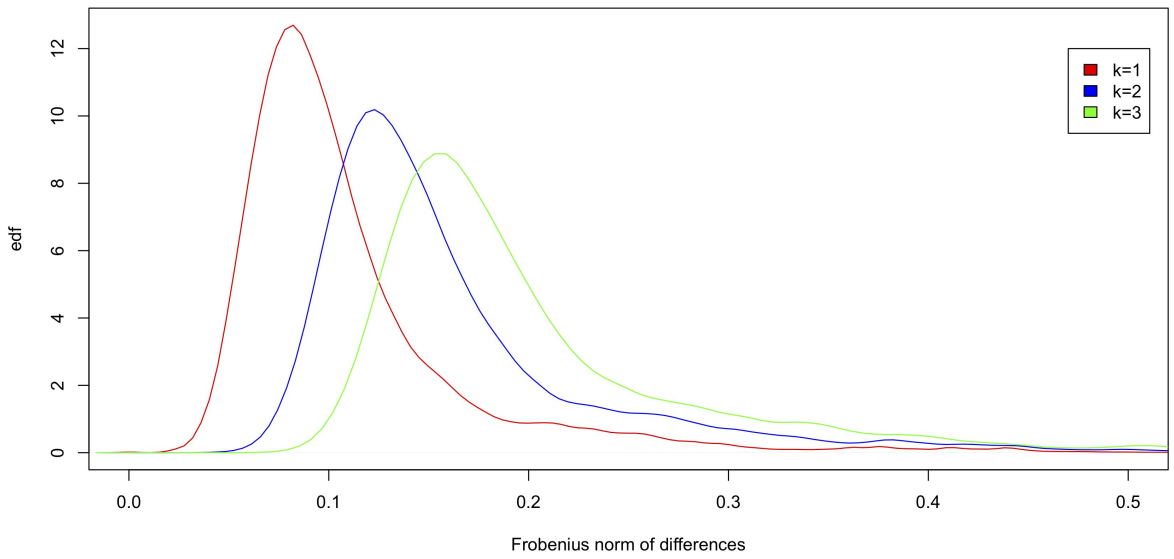


Figure 5.1: The cumulative growth of the individual commodities.



(a) Empirical cumulative density function



(b) Empirical density function (gaussian kernel)

Figure 5.2: ECDF and EDF of the Frobenius norm of pairwise differences of length k between the first 500 data points of the commodities time series.

that we always use pairs of returns leads to this high number. Table 5.2 gives the α-quantiles q_α of these differences in steps of 10%, figure 5.2 shows the empirical cumulative density functions (ecdf) and the empirical density functions (edf) of the given data.

Quantile	$q_{0.1}(k)$	$q_{0.2}(k)$	$q_{0.3}(k)$	$q_{0.4}(k)$	$q_{0.5}(k)$	$q_{0.6}(k)$	$q_{0.7}(k)$	$q_{0.8}(k)$	$q_{0.9}(k)$	$q_{1.0}(k)$
k=1	0.0530	0.0603	0.0656	0.0702	0.0744	0.0784	0.0823	0.0862	0.0903	0.0947
k=2	0.0907	0.0995	0.1060	0.1116	0.1166	0.1215	0.1263	0.1313	0.1365	0.1422
k=3	0.1199	0.1297	0.1370	0.1432	0.1489	0.1545	0.1599	0.1656	0.1717	0.1782

Table 5.2: Quantiles of Frobenius norm of differences between first 500 returns.

5.3.2 Description of the Backtests

To compare the proposed algorithms to each other, we run them with several setups and evaluate their performance on the commodity data set. Note again that we use the whole 3900 trading days in the backtest, but as we used 500 days for the analysis of the assets only the results of the last 3400 days will be evaluated in the performance analysis.
The best constantly rebalanced portfolio algorithm from algorithm 1 is straightforward, as no variables need to be chosen in advance. We will denote this backtest by S1.
The EG method in algorithm 2 depends on η. With the theoretical analysis, we can calculate a good η from the given data by using theorem 4.1. This is done in backtest S2. Still, we also want to try to choose a constant $\eta = 0.05$ (as did [Helmbold et al., 1998]), which leads to backtest S3.
For the universal portfolio, the algorithms differ by the way the integrals are approximated and by the accuracy of the approximation. As argued further above, the trapezoidal rule is not feasible for high dimensions. 17 assets would need an unusual high amount of interpolation nodes, which is why we do not use this algorithm here. For the Monte-Carlo method, we run a backtest with $T = 50000$ (backtest S4) und $T = 200000$ simulated interpolation nodes (backtest S5). From the known theoretical error bounds, the latter approximation should be twice as accurate as the former.
For the kernel method, we now need to choose how many percent of historic differences we want to consider on average when calculating the portfolio. Taking into account that we need to limit the number of experts (in our case we will choose $L = 5$, $L = 10$ or $L = 20$ and $K = 3$) to confine the runtime of the algorithm, we decide to try a linear radius function

$$r_{k,l} = a_1 k + a_2 kl$$

for three different choices of a_1 and a_2 in the backtests S6, S7 and S8 as can be seen from table 5.3. As we analysed the quantiles of the training sample above, we can also

choose a radius function that uses the empirical quantiles $q_{\alpha(l)}(k)$ in figure 5.2 and table 5.2 directly. This is done in the backtests S9 and S10 (with increasing accuracy L), again as explained in table 5.3.

Finally, the nearest neighbour algorithm is applied by using a linear function p_l that chooses 2.5% to 50% of the nearest neighbours for the experts in S11 and S12 with increasing accuracy L. We also try choosing p_l in a way that uses more experts that consider small numbers of nearest neighbours for strategy S13. All those backtests and some of its properties are once again summarized in table 5.3.

Strategy	**S1**	**S2**	**S3**	**S4**	**S5**
Algorithm	1	2	2	6	6
η	-	$\frac{c}{C}\sqrt{\frac{8\ln d}{i}}$	0.05	-	-
T	-	-	-	50000	200000
Strategy	**S6**	**S7**	**S8**	**S9**	**S10**
Algorithm	7	7	7	7	7
(L,K)	(5,3)	(5,3)	(5,3)	(10,3)	(20,3)
$r_{k,l}$	$0.02k + 0.006kl$	$0.04k + 0.008kl$	$0.06k + 0.01kl$	$q_{l/(2L)}(k)$	$q_{l/(2L)}(k)$
Min α of quantile (k=1)	1.00%	3.00%	20.00%	5.00%	2.50%
Max α of quantile (k=1)	4.00%	33.00%	66.00%	50.00%	50.00%
Strategy	**S11**	**S12**	**S13**		
Algorithm	8	8	8		
(L,K)	(10,3)	(20,3)	(10,3)		
p_l	$\frac{1}{2L}l$	$\frac{1}{2L}l$	$\frac{1}{2L-l}$		
Min number of NN	5.00%	2.50%	5.26%		
Max number of NN	50.00%	50.00%	10.00%		

Table 5.3: Choice of parameters for S1 to S13.

5.3.3 Numerical Results for Backtests Related to the Best Constantly Rebalanced Portfolio

Algorithms 1, 2 and 6 can be subsumed as being related to the best constantly rebalanced portfolio, as they are constructed to compete with the (unknown) best constantly rebalanced portfolio. Therefore, a first comparison of the results of the backtests S1 to S5 seems appropriate as they have the same scope. First we are interested in the behaviour of the portfolio vector over time. For this we draw a heat map of the portfolio vectors delivered by the relevant strategies, as can be seen for S1 in figure 5.3. The darker a line in an asset row is, the higher the fraction of wealth invested into this asset is at that time. We can clearly see that this strategy concentrates quickly on a few assets. The performance of this strategy is summarized in figure 5.5. The results here are very disappointing. After 3400 trading days we are approximately at the same wealth level as at the beginning. Calculating the performance measures given in table 5.4, this strategy

seems to be an inferior choice for portfolio selection as its performance is weak in every respect considered here.

Looking at S2 and S3, we see by figure 5.3 that this algorithm leads to a more diversified portfolio selection. Most of the time, the portfolio is almost equally distributed among the assets (the dark line for asset 8 is misleading, as the maximum fraction here is only 0.12). Therefore, it is not surprising that S2 and S3 follow the equally weighted portfolio in the performance charts in figures 5.6 and 5.7. There is nearly no difference between the choice of a constant or variable η.

Finally, the universal portfolios backtests S4 and S5 differ only a little bit from the EG algorithm's results, as can be seen from figures 5.4, 5.8 and 5.9. The most significant difference can be found in the runtime of the backtest, were the EG method is extremely fast with around one second for the complete (!) backtest, while the algorithms 1 and 6 need several hours, as summarized in table 5.4. Enhancing the accuracy of the integral approximation has nearly no effect at all, except an increase of runtime.

As all those backtests closely follow the equally weighted portfolio, it is not surprising that they do not avoid the severe downturn in 2008 to 2010 in the assets. They also fail to recover from that downturn until the end of the time series.

Summarizing, overall performance of S2 to S5 is not too good, but still significantly better than S1. A positive aspect of S2 to S5 is that the portfolio vectors only change rarely and changes are small, so rebalancing is done without much transactions.

	S1	S2	S3	S4	S5
RT (runtime in hours)	3.5000	0.0003	0.0003	2.0000	8.0000
W (daily average growth rate)	0.0000	0.0004	0.0004	0.0004	0.0004
s (standard deviation of daily returns)	0.0283	0.0102	0.0091	0.0091	0.0091
R_{Sharpe} (Sharpe ratio of daily returns)	0.0137	0.0440	0.0462	0.0460	0.0460
R_{Sortino} (Sortino ratio of daily returns)	0.0151	0.0461	0.0468	0.0467	0.0467
$MD(3400)$ (max. drawdown in 3400 days)	0.7425	0.4950	0.4883	0.4889	0.4889
AR (annualized returns)	0.0000	0.1054	0.1002	0.1001	0.1000
PR_{mon} (% of months with positive returns)	0.4936	0.5962	0.5833	0.5833	0.5833
PR_{ann} (% of years with positive returns)	0.6154	0.6923	0.6923	0.6923	0.6923

Table 5.4: Performance measures of S1 to S5.

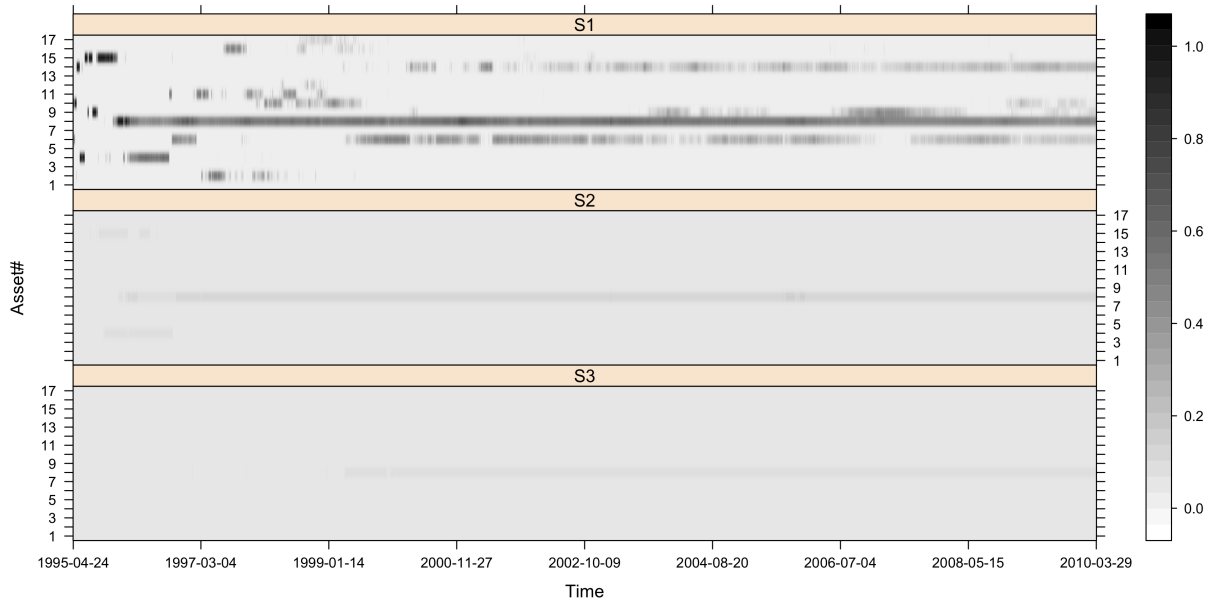


Figure 5.3: Portfolio vectors of S1, S2 and S3.

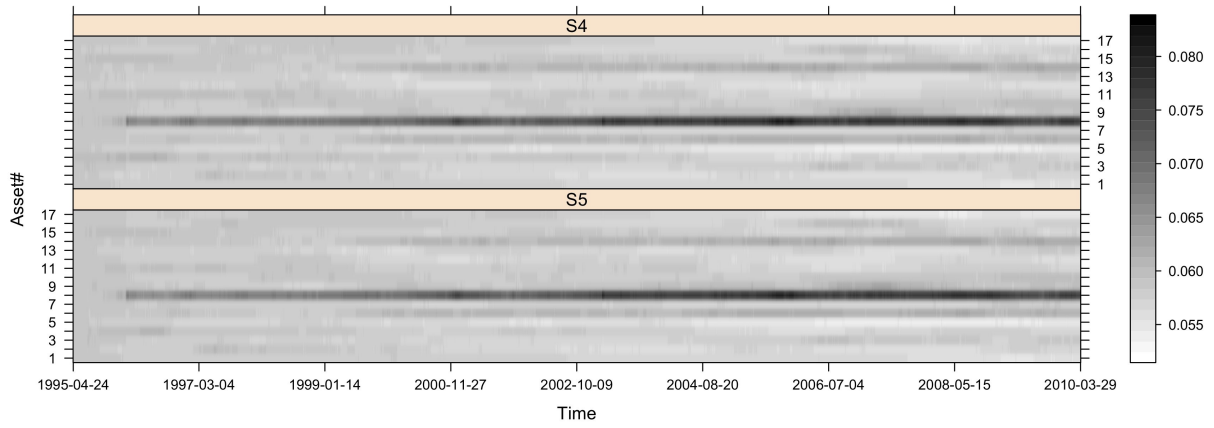


Figure 5.4: Portfolio vectors of S4 and S5.

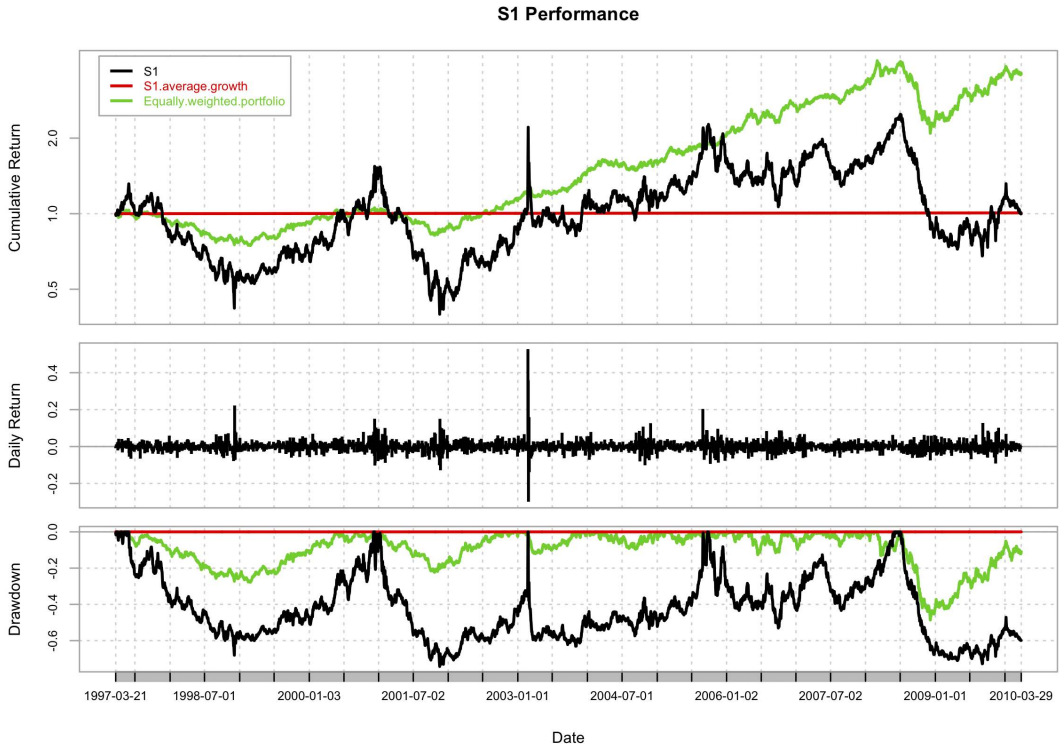


Figure 5.5: Performance chart of S1.

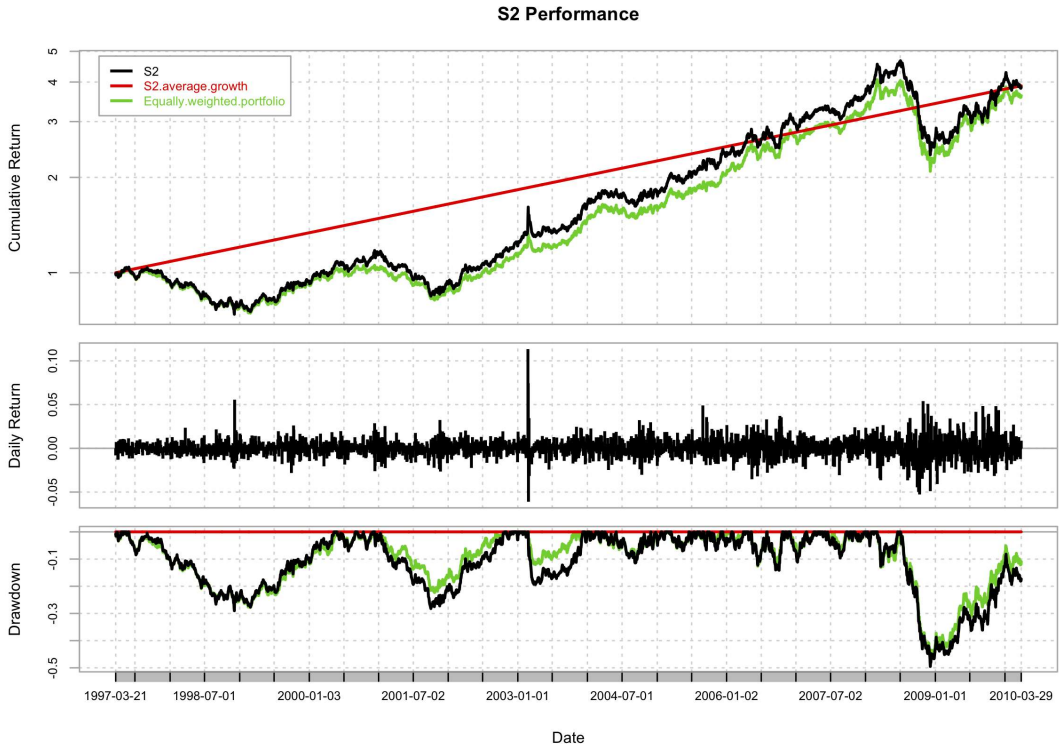


Figure 5.6: Performance chart of S2.

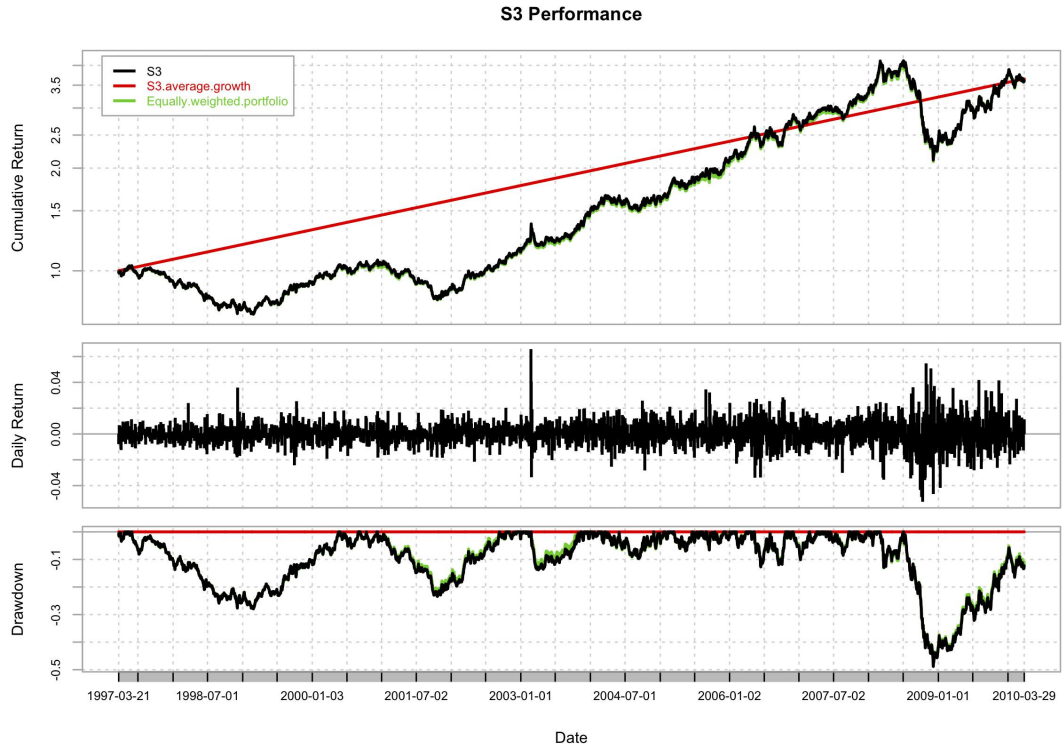


Figure 5.7: Performance chart of S3.

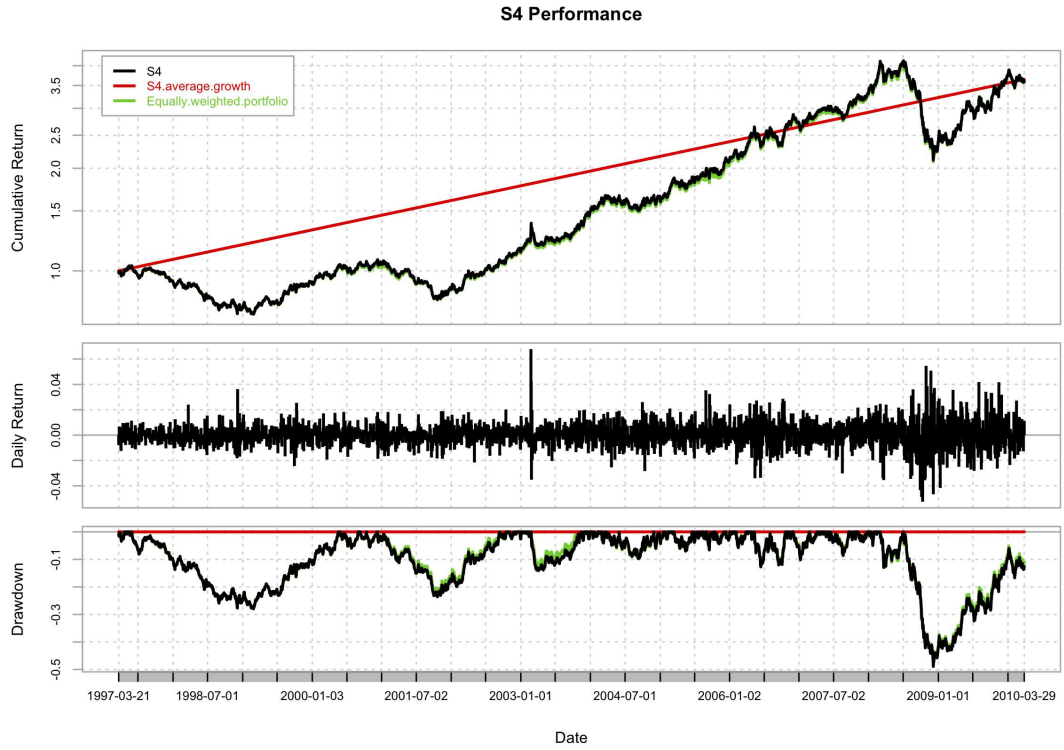


Figure 5.8: Performance chart of S4.

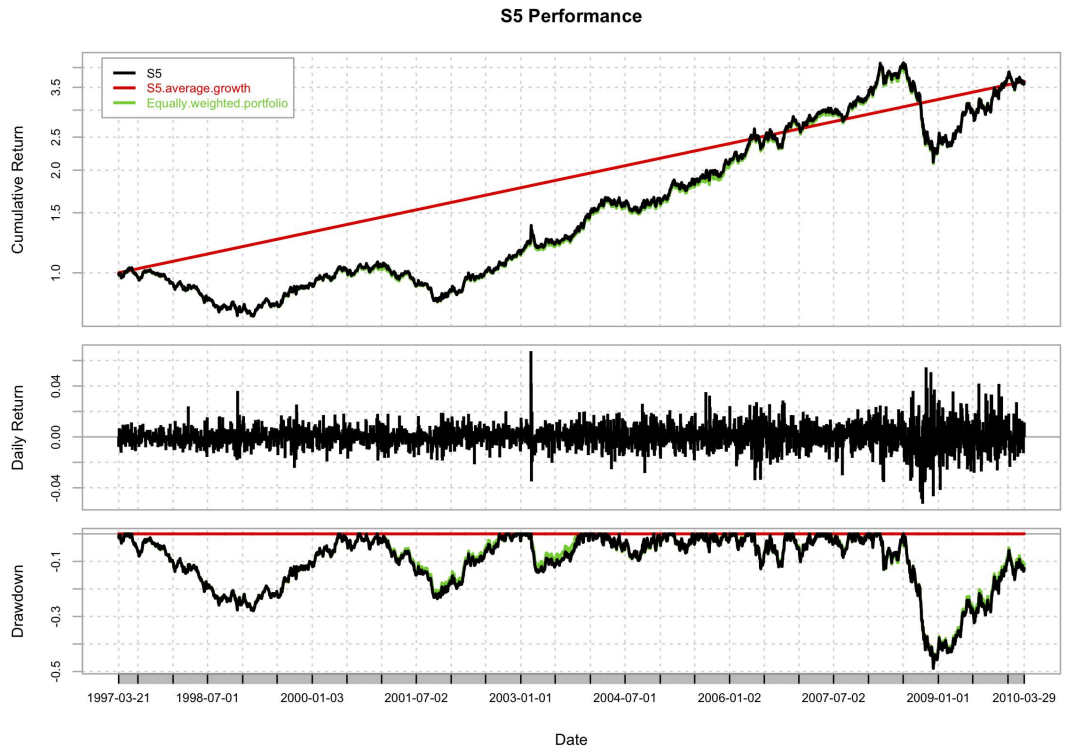


Figure 5.9: Performance chart of S5.

5.3.4 Numerical Results for the Kernel Backtests

	S6	S7	S8	S9	S10
RT (runtime in hours)	0.5000	10.0000	28.0000	11.0000	21.0000
W (daily average growth rate)	0.0010	0.0014	0.0015	0.0015	0.0015
s (standard deviation of daily returns)	0.0168	0.0251	0.0284	0.0245	0.0238
R_{Sharpe} (Sharpe ratio of daily returns)	0.0680	0.0663	0.0657	0.0726	0.0726
R_{Sortino} (Sortino ratio of daily returns)	0.0747	0.0767	0.0780	0.0829	0.0827
$MD(3400)$ (max. drawdown in 3400 days)	0.4544	0.5373	0.5386	0.5194	0.5181
AR (annualized returns)	0.2878	0.4102	0.4558	0.4554	0.4411
PR_{mon} (% of months with positive returns)	0.5577	0.5705	0.6346	0.6026	0.6026
PR_{ann} (% of years with positive returns)	0.7692	0.6923	0.6923	0.6923	0.6923

Table 5.5: Performance measures of S6 to S10.

The kernel algorithm is much more complex than the algorithms considered before. It is not approximating a constantly rebalanced portfolio, but explicitly searching for dependencies in the time series provided. One crucial factor for this method - as already explained before - is the choice of the radius function. A reasonable goal is to select on average a specific amount of nearest points in the time series (for example the 10%, 20% and 30% nearest points). Therefore, we used the analysis of the quantiles to determine radius functions that use little percentages on average and some that use more (S6 to S8). As we know the empirical distribution of the relevant norms from table 5.2, we can also directly relate to those when giving the radius function (S9 and S10). We would expect those strategies to get more and more effective from S6 to S10.

Investigating S6 to S8, this seems to be the case. Considering more historic returns leads to a higher performance. It also leads to higher volatility, but the growth in volatility is not as big as that in the growth rates (as can be seen from the standard deviation s and Sharpe ratio in table 5.5). Cumulated growth is reasonably good, especially for S7 and S8 with a final wealth of 103.82 and 159.14 respectively after the 3400 trading days. This relates to a daily average growth rate of around 0.0015. As mentioned earlier, this is much less than growth rates from backtests on stocks. But commodities should rather be compared to currency investments, as they too have no relevant growth of their own over time. For this, the results are quite good, as the backtest with a kernel algorithm on currencies in [Györfi et al., 2008a] resulted in an annualized return of 26.9% to 58.3% compared to an annualized return of 45.58% for S8.

S9 and S10 directly use the empirical quantiles and one would expect them to perform better than the strategies that rely on a linear interpolation of those quantiles. Still, growth of S9 is slightly worse than that of S8, but significantly better than that of S10, which is surprising at first as S10 has more experts to choose from. But a closer look

at the experts reveals that S9 already had the best experts and S10 was averaging over more, but worse experts. Even though weights should be lower for those, they are still considered, thus leading to an inferior growth rate. This could be called "bad luck" and of course be different for other data. However, one can see from table 5.5 that performance measures that take into account volatility s would prefer both S9 and S10 over S6 to S8. This is most likely the positive consequence of the improved accuracy of those two setups. A closer look at the final wealth of the experts in table 5.6 and 5.7 reveals no systematic best choice of the radius, as the best experts jump from big to small radius and from long to short k. This variation of the best expert is probably due to the fact that the radius function only chooses on average a certain number of neighbours - in the individual case the number of chosen neighbours in the history can vary. Thus, this method is usually not as "stable" as the nearest neighbour based approach that can be seen in the next section. This finding is in line with [Györfi et al., 2008a].
The maximum downturn is again severe and it is disappointing that the kernel method also fails to avoid the downturn in 2008 to 2010. At least the portfolio recovers and reaches a new peak after a short time. It would probably help the performance if such an event would have already appeared in the time series before - in the given time series there is almost constant overall growth of the assets until 2008, thus there is no reference observation for such an event.
Concerning the development of the portfolio vector, we see that a method considering a small radius is more diversified than that with a large radius. This can be linked to the fact that small radius functions only take few values into account when averaging, especially if the history is short. The more observations there are, the more points go into the average. In general, we see that the portfolio vectors change significantly from time to time. More precisely, they begin to jump especially between Asset 8, 14 and 15. Clearly, we can see that the portfolio vectors have not such a stable development as in the backtests with S1 to S5.

	S6			S7			S8		
l/k	1	2	3	1	2	3	1	2	3
1	3.211	4.215	4.594	4.312	40.961	33.372	169.534	*117.515*	9.046
2	3.958	9.937	2.490	12.593	152.379	*158.778*	426.836	34.618	7.404
3	3.787	12.356	24.338	208.208	***470.746***	47.235	732.061	12.794	*12.125*
4	2.413	*76.361*	54.290	122.857	70.199	8.897	***780.966***	16.845	8.616
5	*11.466*	39.149	***414.912***	*426.836*	34.618	7.404	437.833	4.688	6.516

Table 5.6: Cumulated growth of experts for S6 to S8.

	S9				S10		
l/k	1	2	3	l/k	1	2	3
				1	4.571	65.379	11.377
1	14.047	82.580	13.705	2	14.047	82.580	13.705
				3	56.323	19.872	14.068
2	28.410	30.866	17.466	4	28.410	30.866	17.466
				5	137.479	63.510	15.241
3	218.751	83.989	113.910	6	218.751	83.989	113.910
				7	738.789	203.528	54.011
4	159.521	242.962	32.022	8	159.521	242.962	32.022
				9	134.850	210.225	74.081
5	44.024	82.550	*204.635*	10	44.024	82.550	204.635
				11	170.672	170.652	75.041
6	289.370	282.760	118.636	12	289.370	282.760	118.636
				13	237.871	114.160	129.682
7	233.797	*564.200*	128.038	14	233.797	*564.200*	128.038
				15	182.821	236.891	*467.869*
8	111.025	127.392	130.615	16	111.025	127.392	130.615
				17	559.229	159.824	120.964
9	842.858	284.157	189.909	18	842.858	284.157	189.909
				19	1181.850	65.538	180.288
10	***2437.800***	187.200	63.644	20	***2437.800***	187.200	63.644

Table 5.7: Cumulated growth of experts for S9 and S10.

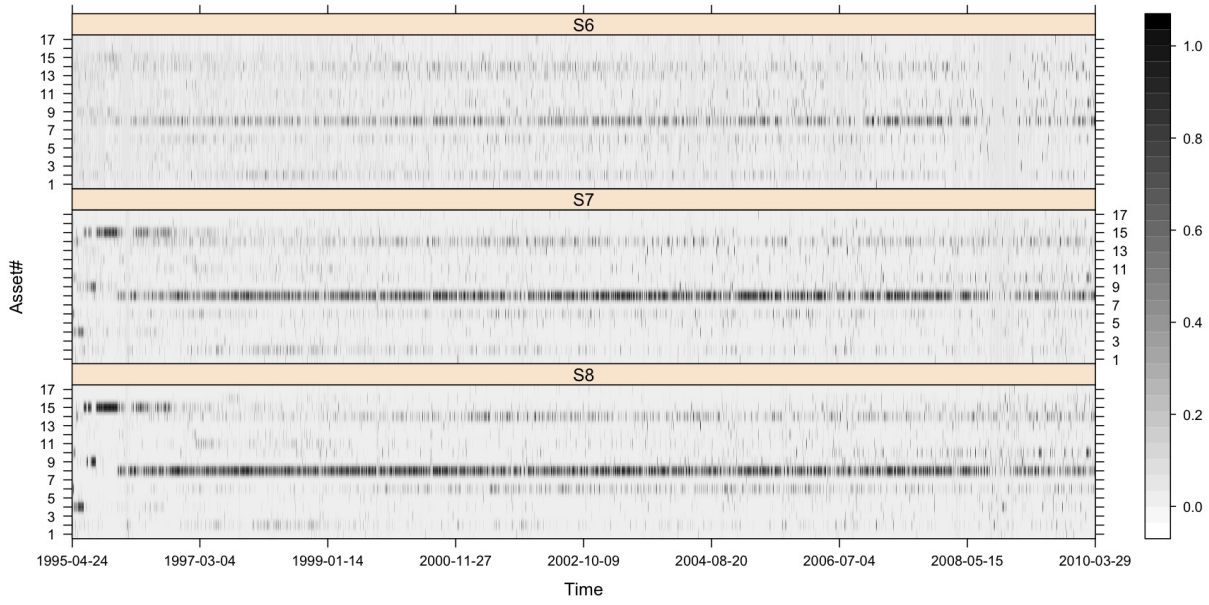


Figure 5.10: Portfolio vectors of S6, S7 and S8.

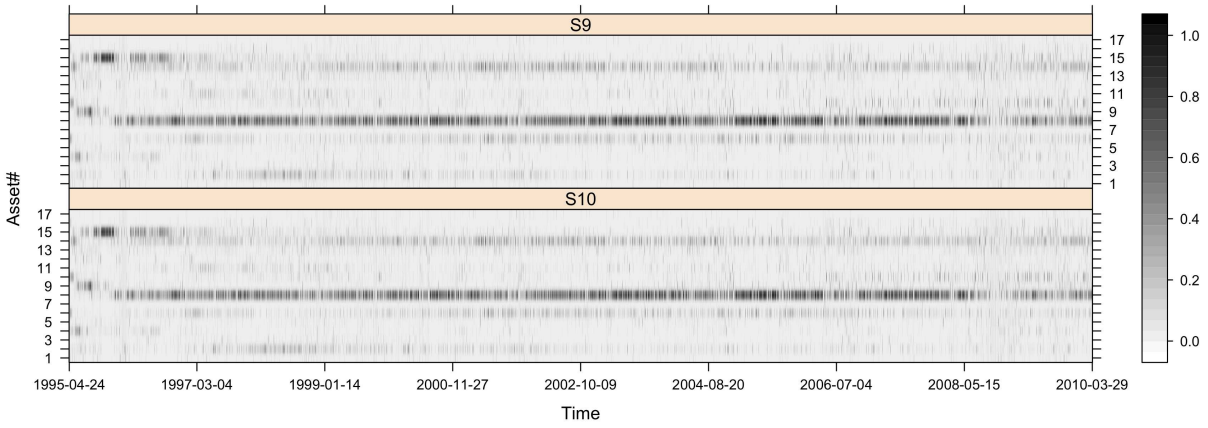


Figure 5.11: Portfolio vectors of S9 and S10.

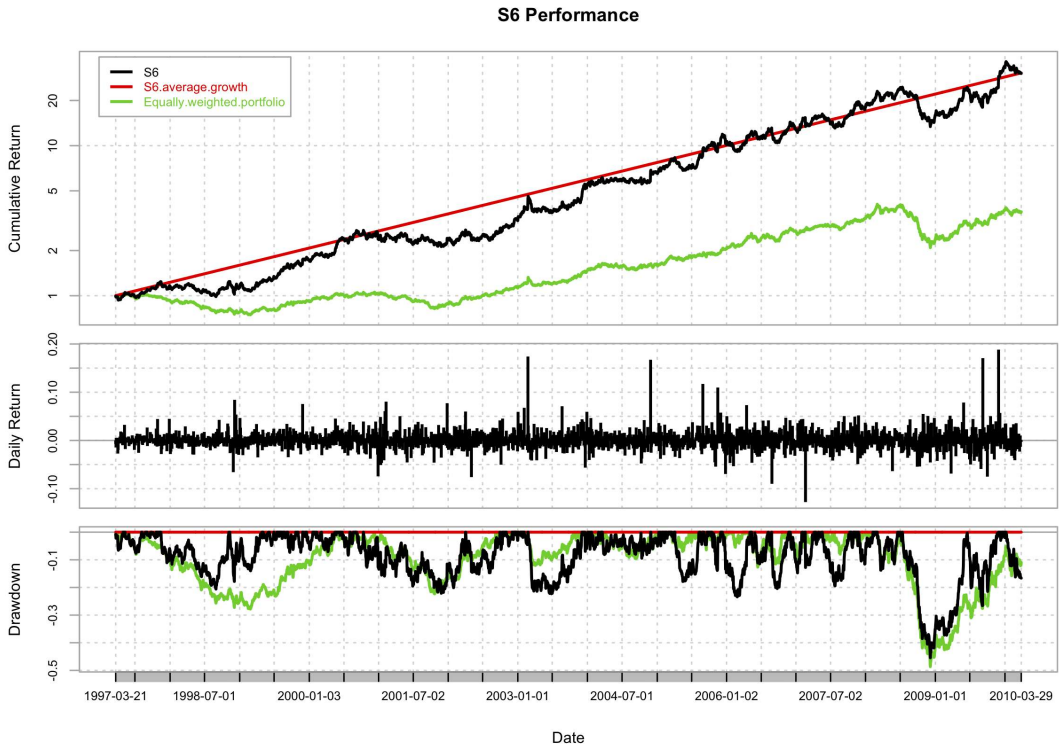


Figure 5.12: Performance chart of S6.

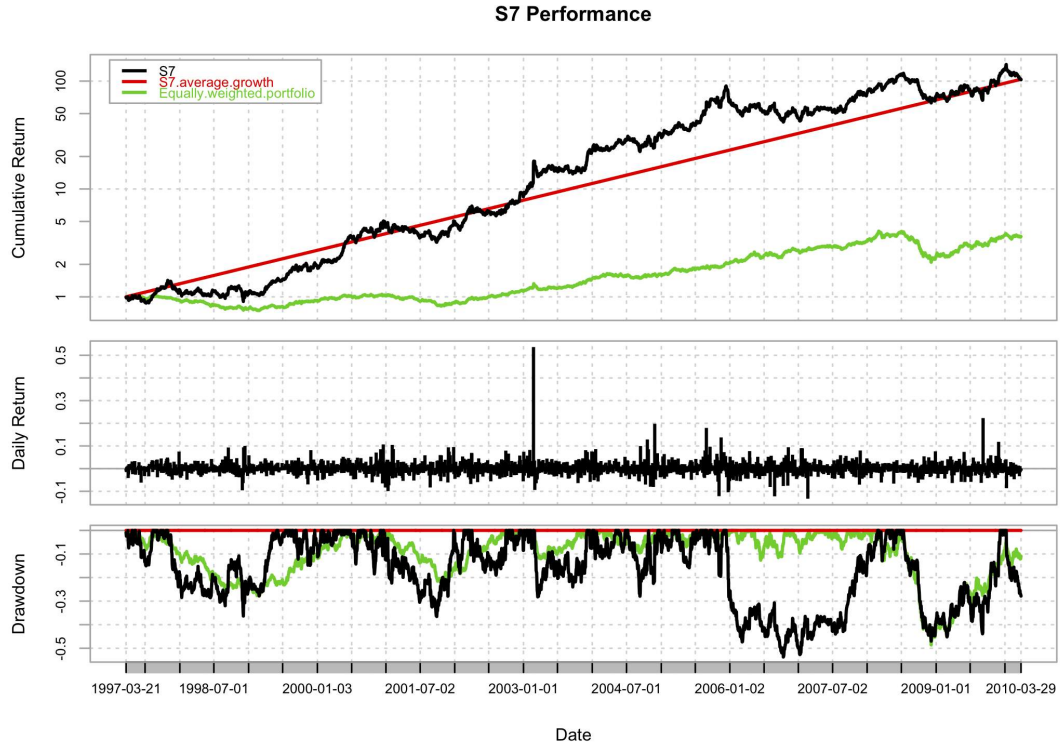


Figure 5.13: Performance chart of S7.

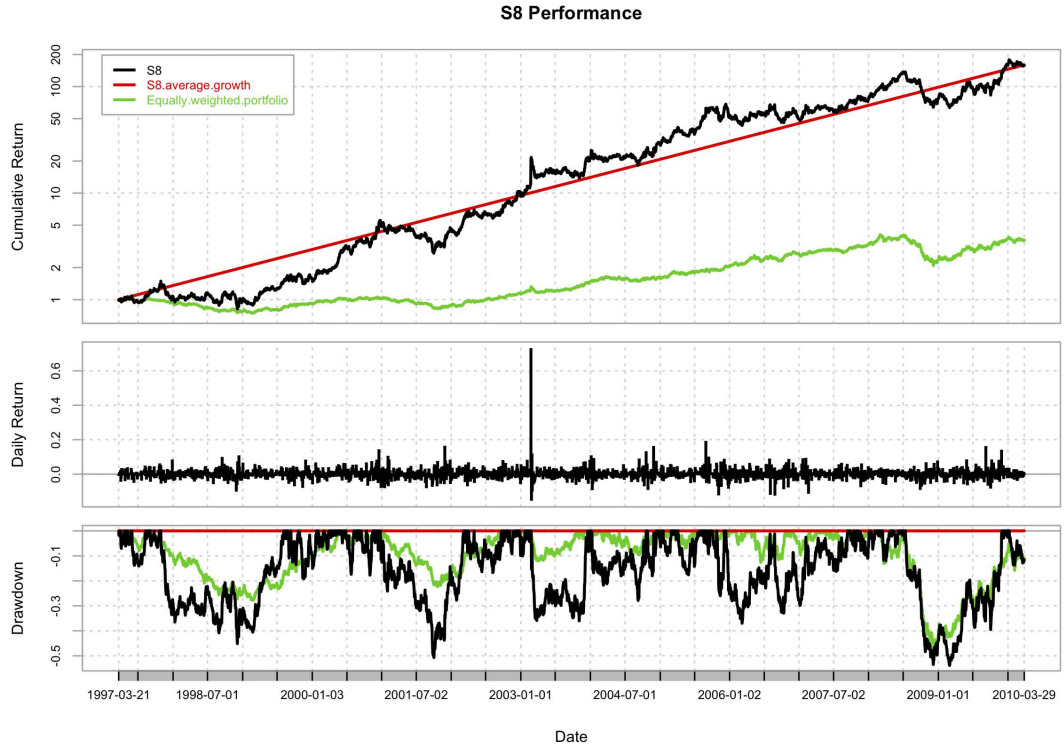


Figure 5.14: Performance chart of S8.

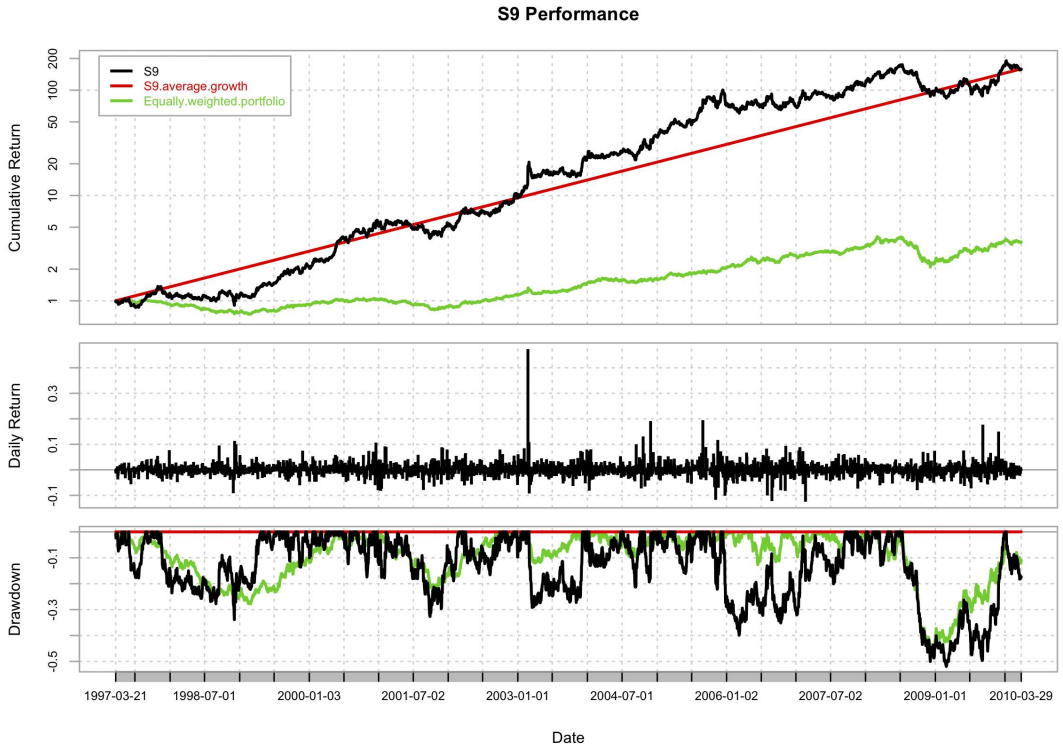


Figure 5.15: Performance chart of S9.

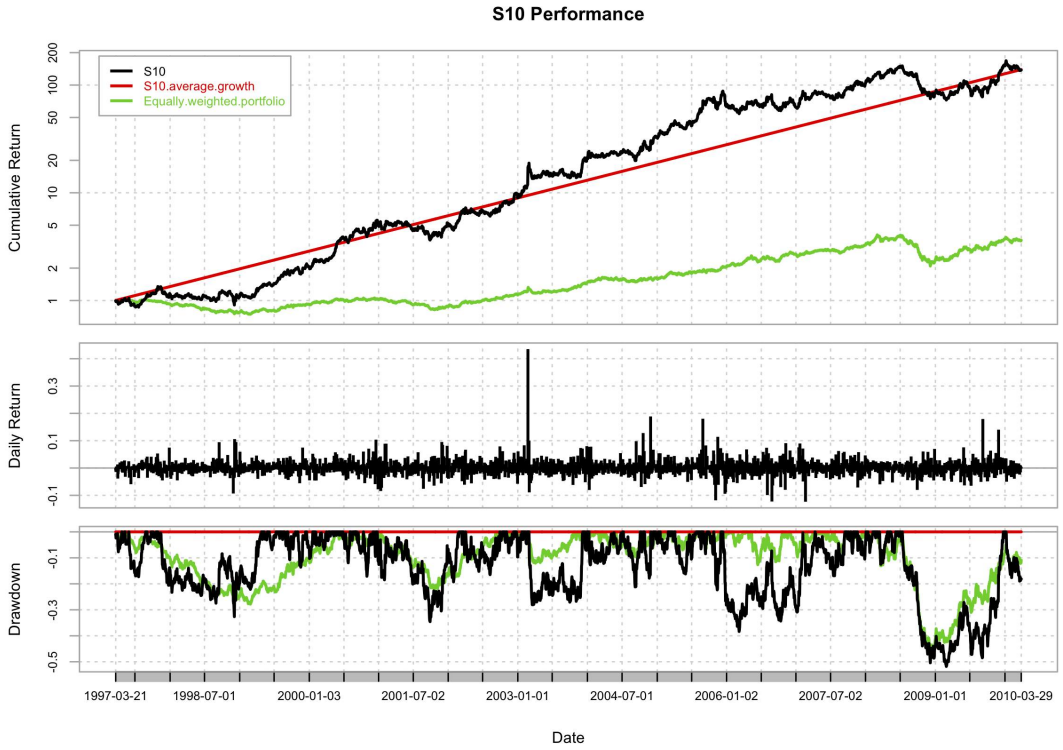


Figure 5.16: Performance chart of S10.

5.3.5 Numerical Results for the Nearest Neighbour Backtests

The nearest neighbour based backtests differ in the choice of the percentage of nearest neighbours considered in the local average. For S11 and S12, we chose two equidistant grids (10 and 20 values) within 2.5% to 50%, whereas we concentrated on a smaller range between 5% and 10% for S13. We relate this to the findings of [Györfi et al., 2008a] that the nearest neighbour based method is more robust and usually has its best experts in a smaller to medium percentage range.

Again as in the kernel analysis, the portfolio vector jumps quite a lot, concentrating on the same three assets (8,14,15) as in S6 to S10. Increasing the accuracy L is again not leading to a higher growth rate, but to less volatility and a better Sharpe and Sortino ratio. This is again in line with the analysis for the kernel method. The concentration on the lower percentages of nearest neighbours does also not really improve the average growth rate too much, but reduces volatility even more. Furthermore, time needed for the backtest is reduced significantly by using S13, as the optimization is done over less points in every step. It should still be noted that this "guess" of p_l is working for the current time series, but need not necessarily be the best choice in other cases.

Looking at the experts, we see that our guess of having the best experts in the lower percentages was correct. In general, the nearest neighbour based experts perform better than the kernel experts, which we relate again to the better robustness of the nearest neighbour method.

Disappointingly, the nearest neighbour based algorithms also fail to avoid the downturn in the years from 2008 to 2010, leading to a high maximum downturn. Again, this is most likely linked to the fact that no such severe downturns occur simultaneously in all assets in the given time series before 2008.

All in all, the results of the nearest neighbour based method are satisfying. Especially the decreasing volatility (that is also reflected by a high percentage of months and years with positive returns) is a benefit of this method, as is its robustness.

	S11	S12	S13
RT (runtime in hours)	22.5000	41.0000	11.0000
W (daily average growth rate)	0.0017	0.0017	0.0018
s (standard deviation of daily returns)	0.0302	0.0294	0.0283
R_{Sharpe} (Sharpe ratio of daily returns)	0.0712	0.0720	0.0780
R_{Sortino} (Sortino ratio of daily returns)	0.0849	0.0858	0.0920
$MD(3400)$ (max. drawdown in 3400 days)	0.5038	0.4981	0.4793
AR (annualized returns)	0.5406	0.5369	0.5815
PR_{mon} (% of months with positive returns)	0.6282	0.6410	0.6090
PR_{ann} (% of years with positive returns)	0.8462	0.8462	0.8462

Table 5.8: Performance measures of S11 to S13.

	S11				S12				13		
l/k	1	2	3	l/k	1	2	3	l/k	1	2	3
				1	185.148	1852.010	630.599	1	174.526	*1868.840*	1314.810
1	181.397	*1283.710*	***2197.020***	2	181.397	1283.710	***2197.020***	2	*392.351*	586.074	***4400.730***
				3	192.314	387.175	520.293	3	223.274	145.688	690.682
2	152.041	904.421	494.871	4	152.041	904.421	494.871	4	149.002	350.973	253.894
				5	48.215	250.469	450.562	5	91.694	985.783	345.734
3	124.640	129.192	353.672	6	124.640	129.192	353.672	6	117.990	921.908	545.870
				7	291.954	449.297	298.316	7	36.369	219.462	104.808
4	*283.544*	942.455	627.947	8	283.544	942.455	627.947	8	164.683	315.504	306.740
				9	154.539	*1580.180*	677.777	9	125.724	193.536	268.308
5	65.850	230.619	329.179	10	65.850	230.619	329.179	10	215.373	16.844	93.434
				11	115.394	230.236	488.746				
6	89.631	263.053	847.906	12	89.631	263.053	847.906				
				13	84.451	216.400	1178.810				
7	92.912	89.862	961.162	14	92.912	89.862	961.162				
				15	184.673	84.015	262.801				
8	179.620	91.206	157.726	16	179.620	91.206	157.726				
				17	179.378	68.543	124.504				
9	441.209	29.178	83.661	18	*441.209*	29.178	83.661				
				19	280.237	37.012	51.822				
10	237.670	45.780	54.062	20	237.670	45.780	54.062				

Table 5.9: Cumulated growth of experts for S11 to S13.

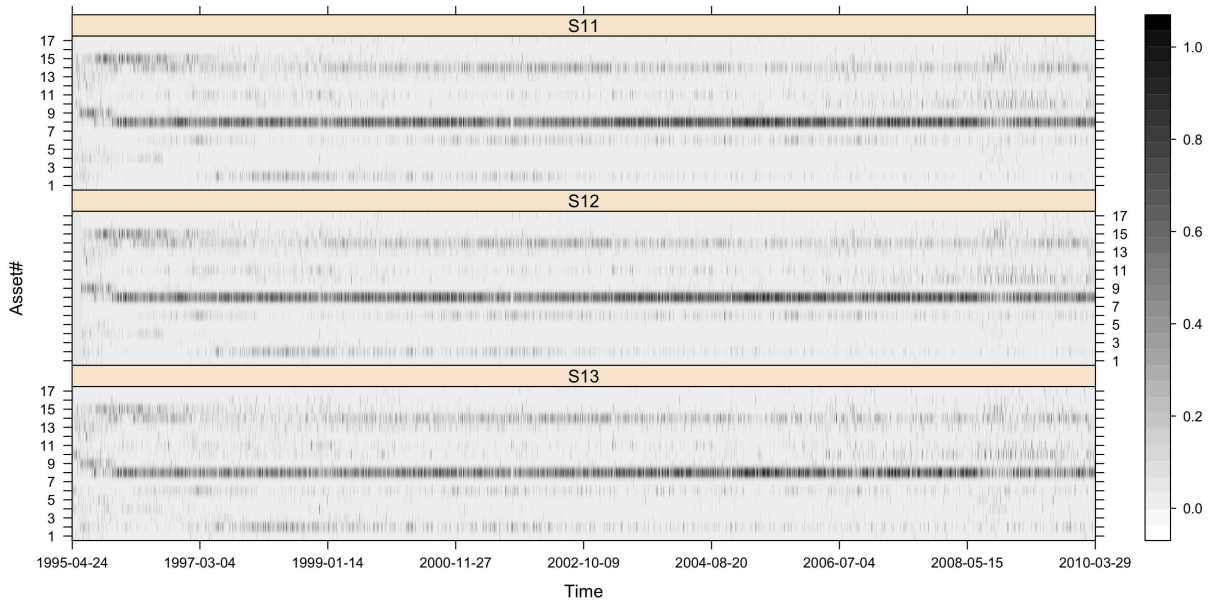


Figure 5.17: Portfolio vectors of S11, S12 and S13.

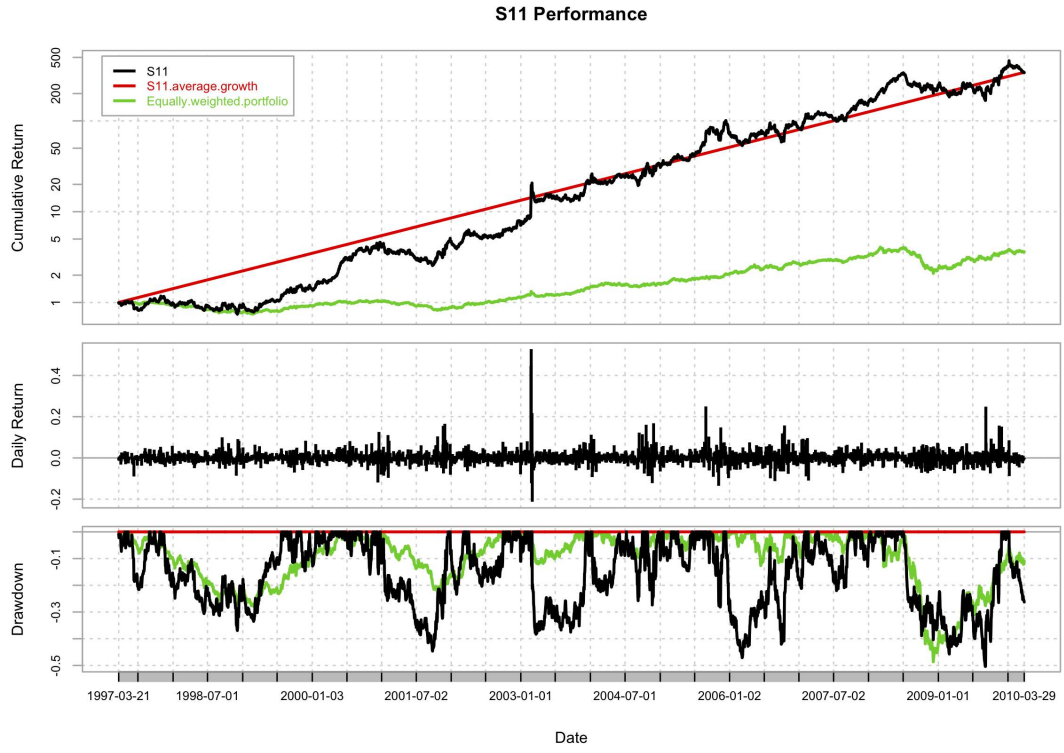


Figure 5.18: Performance chart of S11.

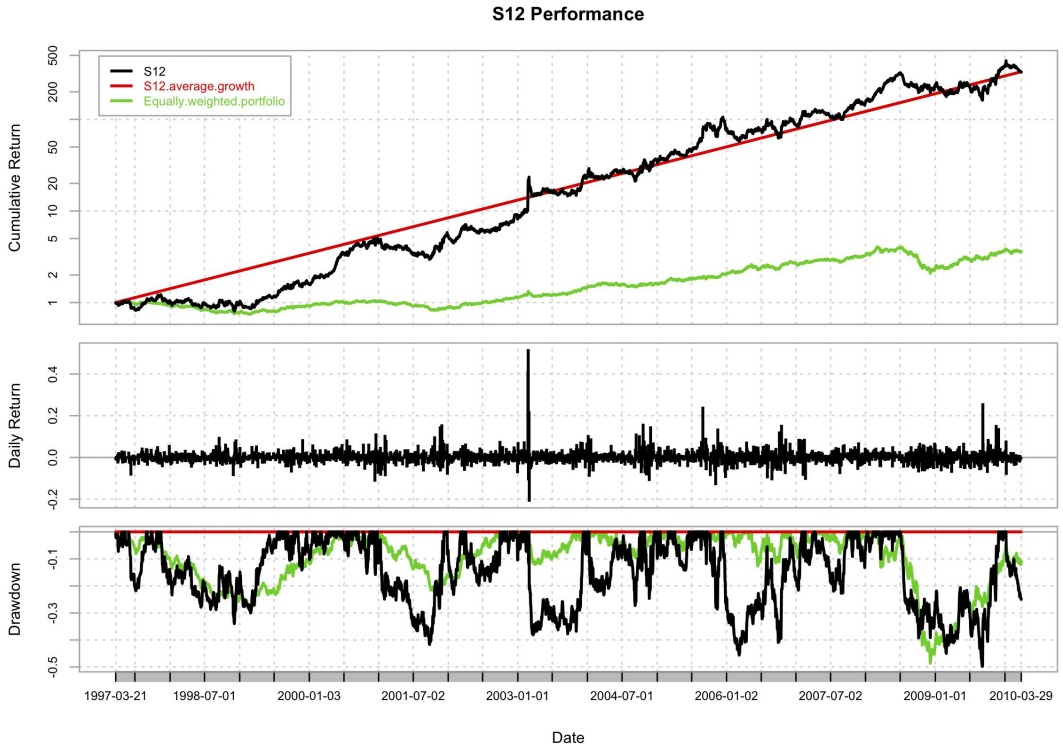


Figure 5.19: Performance chart of S12.

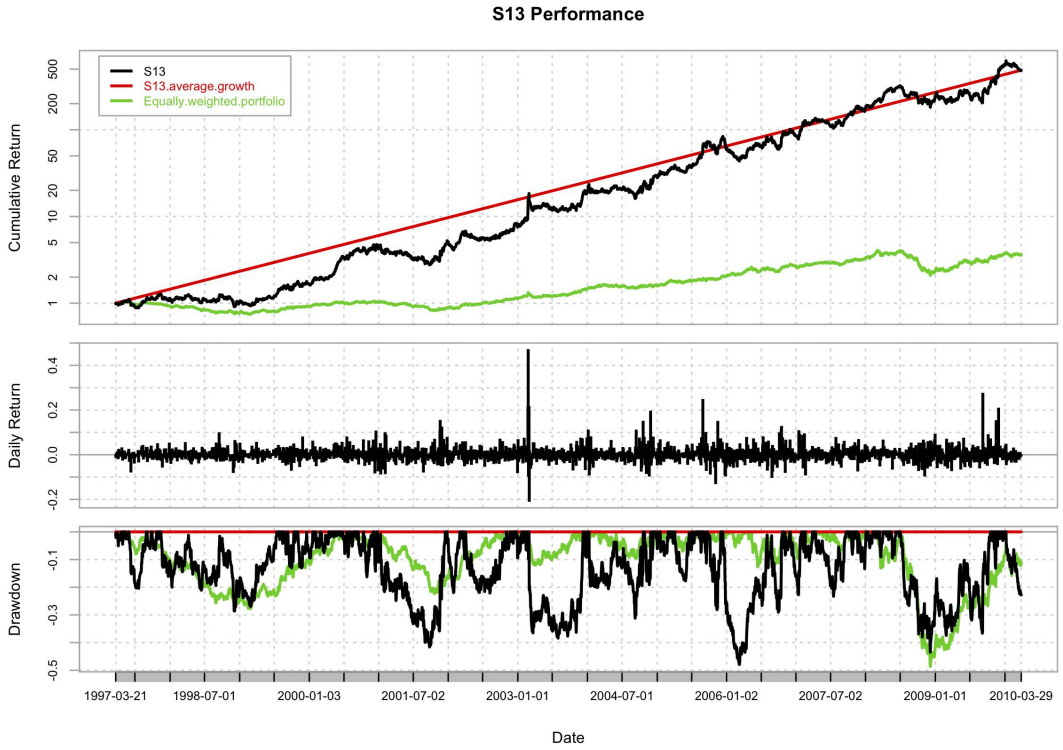


Figure 5.20: Performance chart of S13.

5.3.6 Summary of All Numerical Results

After comparing the numerical results in groups, we now turn to the big picture, represented by the comparison of our measures of all backtests in figure 5.21.
Looking at the runtime first, we see increasing runtime with increasing complexity as expected. However, even the longest run of 41 hours for the whole backtest of S12 is reasonable, as this represents the time needed for 3900 individual prediction steps. Having run the backtest once in a "preparation" phase that may take some time, one can make single new predictions based on the weights resulting from the previous backtest rather fast. This means that even more accuracy (and therefore more runtime) would be appropriate for a real-world application.
Daily average growth rates for S1 to S5 are very small compared to those taking into account dependencies. The difference here is significant. Within the individual methods, parameter choice changes growth only slightly (except for S7 where the radius function was definitely chosen too restrictive).
Standard deviation varies quite a bit, even within the groups. Higher accuracy seems to lead to less volatility, as reflected by the pairs (S9, S10) and (S11, S12). With this, enhanced accuracy also leads to a higher Sharpe and Sortino ratio.
Maximum drawdowns are not significantly different from backtest to backtest. Unfortunately, no setup was able to avoid the largest downturn in the crisis years 2009 and 2010, as was already mentioned several times earlier. Still, if we look at the development of the drawturns in the performance charts of the backtests as given above, we see that the complex algorithms used in S6 to S13 usually have deeper downturns than S1 to S6. On the other hand they have high peaks and usually recover quickly.
The annualized return is just another measurement of the growth rate. It can be interpreted as the annual interest gained on the investment. Therefore, the same consequences can be drawn as for the growth rate.
Looking at the ratio of positive returns, annually and monthly, we see that this ratio in general increases with increasing complexity. This analysis gives an information about the time horizon an investor should have when investing according to the proposed setups. The higher the percentage of positive returns, the more likely a positive return is within one year/one month. Going into even more detail, figure 5.22 and 5.23 show the monthly and yearly returns over the backtest period. For S6 to S13 we can clearly see that negative months are only a bit negative, while positive months have high positive returns. This is even stronger for the annual returns for every year of these strategies. A one or two year investment horizon seems therefore appropriate for these strategies.

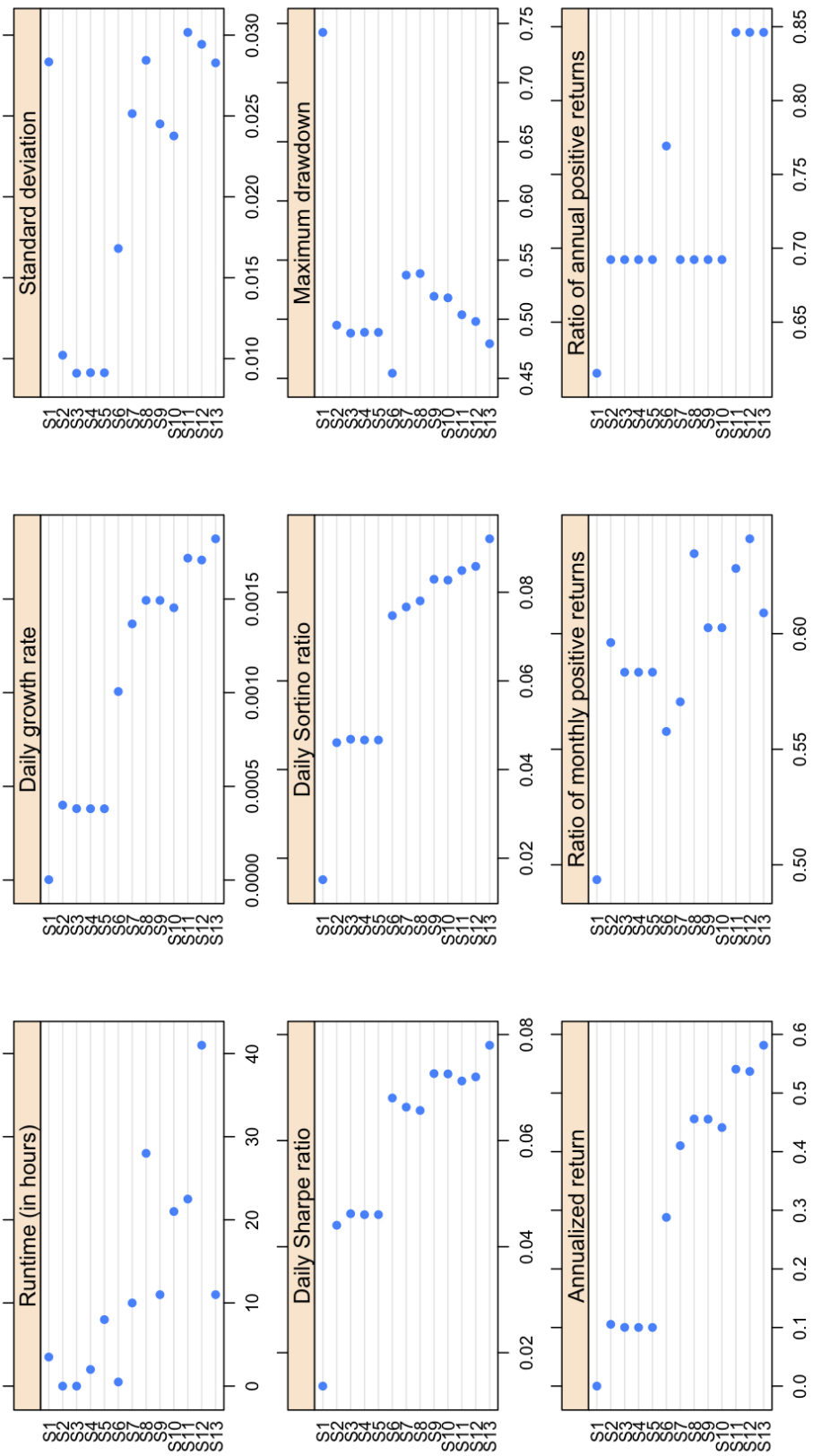


Figure 5.21: Performance comparison of all backtests.

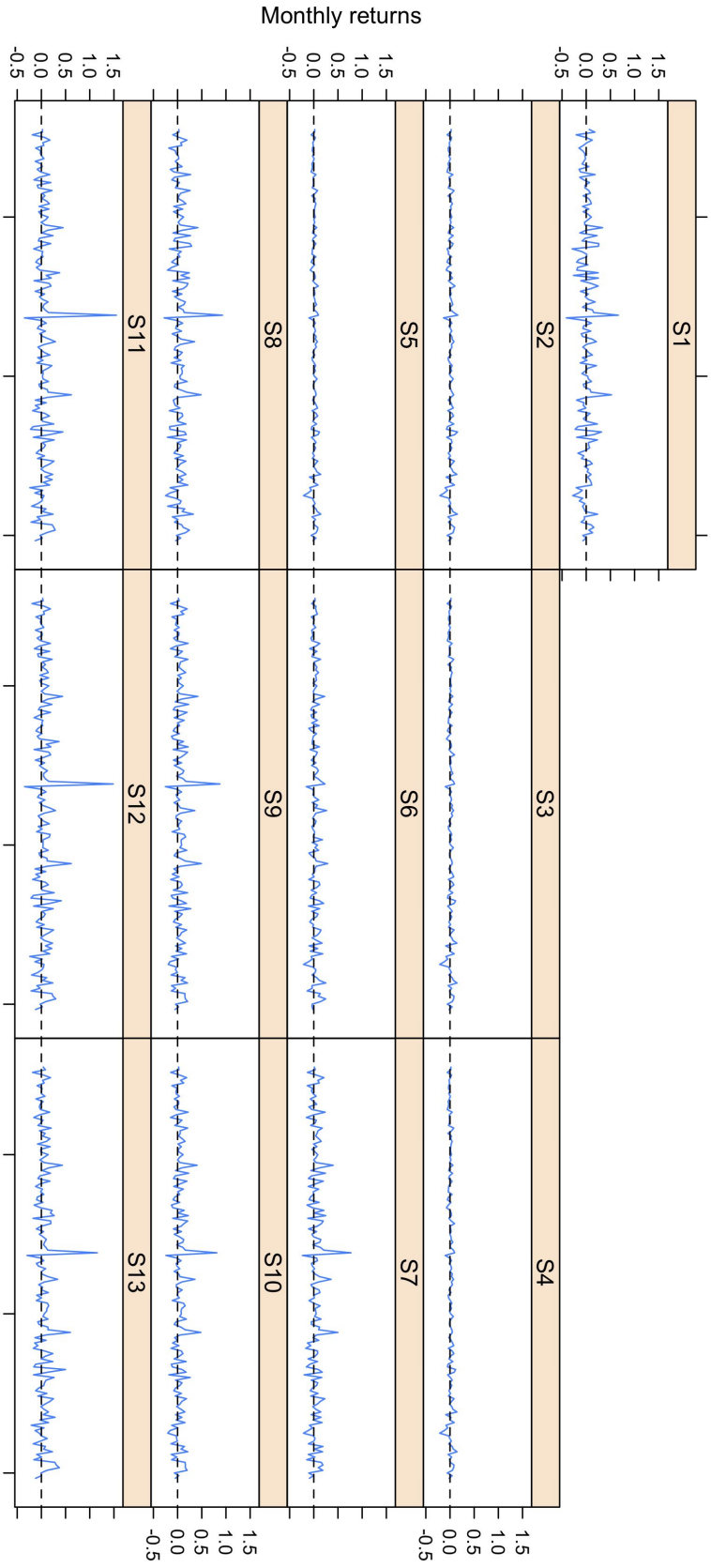


Figure 5.22: Monthly returns of all backtests.

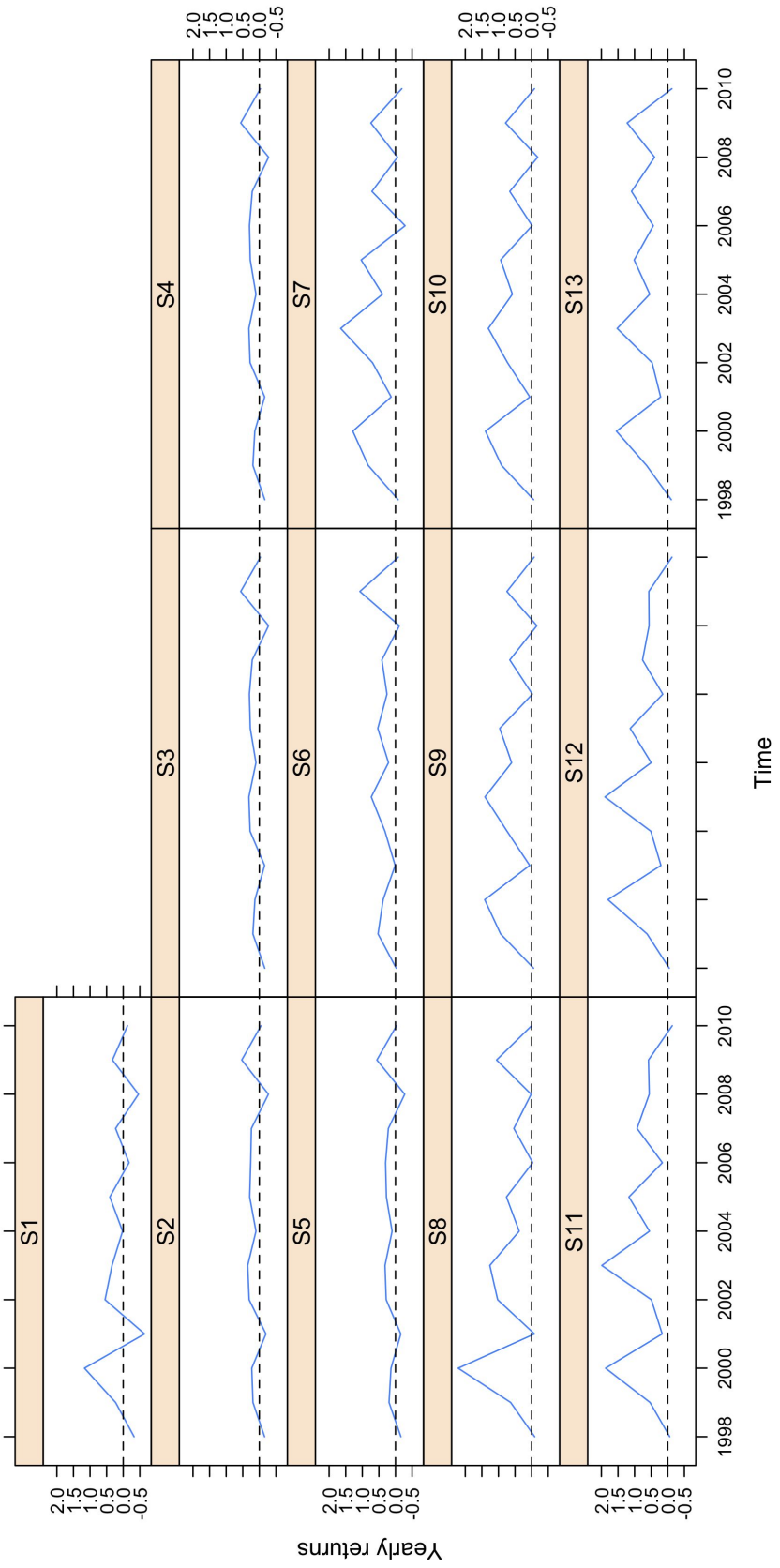


Figure 5.23: Annual returns of all backtests.

Chapter 6

Conclusion

In this thesis, we have investigated results of the growth-optimal portfolio theory. We have shown how to construct sequential growth-optimal portfolio strategies and how to establish five algorithms in this framework. A critical discussion of this idea showed that while the concept of growth optimality is not compatible with some specific economic paradigms, it seems a valuable investment strategy by itself. Following this theoretical analysis, the presented algorithms were applied to a real-world data set, namely returns of 17 commodities over 3900 trading days from April 14th, 1995 to March 23rd, 2010. A thorough analysis of the results showed a lack of effectiveness for the simpler algorithms that are constructed to approximate the best constantly rebalanced portfolio. On the other hand, the more complex kernel and nearest neighbour based algorithms showed a promising performance. These results are mostly in line with findings from other papers for backtests with different assets.

While the concept of growth-optimality has been around for quite some time, the theory still lacks several explanations that would need closer attention. One important aspect to investigate is the question how to (in whatever sense) optimally choose parameters in the kernel and nearest neighbour algorithms. Another possible point would be to assess the rate of convergence of the different methods, as fast convergence towards the best expected growth rate is obviously desirable and therefore preferred. A third important question that is often raised in the literature is how to handle transaction costs (see for example [Györfi et al., 2007]). The performance of the backtests looks fantastic, but daily returns are only in 1/10 of a percent range. Considering that the kernel and nearest neighbour based algorithms require a lot of rebalancing with transactions costs being rarely less than 0.5% of the trading volume, one can easily imagine how fast this profit is vanishing.

Considering these points, growth-optimal portfolio theory looks like a promising field for further research and an interesting aspect for practitioners.

Bibliography

[Algoet, 1994] Algoet, P. (1994). The strong law of large numbers for sequential decisions under uncertainty. *IEEE Transactions on Information Theory*, 40(3):609–633.

[Algoet and Cover, 1988] Algoet, P. and Cover, T. (1988). Asymptotic optimality and asymptotic equipartition properties of log-optimum investment. *The Annals of Probability*, 16(2):876–898.

[Biau et al., 2010] Biau, G., Bleakley, K., Györfi, L., and Ottucsák, G. (2010). Nonparametric sequential prediction of time series. *Journal of Nonparametric Statistics*, 22(3):297–317.

[Breiman, 1957] Breiman, L. (1957). The individual ergodic theorem of information theory. *The Annals of Mathematical Statistics*, 28(3):809–811.

[Breiman, 1960] Breiman, L. (1960). Investment policies for expanding businesses optimal in a long-run sense. *Naval Research Logistics Quarterly*, 7(4):647–651.

[Breiman, 1961] Breiman, L. (1961). Optimal gambling systems for favorable games. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, pages 65–78. University of California Press, Berkeley.

[Carl and Peterson, 2010] Carl, P. and Peterson, B. G. (2010). *PerformanceAnalytics: Econometric tools for performance and risk analysis.* R package version 1.0.3.2.

[Cesa-Bianchi and Lugosi, 2006] Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, learning, and games*. Cambridge University Press, Cambridge.

[Christensen, 2005] Christensen, M. (2005). On the history of the Growth Optimal Portfolio. *Preprint, University of Southern Denmark*, 389.

[Cover, 1991] Cover, T. (1991). Universal portfolios. *Mathematical Finance*, 1(1):1–29.

[Cover, 2002] Cover, T. (2002). An algorithm for maximizing expected log investment return. *IEEE Transactions on Information Theory*, 30(2):369–373.

[Devroye, 1986] Devroye, L. (1986). *Non-uniform random variate generation*. Springer, New York.

[Györfi et al., 2002] Györfi, L., Kohler, M., Krzyzak, A., and Walk, H. (2002). *A distribution-free theory of nonparametric regression*. Springer, New York.

[Györfi et al., 2006] Györfi, L., Lugosi, G., and Udina, F. (2006). Nonparametric kernel-based sequential investment strategies. *Mathematical Finance*, 16(2):337–357.

[Györfi et al., 2007] Györfi, L., Ottucsák, G., and Urbán, A. (2007). Empirical log-optimal portfolio selections: a survey. Machine Learning Summer School 2007, MLSS 2007, Tuebingen.

[Györfi and Schäfer, 2003] Györfi, L. and Schäfer, D. (2003). Nonparametric prediction. In Suykens, J., Horvath, G., Basu, S., Micchelli, C., and Vandewalle, J., editors, *Advances in Learning Theory: Methods, Models and Applications*, pages 341–356. IOS Press, Amsterdam.

[Györfi et al., 2008a] Györfi, L., Udina, F., and Walk, H. (2008a). Experiments on universal portfolio selection using data from real markets. *UPF Working Paper*.

[Györfi et al., 2008b] Györfi, L., Udina, F., and Walk, H. (2008b). Nonparametric nearest neighbor based empirical portfolio selection strategies. *Statistics & Decisions*, 26(2):145–157.

[Hakansson, 1971] Hakansson, N. (1971). Capital growth and the mean-variance approach to portfolio selection. *Journal of Financial and Quantitative Analysis*, 6(1):517–557.

[Hammersley and Handscomb, 1964] Hammersley, J. and Handscomb, D. (1964). *Monte Carlo methods*. Methuens Monographs on Applied Probability and Statistics, London.

[Helmbold et al., 1998] Helmbold, D., Schapire, R., Singer, Y., and Warmuth, M. (1998). On-Line Portfolio Selection Using Multiplicative Updates. *Mathematical Finance*, 8(4):325–347.

[Hull, 2006] Hull, J. (2006). *Optionen, Futures und andere Derivate*. Pearson Studium, Munich.

[Härdle, 1992] Härdle, W. (1992). *Applied nonparametric regression*. Cambridge University Press, Cambridge.

[Kallenberg, 2002] Kallenberg, O. (2002). *Foundations of modern probability*. Springer, New York.

[Kelly, 1956] Kelly, J. (1956). A new interpretation of information rate. *Bell System Technical Journal*, 35(4):917–926.

[Luenberger, 1998] Luenberger, D. (1998). *Investment science*. Oxford University Press, Oxford.

[Markowitz, 1970] Markowitz, H. (1970). *Portfolio selection: Efficient diversification of investments*. Yale University Press, New Haven, Conneticut.

[Markowitz, 1976] Markowitz, H. (1976). Investment for the long run: New evidence for an old rule. *Journal of Finance*, 31(5):1273–1286.

[Pindyck and Rubinfeld, 2009] Pindyck, R. and Rubinfeld, D. (2009). *Microeconomics*. Prentice Hall, Upper Sattle River, New Jersey, 7th edition.

[Rubinstein, 1998] Rubinstein, M. (1998). Continuously rebalanced investment strategies. In *Streetwise: The Best of the Journal of Portfolio Management*, pages 112–115. Princeton University Press, Princeton, New Jersey.

[Samuelson, 1963] Samuelson, P. (1963). Risk and uncertainty: A fallacy of large numbers. *Scientia*, 98(4-5):108–113.

[Samuelson, 1971] Samuelson, P. (1971). The "fallacy" of maximizing the geometric mean in long sequences of investing or gambling. *Proceedings of the National Academy of Sciences of the United States of America*, 68(10):2493.

[Thorp and Beach, 2006] Thorp, E. and Beach, N. (2006). The Kelly criterion in blackjack, sports betting, and the stock market. In Ziemba, W., editor, *Handbook of asset and liability management: Theory and methodology*, volume 1, chapter 9, pages 385–428. North Holland, Amsterdam.