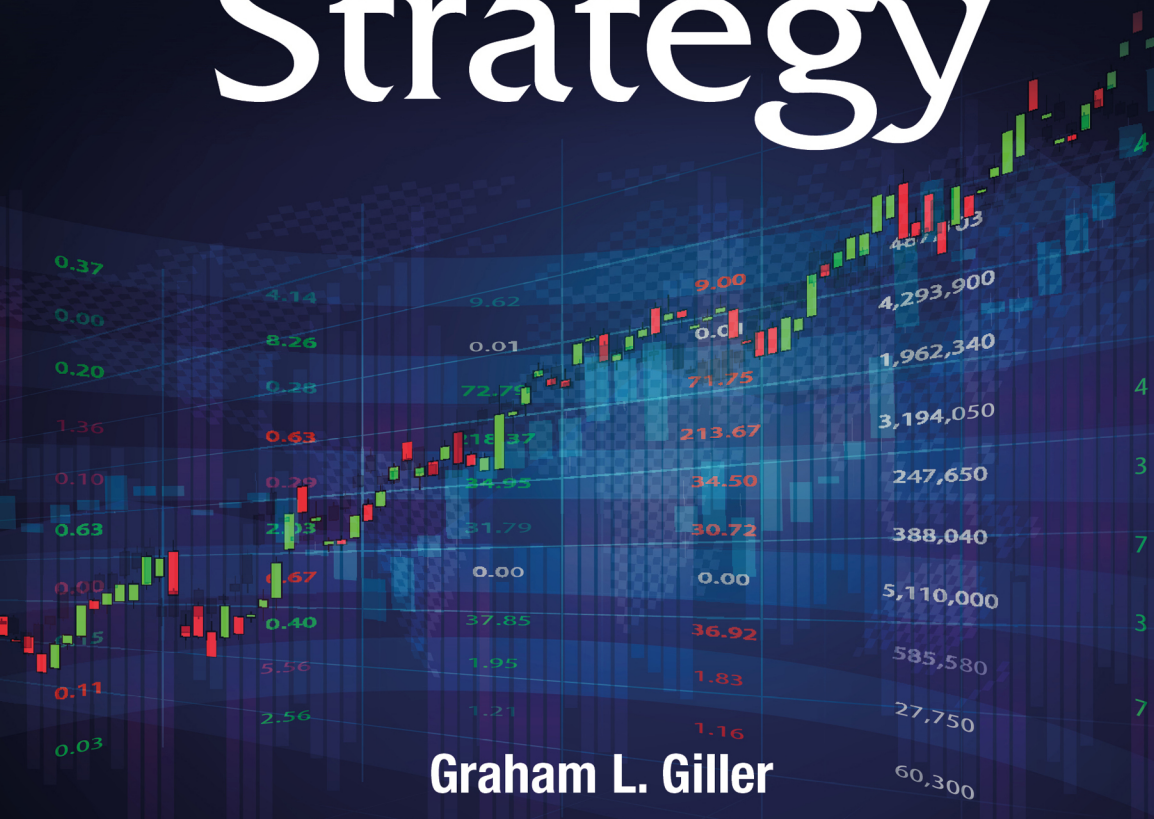
World Scientific Series in FINANCE vol. 20

Essays on Trading Strategy



Graham L. Giller

World Scientific

Essays on Trading Strategy

World Scientific Series in Finance

(ISSN: 2010-1082)

Published:*

Vol. 20 *Essays on Trading Strategy*
by Graham L. Giller (Giller Investments, USA)

Vol. 19 *Adventures in Financial Data Science: The Empirical Properties of Financial and Economic Data*
Second Edition
by Graham L. Giller (Giller Investments, USA)

Vol. 18 *Sports Analytics*
by Leonard C. Maclean (Dalhousie University, Canada) &
William T. Ziemba (University of British Columbia, Canada)

Vol. 17 *Investment in Startups and Small Business Financing*
edited by Farhad Taghizadeh-Hesary (Tokai University, Japan),
Naoyuki Yoshino (Keio University, Japan),
Chul Ju Kim (Asian Development Bank Institute, Japan),
Peter J. Morgan (Asian Development Bank Institute, Japan) &
Daehee Yoon (Korea Credit Guarantee Fund, South Korea)

Vol. 16 *Cultural Finance: A World Map of Risk, Time and Money*
by Thorsten Hens (University of Zurich, Switzerland),
Marc Oliver Rieger (University of Trier, Germany) &
Mei Wang (WHU – Otto Beisheim School of Management, Germany)

Vol. 15 *Exotic Betting at the Racetrack*
by William T. Ziemba (University of British Columbia, Canada)

*To view the complete list of the published volumes in the series, please visit:
www.worldscientific.com/series/wssf

World Scientific Series
in FINANCE vol. 20

Essays on Trading Strategy

Graham L. Giller
Giller Investments, USA

World Scientific

NEW JERSEY • LONDON • SINGAPORE • BEIJING • SHANGHAI • HONG KONG • TAIPEI • CHENNAI • TOKYO

Published by

World Scientific Publishing Co. Pte. Ltd.

5 Toh Tuck Link, Singapore 596224

USA office: 27 Warren Street, Suite 401-402, Hackensack, NJ 07601

UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

Library of Congress Cataloging-in-Publication Data
Names: Giller, Graham L., author.
Title: Essays on trading strategy / Graham L. Giller, Giller Investments, USA.
Description: New Jersey : World Scientific, [2024] | Series: World scientific series in finance, 2010-1082 ; vol. 20 | Includes bibliographical references and index.
Identifiers: LCCN 2023023574 | ISBN 9789811273810 (hardcover) | ISBN 9789811273827 (ebook) | ISBN 9789811273834 (ebook other)
Subjects: LCSH: Investments. | Financial risk management. | Portfolio management.
Classification: LCC HG6041 .G536 2024 | DDC 332.63/2--dc23/eng/20220131
LC record available at https://lccn.loc.gov/2023023574

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

For any available supplementary material, please visit
https://www.worldscientific.com/worldscibooks/10.1142/13413#t=suppl

Desk Editors: Sanjay Varadharajan/Lum Pui Yee

Typeset by Stallion Press
Email: enquiries@stallionpress.com

Printed in Singapore

For Elizabeth

I am grateful to many friends I have found over my years working in the financial services industry, including those who agreed to take a look at drafts of this work and provide feedback, in particular Alex Ribeiro Castro, Nish Gurnani, and Yuri Malitsky. I would also like to take this opportunity to thank the late Bill Ziemba, who was an enthusiastic supporter of my writing.

This page intentionally left blank

Preface

In 1993, I was a graduate student at Oxford and decided to try to follow a career in finance. My background and the context of that decision are explained in my book, *Adventures in Financial Data Science* [20]. As I always did, when I wanted to learn about a subject, I read books. Having read several "popular" volumes on finance including, of course, Michael Lewis's *Liar's Poker* [34], I headed to Blackwell's Bookshop, on Broad Street a short walk from my college, to find an academic book to read. I chose *Portfolio Theory and Investment Management*, by Dobbins, Witt, and Fielding [10], which begins by explaining *Modern Portfolio Theory* and the *Capital Asset Pricing Model* and goes on to include *The Efficient Markets Hypothesis* and *Option Theory*. I was taken by the fact that this appeared to be *real science*[1] and it made logical sense. It's fair to say that this perception was a key part of reassuring me that if I changed careers from Physics to Finance, I would still be doing intellectually legitimate work.

I have later come to believe that some of the assumptions about market data commonly made in the real world are fatally flawed, but I do not believe that this means we should throw all of this work away. As a physicist, my training included *perturbation theory*, which describes how to make modifications to easy results to accommodate actual reality, and I view the approach I've taken in some of these

[1] Physicists can be a little arrogant and assume that they are the only ones who *really* understand the Universe.

fields as a similar application. Science still works, and we can learn from "toy models," but the specific models need some modification to actually work in the real world. We need to understand what the real world does to them so that we can trade successfully. And when they don't work, we need to understand why, so we can replace them with something better.

This smaller book is a collection of "essays" on trading strategy, meaning the procedure of turning expected returns into actionable decisions in the market. The overwhelming majority of the work here is completely original work I have done[2] and a smaller quantity is my "rendering" of the work by others, which is included to provide necessary context. As before, the tone is of an *analytical narrative*, as you might see in a practitioner's seminar. There's quite a lot of mathematics and it is presented worked out organically, as a physicist would, not formal theorems and proofs. I know that this approach might irritate some of my more mathematically trained friends, but this is how I work. I do expect the reader to have more than an elementary understanding of maths, including vector analysis and calculus, but hopefully most of this is straightforward — it requires the sort of skills that the likely reader probably possesses.

Unlike, *Adventures in Financial Data Science* this is an entirely theoretical work. I will not be showing empirical evidence for the performance that derives from using a particular *holding function* versus another. The reason for this is that trading strategies and alpha models do not exist separately in the real world, even if we may analyze them separately. We can't validate a trading strategy without an alpha model that is useful and many viable strategies have low Sharpe ratios which means, as I will show, that judging their performance is problematic at best. I am also not going to be publishing alpha models that I have used recently to make money. So, necessarily, the content here should be judged along the lines of my suggestion to "think about the problem this way" or "consider the effect of the failure of our assumptions[3] as an explanatory factor for the results you observe."

[2]At least it's original to the best of my knowledge: I didn't use others' work and I've not seen it discussed by others.

[3]Which are generally i.i.d. normal distributions of returns.

My initial focus on the rigorous investigation of things like barrier trading systems came from the point of view of a professional trader in a very successful proprietary trading unit at a large investment bank. I think, at this point in time, we should also consider the sophisticated amateur, who also needs analytical support and may not find it online. Professional *retail* finance, it appears, is dominated by nostrums that are not based in science and engineering but in magical thinking and survivorship bias.

Finally, although what I have written here is influenced by conversations I had, and experiments I tried, while I was working in P.D.T. at Morgan Stanley, the contents of this book represent work that I have done while I was running my own firm, Giller Investments, as recently as this year. There is a lot of mathematics in this book, and some readers might claim that the premises it is based upon are invalid. I do not believe this is so, especially as the actual outcomes of the analysis are so straightforward and convincing. It appears that we have to do a lot of analytical work to get to these results but, ultimately, it is worth it.

Graham L. Giller
Holmdel, 2023

This page intentionally left blank

About the Author

Graham Giller is a senior data scientist and financial researcher, with 25 years experience on Wall Street. He has a doctorate from Oxford University in Experimental Elementary Particle Physics, where his field of research was statistical cosmic ray astronomy which featured large scale computer-based data analytics in an era where that was uncommon outside of academia.

He joined Morgan Stanley in London in 1994, where he worked on derivatives pricing and then was an early member of the now famous Process Driven Trading unit (PDT) run by Peter Muller in New York. Within PDT, he managed the Futures Group, developing and managing systematic trading systems for financial futures and futures-options. He also worked on theoretical analysis of optimal trading strategy.

In 2000, he founded his own investment fund, Giller Investments (New York), LLC, as a Commodity Pool Operator/Commodity Trading Advisor and was involved in various quantitative proprietary trading ventures over the next 12 years. In 2012–2013, he built an innovative platform for geospatially aware bidding in Google ads auctions for an Internet marketing agency and in 2013 was recruited to Bloomberg LP to lead the data science effort within the Global Data division. This role included many board-level meetings (including meetings with Mike Bloomberg) to define a vision for how

Bloomberg LP could create value-added premium content using predictive analytics as well as the delivery of systems to enhance operational efficiency through machine learning. Graham joined JPMorgan in 2015 as Chief Data Scientist, New Product Development, and he was appointed Head of Data Science Research in 2016. At JPMorgan, he co-authored the U.S. patent *System and Method for Prediction Preserving Data Obfuscation*. He was recruited to Deutsche Bank as Head of Primary Research to create the unit that incorporated alternative data into analyst research content. While there he built systems for nowcasting important US macroeconomic statistics, including Non-Farm Payrolls, Average Hourly Earnings, Consumer Sentiment, and Inflation Expectations, all based on private survey research and with results available up to three weeks before the public releases. He led the team that delivered the following: Market and Opinion Research, Public Data Capture, Social Media Analytics, and *ad hoc* projects including a system to robotically interrogate Amazons *Alexa* devices for brand positioning intelligence.

He is the author of *Adventures in Financial Data Science* as well as this book and an earlier volume *Electronics: Principles and Applications*, which was written while he was an undergraduate at Oxford and published while he was a graduate student. He currently writes a regular column in Wilmott Magazine. Dr. Giller is a popular public speaker and can be contacted at graham@gillerinvestments.com.

Contents

List of Figures

List of Tables

This page intentionally left blank

Introduction

Outline of This Book

In *Adventures in Financial Data Science*, I systematically investigate the observed distributions of returns in financial markets and demonstrate, I hope convincingly, that the assumption of normal distributions of returns is wholly unsupported by empirical analysis [18]. However, I show additionally that we may find other distributions of returns that provide far more accurate models for financial data while retaining some of the analytical tractability of the normal. Spurred on by the rejection of the normal and the observation of *success for methods not corresponding to canonical theory*, and the failure of that canonical theory in practice, I sought to discover whether these two observations were linked. Was the latter a consequence of the former? This book follows the intellectual journey I took to answer that question and to arrive at solutions for optimal discrete trading strategies. Based on my experiences, and the theory contained within, I advocate that they are the most useful way for small traders to interact with markets.

The work is divided into a set of essays which address important stops on this journey. Although the essays build upon each other, there will be some repetition of key concepts to aid the reader who is "dipping in" to an essay rather than reading from cover to cover. The mathematical frameworks I present will range from conventions about notation to solutions to the problem of composing a policy for a trader in the possession of information about the future distribution of returns. Anticipating that many of the readers of these essays

are not academics or students of finance, I will define, as rigorously as I can, the concepts I am using and demonstrate their consequences. Those definitions will assume the reader has a postgraduate understanding of some mathematical concepts and notations and, to be clear, this subject is one which demands an operational facility in the methods of calculus. However, do not look upon this collection of writing as a formal collection of lectures or a textbook in finance or optimization theory. I have *never* studied either of those subjects formally, and my work probably shows that. In the language of academic publishing, it is a *monograph*, or detailed study addressing a single concept, in this case, optimal trading strategy, and it is done in the way that feels "right" to me as the author.

The concept addressed is *not* alpha building[1] or the practice of building predictions of future distributions of returns. That is a key part of the trader's work, but we assume it has already been completed. Not only that, we assume the alphas we are using are unbiased, meaning that their expected error is zero, and efficient, meaning that they have the smallest possible variance. The topic of study is *necessarily* thoroughly embedded in the domains of stochastic processes, optimal decision-making, probability, and statistics, and I will use the language and terms from those disciplines in this work.

Many quants and studious home traders will encounter concepts such as the *Sharpe ratio* very early on in their journey into quantitative trading as a discipline. It is a topic that makes intuitive sense without the additional context of mean–variance optimization and *Modern Portfolio Theory*. In the first essay, I will examine the Sharpe ratio as a statistic for the analysis of trading performance. To get to that topic, I will follow the conventional route of introducing *Mean–Variance Optimization* and take a slight detour into a discussion of the *Kelly criterion*. The main result of that essay is a demonstration that the Sharpe ratio is a poor statistic to use to grade trading, or investment, performance *ex post*, or on historical data. As a *statistic*,[2] it is particularly inefficient. However, that does not mean that it is not a suitable objective *ex ante*.

[1] See section Alpha building and backtesting on page xxxvi for a more formal definition of that term.

[2] In the sense defined in statistical literature, such as Kendall [30].

The following sections of this essay define some essential terms which are "taken for granted" and necessary for the development of the analysis presented in Essay 1. Essay 2 will extend that language to more formal topics such as policy functions, of which the "holding function" is one of the interests to traders. That essay also includes some notations on the "vectorization" of concepts of norms and absolute values that will, doubtless, not make pure mathematicians happy but are intuitively useful. My goal is not to define a complete formal theory[3] but to provide sufficient tools to get to the end of our journey. I found that one of the things I needed to do was invent a language to express the ideas I had about how to solve the problems I was presented with. The notations and ideas I use arose from thinking about these problems.

I found that I needed a more concrete definition of the information set to make solutions to trading strategies possible so, after defining key concepts, we proceed into a somewhat abstract definition of "information sets," which is a term tossed around by quants in need of a language to describe their decision-making rules. This is expressed in terms of *The Laws of Information for Traders*, two of which are axiomatic in nature and the third, formally, a theorem that arises from those axioms. A key result here is that information sets evolve in a purely stochastic manner. I am hesitant to write "random" in this introduction, as there is a tendency to associate that word with "meaningless," but that is the nature of information and this perspective agrees with the *Information Theory* of Shannon. The information that is a key part of trader's decisions is unpredictable to them and it must be treated that way so that we can solve optimal trading strategies. The essay then develops some key concepts around important functions of the information set. These functions are the mathematical devices that make a somewhat abstract and unknowable concept a reality that can be programmed into a computer.

Essay 2 also includes a relatively deep, and hopefully intuitive, discussion of how *Linear Programming* works, which is important for two reasons. First, some of the problems presented to the trader for solution actually *are* linear programs. Second, and more importantly, understanding how an optimizer gets to a solution operationally

[3]If such a thing is possible.

within that framework is very important in guiding our thoughts about the roles of various terms introduced into the objective functions traders must optimize to extract their best performing trading strategies. Many readers will have mostly encountered "optimization" within the Newtonian world of gradients and second derivatives. The problem with that is the required conditions that the function under optimization be *smooth* is a rather powerful, and global, statement about the mathematical object considered. These powerful features allow Newton's method to deliver "magical" solutions in its ability to exactly locate an extremum and step to it in a single iteration, regardless of where in the solution space the algorithm is started from. The process of designing trading strategy algorithms for the P.D.T. Group at Morgan Stanley in the 1990s took me through the steps outlined in *The Hierarchy of Objective Functions*[4] where the impact of each additional term in an objective function is understandable by consideration of how it differs from simpler structures. To some extent, this reflects my physicist's view of mathematical models of nature developed through the lens of *perturbation theory* [51].

Essay 3 is a branch on our journey designed to lead the reader towards the solution of particular problem: optimal portfolio selection with negative exponential utility and leptokurtotic distributions of returns. I present the full solution to the frictionless multivariate portfolio selection problem that I developed just after leaving Morgan Stanley and which features the multivariate generalized error distribution.[5] This is an important treatment because it features a probability distribution that is closer to reality than the multi-normal distribution but *also* may be smoothly deformed into that distribution. Thus, it provides an excellent laboratory for the study of the effect of leptokurtosis[6] on portfolio selection. A naïve interpretation of this result [48] is that such considerations are not necessary, but this is in error. The effect on portfolio selection may be scaled away for a single horizon investor, but for a perpetual trader, this is not so.

[4] Defined in Section 2.5 on page 41.

[5] I have written at length about this distribution in *Adventures in Financial Data Science* [20] and this derivation is also featured there.

[6] "Fat-tailed" returns.

Essay 4 puts on paper the thought process I went through to figure out *how* to solve trading strategies. I will draw an analogy to option pricing, in particular pricing options on a lattice, to put into context the necessary conditions to permit the solution to trading strategies. This is intended to "smooth the bump" that takes us from the classical world of utility theory[7] to the practical world of trading with an alpha.

The full single horizon solution developed in Essay 3 suggests, heuristically, that a much simpler solution in terms of discrete holdings may be reasonably effective. Essay 5 develops these solutions fully, including optimal trading strategies both with and without transaction costs for normal and leptokurtotic distributions of returns. This essay is the "meat" of the work and presents a practical solution that may be used by traders with alphas in the presence of realistic distributions of returns. It represents the way I trade, and I think it is a reasonable strategy for small traders to adopt. The solution relies heavily on the key concepts developed in Essay 2. These algorithms exhibit key features that resonate with the solutions presented in earlier essays. Of particular note are the following:

(i) When transaction costs exist, a trade should only be done when the alpha covers at least the round-trip cost. This leads to strategies when the "barriers," or trigger thresholds, to trading are separated by twice the transaction cost.
(ii) When risk is considered, trades with a small net expected return should be vetoed. The risk taken on is not related to the trade size (for fixed position trading) and so only trades where the expected return exceeds a "risk penalty" should be done. From a signal processing point of view, this can be thought of as a filter that increases the signal-to-noise ratio of the trading.

The results of Essay 5 represent a full solution to an optimal trading strategy with explicitly worked out parameters for use when returns are described by particular distributions. Of course, reality may not agree with a particular model. In this scenario, again I look back to physics and advocate a "semi-empirical" approach, influenced by the idea of the semi-empirical mass formula in nuclear

[7]It is "classical" in the sense that it is what is discussed in finance classes.

physics [11]. By this, I mean take the structure of the formula as given but allow the parameters to be determined empirically by the data and not use the values fixed by theory. This will deliver a solution which contains the appropriate features but is not wedded to a distributional choice that may be wrong.

Essay 6 represents my thoughts on ways to measure and benchmark trading strategies that differ from the canonical choices. That difference is provoked by the flaws in using the Sharpe ratio as a statistic, as identified in Essay 1. This essay features ideas around the use of counterfactuals, or "trading oracles," to grade performance. The basic idea is that, *ex post*, there is an upper limit to the performance that even a perfectly informed trader can achieve when subject to risk limits and transaction costs and judged according to some benchmark. In forecasting, a presciently informed being who "magically" knows what the future will hold is referred to as an "oracle," and so this essay introduces and explores the concept of "optimal trading oracles." The second part of the essay applies that concept to performance management in an active management fund when asked to evaluate a trading strategy and shows how the process could work. This piece of the work was stimulated by conversations I had with senior management at a multi-billion dollar hedge fund (that is extremely well known within the industry) when challenged to develop a performance management framework that differed from the usual practices.

Finally, I think it's worth considering how the ideas of Essay 6 can be taken into the space of machine learning to create training sets from which deep neural networks can be trained to replicate the holding functions that are derived analytically in Essay 5, and potentially more sophisticated and realistic ones. The idea pursued here is radical in that, if successful, it would make all of the prior work unnecessary. If, as it is suggested in Essay 6, it is possible to compute *ex post* optimal, or counterfactually perfect, sequences of positions for a suitably risk-managed trader to hold, then perhaps it would also be possible to use those sequences as the training set for a deep neural network that will produce as its output a representation of the holding function that, when presented with the appropriate alpha and constraints, is the solution to the optimization problems discussed in the prior essays. This departs from common practice

regarding "A.I.[8] in finance," which seems to be devoted to an exercise which I would characterize as "throw a bunch of time series into a black box and hope a high Sharpe ratio emerges."

Some Essential Concepts

Notations: I have tried to be as consistent as possible with my use of notation and also to avoid departures from "standard notations" that would come across as confusing to the reader. Instead, I leave the reader with the alternate confusion that some notations are used differently depending on the context. I hope the reader can persevere through this, as I believe it actually *aids* exposition. I will (facetiously) rely on Gödel's theorem to justify this usage which is not consistent because it is complete! In Essay 2, I define many of the terms and assumptions I will be working with, but a few priors are needed for Essay 1. This may seem unnecessary, but I'm assuming that some of the readers of this book have, unlike me, not spent the last 25 years thinking about these concepts. This is not a "beginner's guide," and I will assume that all readers have exerted intellectual energy thinking about the problem of trading efficiently, but certain topics need to be carefully defined so that we are all "on-the-same-page," as one might say. This discussion has to begin with the labelling of time.

In-sample, out-of-sample, testing and training, etc.: I will use the terms *in-sample* and *out-of-sample* extensively. In addition, *training set*, *testing set*, *ex ante*, and *ex post* deserve consideration.

***In-sample*:** This generally refers to data in which a model is developed. It is generally viewed as "suspect" as a venue for hypothesis testing because the model development process often takes many bites at the "reduce the errors" cherry and so, almost inevitably, is overfit to some extent. This means that it is predicting individual data, rather than describing more general and "generalizable" *out-of-sample* relationships between data, and that will not serve it well

[8] Whatever that means.

during live trading. In the machine learning communities, the phrase *training set* is often used to mean the same thing, but *in-sample* is pretty standard throughout the finance and economics communities.

***Out-of-sample*:** This refers to data in which a model *has not* been developed and so is viewed as not subject to the biases that result from that process. In a time-series context, this is generally data that are *later*, or *more recent*, than the *training set*. It is called the *testing set* by the machine learning community.

***Ex ante and ex post*:** Swedish economist Gunnar Mydal [3] introduced the terms *ex ante* and *ex post* to economics to refer to decisions made, or quantities known, both "before the fact" and "after the fact," which is an essential distinction in the analysis of both data-generating processes and decision processes that consume and emit time series. Our expected return for an asset is *ex ante* information and the actual, experienced return is *ex post*. Thus, *ex ante* generally refers to time-conditioned expectations whereas *ex post* refers to outcomes that are observed later in sequence.

The reality of the distinction can be made clear from the simple thought experiment of considering how to implement a trading system in production. For example, if model building includes a "demeaning" or "standardization" phase, in which the mean of the entire observed time series is removed from the data to stabilize drift, it should be clear that this procedure uses future information (data point $x_t \to x_t - \bar{x}$) and simply cannot be done in a production system *which does not have access to those future values*. The best that can be done is to remove the mean of the data observed *to date* ($x_t \to x_t - \bar{x}_{t-1}$) or an *ex ante* estimate of what that value will be ($x_t \to x_t - \hat{\mu}_{t-1}$).

Set theory notations: I will be using common set theory notations for collections of quantities of interest or relevance to the problem at hand. I will not be presenting a fully rigorous analysis in the mathematical sense, complete with theorems and proofs. However, I will be using set theory notations for common objects as they are quick and expressive, and I have become used to them. These include the real line, $\mathbb{R}$, the integers, $\mathbb{Z}$, and strictly positive subsets of these sets, denoted as $\mathbb{R}^+$ and $\mathbb{Z}^+$, respectively. In multiple dimensions, an index is added, such as $\mathbb{R}^3$. $x \in \mathcal{X}$ means "x is a member of the

set $\mathcal{X}$" and $\emptyset$ means the empty set.[9] An enumerated set of items is written as $\{a, b, c, \dots\}$ and a range from the integers or reals as $[a, b]$. For the reals, parentheses are used instead of square brackets when the range excludes the limit.[10] Set unions are written as $\mathcal{A} \cup \mathcal{B}$ and set differences may be written as $\mathcal{A} \setminus \mathcal{B}$. Set intersections are $\mathcal{A} \cap \mathcal{B}$. The "size" of a set is written as $|\mathcal{X}|$ which, for countable sets, means the number of members and for uncountable sets, such as $\mathbb{R}$, means their cardinality, which is the extended concept of "size." The symbol $\exists$ means "there exists" and $\forall$ means "for all." A colon is used as a shorthand for "such that." In some sections, I have used the notation $\mathcal{A} \mapsto \mathcal{B}$, or $\mathcal{A}$ "maps to" $\mathcal{B}$, to mean that members of set $\mathcal{A}$ may be put in *one-to-one correspondence* with the members of set $\mathcal{B}$. $\mathcal{A} \subset \mathcal{B}$ means that $\mathcal{A}$ is a "subset" of $\mathcal{B}$, and the direction of the relation may obviously be reversed to imply "superset." Although the standard notation for the subset of members, $\{x\}$, of a set, $\mathcal{X}$, selected according to some condition, $c(x)$, is $\{x \in \mathcal{X} : c(x)\}$, for brevity, I will write this as $\{\mathcal{X}\}_{c(x)}$. For example, the set of all prior returns at time t will be written as $\{r_s\}_{s<t}$ as this is more compact than $\{r_s : s < t\}$.

Notation for extrema and optimization: If $\mathcal{X}$ is a set and the members are *orderable*, then those members possess a maximum, designated $\max \mathcal{X}$, which is the largest member that may be enumerated. If we cannot enumerate the members, but they are still orderable, the convention is to describe a "supremum," designated $\sup \mathcal{X}$, as the smallest value that all members are less than. Essentially, this is just the concept "maximum" rigorously defined for sets, such as the reals, $\mathbb{R}$. The converse expressions for minimum, $\min \mathcal{X}$, and "infinimum," $\inf \mathcal{X}$, are used for the other extreme.

If $f(x)$ is a mapping from $\mathcal{X}$ to some other set of values, $\mathcal{Y}$ (written formally as $f : \mathcal{X} \mapsto \mathcal{Y}$), one might ask what value of x delivers $\max \mathcal{Y}$. The convention I use is to write that as

$$\hat{x} = \arg \max_{x} f(x) \tag{1}$$

[9] I will generally use "caligraphic" letters for sets but not strictly.
[10] That is, $5 \notin (5, 6]$.

which means the specific value, $\hat{x}$, that delivers $f(\hat{x}) = \max \mathcal{Y}$. This equation, equation 1 on page xxix, is the definition of the solution to an optimization problem. Clearly, similar notations exist for the other ways in which extrema may be defined.

This notation, of adding a "hat" to a variable, $x \to \hat{x}$ pronounced "x-hat," is the way of indicating "the estimated value" of a parameter in statistics. Since the estimate is usually the "best estimate," according to criteria such as maximum likelihood or least squares, these are consistent usages. In Essay 6, I will introduce a complementary notation

$$\check{x} = \arg \min_{x} f(x) \tag{2}$$

which is not in common usage but, I think, intuitively useful in the context within which I'm using it. In "math-speak," $\check{x}$ would be referred to as "x-check."

Notation for expectations: Trading is about making a decision based on information I have now that I expect to benefit me in the future. In the English sentence just written, "expect" could be taken to imply something more like "hope," but I will be taking the common usage from science to mean the arithmetic mean of a set of potential future values of a variable weighted by the probabilities with which those outcomes could occur as the meaning of the phrase "expected value." In this book, I will be assuming the reader has a working familiarity with probability theory and that these concepts are not "alien." The entire discussion presented will feature a probabilistic viewpoint on decision theory and the assumption that stochastic process models are the "correct" way to describe price processes. A reader who objects to the use of probability in the context of stock markets will likely not find this work to their taste!

It is quite common to introduce the idea of an expectation "operator" to express this quantity, and I will do so with the symbol $\mathbb{E}[x]$ to mean "the expected value of x," to be interpreted as the *unconditional* expected value, and the symbol $\mathbb{E}[x|y]$ to mean "the expected value of x given y" or the *conditional* expectation. This is not the only symbolic representation of the concept, with other authors writing variations on $E[x]$, $E(x)$, $\boldsymbol{E}x$, and Ex to mean the same thing, but I find it useful to draw out this distinction with the "blackboard

bold" typeface.[11] In my usage, it is a right-associative operator but can also be though of as a "function," since I will use square brackets to indicate the quantity we are taking the expectation of.

In addition, we will want to discuss the variance of quantities, and for consistency, I will use $\mathbb{V}[x]$ for the unconditional variance and $\mathbb{V}[x|y]$ for the conditional variance.[12] The conditioning variable may often be "the information known at time t," and I will write $\mathbb{E}_t[x]$ as a shorthand for this conditional expectation and $\mathbb{V}_t[x]$ as a shorthand for this conditional variance. To complete this notation, I will assert that the unconditional expectation is equal to the expectation conditioned on an empty set: $\mathbb{E}[x] = \mathbb{E}[x|\emptyset]$, with the same applying to variances, and the tautology that the expectation of a known quantity is equal to that quantity, $\forall y\ \mathbb{E}[x|y] = X$, when $x = X$.

Of course, we also have

$$\mathbb{V}[x|y] = \mathbb{E}\Big[(x - \mathbb{E}[x|y])^2 \,|y\Big] \tag{3}$$

by definition. The *Law of Iterated Expectations* is the relationship that

$$\mathbb{E}[\mathbb{E}[x|y]\,|z] = \mathbb{E}[x|z] \quad \text{where } z \subseteq y, \tag{4}$$

from which follows both

$$\mathbb{V}[x|y] = \mathbb{E}[x^2|y] - (\mathbb{E}[x|y])^2 \tag{5}$$

and

$$\mathbb{E}_s[\mathbb{E}_t[x]] = \mathbb{E}_s[x] \quad \text{if } s \leq t. \tag{6}$$

The former equation is just an algebraically simpler expression of the variance but the latter is more profound as it allows us to operationally look into the future and describe decisions based on expectations of the value of quantities that are, themselves, also expectations.

[11] I am not the only author I've found using this particular notation. In my prior book, *Adventures in Financial Data Science*, I used the $E[x]$ notation. [18]

[12] This usage is less common, but I find it appealing.

Trading times

Continuous time: Throughout these essays, I will be assuming that trading *only* occurs at a time within an ordered set of discrete times, $\mathcal{T} \subset \mathbb{R}^+$, and not continuously. In the *real world*, one may not trade "continuously," even if the trade opportunities may exist at *any* future time. Any properties labelled with a temporal subscript, such as P_t, must then be considered to be referred to a particular time and the convention is that they can only refer to either the *beginning* or the *end* of the interval, $(s, t]$, established by *consecutive* times $\{s, t\}$ in the ordered sequence of members of $\mathcal{T}$. Two trading times, s and t, are consecutive if $s < t$ and there is no other trading time, u, between them. That is,

$$\nexists u \in \mathcal{T} : s < u < t. \tag{7}$$

Clearly

$$\exists u \in \mathbb{R}^+ : s < u < t, \tag{8}$$

but, by definition, we cannot *trade* at that time. For u defined by equation 8,

$$u \notin \mathcal{T} \tag{9}$$

$$\therefore \mathcal{T} \mapsto \mathbb{Z}. \tag{10}$$

Discrete time: Returning to the concept of "labelling" properties with times, it is the normal convention that X_t, for consecutive times $\{s, t\}$, means the value of "X," whatever that may be, at the *end* of the interval $(s, t]$. X may, in fact, vary throughout that interval, but mostly it is the values $\{X_t : t \in \mathcal{T}\}$ that are of interest to us. I will use the typographically shorter notation $\{X_t\}_{t\in\mathcal{T}}$ to represent this statement.

As the set of trade times, $t \in \mathcal{T}$, is discrete and ordered, we may also introduce a set of sequence numbers to label those times and these sequence numbers, without loss of generality, may be put in one-to-one correspondence with the trade times i.e.

$$\exists i, j \in \mathbb{Z}^+ : t_i, t_j \in \mathcal{T} \quad \text{and} \quad i < j \qquad \Leftrightarrow \qquad t_i < t_j. \tag{11}$$

This is merely a formal way of saying that if the trade times are discrete, then we can label them by *either* their positions in the

sequence of trade times *or* those times themselves. It is generally more convenient to use those sequence numbers rather than the "wall clock times" they represent, but the mapping to *real* time is always present and implicit. When a quantity is indexed x_t in these essays, it should be assumed that the sequential labels are being used unless it is otherwise noted (or clearly obvious). Thus the value of x before x_t is usually written as x_{t-1}. This is *a lot* easier than writing x_{t_i} and $x_{t_{i-1}}$ to refer to these quantities.

Of course, it need not follow that $t_j - t_i = (j - i)\Delta t$, for some positive *constant* Δt, meaning that we do not have to assume that the trade times are equally spaced in calendar time with a fixed cadence. This is not an esoteric circumstance as something as familiar as "daily data" in fact skips over weekends and holidays as if they never happened.

It is a convenient *choice* to define these sequence numbers as being members of the positive integers, $\mathbb{Z}^+$, as that permits the time labelled by sequence number "0" to be used to label "special" initial conditions that are before time "1" but not equivalent to it. This is a notational convenience that cannot be made if the whole set of integers is used.

It takes time to trade: The convention that sequential labels of measured quantities refer to the *end* of the period will be followed except for a few exceptions. The position taken in an asset, or holding h_i, refers to the value taken at the immediate beginning of the interval defined by consecutive times $\{t_{i-1}, t_i\}$ and held constant throughout the interval $(t_{i-1}, t_i]$. The time at which this transaction occurs is taken to be the time

$$s = \inf u : t_{i-1} < u < t_i. \tag{12}$$

That is, the position is taken at time s *immediately* after time t_{i-1} and held constant until time t_i. I apply the same convention to the expected return, with α_i being used for the expected value of a return over the interval $(t_{i-1}, t_i]$.

Thus both are *effectively* step functions which, for the interval $(t_{i-1}, t_i]$, take the values equal to the values at the *end* of the interval *throughout* the interval and the change in the values of these quantities occurs *immediately before* the interval starts. The purpose of this is to reduce the subscripts in expressions, such as, for example,

$p_i = h_i r_i$, which is the profit accruing to an investor who took a position of h_i immediately after time t_{i-1} in an asset that experienced a return r_i from time t_{i-1} to time t_i.

Finally, this section should be the last one in which the times that are being used to refer to "wall clock times" and not "trade time sequence numbers."

Traders and personal expectations: Traders are people[13] who believe they have superior information to "the crowd"[14] about the future prices of assets and who seek to profit from that information by interacting strategically with other asset holders. Mathematically, we say that a trader has access to information, $\mathcal{I}'_s$ at time s, that indicates that the expected future price of an asset, P_t at time t, is not equal to its current price, P_s. We can write this statement as follows:

$$\mathbb{E}[P_t|\mathcal{I}'_s] \neq P_s \quad \forall\, t > s \tag{13}$$

meaning that the trader *expects* the future price to differ from the current price. The action that the trader can take is to buy, or sell, some quantity, h_t, of the asset so that they may profit from the expected price move. To execute such a trade, however, they're going to have to find another trader who disagrees with them about what the future price will be and who currently possesses the asset, or can arrange to borrow it for a short term, so that they are able to sell it to them.

Note the distinction here between *private information*, upon which expectations are conditioned, and *personal probabilities*, as discussed in texts on *Bayesian analysis*, such as Keynes, Jeffreys *et al.* [28]. I'm presenting the expectation computed as the consequence of applying some mechanistic function to the information provided to a trader and not based on a private personal prior distribution. If this trader is given different information, $\mathcal{I}''_s$, they will compute a different expectation,

$$\mathbb{E}[P_t|\mathcal{I}'_s] \neq \mathbb{E}[P_t|\mathcal{I}''_s] \qquad \Longleftrightarrow \qquad \mathcal{I}'_s \neq \mathcal{I}''_s, \tag{14}$$

[13] And machines built by people.
[14] That is, all the *other* traders.

which will entirely "override" any priors conditioned by their personal experiences. In this, I am assuming that all traders are equally rational, and their differences in opinion about expected prices arise from the differences between the information sets they possess. For traders with no information, the only statement that is possible is one about the unconditional future distribution of returns, as $\mathbb{E}[P_t|\emptyset] = \mathbb{E}[P_t]$, whereas a trader with *relevant* information will compute a different expectation.

The trading crowd: Members of the trading crowd need to have different expectations of the future price of the asset for trading to occur between them and trading is a zero sum[15] game between competitive individuals who disagree about the future price of the asset. My counterparty in a trade is motivated by the fact that they believe me to be wrong in my assessment of the future value and they are willing to sell me the asset for more than they think it will be worth in the future or buy it from me for less than they think it will be worth in the future. This kind of competitive dynamic that leads to price formation is well explored in finance literature, notably by Kyle [33]. However, in public markets, no traders may disagree about what the actual prices of assets are. The price at which an asset trades in a market, P_t, is an observable fact and not a measure of subjective judgement.

Alphas are expected returns conditioned on information: In approaching the concept of personal expectations as rooted in private information sets, we avoid asserting that the actual dynamics of the price process may differ from trader to trader. Those actual dynamics are driven by a unique process of the Universe of which various members of the trading crowd may have differing levels of knowledge and understanding. Thus we may assert there is an information set, $\mathcal{I}_s$, for which[16]

$$\alpha_t = \mathbb{E}[r_t|\mathcal{I}_s] \quad \text{where } r_t = \frac{P_t - P_s}{P_s}. \tag{15}$$

[15]Actually not necessarily zero sum for the entire trading crowd as an aggregate entity, as prices in general may increase or decrease, but zero sum within the trading crowd relative to the aggregate changes.

[16]Sometimes it makes sense to work with price *changes* rather than *returns*.

The labelling of α_t is chosen to align with that of r_t for convenience as many expressions will involve both α_t and r_t.

This equation is not subjective and is the definition of "the alpha" as the expected return of the asset over the interval ending at the specified time. Note that, however, the alpha is knowable at the end of the prior interval to traders in possession of $\mathcal{I}_s$ and able to compute α_t from it. Except for specific cases of "inside information," it is reasonable to assume that most market participants possess $\mathcal{I}_s$. However, their ability to compute α_t from it, or profit from it even if they can compute it, may be compromised in some way. It is also possible that some traders may be indifferent to the value of α_t, as they are compensated for trading via some other mechanism.

Alpha building and backtesting

Alpha building: It is reasonable to assume that most traders have sufficient skills to acquire all public information relevant to their jobs. This is why we assume that returns are conditioned on a single information set that all traders have access to. However, not all traders will process that information efficiently. Each trader, i, may create a function $\alpha_i(\mathcal{I}_s)$ which represents their estimate of the conditional mean of future returns. The difference between this function and the *correct* one, $\alpha(\mathcal{I}_s) = \alpha_t$, represents a trader specific, or idiosyncratic, error. The process of "alpha building" involves the construction of a function that minimizes the error $r_t - \alpha_i(\mathcal{I}_s)$ for *future* returns.

The alpha builder's assumption: As future returns are not generally available for alpha building, traders tend to work with historical data to minimize the *ex post* error on observed data. This procedure is called either "alpha building" or "backtesting," depending on the context. A statistical approach might try to optimize the historical variance of the error in an alpha on the hypothesis that the procedure that achieves this will also deliver a function that minimizes the unconditional variance of future errors:

$$\begin{aligned} &\arg\min_{\boldsymbol{\theta}} \mathbb{V}[\{r_u - \alpha_i(\mathcal{J}_{u-1}, \boldsymbol{\theta})\}_{u \le t}] \\ &\quad = \arg\min_{\boldsymbol{\theta}} \mathbb{V}[\{r_u - \alpha_i(\mathcal{J}_{u-1}, \boldsymbol{\theta})\}_{u > t}], \end{aligned} \tag{16}$$

where[17] $\mathcal{J}_s = \mathcal{I}_s \setminus \boldsymbol{\theta}$ and $\mathbb{V}[\cdot]$ is the variance operator. This is the *Alpha Builder's Assumption*, which often corresponds to standard statistical practice.

The backtester's assumption: *Backtesting* is the simulation of the performance of a trading strategy on historical data. The *Backtester's Assumption* is that, for control parameters $\boldsymbol{\theta}$ and backtest time t,

$$\begin{aligned} &\arg\max_{\boldsymbol{\theta}} Z(\{r_u h_u | \alpha_i(\mathcal{J}_{u-1}, \boldsymbol{\theta})\}_{u>t}) \\ &\quad = \arg\max_{\boldsymbol{\theta}} Z(\{r_u h_u | \alpha_i(\mathcal{J}_{u-1}, \boldsymbol{\theta})\}_{u\leq t}). \end{aligned} \tag{17}$$

Here h_t is the holding in the asset for the interval ending at t, $r_t h_t$ are the associated profits, and Z is the score statistic, such as the Sharpe ratio.[18] Equation 17 states that the parameters that will deliver the best performance in the future are those that delivered the best performance in the past. It is frequently taken as true but almost never *proved*.

The optimal trader's assumption: However, the purpose of this collection of essays is to follow a different path. It is to find out what should be done with information that is known precisely at the time a trade decision is made, even if that information is, itself, stochastic. Instead of equation 17, we assume

$$\begin{aligned} &\arg\max_{\boldsymbol{\theta}} Z(\{r_u h_u | \alpha_i(\mathcal{J}_{u-1}, \boldsymbol{\theta})\}_{u>t}) \\ &\quad = \arg\min_{\boldsymbol{\theta}} \mathbb{V}[\{\alpha_t - \alpha_i(\mathcal{J}_u, \boldsymbol{\theta})\}_{u\leq t}]. \end{aligned} \tag{18}$$

That is, the best trading performance arises from using the most accurate alpha. The relationship between h_t and $\alpha_i(\mathcal{J}_{t-1}, \boldsymbol{\theta})$ is taken as predetermined and not established by data mining in backtests.

[17] This is a tautology, $\mathcal{I}_u = (\mathcal{I}_u \setminus \boldsymbol{\theta}) \cup \boldsymbol{\theta}$, and so is written solely for the purposes of exposition.

[18] See Essay 1 for a definition if needed.

Choice of methodology: Many statistical estimators have a precision that varies inversely with the square root of the sample size or something of that order. My heuristic justification for the choice of alpha building plus optimal strategy design versus simple backtesting is based on the idea that optimal trading requires the veto of trade opportunities that are insufficiently "strong," according to some criteria chosen by the trader. This acts to decrease the sample size available for estimation and leads to biased estimators being used when trading. I believe this argument applies as much to the users of deep neural networks as it does to discretionary traders.

Holdings and transaction costs: To realize value from an alpha, a trader must take risk. That risk is expressed by a position in the security priced and the profit from such a trade is $r_t h_t$, where h_t is the holding taken at the beginning of the interval ending at time t and held constant throughout that interval. Ideally, the *trade* occurs at time s defined by equation 12 on page xxxiii, but in the real world, it will occur at some later time. The expected profit on the trade is $\mathbb{E}[r_t h_t | \mathcal{I}_s]$, which equals $\alpha_t h_t$ under certain reasonable assumptions.[19] One reason for the expected profit to depart from this value is transaction costs, of which there are two principal ones of concern.[20]

Slippage: The difference between the price used to compute the alpha, P_s, and the price at which the trade is executed, P'_s, is called *slippage* and represents a cost to trading because, generally, it will be found that

$$\mathbb{E}_s[P'_s - P_s] \propto \alpha_t. \tag{19}$$

This is not a tautology because the trade occurs "immediately after" time s and so, potentially, at a different price.

Market impact: Another aspect of *real markets* is the fact that trading tends to push the market in the direction of the trade done.

[19] Specifically, that the position taken is not sufficient to *change* the return experienced. Such a scenario, which can be called *overtrading*, is common to traders that "blow up."

[20] I have omitted brokerage costs as we now live in a world in which these are much less relevant than they were in the 1990s.

If the trade is $h_t - h_s$, then market impact will be found to be

$$\mathbb{E}_s[P'_s - P_s] \propto h_t - h_s. \tag{20}$$

Market impact is distinct from slippage because market impact is caused by the trade done whereas slippage represents an opportunity cost due to market moves that occur between the trade decision at time s and its execution at time u.

Brokerage and regulatory fees: Often traders are charged brokerage fees and regulatory fees that depend solely on the magnitude of the trade done. In the case of regulatory fees in the stock market, these may be levied on selling trades only or equally on buying and selling trades. In either case, these fees are proportional to $|h_t - h_s|$.

Expected transaction costs: If one trades in the direction of the alpha i.e. if $\operatorname{sgn} h_t = \operatorname{sgn} \alpha_t$, then, to first order, both slippage and market impact result in expected transaction costs that is proportional to the size of the trade done, $|h_t - h_s|$. In the following, total transaction costs of this nature will be the base assumption. That is, when cost κ per unit of trade must be paid, expected *net* profits will be of the form

$$\alpha_t h_t - \kappa |h_t - h_s|. \tag{21}$$

In this expression, the costs are conditionally deterministic since the position taken to be held to time t is determined by the alpha known at prior time s and acquired shortly afterwards at time u.

This page intentionally left blank

Essay 1

Mean–Variance Optimization and the Sharpe Ratio

For practical purposes, the discipline of quantitative investment strategy was essentially initiated by Harry Markowitz's 1952 paper on *Portfolio Selection* [39]. Markowitz solves a bigger part of the problem than what I will discuss here, but the foundation of the work is his.

1.1. Mean–Variance Optimization

The Markowitz problem as it pertains to trading strategy is to obtain the portfolio that maximizes the objective function of equation 2.53 on page 49, stated here *ex nihilo*:

$$\Omega(\boldsymbol{h}_t) = \mathbb{E}_s[\boldsymbol{h}_t^T \boldsymbol{r}_t] - \lambda \mathbb{V}_s[\boldsymbol{h}_t^T \boldsymbol{r}_t] \quad \text{for } s < t. \tag{1.1}$$

That is, the objective is to maximize the expected returns of the entire portfolio we hold in a manner that is also averse to risk, which is expressed as maximizing the expected profit on the portfolio chosen while minimizing the expected variance of that profit. The trade-off between risk and rewards is determined by the *Lagrange multiplier* λ, which represents the trader's personal "price of risk." Without any reference to Finance Theory, this is clearly a sensible way to approach the problem and one that is defensible for a trader to investigate *a priori*.

1.1.1. The Markowitz holding function

As the desired quantities, $\boldsymbol{h}_t$, are the positions to be held through the time period *ending* at time t and chosen at time s, they are deterministic as far as the conditional expectation $\mathbb{E}_s[\cdot]$ is concerned. Furthermore, if we assume that the returns of assets themselves are not affected by our holdings, a key assumption that we are an *insignificant* participant in the market,[1] then it must follow that

$$\mathbb{E}_s[\boldsymbol{h}_t^T \boldsymbol{r}_t] = \boldsymbol{h}_t \mathbb{E}_s[\boldsymbol{r}_t] \tag{1.2}$$

$$\text{and} \quad \mathbb{V}_s[\boldsymbol{h}_t^T \boldsymbol{r}_t] = \boldsymbol{h}_t^T \left\{\mathbb{V}_s[\boldsymbol{r}_t]\right\} \boldsymbol{h}_t. \tag{1.3}$$

This then allows equation 1.1 on the previous page to be written as follows:

$$\Omega(\boldsymbol{h}_t) = \boldsymbol{h}_t^T \boldsymbol{\alpha}_t - \lambda \boldsymbol{h}_t^T V_t \boldsymbol{h}_t, \tag{1.4}$$

where $\boldsymbol{\alpha}_t = \mathbb{E}_s[\boldsymbol{r}_t]$ and $V_t = \mathbb{V}_s[\boldsymbol{r}_t]$. Assuming that all quantities are known precisely,[2] this is solved by taking the first derivative w.r.t. $\boldsymbol{h}_t$ and equating $\boldsymbol{\nabla}\Omega = \boldsymbol{0}$. A small rearrangement then delivers the holding function

$$\boldsymbol{h}(\boldsymbol{\alpha}_t, V_t) = \hat{\boldsymbol{h}}_t = \frac{V_t^{-1}\boldsymbol{\alpha}_t}{2\lambda}, \tag{1.5}$$

$$\text{where } \hat{\boldsymbol{h}}_t = \arg\max_{\boldsymbol{h}_t} \Omega(\boldsymbol{h}_t). \tag{1.6}$$

As this $\Omega(\boldsymbol{h}_t)$ is a simple quadratic function, we know that the discovered optimum, $\hat{\boldsymbol{h}}_t$, is both unique and global. This means that *anybody* in possession of the *Information Set* $\{\boldsymbol{\alpha}_t, V_t\}$ will seek to hold a portfolio that is proportional to this $\hat{\boldsymbol{h}}_t$: the only thing that differs between such traders is the particular value they use for the *price-of-risk*, λ. This idea is what led Sharpe to propose the *Capital Asset Pricing Model*, which is an equilibrium theory.

When investors, who can also be called *single-horizon traders*, think about Markowitz's problem and its solution, the focus is on

[1] See the comments regarding *The London Whale* in Essay 2.

[2] Which is potentially problematic, but we will continue nevertheless.

discovering the specific portfolio, $\hat{\boldsymbol{h}}_t$, that should be held. The purpose of this book, however, is to introduce the concept of thinking about $\boldsymbol{h}(\boldsymbol{\alpha}_t, V_t)$ as a *policy function* that provides the trader with a recipe for the mapping $\boldsymbol{\alpha}_t \mapsto \boldsymbol{h}_t$. This mapping tells a trader how to get from information to a position. Alphas, which is our name for the concept of *expected returns*, come and go, as time passes, but the functional relationship stays the same: the *Markowitz Holding Function* is linear in the alpha and inversely proportional to the variance of returns. An alpha that is twice as large, *ceteris paribus*, should lead to the trader taking a position that is *twice* as large. That position, due to the nature of variance, will be *four times* as risky!

1.1.2. Eigenportfolios

The matrix, V_t, is the covariance matrix for the returns, $\boldsymbol{r}_t$, and as such it is a symmetric positive definite matrix. Such matrices are interesting, and one property is that if the matrix has dimension d, then there are d solutions to the *Eigenvalue* problem:

$$V\boldsymbol{x} = \sigma^2\boldsymbol{x}. \tag{1.7}$$

(Here, I have not used the standard notation for the eigenvalue, λ, to prevent confusion with the term in equation 1.5 on the facing page.) This arises because the solution to equation 1.7 is obtained by solving for the roots of the determinant:

$$|V - \sigma^2 I|. \tag{1.8}$$

From the properties of determinants and the *Fundamental Theorem of Algebra*, this is a polynomial in σ^2 of order d and so has d roots. (I is the identity matrix.)

Because V is symmetric positive definite, all of the roots, $\{\sigma_i^2\}_{i\in[1,d]}$, are non-negative real numbers.[3] For each specific eigenvalue, σ_i^2, there is an associated vector, $\boldsymbol{x}_i$, and when a matrix is made of those vectors and applied as a "similarity transformation"

[3] And so possess a real square root, σ_i.

to the covariance matrix, it becomes diagonalized:

$$V' = EVE^T = \Delta \tag{1.9}$$

(dropping the t suffixes temporarily for clarity). Here, E is the *eigenmatrix* formed by stacking the *eigenvectors* columnwise.[4] E is an orthogonal matrix, meaning $EE^T = E^TE = I$. Δ is a diagonal matrix with the eigenvalues along the diagonal, i.e., with elements $\Delta_{ii} = \sigma_i^2$. Statisticians and data scientists will recognize this procedure as *Principal Components Analysis*, where the variation observed in a set of vectors $\boldsymbol{x}$ is described by a decomposition into a new, orthogonal basis that replaces the observed coordinates with new ones that are independent.[5]

From the properties of the eigenmatrix, we see that $V = E^T\Delta E$ and $V^{-1} = E^T\Delta^{-1}E$. Substituting this into equation 1.5 on page 2 gives

$$\boldsymbol{h}(\boldsymbol{\alpha}) = \frac{1}{2\lambda}E^T\Delta^{-1}E\boldsymbol{\alpha}. \tag{1.10}$$

How do we parse this equation? First, the matrix E transforms the vector of asset returns into a vector $\boldsymbol{\alpha}' = E\boldsymbol{\alpha}$ of returns for the *eigenportfolios* associated with the covariance matrix, V. These portfolio returns, by virtue of the method of their construction, are statistically independent with covariance matrix Δ. It is important to note, however, that this is not a *factor model*: there is exactly the same number of eigenportfolios as the original assets and their variance captures all of the variances of the assets. There is no partition of asset variance into *factor returns* and *idiosyncratic returns*. In addition, each eigenportfolio contains a position in *every single asset*, and there are no assets excluded from them.[6]

[4]Not to be confused with the expectation operator $\mathbb{E}[\cdot]$.

[5]There is a tendency to think of this decomposition as some kind of "significant" insight into the properties of the observed data, but I hope you can see from the text that such a decomposition is *always* possible due to the nature of V.

[6]This arises from minimizing a quadratic objective function, which almost never picks *sparse* solutions. Terrence Tao has a good heuristic that explains why this is so in the context of *Compressed Sensing*.

The inverse of Δ is trivial

$$\begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_1^2 & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \dots & \sigma_d^2 \end{pmatrix}^{-1} = \begin{pmatrix} 1/\sigma_1^2 & 0 & \dots & 0 \\ 0 & 1/\sigma_1^2 & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \dots & 1/\sigma_d^2 \end{pmatrix} \tag{1.11}$$

and so the vector $\Delta^{-1}\boldsymbol{\alpha}_t'$ is the vector with diagonal elements given by α_i'/σ_i^2. The final factor of E^T simply transforms that portfolio back into the original space indexed by asset.

What has been achieved other than exhibiting a familiarity with linear algebra? This becomes more clear when we compute the expected gains from the whole portfolio:

$$\mathbb{E}_s[\boldsymbol{h}_t^T \boldsymbol{r}_t] = \boldsymbol{h}_t^T \boldsymbol{\alpha}_t = \frac{1}{2\lambda} \sum_{i=1}^{d} \left(\frac{\alpha_{it}'}{\sigma_{it}} \right)^2 = \frac{1}{2\lambda} \sum_{i=1}^{d} Z_{it}^2. \tag{1.12}$$

This equation has the sort of satisfying simplicity and clarity that a physicist has been trained to love.[7] It is (proportional to) the sum of the squares of the Z-scores of the eigenportfolios, their expected returns divided by their expected standard deviation, and contains nothing more than that and the risk-pricing factor λ. To the statistician and empirical scientist, this Z-score feels like a very "natural" property to consider. The mean–variance optimal strategy involves investing into eigenportfolios in such a manner that their contribution to the aggregate portfolio return is in line with the "significance" of the alpha when expressed in terms of the expected returns and expected variance of these independent portfolios.

1.2. The Kelly Criterion

The analysis of Section 1.1 largely follows the path outlined by canonical finance theory. Shortly after the publication of his work by Markowitz [39], Kelly published the analysis of optimal betting

[7] Our only regret being that we didn't introduce the eigenvalues by the label $(\sigma_i')^2$, although that would have been ugly and surprising!

strategy in the *Bell Systems Technical Journal* [29]. Kelly analyzed fixed odds betting and showed the following:

(i) The maximum geometric rate of growth of capital is equal to the information content of the signal that represents information about expected outcomes (where information is defined in terms of Shannon's *Information Theory* [53]).
(ii) A punter wishing to achieve this rate of growth should bet a fixed fraction of their capital for every wager, with that fraction equal to

$$f^* = \frac{\text{Expected payoff to wager}}{\text{Maximum payoff to wager}}. \tag{1.13}$$

This ratio is known as the *Kelly Criterion*.

1.2.1. Application to financial markets

Ed Thorp, one of the pioneers of card counting in blackjack [57], a colleague of Claude Shannon at MIT, and an important quantitative hedge fund manager, applied Kelly's reasoning to market investments [58]. Even though I think there's substantial evidence that market returns are not normally distributed [18], the use of *just* the mean and variance as *sufficient statistics* for the distribution of returns is equivalent to that assumption.

The key difference between the traditional development of the Kelly criterion in the context of fixed odds betting and any potential application to financial markets is that the "maximum payoff" in equation 1.13 is potentially unlimited, and so $f^* = 0$ always if computed that way. However, Kelly and Thorp's objective is to maximize the geometric rate of growth of capital. Thorp defines the growth rate in wealth over interval $(s, t]$ as $W_t = g^{t-s}W_s$ and, in continuous time,[8] $g = e^{\mu(t-s)}$ for the logarithmic rate of return μ.

[8]Which makes for nice math but not necessarily empirically accurate descriptions of data [20].

1.2.2. Solving for the Kelly portfolio of a trader

Following Thorp's development, we then seek to maximize

$$\Omega(\boldsymbol{h}_t) = \mathbb{E}_s\big[\ln\big(1 + \boldsymbol{h}_t^T \boldsymbol{r}_t\big)\big]. \tag{1.14}$$

Using our assumption from Section 2.3.10 on page 37 of increasingly weakly forecastable markets and the Taylor series for $\ln(1 + x)$ for small x, to second order, generates the quadratic objective function

$$\Omega(\boldsymbol{h}_t) \simeq \mathbb{E}_s\left[\boldsymbol{h}_t^T \boldsymbol{r}_t - \frac{1}{2}\big(\boldsymbol{h}_t^T \boldsymbol{r}_t\big)^2\right]. \tag{1.15}$$

Due to the assumption of weakly forecastable markets, this itself may be approximated by

$$\Omega(\boldsymbol{h}_t) \approx \boldsymbol{h}_t^T \boldsymbol{\alpha}_t - \frac{1}{2}\boldsymbol{h}_t^T V_t \boldsymbol{h}_t. \tag{1.16}$$

Equation 1.16 is just the mean–variance objective of equation 1.4 on page 2 with $\lambda = \frac{1}{2}$, and we obtain the limit

$$\boldsymbol{h}^*(\boldsymbol{\alpha}_t, V_t) \simeq V_t^{-1} \boldsymbol{\alpha}_t \tag{1.17}$$

for the Kelly holding function, $\boldsymbol{h}^*(\boldsymbol{\alpha}_t, V_t)$.

In Thorp's development of this expression [58], he includes a risk-free asset that generates a rate of return, b_t, in suitable notation. With extra algebra, this results in the modification of equation 1.17 to

$$\boldsymbol{h}^*(\boldsymbol{\alpha}_t, V_t) = V_t^{-1}(\boldsymbol{\alpha}_t - b_t \mathbf{1})\,. \tag{1.18}$$

All that has happened here is the transformation of our alpha from a return to an *excess* return w.r.t. to the risk-free rate. The covariance matrix is not affected by virtue of the lack of variance in the returns of a risk-free asset.

Taking a step back to compare equations 1.5 on page 2 and 1.17, we see that the only difference between the two is an explicit value for λ. Thorp's development shows that the choice $\lambda = 1/2$ delivers the mean–variance optimal solution that *also* maximizes the logarithmic rate of growth of wealth. They are one and the same, to first order, and share the same flaws and values.

1.2.3. Fractional Kelly and Root Kelly

The direct compatibility between the holding functions of equations 1.17 on the preceding page and 1.5 on page 2 is interesting in the context of "fractional Kelly," which is the practice of betting at a reduced proportion of the optimal Kelly fraction. This is done because Kelly betting involves scaling up bet sizes after winning bets and scaling them down after losing bets, and such behavior leads to large drawdowns that are avoided when betting at a constant size. Scaling up after a win makes one more likely to suffer a loss that wipes out the prior win and scaling down after a loss makes one less likely to experience a gain that erases the prior loss. Thus, *chasing growth* as aggressively as Kelly asks for contributes directly to increasing the size and severity of drawdowns.

Kelly is explicitly not a risk-averse strategy: it is a strategy designed to maximize the rate of growth of capital. The correspondence between the two systems indicates that Markowitz's solution can be characterized as "not risk-averse enough." It is common to deal with these drawdowns by systemically betting less than Kelly would ask a punter to do so, i.e., $f^* \to kf^*$ for some $0 < k \leq 1$. My experience leads me to believe that a better substitution would be $f^* \to (f^*)^k$, again for $0 < k \leq 1$. We might describe this as "root Kelly" or "power law Kelly."

The key underlying question that leads to the Kelly criterion is the following question:

> How much of my *total wealth* should I risk on this wager?

Fractional Kelly essentially amounts to the *mental accounting* [56] step of reserving only a proportion of total wealth for gambling, whereas my proposal involves responding less aggressively to the opportunities than Kelly or Markowitz would want us to do. I will explore this more precisely in Essay 3.

1.3. The Sharpe Ratio

1.3.1. Equilibrium theory

Equations 1.5 *et seq.* are built upon the assumption that the expected returns, $\boldsymbol{\alpha}_t$, and covariance matrix, V_t, are known precisely. In the

real world, this means that all investors, presumably, would have knowledge of them and so there is no concept of a "personal" alpha. This is built into the solution of equation 1.5 on page 2 in which the personal part of the solution, λ, is merely a scale factor for the chosen portfolio, $V_t^{-1}\boldsymbol{\alpha}_t$, and controls just how much of this *market portfolio* each investor would wish to hold.

Sharpe *et al.* observed that all investors would be able to discover the market portfolio, that *any* combination of risk and return could be constructed by combining it with a risk-free asset, which is commonly assumed to be short-term Treasury Bills, and so there should be no other combination of assets that delivers a return of any interest to investors. In the language of equation 1.12 on page 5, there exists some eigenportfolio for which the alpha is non-zero and it is the only portfolio for which this is true. This is the origin of the *Capital Asset Pricing Model* which, Nobel Prizes notwithstanding, does not seem to be an empirically valid description of the data [12].

However, the legacy of the work from the perspective of the trader is the continuation of the use of the *Sharpe Ratio* as a metric of portfolio performance. Adopting the C.A.P.M. assumptions within 1.12 gives

$$\mathbb{E}_s[\boldsymbol{M}^T\boldsymbol{r}_t - R] = \frac{1}{2\lambda}Z^2, \tag{1.19}$$

where $\boldsymbol{M}$ is the market portfolio and Z is its Sharpe Ratio or the ratio of its expected return to the standard deviation of those returns. R is the *risk-free rate* which is explicitly included within the equilibrium theory and subtracted from the portfolio returns to express them as *excess returns*.

1.3.2. Annualization of the Sharpe Ratio

In the above, the Sharpe Ratio that emerges is one computed from *expected annual returns*. It is conventional to express the Sharpe Ratio in terms of annualized *excess returns* to the risk-free rate. The theory developed here does not require this step, but it has become the convention used in finance.

Table 1.1. Arithmetic annualization of the Sharpe Ratio.

Statistic	Observed	Annualized
Expectation	α	αP
Variance	σ^2	$\sigma^2 P$
Sharpe Ratio	α/σ	$(\alpha/\sigma)\sqrt{P}$

Note: Data are observed with P periods per annum.

To traders, the risk-free return is seldom relevant, particularly in these times of exceedingly low interest rates,[9] and annual returns are not the most interesting as few traders rebalance their portfolios just once per year. Thus, the Sharpe Ratio, when measured as a *statistic*, is generally computed for higher frequency returns, such as monthly, weekly, or daily, which are then annualized (Table 1.1). The annualization of the variance is uncontroversial, and one multiplies the sample variance by a scale factor equal to the number of periods within a year. However, the annualization of the mean has two paths. The most mathematically simple one is to just multiply it by the number of periods within a year, just like for variance, leading to a scale factor equal to the square root of that number.

These scalings are shown in Table 1.1 for P periods per year. This is the usage that seems most natural to me as an empirical scientist. However, some authors[10] suggest annualizing *the expected return* in a manner that includes compounded growth. This leads to the changes shown in Table 1.2 on the facing page.

This expression is generally unsatisfactory because it mingles the effect of returns, α, with the effect of risk, σ, in the computation of the variance because the variance after each successive investment period is scaled up by the growth factor $(1+\alpha)^2$. It is also exceedingly complicated and to first order,[11] it is also no different to the prior expression.

[9] In 1996, when I began working in finance, it was conventional to assume a 5% annual risk-free rate. In 2021, it's been *decades* since we saw such rates. Most of the time, since the Global Financial Crisis of 2008, zero rates have been effectively delivered.

[10] This was discussed in working papers circulated by the *Managed Funds Association* in 2000, when I was setting up my first fund.

[11] Using a Taylor series expansion in α about 0.

Table 1.2. Geometric annualization of the Sharpe Ratio.

Statistic	Observed	Annualized
Expectation	α	$(1+\alpha)^P - 1$
Variance	σ^2	$\sigma^2 \frac{(\alpha+1)^{2P}-1}{\alpha(\alpha+2)}$
Sharpe Ratio	α/σ	$\frac{(\alpha+1)^P-1}{\sigma}\sqrt{\frac{\alpha(\alpha+2)}{(\alpha+1)^{2P}-1}}$

Note: Data are observed with P periods per annum.

Many uses of such "annualized" Sharpe Ratios also do not perform the quite messy annualization of the variance as written in Table 1.2. Normal practice has become, it seems, to use compounding in the annualization of the mean but *not* in the annualization of the variance. This statistic

$$\frac{(1+\alpha)^P - 1}{\sigma\sqrt{P}} \tag{1.20}$$

unfairly inflates the measured return w.r.t. the measured risk and leads to Sharpe Ratios that are biased upwards. It is incorrect because the variance following a return α is not σ^2 but $(1+\alpha)^2\sigma^2$ if this reinvestment of profits is permitted.

When using the Sharpe Ratio as a statistic, it is my general practice to follow arithmetic annualization with $P = 252$ when measuring daily returns, $P = 52$ when using weekly returns, and $P = 12$ for monthly. The reason for the daily number departing from the canonical 365 is because there are typically that number of *trading days* in a year for U.S. markets and the assumption is made that there are both zero gains and zero variance on non-trading days.

Some authors chose to use $P = 260$ for annualization of daily returns, but this makes no sense to me. Although there are 260 weekdays in a calendar year, we do not trade on every weekday as all exchanges observe public holidays. If we count a public holiday as a date on which the average daily return might be delivered, which is the consequence of stating that the annualized return is 260α and the annualized variance is $260\sigma^2$, then on what basis do we ignore weekends? We know for a fact that we will not trade on January 1st and that it contributes nothing to the annual returns of strategies, so why should we include it in the annualization? I believe this practice is incorrect.

1.4. The Approximate Statistical Properties of the Sharpe Ratio

In empirical science, no measurement of data is treated as certain. Everything has experimental error and all numbers are quoted with an error bound. In physical sciences, the number $a \pm b$ implies that a is the central estimate in a contiguous[12] 68% confidence region of half-width b. In social sciences, it is common to quote like this but refer to a 95% confidence region. In both cases, we are measuring a *statistic*, meaning a function of random data, and the uncertainty, or *standard error*, is due to *sampling variation* in the values actually used in the computation.

As mentioned above, the Sharpe Ratio has become used not as the forward-looking measure that determines investment size within the holding function but as a statistic of historical performance. In doing this, we are making the assumption that the sample Sharpe Ratio measured from historical data is a sufficient statistic[13] for measuring the true Sharpe Ratio of the underlying strategy which, as we have noted, is defined in terms of *forward-looking* data. Using historic sample Sharpe Ratios to judge strategies means that we are assuming that it is an efficient and unbiased estimator of the Sharpe Ratio that will be measured in the future.

As such, its use *without* a quoted sampling error is meaningless. Does a Sharpe Ratio of 2 mean 2 ± 0.5 or 2 ± 4? In the former case, we probably have a good trading strategy and in the latter, we have no idea whether our strategy is good or not! During the time when I was a member of the P.D.T. Group, all strategy development decisions and performance measurements were made on the basis of the daily Sharpe Ratio computed for either backtests or live trading of strategies, and Peter Muller wrote an article advocating this after I had left [41], yet not once in my experience[14] was a quantitative consideration made as to whether those performance measurements were *reliable*. We did have a general *heuristic* sense that a measurement of,

[12] That is, without gaps or *simply connected*.

[13] Meaning it contains all information necessary to describe the parameter it models.

[14] Including in my own work.

say, 0.5, was not as reliable as one of 3.0, but error bars were never discussed.

When I first learned about the use of the Sharpe Ratio as a statistic, I computed its sampling distribution, and this work is reproduced in the following. My analysis dates from 1994, while I was a graduate student, and I circulated a copy in 1997 which is now online [16]. That document actually contains an error which is corrected here! Andrew Lo also independently published a calculation using the same method in 2002, which was more widely circulated [36].

1.4.1. The sampling distribution of the Sharpe Ratio

Consider a set of T observed portfolio returns, $\{r_t\}_{t=1}^{T}$. We compute *estimates* of the true mean, μ, and true variance, $v = \sigma^2$, using standard formulae:

$$\hat{\mu} = \frac{1}{T}\sum_{t=1}^{T} r_t \quad \text{and} \quad \hat{v} = \frac{1}{T-1}\sum_{t=1}^{T}(r_t - \hat{\mu})^2. \tag{1.21}$$

The sample Sharpe Ratio is then computed as

$$\hat{Z} = \hat{\mu}\sqrt{\frac{P}{\hat{v}}}. \tag{1.22}$$

The variance of $\hat{Z}$ may then be approximated via the delta method [45]:

$$\mathbb{V}[\hat{Z}] \simeq \left|\frac{\partial \hat{Z}}{\partial \hat{\mu}}\right|^2 \mathbb{V}[\hat{\mu}] + \left|\frac{\partial \hat{Z}}{\partial \hat{v}}\right|^2 \mathbb{V}[\hat{v}] + 2\left|\frac{\partial \hat{Z}}{\partial \hat{\mu}}\frac{\partial \hat{Z}}{\partial \hat{v}}\right| \mathbb{C}[\hat{\mu}, \hat{v}], \tag{1.23}$$

with the notation $\mathbb{C}[x, y]$ indicating the covariance between x and y. Substituting in the derivatives of equation 1.22 gives

$$\mathbb{V}[\hat{Z}] \simeq P\left(\frac{1}{\hat{v}}\mathbb{V}[\hat{\mu}] + \frac{\hat{\mu}^2}{4\hat{v}^3}\mathbb{V}[\hat{v}] + \frac{\hat{\mu}}{\hat{v}^2}\mathbb{C}[\hat{\mu}, \hat{v}]\right). \tag{1.24}$$

It is known from *The Law of Large Numbers* that $\mathbb{V}[\hat{\mu}] = v/T$ exactly. In Kendall [30], the following formula is derived for the sampling error of the rth sample *central* moment, m_r, with sample size,

n, in terms of the population moments, μ_r:

$$\mathbb{V}[m_r] \simeq \frac{1}{n}\left(\mu_{2r} - \mu_r^2 + r^2\mu_{r-1}^2\mu_2 - 2r\mu_{r+1}\mu_{r-1}\right). \tag{1.25}$$

If the population density is symmetric about the mean, then the odd moments all vanish giving $\mathbb{V}[\hat{v}] = \mathbb{V}[m_2] = (\mu_4 - \mu_2^2)/n$. A similar, but slightly more complicated expression exists for the covariance of moments:

$$\begin{aligned}\mathbb{C}[m_r, m_q] \simeq \frac{1}{n}\,(&\mu_{r+q} - \mu_r\mu_q + rq\mu_{r-1}\mu_{q-1}\mu_2 - r\mu_{r-1}\mu_{q+1} \\ &- q\mu_{r+1}\mu_{q-1}).\end{aligned} \tag{1.26}$$

For a symmetrical population, this gives $\mathbb{C}[m_1, m_q] = 0$ for even q, and so $\mathbb{C}[\hat{\mu}, \hat{v}] = 0$ in equation 1.24 on the previous page. The variance of the sample Sharpe Ratio is then

$$\mathbb{V}[\hat{Z}] \simeq \frac{P}{T}\left\{1 + \frac{\mu^2(\mu_4 - v^2)}{4v^3}\right\}. \tag{1.27}$$

The only undefined value in equation 1.27 is μ_4. If the returns are drawn from Normal(μ, σ^2), then $\mu_4 = 3\sigma^4$ and

$$\mathbb{V}[\hat{Z}] \simeq \frac{P}{T}\left(1 + \frac{1}{2P}Z^2\right). \tag{1.28}$$

With annual returns, then $P \to 1$ and we get Lo's result.[15] In more general terms, the kurtosis[16] of a distribution $\beta_2 = \mu_4/\mu_2^2$ so $\mu_4 - \mu_2^2 = \mu_2^2(\beta_2 - 1)$, and equation 1.27 becomes[17]

$$\mathbb{V}[\hat{Z}] \simeq \frac{P}{T}\left(1 + \frac{\beta_2 - 1}{4P}Z^2\right). \tag{1.29}$$

[15] In my paper, I forgot to cancel a factor of two leading to $\frac{1}{4}$ in the expression rather than the correct value of $\frac{1}{2}$. I have recently been aware of a derivation by Opdyke in 2007 of a formula like mine but with the effect of skewness included which I have neglected [14,46].

[16] I am using Mardia's notation of β_2 for the kurtosis and $\gamma_2 = \beta_2 - 3$ for the excess kurtosis [38].

[17] The kurtosis of the normal distribution is 3, which recovers equation 1.28.

For the *generalized error distribution* $\mathrm{GED}(\mu, \sigma, \kappa)$, which I extensively explore in my book *Adventures in Financial Data Science* [18], the kurtosis is

$$\beta_2 = \frac{\Gamma(5\kappa)\Gamma(\kappa)}{\Gamma(3\kappa)^2}. \tag{1.30}$$

I consistently find $\hat{\kappa}$ in the middle of the range $[0.5, 1.0]$ for returns of financial assets, giving β_2 between 3 and 6, and I would recommend using

$$\mathbb{V}[\hat{Z}] \approx \frac{P}{T}\left(1 + \frac{3}{4}\frac{Z^2}{P}\right) \tag{1.31}$$

in empirical work with the substitution $Z \to \hat{Z}$ permissible in the large sample limit.

1.4.2. The accuracy of Sharpe Ratios computed from annual data

In equation 1.31, the size of the sample used to compute the estimated period mean return and its variance is critical. Consider a trader who comes to us exhibiting 10 years of annual returns for which they have a Sharpe Ratio of 2.0, the value generally considered to be interesting. Substituting these values into the equation, we get a standard error of 0.63 meaning that $\hat{Z} = 2.0 \pm 0.63$ which is a statistically significant result. However, suppose their reported Sharpe Ratio was lower. After all, a trader with a really good system would not want to sell it to you. How does the relative error, $\sigma_{\hat{Z}}/\hat{Z}$, vary with Z? From Figure 1.1 on the following page, it is clear that the Sharpe Ratio is quite imprecisely measured from annual data, even with relatively large values. For values less than 1, it is barely measurable.

1.4.3. The sample size necessary to measure the Sharpe Ratio

Having concluded that we should not be using Sharpe Ratios computed from a decade of annual returns to choose managers, what data do we actually need? It is common in the alternative investment

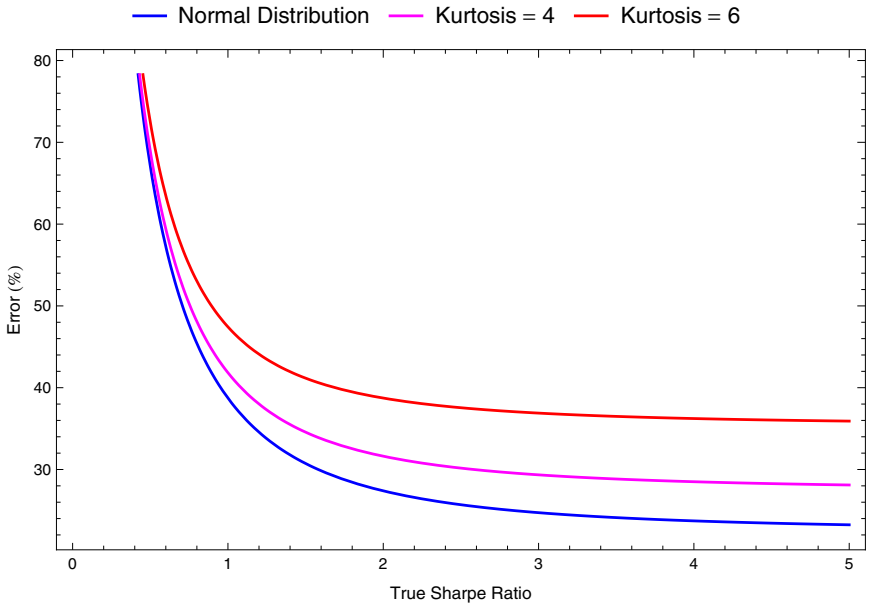


Figure 1.1. The relative error of the sample Sharpe Ratio as a function of the true Sharpe Ratio. The three curves are for the normal distribution (blue), a distribution with kurtosis 4 (magenta), and a distribution with kurtosis 6 (red).

industry to provide 3 years of monthly returns. This means $P = 12$ and $T = 36$ and so the limiting variance is $1/3$ for $Z \to 0$, giving a standard error at $\hat{Z} \approx 0$ of 0.6. In simple terms, with this sample size, any reported Sharpe Ratio less than 1.2 should be regarded as quite suspect.

A logical question is then as follows: How much data do we actually need to reliably measure a Sharpe Ratio of a given size? First, let's define "reliably" as meaning the observed Sharpe Ratio is a "three-sigma" effect, which is when the standard error of the Sharpe Ratio equals one-third of the Sharpe Ratio. When P is larger, the actual effect of the true Sharpe Ratio on the variance is suppressed,[18] and the normal limit is a reasonable approximation *whatever* the actual kurtosis is. Figure 1.2 on the next page shows how many years of data are required to make this determination of strategy quality as

[18]Meaning that higher resolution measurement increases the robustness of the statistic to a failure of the assumption that $\beta_2 = 3$.

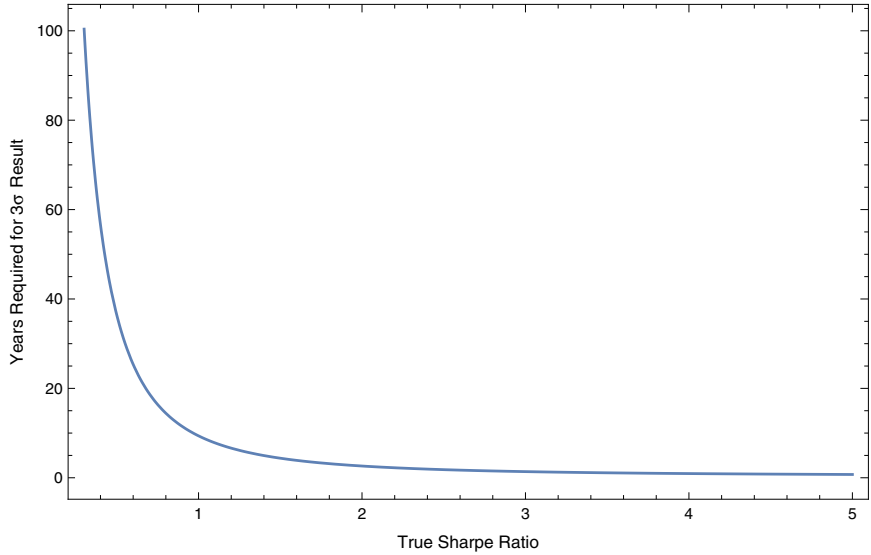


Figure 1.2. The number of years for which monthly data must be sampled to measure a Sharpe Ratio that is a "3σ" result.

a function of the true Sharpe Ratio for normally distributed returns. To call out some specific numbers:

(i) For a true Sharpe Ratio of 3, we need 1.9 years of data.
(ii) For 2, it is 3.2 years of data.
(iii) For 1, it is 9.9 years of data!

1.4.4. The maximum available Sharpe Ratio in public data

The data above represent a striking result. For a Sharpe Ratio of 1, you need almost 10 years of monthly returns just to know that the Sharpe Ratio is 1. Unfortunately, such as system is quite likely to enter a prolonged drawdown during that period and be shut down by its trader or sponsor. Another way to interpret this figure is to suggest that the maximum Sharpe Ratio of any strategy discoverable in monthly returns must be less than that measurable in the 40 odd years that systematic quantitative trading has been a business on Wall Street. That number is just less than 0.5, and a similar calculation is possible for daily data.

1.4.5. The limiting variance of the Sharpe Ratio at higher frequencies

Equation 1.31 on page 15 possesses a useful *Null Hypothesis* limit at higher data frequencies.[19] Concentrating on high-frequency trading permits the approximation

$$\mathbb{V}[\hat{Z}] \simeq \frac{P}{T}, \tag{1.32}$$

in which the dependence of the variance of the Sharpe Ratio on the Sharpe Ratio itself is attenuated in the "large P, small Z" limit. This is analytical support for the advocated use of "daily Sharpe Ratio" by groups, such as P.D.T. and Monroe Trout [42,52]. This expression may be further refined by observing that T/P is the number of years of data used in the sample of trading for which performance is measured via the Sharpe Ratio. Thus, the null hypothesis value of $\mathbb{V}[\hat{Z}]$ is simply given by $1/y$, where $y = T/P$.

1.5. The Sharpe Ratio in Strategy Development and Backtesting

As mentioned at the beginning of Section 1.3, the majority usage of the Sharpe Ratio is as an *ex post* statistic and not as the *ex ante* optimization goal associated with theory. Although this book is about *ex ante* optimal trading strategy development, it is useful to consider how the statistical properties of the Sharpe Ratio *ought* to affect strategy development.

1.5.1. The strategy development workflow

My view of the strategy development workflow is strongly influenced by the time I spent in the P.D.T. Group at Morgan Stanley, but I don't think it is too dissimilar to the processes undertaken by other successful quant trading teams. This strategy development workflow is directed at building systems which maximized the Sharpe Ratio *in-sample* and also produced a sufficiently large Sharpe Ratio when

[19] For active traders, the null hypothesis is always that $Z(\hat{\boldsymbol{\theta}}) = 0$.

tested *out-of-sample*. In both cases, it was computed as the ratio of the sample mean to the sample standard deviation of daily returns and then annualized by multiplying by $\sqrt{252}$. (This is for the average of 252 trading days per year on the calendar of the NYSE.) Development work is divided into four phases:

(i) *Alpha building*: This is the process of constructing a causally legitimate function of past data that was a predictor of future returns when examined in historic data. This was a statistical estimation procedure very similar to that I had gone through in my work in studying the distribution of cosmic rays on the celestial sphere [15].
(ii) *Backtesting*: This is the process of mapping the predictions of future returns created in Step 1.5.1 into simulated trading and scoring the performance of these strategies with the Sharpe Ratio. This often involved several adjustable parameters that were used to tune the trading strategy.
(iii) *Out-of-sample testing*: A "one-shot" test in which the score statistic was evaluated on data not used for either of the first two phases and compared to that from Step 1.5.1.
(iv) *Live trading*: For strategies that passed the out-of-sample test, we would proceed to live trading.

1.5.2. The backtester's assumption

To state this mathematically, let $\boldsymbol{\theta}$ be a set of parameters and $\{\boldsymbol{r}_t, \boldsymbol{h}_t\}_{t\in[1,T]}$ be the associated sets of returns and holdings in the assets traded. The holdings taken are expressed as a function of both the alpha, $\boldsymbol{\alpha}_t$, and the parameters, $\boldsymbol{\theta}$, so the holding function for the strategy under development is defined by $\boldsymbol{h}_t = \boldsymbol{h}(\boldsymbol{\alpha}_t, \boldsymbol{\theta})$.

A set of parameters, $\hat{\boldsymbol{\theta}}$, is chosen to satisfy

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\mathcal{W}} Z(\{\boldsymbol{r}_t, \boldsymbol{h}(\boldsymbol{\alpha}_t, \boldsymbol{\theta})\}_{t\in[1,B]}). \tag{1.33}$$

That is, those that maximize the score statistic, Z (in this case, the Sharpe Ratio), in-sample in the training set defined by $t \in [1, B]$. This choice is made by searching over some sub-region, $\mathcal{W}$, of the entire *feasible space* of $\boldsymbol{\theta}$. The hypersurface, $Z(\boldsymbol{\theta})$, often is neither smooth nor possesses one single maximum, so generally, these searches are

conducted by "grid search" rather than more sophisticated optimization methods.

In the *out-of-sample* testing set, $t \in (B, L)$ for $B < L < T$, the strategy developer computes

$$\hat{Z} = Z(\{\boldsymbol{r}_t, \boldsymbol{h}(\boldsymbol{\alpha}_t, \hat{\boldsymbol{\theta}})\}_{t\in(B,L)}) \tag{1.34}$$

and compares that value to a critical value, Z^*. If sufficiently large,[20] the developer proceeds to live trading, at times $t \in [L, T]$.

The mathematical condition this development workflow is contingent on is what I call the backtester's assumption. In the notation of the current problem, this takes the following form:

$$\arg\max_{\boldsymbol{\theta}\in\mathcal{W}} Z(\{\boldsymbol{r}_t, \boldsymbol{h}(\boldsymbol{\alpha}_t, \boldsymbol{\theta})\}_{t\in[1,B]}) = \arg\sup_{\boldsymbol{\theta}} Z(\{\boldsymbol{r}_t, \boldsymbol{h}(\boldsymbol{\alpha}_t, \boldsymbol{\theta})\}_{t\in[L,T]}). \tag{1.35}$$

The backtester's assumption is that the parameters that give the best performance in-sample are equal to those that will give the best performance in live trading. Note that the actual value of $\hat{Z}$ is not relevant, we just want the maxima of the two optimization phases to coincide, and we seek $\hat{Z}$ that are "sufficiently large" because, to be blunt, trading with a low Sharpe Ratio sucks.

1.5.3. The impact of parameter uncertainty on parameter choice

There are two things to consider when taking on the optimization task represented by equation 0.17 on page xxxv:

(i) the impact of parameter variation on the score statistic,
(ii) the precision with which we know that statistic.

1.5.3.1. *Parameter leverage*

The first item here determines how much *leverage* our parameters have on the trading strategy's performance, as judged through the lens of the chosen optimization statistic, the Sharpe Ratio. During the backtesting phase, as the analyst explores a subset of the space

[20] In my experience, Z^* is somewhat *fuzzily* defined, but it is reasonable to expect it to be $O(\hat{Z})$ and also bigger than, say, 2.

spanned by $\boldsymbol{\theta}$, the *measured* score statistic $Z(\boldsymbol{\theta})$ will change. Unlike least-squares or maximum likelihood estimation, however, in general, it may not be the case that this surface possesses a global supremum. Nevertheless, in general, it must be true that there exists a maximum and a minimum within the space we have explored.[21] The observed dynamic range of the score statistic may be written as follows:

$$\Delta Z = \max_{\boldsymbol{\theta}\in\mathcal{W}} Z(\{\boldsymbol{r}_t, \boldsymbol{h}(\boldsymbol{\alpha}_t, \boldsymbol{\theta})\}_{t\in[1,B]}) - \min_{\boldsymbol{\theta}\in\mathcal{W}} Z(\{\boldsymbol{r}_t, \boldsymbol{h}(\boldsymbol{\alpha}_t, \boldsymbol{\theta})\}_{t\in[1,B]}). \tag{1.36}$$

This quantity reveals the extent to which the trader is able to improve their performance in-sample by varying the parameters we control. If $\Delta Z \approx 0$, then the parameters have no leverage over the performance of the strategy.

1.5.3.2. *Objective reliability*

The second item should be evaluated at the discovered maximum and determines whether the leverage possessed over the strategy is meaningful.[22] Let $\sigma_Z(\hat{\boldsymbol{\theta}})$ represent the standard deviation of the sampling distribution of the statistic, $Z(\boldsymbol{\theta})$, at the discovered maximum, $\hat{Z} = Z(\hat{\boldsymbol{\theta}})$. As the standard error of the statistic at the discovered maximum, this represents the scale of the expected variation in magnitude of the discovered maximum statistic based on random chance alone and so is an expression of how *reliable* the discovered improvement to the strategy is, where "reliable" is interpreted in the sense of the backtester's assumption.

1.5.3.3. *Optimization quality*

It should be clear that the ratio of parameter leverage to objective reliability or

$$Q = \frac{\Delta Z}{\sigma_Z(\hat{\boldsymbol{\theta}})} \tag{1.37}$$

[21] We know this from the extreme value theorem if Z is a continuous function of the parameters defined over a subset of the real volume $\mathbb{R}^k$, for k parameters. For a finite discrete parameter space, the values clearly may be enumerated and that list must contain both a maximum and a minimum.

[22] In my experience, this step is often completely omitted.

is a very important figure of merit for the backtester to know. As the ratio of the maximum strategy improvement attainable by studying the parameter subset to the scale of sampling variation of the statistic at the maximum, it determines whether the discovered variation of performance due to parameter choice is likely to be significant[23] or not.

1.5.3.4. *Optimization quality in theory and practice*

In taking this approach, I am influenced both by ideas from physics, such as the Rayleigh criterion for resolving stars [7], and from statistics, such as Wilks' theorem in maximum likelihood estimation and the Wald test [31]. My training as a scientist and my experience as a strategy developer tell me that unless $Q \gg 1$, it is likely *not true* that parameter choice is having any *useful* effect on the performance of the strategy optimized, and any performance improvement discovered by wandering around in $\boldsymbol{\theta}$ space is likely illusory. If this is the case, the chances of the backtester's assumption being true for the analysis executed are low, and that will lead to losing money in live trading!

Taking the null hypothesis value of $\sigma_Z = 1/\sqrt{y}$ from equation 1.32 on page 18, and an *a priori* quality requirement defined by the analyst, we can immediately calculate the optimization quality metric for a given data set. This tells us that any improvements made to the Sharpe Ratio, made by adjusting hyper-parameters in a backtest, must exceed the minimum permissible improvement of

$$\Delta Z = \frac{Q}{\sqrt{y}} \tag{1.38}$$

to be considered valid.

1.5.4. An example usage from a real trading strategy

Figure 1.4 on page 24 shows the computed Sharpe Ratio, one of the strategies I currently trade for my personal account. This is to invest on the basis of an observed empirical relationship between consumer

[23] Meaning representing a *real* improvement *out-of-sample*.

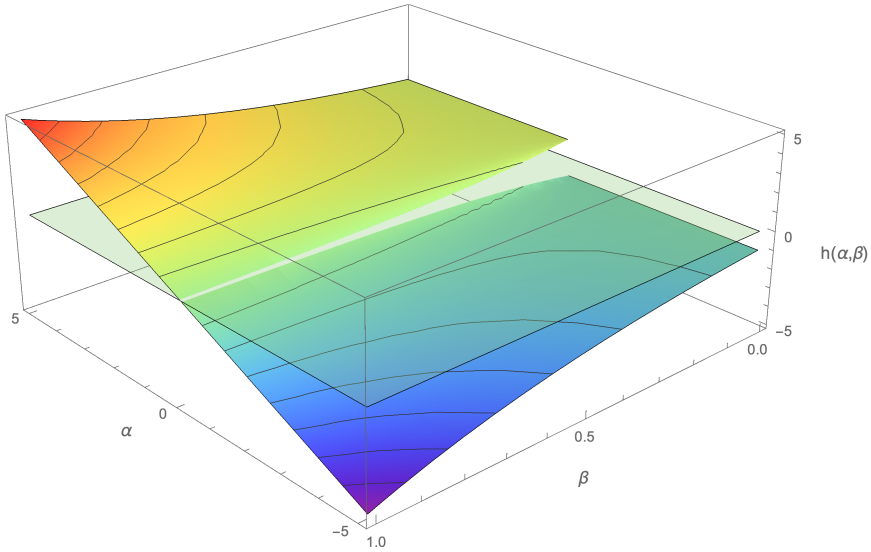


Figure 1.3. The holding function of equation 1.39 as a function of the alpha and spectral index. It transitions from a step function to a linear response as the spectral index parameter, β, is varied from 0 to 1.

expectations of inflation and the monthly returns of the S&P 500 Index from 01/2002 to 11/2002. I sought to examine the "spectral index" parameter β, affecting position size in the nonlinear rule

$$h_t = h(\alpha_t, \beta) \propto |\alpha_t|^{\beta} \operatorname{sgn} \alpha_t. \tag{1.39}$$

This function is illustrated in Figure 1.3. A mean–variance optimal/Kelly criterion approach would suggest $\beta = 1$, whereas my own cube root Kelly approximation[24] would give $\beta \approx 1/3$. With the availability of a backtest, however, we may give the data a voice and study how the Sharpe Ratio varies with this parameter.

This is done by running many backtests with different values of β. Even though the "natural" range of values for this parameter is $\beta \in [0, 1]$, my scan was expanded to the range $[-1, 2]$ "just in case" my view as to what is natural is not, in fact, correct. As this particular strategy trades monthly, I have simplified the analysis by neglecting the (trivial) transaction costs. The results of the optimization

[24]See Essay 3.

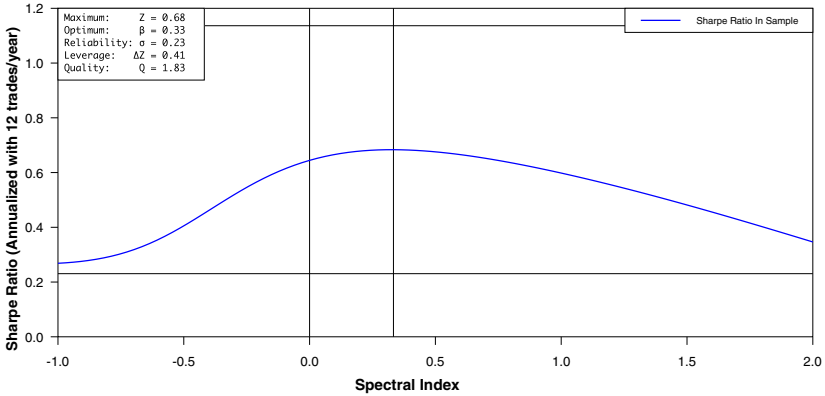


Figure 1.4. Illustration of the examination of parameter leverage and optimization quality for a real trading strategy examined *in-sample*. The blue curve shows the variation of the Sharpe Ratio with spectral index parameter β.
Source: Data Prepared for and © Giller Investments (New Jersey), LLC, 2022. Not investment advice.

scan are shown in Figure 1.4. With the assumption of the so-called "frictionless" trading, I find $\hat{\beta} = 0.33$, $\Delta Z = 0.41$, $\sigma_Z = 0.23$, and $Q = 1.83$. The Sharpe Ratio at the maximum is $\hat{Z} = 0.68 \pm 0.23$, which would be statistically significant were it not for the fact that the strategy was developed from regressions done on the same data set, so it is not a valid *out-of-sample* statistic.

Although potentially statistically significant, in terms of optimization, this is a low quality maximum. It does actually appear close to the expected value under the cube root Kelly approximation. I am confirmed in the use of $\beta = 1/3$, but it really doesn't matter if I did or didn't adopt that choice on the basis of this parameter scan. With a Sharpe Ratio of 0.68, this is not a strategy that I would be comfortable betting my entire wealth upon, but it is certainly a valid part of a larger portfolio of strategies.

1.6. Conclusion

This essay begins with a tour through portfolio selection and performance measurement that is "traditional" in method, although the focus remains on the results that are relevant to the active trader.

It then asks the reader to look again at the practices and assumptions they may have made in work previously executed if they are employed in an quantitative trading role. First, it is demonstrated that the sampling distribution of the Sharpe Ratio makes it a poor tool for performance measurement and then, that the impact of that lack of precision on strategy optimization, as commonly understood in the context of "tuning" a backtest, is reviewed. I think this factor, the unreliability of the Sharpe Ratio as a *statistic*, both in general and as the target of optimization, goes a long way to explain the problems that "backtested" strategies present to the developer of active trading strategies. As in all things, though, the critical question is as follows: "What are the alternatives?" I will present my view of a fruitful and different way to optimize trading strategies in Essay 6 at the end of this book.

This page intentionally left blank

Essay 2

Analytical Framework

2.1. Peter Muller's Rule

Much of how I think about trading is influenced by the P.D.T. framework and what Peter Muller taught me [43]. The way we worked was divided into two parts:

(i) Build a scientifically valid method for *forecasting* future prices of financial assets. This is called an *alpha* or *signal*.
(ii) Use the information generated by these forecasts efficiently in an environment where the information content of your alpha is low and future returns are random.

This seems very natural and obvious to me now, as perhaps it does to you. But the most "systematic trading" I had encountered to this point was about conjuring trades out of arcane recipes and evaluating them in backtests done in worksheets or similar software. There is a legitimate scientific reason for this separation, as many trading systems maximize their performance metric (usually the Sharpe ratio) by not executing a trade every single time a signal is computed. In estimation theory, there are two drivers to the precision with which we can know things: their intrinsic information content *and* the number of measurements we make. By separating item (i) from item (ii), we do not let the potential sparsity of trades interfere with the accuracy with which we can evaluate our forecasting system. I have subsequently referred to this as "Peter Muller's Rule,"

in appreciation of the person I learned it from.[1] My book, *Adventures in Financial Data Science* [18], is mostly about item (i) on the preceding page, and this book is mostly about item (ii) on the previous page. There is a little overlap, but this volume can stand alone from that one.

2.2. The Holding Function

Sometimes I will talk about trading a single asset and sometimes many. In general, the trader seeks a *policy function* that tells them how to act given information about future distributions of returns. To do this, I divide time into a set of discrete intervals with labels t. These labels are positive integers that represent trade opportunities. I assume that we are unable to trade *between* these times. Each trading interval has a duration, Δ_t, which may or may not be constant and which may *elide* over holidays, nights, and other disruptions to trading. Information may be generated through continuous times $(t-1, t]$, and the sum total of the information known by the trader at time t can be written as $\mathcal{I}_t$. Sometimes I refer to *fundamental knowledge* not developed during the trading times as the information set $\mathcal{I}_0$. In general, the value of any property with a time label, P_t, is assumed to refer to the state of the Universe *at the end of the interval* labelled by t unless otherwise noted. Common exceptions are the opening, high, and low prices of the interval, which are defined at sometime after the end of interval $t-1$ and before the end of interval t.

If $\boldsymbol{h}_t$ represents holdings in risk assets *throughout* of period t and $\boldsymbol{\alpha}_t$ expected future profits for those assets for the same time period, then the expected gross profit is

$$\mathbb{E}_{t-1}[\boldsymbol{h}_t^T \boldsymbol{r}_t | \boldsymbol{\alpha}_t] = \boldsymbol{h}_t^T \boldsymbol{\alpha}_t. \tag{2.1}$$

[1] I should note, however, that when I used that terminology at a conference sponsored by Deutsche Bank, the C.O.O. of P.D.T. Partners, my ex-colleague, Amy Wong, called up the next day and demanded that we strike such wording from my talk. Paranoia has always been a key feature of that team, which is one of the reasons why I left.

Here, the notation $\mathbb{E}_t[x|y] = \mathbb{E}[x|y,t]$ means the conditional expectation of x given y at time t. $\boldsymbol{r}_t$ is the returns, or changes, in the asset prices — the choice of which generally depends on the asset set in use. The actual realized profit for the same interval is $\boldsymbol{h}_t^T \boldsymbol{r}_t$. The covariance matrix of asset returns is

$$V_t = \mathbb{E}_{t-1}[(\boldsymbol{r}_t - \boldsymbol{\alpha}_t)(\boldsymbol{r}_t - \boldsymbol{\alpha}_t)^T] \tag{2.2}$$

and the variance of portfolio returns is $\boldsymbol{h}_t^T V_t \boldsymbol{h}_t$.

For many systems, trading creates costs that depend on the magnitude of the trade and they are not random. Often we will encounter transaction costs which take the following form:

$$\kappa ||\boldsymbol{h}_t - \boldsymbol{h}_{t-1}||_1. \tag{2.3}$$

These costs are experienced *at the beginning* of period t and decrease the profits for that period. Their computation involves the "ℓ_1 *pseudo*-norm" of the vector of holding changes, which is defined by

$$||\boldsymbol{x}||_1 = \sum_i |x_i|. \tag{2.4}$$

Often, transaction costs will not be the same for all traded assets and so we must use a cost vector, $\boldsymbol{\kappa}$, rather than the simple constant κ. I will invent an "ℓ_1 *partial*-norm" given by

$$||\boldsymbol{x}||_{\mathbf{1}} = \begin{pmatrix} |x_1| \\ |x_2| \\ \vdots \\ |x_n| \end{pmatrix}, \quad \text{where } \boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}. \tag{2.5}$$

I call this a "partial" norm because it is ready to be "completed" by taking the inner product with another vector e.g. $||\boldsymbol{x}||_1 = \mathbf{1}^T ||\boldsymbol{x}||_{\mathbf{1}}$. The total transaction cost due to trading is then

$$\boldsymbol{\kappa}^T ||\boldsymbol{h}_t - \boldsymbol{h}_{t-1}||_{\mathbf{1}}. \tag{2.6}$$

I will most generally use some statistic, $Z(\{\boldsymbol{h}_t, \boldsymbol{r}_t\}_{t \in \mathcal{T}})$ computed out of the building blocks outlined above, to evaluate the performance of a strategy and our goal is to maximize that statistic. Typically the statistic will be a measure of performance, such as total

net profit, or risk adjusted performance, such as the Sharpe ratio.[2] The goal of trading strategy analysis[3] is to discover a policy function, the holding function $\boldsymbol{h}(\boldsymbol{\alpha}_t, V_t, \boldsymbol{h}_{t-1} \dots)$, that tells us what position to take in response to the information we possess about future returns given the portfolio we currently hold. We will try to chose this function to maximize the expected future value of the performance statistic chosen.

$$\hat{\boldsymbol{h}}(\boldsymbol{\alpha}_t, V_t, \boldsymbol{h}_{t-1} \dots) = \underset{\{\boldsymbol{h}(\boldsymbol{\alpha}_t, V_t, \boldsymbol{h}_{t-1} \dots)\}}{\arg\max} \mathbb{E}_{t-1}[Z(\{\boldsymbol{h}_t, \boldsymbol{r}_t\}_{t \in \mathcal{T}})] \,. \tag{2.7}$$

This optimization is over the *space* of potential holding functions (meaning all of those that can be conceived of), and the one chosen generates a trading program

$$\hat{\boldsymbol{h}}(\boldsymbol{\alpha}_t, V_t, \boldsymbol{h}_{t-1} \dots) - \boldsymbol{h}_{t-1} \tag{2.8}$$

that is executed in the markets to deliver new positions

$$\boldsymbol{h}_t = \hat{\boldsymbol{h}}(\boldsymbol{\alpha}_t, V_t, \boldsymbol{h}_{t-1} \dots) \tag{2.9}$$

ready to experience next period of returns. The holding function tells us what assets to hold and, given our prior holding, what trades to make to get to that desired portfolio.

2.3. Information Sets and Alphas

When discussing trading strategy, I will mention the concept of a trader's "information set" frequently. This is intended to represent everything that is known about the assets under study, or the state of the Universe more generally, at a particular time. Specifically, I will use the notation $\mathcal{I}_t$ to represent the information known at time t. Although this can be treated as a somewhat fuzzy notion, it can also be specifically crystallized into actual knowledge of the state of the variables relevant to a trader's decision.

The expected return, or alpha, is a function of the information set known at the start of the prior interval, $\boldsymbol{\alpha}_t = \boldsymbol{\alpha}(\mathcal{I}_{t-1})$, and the actual returns experienced are a metric of the information set for

[2] Defined in Essay 1.

[3] At least as discussed in this book.

the current time, $\boldsymbol{m}(\mathcal{I}_t)$. The term "information set" is fairly vague, and the notation $\mathcal{I}_t$ merely a label, but the functions $\boldsymbol{\alpha}(\mathcal{I}_t)$ and $\boldsymbol{m}(\mathcal{I}_t)$ dispatch such ambiguity and convert them into mathematical certainty. The metric of the information set that is important to a trader is a concrete and observable thing, usually the returns or changes in asset prices, so $\boldsymbol{m}(\mathcal{I}_t) = \boldsymbol{r}_t$, for example. Furthermore, the "alpha," $\boldsymbol{\alpha}(\mathcal{I}_{t-1})$, is any function of observed data that has been constructed by the trader to create the best possible forecast of future returns.

To make this concrete, consider a trader that has built a momentum model, in which future returns are similar to past returns, then the constructed function might be the conditional mean of an autoregression e.g.

$$\boldsymbol{\alpha}(\mathcal{I}_{t-1}) = \boldsymbol{\mu} + \Phi \boldsymbol{r}_{t-1} \tag{2.10}$$

$$\text{if } \boldsymbol{r}_t = \boldsymbol{\mu} + \Phi \boldsymbol{r}_{t-1} + \boldsymbol{\varepsilon}_t. \tag{2.11}$$

Obviously this simplistic model is only one of the potential forms of alpha function that may be constructed. In this model, the information set is cleanly partitioned into "fundamental constants," meaning $\{\boldsymbol{\mu}, \Phi\}$, relevant but transient information, $\{\boldsymbol{r}_{t-1}\}$, and other, irrelevant information. Thus

$$\mathcal{I}_{t-1} = \{\boldsymbol{\mu}, \Phi\} \cup \{\boldsymbol{r}_{t-1}\} \cup \dots \tag{2.12}$$

2.3.1. The stochastic nature of information

The first law of information for traders: Information about the Universe cannot be destroyed[4] and so information sets themselves are ordered and additive

$$\mathcal{I}_t \supseteq \mathcal{I}_s \quad \text{for } t > s. \tag{2.13}$$

This is the "first law of information for traders." The set difference, $\Delta\mathcal{I}_t = \mathcal{I}_t \setminus \mathcal{I}_s$, is the incremental information acquired during time period $(s, t]$.

[4]Although it may be *forgotten* by some market participants.

If we are willing to make a stronger statement about the accumulation of information, that

$$\mathcal{I}_t \supset \mathcal{I}_s \quad \forall\, t, s \in \mathcal{T} : t > s, \tag{2.14}$$

then we can put the sequence of information set sizes, $\{|\mathcal{I}_t|\}_{t\in\mathcal{T}}$, and their associated times, $\{t\}$, in one-to-one correspondence,[5] $|\mathcal{I}_t| \mapsto t$. Under these circumstances, which rejects the idea that trading should ever occur without the accumulation of new information, we may dispense with a meaning for the labels t other than as a mnemonic for a location within the ordered sequence of distinct sets $\{\mathcal{I}_t\}_{t\in\mathcal{T}}$ and, in a sense, this location is the entire definition of the concept of time that we need.

The second law of information for traders: In the most general terms, $\mathcal{I}_t$ can contain both *deterministic* information, $\mathcal{D}_t$, and *stochastic* (or random) information, $\mathcal{S}_t$, and these sets are disjoint

$$\mathcal{I}_t = \mathcal{D}_t \cup \mathcal{S}_t \quad \text{where } \mathcal{D}_t \cap \mathcal{S}_t = \emptyset. \tag{2.15}$$

By definition $\mathcal{D}_s$ contains all deterministic information known about the Universe at time s. Since deterministic information cannot be forgotten and does not change, it must be true that

$$\mathcal{D}_t = \mathcal{D}_s \quad \text{for } t \geq s. \tag{2.16}$$

This is "the second law of information for traders."

The third law of information for traders: From equations 2.15 and 2.16, it follows that the intertemporal change in information over the interval $(s, t]$ *must be entirely stochastic.*

$$\mathcal{I}_t \setminus \mathcal{I}_s = \mathcal{S}_t \setminus \mathcal{S}_s \quad \text{for } t > s. \tag{2.17}$$

If we define time over a *semi-infinite* domain, such as $\mathbb{R}^+$ or $\mathbb{Z}^+$, then there is a "beginning of time," and all deterministic information that exists must exist at that time. I will label this $\mathcal{I}_0$ for clarity. Apart from this initial information, which represents the "fundamental constants" of the Universe we are studying, the information at any given time is accumulation of all prior changes in stochastic information.

[5] Called a *bijection*.

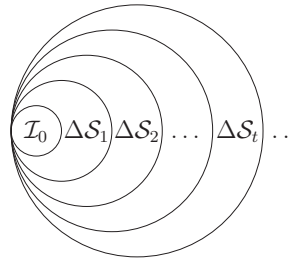


Figure 2.1. Venn diagram illustrating the hierarchy of information sets in discrete time, as used by traders in their strategy design. $\mathcal{I}_0$ are fundamental constants and all incremental information sets, $\Delta\mathcal{S}_t$, are entirely stochastic in content.

In discrete time, $t \in \mathbb{Z}^+$, we may write

$$\mathcal{I}_t \setminus \mathcal{I}_0 = \bigcup_{s=1}^{t} \Delta\mathcal{S}_t \quad \text{where } \Delta\mathcal{S}_t = \mathcal{S}_t \setminus S_{t-1}. \tag{2.18}$$

A Venn diagram illustrating these nested information sets is shown in Figure 2.1. This is 'the third law of information for traders" and follows directly from equations 2.15 and 2.16 on the preceding page. Simply put, it states that a Universe measured in the way we have defined evolves entirely randomly, albeit in a manner that may be influenced by "initial conditions." Although this framework may seem very abstract, it will directly affect how we go about *solving* for optimal trading strategies. A key part of that is the third law.

2.3.2. On the meaning of random

It's important to note that although we are saying that the Universe evolves *randomly* we are not saying it is *meaningless*. Many properties of the Universe that we choose to measure, $\boldsymbol{m}(\mathcal{I}_t)$, may be determined by *unpredictable* incremental information but that does not mean these properties are without meaning; it just means that they are *unpredictable from information known prior to time* t.

2.3.3. Alphas and linear additive noise

All traders create models from the information available to them prior to the trade time and use this information to determine the

position they should hold. Above I introduced the notations $\boldsymbol{m}(\mathcal{I}_t)$ to represent the metrics of the information set we seek to predict and $\boldsymbol{\alpha}(\mathcal{I}_{t-1})$ for a predictive function of some kind we have created. If this function is *useful*, it would not be unreasonable to expect that

$$\mathbb{E}_{t-1}[\boldsymbol{m}(\mathcal{I}_t)] = \boldsymbol{\alpha}(\mathcal{I}_{t-1}), \tag{2.19}$$

and I will take this as the *definition* of a useful forecasting function.

Equation 2.19 is satisfied by processes that may be described by *linear additive noise*

$$\boldsymbol{m}(\mathcal{I}_t) = \boldsymbol{\alpha}(\mathcal{I}_{t-1}) + \boldsymbol{\varepsilon}(\Delta\mathcal{S}_t) \quad \text{where } \mathbb{E}_t[\boldsymbol{\varepsilon}(\Delta\mathcal{S}_t)] = \mathbf{0}. \tag{2.20}$$

This form is used almost exclusively when modelling financial data. The relationship is also satisfied by *multiplicative noise*, such as

$$\boldsymbol{m}(\mathcal{I}_t) = \boldsymbol{\alpha}(\mathcal{I}_{t-1}) \times \varepsilon_t \quad \text{where } \mathbb{E}_t[\varepsilon_t] = 1. \tag{2.21}$$

In both cases, we would describe the alpha as an *unbiased estimator* of the metric of interest. After the trade is made and the market evolves, or *ex post*, we are able to determine the value of the random *innovation* $\boldsymbol{\varepsilon}_t = \boldsymbol{\varepsilon}(\Delta\mathcal{S}_t)$, for it is simply $\boldsymbol{m}(\mathcal{I}_t) - \boldsymbol{\alpha}(\mathcal{I}_{t-1})$ for the case of linear additive noise. Thus, it must be true that $\boldsymbol{\varepsilon}_t$ is knowable at the time of the next trade decision and so $\boldsymbol{\varepsilon}_t \subseteq \Delta\mathcal{I}_t$. To emphasize the stochastic nature of the alpha, we can write it in the form

$$\boldsymbol{\alpha}(\mathcal{I}_{t-1}) = \boldsymbol{\alpha}(\Delta\mathcal{S}_{t-1}, \mathcal{I}_{t-2}). \tag{2.22}$$

2.3.4. Non-additive noise

Equations 2.20 and 2.21 are, of course, not the *only* ways to structure a relationship that satisfies equation 2.19. We can also consider *generalized* relationships in which the alpha controls the parameters of a distribution from which the metrics of the information set are drawn. The most general relationship is

$$\boldsymbol{m}(\mathcal{I}_t) \sim \mathrm{D}[\boldsymbol{\theta}(\mathcal{I}_{t-1})]\,, \tag{2.23}$$

where $\mathrm{D}[\boldsymbol{\theta}]$ is a probability distribution parameterized by prior information from which the metrics of choice are drawn. Some of these distributions, including those in the exponential family, allow the

conditional mean to be extracted and equation 2.20 on the facing page still be used, but for others, for example, the Dirichlet, this is not true. This structure is necessary when the metric is not representable as a continuous random variable.

2.3.5. Alphas are stochastic

Prior to a trade decision, referred to as *ex ante*, we have access to information $\mathcal{I}_{t-1}$ that informs our decision making process. Equation 2.22 on the preceding page shows that $\boldsymbol{\alpha}(\mathcal{I}_{t-1})$ may itself be a stochastic quantity as it is a function of a stochastic information set. This is an important, and subtle, point. Although $\boldsymbol{\alpha}_t$ itself is known exactly at time $t-1$, because it is a function of the trader's own construction built out of data they have observed, the entire time series $\{\boldsymbol{\alpha}_t\}_{t\in\mathcal{T}}$ may be a stochastic process.

2.3.6. Holdings are stochastic

If alphas are stochastic, and holdings are determined by a policy function of the alpha, $\boldsymbol{h}(\boldsymbol{\alpha}_t \dots)$, then holdings themselves are also potentially a stochastic process. This is very important for a trader who wishes to interact with the market repeatedly, as distinct from an investor who may be content to interact with the market only once. If trade decisions are a function of a stochastic alpha and the prior holding $\boldsymbol{h}(\boldsymbol{\alpha}_t, \boldsymbol{h}_{t-1} \dots)$, which is the case when transaction costs are significant, then the position function contains two separate stochastic processes, $\boldsymbol{\alpha}_t$ and $\boldsymbol{h}_{t-1}$. In this analysis, I will assume that they are not correlated: that is, the returns of the market are independent of our positions and we are not *The London Whale* [47].[6] This assumption not only affects the holding function, as above, but also affects the alphas constructed as it means we are generally seeking

[6]Apparently, the traders at JPMorgan decided to "defend" their positions by acquiring more inventory thereby pushing up prices (see Section 2.3.7 on the following page) and increasing the mark-to-market value of their existing positions. This strategy is dangerous because ultimately one will end up owning the entire market and in fire sale liquidation the losses may exceed the gains created by trading.

alphas that may be written as follows:

$$\boldsymbol{\alpha}(\mathcal{I}_{t-1} \setminus \{\boldsymbol{h}_{t-1}, \boldsymbol{h}_{t-2} \dots\}), \tag{2.24}$$

that is, alphas that are not a function of our prior trading history.

2.3.7. Market impact

Market impact is the effect on prices of trades done to enter a position. Generally, market impact is adverse, meaning that trading causes price moves that decrease the available profits to the trader. For a small speculator, it is reasonable to ignore market impact apart from the *slippage* between the prices used to calculate the alpha and the prices we actually execute at. For large traders, market impact can be catastrophic, as both Long-Term Capital Management [37] and JPMorgan can testify.

It is easy to see that permanent market impact of trading in an efficient market ought to be a linear function of trade size. Following Pete Kyle's paper [33], suppose the effect on prices, ΔP_t, of a traded volume, V_t, was something like

$$\Delta P_t = \mu + \lambda |V_t|^\gamma \operatorname{sgn} V_t. \tag{2.25}$$

If $\gamma > 1$, then a trade of size $2V_t$ has more than twice the impact of two trades each sized V_t. Profit could then be consistently made by purchasing, say, $10,000$ shares of a company, forcing the price higher, and then selling the acquired inventory slowly back into the market. This would generate a net profit, although the trader has market risk. The market risk could be reduced by, at the same time, selling short $10,000$ shares of another, different but highly correlated, company, and buying back that inventory slowly over time. Both trades would be possible and the combined risk would be manageable. The same arguments can be made for $\gamma < 1$, with the ordering of the trades reversed. The only scenario in which this market manipulation does not make money is when $\gamma = 1$, leaving "Kyle's lambda" relationship $\Delta P_t = \mu + \lambda V_t$. However, there is substantial empirical evidence to support *temporary* market impact[7] as having a nonlinear relationship, particularly a square-root relationship [35].

[7]Meaning price changes that revert after the trading activity inducing them has been removed.

2.3.8. Weakly forecastable markets

Taking unconditional variances of the linear additive noise model, equation 2.20 on page 34 gives[8]

$$\mathbb{V}[\boldsymbol{m}(\mathcal{I}_t)] = \mathbb{V}[\boldsymbol{\alpha}(\mathcal{I}_{t-1})] + \mathbb{V}[\boldsymbol{\varepsilon}(\Delta\mathcal{S}_t)] \tag{2.26}$$

as the two terms are independent by construction. ($\{\boldsymbol{\varepsilon}_t\} \cap \mathcal{I}_{t-1} = \emptyset \Rightarrow \partial\boldsymbol{\alpha}_t/\partial\boldsymbol{\varepsilon}_t = 0$.) Since most of the time the metric of the information set that we care about is the set of returns of tradable assets, equation 2.26 can also be written in the more familiar form as follows:

$$\mathbb{V}[\boldsymbol{r}_t] = \mathbb{V}[\boldsymbol{\alpha}_t] + \mathbb{V}[\boldsymbol{\varepsilon}_t]. \tag{2.27}$$

A weakly forecastable process is one in which $\mathbb{V}[\boldsymbol{\alpha}_t] \ll \mathbb{V}[\boldsymbol{\varepsilon}_t]$. A *weakly forecastable market* is one in which returns are mostly described by the random innovations $\boldsymbol{\varepsilon}_t$ and only marginally by the alpha, $\boldsymbol{\alpha}_t$, thus $\mathbb{V}[\boldsymbol{r}_t] \approx \mathbb{V}[\boldsymbol{\varepsilon}_t]$.

2.3.9. Efficient markets

An *efficient market* is one in which $\mathbb{V}[\boldsymbol{r}_t] = \mathbb{V}[\boldsymbol{\varepsilon}_t]$. (Note that this *does not* require that $\mathbb{E}[\boldsymbol{\alpha}_t] \equiv \mathbf{0}$, just that the alpha is not stochastic.) Sadly this criterion also applies to a completely irrational market, so lack of predictability should not be taken to confirm either that prices are "correct" or that prices are irrational. In this work, I will not assume that markets are efficient, for if I did, the work would be pointless, but I will also not assume that they are wildly inefficient with large gains available with minimal risk. I will assume that they are weakly forecastable.

2.3.10. Forward alpha and increasingly weakly forecastable returns

We have introduced the concept of alpha to represent the conditional expectation of future returns, $\boldsymbol{\alpha}_t = \mathbb{E}_{t-1}[\boldsymbol{r}_t]$, but we may also be interested in expected returns at larger forecast horizons. In the

[8] Using the notation $\mathbb{V}_t[x|y]$ to represent the conditional variance of x given y at time t and $\mathbb{V}[x]$ to represent the unconditional variance of x.

most general terms, we consider $\boldsymbol{\alpha}_{s,t} = \mathbb{E}_s[\boldsymbol{r}_t]$ and so the previously discussed $\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_{t-1,t}$. Terms such as $\boldsymbol{\alpha}_{t-1,t+1}$ are current expectations of the returns of assets not in the next period but a period after that, which we call "forward alpha."

In close to efficient market, we should expect not only that $\mathbb{V}[\boldsymbol{\alpha}_t] \ll \mathbb{V}[\boldsymbol{\varepsilon}_t]$, or equivalently $\mathbb{V}[\boldsymbol{\alpha}_t] \ll \mathbb{V}[\boldsymbol{r}_t]$, but also

$$\mathbb{V}[\boldsymbol{\alpha}_{s,t}] \gg \mathbb{V}[\boldsymbol{\alpha}_{s,u}] \quad \text{for } s < t < u. \tag{2.28}$$

Laying out the sequence of returns and alphas gives

$$\mathbb{V}[\boldsymbol{r}_t] \gg \mathbb{V}[\boldsymbol{\alpha}_{s,t}] \gg \mathbb{V}[\boldsymbol{\alpha}_{s,t+1}] \gg \mathbb{V}[\boldsymbol{\alpha}_{s,t+2}] \dots \tag{2.29}$$

$$\Rightarrow \quad 1 \gg \frac{\mathbb{V}[\boldsymbol{\alpha}_{s,t}]}{\mathbb{V}[\boldsymbol{r}_t]} \gg \frac{\mathbb{V}[\boldsymbol{\alpha}_{s,t+1}]}{\mathbb{V}[\boldsymbol{r}_t]} \gg \frac{\mathbb{V}[\boldsymbol{\alpha}_{s,t+2}]}{\mathbb{V}[\boldsymbol{r}_t]} \dots \tag{2.30}$$

Thus the R^2 of the forward alphas are smaller than those of the near alpha and get smaller as the forecast period becomes more distant. In real world terms, forecasting deep into the future is not useful and it becomes progressively less useful the further forward one looks. I will refer to a market with this property as an *increasingly weakly forecastable market.*

2.4. Performance Statistics

2.4.1. Stochastic programming

In the most general terms, a trader will choose a performance statistic for future profits, $Z(\{\boldsymbol{h}_u, \boldsymbol{r}_u\}_{u>t-1})$, and seek to determine the set of *future* positions $\{\hat{\boldsymbol{h}}_u\}_{u>t-1}$ such that

$$\{\hat{\boldsymbol{h}}_u\}_{u>t-1} = \underset{\{\boldsymbol{h}_u\}_{u>t-1}}{\arg\max}\, \mathbb{E}_{t-1}[Z(\{\boldsymbol{h}_u, \boldsymbol{r}_u\}_{u>t-1})] \tag{2.31}$$

where the extremum is found over the space of potential future holdings $\{\hat{\boldsymbol{h}}_u\}_{u>t-1}$. Solving such a problem is known as a *stochastic programming* and is described in books such as the one by Birge and Louveaux [5]. Although conceptually straightforward, the analysis of such problems is complicated by the fact that trade action taken now may affect the future value of the statistic. In general terms, we lay out all possible futures and provide each one with a probability weight. The optimal policy is then the one that maximizes

the expected value of the statistic we chose to score those potential futures with.

A good example of this is pricing options with a binomial tree, as developed by Cox, Ross, and Rubinstein [9]. The value of the option is known precisely on the settlement date and so all of these values may be laid out with their associated probabilities, based upon a model for the random evolution of stock prices. All possible paths to those values are laid out with, in this case, the assumption that they follow a recombining binomial tree. At any node of the tree, the value of the option is the discounted future value of the expected payoff in the future states accessible from that node, which is known precisely for the settlement date. The expectations for the day before may then be computed, and the procedure iterates back to the current date. The value on that date is then known, which is the price of the option. This is well explained in books such as Hull's [27].

There are two roadblocks to applying the C–R–R method to trading strategy, however, we have the following:

(i) The tree is not recombining once transaction costs are introduced as the actions taken then depend on the prior actions so the value computed at each node, the trader's wealth, cannot be simply related from one point in time to another.
(ii) Assuming we are successful traders, and so our wealth increases without limit through time, there does not exist a set of "final nodes" at which the value of any given strategy is known.

However, item (ii) can be dealt with *if* we are able to assume that the behavior of the strategy at some distant point in the future is representable by its average properties. More technically, if there exists a $T \gg t$ such that the conditional distribution of the holdings as seen from the initial trade point is "close" to its the unconditional distribution, then this technique will work. Unfortunately, this limit tends to defeat the strategy of enumerating all possible states of the market.

2.4.2. Functional solutions

The solution to equation 2.31 on the preceding page requires enumerating values for all future positions $\{\boldsymbol{h}_u\}_{u>t-1}$ and the search for

a solution is search over the space within which they sit. This potentially requires the examination of millions of variables. However, if the solution could be expressed in terms of a simple *policy function*, $\hat{\boldsymbol{h}}(\mathcal{I}_{t-1})$, then less unwieldy solutions may be expressed by the trader.

It is clearly generally true that

$$\{\hat{\boldsymbol{h}}_u\}_{u>t-1} = \underset{\{\boldsymbol{h}(\mathcal{I}_{u-1})\}_{u>t-1}}{\arg\max} \mathbb{E}_{t-1}[Z(\{\boldsymbol{h}(\mathcal{I}_{u-1}), \boldsymbol{r}_u\}_{u>t-1})], \quad (2.32)$$

where the search is over the space of causal functions $\boldsymbol{h}(\mathcal{I}_{u-1})$ and the solution to equation 2.32 is identical to that of equation 2.31 on page 38 because there must exist a function with the property

$$\hat{\boldsymbol{h}}(\mathcal{I}_{u-1}) = \hat{\boldsymbol{h}}_u \quad (2.33)$$

within the searched space.

2.4.3. Separable statistics

Given equation 2.31 on page 38, and the considerations above, we would like to solve problems where the optimal policy depends on the information available prior the to the decision point and *not* on the specific time itself. This becomes simpler when the statistic to be optimized can be separated by time. For compactness, let

$$Z_t = Z(\{\boldsymbol{h}_u, \boldsymbol{r}_u\}_{u\geq t}) \quad (2.34)$$

for some given trade time t. Then separable statistics of the forms

$$Z_t = Z'_t + Z_{t+1} \qquad \Rightarrow \qquad Z_t = \sum_{u=t}^{\infty} Z'_u \quad (2.35)$$

and

$$Z_t = Z'_t Z_{t+1} \qquad \Rightarrow \qquad Z_t = \prod_{u=t}^{\infty} Z'_u \quad (2.36)$$

will provide more readily solvable systems. For these statistics to converge, it must be true that $\lim_{t\to\infty} |Z'_t| = 0$ or 1, respectively.[9] Statistics built around combinations of moments of a distribution, such as

[9]This is not sufficient to guarantee convergence of Z_t, but it is necessary.

the objective function of Markowitz's mean–variance optimization [39] strategy or utility optimization with negative exponential utility, are separable under addition, in the manner of equation 2.35 on the preceding page. Both the mean and variance are separable statistics, as is the geometric mean, but the Sharpe ratio is not.

2.5. The Hierarchy of Optimization Strategies

2.5.1. Gross profit maximization

There is a hierarchy of economic statistics that we may seek to maximize in deciding our trading strategy and the base of that hierarchy is gross profit maximization. That is,

$$Z_t = \sum_{u=t}^{\infty} \delta^{u-t} \boldsymbol{h}_u^T \boldsymbol{r}_u. \tag{2.37}$$

This is simply the sum of the profits that arise from each future holding $\boldsymbol{h}_u$ for $u \geq t$. The term δ is a discount factor representing the time value of money, which will typically be close to but less than unity.[10]

This is a separable statistic and, if the alpha is known, then

$$\mathbb{E}_{t-1}[Z_t] = \boldsymbol{h}_t^T \boldsymbol{\alpha}_t + \delta\, \mathbb{E}_{t-1}[\boldsymbol{r}_{t+1}^T \boldsymbol{h}_{t+1}] + \cdots \tag{2.38}$$

With the assumption of increasingly weakly forecastable markets, we can discard the discounted forward terms, as their expected value is close to zero, and so the task is merely to maximize a deterministic function

$$\Omega(\boldsymbol{h}_t, \boldsymbol{\alpha}_t) = \boldsymbol{h}_t^T \boldsymbol{\alpha}_t. \tag{2.39}$$

This function is clearly *not* related to the prior position, $\boldsymbol{h}_{t-1}$, so there can be no implied dependence on the "current" position in the holding function. As this expression is linear in the alpha,

[10] Under the reasonable assumptions that neither the position sequence nor the return sequence are divergent, the presence of this term is sufficient for the infinite sum to converge.

it's straightforward to see that aligning the signs of the positions with the alpha will lead to an objective for which all $\partial\Omega/\partial\alpha_{it}$ are strictly positive. There are no other constraints or ways of fixing the function.

If your objective is to maximize gross profits, you should trade an *unlimited amount in the direction of the alpha*. If you're right, you'll make an unlimited amount of money, and if you're wrong, you'll lose an unlimited amount of money, but you're more likely to be right than wrong, so go for it! To be clear, this is not advocated to be a real strategy and, in reality, you cannot do this, but the solution contains a key, albeit fairly obvious, truth: if you want to maximize gross profits, your trades should align with your expectations of returns. Essentially, the best we can say is

$$\boldsymbol{h}(\boldsymbol{\alpha}_t) \propto \operatorname{sgn} \boldsymbol{\alpha}_t, \tag{2.40}$$

with the sign function applied element-wise.

However trivial this may seem, it is, in fact, a major step forward in our analysis. We have gone from predicting returns to suggesting trades that should be should occur when a trader possesses information about the future distribution of returns.

2.5.2. Linear programming

Having removed[11] the stochastic elements of equation 2.38 on the preceding page, we are left with what is called a *linear program*. These are systems which contain a linear objective function, such as equation 2.39 on the previous page, and also (possibly) linear constraints. Such systems are *extensively* studied and described in books such as the one by Murty [44]. They have simple properties which are easy to understand and are immensely useful.

Consider the task of numerically maximizing a linear objective function, such as $\Omega = \boldsymbol{c}^T\boldsymbol{x}$, subject to linear constraints $A\boldsymbol{x} \geq \boldsymbol{b}$. This is a linear program. First, we must chose a feasible solution,

[11] By discarding terms that we assume are likely to be negligible.

which is any of the solutions contained inside the volume defined by the *polytope* that the constraint equation represents.

For simplicity, let there by only two dimensions, (x, y), the constraint equations be

$$x + y \leq 1, \quad x - y \geq -1, \quad x \geq -1, \quad x \leq 2, \quad x + y \geq -2, \quad (2.41)$$

and the objective be $\Omega = 2x - y$. This is compactly represented in matrix notation by

$$\begin{pmatrix} 1 & 1 \\ -1 & 1 \\ -1 & 0 \\ 1 & 0 \\ -1 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \leq \begin{pmatrix} 1 \\ 1 \\ 1 \\ 2 \\ 2 \end{pmatrix} \quad \text{and} \quad \Omega = \begin{pmatrix} 2 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}. \qquad (2.42)$$

In Figure 2.2 on the following page, any point in the blue-shaded region satisfies the constraints. This is called the *feasible region*. The point $(1, -1)$ is potentially a solution because it is within the shaded region and so does satisfy all the constraints.

Our optimization algorithm starts by choosing an initial feasible solution $(x, y) = (1, -1)$. The first step might be to change x by a small amount, δx. The objective then changes by $\partial\Omega/\partial x \; \delta x = 2\,\delta x$. If δx is positive, say 0.1, that increases the objective from 6 to 6.2, so this was clearly a step in the right direction. Maybe we should do it again? As the objective function is *linear* in x, the same step δx will always increase Ω by the same amount. So we keep increasing x until we can't do it anymore, which is when we hit the boundary of the polytope, in this case, the vertex defined by the lines $x + y \leq 1$ and $x \leq 2$ at $(2, -1)$. At that point, it is not possible to increase x any more, but we can change y. As $\partial\Omega/\partial y = -1$, we should decrease y, and we do it until we hit another boundary. That is the vertex defined by the lines $x \leq 2$ and $x + y \geq -2$, at $(2, -4)$. This point gives the most extreme value, $\hat{\Omega} = 8$, and is the solution to the problem.

The objective function and solution are illustrated in Figure 2.3 on page 45. The problem worked out exhibits a fundamental feature of the solutions to linear programs: they are always located at one of the vertices defined by the polytope of constraints. None of the solutions exist within the interior of the feasible region.

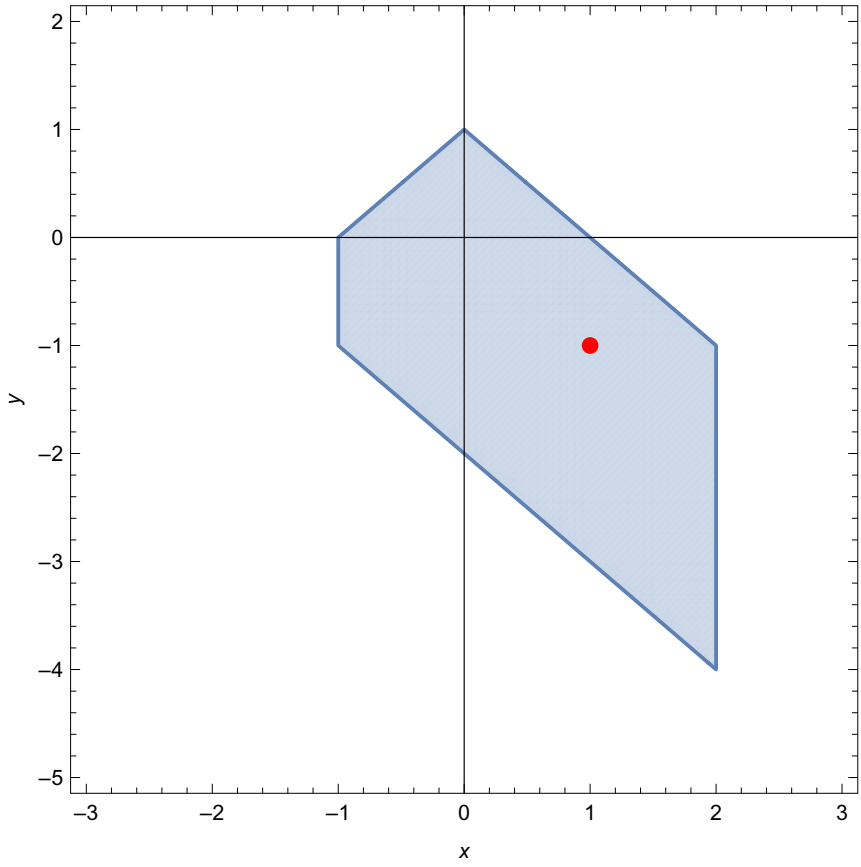


Figure 2.2. The polytope represented by both equations 2.41 and 2.42 on page 43 and a feasible solution $(1, -1)$.

2.5.3. Gross profit maximization with risk limits

Section 2.5.2 has provided the toolkit necessary to solve a more realistic version of Section 2.5.1 on page 41, which is gross profit maximization with position limits. In the real world, there is always some constraint that limits the maximum bet that can be made, whether it is the maximum position that can be taken in any given asset set by an external risk manager, either at your brokerage or by regulators, or the constraint of how much capital may be put at risk across the entire portfolio, a limit on whether you can go short or

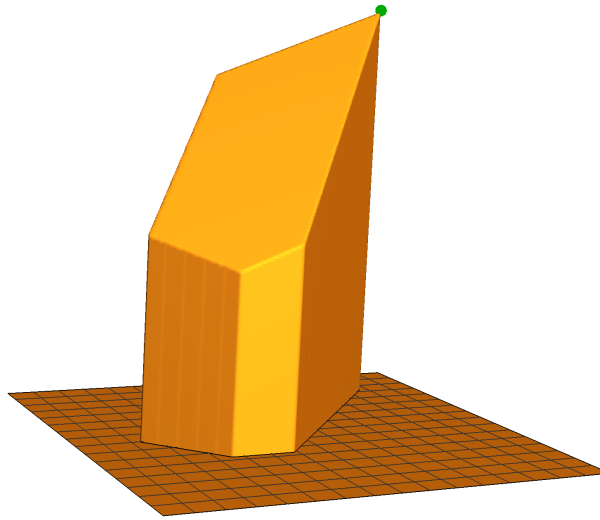


Figure 2.3. The linear program objective value $\Omega(x, y)$ over the feasible region. The xy plane has been rotated relative to that in Figure 2.2 on page 44 to make the shape of the polytope clearly visible. The green dot marks the solution.

not, a requirement that the portfolio have the same equity long or short (zero net investment). All of these kinds of constraints can be written as $A\boldsymbol{h}_t = \boldsymbol{b}$ and incorporated into a linear program.

The most elementary version of these constraints is a simple position limit of, for example, L shares[12] for any tradable asset. We know that a gross profit maximizer when allowed to trade *up to* L shares will always max out and take a position of exactly L shares, so the optimum holding function is now known exactly to be

$$\boldsymbol{h}(\boldsymbol{\alpha}_t) = L \operatorname{sgn} \boldsymbol{\alpha}_t. \tag{2.43}$$

As before, this is essentially a trivial result. However, it is important to know how to get here and what the adoption of such a holding function means about the risk preferences of the trader, especially as it means they are completely indifferent to risk. These are tools that can be deployed to solve more complex problems.

[12]Or contracts, options, etc.

2.5.4. The hierarchy of optimization strategies

So far I have described gross profit maximization (Section 2.5.1) as both "naked" and with constraints (Section 2.5.3). To this I should add net profit maximization, also with constraints.

That is,

$$\Omega = \mathbb{E}_{t-1}\left[\sum_{u=t}^{\infty}\delta^{u-t}\left(\boldsymbol{h}_u^T\boldsymbol{r}_u - \boldsymbol{\kappa}^T||\boldsymbol{h}_u - \boldsymbol{h}_{u-1}||_{\mathbf{1}}\right)\right] \tag{2.44}$$

$$\text{and} \quad A\boldsymbol{h}_u \geq \boldsymbol{b} \quad \text{for } u \geq t. \tag{2.45}$$

Branching in complexity, there is unconstrained mean–variance optimization, which represents risk-averse gross-profit maximization

$$\Omega = \mathbb{E}_{t-1}\left[\sum_{u=t}^{\infty}\delta^{u-t}\left(\boldsymbol{h}_u^T\boldsymbol{r}_u - \lambda\boldsymbol{h}_u^T V_u\boldsymbol{h}_u\right)\right], \tag{2.46}$$

to which can be added risk-averse net-profit maximization

$$\Omega = \mathbb{E}_{t-1}\left[\sum_{u=t}^{\infty}\delta^{u-t}\left(\boldsymbol{h}_u^T\boldsymbol{r}_u - \lambda\boldsymbol{h}_u^T V_u\boldsymbol{h}_u - \boldsymbol{\kappa}^T||\boldsymbol{h}_u - \boldsymbol{h}_{u-1}||_{\mathbf{1}}\right)\right], \text{ etc.} \tag{2.47}$$

All of these strategies represent incremental changes from each other and their relationships are illustrated in Figure 2.4 on the next page.

A key feature of the risk-averse systems is that the objective is no longer linear in the position and that means it is no longer true that the solution must exist on the boundary of the polytope defined by constraints. The solution may be within the *interior* of this region, and such a solution may also exist without any constraints at all. These problems cannot be solved by linear programming but may be tractable to simple calculus in the absence of transaction costs. Some authors, such as Kolm and Ritter [32], use the fact that market impact is an increasing function of trade size to facilitate a cost term which is quadratic, thus delivering a system whereby risk aversion and transaction costs can both be implemented via a simple quadratic term. Unfortunately, this assumption relies on the Kyle model which is not supported by data.

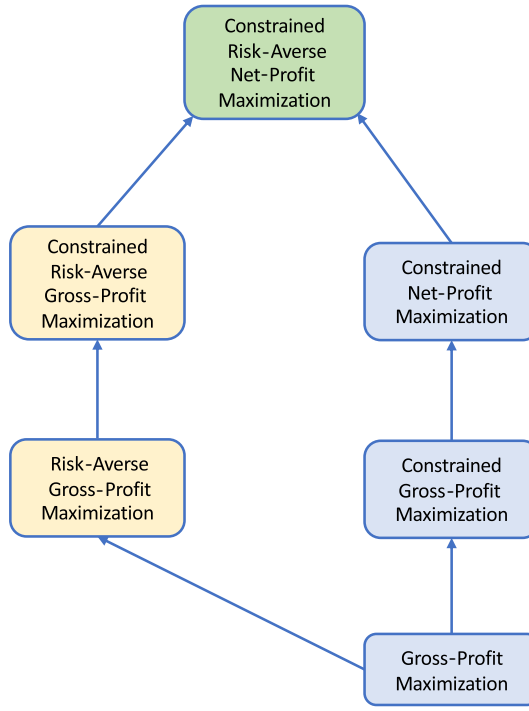


Figure 2.4. The hierarchy of optimization strategies relevant to trading strategies. Blue shading represents linear programming, yellow quadratic programming, and green is linear–quadratic programming.

However, as the absolute value function, $|x|$, is not differentiable at zero, solutions involving piecewise linear transaction costs, which generally depend on the absolute value of the trade size, cannot be obtained through these methods. Remarkably, the absolute value function is not much of a problem to linear programming systems as it can be represented as a new non-negative variable representing the trade size. Adding such a variable, often called a "slack variable" in optimization, is legitimate provided it appears both in the objective function and a constraint.[13] To illustrate this construction, consider

[13]I learned this trick from Ken Nickerson while in P.D.T. at Morgan Stanley. It is well known to those who study operations research as it is a key part of the method of solving linear programs.

a holding, $h_{i,t-1}$, and desired new position, h_{it}. The constraints

$$h_{it} = h_{i,t-1} + b_{it} - s_{it} \tag{2.48}$$

$$b_{it} \geq 0 \tag{2.49}$$

$$s_{it} \geq 0 \tag{2.50}$$

are sufficient to make the new position equal the old position plus a potential buy trade, b_{it}, and sell trade, s_{it}. Since variables in linear programs will bind at constraints, and b_{it} and s_{it} contribute with opposite signs to the total objective, we will never be in a situation where $b_{it} > 0$ and $s_{it} > 0$ or the converse. The solution will always drive one or the other to zero!

2.5.5. Optimizing the Sharpe ratio

The Sharpe ratio[14] is not a separable statistic. Formally, a trader would want to optimize the expected value of the sample Sharpe ratio of future portfolio returns, which is

$$\Omega = \mathbb{E}_{t-1}\left[\lim_{F\to\infty} \frac{\frac{1}{F}\sum_{u=t}^{t+F} \boldsymbol{h}_u^T \boldsymbol{r}_u}{\sqrt{\frac{1}{F-1}\sum_{u=t}^{t+F}(\boldsymbol{h}_u^T \boldsymbol{r}_u)^2 - \frac{1}{F(F-1)}\left(\sum_{u=t}^{t+F} \boldsymbol{h}_u^T \boldsymbol{r}_u\right)^2}}\right]. \tag{2.51}$$

I think we would all agree this is a very nasty expression.[15] In some circumstances, this may be approximated by

$$\Omega \simeq \frac{\mathbb{E}_{t-1}[\boldsymbol{h}_t^T \boldsymbol{r}_t]}{\sqrt{\mathbb{V}_{t-1}[\boldsymbol{h}_t^T \boldsymbol{r}_t]}}, \tag{2.52}$$

which is definitely easier to work with, but in general equation 2.51 is not the direct target of trading strategy optimization.

[14] Examined in more depth in Essay 1, starting on p. 1.

[15] I have omitted the term in the *risk-free rate*, as this is largely irrelevant to traders, and the annualization factor as it is a constant that does not affect the optimization procedure.

2.5.6. Mean–variance optimization

Markowitz's mean–variance optimization elides the entire issue by choosing to maximize a simpler objective, which we would write as follows:

$$\Omega = \mathbb{E}_{t-1}[\boldsymbol{h}_t^T \boldsymbol{r}_t] - \lambda \mathbb{V}_{t-1}[\boldsymbol{h}_t^T \boldsymbol{r}_t]. \tag{2.53}$$

Equations 2.53 and 2.52 on the facing page are clearly not identical, and so optimizing each of them may deliver a slightly different solution.

2.5.7. The Golden Rule of Trading Strategy Design

In *Adventures in Financial Data Science* [18], I introduced The Golden Rule of Prediction, which states that you are free to use any method you want to invent an alpha but you must use legitimate statistical procedures to judge its accuracy. Similarly, we can define *The Golden Rule of Trading Strategy Design* as follows:

> You may use any causally legitimate procedure whatsoever to choose the holding function $\boldsymbol{h}(\boldsymbol{\alpha}_t, V_t, \boldsymbol{h}_{t-1} \dots)$ for your trading strategy but you cannot then use the measured *ex post* performance to optimize the function used.

This page intentionally left blank

Essay 3

Utility Theory-Based Portfolio Choice

In this essay, I will present solutions to the "frictionless" asset allocation problem for several probability distributions that are *leptokurtotic*, meaning that they possess significant density in the tails. In this context, frictionless means without transaction costs. These probability densities all share the trait that the mean and variance are not sufficient statistics for the distribution and we will see that the solutions differ from the linear holding function ($\boldsymbol{h}_t \propto \boldsymbol{\alpha}_t$). I will begin with a summary of utility theory that is intended to provide the context within which the solutions are developed. The goal is not to refine utility theory, or discuss its suitability for the solution of trading strategy problems, but to use it as a tool to exhibit this feature of the holding function.

3.1. Utility Theory and Risk Aversion

In Essay 1, we implicity examined portfolio choice via utility functions without explicitly calling that out. A utility function, $U(W_t)$, provides an idealized model for the way people in the real world judge different states of their wealth and utility based investment choice theorizes that trades are chosen to maximize the expectation of the present value of the utility of future wealth. In our notation,

something like

$$\hat{\boldsymbol{h}}_t = \arg\max_{\boldsymbol{h}_t} \mathbb{E}_{t-1}[U(W_t)]\,. \tag{3.1}$$

This framework was introduced by Daniel Bernoulli to resolve *The St. Petersberg Paradox* [4] in 1783. The criteria on $U(W_t)$ are that it be

(i) a monotonically increasing function of wealth and
(ii) have a first derivative that is a monotonically decreasing function of wealth.

It is straightforward to see that item (i) leads to gain seeking. Item (ii) leads to risk aversion when the second derivative of $U(W)$ is negative, due to Jensen's inequality. To see this additionally assume that the utility is smooth, so it may be replaced by it's Taylor series

$$\begin{aligned} U(W_t) = U(W_{t-1}) + U'(W_{t-1})(W_t - W_{t-1}) \\ + \frac{1}{2}U''(W_{t-1})(W_t - W_{t-1})^2 + \cdots \end{aligned} \tag{3.2}$$

(Using the standard notation $U'(W)$ for dU/dW, $U''(W)$ for d^2U/dW^2, etc.) Then the expected value of the utility may be written as follows:

$$\begin{aligned} \mathbb{E}_{t-1}[U(W_t)] = U(W_{t-1}) + U'(W_{t-1})\mathbb{E}_{t-1}[W_t - W_{t-1}] \\ + \frac{1}{2}U''(W_{t-1})\mathbb{V}_{t-1}[W_t - W_{t-1}] + \cdots \end{aligned} \tag{3.3}$$

As variance is a non-negative quantity, it is clear that the expected utility when future wealth is uncertain is less than that when future wealth is certain, given the same expected return and $U''(W) < 0$. Furthermore, expected utility is a decreasing function of variance so higher risks lead to a higher utility penalty,[1] which is risk aversion. In terms of the notation developed in the prior essays, this may be

[1]These results are presented to aid exposition, as they are well known in the literature.

written as follows:

$$\hat{\boldsymbol{h}}_t = \arg\max_{\boldsymbol{h}_t} \left\{ \boldsymbol{h}_t^T \boldsymbol{\alpha}_t + \frac{1}{2} \frac{U''(W_{t-1})}{U'(W_{t-1})} \boldsymbol{h}_t^T V_t \boldsymbol{h}_t + \cdots \right\} \tag{3.4}$$

as the maximization is not affected by the "constant" $U(W_{t-1})$. If we identify

$$\lambda = -\frac{1}{2} \frac{U''(W_{t-1})}{U'(W_{t-1})}, \tag{3.5}$$

then the form of equation 2.53 on page 49 is recovered. The ratio $-U''(W)/U'(W)$ is known as the *Arrow–Pratt* measure of absolute risk aversion.

3.2. Multi-Horizon Utility

Equation 3.1 on the preceding page is clearly a simplification of the problem we need to solve; in reality, we should consider the expected utility of future positions held more distantly into the future and discount in some way, such as

$$\hat{\boldsymbol{h}}_t = \arg\max_{\{\boldsymbol{h}_u\}_{u \geq t}} \mathbb{E}_{t-1} \left[U(W_{t-1} + \sum_{u=t}^{\infty} \delta^{u-t} \Delta W_u) \right]. \tag{3.6}$$

Here the utility is evaluated for discounted sequences of future changes in wealth and the expectation taken over the possible sequences. The sequences are likely affected in the *short term* by the immediate choice $\boldsymbol{h}_t$ but this has progressively less impact as the forward time recedes from the present. In a way, this is similar to the solutions of a linear differential equation, such as the wave equation, being a response to both boundary conditions, which progressively attenuate, and a steady state behavior, that exhibits itself well away from the boundary.

This is generally going to be a complicated problem to solve, although certain useful simplifications are possible. If the functional form of $U(W)$ is separable, such that $U(W + X) = U(W)U(X)$ or $U(W + X) = U(W) + U(X)$, for example, then the *Law of*

Iterated Expectations

$$\mathbb{E}_{t-1}[\mathbb{E}_t[x]] = \mathbb{E}_{t-1}[x] \tag{3.7}$$

allows equation 3.6 on the preceding page to be expressed as a series of independent decisions to which the same policy may be applied. Given our assumption of increasingly weakly forecastable markets, there may exist a time, $u \geq t$, at which we regard the markets as essentially unforecastable and all variables from that point onwards may be replaced with their unconditional expectations.[2] These values would be taken to be $\mathbb{E}_{t-1}[\boldsymbol{\alpha}_u] = \mathbf{0}$ and $\mathbb{E}_{t-1}[\boldsymbol{h}_u] = \mathbf{0}$ and may be substituted to discover the correct forward policy for $\boldsymbol{h}_{u-1}$. With this policy known, back propagation is used to obtain the solution $\hat{\boldsymbol{h}}_t$.

In the following, I will assume that the markets are sufficiently weakly forecastable, meaning that future alphas and positions are sufficiently random, that the correction to equation 3.1 on page 52 to deliver the "correct" policy under equation 3.6 on the preceding page is negligible, and I will proceed by solving equation 3.1 on page 52 alone without including discounted future terms. To justify this step *heuristically,* I look to the example of the $AR(1)$ time-series model. The expected value of a series, $a_t = \varphi a_{t-1} + \varepsilon_t$, at some future time $u > t$, is given by

$$\mathbb{E}_{t-1}[a_u] = \sum_{s=t}^{u} \varphi^{s-t+1} a_{t-1}. \tag{3.8}$$

Even with $\varphi = 0.1$, which would be a very large number if a_t were to represent the returns of a financial asset, the second-order correction, $\varphi^2 = 0.01$, is negligible.

3.3. Negative Exponential Utility and Moment Generating Functions

Bernoulli himself suggested the functional form $U(W) = \ln W$. This is the function that will deliver the Kelly criterion when $W_t = (1 + g_t)W_{t-1}$, as the utility function is separable by the properties of the

[2] Their "steady state" values.

logarithm. In contrast, I will mostly work with *negative exponential utility*, which has the form

$$U(W) = 1 - e^{-\lambda W} \tag{3.9}$$

for some risk aversion scale factor λ. It is a separable under multiplication as $e^{a+b} = e^a e^b$ and for outcomes described by densities rather than discrete choices, particularly those in the exponential family which I've shown provide an excellent description of markets data, computing the optimal policy involves the integral

$$\mathbb{E}_{t-1}\left[1 - e^{-\lambda(W_{t-1} + \boldsymbol{h}_t^T \boldsymbol{r}_t)}\right] \propto \int \dots \int e^{-\lambda \boldsymbol{h}_t^T \boldsymbol{r}_t} f(\boldsymbol{r}_t)\, d^n \boldsymbol{r}_t. \tag{3.10}$$

This integral is simply related to the *moment generating function* of the probability density[3] and is well studied in the statistical literature [30]. The Arrow–Pratt measure for equation 3.9 is just λ, meaning that risk aversion is constant, whereas it is $1/W$ for the choice of log utility, meaning that risk aversion is a *decreasing* function of wealth. The former seems a better description of the behavior of traders than the latter.

Figure 3.1 on the next page exhibits the shape of the objective functions examined in Essay 1 about the region $\Delta W = 0$, viewing them in this new context as utility functions. It is clear that they are very similar for small changes in wealth, but the tail behavior differs and this is where the higher moments of the distribution have an impact.

3.4. Selected Univariate Distributions

3.4.1. The normal distribution

Portfolio selection results via negative exponential utility are well established for the normal distribution and the results are those already presented in Section 1.3 on page 8. Nevertheless, I will exhibit the univariate solution here for completeness. The generalization to an arbitrary dimensional real space $\boldsymbol{r}_t \in \mathbb{R}^n$ is straightforward, and

[3] It is also the *bilateral Laplace transform* of the density, which is tabulated in resources, such as Gradsteyn and Rhyzik [22].

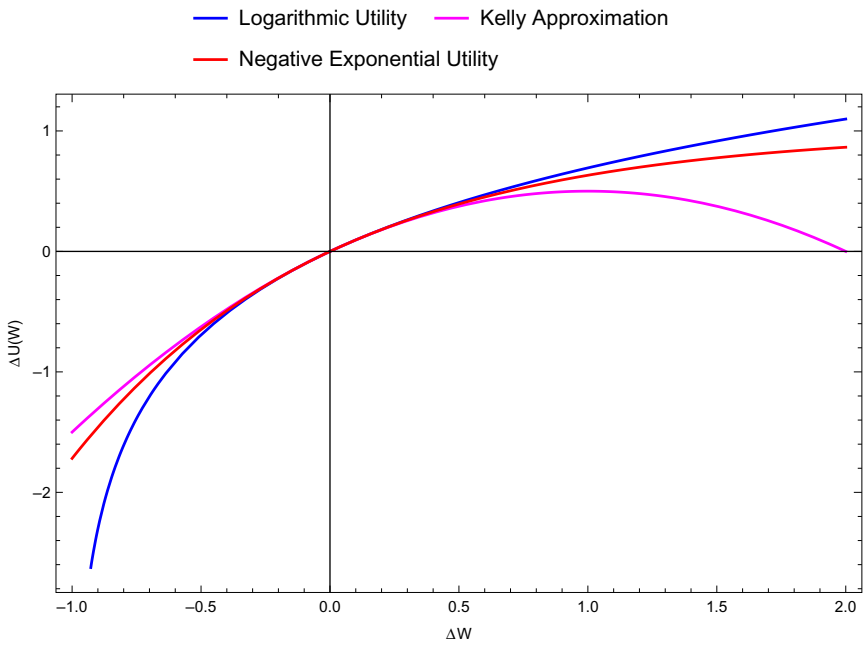


Figure 3.1. The form of the three utility functions examined in Essay 1, in the region of zero change in wealth. The blue line is the logarithmic function $\ln(1+\Delta W)$, magenta is the Kelly approximation of equation 1.16 on page 7, and red is the negative exponential function with $\lambda = 1$.

I will examine in Section 3.5 on page 63 where it is shown to be a particular case of the result for more general ellipsoidal distributions.

With one asset, assuming increasingly weakly predictable markets and zero transaction costs, the objective is to find

$$\hat{h}_t = \arg\min_{h_t} \int_{-\infty}^{\infty} e^{-\lambda h_t r_t} f(r_t, \alpha_t, \sigma_t)\, dr_t, \qquad (3.11)$$

$$\text{where } f(r_t, \alpha_t, \sigma_t) = \frac{1}{\sqrt{2\pi}\sigma_t} e^{-\frac{1}{2}\left(\frac{r_t-\alpha_t}{\sigma_t}\right)^2}. \qquad (3.12)$$

The integral may be evaluated by completing the square and change of variables, giving an objective function

$$\Omega(h_t) = e^{\frac{1}{2}\lambda h_t\left(\lambda h_t \sigma_t^2 - 2\alpha_t\right)}. \qquad (3.13)$$

Differentiating w.r.t. h_t

$$\frac{d\Omega}{dh_t} = \left(\lambda h_t \sigma_t^2 - \alpha_t\right) \lambda \Omega(h_t) \tag{3.14}$$

with single root

$$\hat{h}_t = \frac{\alpha_t}{\lambda \sigma_t^2}. \tag{3.15}$$

This is identical to the solutions derived by other methods in Essay 1 apart from the factor of 2 in the denominator, which may be generated by redefining λ if desired.

3.4.2. The Laplace distribution

The analytic solution: In my work, I've found that symmetric leptokurtotic distributions from the exponential family provide acceptable descriptions of data from financial markets [18]. The easiest distribution to tackle after the normal is the Laplace distribution, given by

$$\text{Laplace}(\alpha_t, \sigma_t) \ : \ f(r_t, \alpha_t, \sigma_t) = \frac{1}{2\sigma_t} e^{-\left|\frac{r_t - \alpha_t}{\sigma_t}\right|}. \tag{3.16}$$

The holding function is the solution of

$$\hat{h}_t = \arg\min_{h_t} \int_{-\infty}^{\infty} \frac{1}{2\sigma_t} e^{-\lambda h_t r_t - \left|\frac{r_t - \alpha_t}{\sigma_t}\right|} \, dr_t \tag{3.17}$$

and convergence at both the upper and lower limits is determined by the argument to the exponential. The integral may then be evaluated as

$$\Omega(h_t) = \frac{e^{-\lambda h_t \alpha_t}}{1 - \lambda^2 h_t^2 \sigma_t^2} \quad \text{where } |\lambda h_t \sigma_t| < 1. \tag{3.18}$$

Differentiating w.r.t. h_t and solving for $d\Omega/dh_t = 0$ give two roots, which are shown in Figure 3.2 on the following page.

$$h_t = \frac{-1 \pm \sqrt{1 + \alpha_t^2/\sigma_t^2}}{\lambda \alpha_t}. \tag{3.19}$$

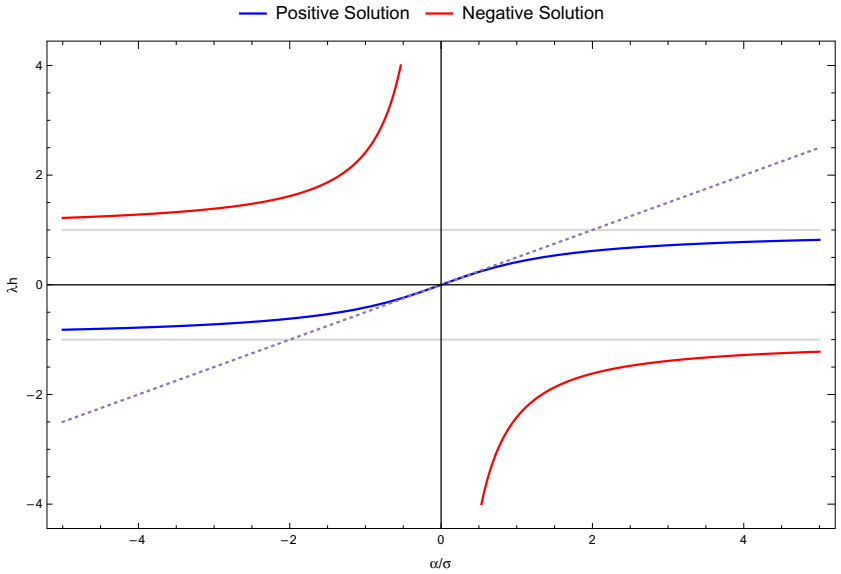


Figure 3.2. The two potential solutions to the holding function for utility maximization with the Laplace distribution. The red curve is the "negative solution" and the blue curve the "positive solution." The dotted line is the Markowitz solution.

From the figure, it's clear that although both solutions are increasing functions of the alpha, only one, the "positive solution" with the blue curve, has the property that $\operatorname{sgn} h_t = \operatorname{sgn} \alpha_t$, which is clearly necessary for a trader. This is the solution that maximizes utility, and so the holding function for the Laplace distribution is

$$h(\alpha_t) = \frac{\sqrt{1 + \alpha_t^2/\sigma_t^2} - 1}{\lambda \alpha_t}. \tag{3.20}$$

In terms of the alpha, this holding function has the Taylor series

$$h(\alpha_t) = \frac{\alpha_t}{2\lambda\sigma_t^2} - \frac{\alpha_t^3}{8\lambda\sigma_t^4} + O\left(\alpha_t^5\right), \tag{3.21}$$

which recovers the Markowitz solution for small alpha but has the limit

$$\lim_{\alpha_t \to \pm\infty} h(\alpha_t) = \pm \frac{1}{\lambda \sigma_t} \tag{3.22}$$

for large alphas.

This result is important on many levels:

(i) It shows that the utility approach can do more than just reproduce mean–variance optimization.
(ii) It shows that an analytic form for a holding function can be derived for a probability distribution which is a more reasonable description of markets data.
(iii) That solution has a "small alpha" limit equal to the Markowitz solution and a "large alpha" limit proportional to the sign of the alpha.

Cube root Kelly approximation: In Section 1.2.3 on page 8, I introduced the concept of "root Kelly" as a potentially superior transformation of the Kelly criterion solution than "fractional Kelly." This is the holding function

$$h^*(\alpha_t) = \left(\frac{|\alpha_t|}{2\lambda\sigma_t^2}\right)^k \operatorname{sgn}\alpha_t, \tag{3.23}$$

for some k such that $0 < k < 1$. Figure 3.3 on the following page shows the use of a cube root, $k = 1/3$, to approximate the exact holding function of equation 3.20 on the preceding page when $\lambda = 1/2$, which is the value that makes the continuous Kelly criterion equal the Markowitz portfolio. Such an approximation is not *necessary* to use the Laplace solution, as it is known exactly for all values of the alpha, but serves to illustrate that a practitioner of the suggested "root Kelly" investing strategy would actually respond to a signal in a manner not much different to a follower of the Laplace solution. Both approaches share the characteristic of "backing away" from a bet when its Z score, α_t/σ_t, is too large.

The optimal root with an ultraviolet cutoff: In quantum field theory, one often encounters integrals that do not converge. These are sometimes dealt with by introducing arbitrary "cutoffs" to the integral at the lower limit or the upper limit. These are referred to as "infrared" and "ultraviolet" cutoffs and often are based on the heuristic that the theory in question is simply not credible at limits of 0 and ∞. Figure 3.3 on the following page clearly shows that, with $k \to 1/3$, $h^*(\alpha_t) \approx h(\alpha_t)$ over a subset of the $Z = \alpha_t/\sigma_t$ axis. It is clearly interesting to ask whether 1/3 is the "best" value of k in some

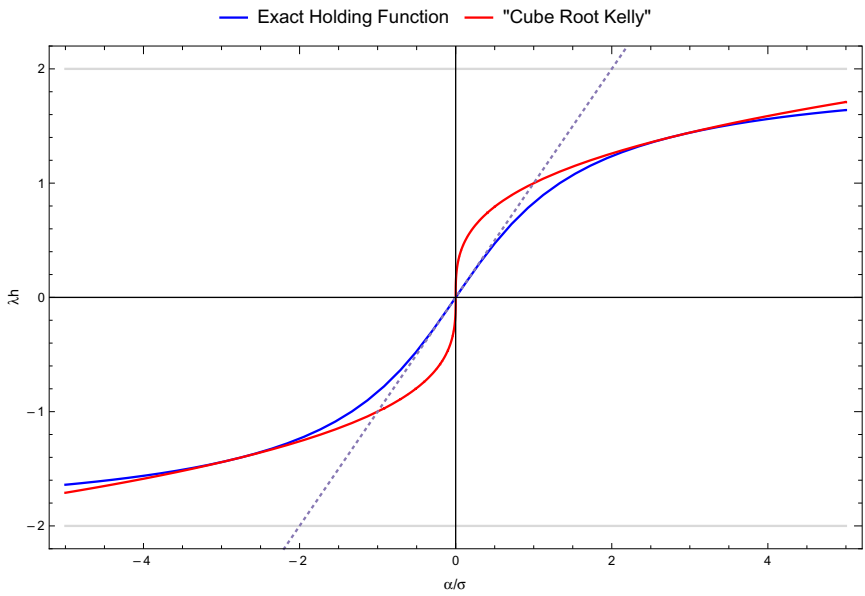


Figure 3.3. Approximation of the exact Laplace distribution holding function with a "cube root Kelly" holding function. The blue line is the Laplace solution, the red line the cube root Kelly function, the light grey lines are the asymptotes of the Laplace solution, and the dotted line is the Markowitz solution.

sense. We can simply write the integral

$$\frac{1}{B}\int_0^B \{h(\alpha_t) - h^*(\alpha_t)\}^2 \, d\alpha_t, \tag{3.24}$$

as a measure of the mean-squared error associated with this assumption, for some ultraviolet cutoff B, however this cutoff must be finite as the root-Kelly function is divergent for $\alpha_t \to \pm\infty$ whereas the analytic solution is not and so the integral will always diverge as $B \to \infty$.

Nevertheless, values of $\alpha_t/\sigma_t \gg 5$ are fairly unlikely, even with the Laplace distribution, and so we can argue that a cutoff of $B \approx 5$–10 is reasonable to impose on the integral. That is, we do not seek the optimal value of k when the error over the *entire* real line is considered but restrict the analysis to a subset of the reals that appear *reasonably likely* to occur. If the cutoff is placed at $B = 5$,

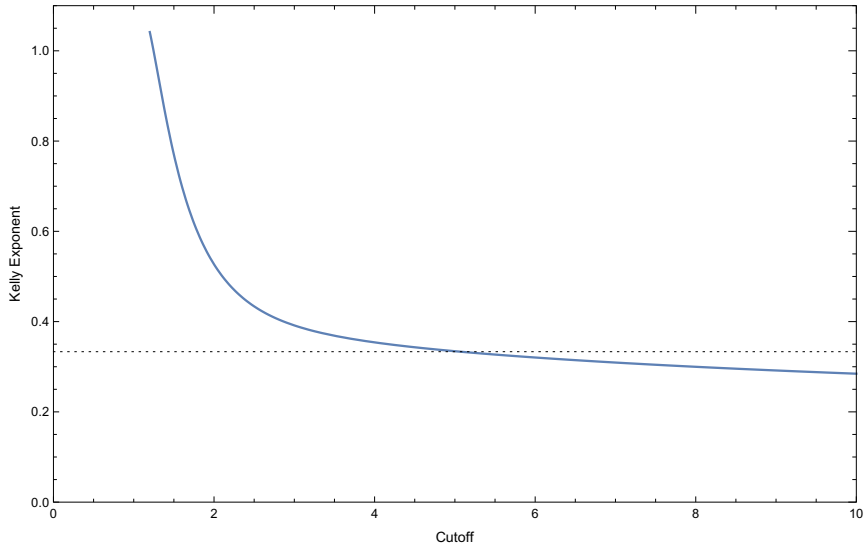


Figure 3.4. The dependence of the mean-squared error optimal Kelly exponent on the "ultraviolet" cutoff of the integral.

then, numerically, the value of k that minimizes the integral is $\hat{k} = 0.334$. It is not a cube root, but it is pretty close to it. Since we expect $\hat{k} \to 0$ as $B \to \infty$, though, it's important to see how sensitive this value is to the cutoff. This is exhibited numerically in Figure 3.4, and it appears to be quite a slowly decreasing function of the cutoff for $B > 4$. Thus the cube root Kelly approximation seems well founded.

One might argue that the *correct* way to approach minimizing the m.s.e. of the root Kelly approximation would be to weight these errors by their probabilities, under the given Laplace distribution of returns. However, this is a little pointless as we know *exactly* the correct holding function for that distribution. The purpose of this analysis is to support the general heuristic solution of cube root Kelly by demonstrating that it is "not far" from that analytical form.

Barrier trading approximation: One aspect of all of the holding functions presented so far is that they assume that the assets to be invested in are infinitely divisible. When trading in real markets, instead of a continuous holding function, it is often practical to use

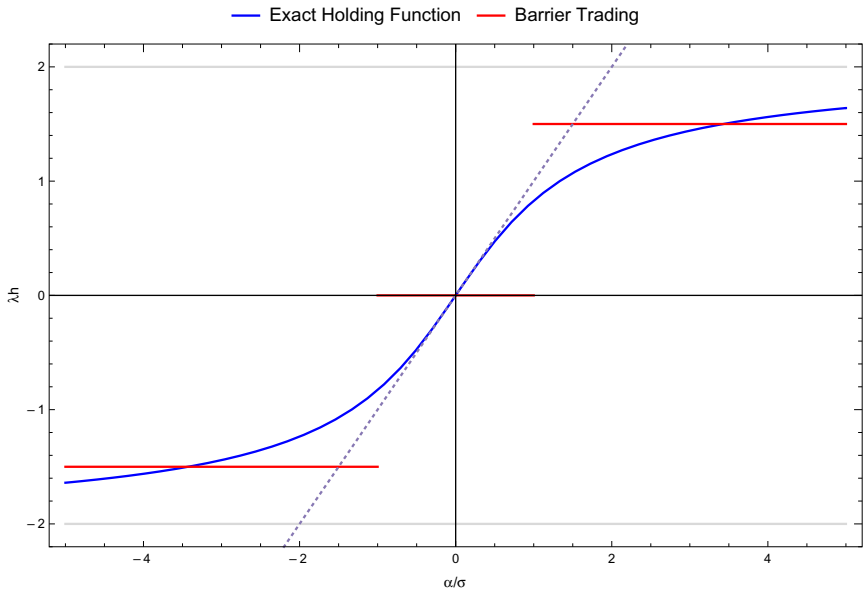


Figure 3.5. Approximation of the exact Laplace distribution holding function with a barrier trading rule. The blue line is the Laplace solution, the red line is the holding function for the barrier rule, the light grey lines are the asymptotes of the Laplace solution, and the dotted line is the Markowitz solution.

instead a discrete holding function. That is one of the form

$$h(\alpha_t) = \begin{cases} +L & \alpha_t \geq b \\ 0 & -b < \alpha_t < b \\ -L & \alpha_t \leq -b \end{cases} \tag{3.25}$$

for risk limited trades of size L and trade entry barrier b. The holding function for such barrier trading rule is shown in Figure 3.5. In the figure, the barrier is taken to be $b = \alpha_t/\sigma_t$ and the position size is $3/2$ whereas the limiting position size is $2\,\mathrm{sgn}\,\alpha_t$. I will explore such barrier trading rules in more detail in Essay 5.

Optimal barrier locations: In the same spirit taken to optimize the root Kelly approximation, it is also possible to optimize the barrier approximation. This involves two free parameters, L and b, and

we seek to optimize a modified version of equation 3.25 on the preceding page:

$$h(\alpha_t) = \begin{cases} \dfrac{L}{\lambda\sigma_t} \operatorname{sgn} \alpha_t & |\alpha_t/\sigma_t| \geq b \\ 0 & \text{otherwise} \end{cases}. \tag{3.26}$$

3.5. Ellipsoidal Distributions

In this section, we extend our analysis back to multivariate distributions but specialize our discussion to the subset of those probability distributions with ellipsoidal symmetry. One recipe[4] for constructing multivariate distributions is to consider the set of continuous multivariate distributions that are obtained from a normalized symmetrical univariate distribution $f(x^2)$ by the substitution $\{x \rightarrow \boldsymbol{x}, f(x^2) \rightarrow \mathcal{A}f(g^2)\}$, where g is the Mahalanobis distance $\Delta_\Sigma(\boldsymbol{x}, \boldsymbol{\mu}) = \sqrt{(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$ and $\mathcal{A}$ is a constant introduced to normalize the constructed distribution over its support, $\mathbb{R}^n$ or some subset thereof. Σ is not the covariance matrix of the distribution but that matrix, V, is proportional to Σ in a distribution dependent manner.[5] These distributions are discussed extensively in my prior work [17,20] and are referred to as possessing *ellipsoidal* symmetry because the equiprobable contours form ellipsoids in the space of the support.

3.5.1. The scaling function

As before, to solve frictionless asset allocation problems with negative exponential utility, we compute the moment generating function of the distribution, differentiate it w.r.t. the holding, $\boldsymbol{h}_t$, and solve for the root of the gradient. The only difference in the multivariate case is the need to compute the vector gradient $\boldsymbol{\nabla}$ rather than ordinary derivatives. For distributions with ellipsoidal symmetry determined

[4]Another method is to use copulas [50].
[5]They are equal for the normal distribution.

by the matrix Σ_t, and expectation $\mathbb{E}_{t-1}[\boldsymbol{r}_t] = \boldsymbol{\alpha}_t$, the moment generating function is proportional to the function

$$\psi_{\frac{n}{2}}(\boldsymbol{k}) = \frac{e^{-\boldsymbol{k}^T\boldsymbol{\alpha}_t}}{\Delta(\boldsymbol{k})^{\frac{n}{2}-1}} \int_0^\infty f(g^2) I_{\frac{n}{2}-1}(g\Delta(\boldsymbol{k}))\; g^{\frac{n}{2}}\, dg, \tag{3.27}$$

where $\boldsymbol{k} = \lambda \boldsymbol{h}_t$, $\Delta(\boldsymbol{k}) = \sqrt{\boldsymbol{k}^T\Sigma_t\boldsymbol{k}}$, n is the number of risky assets, and $I_\nu(x)$ is the modified Bessel function of the first kind.[6]

The holding function is the root of the gradient of equation 3.27, which can be shown to be the solution of

$$\lambda \boldsymbol{h}_t \Psi_{\frac{n}{2}}(\lambda\Delta(\boldsymbol{h}_t)) = \Sigma_t^{-1}\boldsymbol{\alpha}_t, \tag{3.28}$$

where

$$\Psi_\nu(x) = \frac{1}{x}\frac{\int_0^\infty f(g^2) I_\nu(gx)\, g^{\nu+1}\, dg}{\int_0^\infty f(g^2) I_{\nu-1}(gx)\, g^{\nu}\, dg}. \tag{3.29}$$

For the case of the multivariate normal distribution $\Psi_\nu(x) = 1$ for all values of its arguments and so equation 3.28 is solved by the Markowitz portfolio.[7] For other ellipsoidal distributions, $\Psi_\nu(x)$ acts to scale the solution in a manner that is sensitive to the behavior of the distributions of returns in its tails and acts to *reduce* the size of the investment for large $\boldsymbol{\alpha}_t$.

3.5.2. The inverting function

The value of $\lambda\Delta(\boldsymbol{h}_t)$ at the solution is the value $\hat{x}_t$ that solves

$$\hat{x}_t = x \;:\; x\Psi_{\frac{n}{2}}(x) = \sqrt{\boldsymbol{\alpha}_t^T\Sigma_t^{-1}\boldsymbol{\alpha}_t}. \tag{3.30}$$

For arguments where the root exists, we define the "inverting" function $\Phi_\nu(x)$ to be the root with respect to y of $y\Psi_\nu(y) = x$. Thus

$$\hat{x}_t = \Phi_\nu\left(\sqrt{\boldsymbol{\alpha}_t^T\Sigma_t^{-1}\boldsymbol{\alpha}_t}\right) \tag{3.31}$$

[6] For the derivation of this expression, see Ref. [17].

[7] Within a factor of 2 that may be readily absorbed into the definition of λ.

and the value to be used in equation 3.28 on the facing page is

$$\Psi_{\frac{n}{2}}\left(\Phi_{\frac{n}{2}}\left(\sqrt{\boldsymbol{\alpha_t}^T\Sigma_t^{-1}\boldsymbol{\alpha}_t}\right)\right). \tag{3.32}$$

In general, this function must be evaluated numerically.

3.5.3. The holding function

The holding function is then given by

$$\boldsymbol{h}(\boldsymbol{\alpha}_t) = \frac{\Sigma_t^{-1}\boldsymbol{\alpha}_t}{\lambda\Psi_{\frac{n}{2}}(\hat{x}_t)}, \tag{3.33}$$

and the expected return on that portfolio is

$$\mathbb{E}_{t-1}[\hat{\boldsymbol{h}}_t^T\boldsymbol{r}_t] = \frac{\boldsymbol{\alpha}_t^T\Sigma_t^{-1}\boldsymbol{\alpha_t}}{\lambda\Psi_{\frac{n}{2}}(\hat{x}_t)}. \tag{3.34}$$

This expression has several interesting properties:

(i) The optimal portfolio is always proportional to the portfolio $\Sigma_t^{-1}\boldsymbol{\alpha}_t$ which, as we've seen, is proportional to the solution of Markowitz's mean–variance optimization problem [39].

(ii) The dependence on the absolute risk aversion rate, λ, is a simple inverse scaling, which means that all investors with access to public information will be interested in obtaining the same portfolio in some proportion and so a "market portfolio" can exist with these distributions and a C.A.P.M. style model will be constructable.

(iii) The denominator in equation 3.34 is a dynamic function of $\boldsymbol{\alpha}_t^T\Sigma_t^{-1}\boldsymbol{\alpha}_t$ and so itself is a stochastic process. Therefore it is *not possible* to claim that the Markowitz solution is appropriate for all ellipsoidal distributions, with the variation in investment scale "swept" into the definition of λ and this functional dependence ignored in practice [48].

(iv) Although written as subscript in these formulæ because it is not viewed as a dynamic variable, the exact solutions do have a functional dependence on the number of assets, n, traded in the portfolio.

3.6. The Generalized Error Distribution

3.6.1. The distribution

The results of Section 3.5 demonstrate that the characteristics of the solution to trading with a Laplace distribution extend into more general multivariate forms. To make the results less abstract, we can look at them in the context of a multivariate version of the probability distribution that does provide a good fit to financial markets data, the generalized error distribution[18]:

$$f(\boldsymbol{r}_t|\boldsymbol{\alpha}_t,\Sigma_t,\kappa)=\frac{1}{\sqrt{\pi^n|\Sigma_t|}}\frac{\Gamma(1+\frac{n}{2})}{\Gamma(1+n\kappa)}\left\{\frac{\Gamma(3\kappa)}{\Gamma(\kappa)}\right\}^{\frac{n}{2}}e^{-\left\{\frac{\Gamma(3\kappa)}{\Gamma(\kappa)}\Delta^2_{\Sigma_t}(\boldsymbol{r}_t,\boldsymbol{\alpha}_t)\right\}^{\frac{1}{2\kappa}}}. \tag{3.35}$$

This is a symmetric distribution with support $\mathbb{R}^n$, mean $\boldsymbol{\alpha}_t$, and covariance matrix

$$\mathbb{V}_{t-1}[\boldsymbol{r}_t]=\frac{\Gamma\{(n+2)\kappa\}\Gamma(1+\kappa)}{\Gamma(3\kappa)\Gamma(1+n\kappa)}\Sigma_t. \tag{3.36}$$

As before, $\Delta_{\Sigma_t}(\boldsymbol{r}_t,\boldsymbol{\alpha}_t)$ is the Mahalanobis distance. The parameter κ controls the kurtosis of the distribution and $\kappa=1/2$ recovers the normal distribution.

3.6.2. The scaling function

Using the definition of equation 3.35, we may write down an explicit form for the scaling function, equation 3.27

$$\Psi_\nu(x)=\frac{1}{x}\frac{\int_0^\infty e^{-\eta g^{\frac{1}{k}}}I_\nu(gx)g^{\nu+1}\,dg}{\int_0^\infty e^{-\eta g^{\frac{1}{k}}}I_{\nu-1}(gx)g^{\nu}\,dg}\quad\text{where }\eta=\left\{\frac{\Gamma(3\kappa)}{\Gamma(\kappa)}\right\}^{\frac{1}{2\kappa}}. \tag{3.37}$$

Both of the integrands in equation 3.37 contain a modified Bessel function factor and this function converges to $e^{gx}/\sqrt{2\pi gx}$ for large gx (Gradsteyn [22], p. 909). The rate of convergence depends on the order, ν, of the Bessel function but is true for all orders. This means that this Bessel function factor generally leads to exponential divergence of the integral. However, the divergence may be controlled by the exponential term arising from the p.d.f. as this is a convergent factor. Specifically, if $gx-\eta g^{1/\kappa}>0$, then the integral will diverge exponentially, and if this term is negative, then the integral will

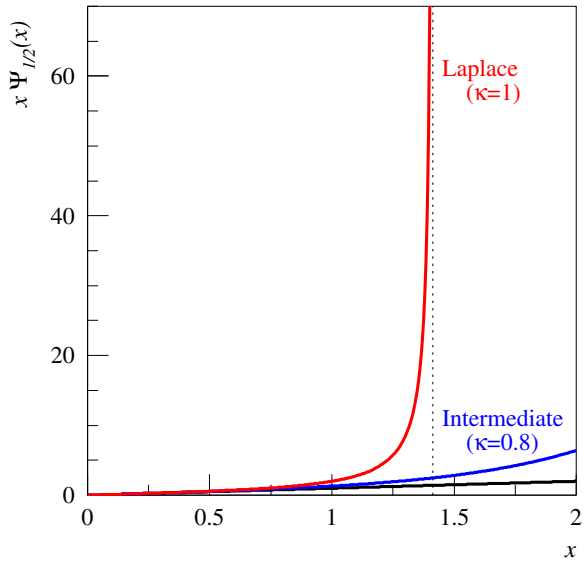


Figure 3.6. Behaviour of the scaling function of the generalized error distribution $x\Psi_{1/2}(x)$ for $\kappa = 0.5, 0.8$, and 1.

converge exponentially. Therefore, we can conclude that the integrals will converge for all $0 < \kappa < 1$ and will converge for $\kappa = 1$ (Laplace distribution) if $x < \sqrt{2}$ but will diverge otherwise. This behavior is illustrated for the univariate distribution ($\nu = 1/2$) in Figure 3.6.

3.6.3. The inverting function

The sharp divergence illustrated for the Laplace distribution as $x \to \sqrt{2}$ has practical consequences for the computation of the inverting function, $\Phi_\nu(x)$. For the "regular" distributions (i.e. $0 < \kappa < 1$), $\Phi_\nu(x)$ is an unbounded increasing function of x. As $\kappa \to 1$, the function converges towards the value for the Laplace distribution, but it is never bounded above. For $\kappa = 1$, the function possesses an asymptote to $\sqrt{2}$ and is bounded below that level. This function is illustrated in Figure 3.7 on the next page.

3.6.4. The holding function

In the following, I will write x for the Mahalanobis distance of the mean, $\Delta_{\Sigma_t}(\boldsymbol{\alpha_t}, \mathbf{0}) = \sqrt{\boldsymbol{\alpha}_t^T \Sigma_t^{-1} \boldsymbol{\alpha}_t}$. This is motivated by the factor

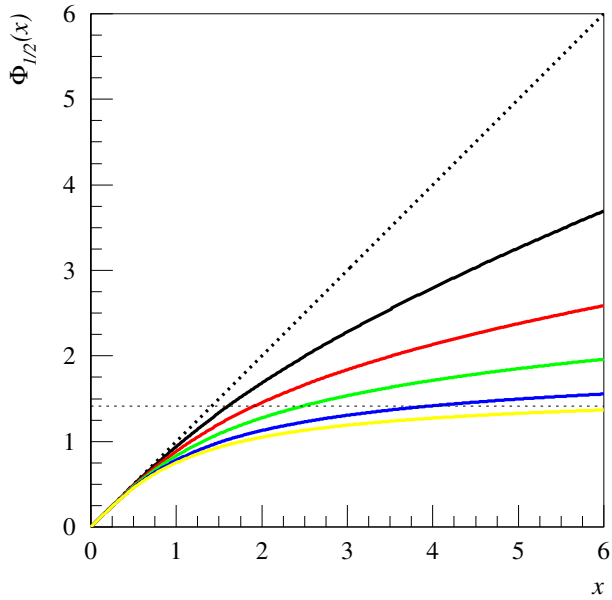


Figure 3.7. Behaviour of the inverting function $\Phi_{1/2}(x)$ as $\kappa \to 1$. The dotted diagonal line represents the normal distribution theory $\Phi_\nu(x) = 1$ and the dotted horizontal line shows the upper bound $\Phi_{1/2}(x) < \sqrt{2}$ for $\kappa = 1$.

that, with this definition, equation 3.30 on page 64 gives[8]

$$\hat{x}_t \Psi_\nu(\hat{x}_t) = x \quad \text{and} \quad \lim_{\kappa \to \frac{1}{2}} \hat{x}_t = x \tag{3.38}$$

and that this x is the argument to the inverting function $\Phi_\nu(x)$.

The holding function for non-normal distributions is computed by scaling the normal distribution-based solution $\Sigma_t^{-1}\boldsymbol{\alpha}_t/\lambda$ by the factor $1/\Psi_{n/2}(\hat{x})$. This scaling factor is illustrated in Figure 3.8 on the next page and shows that for $\kappa > 1/2$ the optimal portfolio is never as heavily invested as the normal distribution would require and is progressively less invested as risk/reward metric increases. The reason for this *scaleback* is clearly shown in Figure 3.9 on the facing page. Here the expected portfolio return (assuming $\lambda = 1$) is plotted

[8]Hopefully this is not too confusing to the reader; it's a notation I've used in private work for over 20 years.

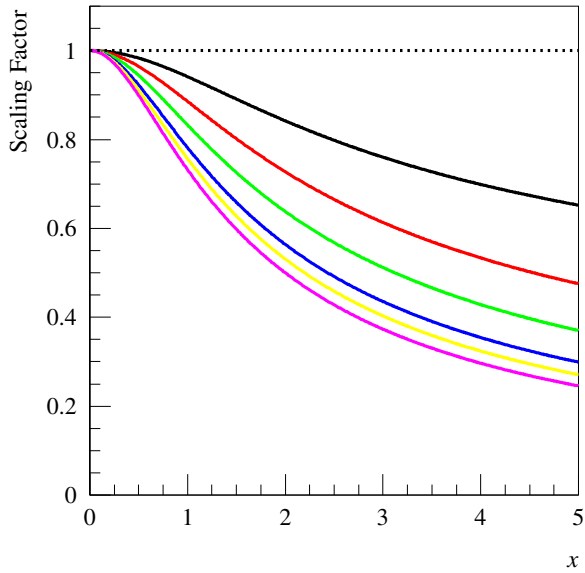


Figure 3.8. Portfolio scaling factors $1/\Psi_{1/2}\{\Phi_{1/2}(x)\}$ for a single asset as $\kappa \to 1$. The dotted line represents the normal distribution theory.

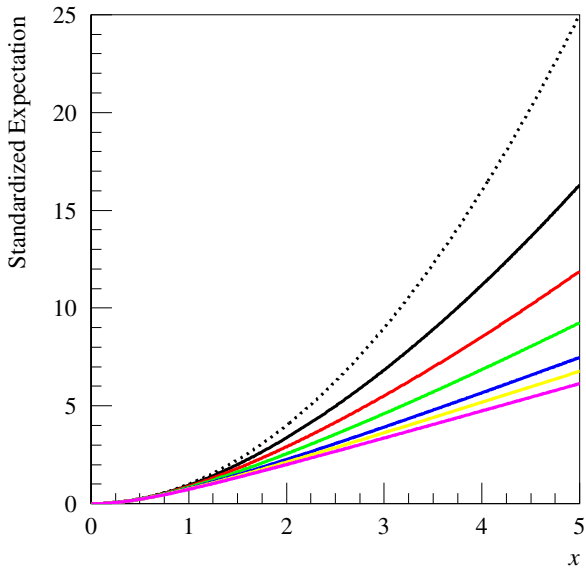


Figure 3.9. Standardized portfolio expected return $x^2/\Psi_{1/2}\{\Phi_{1/2}(x)\}$ for a single asset as $\kappa \to 1$. The dotted line represents the normal distribution theory.

as $\kappa \to 1$. We see that for normal distributions, the expected portfolio return is a quadratically increasing function of the risk/reward metric leading to heavy bets on large expected relative returns. These bets then dominate the profit stream from trading the asset. For non-normal distributions, these bets are dramatically curtailed, due to the progressively less "interesting."[9]

For the normal distribution ($\kappa = 1/2$), a "5σ" expected return is very significant, and the trader's response is to make a heavy bet in those circumstances. For the Laplace distribution ($\kappa = 1$), such an expected return is much less significant and the trader in fact makes a smaller bet on the return nature of high risk/reward portfolios. We also see that a trader that implemented the normal distribution theory-based portfolio in a more leptokurtotic market could be making a substantial overallocation of risk to reward and dramatically increasing their risk of ruin.

3.7. Conclusions

The work presented here should not be taken to indicate that the construction of a holding function via the utility principal, with negative exponential utility and the Laplace or generalized error distribution, is necessarily exactly "correct." The key takeaway is that the introduction of significant kurtosis, which brings us closer to a durably realistic representation of market returns, in fact, has a significant effect on trading strategy computed via a method that is consistent with canonical theory. This causes the trader to favor a holding function that looks more like a "root Kelly," or even a step function, than the linear response model of Markowitz's mean–variance optimization or the maximum geometric growth of the pure Kelly strategy.

Most of the methods studied to this point in this book sit somewhere towards the lower left quadrant of the schematic shown in Figure 3.10 on the next page, whereas we seek methods that fit in the upper right corner. None of the methods considered in this essay feature transaction costs. In Essay 5, holding functions constructed explicitly as step functions will be examined in more detail, and these

[9]By "interesting," we are talking about the nominal statistical significance of the risk/reward metric.

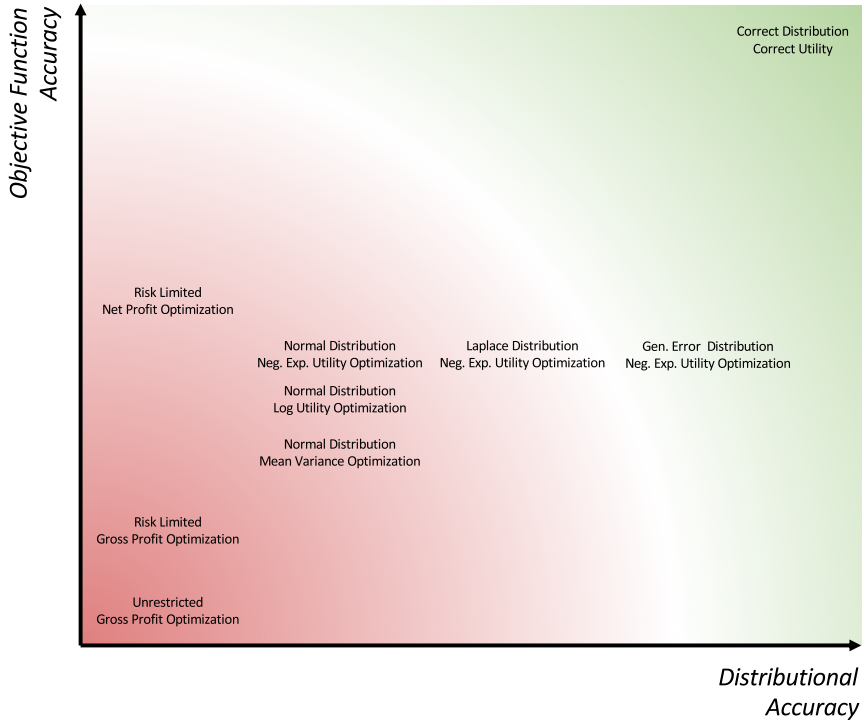


Figure 3.10. The "magic quadrants" of trading strategy theory. Most of the methods that are tractable are towards the lower left of the diagram, whereas we seek to be in the upper right.

allow us to include costs. Before doing that fairly intensive work, however, I will review the thought processes that I went through to figure out how to tackle it in Essay 4.

This page intentionally left blank

Essay 4

Thinking about How to Solve Trading Problems

4.1. The Multiverse of Counterfactuals and Ergodicity

4.1.1. The multiverse in time-series analysis

Time-series analysis carries with it a mathematical philosophy that is remarkably curious. By definition, a time series is a sequence of numbers

$$\ldots, a_{t-2}, a_{t-1}, a_t, a_{t+1}, a_{t+2}, \ldots \tag{4.1}$$

that are ordered in time. It becomes especially interesting, though, when we assert that the numbers are random and drawn from associated distributions $a_t \sim f_t$. This statement is that *all* of the values are randomly drawn from their own distributions and it is not much of an intellectual leap to see that we are really describing a single object with the property that

$$\boldsymbol{a} = \begin{pmatrix} \vdots \\ a_{t-2} \\ a_{t-1} \\ a_t \\ a_{t+1} \\ a_{t+2} \\ \vdots \end{pmatrix} \sim f_{\boldsymbol{a}}, \tag{4.2}$$

where $f_{\boldsymbol{a}}$ represents the *joint* distribution of all possible histories.

The odd thing about such collections of histories,

$$\{\boldsymbol{a}_i : \boldsymbol{a}_i \sim f_{\boldsymbol{a}}\}, \tag{4.3}$$

is that, at a given time t, we have only seen *part of one* of many possible values of $\boldsymbol{a}$,[1] which can be written as

$$P_t\boldsymbol{a}_i = \{a_{is} : s \leq t\}, \tag{4.4}$$

for some "projection matrix," P_t, and particular "history," i. Here $\{a_{it}\}$ represents all values of the time series for any time in any particular history. These other possible histories,

$$\{P_t\boldsymbol{a}_j : j \neq i\}, \tag{4.5}$$

are called *counterfactuals* [54] because their existence is counter to the facts of our observations.

As traders, we are interested in what the *rest* of this $\boldsymbol{a}_i$ might be, which we could label as

$$\bar{P}_t\boldsymbol{a}_i = \{a_{iu} : t < u\} \tag{4.6}$$

because inference about this quantity may lead us to successful investment strategies. In reality, we might only be interested in the *near future*:

$$P_v\bar{P}_t\boldsymbol{a}_i = \{a_{iv} : t < u \leq v\}, \tag{4.7}$$

as the far future does not affect our decision-making processes. Oddly, we are not particularly interested in the counterfactual futures,

$$\{\bar{P}_t\boldsymbol{a}_j : j \neq i\}, \tag{4.8}$$

apart from the what this set tells us about what $\bar{P}_t\boldsymbol{a}_i$ itself might be for *future* times $u > t$. Thus, the analysis of time series can quite rapidly lead us to consider a the existence of a *multiverse* of counterfactual "possible pasts"[2] and probable futures.

[1] This discussion is influenced by the description of time series given by Clive Granger in his Nobel Prize lecture [23].

[2] This term is due to Roger Waters [61].

4.1.2. The multiverse in estimation

In general, if we know or assert some form for $f_{\boldsymbol{a}}$, the unconditional distribution of $\boldsymbol{a}$, then the subject of statistics is very clear on how to proceed. We draw independent samples of these vectors, $\{\boldsymbol{a}_i\}$ for $i \in [1, N]$, and subject to useful assumptions about the nature of $f_{\boldsymbol{a}}$, can estimate its properties from that sample with a precision that, generally, scales as $1/\sqrt{N}$. That knowledge allows us to replace the unknown elements of $P_t\boldsymbol{a}_i$ with their expectations and proceed in making decisions based upon that *expected* future. This theory is all very well developed in standard texts.

However, we cannot do this! We have access to just part of one sample, $P_t\boldsymbol{a}_i$, and no access to the rest of the sample of values $\{P_t\boldsymbol{a}_j\}_{j\neq i}$. These hypothetical, or *counterfactual*, histories are unknowable and so any inference about the future values of the $\boldsymbol{a}_i$ that we will see, such as

$$\left.\begin{array}{c}\mathbb{E}[P_u\bar{P}_t\boldsymbol{a}_i|P_t\boldsymbol{a}_i]\\ \mathbb{V}[P_u\bar{P}_t\boldsymbol{a}_i|P_t\boldsymbol{a}_i]\\ \text{etc.}\end{array}\right\} \quad \text{for } t<u, \tag{4.9}$$

rely on a sample of size one! The only way, it seems, to make any progress in time-series analysis is to add an additional assumption.

4.1.3. Ergodicity in estimation

The whole set of values, $\{a_{it}\}$, can be seen to be elements of a matrix, A, which contains not only the observed history, say $P_{it}A$ for projection matrix P_{it} selecting history i at time t, but also all other counterfactual values for all of its possible pasts and probable futures. If it is true, this necessary assumption, the property of *ergodicity*, is the assumption that the properties of $f_{\boldsymbol{a}}$ estimated from a projection $P_{it}A$ agree in expectation with those that might be estimated from the entire counterfactual matrix A. That is, if $\boldsymbol{\theta}$ are the population values of a set of statistics of interest, then we assume that

$$\mathbb{E}[\hat{\boldsymbol{\theta}}|P_{it}A] = \mathbb{E}[\hat{\boldsymbol{\theta}}|A] = \boldsymbol{\theta}. \tag{4.10}$$

In words this means that the estimates we make from our sole observed sample, $P_{it}A$, are expected to equal the estimates we could

potentially make from the *entire* counterfactual history and, therefore, equal to the population values themselves (since the entire counterfactual history is *conceptually* the population itself).

4.1.4. Counterfactuals in trading strategy analysis

As abstract, and potentially irrelevant, as this discussion on the *philosophy* of time-series analysis may seem, this world-view is a critical part of the math that must be done to solve trading strategy problems. We must consider the counterfactual distributions not just of returns but of alphas and positions as well, and we must assume that what we know of these distributions from a single trading history is valid for all possible futures.

4.2. Traders' Decision-Making Processes

If one possesses information about the future distribution of returns, and the desire to profit from that information, then the action to take is to enter the market and acquire a position sensitive to the difference between the current price of an asset and the expected future price. The optimal strategy is to *trade in the direction of the alpha*. The difference between a "trader" and somebody who has made "an investment" is that the trader expects

(i) to close their position a some future time,
(ii) to re-enter a position in the same asset subsequent to that time,
(iii) that this new position may not match the prior one,
(iv) that they will have to pay some kind of fee to change position,
(v) that their information may potentially not be sufficient for profit,
(vi) that they will repeat these actions time and time again.

Items (i)–(iv), in this list, are a reflection of the fact that, in the real world, *transaction costs* exist. To a trader, meaning somebody who repetitively buys and sells the same security for potential profit, this means that a price must be paid for changing one's mind about the future distribution of returns, and that price is a friction that impacts the decision process. Of course we might, naïvely, suggest that an antidote to transaction costs might be to have a more

accurate alpha and thereby avoid the concept of "changing one's mind" about the optimal position to hold. However, even with perfect information, an observed fact about markets is that they do not trade in the same direction every single day. So even a trader with perfect information, even an oracle, will be changing positions from time to time and generating transaction costs as a tax, or "friction," on this activity. Thus all traders subject to transaction costs must consider whether their alpha exceeds the value necessary to pay for these costs and, potentially, not trade when the alpha is insufficient. This is why we describe costs as a friction because they create an incentive that opposes changing positions.

Item (v) reflects the fact that traders are not oracles: that markets may be correctly modeled *in counterfactual expectation* but that the specific experienced outcomes differ from those values. Bluntly put, this is the statement that trading is a *risky* venture and that actual profits and losses differ from expectations. However, there is more risk than just that the realized return not match its expectation, there is also the fact that the alpha may change unpredictably. To examine this scenario, I often fall back to the simple autoregressive model for returns as a guide to thought experiments.

4.2.1. A simple model for returns

Consider the process

$$r_t = \varphi r_{t-1} + \varepsilon_t \tag{4.11}$$

where the innovation, ε_t, has unconditional variance $\mathbb{V}[\varepsilon_t] = \sigma^2$. For this model, it is straightforward to show that

$$\mathbb{V}[r_t] = \frac{\sigma^2}{1-\varphi^2} \tag{4.12}$$

and that the alpha, $\alpha_t = \varphi r_{t-1}$, has $\mathbb{V}[\alpha_t] = \varphi^2 \mathbb{V}[r_t]$. Stationarity requires that $|\varphi| < 1$, and the system is seen to have an I.C. of

$$R = \sqrt{\frac{\mathbb{V}[\alpha_t]}{\mathbb{V}[r_t]}} = |\varphi|. \tag{4.13}$$

If I, very generously, give this system an I.C. of 10%, we see that $R^2 = 1\%$, which means that 99% of the variance of returns is *not*

due to the alpha. A more realistic I.C., of maybe 2%, gives an R^2 of 0.0004. In practical terms, essentially *none* of the variance of returns is due to the alpha. Thus, even though the constructed system is autoregressive, with realistic values for the I.C. each day's alpha is essentially a *surprise* to the trader!

Now consider a trend follower, with $\varphi = 0.02$. The alpha is known deterministically at trade time and the optimal strategy, *in the absence of transaction costs*, seems to be to trade in the direction of the alpha which is to trade in the direction of today's return when observed at the close of the day's trading. Tomorrow, however, there is close to a 50:50 chance that the position we inherit from today will the be wrong one and, chasing that alpha, there is essentially a 50:50 chance that the position the day after will again require a reversal in trade direction. This stochastic alpha has caused a "whipsaw" and, for a trader subject to transaction costs, money will have been spent to needlessly reverse the position.

4.2.2. The effect of stochastic alphas on optimal strategy

It's clear to see from this toy model that not only should a trader be averse to transaction costs but they should also be averse to *volatility* in the alpha *if* there are consequences to reversing positions frequently and transaction costs, including opportunity costs, are the most obvious of such consequences. A trader who ignores this term will likely trade too frequently and often reverse themselves, leading to sub-optimal performance metrics. This is one of the ways in which the decision-making process for a trader differs from the single horizon investment strategy considered by Markowitz [39] and it shows that repeated decision-making with stochastic signals can result in an optimal policy that differs from static decision-making based on a single, conditionally deterministic, alpha.

4.3. Lattice Methods

4.3.1. Valuing options on a lattice

The familiar algorithm used to value an American Call Option involves the construction of a binary tree where we model each possible future as representing two distinct states: one in which the

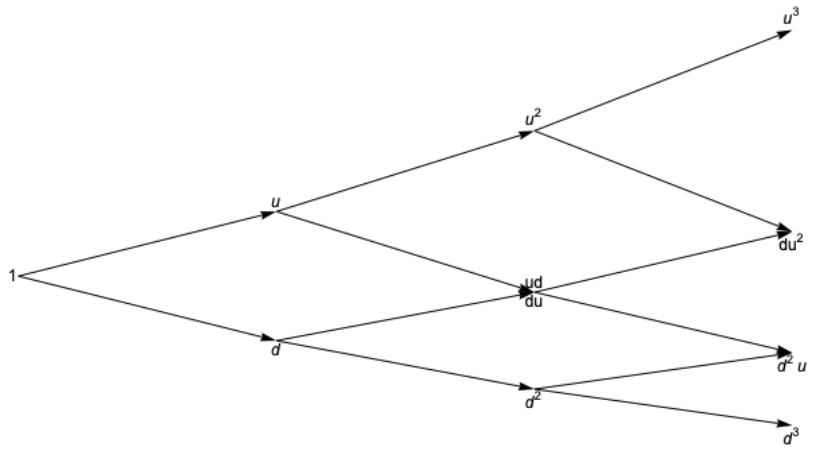


Figure 4.1. Recombining binomial tree used for option valuation by Cox, Ross, and Rubenstein.

asset price went up and one in which the asset price went down. The returns associated with these states are chosen so that the constructed lattice recombines and risk-neutral valuation is used to require that the expected return equal the risk-free rate. This work was pioneered by Cox, Ross, and Rubenstein [27]. Such a lattice is illustrated in Figure 4.1.

The valuation approach is to place a "payoff" value on each node of the lattice, with the final payoff equal to the intrinsic value of the option on the settlement date and thus known unambiguously for each final node, and to establish a valuation at prior nodes by "back propagation," which means making each connected node have a value equal to the conditional expectation of the two payoffs that may be reached from it. Thus the value of the option at the initial node is computable iteratively and the procedure completes. This algorithm has proved to be an extremely useful method for the numerical valuation of complex options.

4.3.2. Evaluating trading strategies on a lattice

Why can't this approach be used to assess the value extracted from a given trading strategy, which is then optimized to deliver the best algorithm?

There are two principal problems attempting to use this method to solve for optimal trading strategies:

(i) A recombining lattice can only be constructed for the special case of constant expected drift in asset prices. This excludes autocorrelation, and many other alphas, as well as the affect of transaction costs, which are all path-dependent.
(ii) A successful trader would never terminate their activity on a given future date — they would trade *perpetually*.

Figure 4.2 illustrates the kind of non-recombining binomial tree that arises from a simple autoregressive system, such as that of equation 4.11 on page 77. One potential approach to fixing this problem might be to ignore the alpha when constructing the pricing lattice, which is just the already introduced assumption that the alphas are small relative to the variance of the traded asset. We could then treat both the alpha and the transaction costs as they were "dividends," payable on each node, and compute the expected payoff of a the strategy this way. Unfortunately, the problem with this approach is that both alphas and transaction costs are path-dependent and so we cannot establish what the value of the dividend on each node ought to be since it can be approached from multiple prior directions

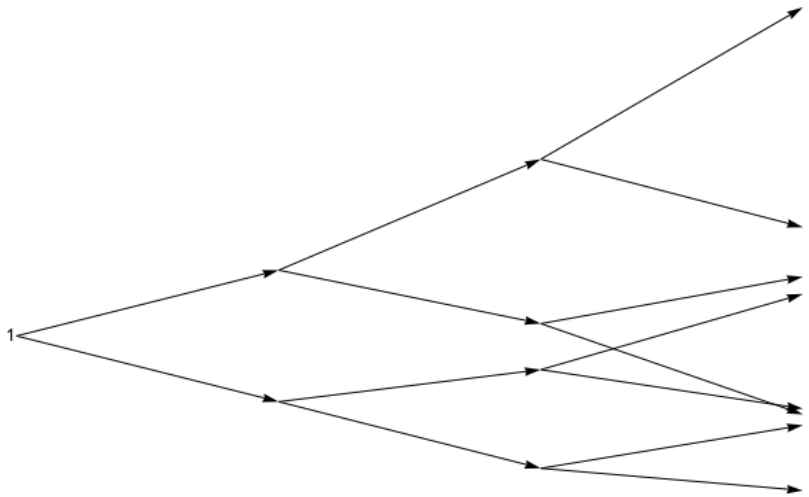


Figure 4.2. Non-recombining binomial tree for a returns process that includes momentum.

and each path through the lattice will ultimately deliver a different "dividend" to be paid at a single node.

4.3.3. Ergodicity and lattice valuation

The requirement that a non-recombining lattice be used, though, is merely a computational complexity. A bigger problem is the lack of a terminal layer to the lattice, since a trader in possession of a valid method for predicting returns and a useful procedure for capturing that value would never terminate their trading. It should be obvious that constructing a pricing lattice involves enumerating all of the potential futures, $\{a_{ju}\}_{u>t}$, that are accessible from the initial time-series state a_{it}, and labelling them with both their value and the probability of them arising given that initial state. However, this labelling cannot be done for a lattice that is infinite in extent i.e. one for which $u \to \infty$, and so the procedure again fails to be usable in reality.

The assumption of ergodicity would suggest that $f_{ju} \to f \; \forall \; j, u$, for unconditional distribution f and all j, as $u \to \infty$ i.e. that the distribution at future time slice u approaches the unconditional distribution as u becomes very large. For an ergodic process, the future becomes progressively disconnected from current conditions and the path dependence may be effectively ignored. This allows us to prune our ever expanding lattice at some $u > t$ and write on those nodes probabilities from the unconditional distributions for key variables, rather than the conditional distributions.

4.3.4. Summary

Although this section might suggest that I see a role for lattice methods in solving for optimal trading strategies, it's not actually as easy as that. The distribution enumerated has to be done in the three dimensions of return, alpha, and position, and the lattice constructed is four-dimensional, non-recombining, and large. I have, in general, just used these lattice ideas as a guide to understand how the mechanics of the process of solving for strategies works, rather than as something that might be executed. Computers are now *much faster* than they used to be, so perhaps there is value to returning to this methodology.

In my mental view of this process, I'm also reminded of solving partial differential equations. In general, the solution to a P.D.E. involves joining the effect of initial conditions and boundary conditions to "free space" solutions. The initial conditions can be thought of a "shocks" that diminish in their effect as that effect is propagated away by the free motion of the medium under discussion. To solve trading strategies, the conditional distributions, built from the initial conditions of alpha, position, and price, are transformed into the unconditional distributions as we look into the fog of the future. In this analogy, the role of Green's function, or a "propagator" of some kind, would be taken by Bayes' theorem.

4.4. Partial Autocorrelation

Time-series analysis is about examining the relationship between measured data and prior values of both that data itself and other series. The "similarity" between data is often addressed through the concepts of covariance and correlation:

$$\sigma_{XY} = \mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \tag{4.14}$$

$$\rho_{XY} = \mathrm{Cor}(X, Y) = \frac{\sigma_{XY}}{\sqrt{\sigma_{XX}\sigma_{YY}}}, \tag{4.15}$$

which I assume are *very familiar* to readers of this essay. The values σ_{XY} being elements of the symmetric positive definite covariance matrix, and the variances are $\sigma_X^2 = \sigma_{XX}$, etc. Such a quantity is referred to as a "product moment" and is discussed extensively in books such as Kendall [30].

4.4.1. Autocorrelation

Autocorrelation is the correlation of a time series with its own historic values. As above, this is built out of the simpler concept of *auto*covariance:

$$\gamma_k = \mathrm{Cov}(a_t, a_{t-k}) = \mathbb{E}[(a_t - \mathbb{E}[a_t])(a_{t-k} - \mathbb{E}[a_{t-k}])] \tag{4.16}$$

$$\rho_k = \mathrm{Cor}(a_t, a_{t-k}) = \frac{\gamma_k}{\gamma_0}. \tag{4.17}$$

This latter step is contingent on the fact that the defining expectations are unconditional and so the two variances in the divisor are equal. In that case, there is no functional difference between the autocorrelation function, ρ_k, and the autocovariance function, γ_k, apart from the scale level γ_0.[3]

4.4.2. Partial autocorrelation

In the definition of correlation given above, *unconditonal* expectations are computed. It is quite valuable, particularly in the context of time-series analysis, to also consider *conditional* expectations of a particular form. The partial covariance and partial correlation can be defined as

$$\omega_{XY|Z} = \mathrm{PCov}(X, Y|Z) = \mathbb{E}[(X - \mathbb{E}[X|Z])(Y - \mathbb{E}[Y|Z])|Z] \quad (4.18)$$

$$\pi_{XY|Z} = \mathrm{PCor}(X, Y|Z) = \frac{\omega_{XY|Z}}{\sqrt{\omega_{XX|Z}\omega_{YY|Z}}}, \quad (4.19)$$

which represent the covariance and correlation between X and Y *controlled* for the effects of Z. It is intuitive, and important, to note that these partial-moments are not unrelated to their unconditional cousins.

It is clear we can also define partial-autocovariance and partial-autocorrelation in a similar manner to the unconditional quantities. However, in time-series analysis, we are particularly interested in a specific formulation. That is, the covariance and correlation between a_t and a_{t-k} corrected for the effects of all of the intermediate values of the series, $\{a_{t-1} \dots a_{t-k+1}\}$. Viewing a time series in the linear additive noise framework, as an accumulation of responses to shocks that take time to dissipate, it's clear that the future values of a series are driven not only by new shocks but also by the lingering effect of prior shocks and, to analyze this data accurately, we need to untangle these responses from each other. Specifically for time series, we can drop the conditioning variable(s) from these expressions and write

$$\omega_k = \mathrm{PCov}(a_t, a_{t-k}|a_{t-1} \dots a_{t-k+1}) = \mathrm{PCov}(a_t, a_{t-k}) \quad (4.20)$$

$$\pi_k = \mathrm{PCor}(a_t, a_{t-k}|a_{t-1} \dots a_{t-k+1}) = \mathrm{PCor}(a_t, a_{t-k}) \quad (4.21)$$

[3]In this common usage, the "functional" dependence is that on the lag index, k.

for the partial covariance and partial correlation functions. These functions represent the covariance between two members of the time-series *after* controlling for the effects due to the intermediate members that sit between them.

4.4.3. Partial autocorrelation and autoregression

In a linear model $Y = \alpha + \beta_X X + \beta_Z Z + \varepsilon$, where Y and Z are independent of each other, it is trivial to show that

$$\beta_X = \frac{\sigma_{XY}}{\sigma_{XX}} \quad \text{and} \quad \beta_Z = \frac{\sigma_{ZY}}{\sigma_{ZZ}}. \tag{4.22}$$

However, when X and Z are dependent on each other, this is not the case. Instead, we have

$$\beta_X = \frac{\sigma_{XY|Z}}{\sigma_{XX|Z}} \quad \text{and} \quad \beta_Z = \frac{\sigma_{ZY|X}}{\sigma_{ZZ|X}}. \tag{4.23}$$

A time series is described as *autoregressive* if one can construct a linear additive noise model in which its values regress onto prior values.[4]

$$a_t = \sum_{i=1}^{p} \varphi_i a_{t-i} + \varepsilon_t \quad \text{for } t > p. \tag{4.24}$$

This may not be the *best* model for the process, meaning that this leads to the least biased and most efficient estimator, but we can *always* construct such a model. Since none of the values of the series within the window of p lags are independent, this model cannot be estimated from the regular *unconditional* regression equations of equation 4.22. Instead we must use the adjusted regression equations of equation 4.23, which *immediately* leads to the identification of the autoregressive lag coefficients φ_i as *identical* to the partial autocorrelations previously defined as follows:

$$\varphi_i = \mathrm{PCor}(r_t, r_{t-i}) = \pi_i. \tag{4.25}$$

[4]This is, of course, the definition of the AR(p) autoregressive time series of order p.

Straightforwardly, equation 4.24 on the facing page may be written down by computing the partial-autocorrelation function of the returns under study. The $\{\varphi_i\}$ may then be literally "read off" the chart.

4.4.4. Partial autocorrelation and autocorrelation

Although this analysis is intended to be intuitively sensible, in my experience, these definitions are potentially quite vague.[5] We may accept the reasoning without having a clear view of *how* to compute the partial covariance matrix for our data. However, this may be analytically realized quite simply and Box and Jenkins provide a simple formula [8].

For an AR(p) process, the vectors and matrix

$$\boldsymbol{\varphi}_p = \begin{pmatrix} \varphi_1 \\ \varphi_2 \\ \vdots \\ \varphi_p \end{pmatrix}, \quad \boldsymbol{\rho}_p = \begin{pmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_p \end{pmatrix}, \quad \text{and}$$

$$\boldsymbol{P}_p = \begin{pmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{p-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{p-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{p-1} & \rho_{p-2} & \rho_{p-3} \cdots & 1 & \end{pmatrix} \tag{4.26}$$

may be defined. The *Yule–Walker equations*, which relate these terms, can be written compactly as

$$\boldsymbol{\varphi}_p = \boldsymbol{P}_p^{-1} \boldsymbol{\rho}_p. \tag{4.27}$$

(This is a standard development covered in most books on time-series analysis, so I will merely just state the result.)

Box and Jenkins then introduce idea of a set of representations of a sample process in terms of AR(k) models, $\{\boldsymbol{\varphi}_k\}$ for $k = 1, 2, \ldots p \ldots$. For this sequence of models, equation 4.27 is solvable algebraically and the kth element of the coefficients vector of

[5] When I first encountered it in Box and Jenkins [8], I found the topic thoroughly confusing. A really good reference is the Wikipedia page [62].

order k, or φ_{kk}, is equal to the π_k partial autocorrelation as defined above. This leads to the following sequence:

$$\pi_1 = \rho_1, \quad \pi_2 = \frac{\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & \rho_2 \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{vmatrix}}, \quad \pi_3 = \frac{\begin{vmatrix} 1 & \rho_1 & \rho_1 \\ \rho_1 & 1 & \rho_2 \\ \rho_2 & \rho_1 & \rho_3 \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho_1 \\ \rho_2 & \rho_1 & 1 \end{vmatrix}} \ldots \tag{4.28}$$

Since the autocorrelation function is *always* readily computable, the partial autocorrelation function may then be obtained numerically. In my experience, few processes in finance have partial autocorrelations that require more than a handful of orders.

4.4.5. The importance of partial autocorrelation to traders

As mentioned above, we can *always* construct an autocorrelation function and a partial autocorrelation function both for the returns we study, $\{r_t\}$, and the alphas used to predict them $\{\alpha_t\}$. The autocorrelation function describes the similarity of a process to its prior values and the partial autocorrelation function delivers a formula to predict these series from their prior values. These predictions may not, necessarily, be the best estimators, but they are always computable so we can consider the behaviors of these time series with the language of Box and Jenkins *even* if we don't actually use those methods to predict returns.

4.5. The Difference Between Static and Stochastic Optimization

Stochastic programming is the discipline of solving optimization problems when inputs to the objective are not known at the decision time. A key feature of the differences between stochastic programming and deterministic programming, such as linear programming or quadratic optimization, is that it is, necessarily, describing how we interact with the future. Systematic trading is an example of stochastic programming. The stochastic elements of the optimization process

become important when actions we take today have an impact on the gains we expect, or the actions we are able to take, in the future. The most obvious way in which trading is subject to this *intertemporal linkage* is when we consider the effect of transaction costs. When I take a position today, I am committing myself to paying to exit that position in the future. For a concrete example, consider that I might be currently long the market but I have an alpha that suggests it will go down. In this circumstance I should short the market, but what if my alpha at the next decision point suggests that it will go up again? Whether I decide to chase the negative alpha or ride it out depends on two factors: how much it costs me to reverse my position relative to my expected gains and how likely I am to want to do that. In the language of traders, is it better to be "whipsawed" by the market or hold the position through the expected losses? To answer the first question requires that I have a model for future returns, an alpha, but to answer the second question depends on how likely my alpha is to change in the future. To understand that, I need to know its autocorrelation function.

4.5.1. Alphas are stochastic

In Essay 2, I emphasize that the alpha itself is a stochastic process and I do this in terms of a, perhaps vaguely defined, information set $\mathcal{I}_t$. This sometimes seems odd, as the alpha we use to make trading decisions is clearly known deterministically. However, when looked at in its entirety, it is a random time series. On a practical level, it's most easy to see this with the AR(p) process.[6] Since

$$r_t = \sum_{i=1}^{p} \varphi_i r_{t-i} + \sigma_t \varepsilon_t \quad \text{for } t > p \tag{4.29}$$

describes the time series, the linear additive noise framework requires that

$$r_t = \alpha_t + \sigma_t \varepsilon_t \qquad \Leftrightarrow \qquad \alpha_t = \sum_{i=1}^{p} \varphi_i r_{t-i} \tag{4.30}$$

[6]And hopefully the discussion of Section 4.4.2 has gone some way to convince the reader that this is always a valid approach, no matter the individual methods in use.

where the $\{\varepsilon_t\}$ are independent random numbers, sometimes called innovations. Following Box and Jenkins [8], equation 4.24 on page 84 can be written in terms of the lag operator, $L : La_t = a_{t-1}$, giving

$$\Phi(L)r_t = \sigma_t\varepsilon_t, \tag{4.31}$$

where $\Phi(L)$ is a pth order polynomial in L with unit constant. Under necessary conditions, there may exist a function, $\Psi(L)$, such that $\Psi(L)\Phi(L) = 1$, permitting equation 4.31 to be written as a moving average process[7] MA(q)

$$r_t = \Psi(L)\sigma_t\varepsilon_t \qquad \Rightarrow \qquad \alpha_t = \{\Psi(L) - 1\}\,\sigma_t\varepsilon_t. \tag{4.32}$$

Since $\Phi(L)$ is, by definition, of the form

$$\Phi(L) = 1 - \sum_{i=1}^{p} \varphi_i L^i, \tag{4.33}$$

it is clear that

$$\Psi(L) = \frac{1}{1 - \sum_{i=1}^{p} \varphi_i L^i} \tag{4.34}$$

which permits expansion via the binomial theorem[8] and may be written as

$$\Psi(L) = 1 + \sum_{i=1}^{\infty} \psi_i L^i, \tag{4.35}$$

in terms of some suitably chosen $\{\psi_i\}$. Therefore,

$$\alpha_t = \sum_{i=1}^{\infty} \psi_i \sigma_{t-i}\varepsilon_{t-i}. \tag{4.36}$$

Thus, α_t is a sum of random innovations and the sum of random numbers is also a random number.[9]

[7] Which may not be of finite order.
[8] Under necessary conditions for the convergence of the sums.
[9] This is all well known in the literature.

4.5.2. Non-autoregressive alphas

Do not take the prior section to mean that I am advocating *solely* for the development of Box–Jenkins style ARIMA models in finance. The purpose of this development is to show that this provides a language to understand the nature of these objects, alphas, that we are constructing from prior information. It is true that all alphas are functions of the prior information set, $\alpha_t = \alpha(\mathcal{I}_{t-1})$, as laid out in Essay 2. Given any alpha constructed in this manner, however, it is *always* possible to consider the relationships between elements of the sequence

$$\alpha_t = \alpha(\mathcal{I}_{t-1}), \alpha_{t-1} = \alpha(\mathcal{I}_{t-2}), \alpha_{t-2} = \alpha(\mathcal{I}_{t-3}) \dots \tag{4.37}$$

and that relationship *is* describable by the partial-autocorrelation function[10]

$$\pi_k = \text{PCor}(\alpha_t, \alpha_{t-k}). \tag{4.38}$$

4.5.3. Forecasting time series

Forecasting time series, or at least forecasting the mean of an autoregressive time series, turns out to be a fairly mundane procedure. Suppose we need to know returns at lead $l > 0$ from an autoregressive model:

$$\begin{aligned}\mathbb{E}_{t-1}[r_{t-1+l}] &= \mathbb{E}_{t-1}\left[\sum_{i=1}^{p} \varphi_i r_{t-1+l-i} + \sigma_{t-1+l}\varepsilon_{t-1+l}\right] \\ &= \sum_{i=1}^{p} \varphi_i \mathbb{E}_{t-1}[r_{t-1+l-i}],\end{aligned} \tag{4.39}$$

since $\mathbb{E}_u[\varepsilon_t] = 0 \; \forall \; u > t$ by definition. Writing

$$\mathbb{E}_{t-1}[r_{t-1+l}] = \begin{cases} \alpha_{t-1,l} & l \geq 1 \\ r_{t-1+l} & l < 1 \end{cases} \tag{4.40}$$

[10] And also the usual autocorrelation function $\rho_k = \text{Cor}(\alpha_t, \alpha_{t-k})$.

for the expected return at lead l computed at time $t-1$, where $\alpha_{t-1,1} = \mathbb{E}_{t-1}[r_t] = \alpha_t$ in our prior notation, $\alpha_{t-1,2} = \mathbb{E}_{t-1}[r_{t+1}]$, etc. The expected return is one of three possible sums:

$$\mathbb{E}_{t-1}[r_{t-1+l}] = \begin{cases} \sum_{i=1}^{p} \varphi_i \alpha_{t-1,l-i} & l > p \\ \sum_{i=1}^{l-1} \varphi_i \alpha_{t,l-i} + \sum_{i=l}^{p} \varphi_i r_{t-i+1} & 1 < l \leq p \\ \sum_{i=1}^{p} \varphi_i r_{t-i} & l = 1. \end{cases} \tag{4.41}$$

Such an expression seems challenging, but it is easier to understand by examining an specific example. Consider an AR(2) model, for leads $l = [1, 2, 3]$, then

$$\begin{aligned} \mathbb{E}_{t-1}[r_t] &= \varphi_1 r_{t-1} + \varphi_2 r_{t-2} &&= \alpha_{t-1,1} \\ \mathbb{E}_{t-1}[r_{t+1}] &= \varphi_1 \alpha_{t-1,1} + \varphi_2 r_{t-1} &&= \alpha_{t-1,2} \\ \mathbb{E}_{t-1}[r_{t+2}] &= \varphi_1 \alpha_{t-1,2} + \varphi_2 \alpha_{t-1,1} &&= \alpha_{t-1,3}\,. \end{aligned} \tag{4.42}$$

This can be seen to be computed from a matrix in which the first row, for lead 1, includes $p = 2$ historic returns, in order. The first alpha is the product of that matrix row with a vector composed of the autoregressive coefficients. As the lead is increased, the next row is obtained by right-shifting the elements of the prior row and inserting the previously computed alpha for the unknown return required on the left. This expression clearly illustrates the recursive nature of the forecasting algorithm where prior forecasts, which are computed from historic returns, are fed back into the system and what results is a convolution of the vector of autoregressive coefficients with itself. For example, with the $AR(2)$ structure of equation 4.42, the forecast at lead 2 is

$$\begin{aligned} \mathbb{E}_{t-1}[r_{t+1}] &= \varphi_1(\varphi_1 r_{t-1} + \varphi_2 r_{t-2}) + \varphi_2 r_{t-1} \\ &= (\varphi_1^2 + \varphi_2) r_{t-1} + \varphi_1 \varphi_2 r_{t-2}. \end{aligned} \tag{4.43}$$

4.5.4. Forward forecasts for any alphas

For any real alpha, it's possible to express that alpha and all forward alphas as functions of the most recently available information:

$$\alpha_t = \alpha_{t-1,1} = A_1(\mathcal{I}_{t-1}) = \mathbb{E}_{t-1}[r_t] \tag{4.44}$$

$$\alpha_{t-1,2} = A_2(\mathcal{I}_{t-1}) = \mathbb{E}_{t-1}[r_{t+1}] \tag{4.45}$$

$$\alpha_{t-1,3} = A_3(\mathcal{I}_{t-1}) = \mathbb{E}_{t-1}[r_{t+2}] \tag{4.46}$$

$$\vdots$$

for forecasting functions A_1, A_2, etc. It is the job of the alpha builder to construct these functions, and they may do it in whatever manner they see fit. This might be by Box–Jenkin's ARIMA method, it might be by the application of traditional regression methods, or it might be machine learning of some form, such as deep neural networks or random forests. Irrespective of the method, however, for these alphas to be *useful*, they should map into expected future returns as specified in equation 4.44 *et seq.*

4.5.5. Expected future alphas

As noted at the beginning of this section, a trader's optimal decision depends not just on what they expect returns to be in the future but also on what they expect *alphas* to be in the future. For any particular set of functions, $\{A_k\}$, the law of iterated expectations means that it must be true that

$$\begin{aligned}
\mathbb{E}_{t-1}[\alpha_{t-1+m,1}] &= \mathbb{E}_{t-1}[A_1(\mathcal{I}_{t-1+m})] \\
&= \mathbb{E}_{t-1}[\mathbb{E}_{t-1+m}[r_{t+m}]] \quad \text{(from equation 4.44)} \\
&= \mathbb{E}_{t-1}[r_{t+m}] \quad \text{(from L.I.E.)} \\
&= A_{m+1}(\mathcal{I}_{t-1}) \\
&= \alpha_{t-1,m+1}. \qquad (4.47)
\end{aligned}$$

That is, expected alphas and forward alphas must be consistent with each other. Clearly a development analogous to that of equation 4.47 may be made for forecasts at other leads computed by the functions

A_2, A_3, etc.

$$\mathbb{E}_{t-1}[\alpha_{t-1+m,k}] = \alpha_{t-1,m+k}. \tag{4.48}$$

It is important to note that, again due to the L.I.E., this result *only* applies to conditional expectations of the alpha. Unconditionally,

$$\begin{aligned} \mathbb{E}[\alpha_{t-1+m,k}] &= \mathbb{E}[A_k(\mathcal{I}_{t-1+m})] \\ &= \mathbb{E}[\mathbb{E}_{t-1+m}[r_{t-1+m+k}]] \\ &= \mathbb{E}[r_{t-1+m+k}] \\ &= \mu, \end{aligned} \tag{4.49}$$

where μ is the unconditional expected return.

4.5.6. The variance of future alphas

Following the same development as above, expressions for the variance of future alphas may be obtained:

$$\mathbb{V}_{t-1}[\alpha_{t-1+m,k}] = \mathbb{V}_{t-1}[r_{t-1+m+k}] \tag{4.50}$$

$$\mathbb{V}[\alpha_{t-1+m,k}] = v, \tag{4.51}$$

where v is the unconditional variance of returns.[11]

4.5.7. Weak and increasingly weak alphas

Consider a returns series with partial autocorrelation function $\{\varphi_k\}$. This means that we may write

$$r_t = \sum_{k=1}^{\infty} \varphi_k r_{t-k} + \varepsilon_t \tag{4.52}$$

and, as pointed out earlier, we may do this irrespective of whether we use an ARIMA model for forecasting. By "weak and increasingly

[11] Not written as σ^2 to prevent collision with earlier notations in this essay.

weak alphas" I mean specifically

$$1 \gg |\varphi_i| \gg |\varphi_j| \quad \forall\, i < j. \tag{4.53}$$

With this condition

$$\alpha_{t-1+m,k} \approx \varphi_{m+k-1} r_{t-1}, \tag{4.54}$$

which may be derived from the expressions for forecasting weights developed by Box and Jenkins [8]. It then follows from equation 4.50 on the facing page that

$$\mathbb{V}[\alpha_{t-1+m,k}] \approx \varphi^2_{m+k-1} v \qquad \Rightarrow \qquad R^2_m \approx \varphi^2_m, \tag{4.55}$$

where R^2_m is the proportion of variance at lead m explained by the alpha and defined by analogy to the traditional expression.

4.5.8. Future alphas are only weakly conditional

For such returns, if the partial autocorrelation function decays towards zero as

$$\varphi_m \propto \phi^m, \tag{4.56}$$

for some $|\phi| \ll 1$, then the variance of forward alphas will decay towards zero as ϕ^{2m}. Thus, if alphas are weak and increasingly weak, the forward alpha will vanish rapidly. If, for example, $\varphi_1 = 0.1$, which represents quite a strong momentum in the returns, then $R^2_1 \approx 0.01$, $R^2_2 \approx 0.0001$, etc. The effect of this alpha vanishes almost immediately.

In Section 4.5.5 on page 91, I demonstrated the required consistency between expected forward alphas and expected future returns. Because of this result, it is tempting to think that the problem must be entirely specified in terms of the forward alphas, however that is not the case. Optimal trading strategy often depends on the costs, in risk terms, of keeping a position in defiance of the alpha and the costs, in terms of transaction costs, of chasing a reversal that the trader expects not to persist. The rapid decay of forward alphas for weak and increasingly weak alphas mean that the trader taking account of the volatility of their alpha in the decision-making process does not need to model the conditional distribution of forward alphas as the simpler unconditional distribution will likely be good enough an approximation almost immediately.

4.5.9. Discounting multi-horizon alphas

As a counterpoint to the evanescent, yet perpetually regenerating, alphas characterized by strongly decaying partial autocorrelation functions, another tractable problem is characterized by weak, yet strongly correlated alphas. For systems with long-memory, it may be the case that the alphas themselves are strongly serially correlated but other factors contribute to position exits. Processes such as the *fractionally integrated* process

$$\mathrm{FI}(d) \ : \ (1-L)^d r_t = \varepsilon_t \quad \text{for } 0 < d < 1, \tag{4.57}$$

which delivers an autocorrelation function that decays hyperbolically as $\rho_k \propto k^{-\alpha}$ rather than the $\rho_k \propto \phi^k$ exponential decay windows delivered by standard ARIMA models, or perhaps even something described by the linear discounting of a term premium over a fixed horizon where the premium is insufficient to pay for trade entry. One could consider a scenario in which an exogenous factor might cause trade exit at a future time with a fixed probability per trade interval. The calling of a bond with a sinking fund by its issuer might be one such scenario. In this case, the expected total premium collected is

$$\sum_{s=t}^{T}(1-p^{s-t})\frac{P}{T-t}, \tag{4.58}$$

where p is the probability of trade exit per interval, t is the current time, T is the maturity date of the bond, and P the initial premium.

I do not believe such circumstances represent the majority of cases to be considered by traders in the markets, but I highlight it to demonstrate that *stochastic alphas* are not the only reason why future positions should be treated as uncertain. The two questions that must be considered are as follows:

(i) What is the future expected return of an asset?
(ii) How likely am I to capture it if I enter a position today?

An answer must be developed for both questions and the product of those answers must be compared to the choice of staying flat at any given decision point in order that optimal trading strategies be developed.

Essay 5

Barrier Trading Algorithms

In Essay 3, I commented that the exact holding function for the Laplace distribution can, in some way, be approximated by a step function. Many traders, in practice, eschew "fancy" portfolio construction methods [19] for such simple holding functions, yet little analytical work is presented on their efficiency. In this essay, I will study these methods analytically and present critical results for traders at any scale.

5.1. Optimal Strategy with Stochastic Alphas and Positions

5.1.1. The joint distribution of future returns

In Sections 2.3.5 and 2.3.6 on page 35, we concluded that both the alpha for an asset and the position of a trader in that asset must be regarded as stochastic processes even though they are deterministic inputs to the holding function. In general terms, due to the strict ordering of the information sets $\mathcal{I}_s$ and $\mathcal{I}_t$, for $s < t$, it is not true that the distribution of some $\boldsymbol{y}_t$ is independent of that of some other $\boldsymbol{x}_s$ at a prior time. The distribution of future returns $\boldsymbol{r}_u$, as computed by trader at time s, has a multivariate density that is a function of all pertinent random variables $(\boldsymbol{r}_t, \boldsymbol{\alpha}_t, \boldsymbol{h}_t)$ for $s < t \leq u$. Such a density is, of course, exceedingly complex.

However, due to our assumption of increasingly weakly forecastable returns, we may write

$$f(\boldsymbol{r}_u, \boldsymbol{\alpha}_u, \boldsymbol{h}_u, \boldsymbol{r}_{u-1}, \boldsymbol{\alpha}_{u-1}, \boldsymbol{h}_{u-1}, \dots) \approx f_{\boldsymbol{h}_u}(\boldsymbol{h}_u) \prod_{t \le u} f_{\boldsymbol{r}_t}(\boldsymbol{r}_t|\boldsymbol{\alpha}_t) f_{\boldsymbol{\alpha}_t}(\boldsymbol{\alpha}_t) f_{\boldsymbol{h}_{t-1}}(\boldsymbol{h}_{t-1}) \tag{5.1}$$

where the notation $f_{\boldsymbol{x}}(\boldsymbol{x}|\boldsymbol{y})$ refers to the marginal distribution of $\boldsymbol{x}$ given $\boldsymbol{y}$. This expression is valid for all $t \le u$ and so includes all $t \le s$, where s is the current time, or the "start" of trading. Yet the observed data *up to* time s are known precisely, so their distributions may be represented by the Dirac delta function[1]

$$f_{\boldsymbol{x}_t}(\boldsymbol{x}_t) = \delta(\boldsymbol{x}_t - \boldsymbol{X}_t), \tag{5.2}$$

where $\boldsymbol{X}_t$ is the observed value of random variable $\boldsymbol{x}_t$ for $t \le s$.

5.1.2. Perpetual homogenous trading

In general, we are interested in the limit $u \to \infty$ in the prior and following equations. This represents *perpetual* trading, meaning that the strategy is successful and trading never terminates nor diverges in scale such that the probability of *instant ruin* at any future time t is essentially zero. As this limit removes the possibility of trading being influenced by an end date, u, we therefore seek a policy function that is homogeneous, meaning that the holding function $\boldsymbol{h}(\boldsymbol{\alpha}_t, \boldsymbol{h}_{t-1} \dots)$ is the same for all t between the start of trading at time s and u.

5.1.3. Homogeneous optimization of separable statistics

Optimization of the expected value of a general statistic involves the integral

$$\Omega = \int \dots \int Z(\boldsymbol{h}_s, \boldsymbol{r}_{s+1}\boldsymbol{h}_{s+1} \dots \boldsymbol{r}_u, \boldsymbol{h}_u) f(\boldsymbol{r}_{s+1}, \boldsymbol{\alpha}_{s+1}, \boldsymbol{h}_{s+1} \dots \boldsymbol{r}_u, \boldsymbol{\alpha}_u, \boldsymbol{h}_u) \, d^n\boldsymbol{r}_{s+1} d^n\boldsymbol{\alpha}_{s+1} d^n\boldsymbol{h}_{s+1} \dots d^n\boldsymbol{r}_u d^n\boldsymbol{\alpha}_u d^n\boldsymbol{h}_u. \tag{5.3}$$

[1] This is defined in many books on *Mathematical methods in Physics*, such as the one by Arfken [2].

For a statistic that is separable by addition i.e.

$$\lim_{u\to\infty} Z(\boldsymbol{h}_s, \boldsymbol{r}_{s+1} \ldots \boldsymbol{h}_{u-1}, \boldsymbol{r}_u, \boldsymbol{h}_u) = \sum_{t=s+1}^{\infty} Z_t(\boldsymbol{h}_{t-1}, \boldsymbol{r}_t, \boldsymbol{h}_t), \tag{5.4}$$

Equation 5.3 may be written as follows:

$$\Omega = \sum_{t=s+1}^{\infty} \left\{ \int_{\boldsymbol{r}_t} \int_{\boldsymbol{\alpha}_t} \int_{\boldsymbol{h}_{t-1}} Z_t(\boldsymbol{h}_t, \boldsymbol{r}_t, \boldsymbol{h}_{t-1}) f_{\boldsymbol{r}_t}(\boldsymbol{r}_t|\boldsymbol{\alpha}_t) f_{\boldsymbol{\alpha}_t}(\boldsymbol{\alpha}_t) f_{\boldsymbol{h}_{t-1}} (\boldsymbol{h}_{t-1}) d^n\boldsymbol{r}_t d^n\boldsymbol{\alpha}_t d^n\boldsymbol{h}_{t-1} \right\}. \tag{5.5}$$

Similarly, for a statistic that is separable by multiplication, we get

$$\Omega = \prod_{t=s+1}^{\infty} \left\{ \int_{\boldsymbol{r}_t} \int_{\boldsymbol{\alpha}_t} \int_{\boldsymbol{h}_{t-1}} Z_t(\boldsymbol{h}_t, \boldsymbol{r}_t, \boldsymbol{h}_{t-1}) f_{\boldsymbol{r}_t}(\boldsymbol{r}_t|\boldsymbol{\alpha}_t) f_{\boldsymbol{\alpha}_t}(\boldsymbol{\alpha}_t) f_{\boldsymbol{h}_{t-1}} (\boldsymbol{h}_{t-1}) d^n\boldsymbol{r}_t d^n\boldsymbol{\alpha}_t d^n\boldsymbol{h}_{t-1} \right\}. \tag{5.6}$$

Although apparently different, for the purposes of optimization, both expressions are equivalent as equation 5.6 may be converted to a summation over logs and, since the logarithm is a strictly monotonic increasing function of its argument, the extremum of the log coincides with the extremum of its argument. So, in both cases, the optimal strategy is determined by taking the extremum of

$$\Omega = \sum_{t=s+1}^{\infty} \lambda_t \left\{ \int_{\boldsymbol{r}_t} \int_{\boldsymbol{\alpha}_t} \int_{\boldsymbol{h}_{t-1}} Z_t(\boldsymbol{h}_t, \boldsymbol{r}_t, \boldsymbol{h}_{t-1}) f_{\boldsymbol{r}_t}(\boldsymbol{r}_t|\boldsymbol{\alpha}_t) f_{\boldsymbol{\alpha}_t}(\boldsymbol{\alpha}_t) f_{\boldsymbol{h}_{t-1}} (\boldsymbol{h}_{t-1}) d^n\boldsymbol{r}_t d^n\boldsymbol{\alpha}_t d^n\boldsymbol{h}_{t-1} \right\}, \tag{5.7}$$

where the $\{\lambda_t\}$ are some set of Lagrange multipliers. These multipliers are "undetermined" and may be chosen in such a manner to ensure that equation 5.7 is convergent if necessary. An homogeneous policy can only be achieved when all of the incremental statistics, $Z_t(\boldsymbol{h}_t, \boldsymbol{r}_t, \boldsymbol{h}_{t-1})$, that the total performance statistic is composed from

all have the same functional form, for all t and independent of the specific value of t, and the probability densities are also homogeneous, which is the case when returns are increasingly weakly predictable.

If the objective is convergent, the incremental statistics are homogeneous, returns are increasingly weakly predictable, and the market is not influenced by our position, then equation 5.7 on the previous page will be at an extremum when the holding function, $\boldsymbol{h}(\boldsymbol{\alpha}_t, \boldsymbol{h}_{t-1} \dots)$, is chosen so that[2]

$$\hat{\boldsymbol{h}}(\boldsymbol{\alpha}_t, \boldsymbol{h}_{t-1} \dots) = \underset{\{\boldsymbol{h}_t = \boldsymbol{h}(\boldsymbol{\alpha}_t, \boldsymbol{h}_{t-1})\}}{\arg\max} \left\{ \int_{\boldsymbol{r}_t} \int_{\boldsymbol{\alpha}_t} \int_{\boldsymbol{h}_{t-1}} Z(\boldsymbol{h}_t, \boldsymbol{r}_t, \boldsymbol{h}_{t-1}) \right.$$
$$\left. f_{\boldsymbol{r}_t}(\boldsymbol{r}_t | \boldsymbol{\alpha}_t) f_{\boldsymbol{\alpha}_t}(\boldsymbol{\alpha}_t) f_{\boldsymbol{h}_{t-1}}(\boldsymbol{h}_{t-1})\, d^n \boldsymbol{r}_t d^n \boldsymbol{\alpha}_t d^n \boldsymbol{h}_{t-1} \right\} \tag{5.8}$$

for all $t > s$ provided that the objective Ω exists (i.e. that the summation is convergent). This is a *functional* optimization over the set of potential holding functions $\{\boldsymbol{h}(\boldsymbol{\alpha}_t, \boldsymbol{h}_{t-1} \dots)\}$.

Equation 5.8 is a key result in our trading strategy theory, as it allows a trader to chose the holding function not for a single, fixed horizon, as the methods explored so far do, but for a trader perpetually involved in the business of extracting value from the markets, provided that trader is "small" and markets are "mostly efficient."

5.2. Trading as a Barrier Crossing Process

In Section 3.4.2 on page 57, I showed how the complex holding function of equation 3.20 could be approximated by a step function such as that of equation 3.25. This solution serves to control risk by limiting the size of positions taken and delivers a response function that is an increasing function of alpha by the way the steps are arranged. I refer to such a trading strategy as "barrier trading" because the trade occurs when the alpha, α_t, exceeds the barrier threshold, b.

[2] Arg may clearly be substituted for arg max should the statistic, Z, require it.

5.2.1. Barrier crossing rates

As the alpha is a stochastic process, the role of a trader using such a system is to watch the time evolution of the alpha and enter a long trade when it crosses $+b$ from below. Similarly, the trader goes short when the alpha crosses $-b$ from above and goes flat when it crosses $\pm b$ from more extreme values.

The fundamental law of active management: In their book *Active Portfolio Management,* Grinold and Kahn introduce a theorem they call *The Fundamental Law of Active Management* [24]. This is expressed as

$$Z = R\sqrt{N}, \tag{5.9}$$

where Z is the *Information Ratio*, which is essentially the Sharpe ratio with the risk-free rate replaced by the rate of return of an active manager's benchmark index, R is the *Information Coefficient* or "I.C." which is the correlation between the trader's predictions of asset returns and their realizations, and N is the *Breadth* or the expected number of trades executed. All quantities are annualized or computed from annual data. For a trader, the benchmark is the return on cash, which is essentially zero over the horizons of most trades, the I.C. is a function of the alpha model in use and, in general, if a market is predictable to some extent, the best a trader can do is to achieve an R^2 for that prediction[3] that matches the extent to which returns are predictable. The rate of trading, however, is under the control of the trader and the more (independent) trades done the better.

The barrier crossing rate: If α_t is a set of independent random variables, then the complementary cumulative distribution function, $\bar{F}_{\alpha_t}(b)$, is the probability that α_t exceeds b. i.e.

$$\bar{F}_{\alpha_t}(b) = \int_b^\infty f_{\alpha_t}(\alpha_t)\, d\alpha_t = 1 - F_{\alpha_t}(b). \tag{5.10}$$

[3]Remember that R^2 is the ratio of the variance of the alpha (the prediction) divided by the variance of returns (the dependent variable) and so measures how much of the variance of the returns are explained by the alpha and, for a linear model, the correlation is the square root of the R^2.

If there are P trade opportunities per annum, clearly the breadth of a strategy is[4]

$$N = 2P\bar{F}_{\alpha_t}(b). \tag{5.11}$$

The Mills ratio and the Q-function: The *Mills ratio*, $m_x(x)$, is the ratio $\bar{F}_x(x)/f_x(x)$ and is the inverse of the *hazard rate* of a distribution, which is often discussed in survival analysis.[5] Thus

$$N = 2Pm_{\alpha_t}(b)f_{\alpha_t}(b). \tag{5.12}$$

If the alpha has a standardized normal distribution,[6] then $\bar{F}_x(x)$ is referred as the "Q-function," $Q(x)$, and it is known not to have a simple analytic form. However, various approximations have been developed for the Mills ratio, and a common one is

$$x \sim \text{Normal}(0,1) \quad \Rightarrow \quad m(x) \approx \frac{1}{x} \quad \Rightarrow \quad Q(x) \approx \frac{1}{\sqrt{2\pi}x} e^{-\frac{1}{2}x^2}. \tag{5.13}$$

The accuracy of this approximation is illustrated in Figure 5.1 on the facing page. Clearly, it is fairly good for $x > 2$ and $Q(x)$ is a *strongly* convergent function with the limit $Q(x) \to 0$ as $x \to \infty$.

The barrier crossing rate for the Laplace distribution: In Section 3.4.2 on page 57, I presented an analysis of trading strategy when returns are drawn from a Laplace distribution. If the assets considered have price processes that are autoregressive in nature, then the distribution of the alpha will similarly have a Laplace distribution for $AR(1)$ and, due to the central limit theorem, somewhere between the Laplace and the normal for $AR(p)$, where $p > 1$. The Mills ratio for a standardized Laplace distribution is just the constant $1/\sqrt{2}$ and the breadth for a barrier B, in units of the standard deviation of the alpha, is simply

$$N = Pe^{-\sqrt{2}B}. \tag{5.14}$$

[4]The "2" is because there are barriers at $\pm b$.

[5]The hazard rate gives the proportion of the population that has survived to time T that will die off at that time, which is of interest in the analysis of mortality rates.

[6]Which is likely *not* to be true but is a baseline.

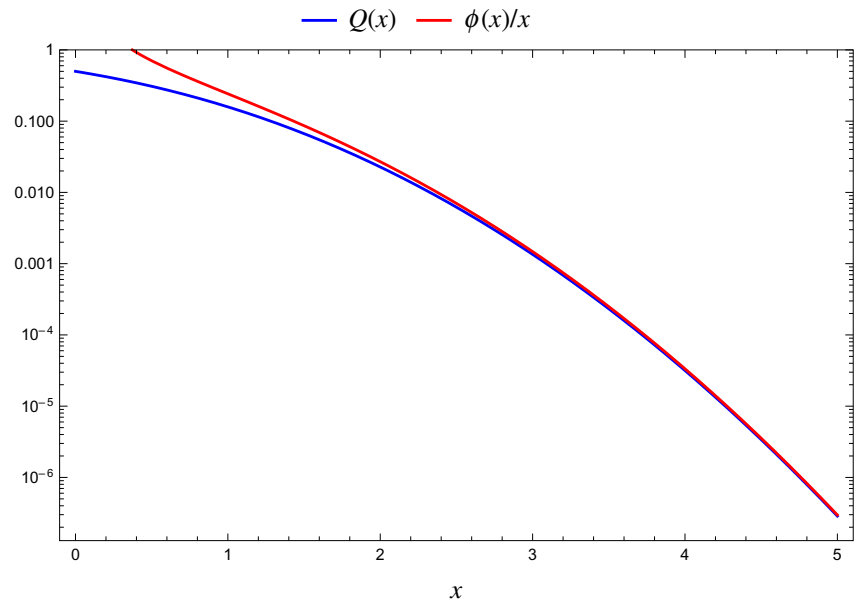


Figure 5.1. The Q-function and its approximation $\phi(x)/x$. $\phi(x)$ is the usual notation for the probability density function of the standard normal distribution.

The barrier crossing rate for the generalized error distribution: With zero mean, the univariate generalized error distribution may be written as follows:

$$\alpha \sim \text{GED}(0, \omega, \lambda) \qquad \Rightarrow \qquad f_\alpha(\alpha) = \frac{e^{-\frac{1}{2}\left|\frac{\alpha}{\omega}\right|^{\frac{1}{\lambda}}}}{2^{\lambda+1}\omega\Gamma(\lambda+1)}, \tag{5.15}$$

where ω and λ are the scale and kurtosis parameters, respectively. With this distribution for the alpha, the breadth for barrier, b, is

$$N = P\frac{\gamma\left(\lambda, \frac{1}{2}\left(\frac{b}{\omega}\right)^{1/\lambda}\right)}{\Gamma(\lambda)}, \tag{5.16}$$

where $\gamma(x, y)$ is the *incomplete* gamma function. For this parameterization, the standard deviation of the alpha is $(2\sqrt{2})\omega$ and this expression is equal to equation 5.14 on the facing page when $\lambda = 1$.

Remarks: For the distributions considered here, the barrier crossing rate of an alpha is an exponentially decreasing function of the

barrier height. This is intuitively reasonable and the linkage between the barrier crossing rate and the Sharpe ratio, according to Grinold and Kahn, shows the importance not only of the accuracy of the alpha but also its variance to a trader who wishes to have a high Sharpe ratio.

5.3. Risk Limited, Frictionless, Barrier Trading

In this section, I will analyze risk limited, frictionless, barrier trading for a single asset. That means the holding function will be of the form

$$h(\alpha_t) = \begin{cases} +L & \alpha_t \geq b \\ 0 & |\alpha_t| < b \\ -L & \alpha_t \leq -b \end{cases}. \tag{5.17}$$

The holding function is not a function of the prior position, h_{t-1}, because as trading is free of costs *any* prior position can be arranged at the decision time, t, without regard to transaction costs or market impact. Thus the prior position cannot influence the holding decision in any way. As the strategy is explicitly *risk limited*, rather than *risk averse*, we also do not need to account for risk reducing trade-offs between positions in correlated assets. Without that complication, we can analyze each asset individually, without regard to the trading in the others, and so the analysis of a single asset is sufficient to solve the multi-asset problem. The objective function then takes the simpler form

$$\Omega(h_t) = \int_{r_t} \int_{\alpha_t} Z(h_t, r_t) f_{r_t}(r_t|\alpha_t) f_{\alpha_t}(\alpha_t)\, dr_t d\alpha_t. \tag{5.18}$$

5.3.1. Computation of the expected profit

With holding function $h(\alpha_t)$, the profit from a position is just $p_t = h(\alpha_t) r_t$. Substituting this for $Z(h_t, r_t)$ in equation 5.18 gives

$$\mathbb{E}[p_t] = \int_{r_t} \int_{\alpha_t} h(\alpha_t) r_t f_{r_t}(r_t|\alpha_t) f_{\alpha_t}(\alpha_t)\, dr_t d\alpha_t. \tag{5.19}$$

Reversing the order of integration and collecting terms

$$\mathbb{E}[p_t] = \int_{\alpha_t} h(\alpha_t) f_{\alpha_t}(\alpha_t) \left\{ \int_{r_t} r_t f_{r_t}(r_t|\alpha_t)\, dr_t \right\} d\alpha_t. \tag{5.20}$$

Clearly, the term within braces $\{\dots\}$ is just the expected value of the return, $\mathbb{E}[r_t|\alpha_t] = \alpha_t$, so equation 5.20 may be written as follows:

$$\mathbb{E}[p_t] = 2L \int_b^\infty \alpha_t f_{\alpha_t}(\alpha_t)\, d\alpha_t \tag{5.21}$$

after substituting in the explicit value for $h(\alpha_t)$ from equation 5.17 on the preceding page. This integral is the incomplete first moment of the distribution of the alpha, which I have assumed to be symmetrical about 0.

The generalized error distribution: If again we consider the case that α_t is independently and identically distributed as $\mathrm{GED}(0, \omega, \lambda)$, equation 5.21 may be written in terms of the gamma functions as follows:

$$\mathbb{E}[p_t] = 2^\lambda L\omega \frac{\gamma\left(2\lambda, \frac{1}{2}\left(\frac{b}{\omega}\right)^{\frac{1}{\lambda}}\right)}{\Gamma(\lambda)}. \tag{5.22}$$

A normal limit may readily be obtained from equation 5.22 for $\lambda = 1/2$, and a Laplace limit for $\lambda = 1$. Note that in equation 5.22 the scale of the expected profit is set by $L\omega$ and not by the alpha. This is because it is an expectation w.r.t. the distribution of α_t, and ω determines the "typical size" of the alpha.

5.3.2. Computation of the variance of profits

The square of the profit is the statistic $Z(h_t, r_t) = p_t^2 = h(\alpha_t)^2 r_t^2$, with expectation

$$\begin{aligned}\mathbb{E}[p_t^2] &= \int_{r_t}\int_{\alpha_t} h(\alpha_t)^2 r_t^2 f_{r_t}(r_t|\alpha_t) f_{\alpha_t}(\alpha_t)\, dr_t d\alpha_t \\ &= \int_{\alpha_t} h(\alpha_t)^2 f_{\alpha_t}(\alpha_t) \left\{ \int_{r_t} r_t^2 f_{r_t}(r_t|\alpha_t)\, dr_t \right\} d\alpha_t. \end{aligned} \tag{5.23}$$

For a distribution of returns with mean α_t and variance σ_t^2, the term in the braces is $\alpha_t^2 + \sigma_t^2$, giving

$$\mathbb{E}[p_t^2] = 2L^2 \left\{ \int_b^\infty \alpha_t^2 f_{\alpha_t}(\alpha_t)\, d\alpha_t + \sigma_t^2 \bar{F}_{\alpha_t}(b) \right\} \tag{5.24}$$

$$\Rightarrow \quad \mathbb{V}[p_t] = 2L^2 \left[\int_b^\infty \alpha_t^2 f_{\alpha_t}(\alpha_t)\, d\alpha_t - 2 \left\{ \int_b^\infty \alpha_t f_{\alpha_t}(\alpha_t)\, d\alpha_t \right\}^2 + \sigma_t^2 \bar{F}_{\alpha_t}(b) \right]. \tag{5.25}$$

In equation 5.25, the first two terms, together, may be written as $\zeta \mathbb{V}[\alpha_t]$, for some $0 \leq \zeta \leq 1/2$, and the third term may be written as $\zeta' \sigma_t^2$, for similar ζ'. With weakly forecastable markets, the third term will dominate, since $\sigma_t^2 \gg \mathbb{V}[\alpha_t]$, and the variance of the profits may be approximated by

$$\mathbb{V}[p_t^2] \approx 2L^2 \sigma_t^2 \bar{F}_{\alpha_t}(b). \tag{5.26}$$

Here we see that the variance of the profits is set by $L\sigma_t$, not $L\omega$. This is a consequence of the assumption of weakly forecastable markets.

The generalized error distribution: Again we consider the specific form of the variance for then generalized error distribution:

$$\mathbb{V}[p_t] = L^2 \sigma_t^2 \frac{\gamma\left(\lambda, \frac{1}{2}\left(\frac{b}{\omega}\right)^{\frac{1}{\lambda}}\right)}{\Gamma(\lambda)}. \tag{5.27}$$

5.3.3. The Sharpe ratio as a function of the barrier height

Computing the Sharpe ratio as approximated by $\mathbb{E}[p_t]/\sqrt{\mathbb{V}[p_t]}$, scaled by the count of N independent trade opportunities in a trading year, gives

$$Z = \sqrt{2N} \frac{\int_b^\infty \alpha_t f_{\alpha_t}(\alpha_t)\, d\alpha_t}{\sigma_t \sqrt{\bar{F}_{\alpha_t}(b)}}. \tag{5.28}$$

Optimization of the Sharpe ratio: To find the extrema of equation 5.28 on the facing page, we differentiate under the integral w.r.t. b using Leibniz's rule [2]. The root of this derivative then satisfies

$$\hat{b} = b \ : \ \int_b^\infty \alpha_t f_{\alpha_t}(\alpha_t)\, d\alpha_t = 2b\bar{F}_{\alpha_t}(b). \tag{5.29}$$

Note that the location of the optimal barrier is independent of both the breadth of trading and the standard deviation of the returns.[7] The value of $\hat{b}$ that solves equation 5.29 is the optimal barrier location for a Sharpe ratio maximizer.

If we assume $\hat{b}$ is close to zero, then equation 5.29 becomes

$$\hat{b} = \frac{1}{2}\int_{-\infty}^{\infty} |\alpha_t| f_{\alpha_t}(\alpha_t)\, d\alpha_t \qquad \Rightarrow \qquad \hat{B} \approx \frac{1}{2}, \tag{5.30}$$

where, as before, $\hat{B}$ is the optimal barrier height in units of the standard deviation of the alpha. This result arises because the integral on the left side of equation 5.29 is one half of the expected value of $|\alpha_t|$, which will evaluate to a constant $O(1)$ times the standard deviation of the alpha. The term on the r.h.s. becomes just $2\bar{F}_{\alpha_t}(0) = 1$ for a symmetric distribution.

Although seemingly trivial, this is an important result as it shows that a Sharpe ratio maximizing trader will exhibit behavior different from a pure profit maximizing trader. Because $\hat{B} > 0$, there will always be some positive expectation trades that are not done. It is by this act of filtering potential trades based on the size of the alpha that the trader enhances the signal to noise ratio of their trading and so lifts their performance (as measured by the Sharpe ratio) above that of a pure profit maximizing trader.

The generalized error distribution: The functional dependence of equation 5.28 on the facing page on the barrier, b, is encapsulated in the lower limits of the incomplete first and zeroth moments of the distribution of the alpha, f_{α_t}, and so the precise location of an optimal barrier will depend on the tail properties of the density as

[7] In fact, it is independent of the *distribution* of returns. It depends only on the distribution of the alpha.

exhibited through this ratio. For the generalized error distribution,

$$Z = \frac{2^{\lambda}\omega}{\sigma_t} \frac{\gamma\left(2\lambda, \frac{1}{2}\left(\frac{b}{\omega}\right)^{\frac{1}{\lambda}}\right)}{\sqrt{\Gamma(\lambda)\gamma\left(\lambda, \frac{1}{2}\left(\frac{b}{\omega}\right)^{\frac{1}{\lambda}}\right)}}\sqrt{N}. \tag{5.31}$$

For this parameterization, the standard deviation of the alpha is $2^{\lambda}\omega\sqrt{\Gamma(3\lambda)/\Gamma(\lambda)}$. Equation 5.31 may then be written as follows:

$$Z = \frac{\gamma\left(2\lambda, \frac{1}{2}\left(\frac{b}{\omega}\right)^{\frac{1}{\lambda}}\right)}{\sqrt{\Gamma(3\lambda)\gamma\left(\lambda, \frac{1}{2}\left(\frac{b}{\omega}\right)^{\frac{1}{\lambda}}\right)}}R\sqrt{N}, \tag{5.32}$$

with R the I.C. when defined as the square root of the coefficient of determination not as the correlation.[8]

Figure 5.2 on the facing page shows the dependence of the Sharpe ratio scale factor, $Z/R\sqrt{N}$, on the barrier location for three values of the kurtosis factor, $\{1/2, 3/4, 1\}$, associated with the normal distribution, a realistic fit to market data, and the Laplace distribution, respectively. For the normal distribution, the optimal value of $\hat{B} = 0.612$, for the Laplace, it is 2 and for the intermediate case, 1.058. We see that our "small barrier" approximation of equation 5.29 on the previous page is close to the solution for the normal distribution and that $\hat{B}$ is an increasing function of the kurtosis parameter, λ.

The functional dependence of $\hat{b}/\omega$ on λ is shown in Figure 5.3 on page 108 for values of λ from 0 to 1. This encompasses the uniform distribution ($\lambda \to 0$) through to the Laplace distribution ($\lambda = 1$). The curve is well approximated by

$$\frac{\hat{b}}{\omega} \approx e^{\lambda^2} - \frac{2}{3} \quad \text{for } 0 \leq \lambda \leq 1. \tag{5.33}$$

Alternatively,

$$\hat{B} \approx \frac{e^{\lambda^2} - \frac{2}{3}}{2^{\lambda}}\sqrt{\frac{\Gamma(\lambda)}{\Gamma(3\lambda)}}. \tag{5.34}$$

[8] That is, $R^2 = \mathbb{V}[\alpha_t]/\mathbb{V}[r_t]$, as it is usually defined for linear regression.

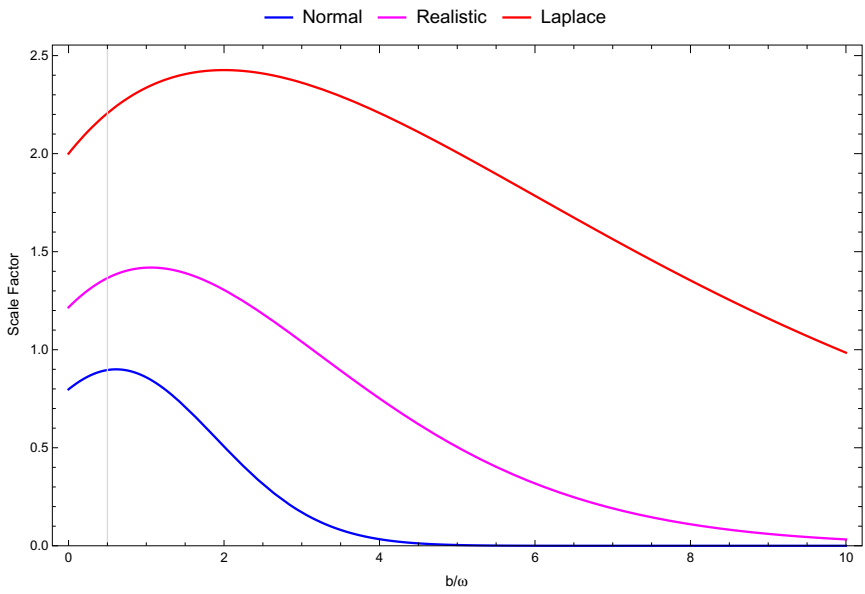


Figure 5.2. The dependence of the Sharpe ratio scale factor, $Z/R\sqrt{N}$, on the barrier location for three values of the kurtosis factor. The blue line is for $\lambda = 1/2$ (normal distribution), magenta for a realistic value of 3/4, and red for $\lambda = 1$ (Laplace distribution).

Conclusions: A risk limited trader who wishes to maximize their Sharpe ratio *ex ante*, and wishes to trade using fixed position sizes (long, short, or flat) to limit risk, should enter trades in the direction of the alpha when the alpha exceeds a threshold which is a strongly increasing function of the kurtosis of the distribution of alphas but is approximately equal to the standard deviation of the alpha for values that would arise from simple autoregressive models of returns such as $AR(1)$. Such a trader should not do all positive expectation trades but should veto trades where the signal-to-noise ratio is too low.

When I first started working with these kinds of trading algorithms, transaction costs were highly significant. The results of Section 5.3 would have been regarded as a "toy model" en route to the main destination, which is "viscous" trading with significant transaction costs. For many retail traders, this is no longer the case, as brokerages have switched to models which transfer the cost of trading from customers to large hedge fund counterparties who pay for

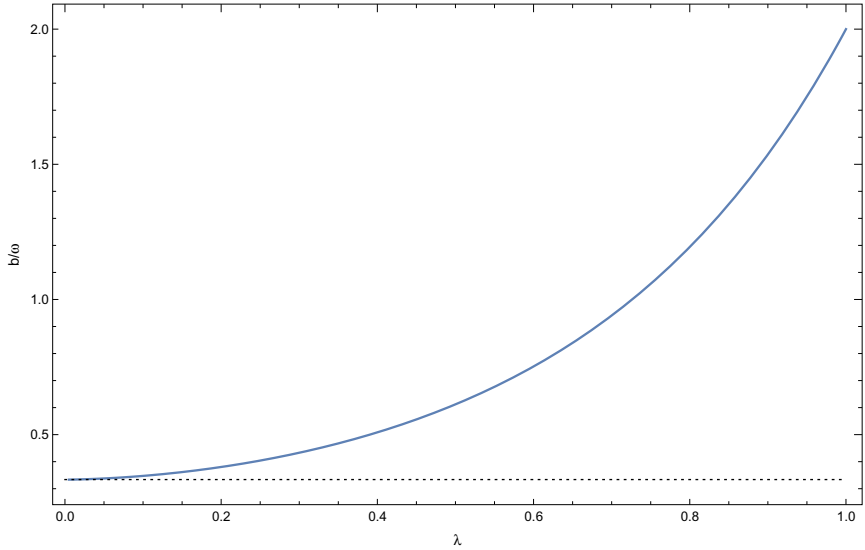


Figure 5.3. The location of the optimal trade entry barrier, $\hat{b}/\omega$, for a range of values for the kurtosis factor, λ. The curve is well approximated by $e^{\lambda^2} - 2/3$ for $\lambda \in [0, 1]$.

order flow because retail traders are generally viewed as uniformed at best.

5.4. Barrier Trading with Transaction Costs

Incorporating transaction costs into the utility formalism of Essay 3 is quite difficult because, for a trader with insignificant market impact, costs generally scale with $|\Delta h_t|$ which leads to non-differentiable functions. Results can always be obtained via numerical methods, but without the structure of an analytical solution it is more difficult to understand *what they mean*. Furthermore, we have shown in Essay 3 that holding functions *for realistic distributions of returns* may often be easily approximated by simple step functions that are considerably easier for the trader to manage. In this section, I will analyze risk limited, viscous, barrier trading which takes account of transaction costs.

5.4.1. The effect of transaction costs

The effect of transaction costs can be understood by ascending the hierarchy of optimization strategies, as outlined in Section 2.5.4 on page 46, while understanding how the changes we make affect the solutions. The step from gross profit maximization to risk limited gross profit maximization is trivial; we merely limit position sizes. The step from risk limited maximization to risk-averse maximization can be understood from the strategy developed in Section 5.3: the Sharpe ratio is improved by choosing to veto some positive expectation trades that do not produce sufficient returns given the risk they require. This can be thought of as introducing a cost of risk and only trading when the alpha exceeds the cost of risk. Similarly, the step from risk limited gross profit maximization to risk limited net profit maximization adds another cost, in this case the transaction cost of trading.

Within the context of barrier trading, the risk aversion has introduced the barrier $b > 0$, and so it is not unreasonable to hypothesize that transaction costs would introduce an additional barrier, b'. To understand how these barriers are related, it helps to follow the process of a trader in detail. If the alpha is positive and exceeds the cost of trading, we want to go long if the *net* profit exceeds the risk costs. The only complication associated with this is that the cost of trading depends on trade size, and so depends on whether we are going to go long from a short position or from a flat position, as the former will cost twice as much as the latter. Clearly the same considerations apply to a negative alpha and shorting the asset. If the position is already long, and the alpha is negative, the trader has two choices: to reverse and go short or to merely exit the long position. If the alpha is sufficiently negative to cover the transaction costs of reversing the position *and* the risk costs, then the trader should go short. If the alpha is negative and covers the cost of flattening the position, the trader should do that, but if the cost of flattening outweighs the risk costs of maintaining a position, they should stay long. Again, the same consideration applies in reverse.

In the above, "costs" don't necessarily mean the literal cost of trading but the "effective" costs of trading, which takes into account the probability that the cost will be incurred. At every potential trade interval, of which there are P per annum, the alpha does not have to

exceed the cost of trading into a position from flat because sometimes the trader will already be in the desired long position "by chance." Thus the expected cost of entering a long position, $h_t = L$, is

$$\hat{\kappa} = 2\kappa \times \Pr(h_{t-1} = -L) + \kappa \times \Pr(h_{t-1} = 0) + 0 \times \Pr(h_{t-1} = L), \tag{5.35}$$

where κ is the actual transaction cost per contract. Clearly the trade should only be done if $\alpha_t \geq \hat{\kappa}$, all other things being equal, which is when the expected return exceeds the expected costs *per contract traded*. In general, there may be scenarios in which $\hat{\kappa} = \kappa$ and others in which it is not.

5.4.2. The Modified trading algorithm

Incorporating the observations of the prior section into the trading algorithm, we modify it as follows:

(i) If the magnitude of alpha exceeds a critical level b_1, then a position equal to the sign of the alpha is taken.
(ii) If the alpha exceeds a second critical level b_2, but not large enough in size to exceed b_1 *in the opposite direction to the position*,[9] then a trade is done to flat.

This gives the holding function:

$$h(\alpha_t, h_{t-1}) = \begin{cases} +L & \text{if } \alpha_t \geq b_1 \\ 0 & \begin{cases} \text{if} & h_{t-1} = +L & \text{and} & -b_1 < \alpha_t < b_2 \\ \text{or} & h_{t-1} = 0 & \text{and} & -b_1 < \alpha_t < b_1 \\ \text{or} & h_{t-1} = -L & \text{and} & -b_2 < \alpha_t < b_1 \end{cases} \\ -L & \text{if } \alpha_t \leq -b_1. \end{cases} \tag{5.36}$$

Equation 5.36 is recognizable as the holding function of equation 5.17 on page 102 with the central case further subdivided according to the prior position. Note that $b_2 \leq b_1$ by definition and if $b_2 \leq -b_1$ the second barrier has no affect on the trading done. Thus we are only interested in $-b_1 \leq b_2 \leq b_1$.

[9] i.e. $-b_1 < \alpha_t < b_2$.

5.4.3. Computation of the expected metrics with discrete holdings

As before we assume the statistic is homogeneous, the returns are increasingly weakly predictable, and returns are independent the holdings. When transaction costs are important, we will also assume that prior holdings are independent of the alpha. The objective we seek to maximize, to obtain the holding function of our trading strategy, is then

$$\Omega = \int_{r_t} \int_{\alpha_t} \sum_{h_{t-1}} Z(h_t, r_t, h_{t-1}) f_{r_t}(r_t|\alpha_t) f_{\alpha_t}(\alpha_t) f_{h_{t-1}}(h_{t-1})\, dr_t d\alpha_t, \tag{5.37}$$

where a *discrete* distribution is used to represent the possible prior holdings $h_{t-1} \in [-L, 0, L]$ and their associated probability mass function $f_{h_{t-1}}(h_{t-1})$.

5.4.4. The distribution of the holdings

For the probability mass function, we just need to consider three states, and so there are only two probabilities. Writing the probability that the prior position is a long one as $p = \Pr[h_{t-1} = +L]$ and, similarly $q = \Pr[h_{t-1} = -L]$ and $n = \Pr[h_{t-1} = 0] = 1 - p - q$, then we can model the p.m.f. by

$$f_{h_{t-1}}(h_{t-1}) = \begin{cases} p & \text{if } h_{t-1} = +L \\ n & \text{if } h_{t-1} = 0 \\ q & \text{if } h_{t-1} = -L. \end{cases} \tag{5.38}$$

With this choice, equation 5.37 becomes

$$\Omega = \int_{r_t} \int_{\alpha_t} \left\{ \begin{array}{c} pZ(h_t, r_t, +L) \\ +nZ(h_t, r_t, 0) \\ +qZ(h_t, r_t, -L) \end{array} \right\} f_{r_t}(r_t|\alpha_t) f_{\alpha_t}(\alpha_t)\, dr_t d\alpha_t. \tag{5.39}$$

5.4.5. Symmetric distributions

We now make the assumption that the distribution of the alphas is symmetric about zero, which is very reasonable given the observed

distributions of returns exhibited in *Adventures in Financial Data Science* [18], which means that $f_{\alpha_t}(\alpha_t) = f_{\alpha_t}(-\alpha_t)$. With a symmetric distribution of alphas *and* a trading algorithm that treats alphas of different signs symmetrically, such as equation 5.36 on page 110, it must also follow that the trader is just as likely to be long as short, so $p = q$ and $n = 1 - 2p$. With this assumption, Equation 5.39 on the preceding page becomes

$$\Omega = \int_{r_t} \int_{\alpha_t} \left\{ \begin{array}{c} pZ(h_t, r_t, +L) \\ +(1-2p)Z(h_t, r_t, 0) \\ +pZ(h_t, r_t, -L) \end{array} \right\} f_{r_t}(r_t|\alpha_t) f_{\alpha_t}(\alpha_t)\, dr_t d\alpha_t. \tag{5.40}$$

5.4.6. The unconditional probability of holding a position

With the trading algorithm of equation 5.36 on page 110, there are only two circumstances that can lead to a trader having a long position:

(1) the alpha is sufficiently large for the trader to be long regardless of the prior position,
(2) the prior position is long and the alpha is sufficiently large to keep that position.

$$f_{h_t}(L) = \int_{b_1}^{\infty} f_{\alpha_t}(\alpha_t)\, d\alpha_t + f_{h_{t-1}}(L) \int_{b_2}^{b_1} f_{\alpha_t}(\alpha)\, d\alpha_t. \tag{5.41}$$

In equation 5.41, the second integral stops at b_1 because the probability mass for $\alpha_t \geq b_1$ is already included in the first integral. Substituting the unconditional probability, p, for both $f_{h_t}(L)$ and $f_{h_{t-1}}(L)$ leads to an expression that may be solved for p in terms of the cumulative distribution function of the alpha,

$$p = \frac{\bar{F}_{\alpha_t}(b_1)}{\bar{F}_{\alpha_t}(b_1) + F_{\alpha_t}(b_2)}. \tag{5.42}$$

For a symmetric distribution of alphas, $b_2 = -b_1 \Rightarrow p = 1/2$ independent of the distribution. An example of this function is shown in Figure 5.4 on the next page for the case of alphas drawn from a Laplace distribution with zero mean and standard deviation ω.

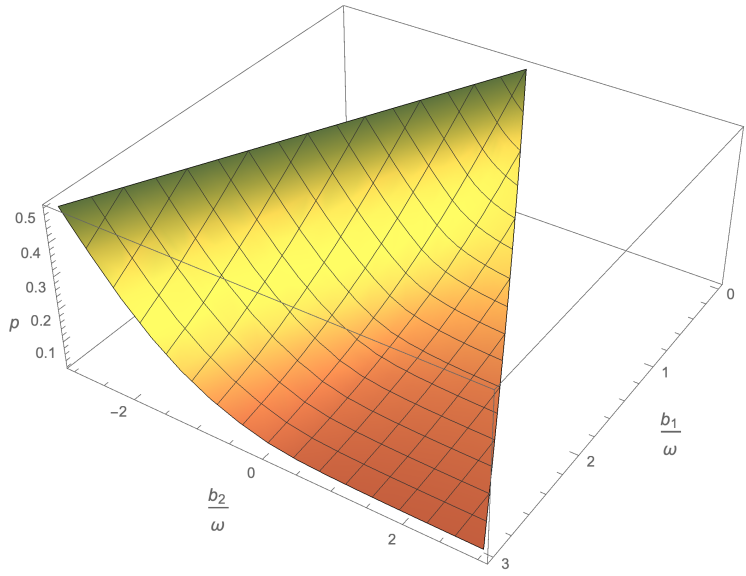


Figure 5.4. The unconditional probability of a long position for the trading algorithm of equation 5.36 on page 110. The function is plotted for a Laplace distribution of alphas with zero mean and standard deviation ω. The infeasible region $|b_2| > b_1$ is excluded from the plot.

5.4.7. Computation of expected profits

In equation 5.39 on page 111, we see that the discreteness of the prior holdings permits the summation w.r.t. to this variable to be explicitly enumerated. In general, the statistic Z acts as a *copula* for the otherwise (assumed) independent random variables r_t, α_t, and h_{t-1}. As we have specified a discrete holding function, with three states, the definition of the *net profit* converts the summation over three terms into a summation over nine terms, now contingent of the value of the alpha.

We have

$$\begin{aligned} Z(h_t, r_t, +L) &= r_t L\mathbb{I}[\alpha_t \geq b_2] - \kappa L\mathbb{I}[-b_1 < \alpha_t < b_2] \\ &\quad - (r_t + 2\kappa)L\mathbb{I}[\alpha_t \leq -b_1] \\ Z(h_t, r_t, 0) &= (r_t - \kappa)L\mathbb{I}[\alpha_t \geq b_1] - (r_t + \kappa)L\mathbb{I}[\alpha_t \leq -b_1] \\ Z(h_t, r_t, -L) &= (r_t - 2\kappa)L\mathbb{I}[\alpha_t \geq b_1] - \kappa L\mathbb{I}[-b_2 < \alpha_t < b_1] \\ &\quad - r_t L\mathbb{I}[\alpha_t \leq -b_2], \end{aligned} \tag{5.43}$$

where $\mathbb{I}[x]$ represents the indicator function, taking unit value if proposition x is true and zero otherwise, and κ the transaction cost[10] *per contract traded*. In equations 5.43, the dependence of the net profit on the return, r_t, is separable from the dependence on the alpha and is either linear in r_t or has no dependence on r_t. Thus we may integrate these terms giving the factors

$$\int_{-\infty}^{\infty} r_t f_{r_t}(r_t|\alpha_t)\, dr_t = \alpha_t \quad \text{and} \quad \int_{-\infty}^{\infty} f_{r_t}(r_t|\alpha_t)\, dr_t = 1 \tag{5.44}$$

within the integral w.r.t. α_t. The summation over three terms in equation 5.39 on page 111 becomes a summation over eight terms[11] and the only integrals left to evaluate involve unit or null powers of α_t over regions of its support.

Taking these term by term leads to[12]

$$\Omega = L \begin{bmatrix} +p \left\{ \int_{b_2}^{\infty} \alpha_t f_{\alpha_t}(\alpha_t)\, d\alpha_t - \kappa \int_{-b_1}^{b_2} f_{\alpha_t}(\alpha_t)\, d\alpha_t \right. \\ \quad \left. - \int_{-\infty}^{-b_1} (\alpha_t + 2\kappa) f_{\alpha_t}(\alpha_t)\, d\alpha_t \right\} \\ \quad + n \left\{ \int_{b_1}^{\infty} (\alpha_t - \kappa) f_{\alpha_t}(\alpha_t)\, d\alpha_t \right. \\ \quad \left. - \int_{-\infty}^{-b_1} (\alpha_t + \kappa) f_{\alpha_t}(\alpha_t)\, d\alpha_t \right\} \\ \quad + q \left\{ \int_{b_1}^{\infty} (\alpha_t - 2\kappa) f_{\alpha_t}(\alpha_t)\, d\alpha_t - \kappa \int_{-b_2}^{b_1} f_{\alpha_t}(\alpha_t)\, d\alpha_t \right. \\ \quad \left. - \int_{-\infty}^{-b_2} \alpha_t f_{\alpha_t}(\alpha_t)\, d\alpha_t \right\} \end{bmatrix}. \tag{5.45}$$

Although this expression is lengthy, it is just the average of the expected net profits from all of the nine possible trades, at time t, when weighted by the unconditional probabilities of the prior positions.[13] Reversing the order of integration for the first term in the

[10]Including *slippage*, or the difference between the trigger price and the execution price, as well as brokerage and regulatory fees.

[11]The "trade" $h_{t-1} = 0 \;\rightarrow\; h_t = 0$ having no impact on the net profit either through potential gains or transaction costs.

[12]This equation is written out as a matrix to emphasize its symmetries.

[13]The trades are, in order from upper left to lower right: stay long, flatten long, reverse long, go long, stay flat, go short, reverse short, cover short, and stay short.

second column and the entire third column, and then making the transformation $\alpha_t \to -\alpha_t$, gives

$$\Omega = L \left[\begin{array}{l} p \left\{ \int_{b_2}^{\infty} \alpha_t f_{\alpha_t}(\alpha_t)\, d\alpha_t - \kappa \int_{-b_2}^{b_1} f_{\alpha_t}(-\alpha_t)\, d\alpha_t \right. \\ \quad \left. + \int_{b_1}^{\infty} (\alpha_t - 2\kappa) f_{\alpha_t}(-\alpha_t)\, d\alpha_t \right\} \\ \quad + n \left\{ \int_{b_1}^{\infty} (\alpha_t - \kappa) f_{\alpha_t}(\alpha_t)\, d\alpha_t \right. \\ \quad \left. + \int_{b_1}^{\infty} (\alpha_t - \kappa) f_{\alpha_t}(-\alpha_t)\, d\alpha_t \right\} \\ \quad + q \left\{ \int_{b_1}^{\infty} (\alpha_t - 2\kappa) f_{\alpha_t}(\alpha_t)\, d\alpha_t - \kappa \int_{-b_2}^{b_1} f_{\alpha_t}(\alpha_t)\, d\alpha_t \right. \\ \quad \left. + \int_{b_2}^{\infty} \alpha_t f_{\alpha_t}(-\alpha_t)\, d\alpha_t \right\} \end{array} \right]. \tag{5.46}$$

Now combining integrals with common domains and integrands, and using the symmetry of the distribution described in Section 5.4.5 on page 111 to substitute $f_{\alpha_t}(-\alpha_t) \to f_{\alpha_t}(\alpha_t)$, $q \to p$ and $n \to 1-2p$, leads to a compact expression

$$\begin{aligned} \Omega = 2L \left[(1-p) \left\{ \int_{b_1}^{\infty} \alpha_t f_{\alpha_t}(\alpha_t)\, d\alpha_t - \kappa \bar{F}_{\alpha_t}(b_1) \right\} \right. \\ \left. + p \left\{ \int_{b_2}^{\infty} \alpha_t f_{\alpha_t}(\alpha_t)\, d\alpha_t - \kappa \bar{F}_{\alpha_t}(-b_2) \right\} \right]. \end{aligned} \tag{5.47}$$

5.4.8. The profit function

Equation 5.47 may be summarized compactly by the introduction of a special function, which I call "the Profit Function." Let $f_{A_t}(A_t)$ be the p.d.f. of a standardized alpha, $A_t = \alpha_t/\omega_t$, where ω_t is the standard deviation of the alpha.[14] Then

$$f_{\alpha_t}(\alpha_t) = \frac{f_{A_t}(\alpha_t/\omega_t)}{\omega_t} \quad \text{and} \quad \bar{F}_{A_t}(b/\omega_t) = \bar{F}_{\alpha_t}(b). \tag{5.48}$$

[14]The mean of the alpha is zero under the symmetry conditions imposed earlier.

Now define[15]

$$\Pi_s(B,K) = \int_B^\infty A_t f_{A_t}(A_t)\, dA_t - K\bar{F}_{A_t}(sB) \tag{5.49}$$

for scale free variables B and K. Equation 5.47 on the preceding page may now be written as follows:

$$\Omega = 2L\omega_t \left\{(1-p)\Pi_1(B_1,K) + p\Pi_{-1}(B_2,K)\right\}, \tag{5.50}$$

where $B_1 = b_1/\omega_t$, $B_2 = b_2/\omega_t$, and $K = \kappa/\omega_t$. The profit function is shown in Figure 5.5 for the special case of the Laplace distribution. In general, the function has a maximum of $\xi - K/2$ at the origin, where ξ is a positive constant $O(1)$ with a value dependent on the specific distribution of the alpha.

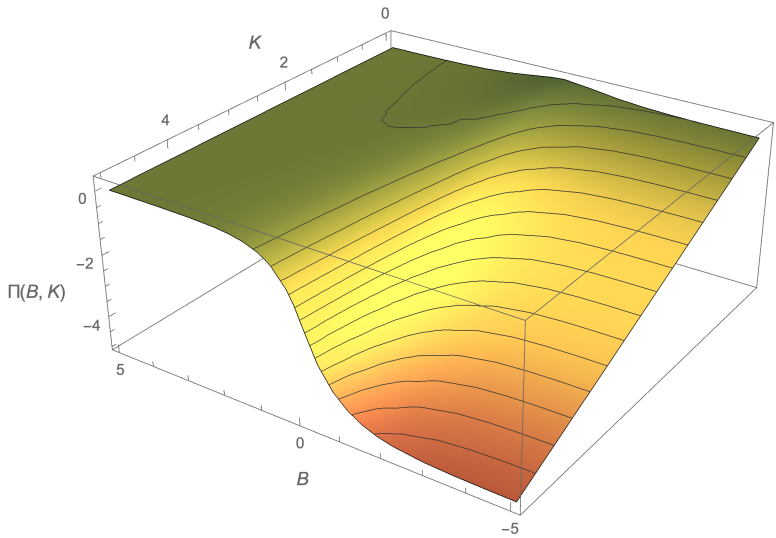


Figure 5.5. The profit function for the Laplace distribution. The maximum of the function is at the origin, which corresponds to the maximum gross profit solution. B is the barrier and K the transaction costs in units of ω_t, the standard deviation of the alpha.

[15] Π does not represent the product operation.

5.4.9. The optimal net profit maximizing algorithm

Should the trader chose simply to maximize their net profit, the values of (b_1, b_2) required are obtained by differentiating Ω with respect to these parameters and solving for the values, $(\hat{b}_1, \hat{b}_2)$, at which those derivatives vanish. The derivatives are as follows:

$$\frac{\partial \Omega}{\partial b_1} = -2L\omega_t \left\{ (1-p)(b_1 - \kappa) f_{\alpha_t}(b_1) + \Pi_1(B_2, K) \frac{\partial p}{\partial b_1} \right\} \tag{5.51}$$

$$\text{and} \quad \frac{\partial \Omega}{\partial b_2} = -2L\omega_t \left\{ p(b_2 + \kappa) f_{\alpha_t}(b_2) - \Pi_{-1}(B_2, K) \frac{\partial p}{\partial b_2} \right\}. \tag{5.52}$$

These expressions appear to be analytically quite difficult to solve, however, from Section 5.4.6 on page 112, we know that there is one constraint under which the partial derivatives of p are zero, which is when $b_2 = -b_1$. With this condition, equations 5.51 and 5.52 are solved by

$$\hat{b}_1 = +\kappa \quad \text{or} \quad \hat{b}_1 \to \pm\infty \tag{5.53}$$

$$\text{and} \quad \hat{b}_2 = -\kappa \quad \text{or} \quad \hat{b}_2 \to \mp\infty. \tag{5.54}$$

The infinite solutions exist because it *must* be true that[16]

$$\lim_{\alpha_t \to \pm\infty} f_{\alpha_t}(\alpha_t) = 0. \tag{5.55}$$

Taking the finite solutions, these results tell us to place our trading barriers 2κ apart at $\pm\kappa$ and that these values are not functions of the distribution of the alpha. We enter a long trade when $\alpha_t \geq \kappa$ and hold it unless $\alpha_t \leq -\kappa$, when we reverse to take a short position. The coincidence of the "cover" barrier, at $b_2 = \mp\kappa$, with the "entry" barrier, at $b_1 = \pm\kappa$, has the effect of keeping the trader in a long position *even when the alpha is negative*, provided it is sufficiently small. Thus the effect of transaction costs has been to introduce *hysteresis* into the holding function, as is seen in Figure 5.6. This is a profound change in comparison to the holding function discovered when transaction costs are zero.

[16] This follows from the definition of a p.d.f.

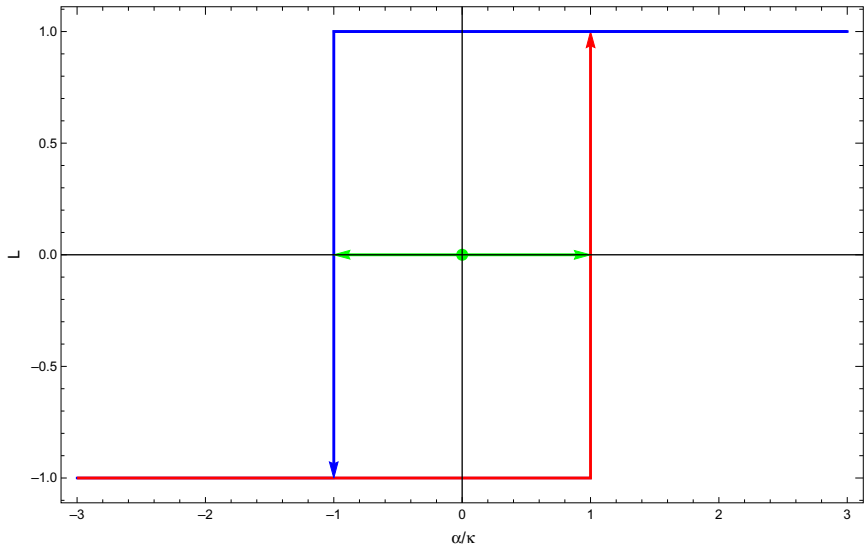


Figure 5.6. The holding function for a risk-limited maximum net profit optimizing trader. The blue line represents the response from a long position and the red line that from a short position. The "initial condition" is represented by the green line.

5.4.10. The effect of hysteresis in signal processing

When a trader uses a barrier to regulate the entry into, and exit from, trades with a stochastic alpha, they often suffer from what electronic engineers would call *jitter* at the threshold. If a signal could be represented as $\alpha_t = \kappa + \varepsilon_t$, where ε_t is i.i.d. according to some distribution with density with both positive and negative support, then there will be a whipsawing effect when trades are entered on a small incremental alpha only to be frequently reversed, destroying the traders equity in transaction costs. If the unconditional mean of the alpha is zero and $\alpha_t \geq \kappa$, then $\Pr(\alpha_{t+1} < \kappa) > 1/2$ and so the effect of the next alpha is more likely to reverse the entered position than to persist it and a trade with an alpha perhaps just marginally larger than the transaction cost would result in a round trip transaction cost of twice that cost and be unprofitable. This problem frequently occurs in circuit design, since electronic signals are always noisy at some level, when we are using a *comparator* to interface between an analogue measurement system and a digital signal processing system.

It is corrected for by the use of the *Schmitt trigger*,[17] a circuit element with hysteresis in its response function [26].

5.4.11. Computation of the variance of the profits

As in Section 5.3, we work towards the computation of the Sharpe ratio[18] of the trading strategy by first computing the expected value of the square of the profits. Equation 5.39 on page 111 is still relevant, but now we will write Ω' for the objective function and $Z'(h_t, r_t, h_{t-1})$ for the statistic to be optimized in expectation.

This function has values

$$\begin{aligned} Z'(h_t, r_t, +L) &= r_t^2 L^2 \mathbb{I}[\alpha_t \geq b_2] + \kappa^2 L^2 \mathbb{I}[-b_1 < \alpha_t < b_2] \\ &\quad + (r_t + 2\kappa)^2 L^2 \mathbb{I}[\alpha_t \leq -b_1] \\ Z'(h_t, r_t, 0) &= (r_t - \kappa)^2 L^2 \mathbb{I}[\alpha_t \geq b_1] + (r_t + \kappa)^2 L^2 \mathbb{I}[\alpha_t \leq -b_1] \\ Z'(h_t, r_t, -L) &= (r_t - 2\kappa)^2 L^2 \mathbb{I}[\alpha_t \geq b_1] + \kappa^2 L^2 \mathbb{I}[-b_2 < \alpha_t < b_1] \\ &\quad + r_t^2 L^2 \mathbb{I}[\alpha_t \leq -b_2]. \end{aligned} \tag{5.56}$$

Marginalizing w.r.t. r_t, *within* the integral w.r.t. α_t, will involve

$$\begin{aligned} &\int_{\alpha_t} \int_{r_t} (r_t + s\kappa)^2 f_{r_t}(r_t|\alpha_t) f_{\alpha_t}(\alpha_t)\, dr_t d\alpha_t \\ &\quad = \int_{\alpha_t} (\alpha_t^2 + \sigma_t^2 + 2s\alpha_t + s^2\kappa^2) f_{\alpha_t}(\alpha_t)\, d\alpha_t, \end{aligned} \tag{5.57}$$

where σ_t^2 is the conditional variance of the returns[19] and $s \in [0, \pm 1, \pm 2]$. We assume that variance of returns is *not* a function of the alpha.

The variance of the trader's profits is $\Omega' - \Omega^2$, where Ω is given by the expressions developed above and we define Ω' to be the unconditional expectation of the square of the profits. As before, taking

[17] Invented by Otto Schmitt in 1938 to deal with these very problems in his work on the working of squid nerves!

[18] Remember, our reservations on the use of the Sharpe ratio are regarding its use as an *ex post* statistic not an *ex ante* objective of optimization.

[19] This follows from the definition of variance, $\mathbb{V}[r_t|\alpha_t] = \mathbb{E}[r_t^2|\alpha_t] - \{\mathbb{E}[r_t|\alpha_t]\}^2$.

these term-by-term leads to equation 5.58. Making the transformation $\alpha_t \to -\alpha_t$ and reversing order of integration then gives equation 5.59

$$\Omega' = L^2 \left[\begin{array}{l} p\left\{\int_{b_2}^{\infty}(\alpha_t^2+\sigma_t^2)f_{\alpha_t}(\alpha_t)\,d\alpha_t + \kappa^2\int_{-b_1}^{b_2} f_{\alpha_t}(\alpha_t)\,d\alpha_t \right. \\ \qquad\qquad \left. + \int_{-\infty}^{-b_1}(\alpha_t^2+\sigma_t^2+4\alpha_t\kappa+4\kappa^2)f_{\alpha_t}(\alpha_t)\,d\alpha_t\right\} \\ +n\left\{\int_{b_1}^{\infty}(\alpha_t^2+\sigma_t^2-2\alpha_t\kappa+\kappa^2)f_{\alpha_t}(\alpha_t)\,d\alpha_t \right. \\ \qquad\qquad \left. + \int_{-\infty}^{-b_1}(\alpha_t^2+\sigma_t^2+2\alpha_t\kappa+\kappa^2)f_{\alpha_t}(\alpha_t)\,d\alpha_t\right\} \\ +q\left\{\int_{b_1}^{\infty}(\alpha_t^2+\sigma_t^2-4\alpha_t\kappa+4\kappa^2)f_{\alpha_t}(\alpha_t)\,d\alpha_t \right. \\ \qquad \left. +\kappa^2\int_{-b_2}^{b_1} f_{\alpha_t}(\alpha_t)\,d\alpha_t + \int_{-\infty}^{-b_2}(\alpha_t^2+\sigma_t^2)f_{\alpha_t}(\alpha_t)\,d\alpha_t\right\} \end{array} \right] \tag{5.58}$$

$$= L^2 \left[\begin{array}{l} p\left\{\int_{b_2}^{\infty}(\alpha_t^2+\sigma_t^2)f_{\alpha_t}(\alpha_t)\,d\alpha_t + \kappa^2\int_{-b_2}^{b_1} f_{\alpha_t}(-\alpha_t)\,d\alpha_t \right. \\ \qquad\qquad \left. + \int_{b_1}^{\infty}(\alpha_t^2+\sigma_t^2-4\alpha_t\kappa+4\kappa^2)f_{\alpha_t}(-\alpha_t)\,d\alpha_t\right\} \\ +n\left\{\int_{b_1}^{\infty}(\alpha_t^2+\sigma_t^2-2\alpha_t\kappa+\kappa^2)f_{\alpha_t}(\alpha_t)\,d\alpha_t \right. \\ \qquad\qquad \left. + \int_{b_1}^{\infty}(\alpha_t^2+\sigma_t^2-2\alpha_t\kappa+\kappa^2)f_{\alpha_t}(-\alpha_t)\,d\alpha_t\right\} \\ +q\left\{\int_{b_1}^{\infty}(\alpha_t^2+\sigma_t^2-4\alpha_t\kappa+4\kappa^2)f_{\alpha_t}(\alpha_t)\,d\alpha_t \right. \\ \qquad \left. +\kappa^2\int_{-b_2}^{b_1} f_{\alpha_t}(\alpha_t)\,d\alpha_t + \int_{b_2}^{\infty}(\alpha_t^2+\sigma_t^2)f_{\alpha_t}(-\alpha_t)\,d\alpha_t\right\} \end{array} \right]. \tag{5.59}$$

Combining integrals with common integrands and limits, and assuming the alpha is symmetric about zero as before, gives

$$\Omega' = 2L^2\left[\int_{b_1}^{\infty}\left\{(1-p)(\sigma_t^2+\alpha_t^2)+(1+2p)\kappa^2-2\kappa\alpha_t\right\}f_{\alpha_t}(\alpha_t)\,d\alpha_t \right. \\ \left. + p\int_{b_2}^{\infty}(\alpha_t^2+\sigma_t^2)f_{\alpha_t}(\alpha_t)\,d\alpha_t + p\kappa^2\left\{F_{\alpha_t}(b_1)+F_{\alpha_t}(b_2)-1\right\}\right]. \tag{5.60}$$

Our assumption of weakly predictable markets means that terms in the integrand involving $\sigma_t^2+\alpha_t^2$ may be approximated by just σ_t^2. If we further assume that the variance of returns is also much larger than the square of the transaction cost i.e. that $\sigma_t^2 \gg \kappa^2$, then we may

drop *all terms* in equation 5.60 on the facing page that don't involve σ_t^2. Therefore

$$\Omega' \approx 2\sigma_t^2 L^2 \left\{(1-p)\bar{F}_{\alpha_t}(b_1) + p\bar{F}_{\alpha_t}(b_2)\right\}, \tag{5.61}$$

and, because of the assumption of weakly forecastable returns, this is also equal to the variance of the net profit.[20] Substituting in the value of $p(b_1, b_2)$ from equation 5.42 on page 112, the variance of the net profit is approximately

$$2L^2\sigma_t^2 \frac{\bar{F}_{\alpha_t}(b_1)}{\bar{F}_{\alpha_t}(b_1) - F_{\alpha_t}(b_2)} = 2L^2\sigma_t^2 \frac{\bar{F}_{A_t}(B_1)}{\bar{F}_{A_t}(B_1) - F_{A_t}(B_2)}. \tag{5.62}$$

5.4.12. The Sharpe ratio for barrier trading with transaction costs

For P trading periods per annum, the Sharpe ratio for choice of barriers (B_1, B_2) and transaction cost K, both expressed in units of the standard deviation of the alpha, ω_t, is

$$Z \approx \frac{\Omega}{\sqrt{\Omega'/P}} = R\frac{(1-p)\Pi_1(B_1, K) + p\Pi_{-1}(B_2, K)}{\sqrt{(1-p)\bar{F}_{\alpha_t}(B_1) + p\bar{F}_{\alpha_t}(B_2)}}\sqrt{2P}, \tag{5.63}$$

where $R = \omega_t/\sigma_t$ is the information ratio of the alpha and the remaining factor is the breadth of this strategy.[21] Substituting in the value of p from equation 5.42 on page 112 gives

$$Z = \frac{F_{A_t}(B_2)\Pi_1(B_1, K) + F_{A_t}(B_1)\Pi_{-1}(B_2, K)}{\sqrt{\bar{F}_{A_t}(B_1)\left\{\bar{F}_{A_t}(B_1) + F_{A_t}(B_2)\right\}/2}} R\sqrt{P}. \tag{5.64}$$

5.4.13. The Sharpe ratio for various distributions of alphas

Inspecting the profit function in Figure 5.5 on page 116, we see that the projection onto the B axis for large K is a sigmoid function. This shape arises from the cumulative distribution function of the

[20] That is $\Omega' - \Omega^2 \approx \Omega'$.

[21] Comparing equation 5.63 to Grinold and Kahn's *Fundamental Law* [24], equation 5.9 on page 99.

alpha used and, due to the nature of continuous probability functions, many distributions will give rise to such sigmoid shapes. Similarly, the projection onto the K axis, for large negative B, is a straight line. This arises from the linear transaction cost model used throughout this work. Although the specific value of the Sharpe ratio given by equation 5.64 on the preceding page will vary as a function of the distribution used, we should not expect it to vary *much* at these limits.

The normal distribution: Figure 5.7 shows the computed Sharpe ratio for a normal distribution and unit transaction cost $K = 1$, or $\kappa = \omega_t$. The surface is restricted to the region of the plane defined by $0 \leq |B_2| \leq B_1$. A clearly defined positive maximum Sharpe ratio point is visible on the plot. For this particular distribution and transaction cost the optimal location of the $\hat{B}_2$ barrier is negative while that of the $\hat{B}_1$ barrier is positive.

The Laplace distribution: Figure 5.8 on the facing page in contrast shows the computed Sharpe ratio for a Laplace distribution with the same standard deviation and transaction cost. The peak

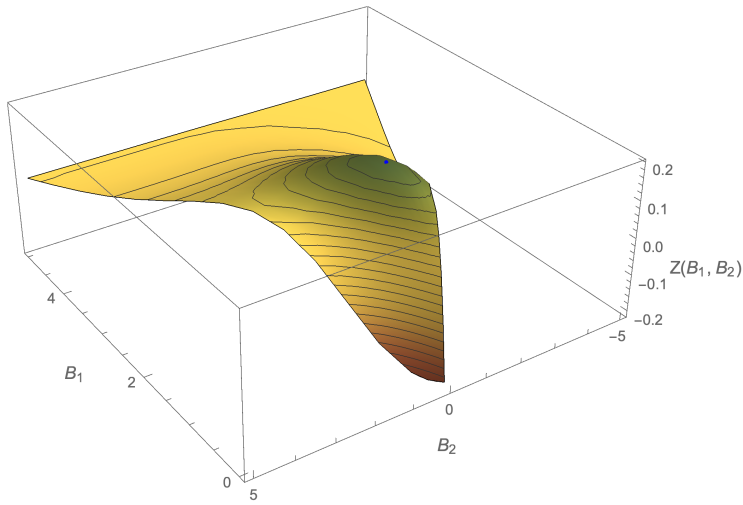


Figure 5.7. The Sharpe ratio for a trading strategy with barriers at B_1 and B_2, and unit transaction cost, with a distribution of alphas drawn from the standard normal distribution. The positive maximum Sharpe ratio is marked with a blue dot.

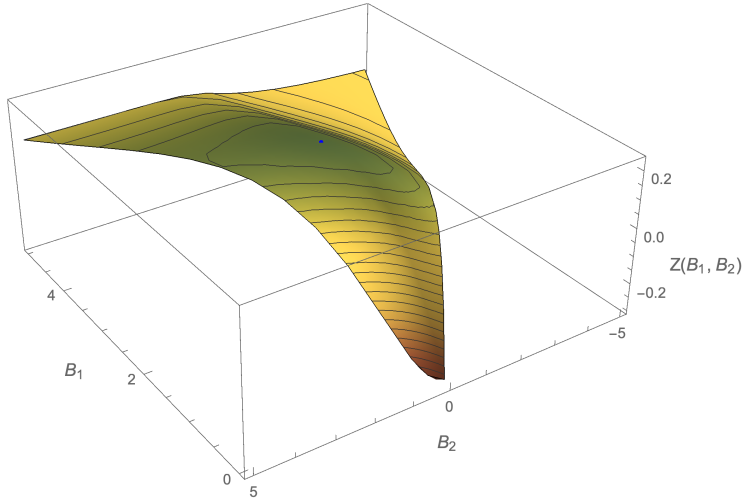


Figure 5.8. The Sharpe ratio for a trading strategy with barriers at B_1 and B_2, and unit transaction cost, with a distribution of alphas drawn from the standardized Laplace distribution. The positive maximum Sharpe ratio is marked with a blue dot.

Sharpe ratio is still defined and positive, but the surface itself is much broader and the peak less distinct.[22] The locations of the $\hat{B}_1$ and $\hat{B}_2$ barriers have also both moved in a positive direction.

5.4.14. The Sharpe ratio maximizing strategy

Both Figures 5.7 and 5.8 show the existence of a choice of barriers, $(\hat{B}_1, \hat{B}_2)$, that maximize the Sharpe ratio, $Z(B_1, B_2)$. The numerical values of these parameters cannot, in general, be expressed analytically, because of the dependence of the profit function on partial moments of the distribution of the alpha that may not have simple functional form.[23] In addition, in the real world, such distributions may not be knowable precisely *at all*. Furthermore, the results presented above have been computed for a particular transaction cost,

[22]Both charts are displayed from the same viewpoint.
[23]For example, the cumulative distribution function of the normal distribution is not computable analytically.

$\kappa = \omega_t$. This is not likely to be a universal constant as transaction costs vary from trader to trader even if all traders have precise knowledge of the true distribution of the alpha.

5.4.15. Numerical and heuristic solutions for various distributions of alphas

The surfaces presented in Figures 5.7 and Figure 5.8 on the previous page have been computed for a particular value of the transaction cost, $\kappa = \omega_t$. Due to the functional form of equation 5.64 on page 121, the particular value of R does not affect the *location* of the optimal barriers $(\hat{B}_1, \hat{B}_2)$, but the distribution of the alpha does. For the trader, it is important to understand how these values scale with K. To do this, we can compute *numerical* solutions and then infer a heuristic rule that describes functional dependence observed in these solutions.

The normal distribution: For the normal distribution, Figure 5.9 shows the trade entry barrier, $\hat{B}_1(K)$, rising close to linearly from an intercept of $\hat{B}_1 \approx 0.6$ while the trade exit barrier, $\hat{B}_2(K)$, falls close to linearly from the same value. Comparison of the numerical solutions indicates that $\hat{B}_1(K) - \hat{B}_2(K) = 2K$ to machine precision. The average, $(\hat{B}_1 + \hat{B}_2)/2$ appears to be slowly trending towards zero from above as K increases. $\hat{B}_2(K)$ is negative for $K > 0.5$. For "large" transaction costs, calculations indicate that $(\hat{B}_1, \hat{B}_2) \to (+K, -K)$.

The Laplace distribution: For the Laplace distribution, Figure 5.10 on page 126 shows the trade entry barrier, $\hat{B}_1(K)$, also rising close to linearly from an intercept of $\hat{B}_1 \approx 0.6$ but rising twice as fast as for the normal distribution and showing more curvature to the eye. In contrast, $\hat{B}_2(K)$ appears to be decaying towards zero from above. The average is an increasing function of K, but the difference between the values, $\hat{B}_1(K) - \hat{B}_2(K)$, is also equal to $2K$ to machine precision. The "large" transaction costs appear to possess the limit $(\hat{B}_1, \hat{B}_2) \to (2K, 0)$.

A heuristic solution that is reasonably accurate: Based on the observations above, and the inference that solutions for the generalized error distribution will lie somewhere between them, I propose

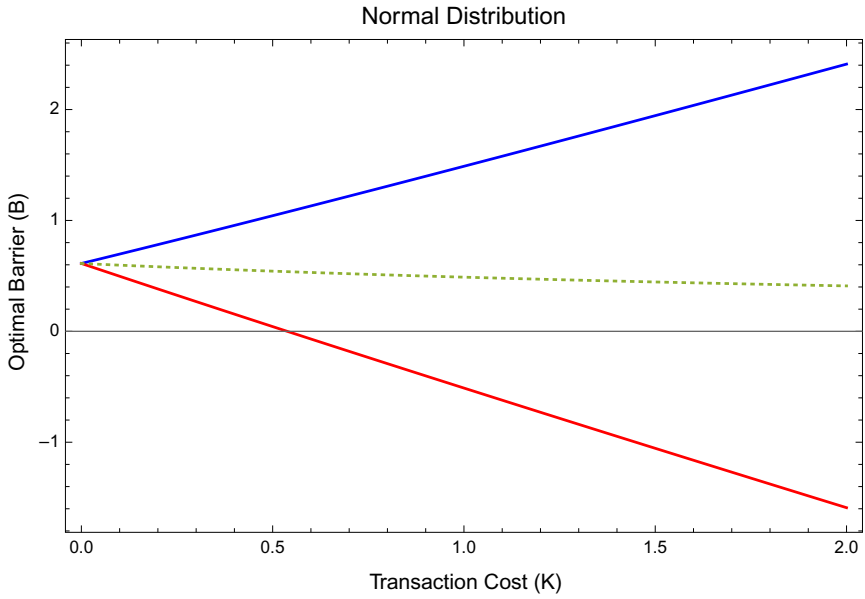


Figure 5.9. The location of the optimal trade entry and exit barriers, $(\hat{B}_1, \hat{B}_2)$, as a function of the transaction cost, K, for a standard normal distribution of alphas. The blue line is $\hat{B}_1(K)$ and the red line is $\hat{B}_2(K)$. The dotted line is their mean.

the following functional forms to be used as *heuristic* solutions to the location of the optimal barriers:

$$\hat{B}_1(K) = \zeta(K) + K \tag{5.65}$$

$$\text{and} \quad \hat{B}_2(K) = \zeta(K) - K, \tag{5.66}$$

where $\zeta(K)$ is a some slowly varying function of K which, in many circumstances, may be replaced with either the constant value of 0.6 or a linear function with that intercept.

For the "constant zeta" version of the heuristic, the corresponding values in units of returns are as follows:

$$\hat{b}_1(\kappa) = \omega_t \zeta(0) + \kappa \tag{5.67}$$

$$\text{and} \quad \hat{b}_2(\kappa) = \omega_t \zeta(0) - \kappa, \tag{5.68}$$

where $\zeta(0) \approx 0.6$. Such an algorithm is illustrated in Figure 5.11 on the next page.

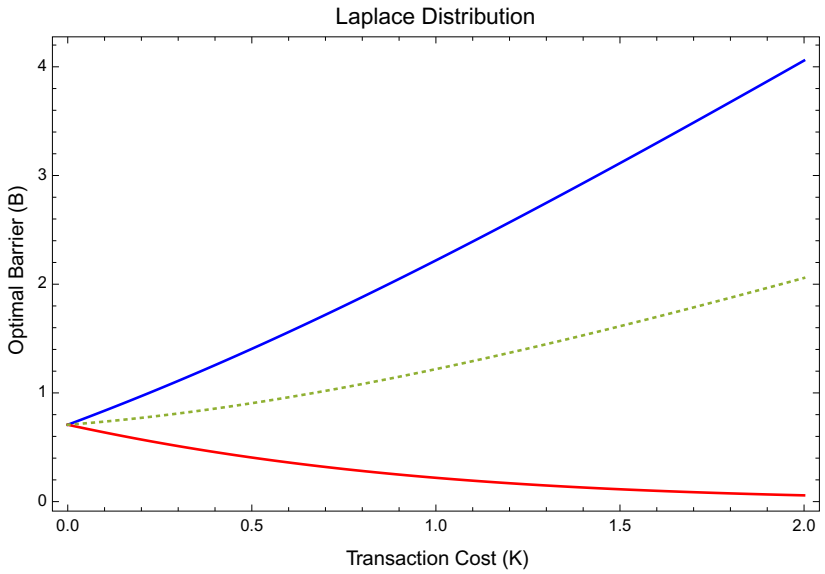


Figure 5.10. The location of the optimal trade entry and exit barriers, $(\hat{B}_1, \hat{B}_2)$, as a function of the transaction cost, K, for a standardized Laplace distribution of alphas. The blue line is $\hat{B}_1(K)$ and the red line is $\hat{B}_2(K)$. The dotted line is their mean.

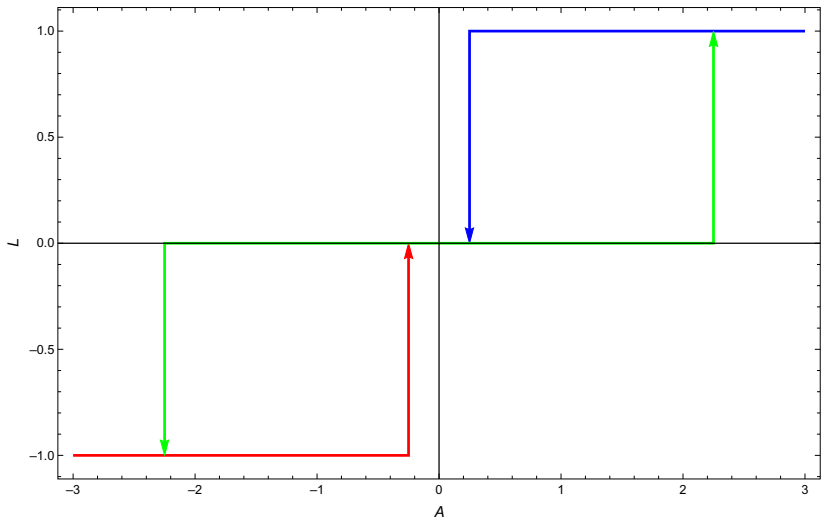


Figure 5.11. Graphical representation of a potential two barrier algorithm for a Laplace-*like* distribution. The specific locations of the branches in the holding function depend on the transaction costs.

5.5. Conclusions

In this essay, I have exhaustively analyzed barrier trading algorithms. I have shown that the effect of risk aversion[24] is to move trade entry barriers away from zero, creating a threshold that the alpha must exceed in order for risk to be taken, and the effect of transaction costs is to split those barriers, introducing hysteresis in the response function. Looked at from a signal processing point of view, the effect of risk aversion is to cause the trader to chose trades with higher signal to noise ratios and the effect of transaction costs is to cause the trader to filter out small changes in the signal that induce unnecessary churn. Although the mathematics we have gone through to get to these remarks has been quite lengthy, I feel that the results are satisfying because they are intuitively sensible *once we know the solution*. As in much of my work, I have shown that the use of leptokurtotic distributions, instead of the normal distribution, has consequential effects on the trader who wishes to pursue an optimal strategy. I caution the reader to ignore these effects at the peril of their bank balance.

[24]Maximizing the Sharpe ratio.

This page intentionally left blank

Essay 6

Ex Post Analysis

6.1. The Value of Counterfactuals

Decision-making under uncertainty is a discipline in which time's arrow is an ever present, and critical, factor. All of the holding functions described in Essays 3 and 5 have been the result of a causal decision process with three distinct divisions:

(1) What do I know now?
(2) What do I expect about the future?
(3) How should I act now to optimize my expectations of some metric of my future performance?

Once time has passed, we are able to look back at the performance that resulted from the causal decision process and compare it to our expectations of performance. The result of this analysis is a statistic that, in the context of machine learning methods, is sometimes called "training loss." The difference between our expected Sharpe ratio of, say, 2.5 ± 1.2 and our experienced Sharpe ratio of 1.2, for example.

Many traders like to talk about "money left on the table," comparing actual performance not to our prior expectations of performance during the period just realized but the "best possible" outcome for such a period. However, sophisticated quants are trained to dismiss such introspection as "counterfactuals" [54] and so, essentially, irrelevant to the measurement of performance.[1] Yet it is the exploration

[1] And indicative of a *lack* of sophistication of the trader.

of such counterfactuals, or the ideas of what could have been, that lead us to modify our methods and improve them. Any disciplined trader should engage in regular *post mortems* of trades to understand why performance was as experienced.[2]

This aversion to counterfactual analysis, however, doesn't extend to other forecasting disciplines. In many retail businesses, the response to the statement

> we made $100 on the deal

would be to ask the question

> well what's the best we could have done?

If the answer is $105, we look like a much better tradesman than if the answer were $500.

6.1.1. Counterfactuals in trading strategy analysis

In finance, the question of "what's the best we could have done" is often difficult to nail down. If I bought 100 shares at $85, and the market closed at $95 after touching $120, then one could argue that I left $2,500 of potential profits on the table by not closing out my long position and shorting into the close. However, would I have done that trade in the same size as the first, and what if I'd bought 200 shares not 100? There are many potential outcomes and the "best possible" performance is to always make an unlimited amount of money. However, that is not a feasible objective to optimize against, and for the concept of counterfactuals to be useful, the outcomes have to be feasible.

Fortunately, looking back to the *Hierarchy of Objective Functions* discussed in Section 2.5.4 on page 46, there are only two cases where considerations of risk are entirely absent, namely gross and net profit maximization without risk limits, and these are the ones that present divergent solutions. All other frameworks will deliver positions that are bounded within finite limits, either explicitly or probabilistically.[3] As nobody has the capacity to stake an unlimited amount of capital on a trade, there is not much value in examining that edge case.

[2]Without restricting such analysis to just "bad trades."
[3]Technically, I should say "almost certainly."

In this essay, I examine the problems of figuring out what the "best possible performance" of a trader is and what to do with that information. I also add consideration of a strategy for training a machine learning algorithm, such as an *artificial neural network*, to replicate the holding function that generates this best possible performance based on causally available data. This is a new procedure that was excluded to us based on prior methodologies and computational capacities, and turns the approach adopted in this book on its head by starting with *ex post* knowledge of holdings. This is an exciting area for current research.

6.1.2. Counterfactuals in linear regression

We can look at the humble R^2 as a measure of performance relative to a counterfactual:

$$R^2 = \frac{\text{Variance of model}}{\text{Variance of target}}. \tag{6.1}$$

As it measures the proportion of the variance explained by our model, the difference between R^2 and 100% is a quantification of the loss in performance of our chosen model relative to the counterfactual of a perfect model. To be more specific, consider a more symbolic definition: in the linear model

$$y_t = \alpha_t + \beta x_t + \varepsilon_t, \tag{6.2}$$

we derive estimates of the parameters, $\hat{\alpha}$ and $\hat{\beta}$, the use of which leads to estimates for each observation

$$\hat{y}_t = \hat{\alpha} + \hat{\beta} x_t, \tag{6.3}$$

and so equation 6.2 is, of course, just an example of a linear additive noise model (equation 2.20 on page 34)

$$y_t = \hat{y}_t + \varepsilon_t. \tag{6.4}$$

With these definitions, we can write R^2 symbolically as

$$R^2 = \frac{\sum_{t=1}^{T} (\hat{y}_t - \bar{\hat{y}})^2}{\sum_{t=1}^{T} (y_t - \bar{y})^2}. \tag{6.5}$$

This statistic must lie between 0% and 100%. One can think of linear regression as a procedure to compute the conditional mean of y when

the conditioning variable is x and, in that framework, the worst possible model would arise when the model is equal to the *unconditional* mean of the dependent variable i.e. when

$$\hat{y}_t = \bar{y} | x_t = \bar{y}. \tag{6.6}$$

With this condition, it's clear to see that[4]

$$\hat{y}_t = \bar{y} \qquad \Rightarrow \qquad \hat{y}_t = \bar{\hat{y}} \qquad \Rightarrow \qquad R^2 = 0. \tag{6.7}$$

By definition as the ratio of sums of squares, it cannot possibly take a lower value and so this is the least useful model we can construct.

6.1.3. Backtesting

Consider a time series of prices separated into three distinct regions by times $\{1, B, L, T\}$:

(i) the *backtesting* period, *in-sample* or *training set*, $t \in [1, B]$,
(ii) the *out-of-sample*, or *testing set*, $t \in (B, L)$,
(iii) the *live trading* period $t \in [L, T]$.

Of course, there's no reason why T must be limited above in general; trading may proceed indefinitely if $T \to \infty$.

The *Backtester's Assumption* is that the parameters chosen to optimize trading performance *in-sample* represent the right parameters to achieve that goal *out-of-sample*. Symbolically,

$$\begin{aligned} &\arg\max_{\boldsymbol{\theta} \in \mathcal{W}} Z(\{\boldsymbol{r}_t, \boldsymbol{h}(\mathcal{I}_{t-1} \backslash \boldsymbol{\theta}, \boldsymbol{\theta})\}_{t \in [1,B]}) \\ &\quad = \arg\sup_{\boldsymbol{\theta}} Z(\{\boldsymbol{r}_t, \boldsymbol{h}(\mathcal{I}_{t-1} \backslash \boldsymbol{\theta}, \boldsymbol{\theta})\}_{t \in [L,T]}), \end{aligned} \tag{6.8}$$

where $\boldsymbol{\theta}$ is a set of parameters that control the trading strategy and $\mathcal{W}$ represents the subset of the domain of valid parameters that has been searched for a maximum by the strategy developer. In equation 6.8, I have written the holding function in its most abstract, but causally legitimate, form possible, but with the parameters separated

[4]This is actually true for any *constant* model, but this is the *particular* constant model that would be delivered by minimizing the residual sum-of-squares of equation 6.2 on the previous page.

from the information set for the purposes of exposition even though, rightly, they are part of the information set[5] i.e.

$$\boldsymbol{h}(\mathcal{I}_{t-1} \setminus \boldsymbol{\theta}, \boldsymbol{\theta}) \equiv \boldsymbol{h}(\mathcal{I}_{t-1}). \tag{6.9}$$

As a side point, I should note that this formalism is *exactly* that advocated by estimation procedures such as ordinary least-squares and maximum likelihood, although it is most explicitly drawn out in discussions of the likelihood principle as advocated by Fisher *et al.* [13]. For statistical procedures, the statements rely on the idea that there is a unknown linking principle to the data, as manifested in the parameters of a data generating process, that is common between the *in-sample* and *out-of-sample* data. Therefore, efficient and unbiased estimation of those parameters leads to the best possible model for the future, yet to be observed, data.

6.1.4. Solving the right problem

The discussion presented in Essays 3 and 5 heavily features the partition of the information set $\mathcal{I}_{t-1}$ into data that specifies the expected distribution of returns e.g. $\boldsymbol{\alpha}_t$ and V_t, the parameters that control the strategy and other, irrelevant, data i.e.

$$\mathcal{I}_{t-1} = \{\boldsymbol{\alpha}_t, V_t, \dots\} \cup \{\boldsymbol{\theta}\} \cup \{\dots\}. \tag{6.10}$$

With this partition, *Peter Muller's rule* [43] is leading us to assert a restricted version of the backtester's assumption

$$\begin{aligned} &\arg\max_{\boldsymbol{\theta} \in \mathcal{W}} Z(\{\boldsymbol{r}_t, \boldsymbol{h}(\boldsymbol{\alpha}_t, V_t \dots, \boldsymbol{\theta})\}_{t \in [1,B]}) \\ &\quad = \arg\sup_{\boldsymbol{\theta}} Z(\{\boldsymbol{r}_t, \boldsymbol{h}(\boldsymbol{\alpha}_t, V_t \dots, \boldsymbol{\theta})\}_{t \in [L,T]}), \end{aligned} \tag{6.11}$$

which is the form discussed in Essay 1. In words, this rule says that, given an estimated model for the conditional distribution of returns, the best trading *out-of-sample* will be achieved by using a holding function that is "controlled" by parameters set to the values discovered by *in-sample* optimization of the performance of a trading simulation.

[5] I have written it less generally elsewhere.

In his book *The Nature of Statistical Learning Theory* [59], Vapnik writes the following:

> one should solve the problem directly and never solve a more general problem as an intermediate step.

Yet this is exactly what we are doing when we follow Peter Muller's rule: first we build a model for expected returns and then we use that model to chose trades. However, as a trader, we don't *really care* what the expected returns *actually are* apart from their use as a vehicle to permit portfolio selection. We seek the optimal holdings *out-of-sample*, and any causally legitimate recipe that allows us to construct them is a valid one.

Symbolically, Vapnik's version of the holding function needs make no mention of expected returns or other measures of the distribution of future returns, and the goal is not to obtain the holding function that performs the best *out-of-sample* but merely the set of holdings that performs the best *out-of-sample*. Vapnik reminds us that the real assumption we should be making is

$$\begin{aligned} &\underset{\boldsymbol{h}(\mathcal{I}_{t-1})\in\mathcal{H}}{\arg\max}\, Z(\{\boldsymbol{r}_t, \boldsymbol{h}(\mathcal{I}_{t-1})\}_{t\in[1,B]}) \\ &\quad = \underset{\boldsymbol{h}(\mathcal{I}_{t-1})}{\arg\sup}\, Z(\{\boldsymbol{r}_t, \boldsymbol{h}(\mathcal{I}_{t-1})\}_{t\in[L,T]}) \end{aligned} \tag{6.12}$$

with the optimization over the space of all causally legitimate holding functions, $\mathcal{H}$, where $\boldsymbol{h}(\mathcal{I}_{t-1}) \in \mathcal{H}$, and not over the space of parameters $\mathcal{W}$, where $\boldsymbol{\theta} \in \mathcal{W}$, which represents a smaller,[6] more restricted set of parametric holding functions.

6.1.5. The leverage of a search

Of course the reason why we adopt proxies like Peter Muller's rule is because we cannot possibly enumerate the entire space of potential holding functions and so the statement "search for all potential $\boldsymbol{h}(\cdot)$

[6] Since it must be true that there are holding functions that are not "parametric" in nature, it seems clear that $|\mathcal{W}| < |\mathcal{H}|$.

that maximizes $Z(\cdot)$" is not feasible to execute.[7] It is only by restricting ourselves to the smaller parametric space, $\mathcal{W}$, that the problem becomes feasible and some kind of solution may be obtained.

Similar to the development presented in Essay 1, define the leverage of the search space over the score statistic by the extremal range within the search space. That is,

$$\begin{aligned}\Delta Z(\mathcal{H}) = &\max_{\boldsymbol{h}(\mathcal{I}_{t-1})\in\mathcal{H}} Z(\{\boldsymbol{r}_t, \boldsymbol{h}(\mathcal{I}_{t-1})\}_{t\in[1,B]}) \\ &- \min_{\boldsymbol{h}(\mathcal{I}_{t-1})\in\mathcal{H}} Z(\{\boldsymbol{r}_t, \boldsymbol{h}(\mathcal{I}_{t-1})\}_{t\in[1,B]})\end{aligned} \tag{6.13}$$

for functional searches *in-sample*, and similarly

$$\begin{aligned}\Delta Z(\boldsymbol{\theta}) = &\max_{\boldsymbol{\theta}\in\mathcal{W}} Z(\{\boldsymbol{r}_t, \boldsymbol{h}(\boldsymbol{\alpha}_t, V_t \ldots, \boldsymbol{\theta})\}_{t\in[1,B]}) \\ &- \min_{\boldsymbol{\theta}\in\mathcal{W}} Z(\{\boldsymbol{r}_t, \boldsymbol{h}(\boldsymbol{\alpha}_t, V_t \ldots, \boldsymbol{\theta})\}_{t\in[1,B]})\end{aligned} \tag{6.14}$$

for parametric searches. Interestingly, these quantities are not merely analogues of each other because, with certainty, there exists a set of positions such that

$$\{\hat{\boldsymbol{h}}_t\}_{t\in[1,B]} = \arg\max_{\{\boldsymbol{h}_t\}} Z(\{\boldsymbol{r}_t, \boldsymbol{h}_t\}_{t\in[1,B]}) \tag{6.15}$$

and, by definition, it must follow that

$$\hat{\boldsymbol{h}}(\mathcal{I}_{t-1}) = \hat{\boldsymbol{h}}_t \text{ where } \hat{\boldsymbol{h}}(\mathcal{I}_{t-1}) = \arg\max_{\boldsymbol{h}(\mathcal{I}_{t-1})\in\mathcal{H}} Z(\{\boldsymbol{r}_t, \boldsymbol{h}(\mathcal{I}_{t-1})\}_{t\in[1,B]}). \tag{6.16}$$

Similarly, there must exist

$$\{\check{\boldsymbol{h}}_t\}_{t\in[1,B]} = \arg\min_{\{\boldsymbol{h}_t\}} Z(\{\boldsymbol{r}_t, \boldsymbol{h}_t\}_{t\in[1,B]}) \tag{6.17}$$

and

$$\check{\boldsymbol{h}}(\mathcal{I}_{t-1}) = \check{\boldsymbol{h}}_t \text{ where } \check{\boldsymbol{h}}(\mathcal{I}_{t-1}) = \arg\min_{\boldsymbol{h}(\mathcal{I}_{t-1})\in\mathcal{H}} Z(\{\boldsymbol{r}_t, \boldsymbol{h}(\mathcal{I}_{t-1})\}_{t\in[1,B]}). \tag{6.18}$$

[7] Because the set is not enumerable.

Thus the leverage of a functional search space over the score statistic, equation 6.13 on the previous page, may be explicitly computed *in-sample* and is given by

$$\Delta Z(\mathcal{H}) = Z(\{\boldsymbol{r}_t, \hat{\boldsymbol{h}}_t\}_{t\in[1,B]}) - Z(\{\boldsymbol{r}_t, \check{\boldsymbol{h}}_t\}_{t\in[1,B]}). \tag{6.19}$$

The same cannot be said of $\Delta Z(\boldsymbol{\theta})$. However, it is straightforward to see that

$$0 \leq \Delta Z(\boldsymbol{\theta}) \leq \Delta Z(\mathcal{H}) \tag{6.20}$$

since the set of all parametric holding functions is strictly a subset of the set of all possible holding functions. Thus the value $\Delta Z(\mathcal{H})$ tells the strategy designer the extent to which their work might impact their chosen performance statistic and the ratio

$$0 \leq \frac{\Delta Z(\boldsymbol{\theta})}{\Delta Z(\mathcal{H})} \leq 100\% \tag{6.21}$$

reveals how much of their value they are able to capture from the parametric family of strategies examined *in-sample*.

6.1.6. The performance of the counterfactually perfect strategy

In the above, I have introduced the idea that, *in-sample*, a trader can always find the holdings which deliver the global maximum performance, as measured by their choice of score statistic.[8] In current times, this previously burdensome exercise in mathematical programming may be achieved in many cases via online resources and open-source packages.[9] This means that the value of the statistic may also be computed for that set of "counterfactually perfect" holdings:

$$\hat{Z} = Z(\{\boldsymbol{r}_t, \hat{\boldsymbol{h}}_t\}_{t\in[1,B]}). \tag{6.22}$$

[8]Omitting "trivial" score statistics, such as constant Z, and those that create unbounded strategies.

[9]Such as `scipy.optimize` [60] in Python.

6.1.7. The quality of an optimum

When engaged in a parameter search *in-sample,* the trader can compute both $\hat{Z}$ and $\Delta Z(\mathcal{H})$, and also some estimate of $\Delta Z(\boldsymbol{\theta})$.[10] It is important not only that the parameter have *leverage* over the statistic but also that the leverage be meaningful, in that it is unlikely that it has arisen by chance. To assess that probability, knowledge of the sampling distribution of the score statistic is required, such as the sampling distribution of the Sharpe ratio as discussed in Section 1.3 on page 8.

As in Essay 1, the *quality* of a discovered optimum may be defined by

$$Q(\hat{\boldsymbol{\theta}}) = \frac{\Delta Z(\boldsymbol{\theta})}{\sigma_Z(\hat{\boldsymbol{\theta}})}, \tag{6.23}$$

where $\sigma_Z(\hat{\boldsymbol{\theta}})$ is the sample standard error of the statistic. If $Q < 3$, the trader should treat the leverage observed with a large degree of skepticism, since the chances of it having arisen by chance are fairly high.[11] In addition, however, the existence of this counterfactually perfect strategy permits knowledge of the upper limit to the quality of that statistic:

$$\hat{Q} = \frac{\Delta Z(\mathcal{H})}{\sigma_Z(\hat{Z})} \geq Q(\hat{\boldsymbol{\theta}}) \geq 0. \tag{6.24}$$

6.2. Optimal Trading Oracles in Theory

An *oracle* is a prescient being that provides the forecaster with accurate information about what would otherwise be unknown future realizations of the target process under study. Suppose that we had access to an oracle that would reveal to us the entire sequence of prices both *ex ante* and *ex post* with respect to our decision time t. From this insight, we might construct an *oracular sequence* of holdings,

$$\mathcal{O}(T) = \{\hat{\boldsymbol{h}}_t\}_{t\in[1,T]} = \arg\max_{\{\boldsymbol{h}_t\}} \Omega(\{\boldsymbol{r}_t, \boldsymbol{h}_t\}_{t\in[1,T]}), \tag{6.25}$$

[10] It is likely an estimate because numerical optimization *in-practice* may not represent an *exhaustive* search over the parameter space.

[11] Figure 1.4 on page 24 illustrates such a low quality optimum.

that deterministically maximizes a chosen objective function, $\Omega(\cdot)$, over the entire price sequence as revealed by the oracle. It is straightforward to see that this oracular sequence is a function of the entire sequence of returns, as the oracle has access to the final information set $\mathcal{I}_T$ that contains all information relevant to decisions made at any time $t \leq T$. This is, of course, the same set of holdings that were described as "counterfactually perfect" in Section 6.1.6 on page 136 when evaluated for the purposes of grading a backtest. That sequence would be $\mathcal{O}(B)$.

6.2.1. Using oracular sequences for forecasting

The oracular sequence would clearly act as an important guide to a trader attempting to construct an algorithmic method to trade assets under study. If the trader could construct a methodology that would generate the oracular sequence, or something close to it, based solely and properly upon the causally available information set $\mathcal{I}_{t-1}$ before time t, then they could confidently assert that they have *solved* the problem of trading the price sequence, as there are not any other strategies worth considering. An analyst with access to the oracular sequence for a given series could use it as the *training set* for a machine learning algorithm constructed to trade based upon the causal information set i.e. they may be interested in minimizing some metric, $M(\cdot)$, of the errors between the holding function adopted by a trader and the position chosen by the oracle

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} M(\{\boldsymbol{h}(\mathcal{I}_{t-1} \setminus \boldsymbol{\theta}, \boldsymbol{\theta}) - \hat{\boldsymbol{h}}_t\}_{t\in[1,B]}). \tag{6.26}$$

In adopting this approach, the trader has abandoned Peter Muller's rule, as they are no longer interested in forecasting returns. Instead, they are strictly interested in forecasting the oracular sequence of holdings which is the only thing that it is absolutely necessary for a trader to do and, as Vapnik suggests, is the approach that should be taken.

Note that the objective function, $\Omega(\cdot)$, under which the oracular trade sequence has been chosen, might differ from the score function, $Z(\cdot)$, used to evaluate performance both *in-sample* and

out-of-sample, although it doesn't have to. It is also defined over a *training set* $t \in [1, B]$, as before, with the reserved data $t \in (B, L)$ used for a *testing set*.[12] There are many sensible reasons as to why a trader might chose that $\Omega(\cdot)$ and $Z(\cdot)$ differ in their functional forms, one of which being the introduction of "Bayesian priors" as to the relationships between parameters in the *training set* versus the *testing set*. One such prior might be to bias hedge ratios, such as the β onto market returns, towards unity. The use of such "priors" means that the strategy for the *testing set*, even if optimal with respect to $Z(\cdot)$, might fail to be optimal with respect to $\Omega(\cdot)$.

However, a key difference between the approach advocated here and that taken in prior essays is that the focus is now to forecast the oracular position sequence and so any changes to the alpha model that don't actually result in different holdings are not relevant.

6.2.2. Irrelevance of *degenerate* oracular sequences

In general, it might not be true that there is a unique oracular sequence, as we may be indifferent between some set of distinct sequences $\{\mathcal{O}_i(T)\}$ based on the values of the objective function computed for them. However, since we *are* indifferent under our metric between those sequences, meaning

$$\Omega(\{\boldsymbol{r}\}_{t\in[1,T]}, \mathcal{O}_i(T)) = \Omega(\{\boldsymbol{r}\}_{t\in[1,T]}, \mathcal{O}_j(T)) \quad \forall\, i, j, \tag{6.27}$$

there is no loss that results from considering only the first discovered oracular sequence as if it were unique. Any subsequently discovered distinct sequences may be described as "degenerate," to follow the language of quantum mechanics [51]. This $\mathcal{O}_1(T)$ may not be the *only* oracular sequence that exists but *none are better* as far as $\Omega(\cdot)$ is concerned! In this context, it is important to note that this sequence is dependent solely on set of returns $\{\boldsymbol{r}\}_{t\in[1,T]}$ and the metric function $\Omega(\cdot)$ and is not dependent in any way in which an actual trader or trading system might go about choosing *their* positions, $\boldsymbol{h}(\mathcal{I}_{t-1})$,

[12]These periods are, exactly, the *in-sample* and *out-of-sample* data noted previously. Here I use the terminology favored by the community to emphasize the direction these thoughts are taking.

or any forecasting methodology, $\boldsymbol{\alpha}(\mathcal{I}_{t-1})$, that might be used to predict the future price sequence given the current information set $\mathcal{I}_t$.

6.2.3. Constraints on the oracular sequence of positions

For the trades proposed by an oracle to be instructive to real world traders, we must ensure that the oracle follows a strategy that is sufficiently constrained as to be realistic. As previously discussed, the objective "make as much money as possible" is not useful as there are a large number of sequences that may be defined all of which may be compactly summarized as "take an infinite position in the right direction" and none of which bear resemblance to a practically realizable strategy. Instead of this cartoon strategy, we must instead say the following: given transaction costs and risk limits, find the strategy with the best possible net profit, for example. Thus, as in Essay 5, I will assume that the maximum positions are strictly limited by an external risk manager i.e. $|\hat{\boldsymbol{h}}_t| \leq \boldsymbol{L}$, with the requirement that inequalities be satisfied *elementwise* for portfolio constraints. For simplicity, I also assume that before trading begins there is no initial position and that the strategy is required to deliver a net flat position at the end of the trading sequence i.e. $\hat{\boldsymbol{h}}_0 = \mathbf{0}$ and $\hat{\boldsymbol{h}}_B = \mathbf{0}$ and that this risk limit is given by the unit vector, $\mathbf{1}$, except for the extremal elements.[13]

6.2.4. Various strategy performance metrics

In addition to position constraints, a trader must also choose a suitable metric to allow the oracular strategy to actually be chosen from the set of all feasible position sequences for the *training set*.[14] In this section, I will enumerate some of the different choices associated with various well known metrics and demonstrate how to obtain solutions for them. For simplicity of exposition, this discussion will be based on trading a single asset and the position limits imposed exogenously

[13] It is simple to adapt to other assumptions.

[14] Which may be finite in size, or countably infinite, or uncountably infinite, depending on the particular process under study.

will be $|h_t| \leq 1$. The reader should be able to generalize to portfolio solutions straightforwardly.

Risk-limited maximum gross profit: The gross profit metric is

$$\Omega_{\text{gp}}(\{r_t, h_t\}_{t\in[1,B]}) = \sum_{t=1}^{B} h_t r_t \tag{6.28}$$

and the associated oracular sequence is

$$\mathcal{O}_T^{\text{gp}} = \{\hat{h}_t\} = \arg\max_{\{h_t\}} \begin{cases} \Omega_{\text{gp}}(\{r_t, h_t\}_{t\in[1,B)}) \\ \text{such that} \quad h_0, h_B = 0 \\ \text{and} \quad |h_t| \leq 1 \quad \forall\, t \in [1, B) \end{cases} . \tag{6.29}$$

It is immediately obvious that the solution to equation 6.29 is

$$\hat{h}_t = \operatorname{sgn} r_t \;\forall\, t \in [1, B). \tag{6.30}$$

That is, the oracle takes a maximum risk position in the direction of the return about to occur, at every trade time.
This can be seen to arise from the fact that

$$\frac{\partial \Omega_{\text{gp}}}{\partial h_t} = r_t \quad \text{and} \quad \frac{\partial^2 \Omega_{\text{gp}}}{\partial h_t^2} = 0. \tag{6.31}$$

Since the first derivative of the objective w.r.t. any potential holding is equal to the return during that period, the oracle will seek to increase the magnitude of the holding in the direction of the return. Since the second derivative is zero, this desire never changes, as the magnitude of the position is increased, and so the holding increases until it binds upon the exogenous position limits. Since

$$\frac{\partial^2 \Omega_{\text{gp}}}{\partial h_s \partial h_t} = 0 \tag{6.32}$$

there is no interaction between any of the holdings in the oracular position sequence, and so the linear programming problem of equation 6.29 is solvable by treating all variables separately. Therefore, the oracular sequence is that given by equation 6.30.[15]

[15]This discussion mirrors the description of linear programming given in Essay 2.

Risk-limited maximum net profit: The net profit metric is related to the gross profit metric of equation 6.28 on the preceding page by the subtraction of the net cost of trading. If this cost is κ per unit of holding,[16] then

$$\Omega_{\mathrm{np}}(\{r_t, h_t\}_{t\in[1,B)}) = \Omega_{\mathrm{gp}}(\{r_t, h_t\}_{t\in[1,B)}) - \kappa \sum_{t=1}^{B} |h_t - h_{t-1}| \quad (6.33)$$

with the same constraints on h_t. Qualitatively, we see that the solution will essentially be similar to that of equation 6.30 on the previous page with the exception that the number of transitions between long and short states (i.e. between $h_{t-1} = +1$ and $h_t = -1$ and between $h_{t-1} = -1$ and $h_t = +1$) will be reduced due to the existence of sequences for which the cost of trading is more than the profit that may be extracted by trading.

This function, equation 6.33, is piecewise linear in the holdings, $\{h_t\}$, but the first derivative may be discontinuous around the vertices $\{(h_{t-1}, h_t)\}$. As before, solution to such problems is well known from the discipline of linear programming to lie on the surface of the polytope defined by the feasible solutions and, specifically, on the vertices of that surface.

Risk-Penalized Metrics: In addition to the piecewise linear functions of equation 6.28 on the previous page and equation 6.33, we may augment the risk neutral $\Omega_{\mathrm{rn}}(\cdot)$ functions (representing either metric discussed above) with a simple risk penalty such as

$$\Omega_{\mathrm{rp}}(\{r_t, h_t\}_{t\in[1,B)}) = \Omega_{\mathrm{rn}}(\{r_t, h_t\}_{t\in[1,B)}) - \lambda \sum_{t=1}^{B} r_t^2 h_t^2 \quad (6.34)$$

to encourage risk-averse behavior. Note that this modification to the objective function can coexist with the previously defined exogenous risk limits, but that it may also allow them to be relaxed. Technically, this change converts the problems into quadratic programming problems from linear programming problems. Qualitatively, if λ is sufficiently large, it moves the solution off the surface of the simplex

[16] Such as per share and per bond.

defined by the constraints to a point within that simplex. The consequences of this solution for the realized trading strategy is to make the solution adjust its position size to take account of the size of the available profit i.e. it will no longer be bound by the risk limits at all times and may sometimes take positions of smaller size or which reduce risk in the absence of sufficiently available profits.

Scale-free metrics: Solutions may also be crafted for scale-free metrics such as the Sharpe ratio[17] of net trading profits and other similar functions. These choices will require nonlinear numerical optimization techniques for which no closed-form solution is easily expressed, but there is no reason why they cannot be used on real data. If the Sharpe ratio itself is chosen as the metric used to choose the oracular sequence, then a byproduct of the solution will be knowledge of the "best possible" Sharpe ratio obtainable by trading the price process under investigation. In general, though, the trader might find numerical optimization of the Sharpe ratio without exogenous risk limits to be dynamically unstable as driving the risk close to zero will be achievable by "unrealistic" trade-offs between different assets in a multivariate optimization framework.

6.2.5. Methods of computing the oracular position sequences

For oracular sequences of positions to be useful to the trader who seeks to train a machine learning algorithm, it must be feasible to compute them in the real world. Fortunately, commodity computation now makes this possible in a way that it wasn't as recently as 10 years ago.[18]

Manual computation of the maximum gross profit sequence: Before connecting any kind of "black box" optimizer to data, and to provide a concrete example of what solutions look like and how they are modified as terms are added to increase the realism of the trader's model and ascend the hierarchy of optimization strategies, it is very instructive to go through such an optimization process *by hand*.

[17]Notwithstanding our discussion of the Sharpe ratio in Essay 1.
[18]The 2010s.

To do this, consider the price sequence for a hypothetical trading day where the trader is prohibited from holding a position overnight by exogenous policy. The prices for times $t \in \{0, 1, 2, 3, 4, 5, 6\}$ are

$$\{P_0 \dots P_6\} = \{\$9, \$10, \$11, \$14, \$11, \$12, \$10\} \tag{6.35}$$

(prices are per share). Corresponding to this price sequence is a sequence of returns:

$$\{r_1 \dots r_6\} \approx \{+11\%, +10\%, +27\%, -21\%, +9\%, -17\%\}. \tag{6.36}$$

With position sizes limited to $-1 \leq h_t \leq +1$, the maximum gross profit is obtained from the position sequence

$$\{\hat{h}_0 \dots \hat{h}_6\} = \{0, +1, +1, +1, -1, +1, 0\}. \tag{6.37}$$

This sequence satisfies the rule that it begin and end with a zero position and be bounded between ± 1. The *ex ante* position, taken immediately after the beginning of the interval that ends at time t, is equal to the sign of the *ex post* gains, and this sequence delivers a gross profit of \$9.

Manual computation of the maximum net profit sequence: However, with an added transaction cost of \$1/share, and as we are requiring that $h_0 = h_6 = 0$, this sequence only delivers a net profit of \$3. In comparison, the associated maximum net profit position sequence is

$$\{\hat{h}_0 \dots \hat{h}_6\} = \{0, +1, +1, +1, -1, 0, 0\}. \tag{6.38}$$

This delivers a lower gross profit of \$8, with a higher net profit of \$4.

Examining the differences between equations 6.37 and 6.38, we see that the effect of the transaction cost is to selectively veto a subset of the trades for which the marginal rate of return does not exceed the cost of trading. It removes the *whipsaw* of taking $\hat{h}_4 = -1$ between $\hat{h}_3 = +1$ and $\hat{h}_5 = +1$ choosing to go flat with $\hat{h}_5 = 0$ and hold that position until the close of trading at $t = 6$, where both oracles are *required* to go flat with $\hat{h}_6 = 0$. This is illustrated in Table 6.1 on the next page.

Table 6.1. Oracular trading sequences for maximum gross profit (m.g.p.) and maximum net profit (m.n.p.) optimizers.

		$\hat{h}_t$	
t	P_t	m.g.p.	m.n.p.
0	\$ 9	0	0
1	\$10	+1	+1
2	\$11	+1	+1
3	\$14	+1	+1
4	\$11	−1	−1
5	\$12	+1	0
6	\$10	0	0
Gross		\$9	\$8
Net		\$3	\$4

Note: Data are for a synthetic "trading day" with transaction costs of \$1 per share and holdings limited to one share long or short.

Understanding the differences between trade sequences is instructive. The maximum net profit (m.n.p.) oracle chose to go flat before the end of this hypothetical trading day because it had foreknowledge that it needed to be flat at the end of the day and the round-trip cost of chasing the whipsaw made it uneconomical to do that, whereas the maximum gross profit oracle (m.g.p.) always follows the market. The m.n.p. oracle underperforms the m.g.p. oracle on a gross profit basis but the situation is reversed on a net profit basis. Optimizing the "wrong" objective, such as the cost free one when costs are present, results in underperformance for the oracle.

Solution as a combinatorial optimization problem: For metrics *without* a risk penalty, from the analysis discussed earlier in this book, we know that the solution will only ever involve the three possible trade positions, $h_t \in [-1, 0, +1]$, and never an intermediate size. Thus, in general, there are 3^{B-1} possible trade sequences to examine for $t \in [1, B)$. Optimization of equation 6.33 on page 142 may be approached by simply enumerating all possible position sequences, scoring them, and selecting the maximum scored sequence. As referred to in Section 6.2.2 on page 139, if there is more than one

solution, we may pick whichever of them we wish — we are indifferent between them.

However, although this naïve solution strategy is legitimate, it is operationally challenging. For just one month of daily trading, there are 3^{20}, or approximately three billion,[19] combinations of positions to consider. For a C.P.U. operating at around 3 GHz, which takes perhaps 20×5 clock cycles to compute the score of each trade sequence, this equates to around two minutes of computation. However, the exponential growth in problem complexity is so rapid that adding just three trading days more results in a computation that takes just under an hour and modelling a month and a half will take more than twice that time itself! On this basis, such solutions are feasible but not usable. Of course, the problem is parallelizable, so can potentially be divided among thousands[20] of C.P.U.s, but ultimately such *brute force* techniques will become inoperable as a practical methodology in the real world. Another computational strategy is needed.

Solution as a linear program: This type of *combinatorial explosion* is well known within the Operations Research community. The *Simplex Algorithm* of Danzig [44] is a method that may be used to solve linear optimization problems with thousands of parameters in entirely reasonable amounts of time and has been in use since the 1940s.[21] The earliest algorithm resembles that described in Section 2.5.2 on page 42 and is efficient because it exploits the fact that the solution will never lie anywhere *other than* on the vertices of the polytope defined by constraints.[22] This dramatically reduces the size of the space that must be searched. Equation 2.48 on page 48 showed elements of how linear programming could be used to find solutions to forward looking decisions, whereas here I will examine how these methods can be used to compute the best possible trade sequence on historic data.

[19] Exactly 3,486,784,401.
[20] Depending on the budget of the trader.
[21] The methods were developed by Danzig while he was working in the U.S. Army Air Force during the Second World War [1].
[22] And may restrict the polytope to be a simplex.

Net profit maximization as a linear program: The objective function for a *net profit maximizing trading oracle* is a piecewise linear function of the position sequence. If we choose to bound the positions both above and below, at ± 1, then the problem is expressible as a linear program and is solvable efficiently via standard methods. This is the form in which it is expressed in equation 6.33 on page 142.

A useful strategy for solving the problem is to think not in terms of positions but in terms of the trades done to arrive at the position. Let b_t and s_t represent quantities of an asset bought and sold at time t, and require that both of these terms be non-negative.[23] Then we have

$$h_t = \sum_{u=1}^{t} (b_u - s_u) \quad \text{where } b_u \geq 0 \quad \text{and} \quad s_u \geq 0 \tag{6.39}$$

as a linear constraint on the position h_t in terms of the trade variables b_t (number bought) and s_t (number sold). This is a "dual"[24] to the original problem in which the goal is to find $\{b_t, s_t\}_{t \in [1,B]}$ rather than $\{h_t\}_{t \in [1,B]}$.

We may also represent the objective entirely in terms of the buy and sell trades, which eliminates the need to include the absolute value function. The total gross profit for a set of trades is

$$\sum_{t=1}^{B} b_t (P_B - P_{t-1}) - \sum_{t=1}^{B} s_t (P_B - P_{t-1}) \tag{6.40}$$
$$\text{where } b_t \geq 0 \; \mathrm{s_t} \geq 0 \; \forall \; t \in [1, B).$$

Remembering that the indexing of positions refers to positions taken immediately after the beginning of the interval ending at time t, this expression gives the profit in terms of buy and sell trades marked from their purchase price to the final prices in the data set.[25]

[23] This constraint is necessary for the representation of the problem to be unique.

[24] Technically, the word *dual* is used to describe the linear program that results from exchanging the roles of the variables and constraints, which is possible, rather than this change-of-variables approach. This is why I'm using the term in quotes.

[25] These worked examples are done in terms of *price changes* rather than *returns* to decrease the notational clutter. There is no practical difference, either way, in the methods followed.

At this point, it makes sense to switch to a matrix notation. Let $\boldsymbol{x}$ be a vector of dimension $2B$ made by stacking the buy and sell trades alternately, and $\boldsymbol{r}$ be a similar vector made by stacking the trade-to-period end profits. Then

$$\boldsymbol{x} = \begin{pmatrix} b_1 \\ s_1 \\ b_2 \\ s_2 \\ \vdots \\ b_B \\ s_B \end{pmatrix} \geq \mathbf{0} \quad \text{and} \quad \boldsymbol{r} = \begin{pmatrix} P_B - P_0 \\ P_0 - P_B \\ P_1 - P_B \\ P_B - P_1 \\ \vdots \\ P_B - P_{B-1} \\ P_{B-1} - P_B \end{pmatrix} \tag{6.41}$$

and the total gross profit is simply $\boldsymbol{r}^T\boldsymbol{x}$. The cost of trading into all of these positions is

$$\sum_{t=1}^{B} \kappa(b_t + s_t) \qquad \Leftrightarrow \qquad \boldsymbol{k}^T\boldsymbol{x} \quad \text{where } \boldsymbol{k} = \kappa\mathbf{1} \tag{6.42}$$

and the cost of trading out, to ensure the final position is $h_B = 0$, is also equal to this quantity. The final objective for a net profit maximizing oracle is then

$$\Omega(\boldsymbol{x}) = (\boldsymbol{r} - 2\boldsymbol{k})^T\boldsymbol{x}. \tag{6.43}$$

An important piece of composing this expression for the transaction costs is to realize that, due to the linearity of $\Omega(\boldsymbol{x})$ and the non-negativity of of the trade sizes, there do not exist any times t where both $b_t > 0$ and $s_t > 0$ so there is no concept of a "net trade" of $b_t - s_t$ with cost $\kappa|b_t - s_t|$. b_t will only be positive when $P_B - P_{t-1} > \kappa$, and $P_{t-1} - P_B < 0$ in that circumstance, so $b_t > 0 \Rightarrow s_t = 0$ and $s_t > 0 \Rightarrow b_t = 0$ by the same reasoning. At all times, the result of the linear objective is to set either or both of b_t and s_t to the lower bound of zero. This is due to the negative costs associated with non-zero values, and so an oracle will not offer a solution where $(b_t + \delta, s_t + \delta)$ is viewed as equivalent to (b_t, s_t) for some non-zero δ. On that basis, the final trade, done just after time $B - 1$ to flatten the position h_{B-1}, must also be either purely a buy trade, $b_B = -h_{B-1}$ and $s_B = 0$ if $h_{B-1} < 0$, or purely a sell trade, $s_B = h_{B-1}$ and $b_B = 0$ if

$h_{B-1} > 0$. In all circumstances, at all times t, the cost of trading will be either κb_t or κs_t and the absolute value operation is eliminated by the primary constraint of non-negativity in the solution, $\boldsymbol{x} \geq \mathbf{0}$.

After composing the objective to a linear program, we must compose the constraints. They are that the net position at any time lies between the limits ± 1. As the net position is the sum of the buy trades minus the sell trades

$$h_t = \sum_{u=1}^{t} (b_u - s_u), \tag{6.44}$$

this may be written as a vector constraint

$$\begin{pmatrix} +1 & -1 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ -1 & +1 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ +1 & -1 & +1 & -1 & \cdots & 0 & 0 & 0 & 0 \\ -1 & +1 & -1 & +1 & \cdots & 0 & 0 & 0 & 0 \\ \vdots & & & & \ddots & & & & \vdots \\ +1 & -1 & +1 & -1 & \cdots & +1 & -1 & 0 & 0 \\ -1 & +1 & -1 & +1 & \cdots & -1 & +1 & 0 & 0 \\ +1 & -1 & +1 & -1 & \cdots & +1 & -1 & +1 & -1 \\ -1 & +1 & -1 & +1 & \cdots & -1 & +1 & -1 & -1 \end{pmatrix} \begin{pmatrix} b_1 \\ s_1 \\ b_2 \\ s_2 \\ \vdots \\ b_{B-1} \\ s_{B-1} \\ b_B \\ s_B \end{pmatrix} \leq \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ \vdots \\ 1 \\ 1 \\ 0 \\ 0 \end{pmatrix} \tag{6.45}$$

or

$$A\boldsymbol{x} \leq \boldsymbol{L} \tag{6.46}$$

in matrix notation. As before there are two rows in the matrix for each constraint: one states that $h_t \leq 1$ and the other that $-h_t \leq 1 \Leftrightarrow h_t \geq -1$.

The position limit vector $\boldsymbol{L}$ is the unit vector in dimension $2B$ with the last two elements set to zero. Since the final two rows of the matrix expression represent

$$\sum_{t=1}^{B} (b_t - s_t) = +h_B \leq 0 \tag{6.47}$$

$$\text{and} \quad \sum_{t=1}^{B} (s_t - b_t) = -h_B \leq 0, \tag{6.48}$$

there is only one possible solution and that is $h_B = 0$, as required by the "trade to zero at time B" constraint.

In aggregate, the oracle's problem is to solve the linear program

$$\hat{\boldsymbol{x}} = \arg\max_{\boldsymbol{x}} \begin{cases} (\boldsymbol{r} - 2\boldsymbol{k})^T \boldsymbol{x} \\ \text{for} \quad \boldsymbol{x} \geq \boldsymbol{0} \\ \text{and} \quad A\boldsymbol{x} \leq \boldsymbol{L} \end{cases} \tag{6.49}$$

for the terms defined above. Although development may look daunting, once one is familiar with solving linear programs, it is actually very straight forward and equation 6.49 is literally a "textbook" problem. The objective is to maximize the total net profit arising from all the long trades and all the short trades done over the trading period. The primary constraints are that the trade sizes are non-negative, which is necessary to define them as "buy trades" and "sell trades." The secondary constraints are that the final position is zero and that the total net position never exceeds the risk limit in either direction. Together, this is sufficient information for the problem to be solved and for this problem there is always a feasible solution.[26]

Transforming between the holdings and trades spaces: To aid in development of these systems, it's helpful to understand how to transform between the "holdings space" and the "trades space" via matrix operations.

First, introduce the "running sum" matrix, S, that transforms from the trades vector, $\boldsymbol{x}$ of dimension $2B$, to an equivalent holdings vector, $\boldsymbol{h}$ of dimension B. The required matrix must have dimension $B \times 2B$ and is written as follows:

$$\begin{pmatrix} h_1 \\ h_2 \\ \vdots \\ h_{B-1} \\ h_B \end{pmatrix} = \begin{pmatrix} +1 & -1 & 0 & 0 & \cdots & 0 & 0 \\ +1 & -1 & +1 & -1 & \cdots & 0 & 0 \\ \vdots & & & & \ddots & & \vdots \\ +1 & -1 & +1 & -1 & \cdots & +1 & -1 \end{pmatrix} \begin{pmatrix} b_1 \\ s_1 \\ b_2 \\ s_2 \\ \vdots \\ b_B \\ s_B \end{pmatrix} \tag{6.50}$$

$$\Leftrightarrow \qquad \boldsymbol{h} = S\boldsymbol{x}. \tag{6.51}$$

[26]The solution $b_t = s_t = 0 \; \forall \; t \in [1, B]$ satisfies all the constraints of equation 6.49 and so is feasible.

Similarly, to transform between the "total returns" vector, $\boldsymbol{r}$ of dimension $2B$, and the desired "period returns" vector, $\boldsymbol{g}$ of dimension B, one needs a $B \times 2B$ "differences" matrix, D. This is defined by

$$\begin{pmatrix} P_1 - P_0 \\ P_2 - P_1 \\ \vdots \\ P_{B-1} - P_{B-2} \\ P_B - P_{B-1} \end{pmatrix} = \begin{pmatrix} +1 & 0 & -1 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & +1 & 0 & -1 & \cdots & 0 & 0 & 0 & 0 \\ \vdots & & & \ddots & & \ddots & & & & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & +1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & +1 & 0 \end{pmatrix}$$

$$\times \begin{pmatrix} P_B - P_0 \\ P_0 - P_B \\ P_B - P_1 \\ P_1 - P_B \\ \vdots \\ P_B - P_{B-1} \\ P_{B-1} - P_B \end{pmatrix}, \tag{6.52}$$

which is written in matrix notation as

$$\boldsymbol{g} = D\boldsymbol{r}. \tag{6.53}$$

The D matrix is a of dimension $B \times 2B$ with zeros everywhere except for $+1$ along the "diagonal" starting in the upper left element and with -1 along the "diagonal" starting on the third column of the first row. These elements do not appear in the (seldom used) matrix for $B = 1$, which is just

$$D = \begin{pmatrix} 1 & 0 \end{pmatrix}, \tag{6.54}$$

and the rightmost column of D is all zeros regardless of dimension.

Non-uniqueness of the D matrix: The nature of the $\boldsymbol{r}$ vector, as defined here, means that there are multiple ways to define a D matrix that will deliver the same "differencing" operation. In addition to

equation 6.52 on the previous page, an alternate form D' may also be defined as follows:

$$\begin{pmatrix} P_1 - P_0 \\ P_2 - P_1 \\ \vdots \\ P_{B-1} - P_{B-2} \\ P_B - P_{B-1} \end{pmatrix} = \begin{pmatrix} 0 & -1 & 0 & +1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & +1 & \cdots & 0 & 0 & 0 \\ \vdots & & & & \ddots & & \ddots & & & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & -1 & 0 & +1 \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & -1 \end{pmatrix}$$

$$\times \begin{pmatrix} P_B - P_0 \\ P_0 - P_B \\ P_B - P_1 \\ P_1 - P_B \\ \vdots \\ P_B - P_{B-1} \\ P_{B-1} - P_B \end{pmatrix} \tag{6.55}$$

$$\Leftrightarrow \qquad \boldsymbol{g} = D'\boldsymbol{r}. \tag{6.56}$$

In fact, any linear combination of D and D' will suitably difference $\boldsymbol{r}$:

$$\boldsymbol{g} = \frac{aD + bD'}{a+b}\boldsymbol{r} \quad \forall\, a, b \in \mathbb{R}, \quad a + b \neq 0. \tag{6.57}$$

Properties of the D and S matrices: The D, D', and S matrices are "inverses," in the sense that[27]

$$DS^T = I \qquad \Rightarrow \qquad DS^T D = D \tag{6.58}$$

$$\text{and} \quad SD^T = I \qquad \Rightarrow \qquad SD^T S = S. \tag{6.59}$$

These equations are true for both D and D' and represent the standard definitions of the generalized inverse in linear algebra [25]. Note that neither $D^T S$ nor $S^T D$ is equal to the identity matrix, I, since the matrices are not square.

[27] Technically, S^T is the generalized inverse of D and D^T is the generalized inverse of S.

Conservation of total profit: Since equation 6.53 on page 151 is a linear system, it may be solved by its generalized inverse:

$$\boldsymbol{g} = D\boldsymbol{r} \qquad \Leftrightarrow \qquad \boldsymbol{r} = D^g\boldsymbol{g}. \tag{6.60}$$

Equation 6.58 on the preceding page defines this generalized inverse, D^g, to be S^T; therefore

$$\boldsymbol{r} = S^T\boldsymbol{g}. \tag{6.61}$$

From this, it follows that the total profit is preserved during the "trades $\leftrightarrow$ holdings" transformation:

$$\begin{aligned} \boldsymbol{r}^T\boldsymbol{x} &= (S^T\boldsymbol{g})^T\boldsymbol{x} \\ &= \boldsymbol{g}^T S\boldsymbol{x} \\ &= \boldsymbol{g}^T\boldsymbol{h}. \end{aligned} \tag{6.62}$$

This is, of course, a necessary condition for both versions of the system solved to be equivalent.

Counterfactual risk aversion: In the counterfactual world, *ex ante* expectations are stripped away and their values replaced with the *ex post* observations. The expectations operator, which implies a sum or integral over potential values weighted by their probabilities, is replaced with a literal sum over observed values. This suggests that the counterfactually perfect risk-averse oracle should optimize an objective including a term such as

$$-\lambda\sum_{t=1}^{B} h_t^2(P_t - P_{t-1})^2 \tag{6.63}$$

or, even, proportional to the variance of profits within the training set[28]

$$-\lambda\left[\frac{1}{B}\sum_{t=1}^{B} h_t^2(P_t - P_{t-1})^2 - \left\{\frac{1}{B}\sum_{b=1}^{B} h_t(P_t - P_{t-1})\right\}^2\right]. \tag{6.64}$$

[28]Bessel's correction [30], $B/(B-1)$, omitted for clarity of exposition [49].

Equation 6.63 on the preceding page may be written in matrix notation with the introduction of a new matrix with the elements of $\boldsymbol{g}$ along the diagonal (and zero elsewhere):

$$G = \operatorname{diag} \boldsymbol{g}. \tag{6.65}$$

The quadratic risk aversion term of equation 6.63 on the previous page then becomes

$$\begin{aligned} -\lambda \sum_{t=1}^{B} h_t^2 (P_t - P_{t-1})^2 &= -\lambda (G\boldsymbol{h})^T G\boldsymbol{h} \\ &= -\lambda \boldsymbol{h}^T G^2 \boldsymbol{h} \\ &= -\lambda \boldsymbol{x}^T S^T G^2 S \boldsymbol{x}. \end{aligned} \tag{6.66}$$

The full, more variance-*like*, term would require the addition of a correction of the form

$$\begin{aligned} +\lambda \frac{(\boldsymbol{h}^T \boldsymbol{g})^2}{B} &= +\lambda \frac{\boldsymbol{h}^T (\boldsymbol{g}\boldsymbol{g}^T) \boldsymbol{h}}{B} \\ &= +\lambda \boldsymbol{x}^T \frac{S^T (\boldsymbol{g}\boldsymbol{g}^T) S}{B} \boldsymbol{x}, \end{aligned} \tag{6.67}$$

although the B in the denominator means that this term will rapidly become irrelevant for real-world sample sizes.

Risk-averse net profit maximization as an *LQ* program: Putting these terms together, an oracle that seeks to maximize net profits in a risk-averse manner must solve the following linear quadratic program:

$$\hat{\boldsymbol{x}} = \arg\max_{\boldsymbol{x}} \begin{cases} (\boldsymbol{r} - 2\boldsymbol{k})^T \boldsymbol{x} - \lambda \boldsymbol{x}^T Q \boldsymbol{x} \\ \text{for} \quad \boldsymbol{x} \geq \boldsymbol{0} \\ \text{and} \quad A\boldsymbol{x} \leq \boldsymbol{L}, \end{cases} \tag{6.68}$$

where

$$Q = S^T G^2 S \quad \text{or} \quad Q = S^T \left(G^2 - \frac{\boldsymbol{g}\boldsymbol{g}^T}{B} \right) S, \tag{6.69}$$

depending on the form of quadratic objective selected. Again, this is a classically formed problem which may be solved, in practice, by a wide variety of software tools.

Overview: In summary, almost all of the systems defined by the *Hierarchy of Objective Functions* in Essay 2 may be expressed as a mathematical programming problem to be solved *ex post* if one wishes to know the counterfactually perfect sequence of trades or holdings that a trader *should* have taken during an historic period. Of course, it is also possible to discover the counterfactually worse sequence as well, replacing our oracle with a "demon of our own design" in the words of Bookstaber [6]. Thus, *ex post* we can absolutely grade any real trading strategy relative to its performance on the scale established by these two limits and, also, use these data to train a machine learning algorithm to execute trades for us.

6.3. Optimal Trading Oracles in Practice

Although this book is intended to be theoretical in nature, I feel that the value of the counterfactual approach can be appreciated when it is executed on real data, and for that we need an alpha, a trading strategy, and a data set.

6.3.1. The alpha model

As I have written in *Adventures in Financial Data Science* [20], I feel that an asymmetric GARCH$(1,1)$ model for returns, with innovations drawn from a fundamentally leptokurtotic distribution, represents a description of financial data that is durably accurate. That is, the process model:

$$\begin{aligned} r_t &= \mu + \varphi r_{t-1} + \sigma_t \varepsilon_t \\ \sigma_t^2 &= C + A r_{t-1}^2 + B\sigma_{t-1}^2 + D\mathbb{I}[r_{t-1} < 0] r_{t-1}^2 \\ \varepsilon_t &\sim \text{GED}(0, 1, \kappa). \end{aligned} \tag{6.70}$$

This model has scope for an alpha, in that the conditional mean return of the traded asset *within the training set* is given by

$$\alpha_t = \mathbb{E}_{t-1}[r_t | \mathcal{I}_B] = \hat{\mu}_B + \hat{\varphi}_B r_{t-1}. \tag{6.71}$$

This represents the conditional mean return calculated at time t by a trader who has access to the entire backtesting data set, $t \in [1, B]$,

and who uses that data to compute estimates of the process parameters μ and φ via an efficient and unbiased method. For this model, that would be by maximum likelihood. *Ab initio*, this is an *acausal* procedure because future information, $\mathcal{I}_B \setminus \mathcal{I}_{t-1}$, is used to produce these estimates $\hat{\mu}_B$ and $\hat{\varphi}_B$. To simulate a *causally valid* procedure, one should use

$$\alpha_t = \mathbb{E}_{t-1}[r_t|\mathcal{I}_{t-1}] = \hat{\mu}_{t-1} + \hat{\varphi}_{t-1} r_{t-1}. \tag{6.72}$$

However, *in-sample*, such usage would intermix estimation error that has temporal dependency[29] into the analysis. As this approach is intended to examine counterfactually perfect strategies, we should also use the most accurate estimates available for the parameters, which are those from the entire data sample.

6.3.2. Model parameters for the S&P 500 ETF (SPY)

The parameters of this model, $\{\mu, \varphi, A, B, C, D, \kappa\}$, are estimated by maximum likelihood, and code to obtain these estimates is available online at my *GitHub* repository [21]. For daily returns of the S&P 500 Index tracking E.T.F. `SPY` [55] during the period 01/03/2019–12/31/2021, inclusive, I obtain the estimates exhibited in Table 6.2 on the facing page.

Although the parameters of the variance process and distribution of innovations are massively significantly different from their null hypothesis values of homoskedastic normal innovations, their specific values do not concern us here. The sole purpose of using this, more complex, model is to obtain more accurate estimate of the returns model to use *in-sample*. That model is

$$\alpha_t = \hat{\mu}_B + \hat{\varphi}_B r_{t-1} = 0.1377 - 0.0890 r_{t-1} \tag{6.73}$$

and so represents a mean reversive process that is mostly long during the data period.

[29] Since the variance of the sampling errors of the parameter estimates scales inversely with sample size.

Table 6.2. Estimated parameters for the predictive model of equation 6.70 on page 155 for the period 01/03/2019–12/31/2021.

Parameter	Estimate	Std. Err.	t statistic	p value
μ	0.1377	0.0022	6.267	Negl.
φ	−0.0890	0.0042	−2.134	3.3×10^{-2}
C	0.0580	0.0017	3.396	6.8×10^{-4}
A	0.1127	0.0056	1.999	4.6×10^{-2}
B	0.6873	0.0057	12.106	Negl.
D	0.3545	0.139	2.550	1.1×10^{-2}
κ	0.7160	0.0515	4.194	1.4×10^{-5}

Note: For $\hat{\kappa}$, the t statistic and p value are quoted based on the distance from the normal distribution value of 1/2.

6.3.3. Trading strategies

In addition to possessing an alpha model, a trader also needs a trading strategy: the recipe of going from $\alpha_t \to h_t$ that is the subject of most of this book. Here I will consider a risk-limited gross profit maximizing trader that is benchmarked against a simple *buy-and-hold* strategy. That is,

$$h_t = \operatorname{sgn} \alpha_t \tag{6.74}$$

for the trader, and, simply,

$$h_t = 1 \quad \forall\, t \in [1, B) \tag{6.75}$$

for the buy-and-hold benchmark. For completeness, I will also consider the *ex ante* mean–variance optimal strategy, or Markowitz function, with $\lambda = 1/2$ after Thorp [58]:

$$h_t = \frac{\alpha_t}{\sigma_t^2}. \tag{6.76}$$

6.3.4. Transaction costs

For simplicity, and because it is no longer so unrealistic for a retail trader, I will neglect transaction costs.

6.3.5. Score statistic

For a score statistic, despite all my commentary in Essay 1 and elsewhere, I will compute the simple daily Sharpe ratio and use that to evaluate the two strategies.[30]

As is my usual practice, I will assume the risk-free rate is not relevant to the decision processes of a daily cadence trader. This may not be correct for long-term investors but it is a perfectly acceptable approximation for daily cadence traders. For the data period studied, the *effective* value of the Federal Funds Target Rate began at around 2.4% but was dropped to less than 0.1% in March 2020, where it was held for the rest of the period.[31] At these rates, the daily return should be reduced by around $2.4\%/360 = 0.0066\%$ at the beginning of the period and 0.0003% during the pandemic. These values hardly have any effect when the standard deviation of daily returns is around 1.4% for the S&P 500.

6.3.6. Counterfactual strategies

The actual, causally feasible,[32] trading strategies will be compared to the counterfactually perfect benchmarks of

(i) maximum gross profit or

$$\{\hat{h}_t\}_{t\in[1,B]} = \underset{\{h_t\}_{t\in[1,B]}}{\arg\max} \sum_{t=1}^{B} h_t r_t \tag{6.77}$$

and

(ii) mean–variance optimal or[33]

$$\{\hat{h}_t\}_{t\in[1,B]} = \underset{\{\boldsymbol{h}_t\}_{t\in[1,B]}}{\arg\max} \left\{ \sum_{t=1}^{B} \left(h_t r_t - \frac{h_t^2 r_t^2}{2} \right) + \frac{1}{2B} \left(\sum_{t=1}^{B} h_t r_t \right)^2 \right\}. \tag{6.78}$$

[30] I will, however, compute standard errors for the Sharpe ratios.

[31] Due to the Coronavirus pandemic that began, as far as monetary policy was concerned, at that time.

[32] *Out-of-sample*, with the appropriate choice of μ_B and φ_B.

[33] In code, I actually use `Series.mean()` and `Series.var()` methods in Pandas [40], which includes Bessel's correction.

Here, the market price of risk, λ, has again been taken to be the asymptotic Kelly optimal value of 1/2.

6.3.7. Performance of the trading strategies

The performance of the three trading strategies is illustrated through their accumulated total return in Figure 6.1. One can immediately see that the performance of the "sign-of-alpha" strategy, of equation 6.74 on page 157, essentially matched the "buy-and-hold" investor during 2019 but really took off during the Coronavirus pandemic period in 2020, and continued to outperform afterwards. Unfortunately, the "Markowitz/Kelly" strategy delivers an order of magnitude less investment size, so its own, relatively good versus buy-and-hold, performance is not apparent from this chart.

However, because I am using the Sharpe ratio as the evaluation score, and that is scale-free, this performance gap will not feature in the ranking of the strategies based on that statistic. The full data are

Figure 6.1. Accumulated, *in-sample*, total profit of the strategies under test for the period 2019–2021. Blue is the "sign of alpha" algorithm of equation 6.74 on page 157, black is "buy-and-hold," and red is the Markowitz/Kelly mean–variance optimal strategy of equation 6.76.

Table 6.3. Performance data for three causally feasible investment strategies for the SPY ETF.

Strategy	Mean	Std. Dev.	Kurtosis	Sharpe Ratio	Std. Err.
Buy-and-hold	0.1016	1.3816	19.7078	1.1671	0.5874
Markowitz/Kelly	0.0182	0.1829	10.5415	1.5818	0.5871
Sign of alpha	0.1897	1.3723	19.5040	2.1944	0.6108
Max gross profit	0.8433	1.0987	30.2288	12.1846	1.4999
Mean–variance optimal	0.6840	0.5306	3.1623	20.4626	0.8565

Note: Data are in percent for daily returns from 2019 to 2021 inclusive and is computed with zero transaction costs.

given in Table 6.3, which gives *in-sample* Sharpe ratios of 1.2 ± 0.6 for the "buy-and-hold" strategy, 1.7 ± 0.6 for the "Markowitz/Kelly" mean–variance optimal strategy, and 2.2 ± 0.6 for the much simpler "sign of alpha" strategy. Thus both active strategies are seen to represent an improvement over buy-and-hold, although the leverage of the parameter "choice of strategy" is not great, at $\Delta Z = 1.0$, and the quality is low at $Q = 1.7$. One can see that the performance differences, however, are dramatic even though, truth be told, we can't really distinguish between these three approaches on the basis of Sharpe ratio alone.

It is also important to point out that, even though I am quoting a standard error for the Sharpe ratio computed via equation 1.29 on page 14, it is quite likely that the sampling errors for these three measures are *not* independent as they are derived from what is basically the same data. This reasoning follows from the assumption of weak stochastic alphas discussed in Essay 2. Therefore forming a t test between the three strategies is likely to be biased upwards in significance as the effective sample size is not three times the sample size of each individual strategy.[34]

6.3.8. Relative performance versus counterfactuals

Turning to the counterfactual strategies, their performances are unsurprisingly considerably better: with a Sharpe ratio of 12.2 ± 1.5

[34] And is not significant, anyway.

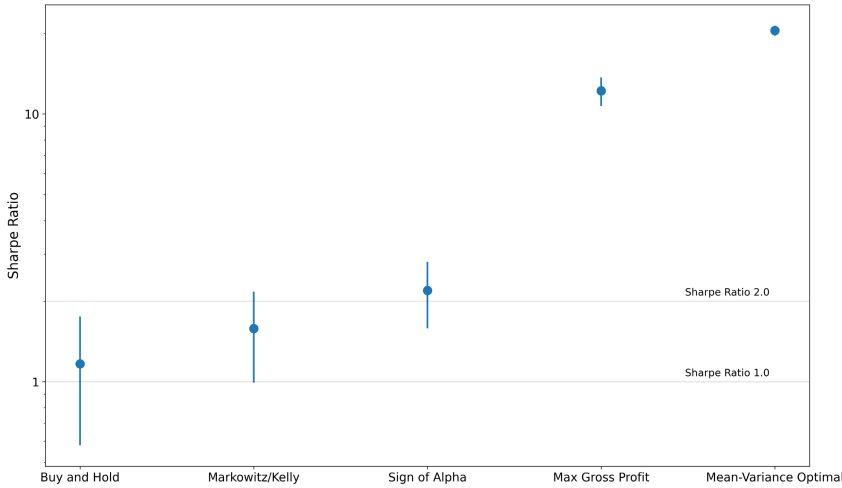


Figure 6.2. Relative performance *in-sample* of three trading strategies and their counterfactually perfect analogues. Data are the Sharpe ratio computed from daily returns of the SPY E.T.F. from 2019 to 2021 inclusive and is computed with zero transaction costs.

for the maximum gross profit counterfactual and 20.5 ± 0.9 for the mean–variance optimal counterfactual. As mere mortals, we may think that an *in-sample* Sharpe ratio of over 2 is a big deal, but our oracle begs to differ, telling us that this strategy could have done as much as six times better!

The data are exhibited in Figure 6.2, and the reader should not take from this that I am implying that there is the potential to actually achieve these performance levels in the real world. Far from it, but it serves to illustrate that neither the profit maximizer nor the portfolio theorist should rest on their laurels of merely beating buy-and-hold with their strategy. This should serve as a stimulus to inquire as to *why they are not doing better.*

6.3.9. Performance analysis from the perspective of machine learning

The existence of a well-defined *training set*, as the counterfactually perfect trade sequence for a given oracle, opens up new avenues to

Table 6.4. Contingency table, or confusion matrix, for the performance of the "sign-of-alpha" strategy versus the maximum gross profit counterfactual.

		Sign-of-alpha	
Strategy	Position	Short	Long
Max gross	Short	3.6%	38.0%
Profit	Long	2.1%	56.3%

Note: Data are the correspondence of positions in the SPY E.T.F. from 2019 to 2021 inclusive.

explore the origins of the performance losses exhibited above, especially with the tool kits provided by the machine learning community.

This method of attack is a current research focus for me. To give the reader a sample of what I'm talking about, consider the fact that both the "sign-of-alpha" strategy and the "max gross profit" oracle deliver outcomes that are at all times either entirely long or entirely short positions.[35]

This means that we can compute a contingency table, or *confusion matrix*, for the positions and use that as a tool to evaluate performance errors. Such an analysis is exhibited in Table 6.4, from which it can be seen that the majority of the errors made be the strategy are to be long when it should have been short. In the language of classification problems, the algorithm has the following: a precision of 59.7%, meaning the proportion of times when it was long when it should have been long; a recall of 96.4%, meaning the proportion of times when the algorithm should have been long and actually was long; the accuracy is 59.9%; and the F_1 Score is 73.7%.

6.3.10. Exploration of the relationship between causal information and oracular trade sequences

With access to established oracular trade sequence provided, there is no reason why we cannot use *any* available inferential method

[35]Subject to the exogenous risk manager that requires $h_t \in \{-1, +1\}$ for both cases.

to understand the functional relationship between causally available information, such as an alpha, and the counterfactually perfect position that should be adopted by a trader interested in maximizing their performance score.

For example, a *k-Nearest Neighbors*, or *kNN*, analysis is one of many approaches that can be used to shed light on the functional form of the optimal holding function, $\hat{h}(\alpha_t)$. This is exhibited in Figure 6.3, with the neighborhood chosen to be of size $k = \lfloor \sqrt{B} \rfloor$. A logistic regression is also illustrated on the same figure.

At a larger scale, the figure shows a "logistic-*like*" response, which is not unsurprising, but it also appears to be exhibiting a lack off monotonicity in the region of an alpha of 0.01%. The fact that it does not saturate to either limiting value (± 1) shows that there do not exist thresholds such that, above or below the threshold in alpha, the oracle is long or short almost certainly. The "jitter" in the curve, meaning the extent to which it is not smooth, is quite likely due to sampling variation in the estimator.

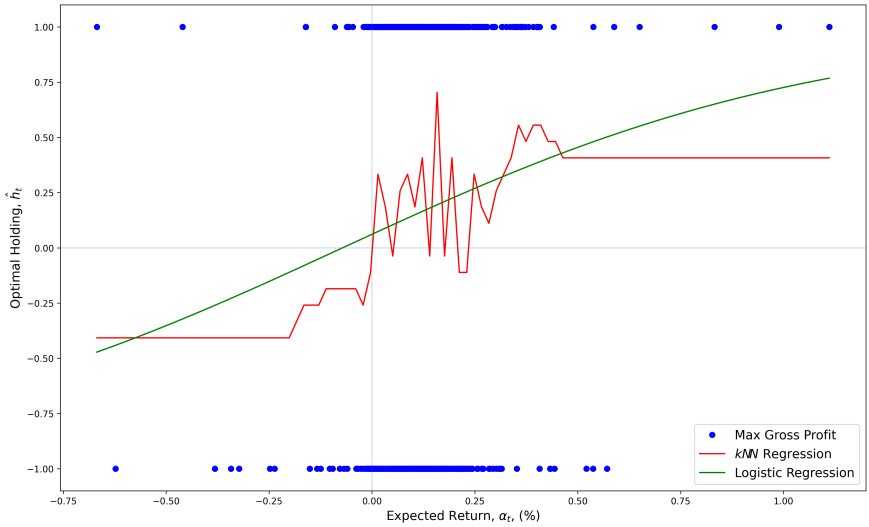


Figure 6.3. The *kNN* estimator of the functional form of the optimal holding function $\hat{h}(\alpha_t)$. The data are illustrated as blue dots, the *kNN* regression as a red line, and a logistic regression as a green line.

Summary: Hopefully the reader can see from the simple work presented here that opening our eyes to counterfactuals permits trading strategy development to enter a powerful new age where the workflow does not require severe approximations to be made in order that solutions be deliverable analytically. With a clean statement as to how a trader judges the relative merits of different trade sequences, it should be possible to derive the counterfactually perfect sequence, on that basis, for any given historical price sequence. With *that* data, we can proceed to train machine learning algorithms to reproduce that sequence from causally available information. This is a distinctly different approach to that in which an amorphous bag of data is fed into a black box "A.I." algorithm in the hope that that may be used to maximize the Sharpe ratio *in-sample* without any reflection as to what the output trade sequence *ought* to look like.

6.4. End Note

The work in this essay was stimulated by a request from a "well-known and large" hedge fund to reconsider the process of portfolio manager evaluation. I made a proposal to them that was the seeds of the ideas that grew into the work presented here but, unfortunately, the executive who would have sponsored the work left before the ideas were fully developed and so I never did, actually, present the work to them in the more rigorously thought out manner with which it is presented here. The prompt was "find a way to think about portfolio manager performance that differs from what we currently do." That workstream was essentially the following:

(1) compute the Sharpe ratios for each manager,
(2) do some *ad hoc* inspection of "bad trades" to learn about how they could have been made less problematic.

I think this approach is common to a lot of funds, and I strongly believe that a more rigorous approach is necessary and that comparison of experienced performance to counterfactuals may open the door to such methods.

References

[1] Donald J. Albers and Constance Reid. An Interview with George B. Dantzig: The Father of Linear Programming. *The College Mathematics Journal*, 17(4):292–314, 1986.

[2] George B. Arfken and Hans J. Weber. Mathematical Methods for Physicists, Academic Press Inc., San Diego, 1985.

[3] William J. Barber. Theorizing about Macro-Economic Instability: Monetary Equilibrium (Versions of 1932, 1933, and 1939). In *Gunnar Myrdal: An Intellectual Biography*, pp. 24–37. Palgrave Macmillan, New York, 2008.

[4] Daniel Bernoulli. Exposition of a New Theory on the Measurement of Risk. *Econometrica*, 22(1):23–36, 1954. Translated from Latin into English by Dr. Louise Sommer, The American University, Washington, DC, from "Specimen Theoriæ Novæ de Mensura Sortis" in *Commentarii Academiae Scientiarum Imperialis Petropolitanæ, Tomus V* [Papers of the Imperial Academy of Sciences in Petersburg, Vol. V], 1738, pp. 175–192.

[5] John R. Birge and Francois Louveaux. *Introduction to Stochastic Programming*. Springer Science & Business Media, New York, 2011.

[6] Richard M. Bookstaber. *A Demon of Our Own Design: Markets, Hedge Funds, and the Perils of Financial Innovation*. John Wiley, New York, 2007.

[7] Max Born and Emil Wolf. *Principles of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction of Light*. Cambridge University Press, Cambridge, 2020.

[8] George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung. *Time Series Analysis: Forecasting and Control*. John Wiley & Sons, New York, 2015.

[9] John C. Cox, Stephen A. Ross, and Mark Rubinstein. Option Pricing: A Simplified Approach. *Journal of Financial Economics*, 7(3):229–263, 1979.

[10] Richard Dobbins, Stephen F. Witt, and John Fielding. *Portfolio Theory and Investment Management*. Blackwell Business, Oxford, 1994.

[11] Robert Eisberg and Robert Resnick. *Quantum Physics of Atoms, Molecules, Solids, Nuclei, and Particles*. John Wiley & Sons, New York, 1985.

[12] Eugene F. Fama and Kenneth R. French. The Capital Asset Pricing Model: Theory and Evidence. *Journal of Economic Perspectives*, 18(3):25–46, Aug. 2004.

[13] R. A. Fisher and Edward John Russell. On the Mathematical Foundations of Theoretical Statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222(594–604):309–368, 1922.

[14] Claudio Fontana. Personal Communication, 2023.

[15] Graham L. Giller. *The Construction and Analysis of a Whole-Sky Map Using Underground Muons*. PhD thesis, University of Oxford, 1994.

[16] Graham L. Giller. Approximate Statistical Properties of the Sharpe Ratio. Available at *SSRN 2361169*, 1997.

[17] Graham L. Giller. The Construction and Properties of Ellipsoidal Probability Density Functions. Available at *SSRN 1300689*, 2003.

[18] Graham L. Giller. *Adventures in Financial Data Science: The Empirical Properties of Financial Data and Some Other Things That Interested Me...* Giller Investments (New Jersey), LLC, Holmdel, 2020.

[19] Graham L. Giller. At Morgan Stanley We Found Simple Trading Rules Outperformed Fancy Portfolio Optimization. After Leaving, I Worked Out Why. *Medium: Adventures in Data Science*, February 2021. bit.ly/3HgN4S5.

[20] Graham L. Giller. *Adventures in Financial Data Science: The Empirical Properties of Financial and Economic Data*. World Scientific, Singapore, 2nd edition, 2022.

[21] Graham L. Giller. *GitHub* Repository, 2022. https://www.github.com/Farmhouse121.

[22] I. S. Gradsteyn and I. M. Ryzhik. *Handbook of Mathematical Functions*. Academic Press, New York, 1965.

[23] Clive Granger. *2003 Nobel Price Lecture: Time Series Analysis, Cointegration, and Applications*. Nobel Prize Foundation, Stockholm, 2003.

[24] Richard C. Grinold and Ronald N. Kahn. *Active Portfolio Management*. McGraw Hill, New York, 2000.

[25] David A. Harville. *Matrix Algebra from a Statistician's Perspective*. Taylor & Francis, New York, 1998.

[26] Paul Horowitz and Winfield Hill. *The Art of Electronics*. Cambridge University Press, Cambridge, 1989.

[27] John Hull *et al. Options, Futures and Other Derivatives*. Prentice Hall, Upper Saddle River, NJ, 2009.

[28] Harold Jeffreys. *The Theory of Probability*. Oxford University Press, Oxford, 1998.

[29] J. L. Kelly. A New Interpretation of the Information Rate. *The Bell System Technical Journal*, 35(4):917–926, 1956.

[30] M. Kendall, A. Stuart, J. Ord, and S. Arnold. *Kendall's Advanced Theory of Statistics*, Volume 1: Distribution Theory. Arnold, London, 1999.

[31] M. Kendall, A. Stuart, J. Ord, and S. Arnold. *Kendall's Advanced Theory of Statistics*, Volume 2A: Classical Inference and the Linear model. Arnold, London, 1999.

[32] Petter N. Kolm and Gordon Ritter. Multiperiod Portfolio Selection and Bayesian Dynamic Models. *Risk*, 28(3):50–54, 2014.

[33] Albert S. Kyle. Continuous Auctions and Insider Trading. *Econometrica: Journal of the Econometric Society*, 1315–1335, 1985.

[34] Michael Lewis. *Liar's Poker*. W.W. Norton & Company, New York, 2010.

[35] Fabrizio Lillo, J. Doyne Farmer, and Rosario N. Mantegna. Master Curve for Price-Impact Function. *Nature*, 421(6919):129–130, 2003.

[36] Andrew W. Lo. The Statistics of Sharpe Ratios. *Financial Analysts Journal*, 58(4):36–52, 2002.

[37] Roger Lowenstein. *When Genius Failed: The Rise and Fall of Long-Term Capital Management*. Random House trade paperbacks, 2000.

[38] K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis*. Academic Press, 1979.

[39] Harry M. Markowitz. Portfolio Selection. *Journal of Finance*, 7(1), 1952.

[40] Wes McKinney *et al.* Data Structures for Statistical Computing in Python. In *Proceedings of the Ninth Python in Science Conference*, Vol. 445, pp. 51–56, 2010.

[41] P. Muller. Proprietary Trading: Truth and Fiction. *Quantitative Finance*, 1(1):6–8, 2001.

[42] Peter Muller. Personal Communication, 1996.

[43] Peter Muller. Personal Communication, 1999.

[44] Katta G. Murty. *Linear Programming*. John Wiley & Sons Inc., New York, 1983.

[45] Gary W. Oehlert. A Note on the Delta Method. *The American Statistician*, 46(1):27–29, 1992.

[46] John Douglas Opdyke. Comparing Sharpe Ratios: So Where Are the p-Values? *Journal of Asset Management*, 8(5):308–336, 2007.

[47] Natalie Packham and Fabian Woebbeking. The London Whale. Available at *SSRN 3210536*, 2018.

[48] Gordon Ritter. Private Communication. LinkedIn chat.

[49] D. J. Rumsey. Let's Just Eliminate the Variance. *Journal of Statistics Education*, 17(3), 2009.

[50] Ludger Rüschendorf. On the Distributional Transform, Sklar's Theorem, and the Empirical Copula Process. *Journal of Statistical Planning and Inference*, 139(11):3921–3927, 2009. Special Issue: The Eighth Tartu Conference on Multivariate Statistics and The Sixth Conference on Multivariate Distributions with Fixed Marginals.

[51] Leonard I. Schiff. *Quantum Mechanics*. McGraw-Hill, New York, 1965.

[52] Jack D. Schwager. *The New Market Wizards: Conversations with America's Top Traders*. John Wiley & Sons, New York, 2012.

[53] Claude Elwood Shannon and Warren Weaver. *The Mathematical Theory of Communication*. University of Illinois Press, Urbana-Champaign, 1962.

[54] W. Starr. Counterfactuals. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Palo Alto, Summer 2021 edition, 2021. https://plato.stanford.edu/archives/sum2021/entries/counterfactuals/.

[55] State Street Global Advisors. SPDR® S&P 500® ETF Trust, 2022. https://www.ssga.com/library-content/products/factsheets/etfs/us/factsheet-us-en-spy.pdf.

[56] Richard H. Thaler. Mental Accounting and Consumer Choice. *Marketing Science*, 27(1):15–25, 2008.

[57] Edward O. Thorp. *Beat the Dealer: A Winning Strategy for the Game of Twenty One*, Vol. 310. Vintage, New York, 1966.

[58] Edward O. Thorp. The Kelly Criterion in Blackjack, Sports Betting, and the Stock Market. In Ziemba, William T. and MacLean, Leonard C. and Thorp, Edward O., editors, *The Kelly Capital Growth Investment Criterion: Theory and Practice*, pp. 789–832. World Scientific, Singapore, 2011.

[59] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 1999.

[60] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C. J. Carey, İlhan

Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[61] Roger Waters. Your Possible Pasts, 1983. *The Final Cut*, track 2.

[62] Wikipedia. Partial Correlation, 2022. https://en.wikipedia.org/wiki/Partial_correlation.

This page intentionally left blank

Index