

Springer Texts in Business and Economics

Ilia Bouchouev

# Virtual Barrels

Quantitative Trading in the Oil Market



Springer

---

## **Springer Texts in Business and Economics**

Springer Texts in Business and Economics (STBE) delivers high-quality instructional content for undergraduates and graduates in all areas of Business/Management Science and Economics. The series is comprised of self-contained books with a broad and comprehensive coverage that are suitable for class as well as for individual self-study. All texts are authored by established experts in their fields and offer a solid methodological background, often accompanied by problems and exercises.

---

Ilia Bouchouev

# Virtual Barrels

Quantitative Trading in the Oil Market



Springer

Ilia Bouchouev  
Pentathlon Investments, LLC  
Westfield, NJ, USA

ISSN 2192-4333                   ISSN 2192-4341 (electronic)  
Springer Texts in Business and Economics  
ISBN 978-3-031-36150-0           ISBN 978-3-031-36151-7 (eBook)  
<https://doi.org/10.1007/978-3-031-36151-7>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG  
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Paper in this product is recyclable.

To our children, Vlad and Valeria

---

## Preface

The public perception of oil trading revolves around counting barrels, predicting outcomes of OPEC meetings, estimating gasoline and jet fuel consumption from high-frequency traffic measures, analyzing Middle Eastern politics, assessing the impact of regulations and trade sanctions, or hiring meteorologists to forecast the trajectory of a hurricane that may impact offshore production and refining centers.

These are fascinating topics to discuss at the dinner table where everyone has an opinion on some of numerous factors that impact the price of oil. But how relevant is this information for professional oil trading, given so much noise and almost no reliable facts that the market does not already know? More importantly, what are the channels that translate such fundamental information into the price action in the futures market where the price of oil is determined, and how do professional oil traders filter out the noise and turn the relevant bits into consistently profitable trading strategies?

The book attempts to answer these questions following the author's own experiences in managing an energy derivatives trading business for over twenty years. The language of the book is rather quantitative, reflecting the demands of modern markets for more data, cutting-edge technology, and advanced analytical modeling. The book tracks the evolution of quantitative oil trading throughout its entire history. Options play a special role throughout the book. With somewhat elevated barriers to entry, the business of options and volatility trading is where quants tend to shine. But, in contrast to many other financial markets, the dynamics of oil options can hardly be separated from the economics of the underlying physical system, the behavior of hedgers, and the crucial role played by storage.

The need for storage is what led to the beginnings of oil trading in the 1860s shortly after the process of oil drilling was discovered in Western Pennsylvania. Oil quickly proved its worth, but the demand for barrels at the well, constrained by complicated logistics, has not always moved in unison with supply. To buy some time and buffer fluctuations between supply and demand, oil had to be kept somewhere. The cheapest storage facility at the time were the dumps. Not surprisingly, the pioneers of oil trading, who were buying excess production at discounts and storing it, were called "dump men."

To move oil from the dumps to consumers, which was done largely with the help of horses, oil was poured into empty whiskey barrels. The whiskey itself helped to negotiate the price. The price of oil was measured in dollars per barrel. Only a

handshake agreement was needed to secure the deal, as honor and trust were held in higher regard than written contracts. The business of oil trading was built as an exclusive private club, largely closed to outsiders.

Little has changed in the first hundred years of the oil business, other than horses being replaced by pipelines and ships, and whiskey by a much wider selection of fine spirits. Membership in the private oil club was not open to newcomers until the 1990s when oil bosses brought in a breed of young scientists, the quants. Hiring geeks for the old-fashioned teams of oil traders was considered somewhat progressive and even entertaining. The mandate for quants was clear. The oil world was becoming more complex, and in a wider web of interlinked prices, they must search for a low-hanging fruit, the strategy called *arbitrage*.

An opportunity to make small but consistent profits from low-risk arbitrage trades had a particular allure for quants. Not only did it provide a more financially lucrative career alternative to teaching mathematics at a university, but it also gave quants a chance to prove that mathematics actually works in the real world. One challenge for quants was the relatively small size of the strategy, which was dwarfed by riskier directional wagers made by old-timers. In the oil market, the size of the bet and the amount of risk taken determine the trading hierarchy, and for a while, quants were largely viewed as second-class citizens in the oil world. To earn their proper seat at the table, more arbitrages had to be found. Fortunately, a new era of oil trading was about to start where quants were destined to play a much larger role.

The world was globalizing and digitalizing. To meet fast-growing oil demand, new sources of supply had to be found. The shale revolution in the US led to unprecedented growth in the global petroleum infrastructure. New energy assets were built with large risks held by financial investors. In contrast to prior decades, when oil supply growth was predominantly driven by sovereign-controlled entities, this time the risks were held by private capital and managed in the market. The risks were quite complex, requiring more sophisticated hedging instruments and more advanced quantitative skills in managing them.

At the same time, oil was also rapidly financializing and moving away from obscure voice negotiations to transparent electronic screens. The private club of oil trading was finally open to the public. Buying the new digital barrel of oil has proven to be particularly handy at times of rising inflation and geopolitical conflicts, when many other financial assets struggle. With new computer technologies coming in handy, quants went on a journey to modernize the oil industry. The journey was fully endorsed by the Wall Street marketing machine, which smelled the potential and scalability of the new business—the business of *oil derivatives*.

This book tells the inside story of this business. This story has been molded into its current form by the energy trading course taught by the author at the Courant Institute of Mathematical Sciences at New York University, and I thank my students for their invaluable feedback. Many of my former colleagues contributed to the ideas

presented in this book. I thank them all sincerely for being such a critical part of the trading business that we have successfully run together for over two decades. And I especially thank my family for their constant encouragement and support of my project in very unique and special ways.

New York, USA  
May 2023

Ilia Bouchouev

---

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	The Ecosystem of Real Options . . . . .	1
1.2	The Structure of the Book . . . . .	4
	Reference . . . . .	7

## Part I Economic Foundations, Markets, and Participants

<b>2</b>	<b>Oil, Money, and Yields</b>	<b>11</b>
2.1	Commodities and Money . . . . .	11
2.2	The Oil Own Rate of Interest . . . . .	15
2.3	Carry and Convenience Yield . . . . .	19
2.4	The Roll Yield . . . . .	24
	References . . . . .	27
<b>3</b>	<b>Fundamentals, Storage, and the Model of the Squeeze</b>	<b>29</b>
3.1	The Invisible Hand of Storage Boundaries . . . . .	29
3.2	The Canonical Theory of Storage . . . . .	32
3.3	A Stylized Model of the Squeeze . . . . .	38
3.4	Cushing: Pipeline Crossroads of the World . . . . .	46
3.5	Negative Oil Prices . . . . .	49
	References . . . . .	53
<b>4</b>	<b>Financialization and the Theory of Hedging Pressure</b>	<b>55</b>
4.1	The Theory of Normal Backwardation . . . . .	55
4.2	The Hedging Equilibrium . . . . .	60
4.3	The Genesis of Oil Financialization . . . . .	64
4.4	Inflation Hedging and Risk Parity . . . . .	69
4.5	Inconvenience Yield, or the Theory of Normal Contango . . . . .	75
	References . . . . .	78

## Part II Quantitative Futures Strategies

<b>5</b>	<b>Systematic Risk Premia Strategies</b>	<b>83</b>
5.1	The Evolution of Algos . . . . .	83
5.2	Myths and Realities of Oil Momentum . . . . .	86
5.3	Carry as a Transmitter of Fundamentals to Prices . . . . .	97

5.4	Value and Mean-Reversion . . . . .	101
5.5	The Reaction Function . . . . .	105
	References . . . . .	107
<b>6</b>	<b>Quantamentals . . . . .</b>	<b>109</b>
6.1	Trading Curve and Convexity . . . . .	109
6.2	Time Spreads and Inventories . . . . .	114
6.3	WTI-Brent Accordion . . . . .	117
6.4	Cointegration and Energy Stat-Arb . . . . .	122
6.5	Disentangling Flows and Positioning . . . . .	128
	References . . . . .	134
<b>7</b>	<b>Macro Trading . . . . .</b>	<b>137</b>
7.1	Dynamic Systems and Feedback Loops . . . . .	137
7.2	Oil, Dollar, and Commodity Terms of Trade Strategies . . . . .	139
7.3	Oil and Energy Equities . . . . .	146
7.4	Oil and Inflation . . . . .	148
7.5	Macro Fair-Value Model . . . . .	155
<b>Part III Volatility Trading</b>		
<b>8</b>	<b>Options and Volatilities . . . . .</b>	<b>161</b>
8.1	Options and “Théorie de la Spéculation” . . . . .	161
8.2	Local Volatility and Diffusions . . . . .	164
8.3	Delta Hedging and Option Replication . . . . .	169
8.4	Realized Volatility . . . . .	175
8.5	Implied Volatility and its Skew . . . . .	180
	References . . . . .	186
<b>9</b>	<b>The Hidden Power of Negative Gamma . . . . .</b>	<b>187</b>
9.1	Options and Insurance . . . . .	187
9.2	The Most Powerful Option Greek . . . . .	191
9.3	The Smile of the Volatility Risk Premium . . . . .	196
9.4	The Art and Science of Delta Hedging . . . . .	202
9.5	The Behavior of Hedgers and Regime Changes . . . . .	205
	References . . . . .	208
<b>10</b>	<b>Volatility Smile Trading . . . . .</b>	<b>209</b>
10.1	Producer Hedging and Volatility Market-Making . . . . .	209
10.2	Skew Delta and Two Types of Stickiness . . . . .	215
10.3	When Black Smirks, Bachelier Smiles . . . . .	221
10.4	Fat Tails and the Quadratic Normal Model . . . . .	226
	References . . . . .	230
<b>Part IV Over-the-Counter Options</b>		
<b>11</b>	<b>Volatility Term Structure and Exotic Options . . . . .</b>	<b>233</b>
11.1	Dark Pools and the Hacienda Hedge . . . . .	233

11.2	The Term Structure of Implied and Local Volatilities . . . . .	237
11.3	Volatility Discount from Price Averaging . . . . .	241
11.4	Early Expiry Options and Swaptions . . . . .	247
11.5	Multi-Factor Models and Other Exotics . . . . .	252
	References . . . . .	255
<b>12</b>	<b>Volatility Arbitrage and Model Calibration . . . . .</b>	<b>257</b>
12.1	The Inverse Problem of Option Pricing . . . . .	257
12.2	Bootstrapping in Time . . . . .	261
12.3	Market-Implied Probability Distribution . . . . .	268
12.4	Local Volatility Smile . . . . .	273
	References . . . . .	279
<b>13</b>	<b>Spread Options and Virtual Storage . . . . .</b>	<b>281</b>
13.1	The Synthetic Storage Strategy . . . . .	281
13.2	Triangular Correlation Arbitrage . . . . .	287
13.3	The Dichotomy of Spread Option Pricing . . . . .	292
13.4	Dealing with Unobservables . . . . .	298
	References . . . . .	303
<b>14</b>	<b>Epilogue . . . . .</b>	<b>305</b>
14.1	The Roadmap for Energy Transition and Virtual Commodities . . . . .	305
	<b>Appendix A: Diffusions and Probabilities . . . . .</b>	<b>309</b>
	<b>Appendix B: Option Pricing under Normal and Lognormal Distributions . . . . .</b>	<b>315</b>
	<b>Appendix C: The Perturbation Method and the Quadratic Normal Model . . . . .</b>	<b>319</b>
	<b>Appendix D: Option Pricing with Time-Dependent Volatility . . . . .</b>	<b>323</b>
	<b>Appendix E: Average Price Options . . . . .</b>	<b>325</b>
	<b>Appendix F: The Inverse Diffusion Problem . . . . .</b>	<b>327</b>
	<b>Glossary . . . . .</b>	<b>331</b>
	<b>References . . . . .</b>	<b>335</b>
	<b>Index . . . . .</b>	<b>341</b>



# Introduction

1

## 1.1 The Ecosystem of Real Options

Global oil trading is an example of a complex dynamic system, where everything depends on everything else, and system components communicate with each other via multiple feedback loops. Other examples of such systems include the human body, the world economy, climate change, or relationships within the family. Such systems are capable of producing their own behavioral patterns which are often governed by quantifiable rules. We will describe some of these rules and refer to the resulting behavior of the system participants as *mental models*.

In the core of the oil trading system lies the concept of optionality. The business of producing oil is a call option on the price of oil. If the price covers all costs of exploration and production, then the business moves ahead. If it does not, then oil remains in the ground. The price of oil is volatile, which makes oil production a valuable option to own, but an expensive one to acquire, as large capital investment is needed at the outset of the project. Investing in such a long-term option is a risky business. By the time oil is found and the necessary infrastructure is built, the price of oil could fall and make the entire project unprofitable. It is possible, of course, to roll the dice and hope for the best in the style of early wildcatters. Alternatively, one can do something about it and secure a profitable forward price for the time when production is expected to commence. This process of managing the forward price risk is called *hedging*. The venue for such transactions is the derivatives market, made up of futures, swaps, options, and highly customized structured products.

When the term derivative is applied to oil futures, it becomes a misnomer. A conventional textbook definition of a commodity derivative suggests that the futures contract derives its value from the spot price of a physical commodity. The oil market, however, is somewhat unconventional. In contrast to many other raw commodities, oil does not have a universal spot price with an instantaneous delivery. There is no one large oil bazaar where all buyers and sellers show up with barrels and gallon jars, negotiate the price with an instant delivery, and take oil home. Oil is too bulky and a special infrastructure is required for its transportation and storage. A

buyer and a seller can only agree on the price of oil to be delivered at a given time to a specified location, typically to a storage tank connected to pipelines. In other words, oil in commercial quantities **can only trade forward**. Even though we use the term spot price throughout the book for brevity, it is meant to **represent the forward price with short delivery time**.

A typical agreement for the delivery of physical barrels is highly customized for the specific grade of oil, its delivery time, and location. To compare simultaneous prices for many such agreements, one needs to peg them to a common anchor. This anchor is usually the futures contract with the nearest expiration, which is also referred to as the **prompt futures**. The prices for physical transactions are then negotiated as a basis to futures. Since the magnitude of the physical basis typically represents only a relatively small adjustment to the futures price, **the primary price determination occurs in the futures market**. Unlike many other commodities, the oil futures contract can hardly be called a derivative of the spot price. In fact, it would be much more appropriate to say that the spot oil price is a financial derivative of oil futures.

The need to transport oil to the desired location highlights an important role that **oil shippers** play in the market. The most effective way of **transporting oil on the ground is via pipelines**. Oil can also be moved by rail cars and even by trucks but at a significantly higher cost. **To transport oil between continents, one must resort to special oil tankers**. The pipeline, the tanker, and other oil-transporting assets are also examples of real options. **They are options on the difference between oil prices in two geographical locations, also known as spread options**. The asset owner will only commit to shipping oil if the price differential between two locations covers the cost of transportation. Spread options play an important role in the functioning of the entire petroleum complex. **They arise at virtually every step within the petroleum value chain**.

Once oil is shipped to the desired destination, it must be placed in storage, since not all of it can be consumed instantaneously. **The storage asset shifts the allocation of the valuable resource in time, from times of plenty to times of relative scarcity**. If the futures price for later delivery is higher, then the rational storage operator will postpone the sale of oil and instead store it, provided that the forward price premium covers the cost of storage. If the forward price premium over the spot price is not sufficient to pay for storage costs, then the operator is better off selling oil in the market than keeping it in storage. **Storage is also a real option**. While the pipeline is **an option on location, the storage asset is an option on time**.

Storage plays a prominent role in global oil trading, and it will feature accordingly throughout the book. In fact, the largest single participant in the futures market is not the producer, the consumer, or even the speculator, **but the hedger of oil inventories**. We will see that the behavior of the storage trader in the futures market translates the fundamental information about supply and demand in the physical market to futures prices, which in turn determines the price of the physical barrel. The presence of storage makes physical and derivatives oil markets operate as a mutually stimulating loop.

The final critical piece in the oil value chain is a refinery. Crude oil has very limited use in its raw form. It must first be processed into a usable product. A refinery that processes crude oil is also an option on the spread. Here, the spread is between the basket of refined products and the basket of crude oil inputs. The power of spread options in the oil business is best exemplified by John D. Rockefeller, who built his empire by monetizing real options to refine, transport, and store oil, while leaving the more glamorous but much riskier option on the price of oil to wildcatters and speculators.

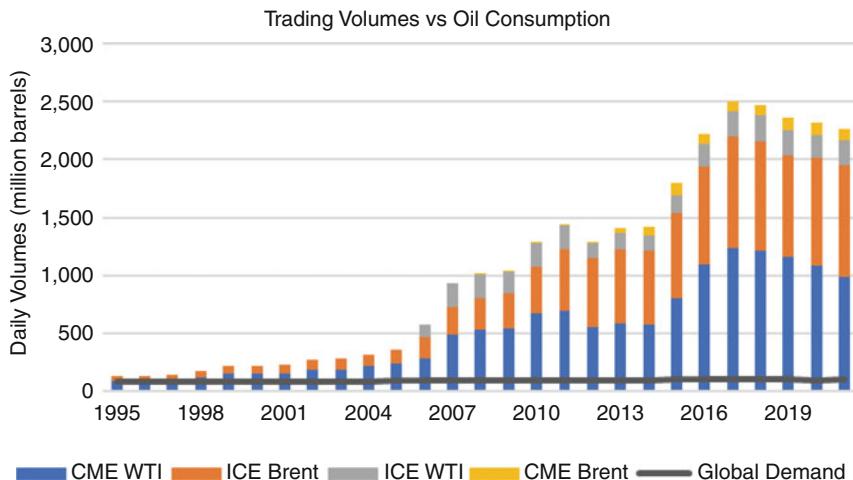
The real optionality embedded in physical assets is the core of the energy market. It transforms trading into a complex dynamic system driven by multiple feedback loops. The system rests on its four primary pillars of production, transportation, storage, and refining. Each of these pillars represents a valuable but expensive option. The high cost and associated risks imply that one cannot pre-build too many of such real options in advance and keep them hanging around just in case. By and large, the energy infrastructure has just enough production, refining, pipelines, tankers, and storage tanks to keep it up and running, but it has very little cushion and no margin for error. The oil ecosystem operates near the edge in a highly optimized fashion with a just-in-time and just-enough mentality. Such systems are characterized by an inherent instability with periodic bursts of extreme volatility.

The economic function of financial derivatives is to provide asset owners with tools to manage such volatility. Given the enormous cost of the energy infrastructure, the demand for risk-transferring services turned out to be so large that the market for oil derivatives stood out not only among all commodities, but it also became one of the deepest derivatives markets across all asset classes. In addition, its liquidity, combined with the high volatility and relatively low correlation to other financial assets, turned oil into a favorite playground of speculators. With so many diverse factors impacting the price of oil, one can always find something in the oil market to bet on, which can only stimulate the marketing machine to make the market even larger.

At present, the world consumes approximately one hundred million barrels of oil per day. While this may sound like a large number, it pales in comparison to over two billion financial barrels that trade every day via two benchmark oil futures contracts, West Texas Intermediate (WTI) and Brent. Such an unprecedented growth of financially-traded barrels is illustrated in Fig. 1.1.<sup>1</sup> If one adds futures on other grades of crude oil, futures on refined products, options, and over-the-counter (OTC) derivatives, then the daily trading volume of petroleum derivatives is at least fifty times larger than daily global oil consumption. While such a comparison should only be interpreted with great caution, it nevertheless shows the growing power of financial markets.

---

<sup>1</sup>From the author's testimony to the U.S. House Subcommittee on Economic Growth, Energy Policy, and Regulatory Affairs of the Committee on Oversight and Accountability. See Bouchouev (2023).



**Fig. 1.1** Daily trading volume of two benchmark oil futures contracts, WTI and Brent, is at least twenty times larger than daily global oil consumption. Source: CME, ICE, EIA

The derivatives market has brought new entrants to the oil marketplace. Besides producers, consumers, storage hedgers, and other traditional market participants, the price of oil became highly sensitive to the behavior and motivation of relative newcomers, including inflation hedgers, systematic quantitative funds, and especially option traders. The goal of this book is to demystify their behaviors that collectively drive the elusive price of oil.

## 1.2 The Structure of the Book

The book consists of four main parts with each split into three chapters. While all parts of the book are interconnected, each one can be read as stand-alone based on individual interests and the level of the reader's mathematical background. The first part introduces market participants and lays down the economic foundations of storage and hedging pressure, upon which many trading mental models will subsequently be built. These mental models and futures trading strategies are presented in Part 2, which is the least technical and can be easily read by anyone with a broad interest in financial markets. Part 3 targets option traders familiar with volatility trading and some basic stochastic modeling. Part 4 is more suitable for professional volatility arbitrageurs and anyone with a stronger interest in exotic oil options and mathematical finance. To keep the main text accessible to a wider audience, technical details are presented in appendices. Throughout the book and unless noted otherwise, we use energy data from the U.S. Energy Information Administration (EIA) and the Chicago Mercantile Exchange (CME), and macroeconomic data from the Federal Reserve Economic Data (FRED).

We begin Chap. 2 by looking at a hypothetical economy where **oil and the US dollar (USD)** function as two alternative standards of money. We relate the interest rate in the oil-denominated economy to the **interest rate on fiat money**, following the first rigorously defined **no-arbitrage carry trade**, formulated by **Irving Fisher**. This formula is now better known as the **Fisher inflation law**, which establishes the close linkage between **oil and inflation**. We revive a largely forgotten **Keynesian concept of commodity own rate of interest** and illustrate it with an example of an oil loan from the **US Strategic Petroleum Reserve (SPR)**. We then relate the oil own rate of interest to the **convenience yield** in the physical market and to the **roll yield** that arises in **trading futures**.

In the following two chapters we look at two alternative approaches to modeling the behavior of **futures prices**. These approaches, known as the **theory of storage** and the **theory of hedging pressure**, loosely correspond to competing trading styles based, respectively, on **fundamentals and flows**.

In Chap. 3, we use the conventional storage theory to illustrate the dynamic feedback loop between prices and inventories and highlight the challenge of its practical applications to the oil market. We then borrow some concepts from the physics of extreme events and develop a more practical alternative approach to the storage problem. We call it a *stylized model of the squeeze*. For an example of such a squeeze, we delve into the infamous episode of negative oil prices.

In Chap. 4, the focus shifts to flow imbalances. Particular attention is paid to causes and consequences of the influx of financial investors to the oil market, the phenomenon dubbed *financialization*. We track the market transition from the early days of *normal backwardation* to the subsequent regime of *normal contango* and combine them into a more general economic framework that describes the *hedging pressure equilibrium*.

In the second part of the book, we present mental models for futures trading strategies deployed by professional speculators, including quantitative hedge funds, sophisticated physical speculators, and global macro traders. We devote one chapter to each of these groups of traders and their marquee strategies. While appearing to be very different in nature, many concepts supporting these strategies can be seen as practical implications of economic theories presented in the first part of the book. Our goal is to focus on modeling frameworks and conditions under which various strategies tend to work, and not on nuances of their technical implementation.

Chapter 5 is devoted to systematic quantitative funds, known as *commodity trading advisors (CTAs)* or simply as *algos*. These traders tend to look at markets through the lenses of risk premia, such as *momentum*, *carry*, and *value*. We discuss how the source of these risk premia in the oil market is ultimately linked to the theory of storage. Important concepts of *signal blending* and the *reaction function* are introduced.

In Chap. 6, we explore the considerably more challenging task of using fundamental data for generating trading signals, and the class of semi-systematic strategies, colloquially referred to as *quantamentals*. We discuss the convergence strategy of *WTI-Brent accordion* and extend the concept to construct a broader *statistical arbitrage* portfolio of energy pairs. We use *fractionation analysis* to

incorporate fundamental data and provide some guidance and ideas for modeling flows and positioning. Such quantamental strategies are more effectively managed with a **discretionary overlay**. In the oil market, the trader equipped with the machine is more powerful than either an oilman or a machine alone.

Chapter 7 looks at oil as a sub-component of a broader **macro trading system**. We illustrate **oil linkages to currency, equity, and fixed income markets** with three **cross-asset relative value strategies**. We conclude by constructing a simple and somewhat **naïve fair-value model for the price of oil**. While this model is built on questionable theoretical grounds, it is a good example of when something that should not work in theory turns out to be helpful in practice.

The third part of the book deals with **vanilla oil options and classic volatility trading**. The market for oil options is one of the most developed options markets in the world. This should not come as a surprise given that the entire oil market is **driven by real options held by owners of physical assets**. However, from the quantitative perspective the dynamics of oil volatility, which determines the price of an option, remains largely undocumented. This book attempts to fill the gap.

When it comes to trading, there is no free money. While trading futures is easier to understand, popular linear strategies quickly become crowded, as barriers to entry are relatively low. In contrast, the barriers to entering the options market are set higher, and lucrative trading opportunities there last longer. The highest barrier is the quantitative skill of the oil trader, so more technical knowledge is essential for success in volatility trading. The goal of this book is to help aspiring option traders to develop a minimum competency level that can then be applied and further improved in a professional trading organization.

In Chap. 8, we summarize the main building blocks that make up the business of volatility trading. We start by giving well-deserved credit to Louis Bachelier, the father of modern quantitative finance, whose *Bachelier pricing formula* is still being used by oil option traders. The classical *Black-Scholes-Merton (BSM) framework* of option replication by *delta hedging* is then derived in a more general setting of *diffusion processes*. We highlight the importance of distinguishing between three commonly used types of volatility: *local volatility*, *realized volatility*, and *implied volatility*.

Chapter 9 looks at writing oil options from the perspective of the insurance product, which is the essence of *gamma trading* and the resulting *volatility risk premium (VRP)*. We dissect the historical performance of various delta-hedging strategies from multiple angles, introduce the concept of the *VRP smile*, and identify regime changes caused by changing behavior of large market participants.

In Chap. 10 we study the problem of the *volatility smile* and the popular market-making strategy of *vega trading*. We demonstrate the shortcoming of tracking oil volatility smile using popular heuristics of *sticky strike* and *sticky moneyness* and replace these naïve methods with a more flexible framework based on general diffusion processes. We apply the technique of *perturbation methods* to develop a novel *quadratic normal (QN) model*. This model corrects the Bachelier formula for skewness and kurtosis. The three parameters of the model are linked to the market prices of three primary option benchmarks: at-the-money straddle, costless collar,

and **out-of-the-money strangle**. The QN model corrects the conventional pricing formula just enough to capture important features of the oil market without introducing excessive complexity.

In the final part of the main text, we describe core components of managing large and often secretive over-the-counter (OTC) deals which remain hidden from public view. We hope to unveil some of the mystery behind these deals for the general public, while helping professional option dealers with somewhat tricky challenges that arise in pricing and calibration of models used for OTC energy derivatives. This part is somewhat more advanced quantitatively, but once again, most technical details are placed in appendices.

In Chap. 11, we start with the example of arguably the most significant oil derivative trade, **the large-scale annual put buying program by the Government of Mexico**. The complexity of OTC deals highlights the importance of handling the **volatility term structure** and the effect of volatility dampening by price averaging. A simplified and more practical model for pricing and hedging **average price options (APO)**, which are very popular among end-users, is presented. *Swaptions* and several other exotic options are also discussed.

Chapter 12 covers the more challenging problem of model calibration. We present the *bootstrapping method* for calibrating volatility time-dependency, including the volatility matrix for non-homogeneous term structure of volatility. To calibrate the model across strikes, we first back out *market-implied probability distribution* from option prices. A more difficult problem is to reconstruct the *market-implied diffusion* process, known as the *inverse problem of option pricing*. Some readers may find it interesting that this important problem is only partially solved, and in its general case, it presents a rare example of an open mathematical problem. This problem arises naturally in calibration of oil options.

Chapter 13 presents a relatively simple but quantitatively elegant strategy based on the idea of *virtual storage*. In this strategy, the physical storage asset is replicated with a financial derivative, a *calendar spread option (CSO)*, and the structural pricing dislocation between the physical and financial market is then exploited. Other spread option and correlation strategies are based on the concept of a *triangular arbitrage*, which links prices of vanilla and spread options. We highlight some challenges with an application of correlation-based models and insist on connecting spread options models to fundamentals of supply and demand.

The book concludes with a short epilogue related to the energy transition. Its focus, however, is on the transition of virtual energy or the market for energy derivatives and its impact on the broader market for virtual commodities.

---

## Reference

- Bouchouev, I. (2023). *Testimony to the House Subcommittee on Economic Growth, Energy Policy, and Regulatory Affairs of the Committee on Oversight and Accountability*, March 8. Reprinted in *Commodity Insights Digest* (2023), Vol. 1.

---

## **Part I**

### **Economic Foundations, Markets, and Participants**



# Oil, Money, and Yields

2

- If oil is so important to the economy, why cannot it be used as an alternative form of money? Some oil transactions are settled directly in barrels, and the US Government even lends oil with the interest also paid in barrels.
- The oil own rate of interest represents the **return on lending oil barrels**. It can be calculated from the **term structure of traded futures**. The own rate is analogous to **the real rate in the famous Fisher inflation law**.
- In fundamental trading, the oil own rate is better known as **convenience yield, net of storage costs**. It represents the trade-off between the intrinsic value of the commodity for its immediate consumption and the cost of storing it.
- Financial traders buy futures to synthesize the ownership of the physical barrel but the investment in futures may lead to very different results. The difference comes from the roll yield, which is a financial equivalent of the oil own rate of interest.

---

## 2.1 Commodities and Money

The **money-rate of interest** is nothing more than the **percentage excess of a sum of money contracted for forward delivery**. For every kind of capital-asset there must be an analogue of the rate of interest on money. Thus for every durable commodity we have a rate of interest in terms of itself, – a wheat-rate of interest, a copper-rate of interest, a house-rate of interest, even a steel-plant-rate of interest.

J. M. Keynes (1936)

Today, we live in the world of **flat money**. For most of our ancestors, however, money was represented by **commodities**. Gold, silver, copper, salt, rice, maize, barley, cocoa beans, tobacco, and many other raw materials have a long history of being used as money. From the very beginnings of human civilization, people were choosing something valuable to them to function as money. Unlike **flat money**, which derives its **value solely from faith in the issuing government**, commodities

possess value within themselves. The intrinsic value of a commodity is its worth for consumption and for various productive processes. Since oil has been the most valuable commodity for nearly a century, why not use it as an alternative standard of money?

We can easily agree to settle all transactions directly in oil, borrow and lend oil barrels, and negotiate interest on such loans, also to be paid in oil. An oil rate of interest will depend on the value of oil today relative to its value at the time when a debt must be settled. If consumers are eager to receive oil barrels quickly, they may agree, for example, to exchange 100 barrels today for 110 barrels to be returned later. This would generate ten percent investment return to the lender of oil barrels. The return comes from the intrinsic value of the commodity, its utility, and the convenience of having it handy if it becomes needed.

However, if immediate demand for oil is tepid, then the owner of oil must find a place to keep it somewhere, while waiting for buyers to come in. The process may involve renting a dedicated storage facility, connecting it to a pipeline, buying insurance, and completing other burdensome and costly logistical operations. Since all payments are agreed to be made in barrels, the lender may need to give up some barrels to remunerate a provider of storage services. For example, the oil lender may agree with a storage manager to exchange 100 barrels today for 95 barrels to be taken back later. In this case, the investment return on lending oil barrels would be negative five percent. Any decision making in such an oil-based economy represents the trade-off between the intrinsic value of oil for its immediate consumption and the cost of storing it.

Now consider a dual-currency economy, where besides oil, there exists an alternative money standard, such as gold, the US dollar (USD), or bitcoin. Let us assume that one can freely substitute between two types of money at the prevailing exchange rate determined by the market. We also assume that the market is efficient to preclude any possibility of arbitrage, which means that it is deemed to be impossible to extract riskless profits from borrowing in one money standard and lending in another. In the absence of such arbitrage, interest rates on two alternative types of money must be related. This relationship was developed by Irving Fisher at the end of the nineteenth century.<sup>1</sup> Today, it is better known as the *Fisher inflation law*, and the transaction that leads to it as the *carry trade*.

Fisher's description of an equivalence between loans in two commodity money standards is arguably the earliest rigorously described example of the quantitative arbitrage trade in financial markets. His work was motivated by the collapse of the bimetallic monetary standard when gold and silver coins could both be used as acceptable forms of money. To illustrate that the rates of interest on two alternative commodity money standards must satisfy an equilibrium no-arbitrage equation,

---

<sup>1</sup>See Fisher (1896). For consistency with contemporary conventions adopted in the economic literature, our definitions of certain terms differ from Fisher's original formulation. For example, we use commodity price appreciation in terms of money to characterize commodity inflation, in contrast to Fisher's choice of looking at inflation as money depreciation in terms of commodities.

Fisher used bushels of corn and ounces of gold. His argument can be easily extended to a hypothetical bi-monetary world where all transactions are settled either in barrels of oil or in USD.

If one borrows  $D(t)$  dollars at time  $t$  and returns  $D(T)$  dollars to the lender at time  $T$ , then the USD rate of interest,  $r$ , is conventionally defined as

$$r = \frac{D(T) - D(t)}{D(t)} = \frac{D(T)}{D(t)} - 1 \quad (2.1)$$

Likewise, if one chooses instead to take a loan in oil, borrow  $B(t)$  barrels at time  $t$ , and return  $B(T)$  barrels at time  $T$ , then the oil rate of interest,  $b$ , is similarly defined as

$$b = \frac{B(T) - B(t)}{B(t)} = \frac{B(T)}{B(t)} - 1 \quad (2.2)$$

Let us assume that we also know the price of the second money standard in terms of the first one. In other words, we know spot oil prices  $S(t)$  and  $S(T)$  at times  $t$  and  $T$ , both expressed in USD per barrel (\$/bbl)<sup>2</sup>:

$$S(t) = \frac{D(t)}{B(t)}, S(T) = \frac{D(T)}{B(T)}$$

From here, we express barrels as the ratio of dollars to oil prices, and substitute into (2.2), which allows us to relate the oil rate of interest to the dollar rate of interest, using the definition (2.1), as follows:

$$b = \frac{B(T)}{B(t)} - 1 = \frac{D(T)S(t)}{S(T)D(t)} - 1 = \frac{S(t)}{S(T)}(1 + r) - 1$$

We then rearrange the terms and obtain that

$$1 + r = (1 + b) \frac{S(T)}{S(t)} = (1 + b)(1 + i) \quad (2.3)$$

where

$$i = \frac{S(T)}{S(t)} - 1$$

represents the rate of change in prices measured by the alternative money standard, which is oil, in terms of the original money standard, which is USD. This formulation is Fisher's original one-period version of the inflation law. He defined  $b$  as the

---

<sup>2</sup>The term spot price is used only for pedagogical clarity. We will explain shortly that oil spot prices with instantaneous delivery do not really exist, and the spot price should be understood as the forward contract with the nearest delivery.

*virtual rate of interest in commodities.* As we will see throughout the book, commodity prices and inflation tend to go side by side.

With the development of commodity futures markets, the concept of *commodity rate of interest was revived by an Italian economist*, Piero Sraffa. Since the price of the futures contract  $F(t, T)$  for the delivery at time  $T$  is observable from the market, one can use it to *replace an unknown future spot price by letting*

$$S(T) = F(t, T)$$

In other words, one no longer needs to *guess what commodity rate of interest is*. It can be calculated directly from the *futures market*. *Sraffa used such market-implied rates for different commodities to challenge the proposal of Friedrich Hayek to fix the money rate of interest at some unique natural commodity rate that can be found in a non-monetary economy*. Since the price of each commodity is driven by its own supply and demand that lead to a *different relationship between spot and futures prices*, Sraffa argued that “*there might be at any one moment as many ‘natural’ rates of interests as there are commodities . . . and there would be no single equilibrium rate*”<sup>3</sup>.

The *idea of commodities having their own intrinsic rates was further developed by John Maynard Keynes*, who labeled the concept the *commodity own rate of interest*. The own rate became the centerpiece in one of Keynes’ most influential economic theories, on how *competition between commodity own rates and the interest rate on money could lead to economic slumps*. His application of this term, however, became highly controversial, causing a considerable amount of confusion among economists.<sup>4</sup> Perhaps for these reasons, the concept of a commodity own rate of interest has largely disappeared from contemporary economic literature.

---

<sup>3</sup>Sraffa (1932) only defined the concept by virtue of the following example of borrowing bales of cotton: “*The rate of interest . . . per hundred bales of cotton, is the number of bales that can be purchased with the following sum of money: the interest on the money required to buy spot 100 bales, plus the excess (or minus the deficiency) of the spot over the forward prices of the 100 bales*”.

<sup>4</sup>See Keynes (1936), chapter 17. Keynes restated Sraffa’s definition algebraically, which triggered a heated debate between the two economists in a series of letters. The debate was caused by their disagreement on whether to use spot prices or forward prices in defining the dollar cost of borrowing the commodity. While it may appear to be a small technical nuance, it led to a more profound philosophical differences in the interpretation of the concept. For details, see, e.g., Naldi (2015). In this book, we accept the definition as it was stated by Keynes, which simply mimics the definition of the interest rate on fiat money.

## 2.2 The Oil Own Rate of Interest

After being forgotten for decades, the concepts of commodity loans and own rates of interest found important practical applications in the **modern oil market**. Settling transactions directly in oil barrels turned out to be very convenient for many market participants. Using oil in lieu of the currency **allows some resource rich but cash poor sovereign producers to attract financial investors to develop oil properties and pay them directly in oil**. It also lets refineries secure **profit margins via netback agreements by exchanging their product output for the crude oil input**. Furthermore, storage owners, including governments, **can lend barrels to the market in exchange for more barrels to be returned later**. In other words, oil is already used as a money substitute by the industry. Before we get to some specific case studies, let us set the stage with some simple examples.

We first consider the scenario, when the demand for oil today at time  $t$  is relatively high, and the spot price  $S(t) = \$60$  exceeds today's futures price for delivery at time  $T > t$ , which is  $F(t, T) = \$50$ . In the jargon of commodity markets, such a decreasing forward curve corresponds to the market being in the state of **backwardation**. Let us assume that the dollar rate of interest is  $r = 2.5\%$ . Then one can use \$60 today to either buy one barrel of oil, or, alternatively, buy  $\$61.50 = \$60 * (1 + 0.025)$  dollars of forward delivery. These \$61.50 of forward dollars can then be converted into  $\frac{\$61.50}{\$50} = 1.23$  oil barrels at the forward oil price. Looking at this transaction from the perspective of an oil lender means that one barrel of oil today is exchanged for 1.23 barrels to be returned later. Therefore, the oil own rate of interest in this example is 23%.

Now consider the second scenario, where the forward price  $F(t, T) = \$70$  exceeds the spot price  $S(t) = \$60$ . The forward curve is now increasing, which defines the market being in the state of **contango**. A lower current price in comparison to the future price indicates that current oil supplies are abundant, and the demand for oil is relatively low. In this case, the same \$61.50 of forward dollars can only buy  $\frac{\$61.50}{\$70} = 0.88$  forward oil barrels. Therefore, the oil own rate of interest is **negative 12%**. While lending barrels at a negative return may sound counterintuitive, it may still be economical if the cost of storing oil is significant.

While the concept of own rate of interest was originally introduced in discrete-time setting using simple returns, the contemporary analysis of financial markets usually operates in continuous-time setting with continuously compounded returns.<sup>5</sup> For consistency, simple returns are usually replaced with logarithmic or log-returns, which makes algebraic calculations simpler and more intuitive. Unlike simple returns, which are not additive, the log-returns can be added both in time and across multiple assets. Since in our book the impact of interest rate plays only a minor role in oil futures strategies, for the most part, we use log-returns in all analyses.

---

<sup>5</sup>The one-year continuously compounded interest rate is conventionally defined as the following limit

$$e^r = \lim_{n \rightarrow \infty} \left(1 + \frac{r}{n}\right)^n, \text{ where } n \text{ is a number of times interest is compounded in a year.}$$

If we replace simple returns with log-returns, defined as

$$r = \ln\left(\frac{D(T)}{D(t)}\right), b = \ln\left(\frac{B(T)}{B(t)}\right), i = \ln\left(\frac{S(T)}{S(t)}\right)$$

then repeating the previous steps and replacing unobservable future spot price  $S(T)$  with the market futures price  $F(t, T)$ , we obtain that

$$b = \ln\left(\frac{D(T)S(t)}{S(T)D(t)}\right) = r - \ln\left(\frac{F(t, T)}{S(t)}\right) = r - i \quad (2.4)$$

This is the continuous version of the Fisher's inflation law. Here,  $b$  is better known as the real rate, which is equal to the difference between the nominal rate of interest  $r$  and the rate of inflation  $i$ .

The Eq. (2.4) is written for a single time period. If  $r$  and  $b$  are viewed as annualized log-returns, then they must be multiplied by time to maturity of the futures contract, so that

$$\ln\left(\frac{F(t, T)}{S(t)}\right) = (r - b)(T - t)$$

Therefore, the futures price can be expressed as the spot price that grows exponentially at the rate equal to the difference between the nominal rate of interest and the oil own rate of interest

$$F(t, T) = S(t)e^{(r - b)(T - t)} \quad (2.5)$$

Alternatively, one can back out the oil own rate of interest from the futures curve

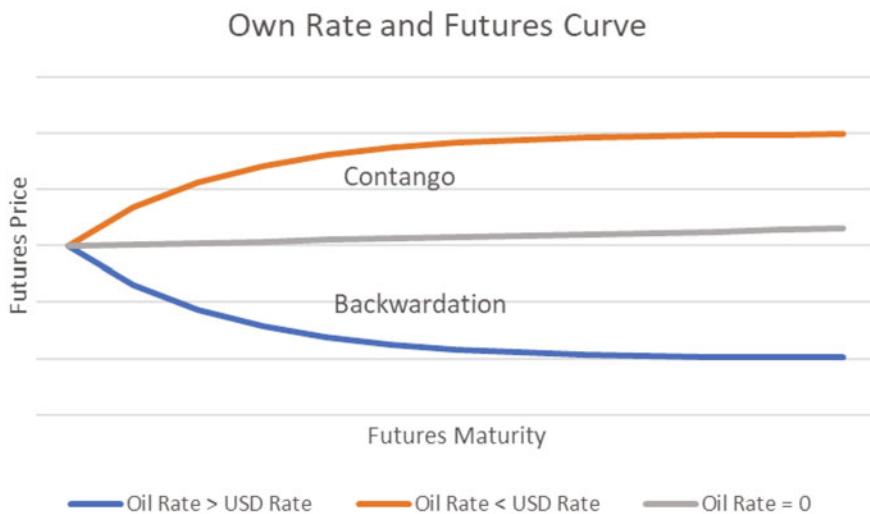
$$b = r + \frac{1}{T - t} \ln\left(\frac{S(t)}{F(t, T)}\right) \quad (2.6)$$

where the spot price is understood as the futures with the shortest maturity.

Figure 2.1 illustrates how the own rate relates to backwardation and contango of the futures curve with different maturities.

If  $b > r$  and the intrinsic worth of the commodity is relatively high, the futures curve is in backwardation. In contrast, if  $b < r$  and the commodity own rate is relatively low, the futures curve is in contango. If  $b = 0$ , then the futures curve is in a slight contango, monotonically increasing at the rate  $r$ .

Having defined the basic terminology of the oil economy, we can now illustrate it with an important case study. We consider an example of oil loans from the US Strategic Petroleum Reserves (SPR), the world's largest supply of emergency oil stocks, held by the US Government. The SPR was established in 1975 by the Energy Policy and Conservation Act (EPCA) following the Oil Embargo of 1973–1974 by the Organization of Arab Petroleum Exporting Countries (OAPEC). The mission of



**Fig. 2.1** If the oil own rate of interest exceeds (falls below) USD rate of interest then the futures curve is in backwardation (contango)

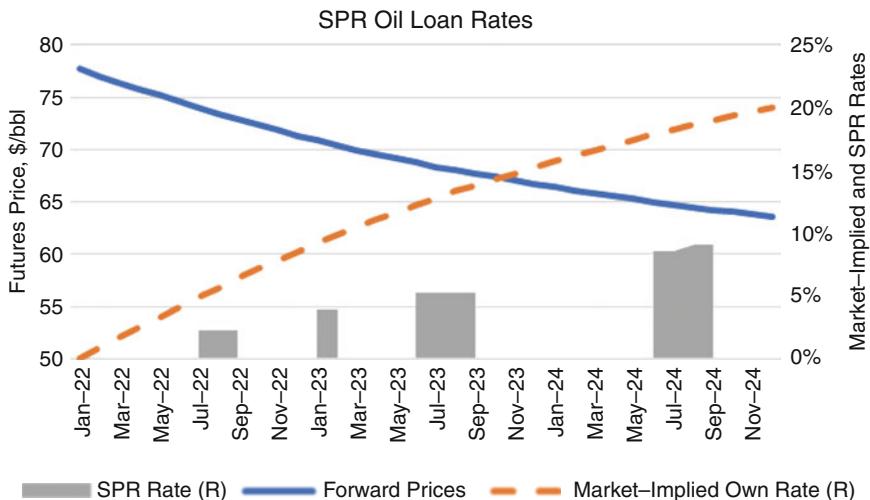
the SPR is to store petroleum to diminish the impact of disruptions on petroleum supplies and to carry out US obligations under the International Energy Program. Oil stocks can be released from the SPR either via competitive sales, or as oil exchanges. Oil exchanges are loans where barrels are released to traders in exchange for a larger quantity of barrels to be returned at some time in the future.<sup>6</sup>

Consider an example of the large SPR oil loan that was authorized in November 2021. At that time, oil demand was quickly recovering from the Covid-19 induced slowdown, while oil supplies from the OPEC+ group of large producers were released to the market at a slower pace.<sup>7</sup> To reduce fundamental supply and demand imbalance, the US Government offered oil from the SPR to traders in exchange for a larger quantity of oil to be returned at a later time. The rates of return on these oil loans were contractually set for five designated loan maturities. The longest maturity loan stipulated 9.1% more barrels in excess of the originally borrowed quantity to be returned to the SPR in approximately thirty months. This allows the US Government not only to replenish the reserve, but also to grow it at no cost.

The question is then why would anyone agree to borrow oil at such a high rate, especially given that the interest rate on fiat money at that time was nearly zero? The answer lies in the steeply backwardated shape of the oil futures curve. To compare loan rates offered by the US Government with the market-implied rates, one can

<sup>6</sup>The history of SPR sales and loans is summarized in Bouchouev (2022).

<sup>7</sup>The Organization of Petroleum Exporting Countries (OPEC) was formed in 1960. The membership of OPEC has varied over the years. As of 2022, OPEC included thirteen countries with the largest producer being Saudi Arabia. A larger but more loosely structured organization, known as OPEC+, which included several other major sovereign producers, was formed in 2016.



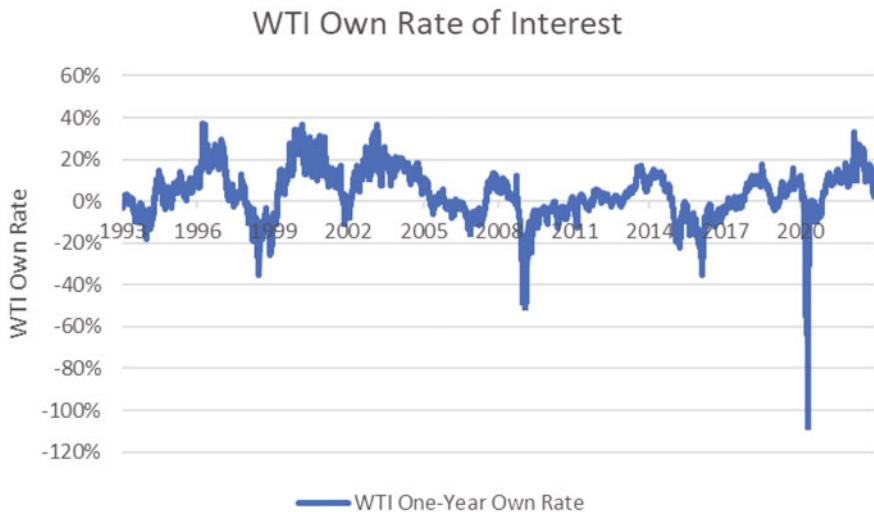
**Fig. 2.2** The rates on SPR oil loans (grey bars) offered by the US government in November 2021 are compared to WTI market implied rates (orange) computed using (2.6)

apply formula (2.6) to calculate oil own rates prevailing at the same time from WTI futures, which is the primary US oil futures contract. It turns out that the thirty-month oil own rate implied by the WTI futures curve at that time was even higher, closer to 20%. Figure 2.2 illustrates.

If oil markets were frictionless and conducted only using the WTI grade of oil with all transactions executed at no cost, then the trader could have theoretically made riskless profits by borrowing barrels from the US Government at 9%, selling them in the spot market, buying thirty-month futures at approximately 20% discount and taking the delivery of physical barrels at that time. Obviously, the real market has plenty of frictions, including high logistical costs associated with physical oil delivery and large price uncertainty in the basis between WTI and specific oil grades that must be returned to the SPR. Nevertheless, for a physical oil trader, participation in such loan transactions could be quite lucrative. This is our first example of the trading strategy where the needs of one market participant, the US Government, creates an arbitrage opportunity for a professional oil trader.

Despite the fact that oil has many attributes of an alternative standard of money, including even having its own rate of interest, we still do not use oil as money. One reason is the complexity of storing oil, but perhaps a more important one is its extreme volatility. To illustrate, Fig. 2.3 shows the history of a one-year own rate of interest implied by WTI futures.

The oil own rate of interest fluctuates wildly, as it is driven by the constantly changing slope of the futures curve. For example, at the onset of the Covid-19 pandemic when oil demand fell rapidly, storage costs increased substantially, culminating in an unprecedented episode of negative prices, which we will discuss in detail in the next chapter. For a negative price, one cannot compute an own rate,



**Fig. 2.3** The history of a one-year oil own rate of interest as implied by WTI futures

but even with this abnormal data point removed, the own rate during that time reached its nadir of negative 100%. However, within two years it swung to positive 30%, as the world recovered from the pandemic and the convenience yield of owning physical barrels skyrocketed. With such a high volatility in a lending rate, it would be extremely challenging to ensure any stability in the oil-denominated economy.

While the oil own rate of interest may sound like an obscure theoretical concept, it underpins many commercial transactions and plays the overall crucial role in the oil market. We will see throughout the book how large its impact is on different trading strategies. For many market participants, however, this concept is better known under different names. In fundamental trading, it is more often associated with the concept of convenience yield.

## 2.3 Carry and Convenience Yield

The own rate of interest reflects the combined effect of the intrinsic value of the commodity and the cost of storing it. It is calculated by comparing the commodity price today to its price in the future. The current price and the future price are connected by the carry trade. The carry arbitrage trade has been well known to commodity merchants from the early days of trading. It simply refers to merchants buying and storing a physical commodity, while hoping to sell it later at a higher price. If the forward price is also locked in, then the carry trade effectively represents arbitrage in time.

Oil started to trade in the 1860s in Western Pennsylvania shortly after the first oil well has been drilled there. The first oil traders were carry arbitrageurs who

developed some primitive forms of oil storage. They were buying excess oil from local producers at discounts and storing it in dumps. These pioneers of oil trading were called “*dump men*”.

The first oil markets were, in my opinion, made by the *dump men* on Oil Creek. Now, the dump men were the first speculators in oil. A dump is a tank of any capacity from 10 barrels to 600. The dump man developed in a refiner later on . . . The dump man visited the small producer in all localities, bought his oil at so much by contract, and if there was one or two feet in the tank bottom he immediately bought in lump sum or so much per barrel. He also bought the good oil, the merchantable oil, from small producers, who could not hold their oil long enough to get a full shipment of 50 or 100 barrels, whose wells had declined to such a point that they could not hold for a month at a time, or were obliged to sell from day to day, or week to week. The dump man was their market.<sup>8</sup>

**The storage carry strategy remains the backbone of modern oil trading.** It gives storage owners an option on time, an option to carry barrels and sell them at a time when they become more valuable. Professional storage traders avoid taking any directional exposure to the price of oil, which is usually hedged by selling futures. The storage business model is entirely driven by the economics of the carry trade. If the spot price is discounted relative to futures, and the discount is sufficiently large to cover the cost of storage, then, by and large, riskless profits can be assured by buying a physical barrel for storage at the spot price while simultaneously selling a higher priced futures contract.

Obviously, in competitive markets, the opportunity to make free money is unlikely to last long. Carry traders will continue to bid up the spot price while selling and pressuring down the futures price until the arbitrage goes away. Therefore, to prevent the existence of a free lunch, the futures price  $F(t, T)$  should not exceed the spot price  $S(t)$  by more than the marginal cost of storage  $U$  and the cost of financing  $R$ :

$$F(t, T) \leq S(t) + U + R \quad (2.7)$$

The existence of this arbitrage boundary has a wide range of consequences. Importantly, it implies that the futures prices cannot be determined solely by market expectations of what the spot oil price will likely be in the future. The arbitrage boundary holds regardless of such expectations. Therefore, the role of expectations in the formation of futures prices is rather limited. This partially explains why futures

---

<sup>8</sup>From the testimony of Patrick C. Boyle, proprietor and publisher of The Oil City Derrick at the Hearing of the US Industrial Commission on Trusts and Industrial Combinations, see Boyle (1899). The testimony was subsequently reprinted in Whiteshot (1905). It describes the origins of oil trading and highlights that dump men were the predecessors of the oil refiners. The dump business is also described in Smiley (1907), who wrote that “the dumps in those days were practically the exchange, and made the market price for oil each day”. Many other interesting facts related to the early days of oil drilling in Pennsylvania, including the construction of first wooden oil tanks, are given by Giddens (1947).

prices are poor predictors of future spot prices, especially when the market is in contango, and prices are dictated by storage economics.

Storage costs represent an important component in the economics of the carry trade. Oil is bulky and relatively expensive. Keeping large quantities of a pricey commodity in storage ties up a lot of capital. One cannot logically leave an oil storage facility empty, and some amount must always be kept there irrespective of the storage economics. In addition, the holder must spend some money on buying insurance and maintaining proper security of the storage facility to protect the high dollar value warehoused there. For these reasons, commercial market participants rarely build oil storage in excess just to have it around in case it ever becomes needed. Instead, they highly optimize the operation of existing facilities. Like many other components of the petroleum ecosystem, the storage business follows a just-in-time and just-enough type of business model. In the end, if consumers run out of oil, it will likely be the governments, and not commercial storage operators, who will be on the hook to prevent an economic catastrophe.

The cheapest way to store oil is in underground salt caverns, which is what the US Government does for the SPR. While salt cavern storage works well as a long-term buffer for national security, it is less suitable for commercial operations, where one needs to take oil in and out of storage frequently in response to changing market prices. While state-owned oil reserves function predominantly as the storage of last resort, the role of managing short-term fluctuations in supply and demand is left to private enterprises. They tend to store oil mostly in specialized aboveground storage tanks that are connected to major production and refining centers via pipelines.

The total storage cost typically includes the cost of transportation. If storage tanks in nearby and easily accessible storage facilities fill up, then oil must be transported to more remote storage locations, preferably by pipelines. Additional shipping expenses increase the total cost of transportation and storage, which raises the level of contango needed for competitive storage owners to stay in business. As more and more storage facilities accessible by pipelines become saturated, more expensive transportation tiers, such as rail cars and trucks, are utilized to ship oil. This further steepens the level of contango that is required to offset higher costs. If all dedicated storage facilities reach their maximum capacity, then the only other practical solution for storing oil in commercial quantities is so-called floating storage, which is a reference to oil being temporarily stored on large idle ships. Floating storage is significantly more expensive, as it must incentivize ship owners to divert tankers away from their primary business of oil transportation.

The cost of storage and the carry trade set the floor on how far the spot price can fall below futures. There is technically no ceiling on how far the spot price can rise relative to futures. In financial markets, such a ceiling is assured by the reverse carry trade. For example, if the price of a common stock today is too high relative to its forward price, then one can always borrow the stock to sell it short and receive it back at the cheaper price by buying futures and taking delivery of the stock at maturity. It is not even necessary to have the stock in possession, as the existence of such riskless profits will force rational investors who own the stock to either conduct such transactions themselves or lend the stock to arbitrageurs. Since the stock is

always owned by someone, the incentive to make riskless profits will make reverse cash-and-carry trading possible in the financial markets, and inequality (2.7) turns into an identity.

In commodities, however, the reverse cash-and-carry strategy cannot be guaranteed. The challenge comes from the consumption side of commodities. If the entire inventory of a commodity has already been used up and resupplies are running behind, then the commodity may not be available for borrowing to enforce the reverse carry arbitrage. Even if it is available, the owner may not be willing to part with the valuable product if its consumption value is perceived to be higher than the expected profit from lending it.

To quantify this additional consumption value of oil, we need to bring in another balancing factor. For now, we simply define it as a fudge factor by introducing an extra term  $Y$  that turns inequality (2.7) into an equation

$$F(t, T) - S(t) = (U - Y) + R \quad (2.8)$$

This balancing term was coined as *convenience yield* by Nicolas Kaldor in 1939, which he interpreted as a negative component of the storage cost.<sup>9</sup> The relationship between futures and spot prices is then determined by the relative magnitude of the total carrying costs and the commodity convenience yield.

While the convenience yield in Eq. (2.8) is understood in dollar terms, it is more common to write this equation in a continuous-time setting, as follows<sup>10</sup>

$$F(t, T) = S(t) e^{(r+u-y)(T-t)} \quad (2.9)$$

Here,  $r$  is the short-term interest rate,  $u$  is the marginal rate of storage costs, and  $y$  is the marginal convenience yield per unit of price. The same no-arbitrage logic applies. At time  $t$ , one can secure the price of oil to be delivered at time  $T$  either by buying the futures contract at the price  $F(t, T)$ , or by borrowing money to buy spot barrels at  $S(t)$  and paying interest rate and storage costs while retaining convenience benefits. The futures price must then be equal to the spot price, accrued at the continuously compounded annualized rate of interest and storage, net of the convenience yield.

The Eq. (2.9) resembles (2.5), which relates futures and spot prices via nominal and oil own rate of interests. By comparing these two equations, one can see that the oil own rate of interest is simply the convenience yield, net of the marginal rate of storage costs

---

<sup>9</sup>Kaldor (1939) writes that “in normal circumstances, stocks of all goods possess a yield, measured in terms of themselves, and this yield which is a compensation to the holder of stocks, must be deducted from carrying costs proper in calculating net carrying costs. The latter can, therefore, be negative or positive.”

<sup>10</sup>See, for example, Geman (2005) and Hull (2018).

$$b = y - u \quad (2.10)$$

In other words, the own rate of a commodity is simply the difference between its benefits for consumption and the cost to store it.

To grasp the meaning of the convenience yield better, it is useful to compare it to the convenience of holding cash in the monetary economy. As suggested by the Keynesian liquidity preference theory of money, three primary motives for holding cash can be categorized as transactional, precautionary, and speculative. First, cash is handy for facilitating immediate transactions. Second, we keep some cash for precautionary reasons, as a safety buffer against unforeseen disruptions. Third, when we make large investments, such as purchasing a house, we partially speculate on the value of money today relative to its value in the future.

The convenience of holding oil inventories can largely be explained by the same three motives. The primary consumer of crude oil is a refinery. The transactional motive is driven by the rigidity of refining operations. To accommodate short-term fluctuations in the end-user demand, it is more efficient for a refinery to hold extra barrels in storage instead of going through the hassle of frequently adjusting output schedules. The precautionary motive, like in the case of money, is simply a safety buffer against unexpected disruptions which could be caused, for example, by hurricanes or pipelines outages. The speculative motive for holding oil stocks though is more widespread than it is for holding fiat money. Being in the very center of the petroleum value chain, many refineries possess superior fundamental knowledge and often attempt to leverage it by adjusting inventories based on their views of supply and demand.

Besides their main processing function of turning crude oil into a usable product, refineries also provide a valuable auxiliary service to deliver the right quantities of each product to customers whenever they need it. This is what a good retail convenience store would do and then include the service charge for warehousing in the price of the product. Pushing storage onto the shoulders of end-users would be impractical. Without an investment in a specialized infrastructure, we can only store as much gasoline as our vehicle's fuel tank allows. While we may call it a convenience yield for the refinery, from the end-user perspective, such an environment causes more of an inconvenience, as low inventories are associated with higher prices.

Neither convenience yield nor storage costs are directly observable in the oil market. What one can infer from the futures curve is only their combined effect, which is the own rate. Nevertheless, such a theoretical decomposition is still useful for understanding that the own rate is driven by the tug of war between the intrinsic value of the commodity and the cost to store it. The own rate of interest is one of the most important drivers of many trading strategies. Futures traders though give it a different name, the *roll yield*.

## 2.4 The Roll Yield

So far, we have only looked at the market from the perspective of a physical trader who has access to a storage facility. However, most financial traders do not have this access. The best that they can do in trying to approximate the behavior of spot oil prices is to buy the nearest maturity futures contract and roll it to the next maturity contract at some time prior to the futures expiration. Such an approximation, however, is rather tricky, as holding and rolling futures may result in a very different outcome from the strategy of holding physical barrels.

Let us assume that one buys the futures contract at price  $F(t, T)$  and holds it until expiration when the futures price converges to the spot price  $S(T)$ . The profit-and-loss (P&L) on this trade can then be decomposed into two terms:

$$P\&L(T) = S(T) - F(t, T) = (S(T) - S(t)) + (S(t) - F(t, T)) \quad (2.11)$$

The first term represents the change in the spot price, which is of course, is unknown at time  $t$ . The second term compares spot and futures prices at time  $t$ , both of which are observable from the shape of the futures curve at that time.

While most oil traders look at P&L in dollar terms, financial investors often prefer to measure the strategy performance using percentage returns which makes comparison easier across asset classes. As explained above, we use log-returns to make analytics less cumbersome. Similar to (2.11), the so-called *excess return* ( $ER$ ) that measures the strategy performance consists of two terms<sup>11</sup>:

$$ER(T) = \ln\left(\frac{S(T)}{F(t, T)}\right) = \ln\left(\frac{S(T)}{S(t)}\right) + \ln\left(\frac{S(t)}{F(t, T)}\right) \quad (2.12)$$

The first term in this decomposition is defined as the *spot return* ( $SR$ ). It represents a hypothetical investment return on buying a physical barrel without incurring any storage costs, which cannot be realized by trading futures. The second term is the *roll return* ( $RR$ ). The roll return also cannot be generated by means of a futures trading strategy, as it is calculated by comparing prices for contracts with different maturities. Both the spot return and the roll return act merely as attribution terms in the decomposition of the actual return on investment in futures. In practice, the roll return is calculated as the difference between the actual realized excess return and the hypothetical spot return.<sup>12</sup>

---

<sup>11</sup>The term excess return differs from the total return as the latter includes the interest on collateral. Since in this book, for the most part, we ignore the interest on collateral for futures, the two terms are often used interchangeably. The distinction will become clearer in Chap. 4 in the context of fully collateralized commodity indices.

<sup>12</sup>Unlike price changes and log-returns, simple percentage returns are not additive. This makes a similar decomposition for simple returns more complex. For simple returns, the roll return is formally defined as the difference between the excess return and the spot return.

It is more common to look at spot and roll returns in annualized terms, in which case all terms in the Eq. (2.12) must be divided by  $T - t$ . We then define the *roll yield* as the annualized roll return on this strategy as follows

$$j = \frac{1}{T-t} \ln \left( \frac{S(t)}{F(t, T)} \right) \quad (2.13)$$

In other words,

$$F(t, T) = S(t) e^{-j(T-t)} \quad (2.14)$$

It follows from (2.5), (2.10), and (2.14) that the roll yield, the own rate, and the convenience yield are related as

$$j = b - r = y - u - r \quad (2.15)$$

In other words, the roll yield can be thought of as the spread between the oil own rate of interest and the USD rate of interest. Alternatively, it can be understood as the convenience yield, net of storage and interest costs. Thus, we have defined an important variable that characterizes the shape of the futures curve and translated it into three different languages used by different market participants.

To illustrate the importance of the roll yield, consider a simple strategy where a financial investor attempts to synthesize the continuous ownership of a physical barrel by holding futures with the nearest maturity. Since futures are listed with monthly expirations, in this simplified strategy an investor buys one-month expiry futures, liquidates the contract when it expires, and buys the next maturity futures at that time. Note that such a strategy is unlikely to be implemented in practice and it is used only for illustration purposes. In the real world, most investors must liquidate futures prior to expiration with more realistic examples of a buy-and-roll strategy analyzed in Chap. 4.

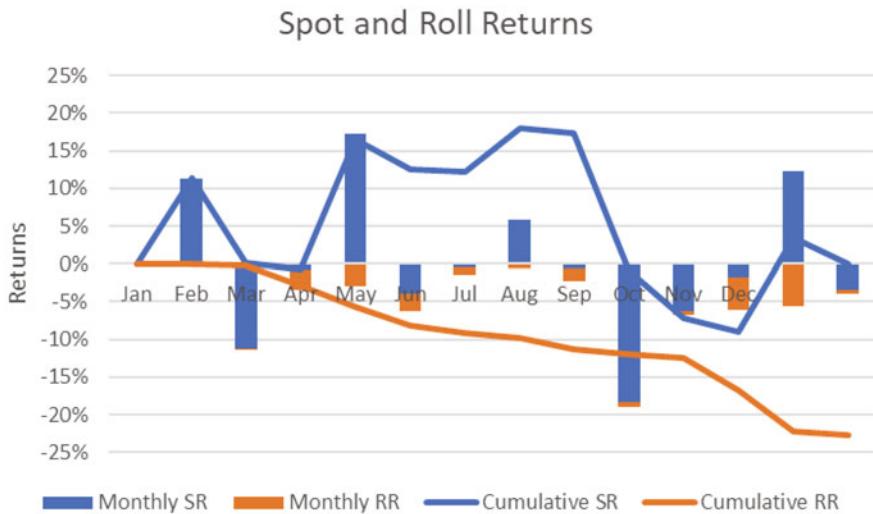
Let  $t = T_0$  and assume that the investor repeats the strategy for  $N$  consecutive months by buying  $T_i$ ,  $i = 1, 2, \dots, N$  maturity futures and successively liquidating them at the spot price at time  $T_i$ . The cumulative excess return (*CER*) of this strategy is

$$CER = \sum_{i=1}^N \ln \left( \frac{S(T_i)}{F(T_{i-1}, T_i)} \right)$$

In each period, we use (2.12) to decompose the excess return into the spot return and the roll return. The cumulative spot return (*CSR*) is then

$$CSR = \sum_{i=1}^N \ln \left( \frac{S(T_i)}{S(T_{i-1})} \right) = \ln \left( \frac{S(T_N)}{S(T_0)} \right)$$

as contributions of all intermediate spot returns cancel out.



**Fig. 2.4** An example of spot returns and roll returns (WTI, Jan-Dec 2006). While cumulative spot return was close to zero, an investment in rolling futures lost 23% because of the accumulation of negative toll returns

The cumulative roll return (*CRR*) is the difference between *CER* and *CSR*

$$CRR = CER - CSR = \sum_{i=1}^N \ln\left(\frac{S(T_{i-1})}{F(T_{i-1}, T_i)}\right)$$

Figure 2.4 highlights the significance of the roll return in futures trading.

It shows the performance of the buy-and-hold futures strategy during 2006. We chose this year for illustration because the spot price at the beginning and at the end of the period was practically the same. In other words, the spot returns oscillated throughout the year but ended up at zero. At the same time, the monthly roll returns, while being smaller in magnitude, were consistently negative. As a result, the investment in futures suffered 23% loss during the period when the spot price did not change at all. The loss was entirely driven by the accumulation of the negative roll yield as the futures curve throughout the year was in contango. We will discuss this topic in a much greater depth in Chap. 4 in the context of financialization and commodity index investments.

It should be clear by now that whatever the name of this magic variable it, the own rate, the convenience yield net of storage or the roll yield, it plays a critical role in trading futures. We now move to analyze the drivers of this key variable. We look at them from two alternative perspectives. In the next chapter, we first take a fundamental view of oil prices being driven by supply, demand, inventories, and the economics of storage. In the following chapter, we take a different approach and study it from the perspective of financial flows and the demand for hedging services. We will see that fundamentals and flows in oil trading are, in fact, deeply intertwined.

## References

- Bouchouev, I. (2022). *The Strategic Petroleum Reserve strategies: Risk-free return or return-free risk?* The Oxford Institute for Energy Studies.
- Boyle, P. C. (1899, September 6). Testimony at the Hearing of the US Industrial Commission on Trusts and Industrial Combination.
- Fisher, I. (1896). Appreciation and interest. *Publications of the American Economic Association*, 11(4), 331–442.
- Geman, H. (2005). *Commodities and commodity derivatives: Modeling and pricing for agriculturals, metals and energy*. Wiley.
- Giddens, P. H. (1947). *Pennsylvania petroleum 1750–1872: A documentary history*. Pennsylvania Historical and Museum Commission.
- Hull, J. C. (2018). *Options, futures, and other derivatives* (10th ed.). Pearson.
- Kaldor, N. (1939). Speculation and economic stability. *The Review of Economic Studies*, 7(1), 1–27.
- Keynes, J. M. (1936). *The general theory of employment, interest, and money*. Macmillan.
- Naldi, N. (2015). Sraffa and Keynes on the concept of commodity rates of interest. *Contributions to Political Economy*, 34(1), 17–30.
- Smiley, A. W. (1907). *A few scraps: Oily and otherwise*. The Derrick Publishing Company.
- Sraffa, P. (1932). Dr. Hayek on money and capital. *The Economic Journal*, 42 (165), 42–53.
- Whiteshot, C. A. (1905). *The oil-well driller: A history of the world's greatest enterprise, the oil industry*. Acme Publishing Company.



# Fundamentals, Storage, and the Model of the Squeeze

3

- The oil market is an example of a complex dynamic system. Such systems operate in **feedback loops** with their behavior driven by important boundaries. In the oil system, one boundary is set by zero inventories. Another boundary is the constraint on the available storage capacity.
- The conventional theory of storage analyzes the joint dynamics of the oil system, including interactions between supply, demand, inventories, and prices. However, its practical use is limited by the low price elasticity of oil supply and demand and challenges in measuring production and consumption in a timely manner.
- In a more practical modeling alternative, prices are viewed as financial derivatives of stochastic mean-reverting inventories that are directly observable. The boundary conditions are defined by the squeeze of futures traders when either inventories are low or storage capacity is scarce.
- The infamous episode of negative oil prices presents an important case study of a financial squeeze. While storage was the initial catalyst for this event, the enforced behavior of financial market participants caused unprecedented chaos.

## 3.1 The Invisible Hand of Storage Boundaries

Storage plays a vital role for any consumable commodity. It allows supplies to be shifted in time, from days when they are plentiful to days when they are scarce. The problem of optimal resource allocation through time in the presence of uncertainty goes back to the very beginning of human civilization. It is a difficult problem to solve. Imagine yourself being Robinson Crusoe on a remote island, where every day the decision must be made on what portion of the existing food supplies to consume today, when you do not know how much more, if anything, you will be able to find tomorrow. The decision how much to eat today impacts the availability of food tomorrow, which also depends on how much you expect to find tomorrow, and

tomorrow's decision will depend on expected supplies and availability of food the day after tomorrow, and so on.

A solution to a problem of this nature **depends on the boundary conditions**. For Robinson Crusoe, the boundary is harsh, as at some point he may run out of food. The mere presence of this zero-inventory boundary and the possibility of it being reached in the future impacts the decision today. The less food remains in stock, the more strongly the presence of the zero boundary is felt. Moreover, if the resource is bulky, or subject to degradation, such as raw fish, then its storage is also capped by the limited warehousing capacity. If the commodity cannot be easily discarded, then the presence of the second boundary on the available storage capacity becomes equally important.

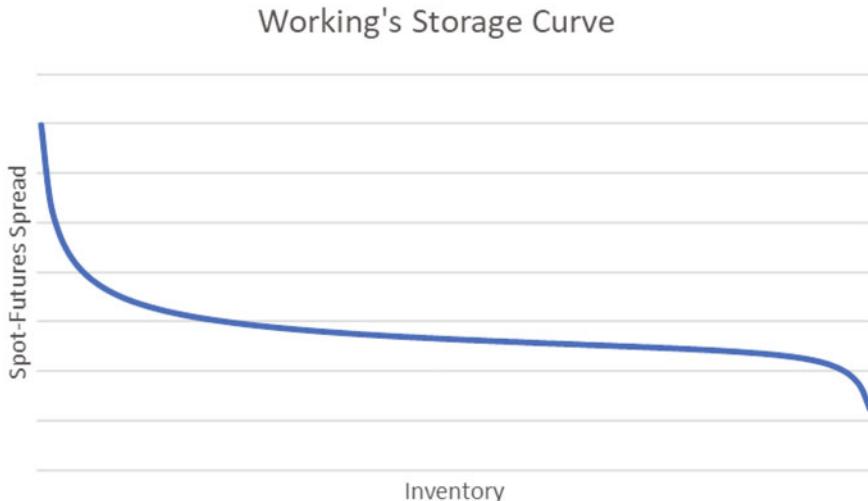
To prevent either boundary from being breached, some changes must happen to commodity supply or demand, which jointly regulate the flow of inventories. In this problem, the critical *level* of the stock affects the *flow* of the stock. Such a property is the hallmark feature of a *complex dynamic system*. The property induces mutual causality between components of the system that operates via multiple feedback loops. Some loops are known as *positive, or reinforcing, loops* that let the system grow. Other loops are *negative, or balancing, loops* that keep the growth from spiraling out of control. The world economy, the human brain, and climate evolution are examples of such dynamic systems. The problem of commodity storage is another one. Its primary components are supply, demand, inventories, and prices. *The dynamic theory of storage* is the study of the behavior of such a system.

Early studies of commodity storage were focused on agricultural markets, as oil prices remained largely regulated until the late 1970s. The goal of these studies was to assist governments with market-based ways to reduce commodity price volatility as an alternative to price controls, quotas, or subsidies. The work on commodity storage and its impact on food prices was pioneered by Holbrook Working during the Great Depression.<sup>1</sup> He looked at a storage operation as a commercial enterprise whose profitability is determined in the marketplace by the competition among professional storage operators for the business of the carry trade. As a consequence, the price of storage and the quantity of storage must be determined jointly in equilibrium by the intersection of the supply and demand curves for the service of storage.

In this framework, the storage reaches its equilibrium state when the marginal revenue earned from selling the product equals the marginal cost of making it. Such equilibrium is effectively spelled out by the Eq. (2.8) of the carry trade, discussed in the previous chapter

---

<sup>1</sup>Working's original observations on commodity storage were published between 1929 and 1933 in a series of short articles issued by *Wheat Studies of the Food Research Institute*. The more complete version of his argument is summarized in Working (1948) and Working (1949). The approach was further extended by Brennan (1958).



**Fig. 3.1** Working's representation of the relationship between the price of storage and inventories. The sign of the spot-futures spread  $S - F$  is flipped for consistency with the convention in the oil market

$$F(t, T) - S(t) = U + R - Y \quad (3.1)$$

The left-hand side of this equation,  $F - S$ , aptly labeled by Working *the price of storage*, is the **marginal revenue received by the storage owner**. The price of storage is **determined by market prices**. The right-hand side represents the total cost. It consists of the **physical cost of storage operations  $U$** , plus the **financing costs  $R$** , but it treats **convenience benefits  $Y$**  as an offset, or as a negative cost. The price of storage, therefore, can be **either positive or negative depending** on the relative magnitude of storage costs, including the costs of financing, and convenience yield.<sup>2</sup>

The core of Working's thesis is the claim that both sides of the Eq. (3.1) are driven primarily by the level of inventories, and that the relationship between prices and inventories becomes highly nonlinear when inventories are particularly low. Low inventories suggest high convenience yield and correspond to a backwardated market when  $S(t) > F(t, T)$ . In the limiting case of inventories approaching zero, the spot price can rise relative to futures without any bound. As inventories accumulate, the contribution of the convenience yield diminishes, and the price of storage is determined more by the cost of storage. The storage cost in this framework is assumed to be mostly constant, increasing only slightly for exceptionally high levels of inventories. The approach is summarized graphically in Fig. 3.1, where for consistency with conventions in the oil market, the sign of  $S(t) - F(t, T)$  is flipped compared to Working's original presentation.

<sup>2</sup>For simplicity, we use the terms convenience benefits and convenience yield interchangeably and refer to the previous chapter for their definitions.

Even though Working's graph is sometimes referred to as the first *theory of storage*, his argument did not follow any analytical theory or model. It only conceptually represented the nonlinearity of convenience yield at low inventories based on empirical observations in several agricultural markets. The objective of a proper storage model is to analytically describe the behavior of the system that generates such nonlinear dynamics of prices versus inventories within the model itself.

---

### 3.2 The Canonical Theory of Storage

**Storage is modeled differently in academic studies and by practitioners.** We start with a more theoretical outline of the problem, known as the *canonical theory of storage*.<sup>3</sup> The limitations of this theory will then lead us to a more practical alternative. The idea behind the dynamic modeling of storage is to consider zero inventory as a critical boundary condition whose impact is propagated backward in time to the present. If inventories fall to zero because of persistent excess of demand relative to supply, then the spot price must continue to rise until the point of demand destruction. The possibility of this zero-inventory boundary being reached in the future impacts the price at all prior times. This price premium resulting from the risk of running out of stock induces the convenience yield. The closer the inventory level is to the zero boundary, the higher and more nonlinear the convenience yield becomes.

Let us specify the components of the oil inventory system. Let  $x(t)$  and  $x(t - dt)$  represent the level of inventories at times  $t$  and  $t - dt$ , where  $dt$  is an increment of time. Inventories change whenever supply and demand are unbalanced. The flow of inventories is then equal to the difference between today's supply  $\tilde{Z}(t)$  and demand  $D(t)$ :

$$x(t) - x(t - dt) = \tilde{Z}(t) - D(t) \quad (3.2)$$

The supply  $\tilde{Z}(t)$  is the sum of the local production and imports from other regions. The total demand  $D(t)$  is made up of the local consumption and exports. We use tilde to highlight that the supply  $\tilde{Z}(t)$  is uncertain, and, for simplicity, we assume the demand to be deterministic. The problem remains conceptually similar if, instead, the uncertainty is attributed to demand, or to both supply and demand. The Eq. (3.2) is an accounting identity; it does not represent any model or assumption.

The integration of supply and demand imbalances over a period  $(t, T)$  results in the aggregate level of inventories accumulated over this period

---

<sup>3</sup>The theory of storage was pioneered by Gustafson (1958) and subsequently extended in multiple directions. The formulation that we present here is largely based on the ideas of Deaton and Laroque (1992). This method has been applied to the oil market by Dvir and Rogoff (2009). For other related methods of solving this problem, see Williams and Wright (1991), Routledge et al. (2000), and Pirrong (2012).

$$x(T) = x(t) + \int_t^T (\tilde{Z}(t) - D(t)) dt$$

The cumulative level of inventories must satisfy two operational constraints. The inventories cannot be negative, or more precisely, they cannot fall below a minimum operating capacity level  $X_{min}$ , as some amount of inventories must be maintained at all times for operational reasons. Similarly, the inventories cannot exceed the maximum operating capacity  $X_{max}$ , which is generally below the nominal shell storage capacity  $X$ :

$$0 < X_{min} \leq x(T) \leq X_{max} < X$$

If demand exceeds supply, then cumulative inventories decrease. They continue to decrease until the lower boundary is reached, a situation sometimes called a *stock-out*. In practice, the boundary is unlikely to be reached. As inventories approach some critically low level, the system sends a signal in the form of a higher price to supply and demand, indicating that the danger is near, and something must change to prevent any further outflows. Either the supply must increase, or the demand must fall.

The opposite occurs when supply exceeds demand. In this case, the inventories can build up until they reach so-called *tank-tops* when the storage is operationally full. The upper boundary of the maximum storage capacity then sends a lower price signal to force either supply to be restricted or demand to increase. If, for whatever reason, the demand disappears altogether, then the supply must also fall to zero to prevent a breach of the storage capacity boundary. The presence of a hard boundary on storage capacity differentiates oil from many other commodities, such as metals and agricultural products, for which storage technology is abundant and relatively straightforward. We will see throughout the book that this boundary induces distinctive price behavior which makes it difficult to apply generic commodity models to trading oil.

We can now add the spot price of oil,  $S$ , to the system. Its role is to convey information about the boundaries to supply and demand. We define the demand to be a decreasing function of price:

$$D = f(S), \frac{\partial f}{\partial S} < 0$$

The price is then expressed in terms of demand as the inverse demand function

$$S = f^{-1}(D)$$

The price elasticity of demand is defined as the ratio of the percentage change in quantity demanded to the percentage change in price.<sup>4</sup> For oil, such elasticity is very low, which means that the inverse demand function is extremely steep and nearly vertical. In other words, the oil price must change by a large amount to induce even a small change in oil consumption. In the hypothetical asymptotic case of a complete stock-out, the price must rise towards infinity if, for a given supply, consumption cannot be restricted.

The presence of storage helps to alleviate short-term deficits and buys the system some time to adjust. Let us define the availability  $a(t)$  as the sum of the supply  $\tilde{Z}(t)$  and the inventory carried from the previous period  $x(t - dt)$ . The problem of resource allocation is the problem of splitting availability between the current demand  $D(t)$  and the inventory to be carried at the end of the period  $x(t)$ , so that:

$$a(t) = \tilde{Z}(t) + x(t - dt) = D(t) + x(t)$$

Likewise, the availability  $a(t + dt)$  for tomorrow is the sum of tomorrow's production and inventory carried from today:

$$a(t + dt) = \tilde{Z}(t + dt) + x(t)$$

The presence of the same variable  $x(t)$  in both equations for today and for tomorrow highlights the intertemporal nature of the storage problem. While this variable represents a carry-out portion for the availability today, it also acts as a carry-in for the availability tomorrow. The availabilities on both days are, therefore, linked as follows

$$a(t + dt) = a(t) - D(t) + \tilde{Z}(t + dt) \quad (3.3)$$

In other words, tomorrow's availability is equal to today's availability minus what is consumed today, plus whatever can be replenished tomorrow, with the latter being uncertain. The theory of storage looks at the price of oil as an implicit function of its availability,  $S(t) = S(a(t))$ . Both price and availability depend on inventory decisions, and they must be determined jointly in equilibrium.

We now differentiate between two states of inventories: one normal state that corresponds to non-zero inventories, and a second stock-out state when inventories are depleted. In the first state, when some oil is available, the spot price must be linked to the forward price for the next period  $F(a(t + dt))$  via the economics of the carry arbitrage trade. In the second case of no inventories, the price is entirely driven by immediate supply and demand and is determined by the inverse demand function. Therefore,

---

<sup>4</sup>More formally, the price elasticity of demand is defined as  $e = \frac{dD/D}{dS/S}$ . While it is difficult to measure this elasticity precisely, it has been steadily declining over time, as increasing overall wealth made consumers less sensitive to energy prices. See, for example, Hamilton (2009) and Kilian (2020).

$$S(a(t)) = \begin{cases} F(a(t + dt)) - C(x), & \text{if } X_{min} < x < X_{max} \\ f^{-1}(a(t)), & \text{if } x = X_{min} \end{cases} \quad (3.4)$$

Here,  $C(x)$  represents the total storage cost, which can also depend on the level of inventories. To simplify the problem, we assume that  $C(x)$  is a known nonlinear function that rises without limit as  $x \rightarrow X_{max}$ . Alternatively, one could have considered a third state where inventories reach the maximum storage capacity,  $x = X_{max}$ . In this simplified two-state formulation of the problem,  $C(x)$  can be viewed as a mirror image of the inverse demand function at low inventories. Note that in the stock-out state, the availability is the same as the demand, and  $D(t)$  in the argument of the inverse demand function is replaced by  $a(t)$ .

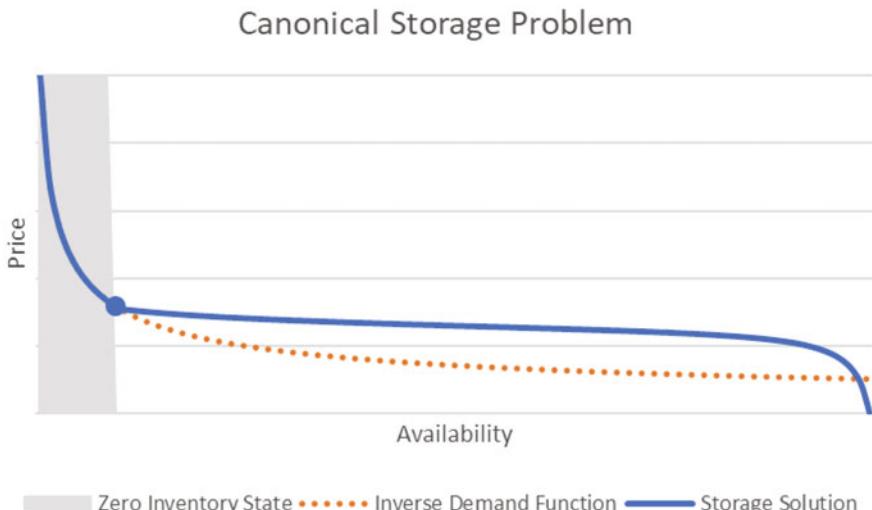
In the traditional formulation of the theory of storage, the futures price in (3.4) is replaced by the expected spot price at time  $t + dt$ , which depends on the availability  $a(t + dt)$ .<sup>5</sup> The expectation is taken over all uncertain realizations of future production  $\tilde{Z}(t)$ , which impacts the availability  $a(t + dt)$  via (3.3). This highlights the circular nature of the problem caused by the feedback loop between prices and inventories. The price today depends on the availability today via the demand function. Likewise, the price tomorrow depends on the availability tomorrow. The prices today and tomorrow are connected via the carry trade. At the same time, the availability today and the availability tomorrow are linked via the Eq. (3.3). This problem is intertemporal, as a similar loop exists for tomorrow, the day after tomorrow, etc. To solve such a problem, one must construct an iterative process which converges to some equilibrium state for both price and availability. The process also needs a boundary condition, which in this case is defined by the possibility of a stock-out.

This is a complex problem. Think about Crusoe's decision-making process as a tree, where at each time period alternative decisions could be made about food allocation between immediate consumption and its storage for tomorrow. As time moves forward, the tree of possible scenarios branches out. Each node represents total supplies that will be held at that time, which is the sum of inventories carried from the prior period and whatever else he can find on that day. The process continues until some random path leads to the case of zero inventories. The stock-out sets the terminal boundary condition for the process, and the problem is solved iteratively backward in time to establish the inventory management rule at each prior step that maximizes the usage of all inventories.

In general, to solve this problem, techniques of *stochastic dynamic programming* must be applied. Fortunately, for the purpose of this book, a precise solution of this complex problem is not needed. Our primary interest is twofold. First is to use this framework to illustrate an important feedback loop that exists between prices and inventories. Second is to explain its limitations when applied to oil markets by

---

<sup>5</sup>The assumption that the futures price is equal to the expected spot price implies that the futures price is fair and unbiased. It is equivalent to the so-called risk-neutral pricing that we will formally introduce in later chapters. In the next chapter, we also consider the case when futures price differs from the expected spot price, as futures may be distorted by imbalances in the hedging market.

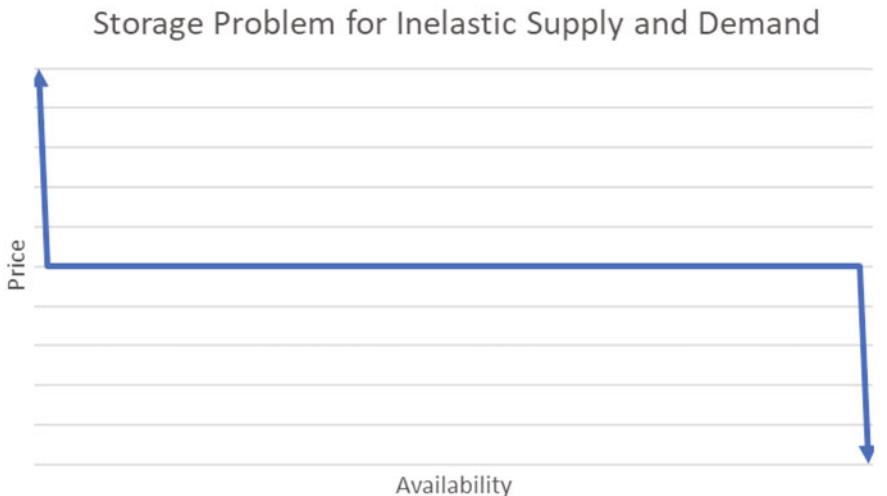


**Fig. 3.2** A typical solution of the canonical storage problem has a characteristic kink that separates the consumption region with zero inventories from the investment region driven by the economics of the carry trade

simply looking at salient properties of a typical solution. We then use some valuable insights resulting from these limitations to develop a simpler and more practical alternative approach in the following section.

A typical solution to this problem shows the equilibrium price as a function of availability, as illustrated in Fig. 3.2. It is characterized by the kink that corresponds to the level of zero inventories separating the two states, like in the Eq. (3.4). This kink symbolizes the dual role of commodities, as an investment asset and as a product for consumption. The location of the kink that marks the danger zone of a stock-out is what the storage problem is effectively solving for. The price in the region to the left of the kink is solely determined by production and consumption. It is computed via an inverse demand function, as there is no storage buffer in this region. To the right of the kink, oil behaves as an investment asset. In this state, the price is mostly determined by the cost of storage, which increases nonlinearly near the boundary on the maximum storage capacity. It should also be noted that conventional models require prices to remain positive. However, for the oil market this assumption, as we will see shortly, is too restrictive.

Despite its popularity in academic studies, the standard approach to the storage theory attracted only very limited interest among oil traders. One practical challenge is the extremely low price elasticity of oil demand. The solution to the problem at zero inventories effectively degenerates into a straight vertical line. It means that if we truly run out of oil stocks, and the supply of oil cannot meet the demand, then the price must rise towards infinity. Therefore, the optimal solution is to always store oil in sufficient quantities so that a stock-out is never reached. However, this solution is impractical because of the high cost of oil and the limitation on the storage capacity.



**Fig. 3.3** The solution to the canonical storage problem degenerates at the boundaries when the price elasticities of demand and supply approach zero

A similar problem occurs when inventories reach maximum storage capacity. Like oil demand, the short-term oil supply is also inelastic with respect to price, as many producers cannot shut down production instantaneously without permanently damaging oil reservoirs. In this case, if a significant portion of the oil demand disappears, like it did, for example, during the early days of the Covid-19 pandemic, then the excess oil must go to storage. Once all storage is filled, then the price must theoretically fall without any boundary to force either production or consumption to change. We will discuss shortly the infamous episode of oil falling to negative forty dollars per barrel, which is probably as close as one can get to negative infinity in financial markets.

To summarize, given that the price elasticity of oil demand and oil supply is extremely low, the optimal solution to the oil storage problem collapses into something that resembles the graph in Fig. 3.3. It is driven by two extremities. The left tail marks the case of zero availability, when the price goes to infinity. The right tail indicates that the storage capacity is full, and the price goes to minus infinity. Anywhere in between, where inventories are deemed to be normal and sufficiently far away from either boundary, the price becomes largely insensitive to the availability of inventories.

We now take this insight from the shortcomings of one model to develop a much simpler and more practical alternative for the relationship between prices and inventories that can actually be used by oil traders. To do this, we borrow some wisdom from the physics of extreme events.

### 3.3 A Stylized Model of the Squeeze

Boundaries play a critical role in the dynamics of any complex dynamic system, such as the oil market. The conventional theory of storage shows how the presence of two storage boundaries affects the behavior of prices even away from the boundaries. One boundary marks the scenario of running out of oil, in which case the spot price moves higher until either supply or demand adjusts. The second boundary corresponds to the scenario of running out of storage, which forces the spot price to drop to prevent further inflows. These boundaries do not need to be reachable. In fact, in most dynamic systems boundaries are understood as unattainable limits, but their mere existence determines the system behavior within the boundaries.

What makes the traditional theory of storage particularly complex is the attempt to explicitly model the response of prices to changes in supply and demand. In practice, the concept of an inverse demand function that plays an important part in the theory has limited applications. In the oil market, neither supply nor demand is easily observable. Supply data is published with a significant lag and frequently revised, sometimes even going back several years. Measuring oil demand is even harder. The demand is typically synthetically reconstructed from estimates of supply and measured inventory data using the identity (3.2). The primary variable that traders can track is inventory. For the storage model to be of any practical use, it must be based on observable inputs. Instead of making assumptions about supply and demand and calculating inventories, the trader is better off modeling the behavior of inventories directly.

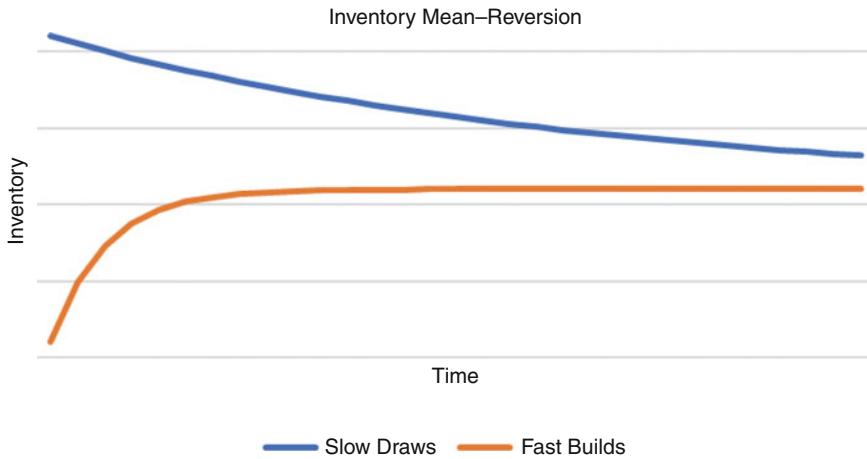
The presence of boundaries causes supply and demand to adjust in a complicated way that we will not even attempt to model. Instead, we model the uncertainty driven by these fundamental adjustments that must cause the *flow* of inventories to reverse when the *level* of inventories approaches the boundary. Such behavior can be characterized by a mean-reverting process, where inventories are pulled towards their long-term equilibrium level. We first illustrate mean-reverting behavior in an idealized case where the inventory path towards equilibrium is certain, or, in other words, deterministic. We then add some noise or uncertainty to the process.

In the deterministic case, the change in the level of inventories  $dx$  over a period  $dt$  can be described by the following equation

$$dx = k(\bar{x} - x)dt \quad (3.5)$$

Here,  $\bar{x}$  represents the long-term equilibrium level of inventories, and the parameter  $k$  defines the speed at which inventories revert to this level. If current inventories are above their long-term level, i.e.,  $x(t) > \bar{x}$ , then the Eq. (3.5) implies that  $dx < 0$ , and, therefore, inventories decrease. Likewise, if current inventories are below the long-term level, i.e.,  $x(t) < \bar{x}$ , then  $dx > 0$  and inventories rise.

If we assume some initial state of the inventories at time  $t = t_0$



**Fig. 3.4** Examples of slowly drawing and rapidly building inventories that converge to a long-term equilibrium

$$x(t_0) = x_0$$

then the solution to the ordinary differential Eq. (3.5) can be easily found, as follows

$$x(t) = \bar{x} + (x_0 - \bar{x})e^{-k(t-t_0)} \quad (3.6)$$

One can easily verify it by direct substitution. As time increases, the inventories  $x(t)$  are pulled from their current state  $x_0$  towards the equilibrium level  $\bar{x}$ . The pull occurs at an exponential rate defined by the parameter  $k$ . As  $t \rightarrow \infty$ , inventories converge towards the equilibrium, i.e.,  $x(t) \rightarrow \bar{x}$ . The larger the parameter  $k$ , the faster the speed of mean-reversion.

The solution (3.6) is illustrated in Fig. 3.4 for two scenarios. In the first example of *slow draws*, the initial inventory level  $x_0$  is high and  $k$  is relatively small, which could be the result of a prior fall in demand, perhaps caused by an economic recession. As demand gradually recovers, inventories decrease and slowly converge to the long-term equilibrium. In the second scenario of *rapid builds*, initial inventories are low but mean-reversion is fast, which could be the result of a short-term supply disruption caused by a geopolitical event, but the supply is assumed to be restored relatively quickly.

The Eq. (3.5) describes an unrealistic dynamic where one can predict future inventories with certainty. In the real world, the path towards an inventory equilibrium state is likely to be noisy with some random fluctuations along the way that can either delay the convergence to the normal state or accelerate it. To capture this noise, a second term is added to the Eq. (3.5):

$$dx = k(\bar{x} - x)dt + \sigma dz \quad (3.7)$$

The noise is assumed to come from an increment  $dz$ , which represents a random variable, drawn from a normal distribution with mean of zero and variance equal to  $dt$ . It is scaled by the volatility parameter  $\sigma$ , which characterizes the magnitude of the uncertainty.

As time  $t$  moves forward, the market may experience large, unexpected shocks to supply and demand that are described by large random values of  $dz(t)$ . During such episodes, the random term in (3.7) dominates the change in inventories. However, over time the steady gravitational pull of the mean-reverting term ensures that inventories drift towards an equilibrium level. The further inventories deviate from the normal level, the stronger the pull towards the long-term equilibrium. The Eq. (3.7) is one of many examples of stochastic processes that are commonly used to describe uncertainty in financial markets. It is known as an *Ornstein-Uhlenbeck process*, which was originally developed in physics to analyze Brownian motion with friction. It has been widely adopted for modeling mean-reversion in commodity prices.

When noise is introduced to the system, the forward dynamics is no longer certain. It can only be understood in a probabilistic sense. Each stochastic process is associated with a probability density function  $p(x, t; x_T, T)$  which describes transition probabilities of the random variable between two states at times  $t$  and  $T$ . If the variable at time  $t < T$  is known to be located at the point  $x$ , then the density function describes the probability of this variable reaching various points  $x_T$  at some future time  $T$ . Alternatively, if the variable is known to be at some target state  $x_T$  at time  $T$ , then the same density function also represents the probability of having reached this target state starting from different levels  $x$  at an earlier time  $t < T$ . The probability density function satisfies the *Kolmogorov forward and backward equations*, with the former also known as the *Fokker-Planck equation*. These equations and other basic facts related to transition probabilities and stochastic processes are summarized in Appendix A.

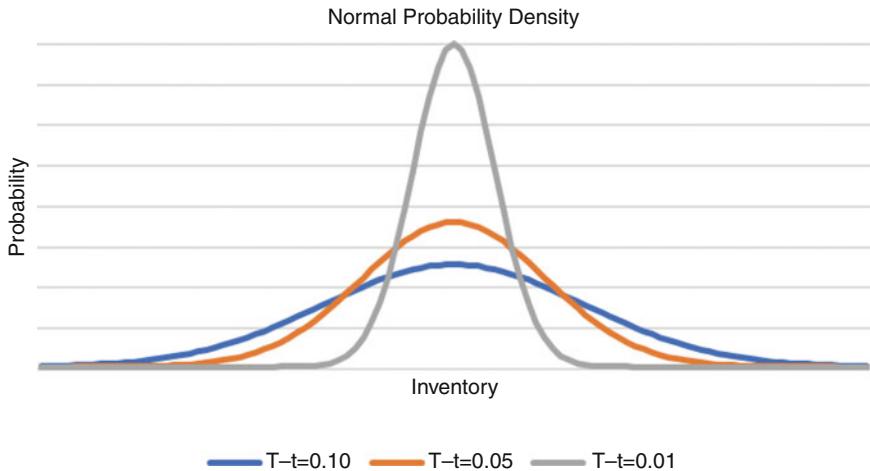
The most well-known probability density function is a Gaussian bell-shaped curve that represents the normal distribution, centered at the point  $x_T = x$  with variance  $\sigma^2(T - t)$ :

$$p_N(x, t; x_T, T) = \frac{1}{\sigma \sqrt{2\pi(T-t)}} e^{-\frac{(x_T-x)^2}{2\sigma^2(T-t)}} \quad (3.8)$$

This normal probability density corresponds to the stochastic process (3.7) without mean-reversion, i.e., for  $k = 0$ , which is known as *Arithmetic Brownian Motion (ABM)*.

Figure 3.5 illustrates the normal probability density functions (3.8) for different  $T - t$ .

As  $T - t$  decreases, the normal probability density function becomes thinner and taller, as the variance is reduced and the probability mass is concentrated around the mean, where  $x_T = x$ . In the limiting case when  $t \rightarrow T$ , the probability density takes a



**Fig. 3.5** Normal probability density for decreasing time variance

very peculiar form. It falls to zero for all values of  $x_T$ , except for a single point  $x_T = x$ , where it goes to infinity.

This limit is known as the *Dirac delta function*  $\delta(x_T - x)$ , which is defined as

$$\delta(x_T - x) = \begin{cases} 0, & x_T \neq x \\ \infty, & x_T = x \end{cases} \quad (3.9)$$

The Dirac delta function describes the boundary of the system. It is frequently used in physics to model extreme but largely unattainable states. One popular example arises in modeling the dissipation of a point heat impulse across a metal bar. As time passes, the heat diffuses along the bar according to the differential equation of heat transfer, for which the Dirac delta function serves as an initial condition. The Dirac delta function possesses some interesting properties. Defined as the limit of probability density functions which must integrate to one, it also integrates to one. Also, when the Dirac delta function concentrated at  $x$  is multiplied by an arbitrary function and integrated over all possible values of  $x_T$ , the integral returns the value of the function at the point  $x$ . These properties are summarized in Appendix A.

We are now in a position to apply the technique of stochastic calculus to describe the behavior of oil inventories and prices. Stochastic processes are frequently used to model oil prices with so-called *reduced-form models*. In such models, one assumes a certain stochastic process for the spot price and for some additional state variables, such as the convenience yield or interest rate. The theoretical futures price is then derived as a function of parameters of stochastic processes for state variables, which are calibrated to market prices of traded futures. This approach has been the dominant modeling framework for nearly three decades, starting from the 1980s,

but its relevance has waned over time.<sup>6</sup> As the futures market developed and became observable, the need to model futures using somewhat nebulous state variables, such as convenience yields, largely disappeared. Perhaps more importantly, reduced-form models ignore the fundamental dynamics of the entire energy system, which made this approach less popular among practitioners.

The canonical storage does take into consideration the dynamics of the system but it is equally unpopular among practitioners, as it is complex, unstable, and based on hard-to-measure variables, such as supply and demand. In this chapter, we propose a more practical approach to the problem by blending the two classical paradigms and applying stochastic calculus instead directly to inventories, which is the primary fundamental variable that traders track.

While the traditional theory of storage attempts to link inventory to the outright price of oil, traders instead typically relate inventories to the spread between the spot and the futures price. The spot-futures spread isolates the exposure to fundamentals, as two legs of the spread provide largely offsetting exposure to many exogenous non-fundamental factors. It is also consistent with Working's original formulation of the storage problem. In practice, the spot oil price is approximated with the nearest-delivery futures contract. The spread between futures with two maturities then serves as a barometer for the economics of the carry trade and the state of inventories.

In our practical formulation of the storage problem, we define the futures time spread as a financial derivative of inventories  $x(t)$ :

$$s(x, t) = F(t, T_1; x) - F(t, T_2; x)$$

The inventory plays the role of an uncertain state variable whose behavior is described by a stochastic process, such as, for example, (3.7).

Ignoring, for simplicity, the effects of discounting, the price of a financial derivative, as shown in Appendix A, must be equal to the expected value of its terminal payoff  $s(x_T, T)$  at the expiration time  $T$ :

$$s(x, t) = \int_{-\infty}^{\infty} p(x, t; x_T, T) s(x_T, T) dx_T \quad (3.10)$$

---

<sup>6</sup>Brennan and Schwartz (1985) applied a one-factor lognormal model for the spot price with deterministic convenience yield to derive futures prices. The model was further extended in Brennan (1991). One-factor models, however, have quickly proven to be too restrictive, as they allow futures across all maturities to move only in the same direction. A popular two-factor model of Gibson and Schwartz (1990), which assumes a stochastic mean-reverting convenience yield, generates a much richer dynamics for the futures curve and volatilities. Miltersen (2003) allowed the equilibrium convenience yield to be time-dependent and showed how to make the model consistent with the futures curve. Several three-factor models have been proposed, such as Schwartz (1997), Casassus and Collin-Dufresne (2005), and Dempster et al. (2012). For surveys of reduced-form models, we refer to Clewlow and Strickland (2000), Eydeland and Wolyniec (2003), and Carmona and Ludkovski (2004).

We will see later in Chap. 8 that, in general, financial derivatives must be priced as the cost of their dynamic replication by trading underlying futures. In this case, the probabilities in formula (3.10) are understood as so-called *risk-neutral probabilities* that correspond to a stochastic process in which the drift term is zeroed out. However, when the underlying instrument represents a non-tradable state variable, such as oil inventories, the derivative of the state variable is simply equal to the expected value of its payoff under regular real-world probabilities.<sup>7</sup>

To find the value of the futures spread,  $s(x, t)$ , we need to specify the boundary condition at time  $T = T_1$  when the first leg of the spread expires, and the carry trade that connects two futures must be closed. Inspired by insights from the conventional storage theory of the previous section, we distinguish among three scenarios. In one normal scenario, the futures spread is determined by the negative of the cost of storage  $-C$ . Here, we do not need to impose any nonlinear cost of storage function, as such nonlinearity will be generated endogenously within the model, driven by the proximity to the storage boundary.

Two other scenarios correspond to two boundary cases when the carry arbitrage cannot be completed, either because of zero inventories or because of zero remaining storage capacity. In the case of zero inventories, the holder of the short futures position has no choice but to bid up the price of the nearby futures contract in the market. However, in this scenario there will be no sellers since no one can deliver physical barrels against the short futures position. The price can then keep rising without any limit. Likewise, the holder of the expiring long futures position who does not have access to storage capacity will try to sell futures to avoid taking the delivery of physical barrels by pushing the price down without any limits.

The spread behavior at these extremes that represent largely unattainable asymptotic states can be described by a pair of Dirac delta functions with opposite signs, concentrated, respectively, at  $X_{min}$  and  $X_{max}$ :

$$s(x_T, T) = \begin{cases} +\infty, & x_T = X_{min} \\ -\infty, & x_T = X_{max} \\ -C, & X_{min} < x_T < X_{max} \end{cases} = \delta(x_T - X_{min}) - \delta(x_T - X_{max}) - C \quad (3.11)$$

One can think about infinite prices as the event of default by futures traders. More practically, the market calls such an event a squeeze, and we refer to this formulation of the storage problem as a *stylized model of the squeeze*.

The solution to the storage problem now becomes simple and intuitive. It is given by the derivative pricing formula (3.10) with the boundary condition (3.11):

---

<sup>7</sup>In Bouchouev (2021), the author applied a slightly different methodology to this problem by representing the futures spread as the solution to the Black-Scholes-Merton (BSM) partial differential equation, where inventory is a state variable. Here, we use a somewhat simplified approach, as BSM equation is formally introduced only later, in Chap. 8.

$$s(x, t) = \int_{-\infty}^{\infty} p(x, t; x_T, T) \delta(x_T - X_{min}, T) dx_T - \int_{-\infty}^{\infty} p(x, t; x_T, T) \delta(x_T - X_{max}, T) dx_T - C \int_{-\infty}^{\infty} p(x, t; x_T, T) dx_T$$

Note that the integration here is effectively performed over the range  $x_T \in (X_{min}, X_{max})$ , outside of which the probability density function is equal to zero. Applying properties of the Dirac delta function (A.4) and (A.5), we obtain that the price of the futures spread is given by

$$s(x, t) = p(x, t; X_{min}, T) - p(x, t; X_{max}, T) - C \quad (3.12)$$

This representation explicitly relates the price of the futures spread to market fundamentals, specifically to the probability distribution function for stochastic inventories. It follows that the value of the futures spread at time  $t$  is determined by the relative probabilities of upside and downside squeezes, adjusted by the normal cost of storage. One squeeze corresponds to the case of zero inventories, and the other one to the case of zero available storage capacity. The proximity to each boundary determines the value of the spread within the boundaries at all times prior to expiration.

Let us now apply the generic formula (3.12) to a mean-reverting stochastic process (3.7). It is shown in Appendix A that the probability density for the mean-reverting process is also described by the normal distribution

$$p_{MR}(x, t; x_T, T) = \frac{1}{\sigma \sqrt{2\pi(T - \hat{t})}} e^{-\frac{(y(T) - y(t))^2}{2\sigma^2(T - \hat{t})}} \quad (3.13)$$

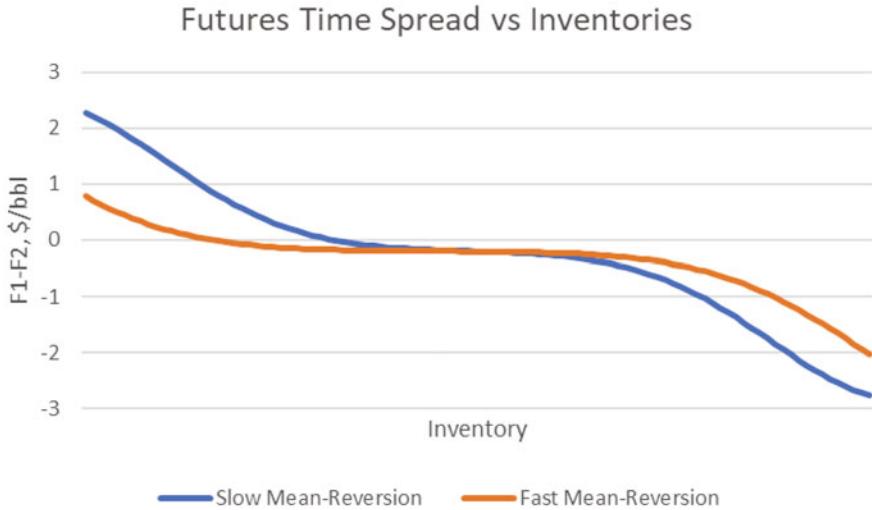
but it is applied to a modified state variable

$$y(t) = \bar{x} + (x - \bar{x})e^{-k(T - t)} \quad (3.14)$$

and to the new time variable  $\hat{t}$ , which is defined by

$$T - \hat{t} = \int_t^T e^{-2k(T - t)} dt = \frac{1 - e^{-2k(T - t)}}{2k} \quad (3.15)$$

The state variable  $y(t)$  resembles the function (3.6) that describes the behavior of inventories without the added noise. The role of the new time variable  $\hat{t}$  is to capture the impact of mean-reversion on the process variance. It reduces the effective variance via an exponential dampening factor that depends on the speed of mean-reversion  $k$ . For example, if the mean-reversion is slow and  $k \rightarrow 0$ , then  $\hat{t} \rightarrow t$  and the variance of the mean-reverting process converges to the variance of the normal process. However, for  $k > 0$  the new effective remaining time to maturity is reduced,



**Fig. 3.6** Examples of one-month futures spread functions (3.16) for fast ( $k = 4$ ) and slow ( $k = 1$ ) mean-reversions with  $0 < x < 1$  and  $\bar{x} = 0.6$

as  $T - \hat{t} < T - t$ . We should also note that at time  $t = T$ , the new time  $\hat{t} = T$ ,  $y(T) = x_T$ , and the boundary condition (3.11) remains intact in new variables.

The substitution of the probability density (3.13) into the representation (3.12) results in the following pricing formula for the futures spread:

$$s(x, t) = \frac{1}{\sqrt{2\pi(T - \hat{t})}\sigma} e^{-\frac{(y(t) - X_{min})^2}{2\sigma^2(T - \hat{t})}} - \frac{1}{\sqrt{2\pi(T - \hat{t})}\sigma} e^{-\frac{(X_{max} - y(t))^2}{2\sigma^2(T - \hat{t})}} - C \quad (3.16)$$

where the variables  $y$  and  $\hat{t}$  are defined above by (3.14) and (3.15).

The solution (3.16) to our stylized model of the squeeze is sufficiently flexible to cover the wide range of plausible spread behaviors versus the level of inventories. Figure 3.6 illustrates that faster mean-reversion makes the range of spread values more contained. In this example, the inventories are expressed as a percentage of total storage capacity, so that  $0 \leq x \leq 1$ , and the long-term equilibrium inventory level is set at  $\bar{x} = 0.60$ . Since  $\bar{x}$  is slightly closer to the upper boundary, this makes the probability of tank-tops higher than the probability of stock-outs, resulting in wider contango than backwardation near the boundaries. The opposite effect can be generated by choosing  $\bar{x}$  to be closer to the lower boundary. One can easily extend this framework and apply (3.12) to the more complex dynamics of inventories characterized by different stochastic processes and probability density functions.

We now move from theory to practice and illustrate the approach for the world's most important storage hub, where the futures market meets the physical barrels. For nearly three decades, all roads in the oil business have led to the rather nondescript

US town of Cushing, Oklahoma, which has often been called the oil capital of the world.

---

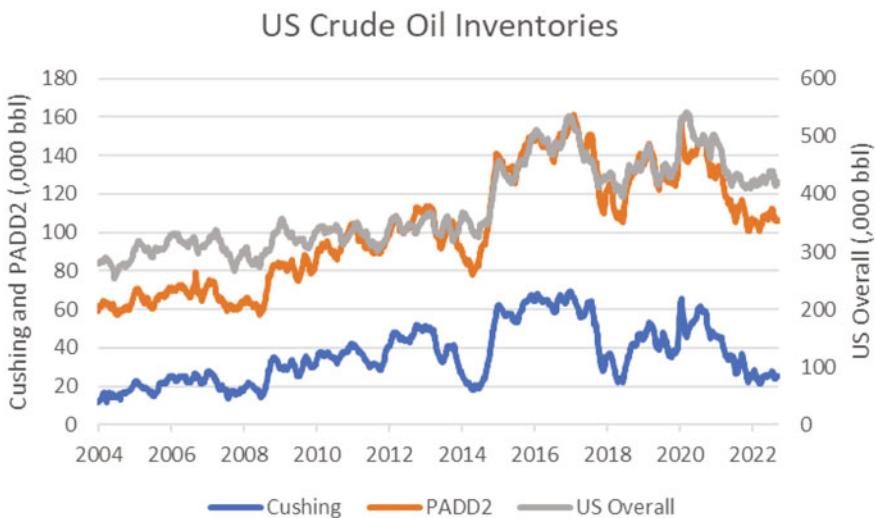
### 3.4 Cushing: Pipeline Crossroads of the World

While the theoretical thesis behind the theory of storage and the nonlinear inverse relationship between inventories and prices is rather intuitive, its validation in practice has been much more elusive. One practical challenge is availability of high-quality inventory data. For the storage model to be tested on real data, the storage facilities must be located within the proximity of the corresponding pricing point. Fortunately, the storage hub at Cushing, the delivery point for the benchmark WTI contract, presents a unique opportunity to test the theory.

Cushing is an important juncture in the oil trading system where the connection between physical and financial markets is very explicit. If a WTI futures trade is not closed before its expiration, then the holder of the short position must deliver physical barrels to one of the designated locations within Cushing tank farm, and the holder of the long futures must take barrels from there. Taking or delivering barrels in Cushing means reserving space on one of the dedicated pipelines. These pipelines are not easily accessible by many financial traders. Further, one cannot really deliver a truckload of oil and unload it there. If you do not have a membership in this elite Cushing club, then you are out of luck. In practice, if you do not close the expiring futures contract, then your clearing house will close the trade without even telling you, then ask one of the Cushing club members to handle physical barrels for a healthy fee and send you a large bill to settle.

This should look sufficiently scary to disincentivize anyone without a Cushing club membership from dabbling in WTI futures contracts near their expiration. The vast majority of rational investors indeed stay away from this game, especially when inventories are within reach of the critical boundaries. However, from time to time, some traders, driven either by greed or more likely by ignorance, decide to take their chances. The brave ones can choose to play this game until the final buzzer, often forgetting that the less time remains on the clock, the more they are exposed to the squeeze. The most infamous example of such a squeeze occurred on the day when oil prices went negative, which we describe shortly.

Cushing is a landlocked location which is connected via pipelines to production and consumption centers and to the Gulf Coast for exports and imports. The supply of oil to Cushing is measured by flows on many inbound pipelines from various production areas in the USA and Canada. The demand is estimated by flows on outbound pipelines delivering barrels to domestic refineries. Unlike many other oil trading hubs located near ports, Cushing is landlocked and better ringfenced from the impact of broader international factors. This relative isolation from global trade factors makes local inventories at the Cushing storage hub more suitable for forecasting local prices. Professional oil traders use a variety of sources and sophisticated technologies, such as aerial imagery and infrared cameras, to outsmart their peers in measuring pipeline flows and the amount of stored inventories.

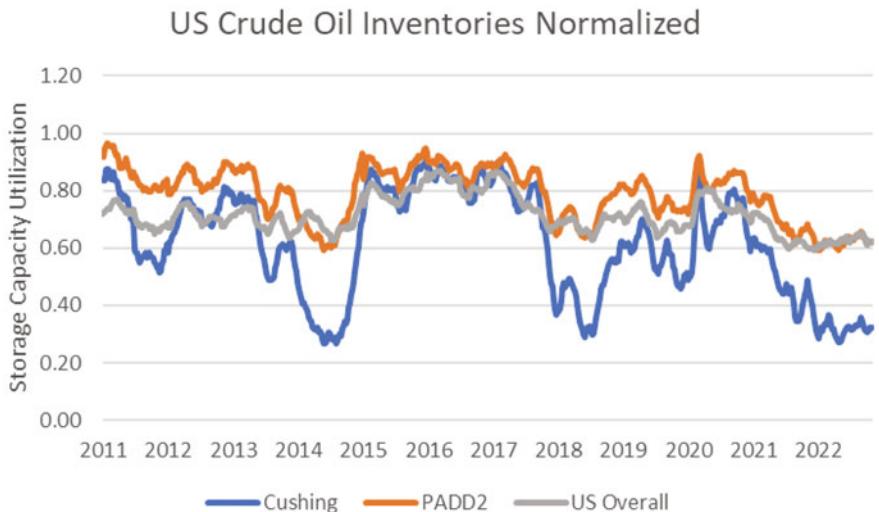


**Fig. 3.7** Inventory levels for Cushing, PADD2, and US overall exhibit the common upward trend driven by the growth of the shale industry

While the proper Cushing model requires higher-frequency private data, the main characteristics of storage modeling can be illustrated with some publicly available data. Such data is provided by the US Energy Information Administration (EIA), the statistical agency of the US Department of Energy (DOE). Every Wednesday, the agency publishes a widely anticipated report with weekly estimates of various fundamental variables, including inventories across various US regions. While the state of inventories at Cushing is perceived to be the most relevant to WTI prices, inventories in broader regions should also be considered, as barrels can be shipped quickly between certain storage locations. For example, the storage in the Midwest area, known as PADD2 (Petroleum Administration for Defense District), is well connected to Cushing and frequently used by analysts, along with overall US inventories.

Figure 3.7 shows a time series of inventories in these three locations. The data exhibits a common trend driven by the growth of US shale production and the need to support the enhanced infrastructure. The presence of the common trend in data significantly contaminates the quality of statistical correlation analysis. It would be incorrect to apply correlations, for example, to oil prices and raw inventories as neither time series is stationary. However, the futures spread is stationary as two legs of the spread negate a large portion of the common trend. To remove such a common trend in raw inventories, one must normalize them as well.

The methods of inventory scaling vary with overall objectives and the availability of data. For the broader markets, where inventory data is harder to collect, one often measures inventories relative to their historical averages, which are often seasonally adjusted. For example, OPEC uses a five-year inventory average as a benchmark for decision making, which makes traders joke that OPEC can simply wait five years to



**Fig. 3.8** Inventory levels for Cushing, PADD2, and US overall normalized by the storage capacity are stationary

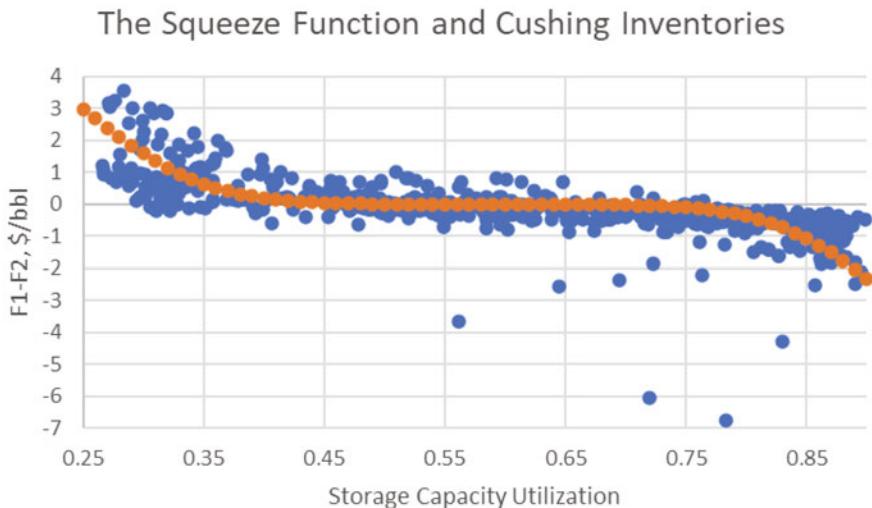
bring distorted inventories to the normal level. Other analysts measure inventories in terms of days of demand, which has been steadily growing over the years. For US markets, where more-granular data are available, a more accurate metric can be constructed if inventories are scaled with the overall storage capacity. The resulting inventory variable,  $x(t)/X_{max}$ , represents the inventory capacity utilization.

In contrast to raw inventories, which are trending, the storage capacity utilization for all three locations is stationary. Figure 3.8 clearly shows more oscillatory or mean-reverting behavior of normalized inventories with no visible trends. This metric is particularly important in modeling crucial nonlinear effects near storage boundaries, which in this case are defined explicitly.

The historical correlation between the futures spreads and normalized Cushing inventories over the entire sample since the storage capacity data became public in 2011 is approximately negative 60%. While such a high correlation validates the basic thesis of the storage theory that futures spread is inversely related to inventories, the relationship is highly nonlinear, as shown in Fig. 3.9.

To illustrate the crux of the storage theory and the nonlinear behavior around the boundaries, we overlay our stylized theoretical solution (3.16) to the storage problem. This squeeze function captures important nonlinear price signals when the market is running out of inventory or out of storage capacity. The specific model parameters can be fitted to sub-samples corresponding to prevailing market conditions. We will leave this exercise to the reader, as econometric analysis of historical data is not the main objective of this book.

So far, we have only illustrated the storage theory with a static snapshot, where the inverse relationship between futures spreads and inventories is measured contemporaneously. While it confirms the thesis of the storage theory, it does not



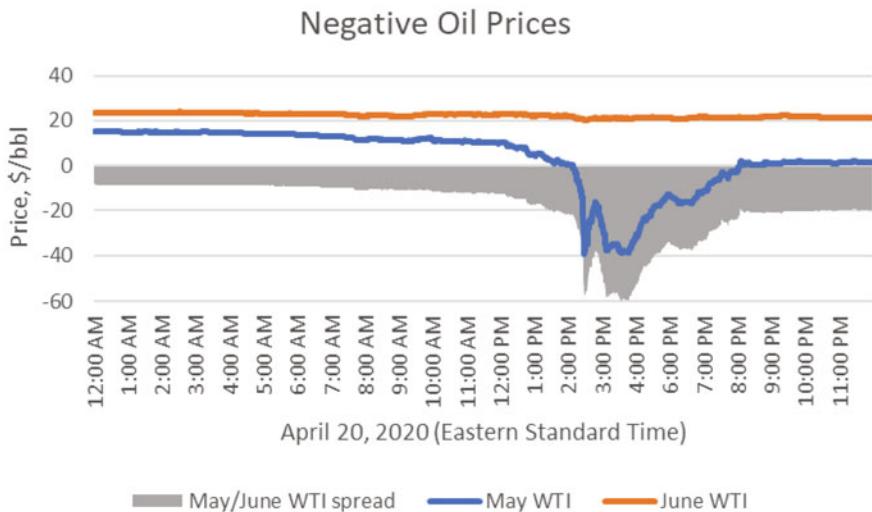
**Fig. 3.9** The inverse relationship between WTI spreads and Cushing inventories is approximated by the model of the squeeze

necessarily help to turn the concept into trading profits, as we do not know yet whether inventories have any *predictive* power for prices. The topic of actually trading futures based on quantitative analysis of inventory data is more nuanced. We will return to quantitative trading strategies based on fundamental data in Chap. 6 when we discuss so-called *quantamentals*.

One can also notice a few outliers in Fig. 3.9, specifically when the market fell into a steeper contango than was justified by prevailing inventories. The cause of nearly all of these outliers is the forward-looking nature of financial markets that trade based not on today's level of inventories, but rather based on expected forward inventories at the time of the delivery of physical barrels. For the most part, the difference between the two is minor, as inventories generally change slowly. However, if something causes inventories to change quickly, then forward expectations play the decisive role. If the development happens to be near one of the two storage boundaries, then the consequences of incorrect expectations could be dramatic. The best way to illustrate such boundary behavior is to look more deeply at its most famous case study, the unprecedented episode of negative oil prices.

### 3.5 Negative Oil Prices

On April 20, 2020, the oil market made history. The price of the prompt WTI futures contract went negative. It did not just dip slightly below zero; it collapsed to a mind-blowing negative forty dollars per barrel. For the first time in the history of financial markets, an asset was priced at the negative of its typical value. The historic price collapse was not limited to futures. The prices for physical barrels also went



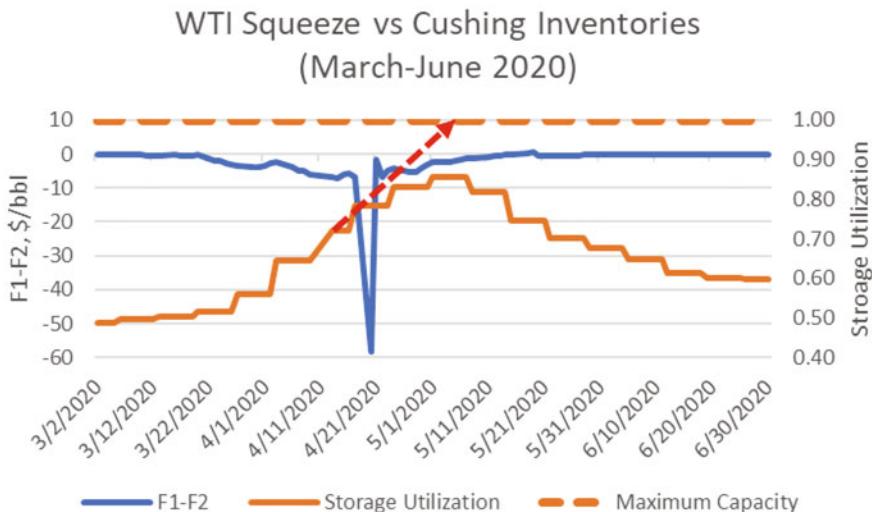
**Fig. 3.10** On April 20, 2020, the price of the May 2020 futures contract traded at negative \$40/bbl and the spread between May and June contracts at negative \$60/bbl

negative, as they were attached to the WTI benchmark via basis spreads which were too slow to adjust to the fast pace of futures.

The speed at which the futures fell was staggering. It is unlikely to be explained by any conventional economic models of rational behavior. It took the prompt oil futures only thirty minutes to move from zero to negative forty dollars per barrel, as illustrated in Fig. 3.10. At the same time, the price of the second nearby contract barely changed. The front spread, the benchmark for the price of storage, widened to negative sixty dollars per barrel, far in excess of even the most expensive way of storing oil; by moving it to the coast and loading it on ships for floating storage. Why did it happen?

In the real world, the storage boundaries work their powers somewhat differently from the theory. Prices are driven not by a magic wand of storage but rather by the behavior of specific market participants reacting to the information conveyed by storage. The market was indeed oversupplied since the onset of the Covid-19 pandemic when oil demand fell virtually overnight. The supply was slow to adjust, with the downward pressure on price further exacerbated by disagreements and a short-term price war between large sovereign oil producers. Inventories were increasing at an alarming pace, but storage was nowhere close to full.<sup>8</sup> Figure 3.11 illustrates.

<sup>8</sup>Some storage was already committed but not yet reflected in inventory data, nevertheless, even including these additional volumes, the total inventory levels were substantially below the maximum operating capacity.



**Fig. 3.11** At the time when WTI fell below zero, Cushing storage was not yet full. However, the market was extrapolating the recent inventory trend, which would have breached the maximum storage capacity

Importantly, what matters for prices are not today's inventories, but rather, inventories at the time of the expected delivery, which in this example corresponds to the following month. Forecasting future inventories during a period of such uncertainty is a difficult task. The most straightforward way to do it is simply to extrapolate the recently observed trend in inventories. Since what goes into storage is the excess of short-term supply over short-term demand, and since supply and demand are slow to adjust, the inventory change tomorrow should be broadly similar to the inventory change today. Therefore, in the short term, the slope of the inventory trend should remain intact. Such naïve extrapolation of the slope would indeed have breached the hard boundary of the maximum storage capacity sometime during the following month, when the delivery of physical barrels must take place.

This is where the important self-regulating property of the complex dynamic system comes into play; the level of the stock approaching the boundary must impact the flow of the stock to prevent the system from overfilling. As discussed earlier, the dynamics of such a system is driven by the interaction between the balancing and reinforcing loops. The balancing loop here is enforced by storage. The presence of the hard boundary forces either supply or demand to adjust and stop the flow of the stock. The demand is always made up of the immediate consumption by refineries and the demand for storage. While the former was constrained by pandemic restrictions, the latter is a function of the cost to transport oil elsewhere and to store it in more remote locations. If the physical system was left to operate alone without the disturbance caused by the financial markets, then the futures spread would have likely traded around \$5/bbl which was the cost of moving oil out of Cushing at that time and storing it elsewhere.

The challenge is that financial futures markets operate via their own feedback loop, which at a time of stress and panic, often acts in a self-reinforcing manner. Let us briefly describe the mechanics of this panic which caused some irrational behavior near the boundary, leading to the squeeze. Recall that financial investors gain exposure to oil prices by buying and rolling futures. It is tempting to think that the closer one can hold futures to expiration, the more closely the future price should mimic the price of the physical barrel. This argument is somewhat dubious though, because the futures market is where the price for the physical barrel is determined, and not the other way around.

Individual investors are rarely allowed by their clearing brokers to hold futures near expiration due to extremely high volatility and the risk of being squeezed. Professional traders, however, such as banks, can hold futures for hedging purposes effectively until the final buzzer. This also allows banks to offer clients OTC replicas of futures that can be held closer to expiration. While most financial products avoid taking this risk by rolling futures a few days or weeks prior to expiration, a handful of dealers were still holding long futures on behalf of their OTC investors on April 20, 2020, one day prior to the contract expiration.

One such dealer still holding a large quantity of long futures during that time was the Bank of China. With oil prices already in a freefall since the start of the Covid-19 pandemic, the bank's investment product named *YuanYouBao*, which was advertised as "*oil cheaper than water*", attracted enormous interest from domestic retail investors. The product was marketed as a safe non-leveraged investment with the entire notional value of oil prepaid upfront, in contrast to leveraged futures contracts, which only require the initial margin to be posted. The structure of the contract, of course, assumed that the notional amount remains positive. From the investor's perspective, it was then reasonable to think that the price of OTC oil-linked product is bounded by zero, like the price of a common stock.

When the bank sells an OTC replica of futures and buys actual futures as a hedge, the bank must either liquidate or roll futures at the futures settlement price on a given day. To make it easier for dealers, the so-called *Trading at Settlement (TAS)* contract has been created by futures exchanges. In this contract, buyers and sellers agree to exchange futures at a price which will only be determined later, after the trading session ends. If the dealer wants a particularly speedy execution, then a TAS contract can be offered at a slight discount or premium to the settlement price to incentivize counterparties to take the other side. In a normal market, an extra one cent per barrel of premium or discount is usually sufficient to attract enough counterparties and ensure prompt execution of a TAS order.

However, given the tremendous uncertainty about the direction of the Covid-19 pandemic, the behavior of the May 2020 WTI contract was anything but normal. Extra demand from oil bargain hunters, further stimulated by OTC dealers, led to a large quantity of long futures held on behalf of individual investors. These futures had to be liquidated or rolled by the dealer at the TAS price on April 20, 2020. The normal TAS liquidity quickly dried up, forcing the dealers to offer progressively larger discounts to sell futures relative to a still undetermined TAS price. Even when the discount reached its maximum allowed \$0.10/bbl, the order to sell futures at the

TAS price remained unfilled. This imbalance telegraphed to the entire market that someone in the market who must sell futures at the closing price is struggling to do it. It meant that the only alternative for the holder of long futures was to sell them instead at the regular market prices as close as possible to the end of the trading session in the hope of matching the settlement price. With dealers suddenly unable to trade TAS, panic kicked in and unfilled futures were sold in a rapid-fire fashion into the vacuum within thirty minutes before the market closed.

Shortly after the market closed at the historical price of negative \$37.63/bbl, the price bounced back to where it was trading a few hours prior, as storage buyers rushed to capitalize on this unique opportunity. The balancing loop of the physical storage came to the rescue and contained the self-reinforcing loop of financial flows that went out of control. Unfortunately, like in many other complex systems, the balancing action came with a delay. The delayed reaction of storage traders was largely driven by operational controls within professional trading shops, such as the need to secure additional risk limits given unprecedented market volatility. In the end, buying prompt futures at nearly \$60/bbl discount to the next-maturity futures was a real gift to anyone with membership in the Cushing storage club. At such a price, with a bit of creativity some could have found a way to store oil in their backyards.

The financial loop of naïve investors had created and effectively subsidized this unique opportunity for storage traders. Contractually, the losses from sales at negative prices were supposed to be passed by the dealer to investors. However, investors revolted, blaming the faulty product design and pointing to an implicit zero price boundary in the contract. After some government intervention, fines were imposed on the dealer, who ended up paying most of the losses. Dealers learned a painful lesson, that the market risk is not the only type of risk that must be managed in the complex world of OTC oil derivatives.<sup>9</sup>

The nature of interactions between physical and financial markets became more visible during this memorable episode of negative oil prices. However, such interactions were not specific to that single day. They happen every single day in the derivatives market for virtual barrels. Similarly to how physical storage with its boundaries drives the dynamics of one complex system, the intricate web of interactions among participants in financial markets drives another. The latter is the subject of our next chapter, the *theory of hedging pressure*.

---

## References

- Bouchouev, I. (2020, April 30). Negative oil prices put spotlight on investors, *Risk.net*.  
Bouchouev, I. (2021). A stylized model of the oil squeeze, *SSRN*.  
Brennan, M. J. (1958). The supply of storage. *The American Economic Review*, 48(1), 50–72.

---

<sup>9</sup>For additional discussions of the episode of negative prices, see Interim Staff Report (2020), Bouchouev (2020), Fernandez-Perez et al. (2021), and Ma (2022).

- Brennan, M. J. (1991). The price of convenience and the valuation of commodity contingent claims. In D. Lund & B. Oksendal (Eds.), *Stochastic models and option values*. North Holland.
- Brennan, M. J., & Schwartz, E. S. (1985). Evaluating natural resource investments. *Journal of Business*, 58(2), 135–157.
- Carmona, R., & Ludkovski, M. (2004). Spot convenience yield models for the energy markets. *Contemporary Mathematics*, 351, 65–79.
- Casassus, J., & Collin-Dufresne, P. (2005). Stochastic convenience yield implied from commodity futures and interest rates. *The Journal of Finance*, 60(5), 2283–2331.
- Clewlow, L., & Strickland, C. (2000). *Energy derivatives: Pricing and risk management*. Lacima Publications.
- Deaton, A., & Laroque, G. (1992). On the behavior of commodity prices. *The Review of Economic Studies*, 59(1), 1–23.
- Dempster, M. A. H., Medova, E., & Tang, K. (2012). Determinants of oil futures prices and convenience yields. *Quantitative Finance*, 12(12), 1795–1809.
- Dvir, E., & Rogoff, K. (2009). Three epochs of oil, NBER Working Paper, 14927.
- Eydeland, A., & Wolyntiec, K. (2003). *Energy and power risk management: New developments in modeling, pricing, and hedging*. Wiley.
- Fernandez-Perez, A., Fuertes, A.-M., & Miffre, J. (2021, Summer). On the negative pricing of WTI crude oil futures. *Global Commodities Applied Research Digest*, 6(1), 36–43.
- Gibson, R., & Schwartz, E. S. (1990). Stochastic convenience yield and the pricing of oil contingent claims. *The Journal of Finance*, 45(3), 959–976.
- Gustafson, R. L. (1958). Carryover levels for grains, *U.S. Department of Agriculture, Technical Bulletin*, 1178.
- Hamilton, J. D. (2009). Understanding crude oil prices. *The Energy Journal*, 30(2), 179–206.
- Interim Staff Report. (2020). *Trading in NYMEX WTI crude oil futures contract leading up to, on, and around April 20, 2020*, Commodity Futures Trading Commission, November 23.
- Kilian, L. (2020). Understanding the estimation of oil demand and oil supply elasticities, Federal Reserve Bank of Dallas Working Paper, 2027.
- Ma, L. (2022). Negative WTI price: What really happened and what can we learn? *The Journal of Derivatives*, 29(3), 9–29.
- Miltersen, K. R. (2003). Commodity price modelling that matches current observables: A new approach. *Quantitative Finance*, 3(1), 51–58.
- Pirrong, C. (2012). *Commodity price dynamics: A structural approach*. Cambridge University Press.
- Routledge, B. R., Seppi, D. J., & Pratt, C. S. (2000). Equilibrium forward curves for commodities. *The Journal of Finance*, 55(3), 1297–1338.
- Schwartz, E. S. (1997). The stochastic behavior of commodity prices: Implications for valuation and hedging. *The Journal of Finance*, 52(3), 923–973.
- Williams, J. C., & Wright, B. D. (1991). *Storage and commodity markets*. Cambridge University Press.
- Working, H. (1948). Theory of the inverse carrying charge in futures markets. *Journal of Farm Economics*, 30(1), 1–28.
- Working, H. (1949). The theory of price of storage. *The American Economic Review*, 39(6), 1254–1262.



# Financialization and the Theory of Hedging Pressure

4

- The Keynesian theory of normal backwardation argues for structural disequilibrium in the futures market caused by producer demand for hedging. As a result, futures must trade at a discount to the expected spot price to incentivize speculators to provide the service of risk absorption.
- The theory of hedging pressure is derived within a conventional mean-variance framework of rational trading behavior. The framework quantifies the hypothesis of normal backwardation and relates it to the risk premium in the oil market.
- The thesis of normal backwardation found strong empirical support in the early days of oil trading. The existence of the structural oil risk premium driven by the roll yield attracted financial investors and spurred the financialization of the oil market.
- Historically, jumps in oil prices preceded most of US economic recessions. Oil also tends to appreciate during periods of high inflation. To hedge inflation risks in highly leveraged risk parity portfolios, large allocations are made to oil futures.
- As investor demand for oil futures overpowered producers' hedging needs, the oil market shifted to a regime of normal contango. The equilibrium hedging framework is generalized to include inflation hedgers, which allows oil risk premium to be negative.

---

## 4.1 The Theory of Normal Backwardation

The question whether commodity futures markets are driven by fundamentals or by financial flows is as old as the market itself. Any attempt to answer this question unambiguously, however, is a fruitless endeavor, as fundamentals and flows are closely intertwined. The ideological underpinnings of fundamental analysts lie in the theory of storage, which seeks an equilibrium between the price of the commodity, its supply, demand, and inventories. Financial traders, on the other hand, look for another equilibrium that arises directly in the futures market. This financial equilibrium is driven by the aggregate demand for hedging services among market

participants and the supply of such services by professional speculators. While the price for a barrel of spot oil is driven by futures, an efficient futures market cannot exist without a deep connection to the physical market. The financial and physical markets are largely inseparable, with the crucial link between them provided by carry traders.

The primary economic function of the futures market is to provide a mechanism for transferring unwanted risks to those better equipped to manage them. The price determined in the futures market should reflect the supply and demand for hedging services, but it can also be significantly impacted by speculative flows. In the oil market, it is rather difficult to draw the line between hedging and speculation. Many hedging programs incorporate subjective views of company executives, at least when it comes to the timing of a trade and the choice of the hedging instrument. Vice versa, some trades that may appear to be speculative are often implemented within the broader mandate to optimize a portfolio of physical or financial assets. The study of the interaction between hedgers and speculators in the futures market is known as *the theory of hedging pressure*.

Much like the theory of storage, the theory of hedging pressure has developed its roots in agricultural markets. It starts with the works of John Maynard Keynes in the 1920s.<sup>1</sup> As an avid speculator in the commodity markets, Keynes was clearly influenced by some painful lessons from his own trading experience. This led him to recognize the importance of psychological and behavioral factors in trading, especially in the market for commodity futures, where hedgers and speculators play a dominant role. These financial traders participate in the market for reasons unrelated to prevailing fundamentals, but they can still have a significant impact on price. Their behavior reflects the willingness of some traders to pay for the services of risk absorption and the remuneration of other traders for providing such services. If supply and demand for the service of risk transfer are unbalanced, then additional price incentives are needed to lure more speculators to the market. The speculators are then compensated by trading futures contracts at a discount or premium to their fair value, which is determined by the supply of and demand for the physical commodity.

To transfer unwanted price risks, commodity producers and holders of inventories sell futures, while consumers of commodities with opposite price risks are expected to buy futures. In an ideal world, consumers should be as eager as producers to get rid of the price risk, in which case the two sides should be able to exchange futures in the market at some fairly negotiated price. In practice, however, the propensity to hedge between producers and consumers of commodities tends to

---

<sup>1</sup> Among many studies of commodity markets in the 1920s, the newspaper article published by Keynes (1923) is credited with a particularly significant contribution. In this article, the author highlighted that the value of agricultural inventories after the harvest is so large relative to financial resources of producers that they cannot themselves bear the risk of inventory value to drop. Thus, the imbalance requires the service of professional speculators. Keynes was careful to distinguish between agricultural and extraction commodities, such as oil, for which the demand for temporary credit is lower due to the ratable nature of production.

be asymmetric. The price risk typically has a more substantial impact on the economics of a commodity producer, whose exposure is less diversified than it is for a typical consumer.

Another important difference between producer and consumer hedging interest lies in the timing of their business decisions. A renowned economist and Nobel Prize winner, John Hicks, wrote that

... supplies in the near future are largely governed by decisions taken in the past, so that if these planned supplies can be covered by forward sales, risk is reduced. But . . . with planned purchases . . . technical conditions give the entrepreneur a much freer hand about the acquisition of inputs . . . than about the completion of outputs. Thus, while there is likely to be some desire to hedge planned purchases, it tends to be less insistent than the desire to hedge planned sales. If forward markets consisted entirely of hedgers, there would always be a tendency for a relative weakness on the demand side; a smaller proportion of planned purchases than of planned sales would be covered by forward contracts.<sup>2</sup>

Given the structural imbalance between producer and consumer demand for hedging, producers have little choice but to offer futures at a discount to the fair value in order to attract professional speculators to take the other side of their hedging needs.

Let us assume that the futures price is deemed to be fair if it is determined by market collective expectations of the spot price at the time of the futures delivery. In other words, if the market is fair, then one cannot consistently make or lose money by trading futures. The Keynes-Hicks hypothesis argues that the futures price should not be fair, as there are more natural sellers of commodity futures than buyers. Therefore, in the natural equilibrium state of the market, the futures price must fall below the expected spot price by an amount equal to *normal backwardation*,<sup>3</sup> defined as:

$$RP = E_t(S(T)) - F(t, T) > 0$$

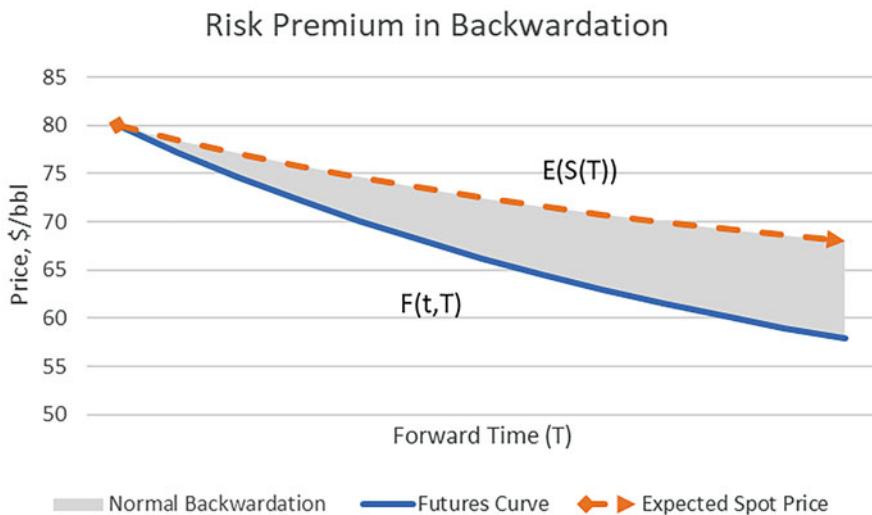
In the modern jargon of financial markets, the Keynesian term of normal backwardation is more commonly known as *risk premium (RP)*. The risk premium provides an incentive to speculators to participate in an otherwise unbalanced market for hedging services. Anyone willing to withstand short-term market fluctuations should be able to buy discounted futures and get rewarded for offering the service of risk absorption. The reward comes in the form of a positive expected return, provided that the speculator has enough financial power to stay in the game.

The concept of normal backwardation should be distinguished from the conventional definition of market backwardation. The latter does not involve expectations; it only refers to spot and futures prices observed at the same time. To see how the two concepts are related, we recall the decomposition of futures profitability (2.11), and split the risk premium in a similar way, as follows

---

<sup>2</sup>Hicks (1939), chapter 10.

<sup>3</sup>The term normal backwardation was introduced in Keynes (1930), volume II, chapter 29.



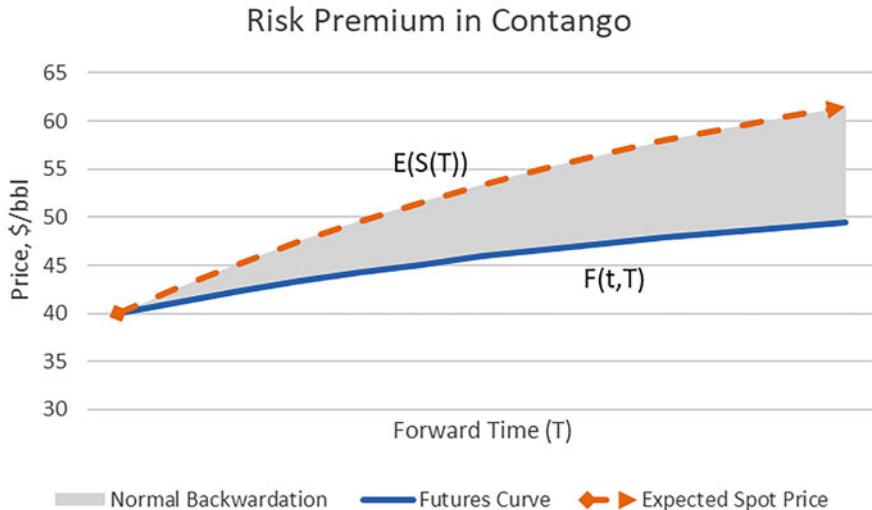
**Fig. 4.1** Normal backwardation, or the risk premium, in a backwardated market

$$RP = (E_t(S(T)) - S(t)) + (S(t) - F(t, T)) \quad (4.1)$$

The first term reflects the expected change in the spot price, which, of course, is not known today. The second term, on the other hand, is directly observable from the current shape of the futures curve. This term is positive when the market is in backwardation, and it is negative when the market is in contango. As explained in Chap. 2, it is the same term that drives the oil own rate of interest, the convenience yield, net of storage costs, and the roll yield.

Keynes argued that normal backwardation, which is equivalent to a positive risk premium, exists regardless of whether the market is in the state of backwardation or contango. His thesis is illustrated in Figs. 4.1 and 4.2. The blue line is the observable futures curve. The dashed orange line represents the expected forward trajectory of the spot price moving from left to right, i.e., from  $t$  to  $T$ . The difference between the two lines is the measure of normal backwardation, or the expected profit to the buyer of futures.

In both examples, the risk premium is positive, and the two terms in the decomposition (4.1) have opposite signs. In the first scenario of market backwardation, shown in Fig. 4.1, a long futures position is expected to make money because the magnitude of the market backwardation exceeds the expected depreciation of the spot price. While the spot price is expected to fall, it falls only by a fraction of what is priced in the futures curve. In the second scenario of market contango, shown in Fig. 4.2, the rise in the spot price is expected to exceed the amount of contango priced in the futures curve. Here, the market contango only partially offsets a more substantial appreciation of the spot price. In both scenarios, the futures price is suppressed by producer demand for hedging services, which results in an expected profit to the buyer.



**Fig. 4.2** Normal backwardation, or the risk premium, in a contango market

The decomposition of the risk premium in (4.1) corresponds to two potential sources of revenue for the commodity speculator. The first one is highly uncertain as it comes from speculating on the volatile spot price of a commodity. The second term does not require any forecasting. It simply reflects the discount or premium, which is measured by the current shape of the futures curve. By and large, the two terms correspond to two types of speculators, eloquently labeled by Keynes as *prophets* and *risk-bearers*. Being a prophet requires a special talent for predicting future spot prices, something that Keynes himself was highly skeptical about, as

... it presumes that the speculator is better informed on the average than the producers and the consumers themselves, which, speaking generally, is rather a dubious proposition.<sup>4</sup>

In contrast, the second type of speculator, who does not pretend to possess any unique trading insights, is not involved in the game of price forecasting, as

... he is not so much a prophet (though it may be a belief in his own gifts of prophecy that tempts him into the business), as a risk-bearer.<sup>5</sup>

The business model of the risk bearer is very different from the business of price forecasting. Risk bearing resembles the business of a casino dealer. The primary requirement for the risk bearer is an ability to provide capital and withstand short-term losses resulting from random fluctuations in spot prices. Since the spot price is

<sup>4</sup>Keynes (1923).

<sup>5</sup>Keynes (1923).

likely to mean-revert, as implied by the theory of storage, cumulative spot price changes in the first term of (4.1) are expected to net out. In the meantime, the contribution of the second term steadily accumulates as long as the market remains backwardated. This is the reason why casinos tend to win; they stay in the game and repeat the same bet with a tiny statistical edge over and over. For them, wins and losses of individual gamblers, or aspiring prophets, only represent noise that eventually cancels out. By the law of large numbers, the casino's statistical edge turns into profits over time. For risk bearers in commodity markets, this casino-like edge comes from the backwardated shape of the futures curve.

We next quantify the Keynesian hypothesis of normal backwardation and expected profits for risk absorbers with a simple analytical framework based on the rational behavior of market participants.

## 4.2 The Hedging Equilibrium

Let us consider a simplified model of the futures market which is made up of three participants: a producer, a consumer, and a speculator. We assume that they follow a conventional one-period mean-variance decision-making framework, where participants are maximizing their expected wealth,  $E(\tilde{W})$ , while being penalized for risks measured by the variance of wealth,  $\sigma^2(\tilde{W})$ . In other words, all participants maximize their quadratic utility function of the form

$$\max \left\{ E(\tilde{W}) - \frac{\alpha}{2} \sigma^2(\tilde{W}) \right\} \quad (4.2)$$

where  $\alpha$  represents the risk aversion coefficient.<sup>6</sup>

We denote wealth functions  $\tilde{W}_p$ ,  $\tilde{W}_c$ ,  $\tilde{W}_s$  and risk aversion coefficients  $\alpha_p$ ,  $\alpha_c$ ,  $\alpha_s$  with corresponding subscripts for producer, consumer, and speculator, and use tilde to represent random variables whose values are unknown at the beginning of the period.

Assume that the producer is endowed with  $Q_p$  barrels of oil and, thus, is exposed to potential losses if the spot oil price  $\tilde{S}$  falls. The producer also participates in the futures market and trades  $N_p$  units of futures that expire at the end of the period. Since the producer is long physical barrels of oil, one should expect the producer to be net short futures, so that  $N_p < 0$ . We do not differentiate between the producer of oil and the owner of oil inventory, as both have similar exposure to oil prices. The main practical difference in their market participation is in the duration of their typical hedges. Producers usually seek a longer-term price hedge against the value of their expected production and reserves, while inventory hedgers trade only a few

---

<sup>6</sup>This approach follows the mean-variance hedging framework originally developed by Stoll (1979) and Hirshleifer (1988).

months forward to protect the value of oil that has already been taken out of the ground. In our simplified one-period framework, their exposure is largely identical.<sup>7</sup>

The producer's wealth is determined by the dollar value of the oil endowment, plus the profit or loss on  $N_p$  futures traded at price  $F$ , less some fixed operational cost  $U_p$ :

$$\tilde{W}_p = Q_p \tilde{S} + N_p (\tilde{S} - F) - U_p$$

The only uncertain variable here is the spot oil price  $\tilde{S}$ . The expected value of the wealth function is then given by

$$E(\tilde{W}_p) = Q_p E(\tilde{S}) + N_p (E(\tilde{S}) - F) - U_p$$

and the variance of wealth is proportional to the variance of the oil price  $\sigma^2(\tilde{S})$

$$\sigma^2(\tilde{W}_p) = (Q_p + N_p)^2 \sigma^2(\tilde{S})$$

The producer wealth maximization problem is to determine the optimal position in the futures market  $N_p$  that maximizes the quadratic utility (4.2):

$$\max \left\{ Q_p E(\tilde{S}) + N_p (E(\tilde{S}) - F) - U_p - \frac{\alpha_p}{2} (Q_p + N_p)^2 \sigma^2(\tilde{S}) \right\}$$

We apply the standard first-order optimality condition, differentiate this expression with respect to  $N_p$ , and equate the derivative to zero

$$E(\tilde{S}) - F - \alpha_p \sigma^2(\tilde{S}) (Q_p + N_p) = 0$$

Solving this equation for  $N_p$ , we obtain that the producer optimal position in the futures market is given by

$$N_p = -Q_p + \frac{E(\tilde{S}) - F}{\alpha_p \sigma^2(\tilde{S})} \quad (4.3)$$

As anticipated, the producer is likely to be short futures to hedge the value of the oil endowment against the risk of falling prices. If the producer is highly risk averse, so that  $\alpha_p \rightarrow \infty$ , then the optimal hedging decision is to cover the entire value of  $Q_p$ . However, a less risk averse producer may reduce the size of the short futures position if there is a substantial positive risk premium in the market and the hedge is expected to lose money. This adjustment depends on the market volatility and the individual producer's risk tolerance. The optimal hedging decision, therefore, includes a speculative component.

---

<sup>7</sup>The theory of hedging pressure has been combined with the canonical theory of storage by Gorton et al. (2012), Acharya et al. (2013), and Baker (2021).

We repeat the same argument for the consumer, but while doing so, we also illustrate how to handle the basis risk in hedging. Since oil can only be consumed in its processed form, the end-user is usually exposed to the price of the refined consumer product  $\tilde{P}$ , and not to the price of crude oil  $\tilde{S}$ . We assume that  $\tilde{P}$  is correlated to the price of oil  $\tilde{S}$  with the correlation coefficient  $\rho_{p,s} = \rho(\tilde{P}, \tilde{S})$ . The consumer, whose risk aversion coefficient is  $\alpha_c$ , needs to purchase  $Q_c$  units of the refined product at price  $\tilde{P}$ , which is used as an input into a manufacturing process that generates fixed revenue  $U_c$ . The consumer trades  $N_c$  futures contracts which are expected to be long positions, i.e.,  $N_c > 0$ .

The consumer net wealth is then defined as

$$\tilde{W}_c = -Q_c\tilde{P} + (\tilde{S} - F)N_c + U_c$$

Here, we have two random variables,  $\tilde{S}$  and  $\tilde{P}$ . The variance of the sum of two random variables is equal to the sum of their variances, plus the covariance between them.

To maximize the expected utility of wealth, we substitute  $E(\tilde{W}_c)$  and  $\sigma^2(\tilde{W}_c)$  into (4.2)

$$\max \left\{ -Q_c E(\tilde{P}) + N_c (E(\tilde{S}) - F) + U_c - \frac{\alpha_c}{2} (Q_c^2 \sigma^2(\tilde{P}) + N_c^2 \sigma^2(\tilde{S}) - 2 Q_c N_c \rho_{p,s} \sigma(\tilde{P}) \sigma(\tilde{S})) \right\}$$

To find the optimal futures hedge, we differentiate with respect to  $N_c$  and equate the derivative to zero:

$$E(\tilde{S}) - F - \alpha_c (N_c \sigma^2(\tilde{S}) - Q_c \rho_{p,s} \sigma(\tilde{P}) \sigma(\tilde{S})) = 0$$

Solving for  $N_c$ , we obtain that

$$N_c = \beta_c Q_c + \frac{E(\tilde{S}) - F}{\alpha_c \sigma^2(\tilde{S})} \quad (4.4)$$

Here,

$$\beta_c = \rho_{p,s} \frac{\sigma(\tilde{P})}{\sigma(\tilde{S})}$$

is the standard beta sensitivity of the asset with respect to the benchmark index, commonly used in the equity market. The consumer optimal hedge is based on the beta-weighted exposure of the refined product to crude oil, further adjusted for a speculative view. The beta hedge ratio increases with higher correlation between the refined product and crude oil, while it decreases when crude oil volatility rises. If necessary, a similar beta adjustment can be applied to the producer hedging problem if the type of oil being produced is different from the benchmark used for hedging. In practice, however, the correlation among various grades of crude oil is much higher

than the correlation between crude oil and refined products, so, for simplicity, we let the producer beta to the benchmark be equal to one.

Finally, the same logic is applied to the speculator with no endowment, liability, or any other hedging mandate. The only motive of the speculator is to make money by trading futures and to capture the risk premium caused by the hedging imbalance. Repeating the previous steps, we obtain that the optimal position of the speculator is simply dictated by the market risk premium, adjusted for volatility and the speculator risk aversion

$$N_s = \frac{E(\tilde{S}) - F}{\alpha_s \sigma^2(\tilde{S})} \quad (4.5)$$

We can now construct the financial equilibrium in the futures market. The futures market is a zero-sum game, so if the producer, the consumer, and the speculator are the only three participants in the market, then for futures to clear, their net position must be zero

$$N_p + N_c + N_s = 0$$

Adding up Eqs. (4.3), (4.4) and (4.5) for optimal futures positions held by producer, consumer and speculator, we obtain that

$$-Q_p + \beta_c Q_c + \left( \frac{1}{\alpha_p} + \frac{1}{\alpha_c} + \frac{1}{\alpha_s} \right) \frac{(E(\tilde{S}) - F)}{\sigma^2(\tilde{S})} = 0$$

This equation can be solved for the risk premium, as follows

$$E(\tilde{S}) - F = \alpha \sigma^2(\tilde{S}) (Q_p - \beta_c Q_c) \quad (4.6)$$

where the coefficient  $\alpha$  is given by

$$\alpha = \frac{1}{\frac{1}{\alpha_p} + \frac{1}{\alpha_c} + \frac{1}{\alpha_s}}$$

The Eq. (4.6) explicitly relates the Keynesian risk premium to the aggregate hedging imbalance between producers and consumers. Given that the producer has a much more concentrated price exposure than the consumer, the producer propensity to hedge is likely to be higher, and  $Q_p \gg Q_c$ . This implies that in this simple market model, the resulting oil risk premium in the Eq. (4.6) is indeed expected to be positive, as is suggested by the theory of normal backwardation.

The size of the risk premium depends on the volatility of the market and the risk aversion of all market participants. The risk tolerance of speculators cannot change the sign of the risk premium, but the speculators' lower risk aversion can dilute the magnitude of the imbalance. The risk in this hedging equilibrium model is shared

among all market participants. The parameter  $\alpha$  can be thought of as the market collective risk aversion coefficient. It is proportional to the harmonic mean of the individual risk aversions. Remarkably, the idea of using a harmonic mean in risk sharing goes back to Ancient Greece and Aristotle, who considered the price of voluntarily exchanged goods to be fair if it is calculated as the harmonic mean of the buyer's bid and the seller's offer.<sup>8</sup>

To validate the existence of the structural risk premium in the oil market, we now test the hypothesis of normal backwardation empirically.

---

### 4.3 The Genesis of Oil Financialization

The search for a Keynesian free lunch for risk absorbers in commodity futures got off to a rough start in agricultural markets. Numerous attempts have been made to trace any long-term structural bias, but little empirical evidence was found to support the hypothesis of normal backwardation. In fact, the term structure of many agricultural futures was more often in contango than it was in backwardation. It turns out that the treasure hunt was going on in the wrong place. Only after the introduction of oil futures in the 1980s did the thesis of normal backwardation get its second wind.

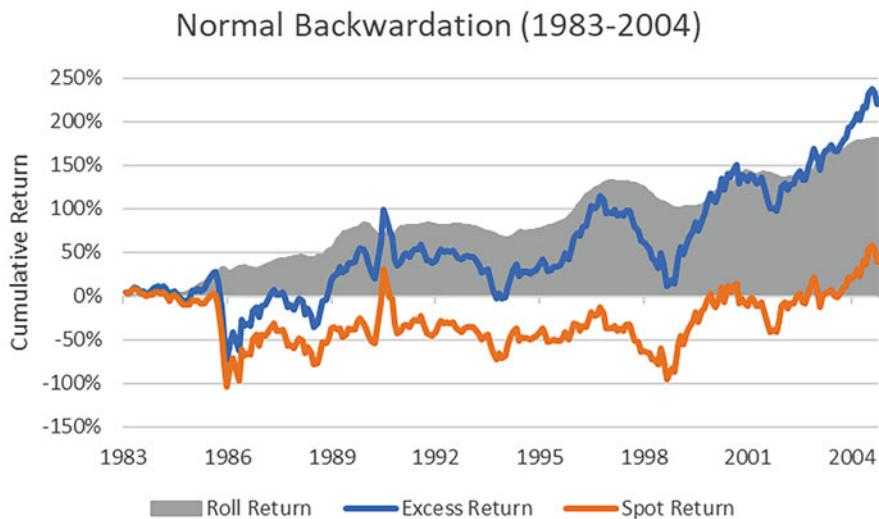
The structure of petroleum markets appears to fit the Keynes-Hicks consumer-producer hedging paradigm quite well. Investments in oil exploration and production are extremely capital intensive and are accompanied by large price uncertainty with a multi-year payback time. These decisions are typically irreversible. Once an investment decision is made, hedging the value of the expected output becomes the only practical way to mitigate the impact of price fluctuations. Things work differently for many oil consumers. The primary usage of petroleum products is in transportation, where the demand mostly comes from individuals buying gasoline and diesel to drive their cars. While collectively, we all share a price risk of a similar magnitude to oil producers, this risk is spread out across millions of us. The fuel cost represents a relatively small portion of our routine expenses, and very few of us would seriously contemplate the idea of hedging personal gasoline consumption.<sup>9</sup> The much higher concentration of the energy price risk makes hedging more important for producers than for consumers.

One notable exception of an active consumer hedger is the airline industry. For an airline, the purchases of jet fuel represent the second largest cost after the cost of labor. Since jet fuel, which is tied to the price of oil, is much more volatile than salaries, higher oil prices could easily dwarf the slim operational profit margins of the industry. While to a certain degree an airline might be able to pass on some fuel costs to travelers by raising ticket prices, such pass-throughs are often limited by

---

<sup>8</sup>See Backhouse (2002).

<sup>9</sup>Several attempts have been made to provide individual drivers with hedging instruments, typically by selling them prepaid gasoline cards to be used at retail gas stations, but such attempts never gained the economy of scale.



**Fig. 4.3** Cumulative performance of fully funded long WTI futures during the regime of normal backwardation

competitive pressures. Overall, however, airline hedging is substantially smaller than producer hedging. The first part of the Keynes-Hicks hypothesis, which argues for the structural imbalance between producers' and consumers' propensity to hedge, clearly applies to the oil market. The second part, which claims the existence of a structural risk premium that rewards the buyers of futures, is, of course, a more interesting question.

In contrast to agricultural markets dominated by contango, in the early days of oil trading the futures curve exhibited a more persistent backwardation. To test the Keynesian hypothesis of a free lunch for futures buyers, we look at the passive long strategy that buys prompt WTI futures contract and rolls it at the end of each calendar month to the second nearby contract. Figure 4.3 presents the performance of this strategy since the inception of the WTI futures contract until 2004. It shows the picture as it was seen by prospective oil investors at that time.

Driven by the preponderance of backwardation in the oil market, the strategy of buying futures at a discount relative to spot prices was indeed spectacularly profitable. During this period, the passive oil investor would have realized over 10% annualized logarithmic return on an unleveraged, or fully funded, investment in oil futures.<sup>10</sup> Such returns were comparable, if not superior, to long-term investment

<sup>10</sup>The term fully funded means that the entire notional value of the futures contract is held as collateral. We use log-returns which are additive for the reasons of analytical tractability, as explained in Chap. 2. While it is not recommended to average simple returns, the annualized arithmetic average of monthly returns over this period would have been even higher at 15.6%. These returns do not include any additional return on collateral.

returns in the stock market. Moreover, since in practice only a small portion of the notional value of the futures contract is required to be paid as the initial margin, the returns on the actual cash deployed in highly leveraged oil futures contracts were five to ten times higher.<sup>11</sup>

The source of such impressive returns turned out to be even more striking. Figure 4.3 also shows the decomposition of the excess return into the spot return and the roll return based on the identity (2.12). As one can see, nearly entire profitability of the strategy came from the roll return, as cumulatively WTI spot price did not change by much over this period. While the spot price fluctuated, which is what one should expect from the supply and demand induced price mean-reversion, the positive roll return steadily accumulated as the oil market was predominantly backwardated.

The superb performance of an investment in oil futures has finally vindicated the Keynesian theory of normal backwardation. The profits were driven not by a highly questionable ability to forecast the direction of the spot price, but rather by providing liquidity and taking the other side of the dominant producer hedging at advantaged prices. The strategy of holding oil futures looked very compelling. It required no unique skills, just enough money to withstand occasional losses. Its appeal for financial investors has become so powerful that it marked the beginning of a new era in the history of oil trading, the era of oil *financialization*.

The opportunity for a free lunch promised by the theory of normal backwardation became evident. The lunch was not entirely free though. It was a compensation for committing large amount of capital for a long period of time, which was needed to stay in this game of risk absorption. To bring such capital to the market for oil futures, the industry had to overcome significant challenges. A typical oil speculator at that time neither had much capital, nor any desire to lock it up for more than a few months ahead. A better candidate for the role of a risk bearer was a longer-term investor with large financial coffers, like a pension fund or an insurance company.

Bringing pension funds to trade oil was not straightforward since many institutional investors were explicitly prohibited from trading futures. Moreover, oil, along with many other commodity futures, suffered from an image problem. For a long time, trading commodity futures has been perceived by the public as a speculative, borderline gambling enterprise, which made it difficult for large strictly regulated asset managers to participate in this market. Traditional investments, like stocks and bonds, serve a clear investment purpose; they provide capital that allows businesses to operate and grow their enterprises. In contrast, futures were viewed as highly leveraged speculative bets that do not even require as much capital to be put upfront. Oil futures did not appear to be suitable for deep-pocketed conservative financial investors who were looking for steady long-term returns. The breakthrough for

---

<sup>11</sup> A typical initial cash margin requirement for energy futures is 10–20% of the contract's notional value. Therefore, the returns on cash required to hold futures are 5–10 times higher than returns on a fully funded position.

turning commodity futures into a new asset class came from the clever packaging of futures contracts into investable indices by the banks.

Commodity indices were designed to function as a bridge that connects the capital of long-term financial investors to the Keynesian opportunity in highly speculative commodity futures markets. The idea was for the banks to step in between and take care of futures trading themselves. The resulting economics of the futures trade can then be transferred to investors via an equity-like product, whose returns are determined by profit and loss on futures. Since only a small portion of the notional value of the futures contract is required as collateral, the remaining cash can be invested in safe securities, such as US Treasury bills. Not only does this cash provide an additional return in the form of interest, but it also eliminates unwanted leverage, as the entire notional value of commodity futures is effectively prepaid. Such fully collateralized commodity indices look a lot closer to traditional financial and capital investments like stocks and bonds.

To address concerns about the role of investor capital in commodity investment, the Keynesian thesis has once again been recruited to help. While in equity and bond markets capital provides explicit funding of business operations, commodity index providers argued that in their products it does so implicitly. The capital invested in commodity indices can also be perceived as productive if it lets producers free up their own capital. Otherwise, producers' own capital would be tied up unproductively and held as a buffer against price fluctuations.

Such an interpretation would allow producers to view oil hedges as an integral part of their new capital structure along with equities and bonds. The ability to hedge reduces volatility of producer earnings and lowers borrowing costs, thus allowing more capital to be redeployed towards expanding the primary business of producing oil. Producer hedging and the credit market become intertwined. While hedging was explicitly mandated by credit departments of lending banks, the deals were also often executed by the derivatives desks of the same banks, allowing them to capture profits on both sides. In contrast to many tightly regulated financial markets, there were no restrictions on such an intimate arrangement between oil and credit markets, which is one of the reason why oil OTC market grew to become so large.

From the investor's perspective, the capital allocated to commodity indices is rewarded for risk bearing, much like the capital invested in insurance products. The reward is expected to accrue over time from the passive accumulation of normal backwardation, which does not require any price forecasting. The opportunity to generate long-term casino-like returns on capital via investments in commodity indices without any need to develop unique commodity expertise quickly became a home run with pension funds and other long-term financial investors.

The commodity index business was designed for a diversified basket of commodities but crude oil and other petroleum products were its crown jewels. The Goldman Sachs Commodity Index (GSCI), the largest and the most successful commodity index at that time, epitomizes it well. In this index, seventy percent of its assets was allocated to the energy sector. GSCI weights for individual commodities are calculated based on the total value of commodity annual production, where petroleum dominates by a wide margin. It leaves little doubt that such a choice for

the index construction was influenced by the superb historical performance of energy futures, which happened to be more often in backwardation than in contango. Unlike many other commodities, oil had a high convenience yield that dominated the cost of storage, resulting in a positive own rate of interest. This intrinsic interest, or the roll yield, was what financial investors were after.<sup>12</sup>

While energy futures were primary contributors to the index returns, other commodities provided valuable diversification. Individual commodities tend to have low correlations with each other given their idiosyncratic risks, such as, for example, their sensitivity to weather. Commodity prices have vastly different seasonal production and consumption profiles. Even within the energy sector, gasoline has higher upside risks during the summer driving season, while heating oil and natural gas are more susceptible to price spikes during the winter. An investment in a broader commodity index isolates the contribution from the roll, and it dilutes the impact of random fluctuations in spot prices. While backwardated commodities, such as oil, are expected to drive the index return, other commodities help to dampen the volatility of unpredictable spot returns. In addition, marketing the diversified index for the commodities, as a new asset class, is much easier than marketing oil futures, which would have been seen as too speculative.

One other substantial benefit that results from an investment in a diversified commodity basket is the so-called *rebalancing effect*. Some commodity indices target allocations in dollar terms that must be periodically rebalanced as commodity prices change. Since commodity spot prices tend to mean-revert over time, the rebalanced index would buy more futures of relatively depreciated commodities and sell the appreciated ones to keep their dollar notional contract value the same. While GSCI does not rebalance frequently, with rebalancing impact muted by its high concentration of energy futures, many other sector- diversified commodity indices are able to capture additional returns from this rebalancing effect by buying lows and selling highs.

The Keynesian normal backwardation revived by the energy market and adopted by long-term passive investors was only the beginning of oil financialization. The pace of financialization accelerated as another even more powerful financial force, fueled by leverage, was about to enter the oil market.

---

<sup>12</sup>The idea of investments in diversified commodity indices has been covered extensively in the literature. For earlier studies of this subject that predate the introduction of energy futures, see Greer (1978), Bodie and Rosansky (1980), and Fama and French (1987). The addition of petroleum futures markedly improved the performance and the overall attractiveness of commodity indices for investors, as documented in the influential work of Gorton and Rouwenhorst (2006), and Erb and Harvey (2006). See also Till and Eagleeye (2007), Ashton and Greer (2008), Tang and Xiong (2012), Fattouh and Mahadeva (2014), Büyüksahin and Robe (2014), Hamilton and Wu (2014), Cheng et al. (2015), and references therein.

## 4.4 Inflation Hedging and Risk Parity

The introduction of commodity indices turned commodities into a new asset class. The precise definition of what constitutes an asset class is somewhat murky and highly debatable. In a very broad sense, one can define an asset class as any type of investment that embeds a structural risk premium which cannot be replicated by a portfolio of traditional assets. An investment in commodity indices appeared to carry such a risk premium in the form of normal backwardation, predominantly driven by energy futures. Among all commodities, oil stood out not only as the most valuable commodity in terms of its global production value but also as the one that has the largest futures market. Since oil also happened to have higher historical returns with low and often even negative correlation to other financial asset classes, oil by itself was often perceived to be a convenient proxy for commodities asset class.

Cross-asset correlations are crucial inputs into conventional asset allocation frameworks, many of which are based on the Capital Asset Pricing Model (CAPM). According to the CAPM, the expected returns on all assets are proportional to the asset's contribution to non-diversifiable risk in the overall market portfolio that includes all assets. The optimal asset allocations are then constructed based on assets' expected returns, volatilities, and correlations, which are typically estimated from historical data. Assets with lower correlation to the rest of the portfolio warrant higher allocations. Based on the data available in the early 2000s, the CAPM-optimal investment portfolio could have justified up to 15–25% allocation of all assets to the petroleum-heavy GSCI index.

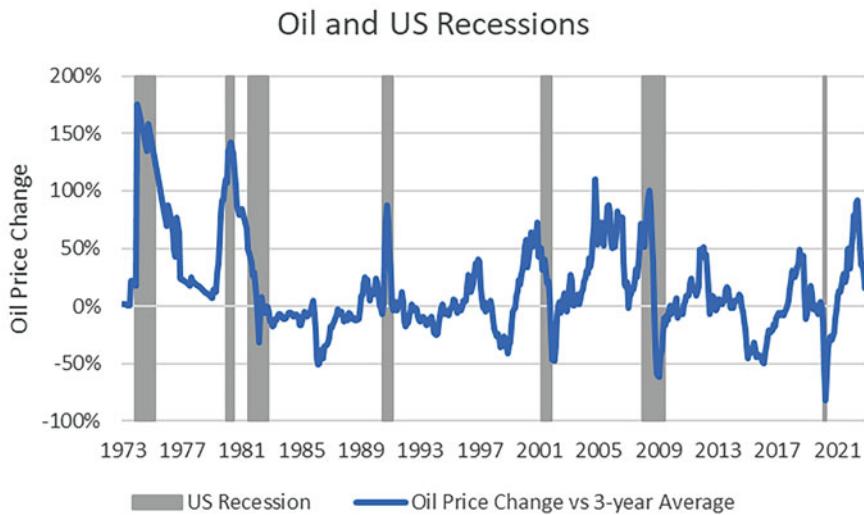
Unfortunately, the concept of linear correlation that underpins CAPM-based asset allocation methodologies is not well suited for analyzing highly nonlinear systems such as the oil market. Any linear measure of dependency is inevitably prone to instability. Correlations computed using historical time series are extremely sensitive to outliers and to the selection of the lookback period, which is often cherry-picked by analysts to improve the optics of the desired outcome. The reliance on correlation estimates in the presence of outliers does not let one see the forest for the trees. Instead, one is better off focusing on specific events that might have caused these outliers. As we have already seen in the study of storage constraints in Chap. 3, the outliers tend to reflect the presence of natural boundaries in the nonlinear system. For the overall economy and financial markets, one such boundary is an economic recession.<sup>13</sup> It turns out that the behavior of oil prices near this boundary is what gave oil its prominent role in investment portfolios.

It has been long observed that most of US postwar economic recessions were preceded by significant increases in oil prices which became known as oil shocks.<sup>14</sup>

---

<sup>13</sup>In the USA, a recession is formally defined by the National Bureau of Economic Research (NBER) as “a significant decline in economic activity spread across the market, lasting more than a few months, normally visible in real gross domestic product (GDP), real income, employment, industrial production, and wholesale-retail sales.” In practice, two consecutive quarters of negative GDP growth is often taken as an indicator of a recession.

<sup>14</sup>This observation was first made in Hamilton (1983).



**Fig. 4.4** Oil price appreciation by at least 50% either preceded or coincided with most US recessions since the 1970s

Prior to the 1970s, price increases were fairly modest, as during those years oil prices were largely regulated. The first large oil shock occurred in 1973 when the price of oil tripled after an oil embargo was imposed by the Organization of Arab Petroleum Exporting Countries (OAPEC). The second energy crisis came in 1979 when oil price doubled in the aftermath of the Iranian Revolution followed by the Iran-Iraq War. In both episodes oil was blamed for severe economic recessions. The next oil spike of 1990 was initially driven by the Persian Gulf War. While it happened shortly after the US recession has already been announced, it also contributed to the economic slowdown. Figure 4.4 illustrates.

These episodes of large disruptions in oil supplies with severe consequences secured a very special place for oil among key macroeconomic variables. Furthermore, the pattern continued to hold with oil prices appreciated again by more than 50% versus its prior three-year average just before the recession of 2001 and then again in the beginning of the Global Financial Crisis. One exception from this statistical anomaly is a brief recessionary period at the onset of the Covid-19 pandemic when oil prices also weakened.

Does it mean that rising oil prices *cause* economic recessions? This question has been the subject of heated debates for nearly forty years. Undoubtedly, the answer depends on whether higher oil prices were driven by supply disruptions, unexpectedly strong demand, or other factors, such as financial speculation.<sup>15</sup> Our preference

<sup>15</sup> Hamilton (2003) developed a nonlinear metric for net oil price increases and applied it to explain oil price shocks mostly with supply disruptions driven by geopolitical events. In contrast, Kilian (2009) attributed a much larger role in many oil shocks to growth in global demand for commodities

is to stay away from this debate as the data is just too noisy and results are too sensitive to the sample selection and to the choice of the econometric methodology. Regardless of whether this pattern of oil being a precursor of economic recessions is just a statistical fluke or created by some invisible market force, its mere existence made oil very special in the eyes of financial investors. An investment in oil provided diversification to financial portfolios precisely when it was most needed.

While recessions tend to be bad for the economy, periods of high inflation are even more challenging for financial investors. The prices of all financial assets are generally determined by the present value of their future cash flows. If inflation unexpectedly rises, causing higher interest rates, then discounting decreases the present value of all future cash flows. For bonds, the negative impact is straightforward as future cash flows are fixed. For equities, the value of real assets owned by companies generally also rises with inflation, but it is offset by higher costs of labor and capital. When inflation moves up quickly and unexpectedly, input costs tend to increase at a faster pace than revenues. In general, rising inflation hurts both stocks and bonds, and oil futures happen to be one among very few investments that tends to be profitable during periods of high inflation.

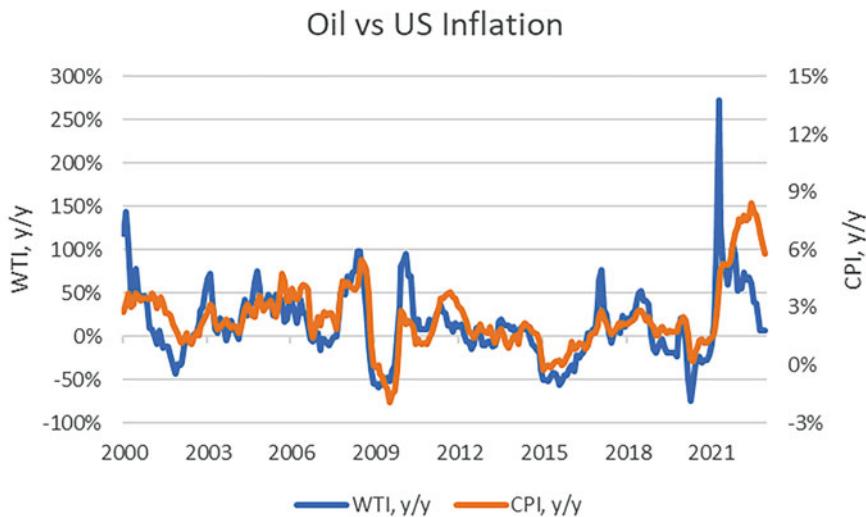
US inflation is typically measured by changes in the Consumer Price Index (CPI), which represents the price for a basket of goods and services paid by a typical consumer. Figure 4.5 shows how closely US annual headline inflation tracks the annual change in the price of WTI. The only visible dislocation between two variables occurred during the recovery period after the Covid-19 pandemic when inflation lagged the change in oil prices.

This relationship is simply too good to be purely statistical. In fact, it is much more mechanical and driven by the algebra of oil pass-through into the consumer cost basket. The price of oil is the primary driver of retail gasoline prices, which make up the majority of the motor fuel component of the CPI. Even though this sub-component represents less than 4% of the CPI basket, it explains the large portion of the CPI monthly variance simply because other components of the CPI move much more slowly. We will return to this topic in Chap. 7 at greater length, where we develop a relative-value trading strategy between energy futures and inflation swaps which is centered around the pass-through of futures into the CPI.

Besides direct mechanical impact on the CPI via retail gasoline prices, oil also affects inflation indirectly. Higher oil price raises the cost of feedstocks for many manufacturing processes, often leading to higher prices of many consumer goods. In addition, rising prices for consumer goods may also force workers to demand higher wages. This could lead to a potentially dangerous inflationary spiral, which is understandably a major concern of policymakers.

---

and to the so-called precautionary demand, which is associated with speculative buying in anticipation of rising uncertainty and shifts in future expectations. The latter approach was formalized via a structural vector autoregressive model (VAR) in Kilian and Murphy (2014) that decomposes contribution of supply disruptions, business cycle-related demand, and speculative oil-specific demand for inventories to the real price of oil. Another VAR model was used by Kilian and Vigfusson (2017) to quantify the contribution of oil shocks to past US recessions.



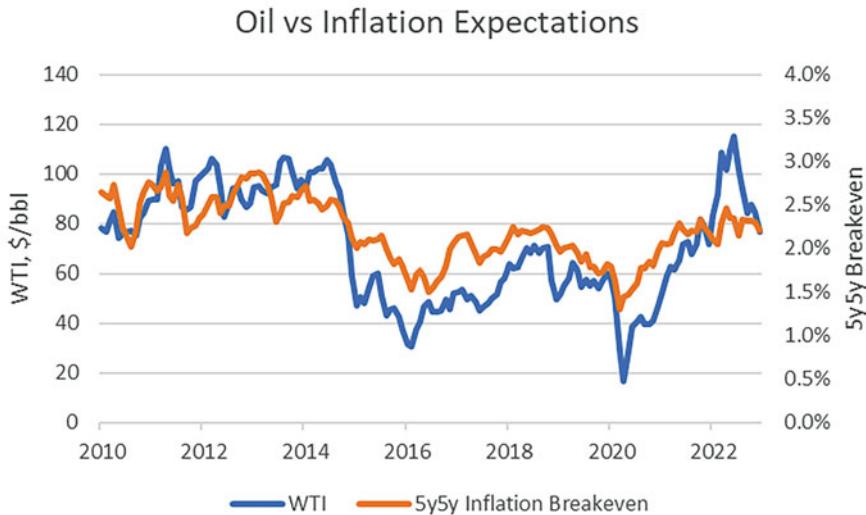
**Fig. 4.5** US inflation measured by year-over-year changes in CPI is largely explained by annual changes in the price of oil

Since the indirect impact of oil on future inflation is difficult to observe quickly, it has been long suggested that consumers often use highly visible and frequently changing gasoline prices to form their expectations about future inflation. Observed trends in gasoline prices are often extrapolated by consumers as trends in broader inflationary pressures. Measuring consumer expectations via surveys, however, is a rather daunting task, so it is not surprising that analysts and policymakers switched to an alternative market-based measure of inflation expectations.

To gain exposure to US inflation in the market, one buys inflation-linked bonds, called *Treasury Inflation-Protected Securities (TIPS)*, which pay the real yield, and simultaneously sells nominal Treasury bonds with the same maturity. The difference between the nominal yield of the Treasury bond and the real yield paid by TIPS is known as the *inflation breakeven rate*. Five-year and ten-year breakevens are two primary inflation benchmarks in the marketplace. They are often combined to form the so-called *5y5y forward breakeven* rate, which represents five-year inflation measured five years forward. One can approximate the 5y5y forward inflation rate as twice the ten-year inflation rate minus the five-year inflation rate. This metric became the de facto standard used by policymakers in estimating market-based measures of inflation expectations.

Figure 4.6 shows a rather surprising picture of how closely the 5y5y breakeven rate follows the price of WTI futures. Even though, theoretically, the oil price today should not have anything to do with such long-dated inflation expectations, the 5y5y inflation breakeven rate tracks the price of WTI quite well.

This co-movement is very bizarre from the statistical perspective. It shows that the *level* of the commodity co-moves with the *rate of change* in the consumer basket.



**Fig. 4.6** 5y5y inflation breakeven closely tracks the price of WTI

The statistical relationship is usually properly measured only when both time series are stationary, and for most financial variables the price levels are not stationary, but price changes or returns are.<sup>16</sup> On the one hand, this relationship can be viewed purely as a coincidence and discarded. On the other hand, as we will see in Chap. 7, the connection between the two markets may indeed be reinforced by cross-asset arbitrageurs. While its merit is still subject to debate, its optics clearly contributes to the belief in the oil's special role as a valuable hedge against the risk of inflation.

Furthermore, it turns out that oil is also somewhat unique in its ability to hedge against inflation surprises, or unexpected shifts in inflation. Obviously, inflation surprises are hard to measure statistically. Measuring them by comparing realized inflation against prior surveys of forecasts by economists has proven to be highly unreliable. A more objective measure of historical inflation surprises can be given by the change in the rate of CPI. Using such a definition, one can then analyze statistically the performance of different assets during such unexpected changes in inflation. Contrary to common beliefs, oil performs much better than gold, real estate, or virtually any other hedging alternative in countering jumps in unexpected inflation, and this is what pension funds that own a large portfolio of stocks and bonds care about.<sup>17</sup>

<sup>16</sup> See, for example, Alexander (2001).

<sup>17</sup> Many of the previous references on commodity indices also discuss superior investment performance of commodity futures during periods of high inflation. Neville et al. (2021) conducted the comprehensive empirical study of hedging against inflation surprises using not only various financial assets but also some systematic trading strategies. Passive investment in petroleum futures

While the existence of normal backwardation and commodity indices brought large institutional investors to the oil market, the next wave of oil financialization was spurred by growing interest in hedging the risk of inflation. A large role in this growth is attributed to the popular investment strategy dubbed *risk parity*. The earliest versions of the risk parity framework are usually credited to the so-called “All Weather” strategy launched by Ray Dalio at a then relatively unknown hedge fund called Bridgewater Associates. Subsequently, it became one of the most profitable hedge funds in the world, and somewhat surprisingly its risk parity strategy turned the fund into one of the world’s largest oil traders.

The “All Weather” strategy was designed as an alternative to CAPM-based asset allocation. It looks at returns on all assets through shifts in two primary driving factors, growth and inflation. This rationale is intuitive since all expected future cash flows are driven by the volume of economic activity, which is growth, and how this activity is discounted today relative to the future. The latter is largely based on future inflation expectations. The standard portfolio of stocks and bonds is only diversified with respect to the growth factor. If the economy is strong, then equity markets should perform well, as stocks represent claims on future company earnings, which are expected to increase when the economy is booming. On the other hand, bonds tend to do better when the economy is weaker, as bonds provide investors with fixed cash flows whose present value increases when interest rates drop. While a conventional portfolio of stocks and bonds provides investors with some balance with respect to shifts in the economic regime, it is not diversified with respect to shifts in inflation expectations. To hedge this risk, commodities and, in particular, oil with its optically impressive inflation-hedging properties, must be added to the portfolio.

The crucial part of the risk parity methodology is in how assets are weighted. Since equities are more volatile than bonds, the traditional 60–40 equity-bond portfolio translates into an approximately 90–10 allocation in risk terms, which provides little real diversification. Instead, in the risk parity allocation framework, all assets are weighted equally based on their risks. Lower-volatility assets, such as bonds, must then be purchased in larger quantities to contribute the same amount of risk as higher-volatility assets. This forces risk parity managers to heavily invest in bonds and leveraged fixed income derivatives. Since inflation is the major risk for bonds, commodities must also be purchased in large quantities, which in practice can only be done in the futures market. While banks brought commodities to unlevered investors via fully funded indices, risk parity went a step further and offered the coveted exposure to commodities on a highly leveraged basis. Among all commodities, oil with its superior inflation-hedging properties and the best liquidity across all commodity futures received the largest slice in the risk parity allocation pie.

Bridgewater’s original idea was adopted by many other providers of risk parity investment strategies. While these strategies vary in implementation, the core

---

is shown to be by far the best hedge. Similar conclusions have been reached in many publications by sell-side research analysts.

premise of allocating larger notional investments to lower-volatility bonds requires fund managers to also buy large amounts of inflation sensitive commodity futures, among which oil stands out. The arrival of risk parity funds not only restored the Keynesian imbalance between sellers and buyers in the futures market, but it also started to shift the disequilibrium in the opposite direction. If positive expected returns from risk bearing alone were not sufficient to induce buying of oil futures, the additional benefit of hedging the risk of inflation became the final straw that broke the camel's back.

## 4.5 Inconvenience Yield, or the Theory of Normal Contango

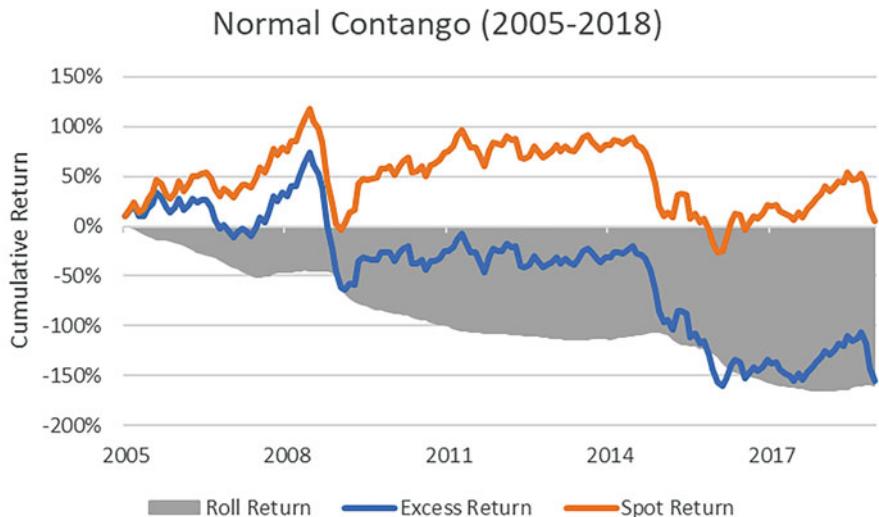
Keynesian risk bearers came to the oil futures market with additional capital to bring it back to equilibrium. They did their job, in fact they did even more than what was needed. They did not stop when the hedging equilibrium was restored, as diversification benefits of oil futures brought further benefits to their broader financial portfolios. They continued to buy oil futures even as their own activity started to push futures prices above the expected spot price. As a result, the oil market shifted into a new structural regime of normal contango.

Extending the concept of normal backwardation, we define normal contango when the *expected* spot price falls below the current futures. This is not the same as market contango. The existence of normal contango implies that investors are expected to lose money by buying and rolling oil futures. These expected losses have indeed materialized. The sign of the risk premium has flipped from positive to negative.<sup>18</sup>

Figure 4.7 extends the graph presented in Fig. 4.3. The strategy performances over two periods are nearly mirror images of each other. Starting from 2005 the strategy of buying and rolling oil futures began to steadily lose money. By the end of 2018, it lost nearly all of its entire gain accumulated during the prior period of normal backwardation, as the normal state of the market was contango. Remarkably, the spot price at the beginning and at the end of this entire period was basically the same. In other words, the entire loss came again from the accumulation of the roll yield which during that period was consistently negative.

The futures market is a zero-sum game. For every loser, there must be a winner. Where did the negative risk premium go to? The answer has already been given in the previous chapter. Selling futures at a premium to financial investors and buying physical barrels is the business of the professional storage trader. From being a provider of a service to oil producers, investors themselves turned into hedgers of financial risks. Investors pay for this service with the risk premium by rolling futures in the contango market with profits accruing to the storage owner. The investor looking to buy futures must incentivize new marginal suppliers of short futures to come to the market, if existing short futures offered by producer hedgers are not

<sup>18</sup>The thesis of normal contango was suggested in Bouchouev (2012).



**Fig. 4.7** Cumulative performance of long WTI futures during the regime of normal contango

sufficient. These incentives have also provided the impetus for building new storage capacity, which we discuss later in the context of specific trading strategies in Chaps. 6 and 13.

Since investors cannot easily buy and store physical barrels themselves, they have no choice but to pay for the privilege to invest in oil. These payments are not explicit. Every month when the futures curve is contango, investors must sell prompt futures before their expiration, and reset the long position at a higher price by buying the next available futures contract. Even though no money is exchanged on the day of the roll, losses are accumulated over time as futures tend to roll down towards the lower spot price. The convenience yield of owning a physical commodity, net of storage costs, has turned negative. The intrinsic yield of oil became much more of an inconvenience. While the negative roll yield acted as the synthetic cost of storage, it also compensated storage owners for providing the service of essentially storing the bulky commodity on behalf of investors.

With the addition of another major market participant, the manager of the broader financial portfolio, the hedging equilibrium Eq. (4.6) must be revisited. An investor, such as a risk parity fund, now has an inflation-hedging mandate that counters the producer's need to sell futures. Let us assume that such an investor is tasked to manage the notional inflation exposure  $Q_i$  in a financial portfolio that generates income  $U_i$  while being exposed to inflation uncertainty  $\tilde{I}$ . The investor follows the same mean-variance framework with risk aversion coefficient  $\alpha_i$ , and trades  $N_i$  oil futures to hedge the inflation risk. The inflation hedger's wealth is then given by<sup>19</sup>

<sup>19</sup>This approach to the derivation of the hedging equilibrium was proposed in Bouchouev (2020).

$$\tilde{W}_i = -Q_i \tilde{I} + (\tilde{S} - F) N_i + U_i$$

Following the same steps as before, we find the futures position  $N_i$  that maximizes the following expression for the investor wealth function

$$\begin{aligned} \max & \left\{ -Q_i E(\tilde{I}) + N_i (E(\tilde{S}) - F) + U_i \right. \\ & \left. - \frac{\alpha_i}{2} \left( Q_i^2 \sigma^2(\tilde{I}) + N_i^2 \sigma^2(\tilde{S}) - 2Q_i N_i \rho_{i,s} \sigma(\tilde{I}) \sigma(\tilde{S}) \right) \right\} \end{aligned}$$

Like in the case of a consumer cross-hedging refined products exposure with crude oil, the variance of the sum of two random variables contains the covariance term. Here,  $\rho_{i,s} = \rho(\tilde{I}, \tilde{S})$  represents the correlation coefficient between inflation and oil.

As before, differentiating with respect to  $N_i$  and equating the partial derivative to zero results in

$$E(\tilde{S}) - F - \alpha_i (N_i \sigma^2(\tilde{S}) - Q_i \rho_{i,s} \sigma(\tilde{I}) \sigma(\tilde{S})) = 0$$

The optimal number of futures held by the inflation manager is then given by

$$N_i = \beta_i Q_i + \frac{E(\tilde{S}) - F}{\alpha_i \sigma^2(\tilde{S})} \quad (4.7)$$

where

$$\beta_i = \rho_{i,s} \frac{\sigma(\tilde{I})}{\sigma(\tilde{S})}$$

represents inflation beta to oil. This is analogous to consumer hedging of refined product exposure by trading crude oil futures, as in (4.4).

The futures now must clear among the producer, the consumer, the inflation hedger, and the traditional speculator:

$$N_p + N_c + N_i + N_s = 0.$$

Adding up Eqs. (4.3), (4.4), (4.5) and (4.7) for optimal hedging positions held by each participant and solving for the risk premium, we obtain that

$$E(\tilde{S}) - F = \alpha \sigma^2(\tilde{S}) (Q_p - \beta_c Q_c - \beta_i Q_i) \quad (4.8)$$

where the risk is shared among all futures traders, as

$$\alpha = \frac{1}{\frac{1}{\alpha_p} + \frac{1}{\alpha_c} + \frac{1}{\alpha_i} + \frac{1}{\alpha_s}}$$

The theory of hedging pressure is now generalized to cover all the main market participants. The resulting risk premium is a function of the net imbalance between producers, consumers, and inflation hedgers. Traditional speculators are incentivized to take either long or short positions depending on the sign of the net hedging imbalance. Since financial markets are significantly larger than oil markets and the demand for inflation hedging has steadily increased over time, the structural risk premium during the era of financialization has become negative. In this case, systematically buying and rolling oil futures is a losing value proposition.

All good things in financial markets eventually come to an end. Storage operators and short speculators enjoyed a long run of spectacular performance during the regime of normal contango. Financial investors, in turn, have learned their lessons and became more dynamic with their oil investment strategies. By and large, the oil markets have largely matured, and the structural directional oil risk premium has vanished. It became nimbler, responding to faster changes in fundamentals of storage and hedging imbalances. To capture it, investment strategies must become dynamic as well. These strategies employed by different types of investors and speculators are presented in the following three chapters of the book.

---

## References

- Acharya, V. V., Lochstoer, L. A., & Ramadorai, T. (2013). Limits to arbitrage and hedging: Evidence from commodity markets. *Journal of Financial Economics*, 109(2), 441–465.
- Alexander, C. (2001). *Market models*. Wiley.
- Ashton, M., & Greer, R. (2008). History of commodities as the original real return asset class. In *Inflation risk and products* (pp. 85–109). Risk Books.
- Backhouse, R. E. (2002). *The ordinary business of life*. Princeton University Press.
- Baker, S. D. (2021). The financialization of storable commodities. *Management Science*, 67(1), 471–499.
- Bodie, Z., & Rosansky, V. I. (1980, May–June). Risk and return in commodity futures. *Financial Analysts Journal*, 36(3), 27–39.
- Bouchouev, I. (2012). Inconvenience yield, or the theory of normal contango. *Quantitative Finance*, 12(12), 1773–1777.
- Bouchouev, I. (2020). From risk bearing to propheteering. *Quantitative Finance*, 20(6), 887–894.
- Büyüksahin, B., & Robe, M. A. (2014). Speculators, commodities and cross-market linkages. *Journal of International Money and Finance*, 42, 48–70.
- Cheng, I.-H., Kirilenko, A., & Xiong, W. (2015). Convective risk flows in commodity futures markets. *Review of Finance*, 19(5), 1733–1781.
- Erb, C. B., & Harvey, C. R. (2006). The strategic and tactical value of commodity futures. *Financial Analysts Journal*, 62(2), 69–97.
- Fama, E. F., & French, K. R. (1987). Commodity futures prices: Some evidence on forecast power, premiums, and the theory of storage. *Journal of Business*, 60(1), 55–73.
- Fattouh, B., & Mahadeva, L. (2014). Causes and implications of shifts in financial participation in commodity markets. *The Journal of Futures Market*, 34(8), 757–787.
- Gorton, G. B., Hayashi, F., & Rouwenhorst, K. G. (2012). The fundamentals of commodity futures returns. *Review of Finance*, 17(1), 35–105.

- Gorton, G., & Rouwenhorst, K. G. (2006). Facts and fantasies about commodity futures. *Financial Analysts Journal*, 62(2), 47–68.
- Greer, R. J. (1978, Summer). Conservative commodities: A key inflation hedge. *Journal of Portfolio Management*, 4(4), 26–29.
- Hamilton, J. D. (1983). Oil and the macroeconomy since World War II. *Journal of Political Economy*, 91(2), 228–248.
- Hamilton, J. D. (2003). What is an oil shock? *Journal of Econometrics*, 113(2), 363–398.
- Hamilton, J. D., & Wu, J. C. (2014). Risk premia in crude oil futures prices. *Journal of International Money and Finance*, 42, 9–37.
- Hicks, J. R. (1939). *Value and capital: An inquiry into some fundamental principles of economic theory*. Oxford University Press.
- Hirshleifer, D. (1988). Residual risk, trading costs, and commodity futures risk premia. *The Review of Financial Studies*, 1(2), 173–193.
- Keynes, J. M. (1923, March 29). Some aspects of commodity markets. The Manchester Guardian Commercial, Reconstruction Supplement.
- Keynes, J. M. (1930). *A treatise on money* (Vol. II). Macmillan.
- Kilian, L. (2009). Not all oil price shocks are alike: Disentangling demand and supply shocks in the crude oil market. *American Economic Review*, 99(3), 1053–1069.
- Kilian, L., & Murphy, D. P. (2014). The role of inventories and speculative trading in the global market for crude oil. *Journal of Applied Econometrics*, 29(3), 454–478.
- Kilian, L., & Vigfusson, R. J. (2017). The role of oil price shocks in causing U.S. recessions. *Journal of Money, Credit and Banking*, 49 (8), 1747–1776.
- Neville, H., Draaisma, T., Funnell, B., Harvey, C. R., & Van Hemert, O. (2021). The best strategies for inflationary times, SSRN.
- Stoll, H. R. (1979). Commodity futures and spot price determination and hedging in capital market equilibrium. *The Journal of Financial and Quantitative Analysis*, 14(4), 873–894.
- Tang, K., & Xiong, W. (2012). Index investment and the financialization of commodities. *Financial Analysts Journal*, 68(6), 54–74.
- Till, H., & Eagleeye, J. (Eds.). (2007). *Intelligent commodity investing*. Risk Books.

---

## **Part II**

### **Quantitative Futures Strategies**



# Systematic Risk Premia Strategies

5

- Systematic commodity trading attempts to identify and harvest alternative risk premia caused by structural market imbalances and behavioral biases of market participants. The primary risk premia are momentum, carry, and value.
- Oil momentum has its roots in the theory of storage. Oil supply and demand are slow to adjust, inducing persistent trends in inventories. Many conventional formulations of momentum strategies, however, are sensitive to parameter selection and prone to overfitting.
- Carry strategies also follow from the theory of storage but, unlike momentum, the carry signal is forward-looking and largely model-independent. It transmits fundamental information about supply and demand to the futures market via the behavior of inventory hedgers.
- Multiple risk premia can be combined to strengthen the predictive power of the signal. While carry can be viewed as describing the prevailing state of inventories, momentum applied to carry is associated with expected changes in inventories.
- The value risk premium is a mean-reverting strategy of buying low and selling high. Even though the spot price of oil fluctuates cyclically, betting on price mean-reversion in the futures market is not a viable strategy due to the headwinds of negative carry.
- An important component of systematic trading is dynamic position sizing. The strength of the signal can be transformed into the optimal risk exposure via the reaction function. This function is often characterized by an inflection point beyond which the position is reduced.

---

## 5.1 The Evolution of Algos

Traditional technical analysis can be viewed as the predecessor of quantitative futures trading. The idea of identifying and betting on recurrent patterns in commodity prices is as old as commodity trading itself. Japanese rice traders were known

to use candlestick charts in the eighteenth century, and many other techniques developed by commodity traders remained secret and largely undocumented. It is impossible to give the entire credit for pioneering technical trading of commodity futures to a single person or trade shop. The secret sauce has been built by generations of traders learning as apprentices from their masters and passing new tricks to their successors.

The quantitative trading industry grew side by side with the development of the computing power required to handle large sets of data. By the late 1950s, several trading units that resembled modern quantitative hedge funds were already in existence. Perhaps, an honorary mention should go to Commodities Corp., the fund set up in 1969 in an old farmhouse near Princeton University by Helmut Weymar. His MIT dissertation on modeling cocoa prices was arguably the first blueprint for the quantitative analysis of commodity markets.<sup>1</sup> Among the founding partners of Commodities Corp. were other MIT luminaries, including Nobel Prize winner Paul Samuelson and Paul Cootner, best known for his bestseller on random behavior of stock prices. Much like Keynes, the legends of the academic world were ready to put their money where their mouth was.<sup>2</sup>

The original vision of Commodities Corp. was to collect unique fundamental data and use it to run sophisticated statistical prediction models. To obtain the data, the fund even hired agents to monitor the state of cocoa trees in Africa. While the theoretical rationale behind fundamental models appeared to be sound, the actual trading performance turned into a fiasco. The markets stubbornly defied rationality and often moved further away from fundamentally justifiable fair values. To mitigate mounting losses, the fund introduced risk management filters to penalize traders for betting against the prevailing trend. Surprisingly, the risk management system itself generated consistent profits. It was then applied across a wider universe of futures markets, which gave rise to what was possibly the first diversified commodity trend-following system. Besides, its systematic nature brought significant cost savings, as trading decisions were entirely based on quantitative rules with no need to employ agents to track the state of the crops.

Today, quantitative futures traders are usually called commodity trading advisors (CTAs), even though this term is somewhat of a misnomer. A CTA is a broad regulatory definition that applies to anyone who provides investment advice on trading commodity futures. In addition, the regulatory definition of the term commodity is so broad that it even includes interest rate and currency futures as financial

---

<sup>1</sup>See Weymar (1965). In addition to empirical study of cocoa prices using fundamental data, Weymar developed one of the first theoretical models of storage, which links prices and inventories via coupled differential equations.

<sup>2</sup>The fund's trading roster included the names of Paul Tudor Jones, Louis Bacon, and Bruce Kovner, who subsequently became trading legends in their own right, creating multi-billion dollar hedge funds known to be among the largest commodity traders. Jones founded Tudor Investment Corporation, Bacon is the founder of Moore Capital Management, and Kovner established Caxton Associates. More details about the history of Commodity Corp. are provided in Tully (1981) and Lux (2003).

commodities. Most money managers must be registered as CTAs, regardless of how they trade, and vice versa, many CTAs are discretionary traders. In practice, however, the term CTA became synonymous with the business of managed futures, or any systematic futures trading based on quantitative algorithms.

The business of a quantitative CTA is based on the science of statistical inference, which identifies repeatable historical patterns and attempts to extrapolate them into the future. A hypothesis is made about the relationship among certain variables and a prediction is formed that this relationship will continue to hold with a certain probability. Trades are placed based on such predictions, generally with no discretionary intervention. The goal is to eliminate human biases, such as the avoidance of buying high or selling low, or the reluctance to size up when it feels the least comfortable. Since the trading rules are known, the performance of the strategy can be evaluated by backtesting, or, in other words, by simulating the strategy's historical results. To validate the hypothesis with the arrival of new data, the simulation process must always be tested out of sample. The statistical edge in each prediction is rather small, so one needs to apply it either many times repeatedly or diversify across many assets, much like in the business of running a casino.

Commodity futures turned into fertile ground for CTAs for a number of reasons. Futures trading provides high leverage as the initial cost is only a small fraction of the notional value of the asset. Furthermore, different commodities are impacted by their idiosyncratic factors that make strategies less correlated and the overall portfolio better diversified. Unlike equities, there are no restrictions on shorting futures, and one does not need to pay the dealer to borrow the stock; selling futures is as easy as buying. Finally, commodities have larger probabilities of extreme moves. These moves can be captured with systematic momentum and breakout strategies. It would be more difficult to execute such a strategy discretionarily as buying high and selling low goes against human contrarian biases.

In addition, commodity momentum strategies bring valuable diversification to broader investment portfolios. This diversification comes from the negative correlation during major market selloffs, when many financial portfolios struggle but momentum strategies tend to do well. Much like oil investments functioning as a hedge against inflation, the momentum strategy performs well precisely when it is the most needed. Some hedge funds even developed investable products to capture a *crisis alpha* with an asymmetric momentum signal which is only triggered on the downside. Such one-sided momentum strategies are marketed as an alternative to buying insurance in the form of put options. In addition, investors in a systematic strategy may feel that they are in control since trading decisions can be anticipated, in contrast to less predictable judgments of discretionary managers.

As normal backwardation largely disappeared from many commodity markets and passive long-only indices stopped being profitable, banks had a desperate need to find new investable commodity products to retain large financial investors. The catalyst came from equity markets, which were enthusiastically exploring the novel concept of factor investing. The idea was to replace traditional portfolio construction by asset class with allocations to common risk factors that can explain price behaviors across different markets.

The concept of risk factors was quickly adopted and subsequently enhanced by commodity futures traders. It was relatively straightforward to replace struggling long commodity indices with more dynamic commodity portfolios that combine both long and short positions. The signals would be based on the common risk factors that investors were familiar with from equity markets but applied across a broader universe of commodities. The goal once again was to extract a small risk premium from each market and reap diversification benefits from the large number of mostly uncorrelated commodity sub-sectors.

The three most powerful commodity risk factors are momentum, carry, and value, which can be combined in many different ways. A diversified systematic commodity portfolio is typically constructed by ranking its constituents on the cross-asset basis. A common method to do this is to sort all commodities within the portfolio based on the strength of various factors and assign them some sort of score. Commodities with the highest scores are then bought, and the ones that have the lowest scores are sold. Many alternative risk factors, such as inventory, speculative positioning, volatility, skewness, and open interest have also been considered, but as we explain in the next chapter, these factors in the oil market are better used in a less systematic manner.<sup>3</sup>

It turns out that despite the complexity of a diversified commodity risk factor portfolio, the largest portion of the profits is typically generated by energy futures. The role of many other commodity sectors is mostly to provide diversification and to reduce the denominator of the risk-adjusted returns. In the following sections we will explain that the success of these strategies in the energy market can be explained by the special role of storage and the related behavior of large market participants.

---

## 5.2 Myths and Realities of Oil Momentum

The basic thesis of momentum or trend-following strategies is buying what has already gone up and selling what has gone down. One can use a popular analogy that the market, like a horse, is easier to ride in the direction that it is already going. For many discretionary traders, however, it is undoubtedly challenging to embrace the counterintuitive concept of buying high and selling low. Well-known human behavioral biases direct many of us to do exactly the opposite and attempt to buy on dips and sell on rallies. We are reluctant to lock in losses and often retain falling assets for too long in the hope of their eventual recovery, and we tend to take profits prematurely to enjoy faster gratification. Momentum trading takes the other side of such human biases. To eliminate these emotions from decision making, the momentum strategy is typically implemented in a rule-based systematic manner that does not allow for any human intervention.

---

<sup>3</sup>While the literature on long-short commodity risk premia portfolios is broad, we highlight the following work by Szymanowska et al. (2014), Daskalaki et al. (2014), Fernandez-Perez et al. (2018), Bakshi et al. (2019), and Boons and Prado (2019). See also Miffre (2016) for a comprehensive survey of empirical results and additional references on long-short commodity portfolios.

There is no shortage of behavioral theories trying to explain the root cause of momentum observed across many financial markets. Some argue that momentum is created by investors underreacting to new information which is slow to diffuse. Others think that it is driven by investors extrapolating past returns due to inertia. One can also add overconfidence bias as a factor, as we are often reluctant to adjust our subjective views quickly if new information does not support them. We also tend to chase the desired investment more aggressively for the fear of missing out when fresh news confirms our beliefs. More recently, herding started to play a larger role in short-term momentum, ignited by interactions and mutual reinforcement on social media.

While such behavioral explanations may have some merit for diversified equity markets with broad participation of retail investors, a more plausible explanation for momentum in the oil market lies in its fundamentals. Oil supply and demand are price inelastic, and they are very slow to adjust. If oil consumption exceeds oil production today, then it is more likely that the status quo will prevail tomorrow, which results in drawing inventories. Oil consumption tends to follow longer-term business cycles, but relatively rigid production cannot adjust quickly in response to unexpected demand shocks. As a consequence of the theory of storage, such persistence in the dynamics of oil inventories must then translate into persistence of prices.

We will use the concept of momentum synonymously with trend following. In the context of a single market, the two terms are identical. More generally, trend following should be identified as the time series momentum that generates trading signals solely based on the asset's own price history. In contrast, cross-sectional momentum, which is often used in the construction of broad commodity momentum portfolios, ranks assets based on the relative strength of their corresponding momentum signals. In such portfolios, one buys commodities with the highest momentum rank and sells the ones with the lowest rank, while maintaining overall relatively neutral exposure to an asset class. If desired, one can also retain the directional bias of the overall portfolio by buying all prior winners and selling all prior losers within the asset class.

As a first-order approximation, one can interpret a portfolio of time series momentum for individual commodities as the sum of the cross-sectional portfolio and some common trend across all commodities. The contribution of a common trend, however, is much smaller for commodities than it is for equities. In normal market environments individual commodities are not as highly correlated with each other. These low correlations dilute the commodities' common trend, except for occasional episodes of high inflation and global demand shocks that may have similar impact across many commodities. One example of such a common trend is the negative demand shock at the onset of the Covid-19 pandemic, followed by the positive inflationary shock driven by the subsequent recovery. If there is no strong trend in the commodities sector overall, then the performance of a cross-sectional momentum portfolio will be similar to the combined performance of trend-following strategies for its constituents. If, in addition, all commodities trend together, then the

portfolio of time series momentum will outperform the sector-neutral portfolio of cross-asset momentum.

An important, but often overlooked, feature of the long-short momentum portfolio is the disproportionately large contribution of energy futures that tend to exhibit stronger momentum properties. As discussed previously, the complexity of oil and gas storage leads to structurally different return distributions for energy commodities. In contrast to many agricultural and other seasonal commodities that tend to have a price distribution with a larger upside tail, oil follows an *up the stairs, down the elevator* dynamics. Such a behavior is more typical for a financial asset than for a typical consumption commodity. Large periodic breakdowns of the oil market make a diversified momentum strategy highly desirable to financial investors. As we will see later in the discussion of oil options, the downside of oil momentum is also exacerbated by delta hedgers of short put options.

There are many ways to define the time series momentum. The most commonly used approach is to compare the latest price to some of its recent history. The latter is often measured by the moving average of prices  $P_t$  over the prior  $n$  days, defined as

$$MA_t(P_t; n) = \frac{P_t + \dots + P_{t-n+1}}{n} = \frac{1}{n} \sum_{i=1}^n P_{t-i+1}$$

We should note that for consistency with standard academic literature on time series analysis, time variable in this part of the book is indicated by a subscript, so that  $P(t) = P_t$ . Likewise, for the futures price  $F(t, T) = F_t(T) = F_t$ , where the maturity  $T$  is sometimes omitted when it is clear from the context.

We then define basic momentum  $M_t$  as the spread between the latest price and its moving average

$$M_t(P_t; n) = P_t - MA_t(P_t; n)$$

The momentum trading signal,  $\pi_M(F_t)$ , buys and sells futures  $F_t$  based on the sign of  $M_t$ <sup>4</sup>

$$\pi_M(F_t) = \begin{cases} +1, & \text{if } M_t(P_t; n) \geq 0 \\ -1, & \text{if } M_t(P_t; n) < 0 \end{cases} = sign(M_t(P_t; n)) \quad (5.1)$$

The strategy P&L at time  $t$  is calculated using the momentum signal at time  $t - 1$ , i.e.,

$$P\&L(t) = \pi_M(F_{t-1})(F_t - F_{t-1})$$

---

<sup>4</sup>For brevity, when defining the trading signal with the sign function we assume that  $sign(0) = +1$ , which ensures that either a long or a short position is held on each trading day.

While this clarification may appear to be obvious, we mention it explicitly to prevent analysts from making a grave mistake of calculating daily P&L at time  $t$  using the signal calculated at the end of the same trading day.

Note the strategy specification (5.1) trades futures  $F_t$ , even though the momentum signal could be defined for a different time series  $P_t$ . In other words, the signal generation does not have to apply to the same instrument that one trades. For example, momentum signal can be calculated using the time series for fixed nearby or for fixed maturity month futures contract. In the following chapter, we also provide examples of applying the momentum signal to non-price variables, such as fundamentals and flows. Whenever the price series  $P_t$  is clear from the context, we do not reference it explicitly and simply use  $MA_t(n)$  and  $M_t(n)$  to denote, respectively,  $n$ -day moving average and  $n$ -day momentum. In this chapter, we only use a simple price momentum, but even in the basic case, some implementation nuances of the momentum strategy must be handled with caution.

Consider, for example, the impact of futures rolls. CTAs tend to keep their positions in the most liquid contract, which is typically the nearby futures. Prompt futures are usually rolled within one or two weeks prior to their expiration, when the liquidity starts shifting to the second nearby futures contract. Handling the rolls in the definition of the price momentum based on moving averages could be tricky. If one uses a continuous time series of prompt futures, then the impact of the rolls would generate artificial jumps in the time series which could produce false momentum signals. Instead, it is more appropriate to calculate moving averages using the roll-adjusted time series, which is a synthetic time series constructed from cumulative price changes for the appropriate contract.

The proper adjustment for rolls in signal generation is particularly important for seasonal commodities. Consider, for example, RBOB futures.<sup>5</sup> The prompt futures price jumps when RBOB transitions from its winter specification, which ends with the March contract, to the summer specification, which starts with the April contract. Winter and summer RBOB represent two rather different molecules, and the latter always trades at a significant premium to the former due to additional blending expenses required for compliance with stricter environmental regulations during summer. If momentum is defined for the continuous time series of the prompt contract, then a seasonal jump will generate a long signal on the day of the March–April roll and a short signal in the fall when the contract specification reverts to the winter grade. Such jumps will generate false signals that have nothing to do with the trendiness of a given futures contract. If moving averages are calculated using different contracts, then additional seasonality adjustments are usually required.

---

<sup>5</sup>RBOB stands for Reformulated Blendstock for Oxygenate Blending. It represents a primary blending component used to create the finished gasoline product. Even though RBOB does not represent the finished product, it is widely used as the most liquid benchmark for gasoline prices in the futures market.

An alternative way that bypasses the issue of rolls is to redefine the momentum signal in terms of price changes or roll-adjusted returns instead of moving averages of price levels. Momentum then means buying futures when past return over a given period are net positive and selling futures if past returns are net negative. We show next that these two definitions are similar, as moving averages can be expressed as time-weighted price changes.

To illustrate the equivalence between the two alternative definitions, consider an example of a simple short-term momentum based on a five-day moving average which does not cross the rollover day:

$$M_t(5) = P_t - \frac{P_t + P_{t-1} + P_{t-2} + P_{t-3} + P_{t-4}}{5}$$

This formula can be rewritten as the linear combination of four previous price changes, where the weights decrease linearly for older observations:

$$M_t(5) = \frac{4}{5}(P_t - P_{t-1}) + \frac{3}{5}(P_{t-1} - P_{t-2}) + \frac{2}{5}(P_{t-2} - P_{t-3}) + \frac{1}{5}(P_{t-3} - P_{t-4})$$

This formulation is intuitive, as the largest weight is attributed to the latest and arguably most relevant observation. The weights on past price changes gradually decrease, which makes the signal less sensitive to prices that drop out of the sample.

The same transformation can be repeated for any lookback period  $n$ :

$$\begin{aligned} M_t(n) &= \frac{n-1}{n}(P_t - P_{t-1}) + \frac{n-2}{n}(P_{t-1} - P_{t-2}) + \dots + \frac{1}{n}(P_{t-n+2} - P_{t-n+1}) \\ &= \sum_{i=1}^{n-1} \left(\frac{n-i}{n}\right) dP_{t-i+1} \end{aligned}$$

where

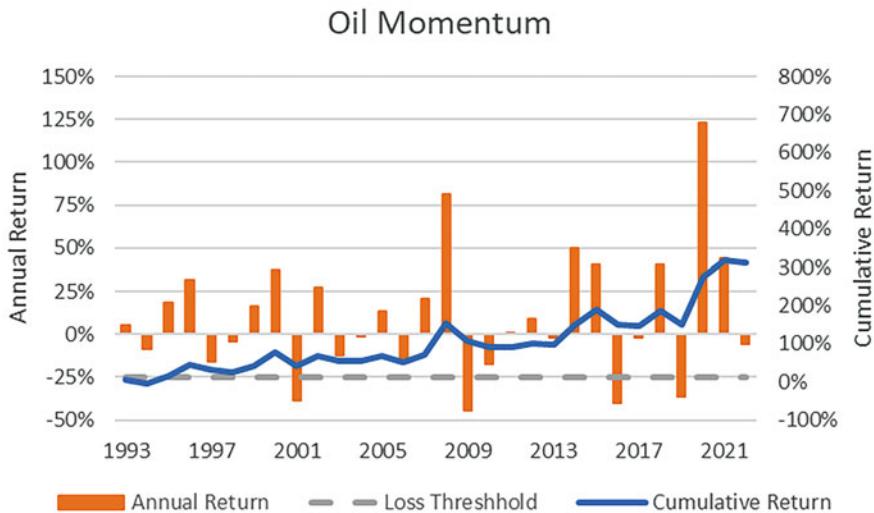
$$dP_t = P_t - P_{t-1}$$

defines daily price changes. The weights on price changes again monotonically decrease with time, and the largest weight is assigned to the latest price change. This formulation is also convenient because price changes can easily be used for the specific contract regardless of what nearby futures it represents at different times.

One can further modify the definition of momentum and specify it directly in terms of a weighted average of prior price changes:

$$M_t(n) = \sum_{i=1}^{n-1} \omega_i dP_{t-i+1}$$

For a very specific choice of weights given by  $\omega_i = \frac{n-i}{n}$ , this definition is equivalent to the one based on the moving average of prices. However, the weights  $\omega_i$  can also



**Fig. 5.1** Historical returns of a 20-day momentum strategy for WTI futures (1993–2022)

be chosen in many other ways. For example, exponentially decreasing weights can be used to assign higher relevance to the recent data. One can also use equal weights  $\omega_i = \frac{1}{n-1}$ , in which case the definition of momentum reduces to the difference between the latest price  $P_t$  and the roll-adjusted price  $n - 1$  days ago  $P_{t-n+1}$ .

Figure 5.1 illustrates the long-term performance for the benchmark 20-day momentum strategy  $M_t(20)$  for WTI futures, which corresponds to approximately one-month lookback. In this example, we use a prompt futures contract which is rolled on the last business day of the calendar month. Such a rolling schedule simplifies the analysis of diversified portfolios of commodity futures, as it avoids dealing with expiration calendars for individual commodities. It also makes it easier for interested readers to replicate the results as an exercise. As before, for simplicity we use log-returns which are additive.

At first glance the average annual return of 10.4% generated by this simple strategy over such a long time may look appealing. However, it is less attractive on a risk-adjusted basis, with an information ratio of only 0.27.<sup>6</sup> In four years, the strategy drawdowns substantially exceeded 25%, which often marks the maximum loss that investors are willing to tolerate for any hedge fund investment strategy. While the momentum signal clearly has some informational content, the strategy is unlikely to be good enough to be traded on a stand-alone basis.

<sup>6</sup>In general, the term information ratio is used to measure the performance of the strategy relative to a certain benchmark. It is analogous to the conventional Sharpe ratio if the risk-free return is replaced with the return on the benchmark. Since trading futures does not require much initial capital, the benchmark for most strategies is typically set at zero. Therefore, here the information ratio is simply defined as the ratio of the strategy's annualized return to its annualized volatility, and, for convenience, we use the term information ratio interchangeably with the Sharpe ratio.

One of the main objectives of this book is not only to share ideas that have a proven track record, but also to highlight their pitfalls. While momentum is a useful trading concept, it is, unfortunately, one that is often abused. In systematic trading, such abuse comes from data mining and overfitting when the trading signal is allowed to have too many degrees of freedom that are fitted to produce better-looking backtests. The more parameters a trading strategy has, the more likely it is to fall apart when some of these parameters are modified. Momentum strategies provide a good case study.

The most common extension of basic momentum is to smooth the signal sensitivity to the latest data point. Financial markets are full of noise, and the momentum trader may want to avoid being whipsawed by false signals and instead to wait for some confirmation of the trend to be firmly established. Such smoothing is typically done by replacing the latest price in the definition of momentum with another, shorter-term moving average. The crossover momentum signal is generated by the difference between moving averages

$$M_t(m, n) = MA_t(m) - MA_t(n), m < n$$

The trading signal  $\pi_M$  of the momentum strategy then switches between long and short futures  $F_t$  when the two moving averages cross over

$$\pi_M(F_t) = \text{sign}(M_t(m, n))$$

While an introduction of additional degrees of freedom provides traders with more flexibility to search for optimal parameters to improve the historical backtest, it adversely effects the stability of the output. Here, one effectively searches for local minima in a multidimensional space of parameters, and such local minima are rarely unique.

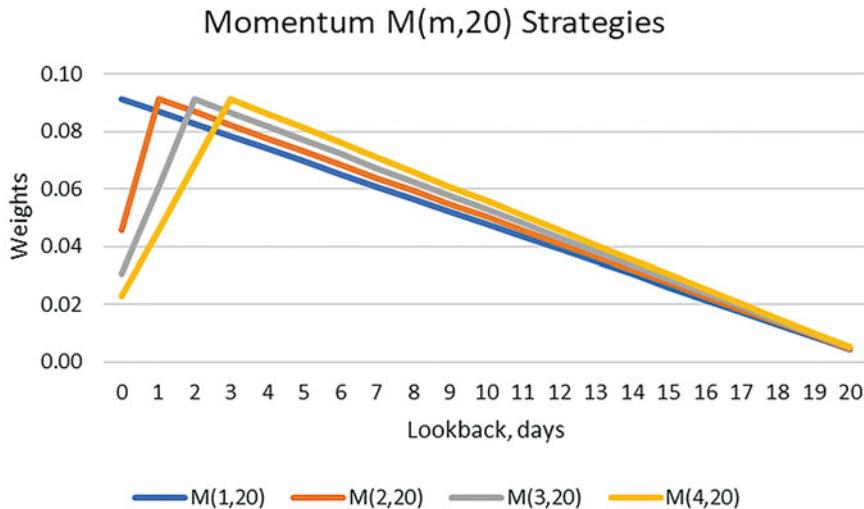
To illustrate the nature of instability inherent in many momentum strategies, we can again translate the crossover of moving averages to an equivalent definition based on price changes. Following the previous transformation, we represent both moving averages of prices in terms of price changes:

$$M_t(m) = P_t - \frac{1}{m} \sum_{i=1}^m P_{t-i+1} = \sum_{i=1}^{m-1} \left( \frac{m-i}{m} \right) dP_{t-i+1}$$

and

$$M_t(n) = P_t - \frac{1}{n} \sum_{i=1}^n P_{t-i+1} = \sum_{i=1}^{n-1} \left( \frac{n-i}{n} \right) dP_{t-i+1}$$

Then using these representations, we can write the crossover momentum as follows



**Fig. 5.2** The weights for the  $M(m, 20)$  momentum strategy peak for the price change that occurred exactly  $m - 1$  days ago

$$\begin{aligned}
 M_t(m, n) &= MA_t(m) - MA_t(n) = \frac{1}{m} \sum_{i=1}^m P_{t-i+1} - \frac{1}{n} \sum_{i=1}^n P_{t-i+1} \\
 &= P_t - \sum_{i=1}^{m-1} \left( \frac{m-i}{m} \right) dP_{t-i+1} - P_t + \sum_{i=1}^{n-1} \left( \frac{n-i}{n} \right) dP_{t-i+1}
 \end{aligned}$$

where the two  $P_t$  terms in the last line cancel out.

We then split the second summation into the first  $m - 1$  and remaining  $n - m$  data points, and combine the former with the first summation to obtain

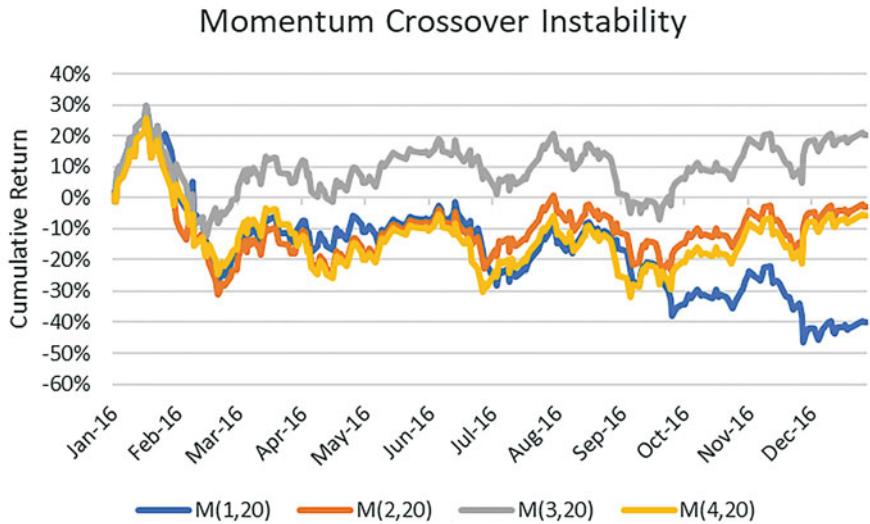
$$M_t(m, n) = \sum_{i=1}^{m-1} \frac{(n-m)i}{mn} dP_{t-i+1} + \sum_{i=m}^{n-1} \frac{(n-i)}{n} dP_{t-i+1} = \sum_{i=1}^{n-1} \omega_i dP_{t-i+1}$$

The crossover momentum is still represented as a weighted sum of price changes, but the weights  $\omega_i$  are no longer monotonically decreasing. Since

$$\frac{n-m}{mn} < \frac{2(n-m)}{mn} < \dots < \frac{(m-1)(n-m)}{mn} < \frac{m(n-m)}{mn} > \frac{(n-m-1)}{n} > \dots > \frac{1}{n}$$

the largest weight is assigned to the date which corresponds to the start of the short lookback window.

Figure 5.2 illustrates these weights for the  $M(m, 20)$  momentum strategy. The weights reach their corresponding peaks for price changes that occurred exactly  $m - 1$  days ago.



**Fig. 5.3** An example of instability of crossover momentum strategies with respect to short-term moving average. In 2016,  $M(1,20)$  lost 40%,  $M(3,20)$  made 20%, while returns on  $M(2,20)$  and  $M(4,20)$  strategies were close to zero

This simple illustration exposes how easy it is to overfit the crossover momentum strategy. By varying  $m$  in the signal construction, one can be fooled by seeking to improve the historical performance, while essentially fitting to the noise. An optimizer will always pick the best combination of parameters, which will likely artificially overweight the contribution of some large favorable price move that happened to occur precisely  $m - 1$  days ago. The model becomes unstable and any attempt to dynamically adjust the set of optimal parameters will result in nothing but fitting to randomness.

Figure 5.3 illustrates how sensitive the results of such momentum strategies can be to even a small perturbation of parameters. For example, in 2016 the base momentum strategy would have lost 40%, a remarkably similar strategy where the short-term momentum averages prices over the past three days would have made 20%, and returns on strategies with price averaging over two and four days were close to zero.

There is nothing structural about this choice of parameters and one can easily find another period when the reverse was true. The difference is typically driven by how well the parameter optimizer captures a small number of large returns that tend to dominate the overall strategy performance. On the bright side, the sensitivity to the longer-term moving average,  $n$ , becomes significantly smaller as the weights on the corresponding price changes monotonically decrease.

Our goal here is not to find the best combination of parameters for the oil momentum strategy. We believe that this would be an ill-intentioned task given the non-stationarity of the financial time series. It is clear that the oil momentum does

exist. It contains some useful information that one can see even in the historical performance of the basic momentum strategy. A few other formulations of momentum strategies are also worth mentioning, not because they are better, but more because they are often cited by researchers who claim to have found some other magic versions of momentum.

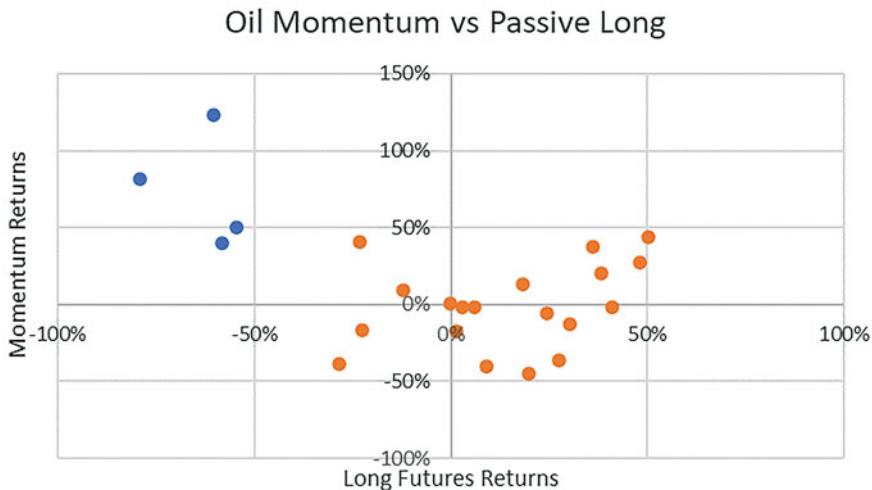
One popular variation of momentum is a class of breakout strategies where the asset is bought or sold only when the latest price moves beyond its prior maximum or minimum over some given lookback period. In contrast to the basic momentum, the breakout strategy is likely to remain on the sidelines for substantial periods of time, which, unfortunately, makes it even more prone to overfitting. Further variations could include filters to keep the moving averages crossover intact for a certain number of days to avoid being whipsawed before taking a position. Alternatively, one can impose the filter to enter the trade only when the spread between two moving averages exceeds a certain threshold, which becomes an additional parameter as well. Similar conditions can be imposed around the exit of the position.

There are a few other general techniques that are helpful for momentum trading regardless of its exact specification. One can often benefit from combining momentum signals across multiple trading frequencies. For example, one can use the crossover of the latest price and its weekly price average as the short-term momentum signal, the weekly versus monthly averages as the medium-term signal, and the monthly versus annual moving averages as the longer-term momentum signal. Then the three signals can be combined into an aggregate momentum score which determines the overall size of the trading position. The weights for each of the three momentum frequencies can also be customized.

Undoubtedly, the more parameters the strategy backtest is allowed to have, the better the fitted results will inevitably be. However, if one normalizes numerous permutations of momentum strategies for the same degrees of freedom, then their long-term performances tested out-of-sample become broadly similar as well. Any claim of the discovery of some magic momentum specification that works much better than any other momentum strategies should be treated with a great amount of skepticism. If one is sufficiently excited about the rationale of the oil momentum strategy, then the best advice would be not to chase the best backtest but to keep the strategy simple and to avoid overfitting.

One of the reasons that explains the popularity of the oil momentum strategy is its negative correlation to directional financial investments during synchronized risk-off events when any benefits of portfolio diversification are particularly appreciated. For example, a strategy of buying and rolling oil futures lost over 50% in four years since 2000, but simultaneous investments in a momentum strategy would have largely offset these losses. Figure 5.4 highlights these years with blue dots. Since momentum strategies tend to do well when prices make large moves in either direction but underperform when prices are range bound, the graph of momentum returns versus price returns is sometimes referred to as *momentum smile*.

One important topic for systematic trading that we have not yet discussed is the sensitivity of the strategy to transaction costs. Fortunately, for liquid oil futures, such as WTI and Brent, the existence of the TAS mechanism described in Chap. 3 makes



**Fig. 5.4** Momentum strategies would have mostly offset large losses from directional investments in oil futures (WTI, 2000–2022)

the task much easier as it allows systematic traders to lock in the settlement price practically without crossing the bid-ask spread. We should note that similarly set up minute marker contracts also allow traders to lock in prices during pricing windows for physical markets in the European and Asian time zones. The European prices are often used by managers of global systematic portfolios to capture the best simultaneous liquidity across the three main time zones. One can also run faster intraday momentum strategies based on higher-frequency data. Such strategies are outside the scope of this book, as they depend more on the market microstructure than on the behavior of participants.

To summarize, one can consider practically infinite permutations for different momentum strategies for oil futures, but in the long run, most exhibit similar predictive power once the somewhat artificial contribution driven by additional degrees of freedom is removed. Oil momentum does add some useful information content, but its strength appears to be declining as markets become more liquid and mature. Practitioners observed that momentum strategies generally work better in less liquid markets, where the information diffuses more slowly. However, the execution of such strategies in illiquid markets is more challenging due to higher transaction costs.

The oil momentum can be used to enhance other strategies with more solid economic foundations, but it should not be abused by data mining. As we will see later, the momentum in liquid oil markets can still add significant value when it is combined with other drivers that have stronger economic foundation. Momentum is a great optimizer of other trading signals and should be used as such. As a stand-alone trading strategy, it is not sufficiently robust and the economic foundation

behind the concept also remains rather weak. As traders often quip, momentum works until it does not.

### 5.3 Carry as a Transmitter of Fundamentals to Prices

One pivotal factor that threads through virtually every aspect of oil trading is the shape of the futures curve. It arises everywhere while often hiding under different names in specific applications. Some economists may recognize it as the own rate of interest, fundamental traders know it as the convenience yield, and financial investors call it the roll yield. Systematic traders refer to the same term by yet another name, the commodity carry.

In the world of systematic trading, the term carry is often understood in a broader sense, which allows it to be applied across many markets of different nature. Simplistically, carry can be defined as P&L if nothing changes. When applied to commodity futures, this definition means P&L of the futures trading strategy under the assumption that the spot price does not change, and the term structure of futures with fixed time to expiration retains its shape. The P&L is then determined by how futures roll up or down along the curve, as time passes. In this static spot price scenario, the expected spot price is the same as the current spot price. Therefore, the first term in the decomposition (4.1) is zero:

$$RP = (E_t(S(T)) - S(t)) + (S(t) - F(t, T)) = S(t) - F(t, T)$$

and the risk premium is determined by the carry, which is the value of the spot-futures spread, as observed at time  $t$ .

The concept of systematic carry trading came from foreign exchange markets. It hinges on the same principle of no-arbitrage between two alternative money standards that Irving Fisher illustrated for the case of the commodity money, as it was explained in Chap. 2. The argument applies to fiat money of different countries as well. Each currency pays its own domestic rate of interest. The forward value of a currency with a relatively high interest rate must be lower relative to a currency with a lower interest rate to make investors indifferent in which currency to hold their money. According to an economic theory, the magnitude of the forward discount, or an expected currency depreciation, must negate the value that can be gained from interest rate differentials, the hypothesis known as an *uncovered interest rate parity*. In practice, however, this economic theory does not hold. In the long run, holding riskier currencies with higher yields is often more profitable, as on average, expected currency depreciation is not fully realized. Obviously, in the short-term such a strategy can experience large losses, but over time investors are generally able to extract positive risk premium from the foreign exchange carry trade.

Using the broad definition of carry as the strategy return when nothing changes, the idea of carry trading can be extended to almost every asset class. In equities, carry is simply given by the dividend yield. The future stock price is then discounted relative to the current stock price by the amount of the expected dividend, net of the

risk-free rate. The carry trade in the equity market buys single-stock futures that have larger discounts relative to the current stock price, which is equivalent to buying high-dividend stocks. The dividend yield behaves like the foreign interest rate in the currency market, or like the convenience yield of holding oil inventories. In the fixed income market, carry is simply the coupon received from holding the bond, or the roll down in the interest rate swap curve. The fixed income carry strategy then buys bonds with higher coupons and steeper interest rate curves, where short-term rates are typically lower than long-term rates.<sup>7</sup>

The power of carry in the oil market has already been shown in the previous chapter. When carry is positive and the oil market is in backwardation, buying oil futures tends to be a profitable strategy. As the futures curve moves into contango and the carry becomes negative, the investor returns deteriorate. Since, in the futures market, selling is as easy as buying, one can replace the long-only futures investment with the dynamic strategy of buying and selling based on the direction of carry. Like momentum, the same idea can be applied across all commodities to achieve some diversification, but the strategy again works better for energy commodities. Oil carry is also rooted in the theory of storage. Unlike momentum, however, the transmission of the carry theory to the futures market is much more explicit. An oil carry plays a unique role in the market as it transforms the fundamentals of supply and demand into pressure on prices via the behavior of inventory hedgers.

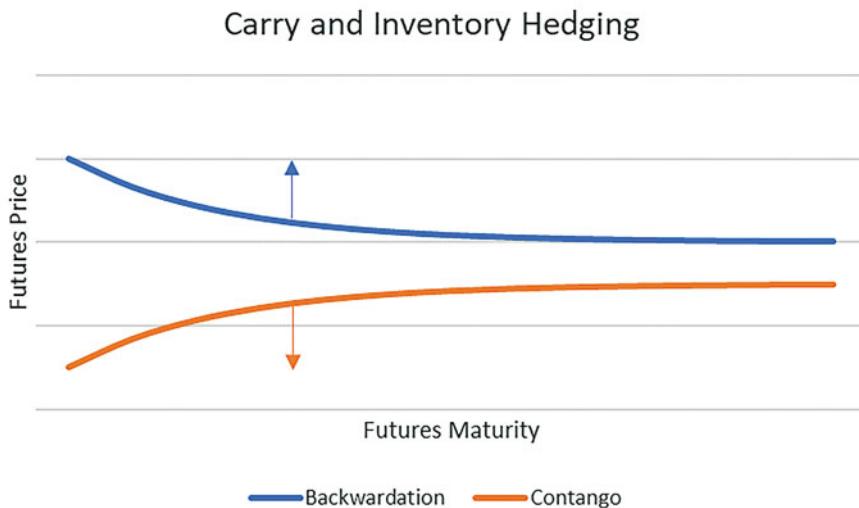
The business model of the storage trader is to generate low-risk steady returns by providing the service of storing oil. In such a business model, the risk of volatile oil prices must be eliminated by hedging in the futures market. As soon as a physical barrel is purchased for storage, a financial barrel is sold in the futures market to reduce the risk. This downward pressure on futures price occurs when the contango is steep enough to cover the cost of storage. If the market flips to backwardation or the contango becomes too narrow to economically hold inventories, then the storage hedger is incentivized to pull oil out of storage and buy back short futures hedges. This creates an upward pressure on the futures price. The carry transmission channel is illustrated in Fig. 5.5.

Such buying and selling of futures by inventory hedgers driven by the shape of the futures curve is rather mechanical and somewhat predictable. The cost of carry, which includes transportation, logistics, and borrowing costs, and the convenience yield of holding inventories depend on the economics of individual traders, which makes hedging pressure vary with the steepness of the curve. In general, the steeper the contango, the more widespread hedging becomes among inventory managers, which increases the selling pressure in the futures market. However, when carry becomes too steep, most of the hedging is already finished, and the pressure on price starts waning. We address this important transition point in more detail in the last section of this chapter.

To define the systematic carry signal more formally, let

---

<sup>7</sup>The empirical performance of the broad cross-asset carry portfolio is presented in Kojen et al. (2018).



**Fig. 5.5** An inventory hedger sells futures when contango covers the cost of storage and buys them back otherwise

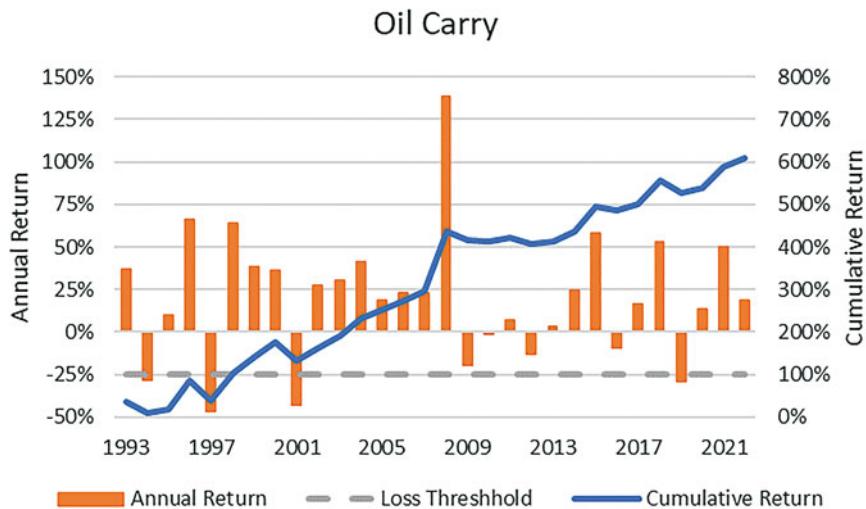
$$C_t(M, N) = F_{t,M} - F_{t,N}, \quad M < N$$

represent the spread between the two futures contracts with  $M$  and  $N$  months to expiration. The basic carry strategy buys futures when carry  $C_t$  is positive, and sells futures when it is negative, i.e., the trading signal  $\pi_c$  can be defined as

$$\pi_C(F_t) = \text{sign}(C_t(M, N)) \quad (5.2)$$

The definition of the carry signal again does not have to use the same contract that is traded. Like in the case of the momentum strategy, multiple carry signals computed from different parts of the futures curve can be combined to generate a single aggregate carry signal. However, in the case of crude oil multiple carry signals measured from different parts of the futures curve tend to have the same sign and become largely redundant. If the futures curve is monotonically decreasing or increasing, then the sign of each monthly carry signal is the same for all  $M$  and  $N$ .

Figure 5.6 shows the historical returns of the benchmark WTI carry strategy that buys and sells prompt futures contracts based on the sign of the annual carry  $C_t(1, 13)$ . The twelve-month spread is commonly used to define carry to eliminate seasonal effects, which are more substantial for refined products and natural gas. As before, the futures are rolled at the end of the month. Since both momentum and carry strategies maintain either a long or a short position at all times, the two strategies have the same volatility. The carry strategy, however, generated a much higher annualized log-return of 20.3% over the thirty-year period with impressive Sharpe ratio of 0.53. In addition, the drawdowns of the carry strategy are significantly smaller than the drawdowns of the momentum strategy.



**Fig. 5.6** Historical returns of the C(1,13) carry strategy for WTI futures (1993–2022)

Carry, or the shape of the futures curve, is the ultimate mechanism that transmits fundamental information into futures prices. When we read media reports that the price of oil went up because of stronger Asian demand, or it went down because of increased OPEC production, these reports do not tell the full story. OPEC itself does not buy futures and very few oil consumers do. However, any change in supply and demand impacts the spot-futures spread via variations in convenience yields and storage costs. The actual futures are bought and sold by storage traders, who react to the economics determined by the carry. We will see in the following chapter that an inventory hedger is indeed the largest trader in the oil futures market. The systematic carry strategy described in this section essentially trades ahead of anticipated behavior of the largest market participant, who is acting in response to the arrival of new fundamental information.

Compared with the momentum strategy, carry trading has one tremendous advantage. The strategy is essentially model-free. Unlike momentum, carry is a forward-looking signal. It does not depend on history and does not require any parameter estimation. The trading signal is directly observable from the current shape of the futures curve. It measures the economic incentives of inventory hedgers and, by and large, the strategy front-runs the expected flows of hedgers in the futures market.

The idea of oil carry is probably one of the most powerful and widely used indicators for directional oil trading. Many physical traders simply refuse to take any positions that go against the direction of carry. A similar idea of investing in oil only

when the carry is positive has found some interesting applications in broader financial portfolios.<sup>8</sup>

The primary drawback of the carry strategy is its inability to react fast enough to rapidly changing market conditions. Since the shape of the futures curve does not flip frequently between contango and backwardation, the carry signal often retains the same sign for a long time. Even though it may capture early gains quickly when the shape of the curve changes, it then often gives up a portion of the gains when the market starts correcting. By holding and rolling the same directional position until the time spread crosses zero, or some pre-defined threshold, the carry strategy often misses early warning signs that fundamentals are beginning to change. To capture an anticipated change in fundamentals, traders often look at the change in carry instead of its level. One can think about carry being a measure of the current state of supply and demand, and the change in carry as a measure of expected changes in future supply and demand.

A popular way to make the carry strategy more dynamic is to blend it with the basic momentum signal. This can be accomplished by applying momentum not to the price of oil, but instead directly to carry, or to the shape of the futures curve. The carry-momentum trading signal  $\pi_{CM}$  is then defined as

$$\pi_{CM}(F_t) = \text{sign}(C_t - MA_t(C_t; n)) \quad (5.3)$$

The results from such simple signal blending, shown in Fig. 5.7, are quite remarkable. The strategy, which we call the dynamic carry, or *carry-momentum*, generated 24.7% annualized returns with the Sharpe ratio of 0.64 over the thirty-year period. Out of hundreds of different blends of systematic signals that the author has traded over many years, the carry-momentum signal stood out in its long-term robustness. It highlights how a technical indicator, such as momentum, can improve a trading concept that has stronger links to the market fundamentals.

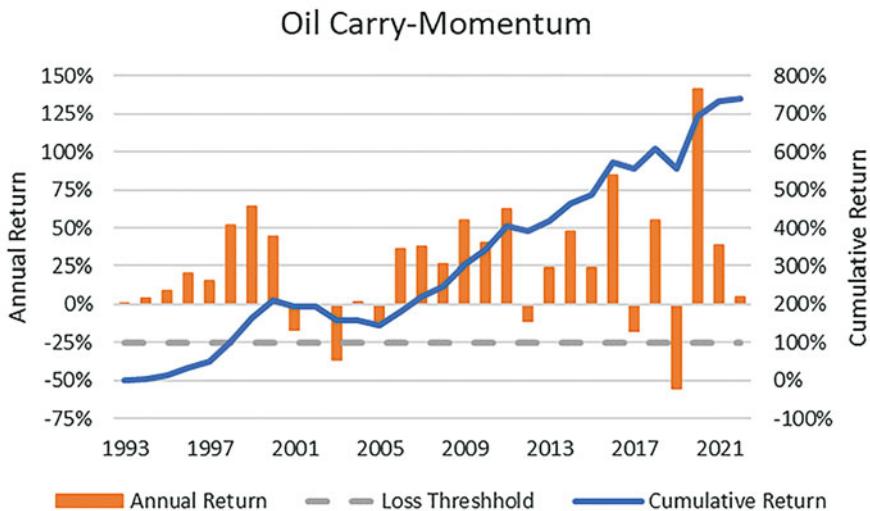
---

## 5.4 Value and Mean-Reversion

The success of momentum and carry strategies in the oil market may surprise traders who are more accustomed to a bargain-hunting value trading style. Since oil prices are cyclical and expected to mean-revert, the natural temptation for many discretionary traders would be to do the opposite of momentum, and instead attempt to buy low and sell high. The convergence strategy of betting against price deviations from some long-term equilibrium level is an example of the value risk premium. Value is a contrarian strategy where one buys what is deemed to be underpriced and sells what is overpriced relative to some price level that is considered to be fair.

---

<sup>8</sup>Till (2022) showed that a dynamic strategy that replaces a conventional 60–40 equity-bond portfolio with a 30–40–30 equity-bond-oil futures portfolio when oil carry is positive significantly outperforms the static 60–40 portfolio.



**Fig. 5.7** Historical returns of the 20-day carry-momentum  $CM(1,13)$  strategy for WTI futures (1993–2022)

Establishing the fair value for oil is difficult. Fundamentally, it only makes sense for contracts with long maturities where the fair price can be approximated by the marginal cost of production. Short-term futures can then be viewed as the spread to the long-term fair value. However, the volatility of this spread driven by short-term fluctuations in supply and demand is too high for any meaningful fundamental definition of the short-term fair value. As it was demonstrated in Chap. 3, oil price can rise or fall without any limits when inventories approach zero or reach the maximum level of the storage capacity. In Chap. 7, we will take a different approach and construct another example of the fair value of oil by calculating it as a function of macroeconomic variables.

In practice, one often estimates the fair value statistically by comparing the current price to its moving averages, like in the definition of momentum. The purpose of moving averages is to incorporate ongoing structural changes that constantly occur in the oil market. In its simplest form, one can define the value trading signal  $\pi_V$  as the mirror image of the momentum indicator, such as

$$\pi_V(F_t) = -\pi_M(F_t) = -\text{sign}(M_t(n))$$

This strategy buys futures when the price falls below its moving average and sells when it rises above. Obviously, since the momentum strategy is generally profitable, systematically running the opposite value strategy would be a loser.

To capture the value risk premium, traders usually run their strategies on slower frequencies. While it is critical for momentum traders to react fast and trade as soon as the price crosses some trigger, value traders generally prefer to remain patient and avoid rushing into a trade. They often let the market run in the same direction for

some distance and only trade when the deviation exceeds a certain threshold,  $\varepsilon$ . Such a threshold-based value signal can be defined as:

$$\pi_V(F_t) = \begin{cases} -1, & \text{if } M_t(n) > \varepsilon \\ +1, & \text{if } M_t(n) < -\varepsilon \\ 0, & \text{if } |M_t(n)| \leq \varepsilon \end{cases} \quad (5.4)$$

In this variation of the value strategy, the trader may remain on the sidelines for a while and bet only when the deck is rich, i.e., when the price deviates sufficiently far away from its normal range.

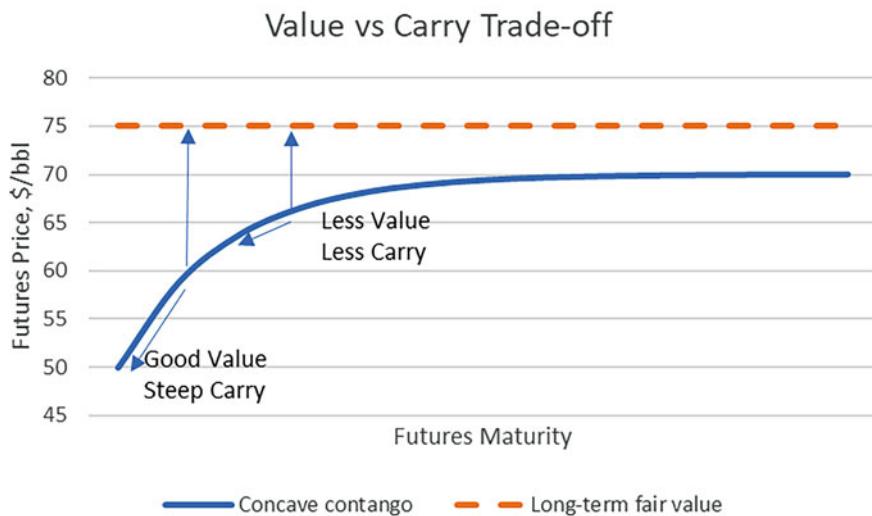
In general, calling the tops and the bottoms for the price of oil is not a prudent strategy in the futures market, as one must be very precise on the timing of anticipated price reversals. If one could trade physical oil and keep some oil in storage without incurring any cost, then selling high and buying low would undoubtedly work well, as spot price does eventually mean-revert. However, storage is not free, and it is rarely accessible by financial traders, so that the closest proxy that one can trade is the futures contract. Buying low and selling high in the futures market, however, is very different from doing it in the spot market. In the futures market, the value trader constantly fights against the punitive cost of carry, which is determined by the slope and the convexity of the futures curve.

If the market is oversupplied, then the price of prompt futures is likely to be lower than the forward price. The slope of the futures curve reflects the cost of storage, which is typically higher for shorter-maturity futures, where the fundamental imbalance is the most acute. Storage buys time either for consumption to increase or for production to be curtailed. As time moves forward and fundamentals normalize, the cost of storage decreases, and consequently the slope of the futures curve flattens. Since the slope, which is the first derivative with respect to futures maturity, is decreasing, then the futures curve in an oversupplied market is expected to be concave. We call this a state of *concave contango*.

Likewise, at times when inventories are low, the slope of the futures curve is the steepest in the front-end of the curve, which is dominated by the highest convenience yield of owning a physical barrel. As inventories normalize over time, the convenience yield decreases, which makes the forward slopes flatter. This is a state of *convex backwardation*. The futures curves that are shown in Fig. 5.5 are in their normal states of concave contango and convex backwardation.

The convexity of the futures curve creates a dilemma for the value trader: whether to use shorter-term futures and lock in better value at the expense of a steeper carry, or trade further out on the curve where the carry is less punitive, but the value is also less attractive. Figure 5.8 illustrates the challenge of buying cheap futures when the market is in concave contango.

The case of selling futures in convex backwardation is handled similarly, as the mirror image of concave contango. In either case, for the value strategy to be profitable the magnitude of the expected price mean-reversion must dominate the accumulation of carry over the holding period. One should avoid using short-term futures for capturing value unless the reversal is deemed to be imminent as indicated



**Fig. 5.8** Buying cheaper short-term futures in concave contango is offset by steeper carry, while buying long-term futures has less value

by other fundamental factors. Otherwise, time is not on the side of the value trader. The apparent cheapness of the futures contract could quickly become overpowered by the adverse buildup of the negative carry. These types of strategies where a systematic signal is best combined with a discretionary overlay based on fundamental and flow factors are discussed in the next chapter.

In theory, it is possible for both systematic momentum and value strategies to generate positive returns if the two are traded on different frequencies. Momentum tends to work better on the shorter time scale, while the contrarian value signal works better in the longer run, especially when an additional buffer is provided by a threshold  $\varepsilon$ . One can also combine momentum, carry and value into signal-integrated portfolios. This is typically done on a cross-sectional basis within broader commodities portfolio.

Overall, momentum and carry signals tend to overpower value in directional oil trading, but in the next chapter we will show that value strategies are more robust in trading futures spreads and construct one of such diversified energy value portfolios. For now, we continue to focus on single-asset strategies and conclude this chapter by showing another way of blending multiple signals that can help systematic traders to determine the optimal size of their bet.

## 5.5 The Reaction Function

The simplest way to size positions within broad systematic portfolios is to apply so-called *volatility targeting*. Similarly to the risk parity strategy, every asset or strategy in the systematic portfolio is often allocated the same amount of risk. The risk is measured as the product of the notional size of the strategy and its volatility. Therefore, to keep the dollar risk the same across all assets, the size of the position must be inversely proportional to volatility. If the asset volatility increases, then the risk capital is withdrawn and the position must be reduced. Such a forced liquidation, however, is suboptimal as it does not take into consideration the strength of the systematic signal.

More advanced systematic traders incorporate not only volatility but also the magnitude of multiple signals and combine them to determine an optimal position size. To illustrate this for a single commodity, one can scale the position by blending various risk premia, such as momentum and value. While value is unlikely to be an attractive directional strategy on its own, it can be used to optimize momentum and carry signals. The idea is for the value metric to act as a control check to indicate that momentum and carry might have gone too far in one direction and the status quo is unlikely to remain for much longer.

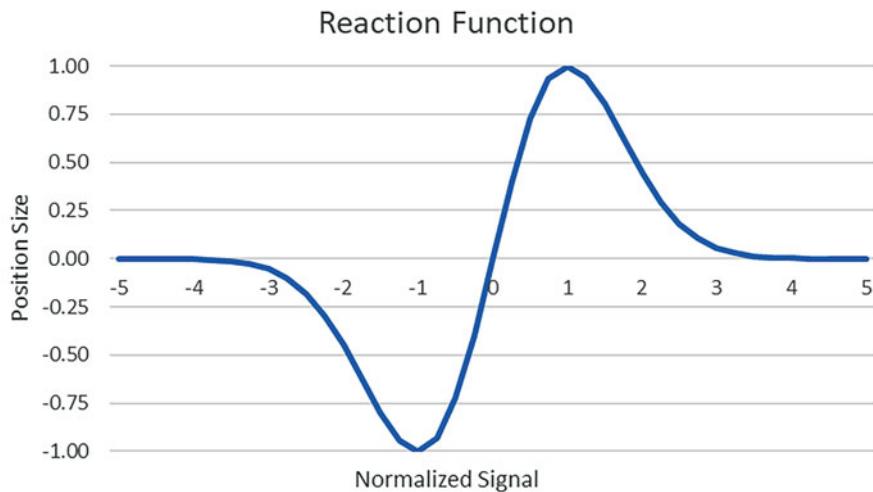
In Chap. 7 we will discuss the fact that this dynamic is typical for a complex dynamic system. In such a system the presence of natural boundaries eventually limits its growth, and the value signal in a systematic trading strategy is akin to measuring the distance from the system fundamental boundary. If momentum or carry signals become too strong, then the market might have already moved too far outside of its normal range, and the likelihood of a price reversal increases. In such cases, it would be prudent to reduce the size of the position inversely to the strength of the momentum or carry signals.

To quantify the blending of directional momentum and carry signals with value, we use a so-called *signal transformation* or *reaction function*. This is a function that maps the strength of the systematic signal to the size of the bet. This function typically increases along with the strength of the signal but only up to a certain point, beyond which any further position increases are no longer justified. We call this an *inflection point*. For example, for the momentum signal such an inflection point may indicate that the trade is too crowded, and the momentum has high probability to subside, as everyone who wanted to be in this trade is already in this trade.

The reaction function can be specified parametrically. Figure 5.9 provides one example of such a function defined as

$$R(x) = xe^{0.5(1-x^2)} \quad (5.5)$$

The dependent variable in Fig. 5.9 is some metric of the strength of the momentum. In diversified systematic portfolios the metric is often normalized. This makes it easier to compare signals across different markets and to ensure more balanced risk



**Fig. 5.9** An example of a reaction function for the momentum strategy

deployment for assets with different volatilities. Without such normalization the strategy performance will be dominated by more volatile futures. The standard way to normalize systematic signals is to divide raw signals by the standard deviation of the asset returns. The resulting normalized metric is known as the *z-score*, which represents the strength of the trading signal in terms of its number of standard deviations. In the reaction function above, parameters were chosen to set the maximum position of  $\pm 1$  when the normalized signal is equal to  $\pm 1$ .

Many other reaction functions are used by systematic traders, but the general idea is to grow the position size up to a certain maximum level typically defined by the trader's risk limit, and then gradually reduce it when the signal becomes too strong. This avoids taking excessive risks when prices move too far away from their normal range and the likelihood of a reversal increases. Some traders prefer to use piecewise linear reaction functions that allow for the fixed-size maximum position to be held for a wider range of the signal. The reaction function can be easily parametrized to optimize its slopes and widths, but one must be careful not to overfit parameters, especially if they are fitted to a single time series. It is also easy to modify a reaction function that changes its sign at the extremes when the momentum strategy is replaced with a mean-reverting strategy. This would effectively blend momentum and value into a single strategy.

Similar reaction functions can also be applied to carry and carry-momentum signals and these strategies can be optimized by using a reaction function with an inflection point. In the case of carry, the existence of an inflection point is consistent with the behavior of inventory hedgers. When the negative carry is too large, then most of the storage is already full and there is nothing left for inventory traders to hedge. Likewise, when the backwardation is extreme, most short futures held by inventory hedgers are already covered. Therefore, beyond a certain threshold the

pressure on price from storage hedgers starts waning and other corrective mechanisms may play a larger role, causing the market to mean-revert. For example, extreme contango could force producers to shut down drilling, and extreme backwardation may lead to product substitution among consumers. The reaction function attempts to quantify an important rule of thumb in oil trading, where short-term momentum is better combined with mean-reversion.

So far, we have only considered trading signals that are based on the information contained in prices, and all strategies were designed to be entirely rule-based and executed systematically without any human intervention. As we have already seen, the power of good trading often lies in blending signals of different natures, and also in letting humans contribute to the decision making. In the next chapter, we look at how some systematic concepts can be improved with non-price fundamental and flow information, which require additional inputs from the discretionary trader.

---

## References

- Bakshi, G., Gao, X., & Rossi, A. G. (2019). Understanding the sources of risk underlying the cross section of commodity returns. *Management Science*, 65(2), 619–641.
- Boons, M., & Prado, M. P. (2019). Basis-momentum. *The Journal of Finance*, 74(1), 239–279.
- Daskalaki, C., Kostakis, A., & Skiadopoulos, G. (2014). Are there common factors in individual commodity futures returns? *Journal of Banking and Finance*, 40, 346–363.
- Fernandez-Perez, A., Frijns, B., Fuertes, A.-M., & Miffre, J. (2018). The skewness of commodity futures returns. *The Journal of Banking and Finance*, 86, 143–158.
- Koijen, R. S. J., Moskowitz, T. J., Pedersen, L. H., & Vrugt, E. B. (2018). Carry. *Journal of Financial Economics*, 127, 197–225.
- Lux, H. (2003, February 1). What becomes a legend? *Institutional Investor*.
- Miffre, J. (2016). Long-short commodity investing: A review of the literature. *Journal of Commodity Markets*, 1(1), 3–13.
- Szymanowska, M., De Roon, F., Nijman, T., & Van Den Goorbergh, R. (2014). An anatomy of commodity futures risk premia. *The Journal of Finance*, 69(1), 453–482.
- Till, H. (2022, Winter). Commodities, crude oil, and diversified portfolios. *Global Commodities Applied Research Digest*, 7(2), 65–74.
- Tully, S. (1981, February 9). Princeton's rich commodity scholars. *Fortune*.
- Weymar, F. H. (1965). *The dynamics of the world cocoa market*. Ph.D. Thesis, Massachusetts Institute of Technology.



- Trading time spreads is an example of a quantamental trading style where quantitative curve signals are combined with fundamental information about inventories and financial flows.
- A discretionary overlay is particularly important in identification of fundamental regimes perceived to be favorable for the performance of systematic signals. One important case study is a regime-dependent WTI-Brent convergence strategy.
- Many energy assets represent a real option on the output-input spread. Asset owners hedge profit margins, which induces spread mean-reversion. A portfolio of mean-reverting energy spreads is akin to a statistical arbitrage strategy.
- The analysis of flows and positioning is another important strategy overlay. It is effective in identifying crowded trades that can be used as a reversal indicator to optimize position sizing and risk management.

---

## 6.1 Trading Curve and Convexity

The systematic risk premia presented in the previous chapter are based on rather generic concepts that can be applied across many different markets and traded within a broader portfolio of commodity futures. While risk premia strategies have proven to work better for harder-to-store energy commodities, adding other commodities to the portfolio brings a much-needed diversification that reduces the overall risk of the strategy. In such portfolios, distinct properties of individual commodities are largely ignored. In this chapter, we focus instead on important idiosyncratic features of oil price behavior and show how a generic cookie-cutter strategy can be turned into a more elaborate oil-specific bet.

Many oil traders do leverage the ideas of risk premia, but applying these strategies to a single asset without the diversification benefits of a broader portfolio would be too risky. To make the strategy viable on its own, systematic price-based signals are often enhanced with additional information about market fundamentals and flows. Fundamental and flow data, however, are much noisier, more difficult to

systematize, and better left to a discretionary interpretation. In such trading, a human equipped with a machine tends to do better than what either a machine or human can achieve alone. An application of quantitative signals to fundamental and flow data, or, more generally, any strategy based on quantitative signals with a fundamental overlay, is described as *quantamental trading*.

It may come as a surprise, but professional oil traders do not like to speculate on the price of oil. The oil price is driven by too many factors that are difficult to get a handle on. In the short run, the price does not have any well-defined boundaries, and the cost of being wrong could be catastrophic. To keep the risks contained, oil specialists instead focus their attention on trading futures spreads. The spreads are constructed to isolate the desired exposure by eliminating a portion of the risk that the trader is not willing to bet on. While managers of broad portfolios reduce risk by diversification, oil traders rely on trade structuring and supplementary fundamental information for risk mitigation. Spreads are significantly less volatile than outright prices, and to generate the same amount of profits, spread strategies must be leveraged and traded in larger volumes. A very significant portion of the trading volume in oil futures represents components of various spread-trading strategies.

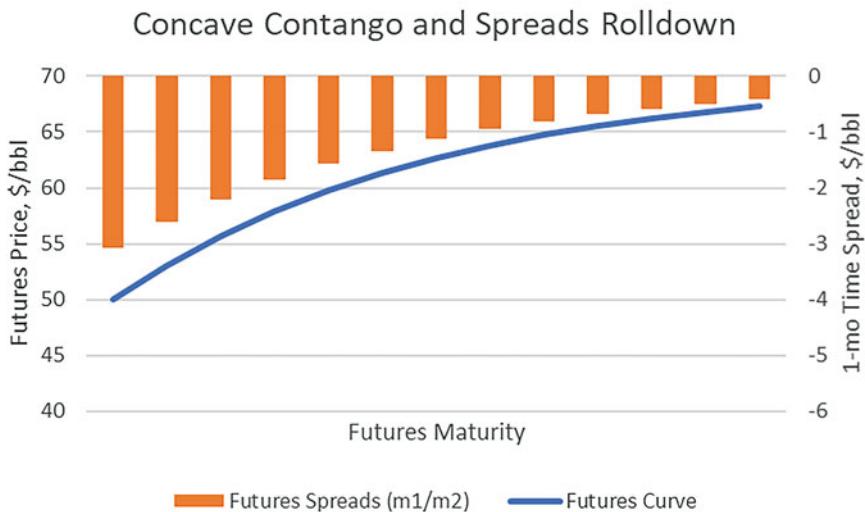
The single most important and the most widely traded spread in the oil market is the spread between the first- and the second-nearest maturity futures. This is the spread that determines the own rate of interest, the convenience yield, the roll yield, and the commodity carry. In the previous chapter, we have used the spread as an indicator for directional price trading. Now, we consider the spread itself as a trading instrument. Trading short-term time spreads is not a trivial task. It is driven by a complex interplay between fundamentals of supply and demand in the physical market and financial flows in the futures market, as traders must roll their positions before the expiration of the contract. The previously discussed episode of negative oil prices is perhaps the most famous example of such an interplay.

Much as price traders use the slope of the curve for decision making, curve traders often start their analysis by looking at the convexity of the futures curve. The idea is to use the same concept of carry but to apply it in a higher dimension. *Convexity* is viewed by practitioners as the spread between two adjacent futures spreads. For example, for the first three nearest maturity contracts, the convexity can be defined by

$$CV_{1-2-3,t} = (F(t, T_1) - F(t, T_2)) - (F(t, T_2) - F(t, T_3)) = F(t, T_1) - 2F(t, T_2) + F(t, T_3)$$

In other words, convexity is simply the carry of the carry.

As discussed in the previous chapter, in the absence of any hedging distortions, the natural shape of the futures curve is either *concave contango* or *convex backwardation*. This follows directly from the theory of storage. The role of storage is to smooth out short-term variations in supply and demand and to dilute the magnitude of the fundamental imbalance by shifting some of its burden forward. Storage literally buys time. Therefore, nearby time spreads that bear the strongest



**Fig. 6.1** If the futures curve is in concave contango then time spreads are expected to roll down

immediate effect of fundamental imbalances are likely to be wider than forward time spreads, as the futures curve is expected to gradually normalize.

Figure 6.1 illustrates monotonicity of time spreads for the futures curve in the state of concave contango. Similarly, in the state of convex backwardation, which would be the mirror image of Fig. 6.1, time spreads are positive and monotonically decreasing.

Since carry, or the slope of the futures curve, has proven to be useful in predicting the direction of the price, one can speculate that carry of the carry, which is the curve convexity, could be relevant for the dynamics of the curve slope. In other words, if the first derivative of futures with respect to maturity has some predictive ability for the price level, then the second derivative may be useful in predicting the direction of the slope. For example, if the futures curve is in a state of concave contango, as pictured in Fig. 6.1, then the spread carry strategy is to sell time spreads in anticipation of them rolling down the curve. Likewise, if the curve is in a state of convex backwardation, then the spread carry strategy is to buy time spreads, which are expected to roll up.

Another way to think about the convexity-trading mental model is to assume that the fair value of forward time spreads is determined by the spot spread, which contains the most up-to-date information about the current state of inventories. Since inventories are slow to adjust, forward spreads are expected to roll towards the value of the spot spread. Therefore, it might be beneficial to buy forward time spreads if they trade below the prompt spread and to sell forward time spreads if they are above the prompt spread. It should be noted that such a simple formulation of the convexity strategy can only apply to commodities that do not exhibit strong seasonality, such as crude oil.

Even though convex backwardation and concave contango can be viewed as two fundamentally normal states of the futures curve, in practice, the curve is distorted by imbalances between buyers and sellers in the futures market. The largest imbalance in the spot time spreads is caused by financial investors holding large quantities of the nearest-maturity futures, which are held as a substitute for owning the physical commodity. Every month, prompt futures positions must be rolled ahead of their expiration, which is done by selling the time spread. To the big disadvantage of commodity investors, exact roll schedules for main commodity indices and exchange-traded funds (ETFs) are publicly available and well known to the market. Such transparency incentivizes professional traders to sell time spreads ahead of anticipated rolls and buy them back when spread values are expected to weaken from the negative rolling pressure.

When the market is oversupplied and is in a state of contango, then the pressure from investor rolls tends to exacerbate the degree of the curve concavity. In fundamentally weak markets, the combination of such rolls and storage capacity constraints can lead to the situation known as a *super-contango*, when the prompt futures contract dislocates and falls substantially below the contract with the next maturity.

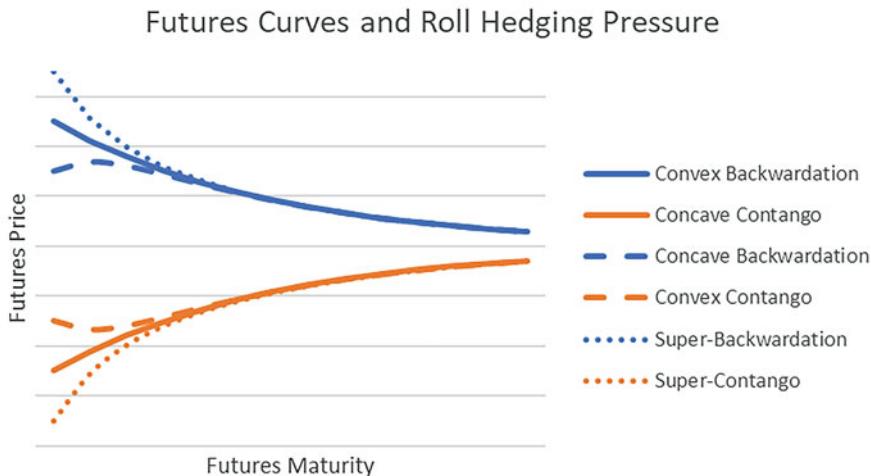
In the case of backwardation, the rolling pressure on the prompt spread may result in a humped futures curve, which corresponds to the case of *concave backwardation*. The hedging pressure on spreads can also be positive, which occurs when inventory hedgers and refineries roll their short futures positions. This can lead to the scenario of *convex contango*. However, such episodes occur rather infrequently because commercial hedgers are not driven by any fixed roll schedule, which gives them more flexibility on when to roll. In fact, many refineries and storage hedgers view rolling their short futures as a way to take advantage of the more rigid schedule that must be followed by long futures holders, such as commodity index investors. Finally, the scenario of *super-backwardation* arises during unexpected supply disruptions when refineries have to pay up for spot barrels, overwhelming the negative hedging pressure from investor rolls.

Various shapes of the futures curve are illustrated in Fig. 6.2. Note that here we define convexity locally by using the shape of the front-end of the futures curve, while describing backwardation and contango by comparing the spot price to long-dated futures.

The structural hedging pressure caused by investor rolls makes the returns of buying and selling short-term WTI time spreads highly asymmetric. The odds of making money are generally higher by systematically selling prompt WTI time spreads than buying them. To illustrate, consider a simple strategy of selling the spot spread and rolling it to the next spread five business days prior to expiration.<sup>1</sup>

---

<sup>1</sup>Many financial speculators are required to roll their position at least five days prior to the expiration of the futures contract to avoid regulatory position limits and risks of physical delivery. The static short spread strategy is relatively insensitive to the exact rolling schedule, provided that a short position is entered prior to the scheduled index rolls.



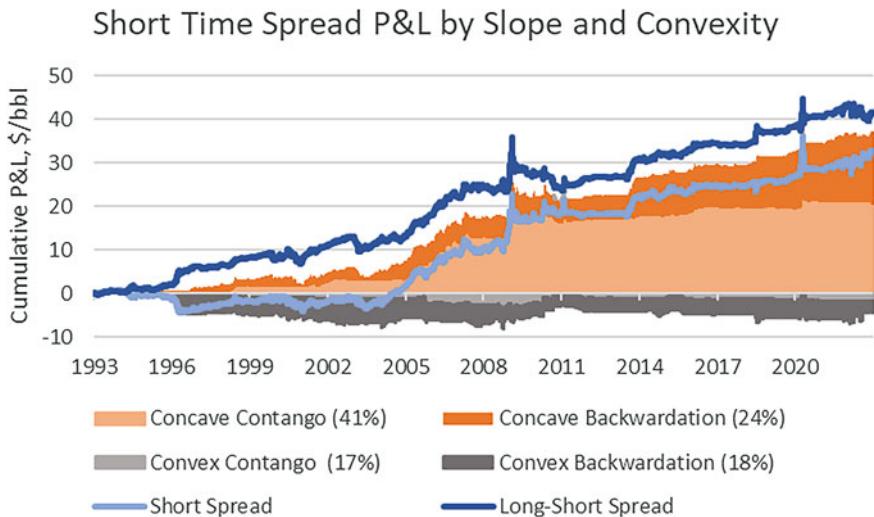
**Fig. 6.2** The impact of roll hedging pressure on the shape of the futures curve

Over the thirty-year period starting from 1993, such a naïve strategy would have generated a Sharpe ratio of 0.50. The strategy would have done even better if one started it in 2005, which we previously earmarked for the start of the regime of normal contango. Not surprisingly, the strategy of selling WTI time spreads became the darling of Wall Street. It was packaged into investable indices and sold to investors as an alternative risk premium specific to WTI market.

To illustrate the likely source of such an incredibly strong bias for the static strategy, we decompose the historical P&L of the strategy by the shape of the futures curve. Figure 6.3 splits the cumulative rolling strategy P&L into four constituents, determined by the slope of the futures curve and its convexity.

The decomposition of P&L with respect to a certain explanatory variable is often referred to as *fractionation analysis*. For the short time spread strategy the fractionation analysis reveals that strategy profits were generated mostly when the curve was concave, either when it was in its normal state of concave contango, or in a state of concave backwardation where flow distortions lead to a humped curve. A significant contribution to positive P&L during a somewhat abnormal state of concave backwardation indicates the importance of financial rolls, the phenomenon that accelerated since the beginning of financialization. In contrast, when the curve was convex, then selling time spreads was a losing trade, irrespective of contango or backwardation, but the losses were relatively small. It is interesting to see that it is the convexity of the curve rather than its slope is the main driving force for time spreads.

These observations can be combined into a dynamic long-short spread strategy based on the curve convexity. Let us define the prompt time spreads as  $S_{1-2,t} = F(t, T_1) - F(t, T_2)$ . Then the trading signal can be specified as



**Fig. 6.3** Decomposition of short time spread strategy by slope and convexity. The number in parenthesis indicates the percentage of time that the futures curve was in a given regime

$$\pi(S_{1-2,t}) = \text{sign}(CV_{1-2-3,t})$$

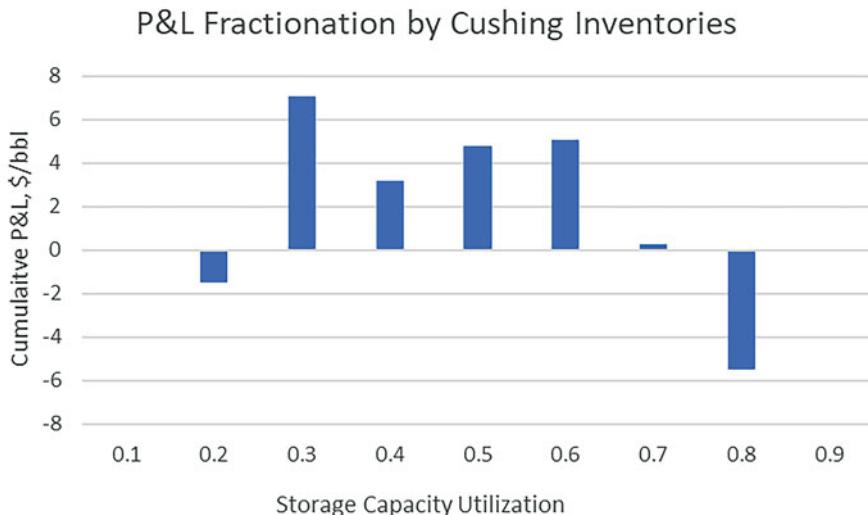
and the spread carry strategy is to sell and roll the prompt spread when the curve is concave and to buy it when the curve is convex.

This long-short strategy further improves the cumulative strategy performance, mostly from having long spread positions, prior to the beginning of financialization. Since 2005 nearly all profits from the spread carry strategy came from selling spreads when the futures curve was concave. Is the strategy performance then entirely driven by financialization, or some other fundamental factors, such as rapid growth of US shale production, also played a role in the structural rolldown of WTI time spreads? To assess the impact of fundamental factors, we next look at another important example of fractionation analysis and decompose P&L of the time spread strategy by the level of inventories.

## 6.2 Time Spreads and Inventories

An analysis of time spreads behavior versus inventories is motivated by the theory of storage, presented in Chap. 3. In that chapter, we have only looked at *contemporaneous* correlations between time spreads and inventories. Now we address the more difficult question of whether the level of inventories have any *predictive power* for the direction of time spreads.

We fractionate P&L of the same short WTI time spread strategy by the level of Cushing inventories. As discussed in Chap. 3, one can also use PADD2 and overall



**Fig. 6.4** P&L of short WTI time spread strategy (2011–2022) split into deciles based on utilized storage capacity in Cushing

US inventories as alternatives. In practice, traders tend to look at all three inventory metrics to get a broader assessment of the fundamental state of the market. In this analysis, we compare the levels of Cushing inventories, which are reported weekly, to P&L of the strategy during the following week. As before, we use the normalized storage capacity utilization as our preferred measure of inventories to better assess proximity to the upper storage boundary.<sup>2</sup> The results are shown in Fig. 6.4.

At first glance, these results may look surprising. There is no visible correlation between forward P&L of the short time spread strategy and the previously observed level of inventories. The conventional theory of storage only explains contemporaneous correlation between time spreads and the level of inventories. However, contrary to a common perception, neither inventory level nor recent changes in inventories show any strong predictive power for the direction of time spreads.

In fact, a deeper econometric analysis shows that the reverse is true, and time spreads have some predictive power for the future change in inventories.<sup>3</sup> This happens because storage traders often lock in contango for several months ahead when the steepness of the curve exceeds the net cost of storage for a longer time period. Consider a trader who buys a heavily contangoed spread between one-month and six-month futures, then takes a physical delivery on the long futures contract and keeps oil in storage for six months. These barrels increase oil inventory next month

<sup>2</sup>The choice of this metric does shorten the lookback period as Cushing capacity data is only available since 2011. However, the conclusions are substantially similar if the analysis is repeated for longer lookbacks using private estimates for the storage capacity prior to 2011.

<sup>3</sup>See Ederington et al. (2021).

and decrease inventory in six months, when oil is pulled out of storage and delivered against a short futures contract. Therefore, the storage decision, which is made based on the slope of the futures curve today, affects the level of inventories six months from now. Moreover, the level of inventories in six months depends on all prior storage decisions made based on the prevailing shape of the futures curve.

P&L fractionation by inventory level does highlight an important feature of the spread strategy, the impact of inventory boundaries. When inventories drop to a low level then selling time spreads becomes particularly dangerous as the probability of an upside squeeze increases. In such environments, the strategy generated a small loss, shown in the leftmost bucket of Fig. 6.4. A much larger loss in the case of high inventories is nearly exclusively explained by an isolated event that occurred when oil price and time spreads recovered from super-contango levels set at the onset of the Covid-19 pandemic. In this episode, discussed in detail in Chap. 3 and illustrated by Fig. 3.11, the market was extrapolating a trend in inventories that would have breached the maximum available storage capacity. Since the boundary on storage capacity cannot be violated, supply and demand have no choice but to adjust, and the inventory trend must stop and reverse. This boundary phenomenon resembles the structure of the reaction function, presented in the previous chapter. The trend in inventories and time spreads can continue up to a certain inflection point, beyond which the presence of the hard boundary induces the reversal.

One lesson that a quantamental trader can learn from the fractionation analysis is that it might be better not to trade the time spread strategy when Cushing inventories approach either boundary. The proximity of boundaries clearly introduces additional uncertainty and makes spread trading particularly risky. Unfortunately, the useful insight obtained from fractionation analysis is sometimes abused by systematic traders, who are often driven by an overarching desire to improve the historical backtest. Imagine a conditional systematic strategy which is short a time spread for structural reasons but the strategy switches to a long spread position when inventories are less than 20% or more than 80% of the storage capacity. Obviously, such a rule-based strategy conditioned on inventories produces a substantial improvement to the backtest, as the two negative buckets in Fig. 6.4 are flipped from losses to gains. If the purpose of this book was to sell a systematic strategy to an investor, we might have even presented an optimized backtest for such a systematic strategy. Our goal, however, is to describe how the sausage is made, which is what the fractionation analysis reveals.

The fractionation analysis highlights the danger of data mining and the sensitivity of any systematic strategy to additional parameters, much like an optimized momentum crossover strategy. For a quantamental trader, however, such analysis is a valuable tool for decision making. It confirms the structurally short bias of the naïve static strategy that results from the joint impact of financial rolls and the growth of US shale production. When the two forces work in tandem, the spreads steadily roll down. However, when inventories decrease towards particularly low levels, then fundamental and flow forces pull the spread in opposite directions. During such periods, the strategy performance deteriorates. If a trading strategy makes money when two primary driving factors point in the same direction, but does

not lose much when factors diverge, then it is definitely an attractive strategy. In addition, the analysis emphasizes the important role of storage boundaries.

In this section we only illustrated the curve convexity and inventory analysis for WTI futures. While similar techniques can be used for other energy futures, the strategy implementations are more nuanced, which makes it difficult to apply these concepts across broader systematic portfolios. Recall from Chap. 3 that the value of the time spread is driven by the relative probabilities of downside and upside squeezes. For example, in contrast to crude oil, time spreads for refined products have an upside skew. In the case of many refined products, fractionation of spread P&L by inventories often shows more discernible impact of zero inventory boundary, when time spreads tend to appreciate. The downside for time spreads on refined products is more limited as refineries usually can cut product runs quickly in response to rising inventories and declining profit margins. The upside, on the other hand, does not have a well-defined boundary due to potentially unlimited convenience yield if the market runs out of an essential product during its peak consumption period, such as heating oil during cold weather. The statistical analysis of refined products inventories is, however, more challenging, due to their strong seasonality, which significantly limits the quantity of relevant historical data.

The case study of inventories is only one of many tools that can be utilized by a quantamental trader. Additional insights can be gained from a traditional fundamental analysis that involves so-called *barrel counting*. Since the spreads are only contemporaneously correlated to inventories, to forecast where spreads will be in the future, one needs to know where inventories will be in the future. Such forecasting of future inventories is based on counting physical barrels of oil that are being moved by pipelines and ships and by extrapolating the data forward using estimates for production and refinery inputs. While this analysis is outside the scope of the book, a good quantamental trader would always attempt to incorporate such information in a discretionary manner to improve a rule-based trading strategy.

We next discuss another popular quantamental strategy that trades the most important locational spread between two primary oil benchmarks, WTI and Brent.

---

### 6.3 WTI-Brent Accordion

So far, we have been focusing primarily on the WTI futures contract for several reasons. This contract has the deepest price history, which is particularly useful for long-term analyses. WTI also serves as a primary anchor for the large OTC hedging market dominated by independent North American oil producers. And, perhaps most importantly, the physically deliverable nature of the WTI contract allows futures traders to convert their financial bets into physical barrels when the futures contract expires. Such conversion, which can occur in Cushing, Oklahoma, ties the pricing of the WTI contract to supply and demand for inventories in the regional storage hub.

The second equally important financial oil contract is Brent, which represents the basket of crude oils originating from the North Sea.<sup>4</sup> Unlike WTI, Brent is a waterborne contract, which allows barrels to be easily shipped anywhere in the world. Most of the world's physical barrels are priced off Brent futures. Given the global scope of this market, storage-based trading strategies are more difficult for Brent, as these strategies work better for isolated storage hubs. To trade such a strategy globally, one would need to count barrels held in storage and on ships around the world, which is rather difficult to do in real time. While many WTI strategies are linked to storage, a significant portion of Brent trading revolves around its location and the connectivity of this benchmark to other pricing locations around the world.

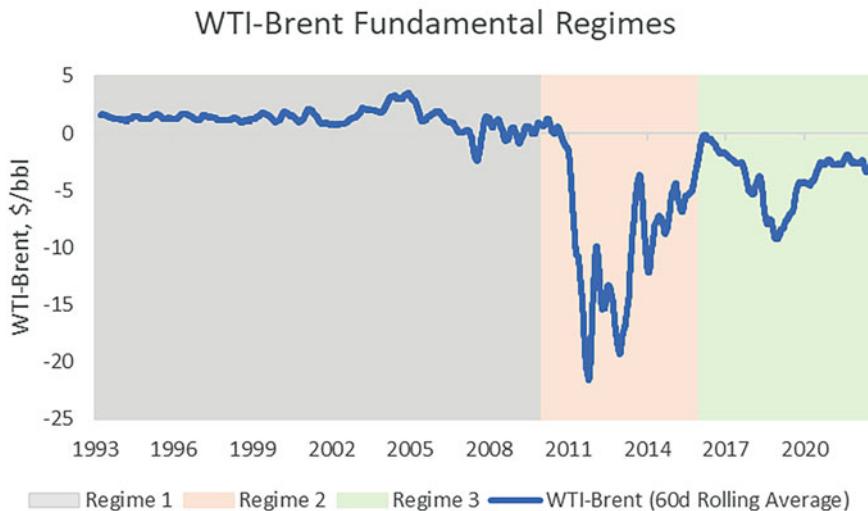
WTI and Brent are linked to each other by the economics of the transportation arbitrage. Chartering a ship or leasing pipeline space gives traders a real option to buy oil at a cheaper price in one location and sell it at a higher price at another location, provided that the spread between the two prices covers the cost of shipping. To monetize the real optionality, physical traders attempt to sell the locational spread high and buy it back low, having an ability to ship physical barrels as a backstop. Such spread-trading behavior is analogous to delta hedging of a long financial spread option, the topic that we will discuss in Chap. 13. The existence of the physical arbitrage does induce some degree of mean-reversion in the spread behavior, which prevents the two benchmarks from diverging too far from each other. This makes the WTI-Brent spread a suitable candidate for a convergence strategy, or using the terminology of the previous chapter, it is the value risk premium but applied to the locational spread.

As we have seen in the previous chapter, value strategies do not work well for the price of oil due to the persistent headwind of the negative carry, but they work much better for fundamentally linked energy spreads. In cross-product spread trading, carry is less of a factor because two forward curves tend to have similar shapes, where carry on one leg of the spread is partially mitigated by an opposite carry on the second leg. With less punitive cumulative impact of carry, one has more time to wait for convergence to occur after selling on rallies and buying on dips for the spread. Such a spread convergence strategy implemented purely in the futures market is known as a *paper arbitrage*.

The thesis of the paper arbitrage strategy is to trade ahead of an anticipated behavior of physical traders. Since the economics of shipping is generally known, a financial trader can attempt to mimic a physical arbitrage strategy in the futures market, but without the delivery of physical barrels. The financial convergence strategy generally works well as long as the physical arbitrage flows remain uninterrupted. The risk of running a spread arbitrage strategy on paper is, of course, an

---

<sup>4</sup>Brent futures are based on the so-called BFOET basket, which includes Brent, Forties, Oseberg, Ekofisk, and Troll, all produced in the North Sea. US Midland oil produced in the Permian Basin was added to the basket in 2023. For a detailed description of the price setting mechanism in the Brent market, we refer to Fattouh (2011). See also Imsirovic (2021).



**Fig. 6.5** Three fundamental regimes for WTI-Brent: the regime of the USA being a large oil importer, the regime of shale growth and logistical bottlenecks, and the regime of global interconnectedness

unexpected disruption in physical flows that can no longer keep the spread contained. While the physical trader has a hedge in place secured with physical barrels, the financial trader does not.

Spread convergence strategies of this nature, which generate steady income but with a low probability of a large loss, are difficult to trade purely on a systematic basis. The most important discretionary overlay in the spread convergence strategy is the decision of when not to trade. Spread strategies are highly dependent on a particular fundamental regime. For systematic traders, identifying and forecasting such regimes without having access to the physical market is rather difficult, as the entire petroleum infrastructure is constantly adjusting to new developments in supply and new trends in demand.

To illustrate this for the spread between the two primary oil benchmarks, WTI and Brent, we identify three drastically different fundamental regimes, illustrated in Fig. 6.5.

The first regime covers several decades of US dependency on oil imports when the country consumed substantially more oil than it produced. To incentivize non-US producers to ship barrels to the USA versus alternative destinations, WTI at that time was trading at a small premium relative to Brent. By and large, this premium reflected a relatively stable cost of shipping oil from the Middle East to the US Gulf Coast. During this period, sufficient pipeline capacity was also available to further distribute imported oil to US inland refineries for processing, or to Cushing for storage. This low-volatility price regime for the WTI-Brent spread started to

change with the rapid growth of shale production, which eventually made the USA self-sufficient.<sup>5</sup>

While US shale oil production grew rapidly, the infrastructure needed to support such growth lagged. This marked the beginning of a fundamentally new regime for WTI-Brent behavior, the regime of logistical bottlenecks, characterized by bouts of extreme volatility. As US domestic production exceeded domestic consumption, oil had to go to storage, which quickly saturated storage capacity near production centers. Moving oil to alternative storage locations around the country was also constrained by an insufficient pipeline infrastructure that could not keep up with the fast pace of production growth. Even when oil found its way from inland production centers towards the coasts, excess barrels could not be exported, as at that time crude oil exports were prohibited by US regulations. Running the financial convergence strategy during these years would have been a disaster. The spread was no longer capped by any fundamental boundaries. WTI and Brent traded as two disjoint prices with fundamental connections between US and global oil benchmarks temporarily broken.

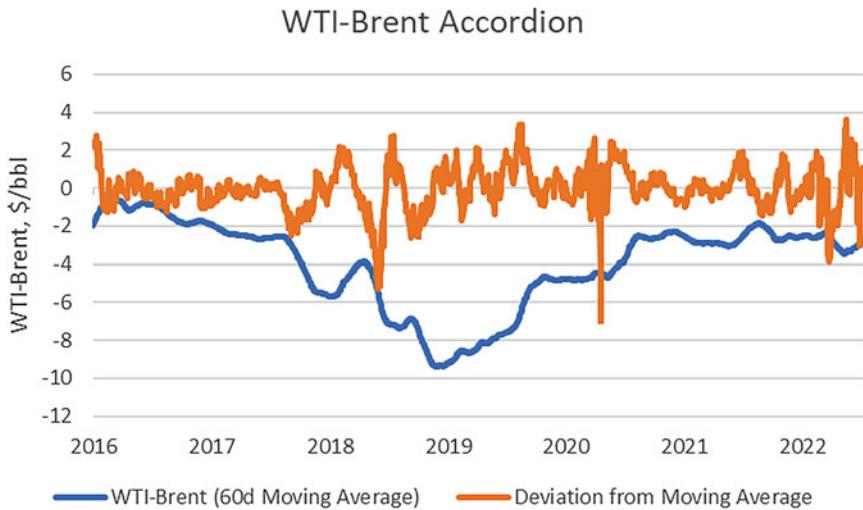
The ban on US crude oil exports was lifted in December 2015, starting the third and the current regime of global interconnectedness. The economic link between the two primary oil benchmarks has been re-established. Moreover, exports and imports started to flow in both directions driven by the relative prices of the two regional benchmarks. In the absence of any regulatory restrictions, any abnormal deviations in the spread between the two prices beyond the cost of shipping can be corrected by physical arbitrageurs. As a result, the behavior of WTI-Brent became more oscillatory and mean-reverting. Given its constant compressions and expansions, the convergence strategy was dubbed by traders the *WTI-Brent accordion*.

The actual economics of the physical arbitrage is rather intricate. Most importantly, the cost of waterborne shipping is highly uncertain, as it fluctuates along with volatile market prices of freight futures. As part of an arbitrage trade, a physical trader must not only lock in the spread between WTI and Brent, but also fix the freight price. Moreover, the impact of regional bases must be carefully taken into consideration. Such bases reflect transportation costs between specific inland production, refining centers, and export-import hubs. For a physical trader involved in the shipping of oil, the precise calculation and hedging of all moving parts is required to secure riskless profits. A financial trader usually takes a shortcut. Instead of calculating the true fundamental arbitrage boundary, which depends on freight and regional bases, the futures trader often approximates it with a statistical boundary.

The simplest way to define a paper arbitrage strategy is to take positions against large statistical deviations of the spread from its historical norm. The signal can be

---

<sup>5</sup>The USA continues to import some oil to better match refinery needs that are optimized to run on heavier crude oil. While the USA imports heavy oil, the light shale oil is exported to less complex refineries located mostly in Asia and Latin America. With the growth of shale production, US oil exports started to exceed its imports.



**Fig. 6.6** The deviation of WTI-Brent ( $m_3$ ) from its moving average

specified using moving averages, as was done in the previous chapter for momentum strategies. Specifically, if  $S_t$  is the spread between two futures, then a financial trader could take a contrarian or value position when the spread deviates by more than a certain threshold  $\varepsilon$  from its  $n$ -day moving average  $MA_t(S_t; n)$ <sup>6</sup>:

$$\pi_V(S_t) = \begin{cases} -1, & \text{if } S_t - MA_t(S_t; n) > \varepsilon \\ +1, & \text{if } S_t - MA_t(S_t; n) < -\varepsilon \\ 0, & \text{if } |S_t - MA_t(S_t; n)| \leq \varepsilon \end{cases} \quad (6.1)$$

The purpose of the moving average is to approximate the equilibrium level of the spread, which might be changing dynamically to reflect ongoing changes in the physical infrastructure. The threshold  $\varepsilon$  is designed to function as a buffer that covers uncertainty in freight and other logistical costs. This specification means that the paper arbitrage strategy is implicitly betting on some mean-reversion in freight rates. The risk is that if the freight rate spikes, then the paper arbitrage strategy will generate a trading signal even though the physical export-import arbitrage remains closed due to the higher cost of freight. A more advanced trader can make  $\varepsilon$  explicitly depend on market prices of freight futures. The only other parameter in this strategy is the length of the lookback period,  $n$ .

Figure 6.6 illustrates the moving average of the spread for the third-nearby WTI and Brent futures along with a daily deviation from such a moving average, which

<sup>6</sup>In practice, arbitrage traders often shift one leg of the spread by one month to account for the shipping time. For example, to approximate the economics of oil exports from the USA to Europe, traders are more likely to use the spread between Brent futures that expire at time  $T + 1$  and WTI futures that expire at time  $T$ .

defines the trading signal. This simple WTI-Brent convergence strategy effectively trades the residuals by taking positions against this deviation.

This strategy has performed quite well since the beginning of the third regime of global interconnectedness, reaching a Sharpe ratio of nearly 1.0 with a fairly stable combination of model parameters.<sup>7</sup> As before, our preference is to avoid showing backtests for quantamental strategies as it only creates an urge to improve them by introducing additional conditions and filters. These filters can be identified by fractionation analysis with respect to various fundamental factors, an example of which was presented in the previous section. More importantly, these strategies thrive only during specific fundamental regimes, the identification of which is largely discretionary. Since oil regimes do change rather frequently, any backtest for a quantamental strategy can quickly become obsolete.

Certain regime features, however, are likely to be more long-lasting and even irreversible. For example, increasing global interconnectedness of energy markets is unlikely to disappear once appropriate infrastructure is built. This latest regime opened a new trading opportunity not only for WTI-Brent convergence strategies, but also for many other fundamentally linked petroleum spreads. It allowed the construction of a much broader mean-reverting spread portfolio that can diversify idiosyncratic risks associated with individual spreads. In the equity markets, such portfolios fall into the category of *statistical arbitrage*, or *stat-arb*, strategies. In the next section, we explain how to construct its analogue in the energy market.

---

## 6.4 Cointegration and Energy Stat-Arb

The mathematical foundation behind spread convergence strategies and the stat-arb portfolio is based on the concept of *cointegration*. As described in many statistical textbooks, cointegration between two time series is often compared to a person walking a dog on a leash. The two can wander around, but ultimately, they cannot deviate too far from each other and tend to move together. In more precise statistical language, a cointegration property means that a certain linear combination of time series is stationary. If the process is stationary, then it has a finite variance that does not allow it to move too far from its mean. Thus, it exhibits a mean-reverting behavior. Oil prices are not stationary, but many petroleum spreads are.

Cointegration should be distinguished from the better-known concept of correlation. Correlation is only defined for stationary variables, so in most cases applying it to the price of oil is rather meaningless. Instead, it is applied to price changes or to percentage returns, as taking the difference between prices generally makes the time series stationary. For example, measuring correlation between the spot-futures spread and the normalized inventory storage capacity, as defined in Chap. 3, is meaningful because both time series are stationary.

---

<sup>7</sup>The WTI-Brent convergence strategy and its sensitivity to model parameters are analyzed in Bouchouev and Zuo (2020).

When applying correlation analysis to price changes, one effectively eliminates the entire memory of price levels. While this might be acceptable for some financial markets, in commodities the memory contains valuable information about the long-term equilibrium price level driven by supply and demand. Therefore, cointegration is a preferred starting point for analyzing commodity spreads, as the method applies directly to price levels rather than price changes or investment returns. Only when prices are cointegrated does measuring the correlation between them become statistically meaningful.

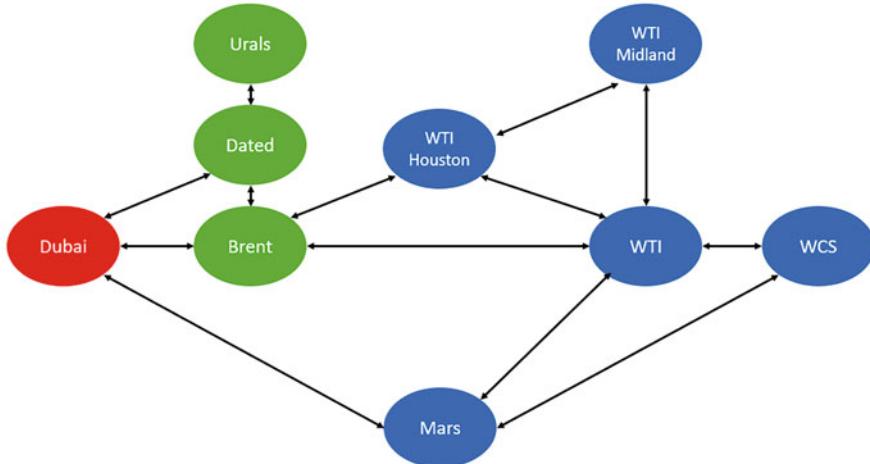
The concept of cointegration extends to the multivariate case. For example, it can be applied to the vector of petroleum futures across various geographical regions. Similarly, the vector is said to be cointegrated if there exists a linear combination of prices that is stationary or mean-reverting. In other words, if multiple assets are connected with each other via various fundamental linkages, they all tend to fluctuate around some equilibrium state for the entire system but eventually they gravitate towards such an equilibrium state. It is not our objective to provide rigorous statistical details for various cointegration tests. The topic is well covered in many statistics books, and most statistical packages have cointegration tests built-in.<sup>8</sup>

While one can indeed construct a trading strategy based on multivariate cointegration relationship among many petroleum futures that fluctuate around some natural equilibrium for the entire system, practitioners tend to simplify the problem and look instead at a portfolio of energy spreads. Like the WTI-Brent accordion strategy, many other petroleum spreads exhibit stronger mean-reversion only during certain fundamental regimes. The prevalent regime of global interconnectedness is particularly favorable for spread convergence strategies which can be combined into a broader portfolio of energy pairs. The growth of shale production created new fundamental linkages, and the removal of the US ban on crude oil exports reconnected US petroleum markets with the rest of the world. For example, shale oil spurred the rapid growth of US natural gas liquids, which became an important alternative feedstock for the petrochemical complex, competing with other refined products around the world.

The idea behind trading a diversified energy stat-arb portfolio is again based on the concept of monetization of the real optionality embedded in the physical asset. Pipeline and oil tankers are locational spread options. A refinery is an option on the spread between a basket of refined products and another basket of crude oil inputs. A storage tank is not only an option on time, but it also represents a valuable substitution option to blend different grades of crude oil. The owners of these real options are driven by economic incentives to lock in the appreciated value of an asset when the spread that drives the profit margin widens. Likewise, when the value of the spread falls, the asset owner can monetize an optionality by reducing the productive capacity of the asset and taking profits on financial hedges.

---

<sup>8</sup>For a good introductory discussion of cointegration and its application to financial markets, we refer to Alexander (2001).



**Fig. 6.7** Physical arbitrage relationships actively traded within a global crude oil portfolio

In contrast to popular equity stat-arb strategies, where the secret sauce typically lies in technical details of the trading signal, the keys to the successful implementation of the strategy in the energy market are different. For the most part, regardless of whether one uses a simple mean-reverting rule, such as (6.1), or more elaborate quantitative signals, the performance of a strategy is driven by other factors. What matters more for the energy stat-arb is the selection of what spreads to trade and when to trade them. The spreads must be carefully picked when the trading behavior of physical arbitrageurs is deemed to be significant relative to the size of the market, and when there are no impediments constraining hedging flows that keep the pair together.

As in many other quantamental energy strategies, the composition of a stat-arb portfolio is likely to change dynamically, as the energy infrastructure always evolves. However, even a snapshot example of such a portfolio constructed at the time of writing this book might be a useful starting point for its future iterations. Let us first describe the composition of a sub-portfolio made up of different grades of crude oil, which is illustrated in Fig. 6.7.

This diagram represents oil spreads that are typically traded by a physical arbitrage desk. It only shows pairs that meet certain minimum liquidity thresholds, measured by volumes and open interest, and the convergence of which is not impeded by regulatory or other restrictions. For example, one notable exclusion is a recently developed oil contract listed on the Shanghai International Energy Exchange (INE). Even though it has quickly grown to become the third most actively traded oil futures after WTI and Brent, the contract has not yet been fully integrated into global arbitrage trading, as local regulatory restrictions may preclude its convergence to other oil benchmarks.

In North America, oil grades are traded as differentials to the WTI futures contract, which is often associated with the Cushing storage hub. Unfortunately,

many traders do not realize that the reference to the name WTI in the futures contract is a misnomer. As it is clear from the name WTI, which stands for West Texas Intermediate, the contract is expected to represent oil produced in West Texas. However, oil that is delivered against the WTI contract could be very different from that produced in West Texas. Any blend of various grades of oils, whose chemical characteristics, such as density, viscosity, and sulfur content, are within a certain range, can be delivered against the futures contract. These blends are collectively referred to as domestic sweet (DSW) oil. Not surprisingly, the business of blending has become such a valuable real option for the owners of storage tanks in Cushing. If a futures buyer decides to take physical delivery of a WTI futures contract, then the exact origin of such a blended barrel delivered by the seller may not even be known to the buyer.

The price of a genuine WTI barrel is better represented by the WTI Midland contract, which explicitly references one of its main production centers. The same oil barrel also trades via the WTI Houston contract, whose price includes the transportation cost to US Gulf Coast export terminals.<sup>9</sup> The three WTI contracts at Cushing, Midland, and Houston form a fully integrated triangle of cointegrated prices with corresponding spreads fluctuating around their pipeline tariffs. Such linkages only stabilized after a sufficient pipeline infrastructure was built that connected shale production areas to storage tanks and export terminals. In addition, WTI, which represents a light sweet barrel of oil, is also linked to prices of heavy sour grades.<sup>10</sup> One such grade, Western Canadian Select (WCS), is the benchmark for oil produced in Canada that can be shipped to Cushing for blending to produce a futures-deliverable grade. Alternatively, WCS can be shipped directly to US Gulf Coast refineries, where it competes with Mars, which is the benchmark for US heavy oil produced in the Gulf of Mexico.

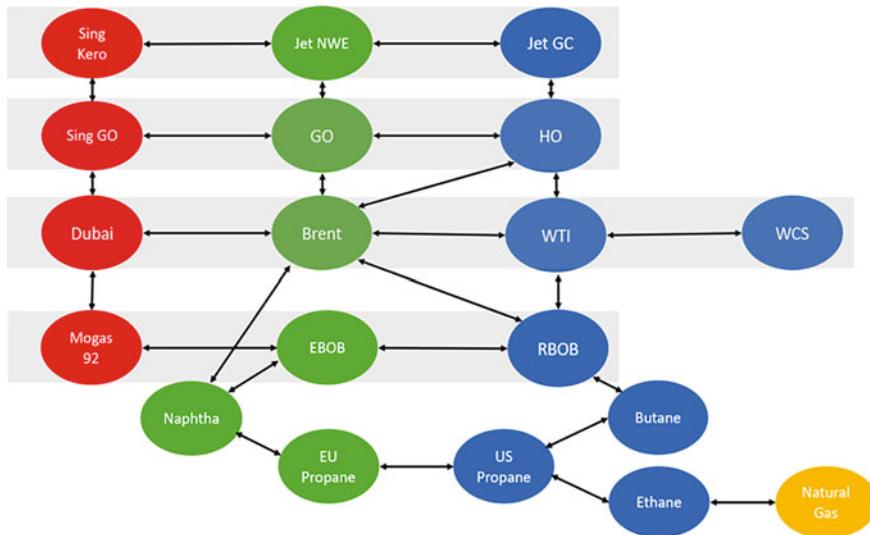
Internationally, as discussed above, WTI is connected to Brent via the transportation arbitrage. In Asia, the primary oil benchmark is Dubai, which represents another grade of heavy oil. It trades at a differential to Brent that reflects not only the cost of transportation, but also a discount for the lower quality of heavy sour oil relative to light sweet. Likewise, the WTI-Dubai spread captures both transportation and quality differentials. At the same time, Mars-Dubai represents the spread between two similar grades of oil and this spread reflects mostly the cost of transportation. Two other European oil spreads to Brent that play an important role in the global oil ecosystem include dated Brent, which differs from Brent futures in timing for the physical delivery of barrels, and Urals.

The construction of the petroleum stat-arb portfolio does not stop with crude oil. One can picture the petroleum complex as a large shopping mall, where all crude oils trade on the first floor, and refined products for each brand of crude oil trade on the second floor. One can walk on either floor and shop by product, or, alternatively, one

---

<sup>9</sup>There are, in fact, several competing US Gulf Coast WTI contracts listed by two major exchanges, CME and ICE, but for simplicity, we do not differentiate between them.

<sup>10</sup>Light and heavy oil correspond to the oil density and sweet and sour to its sulfur content.



**Fig. 6.8** Physical arbitrage relationships actively traded within a global refined-products portfolio

can shop for both crude oil and its affiliated refined product in the same branded store. On the one hand, the price of each refined product is linked to the local crude oil by the regional economics of refining. On the other hand, all refined products, like all crude oils, are also connected with each other via the transportation arbitrage. For example, US ultra-low-sulfur diesel (ULSD) impacts the profitability of US refineries, and its European equivalent, called gasoil, is crucial for the profitability of European refineries. At the same time, ULSD and gasoil are also linked via the shipping arbitrage, as they are comparable refined products. Similar multifaceted relationships driven by substitution and transportation options exist for many other refined products.

Figure 6.8 displays the map of the second floor of this petroleum shopping mall, where for transparency, only three major regional crude oil benchmarks are shown, among which Brent plays the central role.

To navigate through this diagram, one can follow contracts by the type of refined product. In the middle distillate complex, European gasoil connects with ULSD and Singapore gasoil. In addition, each of them is linked to corresponding jet fuel markets that trade as basis spreads to regional distillate benchmarks. In the gasoline complex, the Eurobob (EBOB) swap contract, which is a European counterpart of US RBOB futures, and Singapore Mogas swaps, form the triplet of gasoline products. A similar chain exists for the market for another refined product, fuel oil. However, it is omitted from the graph, as the fuel oil market has undergone significant structural changes in 2020 driven by the introduction of a new sulfur specification, which makes the historical data required for a cointegration analysis less relevant.

One can further expand the petroleum stat-arb portfolio by adding natural gas liquids (NGLs) which include ethane, propane, and butane. NGLs not only have strong fundamental linkages among themselves, but they also compete as the feedstock for the petrochemical business with naphtha, another European refined product linked to European gasoline. Going beyond petroleum futures, one can add natural gas markets, where in the US alone there are over thirty different pricing locations that can be grouped together and traded for convergence towards an equilibrium based on their pipeline connectivity. With the expansion of liquefied natural gas shipping facilities, US natural gas is expected to be more closely linked to natural gas prices in Europe and Asia. Furthermore, natural gas competes against coal for local power generation, allowing utilities to switch their inputs based on the relative prices of competing fuels, which could bring coal markets along with emission credits into the global energy stat-arb portfolio as well.

Trading energy pairs is an example of a complex dynamic system, which will be revisited in more detail in the next chapter in the context of broader financial markets. Such systems are characterized by multiple nonlinear interdependencies. In the world of oil spreads, these interdependencies are driven by the constantly evolving economic incentives of arbitrageurs and asset owners that drive their decisions where to ship barrels and what products to make out of them. For example, if additional pipelines are built to move oil from US production areas to the Gulf Coast for exports, then less oil will be sent to Cushing, decreasing the safety buffer of oil inventories available for delivery against WTI futures. A single change within the infrastructure will impact many other spreads tied to WTI. The incentives for producers and shippers change over time, as local production and pipeline capacity do not grow synchronously. While production, once it is established, tends to grow gradually, the takeaway pipeline capacity moves in steps as new pipelines become operational. As a result, during some periods production overwhelms the pipeline capacity, but during other periods production lags, leaving some pipelines unfilled. One should always carefully distinguish between such fundamentally different regimes.

As it has been previously highlighted, because of its high cost, the petroleum ecosystem tends to operate with a just-enough and just-in-time mentality, leaving very little room for error. Such a highly optimized complex system is susceptible to periodic extreme events that could be caused by unexpected disruptions. Consider, for example, Canadian WCS oil, which is linked to WTI via pipelines, but has little buffer in terms of spare pipeline capacity. If the pipeline capacity is sufficient to handle variations in oil flows, then a mean-reverting strategy generally works. However, in the absence of an adequate capacity buffer, a single event that disrupts the critical pipeline flow could cause a downward jump in the WCS-WTI spread since no other comparable transportation alternatives are readily available to move oil out of Canada. The spread then must immediately widen to reflect the economics of the next available pricing tier, transportation by rail, which is much more expensive than moving oil by a pipeline. Such jumps will inevitably lead to trading losses in the convergence strategy for this pair.

The energy stat-arb portfolio is an example of a short volatility strategy characterized by the negative skewness in its returns. Strategies of this type tend to generate steady income, but at the expense of taking a large tail risk. The art of managing the systematic spread convergence portfolio is in knowing when not to trade individual pairs. Such warnings are best seen in the fundamental data, such as the level of inventories. For example, selling elevated refining margin spreads ahead of a hurricane may be a much safer bet when refined product inventories are high, but one would be better off skipping the convergence signal when inventories are low. The fractionation analysis by inventories presented earlier has proven to be a particularly useful tool in making such decisions.

In the oil market, physical arbitrageurs and stat-arb traders mimicking the behavior of asset owners can be viewed as liquidity providers of last resort. Their incentives are typically to trade in the direction opposite to systematic traders, who follow trend and carry signals. To some degree, momentum and carry traders exacerbate volatility and add fuel to the fire by buying when the price is already high and selling when the price is low. Asset owners take the other side of these flows and attempt to extinguish the fire. Despite being on opposite sides of the trade, trend followers and cross-asset arbitrageurs can both make money, as they trade different strategies with different investment horizons. Systematic traders focus on shorter-term directional price movements, while physical and paper arbitrageurs trade longer-term relative value spread strategies which are often structured to be market neutral.

Large negative tail risks coupled with constraints on human intervention prevent traditional CTAs from fully embracing spread convergence strategies, as quants often lack the fundamental insights that are essential for forecasting the likelihood and potential impact of short-term dislocations. However, with technological advances and the proliferation of unique fundamental data, quantitative traders are expanding their participation in petroleum spread trading. Quants can now use up-to-date inventory data from satellites, track tankers carrying oil in transit, and even estimate gasoline demand from downloads of driving maps on mobile devices. While such fundamental data is useful in determining when not to trade the strategy, implementing decision rules in a systematic manner remains challenging.

---

## 6.5 Disentangling Flows and Positioning

Traditional methods of analyzing and forecasting the price of oil are based on fundamental arguments of supply and demand. As the oil market matured, it attracted many new participants, including financial investors, corporate risk managers, and systematic hedge funds, whose trading motives have little to do with fundamentals. Even though their trading strategies cannot always be rationalized by market fundamentals, what these traders do in the market critically impacts the price. In the end, the price rises when there are more buyers than sellers, and it falls when there are more sellers than buyers, regardless of their rationality or motivation. In much the same way that fundamental analysts count the supply and

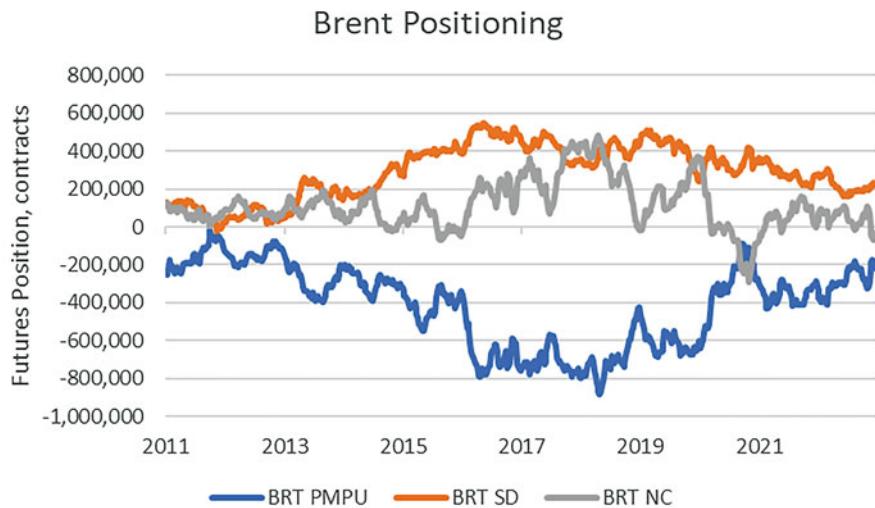
demand for physical barrels, professional traders also count the supply and demand for financial barrels, which are typically measured by futures positions held by different groups of market participants.

Fortunately, some information about behavioral patterns followed by various types of traders can be extracted from the data on their aggregate holdings reported by regulators. Such positioning reports originated in the 1920s for certain agricultural commodities and subsequently morphed into the so-called *Commitments of Traders (CoT)* report. The reporting format evolved over time, but its basic idea is to segregate hedgers and speculators, which are categorized, respectively, as *commercial (CM)* and *non-commercial (NC)* market participants. All traders carrying positions above a certain threshold are categorized. Separate reports are published for futures, and for futures and options combined where options are counted on a delta-adjusted basis. In addition, the CoT report also includes the number of large traders, which is sometimes helpful in the analysis of the position concentration.

According to the CoT definition, the CM designation applies to traders engaged in business activities hedged with the use of futures and options. Therefore, any trading entity that owns a physical asset, or a bank that hedges an OTC market-making book with exchange-listed products, can be classified as a commercial market participant. This definition was originally designed for agricultural commodities where the primary CM is a farmer, and all other traders can be deemed to be speculators. However, in the oil market the categorization is rather confusing. The ownership of a storage asset puts many large physical speculators under the CM designation. In addition, the existence of a large OTC market and the need for market-makers to hedge risks from bilateral derivatives deals puts many banks into the CM category as well, despite the fact that the same banks are also large speculators.

In an attempt to provide additional granularity, in 2006 CoT started to publish the so-called *disaggregated report*, with ICE adopting a similar format five years later. In this report, which became the primary source of positioning information, CM traders are separated into two sub-categories that distinguish between end-users and market-makers. One sub-category includes *producers, merchants, processors, and users (PMPU)*. The other sub-category captures *swap dealers (SD)* and applies mostly to banks, even though several registered market-making entities are also owned by oil majors.

Unfortunately, the reference to producers within the PMPU category is a misnomer, as oil producers and many consumers prefer to trade bilaterally with banks to avoid posting margins for exchange-listed products. Therefore, to make some sense of positions held by corporate hedgers, one instead should be looking at the SD category, as banks effectively hold futures on behalf of end-users. Unlike many other commodities, the PMPU category for the oil market is dominated by large trading houses that own storage and other assets. Their primary business model is based on a physical arbitrage, where the ownership of a physical barrel is promptly hedged with futures. However, this business model is supplemented by aggressive speculation as traders try to take advantage of unique insights derived from the physical market by trading highly leveraged futures and options strategies.



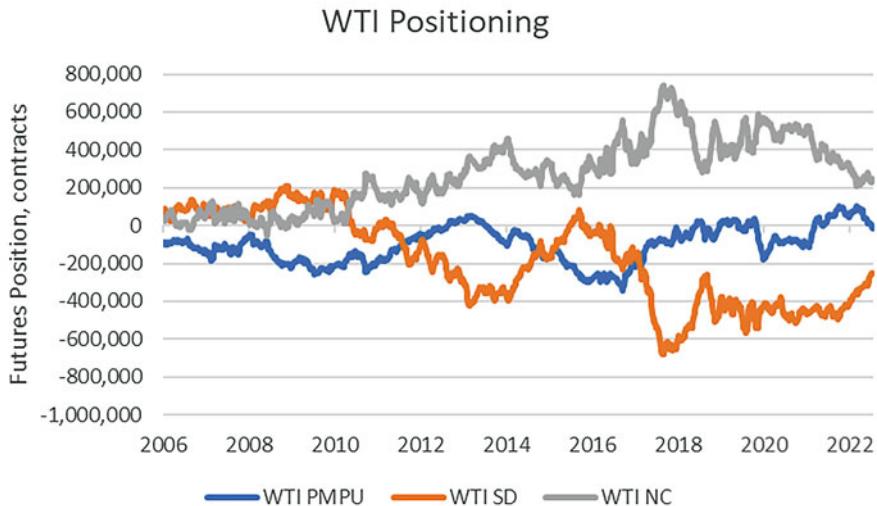
**Fig. 6.9** Net ICE Brent futures positions held by PMPUs, SDs, and NC traders. Source: ICE

The dynamics of positions held by PMPUs and SDs are substantially different for the WTI and Brent markets.<sup>11</sup> Fig. 6.9 shows that the PMPU category is structurally short Brent futures, as most of the world's physical barrels are priced referencing the waterborne Brent contract, and nearly all oil in transit is diligently hedged by selling futures. In fact, it is widely believed that Brent inventory hedgers are the largest participants in the entire oil futures market.

In contrast to Brent, net WTI positions of PMPUs, shown in Fig. 6.10, are more mixed. While some short WTI futures are held to hedge barrels held in domestic storage facilities, such shorts are partially offset by the long leg of WTI-Brent spread arbitrageurs. Physical arbitrageurs buy WTI-Brent spread to hedge US exports or to lock in the economics of pipelines that connect WTI production centers and inland storage facilities to the US Gulf Coast export terminals, where pricing is determined by a waterborne Brent contract.

Positions held by SDs also differ for WTI and Brent, reflecting regional imbalances between producer and consumer hedgers. In North America, the aggregate position held by dealers is heavily dominated by futures shorts held against OTC producer hedging. If one is interested in modeling the hedging behavior of genuine oil producers, then holdings of WTI SDs are likely to be the best proxy. In contrast to WTI, OTC Brent flows are more skewed to the consumer side, as European and Asian airlines and some industrial consumers are more active hedgers

<sup>11</sup>For illustration, we use holdings of two main futures benchmarks, WTI traded on CME and Brent traded on ICE. Adding options and other less liquid exchange-traded WTI and Brent futures does not change the conclusions.



**Fig. 6.10** Net CME WTI futures positions held by PMPU, SD, and NC traders. Source: Commodity Futures Trading Commission (CFTC)

than their US counterparts.<sup>12</sup> In addition, some European utilities purchase natural gas via long-term supply contracts, the pricing of which is formulaically linked to Brent. This exposure is also hedged via swaps with OTC dealers who then buy Brent futures on the exchanges.

The disaggregated CoT report also splits the NC category into two sub-categories. One sub-category corresponds to positions held by *managed money (MM)*, which refers to all investors registered with US regulators. The second sub-category, referred to as *other reportables (OTH)*, is essentially the residual. However, this additional granularity in splitting NC category brings more confusion than any informational value. Many analysts focus only on positions held by MMs and overlook the important OTH sub-category, mistakenly assuming that this residual is rather small. However, the OTH captures all financial institutions that are exempt for various reasons from US registration requirements, including some affiliates of non-US state-owned wealth funds, and private investment vehicles of large family offices. For all practical purposes, MMs and OTHs must be combined and analyzed as overall NCs.

Unfortunately, NC positions are also a mixed bag. This category combines the aggregate holdings of many different types of financial investors, including pension funds that own futures as part of their portfolio and more dynamically managed hedge funds that can trade futures in either direction. The contribution of passive

<sup>12</sup>To simplify the exposition, we do not include positions held in refined products, such as diesel, gasoil, or RBOB. The notional size of such positions is smaller than it is for crude oil futures, even though it is comparable and sometimes even larger if measured relative to the size of the market for refined products.

investments in oil futures makes the NC category structurally long. On aggregate, passive investors take the other side of producers and inventory hedgers, as the sum of all futures positions must be zero. Net NC positions in WTI are substantially longer than in Brent due to the contribution of hedges against ETF flows, the largest of which are linked to WTI. Finally, what makes the deciphering of the oil-positioning puzzle particularly difficult is that only a portion of financial flows is reported under the NC category. Since some investors are using commodity indices intermediated by banks, these flows are reported under the SD, but exact split of financial investments between the two categories is difficult to estimate.

Given the complexity of oil-positioning data with multiple overlaps, it would be very difficult to use this data in a systematic manner, as any signal can hardly be separated from noise. For this reason, we again avoid presenting any historical backtests, which, unfortunately, are frequently seen in marketing presentations by sell-side analysts. Instead, we follow the same path as before, and highlight a few useful mental models that can be applied to positioning data. These concepts are unlikely to be tradable as a stand-alone strategy based on flows, but they are sufficiently robust with respect to the sample selection to be a useful add-on to other trading strategies.

One popular idea for trading based on positioning is the *follow the flow* or to *follow the smart money* strategy. In many financial markets, smart money is generally associated with hedged funds, or the NC category. In the oil market, the answer is more subtle. In fact, as we have seen in Chap. 4, over long periods of time the buyers of oil futures have lost money and the sellers have made money. Since the longs are generally NCs and the shorts are CMs, one could argue that in contrast to many other asset classes, the smart oil money is structural shorts held by inventory hedgers and physical arbitrageurs. However, this argument is also debatable, as over the long run a large portion of P&L from futures comes from the roll yield and not from forecasting the direction of prices. The answer on where the smart money is depends on the investment horizon. Hedge funds generally have an edge in short-term momentum-type trading and use liquidity provided by commercial traders that use futures to monetize optionality embedded in physical assets. In contrast, physical traders tend to do better in capturing long-term value via mean-reverting strategies, where liquidity is provided by long-term investors.<sup>13</sup>

To approximate hedge funds' short-term behavior, one can consider a trading strategy where the standard momentum indicator, introduced in the previous chapter, is applied not to the price, but instead to hedge fund positions. Buy and sell signals  $\pi_{NC}$  can be defined by the momentum in positioning as the crossover of moving averages

---

<sup>13</sup>The question whether speculators or hedgers make money by trading commodity futures has been studied extensively in the academic literature but with largely inconclusive results. Given the complexity of differentiating between hedgers and speculators in the oil market and the lack of available data, establishing such a causality using purely statistical tools is extremely difficult. One interesting attempt to separate the impact of hedgers and speculators on prices by the investment horizon was made by Kang et al. (2020) for a broader commodity portfolio.

$$\pi_{NC}(F_t) = \text{sign}(M_t(NC_t; m, n))$$

where moving averages are calculated for positions held by NCs

$$M_t(NC_t; m, n) = MA_t(NC_t; m) - MA_t(NC_t; n), \quad m < n$$

Our intent is only to provide the generic signal specification but avoid fitting specific parameters. Given the high sensitivity of momentum signals to the lookback period, to avoid the risk of overfitting, we would rather steer away from optimization and let the parameters be customized by the user.

In practice, any trading based on positioning indicators can only be implemented with a lag. Each weekly report reflects positions held as of Tuesday, but the report is released on Friday afternoon after the market is already closed. If one backtests the strategy using daily closing prices, then the earliest day when the trader can react to the latest positioning information would be the following Monday. The existence of such a lag is not specific to strategies based on positioning. Similar issues arise with any non-price-based trading signals where fundamental and economic data become publicly available only with a lag.

Another important concept in trading based on positioning resembles the idea of an inflection point used in the reaction function in the previous chapter. While it may pay to follow hedge funds up to a certain point, if too many funds end up in the same bandwagon, then their own buying or selling can bring the price to an unsustainable level and the trend can quickly reverse. As this turning point is reached, it might be better to enter into a contrarian trade. This strategy is known as *fading extreme positioning* or *fading the crowded trade*. This concept is popular among physical traders, as crowded trades are often accompanied by large fundamentally unjustifiable price moves that may open opportunities for physical arbitrages.

There are numerous ways of measuring the crowdedness of the trade. One metric that has proven to be particularly useful is the so-called *sentiment index* (*SI*). The index normalizes raw positioning data for a certain category, for example the NC, by measuring it relative to its previous minimum and maximum levels over a specified lookback period, as follows

$$SI_t = \frac{NC_t - \min_t(NC_t)}{\max_t(NC_t) - \min_t(NC_t)} \quad (6.2)$$

The sentiment index takes values between 0 and 1. Such normalization removes the overall long bias of the NC category that arises from long futures held by passive investors and focuses only on more dynamic funds.

The trading strategy takes a contrarian position defined by

$$\pi(F_t) = \begin{cases} +1, & \text{if } SI_t \leq \varepsilon \\ -1, & \text{if } SI_t \geq (1 - \varepsilon) \end{cases} \quad (6.3)$$

For example, one can let  $\epsilon = 0.20$  and take long and short positions, respectively, in lower and upper quintiles of the sentiment index.

Like in previously discussed systematic strategies, multiple signals can be blended. One can easily combine a *follow the flow* strategy with *fading extreme positioning* using the idea similar to the reaction function from the previous chapter. Here, one would follow hedge funds up to a certain inflection point, beyond which the position is reduced and then switched to a contrarian direction as the trade gradually becomes crowded. Obviously, the inflection point is hard to pinpoint, and here it becomes a strategy parameter that can be calibrated to historical data.

Even more powerful signals can be constructed if positioning indicators are further confirmed by other signals. For example, one can strengthen the idea of fading the crowded momentum trade by taking a contrarian signal only when some combination of positioning and the price moves is extreme. One way to do this is to apply the algebraic structure (6.2) of the sentiment index to construct a normalized price (NP) indicator, as follows

$$NP_t = \frac{P_t - \min_t(P_t)}{\max_t(P_t) - \min_t(P_t)}$$

Then the two scaled metrics for positioning and price can be combined, using, for example, the Euclidean distance formula

$$d_t = \sqrt{(SI_t)^2 + (NP_t)^2}$$

and the same trading signal (6.3) can be applied where  $SI_t$  is replaced with  $d_t$ .

Any information that can be deduced from the positioning analysis is highly sought after by oil traders. It is analogous to playing a poker game with an educated guess about the opponent's cards. Complete knowledge of the opponent's hand is unlikely to be possible, but even a quick glimpse could lead to a powerful edge. However, given the incompleteness of the overall picture, which is contaminated by overlapping categories in reported data, running a systematic trading strategy purely based on positioning data is unlikely to be sustainable. Like any other quantamental strategy, this concept works better when it is combined with other discretionary inputs. In the next chapter, we extend the idea of quantamental trading to macro factors that connect oil to the broader world of financial markets.

## References

- Alexander, C. (2001). *Market models*. Wiley.
- Bouchouev, I., & Zuo, L. (2020, Winter). Oil risk premia under changing regimes. *Global Commodities Applied Research Digest*, 5(2), 49–59.
- Ederington, L. H., Fernando, C. S., Holland, K. V., Lee, T. K., & Linn, S. C. (2021). The dynamics of arbitrage. *Journal of Financial and Quantitative Analysis*, 56(4), 1350–1380.

- Fattouh, B. (2011). An anatomy of the crude oil pricing system. *Oxford Institute for Energy Studies, Working Paper*, 40.
- Imsirovic, A. (2021). *Trading and price discovery for crude oils: Growth and development of international oil markets*. Palgrave Macmillan.
- Kang, W., Rouwenhorst, K. G., & Tang, K. (2020). A tale of two premiums: The role of hedgers and speculators in commodity futures markets. *The Journal of Finance*, 75(1), 377–417.



# Macro Trading

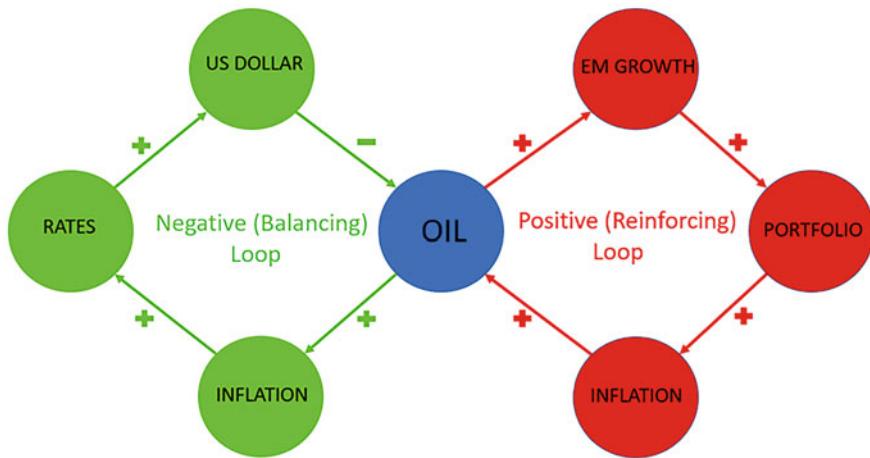
7

- The oil ecosystem is a sub-component of a larger complex dynamic system of financial markets. Such systems are characterized by feedback loops with two-way causality and a critical role of boundaries that keep the system from falling apart.
- The relationship between oil and the US dollar (USD) is an aggregate of multiple feedback loops competing for dominance. While the joint dynamics of oil and USD is too unstable for trading, the linkages between oil and currencies of oil-exporting countries are more robust.
- The idea of cross-asset stat-arb trading can also be applied to the equity market, where oil is traded as a spread to the basket of stocks of independent oil producers. The stock selection can be optimized using hedging profiles of individual companies.
- The relative value strategy between gasoline futures and short-term inflation swaps is another example of the cross-asset stat-arb. The two markets operate at different speeds, creating trading opportunities when the needs of respective market participants diverge.
- An attempt is made to estimate the fair value of oil as a function of macroeconomic variables. While the model is based on shaky theoretical grounds, it has still proven to be helpful to macro traders. Its modifications include dynamic factor selection and methods of machine learning.

---

## 7.1 Dynamic Systems and Feedback Loops

In this final chapter dedicated to trading linear instruments, such as futures and swaps, we look at how oil trading connects to other financial asset classes. So far, we have only considered strategies confined to the petroleum complex, implicitly assuming that these trades are immune to exogenous financial factors. However, the entire petroleum ecosystem is a sub-component of a larger complex dynamic system of financial markets, where everything impacts everything else, the



**Fig. 7.1** Examples of negative (balancing) and positive (reinforcing) macroeconomic feedback loops that involve oil

relationships between system components are reciprocal, and information flows in both directions in the form of multiple feedback loops.

Financial markets are filled with numerous loops between interconnected components. Consider, for example, one of many causal macroeconomic chain reactions that could result from an exogenous rise in oil prices. As explained in Chap. 4, an increase in the price of oil raises the headline inflation mechanically via oil pass-through into retail gasoline prices. Rising inflation impacts monetary policy, increasing the likelihood of interest rate hikes. Higher US interest rates then induce capital inflows to the USA from countries that have lower domestic interest rates, leading to a stronger USD. However, since the price of oil is denominated in USD, dollar appreciation creates downward pressure on the price of oil, which contains its initial rise.

This chain is an example of a *negative, or balancing, feedback loop*, shown on the left side of Fig. 7.1. In a balancing loop, any change in a given variable forces other system components to react in a way that dampens the initial change. Such balancing loops are generally associated with mean-reversion.

Consider now an example of a different causal loop. The rise in oil prices can also be associated with stronger demand in emerging markets (EM). Any indication of EM growth then stimulates capital flows to move from stable assets, such as US Treasuries, towards riskier investments, triggering a so-called risk-on move across broader financial markets. Weakening USD can lead to higher US inflation via more expensive imports of consumer goods. This brings additional demand for financial oil products as a hedge against inflation, further contributing to the rise in oil prices. This is a *positive, or self-reinforcing feedback loop*. In contrast to the balancing loop, the positive feedback loop amplifies the initial change in the given variable. Self-reinforcing loops are associated with momentum.

The behavior of any complex dynamic system is characterized by multiple positive and negative feedback loops, all acting concurrently. If the positive loop dominates, then the system maintains its current direction. However, it cannot run in the same direction forever. There are certain hard boundaries that must eventually limit the move, much like storage boundaries do in the case of trending oil inventories. The presence of system boundaries forces a shift in dominance when the balancing loop takes over from the reinforcing one. Obviously, the crucial element in the analysis of such systems is identification of the turning point when the dominance shifts. This point is what drives the trader's decision whether to buy or to sell. The reaction function with an inflection point, introduced in Chap. 5, provides an example of how traders can quantify such a shift in dominance.

Modeling the dynamics of the entire system is a gargantuan task. Conventional statistical methods, such as linear regression with a pre-defined direction of causality, are unlikely to be of much value here. A more promising approach is to use the concept of cointegration, introduced in the previous chapter. It searches for some stable equilibrium among system components. While the system revolves around the equilibrium state, it eventually converges to it. Deriving a universal equilibrium for the entire dynamic system of financial markets, however, is more interesting for economic theorists than for traders. In practice, such an equilibrium is impossible to quantify given the lack of timely data and the large amount of noise that surrounds it. Practitioners instead tend to focus their attention only on specific components of the system, where mutual feedback loops are more visible and sufficiently ring-fenced from unrelated noise.

In this chapter, we briefly explore pairwise relationships between oil and three major financial markets. First, we look at the foreign exchange market and analyze the widely debated but very unstable relationship between oil and USD. We eliminate some USD-specific noise by focusing on the relative value strategy between commodity futures and currencies of commodity-exporting countries. We then extend the concept of cross-asset stat-arb to the equity market and consider a strategy of trading oil futures as the spread to the basket of stocks of independent oil and gas producers. Finally, we revisit an important linkage between oil and inflation that has already come up on several occasions in the book. This time, we outline a specific arbitrage-like trading strategy between petroleum futures and short-term inflation swaps. To conclude, we combine some of these observations into a directional macro model that trades based on the estimate of the fair value for the price of oil derived predominantly from various macroeconomic variables.

---

## 7.2 Oil, Dollar, and Commodity Terms of Trade Strategies

The relationship between oil and USD is too multifaceted to be molded into any prescriptive format. It is one of the most extensively studied economic relationships that, nevertheless, still leaves traders scratching their heads when its dominant driver suddenly changes. One common pitfall of such analyses is the application of conventional regression methods that implicitly assume the direction of causality

from one asset to another, which is not consistent with the dynamics of a highly nonlinear system. The oil-USD link in the complex system of financial markets is the result of the confluence of many forces and feedback loops acting with varying levels of strength at different times while fighting for dominance. The dominant factor shifts quite frequently, leading to drastic changes in correlations, which makes regression-based estimates extremely unstable and sensitive to the sample selection.

Like members of a family or parts of the universe, oil and USD impact each other in many different ways. There is little doubt that in the long run oil and USD are related. However, to turn this relationship into a viable strategy on shorter-term trading frequencies, one must find a way to eliminate less predictable factors and zoom in on the behavior of more stable residuals. This is the same principle that underlies previously studied energy stat-arb strategies. Now we take another step and extend similar ideas across different asset classes, starting with the elusive relationship between oil and USD.

The question of how the price of oil impacts USD has been actively debated ever since the dollar was de-pegged from gold in 1971. Interest in this topic has gone way beyond the basics of macroeconomics. It has been a question of US national security that played a central role in geopolitics during previous episodes of high oil prices. Following the steep rise of oil prices and fuel shortages in the aftermath of the Oil Embargo of 1973–1974 by the Organization of Arab Petroleum Exporting Countries (OAPEC) that sent the USA into a recession, Washington convinced Saudi Arabia to invest excess oil revenues into US Treasuries in exchange for military aid. These investments were subsequently dubbed *petrodollar recycling*. The deal was so clandestine that its details only became known to the public several decades later. Effectively, Saudi Arabia was financing a significant portion of American spending, while at the same time higher oil prices were increasing demand for USD from many other petroleum-importing countries. This *medium of exchange* transmission channel contributed to a positive correlation between oil and USD.

For years, the positive linkage from the medium of exchange channel was largely offset by an inverse relationship between oil and USD from the *US terms of trade* channel. Since the USA was the largest importer of oil, rising oil prices negatively affected the US overall trade deficit, which, in turn, put downward pressure on the US currency. Over time, the contribution of both petrodollar recycling and the US terms of trade channel became muted. The impact of petrodollar recycling on USD even turned negative when sovereign wealth funds of oil-producing countries started to diversify their petrodollars by selling USD and buying other currencies. The US terms of trade linkage also vanished nearly entirely, as the rapid growth of US shale production made the country energy independent.

For completeness, however, we list both the medium of exchange and the US terms of trade transmission channels as legacy factors in Fig. 7.2, as oil analysts must be aware of their past contribution when backtesting trading strategies.

As demand for oil grew globally, particularly outside of the USA, the dominant role in oil-USD dynamics shifted to *the denomination, or the numeraire effect*, which is the consequence of oil being priced predominantly in USD. The

Oil-USD Linkages		Positive	Negative
Macroeconomic	Legacy	Petrodollar recycling / Medium of exchange	US terms of trade
		Monetary policy	Denomination / Non-US demand
		US growth/weakness	Denomination / Non-US supply
Flow-based		Inflation expectations	Risk aversion
		Geopolitical uncertainty	Portfolio rebalancing

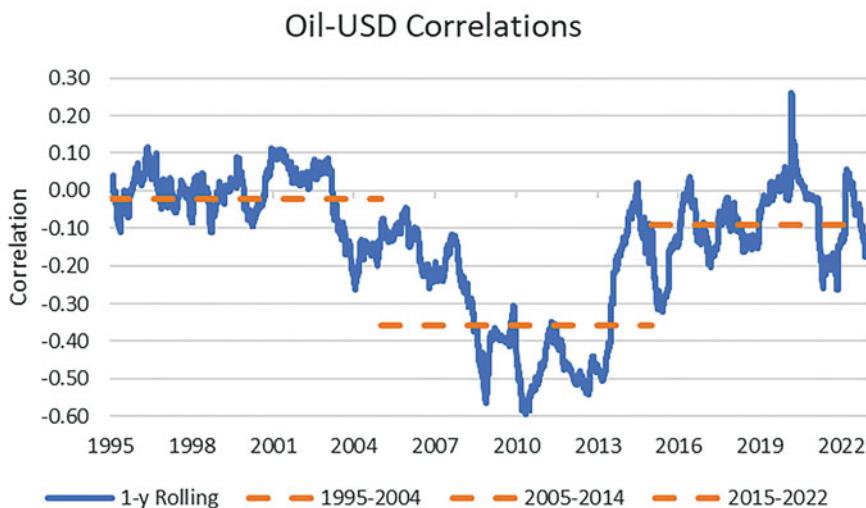
**Fig. 7.2** Multiple transmission channels between oil and USD

denomination effect leads to an inverse relationship, as stronger USD implies that it takes fewer USD to buy a barrel of oil.

This inverse relationship can be visualized from the supply and demand perspective. A stronger dollar translates into higher domestic prices in countries outside of the USA, which over the last decade were the primary drivers of oil demand growth. Higher prices have a negative impact on demand, and, therefore, create downward pressure on the price of oil. In addition, USD appreciation also causes a negative impact on price via increasing supply by oil-producing countries with free floating currencies. Since a portion of production costs is denominated in local currencies, stronger USD improves profit margins and incentivizes oil exports. The supply-side channel is more complicated though due to secondary factors, such as the value of USD-denominated debt issued by foreign producers.

There are several macroeconomic factors that counter the predominantly negative oil-USD relationship. One is driven by oil direct pass-through into inflation and its impact on changes in monetary policy. As we have already mentioned, rising inflation increases the likelihood of interest rate hikes, which, in turn, positively impact demand for short-term US Treasury bonds. Another positive contribution to the oil-USD relationship comes from the US growth effect. More favorable economic prospects in the USA relative to the rest of the world could result in stronger USD given the higher likelihood of monetary tightening and stronger oil prices, as the USA is still the world's largest petroleum consumer. Likewise, both USD and oil can falter when the USA experiences an idiosyncratic slowdown. This happened, for example, in the early days of the Covid-19 pandemic in the USA when both oil demand and USD fell, and for a short period of time the correlation between the two falling assets was positive. Subsequently, as the pandemic escalated around the world, USD regained its stability, while oil continued to fall because of the reduction in global demand. As a result, the correlation returned to negative.

Numerous academic studies of macroeconomic links between oil and USD have been conducted, but they only produced conflicting and inconclusive results. This highlights the time-varying nature of the relationship and the high sensitivity of econometric analyses to the sample selection. Many attempts have also been made to establish an aggregate direction of causality. However, given so many comingled channels acting at once, disentangling causality in any robust way using statistical methods is practically impossible. While slow-moving macroeconomic forces remain more relevant for a long-run relationship, in the short-run their combined impact is indistinguishable from non-tradable noise. In fact, prior to the financialization of oil markets that began in the 2000s, the dynamics of short-run



**Fig. 7.3** Rolling one-year and sub-sample WTI-USD correlations during three different regimes

oil-USD correlation largely resembled a random walk around zero mean. Financialization has brought drastic changes to this relationship. Markets started to move much faster, and the dominance of fundamental forces has been challenged and superseded by fast-moving financial flows and cross-asset spillovers.

Perhaps the most important driver of the oil-USD relationship in the modern financialized world of oil trading is the *risk aversion or safe haven channel*, which is driven by risk-on, risk-off sentiment shifts. When adverse economic news triggers a large selloff in the equity market, many other risky assets, including oil, are also liquidated by financial portfolio managers tasked to maintain fixed allocation across all risky assets. During such a risk-off event, investors reallocate capital towards safer securities, such as US Treasuries which, in turn, increases demand for USD.

The financial safe haven channel has become the dominant driver of the predominantly inverse oil-USD relationship. Since approximately 2004, the correlation between returns on oil and a broad USD index has had a strong negative bias, as illustrated in Fig. 7.3. The correlation was the strongest, i.e., the most negative, during the years of quantitative easing following the global financial crisis. To stimulate investment demand, large bond purchases were made by the US Federal Reserve Bank, which led to low interest rates and downward pressure on USD, pushing investors into riskier assets, including oil.

Occasionally, the risk aversion channel also contributes to a positive relationship between oil and USD. This typically occurs either during periods of high inflation caused by supply-side constraints, or when geopolitical tensions jeopardize the stability of global oil production. During such episodes, oil itself, along with USD, is perceived by the market to be a safe haven instrument, and the dominant negative relationship is interrupted by shorter spells of positive correlation. For example, the

inflation factor became more visible when the economy was recovering from the Covid-19 pandemic and higher inflation was driven by supply disruptions and rising commodity prices.

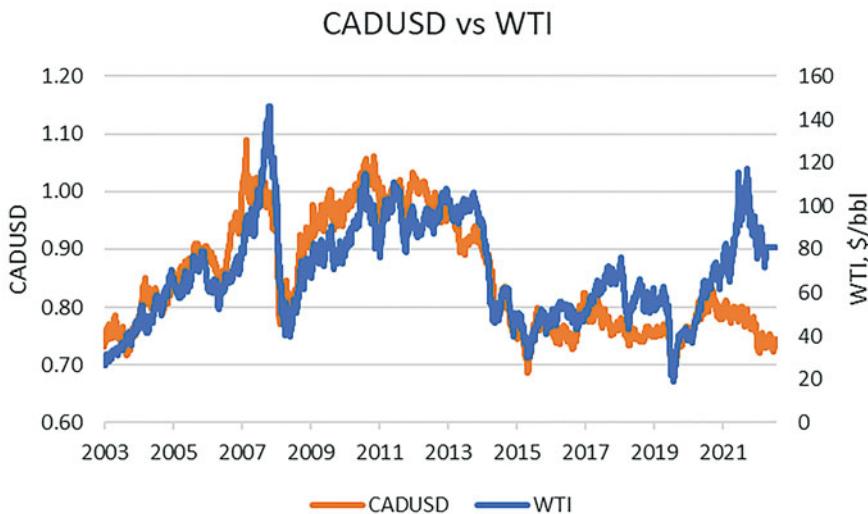
Finally, the conventional denomination effect also started to play a larger role in financial oil markets. Many international investors have oil allocation mandates set in their local currency, such as EUR. If USD strengthens, i.e., the EURUSD exchange rate weakens, then the notional value of oil holdings measured in EUR exceeds the investor's target allocation. Oil positions must then be reduced to bring the allocation back to the target, creating downward pressure on the price of oil. The numeraire effect is even more pronounced for highly leveraged risk party funds and CTAs that manage allocations based on the asset volatility. If oil and USD are negatively correlated, then oil volatility measured in EUR is generally smaller than oil volatility measured in USD. Consider, for example, oil trading at 90 USD per barrel and a EURUSD exchange rate at 1.20, which translates to an oil price of 75 EUR per barrel. If the oil price rises 10% to 99 USD per barrel and EUR vs USD strengthens 10% to 1.32, then the oil price measured in EUR remains unchanged at 75 EUR per barrel despite its 10% change in USD.

With so many overlapping driving factors of the oil-USD relationship, trying to use one of the two variables to predict the other rarely results in a viable trading strategy. One is better off trading them as a spread that does not depend on the direction of causality. For the spread to be a viable candidate for the convergence strategy, some cointegration criteria must be satisfied. While such criteria usually fail for the spread between oil and a broad USD index, they are often met for a subset of the so-called *commodity currencies* of producing countries, such as Canada or Norway. In contrast to petrodollar recycling by producers whose currencies are pegged to USD, many commodity exporters with free-floating currencies convert excess USD-denominated oil revenues into domestic currencies, which makes them highly sensitive to the price of oil. This transmission channel is sometimes referred to as a *wealth effect*.

In fact, oil plays such an important role in the economics of these countries that the price of oil is frequently used by Central Banks as an input for modeling the fair value of their currencies. Similar models for commodity currencies have also been developed and widely adopted by foreign exchange traders. This opened an opportunity for a cross-asset stat-arb-like strategy to trade certain commodities as the spread relative to their currencies whenever the two markets dislocate. We illustrate this with a popular convergence strategy between CADUSD, a Canadian dollar expressed in USD, and WTI.

Since the CADUSD exchange rate and WTI represent values of two assets both denominated in USD, trading one as the spread against the other eliminates a large portion of the idiosyncratic USD-specific noise. As a result, the co-movement between CADUSD and WTI is more visible, as illustrated in Fig. 7.4. In other words, CADUSD and WTI are generally considered to be cointegrated.

There are effectively two primary factors that drive CADUSD. One factor is the relative monetary policy in Canada versus the USA. This factor is typically modeled by interest rate differentials, which steer financial flows towards the direction of



**Fig. 7.4** Co-movement between CADUSD and WTI futures prices (m3)

higher interest rates. The other primary factor is the price of oil. When monetary policies in both countries are on hold, then the contribution of the first factor weakens, and the price of oil starts playing the dominant role. However, when monetary policies are changing, the exchange rate becomes more driven by interest rate differentials and the relationship with oil weakens. One can clearly see this during uneven recovery after the Covid-19 pandemic and rapidly rising inflation.

If the monetary policies in both countries are expected to remain relatively stable, then the technical implementation of the convergence strategy becomes very simple, as in the energy convergence strategies discussed in the previous chapter. We define the cross-asset spread as

$$S = CADUSD - \beta \cdot WTI$$

and then buy and sell the pair when the spread moves outside of a certain band. The band can be specified, for example, by the spread deviation from its moving average, as follows:

$$\pi(S_t) = \begin{cases} -1, & \text{if } S_t - MA_t(S_t; n) > \varepsilon \\ +1, & \text{if } S_t - MA_t(S_t; n) < -\varepsilon \\ 0, & \text{if } |S_t - MA_t(S_t; n)| \leq \varepsilon \end{cases} \quad (7.1)$$

The most challenging technical part of this strategy is in defining  $\beta$ , which is generally much more of an art than it is a science. Since the idea of cross-asset stat-arb is based on the concept of cointegration, which measures the long-term relationship between price levels, in theory, one should be estimating betas using the price-level regression with a fairly long lookback period.

**Fig. 7.5** An example of pair selection for a CToT portfolio

Commodity Terms of Trade (CToT) Pair Selection	
Currency	Commodity
CADUSD	WTI
NOKUSD	Brent
CLPUSD	Copper
AUDUSD	Iron Ore, Coal, LNG
BRLUSD	Brent, Soybeans, Sugar
RUBUSD	Brent, Nickel, Platinum
ZARUSD	Platinum, Gold, Coal

In practice, cross-arb traders pay particular attention to the choice of the lookback period. It must be chosen carefully, corresponding to regimes that are less contaminated by changes in monetary policies. As in the case of the energy stat-arb portfolio, the crucial decision here is when not to trade the strategy. In a more sophisticated version of the strategy, one can consider hedging against the uncertainty resulting from changes in monetary policies. This can be done by supplementing CADUSD versus WTI spread trade with additional positions in the spread between Canadian and US interest rate futures.

Like energy stat-arb, the cross-asset stat-arb strategy is also better traded within a broader portfolio. Besides the CADUSD and WTI pair, a similar statistically robust relationship exists between NOKUSD, a Norwegian krone expressed in USD, and Brent. Trading this pair together with CADUSD and WTI as a mini portfolio can substantially improve the overall performance, as it provides some degree of geographical diversification both across currencies and in the oil market. One can further diversify the strategy by adding EM currencies of oil producers, but it should be done with care due to higher unrelated risks in these countries. Furthermore, it could be beneficial to replace benchmark WTI and Brent futures with oil grades that are specific to the country's exports, such as WCS (West Canadian Select) for Canada. The addition of less liquid oil contracts, however, increases transaction costs and requires more advanced execution capabilities.

The terms of trade strategy is not limited to energy futures. It can also be applied to some metals and agricultural products. For example, Chile is the largest producer of copper, with the Chilean peso, CLPUSD, highly correlated to the price of copper. Figure 7.5 lists commodity currency pairs commonly used for constructing more diversified *commodity terms of trade (CToT)* portfolios. This pair selection is based on various correlation and cointegration thresholds with an additional adjustment for liquidity and geographical diversification. For some countries that export many different commodities, the strategy can be improved by using the basket of corresponding commodity futures that mimic the basket of exports.

To conclude, we highlight the primary risk embedded in such strategies. Since many commodities are produced in countries known for their geopolitical instability, challenges arise when domestic commodity production is disrupted. The strategy can then experience a double whammy loss if the commodity price spikes due to the

reduction of supply from the producing country, while the local currency simultaneously weakens, as financial investors reduce their financial holdings in the country in response to increasing domestic uncertainty.

We next extend the concept of cross-asset stat-arb to the equity market, where oil futures can be traded as a spread to a portfolio of energy stocks.

### 7.3 Oil and Energy Equities

The relationship between oil and broad equity indices is no less complex than the relationship between oil and USD. Many of the macroeconomic transmission channels listed in Fig. 7.2 stem from overall financial conditions that impact both equities and oil markets. Transmission linkages between broad equity markets and oil are also time-varying and have low signal-to-noise ratios, which makes it difficult to trade the relationship. The idea behind a cross-asset stat-arb strategy is again to eliminate a portion of the macro noise and trade oil futures relative to the subset of the equity market where connections are more direct. This subset is the equity basket of independent oil and gas producers, whose profitability is highly dependent on the price of oil. With the exception of a few gas-focused companies, the majority of the revenue in this sector is derived from oil production.

While professional oil-equity arbitrageurs are more likely to design their own customized equity baskets, the strategy works quite well even for standardized baskets of stocks, using, for example, ETFs. For illustration, we use XOP, the most popular and the most liquid ETF of US independent oil and gas producers. XOP has been in existence since 2006, the period that covers the growth in US oil production. Figure 7.6 shows that the correlation between WTI and XOP is naturally much higher than the correlation between WTI and the broader US equity market.

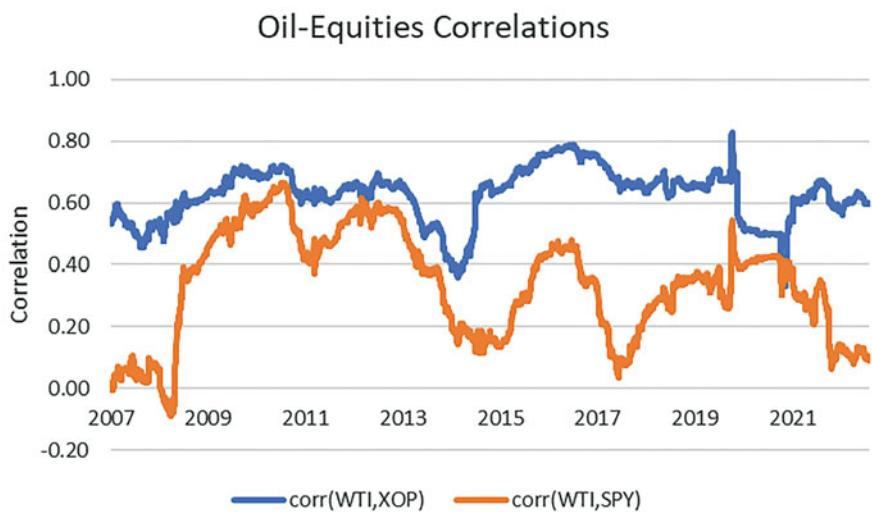
The XOP-WTI spread satisfies conventional cointegration criteria, making this pair a suitable candidate for the cross-asset convergence strategy. Figure 7.7 illustrates how closely the two assets tracked each other over time.

The actual strategy implementation is again kept simple. We apply the basic trading rule (7.1) to the spread, defined by

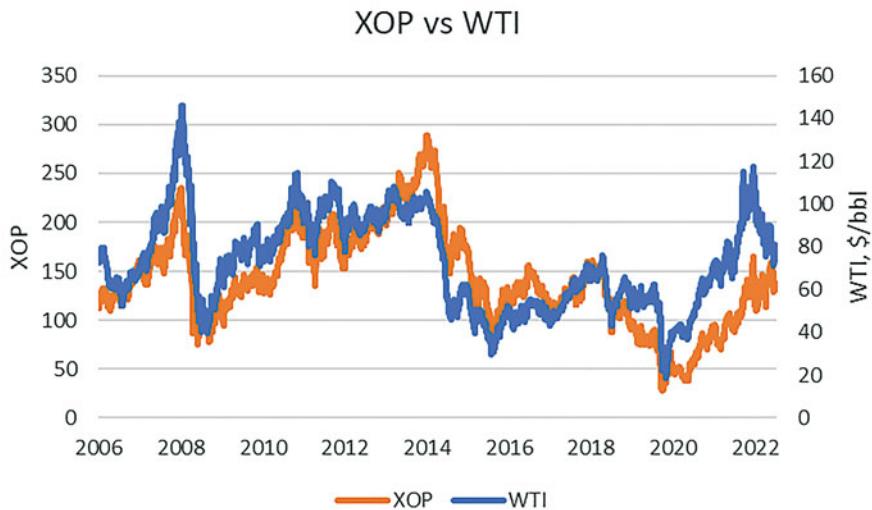
$$S = XOP - \beta \cdot WTI$$

To avoid the hassle with the estimation of hedging beta, practitioners often construct the equity basket, which has similar volatility to WTI. Since the volatilities of XOP and WTI happen to be roughly similar, in this example one can trade the strategy with  $\beta = 1$ , which means that the pair is notionally weighted.

We should also caution readers against backtesting the strategy using publicly available closing prices for XOP and WTI. All stat-arb trading strategies critically rely on simultaneity of price data for both legs. While the WTI market closes at 14:30 Eastern Standard Time (EST), US equity markets remain open for another ninety minutes. Using both settlement prices for backtesting introduces a look-ahead



**Fig. 7.6** Rolling one-year returns correlation of WTI futures ( $m^3$ ) to XOP and SPY equity ETFs



**Fig. 7.7** Co-movement between XOP and WTI futures prices ( $m^3$ )

bias that could overestimate the historical performance of the strategy. To properly analyze this strategy, one must use intra-day data.

Obviously, like in all other trades, the basic oil-equity strategy can be enhanced. Perhaps the most unique approach which is specific to this strategy is to replace the standardized XOP ETF with a customized basket of energy stocks. One can also run the strategy for each individual stock and then construct an optimal basket of stocks

that exhibit the strongest cointegration with oil. However, such cherry-picking introduces additional degrees of freedom and must be used carefully to avoid overfitting. It is also possible to construct a portfolio of oil and natural gas futures weighted based on the share of oil and gas revenues for companies in the chosen basket.

One important factor to consider is the impact of hedging by producers, which could reduce the sensitivity of the company's stock to the price of oil. A particularly novel concept is to design a customized equity basket with weights that explicitly incorporate producers' individual hedging ratios. Alternatively, all producer stocks can be split into basket of hedgers and non-hedgers, which theoretically must have different oil betas but the market may not fully recognize such a distinction. While this approach is quite elegant, its implementation is very complex. The hedging data are generally available only from regulatory producer filings, which are published quarterly with a lag. However, this data can be combined with information about hedging deals that are reported live by Swap Data Repositories (SDRs). Even though real-time SDR reporting does not disclose names of trading counterparties, one can attempt to deduce them by establishing hedging patterns using quarterly filings. We will leave this idea for future research that can probably qualify for a doctoral thesis in finance.

Having analyzed cross-asset arbitrage between two fast-moving markets, we next look at another interesting example of the cross-asset lead-lag strategy, between fast-moving energy futures and the slower-moving OTC market for inflation swaps.

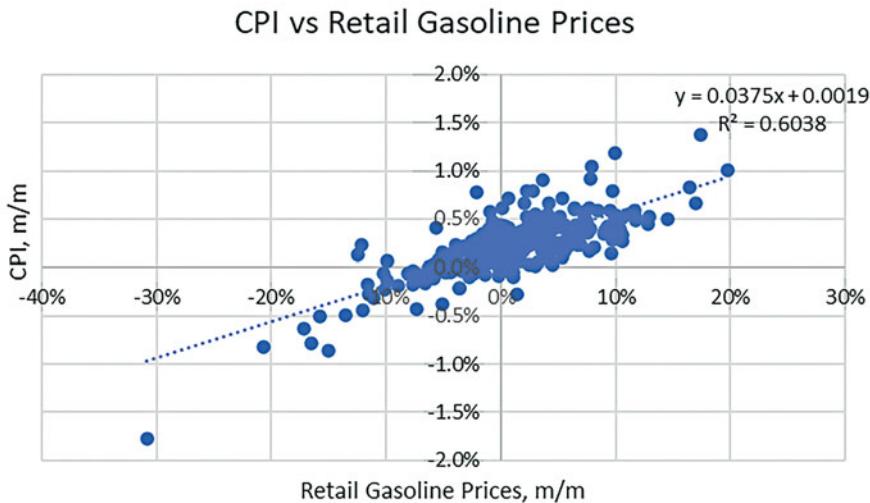
---

## 7.4 Oil and Inflation

The relationship between oil and inflation that we have already encountered on multiple occasions provides another example of a bidirectional causality, a typical attribute of complex dynamic systems. As discussed in Chap. 4, oil causes inflation directly via its pass-through to retail gasoline prices and indirectly as the feedstock cost of other goods. Inflation, on the other hand, impacts oil more via forward expectations, which generates demand for hedging, in particular, from risk parity funds. The strength of both transmission channels depends on the time horizon.

One simple way to trade inflation relative to oil would be to replicate the statistical convergence strategies presented above for commodity currencies and for oil and gas producer equities. This indeed can be done but a better trading opportunity exists, which is more specific to the structure of the inflation market. This opportunity is to trade short-term inflation swaps, whose linkages to energy futures happen to be more algebraic than statistical. While this strategy is not free money, it is much closer to being a model-free arbitrage than to a statistical arbitrage.

The world's most developed inflation market is in the USA, where the primary trading instruments are inflation-linked bonds, or TIPS, introduced in Chap. 4, and bilaterally negotiated inflation swaps. Both TIPS and inflation swaps are settled based on the CPI, which was also introduced in Chap. 4. The index represents the weighted average of prices for the aggregate consumer basket of goods and services,



**Fig. 7.8** A large portion of CPI monthly variance is explained by retail gasoline prices (1993–2022)

with the reference base set at 100. Inflation is quoted as the **rate of change in the CPI typically on a month-over-month or year-over-year basis**. While TIPS are only issued with maturities in January, April, and July, **inflation swaps exist for each monthly CPI print, which is also called a CPI fixing**. Such a finer granularity of swaps for monthly CPI fixings makes them a better match to trade against energy futures, which are also listed with monthly expirations.

As it was shown in Fig. 4.5, the **headline CPI is driven by volatile oil prices via their pass-through to retail gasoline prices**. While consumer gasoline expenditure, which is reported under the motor fuel component of the CPI, represents less than **4% of the total basket, it explains approximately 60% percent of the CPI monthly variance**.<sup>1</sup> Fig. 7.8 illustrates this relationship over a long period of thirty years. In fact, during periods of low inflation the explanatory power of retail gasoline prices tends to be even larger, as other CPI components exhibit low volatility.

To describe the trading strategy, we use RBOB futures, which are most closely related to retail gasoline prices, even though a similar strategy can be constructed using more liquid WTI futures, or a basket of petroleum futures that may also include diesel futures and Brent.

The trading opportunity exists largely because energy futures and inflation markets operate at different speeds. While the electronic market for futures moves very fast, inflation swaps largely trade OTC, where transactions are still negotiated in the old-fashioned manner bilaterally between clients and their bank dealers. If the

<sup>1</sup> As of December 2022, the relative importance of the motor fuel category was 3.275%, out of which 3.172% was gasoline. Source: US Bureau of Labor Statistics (BLS).

inflation market were efficient, then one would expect every move in energy futures to be instantaneously repriced in CPI swaps. In practice, however, this does not always happen, as the slow-moving OTC inflation market often exhibits some inertia. This market is dominated by a different set of market participants, predominantly by fixed income portfolio managers whose decisions and investment mandates are set based on longer-term structural views. They largely ignore short-term fluctuations in energy prices as many of them are not even authorized to trade RBOB futures. Energy traders, on the other hand, tend to be nimbler and more flexible with a better eye on short-term arbitrage-type profits.

When the inflation market lags the move in energy futures, the trader can take a position against the dislocation between RBOB futures and inflation swaps and trade the pair as a cross-asset spread. The strategy bets on the residual which we can define as *ex-gasoline inflation*. The dislocation between the two prices may allow the trade to be entered at sufficiently advantageous levels, so that other components of the CPI do not have to be estimated with high precision. The crux of the strategy is in separation and careful hedging of the CPI gasoline component. The ex-gasoline component is effectively treated as residual noise. To implement the strategy, however, one needs to understand the somewhat intricate plumbing of inflation derivatives.

One important feature of the inflation market that often causes a considerable amount of confusion to outsiders is the so-called *base effect*. Another peculiar quirk of inflation derivatives is the lag between the valuation date and the period over which inflation is measured. We will briefly explain these features but without getting too deeply into details. For simplicity, we use a benchmark one-year inflation swap.

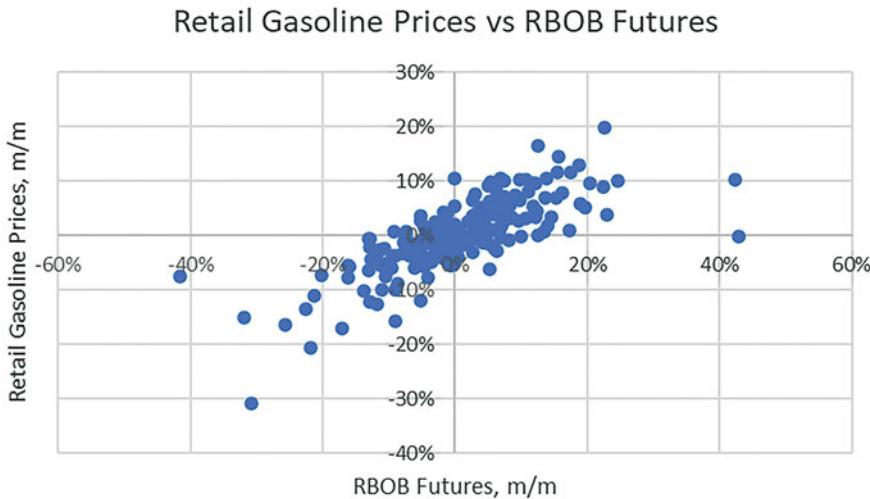
The fixed rate of the swap is quoted as an annual expected rate of inflation. This rate is swapped for the floating rate of inflation  $i(T)$ , which is determined by the ratio of the reference CPI index  $I(T)$  at the future time  $T > t$  to the reference index  $I(t_0)$  at past time  $t_0 < t$ , as follows:

$$i(T) = \frac{I(T)}{I(t_0)} - 1$$

The reference index is calculated using monthly CPI prints but with a lag of between two and three months. If the observation day  $t$  is the first day of a calendar month, then the reference index uses a monthly CPI print for the third preceding calendar month. In other words, in this case  $t_0 = t - 3/12$  and the annual inflation rate will be determined by the CPI print nine months forward, i.e., at time  $T = t + 9/12$ . If the observation date  $t$  is any other day of the calendar month, then the reference index is determined by a linear interpolation between CPI prints two and three months prior.<sup>2</sup>

---

<sup>2</sup>The interpolation formula is  $I(t) = I(t_M) + \frac{t-t_M}{D}(I(t_{M+1}) - I(t_M))$ , where  $t_M$  is the reference index for the first day of the calendar month in which  $t$  falls,  $t_{M+1}$  is the reference index for the first day of the calendar month immediately following  $t$ , and  $D$  is the number of days in the month in which  $t$  falls.



**Fig. 7.9** Retail gasoline prices are highly correlated to RBOB futures (2005–2022) (RBOB futures were introduced in October 2005)

If another one-year swap is traded tomorrow at time  $t + dt$ , then it will reference a new base index  $I(t_0 + dt)$  which could be substantially different from  $I(t_0)$  if energy futures happened to experience a large move at time  $t_0 + dt$ . This makes it rather difficult to compare market inflation swap rates on a rolling basis and visualize the price dynamics in the form of a time series. As the base index constantly rolls in time, the quoted swap level changes daily even if market expectations about the future inflation index remain the same.

The skill of trading short-term inflation relative to energy futures hinges on the trader's ability to separate and hedge out the volatile gasoline contribution from the residual ex-gasoline inflation component of the index, which we denote by  $i_x$ . Let  $R(T)$  represent retail gasoline prices, and  $\omega$  denote the weight of gasoline expenditures in the CPI basket. For simplicity, we assume that  $\omega$  is a known constant, even though, more precisely, it can also vary with energy prices. Then the quoted rate of inflation can be decomposed as

$$i = \omega \left( \frac{R(T)}{R(t_0)} - 1 \right) + (1 - \omega)i_x \quad (7.2)$$

Let us define the *pass-through*  $p$  of energy futures into retail gasoline. This is typically estimated by the slope of the linear regression of percentage changes in retail gasoline and RBOB prices. An example for monthly data is given by Fig. 7.9.

The magnitude of the pass-through increases if one aggregates data over several months or if one uses year-over-year calculations, which eliminates short-term noise caused by seasonal differences. However, the results of such regressions should be interpreted with caution given the presence of autocorrelation effects.

Furthermore, the goodness of the fit can be improved by allowing for a lag between futures and retail prices. This lag, however, is not trivial. In the short run, retail prices are stickier than futures, which is one of the reasons that this trading opportunity exists. Moreover, the retail lag is also asymmetric. Retailers tend to raise prices at the pump quickly in response to rising futures, but they reduce pump prices at a much slower rate when futures fall, as they are trying to maximize profit margins. Without getting too sidetracked with the econometrics of pass-through calculations, which is not the goal of this book, we can generally use the industry standard assumption of approximately 60% pass-through for RBOB futures into retail gasoline prices.

We then replace retail gasoline inflation in (7.2) with RBOB using the pass-through of traded futures as follows

$$i = \omega p \left( \frac{F(T)}{F(t_0)} - 1 \right) + (1 - \omega)i_x \quad (7.3)$$

where  $F(t_0)$  represents prompt RBOB futures prices observable at time  $t_0$ .<sup>3</sup>

Alternatively, the market-implied rate of ex-gasoline inflation can be expressed as follows

$$i_x = \frac{i - \omega p \left( \frac{F(T)}{F(t_0)} - 1 \right)}{1 - \omega} \quad (7.4)$$

If one were able to trade futures directly on retail gasoline prices, then (7.4) would have applied with  $p = 100\%$ . In fact, there exists an OTC market for retail gasoline swaps, but, unfortunately, it is extremely illiquid. A similar market for retail diesel swaps is more developed, and some energy traders use the market forward curve for retail diesel prices as a guide to estimate the forward price of retail gasoline, which is the most uncertain variable in the CPI-RBOB convergence strategy.

Let us consider a trading position in CPI swap  $i$  hedged with  $n$  contracts of RBOB futures

$$S = i - nF$$

What makes this strategy appealing is that the crucial hedging ratio can be calculated algebraically rather than estimated statistically.

Let us assume that the futures price experiences an instantaneous shock  $dF$ . Then from (7.3) we obtain that the value of the inflation swap contract changes by

---

<sup>3</sup>Since inflation is measured on a calendar monthly basis, future prices must be interpolated between two contracts with surrounding maturities. To calculate the base gasoline price, one must also use linear interpolation similar to that described in the prior footnote for the reference inflation index.

$$di = \omega p \frac{dF}{F(t_0)}$$

To illustrate this with a numerical example, let  $p = 0.60$ ,  $\omega = 0.04$ , and  $F(t_0) = 3.00$  per gallon. Then a ten cent per gallon move in the RBOB futures price changes the annual rate of inflation by  $0.60 * 0.04 * \frac{0.10}{3.00} = 0.08\%$ . Since one RBOB futures contract is for 42,000 gallons, to hedge a \$100 million notional position in the inflation swap, the trader needs approximately  $\frac{0.08\% * 100mm}{42,000 * 0.10} \approx 19$  RBOB futures contracts. In contrast to previous strategies, here the hedging ratio is calculated fundamentally instead of using rolling regressions, even though some statistical uncertainty is still hidden in the assumption about the pass-through  $p$ .

Once the most volatile component of the headline inflation is removed, the trading signal can be defined in a number of alternative ways. Fixed income specialists tend to look at it more in inflation terms. They calculate  $i_x$  as in (7.4) and then compare it to some measure of historical inflation  $i_h$ . The latter is usually estimated from a range of inflation realizations over prior years, adjusted for seasonality and recent trends. The trading signal is to buy and sell inflation-RBOB spread  $S_t$  when  $i_x$  measured at time  $t$ , which we denote by  $i_{x,t}$ , deviates from its trend-adjusted historical seasonal norm, i.e.,

$$\pi(S_t) = \begin{cases} -1, & \text{if } i_{x,t} - i_h > \varepsilon \\ 1, & \text{if } i_{x,t} - i_h < -\varepsilon \\ 0, & \text{if } |i_{x,t} - i_h| \leq \varepsilon \end{cases}$$

Since the idea of the strategy is to trade it only when the deviation becomes particularly significant, one can avoid the need for high precision in the estimation of  $i_h$ . If the dislocation between inflation and futures markets is very large, then ex-gasoline inflation  $i_{x,t}$  can even fall entirely outside of the range of previous historical observations. This would suggest that the trading opportunity is so good that it would have made money under all previous paths of realized inflation.

In contrast to inflation specialists, oil traders like to turn the problem upside down and map all variables instead into energy terms that they are more familiar with. The primary uncertain variable in this strategy is the forward price of retail gasoline  $R_t(T)$ , which at time  $t$  is approximated using the shape of observable futures curves for RBOB and, if necessary for longer maturities, crude oil futures. Therefore, by and large, this strategy amounts to trading the spread between retail gasoline prices implied by the inflation market and RBOB futures.

Instead of expressing market-implied  $i_{x,t}$  via pass-through of observable futures prices as in (7.4) and comparing it to  $i_h$ , energy traders assume that the inflation market already embeds some consensus about historical average ex-gasoline inflation  $i_h$ . In this case, one can replace  $i_x$  with  $i_h$  in the Eq. (7.2), and instead back out the market-implied future retail gasoline price  $\hat{R}_t(T)$  from the headline inflation rate  $i_t$  as follows

$$\hat{R}_t(T) = R(t_0) \left( 1 + i_h + \frac{i_t - i_h}{\omega} \right)$$

This market-implied retail gasoline price is then compared to observable prices of RBOB futures. If inflation and futures are misaligned, then the trader buys or sells retail-futures basis at a level which falls outside of the historical seasonal norm for this basis. In other words, using energy variables one can also consider an alternative trading rule, specified as

$$\pi(S) = \begin{cases} -1, & \text{if } \hat{R}_t(T) - F_t(T) > b(T) + \varepsilon \\ 1, & \text{if } \hat{R}_t(T) - F_t(T) < b(T) - \varepsilon \\ 0, & \text{if } b - \varepsilon \leq |\hat{R}_t(T) - F_t(T)| \leq b(T) + \varepsilon \end{cases}$$

where  $b(T)$  represents some historically normal retail-futures basis for the month  $T$ . The oil trader can also recalculate  $\hat{R}_t(T)$  for a range of different  $i_h$  to check the sensitivity of this trading rule to the assumption about historical ex-gasoline inflation.

The dynamics of the retail-futures basis is the core component of this trade. Besides the non-trivial lag mentioned earlier, the basis can also be affected by regional anomalies in gasoline prices. The gasoline component of the CPI represents the national average for finished gasoline, but RBOB is only a particular blending component to be delivered in the New York area against the futures contract. The spread, therefore, could periodically dislocate during isolated regional events, such as Gulf Coast hurricanes, or supply disruptions of gasoline-blending components on the US East Coast. Such energy-specific knowledge, which financial investors lack, creates an opportunity for energy specialists to function as arbitrageurs in the short-term inflation swap market. Given so many important nuances in the behavior of the gasoline basis, this strategy is unlikely to be successful if it is managed purely systematically. It presents another example of quantamental trading where a human equipped with a machine is more powerful than either a human or machine alone.

So far, we have focused our attention only on the most volatile gasoline component of the CPI. One can also eliminate an even larger portion of the CPI variance by hedging some other energy components of the basket reported under energy commodities or energy services which are also correlated to diesel, natural gas and electricity prices. The algebra of the previous calculations remains the same with  $\omega$  now representing a larger weight of approximately 6%, which explains a larger portion of the CPI monthly variance, but the pass-through must be calculated with respect to the basket of energy futures.

Taking it one step further, one can extend the framework and attempt to hedge out the second most volatile CPI category, food prices. Let  $\omega_1$  and  $\omega_2$  represent relative importance weights for the energy and food sub-components of inflation, and  $p_1$  and  $p_2$  represent pass-through sensitivities of energy futures  $F_1$  and agricultural futures  $F_2$  to the energy and food components of the CPI, respectively. Then one can introduce the ex-energy-ex-food component of inflation, which is analogous to the market-implied *core inflation*,  $i_{\text{core}}$ , as follows

$$i = \omega_1 p_1 \left( \frac{F_1(T)}{F_1(t_0)} - 1 \right) + \omega_2 p_2 \left( \frac{F_2(T)}{F_2(t_0)} - 1 \right) + (1 - \omega_1 - \omega_2) i_{core}$$

where  $F_1(t_0)$  and  $F_2(t_0)$  represent two baskets of energy and agricultural prompt futures prices at time  $t_0$ .

The market-implied core inflation is then calculated as follows

$$i_{core} = \frac{i - \omega_1 p_1 \left( \frac{F_1(T)}{F_1(t_0)} - 1 \right) - \omega_2 p_2 \left( \frac{F_2(T)}{F_2(t_0)} - 1 \right)}{1 - \omega_1 - \omega_2}$$

The practical challenge with this more advanced approach lies in the estimation of the pass-through of agricultural futures into the food sub-component of inflation. Given the larger diversity of products that make up the food basket, such pass-through estimates for agricultural futures are less reliable.

Having developed some tradable linkages between oil and other main financial asset classes, we conclude this chapter by building a simple model that aggregates the information from various macro variables and attempts to use this information to estimate the fair value for oil.

## 7.5 Macro Fair-Value Model

As we have highlighted throughout the book, professional oil traders tend to focus more on capturing market inefficiencies and trading spreads than on forecasting the direction of oil prices. Many successful oil traders are effectively arbitrageurs. They tend to buy relatively cheap assets and sell correlated but more expensive ones, trading them as spreads and minimizing directional exposure to the price of oil. Such relative value trades almost always have better risk-adjusted returns than outright bets on oil prices.

However, for the vast majority of financial traders oil represents only one of many assets in their portfolio. Macro traders are unlikely to make a living by arbitraging energy prices, and instead they may be content with a high-level opinion on whether the price of oil is cheap or expensive. To help them navigate in the ocean of macroeconomic variables that could affect the price of oil, we conclude this chapter with one somewhat naïve model. It is based on rather shaky theoretical grounds, so professional oil traders are unlikely to take it seriously. However, as we stated previously, all models are wrong, but some are, nevertheless, useful. This model attempts to estimate the fair value for oil by applying some elementary statistical techniques to macro variables.

The idea behind the model comes from the foreign exchange market. Historically, currencies have been viewed as balancing valves that adjust in response to various macroeconomic factors. Foreign exchange analysts spend a considerable amount of time valuing currencies using various techniques, ranging from purchasing power parity and long-term equilibrium models, to faster market models, which are based

on short-term currency drivers and capital flows. For example, to estimate the fair value of CADUSD, analysts often apply a multidimensional regression, where explanatory variables are interest rate differentials that impact cross-border capital flows, some measure of the global risk appetite, such as the equity index or VIX, and the price of oil. Traders then buy and sell the currency if the market price deviates too far from the theoretical value generated by the fitted model.

A similar approach can be applied to estimate the fair price for oil  $\hat{P}_t$ , which is viewed as a linear combination of  $i = 1, \dots, N$  macro factors  $x_{i,t}$  observed at time  $t$ . In other words, we let

$$\hat{P}_t = \beta_0 + \sum_{i=1}^N \beta_i x_{i,t} \quad (7.5)$$

For example, we can choose the three-factor model ( $N = 3$ ) with factors representing financial legs of the three previously discussed cross-asset spread trading strategies. In this case,  $x_1 = CADUSD$  exchange rate,  $x_2 = XOP$  equity ETF of oil producers, and  $x_3 = CPI$  swap price. These are representative factors from foreign exchange, equity, and fixed income markets, which have close economic links to the price of oil.

In the cross-asset spread strategies, we were buying and selling oil and taking the opposite position in these three financial markets. Here, we consider a strategy which uses financial variables only as a signal and not as trading instruments. For example, if oil is deemed to be cheap relative to all three factors, or at least relative to some combination of them, then it may indicate that oil is generally undervalued. In this case, one may want just to invest in oil without taking any offsetting position in other financial assets. Obviously, such a directional position is much riskier, as oil along with other financial factors could get even cheaper in the case of a macro-driven risk-off event.

To determine the cheapness or expensiveness of oil relative to financial factors, we use the simple price-level regression defined by (7.5). As we discussed in the previous chapter, for the most part running linear regressions on non-stationary prices is meaningless, which is why we warned in advance about the shakiness of the theoretical foundation of this approach. However, while being undoubtably questionable, this approach is not necessarily wrong. Applying regressions to price levels may be appropriate if variables are cointegrated. Recall that cointegration means that while each variable may not necessarily be stationary, there exists a linear combination of them that makes the entire basket stationary. The Eq. (7.5) effectively makes such an assumption.

In truth, for many other financial factors an Eq. (7.5) is an ambitious assumption, but the method is still used by traders. We all know plenty of examples where something that works in theory does not work in practice. This model is somewhat the opposite. It should not work in theory, but it has proven to be helpful in practice. In practice, traders often use (7.5) by letting beta coefficients be time-varying and estimating them by running rolling regressions. The lookback for such regressions

Macro Tradable	Macro Fundamental	Oil Specific
Broad USD Index	Global PMI	Inventory
Commodity <b>FX</b> Basket	GDP Nowcast	Carry
5y5y Breakeven	Economic Surprise Index	Hedge Fund Positioning
2y-10y Interest Rate Spread	Financial Conditions Index	Refinery Margins
VIX	Chinese Demand Index	Option Skews
Emerging Market Equity Index		
Credit Spreads		

**Fig. 7.10** Additional factors that can be included in the oil fair-value model

must be relatively long as the idea of cointegration is to find some long-term equilibrium among the variables. At the same time, running regressions on the rolling basis allows traders to incorporate the latest data points. The output of the regression  $\hat{P}_t$  can be viewed as representing the fair price of oil, which is then compared to the market price  $F_t$ .

A simple trading strategy then buys oil futures  $F_t$  when they trade below the estimated value by more than a certain threshold  $\epsilon$  and sell them when the market price exceeds the model price by this threshold:

$$\pi(F_t) = \begin{cases} -1, & \text{if } F_t - \hat{P}_t > \epsilon \\ +1, & \text{if } F_t - \hat{P}_t < -\epsilon \\ 0, & \text{if } |F_t - \hat{P}_t| \leq \epsilon \end{cases}$$

This three-factor fair-value strategy based on USDCAD, XOP and CPI swaps has been working well since approximately 2016, when the US ban on oil exports was lifted and WTI became more exposed to global macroeconomic forces. As with all other quantamental strategies, we prefer to stay away from delving into backtests to avoid the natural temptation to make them better with optimized parameters. When working with daily settlement prices, one should also be careful not to introduce the look-ahead bias in backtesting, as oil daily settlement prices are published before prices of some financial factors, such as equities. In these backtests, one must either wait until the following day before taking a position, or better calculate signals using simultaneous intra-day data for all factors.

The static factor-based strategy does have some merit, but it also carries a significant tail risk. There are many factors not included in the model specification that could cause large shifts in oil prices. OPEC decisions, geopolitical events, or weather-driven supply disruptions can have a large impact on price, but they cannot easily be modeled systematically. The natural extension of this model would be to include additional risk factors. Figure 7.10 lists a much wider set of potential factors, which at various times have been important drivers of the price of oil.

The factors are grouped into three categories that represent tradable market prices of different assets, macroeconomic fundamental factors, and oil-specific factors. Some non-tradable fundamental factors may not be available on the daily basis, so a broader macro model can only be implemented at a lower frequency, which does decrease its performance characteristics. Another challenge comes from the fact that

very few of these factors are cointegrated with oil, making it difficult to use price-level regressions of the form (7.5). In this case, one can switch to using regressions on price changes which makes all variables stationary.

One can take this idea a step further and make the factor-selection process more dynamic in an attempt to identify the most relevant factors during each period. This can be done by running individual single-factor regressions on price differences and ranking factors based on corresponding  $R^2$ . One can then estimate the fair value of oil using a multidimensional regression against several factors that have the highest  $R^2$  in the previous period. However, we would caution against the mechanical application of such a factor-selection approach given the high degree of collinearity among them. At the end, a quantamental trader still has the final say on which factors to use in any given period.

This problem of dynamic factor selection appears to be a good candidate for using more sophisticated statistical techniques, such as the methods of machine learning. While the author spent a considerable amount of time in trying to apply these methods, so far, the results produced by more advanced statistical models are only marginally better as compared to a simple static three-factor model. However, with recent advances in data science, this area of research continues to evolve. If this book ever makes it to its second edition, the chapter on application of machine learning methods will most certainly be expanded. Until then, we can say that our analysis of relatively simple linear trading strategies is now completed. We can move to a more advanced, but also a more lucrative universe of nonlinear option strategies.

---

## **Part III**

### **Volatility Trading**



# Options and Volatilities

# 8

- Options are nonlinear derivatives that inspired the development of modern probability theory. An option price is driven by uncertainty, which is measured by volatility. It is essential to distinguish between local volatility, realized volatility, and implied volatility.
- Local volatility is a functional parameter of the diffusion process that describes the behavior of futures prices. In contrast to one-parameter normal and lognormal models, local volatility of a general diffusion process is a function that depends on time and the futures price.
- Options can be replicated by dynamically trading futures. The original Black-Scholes-Merton framework extends to general diffusions. The price of an option satisfies the equation of heat transfer where conductivity of a non-homogeneous medium is replaced with the local volatility function.
- Realized volatility is typically measured by the standard deviation of asset returns for a particular realization of a diffusion process. It is a noisy, backward-looking, and skewed statistic. In the oil market, volatility of price changes is more meaningful than volatility of percentage returns.
- Implied volatility is a forward-looking plug designed to tweak an inaccurate pricing formula to make it match the market price. Its sole purpose is to convert option prices into a more convenient coordinate system. Volatility smile measures how far the model deviates from the market.

---

## 8.1 Options and “Théorie de la Spéculation”

By some accounts, options embedded in commodity transactions predate any other financial derivatives. Perhaps the most famous example of an early commodity option trade is attributed to Thales of Miletus, who around 550 BC, according to Aristotle, paid a little money to secure the exclusive use of all olive presses in the towns of Chios and Miletus. After an unexpectedly bumper harvest that the philosopher claimed to have predicted using weather patterns, the option paid off

handsomely, as the demand for olive presses surged. The profit from the option not only took the philosopher out of poverty, but it also proved to sceptics that science may indeed be a worthwhile endeavor.

The payoff of an option contract is asymmetric, somewhat analogous to a lottery ticket. The buyer pays a relatively small premium and gets compensated if the price rises above a certain threshold in the case of a *call option*, or if the price falls below a given threshold in the case of a *put option*. This threshold, which is crucial for valuing an option, is called the *strike price*.

More formally, a *European call option*  $C(F, t)$  pays the difference between the futures price  $F(T)$  at time  $t = T$  and the contractually specified strike price  $K$  if this difference is positive, and, otherwise, it expires worthless:

$$C(F, T) = \max(0, F(T) - K) \quad (8.1)$$

Likewise, a *European put option*  $P(F, t)$  at time  $t = T$  pays the amount by which the future falls below the strike price if this amount is positive, and it pays zero, otherwise:

$$P(F, T) = \max(0, K - F(T)) \quad (8.2)$$

The payoffs of call and put options are nonlinear due to the pronounced kink defined by the strike price. The option's nonlinearity, or its convexity, is the highest when the strike price is located near the current futures price. An option with the strike  $K = F(t)$  is called *at-the-money (ATM)*. It is typically the most liquid option which serves as a primary anchor for option prices with other strikes. Many options are initially struck *out-of-the-money (OTM)* with  $K > F(t)$  for calls, and  $K < F(t)$  for puts. OTM options are cheaper, as futures must cross some distance before the strike price is reached for the option to start paying off. If the current futures price exceeds the strike price, i.e.,  $F(t) > K$ , then the call is *in-the-money (ITM)*, and likewise, the put is ITM when  $F(t) < K$ .

Nonlinearity is what makes the mathematics more challenging for options relative to futures. The value of a call option depends on the probability that the futures price exceeds the strike price at the expiration of the option. Since the magnitude of the option payoff varies with the level of futures, one must know not only the cumulative probability of reaching the strike price, but rather the collection of all probabilities of futures reaching every possible price level. The sum of such probability-weighted payoffs for all possible futures prices then determines the value of the option. The tricky part, of course, is how to get such probabilities.

The search for probabilities that determine the value of an option contract created a foundation for the entire modern theory of probability. Arguably, such a theory was born in 1900 when an extraordinary “*Théorie de la Spéculation*” was presented by a French mathematician, Louis Bachelier, as his doctoral thesis at the Sorbonne.<sup>1</sup> For the first time in the history of science, the process of Brownian motion, or the

---

<sup>1</sup>Bachelier (1900).

random motion of particles in a medium, was described mathematically. Bachelier's idea was to use Brownian motion to characterize the random behavior of stocks and options in the financial market. Remarkably, one of the most groundbreaking scientific discoveries of all time was made by an options analyst. Only five years later, Albert Einstein, unaware of Bachelier's pioneering work, applied a similar framework to describe the erratic motion of pollen particles caused by water molecules.

The significance of Bachelier's work cannot be overstated. Its reach extends beyond the derivation of the first option pricing formula, which is still frequently used today in the oil market. More importantly, while trying to tackle the problem of option pricing, Bachelier introduced the so-called *law of radiation of probability*, and derived that the normal probability density, the famous bell-shaped Gaussian curve, satisfies the Fourier equation of heat transfer. Thus, the first connection between the concept of probability and the physics of diffusions that describes uncertainty in many scientific disciplines was established in the derivatives market. This connection subsequently gave rise to the Fokker-Planck equation in physics, the Chapman-Kolmogorov equation in probability and statistics, and the Black-Scholes-Merton (BSM) equation in finance. The basics of the diffusion process and the corresponding probabilities are summarized in Appendix A.

To price an option, Bachelier assumed that the speculator is not expected to make or lose any money by trading the asset that underlies the option contract. While individual expectations of buyers and sellers of stocks or futures could differ, the market's aggregate expectation, he argued, must be zero, since there are as many willing buyers as willing sellers. This *principle of zero expectations* for the speculator's profit can be interpreted as the earliest formulation of the efficient market hypothesis. The fair price of an option is then established under this assumption that the speculator's own view about the direction of the underlying asset price is irrelevant.

Bachelier paid particular attention to an ATM option which he called a *simple option*. Interestingly, in his days, only simple options were available for trading in commodity markets. The search for the fair value of such an option led Bachelier to what he viewed as the most important contribution of his entire study, summarized in the following statement:

The value of a simple option must be proportional to the square root of time.

The proportionality is characterized by what Bachelier defined as the *coefficient of instability*, or what we know today as *volatility*. His insight inspired the foundation of stochastic calculus, where the variance, or the volatility squared, of the variable following the random walk is proportional to an increment of time.

Bachelier was clearly ahead of his time. Back then traders did not rule the world, and the science of financial markets was not perceived to be a real science. Despite such a revolutionary invention, it took Bachelier more than a decade to secure a permanent academic position. His remarkable contribution remained largely

forgotten until the 1960s, when the modern theory of stochastic calculus was developed and applied to the analysis of financial markets.

The second big innovation in the theory of options came with seminal papers by Black and Scholes and by Merton.<sup>2</sup> They formally derived the differential equation for the option price, by showing that the option can be replicated by dynamic trading of the underlying asset. In the next section, we apply their argument to a more generalized setting of diffusion processes that have proven to be particularly useful for describing the dynamics of oil futures. The derivation of the *Black-Scholes-Merton (BSM)* equation confirmed that the expected value of the underlying asset is indeed irrelevant for the valuation of derivatives. Under certain additional assumptions, the risk carried by an option contract can be offset by holding some variable quantity of futures. As a result, options should be priced in the so-called *risk-neutral* world, which is analogous to Bachelier's principle of zero speculator expectations. The only parameter that matters for option pricing is the volatility of the asset, or Bachelier's coefficient of instability.

More broadly, volatility in financial markets is defined as the measure of dispersion of uncertain prices around the mean. In other words, it is the measure of risk. What makes the concept of volatility often confusing to outsiders is the somewhat broad usage of the term volatility. In contrast to the price of a financial asset, which is generally observable, the volatility of the asset is not observable. Volatility can only be understood and calculated in connection to a specific model or assumption. Since the term volatility could mean different things in different settings, the best way to remove any ambiguity is to reference it with a clarifying adjective. In this chapter, we introduce three types of volatility, the *local volatility (LV)*, the *realized volatility (RV)*, and the *implied volatility (IV)*.

---

## 8.2 Local Volatility and Diffusions

Prices of financial assets, including oil futures, are stochastic, which means that some part of their behavior is uncertain or, in other words, random. The randomness is described by the probability distribution for futures prices. For example, a random component of the price change can be pulled from a normal distribution, whose probability density is described by the well-known bell-shaped Gaussian curve. The dynamics of the asset price could also carry a deterministic component, for example, to characterize the growth trend or mean-reversion towards a long-term equilibrium price level. Together, deterministic and stochastic components form the stochastic differential equation that describes the behavior of prices. We have already encountered one example of a stochastic equation in Chap. 3 when we modeled oil inventories. To keep this part of the book self-contained, we reintroduce the topic here in somewhat greater depth and refer for more technical details to Appendix A.

---

<sup>2</sup>Black and Scholes (1973), and Merton (1973).

Perhaps the simplest example of a stochastic process, which was also used by Bachelier in his thesis, is *arithmetic Brownian motion (ABM)*. It assumes that future changes in the asset price are normally distributed. The stochastic differential equation for ABM specifies changes in futures prices  $dF$  over a small increment of time  $dt$  in the following form:

$$dF = \mu_A dt + \sigma_A dz \quad (8.3)$$

The first term, denoted by  $\mu_A$ , is deterministic. It represents the drift, or the expected futures price change per unit of time. If it were the only term in the equation, then futures prices would drift linearly in time with a constant slope  $\mu_A$ .

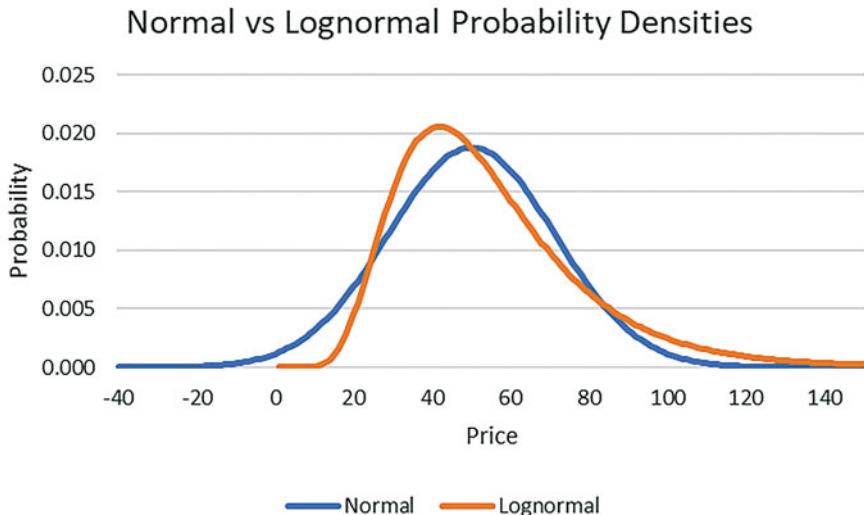
The second term is stochastic. It describes the uncertainty via independent random increments  $dz$ , which is drawn from a normal distribution with zero mean and whose variance is equal to an increment of time  $dt$ . One can also write the random component  $dz$  as

$$dz = \varepsilon \sqrt{dt}$$

where  $\varepsilon$  represents noise, or a random number taken from the normal probability distribution with zero mean and variance equal to one. The scaling by  $\sqrt{dt}$  ensures that the variance of  $dz$  is equal to  $dt$ . Since increments  $dz$  are independent and variances of normally distributed variables are additive, the variance of uncertainty grows linearly with time, and the standard deviation, which is the square root of the variance, grows as the square root of time. It should also be noted that for small increments of time the magnitude of  $\sqrt{dt}$  is larger than  $dt$ , and therefore, the second, random term in the stochastic equation dominates the short-term deterministic component of  $dF$ .

The magnitude of the uncertainty in the ABM process is controlled by a constant parameter  $\sigma_A$  which is often referred to as *normal volatility*, or *dollar volatility* to emphasize its unit of measure. If  $dz$  represents the source of risk, then normal volatility can be thought of as representing the quantity of risk. To indicate the arithmetical nature of the Brownian motion, we use corresponding subscripts for drift and volatility coefficients. Under the ABM assumption, the futures price drifts at a constant rate  $\mu_A$  with uncertainty specified by a constant volatility of price changes  $\sigma_A$ . Since, for the oil market,  $dF$  is measured in dollars per barrel, the normal oil volatility  $\sigma_A$  is measured in dollars per barrel per unit of time. The distribution of prices generated by ABM process (8.3) is described by the normal probability density given by (A.12).

The assumption of normality to describe price changes is often problematic for certain financial markets, such as equities, for which many traditional derivatives models have been tailored. The ABM process allows asset prices to move indefinitely in either direction and does not prevent prices even from falling below zero, something that is impossible for stocks and bonds. However, as we have seen previously, negative prices cannot be ruled out for a storable commodity, such as oil. Another challenge with an application of ABM to the equity market comes from



**Fig. 8.1** An example of normal and lognormal probability densities

its inconvenient scaling, as a five-dollar price change is much more impactful for a ten-dollar stock than it is for a hundred-dollar stock. For equity investors, percentage returns are more intuitive than absolute changes in stock prices.

To keep modeling standardized across assets and to bypass the issue of negative stock prices, the financial industry applied the assumption of normality instead to investment returns, which are measured in percentages. Such a price dynamics is described by the following stochastic differential equation, known as *geometric Brownian motion (GBM)*

$$\frac{dF}{F} = \mu_G dt + \sigma_G dz \quad (8.4)$$

In the GBM case, asset returns are normally distributed, or alternatively, the logarithm of prices is normally distributed.

Under these assumptions, the percentage return on a fully funded futures position is expected to grow with a constant rate of return  $\mu_G$  and uncertainty characterized by a constant percentage volatility of returns  $\sigma_G$ . For the most part, when a generic reference is made to the term volatility, it is typically assumed to be  $\sigma_G$ . It is important to highlight that unlike the price of the financial asset, the term volatility can only be understood in reference to a specific model. Here, volatility is understood as the standard deviation of percentage returns under the assumption that prices are lognormally distributed. The corresponding lognormal probability density function is given by (A.14).

The lognormal probability density is inherently asymmetric. It is only defined for positive prices, and it has a much larger right-side tail compared to the symmetric bell-shaped normal curve. The shapes of two probability densities are illustrated in

**Fig. 8.1.** While one can justify using an asymmetric distribution in financial markets that tend to grow over time, applying it to mean-reverting commodity markets brings some adverse side-effects. As we explained in the chapters on storage and hedging pressure, in the oil market, the frequency and the magnitude of upward and downward price movements tend to be more symmetric, if not even skewed to the downside.

By and large, the commodity industry took the path of least resistance and followed the lognormal modeling paradigm prevalent in the equity markets. Without any doubt, this assumption is extremely convenient when comparing strategies across different commodities using a standardized percentage-based volatility metric. No risk manager would want to deal with volatility measured in dollars per barrel for oil, in dollars per bushel for corn, or in dollars per ounce for gold.

The convenience of such a one-size-fits-all measure comes at a price. It forces traders to look at the oil market through an incorrect frame of reference imposed by inflexible and highly standardized operational setups of their trading systems. In contrast, dedicated oil specialists customize their metrics, which allows them to capture important salient features of the market and to take advantage of the framing bias often suffered by generalists. This modeling duality adds certain pedagogical challenges, as many market phenomena must be presented in two alternative ways, first as they are seen by the general crowd through standard but poor-quality lenses, and then, as perceived by oil professionals equipped with more accurate viewing tools.

Both ABM and GBM assumptions are rather simplistic and chosen primarily for their analytical tractability. It would be naïve to think that oil price behavior, which is influenced by a multitude of driving factors, can be fully characterized by only two parameters, the constant drift, and the constant volatility. The real world of oil prices is much more complex. Fortunately, as we will see in the following section, the drift term in the stochastic equation plays no role in modeling options, and our entire focus will be on volatility. However, volatility of oil futures is anything but constant. It varies not only with time, but also with the level of futures prices. For example, as futures move outside of some normal price range, absolute volatility tends to rise. This indicates much higher probabilities of extreme events that are nearly impossible under either normal or lognormal distributions. In other words, the distribution of oil prices has fat tails.

To allow ourselves more flexibility in modeling notorious tails of the price distribution, we assume that the drift and volatility parameters of the stochastic process are deterministic, but not yet specified, functions of time and the futures price. Futures then follow a more general stochastic process described by the following equation:

$$dF = \mu(F, t)dt + \sigma(F, t)dz \quad (8.5)$$

Such stochastic processes, which arise in many natural sciences and applications, are called *diffusions*. In finance, the function  $\sigma(F, t)$  that characterizes the volatility of the diffusion process is known as *the local volatility* function. ABM and GBM

processes represent only two special cases of the more general class of diffusions. The local volatility of the ABM process is constant  $\sigma_A$ . For the GBM process, the local volatility  $\sigma_G F$  is proportional to the futures prices. All price changes for diffusions are still driven by the same normally distributed single source of noise  $dz$ .

The dependency of the volatility function on the futures price, which itself is stochastic, allows diffusions to produce a much wider universe of probability distributions for future price changes. One simple way to generate a probability distribution by a diffusion process is to use Monte Carlo simulations, where multiple sequences of random variables  $dz(t)$  are drawn from the normal distribution. The corresponding futures prices are then computed using the discretized version of (8.5) and presented in the form of a histogram.

The science of diffusions largely hinges on one important result, known as *Itô's lemma*. It describes how a function of a stochastic variable changes for a given change in the variable itself. In the ordinary calculus, a small change in the value of the function  $G(F, t)$  is approximated by its partial derivatives multiplied by a change in the function's arguments, the result known as *Taylor's formula*. When one of the variables  $F$  is stochastic, the analogous approximation is given by the following *Itô's lemma*:

$$dG = \left( \frac{\partial G}{\partial t} + \frac{1}{2} \sigma^2(F, t) \frac{\partial^2 G}{\partial F^2} \right) dt + \frac{\partial G}{\partial F} dF \quad (8.6)$$

This equation resembles the first-order approximation of a regular function of two variables  $F$  and  $t$ , except for an additional term that contains the second derivative with respect to  $F$ . In the regular calculus, this term would have been multiplied by  $(dF)^2$  and discarded as a lower second-order term, but in the stochastic calculus it can no longer be omitted. The reason for keeping it in the approximation comes from the property of the random noise  $dz$ . Since, by definition,  $dz$  is proportional to the square root of time, the variance of  $dz$  is equal to  $dt$ . Therefore,  $(dF)^2$  is also of the same order of magnitude as  $dt$ , which makes it necessary to retain this term in (8.6). A rigorous proof of Itô's lemma is rather complicated and is not needed for our purposes.

Diffusion processes play a special role in modeling oil prices. They provide a good balance between the complexity of the model and its tractability. On the one hand, simple ABM and GBM models are not sufficiently flexible to capture some important nuances of the oil price dynamics. On the other hand, many sophisticated models inevitably introduce too much complexity. As a result, their marginal value added is dwarfed by detrimental side-effects of an excessive parametrization that causes instability and often detracts from gaining valuable intuition.<sup>3</sup> The general diffusion framework outlined here appears to be a sweet spot in modeling oil prices.

---

<sup>3</sup>Some energy commodities, such as natural gas or power prices, are more susceptible to large short-term spikes where diffusions are often combined with jump processes, but this modeling paradigm is more complex as options can no longer be dynamically replicated with futures. We discuss some limitations of diffusions in Chap. 12.

It presents a relatively minor but extremely powerful extension of simple ABM and GBM modeling choices. Perhaps most importantly, the seminal argument of option replication by delta hedging easily extends to this more flexible class of diffusion models.

---

### 8.3 Delta Hedging and Option Replication

The breakthrough idea developed by Black and Scholes and independently by Merton was based on an important insight that follows from Itô's lemma. Since a financial derivative, such as an option, is a function of a random variable, Itô's lemma provides the rule relating a small change in the option price to a small change in the stock price. Importantly, both the stock and the option, which is a function of the stock, are driven by the same single source of uncertainty  $dz$ . Therefore, one should be able to combine the option and the stock in some smart way that eliminates this uncertainty, at least for a short period of time. If the entire risk can indeed be eliminated, then in the absence of an arbitrage, which rules out the existence of riskless profits, the value of a combined portfolio that includes an option and some quantity of the stock can only grow at the risk-free interest rate accrued on the initial investment.

The BSM framework was initially designed for the equity market and developed under the lognormal GBM assumption. Subsequently, it was tailored by one of the authors to the futures market.<sup>4</sup> The BSM argument, however, remains intact for all diffusions of the form (8.5), and we now replicate their framework in a more general setting.

Let  $C(F, t)$  denote the price of a call option struck at  $K$  that expires at time  $T$ . The call depends on the stochastic futures price  $F$  and time  $t$ . To simplify the notation, in this chapter we suppress the price dependency on  $K$  and  $T$ , which are set contractually. To construct a mini portfolio of an option and futures, we need to know how  $C(F, t)$  changes over a small increment  $dt$  in response to the change in the futures price  $dF$ . This change is described by Itô's formula (8.6). Applying it to the diffusion specification (8.5) results in the following stochastic equation for the option price:

$$dC = \left( \frac{\partial C}{\partial t} + \mu(F, t) \frac{\partial C}{\partial F} + \frac{1}{2} \sigma^2(F, t) \frac{\partial^2 C}{\partial F^2} \right) dt + \sigma(F, t) \frac{\partial C}{\partial F} dz \quad (8.7)$$

Note that both the futures price and the option price are driven by the same source of uncertainty  $dz$ . This allows us to combine the option and the future in a particular way that eliminates the source of randomness.

Assume that we bought a call option and want to offset some risks by selling a yet to be determined quantity of futures, denoted by the Greek letter delta,  $\Delta$ . We now

---

<sup>4</sup>Black (1976).

have a portfolio  $\Pi$  that consists of a long call option and a short position in  $\Delta$  units of futures:

$$\Pi(F, t) = C(F, t) - \Delta \cdot F$$

To calculate the change in the portfolio value  $d\Pi$  over time increment  $dt$ , we substitute  $dF$  and  $dC$  with their corresponding dynamics given by (8.5) and (8.7) and obtain that

$$d\Pi = \left( \frac{\partial C}{\partial t} + \mu(F, t) \left( \frac{\partial C}{\partial F} - \Delta \right) + \frac{1}{2} \sigma^2(F, t) \frac{\partial^2 C}{\partial F^2} \right) dt + \sigma(F, t) \left( \frac{\partial C}{\partial F} - \Delta \right) dz$$

We can now choose the number of futures hedges,  $\Delta$ , in a very special way that makes the stochastic term  $dz$  disappear from the equation. This can be accomplished by letting the hedging delta be:

$$\Delta = \frac{\partial C}{\partial F}$$

Such a special choice of delta does more than the elimination of randomness. It also removes the drift term  $\mu(F, t)$  from the first, deterministic part of the equation. This insight is the central part of the options theory. It shows that the price of an option does not depend on investor expectations, which are ultimately linked to individual risk preferences. The expectation term  $\mu(F, t)$  drops out from the equation and plays no role in the option pricing. This argument developed by Black, Scholes, and Merton could be viewed as a formal proof of Bachelier's original hypothesis that options must be evaluated under the principle of zero expectations to the speculator. Since the drift term is no longer present in the equation, the option price remains the same even if the expected value of the future price changes is set to zero. This approach became known as *risk-neutral* pricing of derivatives.

Without the random component  $dz$ , the resulting portfolio  $\Pi$  carries no risk, at least, instantaneously. In the absence of an arbitrage, such a portfolio can only accrue the risk-free interest rate  $r$ . If the portfolio were growing at any other rate, then traders would have been able to either borrow money and invest in the portfolio or short the portfolio and invest the proceeds, making money without any risk, which is deemed to be impossible. Therefore, over a small period  $dt$ , the change in the value of this riskless portfolio should be equal to the interest received on the initial investment:

$$d\Pi = \left( \frac{\partial C}{\partial t} + \frac{1}{2} \sigma^2(F, t) \frac{\partial^2 C}{\partial F^2} \right) dt = r\Pi dt = rC dt$$

Note that in the last term only the option premium  $C$  accrues interest. The futures contract does not require any initial investment, besides a small collateral, the impact

of which for simplicity is ignored here. Since it does not cost anything to enter into the futures contract, there is no interest accrued on the futures position either.

Cancelling the  $dt$  term, we obtain the generalized BSM equation for the option price written on futures, which follows a diffusion process with local volatility  $\sigma(F, t)$ :

$$\frac{\partial C}{\partial t} + \frac{1}{2} \sigma^2(F, t) \frac{\partial^2 C}{\partial F^2} - rC = 0 \quad (8.8)$$

This equation applies to any financial derivative on futures. The specific nature of the derivative is defined by its boundary condition at expiration  $T$ . For example, if the Eq. (8.8) is combined with the boundary condition defined by the payoff (8.1), then it always admits a unique solution that determines the price of the call option.

The Eq. (8.8) is analogous to the well-known equation of heat transfer that describes the dissipation of heat impulse over time with respect to the spatial variable that characterizes the medium. Here, the spatial variable is given by the futures price, real time is replaced with the time remaining to maturity of the option, and the initial condition is specified by the option's payoff at maturity. The behavior of option prices versus futures with the impulse provided by the strike price is analogous to the dissipation of temperature within the medium. The local volatility plays the role of the thermal conductivity of a non-homogeneous medium.

To solve the equation, in general, one must apply numerical algorithms, such as finite difference methods. Simple analytic solutions exist only for a few special but important cases. One special case, which underpins Bachelier's thesis, is the ABM process for which the local volatility is constant

$$\sigma(F, t) = \sigma_A$$

The solution to the Eq. (8.8) with the boundary condition (8.1) is then given by the following *Bachelier formula*:

$$C_{BC}(F, t) = e^{-rt} \{ (F - K)N(m_A) + \sigma_A \sqrt{\tau} n(m_A) \} \quad (8.9)$$

Here, standard notations are used for the normal Gaussian probability density with zero mean and variance equal to one:

$$n(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

and the cumulative normal distribution function, which is

$$N(x) = \int_{-\infty}^x n(y) dy$$

The quantity

$$\tau = T - t$$

represents time remaining to maturity, and

$$m_A = \frac{F - K}{\sigma_A \sqrt{\tau}}$$

defines the normalized moneyness of the option, which scales the distance between the futures price and the strike price by the total volatility over the life of the option. The normalized moneyness  $m_A$  can be understood as the number of standard deviations that an option is either ITM or OTM.

The Bachelier formula reduces to a particularly simple form for an ATM option, for which  $F = K$ :

$$C_{BC}(K, t) = e^{-rt} \frac{\sigma_A \sqrt{\tau}}{\sqrt{2\pi}}$$

This formula reflects Bachelier's original discovery that the value of a simple option must be proportional to the square root of time.<sup>5</sup> Without any doubt, the development of the entire modern probability theory was inspired by this result.

In Appendix B, we provide more details on deriving the Bachelier formula as the solution to the differential Eq. (8.8) by integrating the option's payoff with the normal probability density. Alternatively, one can simply verify that it indeed solves (8.8) by calculating partial derivatives of (8.9), which are also given in Appendix B, and substituting them directly into the Eq. (8.8). These derivatives play an important role in understanding the dynamics of option pricing. They are known as Greeks, as they are typically labeled by letters of the Greek alphabet.<sup>6</sup> We have already introduced *delta* in the derivation of the BSM equation, which is the first derivative of the option price with respect to the futures price. The second partial derivative of the option with respect to futures is *gamma*, which measures the degree of convexity in the option's payoff. *Theta*, the derivative with respect to time, shows how quickly an option value decays as time passes. *Vega* is the derivative with respect to the most important input into the pricing model, the volatility of the stochastic process.<sup>7</sup>

The second important special solution to the pricing equation for diffusions is given for the GBM case, where the local volatility is assumed to be proportional to the futures price:

<sup>5</sup>For the type of options studied by Bachelier, the discounting was not necessary as the option premium was netted against settlement at expiry. As a result, in his original derivation the interest rate was ignored. We will explain shortly why a similar assumption can be made for pricing many oil options. The discounting factor brings in an additional time dependency resulting from the present value of money, rather than from the evolution of the variance.

<sup>6</sup>Greeks are discussed in many standard derivatives textbooks, such as Alexander (2008) and Hull (2018). For their practical interpretation, see also Leoni (2014).

<sup>7</sup>Vega is not a letter of a Greek alphabet but it was adopted by option traders for its phonetic similarity.

$$\sigma(F, t) = \sigma_G F$$

The solution to the Eq. (8.8) with the boundary condition (8.1) is then given by the *Black formula*

$$C_{BL}(F, t) = e^{-r\tau} \left\{ FN\left(m_G + \frac{1}{2}\sigma_G\sqrt{\tau}\right) - KN\left(m_G - \frac{1}{2}\sigma_G\sqrt{\tau}\right) \right\} \quad (8.10)$$

The Black formula can also be derived by integrating the payoff with the lognormal probability density or, alternatively, it can be verified by the direct substitution of its partial derivatives into (8.8). The details are provided in Appendix B.

The normalized log-moneyness term  $m_G$  in (8.10) is defined as the logarithmic ratio of the futures price to the strike price scaled by the geometric volatility over the life of the option:

$$m_G = \frac{\ln(F/K)}{\sigma_G\sqrt{\tau}}$$

So far, we have only dealt with the valuation of a call option. However, the Eq. (8.8) applies to any derivative of the futures price. The fact that it is a call option was only specified by its boundary condition at maturity. If the boundary condition (8.1) is replaced with (8.2), then similar formulas can be obtained for a put option. However, an easier way to derive pricing formulas for puts would be to utilize an important no-arbitrage relationship that links prices of call and put options.

If we construct a portfolio which is long a call option and short a put option with the same strike  $K$ , then the payoff of this portfolio at expiration time  $T$  can be written as

$$C(F, T) - P(F, T) = F - K \quad (8.11)$$

In other words, holding this portfolio is identical to holding a long futures position established at the price  $K$ . Since the portfolio of a long call and a short put is itself a financial derivative of the futures, then the solution to (8.8) with the boundary condition (8.11) is simply the discounted futures payoff, reflecting the fact that option premia are typically paid upfront at the initiation of the trade:

$$C(F, t) - P(F, t) = e^{-r\tau}(F - K) \quad (8.12)$$

This formula, known as a *put-call parity*, allows one to determine the price of a European put option given the price of a European call and the price of futures. Alternatively, one can determine the price of a call given the price of a put along with futures.

We have chosen to repeat the standard BSM replication algorithm not because of its mathematical elegance, but rather because it provides an explicit recipe for trading volatility. This book is not meant to be a comprehensive reference of

derivatives pricing models. Our goal is to highlight certain features of these models that can be turned into profitable strategies, specifically in the oil market. Not all models, of course, present such unique trading opportunities. As such, to gain more clarity in already complex topics, we make some simplifying assumptions in areas where we do not see any particularly unique trading opportunities.

For the most part, we ignore the impact of the interest rate, assuming it to be zero. Unlike stocks and bonds, the interest rate plays only a relatively minor role in the valuation of options on oil futures. For European options that can only be exercised at expiration, the impact of the interest rate amounts merely to the discounting factor that appears as the multiplier in pricing formulas. The discounting comes from the fact that an option premium is assumed to be paid upfront, but the settlement of the option occurs later, at the expiration. In the oil market, the premium is paid upfront only for WTI options, while for exchange-traded Brent options, the premium is netted against the final settlement. Therefore, Brent options are marked-to-market like futures, in which case the discounting factor should be removed. It is interesting that this was also the case for options considered by Bachelier, which is why the interest rate was also omitted in his study.

Regular exchange-traded oil options are *American options* that can be exercised at any time prior to the expiration. Thus, an American option must be slightly more expensive than an equivalent European option. The precise calculation of an early exercise premium is rather complicated as the partial differential equation for the price of an option turns into an inequality, where one needs to solve a complex free boundary value problem or to apply other optimization techniques.<sup>8</sup> To simplify this, several analytical approximations have been developed, and the one developed by Barone-Adesi and Whaley has proven to be adequate for the oil market.<sup>9</sup> These models show that it is optimal to exercise an American option early only if the benefits from receiving a non-discounted cash settlement sooner outweigh an additional value from holding an option longer. The latter is driven by the remaining volatility. Since the impact of oil volatility on the option price is much larger than the contribution of the interest rate, it is rarely optimal to exercise an American option early, except for deep ITM options at times of high interest rates. Furthermore, under our simplified assumption of zero interest rates, the early exercise premium is equal to zero.

While the interest rate does matter for oil trading, using it generically in pricing formulas can cause more harm than it adds value. The interest rate for all traders is nearly always tied to the cost of funding specific to the trading company and to customized collateral agreements with clearing brokers and counterparties. Some strategies can indeed be enhanced by optimizing financing costs, but they are ultimately inseparable from the management of the trader's balance sheet. For example, when a long-dated ITM option has a significant intrinsic value that cannot be withdrawn, which is sometimes called the *trapped option value*, an auxiliary

---

<sup>8</sup>See, for example, Wilmott et al. (1993).

<sup>9</sup>Barone-Adesi and Whaley (1987).

agreement is often structured to use it as the collateral for other trades. The true value of the early exercise premium becomes heavily dependent on the details of such agreements and remains highly tailored to the needs of individual traders. In other words, this area is much closer to the field of structured finance than to the topic of oil trading, and for this reason we do not consider it in the book, mostly assuming zero interest rate, unless noted otherwise.

Finally, for some theoretical arguments we generally assume that the maturity of the option coincides with the maturity of the underlying futures. In practice, however, standard exchange-traded oil options expire three business days prior to the expiration of the corresponding futures. To simplify the exposition, in this chapter we ignore the relatively minor impact of the three-day maturity mismatch between options and futures and assume that both expire at time  $T$ . In Chap. 11, we will cover a more general case of options whose expiration is decoupled from the expiration of the underlying futures.

In the remainder of this chapter, we discuss two other types of volatility, realized and implied, both of which can only be understood in the context of particular simple modeling assumptions. Our initial goal is to highlight potential challenges and pitfalls with blind application of such simplified frameworks to the oil market. The actual recipes for trading that feature models specifically tailored to opportunities in the oil markets will be provided in the following five chapters of the book.

---

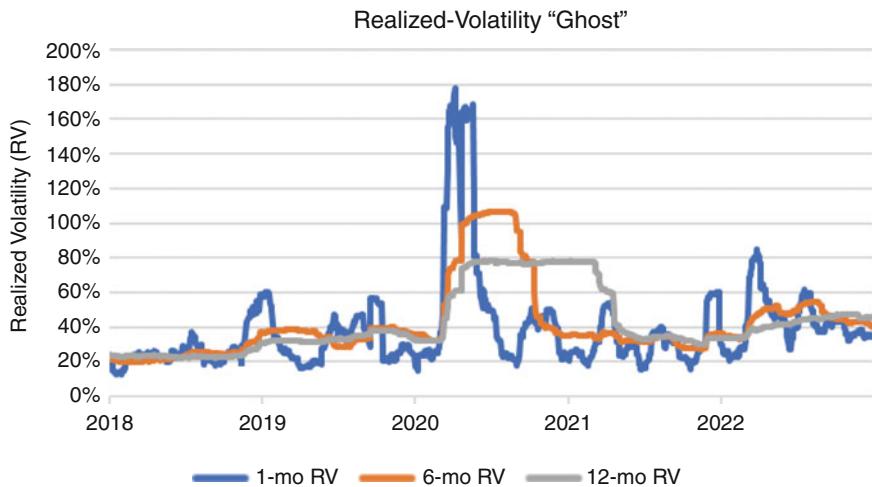
## 8.4 Realized Volatility

To price an option, we need to know the local volatility function of the underlying diffusion process. The local volatility is not directly observable, and it is rather difficult to estimate. Many traders start with an easier route and attempt to get some sense of how this function might look by measuring its traces from the historical time series of futures prices. Unfortunately, this route is prone to dangerous pitfalls. The purpose of this section is to caution traders against relying too much on backward-looking volatility estimates in valuation of forward-looking options. One should always remember that history is only one particular realization of an unknown stochastic process. History does not have information about the entire process. One can think of history as a single random path among thousands of other paths generated by a Monte Carlo simulation.

The volatility estimated statistically from the historical time series is known as the *realized volatility*. It is usually defined as the standard deviation of historical percentage returns over a particular lookback period. Since historical returns could be calculated over different frequencies and time horizons, the realized volatility is annualized.<sup>10</sup> The realized volatility is typically measured on the rolling basis, where

---

<sup>10</sup>For example, if daily prices are used then realized volatility is the standard deviation of returns multiplied by  $\sqrt{250}$  given approximately 250 trading days in a year, and for weekly prices, it is multiplied by  $\sqrt{52}$ .



**Fig. 8.2** Realized volatilities of third nearby WTI futures computed for one-month, six-month, and one-year lookback periods with highly visible “volatility ghost” in 2020

the size of the data sample is fixed, but each day the new return is added and the oldest one is removed.

The realized volatility is extremely sensitive to the choice of the lookback period. The shorter the lookback period, the more volatile the realized volatility is. The short-term realized volatility is highly variable due to the large sampling error, which makes it difficult to use for any investment decisions. The longer-term realized volatility is more stable, but such a slow-moving estimate is rarely suitable for traders, whose investment horizon tends to be much shorter.

Figure 8.2 illustrates different measures of the realized volatility for the third nearby WTI futures calculated using one-month, three-month, and one-year rolling windows.

The calculation of the realized volatility is also very sensitive to outliers, or particularly large price moves that have occurred in the past. The realized volatility calculated using a rolling window could drop nearly instantaneously when a single large historical return moves out of sample. To illustrate, consider different calculations of the realized volatility in 2020, shown in Fig. 8.2. The realized volatility for all futures contracts spiked during several consecutive days in April 2020 after spot oil prices went negative, and futures across all maturities moved down sharply. Subsequently, this event continued to haunt six-month realized volatility for exactly six months, and one-year realized volatility for exactly one year, leading to an abnormally high volatility estimate until it suddenly dropped when the event fell out of sample. Traders often refer to this phenomenon as *volatility ghost*.

This naïve method of estimating realized volatility does not differentiate between an event that occurred a while ago at the beginning of the sample and one that just happened yesterday. The calculation of the standard deviation weighs all historical

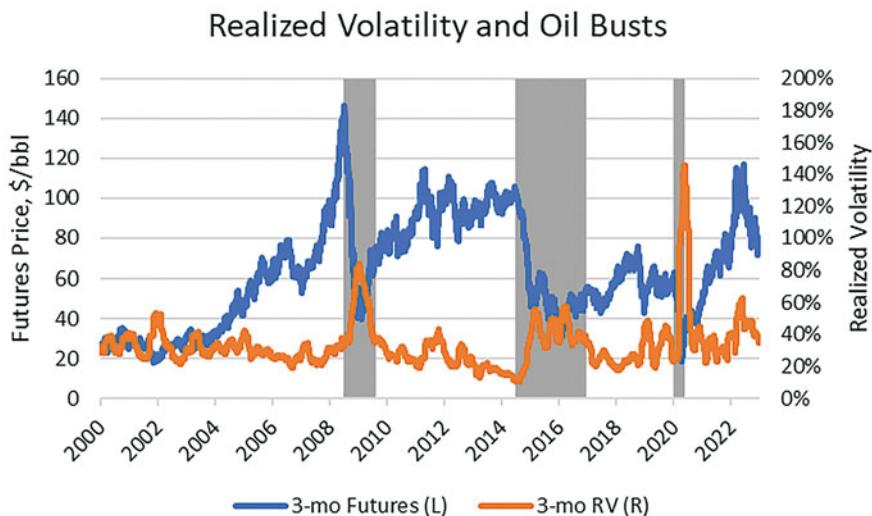
returns equally, which creates a dangerous problem if such an estimate is used for pricing an option. Some traders like to use the realized volatility calculated over the lookback period that matches the investment horizon, which for options is typically determined by time remaining to maturity. For example, to value an option that expires  $N$  days from today, they may use  $N$ -day realized volatility. This approach is very problematic. If a large price move happened exactly  $N$  days ago, then the realized volatility calculated tomorrow will be drastically different from the one calculated today, as this large price move drops out of sample. While the historical volatility calculation changes when the sample shifts by a day, the market expectations of the future volatility are unlikely to be very different from expectations the day prior.

The problem with equally weighted returns can be somewhat mitigated by using instead an exponentially weighted average of historical returns, where the largest weight is assigned to the latest return and prior returns enter the calculation with exponentially decreasing weights. In this case, the size of the lookback window becomes less relevant as the contribution of past returns exponentially approaches zero. The effect of gradually dissipating large prior returns makes the volatility estimate more representative of the current market conditions. The exponential parameter provides an additional degree of freedom that controls the persistence of the short-term shock.

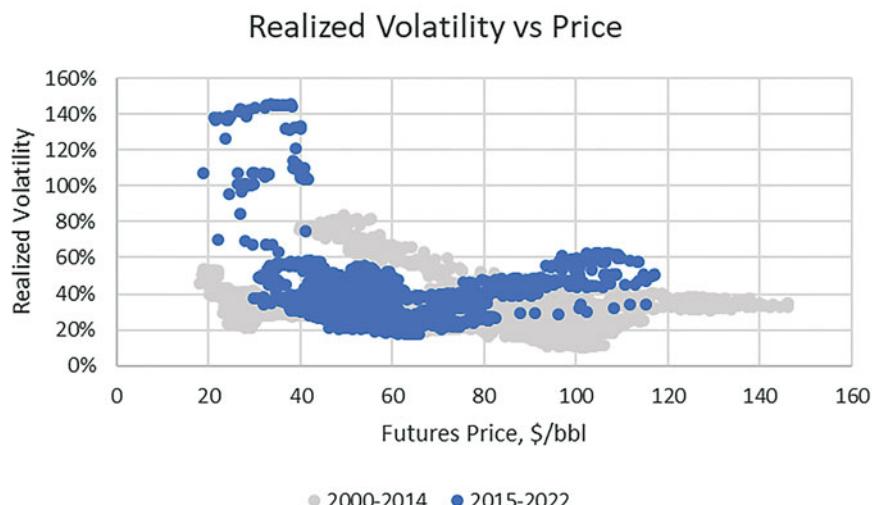
Many more sophisticated techniques have been proposed to improve volatility estimation and forecasting using historical futures data. However, these methods are not widely used by option traders in practice. At the end, the realized volatility is only an estimation method of the local volatility of the unknown stochastic process. For the most part, historical volatility estimates tend to be poor predictors of future volatility. An estimate of the realized volatility could mean different things for different price distributions. If the nature of the stochastic process that drives the underlying price distribution is not known, then any interpretation of the realized volatility could be rather dubious. Studying one realization of the unobservable stochastic process simply does not get us too far in terms of volatility forecasting regardless of how sophisticated the statistical model is.

Another challenge with the conventional application of a realized volatility estimate to oil options comes from measuring uncertainty in terms of percentage returns. Figure 8.3 shows that the volatility of oil returns has strong inverse dependency on the price level.

The inverse relationship between prices and percentage-based volatility measures is particularly vivid during the so-called oil busts. Simplistically, we define oil busts as periods during which the price of oil fell by more than 50% from its recent peak. There were three such periods since the beginning of the century. First, it happened in the aftermath of the global financial crisis, then at the time when OPEC unexpectedly increased production to protect its market share from the rapid growth of US shale, and finally when oil demand collapsed in the beginning of the Covid-19 pandemic. The realized volatility jumped during these periods of falling prices. This highlights an important difference between oil and many other consumption commodities, which tend to be more volatile when prices are high, and risks of



**Fig. 8.3** Realized 3-month volatility of third nearby futures (WTI, 2000–2022) spiked during the periods of oil busts during which the price decreased by more than 50% from the previous peak



**Fig. 8.4** Realized 3-month volatility of percentage returns versus third nearby futures (WTI, 2000–2022)

supply disruptions rise. In contrast, crude oil tends to be more volatile when prices are lower, at least, if the volatility is measured in percentage terms.

To see this effect even more clearly, Fig. 8.4 shows the relationship between realized volatility and oil prices in the form of a scattergram. The realized volatility as a function of price exhibits a strong negative skew, something which is more

typical for financial markets, such as equities. As mentioned previously, these markets follow an *up the stairs, down the elevator* dynamics which is characterized by frequent small gains and less frequent but large losses. When it comes to oil trading, measuring risk with volatility of percentage returns creates a strong artificial bias which is caused by looking at the uncertainty through the wrong lenses.

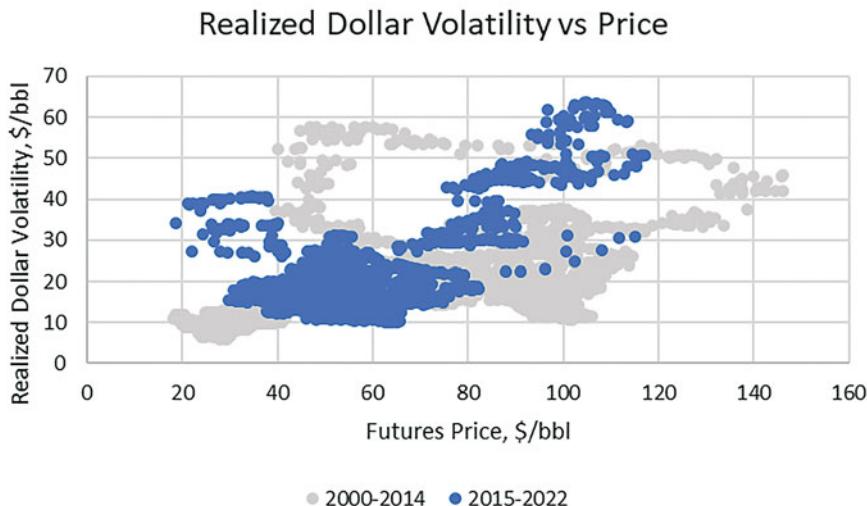
The problem with traditional percentage-based measures of volatility starts with the definition of an investment return, which is more ambiguous for trading futures. Since the only investment required to enter into a futures contract is a relatively small initial margin posted with the clearing broker, in theory, this posted margin should be used as an initial investment in the definition of return. In practice, margin requirements fluctuate along with prevailing risks and using the margin for the standardized denominator of a return would be rather cumbersome. Many financial analysts simply ignore this specificity of futures trading, and for convenience, use the notional value of futures in the definition of returns and subsequent calculation of volatility. While such percentage-based metrics make sense for financial assets that tend to grow over time, applying them to mean-reverting commodity prices creates an artificial framing bias that could adversely impact decision making. When oil prices are low, the division by a small number blows up the magnitude of percentage returns, and even makes calculation mathematically impossible if the asset price becomes negative.

Since the profitability of highly leveraged futures trading is measured by professional oil traders nearly exclusively in dollar terms, corresponding risks are better measured accordingly. What oil traders care about is the volatility of profits and losses driven by price changes, not the volatility of somewhat artificially constructed investment returns. If realized volatility is calculated instead as the standard deviation of price changes instead of percentage returns, then this results in a more accurate measure of uncertainty. We refer to such a measure as the *realized dollar volatility*.

As shown in Fig. 8.5, the artificial negative skewness of the volatility with respect to the price level disappears. The relationship between volatility of price changes and futures prices becomes more intuitive. The volatility tends to be relatively low when oil prices remain in some normal range, but it increases sharply when prices move away from such a range in either direction. In other words, the relationship between the realized dollar volatility and prices is somewhat parabolic, the insight that we will use later in developing an appropriate model for pricing oil options.

Figures 8.4 and 8.5 highlight the importance of choosing the correct lenses through which to look at oil volatility. One measure is rather poor, but it is often chosen for its operational convenience and for its easy standardization across asset classes. Another one is a more customized metric tailored to the specifics of the oil market. Fortunately, for many practical purposes it is rather straightforward to go back and forth between the two metrics. Since an investment return is defined as the ratio of a price change to the initial futures price, the volatility of returns can often be approximated by the ratio of the volatility of price changes to the price level.

This distinction between alternative ways of measuring volatility turns out to be vital for trading oil options. Switching the calculation of the historical volatility from



**Fig. 8.5** Realized 3-month volatility of price changes measured in dollars per barrel versus third nearby futures (WTI, 2000–2022)

percentage returns to price changes does not, of course, resolve any conceptual problems with the realized volatility concept, but, at least, it removes an artificial skewness of risk created by volatility of percentage returns. This skewness becomes even more visible and problematic when statistical estimates of the volatility of stochastic process are replaced with an alternative method of estimating volatility by deducing it from the options market.

## 8.5 Implied Volatility and its Skew

Any model of financial markets is only a theoretical construct designed to approximate the real behavior of prices. A typical financial model transforms a certain set of inputs by means of some quantitative machinery into a description of market prices. Regardless of how good the machinery is, the output of the model can only be as good as the quality of its inputs. The primary input to the models for option prices is the volatility of the stochastic process that governs the dynamics of the underlying futures contract. As discussed in the previous section, estimating this input from history leaves a lot to be desired, as historical realized volatility is all over the place. One can, of course, build more complex statistical models to forecast volatility, but regardless of the econometric technique, the resulting option price based on the volatility estimate is likely to be different from the one observed in the market.

When a choice is to be made between the model and the market, practitioners tend to assign higher powers to the market. To reconcile the model with the market, traders came up with a clever workaround and turned the option pricing problem

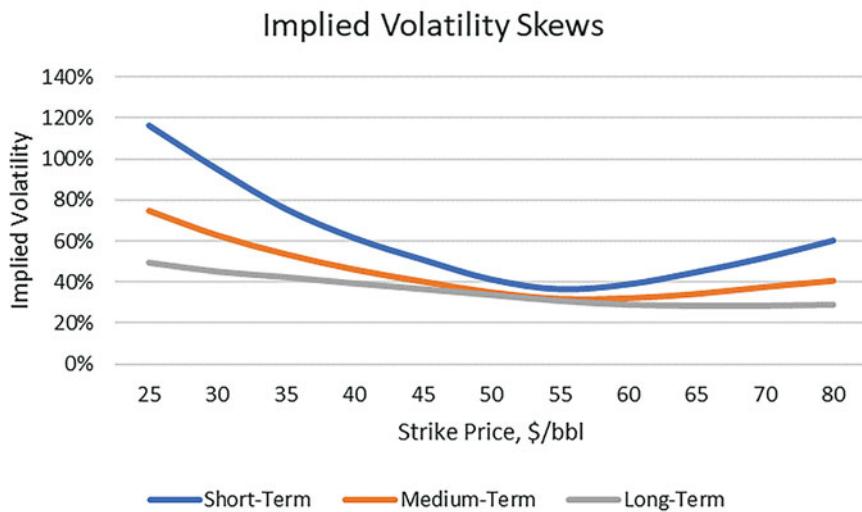
upside down. Instead of betting on the model driven by a noisy input, they use the model in reverse and back out the missing volatility input from the market price of an option. In other words, the pricing formula is matched to the observable option price and then inverted for volatility. The resulting measure of volatility is known as the *market-implied volatility*.

The general pricing formulas (8.9) and (8.10) cannot be analytically inverted for volatility, and implied volatility must be calculated numerically using a root-finding algorithm. Since the option's vega, which is the derivative of an option price with respect to its volatility input, is strictly positive, and an option price is a monotonically increasing function of volatility, such numerical inversion is always possible. Therefore, every option price is uniquely mapped to its corresponding implied volatility, and one can view implied volatilities as an alternative metric for option prices.

The market-implied volatility presents an alternative method of estimating some properties of unobservable local volatility of the stochastic process. In contrast to a backward-looking realized volatility estimate, implied volatility is a forward-looking measure. It is simply a plug to a particular option pricing formula that forces the model to match the market. Only in a very special and somewhat unrealistic case, when the futures market happens to behave exactly as prescribed by the model, can implied and realized volatilities be meaningfully compared. The implied volatility can then be interpreted as the market expectation of the future realized volatility. But even in this case, the two volatilities can still be very different if the option price contains an additional risk premium caused by hedging imbalances. One can also think of the implied volatility as the average of the expected local volatility, adjusted by the volatility risk premium. We will discuss these topics in more detail in subsequent chapters.

To distinguish the market-implied volatility from other types of volatility, we denote it by a different letter. For an option with the strike price  $K$  and expiration  $T$ , we use  $v(K, T)$  to represent the market standard *implied Black volatility* (IBV), which is computed by inverting the Black pricing formula (8.10). We will also use the less common, but arguably more important measure of *implied normal volatility* (INV) or, alternatively, *implied dollar volatility*, which we denote by  $v_N(K, T)$ . The INV is backed out from the same market price of an option by inverting the Bachelier formula (8.9). Regrettably, the market rarely explicitly attaches Bachelier's name to the normal volatility, and we reluctantly follow the market lingo and adopt the term normal volatility. In contrast to IBV, which is measured in percent, INV is expressed in dollars per barrel.

The irony of the market standard IBV metric is that it debunks the main assumption of the model that is responsible for its own creation. For a given maturity, there are many options with different strikes, but there is only one futures contract. The lognormal assumption of constant proportional volatility is a property of the futures contract, and it has nothing to do with options. If the Black model were correct, then the inversion of any option price should result in the same number, the same constant volatility of the futures that the model is based on. In practice, implied Black



**Fig. 8.6** Representative implied volatility skews for short-term, medium-term, and long-term oil options

volatilities computed for options with different strikes and maturities are nearly always different.

A typical dependency of implied Black volatilities  $v(K)$  on the strike price is shown in Fig. 8.6 for short-term, medium-term, and long-term oil options. The graph is called the *volatility smile*. The reference to a smile comes from its origins in the foreign exchange market, where the plot is often more symmetric with similar curvature on both ends, making it look like a smile. In the oil market, this graph is also known as the *volatility skew* to highlight its predominantly negative slope, except for the utmost right tail for shorter-term options, which makes the graph look more like a smirk. The curvature of the smile on both ends indicates that the options market expects a higher frequency and magnitude of extreme events than can be generated by the lognormal distribution. The market adjusts the pricing model by using higher volatility input to capture such events. We will discuss alternative normal volatility smiles and develop a more accurate model that captures the fat tails of the price distribution in Chap. 10.

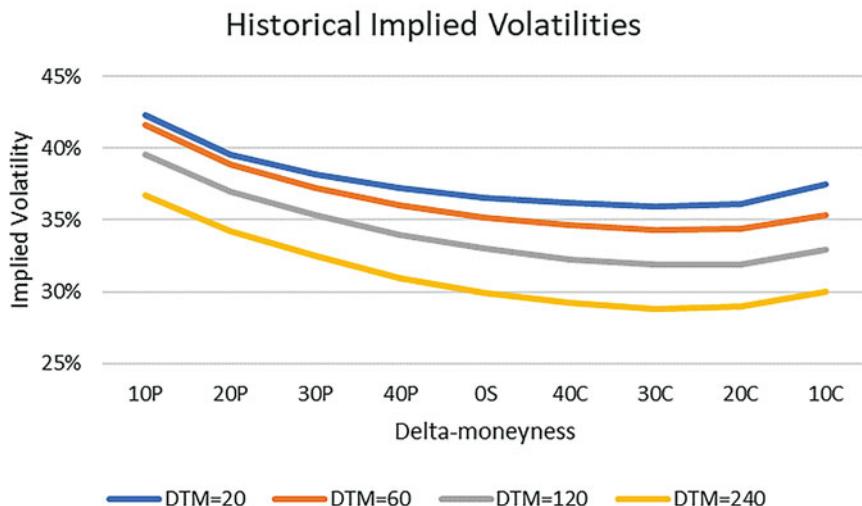
It is important to emphasize that the implied volatility smile is nothing more than a collection of option prices presented in a more convenient coordinate system. Simultaneous tracking of prices for all options with different strikes and maturities in dollars per barrel would be next to impossible, as option prices are constantly changing along with futures based on their corresponding deltas. In contrast to many statistical forecasting models, the main purpose of the Black model is to translate market observable option prices to a different and more convenient scale. Such a transformation does not require any statistical methods.

Even though the implied volatility smile has a negative skew with puts generally having higher implied volatilities than calls, this does not mean that puts are overpriced relative to calls. It is simply another artefact of looking at the market through incorrect lenses. The skewness of the IBV curve versus strikes is a side-effect of the same wrong unit of measure, previously illustrated by Fig. 8.4, where the realized volatility was plotted versus futures. If the true oil price distribution happened to be more symmetric, then observing a phenomenon through skewed lognormal lenses would force us to adjust our perception of option prices by raising percentage volatility for lower strikes and lowering it for higher strikes. The lognormal skewness is akin to prescribing lenses for astigmatism to someone who does not need them, which will result in blurry vision. In Chap. 10, we will construct more appropriate lenses to view the oil options market, but for now, we proceed down the conventional route of expressing option prices in terms of their IBVs.

Choosing the right metric is critical in trading as the frame of reference has a strong impact on traders' decision making. The volatility smile, shown in Fig. 8.6 as a function of strike  $K$ , works well as a static snapshot of option prices. However, this smile is more difficult to track dynamically once futures move, as the range of actively traded strikes also shifts. The most liquid option is typically ATM and the strikes for many other options are often chosen by traders relative to ATM, which makes it more practical to maintain the smile as a function of option moneyness instead of fixed strikes. The simplest way to define moneyness would be to use either the difference between the strike price and futures,  $K - F$ , measured in dollars per barrel, or their ratio,  $K/F$ , in percent. Such choices are intuitive and indeed both are often utilized for short-term analysis and for options with the same expiration. However, simple moneyness metrics become more problematic when comparing implied volatilities across multiple time horizons.

In both the Bachelier and Black formulas, the volatility is scaled with the square root of time to maturity,  $\sqrt{\tau}$ . Therefore, when we back out implied volatility from option prices, we always divide by  $\sqrt{\tau}$ . This magnifies the resulting smile for shorter-term options and flattens it for longer-term options, as clearly seen in Fig. 8.6. However, what matters for option pricing is volatility-normalized moneyness, such as  $m_A$  for the normal distribution, and  $m_G$  for the lognormal one. If we use such a normalized moneyness as the primary unit of measure, then volatility smiles will be more accurately compared across all maturities.

We could have stopped here and accepted a normalized moneyness setup as the base case for tracking the smile, but option traders often take it one step further and define moneyness instead directly in terms of their hedging deltas. The deltas for Bachelier and Black models, which are given in Appendix B, effectively transform volatility-normalized moneyness to make it vary between minus one and plus one. Keeping volatility as a function of delta has proven to be handy for traders as the required hedging ratio comes directly as a biproduct of the smile setup. For example, to hedge the sale of 100 units of 25-delta calls, one would need to buy 25 futures. Black deltas are often used as the market primary communication tool, even though traders typically make further adjustments to the actual delta hedging, which we discuss in more detail in Chap. 10.

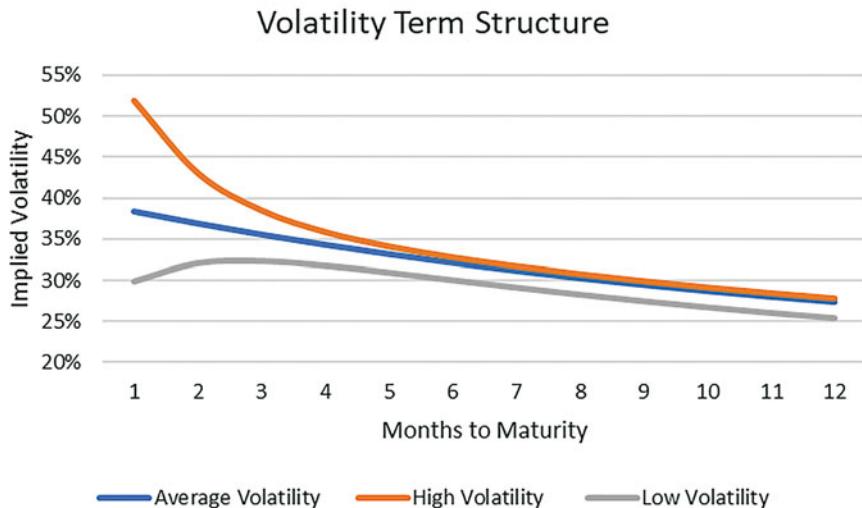


**Fig. 8.7** Average IBV versus delta-moneyness for different maturities (WTI, 2000–2022)

To illustrate historically observed oil smiles, Fig. 8.7 shows the average of IBV smiles for 10-, 20-, 30-, and 40-delta OTM puts and calls and zero-delta straddle since 2000. Note that because of an asymmetry embedded in the lognormal distribution, the Black delta for an ATM straddle is counterintuitively not equal to zero. To preserve the uniform spacing in the graph of the smile, we use a zero-delta straddle instead of an ATM straddle.

Average IBV smiles are shown for different days to maturity (DTM) that correspond to one-, three-, six-, and twelve-month options. Implied volatilities are always calculated using more liquid OTM options. Less liquid ITM options trade infrequently, and their prices are usually synthetically derived from corresponding OTM options and the put-call parity relationship (8.12). We should also acknowledge that maintaining volatilities as a function of delta has its own challenges, as such a setup is somewhat circular. While volatility is commonly measured versus delta, the delta itself depends on volatility. In practice, implied volatilities are first calculated for fixed strikes along with corresponding deltas, and then subsequently interpolated for the desired grid of deltas.

It is also clear from Fig. 8.7 that for each moneyness, implied volatility declines with increasing time to maturity. However, unlike the contradictory presence of the skew for any given futures, the decreasing term structure pattern of implied volatilities does not cause any alarms, as it reflects volatilities of different futures which do not have to be the same. In fact, the volatility should naturally decline for longer maturities futures as short-term fundamental uncertainty fades away while being smoothed out by the balancing role of storage. This phenomenon is known as



**Fig. 8.8** While oil volatility generally declines with time to maturity, the short-term implied volatility is more affected by realized volatility

the *Samuelson effect*.<sup>11</sup> The Samuelson effect, which was originally observed for the realized volatility, applies to the implied volatility as well. Under the normal market conditions, the term structure of implied volatilities is expected to be a decreasing function of time to maturity, as illustrated by the middle line in Fig. 8.8.

If short-term realized volatility is particularly high, which, for example, could be driven by falling oil prices, then the slope of the implied volatility term structure steepens, as long-maturity futures move less. This effect is exacerbated by a percentage-based volatility metric, as weaker prices generally result in a contango market when short-term futures prices fall below long-term futures. To compensate, the percentage volatility must be raised for contracts with lower prices. When volatility is measured instead in dollar terms, the volatility term structure is also typically decreasing but with a flatter slope. During periods of low realized volatility, the implied volatility term structure may also exhibit a visible hump. The hump may reflect not only market expectations of eventual increase in volatility, but also the risk premium embedded in oil options, which we will study in the following chapter. In Chap. 11, we will also see how critical the shape of the volatility term structure is for pricing many exotic options.

To summarize, so far, we have only adapted some standard methodologies for pricing options in the oil market. We chose to model futures prices in the setting of general diffusions characterized by the local volatility function. This framework has proven to be the sweet spot in the oil market, providing a reasonable trade-off between model accuracy and complexity. We introduced and discussed two

<sup>11</sup> See Samuelson (1965).

essentially defective, but, nevertheless, commonly used estimates of the local volatility of the diffusion process. The realized volatility estimates it by taking a look at the history, while the market-implied one attempts to extract it from an unknowable future. Both volatilities can only be understood in the context of simplified and largely unrealistic assumptions about the distribution of futures prices.

The next step often taken by many amateur traders and occasionally by some textbook writers is a comparison of one defective metric to another, as an attempt to determine whether an option is cheap or expensive. Our preference is to stay away from making such a naïve comparison, as in the volatile oil markets, this simplified approach can potentially bring more harm than value. While the difference between implied and realized volatilities could indeed provide some information about the richness of the option, the question of the fairness of the option price cannot be properly answered without a careful examination of the option's gamma, the measure of its convexity.

---

## References

- Alexander, C. (2008). *Market risk analysis, Vol. III: Pricing, hedging and trading financial instruments*. Wiley.
- Bachelier, L. (1900). *Théorie de la Spéculation*, Annales scientifiques de l'École Normale Supérieure, Serie 3, 17, 21–86.
- Barone-Adesi, G., & Whaley, R. E. (1987). Efficient analytic approximation of American option values. *Journal of Finance*, 42(2), 301–320.
- Black, F. (1976). The pricing of commodity contracts. *Journal of Financial Economics*, 3(1/2), 167–179.
- Black, F., & Scholes, M. (1973). The pricing of options and corporate liabilities. *The Journal of Political Economy*, 81(3), 637–654.
- Hull, J. C. (2018). *Options, futures, and other derivatives* (10th ed.). Pearson.
- Leoni, P. (2014). *The Greeks and hedging explained*. Palgrave Macmillan.
- Merton, R. C. (1973). Theory of rational option pricing. *The Bell Journal of Economics and Management Science*, 4(1), 141–183.
- Samuelson, P. A. (1965). Proof that properly anticipated prices fluctuate randomly. *Industrial Management Review*, 6(2), 41–49.
- Wilmott, P., Dewynne, J., & Howison, S. (1993). *Option pricing: Mathematical models and computation*. Oxford Financial Press.



# The Hidden Power of Negative Gamma

9

- End-users and speculators often evaluate options actuarially, like an insurance contract. Dealers, on the other hand, price options based on the cost of their dynamic replication. The pricing dichotomy results in alternative valuations and motivates trading among different market participants.
- Option delta hedging is riskless only in an idealized setting. In practice, the replication strategy is driven by the option's gamma. Oil gamma is dominated by producer demand for downside price insurance, which contributes to a negative skewness of futures returns.
- The volatility risk premium (VRP) can be extracted by selling and dynamically delta hedging oil options. Its magnitude depends on option moneyness and time to maturity. The strategy profitability is characterized by the VRP smile and the VRP term structure.
- Option traders say that they do not need a model for pricing. The price is given by a broker, but they need a model for hedging. The option replication model is particularly sensitive to the hedging frequency and to the volatility input used for delta calculation.
- VRP strategies evolve along with the changing needs of large market participants. As the market matured, directional VRP became less appealing, but new opportunities developed in trading relative VRP strategies across moneyness and maturities.

---

## 9.1 Options and Insurance

Options can be viewed as insurance contracts. A relatively small premium is paid upfront to ensure financial protection against some adverse events in the future. A typical insurance contract pays off only after a certain deductible is met, which is the responsibility of the buyer. The insurer and the insured are effectively sharing the risks, where the deductible sets the threshold for splitting their joint liability. A smaller deductible means that larger risks are taken by the insurer and a higher

insurance premium is charged for taking on such risks. In the jargon of commodity options, the equivalent of the deductible is the moneyness of the option, determined by its strike price.

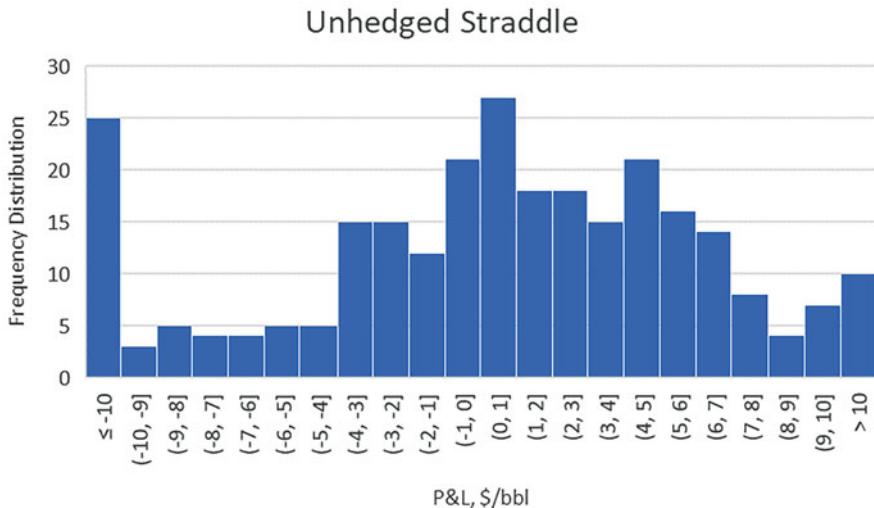
The seller of a commodity option, like the seller of an insurance contract, is exposed to highly asymmetric risks. The seller receives a steady income in the form of insurance premiums paid by the buyers. To motivate the seller to take on undesired risks, the premium includes some service surcharge on top of the fair price that reflects the risks. In the long run, selling insurance is generally a profitable business, as the willingness of buyers to pay up for the comfort of certainty allows sellers to dictate the price that provides sufficient incentive to be in this business. The insurance business thrives on diversification and the underwriter's ability to create sufficiently broad portfolios that can withstand large periodic losses. The competition among sellers is often constrained by the availability of capital and the limited appetite to assume such highly asymmetric risks.

The market price of an insurance product can be viewed as the sum of its fair price and the risk premium that compensates the seller for being in a business with a negatively skewed payoff. The fair price is determined by the amount that, in the long-run, is expected to offset periodic payouts to the buyers. The entire insurance business model hinges on its ability to accurately estimate the fair price by quantifying the risks of adverse events. The fair value of insurance is typically determined actuarially. A simple way to price an insurance contract actuarially would be to average historical payoffs to buyers from writing the same policy in the past.

Let us illustrate an application of this actuarial approach to the valuation of oil options. We consider the so-called naked straddle strategy that sells the benchmark three-month ATM straddle once a month with 60 days to maturity and holds the trade until its expiration without doing any hedging. The actuarial value of the straddle is then equal to its average historical payoff. The payoff of an ATM straddle is given by the distance between the initial strike level, determined by the futures price on the day of the trade, and the futures price on the day when the option expires. The risk premium is the P&L of this strategy. It is the difference between the option premium collected on entry and its actuarially determined fair value.

The frequency distribution of the P&L for the strategy of selling unhedged ATM straddles is shown in Fig. 9.1. Its summary statistics are also presented in the first column of Table 9.1.

The profile of this P&L distribution can hardly excite anyone to be in the business of selling naked oil straddles. On average, the seller would have collected \$8.36/bbl at the trade entry. Its actuarial value, or the average historical payout, however, is higher, at \$8.69/bbl, meaning that on average the futures moved more over the life of the option than the premium collected by the seller. Contrary to expectations of selling insurance being a profitable business, selling naked oil straddles would have lost, on average, \$0.33/bbl. Moreover, the P&L distribution is highly negatively skewed, with the maximum loss being much larger than the maximum gain. This is puzzling, as nobody should be writing insurance policies at a loss. In fact, the buyer, rather than the seller, of such insurance would have generated a 4% three-month



**Fig. 9.1** Frequency distribution of P&L (\$/bbl) for the strategy of selling unhedged 3-month ATM WTI straddles (2000–2022)

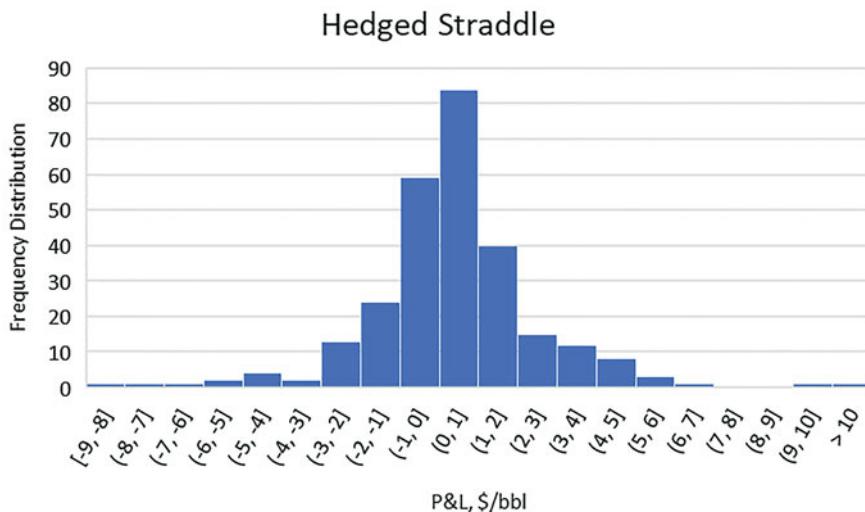
**Table 9.1** The summary statistics of the P&L distribution for two strategies of selling unhedged and daily delta-hedged 3-month ATM WTI straddles (2000–2022)

	Actuarial (Unhedged)	Volatility (Hedged)
Premium Collected	8.36	8.36
Fair Value	8.69	7.76
Mean P&L	(0.33)	0.60
Return on Premium	-4.0%	7.1%
Standard Deviation	9.86	2.19
Skewness	(2.09)	0.08
Excess Kurtosis	4.61	1.01
Minimum P&L	(47.71)	(8.81)
Maximum P&L	21.43	10.54

return on the option premium invested by systematically investing in three-month ATM oil straddles and holding them to expiration.<sup>1</sup> Why then do oil options appear to be so cheap?

The answer is that unlike traditional insurance, oil options are not priced actuarially by their dealers. The delta-hedging technique presented in the previous chapter allows volatility traders to significantly alter the shape of the resulting P&L distribution. Figure 9.2 shows the P&L distribution from selling the same ATM straddle

<sup>1</sup>Note that this analysis is presented on the trade level. Since the trades are entered once a month and held for three months, multiple trades are open at the same time. The analysis of the continuously managed strategy of selling options is presented later in this chapter.



**Fig. 9.2** Frequency distribution of P&L (\$/bbl) for the strategy of selling 3-month ATM WTI straddles (2000–2022) and delta hedging them daily

but instead delta hedging it. The summary statistics are presented in the second column of Table 9.1.

The difference between the two strategy profiles is striking. The losing strategy of selling naked straddles turns into a winning one. It generates an average profit of \$0.60/bbl, as the fair value of a delta-hedged straddle drops to \$7.76/bbl. On average, the option seller would have retained approximately 7% of three-month premiums collected, which is a reasonable compensation for the risks of writing an insurance-like product. Moreover, the overall P&L variance, the negative skewness, and the excess kurtosis are significantly reduced. Despite the contractual asymmetry in the option's payoff, the resulting P&L profile is much closer to a symmetric normal distribution.

We have a rather unique and a nearly magical situation here. Both the buyer and the seller of a three-month ATM straddle would have made good amounts of money by trading with each other, provided that they were prudent enough to manage the same trade differently. The buyer should have just left the option alone until its expiration, while the seller should have hedged it daily with futures. But what if a third trader comes to the market with another hedging strategy, for example, by hedging the same option at a different frequency (every other day, once a week, etc.). A fourth trader can come to the market and for some proprietary reasons choose something more esoteric, such as hedging the option every Thursday afternoon only if it rains. When it comes to options, different hedging strategies lead to different fair values, and the concept of the fair value is no longer uniquely defined. The existence of such alternative valuations is the essence of volatility trading. What appears to be fair to one trader may not look so fair to another, and such disagreement is what motivates them to trade.

The P&L of the naked straddle strategy is easy to understand. It is conceptually similar to the outcome of momentum and trend-following strategies. If the futures price moves sufficiently far away from its initial level determined by an ATM strike, then the buyer wins, and the seller loses. On the other hand, if futures prices remain range-bound, then the seller is more likely to win, and the buyer is more likely to lose. For the delta-hedged strategy, understanding P&L is trickier, as it becomes model dependent. Delta hedging only eliminates the futures price risk instantaneously based on the slope, or the first derivative, of the option pricing function with respect to the futures price. This is akin to approximating a nonlinear function with a straight line. The accuracy of such an approximation depends on the degree of the option's convexity. The convexity is measured by the second derivative of the option price with respect to the futures price, which is known as the option's gamma.

---

## 9.2 The Most Powerful Option Greek

As elegant as the idea of dynamic option replication is, we should accept that it is an idealization. It is derived under two major simplifications, which are not feasible in the real market. The theory assumes that trading can be done continuously without any friction, and that the option's risk can always be instantaneously offset by the futures position, which is determined by some perfect foresight of the option's delta. In the real world of financial markets, nothing can be done continuously, which results in some slippage error and transaction cost. An even larger error comes from the uncertainty in the calculation of delta. The only delta that truly eliminates the option risk is the one computed using the future local volatility, something that one can only guess ahead of time. In other words, the practical implementation of the strategy cannot be riskless. We now move from the theoretical lab of perfect delta hedging to the real world of running such a strategy.

Consider a strategy of selling a put option to an oil producer and delta hedging it until expiration. Its market price  $P(F, t; v)$ , which can also be quoted in terms of its implied Black volatility  $v$ , is determined by supply and demand between buyers and sellers of this option. Let us assume that the futures market follows the diffusion process (8.5) with the local volatility function  $\sigma(F, t)$ . The market, of course, does not know  $\sigma(F, t)$ , as it is unobservable. In fact, if there are more buyers than sellers, then the market might be willing to pay more than the option's fair price, which is determined by the solution to the diffusion Eq. (8.8) with the boundary condition (8.2).

Since the risk of writing a put option comes from falling futures prices, we can reduce this risk by also selling some quantity of futures  $\Delta < 0$ , so that the total portfolio is given by

$$\Pi = -P(F, t; v) + \Delta \cdot F$$

We immediately face a challenge. In the traditional delta-hedging framework that we are planning to use, delta depends on future realized volatility, which we, like the

rest of the market, do not know. Since we need to know delta to hedge the option and delta depends on volatility, if we are wrong on the future volatility then we will be wrong on delta, and therefore, our hedging strategy is no longer riskless.

Let us see what happens to P&L if the actual volatility  $\sigma(F, t)$  deviates from the constant implied volatility  $v$ . The option price  $P(F, t; v)$  is a known Black formula that was made to fit the market price of the option by tweaking the model parameter  $v$ . The formula is a function of the stochastic variable  $F$ . Therefore, like in the previous chapter, the small change  $dP$  in the value of this function over a time increment  $dt$  is determined by Ito's lemma. The P&L of the portfolio during this time period is then given by

$$\begin{aligned} d\Pi = -dP(v) + \Delta \cdot dF &= -\left( \frac{\partial P(v)}{\partial t} + \frac{1}{2} \sigma^2(F, t) \frac{\partial^2 P(v)}{\partial F^2} \right) dt \\ &\quad + \left( \Delta - \frac{\partial P(v)}{\partial F} \right) dF \end{aligned} \tag{9.1}$$

Now we need to make the choice of what delta to use to hedge the option. Even though we believe that the implied Black volatility  $v$  may not be accurate, we also notice that the only way for us to eliminate the futures risk  $dF$  in (9.1) would be to hedge the option with deltas calculated using the same implied Black volatility. Thus, we choose our delta as

$$\Delta = \frac{\partial P(v)}{\partial F}$$

The fact that the only delta that instantaneously removes the futures risk  $dF$  happens to be the one that is calculated using the wrong volatility may sound a bit odd. This subtle but important point deserves a short diversion. If somehow, we knew in advance what the future actual volatility  $\sigma(F, t)$  would be and hedged the option continuously using the corresponding deltas, then the option payoff could be perfectly replicated. In such a purely hypothetical case, we would not even care what the market thinks about daily volatility and what our P&L is on a daily basis, because the *terminal* P&L would be known with certainty. The terminal P&L is the difference between the sale price  $P(v)$  and the cost of the option's replication given by the diffusion (8.8) with the local volatility  $\sigma(F, t)$ . However, even though hedging based on perfect foresight ensures certainty of the terminal P&L, on a daily basis P&L will fluctuate along with  $dF$ .<sup>2</sup> This is because the market does not know anything about our unique and exclusive foresight of the future volatility, and it is not pricing the option correctly. We will illustrate this argument shortly with specific strategy examples that use different hedging deltas. For now, we take the easiest conventional route and hedge the option using deltas computed with the implied volatility  $v$ .

---

<sup>2</sup>This point is well covered in Derman and Miller (2016).

With the contribution of  $dF$  instantaneously removed by the delta hedge, we now have that

$$d\Pi = - \left( \frac{\partial P(v)}{\partial t} + \frac{\sigma^2(F, t)}{2} \frac{\partial^2 P(v)}{\partial F^2} \right) dt$$

This equation is similar to the one obtained in the derivation of the pricing equation in the previous chapter with the exception of one important difference. Here,  $P(v)$  represents a particular theoretical formula that does not know anything about the real market dynamics characterized by  $\sigma(F, t)$ . However, what we also know is that this theoretical formula  $P(v)$  happens to satisfy the BSM equation with constant geometric volatility  $v$ , which can be written as follows:

$$\frac{\partial P(v)}{\partial t} = - \frac{v^2 F^2}{2} \frac{\partial^2 P(v)}{\partial F^2}$$

We then substitute this expression into the formula for  $d\Pi$ , which leads to the following P&L of the portfolio over period  $dt$ :

$$d\Pi = \frac{1}{2} (v^2 F^2 - \sigma^2(F, t)) \Gamma_{BL} dt \quad (9.2)$$

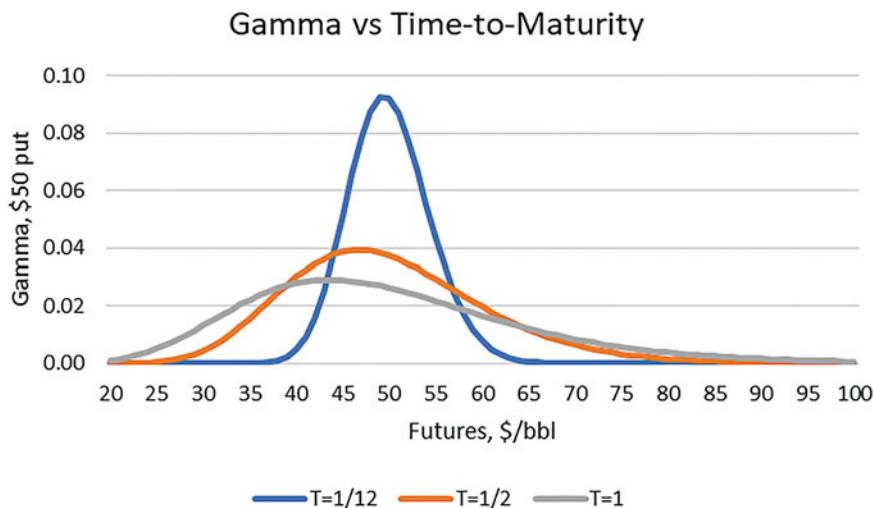
Here,  $\Gamma_{BL}$  denotes Black gamma, or the second derivative of the Black formula for the put option with respect to the futures price, which is given by:

$$\Gamma_{BL} = \frac{\partial^2 P(v)}{\partial F^2} = \frac{n \left( m_G + \frac{v\sqrt{\tau}}{2} \right)}{v F \sqrt{\tau}}$$

We should also note that call and put options with the same strike have the same gamma. This can be easily seen by differentiating the put-call parity relationship (8.12) twice with respect to the futures price. For volatility dealers, puts and calls with the same strike carry essentially the same risk.

We are now in a good position to see why option sellers tremble when they hear the term gamma. If the actual price moves on a given day happens to be smaller than the one implied by the option market, then the seller makes some money, but the gain is limited to the daily option's time decay, i.e., its theta. If the market moves exactly one standard deviation, as measured by the option's implied volatility, which is often referred to as the daily breakeven point, then the daily P&L is zero. However, if futures move  $N$  standard deviations, then the option seller's losses are proportional to  $N^2 - 1$  multiplied by the option's gamma. Things could quickly become troublesome for the option seller if a large market move occurs at a time when the options gamma is particularly large.

The option's gamma is very dynamic, as it varies significantly with the moneyness of the option and time remaining to maturity. To illustrate, Fig. 9.3

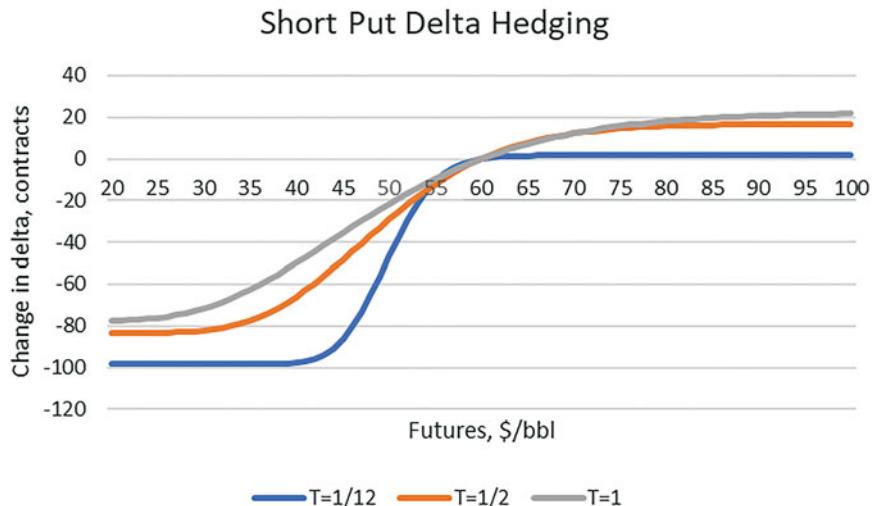


**Fig. 9.3** Black gamma for the \$50/bbl put option with  $v = 0.30$  and different times to expiration

shows Black gamma for one-, three-, and twelve-month put options struck at \$50/bbl.

Algebraically, Black gamma is simply the lognormal probability density function with the mean at the strike price. It is highest for an option near ATM, and the peak gamma rises as time to maturity shrinks. In the limiting case, the gamma of an ATM option right before its expiration approaches infinity. Imagine now the scenario of a large futures price move that unexpectedly crosses the strike price of the option at expiration. The losses for the option seller determined by the difference between realized and implied volatility are magnified by a large gamma. This is where the big money in option trading is made or lost, and why short-term volatility trading strategies are deservedly named after this powerful Greek letter.

In addition to suffering losses from a large price move, an initially delta-hedged option portfolio quickly loses its directional immunity, and the residual delta must be promptly rebalanced. Delta hedging can only keep the portfolio neutral with respect to futures instantaneously. As soon as futures move, the desired risk neutrality disappears, as delta itself changes according to its gamma profile. Gamma is the second derivative of the option's price, or alternatively, it is the first derivative of the option's delta. If the futures price falls and moves towards the strike of a short OTM put option, then the put delta becomes more negative, making the portfolio directionally longer futures. Additional futures must be sold to bring net delta back to zero. Likewise, if the futures price rises and moves away from the strike, then the put delta becomes less negative, making the portfolio overall short. Some short futures hedges which are no longer needed must then be bought to bring net delta back to zero.



**Fig. 9.4** Incremental futures required for rebalancing 100 short put options with  $K = 50$ ,  $v = 0.30$  and different times to expiration. Initially,  $F = 60$  and the portfolio is instantaneously delta neutral

The short gamma trader must always sell futures on the way down and buy them back on the way up to keep the portfolio delta neutral. The quantity of futures needed for rebalancing is determined by the combined impact of the option's gamma and the magnitude of the futures price move. Figure 9.4 shows the number of futures contracts needed for rebalancing the delta of 100 contracts of short put option struck at  $K = 50$  with different times to maturity. The options have gamma profiles, as in Fig. 9.3, but taken with the negative sign to represent the short position. The initial futures price is  $F = 60$  when the hedged portfolios are delta neutral.

The gamma here is observed as the change in the slope of the delta function. If the market moves down, then gamma increases as futures move towards the strike. In this case, dealers have no choice but to sell more futures to remain delta neutral. Larger gamma leads to more aggressive selling of futures by volatility dealers as the futures price drops. On the other hand, if futures move up, drifting further away from the short strike, then the need for rebalancing becomes more muted. Such an asymmetric gamma profile is very typical for the oil options market, which is dominated by producer hedging demand for downside protection. On the upside, the overall industry gamma profile is more balanced, as some producers gain leverage by selling options to the market, a topic that we discuss in more detail in the next chapter.

The rebalancing of short gamma portfolios is one of the most powerful forces that drives the market for the underlying futures. This process is completely mechanical. It has little to do with volatility dealers' own opinion about the direction of oil prices. When futures fall, dealers who are short puts to producers must sell futures to remain within tight volumetric risk limits on residual deltas, prudently imposed by their risk

managers. The larger the gamma, the more futures they need to sell, pushing the price further down. This creates a vicious cycle when prices fall in a downward spiral until strikes are crossed and dealers' short gamma exposure starts subsiding. The gamma hedging profile can also be viewed as an alternative to the reaction function previously introduced for the momentum strategy.

The insurance-like business of selling delta-hedged options, which mandates selling futures on the lows and buying them back on the highs with potentially unlimited losses, may appear to be a questionable-value proposition for investors. This perception is driven by cognitive biases that cause painful financial losses to be overweighted relative to their frequency and magnitude. In the meantime, every day when nothing major happens in the market, and the futures price moves less than its daily breakeven, the option seller retains a small portion of the option premium. The positive P&L on uneventful days could easily add up to a substantial buffer which more than offsets periodic large losses.

Is this strategy just an example of picking up pennies in front of the steamroller, or can it be turned into a viable investment with its risk kept under control? The answer is not so straightforward. It turns out that one needs to be more selective on which options to sell, and which ones are better to stay away from. We address this important topic in the following sections with a comprehensive empirical study of the *volatility risk premium (VRP)* strategy in the oil market.<sup>3</sup>

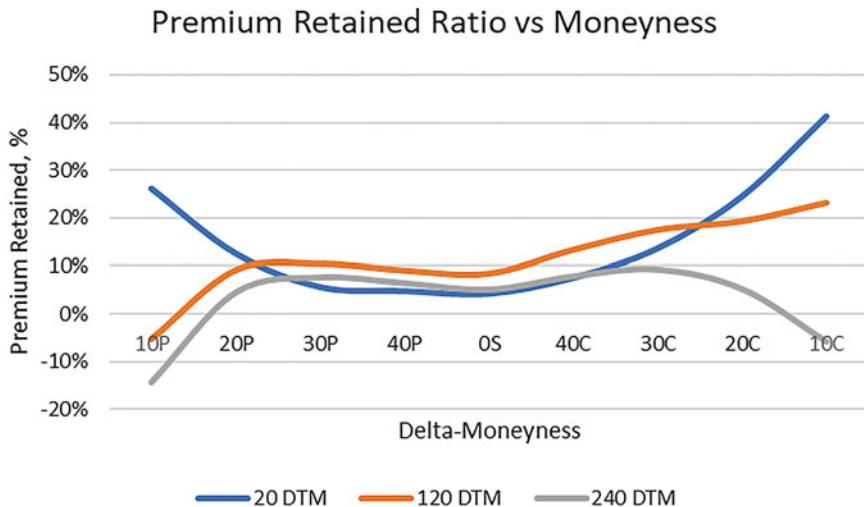
### 9.3 The Smile of the Volatility Risk Premium

To test the existence of the structural risk premium embedded in prices of oil options, we look at the long-term performance of the VRP strategy, starting from its earliest days when oil options gained sufficient liquidity. In the base version of the VRP strategy, one sells a particular option and hedges it daily using deltas computed with conventional implied Black volatilities. The VRP profitability is analyzed for options with different moneyness and times to maturity. Then VRP trades are combined into a portfolio that can be analyzed as a continuously managed systematic risk premium strategy. In the following section, we present VRP implementations that are based on alternative delta-hedging techniques.

When analyzing results of option-based strategies one must start with a careful choice of appropriate performance metrics, which could differ from conventional metrics used for studying stocks, bonds, and futures. Consider first the investment return on an option trade. If an investor buys an option, then the return is defined in

---

<sup>3</sup>The existence of the volatility or variance risk premium in oil options has been documented in several studies, see Doran and Ronn (2006, 2008), Trolle and Schwartz (2010), Kang and Pan (2015), Prokopczuk et al. (2017), Ellwanger (2017), and Jacobs and Li (2023). The analysis of VRP as the actual trading strategy, which we present here, is a much more difficult task due to its extreme path-dependency and non-straightforward measurements of risks. The remainder of this chapter is based on empirical results of Bouchouev and Johnson (2022). The author would like to thank Brett Johnson for his highly valuable contribution to the data analysis.



**Fig. 9.5** VRP smile, as the premium retained ratio versus moneyness with daily delta hedging (WTI, 2000–2022)

the traditional way, where the denominator of the return reflects the initial investment, i.e., the option premium. This is the maximum amount that the buyer can lose.

For selling options, the maximum loss is technically unlimited, and additional capital must be set aside by the seller with a clearing broker to ensure that all contractual obligations will always be met. However, incorporating such highly variable and trader-specific funding requirements would significantly distract us from our primary objectives. To keep the presentation transparent and intuitive, we keep the simple definition of the return, but address the risks of an option-selling strategy that drive additional capital requirements separately. We define the return on selling an option as the negative of the return on buying the option. This also means that the return on the VRP trade is the percentage of the premium retained after hedging.

Since selling options is akin to selling insurance, and option moneyness corresponds to an insurance deductible, it is important to analyze VRP performance for options with different moneyness. Using moneyness as a function of an option's delta, as defined in the previous chapter, the performance of VRP strategies is measured for 0.10-, 0.20-, 0.30-, and 0.40-delta OTM puts and calls and for zero-delta straddle. Option tenors are defined by the number of trading days remaining to maturity (DTM) counting from the day when the option is sold. We use increments of 20 DTM that approximately correspond to a monthly option expiration schedule with roughly 20 trading days per month.

Figure 9.5 presents historical VRP returns, measured by the percentage of the option premium retained with daily delta hedging across different delta-moneyness for various investment horizons. In the spirit of an implied volatility smile that

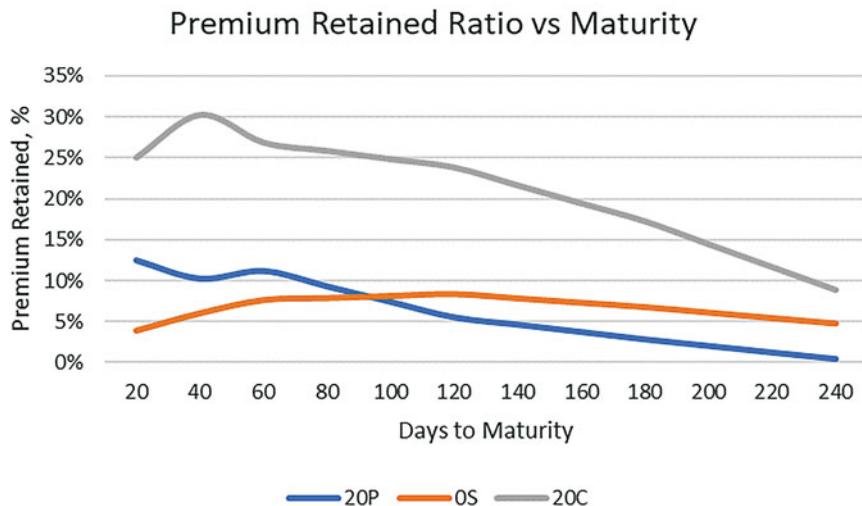
describes option prices for different moneyness, we label Fig. 9.5 the *VRP smile*. The VRP smile characterizes the long-term historical profitability of selling delta-hedged options for different moneyness, analogous to returns on selling insurance with different deductibles.

For short-term one-month options, the graph does indeed look like a smile. It confirms that a catastrophe-type oil insurance with the highest deductible that only pays off in the case of an extreme price move is the most overpriced one, at least, if the profit is measured as a fraction of the option premium collected. In other words, the frequency of large futures moves implied by the options market is much higher than the frequency of their actual occurrences. Either market participants overestimate the likelihood and the magnitude of such events, or, more likely, they have no choice but to compensate writers of deep OTM options for the higher risk taken per unit of premium collected. The short-term VRP smile appears to be relatively symmetric, confirming the higher risk premiums embedded in market prices for both OTM calls and OTM puts.

For medium-term options, the VRP smile exhibits a higher degree of asymmetry. The smile flattens for OTM puts but it remains steep for OTM calls. The call risk premium is often driven by geopolitical uncertainty and risks of supply disruptions. However, this risk is not realized to the extent that it is priced in the options market. In contrast, the risk premium embedded in put options, which is typically driven more by macroeconomic weakness and occasional spillover from falling equity markets, turns out to be more modest.

For longer-term options, VRP profitability erodes for both puts and calls. For the most part, this is explained by the increasing denominator in the definition of the return, as option values rise with time to maturity. However, the profitability also declines even if the performance is measured in absolute terms, i.e., in dollars per barrel. In other words, it does not pay to sell options too early when the options' time decay is too small to adequately compensate for the prolonged risk exposure. Note also that selling deep OTM medium and long-term put options results in trading losses, or, alternatively, buying such options becomes profitable. This is another manifestation of the negatively skewed realized price distribution. However, the conclusions for longer-dated deep OTM options should be interpreted with a greater degree of caution given the lower liquidity and the noisier historical option data.

It is useful to contrast the shape of the VRP smile to the typical shape of the implied volatility smile. The latter, as seen in Fig. 8.7, exhibits a distinctively negative put skew. At the first glance, such skew may create an illusion that oil puts are overpriced, like they are, for example, in the equity market. However, a positively skewed VRP smile shows that oil puts are surprisingly much more reasonably priced than oil calls if they are measured relative to their corresponding fair values. Despite the fact that implied volatilities for most calls trade at discounts to ATM volatility, call options are, nevertheless, still expensive. For calls to be priced fairly, the implied call skew should be even more negative. This is because the presence of an implied volatility skew is a side-effect of looking at option prices through the lens of the lognormal framework, which, as illustrated in the previous chapter, fails to capture important properties of the oil price distribution. A more



**Fig. 9.6** VRP term structure, as the premium retained ratio versus maturity with daily hedging for 0.20-delta puts (20P), zero-delta straddles (0S), and 0.20-delta calls (20C) (WTI, 2000–2022)

accurate pricing model that corrects for such distortion will be introduced in the next chapter.

Figure 9.6 presents results of the historical VRP performance versus time remaining to maturity, which we define as the *VRP term structure*. For both OTM calls and OTM puts, the term structure of VRP, measured again by the premium retained ratio, resembles the term structure of implied volatilities. Both curves generally decline with increasing DTM. However, for zero-delta straddles the VRP term structure exhibits a noticeable hump. This indicates that selling medium-term ATM options is more profitable than selling either short-term ATM options that have higher gamma or long-term options that maintain risk exposure for too long. We will see shortly that this hump is magnified when the analysis is extended to the portfolio of short oil options.

While the historical returns of VRP strategies appear to be attractive, returns only represent one side of the investment analysis. The other side, which is arguably even more important for short option strategies, is the amount of risk that must be taken to achieve such returns. Here, we need to make an important distinction between the risk characteristics of individual trades, which we define as positions, and the risks of the portfolio made up of such positions. Since standard oil options are only available with monthly maturities, this adds some complexity to the analysis of risk-adjusted returns because of the overlapping nature of multiple positions within each portfolio that remain open at the same time. While returns generated by portfolios with different maturities can be easily annualized and compared to each other, the

annualization of risks resulting from individual trades with different maturities is prone to statistical anomalies, especially when trades are highly correlated.<sup>4</sup>

To properly measure the relative performance of VRP strategies across different time horizons, we construct continuously run systematic strategies made up of corresponding VRP positions with the same moneyness and the same DTM at the time of the option sale. Each one-month portfolio (20 DTM) consists of only one position, as the new trade is initiated within a few days of the expiration of the previous one. However, it is not the same for portfolios with longer DTM. The two-month (40 DTM) portfolio has two open trades with expirations approximately twenty days apart, the three-month (60 DTM) portfolio has three open trades, etc., and the twelve-month (240 DTM) portfolio consists of twelve open trades. Only one option in a portfolio expires each month. Therefore, all portfolios are exposed to the same futures risk at the expiration of this option. However, since a new option is added to the portfolio every month at the moneyness level prevalent at the time, the strikes of options within the portfolio are spread out, providing some valuable risk diversification benefits to the portfolio of options.

For a proper apples-to-apples comparison of VRP investments across time horizons, we construct continuous equity lines by cumulating daily P&L from all open positions within the portfolio. We can then calculate traditional investment characteristics for each portfolio, such as the Sharpe ratio.<sup>5</sup> Figures 9.7 and 9.8 display Sharpe ratios for various VRP portfolios by moneyness and maturity.

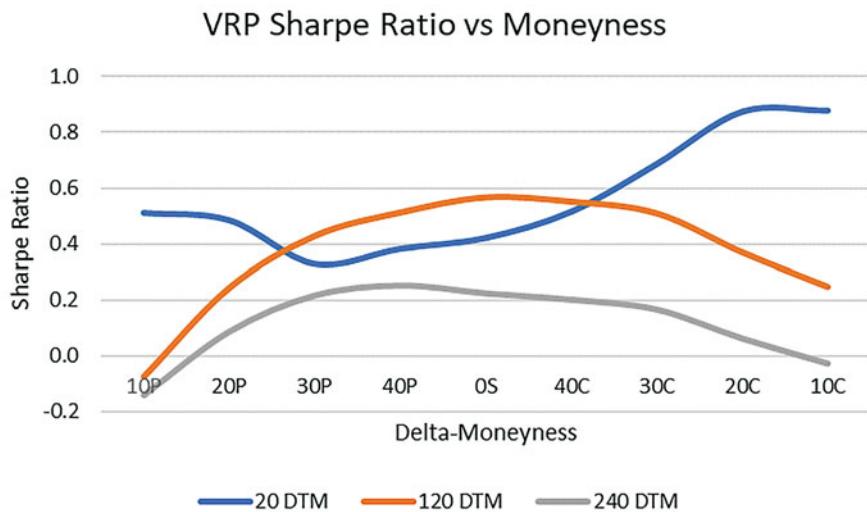
The main difference between the portfolio VRP curves, shown in Figs. 9.7 and 9.8, and position VRP curves, shown in Figs. 9.5 and 9.6, is the materially improved performance of zero-delta straddles. ATM options, being highly exposed to large gamma risks, benefit the most from the strike diversification. The benefits of diversification for OTM options are more muted. The reason again lies in the crucial role of the option's gamma, which becomes infinite if the futures cross the strike price at the option's expiration. Since strikes near ATM are more likely to be crossed than strikes further OTM, the benefits from diversification of the strike risk are more substantial for ATM options.

Similarly to other systematic risk premia strategies, oil VRP clearly demonstrates the existence of an investment edge, but its actual implementation is highly customized by professional volatility traders. In the next section, we provide some examples of such practical implementations. These implementations explicitly incorporate the impact of transaction costs and optimize VRP by combining it with short-term biases in the behavior of the underlying futures.

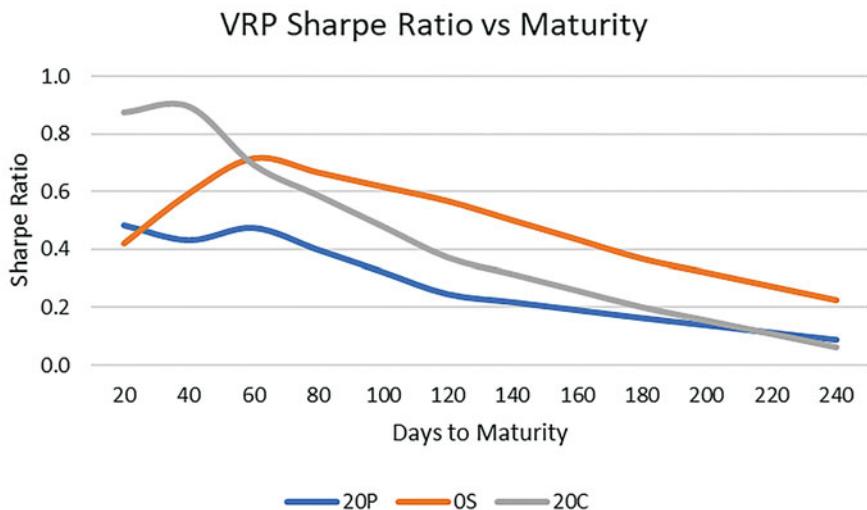
---

<sup>4</sup>The challenges that arise with annualization of risk-adjusted returns are well explained by Lo (2002).

<sup>5</sup>As before, we calculate the Sharpe ratio assuming zero risk-free interest rate. Since we do not include the cost of capital and exchange margin in our calculations, we do not include any interest that could be received on the initial margin either. Overall, the impact of financing costs on the performance of VRP strategies is relatively minor. In addition, both the numerator and the denominator of the Sharpe ratio here are computed in dollars per barrel, not in percent.



**Fig. 9.7** Sharpe ratios of VRP portfolios by moneyness (WTI, 2000–2022)



**Fig. 9.8** Sharpe ratios of VRP portfolios by maturity for 0.20-delta puts (20P), zero-delta straddles (OS), and 0.20-delta calls (20C) (WTI, 2000–2022)

## 9.4 The Art and Science of Delta Hedging

Traders like to say that they do not need a model to price an option, as the price is given to them by a broker, but they do need a model to hedge an option. The hedging model is what determines the cost of the dynamic replication, or the option's value. The option fair value is not uniquely determined, as different hedging models imply different values. In this section, we test the robustness of VRP strategies with respect to important parameters of the hedging model. As highlighted earlier, the theory of dynamic option replication is developed for an idealized world, where trading can be done continuously without any friction, and delta can be calculated using the perfect foresight of the future realized volatility. The real world is far from perfect, and the trader constantly faces many practical challenges. The two most commonly asked questions by volatility traders are how frequently to hedge and what volatility to use to calculate the options' delta.

For any other systematic strategy that trades relatively frequently, the question of bid-ask and transaction costs comes to the fore. In many academic studies, this important topic is often either skipped entirely or sidestepped with simplistic assumptions, the impact of which is buried within the strategy performance metrics. For traders, however, the devil is in details. They only care about the existence of a theoretical risk premium if it can be turned into real profits, net of all frictions from market imperfections. A big question is whether transaction costs associated with futures hedging chip away so much of the trading edge that they could make the entire VRP strategy less appealing. We address this important topic explicitly and illustrate how delta hedging can be optimized to make the impact of execution costs more palatable.

One important instrument in the oil market that helps to reduce the cost of hedging is the previously described *Trading at Settlement (TAS)* contract. This contract allows traders to execute futures at a price to be determined later during the daily settlement window. It is very liquid for the nearest maturity futures, and for the most part, the trader can execute any given number of futures effectively with zero bid-ask. The TAS contract has proven to be quite useful for option traders, who often choose to rebalance their portfolio deltas at settlement prices. One caveat, however, is some uncertainty in the futures quantity required for rebalancing, as the settlement delta is not known until the settlement price is known. To tackle this issue, the option trader usually adjusts the futures quantity for the TAS order throughout the day based on updated estimates of the end-of-the-day portfolio delta. Any residual exposure is then promptly cleaned up once the futures settlement is published. This technique usually allows the trader to keep the average delta-hedging slippage to a minimum, typically to less than \$0.01/bbl.

Delta hedging of the short gamma strategy requires a trader to buy futures after the price rises and sell futures after the price drops. Many traders, reluctant to buy high and sell low, resist full delta rebalancing and utilize instead strategies reduced-hedging strategies. One popular choice is to hedge less frequently.<sup>6</sup> This approach not only saves on transaction costs, but also attempts to capitalize on short-term price

---

<sup>6</sup>Another popular reduced-hedging strategy is to trade fewer futures than required by the model. The details of this strategy can be found in Bouchouev and Johnson (2022).

reversals that decrease the need to hedge. Obviously, reduced-hedging strategies come with significantly higher risks if the price trend continues. The question is whether such risks are worth taking.

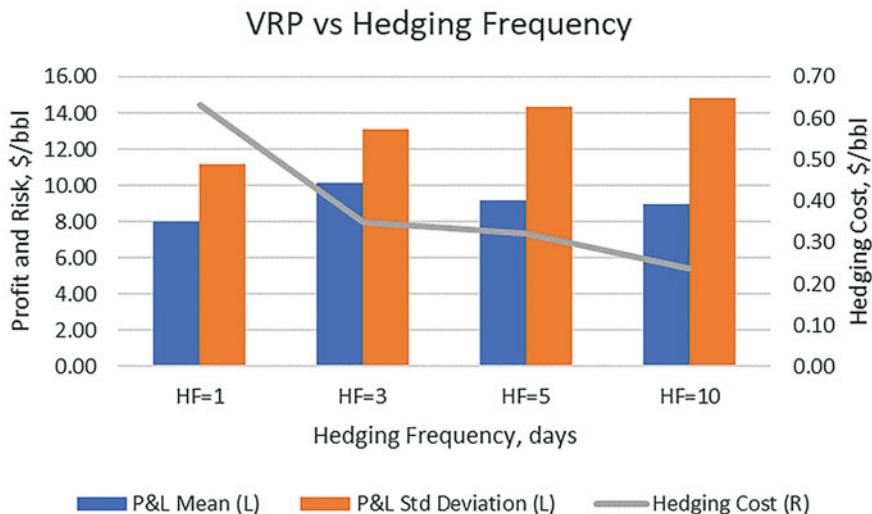
Consider the hedging strategy where daily portfolio delta rebalancing is replaced with re-hedging it only every  $N$  days. To keep the presentation more concise, we only look at a representative case of an ATM option with 60 DTM, but the conclusions for options with other moneyness and maturities are broadly similar. Figure 9.9 shows the performance of VRP strategies hedged every  $N = 1, 3, 5$ , and 10 days, where  $N = 1$  corresponds to the base case of daily hedging. Transaction costs are assumed to be \$0.01/bbl, which can be easily scaled to reflect an individual trader's assessment of the execution slippage.

It is not surprising that with less frequent hedging, the risks measured by the annualized standard deviation of the portfolio's P&L steadily increase. More interesting is the observation that annualized profits also increase slightly if the portfolio does not rebalance the delta for at least two days. This confirms that some additional gains can indeed be captured from short-term price reversals by leaving the portfolio unhedged for a few days. If instead of hedging every day, the trader hedges only every two or three days, then higher profitability adequately compensates for taking larger risks while simultaneously saving on transaction costs. However, leaving the strategy unhedged for more than a week makes it less attractive on a risk-adjusted basis. While such a strategy allows traders to benefit from short-term price reversals, it can suffer significant losses if the portfolio is left unhedged for too long.

The second important driver of VRP profitability is the choice of the hedging delta. Up until now, VRP strategies have been hedged with deltas calculated using implied volatilities. This choice was motivated by the desire to eliminate the stochastic term in (9.1). However, this choice is somewhat inconsistent with the entire investment concept behind VRP. We have already highlighted that the option's payoff can be perfectly replicated by trading futures only if the hedging delta is calculated using the future actual volatility, which, of course, is not known in advance. If the whole reason to trade VRP is driven by the view that options are overpriced, which means that implied volatility is too high relative to future realized volatility, then why would we be hedging using a volatility that we ourselves do not even believe to be correct?

To illustrate VRP sensitivity to the choice of the hedging delta, we scale the hedging volatility by multiples of 0.5, 0.75, 1.25, and 1.5 of the prevalent implied volatility. The multiple of 1.0 represents the base case. The results of this experiment are summarized in Fig. 9.10.

The most interesting takeaway from this analysis is P&L improvement from hedging with low volatility and P&L decline for hedging with high volatility. To interpret this result, recall from Fig. 9.1 and Table 9.1 how poorly the strategy of selling the naked straddle performs. Selling unhedged straddles suffers from large losses when futures periodically deviate too far from the initial ATM strike. This is consistent with the overall trendiness of oil futures already established in directional risk premia strategies. In the VRP strategy, when the option crosses the strike and moves ITM, the sooner the seller hedges changing delta, the better the protection

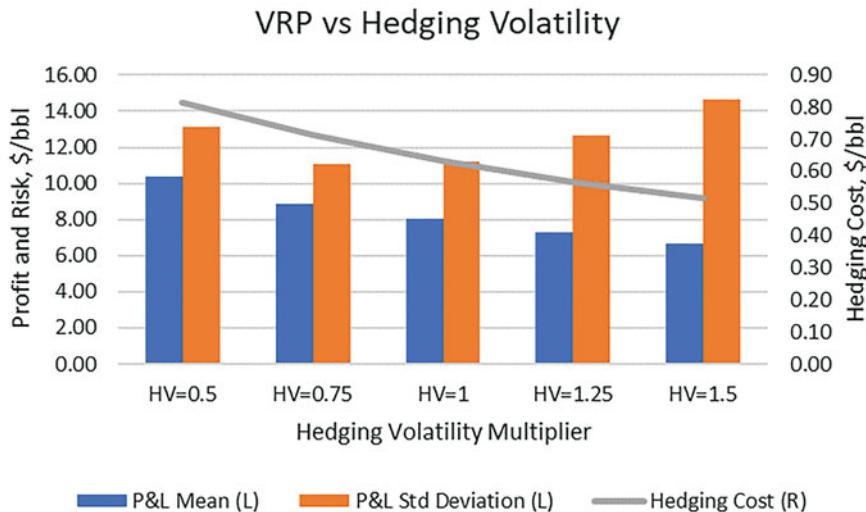


**Fig. 9.9** The performance of VRP strategies for 60 DTM ATM straddle for different hedging frequencies (WTI, 2000–2022)

against losses if the trend persists. Using the lower volatility to calculate deltas does indeed induce such a faster reaction. Lower input for the implied volatility reduces the remaining variance in the price of the option and brings the option payoff function closer to its terminal hockey stick-like payoff. The deltas computed in this manner, therefore, increase for calls and become more negative for puts when options are ITM. This forces the dealer to hedge more aggressively in the direction of the trend. In contrast to the reduced-hedging example, however, such aggressive hedging increases transaction costs.

Even though one might easily jump to the conclusion that hedging with a lower volatility is a better way to go, this approach has its own challenges. If we hedge conventionally, using what we perceive to be an incorrect implied volatility, then the second term in the Eq. (9.1) drops out and the mark-to-market of the P&L becomes less volatile. In contrast, if we hedge with any other volatility, even with a more accurate one, then daily P&L will fluctuate along with  $dF$ , and the standard deviation of P&L increases. This can also be seen from Fig. 9.10. Should we hedge then with a volatility that we believe to be incorrect and benefit from relatively stable daily P&L, or is it better to stick with our own opinion about future volatility and accept larger daily swings?

While we may believe that we are right and the market is wrong, our risk managers and controllers may have a different opinion. From their independent perspective, the market is fair. Therefore, whatever delta is implied by the market volatility should be the right one, and if our creative delta-hedging strategy results in additional P&L from futures, then it should be treated as a speculative position. Obviously, if risk managers view an option price to be fair, then they should not be



**Fig. 9.10** The performance of VRP strategy for 60 DTM ATM straddle versus scaled implied volatility used for calculating hedging deltas (WTI, 2000–2022)

letting the volatility trader sell an option to begin with. Some debate with managers becomes inevitable, and to avoid it, many gamma traders opt for a less controversial route and use conventional Black deltas based on exchange-published futures settlement prices for the base hedging case.

Despite the evidence of systematic biases highlighted by VRP hedging alternatives presented in this section, actual optimization decisions are usually made more tactically only when the volatility trader has particularly strong views about the short-term expected behavior of the market. To a certain degree, the problem of delta hedging also has elements of quantamental trading, where a trading algorithm is combined with human intervention. It would be a disservice to attempt to provide readers with any more precise systematic guidance on how frequently to rebalance a short VRP portfolio, and what volatility to hedge it with, but these case studies may steer hedgers towards better independent decisions.

The primary subjective decision in trading VRP, like in any other systematic oil strategy, is the identification of regimes in which the strategy is more likely to perform. For the oil VRP, such regimes are mostly driven by the demand for hedging and the behavior of large market participants.

## 9.5 The Behavior of Hedgers and Regime Changes

Additional insights can be gained from looking at the performance of VRP strategies over time. Many systematic risk premia strategies in energy markets are known to be sensitive to regimes. The world of energy constantly evolves, adjusting to new

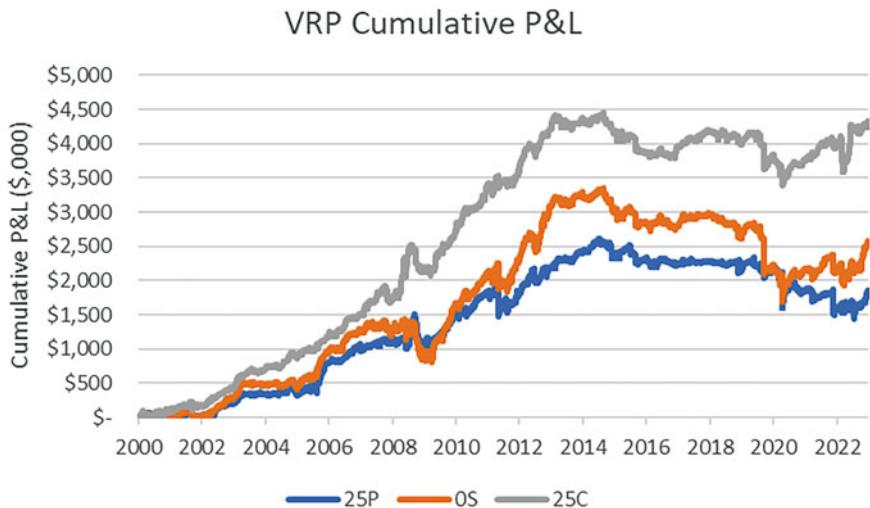
fundamental drivers, such as the growth of shale oil, or flow factors resulting from market financialization. These factors lead to changing behavior among hedgers and speculators. Such behavioral changes impact the supply and demand for risk management services and resulting risk premia. The VRP strategy is not an exception. Figure 9.11 presents the cumulative P&L of one-month VRP strategies for 0.25-delta OTM puts, 0.25-delta OTM calls, and zero-delta straddles. It clearly shows the presence of the structural break that separates two distinct regimes.

The idea of selling oil options as an overpriced insurance has proven to work remarkably well for a while. Between 2000 and 2014, VRP strategies generated impressive risk-adjusted returns with Sharpe ratios exceeding 1.0 for many moneyness and maturity configurations. After that, by and large, the risk premium contained in oil options followed the path of Keynesian normal backwardation. It gradually disappeared, like the structural futures risk premium did approximately a decade earlier. The opportunity to make easy money in any competitive market rarely lasts long. At some point, the strategy's consistent profitability motivates enough participants to invest in the capability needed to manage the strategy, which helps to bring the market towards an equilibrium.

The VRP strategy became a victim of its own success. As the business of passive commodity investments struggled under the pressure of contango and punitive rolling costs, capital shifted towards more dynamic strategies designed to capture alternative risk premia. To make it easier for investors, the VRP concept was packaged into investable indices, and the cumbersome task of delta hedging was effectively outsourced to index providers. Such investable volatility indices provided large pools of capital held by institutional investors with access to what used to be an obscure investment opportunity previously dominated by oil quants. The barriers to entry were lifted, and once again risk-bearing investors were compensated for providing capital rather than for having any particularly unique trading skills. As a consequence, the reward for offering relatively simple services of risk absorption cratered. Another factor that contributed to the structural break in VRP is a significant change in hedging strategies run by US shale producers, a topic that we will discuss in more detail in the next chapter.

The profitability of the VRP strategy will likely increase again, when more buyers come to the market to buy oil insurance. This usually happens shortly after the crisis that often forces insurance writers out of business, resulting in higher premiums. At the time of writing this book, there are indeed some indications that the Covid-19 pandemic could well be another turning point, as the VRP performance in the following two years improved substantially. This is driven by increasing demand for hedging and the exodus of many option sellers after the period of extreme volatility.

Even though the simple VRP strategy became less attractive to financial investors, the aggregate risk premium in oil options has not vanished. Instead, it spread out across various strikes and maturities, but not in a uniform manner. Some options became consistently more expensive than others. Like futures traders, many professional volatility traders have also switched from outright directional bets to



**Fig. 9.11** 20 DTM VRP equity history for 0.25-delta puts (25P), zero-delta straddles (OS), and 0.25-delta calls (25C) (WTI, 2000–2022)

relative value trades. In relative VRP strategies, one aims to sell options with higher risk premia and hedge them by buying options with lower risk premia.

A professional oil volatility portfolio is typically made up of many relative value strategies based on the primary VRP building blocks presented in this chapter. For example, the presence of the VRP smile suggests that OTM calls, while being cheaper in dollar terms than ATM calls, appear to contain a larger risk premium. One can then construct a relative value strategy that takes advantage of the VRP call skew by selling a larger quantity of OTM calls versus buying a smaller quantity of ATM calls. Such a trade, known as a *ratio call spread*, is very common in the marketplace. The weights between two legs are often chosen for the trade to be either volatility-neutral or premium-neutral. However, to successfully trade options across different moneyness, one needs to better understand important quantitative linkages that connect prices, which we discuss in the next chapter.

A relative value option portfolio typically includes options with multiple expirations. For example, the decreasing term structure of VRP may suggest hedging the sale of a short-term ATM option by buying a longer-term ATM option that is less overpriced or even underpriced. If the two options are traded in equal volumes, then the resulting calendar spread trade is net short gamma, since gamma of the short leg exceeds gamma of the longer-dated leg. However, the trade is net long vega, since vega of the longer-dated leg exceeds vega of the short leg. Buying deferred options can also be used as a valuable risk mitigant to the base VRP strategy, as it provides significant mark-to-market offsets when futures experience large moves. The weights between the two legs can be adjusted to make the trade either gamma or vega neutral. The proper construction of such strategies requires a more elaborate modeling of the local volatility term structure, which we focus on in Chap. 11.

The number of permutations that can be constructed from individual options and distinct hedging strategies is nearly infinite, but they are often based on the main features of VRP described in this chapter. In addition, professional volatility traders rarely even hold options until expiration. They might enter the trade when the opportunity caused by end-user flows is particularly attractive and then attempt to exit the trade once imbalances normalize. In the meantime, they would delta hedge based on a particular model to preserve the value of the option until the opportunity comes to exit the trade. This brings an important additional dimension to volatility trading since the trader is no longer betting on implied volatility versus gamma-scaled realized volatility during the entire lifespan of the option. The trader must also manage P&L fluctuations that come from changes in implied volatility itself, which is often referred to as *vega trading*, the topic we discuss next.

---

## References

- Bouchouev, I., & Johnson, B. (2022). The volatility risk premium in the oil market. *Quantitative Finance*, 22(8), 1561–1578.
- Derman, E., & Miller, M. B. (2016). *The volatility smile*. Wiley.
- Doran, J. S., & Ronn, E. I. (2006). The bias in Black-Scholes/Black implied volatility: An analysis of equity and energy markets. *Review of Derivatives Research*, 8(3), 177–198.
- Doran, J. S., & Ronn, E. I. (2008). Computing the market price of volatility risk in the energy commodity markets. *Journal of Banking and Finance*, 32(12), 2541–2552.
- Ellwanger, R. (2017). On the tail risk premium in the oil market, Bank of Canada Working Paper, 46.
- Jacobs, K., & Li, B. (2023). Option returns, risk premiums, and demand pressure in energy markets. *Journal of Banking and Finance*, 146, 1–26.
- Kang, S. B., & Pan, X. (2015). Commodity variance risk premia and expected futures returns: Evidence from the crude oil market, *SSRN*.
- Lo, A. W. (2002). The statistics of Sharpe ratios. *Financial Analysts Journal*, 58(4), 36–52.
- Prokopczuk, M., Symeonidis, L., & Simen, C. W. (2017). Variance risk in commodity markets. *Journal of Banking and Finance*, 81, 136–149.
- Trolle, A. B., & Schwartz, E. S. (2010, Spring). Variance risk premia in energy commodities. *The Journal of Derivatives*, 17(3), 15–32.



# Volatility Smile Trading

10

- Oil volatility traders like to quip that if you do not know who the fool in the market is, then it is probably you. This does not literally mean that someone is acting foolishly. Rather, it highlights the importance of understanding the behavior of large market participants.
- Vega-trading strategies require modeling the evolution of the entire volatility smile. Common smile heuristics, such as sticky moneyness and sticky strike, perform poorly in the oil market. Instead, a more realistic dynamics for the underlying futures contract must be imposed.
- Empirical study of the relationship between futures prices and implied volatilities guides model selection. The normal model removes an artificial skewness embedded in the conventional lognormal framework, but neither of the two models is capable of capturing extreme events.
- A novel quadratic normal model is developed based on the perturbation method for the general diffusion equation. The option pricing formula incorporates skewness and fat tails and relates model parameters to market prices for benchmark collars and strangles.

---

## 10.1 Producer Hedging and Volatility Market-Making

The strategy of harvesting an insurance-like risk premium by systematically selling oil options turned out to be rather fragile. The volatility risk premium that was clearly visible for nearly two decades became scantier over the years, as the oil market transitioned to a fundamentally different regime. The overall balance between buyers and sellers of oil volatility has been largely restored. The supply of volatility was increased through crafty financial products that allowed investors to provide risk-bearing capital while outsourcing the complexity of delta hedging to professional dealers. In addition, many corporate hedgers, reluctant to pay up for overpriced options, switched to more advanced risk management programs, many of which are volatility neutral. The adoption of diverse and sophisticated hedging

strategies by end-users, however, opened new opportunities for volatility market-makers in *relative vega trading*.

In the volatility market, traders say that if you do not know who the fool in the market is, then it is probably you. This does not literally mean that the counterparty of your trade is acting foolishly. Rather, it highlights the importance of understanding the motivation of market participants whose behavior may temporarily distort prices of certain options and present attractive trading opportunities for the dealer. Such understanding is particularly crucial in trading oil volatility.

The oil market is full of real options, many of which we have already discussed. The volatility embedded in real options motivates asset owners to utilize financial options in their hedging programs. The largest hedger in the options market is the producer. Producers are unlikely to be fools in making directional bets on oil prices. They are arguably even more knowledgeable about market fundamentals than dealers themselves. However, when it comes to volatility, end-users are generally prepared to pay up for the service of tailoring risk management products to their specific needs.

The risk management tools used by oil producers vary with the ownership structure of the company, its overall business model, and the level of indebtedness. The governments of oil-producing counties and state-owned entities rarely get involved in the derivatives market. One notable exception is a large-scale hedging program administered by the Government of Mexico, which we will cover in more depth in the next chapter. For the most part, other large sovereign producers manage price risks with policy tools, such as taxation, subsidies, and rainy-day funds. Large vertically integrated oil majors also rarely hedge in the derivatives market. They tend to benefit more from related energy businesses, such as refining, and often prefer to retain their exposure to oil prices, which is the primary reason why some investors buy the company's stock. The most active participants in the oil hedging market are independent oil producers, particularly the ones specializing in shale oil.

The economics of shale drilling differs from conventional oil production in several important ways. It is generally more expensive to extract oil from shale than to produce it via traditional means in oil-rich parts of the world. However, shale has a much shorter production cycle, which makes its operation more like a mining or manufacturing business. Any oil production represents a real call option on the price of oil with the strike price determined by the cost of production. The real option held by a shale producer can be thought of as being closer to ATM because of its higher production cost, and ATM options have the largest exposure to volatility.

Shale's shorter production cycle is valuable for capturing short-term volatility and monetizing the real option. It allows producers to be nimbler in the market by quickly adjusting production levels in response to market movements, while simultaneously locking in favorable prices in the derivatives market when futures rise. If futures subsequently fall, then producers can take profits on existing financial hedges and reduce production. In contrast, real options held by low-cost producers are effectively in-the-money. While these options have larger intrinsic value, the producer's decision making is less affected by volatility. It is nearly always rational for any individual low-cost producer to continue pumping oil, regardless of the price.

The only exception is the situation when large producers collectively agree to reduce output with the specific goal to impact the price.

Unlike national oil companies and oil majors, independent shale producers are notorious for their high leverage. Since shale production has faster decline rates, producers must constantly drill just to maintain the same output level. This strategy requires a persistent inflow of capital, which is typically borrowed from the banks. The lenders, unwilling to take the risk of a producer failing in the event of lower oil prices, typically insist on hedging the price risk to support the loan. This is not too different from the mandate to purchase house insurance when the property is mortgaged to a bank. Shale lenders may also hold a lien on the producer's main property, oil reserves, as collateral for the loan. However, bankers are not keen on ever taking ownership of the physical oil. They would rather require producers to protect themselves in the derivatives market against the adverse price scenario, which can be enforced via lending covenants. Besides, the same lending bank often acts as the counterparty on the derivatives hedge as well, thus, making money on both transactions.

While the need for shale producers to hedge is apparent, their financial resources to do so are rather limited. Producers can rarely afford to buy traditional insurance and shell out upfront cash for put options. With the strong competition for market share in the fast-growing shale business, cash is usually preserved and used to maintain and scale up production. Besides, producers often perceive an initial investment in the oil property as a form of option premium that has already been paid. Paying yet another premium to acquire a financial option to hedge is a rather tough pill for many of them to swallow. What the producer owns is a real call option; what is needed for a bank loan is a financial put option. Thus, a more efficient solution for the producer would be to swap one option for another and avoid additional cash outflows. This is where the derivatives market with its flexibility comes in handy.

The most popular producer hedging structure in the oil market is a *costless collar*, sometimes also referred to as a *fence*. In this trade, a producer buys a put option with a strike price that ensures some minimum investment return on borrowed capital. The producer does not pay any cash for the put option and instead gives the dealer a call option of the same value. The call strike is iterated until the value of the call matches the value of the put. In practice, the traded call strike is shifted down to make the call slightly more valuable than the put to incentivize the dealer to participate in the trade. The dealer's compensation comes not from retaining a portion of the option premium, which is zero, but rather from receiving at zero cost the collar, which has several cents of a positive value, at least according to the dealer's pricing model.

Since the iterated call strike in the costless collar trade is customized specifically for a given deal, it is unlikely that exactly the same collar will be quoted in the open market. In the jargon of option traders, an end-user option trade is rarely a back-to-back transaction, meaning that it cannot be hedged perfectly. To turn a theoretical profit margin into real dollars, the dealer must reduce risks by trading more liquid options and managing the residual exposure dynamically. With many option trades

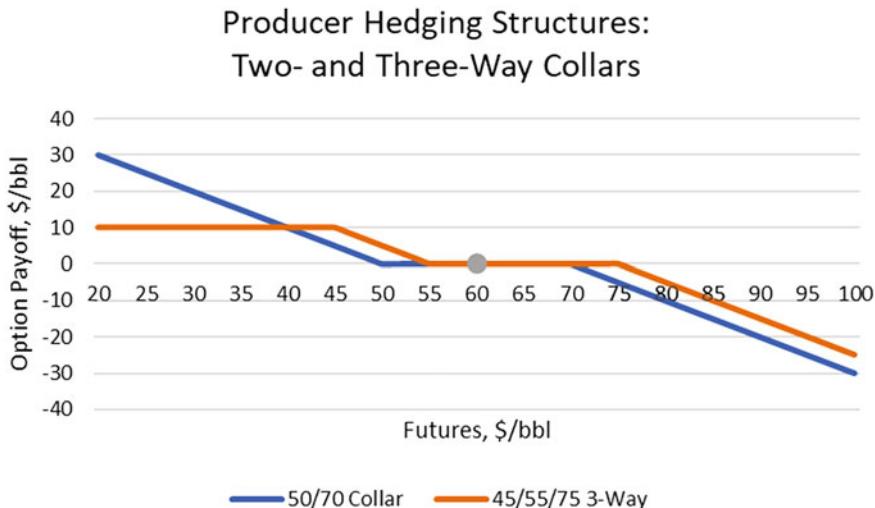
added to the volatility market-making portfolio throughout the day, the residual exposure is always managed on an aggregate level. This can only be done if the pricing model used by the dealer can handle options across all strikes in the portfolio consistently with the market.

The costless collar is approximately volatility neutral. The vega from the call largely offsets the opposite vega from the put, provided that both options are struck with similar moneyness. Collar transactions are, however, primary drivers of the skewness and the overall dynamics of the volatility smile. If the end-user hedging program is executed ratably, then the put strike is often kept the same, as it is often chosen to meet lending requirements. However, call strikes are spread out based on whatever strike makes the collar transaction costless for a given futures price. This results in a rather convoluted dynamics for the volatility smile. On the put side, the peak hedging demand is often tied to the specific strike, while the supply of calls sold by a producer to finance the put is more driven by the moneyness of the option. This brings additional challenges to modeling the dynamics of the oil smile using conventional techniques often employed in other financial markets.

The costless collar is only one example that illustrates the leverage that a producer can gain by selling optionality in the derivative market. Some producers take the idea of monetizing their real option further by selling a financial call option to the lender and collecting an option premium without buying any downside price protection. Such deals can hardly be classified as proper hedges. In fact, they are much closer to hidden loans. The outright selling of options by producers is generally frowned upon by lenders due to their large credit exposure. However, as we will see in the next chapter, short calls are often embedded into more complex exotic derivatives. These over-the-counter derivatives are less transparent and sufficiently profitable for dealers to justify taking on additional credit risks.

Another popular leveraged hedge in the oil market is a *three-way collar*. In this trade, instead of buying a put option financed by selling a call, the producer purchases a put spread. The put spread is essentially a put option with capped payoff, where the purchased put is combined with the sale of another put that has a lower strike. The put spread is again financed by selling an OTM call that makes the entire structure of three options costless. Since the put spread is generally cheaper than an outright put option, less premium needs to be raised from the sale of the call to make the net transaction cost zero. This lets the producer push the call strike further OTM and retain larger potential upside from higher prices. Alternatively, the producer can sell the call with the same strike as in the regular collar, collect a larger premium and use it to buy a more valuable put spread with a higher upper strike. The strikes are again selected iteratively to make the three-way collar theoretically costless, and then shifted slightly in favor of the dealer, as an incentive to trade.

Terminal payoffs of a two-way collar and a three-way collar are illustrated in Fig. 10.1. In this example futures trade at \$60, and a producer can consider buying \$50 puts, which are financed by selling \$70 calls. If the distribution of oil prices happens to be perfectly symmetric around its mean, which is determined by the futures price, then such a collar trade should be theoretically costless. More generally, the price for an equidistant collar represents the market measure of asymmetry,



**Fig. 10.1** The payoff profile for \$50/\$70 collar and \$45/\$55/\$75 three-way collar with \$60 futures price

or the skewness of the price distribution. If the distribution were lognormal with a larger upside tail, then an equidistant collar trade would have had positive premium to the call. Similarly, if the market had a heavier left tail relative to the normal distribution, then an equidistant collar would have traded at a premium to the put, which is indeed more often the case in the oil market.

In the case of a three-way collar, a producer can buy a \$45/\$55 put spread with a maximum payout of \$10 and retain larger unhedged price upside by selling further OTM \$75 calls. In return for raising the effective floor to \$55, the producer gives up any additional hedging benefits if futures fall below \$45. The producer often justifies the sale of a lower strike put by the argument of already owning some real optionality. If the strike of the short put is set at a sufficiently low level that the shale production is no longer economical, then the producer can monetize the physical option to close down production instead of using the financial hedge. Such a rationale is highly debatable, and the usage of three-way collars is widely considered to be a borderline speculation by oil producers. It would be fair to accept that this transaction lies somewhere in between a legitimate risk management instrument and a speculative bet by a well-informed market participant.

Another way to look at a three-way collar transaction is to decompose it as a purchased put option financed by selling two other options, an OTM call and an OTM put, collectively defined as a *strangle*. Even though the transaction retains its costless structure in terms of an aggregate option premium, it is unlikely to be neutral in terms of volatility. The vega of the short strangle typically exceeds the vega of the purchased put option, especially for longer-term deals, as two options are sold for any one option bought. This means that in a three-way collar trade the producer can hedge by selling volatility to the market rather than by buying volatility as in the case

of traditional insurance. The proliferation of such leveraged hedging structures increased the supply of volatility in the oil market by producers, which led to a decreasing volatility risk premium, as documented in the previous chapter.

Put options, two-, and three-way collars are the most common hedging structures in the oil options market. Their market prices are converted by option dealers into three primary volatility benchmarks: an ATM straddle, an equidistant collar, and an OTM strangle. These three benchmarks correspond to three moments of the underlying price distribution. The price of an ATM straddle captures the market-implied variance. An equidistant collar characterizes the skewness of the distribution. An OTM strangle, which is embedded in the price of a three-way collar, provides the market assessment of the tails of the price distribution, or its kurtosis. At the end of this chapter, we develop a new quadratic option pricing model whose three parameters are mapped to the prices of three derivatives benchmarks.

The above-mentioned derivatives structures dominate the hedging market, but the choice of strikes and maturities remains highly dependent on the economics of an individual producer. The strike for the purchased put usually depends on the cost of production, which varies substantially across specific oil properties. The selection for short options, on the other hand, is often driven by the hedger's overall business leverage and creditworthiness. The lender can also reduce the leverage and credit risk by requiring the producer to pay some fixed amount of cash premium, such as \$1/bbl, for the collar transaction. In these so-called *premium collars*, the transaction is no longer costless, and strikes are iterated to match the target value for the entire package. As a consequence, a typical oil option market-making portfolio has open positions with numerous strikes resulting from various transactions with end-users.

The additional selling of optionality by producers made the oil volatility market more balanced. At the same time, the supply and demand for individual strikes became more unbalanced and dynamic, constantly changing as new hedging deals with different strikes come to the market. Such an environment is a paradise for quantitatively minded volatility traders. They could often charge one customer an extra premium for an outright put option with a popular strike and acquire potentially at a discount another option with a nearby strike from a different client via a three-way collar or some other carefully designed option package. Not only is the surcharge collected by the dealer on both trades for providing the service, the dealer's own market risk is also simultaneously reduced.

To succeed in the competitive business of volatility market-making, the trader must be ready to respond to any pricing request by end-users and be able to act quickly in the market when an option structure is perceived to be mispriced. What is mispriced, however, is model dependent. If the trader can sell one option above its theoretical price generated by a certain model and buy a similar option below its model price, then it could be called mispricing only in the context of a specific model. We would classify this as a *model arbitrage*. The question is, of course, what model to use for tracking the constantly changing prices of many different options, which are represented by the volatility smile. However, before we proceed to a recommendation for such a model, we first provide a cautionary warning about the application of conventional approaches to this problem in the oil market.

## 10.2 Skew Delta and Two Types of Stickiness

The business of vega trading is very different from the insurance-like business of gamma trading described in the previous chapter. A gamma trader speculates on what the realized volatility will likely be during some time in the future, typically the entire lifespan of the option. In contrast, a vega trader is an arbitrageur who cares more about *relative* pricing of different options and their corresponding short-term price dynamics. The profits in volatility market-making are expected to accumulate from the slight edge in original customer transactions. The goal of the trader is to buy enough time to rebalance strikes in the portfolio. In the meantime, the portfolio must be neutralized in a way that isolates and protects the mispricing premium which was embedded in the original trade. This can only be accomplished if the portfolio is hedged using a theoretical model that governs the joint price dynamics of options across all strikes, and prices produced by the model are consistent with market prices of options that are used for hedging.

As discussed in Chap. 8, relative prices of options with different strikes are maintained in the form of the volatility smile. The smile converts the assortment of prices into more convenient standardized units of measure mostly to facilitate price tracking. Option implied volatilities are maintained as a function of either strike or moneyness. The moneyness can also be defined in a variety of ways, such as the strike's distance to ATM, its ratio to ATM, the same ratio scaled with time to maturity, or as an option's delta. The smile itself does not say much about the richness or the cheapness of the option. Its presence only indicates that the futures price does not behave according to the assumptions of the model. If the smile is steep, it simply means that the chosen model for the dynamics of futures prices is a poor approximation of the real price behavior. The smile is merely a communication tool, a sort of universal language which all options traders agreed to use to efficiently liaise with each other.

The smile presents only a static snapshot of prices taken at a given instant. But what a volatility dealer needs is a movie of how the smile evolves when futures move. The static smile does not tell the trader what option prices will be tomorrow, or how to calculate option deltas to keep the portfolio price neutral while looking for opportunities to unload unwanted strikes. To compare option prices today and tomorrow, the trader needs to understand what drives the dynamics of the entire smile. Since the smile itself is an artefact of the questionable assumption of log-normality, modeling the behavior of something that does not correspond to reality may look a bit odd. However, this is what many traders do for operational convenience. The traditional approach to describing the behavior of the smile relies on ad hoc heuristics that dynamically adjust model inputs in an attempt to artificially preserve its validity. We first demonstrate the shortcomings of applying this approach to the oil market before replacing it with a more accurate pricing framework.

Consider the snapshot of option prices for a given maturity, which are expressed via implied Black volatilities  $v(K, F)$

$$C = C_{BL}(F, t; v(K, F))$$

We explicitly write the smile  $v(K, F)$  as a function of two variables, the strike price  $K$  and the futures price  $F$ . Since the smile is simply a plug-in to the Black formula designed to match market prices of options for a given futures price, the shape of the smile is likely to change once the futures price moves. This makes the life of a volatility dealer more difficult, as a change in the smile also impacts the option's delta. Finding the delta that keep the overall portfolio of options neutral with respect to futures is critical for the volatility trader. The impact of any slippage in the calculation of delta can easily dominate the initial mark-up in pricing the option, thus ruining the entire arbitrage opportunity.

When the smile moves, the option's total delta, which is the partial derivative of the option price with respect to futures, must be calculated using the chain rule, as follows:

$$\Delta = \frac{\partial C_{BL}}{\partial F} + \frac{\partial C_{BL}}{\partial v} \frac{\partial v}{\partial F} = \Delta_{BL} + V_{BL} \frac{\partial v}{\partial F} \quad (10.1)$$

where  $V_{BL}$  is Black vega. In addition to the standard Black delta, the total delta in (10.1) acquired another component, which is equal to the product of Black vega and the slope of the implied volatility function with respect to futures. This second term in (10.1) is colloquially referred to as a *skew delta*. It represents the change in the value of the option resulting from the change in implied volatility, induced by the change in futures.

Unfortunately, the smile today does not tell us much about its possible shape tomorrow, so we do not really know the true hedging delta. What we observe today is the slope of the smile with respect to  $K$ . However, what we need for hedging is the slope of the smile with respect to  $F$ , which arises in the second term of (10.1). To calculate the latter, one must impose additional assumptions about the evolution of the smile. Since the smile is predominantly maintained as a function of option moneyness, the path of least resistance is to assume that implied volatilities remain the same for a given moneyness.

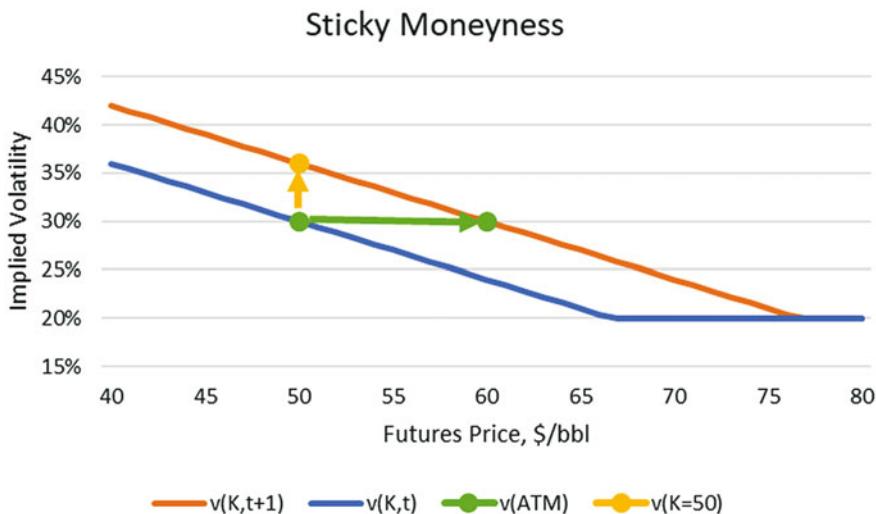
Let us illustrate the dynamics of the implied volatility in the simplified case when the smile is linearly decreasing with slope  $\beta > 0$ , so that

$$\frac{\partial v}{\partial K} = -\beta < 0$$

Such a linear approximation is reasonable in the oil market for strikes in the vicinity of ATM<sup>1</sup>. To make sure that implied volatility in this example does not drop below zero given its negative slope, we flatten the smile once volatility reaches 0.20,

---

<sup>1</sup>We only provide a simple illustration of this topic. For a more advanced exposition, we refer to Rebonato (2004), Gatheral (2006), and Derman and Miller (2016).



**Fig. 10.2** The volatility smile behavior under the sticky moneyness rule

as our interest is predominantly in the linearly decreasing portion of the graph. Figure 10.2 illustrates it.

We first consider the scenario of *sticky moneyness* for which the shape of the smile stays intact for a given moneyness of the option. Assume that at time  $t$ , the futures price  $F(t) = \$50$ , and ATM volatility  $v_0 = 0.30$ . Assume that in the next period, futures move to  $F(t + 1) = \$60$ , and the entire smile shifts in parallel while retaining the same shape versus moneyness. The graph in Fig. 10.2 simply slides to the right.

Algebraically, the dynamics of sticky moneyness with a linear volatility skew can be written as follows:

$$v(K, F) = v_0 - \beta(K - F)$$

where  $v_0$  is ATM volatility for the strike  $K = F$ . When futures move up, then ATM volatility  $v_0$  always stays the same, but the implied volatility for any fixed strike  $K$  rises. This is the consequence of a negatively sloped volatility smile and the positivity of the second term in (10.1) since

$$\frac{\partial v}{\partial F} = \beta > 0$$

Likewise, when futures fall, then implied volatilities for all strikes decrease making options somewhat less expensive. Such behavior is counterintuitive and potentially indicative of a problem with the assumption of sticky moneyness. Since oil futures tend to have larger downside gaps and percentage volatility is generally inversely related to the futures price, it would be strange for the optionality to

become less valuable following the futures move in the direction of greater uncertainty.

The problem manifests itself more clearly when one considers its impact on the option's delta. Since for a negatively sloped smile, the skew delta is positive, the total hedging delta for sticky moneyness always exceeds a conventional Black delta:

$$\Delta = \Delta_{BL} + \beta V_{BL} > \Delta_{BL}$$

This inequality holds for all options, calls, and puts.

Let us illustrate the hedging delta with a simple numerical example. Consider a one-year ATM option struck at  $K = F = \$50$  with an implied volatility  $v_0 = 0.3$ , and  $\beta = 0.004$ , which corresponds to a typical slope of the Black skew near ATM. Using formulas for Black delta and vega from Appendix B, we calculate the total hedging delta for an ATM call option

$$\Delta_{Call} = \Delta_{BL, Call} + \beta V_{BL} \approx 0.56 + 0.08 = 0.64$$

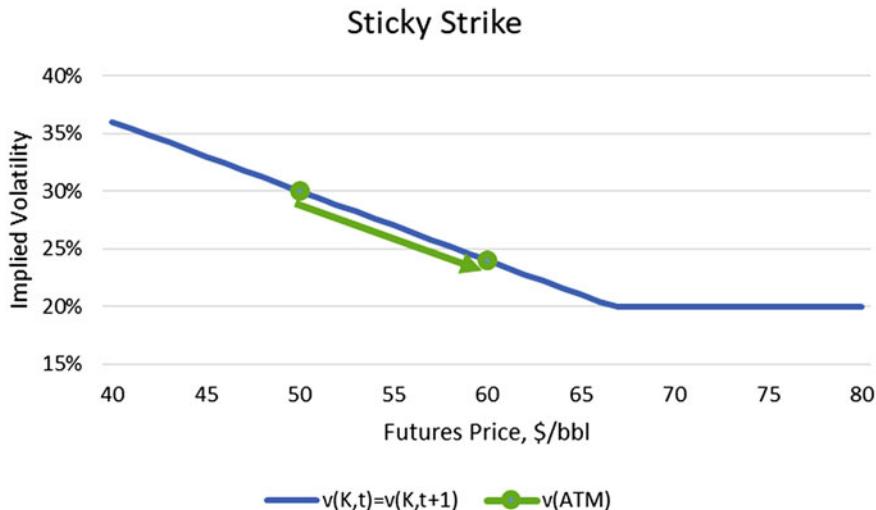
and for an ATM put option

$$\Delta_{Put} = \Delta_{BL, Put} + \beta V_{BL} \approx -0.44 + 0.08 = -0.36$$

In this example, the assumption of sticky moneyness results in 28% delta for an ATM straddle. Such a high delta is counterintuitive, as one would have expected an ATM straddle with a symmetric payoff function to have delta closer to zero. Part of the problem here comes, of course, from Black delta itself, which is around 12% for a one-year ATM straddle, as a consequence of the skewed lognormal assumption. Sticky moneyness makes the hedging problem worse, as it magnifies the artificially skewed positive delta bias. This suggests that the seller of 100 contracts of ATM straddle should buy 28 futures as a hedge. Such an asymmetric hedge would be very bizarre in the oil market, especially given its large downside price gaps.

Hedging a portfolio of options using the operationally convenient setup of sticky moneyness often leads to catastrophic consequences. Since the skew delta is positive for all options, managing a portfolio of short options requires holding additional long futures. This means that for any put option sold to a producer, sticky moneyness forces the dealer to sell fewer futures than truly needed for delta hedging. A portfolio which may appear to be delta neutral will behave with a long futures bias. In a market that follows an *up the stairs, down the elevator* dynamics, such a hedge could lead to large losses. Once sticky moneyness traders realize their mistake, they often panic to unload the excess of long futures. This often coincides with heavy pressure from the negative gamma hedgers which periodically sends futures into a downward spiral. Unfortunately, this is how many unsuccessful volatility traders were forced to end their trading careers in the oil market.

The conceptual problem with sticky moneyness is that implied volatilities for all strikes are shifted in the wrong direction. Falling implied volatility for each option in a falling futures market does not feel right, as one should expect precisely the



**Fig. 10.3** Volatility smile behavior under the sticky strike rule

opposite to happen, at least, if volatility is measured in percentage terms (recall Fig. 8.4). Since collectively dealers share the downside risk of a price collapse, they will attempt to reduce their volatility exposure when futures fall closer to put strikes sold to producers. The proximity to short strikes makes dealers shorter volatility. To stay within allocated risk limits, dealers have no choice but to buy back some short options, which will inevitably push implied volatility higher. However, the naturally rising percentage volatility in a falling futures market would be out of sync with the dynamics imposed by the sticky moneyness assumption.

The problem with sticky moneyness does not go away for markets with larger upside risks, such as markets for natural gas and refined products. In these right-tailed markets, the implied volatility skew is more likely to have a positive slope. If futures rise and the smile slides in the sticky moneyness fashion with the same ATM volatility, then given the positive slope of the smile, implied volatilities for each strike decrease. This is again counterintuitive, as specific strike volatilities are unlikely to fall when futures move in the direction of larger uncertainty, which is now on the upside. In this setup, the skew delta is negative. This forces the dealer to mistakenly under-hedge short OTM calls, which is highly dangerous in positively skewed markets.

One simple way to mitigate the counterintuitive effects of sticky moneyness is to assume an alternative heuristic of a *sticky strike*. Under the sticky strike assumption, the volatility smile is assumed to retain its shape across all given strikes, regardless of their moneyness. Since the smile is still operationally maintained in terms of moneyness, to ensure that implied volatilities stick to the same strikes, one must perpetually adjust the shape of the smile by moneyness.

Figure 10.3 illustrates the sticky strike behavior. The smile graphed versus specific strikes is static. The sticky strike essentially preserves the Black formula

with a one-off volatility adjustment for each strike. In this case, ATM volatility does not shift in parallel when futures move, but instead it slides up and down the skew.

To characterize the linear sticky strike rule, one only needs to specify the slope  $\beta$  along with an initial futures price  $F_0$  for which the corresponding volatility is  $v_0$ . Its dynamics in the linear case is described as

$$v(K) = v_0 - \beta(K - F_0)$$

The second term in (10.1) then vanishes because the smile is static; it depends only on fixed  $F_0$  but does not depend on future price  $F$ . The total hedging delta in the sticky strike case is identical to Black delta for calls:

$$\Delta_{Call} = \Delta_{BL, Call} + 0 = \Delta_{BL, Call}$$

and likewise, for puts

$$\Delta_{Put} = \Delta_{BL, Put} + 0 = \Delta_{BL, Put}$$

The sticky strike heuristic forces the trader to depart from the path of least operational resistance determined by sticky moneyness. Instead of doing nothing and letting the smile slide in parallel, the sticky strike requires the trader to constantly repivot the smile to the new ATM level and then tweak its shape by moneyness, as required by operational standards.

At this point, one may wonder why not just keep the entire smile in terms of fixed strikes? Theoretically, it is doable, especially for short time periods when the range of popular strikes does not change. In the long run, however, keeping the skew by strike is impractical. The market always talks in terms of moneyness, and it uses ATM volatility as a pivot point for the construction of the smile. Nearly all risk and accounting systems require smiles to be saved by moneyness to facilitate standardization across multiple assets. While the trader can maintain the sticky strike smile for decision making, its system-required shape by moneyness must be constantly bent. The reward for this hassle comes from the elimination of undesirable and counterintuitive skew delta corrections. The improvement, however, is marginal. The delta of an ATM straddle is still positive, which is caused by the assumption of lognormality. In the previous numerical example even with no contribution from the skew delta, the ATM straddle still has 12% delta. It is half as bad compared to the sticky moneyness case, but it is still not a viable hedging solution as one should expect ATM straddle delta to be closer to zero.

The obvious question arises now whether it is possible to come up with other heuristics that shift artificially skewed Black deltas in the right direction. Yes, it can be done. Nothing can stop the trader from specifying any other arbitrary dynamics of the smile. The trader can draw a certain glide path for an ATM volatility that does not simply slide along the same shape, but instead increases or decreases with any chosen slope. An equally arbitrary skew can then be attached to this glide path of ATM volatility to complete the smile dynamics. However, all such heuristics amount

to more complex but still artificial tweaks to a base pricing model that does not represent the real market behavior. The problem can only be solved by treating its cause and not the symptoms.

One challenge is the natural asymmetry embedded in the lognormal distribution. While the assumption of lognormality might be reasonable for some financial markets that tend to grow over time, it becomes problematic for mean-reverting oil markets, where the underlying price distribution is much more symmetric. In Chap. 8, we illustrated how the desired symmetry can be restored if one switches from lognormal to normal measures of the realized volatility. In the following section, we take a similar route for the implied volatility and compare the behavior of volatility smiles, as they are seen from two alternative angles.

---

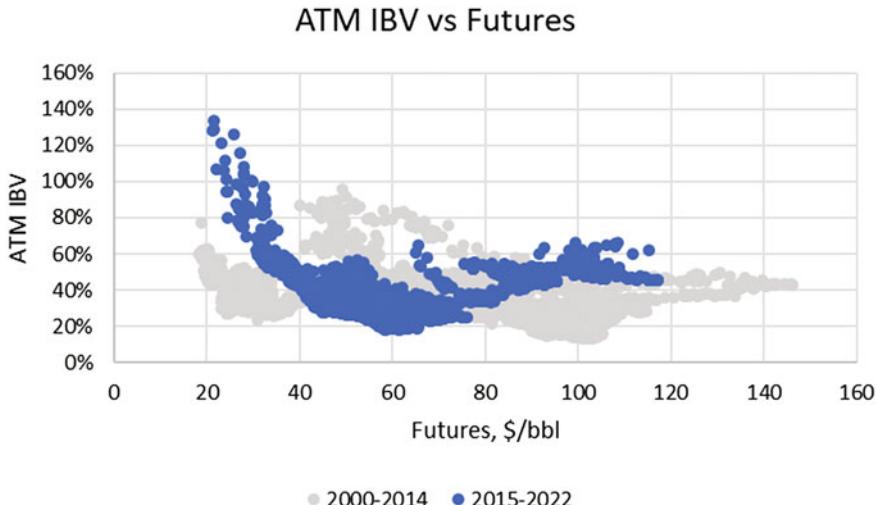
## 10.3 When Black Smirks, Bachelier Smiles

The history of oil implied volatility, like many other topics related to oil trading, is very colorful and multifaceted. Its complete story might even warrant a separate publication to be filled up with fascinating anecdotes of extreme volatility spikes caused by a panic of option sellers running for cover in the aftermath of squeezes, wars, hurricanes, and other unexpected market disruptions. Somewhat regrettfully, we leave these anecdotes for another time and only look at the history of volatility through some observations from implied volatility data. Our goal is to highlight some salient properties of oil volatility that can guide us towards a better model, which will be developed in the following section.

Certain key characteristics of the oil markets have already become apparent from Chap. 8 where we looked at the historical behavior of realized volatility versus the futures price. We now extend the same view to implied volatility. We start with the benchmark ATM implied Black volatility (IBV) and look at the history of implied volatility in the form of a scatterplot versus the futures price, shown in Fig. 10.4.

Like realized volatility, IBV also exhibits a strong inverse relationship with futures. To see more clearly the evolution of this relationship, we split the historical sample into two sub-periods that correspond to different regimes outlined in the previous chapter. In the earlier period, IBV showed little dependency on the futures price, providing some support to the assumption of lognormality. However, in the latest regime of financialization and shale, the inverse relationship became much stronger and highly nonlinear.

By now, a careful reader must have recognized that these observations are also consequences of the skewed lognormal volatility metric. The volatility of percentage returns rises algebraically when the denominator of the ratio defining the return decreases. Recall our analogy of looking at the market through lognormal lenses to glasses for astigmatism worn by someone who does not need them. Imagine now that over the years the overall vision of this patient somehow normalized, but the patient is still wearing the same old glasses with the wrong prescription. The perception that was blurry to begin with, may not even let the person see anything at all.



**Fig. 10.4** ATM IBV and futures prices for third nearby contract (WTI, 2000–2022)

Let us now switch to better lenses and look at the same historical picture using *implied normal volatility (INV)*. Thankfully, we do not need to recalculate another set of implied volatilities, as the two metrics can be easily transformed from one into the other. To relate IBV and INV, we equate the Bachelier formula (8.9) with normal volatility  $v_N$  and the Black formula (8.10) with lognormal volatility  $v$ . By definition of implied volatilities, the two formulas must produce the same observable market price:

$$C_{BC}(F, t; v_N) = C_{BL}(F, t; v)$$

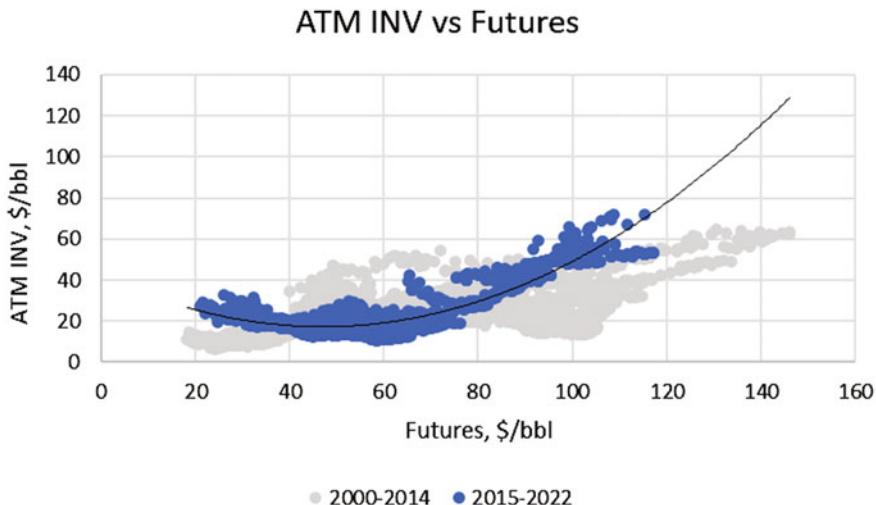
For an ATM option, when  $K = F$ , this pricing identity simplifies considerably, as follows:

$$v_N \sqrt{\tau} n(0) = F \left\{ N \left( \frac{v \sqrt{\tau}}{2} \right) - N \left( -\frac{v \sqrt{\tau}}{2} \right) \right\}$$

Since the normal probability density is the derivative of the cumulative normal function, i.e.,  $n(x) = \frac{\partial N}{\partial x}$ , we can apply the Taylor rule to the right-hand side of this identity, and obtain that for short maturities  $\tau$ :

$$v_N \sqrt{\tau} n(0) \approx F v \sqrt{\tau} n(0)$$

It follows that for ATM options with short maturities INV is approximately equal to IBV multiplied by the futures price



**Fig. 10.5** ATM INV and futures prices for third nearby contract (WTI, 2000–2022)

$$\nu_N \approx \nu F, \quad \text{if } K = F$$

This approximation is very accurate and commonly used by traders to switch between the two alternative implied volatility metrics.

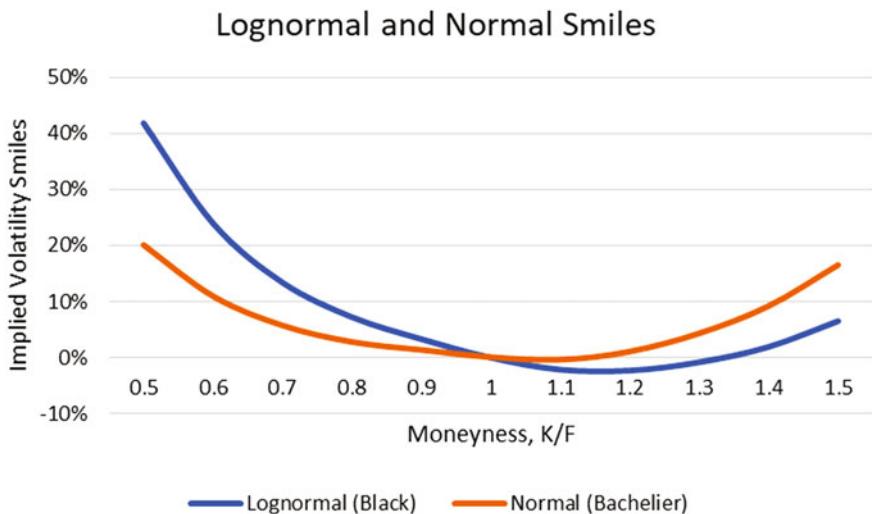
The shift from skewed lognormal lenses to normal ones provides a different look at the historic relationship between prices and implied volatilities. This alternative view is presented in Fig. 10.5, where INV is shown versus the futures price.

In the first period, ATM INV exhibited little convexity, generally moving in the same direction as the futures price, which is consistent with a more lognormal type of behavior. If IBV remains unchanged in percentage terms and futures rise, then INV and the price of an ATM option in dollars per barrel must also rise. In the second period, the curvature on both sides of the graph is more visible. One can see increasing volatility at lower prices, which is driven by financialization and shale growth with increasing spillovers from equity markets and periodic constraints on maximum storage capacity. The curvature of the right tail can be attributed to larger impact of supply disruptions, given a reduced buffer of spare production capacity and reluctance by producers to commit to long-term investments in anticipation of energy transition. As a result, the relationship between INV and futures became somewhat parabolic.

While the analysis of ATM volatility sheds some light on the underlying price distribution, a better diagnostic of the volatility behavior can be obtained by looking at the entire volatility smile. Conveniently, IBV and INV smiles can also be transformed from one to the other by using the following simple approximation<sup>2</sup>:

---

<sup>2</sup>For brevity, we omit the rigorous derivation of this approximation. It is again based on the Taylor series expansion for both formulas, but with more cumbersome mathematical details which are less



**Fig. 10.6** Historical average IBV and INV/F by moneyness for third nearby contract (WTI, 2015–2022)

$$v_N \approx \begin{cases} vF, K = F \\ v \frac{F - K}{\ln\left(\frac{F}{K}\right)}, K \neq F \end{cases} \quad (10.2)$$

It follows that for OTM options, the ratio of INV to IBV is approximately equal to the ratio of dollar moneyness to the logarithmic moneyness. This approximation has also proven to be remarkably accurate. The average approximation error over the entire data set is much smaller than 0.1% if measured in terms of Black volatilities. Obviously, the approximation cannot be used for negative strike prices, when the Black formula becomes meaningless and market participants use INV as a primary volatility benchmark.

To complete the historical study, Fig. 10.6 shows the average Black and Bachelier smiles calculated over the second period, which is more reflective of the current market conditions. To make the two curves visually comparable, the normal smile is scaled by the corresponding futures price, so that the shapes of both volatility curves can be displayed side by side in percent. Since the smile is shown here for fixed maturity, for transparency, the moneyness is defined by a simple ratio  $K/F$ .

It is important to remember that smiles are nothing more than option prices displayed in different coordinate systems. Smiles attempt to correct the flaws of simple models by tweaking their inputs to match observable prices. Only a

---

essential for our goals. We refer to Schachermayer and Teichmann (2008) and Grunspan (2011) for a more detailed comparison of the two formulas, and more advanced approximation formulas.

horizontal line for the smile indicates model perfection. The magnitude of a smile's deviation from a straight line is, therefore, a measure of the model error, or the degree to which the model misrepresents reality.

The Black smile in Fig. 10.6 is highly asymmetric, which is caused by the natural skewness of the lognormal distribution. The presence of such a steep skew confirms that the market disagrees with the assumption of lognormality. This assumption underestimates the probability of large downside price moves (*down the elevator*), while it overestimates the likelihood of modest upside moves (*up the stairs*). The market is trying to compensate for this shortcoming of the Black model by assigning higher IBV parameters for lower strikes and lower IBV for prices slightly above ATM.

In contrast, the INV smile is much more symmetric, and it appears to be nearly parabolic. This symmetry is intuitive, as oil is, at least, as likely to move down as it is to move up by the same dollar amount. Note, however, that the vertex of the normal parabolic smile is also typically located slightly above ATM. It often corresponds to the point of maximum hedging pressure resulting from selling of slightly OTM calls by oil producers to finance mandatory purchases of puts and put spreads.

The benefits of using the more symmetrical INV are clearly seen in managing the problem of deltas, which could not be rectified with sticky moneyness and sticky strike band-aids applied to the naturally skewed IBV metric. In the normal framework, the deltas for ATM calls and puts are, respectively, positive and negative 50%, and the delta of ATM straddle is zero.<sup>3</sup> This is much more intuitive than the 12% ATM straddle delta in the previous example of the lognormal sticky strike model, or the 28% delta under the assumption of sticky moneyness. The ATM normal delta simply reflects the probability of futures moving up or down from the current level. In other words, the ATM delta also represents the fair price for playing a coin-flipping game to receive one dollar or nothing, depending on whether futures rise or fall. If a random draw that determines the outcome of the game is taken from the symmetric normal distribution, then nobody should be paying more than fifty cents on a dollar to participate in such game.

Having addressed the issue of artificial lognormality-induced skewness, we can now move to handling the more challenging problem of capturing fat tails of the oil price distribution, or large price moves that occur way more frequently than allowed by either the lognormal or the normal assumption. The problem is evident from the wings of both smiles in Fig. 10.6. Since neither assumption can generate sufficiently large probabilities of extreme moves, these models require larger volatility parameters to match prices for deep OTM options. This creates the curvature on both ends of the smile. The upside tail is driven by the risk of supply disruptions caused by unexpected geopolitical or weather events. The downside tail reflects the negative producer gamma and the risks of macroeconomic spillovers, which are further amplified by periodic constraints on the available storage capacity.

---

<sup>3</sup>Recall that we are using zero interest rate. In the general case, deltas must be discounted.

The simple one-parameter models of Bachelier and Black are too rigid to be able to capture the important effects of skewness and fat tails, which are typical for the oil price distribution. At a bare minimum, at least two more parameters are needed to better characterize the joint dynamics of prices and volatilities. Taking some inspiration from the empirical evidence of parabolic normal smiles, we now develop a novel extension of the option pricing framework based on a diffusion process with quadratic local volatility.

## 10.4 Fat Tails and the Quadratic Normal Model

Modeling prices of financial instruments always involves a trade-off between the desired accuracy of the model and its stability. At the one extreme, there are simple one-parameter models, such as the ones developed by Bachelier and Black. They are convenient to use, but, as evidenced by the presence of volatility smiles, these models are not flexible enough to describe the richer dynamics of oil prices. At the other extreme, one can construct very complex models with numerous parameters that can always be tweaked to fit the data, but such fits are unlikely to last long. Once the market moves, overparametrized models quickly fall apart, ruined by the instability of fitted parameters. Such models are characterized by multiple minima in the space of parameters, where various permutations of inputs compete to produce the best fit. Besides, the more parameters the model has, the less intuitive it is for traders. In this section, we propose a sweet spot and modify a simple model just enough to capture the most important features of the oil price dynamics, while retaining transparency and intuition.

At this point, we should elaborate on the difference between what we define as a proper model and ad hoc heuristics, such as sticky strike and sticky moneyness. A heuristic is a certain pattern in the behavior of outputs produced by some unknown pricing engine. These outputs are option prices for different strikes, typically expressed in IBV terms, to which a heuristic rule assigns a certain dynamic without much regard to where such dynamics might have come from. Since the root cause behind such dynamics is not even considered, heuristic rules often fail in practical applications, for example, in the calculation of hedging deltas. In contrast, a proper pricing model should not depend on any option-specific characteristics, such as its strike price, as options across all strikes are functions of the same futures contract. What needs to be modeled is not the evolution of outputs generated by some undefined pricing engine, but the pricing engine itself that describes the dynamics of the underlying futures. Then one no longer needs to guess option deltas and speculate on the dynamics of the smile which is implied by the model.

In our preferred setting of generalized diffusions, the model is described by the local volatility function, as in (8.5). To price an option, one needs to solve the differential Eq. (8.8), where the volatility coefficient is given by a general deterministic function. This equation can be solved analytically only for a few special cases, including constant and proportional volatilities, which we have already considered, and several other volatility specifications that have not proven to be particularly

useful for the oil market. To solve the diffusion equation with any other local volatility functions, one must resort to numerical techniques, such as finite difference methods. However, traders are always a bit leery of numerical black boxes built by quants. Traders like to be in charge of their own destiny and often prefer to use more intuitive analytical approximations rather than numerical methods. The practical benefits of an analytical representation are also evident in the calculation of Greeks, which can be computed by a straightforward differentiation of the closed-form solution for the option price.

A popular technique used in applied sciences for deriving analytical approximations to solutions of complex nonlinear problems is based on the *method of perturbation or linearization*. The idea is to represent a solution to a complex problem as the sum of the solution to a simpler problem and a first-order perturbation of the latter. Conceptually, this is not too different from the idea of a Taylor series expansion, but it is applied to the much more complex functional relationship between option prices and local volatility governed by the diffusion equation.<sup>4</sup>

Given the substantial evidence of the superiority of the normal assumption over the lognormal one, we use the former for the base solution to more complex diffusion problems. The solution to a simplified problem is then given by the Bachelier formula with constant normal volatility. To correct it for its main flaw of underpricing extreme events, a nonlinear perturbation is applied to constant normal volatility. More precisely, we let the local volatility function to be described by a relatively small perturbation of the constant normal volatility, specified by the function  $\varepsilon(F)$

$$\sigma(F) = \sigma_A + \varepsilon(F)$$

The solution to (8.8) is then constructed as the sum of the base solution given by the Bachelier formula,  $C_{BC}(F, t)$ , and the *skew correction function*  $U(F, t)$ , which is designed to capture the impact of non-normality:

$$C(F, t) = C_{BC}(F, t) + U(F, t)$$

Since the function  $C(F, t)$  must solve the diffusion equation with local volatility  $\sigma(F)$ , we substitute this decomposition into (8.8), and make use of the fact that  $C_{BC}(F, t)$  also solves the same equation but with constant coefficient  $\sigma_A$ . The main idea of the perturbation method is to retain only the terms of order  $\varepsilon(F)$  and discard higher-order terms, whose contribution is smaller for a relatively small  $\varepsilon(F)$ .

It is shown in Appendix C that this substitution leads to the following equation for  $U(F, t)$ :

---

<sup>4</sup>The method of linearization and the related parametrix method for the diffusion equation were originally applied to option pricing in Bouchouev (1998, 2000).

$$\frac{\partial U}{\partial t} + \frac{\sigma_A^2}{2} \frac{\partial^2 U}{\partial F^2} = -\varepsilon(F) \frac{n(m_A)}{\sqrt{T-t}} \quad (10.3)$$

where, as before,

$$m_A = \frac{F - K}{\sigma_A \sqrt{T-t}}$$

denotes the scaled normal moneyness. It follows that  $U(F, t)$  also solves the BSM equation with constant coefficient  $\sigma_A$ , but its right-hand side has an additional term which is proportional to the perturbation function  $\varepsilon(F)$  and Bachelier gamma.

To solve (10.3), we need to supplement it with a boundary condition. At the expiration time,  $C(F, T)$  is defined by the option's terminal payoff. The same payoff is also produced by the Bachelier formula  $C_{BC}(F, T)$  at expiration. Therefore, the terminal boundary condition for the skew correction function is zero:

$$U(F, T) = C(F, T) - C_{BC}(F, T) = 0$$

In the regular diffusion Eq. (8.8), the impulse to an option price is sent by its boundary condition at expiration. Here, instead the diffusion is being driven by the new term in the right-hand side of the equation, which impacts the dynamics throughout the life of the option.

The Eq. (10.3) is written for a yet to be specified function  $\varepsilon(F)$  which can be chosen based on properties of the market and observed implied volatility skews. Motivated by empirical observations of parabolic normal smiles from the previous section, we specify the perturbation function in the quadratic form:

$$\varepsilon(F) = a + bF + cF^2$$

Note that while constant parameter  $a$  may appear to be redundant as it is added to another constant  $\sigma_A$ , in practice, it is convenient to write it explicitly. Here, the perturbation is applied to constant volatility  $\sigma_A$ , which is usually taken to be ATM volatility. It only plays the role of the starting point to characterize the approximating normal distribution. For any given  $\sigma_A$ , parameters  $a, b, c$  adjust base option prices for variance, skewness and kurtosis of the underlying market price distribution. Such a parametrization of the local volatility function is known as a *quadratic normal (QN)* model.<sup>5</sup>

The Eq. (10.3) with a quadratic perturbation function is solved analytically in Appendix C. The price of a call option in the presence of skewness and fat tails is given by a simple formula

---

<sup>5</sup>Quadratic parametrizations of the local volatility function have been considered by several authors, including Ingersoll (1997), Zühlsdorff (2001), Andersen (2011), and Carr et al. (2013), but resulting option pricing formulas are considerably more complicated.

$$C(F, t) = C_{BC}(F, t) + V_N \left( v_{QN}(K, F) + \frac{c}{6} \sigma_A^2 (T - t) \right) \quad (10.4)$$

where

$$V_N = \sqrt{T - t} n \left( \frac{F - K}{\sigma_A \sqrt{T - t}} \right)$$

denotes Bachelier vega with the constant normal volatility  $\sigma_A$ , and

$$v_{QN}(K, F) = a + \frac{b}{2}(K + F) + \frac{c}{3}(K^2 + KF + F^2) \quad (10.5)$$

provides corrections to prices for OTM options.

The three-parameter parabolic model appears to represent a sweet spot in modeling oil options that we were seeking. It is significantly more flexible than single-parameter normal and lognormal models. At the same time, the QN model preserves valuable intuition as its three parameters are directly linked to moments of the probability distribution of futures prices. Such intuition is very appealing to traders, as these distributional characteristics are mapped to three benchmark option trades in the market: ATM straddle, costless collar, and OTM strangle.<sup>6</sup> The closed-form nature of the formula (10.4 and 10.5) allows for explicit calculation of all Greeks that traders use for hedging. The model generates a much richer set of deltas depending on the interplay between the skewness and the curvature. It has proven to be able to capture the actual dynamics of oil option prices very accurately.

The method of perturbation is very powerful and can be extended in several directions. If more accuracy is desired, then the perturbation function  $\epsilon(F)$  can be specified in the form of a higher-order polynomial in which case the Eq. (10.3) can still be solved analytically. The perturbation technique can also be applied to the Black formula instead of Bachelier. The corresponding solution retains a structure similar to (10.4) where  $\sigma_A$  is replaced with  $\sigma_G F$ , and normal vega  $V_N$  is replaced with the lognormal vega  $V$ . However, the exact formula is more cumbersome as parabolas are naturally less compatible with geometric volatilities. The formula simplifies only if the local volatility is assumed to be quadratic with respect to  $\ln F$  instead of  $F$ . Furthermore, in the lognormal case, one would be correcting a base model which is worse off to begin with, which makes the perturbation method less robust. Given the inferiority of the quadratic lognormal model, we limit the presentation in this book to a more practical quadratic normal model that works particularly well in the oil market.

In this part of the book, we dealt with vanilla options that depend only on a single futures contract. This allowed us to fix the time remaining to maturity of the option

---

<sup>6</sup>The use of three parameters that correspond to the first three moments of the price distributions is often used by market-makers to characterize the dynamics of the volatility smile in financial markets. For example, Castagna and Mercurio (2007) developed a popular vanna-volga model for options in foreign exchange markets.

and focus on the relationship among option prices across different strikes. In the following three chapters, we add more advanced features that commonly arise in OTC oil options that depend on multiple futures contracts. Trading such options brings an extra dimension. In addition to linkages between strikes, no-arbitrage boundaries must also be established for option contracts with different maturities. Furthermore, the idea of volatility arbitrage presented in this chapter is not complete until we show how to choose the parameters of the pricing model, the topic that will be covered in the chapter on model calibration.

---

## References

- Andersen, L. (2011). Option pricing with quadratic volatility: A revisit. *Finance and Stochastics*, 15(2), 191–219.
- Bouchouev, I. (1998). Derivatives valuation for general diffusion processes, *Proceedings of the Annual Conference of the International Association of Financial Engineers*, New York, USA, pp. 91–104.
- Bouchouev, I. (2000, August). *Black-Scholes with a smile*. Energy and Power Risk Management, pp. 28–29.
- Carr, P., Fisher, T., & Ruf, J. (2013). Why are quadratic normal volatility models analytically tractable? *SIAM Journal on Financial Mathematics*, 4(1), 185–202.
- Castagna, A., & Mercurio, F. (2007). The vanna-volga method for implied volatilities. *Risk*, 20(1), 106–111.
- Derman, E., & Miller, M. B. (2016). *The volatility smile*. Wiley.
- Gatheral, J. (2006). *The volatility surface: A Practitioner's guide*. Wiley.
- Grunspan, C. (2011). A note on the equivalence between the normal and the lognormal implied volatility: A model free approach, *SSRN*.
- Ingersoll, J. E., Jr. (1997). Valuing foreign exchange rate derivatives with a bounded exchange process. *Review of Derivatives Research*, 1, 159–181.
- Rebonato, R. (2004). *Volatility and correlation: The perfect hedger and the fox*. Wiley.
- Schachermayer, W., & Teichmann, J. (2008). How close are the option pricing formulas of Bachelier and Black-Merton-Scholes? *Mathematical Finance*, 18(1), 155–170.
- Zühlsdorff, C. (2001). The pricing of derivatives on assets with quadratic volatility. *Applied Mathematics Finance*, 8(4), 235–262.

---

## **Part IV**

### **Over-the-Counter Options**



# Volatility Term Structure and Exotic Options

11

- Exchange-traded contracts constitute only one side of the oil derivatives market. The other one is hidden from public eyes in the obscure world of over-the-counter (OTC) trading, where more complex options trade bilaterally. Among thousands of OTC deals, one particularly secretive sovereign hedging program stands out.
- Many OTC options depend on contracts with multiple maturities. The pricing of such options is driven by the forward dynamics of the local volatility. The quadratic mean of the local volatility over the lifespan of an option determines the implied volatility quoted in the market.
- An average price option (APO) is the preferred hedging instrument among end-users whose exposure to oil prices is ratable. While precise APO valuation is not possible in the conventional lognormal setting, a simple transformation inspired by the normal model produces an accurate approximation.
- The expiration of certain OTC derivatives is decoupled from the expiration of the underlying futures. One example of such a derivative is a swaption, which represents a strip of early-expiring European options. The pricing of swaptions and many other exotic oil derivatives requires a multi-factor framework.

---

## 11.1 Dark Pools and the Hacienda Hedge

When the media talk about oil options, they almost certainly reference options traded on organized exchanges. These are the only oil options that individual investors can easily monitor and trade. Exchange-listed options are relatively transparent. All transactions, prices, volumes, and open interest are visible to the public. Professional volatility traders call them *vanilla options*. These options are tied to corresponding monthly futures contracts, expiring three business days prior to the futures expiration. Vanilla options are listed for a wide range of strike prices with \$0.50/bbl increments. Such a fine granularity and transparency makes them popular among speculators, who can pinpoint their wagers and easily track their performance.

However, these options represent only one visible side of the oil options market; the other one, which operates OTC, is covered with mysteries.

An OTC market is bilateral, where prices are privately negotiated between professional dealers and end-users. The market for private oil transactions developed long before any of the modern organized oil exchanges.<sup>1</sup> The introduction of oil exchanges did not replace the OTC market, which is still very significant. One large incentive for corporate hedgers to trade directly with banks is the need for credit. While vanilla options are tightly margined by clearing houses, privately negotiated collateral agreements that support OTC trades are more lenient. For end-users, securing favorable margining terms to support their hedging program is crucial. An oil producer can hardly afford to take the risk of posting cash collateral against mark-to-market losses on forward hedging. Since collateral must be posted with the exchange immediately, while the production is still to come, any short-term price spike can easily put the producer out of business. In contrast, banks have proven to be very accommodating to producers, often waiving cash margin requirements and instead accepting oil reserves as a form of collateral against hedging deals. Likewise, in the case of consumer hedging, there was even a precedent of an airline using an aircraft as collateral on OTC oil hedging. Obviously, such credit accommodation by dealers comes with a price. Banks usually act in a dual capacity as a lender and as the dealer for derivatives hedges, which are sufficiently profitable to compensate for carrying additional credit risk.

Another advantage of the OTC market is its versatility. In contrast to vanilla options, bilateral oil derivatives are highly customized. Standardized exchange-traded options are not suitable for many producers and consumers whose exposure to oil prices is ratable. Buying an option contract that settles based on the futures price on a given day is unlikely to be a good hedge for the end-user. The futures price on any single day can be easily impacted by a multitude of one-off factors. Instead, corporate hedgers resort to the services of OTC dealers, who can provide more tailored protection against unwanted risks. In addition, the end-user can benefit from the credit line extended by the dealer where no collateral is required against mark-to-market hedging losses up to a certain limit.

Since end-users are usually exposed to oil prices daily, what they need is a hedge against fluctuation in the average price of oil. End-users can usually live with daily price volatility if the average oil price remains within their budget. A better hedge for them would be an option that settles based on the average futures price over a certain period rather than on a particular day, when the price can be easily distorted. Such OTC options are known as *average price options (APOs)*, or *Asian options*. A typical OTC oil option transaction involves a quarterly or an annual strip of monthly settled APOs. Each day the settlement price of the prompt futures contract contributes to the calculation of the monthly average. If futures expire sometime

---

<sup>1</sup>It should be noted that oil exchanges did exist in Pennsylvania and New York as early as in the 1870s, see Giddens (1947). However, all of them were closed within a decade under monopolistic pressure from the Standard Oil Company.

---

during the month, like in the case of a WTI contract, then the monthly settlement is the weighted average of two futures contracts with consecutive maturities. The corresponding weights are based on the number of days that each futures represent the prompt contract.

Intuitively, price averaging must also reduce volatility and make APOs cheaper in comparison to similar maturity vanilla options. This additional benefit of volatility dampening has made APOs particularly attractive for end-users. Conventional APOs are settled monthly, but a hedger can achieve even larger volatility discount by trading so-called *term APOs*, which settle based on the average price over a longer time period, such as the entire year. A term APO, however, provides less protection than a strip of monthly APOs. This is because the latter could well have positive payoffs for certain months, even when the price average over the entire term does not reach the strike price. The longer the averaging period, the more significant the savings are from the APO volatility discount.

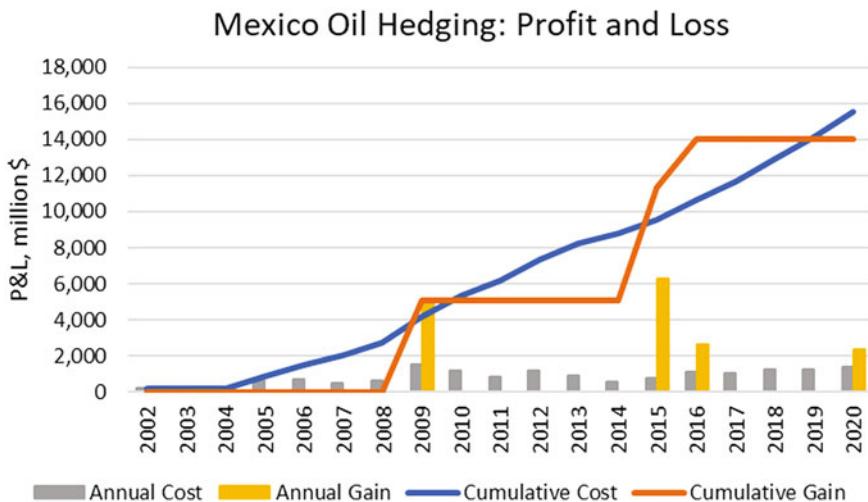
The largest, and arguably the single most important derivative transaction in the oil market is structured as a term APO. No book on oil trading can be complete without describing this unique large-scale sovereign hedging program executed annually by the Government of Mexico. It is designed as an insurance policy to protect the country budget, which is heavily dependent on revenues from oil exports. For over a decade, this hedging program has rocked and rolled oil markets with fortunes made by some traders and large losses suffered by others.<sup>2</sup> The market colloquially refers to it as the *Hacienda hedge*. The program grew to be so important for the overall oil market that the Government of Mexico stopped providing any information about hedging transactions after 2020, unofficially treating it as a state secret.

The concept of hedging the country's oil export revenues was first introduced by Mexico in 1990 in response to the oil price spike at the beginning of the Gulf War. The pilot program was largely successful and ensured higher export prices for the following year. However, it took the country over ten years to fully develop the hedging framework and support it with in-house derivatives expertise. Convincing the lawmakers to spend taxpayers' money on buying oil put options has proven to be challenging given the high political risks in making such decisions. Finally, in 2001 a special structure was created in the form of an oil stabilization fund with an annual budget allocation specifically earmarked for paying the put option premium.

The program evolved over the years, but its core components remained largely intact. The put option is settled as a term APO based on the average futures price over a twelve-month period, usually ending in November, to allow for one month to finalize the option settlement before the end of the calendar year. The strike price, which is typically set slightly below the forward price for the following year, is linked to the price set in the country's budget for the exports of a Mexican oil grade, called Maya. The market price of Maya is determined by a basket of more liquid

---

<sup>2</sup>For a colorful description of the history of this program, see Blas (2017).



**Fig. 11.1** Profit and Loss of hedging program by the Finance Ministry of Mexico. Sources: Auditoría Superior de la Federación, Secretaría de Hacienda y Crédito Público, Oxford Institute for Energy Studies

futures with an additional adjustment by a non-tradable factor set by Mexico itself.<sup>3</sup> The presence of such an unhedgeable risk complicates the task for the dealer, as proxy hedges must be constructed to approximate the dynamics of Maya with liquid futures. The price for this extra hassle and the compensation for bearing an additional basis risk is embedded into the option premium.

Since the government's primary objective is to protect the average export price for the entire fiscal year, it does not necessarily need to pay up for purchasing a more expensive protection against monthly price fluctuations. It should be perfectly content with buying an option settled based on the average price over the entire year. The long-term performance of the Mexican hedging program is representative of a typical insurance payout characterized by regular annual payments and occasional large compensation. Figure 11.1 shows that large payouts occurred in four years when oil prices collapsed. In aggregate, these payouts approximately add up to the cumulative premium paid over the years, confirming that the money was well spent. By using the derivatives market, the government was able to reduce the volatility of the country's budget and accomplish this without paying any excess risk premium. In other words, by and large, the program was executed at a fair actuarial price.

<sup>3</sup>Prior to 2019, the Maya formula was calculated as 40% WTS (West Texas Sour), 40% HSFO (High-Sulfur Fuel Oil), 10% LLS (Louisiana Light Sweet), 10% Dated Brent, plus a K-factor set by Mexico. To reflect the growth in US shale and changes in the sulfur specification for marine fuel, in 2019 the formula changed to 65% WTI Houston, 35% ICE Brent, plus a K-factor.

The fact that Mexico was able to avoid paying any structural VRP over the years by timing the execution well does not mean that option dealers did not make any money. They sure did. The deal highlights the importance of the difference between the actuarial value of an option and its fair value determined by delta hedging, as discussed in Chap. 9. For volatility market-makers, this hedging program turned into a high-stakes skill competition for a slice of the approximately one-billion-dollar premium paid by the government annually. The winners and losers in this competition are often determined by how accurately they can estimate the APO volatility discount for pricing the deal and corresponding deltas for managing the risks.

This sovereign hedge by Mexico is only one among many other examples of a customized OTC derivative, but it illustrates well the complexity of the task faced by volatility traders. These complex deals bring unparalleled profit opportunities to the dealers, but they are inevitably accompanied by equally large risks of mis-hedging if the price dynamics and resulting hedging ratios are not captured correctly. The quantitative challenges here are substantial. In addition to tricky handling of price averaging and basis risks, the trader must also incorporate the volatility term structure, as the tenor of many OTC derivatives spans periods covered by multiple futures contracts.

Up until now, we have dealt only with options written on a single futures contract that allowed us to simplify the problem by fixing the corresponding maturity. However, even a typical monthly APO volatility for WTI depends on the volatilities of two different futures due to the rolls that must occur sometime during the month. Likewise, the volatility of an annual term APO is determined by volatilities of thirteen vanilla options that correspond to prompt futures throughout the calendar year. The volatility of a term APO depends on the slope of the implied volatility term structure, which is particularly sensitive to short-term fundamental uncertainty in the physical market. The analysis of the volatility term structure is a prerequisite for valuing not only APOs, but also many other OTC options which will be introduced later in this chapter.

---

## 11.2 The Term Structure of Implied and Local Volatilities

Modeling non-standard oil options can be a rather daunting exercise. As in many other modeling choices that we have already encountered, one should strive for some reasonable trade-off between the complexity of the model and its robustness. This trade-off depends on the nature of the specific trading opportunity. What works for one trade may not be suitable for another. Traders must always pick their modeling battles based on the dominant risk in the portfolio.

For example, the volatility smile trader focuses on capturing relative value opportunities across options for various strikes, but typically for the same maturity. For such a trader, modeling distributional properties of a given futures contract is more important than linking up contracts with different maturities. Since each vanilla option depends only on a single futures contract, the smile for each maturity can be analyzed in isolation. In contrast, an OTC trader is more concerned about volatility

over a period that spreads across futures with multiple expirations. While capturing key distributional properties is still important for some OTC trades, a more pressing task is to understand how the volatility evolves with time. To get a better sense of the essential time-varying properties of volatility, the term structure of volatility is usually analyzed under simpler distributional assumptions.

In this chapter, we mostly focus on the conventional lognormal framework and the GBM process. We let the geometric local volatility  $\sigma_G(t, T)$  that drives this process depend on the current time  $t$  and on the expiration of the futures contract  $T$  but we make it independent of the futures price:

$$\sigma(F, t, T) = \sigma_G(t, T)F \quad (11.1)$$

In parallel, we also consider the ABM process, defined by the time-varying arithmetic local volatility:

$$\sigma(F, t, T) = \sigma_A(t, T) \quad (11.2)$$

The normal framework turns out to be particularly useful for deriving closed-form option pricing formulas. These analytic formulas are then used as convenient approximations in a more conventional lognormal setting.

The diffusion differential Eq. (8.8) applied to the GBM with time-varying volatility is

$$\frac{\partial C}{\partial t} + \frac{1}{2} \sigma_G^2(t, T) F^2 \frac{\partial^2 C}{\partial F^2} = 0 \quad (11.3)$$

Fortunately, one can easily eliminate volatility time-dependency in (11.3) by switching to a new time variable scaled with the local volatility. It is shown in Appendix D that the solution to (11.3) is given by the regular Black formula with volatility  $v(T)$ , which is calculated as the quadratic mean of the local volatility  $\sigma_G(t, T)$  over the life of the option:

$$v(T) = \sqrt{\frac{1}{T-t} \int_t^T \sigma_G^2(s, T) ds} \quad (11.4)$$

This well-known extension of the BSM pricing framework, originally developed by Merton,<sup>4</sup> has some important implications. It shows that time-dependent volatility does not bring any serious obstacles to the valuation of options. The same Black pricing formula still applies for options on futures. Its single constant volatility parameter is simply replaced with the quadratic mean of the local volatility over the lifespan of the option. Since we have chosen to ignore the impact of the volatility

---

<sup>4</sup>Merton (1973).

smile for this analysis,  $v(T)$  can be taken to represent the term structure of the market implied ATM volatility.

In general, the local volatility  $\sigma(t, T)$  represents a two-dimensional function that describes the volatility behavior at time  $t$  of the futures contract expiring at time  $T$ . In contrast, implied volatility  $v(T)$  is a one-dimensional function of  $T$ . It only contains information about the quadratic mean of the local volatility function over the entire period, and it does not say anything about its path over time. Obviously, there are many local volatility time paths that have the same quadratic mean, i.e., the same implied volatility. This makes it impossible to create a one-to-one mapping between local and implied volatilities without additional assumptions.

To match the dimensionality of local volatility and implied volatility, we impose an extra restriction on the local volatility. Specifically, we assume that it only varies with time remaining to maturity  $\tau = T - t$ . This constraint turns a two-dimensional local volatility function into a one-dimensional curve

$$\sigma_G(t, T) = \sigma_G(T - t) = \sigma_G(\tau)$$

This simplification is perfectly reasonable for non-seasonal commodities, such as crude oil. However, it is not applicable to the market for natural gas and refined products, where volatilities during winter and summer months are markedly different. In the next chapter, we will make some adjustments to incorporate the more complex case of non-homogeneous volatility with respect to time to maturity. For now, we also let the option expiration be the same as the futures expiration  $T$ , an assumption that will also be relaxed shortly. Under these assumptions, the mathematics simplifies considerably, allowing us to identify some essential properties of time-dependent volatility more clearly.

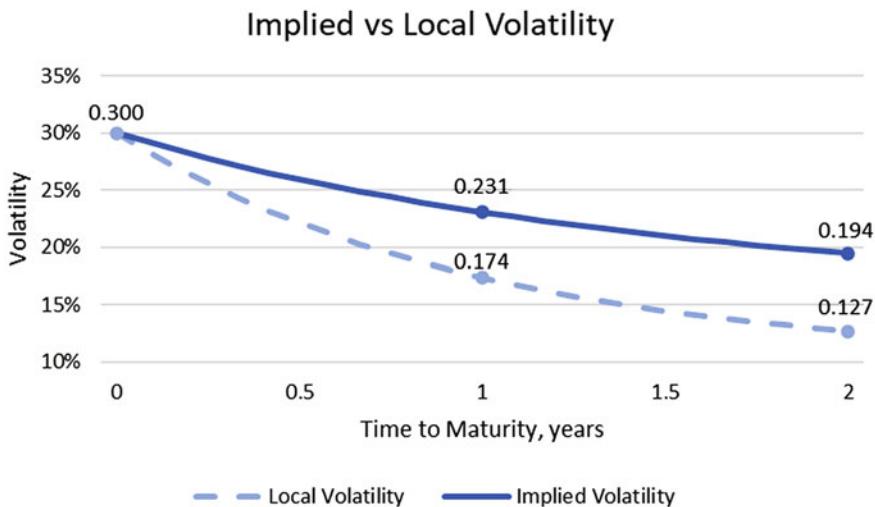
For simplicity, we let the current time to be equal to zero. Then the implied volatility in (11.4) is written as the quadratic mean of the local volatility, which is a function of time remaining to maturity:

$$v(T) = \sqrt{\frac{1}{T} \int_0^T \sigma_G^2(\tau) d\tau} \quad (11.5)$$

To illustrate the important analytical properties implied by this relationship with a specific example, we use the following frequently used exponential parametrization of the local volatility function

$$\sigma_G(\tau) = \sigma_\infty + \sigma_0 e^{-k\tau} \quad (11.6)$$

Here,  $\sigma_\infty$  is the asymptotic volatility of a hypothetical futures contract with infinite maturity. This long-term volatility can be associated with macroeconomic and structural factors that affect the entire futures curve. The short-term volatility  $\sigma_0$  represents additional volatility driven by fundamental factors of supply and demand. The sum of the long-term and short-term volatilities represents the volatility of a hypothetical futures contract with instantaneous oil delivery. The closer the



**Fig. 11.2** An example of implied and exponentially decreasing local volatility with  $\sigma_\infty = 0.1$ ,  $\sigma_0 = 0.2$ ,  $k = 1$

futures contract is to expiration, the larger the contribution of the short-term volatility to the total volatility. The volatility decay factor  $k$  measures how long it takes for the short-term fundamental uncertainty to dissipate. Under normal market conditions, the volatility function (11.6) is consistent with the Samuelson effect, illustrated in Fig. 8.8.

Implied volatility  $v(T)$  for each futures contract that corresponds to the local volatility  $\sigma_G(\tau)$  specification (11.6) is then easily calculated by straightforward integration of (11.5), as follows:

$$v(T) = \sqrt{\sigma_\infty^2 + 2\sigma_\infty\sigma_0 \frac{(1 - e^{-kT})}{kT} + \sigma_0^2 \frac{(1 - e^{-2kT})}{2kT}}$$

Figure 11.2 illustrates the relationship between  $\sigma_G(\tau)$  and  $v(T)$  for options expiring up to two years forward.

For each maturity  $T$ , the implied volatility represents the quadratic mean of the local volatility over the lifespan of the option. For example, for a two-year option, implied volatility  $v(2) = 0.194$ , which is calculated by averaging the local volatility squared along its path from  $\sigma_G(2) = 0.127$  until the option expiration, when the local volatility rises to  $\sigma_G(0) = 0.30$ . Likewise,  $v(1) = 0.231$  is the quadratic average of the local volatility along its path from  $\sigma_G(1) = 0.174$  to the volatility at expiration  $\sigma_G(0) = 0.30$ . Since the implied volatility represents a certain average of the local volatility, it is smoother, and its slope is flatter than the slope of the local volatility.

The reduced slope of the implied volatility term structure relative to the local volatility is characteristic of a cumulative quantity, being an attenuated version of the local building blocks that it is constructed with. The same smoothing effect can be

illustrated with a trivial example from the interest rate market. The two-year interest rate is effectively the average of two consecutive one-year forward interest rates. If such one-year forward rates are 2% and 3%, respectively, then their average, or the two-year rate, is 2.5%. If the second one-year forward rate suddenly jumps from 3% to 4% while the first one remains unchanged, then the two-year average rate only moves by half as much, from 2.5% to 3%. Any move in a local quantity is smoothed by averaging. In the case of implied volatility, such smoothing is done by integrating the local variance.

The methodology for elimination of time-dependent volatility, presented in Appendix D for the conventional lognormal assumption, applies identically to time-dependent normal volatility. In the normal case, the Bachelier formula still applies with its volatility input calculated as the quadratic mean of the arithmetic local volatility

$$\nu_N(T) = \sqrt{\frac{1}{T-t} \int_t^T \sigma_A^2(s, T) ds} \quad (11.7)$$

Having established this important linkage between implied and local volatilities, we now proceed to the problem of calculating the volatility discount that results from price averaging, which is crucial in the valuation of APOs.

### 11.3 Volatility Discount from Price Averaging

In financial markets, APOs are considered to be exotic derivatives that trade rather infrequently. In contrast, in the oil market an APO is the primary derivative instrument used by corporate hedgers to manage ratable exposure to prices. The impact of large APO deals conducted in the OTC market is an important driver of volatility and skews. These APOs are hedged using vanilla options with any arbitrage opportunities between the two instruments quickly captured by market-makers. However, these trading opportunities are not trivial as pricing and hedging of APOs is model-dependent.

The challenge of handling APOs in the traditional lognormal framework comes from the fact that the average of lognormal variables is not lognormal. Conventional pricing approaches tend to focus on approximation methods for the sum of lognormal variables and various numerical techniques.<sup>5</sup> They usually do not cover the term structure of the local volatility, assuming it to be constant. In contrast, in the oil

---

<sup>5</sup>There is a large body of literature on pricing APOs. Kemna and Vorst (1990) formulated the partial differential equation for the price of an APO and found a closed-form solution for an option on the geometric average of prices. Such a solution is possible because the geometric average of lognormal variables is also lognormal. Subsequently, many authors used the idea of geometric averages to develop approximations of the probability density function for the arithmetic average of lognormal variables and corresponding approximations for APO prices, such as the formulas of Turnbull and Wakeman (1991) and Levy (1992). For other APO pricing methods, see also Wilmott et al. (1993),

market the challenge is rather different. Since the underlying price distribution is much closer to being normal than lognormal, one can bypass the analytical complexity by pricing them in the normal world, where the average of normally distributed variables is also normal. Instead, in the oil market it is much more essential to properly capture the time-dependent nature of volatility, which also leads to some important, but often overlooked model modifications.

What makes APOs unusual and somewhat cumbersome to price is their path-dependent nature, as the value of an APO depends on the history of prices during the averaging period. Let  $(T_a, T)$  denote the period over which the futures prices are being averaged. Then once an APO moves inside the averaging period, i.e., when  $t \geq T_a$  then one needs to track the history of prices that have already contributed to the running average. To capture the running average of futures, we introduce a new variable

$$A(t) = \frac{1}{t - T_a} \int_{T_a}^t F(s, T) ds$$

For simplicity, we assume that the price averaging is applied only to a single futures contract. The extension to a more general case where the average is calculated over the rolling prompt contract conceptually follows the same argument.

The payoff of an average price call option that expires at time  $T$  is determined by the terminal value of the running average at time  $T$ :

$$C_{APO}(F, A, T) = \max(0, A(T) - K) \quad (11.8)$$

When an APO moves inside the averaging period, i.e.,  $t \geq T_a$ , then the problem becomes more complex as the option price depends not only on the futures price  $F(t, T)$  but also on the running average of accumulated prices  $A(t)$ . At a first glance, one might be tempted to discard this challenge since APOs usually trade before the pricing period even starts and one may not even care about pricing it within the averaging period. However, as we have seen in the previous section, the price of an option depends on the volatility during the entire lifespan of the option, which includes periods both before and after the averaging starts. Thus, volatility reduction that occurs only during the averaging period still affects the option price at all prior times.

Before the APO moves into the pricing period, i.e., when  $t < T_a$ , the standard delta-hedging argument of Chap. 8 applies and the APO price  $C_{APO}(F, t)$  satisfies the regular diffusion Eq. (8.8):

---

Lipton (2001), and Geman (2005). The method described in this book was originally suggested in Bouchouev (2000).

$$\frac{\partial C_{APO}}{\partial t} + \frac{1}{2} \sigma^2(F, t) \frac{\partial^2 C_{APO}}{\partial F^2} = 0, \quad t < T_a \quad (11.9)$$

However, to solve this equation for  $t < T_a$  one needs a boundary condition at time  $t = T_a$ . This boundary condition can only be obtained by first solving the problem for the subsequent period when  $T_a < t < T$ , for which the boundary condition is given by the option's terminal payoff (11.8) at time  $t = T$ . Note that for brevity, we write local volatility as  $\sigma(F, t)$  omitting the reference to  $T$  which is fixed in this problem.

The pricing problem within the averaging period is multi-dimensional, and, thus, it is much more complicated. In Appendix E we show that an APO price must generally solve a partial differential equation with two spatial variables  $F$  and  $A$ . Fortunately, the specificity of this problem allows the dimensionality to be reduced, resulting in the following pricing equation for an APO

$$\frac{\partial C_{APO}}{\partial t} + \frac{1}{2} \left( \frac{T-t}{T-T_a} \right)^2 \sigma^2(F, t) \frac{\partial^2 C_{APO}}{\partial x^2} = 0, \quad T_a < t < T \quad (11.10)$$

This equation is written with respect to an auxiliary spatial variable

$$x(t) = \left( \frac{t-T_a}{T-T_a} \right) A(t) + \left( \frac{T-t}{T-T_a} \right) F(t, T)$$

which represents the time-weighted average of the accumulated average  $A(t)$  and the futures price  $F(t, T)$  for the remainder of the pricing period.

Note that at the beginning of the averaging period  $x(T_a) = F(T_a, T)$ . As the option moves through the pricing period, the running average accumulates and weights in the calculation of  $x(t)$  gradually shift from  $F(t, T)$  to  $A(t)$ . At the end of the period, the entire average becomes known and  $x(T) = A(T)$ . Therefore, the boundary condition (11.8) for the option's payoff can also be written in term of  $x(T)$

$$C_{APO}(F, A, T) = \max(0, x(T) - K)$$

The second important feature of the Eq. (11.10) is the presence of the multiplier  $\left( \frac{T-t}{T-T_a} \right)$  that applies to the local volatility function. This volatility adjustment is intuitive. As time to maturity of the option shrinks and a larger portion of the average becomes known, the remaining uncertainty decreases accordingly. At the time when the option expires and the calculation of the entire average is completed, the local volatility of an APO reduces to zero. Such volatility reduction near the expiration of the option is very desirable for the volatility dealer. The smoothing dilutes the high uncertainty associated with the gamma risk at the expiration. In contrast to a vanilla option, an APO has very little risk left when  $t \rightarrow T$ , as by then, most of the average is already determined.

While the Eq. (11.10) considerably simplifies the problem, it still cannot be easily solved because the local volatility  $\sigma(F, t)$  in (11.10) for the general diffusion process depends on  $F$ , while the equation itself is written with respect to  $x$ . However, this

equation can be solved for the special case of the ABM process for which the volatility (11.2) does not depend on futures. Applying the argument of the previous section, the solution for the ABM process is given by the Bachelier formula, where the normal APO volatility is calculated as the quadratic mean of its local volatility, as in (11.7):

$$v_{APO}(T) = \sqrt{\frac{1}{T-t} \int_t^T \sigma_{APO}^2(s) ds} \quad (11.11)$$

In (11.11), we omitted the subscript that corresponds to the normal volatility as we will see shortly that the same formula can also be used for lognormal volatility.

The local volatility of an APO is the same as local volatility of futures outside of the pricing period, but it follows from (11.10) that within the pricing period it is adjusted by the linear multiplier

$$\sigma_{APO}(t) = \begin{cases} \sigma(t), & t < T_a \\ \left(\frac{T-t}{T-T_a}\right)\sigma(t), & T_a \leq t \leq T \end{cases} \quad (11.12)$$

Since the integration of the local variance must cover the entire lifespan of the option, one needs to separate two periods when the option is within and outside of the pricing period. If an APO is already pricing out, i.e.,  $T_a \leq t \leq T$ , then only a single integral is calculated from current time  $t$  until maturity  $T$ . However, if the option is yet to start pricing and  $t < T_a$ , then the integration period  $(t, T)$  must be split into two sub-periods  $(t, T_a)$  and  $(T_a, T)$  as the local volatilities in the two periods are different.

In other words, (11.11) and (11.12) can be combined as

$$v_{APO}(t) = \begin{cases} \sqrt{\frac{1}{T-t} \int_t^{T_a} \sigma^2(s) ds + \frac{1}{T-t} \int_{T_a}^T \left(\frac{T-s}{T-T_a}\right)^2 \sigma^2(s) ds}, & t < T_a \\ \sqrt{\frac{1}{T-t} \int_t^T \left(\frac{T-s}{T-T_a}\right)^2 \sigma^2(s) ds}, & T_a \leq t \leq T \end{cases} \quad (11.13)$$

What made this transformation possible for the ABM process is the fact that the average of normally distributed variables is itself normal. If  $F(t, T)$  is normal, then  $A(t)$  and  $x(t)$  are also normal, and the volatility coefficient in (11.10) does not depend on any spatial variable. This property does not hold for the GBM process or other diffusions, and for the general volatility function a closed-form expression for the price of an APO is not possible. However, the same volatility transformation (11.13) can be used as a high-quality analytic approximation for other diffusion processes.

By using this approximation for other diffusions, one effectively assumes that the local volatility function behaves similarly with respect to  $x$  and  $F$ , i.e.,  $\sigma(x, t) \approx \sigma(F, t)$ , in which case (11.10) turns into a regular pricing equation with respect to  $x$ .

For example, under the lognormal assumption it assumes that the average of lognormal variables is approximately lognormal. From the practical standpoint, this is a relatively benign assumption because the real oil price distribution, as we have seen before, is closer to normal than to lognormal anyway, and in the normal case the formula holds exactly. It means that one can still use the Black formula to price APOs with the volatility given by (11.13), which is now understood in percentage terms.

Let us illustrate this first for the constant local volatility

$$\sigma(F, t) = v$$

which can be understood either as normal or lognormal volatility. In this case, the integration of (11.13) is straightforward. After some simple algebra, we obtain the following formula that relates the volatility of an APO to the implied volatility of a vanilla option:

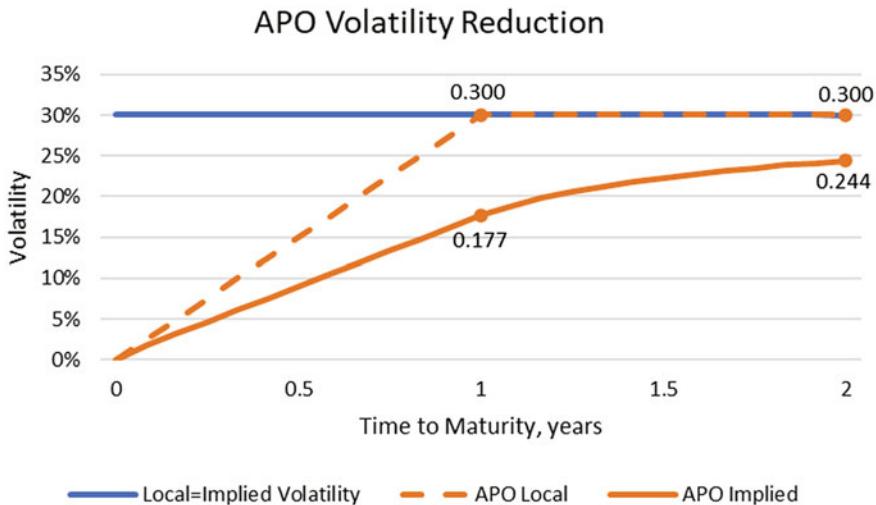
$$v_{APO}(t) = \begin{cases} v\sqrt{1 - \frac{2}{3}\frac{(T - T_a)}{(T - t)}}, & t < T_a \\ v\frac{T - t}{\sqrt{3}(T - T_a)}, & T_a \leq t \leq T \end{cases} \quad (11.14)$$

The nature of the volatility reduction resulting from the price averaging becomes very transparent. One particularly useful rule of thumb can be obtained for pricing an APO precisely at the beginning of the averaging period. When  $t = T_a$ , the formula (11.14) reduces to

$$v_{APO}(T_a) = \frac{v}{\sqrt{3}}$$

It follows that when the averaging is about to start, the implied volatility of an APO should be approximately 60% of the implied volatility of a vanilla option with the same maturity. For the normal Bachelier volatility, this rule of thumb is an exact analytical formula, but for the lognormal Black model it represents an accurate approximation.

Figure 11.3 illustrates an APO volatility discount graphically for a two-year option, where price averaging occurs over the second year with constant Black volatility  $\sigma(t) = v = 0.30$ . Since the local volatility is constant, the vanilla option must be priced with the same constant volatility 0.30 regardless of time remaining to maturity. However, the local volatility for an APO, marked by the dashed orange line, starts decreasing when the option moves into the pricing period and time to maturity falls under one year. Integrating the local APO variance over two years, including the second year where it is linearly reduced, we obtain that APO pricing volatility at the beginning of the averaging period  $v_{APO}(1)$  is only 0.177. This is consistent with the 60% rule of thumb derived above. Likewise, the APO volatility



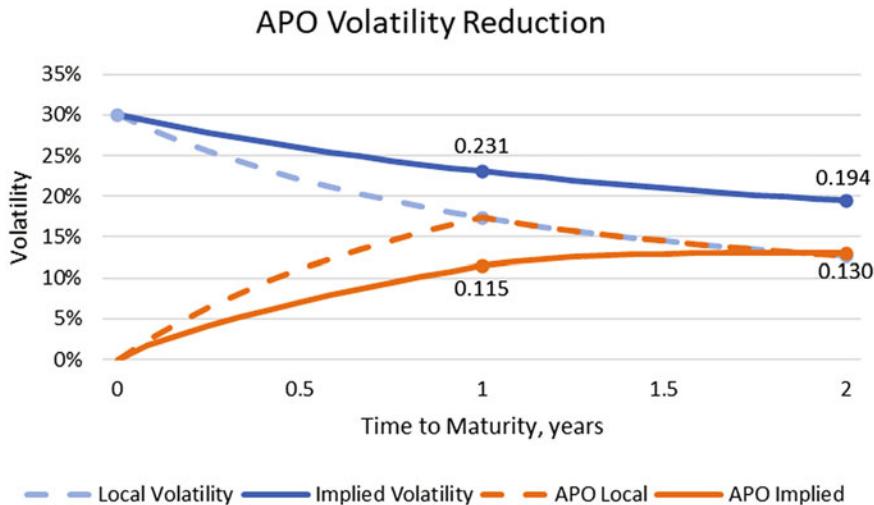
**Fig. 11.3** Volatility reduction for an APO with  $T - T_a = 1$  and constant local volatility

with two years to expiration,  $v_{APO}(2) = 0.244$  is also discounted relative to the Black volatility, but the discount is smaller because outside of the averaging period the local volatilities for the two options coincide.

The volatility discount for an APO is not too sensitive to the choice of normality or lognormality, but it varies substantially with the steepness of the volatility term structure in either framework. To illustrate this, let us modify the example above. Instead of using constant volatility, we calculate the APO volatility discount for an exponentially decreasing local volatility specified in (11.6) with the same set of parameters that we used above. The corresponding volatility behavior is shown in Fig. 11.4.

The local volatilities for a vanilla option and for an APO are represented by the corresponding dashed lines, which coincide prior to the pricing period. The solid lines are implied volatilities which are the quadratic means of the corresponding local volatilities for the same time to maturity. As before, within the pricing period, a volatility multiplier that linearly decreases with time to maturity applies to an APO. For an exponential volatility function, however, volatility attenuation is stronger as compared to the case of constant volatility, as the multiplier dampens local volatility when it is at its highest near the option expiration. For example, at the start of the averaging period APO volatility  $v_{APO}(1)$  is approximately 50% of the corresponding vanilla volatility in contrast to 60% for the constant volatility assumption. Likewise, the  $v_{APO}(2)$  discount relative to the volatility of vanilla option is also larger in comparison to the case of constant volatility. The steeper the term structure of the local volatility curve, the larger the discount for APOs relative to vanilla options across all maturities.

If the option dealer can capture even a fraction of the APO price discrepancy driven by a more accurate modeling of the volatility term structure, it will definitely



**Fig. 11.4** Volatility reduction for an APO with  $T - T_a = 1$  and exponential decreasing local volatility

lead to a stellar career in the oil market. Since many less sophisticated market participants use off-the-shelf software packages for APO pricing that are often constrained to constant volatility specification, they could easily misprice an APO. The APO volatility discount is sensitive to the slope of the volatility term structure, which in turn, is linked to the speed of price mean-reversion. The impact of the volatility term structure becomes even more significant for another important type of OTC options that expire prior to the expiration of the underlying futures.

## 11.4 Early Expiry Options and Swaptions

In this section, we describe another important feature often encountered in OTC oil contracts, which brings additional flexibility to an option buyer by decoupling the option expiration from the expiration of the underlying asset. In contrast, expiration of a vanilla option is always contractually tied to a futures contract that expires during the same month. For example, one can only trade June options written on June futures both expiring in May, or December options written on December futures expiring in November, but an option on December futures that expires in May is not available on exchanges. The OTC market provides this flexibility, where an option can be chosen to expire on any business day, regardless of the expiration schedule for the underlying futures. Such an option in the oil market is known as an *early expiry option (EEO)*.

The choice of the option expiration date is often driven by the timing of certain events that an end-user or a speculator is exposed to. For example, an oil producer may be looking to hedge the long-term oil price risk only during the negotiation

process to acquire another oil company. Such a producer or investor would benefit from buying some protection on long-dated oil prices that impact the value of the target company, but the hedge may only be needed for the duration of a few months while negotiations are taken place. Likewise, a speculator may buy a short-term option on long-term futures specifically to bet on the outcome of a certain event, such as a possible new regulation that is expected to affect forward prices more than spot prices.

Buying a short-term option on long-term futures is significantly cheaper than buying a vanilla long-term option on the same long-term futures because the former has a shorter lifespan. Furthermore, buying a short-term option on long-term futures is also cheaper than buying a vanilla short-term option on short-term futures. Here, the lifespans of the two options are the same, but because the volatility of long-term futures is lower than the volatility of short-term futures, an EEO is worth less. For option pricing, volatility is always determined by integrating the local variance of the underlying futures over the lifespan of the option. For an EEO, this integration only covers a period when the futures are still relatively far from their expiration, and, therefore, they are less volatile during that period. As a result, the integration chops off the most volatile segment of the local volatility, which makes an EEO cheaper.

To quantify this effect, we show in Appendix D that the implied Black volatility of a European option that expires at some time  $T_0 < T$  is given by the quadratic mean of the local volatility during the period  $(t, T_0)$ , which is now shorter than the time remaining to the expiration of the futures:

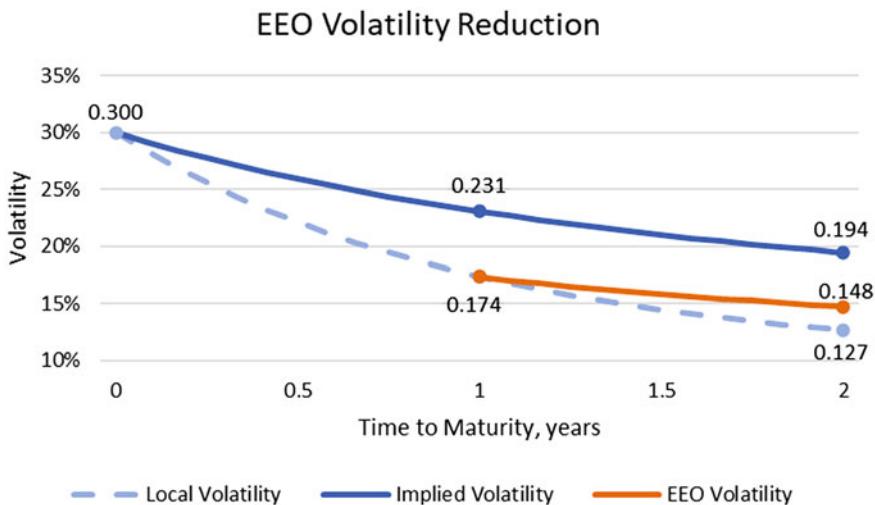
$$v_{EEO}(T_0, T) = \sqrt{\frac{1}{T_0 - t} \int_t^{T_0} \sigma_G^2(s, T) ds}$$

For clarity, we explicitly show the expirations of both the option and the underlying futures as two arguments of the volatility function, and, as before, we omit the subscript that differentiates between lognormal and normal cases. For a vanilla option, where the option expiration is tied to the expiration of futures, one can assume that  $T_0 \approx T$  and, therefore,

$$v(T) = \sqrt{\frac{1}{T - t} \int_t^T \sigma_G^2(s, T) ds}$$

Compared to an EEO, where we integrate over  $(t, T_0)$ , for a vanilla option the same local variance is integrated over a longer period  $(t, T)$  which includes a particularly volatile period near the expiration of the futures. We illustrate the difference between the two volatilities using the same example of an exponentially decreasing local volatility function specified in (11.6) for a two-year option that expires in one year.

The local and implied volatilities shown in Fig. 11.5 are the same as in Fig. 11.2. The rightmost points of the graph correspond to the current time, which is two years until the maturity of the vanilla option. As before, the implied volatility of a vanilla



**Fig. 11.5** Volatility reduction for an EEO with  $T_0 - t = 1$  and exponential decreasing local volatility

two-year option is  $v(2) = 0.194$ . The implied volatility of an EEO, however, is only  $v_{EEO}(1, 2) = 0.148$ . It is calculated as the quadratic average of the local volatility along its path going from right to left, starting from the current time, when  $\sigma_G(2) = 0.127$ , and ending in one year when  $\sigma_G(1) = 0.174$ . The orange line shows that the implied volatility of an EEO increases only modestly as the option approaches its expiration. It does not exceed the maximum local volatility  $\sigma_G(1)$  during the lifespan of an EEO. The higher local volatility of the same futures that occurs later near the expiration of the futures contract does not contribute to the volatility and the price of an EEO.

One can see some similarities between EEOs and APOs. For an APO, the local volatility is damped by a linear multiplier once the option moves into the pricing period. For an EEO, the local volatility for  $t > T_0$  vanishes entirely, as the option no longer exists then. However, direct comparison between the prices of these two types of OTC options is not straightforward, as prices depend on the aggregate effect of the different times to maturity and the slope of the volatility term structure.

While EEOs do trade from time to time in the OTC market, they primarily function as a building block for a more commonly traded OTC derivative, known as a *swaption*, or an option on a *swap*. The standard oil swap settles based on the calendar month average of prices for the prompt futures contract. Like an APO, an oil swap typically trades in the form of quarterly and annual strips with monthly settlements. This means that the swap price can effectively be represented as the linear combination of the  $M$  futures that make up the swap, which are also referred to as *swaptlets*:

$$S = \sum_{i=1}^M \omega_i F_i \quad (11.15)$$

The weights  $\omega_i$  on the futures are determined by the count of days that correspond to each futures contract being a prompt contract and by discounting factors. If we again ignore, for simplicity, the effect of discounting and assume zero interest rate, then the futures weights must add up to one<sup>6</sup>

$$\sum_{i=1}^M \omega_i = 1$$

A swaption is a European option that at time  $T_0$  can be exercised into the strip of monthly swaps with expiries  $T_1 < \dots < T_M$  at the predetermined strike  $K$ . Therefore, at the time of its expiration  $T_0$  that occurs prior to  $T_1$ , the call swaption payoff is given by

$$C_{SW}(S, T_0) = \max(0, S(T_0) - K)$$

Unlike vanilla options or APOs that are settled financially, an oil swaption gives its owner the right to enter into a swap at the contractually specified fixed strike price. This right, of course, is exercised only if the swaption is in-the-money at its expiration. The swaption can be viewed as an EEO whose underlying contract is not a single futures contract, but rather a basket of futures that comprises the swap. To price a swaption, one still needs to integrate and average the local variance of each swaplet over the lifespan of the swaption  $(t, T_0)$ .

Since a swap is just a linear combination of futures, pricing swaptions in the lognormal setting faces a similar challenge to pricing an APO, namely the fact that the weighted average of lognormal variables is not lognormal. The problem again simplifies if it is first solved for normally distributed variables, whose variances are additive, and then extended to the lognormal framework in the form of an approximation. Let us assume that futures for all expirations  $T_i$  are normally distributed and driven by the same single source of uncertainty  $dz$  with corresponding local volatilities  $\sigma_A(t, T_i)$ . In other words, in a risk-neutral world, where the drift term is replaced with zero, the dynamics of futures is

$$dF(t, T_i) = \sigma_A(t, T_i) dz$$

Then the swap, which is the linear combination of futures, also follows the normal process driven by the same  $dz$ :

---

<sup>6</sup>For non-zero interest rates  $r_i$  that correspond to times  $T_i$ , the swap value  $S$  is determined by equating the present value of the future cash flows to zero:  $\sum_{i=1}^M e^{-r_i(T_i-t)}(S - F_i) = 0$ . Solving it for  $S$ , we obtain that  $S$  is given by (11.15) with  $\omega_i = e^{-r_i(T_i-t)} / \sum_{i=1}^M e^{-r_i(T_i-t)}$ .

$$dS = \sum_{i=1}^M \omega_i \sigma_A(t, T_i) dz$$

Therefore, the swaption in the normal setting can be priced using the Bachelier formula. The implied normal volatility of the swaption  $\nu_{SW}$  is calculated as the quadratic average of swaplet volatilities during the lifespan of the swaption  $(t, T_0)$ :

$$\nu_{SW} = \sqrt{\frac{1}{T_0 - t} \int_t^{T_0} \left( \sum_{i=1}^M \omega_i \sigma(s, T_i) \right)^2 ds} \quad (11.16)$$

Here, we again omitted the subscript that corresponds to the normal volatility, because the same formula can be used for the lognormal volatility as well. As in the previous example of an APO, the formula (11.16) is exact for normal variables, the average of which is also normal, but it produces an accurate approximation for the conventional lognormal volatility. In other words, if

$$\frac{dF(t, T_i)}{F(t, T_i)} = \sigma_G(t, T_i) dz$$

then we assume that the swap, which is a linear combination of lognormal variables, is approximately lognormal, and, therefore, it follows the following stochastic dynamics:

$$\frac{dS}{S} \approx \sum_{i=1}^M \omega_i \sigma_G(t, T_i) dz$$

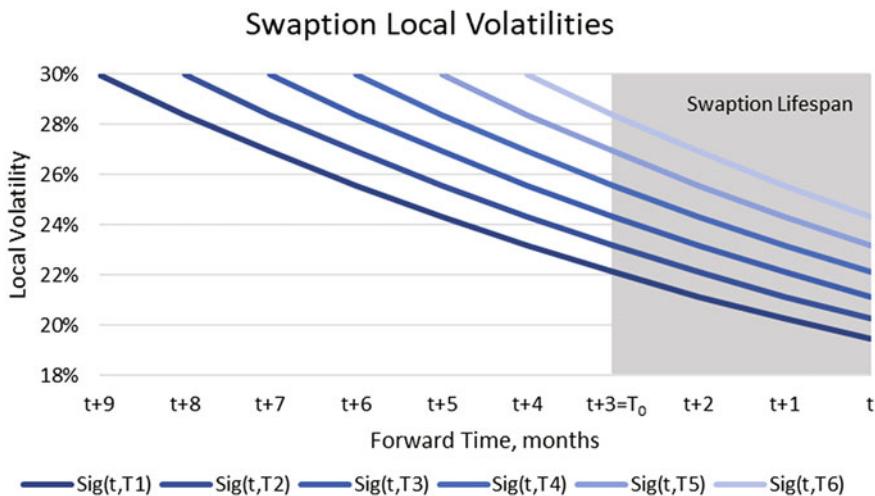
The swaption can then be priced using the conventional Black formula with its lognormal volatility given by (11.16).

For example, for the exponential local volatility of the form (11.6), the corresponding Black swaption volatility is given by

$$\nu_{SW} = \sqrt{\frac{1}{T_0 - t} \int_t^{T_0} \left( \sigma_\infty + \sigma_0 \sum_{i=1}^M \omega_i e^{-k(T_i - s)} \right)^2 ds}$$

which can be evaluated explicitly. Similarly to the case of an EEO, the integration of the local swaption variance does not extend to more volatile periods when futures that make up the swap approach their corresponding expirations. Figure 11.6 illustrates this with an example of a three-month option written on a strip of six monthly swaps, sometimes referred to as a 3x6 swaption.

The shaded area in Fig. 11.6 represents the swaption's three-month lifespan over which the integration of the local variances of the swap constituents is calculated. The most volatile segments for each component near the expiration of the corresponding futures fall outside of this integration area. This results in a much lower implied volatility for the swaption compared to the corresponding vanilla



**Fig. 11.6** Local volatilities for a 3x6 swaption. The shaded area represents the area over which local variances must be integrated

option with the same expiration. We will return to this example in the next chapter on model calibration. In the remainder of this chapter, we provide a brief introduction to multi-factor models, which are more accurate for pricing swaptions, many other complex OTC derivatives, and structured deals that depend on multiple futures contracts.

## 11.5 Multi-Factor Models and Other Exotics

So far, we have only looked at so-called one-factor models, where all futures are driven by the same single source of uncertainty  $dz$ . Such models are generally sufficient for options written on a single futures contract. However, for options that depend on multiple futures contracts, such as swaptions, one-factor models are too restrictive. One-factor models implicitly assume that futures across all maturities are perfectly correlated. While futures with different expirations can move by different amounts as determined by their corresponding local volatilities, in one-factor models all futures can only move in the same direction. The direction is always determined by the sign of their common random component  $dz$ . This behavior is problematic for swaptions and other options that depend on multiple futures. It is easy to imagine a scenario where some futures that make up the swap move in opposite directions in a way that keeps their linear combination or the swap price unchanged. To allow futures to decorrelate, one must introduce a second source of uncertainty.

Fortunately, the volatility algebra for many OTC options remains essentially identical if we model the behavior of futures with a *multi-factor model*. To keep

the presentation simple, we only illustrate this with the industry standard base case of a two-factor lognormal model with exponentially decreasing volatility. The two-factor volatility specification is analogous to (11.6). The only difference is that short-term and long-term futures are now driven by two different but correlated sources of uncertainty  $dz_1$  and  $dz_2$  with correlation coefficient  $\rho$ :

$$\frac{dF(t, T)}{F(t, T)} = \sigma_\infty dz_1 + \sigma_0 e^{-k(T-t)} dz_2 \quad (11.17)$$

The first factor is a hypothetical contract with an infinite maturity, whose volatility  $\sigma_\infty$  is assumed to be constant. The second factor captures the additional volatility associated with short-term fundamental uncertainty that dissipates exponentially with time. Since the variance of the sum of two normal variables is equal to the sum of the variances plus the covariance between them, the total lognormal volatility can be written as

$$\sigma(t, T) = \sqrt{\sigma_\infty^2 + \sigma_0^2 e^{-2k(T-t)} + 2\rho\sigma_\infty\sigma_0 e^{-k(T-t)}}$$

The pricing of a swaption in this two-factor lognormal framework is a straightforward extension of the previous argument. Assuming again that the swap itself is approximately lognormal, then

$$\frac{dS}{S} \approx \sigma_\infty dz_1 + \sigma_0 \sum_{i=1}^M \omega_i e^{-k(T_i-t)} dz_2$$

The usual Black formula applies, with the swaption volatility given by the quadratic average of the corresponding local volatilities over the lifespan of the swaption

$$v_{sw} = \sqrt{\frac{1}{T_0-t} \int_t^{T_0} \left( \sigma_\infty^2 + \sigma_0^2 \left( \sum_{i=1}^M \omega_i e^{-k(T_i-s)} \right)^2 + 2\rho\sigma_\infty\sigma_0 \sum_{i=1}^M \omega_i e^{-k(T_i-s)} \right) ds}$$

Importantly, when  $\rho < 1$ , the introduction of the second factor reduces the overall implied volatility of the swaption in comparison to its volatility in a one-factor model. The lower the correlation between the factors, the lower the value of the swaption, as the likelihood of some swaptlets moving in opposite directions increases, which dampens the volatility of the swap. If the correlation  $\rho = 1$ , then the two-factor model reduces to the one-factor model with the local volatility specified in (11.6).

Occasionally, two-factor lognormal models are further extended into  $N$ -factor models of the form:

$$\frac{dF(t, T)}{F(t, T)} = \sum_{i=1}^N \sigma_i(t, T) e^{-k_i(T-t)} dz_i$$

These more complex specifications are more useful for modeling natural gas and power prices. The purpose of the additional factors and their volatility loadings  $\sigma_i(t, T)$  is to capture strong seasonal effects, as natural gas volatility in the winter is substantially higher than it is in the summer. In the power market, the usage of multiple exponential decay factors  $k_i$  also allows the model to separate the time scales of short-term and long-term mean-reversions, which are typically driven by different forces. Since the efficient storage of power is much more difficult, at least at the time of writing this book, the power market is characterized by large spikes with an extremely fast mean-reversion, which requires a large parameter  $k_i$ . At the same time, for long-term natural gas and power contracts, price mean-reversion is driven by structural factors that evolve much more slowly. Using two exponential factors with different volatility decay parameters captures the dynamics of such markets better.<sup>7</sup>

Multi-factor models are also used for pricing more complex exotic options and structured derivatives in the oil market. In a structured transaction, the dealer often attempts to embed some optionality and present it in a less transparent, but optically attractive way for the end-user. This is, of course, done to maximize the dealer's expected profit margins. For example, swaptions are often embedded in so-called *extendable swaps* that are marketed to producers as a way to achieve higher swap prices for hedges.

Consider a producer who wants to hedge an oil price for one year forward but hesitates whether to hedge or not for the second year. If the producer chooses to hedge for both years, then ignoring again the discounting factor, the hedge price will be the average of two one-year swap prices. Alternatively, the producer can commit to hedging only for the first year and instead boosting the swap price by giving the dealer a unilateral right rather than an obligation to extend the one-year swap agreement for the second year. The producer, therefore, sells a call swaption on the second-year swap that expires at the end of the first year. The premium for selling a call swaption is not paid in cash, but instead it is used to increase the swap price. In this case, the producer improves the hedging price but loses the assurance that the second year risk will be hedged at all. If the second-year price drops by the end of the first year, the dealer will likely waive the option to extend the deal, but at least, the producer receives the higher price in the first year. If the oil price rallies by the end of the first year, then the producer obtains higher price for both years, as compared to the average two-year swap price at the time when the deal was struck.

The flexibility and the complexity of OTC options is practically limitless. Many such options depend on the term structure of futures and volatilities. Extendable options can be quoted with multiple nested extension periods, where the decision for

---

<sup>7</sup>For more details on the time-scale separation for pricing natural gas and power options, see Swindle (2014).

one period may depend on the decision made in the previous period, which makes pricing incredibly complex. Further variations could include *expandable options*, where instead of owning the right to extend the deal in time, the dealer acquires an option to unilaterally expand the deal or, in other words, increase its original volume. For the most part, the hidden objective of the dealer is to buy a relatively cheap optionality from an end-user without making its fair value too apparent. The additional value brought up by such optionality is often used to bridge the gap in the negotiation of the hedging price.

An important component of pricing optionality embedded in complex OTC deals is modeling the creditworthiness of a seller. The higher the leverage in the transaction, the larger the risk of credit default for the dealer. In fact, counterparties with lower credit and limited free cash flows tend to have stronger motivation to sell optionality to pay for the hedge. The value of the collateral often comes into the valuation of such derivatives deals. To limit the credit risk, dealers sometimes introduce caps on the maximum payoff of the deal. One derivative structure with such a capped payout, called a *target redemption swap*, was particularly popular during the global financial crisis. The structuring possibilities in oil derivatives are endless, but many of these complex deals heavily rely on the two key modeling pillars described in this chapter, the impact on volatility from price averaging and from an early expiration. Both effects are captured by certain tricks applied to the local volatility function.

Without any doubt, one can build more sophisticated models to price more complex deals. Analytic valuation for such deals is rarely possible. The standard solution is to price them numerically using a Monte Carlo simulation for the underlying stochastic process. The question is what process to simulate. The results of such simulations are highly sensitive to the chosen specification of the local volatility of the stochastic process. Since all OTC options are ultimately hedged with vanilla options, the simulated process for futures must also generate prices for vanilla options that are consistent with observable market prices. For the professional volatility trader, this process of reconciling theoretical models with market prices is known as *model calibration*. It is the core engine that drives OTC arbitrage, and we devote the next chapter to this important topic.

---

## References

- Blas, J. (2017, April 4). Uncovering the secret history of Wall Street's largest oil trade, *Bloomberg*.
- Bouchouev, I. (2000, July). *Demystifying Asian options*. Energy and Power Risk Management, pp. 26–27.
- Geman, H. (2005). *Commodities and commodity derivatives: Modeling and pricing for agriculturals, metals and energy*. Wiley.
- Giddens, P. H. (1947). *Pennsylvania petroleum 1750–1872: A documentary history*. Pennsylvania Historical and Museum Commission.
- Kemna, A. G. Z., & Vorst, A. C. F. (1990). A pricing method for options based on average asset values. *Journal of Banking and Finance*, 14(1), 113–129.

- Levy, E. (1992). Pricing European average rate currency options. *Journal of International Money and Finance*, 11(5), 474–491.
- Lipton, A. (2001). *Mathematical methods for foreign exchange: A financial engineer's approach*. World Scientific.
- Merton, R. C. (1973). Theory of rational option pricing. *The Bell Journal of Economics and Management Science*, 4(1), 141–183.
- Swindle, G. (2014). *Valuation and risk management in energy markets*. Cambridge University Press.
- Turnbull, S. M., & Wakeman, L. M. (1991). A quick algorithm for pricing European average price options. *Journal of Financial and Quantitative Analysis*, 26(3), 377–389.
- Wilmott, P., Dewynne, J., & Howison, S. (1993). *Option pricing: Mathematical models and computation*. Oxford Financial Press.



# Volatility Arbitrage and Model Calibration 12

- The problem of option pricing is turned upside down. Instead of making assumptions about the stochastic behavior of oil futures, this behavior is reconstructed from market prices of oil options. This inverse problem of model calibration is a blueprint for the volatility arbitrage.
- The local volatility term structure for normal and lognormal processes can be recovered from implied volatilities via a bootstrapping algorithm. However, due to the incompleteness of the oil option market with respect to maturities, this reconstruction is uniquely defined only under additional assumptions on the local volatility.
- In contrast, for a given maturity, the oil option market is relatively complete with prices available for a wide range of strikes, allowing recovery of the entire market-implied probability distribution. Distortions in the shape of this distribution indicate potential arbitrage opportunities.
- A more difficult problem is to reconstruct a diffusion process that generates a given market-implied probability distribution. Simple approximations can be used to extract the local volatility of the diffusion directly from the implied volatility smile. The relationship between the two functions is particularly intuitive for the quadratic normal model.

---

## 12.1 The Inverse Problem of Option Pricing

A typical process of modeling system behavior starts with the description of a phenomenon with certain characteristic variables. The relationship between these variables is then established in the form of mathematical equations. The equations are characterized by a set of parameters that are assumed to be known and estimated from empirical studies. If the equation describes an evolutionary process, then for the solution to be uniquely defined, it must also be supplemented with a boundary condition that defines the initial state of the system. Such a boundary value problem is an example of a *direct problem*. If the solution to the direct problem is unique and

stable with respect to small changes in the boundary condition or with respect to parameters of the process, then the problem is *well-posed*. Most direct problems, including the problem of solving the diffusion equation with a given initial condition, are well-posed.

We have previously highlighted that the problem of option pricing is mathematically analogous to the problem of heat dissipation within a medium after an application of an initial point impulse. The diffusion process is described by the partial differential equation of heat transfer. The properties of the medium are characterized by its thermal conductivity. If the medium is non-homogeneous then its conductivity is modeled as a function of the spatial variable. The solution to the heat equation for a given conductivity function describes the temperature distribution across the medium over time.

The problem of heat dissipation is an example of a direct problem. Such a problem can be written in the form  $y = f(x)$ , where an input  $x$  is transformed by a given model  $f$  into an output  $y$ . If small changes in  $x$  produce small changes in  $y$ , then the problem is well-posed and stable. The diffusion problem is indeed stable with respect to the initial impulse and also with respect to properties of the medium.

Now consider the problem of heat transfer in a composite non-homogeneous medium, whose properties, such as its thermal conductivity, are not well understood. In fact, we would like to deduce these properties by applying multiple heat impulses and measuring how the medium responds to them. Mathematically, the process is still described by the same equation, but the conductivity of the medium is now an undetermined function. However, the solution itself, which is the temperature distribution, can be measured for various boundary conditions. The problem is turned upside down: by measuring the output of the model, what can we say about the property of the material, or its conductivity?

In such a formulation of the relationship defined by  $y = f(x)$ , we now observe both the input  $x$  and its measurable output  $y$ . However, the properties of the engine itself  $f$  that is responsible for the transformation are not fully specified. The question is whether it is possible to observe enough input-output pairs  $(x, y)$ , perhaps for different initial impulses, to recover some information about an unobservable engine  $f$ ? This would be an example of an *inverse problem*.

Inverse problems constitute an important branch of applied mathematics that has contributed to numerous scientific discoveries. In many real-life problems, we cannot easily see what is inside the closed surface, such as deep inside the earth or inside the human body. We can only reconstruct some unobservable properties by sending various signals and measuring the response of the system on the surface that we can easily access. Inverse problems arise everywhere, from medical tomography and the geophysics of oil exploration to military submarine tracking and image recognition by machine learning algorithms. The problem of volatility reconstruction is an example of such an inverse problem in financial markets.

The analogy between the inverse conductivity problem in heat transfer and the inverse problem of volatility reconstruction in financial markets is striking. The spatial variable  $x$  that describes the physical medium represents the range of possible futures prices in the market. The conductivity of the medium is an unobservable

local volatility of the diffusion process for futures prices. The real-time variable is replaced with time remaining to maturity, which turns the option's terminal payoff at expiration into an initial condition. The strike price of a given option plays the role of an initial impulse sent by the option to the market medium of futures. The price of an option is then the reaction to this impulse, which reflects the temperature of the underlying futures market. The response is easily measured, as the market price of an option is provided by a broker. If the market gives us enough option quotes for different strikes and times to expiration, can we then reconstruct the entire engine, the diffusion process with its local volatility that describes the underlying futures market?

Financial analysts are more accustomed to statistical models of the form  $y = f(x)$ , where an empirically estimated input  $x$  is plugged into an econometric model  $f$  to produce a forecast for a variable  $y$ . In contrast, volatility traders are not forecasters, they are arbitrageurs seeking to profit from relative mispricing opportunities. They do not care as much about the historical behavior of prices; for them the purpose of the model is reversed. Volatility traders do not pretend to know what is inside the body of the market, but they can observe how the body reacts to various impulses by measuring its responses via option prices. Both inputs  $x$  and outputs  $y$  in a stylistic representation  $y = f(x)$  are observable, as they represent market prices of futures and options. The goal of the volatility arbitrageur is to reconstruct some unknown properties of the machine  $f$  that turns an observable  $x$  into an observable  $y$  for many  $(x, y)$  pairs. We call this process of backing out unknown properties of the futures market from market prices of traded options the *inverse problem of option pricing*. More colloquially, it is known as the problem of *volatility calibration*.

Why does the volatility trader need this calibration? The volatility trader makes a living by arbitraging prices of options and hedging out residual risks. Since potentially mispriced options are hedged with the liquid ones, the model must reproduce the prices of liquid options correctly. The residual risks are then reduced by delta hedging, and all options within a portfolio must be delta hedged consistently using the same calibrated model for the underlying futures prices. Recall that an option can be dynamically replicated only if its delta is computed using the correct diffusion process for futures. Even though nobody knows with certainty what the correct process is, the one implied by the market produces deltas which are likely to be reflective of the actual changes in option prices, at least over short time periods.

In addition, lucrative arbitrage opportunities exist in trading exotic options and hedging them with vanilla options. As we have seen in the previous chapter, the world of OTC oil options is quite diverse and complex. The pricing of these options is not transparent with closed-form solutions rarely possible for many OTC derivatives products. These deals are usually priced with Monte Carlo simulation, but the trader needs to know what dynamics for futures prices to simulate. The simulated process for futures must also produce prices for vanilla options, which are consistent with market prices, as vanilla options are used for hedging an exotic option. Once the stochastic process is calibrated, many exotic options can be priced at once, using the same Monte Carlo simulation.

Unfortunately, inverse problems, such as the problem of volatility calibration, are generally much harder to solve. In contrast to direct problems, many inverse problems tend to be unstable and *ill-posed*. The ill-posedness of the problem means that a small change in the input could lead to a large change in the solution. In our case the solution to the inverse problem is the property of the medium that we are trying to reconstruct. If the signal happens to be noisy, and many option prices are, then it can only produce a blurry reconstructed image of the market medium.

To solve an inverse problem, one usually needs to smooth, or regularize, the problem. Regularization comes in different shapes and forms, but the main objective of this technique is to find an optimal trade-off between the accuracy of the reconstruction and stability with respect to noisy data. By and large, finding an optimal balance between accuracy and stability is the core premise behind many machine learning algorithms. On the one hand, there are very flexible models that can be tweaked to match almost any data set, but such models suffer from extreme instability. On the other hand, one can build robust and relatively rigid models which are only capable of producing a poor resolution of the desired picture.

The same challenge applies to the inverse problem of volatility calibration. If the model is chosen to be too complex, then its calibration is inevitably more difficult, as many model parameters become less stable with respect to small changes in input data. Finding a sweet spot between model accuracy and stability is an art of option modeling. At the one extreme would be the calibration of the one-parameter Black model. The calculation of implied volatility is robust, but the model description of the real world is rather poor, as the lognormal assumption cannot capture the much richer dynamics of the oil market. At the opposite extreme is the calibration of a multifactor model with volatility smiles. Such a model has too much flexibility, and the solution to the calibration problem is unlikely to be stable, as many combinations of model parameters can produce similar results. An optimally calibrated local volatility can then change drastically in response to even small changes in observable option prices. This would certainly create chaos in delta hedging of derivatives portfolios.

As much as mathematicians dream of developing a grand unified model for pricing everything, such an approach does not work in financial markets. The markets are driven by the behavior of humans and not by the natural laws of any physical process. The oil market is too complex and constantly evolving to be described by a single model. The model and its calibration must always be tailored to the specific nature of the trading strategy. For example, a calibration technique that works for a smile arbitrageur is unlikely to be as useful to an OTC term structure trader, and vice versa. To handle this complex task of volatility calibration, we parse the problem into pieces and develop several foundational building blocks upon which the proper calibration engine can be built. The assembly part of this engine is left to the reader, as it must be tailored to a particular trading opportunity.

## 12.2 Bootstrapping in Time

The first building block of the calibration process deals with the relatively simple problem of reconstructing time-dependent local volatility from the term structure of implied volatilities under simple lognormal and normal distributional assumptions. This part of the calibration process is based on the relationship between the two functions developed in the previous chapter. We now invert this relationship.

Let the dynamics of futures prices be described by the GBM process with lognormal time-dependent volatility of the form (11.1). Following the assumption made in the previous chapter, we first consider the case of local volatility that depends only on time remaining to maturity

$$\sigma_G(t, T) = \sigma_G(T - t) = \sigma(\tau)$$

To simplify the notation, we drop the subscript that corresponds to the geometric volatility since the methodology presented in this section is identical for geometric and arithmetic volatilities. The goal is to recover the term structure of the local volatility  $\sigma(\tau)$  from the term structure of the market-implied volatility  $v(T)$ . As before, if we ignore the presence of the smile, then  $v(T)$  can be taken to represent the term structure of the implied ATM volatilities.

For convenience, we restate the formula (11.5), which represents the implied volatility as the quadratic mean of the local volatility over the lifespan of the option:

$$v^2(T) = \frac{1}{T} \int_0^T \sigma^2(\tau) d\tau \quad (12.1)$$

The calibration problem reduces to an inversion of (12.1) for the local volatility. This can be done by differentiating (12.1) with respect to the upper limit of integration, resulting in:

$$\sigma^2(T) = \frac{\partial(Tv^2(T))}{\partial T} \geq 0 \quad (12.2)$$

Since the local variance on the left-hand side of (12.2) cannot be negative, it sets an important restriction on the shape of the implied volatility function. While the term structure of implied volatilities is expected to decrease with time remaining to maturity as short-term fundamental uncertainty reduces with time, there is a limit on how fast it can decrease.

Any violation of the restriction (12.2) could indicate possible mispricing of options with certain maturities. However, such mispricing turns into an arbitrage only if the assumption of the model holds, specifically, the assumption that volatility depends solely on time remaining to maturity. To distinguish this opportunity from unconditionally riskless profits, we refer to any violation of this boundary as a *model arbitrage*.

Let us illustrate the restriction on implied volatilities, stated by (12.2) for the continuous-time case, with a more practical example of discrete times to expiration. In the oil market, vanilla options are available only with monthly expirations and the partial derivative with respect to time in (12.2) must be discretized. In the discrete case, we continue to use the same letters  $v(i)$  and  $\sigma(i)$  to denote implied and local volatilities but  $i = 1, \dots, N$  now refers to the  $i$ -th nearby futures contract and the volatility during the corresponding month.

The integral in (12.1) that represents implied volatility for  $N$ -th nearby futures contract is then replaced with the sum of local monthly variances over  $N$  prior months

$$v^2(N) = \frac{1}{N} \sum_{i=1}^N \sigma^2(i)$$

As above, volatility is assumed to be time-homogeneous depending only on time remaining to maturity.

One can also express the same implied variance  $v^2(N)$  recursively, as the time-weighted average of the implied variance for the previous period  $v^2(N - 1)$  and the local variance during the latest period  $\sigma^2(N)$ . It is useful to write this recursive relationship explicitly, as follows:

$$\begin{aligned} v^2(1) &= \sigma^2(1) \\ v^2(2) &= \frac{\sigma^2(1) + \sigma^2(2)}{2} = \frac{v^2(1) + \sigma^2(2)}{2} \\ v^2(3) &= \frac{\sigma^2(1) + \sigma^2(2) + \sigma^2(3)}{3} = \frac{2v^2(2) + \sigma^2(3)}{3} \\ v^2(4) &= \frac{\sigma^2(1) + \sigma^2(2) + \sigma^2(3) + \sigma^2(4)}{4} = \frac{3v^2(3) + \sigma^2(4)}{4} \\ &\dots \\ v^2(N) &= \frac{(N-1)v^2(N-1) + \sigma^2(N)}{N} \end{aligned}$$

In the final formula, the implied variance  $v^2(N - 1)$  for the periods  $N - 1$  is combined with the local variance  $\sigma^2(N)$  in the last period to produce the variance  $v^2(N)$  for  $N$  periods.

These recursive relationships can be inverted for the local variance  $\sigma^2(N)$ , which is expressed as the weighted difference between two implied variances for consecutive periods:

$$\begin{aligned}
\sigma^2(1) &= v^2(1) \\
\sigma^2(2) &= 2v^2(2) - v^2(1) \\
\sigma^2(3) &= 3v^2(3) - 2v^2(2) \\
\sigma^2(4) &= 4v^2(4) - 3v^2(3) \\
&\dots \\
\sigma^2(N) &= Nv^2(N) - (N-1)v^2(N-1)
\end{aligned}$$

The last formula is the discrete version of (12.2). This simple recursive algorithm to recover the local volatility term structure from the implied volatility curve is often referred to as *volatility bootstrapping*.

The local volatility can only be calculated if the implied volatility does not decline too fast with increasing time to maturity. To illustrate non-negativity restriction (12.2) with a numerical example, let us assume that the implied volatilities observed in the market for the first three monthly futures contracts are

$$v(1) = 0.30, v(2) = 0.25, v(3) = 0.22$$

Then corresponding local volatilities are computed as follows:

$$\begin{aligned}
\sigma(1) &= 0.30, \sigma(2) = \sqrt{2 * 0.25^2 - 0.30^2} = 0.187, \\
\sigma(3) &= \sqrt{3 * 0.22^2 - 2 * 0.25^2} = 0.142
\end{aligned}$$

So far, the bootstrapping algorithm seems to be working fine, but let us consider the scenario when the implied volatility for the fourth month  $v(4) = 0.19$ . Then the local volatility during the fourth month  $\sigma(4)$  cannot be computed because its variance becomes negative, which is mathematically impossible. In this example, the restriction imposed by (12.2) is violated. The implied volatility  $v(4)$  is too low relative to  $v(3)$ , and its term structure declines with a steeper slope than the maximum allowed by the model. This violation suggests that something is wrong either with the model or with the market.

If the trader believes in the validity of the model, then mispricing could indicate an attractive trading opportunity. Since the implied volatility curve is perceived to be unreasonably steep, the trader could buy a relatively underpriced option on the fourth nearby futures contract and sell a relatively expensive option on the third nearby contract. If the volatility varies only with time remaining to maturity, as specified by the model, then these two options share the same variance during the first three months. However, the longer-dated option must have an additional non-negative value that arises from the price variance during the fourth month. The violation of the restriction (12.2) implies that by selling a short-term option and buying a longer-term option, the trader acquires the local volatility  $\sigma(4)$  during the fourth month for free or even gets paid to own it.

	N	1	2	3	4	5	6	7	8	9	10	11	12	
k	V(N)	0.300	0.310	0.320	0.320	0.316	0.312	0.309	0.306	0.303	0.300	0.298	0.296	
1	$\infty$	sig(1,N)	0.300	0.284	0.269	0.256	0.243	0.232	0.221	0.212	0.203	0.194	0.187	0.180
1	0.1	sig(2,N)		0.334	0.315	0.298	0.282	0.268	0.254	0.242	0.231	0.220	0.211	0.202
1	0.1	sig(3,N)			0.368	0.346	0.327	0.309	0.292	0.277	0.262	0.249	0.237	0.227
1	0.1	sig(4,N)				0.368	0.346	0.327	0.309	0.292	0.277	0.262	0.250	0.238
1	0.1	sig(5,N)					0.366	0.344	0.325	0.307	0.290	0.275	0.261	0.248
1	0.1	sig(6,N)						0.372	0.350	0.330	0.312	0.295	0.279	0.265
1	0.1	sig(7,N)							0.383	0.360	0.339	0.320	0.302	0.286
1	0.1	sig(8,N)								0.389	0.366	0.345	0.325	0.307
1	0.1	sig(9,N)									0.395	0.372	0.350	0.330
1	0.1	sig(10,N)										0.401	0.377	0.355
1	0.1	sig(11,N)											0.414	0.389
1	0.1	sig(12,N)												0.421

0.267 Swaption (3,4:9) Volatility  
0.311 Average Implied (4:9) Volatility

**Fig. 12.1** An example of a local volatility matrix

Such a strategy would indeed be an arbitrage, but only if volatility is time-homogeneous, changing only with time remaining to maturity. If it does not, and if the slope of the implied volatility curve is impacted by uncertainty specific to a particular futures contract, then the trader can easily end up losing money by trying to capture this pseudo-arbitrage opportunity. For example, the drop in implied volatility for the fourth contract could represent a justifiable reduction of uncertainty after the passage of a certain important event, such as an OPEC meeting, which occurred during the previous period. In this case, there is no arbitrage, as both implied and realized volatility will likely decrease during the fourth period. To capture such a scenario, one needs a more flexible modeling framework that incorporates a non-homogenous time-dependency of the local volatility. In fact, the market-implied volatility curve does not even have to monotonically decrease. It can easily take a humped shape with the peak corresponding to the timing of the event after which the market expects the uncertainty to decline.

To incorporate such events into the calibration process, we extend the bootstrapping algorithm to a more general specification of the local volatility  $\sigma(t, T)$  that now depends on both the real time  $t$  and the contract expiration  $T$ . In the discretized setting, such a model is often maintained in the form of a local volatility matrix, an example of which is shown in Fig. 12.1.

The second row above the matrix shows the implied Black volatilities  $v(N)$  for the  $N$ -th maturity contract, which are observable in the market. The implied volatility curve in this example is chosen to be humped. The hump could reflect either some uncertainty associated with a particular event or an abnormal risk premium in the specific option contract caused by hedging imbalances.

A non-homogeneous local volatility in the discrete setting is represented by an upper-diagonal matrix  $\sigma(i, N)$ , where each cell represents volatility measured  $i$  months forward for the contract, which is the  $N$ -th nearby as of today, with  $i \leq N$ . Today is defined by  $i = 1$ . Using this notation,  $\sigma(i, i)$  is volatility of a rolling prompt contract measured  $i$  months forward,  $\sigma(i, i + 1)$  is volatility of the second-nearby contract  $i$  months forward,  $\sigma(i, i + 2)$  is volatility of the third-nearby contract  $i$  months forward, etc. In other words, each row defines local volatility of various

maturity futures at a given forward time. Each column, on the other hand, describes the evolution of local volatilities of the same contract during its lifespan.

The matrix  $\sigma(i, N)$  of local volatilities must be constructed in such a way that it is consistent with the implied volatilities  $v(N)$  observed in the market. Recall that the implied volatility of each option is the quadratic mean of the local volatilities of the same contract, which are represented by the corresponding column in the matrix. Obviously, there are infinitely many ways to fill in a two-dimensional local volatility matrix and still ensure that the quadratic average of each column matches the implied volatility specified in the second row of the same column. We have some additional degrees of freedom at our discretion. We can use them to impose a supplementary structure on our matrix that acts as a form of regularization for this inverse problem. One way to do this is to restrict the shape of the local volatility to a certain functional form.

To illustrate, we use our standard example of an exponentially decreasing volatility function, but introduce an extra degree of freedom by letting the short-term volatility depend on time  $t$

$$\sigma(t, T) = \sigma_\infty + \sigma_0(t)e^{-k(T-t)}$$

If  $T = t$ , then the futures contract represents the spot price with an instantaneous delivery. Thus,  $\sigma_0(t)$  can be understood as the excess of total spot volatility over constant long-term volatility driven by short-term fundamental uncertainty that now depends on time:

$$\sigma_0(t) = \sigma(t, t) - \sigma_\infty$$

In the discrete setting of monthly local volatilities  $\sigma(i, N)$  this assumption is equivalent to

$$\sigma(i, N) = \sigma_\infty + (\sigma(i, i) - \sigma_\infty)e^{-k(N-i)/12}$$

Allowing spot volatility to be time-dependent is particularly helpful for modeling seasonal commodities, such as natural gas and heating oil, for which volatility during winter months is substantially higher. The other two model parameters, the long-term asymptotic volatility  $\sigma_\infty$  and the decay factor  $k$ , are kept fixed. As in the previous examples, we assume that  $\sigma_\infty = 0.10$  and  $k = 1$ . This specification aligns the degrees of freedom between the inputs and the outputs of the model and allows us to construct a one-to-one mapping between implied and local volatilities.

To extend the bootstrapping algorithm to a more general volatility specification and to fill in a non-homogeneous time-dependent local volatility matrix recursively in a unique way, we proceed as follows. The local volatility for the first discrete period is the same as the implied volatility for the first contract:

$$\sigma(1, 1) = v(1)$$

The first row of the matrix, which represents local volatilities for various futures during the first month  $\sigma(1, N)$ , is calculated using the exponential volatility specification, as follows:

$$\sigma(1, N) = \sigma_\infty + (\sigma(1, 1) - \sigma_\infty)e^{-k(N-1)/12}$$

The second row starts with  $\sigma(2, 2)$ , which is the volatility of the contract that is second nearby today, but its volatility is measured during the following month when this contract becomes the prompt. Since the total two-month implied variance for the contract is the sum of its variances during the first month and the second month,

$$v^2(2) = \frac{\sigma^2(1, 2) + \sigma^2(2, 2)}{2}$$

We can then apply the bootstrapping technique and back out the local volatility  $\sigma(2, 2)$  from the implied volatility  $v(2)$  and the local volatility during the first month  $\sigma(1, 2)$  calculated in the previous step:

$$\sigma(2, 2) = \sqrt{2v^2(2) - \sigma^2(1, 2)}$$

The knowledge of  $\sigma(2, 2)$  gives us the pivot spot volatility for the second row, and we can fill in the rest of the second row for all  $N > 2$ , using the specified exponential function, as follows

$$\sigma(2, N) = \sigma_\infty + (\sigma(2, 2) - \sigma_\infty)e^{-k(N-2)/12}$$

Moving on to the third row, the implied variance for the third contract is the average of the local variances over three periods

$$v^2(3) = \frac{\sigma^2(1, 3) + \sigma^2(2, 3) + \sigma^2(3, 3)}{3}$$

Since we already know the local variance for the first two rows and we also know the implied variance that contains the cumulative information from all three periods, the local volatility during the third month can be bootstrapped, as follows:

$$\sigma(3, 3) = \sqrt{3v^2(3) - \sigma^2(1, 3) - \sigma^2(2, 3)}$$

After obtaining the spot volatility during the third month  $\sigma(3, 3)$ , we again calculate  $\sigma(3, N)$  for all  $N > 3$  using the exponential specification for the local volatility. The process continues until the entire local volatility matrix is filled in a unique way.

Having such a local volatility matrix in place is important for pricing swaptions and other exotic options described in the previous chapter. To illustrate its

application to swaptions, we assume that the volatility matrix in Fig. 12.1 is constructed for calendar months, rather than for futures contracts, as one can easily convert between the two. To price a swaption, the trader must calculate the quadratic mean of local swaplet volatilities over the life of the swaption. For example, for a three-month swaption into a six-month swap used in the previous chapter, the volatility of the swaption spans the highlighted area in Fig. 12.1, which is made up of eighteen discrete local volatility blocks. The quadratic average of swaplet volatilities with maturities from four to nine months over next three months produces the swaption volatility. In this example, the swaption volatility is only 0.267, which is lower than any of the implied volatilities for swaplets. It would be a major mistake by the trader to price a swaption by calculating the quadratic mean of implied volatilities for swap components. What goes into averaging are the local variances of swaplets, not the implied ones.

In addition, if some swaption prices are observable in the broker market, then this flexible format of the local volatility matrix also allows its parameters to be further calibrated to the market prices of these swaptions. For example, we still have parameters  $\sigma_\infty$  and  $k$  that remain entirely under our control. We can choose them to approximate market prices of swaptions, EEOs, or some other exotic derivatives that may be quoted in the market. This is typically done in an ad hoc manner by trying out different combinations of  $\sigma_\infty$  and  $k$ , while monitoring how the entire local volatility matrix responds to different combination of inputs. If more flexibility is desired, the decay parameter  $k(t)$  can also be made time-dependent.<sup>1</sup>

A similar volatility matrix can also be built for the two-factor model specification (11.17). Its construction remains largely identical, except for an additional degree of freedom introduced by the factor correlation  $\rho$ . Using it, however, often causes more harm than it adds value as many combinations of parameters may end up producing similar results. For example, one can achieve a drop in the swaption volatility either by lowering the factor correlation, or alternatively, by steepening the local volatility function. The market, however, does not give us enough observable information to distinguish between these two alternatives. The problem becomes over-parametrized, and unless some model parameters are fixed, it becomes prone to instability.

The inverse calibration problem in the time direction is relatively simple analytically. However, this problem is inherently incomplete, as there are too many degrees of freedom relative to the number of available observations. The oil option market does not contain enough information to allow us to deduce the rich dynamics of the volatility time-dependency, so the term structure modeling is often supplemented by econometric analysis of futures prices. In contrast, an inverse problem in the strike dimension where the local volatility is allowed to be space-dependent faces the

---

<sup>1</sup> Another example of a discrete volatility matrix is constructed in Pilipović (2007), where instead of imposing the exponential structure on local volatilities, an additional constraint is set by equating long-term local volatilities to historical realized volatilities. Our preference is to avoid explicitly tying market-implied volatility matrix to realized volatilities due to hedging imbalances and the presence of the volatility risk premium documented in Chap. 9.

opposite challenge. It is relatively complete in terms of the availability of granular data for option prices with different strikes, but it is more challenging from the mathematical point of view. We turn to this problem next.

### 12.3 Market-Implied Probability Distribution

So far, we have only analyzed a rather simple inverse problem of reverse engineering the local variance of a lognormal process from its cumulative variance. What made the inversion possible is a special property of normally distributed random variables whose variances are additive. We only had to express the cumulative variance as the sum of its constituent local variances, and then bootstrap the local variance recursively by taking differences between cumulative variances for options with nearby expirations. Such a bootstrapping requires only ATM volatilities that are uniquely mapped to an implied volatility parameter characterizing the width of the price distribution, which is assumed to be either normal or lognormal. In general, the nature of the price distribution is unknown. It turns out that by taking options prices for all strikes, one can reconstruct the distribution itself. Such a distribution is called a *market-implied probability distribution*.

Recall from (A.10) that the price of a call option can be expressed in terms of the probability density function and the option's terminal payoff as follows:

$$C(F, t; K, T) = \int_0^\infty p(F, t; F_T, T) \max(0, F_T - K) dF_T \quad (12.3)$$

Here, we again ignore the impact of interest rate, assuming it to be zero. The function  $p(F, t; F_T, T)$  represents the probability of the futures price reaching the level  $F_T$  at time  $T$ , given that its price at time  $t$  is  $F$ . This function is also the same as the fundamental solution to the pricing Eq. (8.8). As explained in Appendix A, the fundamental solution solves the Eq. (8.8) with the special boundary condition given by the Dirac delta function.

Importantly, what we refer to as probability is not the real-world probability, but rather it is the so-called *risk-neutral probability*. This is because the real-world stochastic process (8.5) has a non-zero drift  $\mu(F, t)$  and its corresponding probability density function satisfies (A.8). The pricing Eq. (8.8), however, does not have any drift. Even though we still refer to  $p(F, t; F_T, T)$  as a probability, it should be distinguished from the statistical probability of the historical price moves, which can only be measured in the real-world. The Eq. (12.3) relates an option price to the market-implied risk-neutral probabilities generated by an imaginary stochastic process in which the real-world drift term is replaced with zero. Much as the realized volatility is distinct from the implied volatility, the historical distribution of prices is not the same as the market-implied risk-neutral distribution.

To calculate the market-implied probability distribution, we need to invert formula (12.3) for the function  $p(F, t; F_T, T)$ . Since the integration in (12.3) is

performed only over the range of  $F_T$  for which the call option payoff is non-zero, we can rewrite it as

$$C(F, t; K, T) = \int_K^\infty p(F, t; F_T, T)(F_T - K)dF_T$$

We differentiate this formula with respect to  $K$  and obtain that

$$\frac{\partial C}{\partial K} = -p(F, t; K, T)(K - K) - \int_K^\infty p(F, t; F_T, T)dF_T = - \int_K^\infty p(F, t; F_T, T)dF_T$$

The first term comes from differentiation with respect to the lower limit of integration. It is equal to the value of the integrand evaluated at the point  $K$ , which is zero, and, therefore, this term vanishes.

We then take the second derivative with respect to  $K$ , differentiating it again with respect to the lower limit of integration, which results in

$$\frac{\partial^2 C(F, t; K, T)}{\partial K^2} = p(F, t; K, T) \quad (12.4)$$

This is a rather remarkable result. The given set of option prices can be transformed into the set of market-implied risk-neutral probabilities simply by differentiating the option prices twice with respect to the strike price. This relationship was once again first discovered by Bachelier in his seminal doctoral thesis.<sup>2</sup>

The transformation (12.4) is possible only because of the very special terminal payoff of the call option, which is given by  $\max(0, F_T - K)$ . The first derivative of the payoff with respect to  $K$  taken with the negative sign is given by the so-called *digital*, or the *Heaviside, function*. This function is equal to either one or zero, depending on whether  $F_T$  is above or below  $K$ :

$$-\frac{\partial}{\partial K} (\max(0, F_T - K)) = D(F_T - K) = \begin{cases} 1, & F_T > K \\ 0, & F_T \leq K \end{cases} \quad (12.5)$$

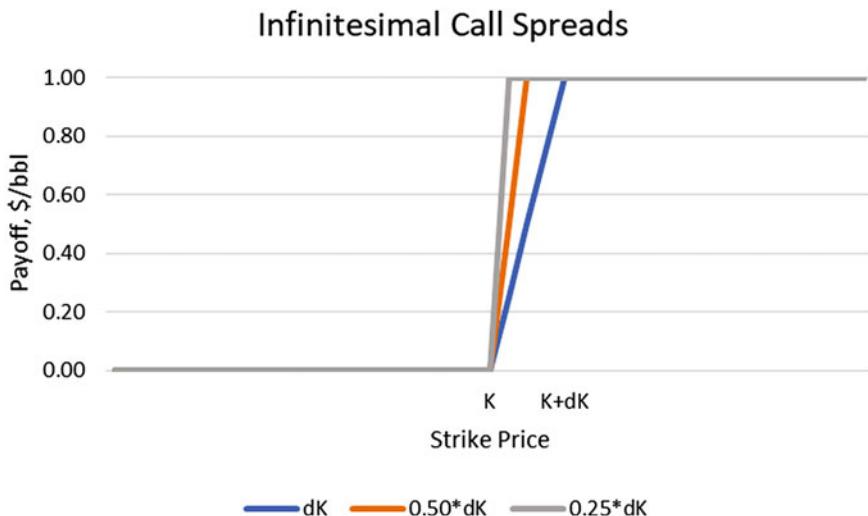
Furthermore, the second derivative of the payoff is equal to the Dirac delta function centered at  $K$

$$\frac{\partial^2}{\partial K^2} (\max(0, F_T - K)) = \delta(F_T - K) = \begin{cases} \infty, & F_T = K \\ 0, & F_T \neq K \end{cases} \quad (12.6)$$

In other words, the second derivative of the terminal option payoff is a point impulse function which represents the terminal state of the probability density

---

<sup>2</sup>See Bachelier (1900). Bachelier's brief statement of this formula is often overlooked in the literature, where the formula is typically attributed to a more comprehensive work on this topic by Breeden and Litzenberger (1978).



**Fig. 12.2** The payoff of the call spread with strikes  $K$  and  $K + dK$  approaches the payoff of the digital option when  $dK \rightarrow 0$

function  $p(F, T; K, T)$ . As the impulse diffuses backward in time for  $t < T$ , the probability density function  $p(F, t; K, T)$  is described by the second derivative of the option price, as in (12.4). For further background on the Dirac delta function, probabilities, and related differential equations, we refer to Appendix A.

Let us illustrate the concept of market-implied probabilities in a more practical discrete case. Since in the real market we only have option prices for a finite set of strikes, the second derivative must be discretized. In the previous section, when we applied the bootstrapping technique to time-dependent volatility, we were able to extract some information about the local quantity from the difference between two closely related global quantities. We will do something similar here and extract the local probability at each point by taking the difference between option prices with nearby strikes, which contains some information about the likelihood of futures being between two strike prices.

Let us construct a call spread trade, where we purchase a call option  $C(K)$  struck at  $K$  and sell a slightly further OTM call option  $C(K + dK)$  struck at  $K + dK$  for some small increment  $dK$ . Both options have the same expiration  $T$ , and, for brevity, we omit explicit references to other option parameters. If we now take  $1/dK$  units of these call spreads and let  $dK$  be infinitesimally small, then this trading position represents the first derivative of the option price with respect to the strike price taken with the negative sign

$$-\frac{\partial C(K)}{\partial K} = \lim_{dK \rightarrow 0} \frac{C(K) - C(K + dK)}{dK}$$

The payoff of such call spreads are shown in Fig. 12.2 for decreasing strike increments.

The maximum payoff of one call spread is  $dK$ . Therefore, the maximum payoff of holding  $1/dK$  units of such infinitesimally tight call spreads is equal to one dollar. In the limit when  $dK \rightarrow 0$ , the call spread converges to a *digital option*  $D(F_T - K)$ , which is defined by (12.5). The digital option pays one dollar if the futures price is above the strike  $K$  at expiration and zero if the futures price is below the strike:

$$-\frac{\partial C(K)}{\partial K} = D(F_T - K) = \begin{cases} 1, & F_T > K \\ 0, & F_T \leq K \end{cases}$$

Digital options do trade in the OTC oil market. Volatility traders typically manage the risk of selling a digital option by buying the tightest possible call spreads to approximate the discontinuous payoff. The first derivative of the option price is itself a popular financial instrument. Many traders appreciate the simplicity of the digital payoff, which is similar to a lottery ticket that pays a fixed amount in the event that the futures price exceeds the strike price at expiration. The digital payoff is also identical to the option's delta, and prices of the digital option under normal and lognormal assumptions are simply given by the formulas for deltas provided in Appendix B.

We now consider the so-called *butterfly spread*, where we purchase an equal quantity of calls struck at  $K - dK$  and  $K + dK$  and sell twice as many calls struck at  $K$ . The butterfly can also be understood as the spread of two call spreads, where one buys the  $C(K - dK) - C(K)$  call spread and sells the  $C(K) - C(K + dK)$  call spread. If we scale the call butterfly position by holding  $1/(dK)^2$  units of this trade and make  $dK$  again infinitesimally small, then this portfolio is equal to the second derivative of the option price with respect to the strike  $K$ :

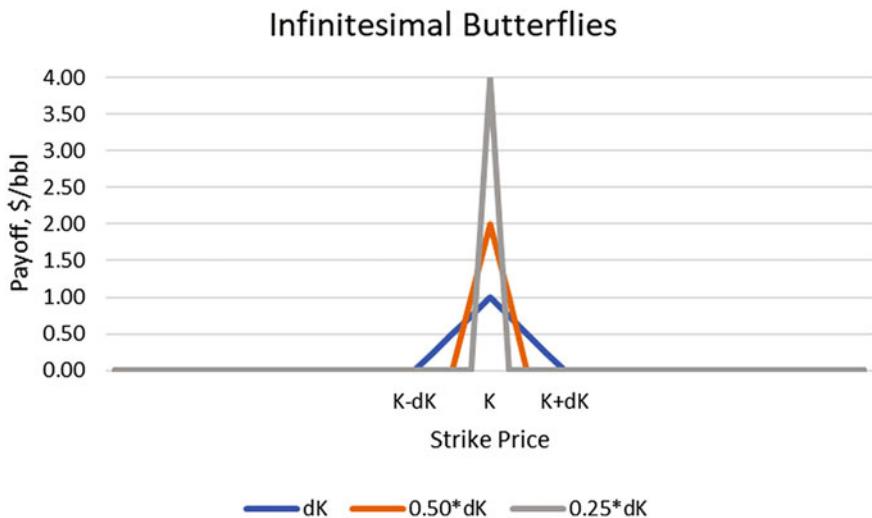
$$\frac{\partial^2 C(K)}{\partial K^2} = \lim_{dK \rightarrow 0} \frac{C(K + dK) - 2C(K) + C(K - dK)}{dK^2}$$

The terminal payoff of the butterfly call spread is shown in Fig. 12.3 for decreasing strike increments.

The butterfly trade is certain to generate a non-negative payoff which is defined by the triangular region with a width of  $2dK$  and peak height  $dK$ . Therefore, if we hold  $1/(dK)^2$  units of such butterflies, then the area of the triangle is always equal to one, regardless of the size of  $dK$ . If we let  $dK \rightarrow 0$ , then the second derivative of the option price with respect to  $K$  turns into the Dirac delta function, centered at  $K$ :

$$\frac{\partial^2 C(K)}{\partial K^2} = \delta(F_T - K) = \begin{cases} \infty, & F_T = K \\ 0, & F_T \neq K \end{cases}$$

Note that since the area of the triangle in the butterfly payoff is equal to one for any  $dK$ , the normalization property (A.4) of the Dirac delta function that integrates to one is also satisfied. The butterfly payoff is the discrete analogue of (12.6).



**Fig. 12.3** The terminal payoff of  $1/(dK)^2$  units of call butterflies with strikes  $K - dK$ ,  $K$ , and  $K + dK$  approaches the Dirac delta function when  $dK \rightarrow 0$

Obviously, the second derivative of the option price does not explicitly trade in the market as it would have an impossible infinite payoff, but its approximations in the form of tight butterflies with small  $dK$  are quite common even on the exchanges, where the smallest strike increment is only fifty cents per barrel. The maximum payoff of the infinitesimally tight butterfly occurs when  $F_T = K$ , and traders use butterflies to explicitly bet on the probability of the price reaching the specific level at expiration. Since at expiration the payoff of the butterfly approximates the Dirac delta function centered at  $K$ , the market prices of various butterflies with various strikes make up the probability density function for the futures price to be at the level  $F_T = K$  at expiration.

Market-implied probability distributions are very helpful for volatility traders in spotting arbitrage opportunities in option prices across different strikes. The probability density function is generally expected to be smooth. Any visible kink in the density function indicates that prices for some options with nearby strikes might be inconsistent. It is rather difficult to identify such an inconsistency by looking at option prices or at their implied volatilities, which are cumulative quantities, because averaging tends to smooth the effects of any mispricing. In contrast, for local quantities, such as probability densities, there is nowhere to hide, and any distortion in its shape exposes a pricing anomaly. Looking at the implied probability distribution is like looking at option prices under the microscope.

The recovery of the market-implied distribution from option prices is undoubtably useful in identifying potential trading opportunities. However, to capture such opportunities, additional information is needed. The market-implied distribution provides us only with a snapshot of the price distribution taken at a given time, but it tells us little about the evolution of prices between today and the option's

maturity. Since the opportunity in trading options is captured by delta hedging, which must be done throughout the life of the option, to calculate the deltas, the trader needs to model the evolution of the entire stochastic process that produces this distribution. This information is contained in the local volatility function, which is also needed for pricing and hedging OTC options.

## 12.4 Local Volatility Smile

The conditional price distribution contains much less information than the stochastic process itself. In fact, many different stochastic processes could produce the same distribution at a given time horizon. Consider, for example, the trivial case of the normal distribution. Any time-dependent local volatility with the same quadratic mean produces the same implied volatility, and, therefore, the same terminal distribution. Moreover, a mean-reverting process, such as the one used for modeling oil inventories in Chap. 3, also has a normal probability density (A.17) which can alternatively be generated by a different ABM process without mean-reversion. The problem of reconstructing the entire diffusion process from a snapshot of the terminal distribution is ill-posed as it admits multiple solutions. To create a one-to-one mapping between the price distribution and the underlying stochastic process, one must restrict the degrees of freedoms in the model specification for the process.

Let us first consider a hypothetical case and assume that at present time  $t$  when the futures price is  $F$ , all option prices  $C(F, t; K, T)$  are available for the continuum of strikes  $K$  and maturities  $T$ . Then these option prices are related to the local volatility function of the diffusion process via the following equation:

$$\frac{\partial C}{\partial T} - \frac{1}{2} \sigma^2(K, T) \frac{\partial^2 C}{\partial K^2} = 0 \quad (12.7)$$

This equation is known as the *Dupire equation*.<sup>3</sup> In the Appendix F we show how it can be easily derived by integrating the Fokker-Planck equation for the probability density function.

While such an assumption about option prices may be unrealistic, the Dupire equation plays an important role in solving inverse calibration problems, much like the BSM equation does in solving direct problems of option pricing. The structure of the Dupire equation resembles the BSM Eq. (8.8) which is written with respect to futures  $F$  and current time  $t$ . However, in (12.7), the futures price  $F$  is replaced with the strike price  $K$ , and real time  $t$  is replaced with option expiration  $T$ . In addition, the direction of time is reversed. While the BSM equation moves backward in real time for  $t < T$  starting from the terminal boundary condition when  $t = T$ , the Dupire

---

<sup>3</sup>The equation was presented at several conferences in 1993 and subsequently published in Dupire (1994).

equation moves forward with respect to maturities  $T > t$  with an initial boundary condition at  $T = t$ .

The BSM and Dupire equations are complementary to each other. They are somewhat analogous to the Kolmogorov backward and forward (Fokker-Planck) Eqs. (A.8) and (A.6) for the probability density function. However, the BSM equation holds for any derivative, but what makes the existence of the dual Dupire equation possible is the unique nature of the call option payoff, whose second derivative with respect to the strike price happens to represent the probability density function, as shown by (12.4). If the option had any other payoff profile, then a complementary dual equation for option prices would not have been possible.

The Dupire equation provides a stylish theoretical path towards solving the inverse calibration problem. The local volatility function, which generates theoretical option prices that match market prices with all strikes and maturities, can be backed out from option prices as follows:

$$\sigma(K, T) = \sqrt{\frac{2 \frac{\partial C}{\partial T}}{\frac{\partial^2 C}{\partial K^2}}} \quad (12.8)$$

Note that while the actual local volatility function depends on  $F$  and  $t$ , when it is reconstructed from option prices, its functional arguments are replaced with  $K$  and  $T$ .

To understand this representation better, let us revert to a discrete approximation of partial derivatives of option prices with respect to strikes and maturities. We know from the previous section that the denominator of (12.8) is the probability density function, which can be approximated by market prices of call butterflies. Similarly, the numerator of (12.8) can be approximated by an infinitesimally tight call spread with respect to  $T$ .

Consider a calendar call spread which consists of long and short call options with the same strike  $K$  and corresponding expiration times  $T + dT$  and  $T$ . If we hold  $1/(dT)$  units of this calendar call spread and let  $dT \rightarrow 0$ , then such a trading position becomes the partial derivative of the option price with respect to  $T$ :

$$\lim_{dT \rightarrow 0} \frac{C(K, T + dT) - C(K, T)}{dT} = \frac{\partial C}{\partial T}$$

Therefore, the local volatility at the point  $(K, T)$  can be approximated by the square root of the ratio of  $2/(dT)$  units of calendar call spreads to  $1/(dK)^2$  units of call butterflies. This relationship shows how the local volatility function ties together option prices with nearby strikes and maturities. In this discretized example, the local volatility at each point is determined by a portfolio of four options, two from the calendar spread and two more from the wings of the call butterfly. Since these options also impact neighboring local volatility points, all option prices across strikes and expirations are deeply intertwined via this not so obvious relationship between their partial derivatives.

Unfortunately, direct application of the Dupire equation to the oil market faces some limitations. While the assumption about availability of options with the continuum of strikes is reasonable given that vanilla options are listed with granular strike increments, a similar assumption does not hold for options with different maturities. For a given futures contract, vanilla oil options exist only for one maturity date in the same month. It would be difficult to calculate the  $T$ -derivative in the Dupire formula as options on the same contract with other maturities do not exist. One would need to know prices for all EEOs on the same futures, which are clearly not available. The only way to apply Dupire's equation to volatility calibration in the oil market is to estimate the  $T$ -derivative from the market price of a calendar call spread with the same strike but on *different* maturity futures. However, such an approach would implicitly assume that all futures are identically distributed, which is very precarious in the oil market where different maturity futures reference different physical barrels.

In practice, the volatility calibration problem in the oil market must be simplified. Instead of trying to recover two-dimensional local volatility function  $\sigma(F, t)$  from option prices with all strikes  $K$  and all maturities  $T$ , one usually fixes the maturity and focuses on recovering time-independent local volatility  $\sigma(F)$  for a given futures contract using option prices with various strikes. Unfortunately, for such a problem the Dupire equation cannot be used as the partial derivative with respect to  $T$  is not known.

It turns out that such an inverse problem has very unique and important scientific applications. It is analogous to the problem of reconstructing the conductivity of a non-homogenous medium from temperature observations at a given time. It also represents the problem of reconstructing a time-independent diffusion process from a snapshot of the probability density function. Surprisingly, such a problem has been solved only under additional assumptions, but in its general case, at least to the best of the author's knowledge, it presents a rare example of an open mathematical problem. We formulate this problem more precisely in Appendix F. It is fascinating how the problem of option pricing that gave rise to an entire branch of modern mathematics continues to push science to new frontiers. If any readers choose to solve this inverse problem instead of taking advantage of volatility arbitrage opportunities in the oil derivatives market, then the author would consider the mission of this book to be largely accomplished.

Despite its theoretical shortcomings, practitioners have no choice but to come up with a way to back out the local volatility numerically. Many computational techniques have been developed for recovering  $\sigma(F)$  from option prices with fixed maturity, but nearly all of them suffer from numerical instability. This instability is not a flaw of the algorithms, but rather it is a salient property of this inverse problem.<sup>4</sup>

---

<sup>4</sup>This inverse problem has been extensively studied in academic literature. See, among many others, Avellaneda et al. (1997), Bouchouev and Isakov (1997, 1999), Dempster and Richards (2000), Carr

One way to simplify and tame an instable inverse problem is to impose a certain parametric structure on the local volatility which acts as its automatic stabilizer. Instead of trying to bootstrap the entire function  $\sigma(F)$ , one can assume that  $\sigma(F; \theta_i)$  is defined by parameters  $\theta_i$  and then minimize the distance between theoretical  $C$  ( $\sigma(K; \theta_i)$ ) prices that depend on  $\theta_i$  and observable market prices  $C^*(K)$ :

$$\min_{\theta_i} \sum_K (C(\sigma(K; \theta_i)) - C^*(K))^2$$

For example, in the QN volatility model of Chap. 10, parameters characterize variance, skewness, and kurtosis of the underlying price distribution. These parameters can be fitted to market price of options with many different strikes. Alternatively, one can choose to use only three liquid options benchmarks for ATM straddle, a costless collar and a strangle and match them to three parameters of the QN model. In this case, the least squares minimization is replaced with three nonlinear equations with respect to three model parameters. In fact, many traders prefer not to fit the model to all strikes, arguing that the three-parameter model provides a proper balance by imposing enough structure on the model while leaving the residuals from this parametric fit as an indication of mispricing.

While parameter fitting approach is widely used by quants, professional market-makers always look for some shortcuts that may help them to gain some intuition into the cumbersome process of volatility calibration. What the trader sees in the market is an implied volatility smile  $v(K)$ ; what is needed for calibration is the local volatility function of the diffusion process  $\sigma(F)$ . Recall that for lognormal and normal processes, implied and local volatilities are related via (12.1). Since for these processes implied volatility is represented as some average of instantaneous local volatility in the time direction, the natural question is whether a similar relationship between implied and local volatilities may also exist in the price direction?

Unfortunately, there are no simple analytic formulas that relate local and cumulative variances for more complex price distributions. Conceptually, the implied volatility still represents some average of the local volatility, but this average is calculated along certain stochastic paths taken by the futures price. Imagine pricing an option using a Monte Carlo simulation, where many different paths for the futures price are generated by the diffusion process with the calibrated local volatility. The option value is then determined as the average payoff across all simulated paths. Each path starts at the current futures price  $F$ , but only certain paths that cross the strike price  $K$  contribute to the option value because the payoff from the remaining ones is zero.

---

and Madan (2001), Bouchouev et al. (2002), Chiarella et al. (2003), Alexander (2008), Lipton and Sepp (2011), and references therein.

We can, therefore, internalize the implied volatility as an average of the local volatility but compute the average only along such paths from  $F$  to  $K$  that contribute non-zero values to the option price.<sup>5</sup> However, in the special case of the QN model we can derive a more elegant solution for the relationship between implied and local volatilities. If we equate the formula (10.4) to option prices expressed in terms of implied normal volatilities  $v_N(K)$  and apply the Taylor formula, we obtain that

$$\begin{aligned} C_{BC}(\sigma_A) + V_N(\sigma_A) \left( v_{QN}(K, F) + \frac{c}{6} \sigma_A^2 (T - t) \right) \\ = C_{BC}(v_N(K)) \approx C_{BC}(\sigma_A) + \frac{\partial C_{BC}(\sigma_A)}{\partial \sigma_A} (v_N(K) - \sigma_A) \end{aligned}$$

Therefore,

$$v_N(K) \approx \sigma_A + v_{QN}(K, F) + \frac{c}{6} \sigma_A^2 (T - t) \quad (12.9)$$

Furthermore,  $v_{QN}(K, F)$  represents the average of the local volatility between  $F$  and  $K$ , which is verified by direct integration:

$$\frac{1}{K - F} \int_F^K (a + bx + cx^2) dx = a + \frac{b}{2} (K + F) + \frac{c}{3} (K^2 + KF + F^2) = v_{QN}(K, F)$$

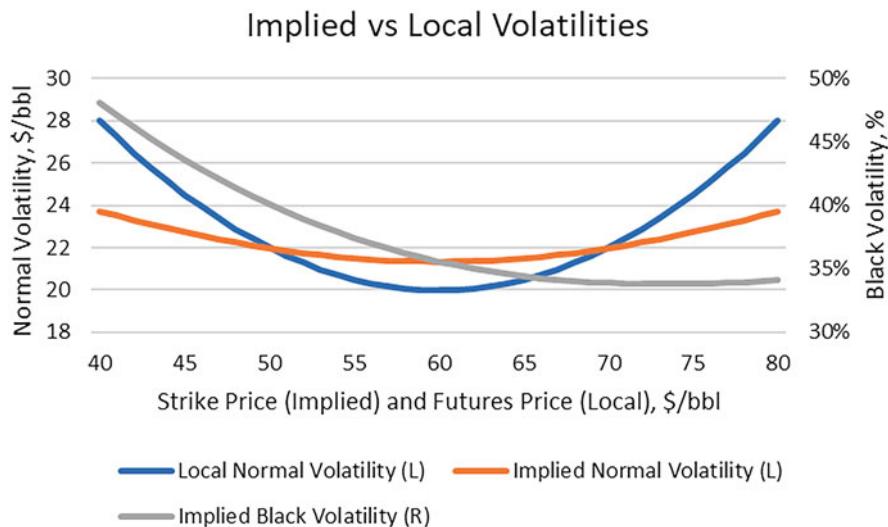
Figure 12.4 illustrates the relationship between the quadratic local volatility with respect to futures and the corresponding implied normal and Black volatilities with respect to strike prices. Implied normal volatility is also approximately quadratic. However, its curvature is only one-third of the local volatility curvature, as implied volatility smooths the perturbation of the local quantity via averaging.

Another useful example is given by a linear local volatility function for which the curvature  $c = 0$ . In this case, the implied volatility is also approximately linear, but its slope is only one-half of the slope of the local volatility. Again, the implied volatility is flatter than the corresponding local volatility as the cumulative function smooths the average of its local constituents.

Figure 12.4 also shows how option prices produced by the QN model would be seen through the incorrect lenses of the lognormal model. Equivalent Black volatilities that correspond to QN volatilities can be calculated either numerically or by using the approximation formula (10.2). The QN model produces Black volatility skew which is representative for the crude oil option market. The bottom of the skew, which corresponds to slightly OTM calls, indicates the impact of producer pressure that results from selling calls via two- and three-way collars. The skew exhibits positive curvature on both ends, which is consistent with fat-tailed distributions and the larger volatility risk premium for OTM options.

---

<sup>5</sup>For a more detailed discussion of this topic, we refer to Gatheral (2006), and Derman and Miller (2016).



**Fig. 12.4** Implied volatilities generated by the local volatility  $\sigma(F) = 20 + 0.02(F - 60)^2$

The reconstruction of the local volatility and the corresponding stochastic process for future prices forms the core of the pricing framework that must be built by professional volatility traders. With calibrated volatility, all hedging ratios are consistently calculated within the same framework, and traders no longer need to guess whether the smile is fixed or floating. The smile behavior is entirely determined by the diffusion process with the calibrated local volatility function. In addition, the local volatility allows option traders to price many exotic derivatives of high complexity all at once. Once the local volatility is reconstructed, complex options can be priced by using a Monte Carlo simulation. Thousands of forward price paths are simulated by the diffusion process with the reconstructed local volatility, and the value of the complex derivative is computed as the derivative's average payout for such simulated paths.

The diffusion framework is, by far, our preferred choice for modeling oil options. However, we should also recognize that no model is perfect, and we briefly mention its shortcomings and alternatives often proposed in the literature. One drawback of the diffusion assumption is the difficulty of calibration to short-term maturity options. The implied volatility smile is often very steep because of the shortage of natural sellers for deep OTM options. The local volatility, therefore, must be even steeper. To match observed market smiles, the local volatility is then forced to rise with unrealistic slopes. This comes from the continuous nature of the diffusion process, which does not allow for jumps. Allowing local volatility to rise sharply for short-term maturity options can also be viewed as an attempt to incorporate such jumps. It should also be understood that since the short-term implied smile contains a

larger volatility risk premium, the corresponding local volatility may not accurately represent the actual dynamics of the futures.

Another popular alternative to the diffusion framework is allowing volatility to be stochastic. Stochastic volatility processes can also be calibrated to match the market observed smile. However, the forward smile dynamics implied by such models is less intuitive than the one generated by local volatility models. Many popular stochastic volatility models produce inadequate deltas which resemble deltas from the sticky moneyness heuristics. Even though for many other financial markets, such as equities, stochastic volatility models are widely used, as far as the oil market is concerned, diffusion seems to represent a better choice.<sup>6</sup>

Perhaps most importantly, the diffusion pricing framework guarantees the absence of arbitrage, as the risks of the option and the futures are driven by the same source of uncertainty, which can be hedged. This is not the case in jump and stochastic volatility models, which introduce other non-tradable sources of risks that require more assumptions and more unobservable parameters. Diffusion models are complete in the sense of their ability to eliminate risk by hedging, while more complex jump and stochastic volatility models are not, and for that reason we chose not to cover them in this book. For the volatility arbitrageur, the model completeness that allows for consistent pricing and hedging is far more important than complex multi-factor frameworks with unobservable parameters.

The inverse problem of volatility calibration is a non-trivial task. Its implementation is often the primary assignment for quants on option trading desks. Calibration techniques are typically customized to individual markets and even to individual large deals, such as the hedging program of the Government of Mexico. Going into more specifics here would significantly distract us from the overall purpose of this book. All traders, however, need to be aware of the importance of model calibration, and reconstruction of the volatility engine from observable prices of liquid options. The volatility calibration creates an ultimate roadmap for identifying and capturing arbitrage opportunities in the options market.

---

## References

- Alexander, C. (2008). *Market risk analysis, Vol. III: Pricing, hedging and trading financial instruments*. Wiley.
- Avellaneda, M., Friedman, C., Holmes, R., & Samperi, D. (1997). Calibrating volatility surfaces via relative entropy minimization. *Applied Mathematical Finance*, 4(1), 37–64.
- Bachelier, L. (1900). *Théorie de la Spéculation*, Annales scientifiques de l'École Normale Supérieure, Série 3 17, 21–86.
- Bouchouev, I., & Isakov, V. (1997). The inverse problem of option pricing. *Inverse Problems*, 13(5), L11–L17.
- Bouchouev, I., & Isakov, V. (1999). Uniqueness, stability and numerical methods for the inverse problem that arises in financial markets. *Inverse Problems*, 15(3), R95–R116.

---

<sup>6</sup>For further reading on jump-diffusions and stochastic volatility models, we refer to Rebonato (2004), Javaheri (2005), Gatheral (2006), and Derman and Miller (2016).

- Bouchouev, I., Isakov, V., & Valdivia, N. (2002). Recovery of volatility coefficient by linearization. *Quantitative Finance*, 2(4), 257–263.
- Breeden, D. T., & Litzenberger, R. H. (1978). Prices of state contingent claims implicit in option prices. *Journal of Business*, 51(4), 621–651.
- Carr, P., & Madan, D. (2001). Determining volatility surfaces and option values from an implied volatility smile. In M. Avellaneda (Ed.), *Quantitative analysis in financial markets* (Vol. II, pp. 163–191). World Scientific.
- Chiarella, C., Craddock, M., & El-Hassan, N. (2003). An implementation of Bouchouev's method for short time calibration of option pricing models. *Computational Economics*, 22, 113–138.
- Dempster, M. A. H., & Richards, D. G. (2000). Pricing American options fitting the smile. *Mathematical Finance*, 10(2), 157–177.
- Derman, E., & Miller, M. B. (2016). *The volatility smile*. Wiley.
- Dupire, B. (1994). Pricing with a smile. *Risk*, 7(1), 18–20.
- Gatheral, J. (2006). *The volatility surface: A Practitioner's guide*. Wiley.
- Javaheri, A. (2005). *Inside volatility arbitrage: The secrets of skewness*. Wiley.
- Lipton, A., & Sepp, A. (2011, October). Filling the gaps. *Risk*, 24(10), 78–83.
- Pilipović, D. (2007). *Energy risk: Valuing and managing energy derivatives*. McGraw-Hill.
- Rebonato, R. (2004). *Volatility and correlation: The perfect hedger and the fox*. Wiley.



- The payoff of a storage asset can be replicated with calendar spread options (CSOs). A synthetic storage strategy buys long-dated CSOs at discounted “wholesale” prices from asset owners, monetizes value by delta hedging, and resells options at premium “retail” prices to financial investors.
- The volatility of the spread cannot be lower than the difference between the volatilities of the spread components. The relationship between vanilla and spread options is bound by the triangular arbitrage. Violation of the arbitrage boundary can generate riskless profits that do not depend on any models or assumptions.
- While many energy spreads trade as independent assets, certain pairs can also be constructed as spreads between two correlated prices. This dichotomy leads to two alternative approaches to pricing and hedging of spread options.
- Correlation-based spread option models are unstable at high levels of correlation, which is typical for petroleum spreads. The volatility of energy spreads is better estimated using analogue fundamental regimes that are indicative of potential disruption risks.

---

## 13.1 The Synthetic Storage Strategy

Our last chapter on trading oil options can be viewed as its grand finale. It is a remarkably simple strategy, but it encompasses many concepts presented throughout the book, starting with the critical role of oil storage. This trading opportunity has played a special role in the author’s own trading career, which started with the launch of this strategy in the 1990s when quantitative oil trading was yet in its infancy.<sup>1</sup> Importantly, this storage strategy may find new applications during the

---

<sup>1</sup>The story of how the strategy was originated and developed is described in Johnson (2022).

energy transition, which has shown the desperate need for new alternative forms of storage, including batteries.

Storage is arguably the most valuable asset for a professional oil trading shop. It gives the owner a powerful option to shift a limited resource from times of plenty to times of relative scarcity. This option is particularly valuable if supply and demand balances are expected to change. Storage pays off only when the market is in contango, and when the trader can cover the cost of storage and financing by buying physical barrels at a sufficiently large discount relative to the futures price. When the market is backwardated, the storage option is effectively out-of-the-money, and the option premium is largely wasted.<sup>2</sup> In other words, a storage asset represents a real put option on the shape of the futures curve with the strike price determined by the cost of storage.<sup>3</sup>

In the early days of the oil derivatives market, quantitative traders were not part of the elite group of physical traders who had access to physical storage facilities. To substitute, quants created virtual storage and synthesized the storage asset on paper. This was done by means of a financial put option written on the futures time spread. Today this option, known as a *calendar spread option (CSO)*, represents the most liquid exotic oil option, with its trading volumes rapidly growing not only OTC but also on the exchanges.

More formally, a call option on the spread between the two futures  $F_1$  and  $F_2$  is defined by the following payoff at the expiration time  $T$

$$C_{SP}(F_1, F_2, T) = \max(0, F_1 - F_2 - K) \quad (13.1)$$

where  $K$  is the strike price.

Similarly, the terminal payoff of a put option on the spread is given by

$$P_{SP}(F_1, F_2, T) = \max(0, K - (F_1 - F_2)) = \max(0, F_2 - F_1 - (-K)) \quad (13.2)$$

In other words, the put option on  $F_1 - F_2$  with the strike  $K$  is the same as the call option on  $F_2 - F_1$  with the strike  $-K$ . Here, we keep generic notations  $F_1$  and  $F_2$  for the two futures that may represent either contracts on the same commodity with different expirations or futures on two different commodities. In the former case, which is a CSO,  $F_1 = F(t, T_1)$  and  $F_2 = F(t, T_2)$ , with  $T_1 < T_2$ . Standard CSOs contractually expire on the so-called *penultimate day*, which is one business day before the expiration of the futures, in this case the futures with the shorter maturity  $T_1$ .

---

<sup>2</sup>In this chapter, we ignore blending optionality of the storage asset, which allows the owner to mix different grades of crude oil and to sell the blend at a premium by customizing it to the rigorous specifications of a refinery.

<sup>3</sup>Likewise, the cost of freight determines the strike price of the locational spread option on the difference between oil prices in two different regions. Considine et al. (2022) relates the value of such an option to inventories.

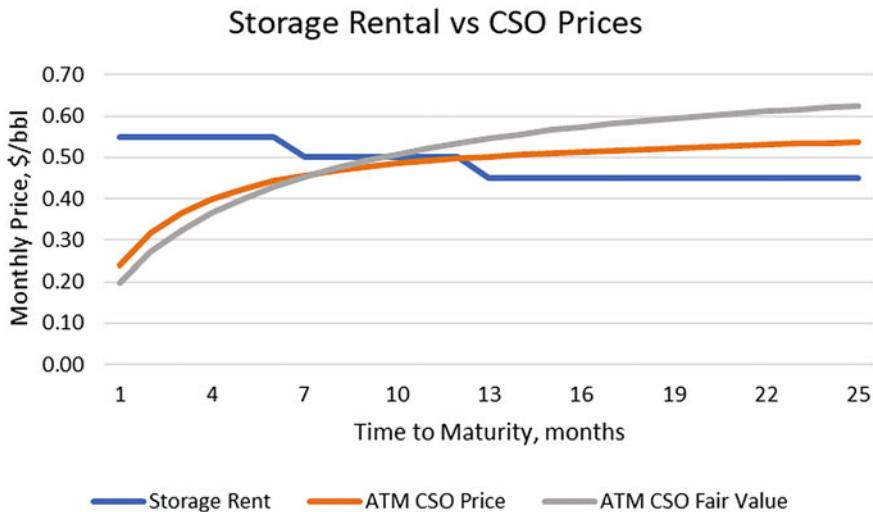
Two types of CSOs trade both on exchanges and OTC. One type is physically exercised into futures spreads at the expiration, similarly to vanilla options. The other one is settled financially, based on the settlement price of the spread on the penultimate day. When a physically settled CSO is in-the-money at the expiration and is 100% hedged with the futures spread, then the hedge automatically offsets futures received from exercising the CSO. In contrast, a fully-hedged financially settled in-the-money CSO requires the trader to liquidate futures in the market as close as possible to the settlement price for the spread, which is typically done using the previously discussed TAS contract. The liquidation of these futures, which are held as hedges against expiring CSOs, can cause additional spread volatility on the penultimate day. For example, such a liquidation played an important role in the episode of negative oil prices discussed in Chap. 3.

With the use of CSOs, financial traders can closely replicate cash flows of physical storage without getting their hands dirty in the somewhat cumbersome operation of a storage facility. The option premium is equivalent to the rental fee which is paid by physical traders to lease the storage facility. In the physical market, the storage lease is typically set for the fixed term, such as one year, with payments made monthly. A strip of monthly put options on the spread can then be used to mimic the revenues of the storage asset.

In theory, the value of any option must increase if the option has more time before its expiration. Surprisingly, this is not always the case in the market for physical oil storage. Storage facilities are often owned and operated by independent companies. The owner then leases storage to professional traders, which effectively means selling asset optionality in exchange for the monthly rent. The longer the rental commitment, the more attractive the deal is to the risk-averse operator, who places higher value on the certainty of securing fixed payments for a longer time period. For a physical trader, however, it is rather difficult to commit to a long-term storage lease due to the fluid nature of the trading business with its relatively short-term vision and frequent changes in personnel and investors' risk appetite. As a result, the term structure of rental fees in the storage market is often relatively flat. It may even decrease slightly for long-term deals, which are often discounted by operators to secure a longer-term income stream. Such a pricing structure is similar to the real estate market and apartment leases.

This is where an opportunity arises for quantitative traders who can construct a more liquid version of synthetic storage in the derivatives market. Ignoring the impact of interest rates, the value of a financial option must increase with time remaining to maturity. If it did not, then one would sell a shorter-term option, buy a longer-dated option, and enjoy holding the latter with zero theta since the option only increases in value even as its time to expiration shrinks. This is in contrast to the much flatter term structure of rental fees in the physical market. It gives a derivatives trader an opportunity to pay a higher premium to the storage owner relative to the owner's alternative of renting out the asset to a physical trader.

From the operator's perspective, selling an option to a quantitatively minded trader instead of leasing it in the physical market does require a bit more trading savviness. The physical barrels now must be moved in and out of storage efficiently

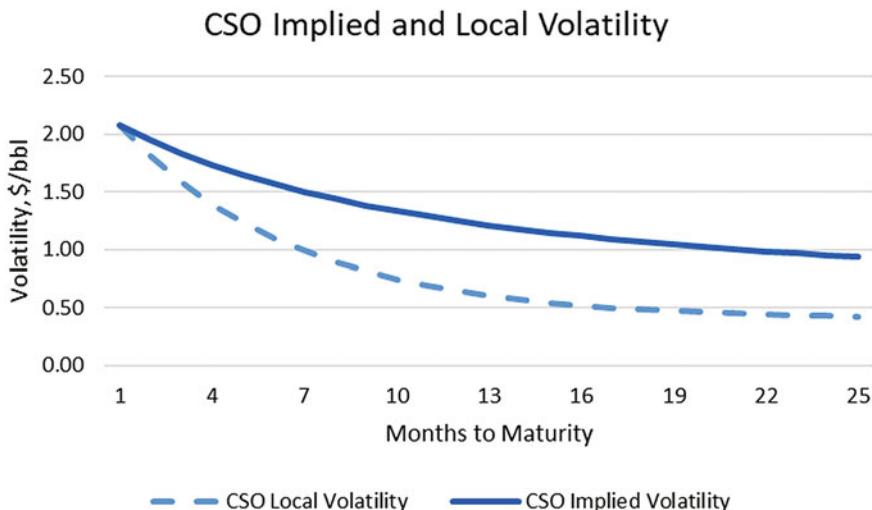


**Fig. 13.1** Storage rents and CSO market prices with their estimated fair value

as storage economics change. The owner must be able to monetize a real option to offset the liability that arises from selling a financial option. The opportunity to sell the storage option in the financial market at a higher price usually provides sufficient incentives to the storage operator to develop such basic trading capabilities. In addition, collecting an option premium upfront from selling long-dated CSOs allows the storage operator to use this cash in lieu of a bank loan and deploy it towards the construction of additional storage facilities. The US storage capacity has been steadily growing over the years to accommodate the corresponding growth of shale production. A portion of this storage capacity has been implicitly financed by the option premium that operators collected from selling CSOs in the derivatives market.

As we have already seen in Chap. 9, oil options with longer maturities tend to be somewhat underpriced from the volatility perspective. For vanilla options, the excess supply of long-dated volatility comes from leveraged producer hedgers that attempt to monetize the real asset optionality in the derivatives market. For long-dated spread options, such a discount to the fair value is provided by storage operators, and its magnitude happens to be even more significant. Taking the other side of this imbalance by buying a discounted long-dated CSOs and delta hedging them to protect the option value has become an attractive investment strategy. We call this strategy a *synthetic storage warehouse*.

Figure 13.1 helps to visualize the rationale behind this trading strategy. It shows a typical term structure of rental premia for physical storage relative to market prices of equivalent CSO puts. It also shows the realized fair value for the CSO, which is calculated here using the realized volatility during the lifespan of such options. The fair value is what a CSO owner might expect to generate by delta hedging a long option position during the lifespan of the option. This example is stylized for



**Fig. 13.2** Implied and corresponding normal local volatility for a CSO computed using the Bachelier formula

pedagogical convenience, where, for simplicity, we assume that all options are ATM. The nature of the trading opportunity remains the same for OTM options, where the fixed strike represents the cost of physical storage.

One can see that rental fees vary only slightly with the duration of the rental commitment. The term structure of fees even declines for long-dated contracts. Like in the real estate market, the highest premium is paid for a short-term lease. Long-dated leases may be cheaper as the operator is willing to provide incentives by discounting the price in return for stable longer-term cash flows. Short-term physical deals are also favored by trading houses, looking to capture additional value from blending different grades of oil. However, it would be difficult for the blender to justify long-term commitments, as imbalances among various oil grades that make the business of blending profitable often disappear quickly.

The market price for a long-dated financial CSO is also typically pressured down by discounted forward physical storage rates. This makes the CSO an attractive buy versus its fair value, which is the expected cost of its dynamic replication by trading futures spreads. In some sense, the storage operator and the CSO buyer split the expected profit. The operator gains by selling a financial option at a price higher than the one offered in the physical market. The volatility trader, in turn, extracts profits by buying and delta hedging the discounted optionality in the derivatives market.

Figure 13.2 shows term structures of implied and local volatility that correspond to the CSO prices in Fig. 13.1. For simplicity, the spread volatility is assumed to be time-homogeneous, depending only on time remaining to maturity. Implied volatility is computed using the Bachelier formula, which is the market standard model for pricing CSOs, and the units of implied normal volatility are dollars per barrel. In contrast to the Black model, the normal assumption for spread options allows

spreads to be positive and negative as the market flips between backwardation and contango. We will discuss pros and cons of different pricing models for spread options later in this chapter.

The local volatility curve for time spreads is typically very steep. As a result, the spread volatility rolls up the curve so fast that the gain from vega negates a large portion of the option theta decay. The steepness of the curve is driven by a combination of the much higher fundamental risks embedded in short-term options and discounted prices offered by storage operators for longer-term deals. Occasionally, the discount for forward CSOs becomes so large that it even violates the no-arbitrage restriction (12.2) on how fast implied spread volatility term structure can decline.

The long-dated CSO that the derivatives trader acquires does not stay cheap forever. In contrast to a physical storage facility, the lifespan of the virtual storage is shorter. When the CSO expires, its time value must decay to zero. It means that at some point prior to its expiration, the CSO is likely to become expensive relative to its fair value that the buyer expects to generate by delta hedging. Therefore, one challenge for a synthetic storage strategy is to liquidate an option before the escalation of theta turns into a real burden. But why would anyone in the market at that time buy such an overpriced and rapidly decaying financial asset?

Recall from Chap. 9 our study of options as insurance contracts that carry a volatility risk premium paid by hedgers. The same argument applies to CSOs. Here, the primary demand for hedging comes not from end-users, but rather from financial investors. We have already seen how punitive the negative roll yield is for long-only investors holding futures in contango markets. By paying this roll yield, financial investors effectively outsource the function of storage to carry traders, but the magnitude of the cost to roll is subject to the vagaries of the market. Buying a CSO provides financial investors with an interesting alternative. Instead of taking the risk of an excessive roll cost in the event of a super-contango, investors can hedge such a cost in the financial market. They can pay the premium to buy CSO puts and eliminate potentially the largest drag on the performance of their long futures investment. Such an incremental demand for short-term CSO puts from hedgers of the roll yield turns out to be substantial, especially during times when the market is in a state of contango.

In addition, professional physical trading houses also like buying short-term CSOs to place highly leveraged speculative bets on time spreads. These options are used as effective wagers on the squeeze probability that the market may either run out of oil or run out of capacity to store it. CSOs are also routinely used for risk management by trading desks to protect large speculative positions taken in the futures market. Physical speculators trade time spreads based on fundamental models that rarely look beyond just a few months ahead. Their demand for spread volatility tends to be higher when futures approach expiration. This hedging imbalance allows CSO holders to exit the ownership of virtual storage just in time before its theta decay starts to accelerate.

To summarize, the CSO strategy described here is akin to a three-step virtual storage warehouse. First, one acquires a long-dated strip of monthly CSOs from

storage operators at a discounted “wholesale” price. The time value of financial options is then protected by delta hedging. Given the steepness of the volatility term structure and only a minimal theta to cover, the bar for delta hedging gains is set rather low. Finally, the trader parses a long-dated strip into individual month CSOs and resells them to financial investors looking to hedge against the negative roll yield, or to short-term physical speculators. An embedded volatility risk premium allows these short-term sales to be made at higher “retail” prices.

This strategy may sound like a remarkably simple way to make money. One challenge is, of course, in knowing how much to pay for such a long-dated optionality at the initiation of the trade. Another one is the calculation of delta, which heavily depends on the pricing model to be covered later in this chapter. But before we get into details of spread option pricing, an important arbitrage boundary that plays the central role in many spread option strategies must be introduced. This boundary is somewhat easier to visualize first in the context of cross-product and locational spread options.

---

## 13.2 Triangular Correlation Arbitrage

The idea behind the virtual storage strategy can be applied to cross-asset spread options, many of which also represent financial replicas of real options. We have already highlighted on several occasions that an oil refinery represents a call option on the spread between a refined product and crude oil, called the *crack spread*. Similarly, a pipeline asset or an oil tanker is an option on the spread between prices of the same commodity but in two geographical locations. A natural question would be whether the virtual asset strategy that buys discounted financial optionality from an asset owner can also be applied to a refinery or a pipeline whose owner may wish to monetize the physical optionality. This turns out to be more difficult, as in contrast to a storage facility, refineries and pipelines are simply too expensive to be financed by option premiums in the derivatives market.

The trading opportunity in crack options and locational spread options, such as options on the WTI-Brent spread, comes with a slightly different twist. While it is plausible for the volatility trader to be able to buy such a spread option at a slight discount to the fair value and effectively own a tiny refinery or a pipeline on paper, the owner of a large asset is unlikely to be sufficiently motivated to offer any substantial discounts to relatively small buyers. As a result, the implied volatility curve for a cross-product spread option is usually much flatter, and the volatility roll up does not cover as much of the theta decay. Thus, a cross-product spread option represents a much faster decaying financial asset. However, such a spread option is a vital component of a more advanced strategy, known as *triangular correlation arbitrage*.

In the triangular correlation strategy one attempts to isolate correlation exposure between two assets while neutralizing overall exposure to volatility by constructing a mini-portfolio of three options. This is accomplished by combining an option on the spread with the spread between two vanilla options on each leg of the spread. The

spread between two options is also known as a *synthetic spread option*, where one sells a relatively expensive option on one commodity and hedges it by buying a cheaper option on a correlated commodity. For example, the trader can sell a call option on diesel at a premium relative to a similar call option on crude oil. The higher price of a diesel call may reflect not only the higher realized dollar volatility of diesel, but also some imbalances in the hedging market. Diesel calls are often relatively expensive as they are used by airlines to substitute for jet fuel hedging, while crude oil calls are well supplied by oil producers via two- and three-way collars. The term synthetic spread option is somewhat misleading as it represents the spread of options, and not the option on the spread.

In the synthetic spread option trade of selling diesel calls versus buying crude oil calls, the dealer collects net premium in return for taking diesel idiosyncratic risks that cannot be mitigated by crude oil options. If diesel and crude oil futures happen to move up and down in parallel by a similar amount in dollars per barrel, then gamma risks in both markets largely offset each other. In this case, the dealer can retain most of the premium collected upfront by delta hedging both options. However, if the diesel happens to be more volatile than crude oil, perhaps driven by an unforeseen interruption of refining processes, then the dealer may end up suffering significant losses. Such refinery outages could be caused by hurricanes or unexpectedly cold weather which explains more pronounced seasonal patterns in the volatility of refined products.

One way for the seller of a synthetic spread option to protect the relative value trade against unexpected moves in the spread is to buy an option on the spread itself, perhaps from a refinery that naturally owns it. Then the strategy involves selling a relatively expensive call option on diesel to an airline and buying two relatively cheap call options, one option on crude oil from a producer, and another option on the spread from a refinery. If the trader somehow manages to buy two cheaper options at the price of the expensive one, then it will set an important triangular arbitrage boundary. While this arbitrage boundary is unlikely to be breached in practice as it is difficult to buy two options for the price of one, it still plays a pivotal role in trading correlation and spread options.

To construct the triangular boundary more precisely, we express the variance of the spread between two random variables  $X$  and  $Y$  in terms of the variances of each leg, net of the covariance between them:

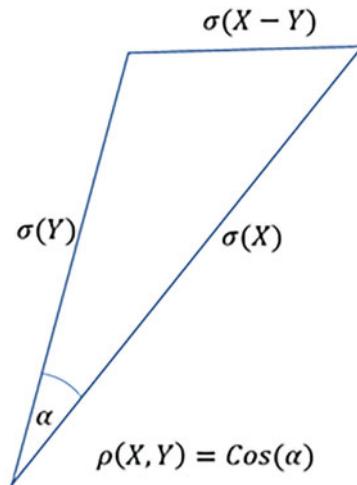
$$\sigma^2(X - Y) = \sigma^2(X) + \sigma^2(Y) - 2\rho\sigma(X)\sigma(Y) \quad (13.3)$$

One can also rewrite this formula as follows

$$\sigma^2(X - Y) = (\sigma(X) - \sigma(Y))^2 + 2(1 - \rho)\sigma(X)\sigma(Y)$$

It is clear from this decomposition that the volatility of the spread is made up of two components, the spread between two volatilities, and the contribution resulting from non-perfect correlation. The volatility of the spread increases either when one

**Fig. 13.3** Triangular relationship between volatilities and correlation



commodity becomes more volatile than the other, or when the correlation between the two assets declines.

Since the correlation coefficient  $\rho \leq 1$ , it follows that the volatility of the spread cannot be lower than the spread of volatilities

$$\sigma(X - Y) \geq |\sigma(X) - \sigma(Y)| \quad (13.4)$$

Only in the very special case of two perfectly correlated assets, this inequality becomes an identity.

Figure 13.3 illustrates this relationship between volatilities and correlation geometrically. It is presented in the form of a triangle whose sides are drawn in proportion to volatilities  $\sigma(X)$ ,  $\sigma(Y)$ , and  $\sigma(X - Y)$  of three options. The Eq. (13.3) is effectively the law of cosines if we think of the correlation coefficient  $\rho$  as the cosine of the angle formed by  $\sigma(X)$  and  $\sigma(Y)$ .

The size of the angle  $\alpha$  affects the length of the opposite side, which represents the volatility of the spread  $\sigma(X - Y)$ . If the two assets are perfectly correlated, then the triangle collapses into a straight line as the correlation, or the cosine of a zero angle, is equal to one. In this case of perfect correlation, the volatility of the spread is determined entirely by the spread between the two volatilities. If we keep the volatilities of each leg unchanged, but instead widen the angle, which means reducing correlation, then the length of the opposite side, which corresponds to the volatility of the spread, increases. The lower the correlation, the more volatile the spread, and, therefore, the more expensive the option on the spread is.

This triangular relationship sets up an important boundary not only for volatilities, but for option prices as well. It follows, for example, from the Bachelier formula that prices of ATM options with the same maturity are directly proportional to their dollar volatilities, and, therefore, the same inequality must also hold for ATM call prices:

$$C_{SP,ATM}(X - Y) \geq |C_{ATM}(X) - C_{ATM}(Y)| \quad (13.5)$$

In other words, an option on the spread  $C_{SP}$  cannot be worth less than the spread of two options with the same moneyness and maturity. If market prices for three options violate this inequality, then it is an indication of a potential arbitrage opportunity.

Let us illustrate this with a simple numerical example. Let  $X(t) = 60$  and  $Y(t) = 55$  represent diesel and crude oil prices in dollars per barrel (\$/bbl), as observed today at time  $t$ . Assume that their implied ATM Black volatilities for options with one-year maturity are, respectively, 30% and 29%. Then call options prices, obtained from the standard Black formula, are \$7.18/bbl for diesel and \$6.36/bbl for crude oil. By selling the diesel option and buying the crude oil option, the dealer collects \$0.82/bbl in exchange for taking unforeseen diesel-specific risks. It is quite possible that the net hedging cost could exceed the premium collected if diesel prices happen to be highly volatile. To mitigate this risk, a prudent dealer would attempt to buy the same maturity ATM option on the crack spread from a refinery.

If the dealer does buy the spread option, then the overall portfolio consists of three options, one short option and two long:

$$\pi = -C_{ATM}(X) + C_{ATM}(Y) + C_{SP,ATM}(X - Y) \geq 0$$

Given the relationship (13.5), such a portfolio has a non-negative payoff for any values of  $X$  and  $Y$ . In other words, despite being short one option, the overall portfolio  $\pi$  itself can be viewed as owning an option, as it cannot lose money under any scenario. If somehow the dealer can convince a refinery to sell an ATM spread option at, say, \$0.70/bbl, then the dealer could even get paid \$0.12/bbl to own such an option  $\pi$ ! Since the volatility of the crack spread during normal market conditions is often low, selling a spread option may still look appealing to a refiner, as it brings some cash and turns the naturally long asset exposure into a covered call. At the same time, the dealer not only gets paid to own an option  $\pi$ , but can even reap additional benefits in some particularly favorable price scenarios. Under no circumstances does the dealer stand to suffer any loss at expiration even if all three options are left unhedged.

To illustrate, consider several price scenarios at the expiration of these options.

**Scenario 1** Both diesel and oil markets rally, and diesel outperforms crude oil, so that  $X(T) = 70$ ,  $Y(T) = 60$ . Then a ten-dollar loss from the short diesel call is precisely offset by the sum of a five-dollar gain from owning the oil call, and another five-dollar gain from owning an ATM crack call struck at \$5/bbl:

$$\pi(T) = -10 + 5 + 5 = 0$$

**Scenario 2** Both diesel and oil markets rally, and the oil price outperforms diesel, so that  $X(T) = 70$ ,  $Y(T) = 67$ . Here, the gain from owning the oil option exceeds the ten-dollar liability on the diesel call, while the crack call expires worthless:

$$\pi(T) = -10 + 12 + 0 = 2$$

**Scenario 3** Diesel rallies, but oil falls as a result, perhaps, of an unexpected weather event that temporarily shuts the refinery, but simultaneously reduces demand for crude oil. Let  $X(T) = 70$ ,  $Y(T) = 50$ . In this case, the oil call expires out-of-the-money, but the loss on the diesel call is more than offset by the gain on the ATM crack call which is \$15/bbl in-the-money:

$$\pi(T) = -10 + 0 + 15 = 5$$

**Scenario 4** The diesel price remains unchanged as inventories are abundant, but oil rallies due to an unexpected OPEC production cut, so that  $X(T) = 60$ ,  $Y(T) = 60$ . Then the only option among the three that has a non-zero value at expiration is the oil call:

$$\pi(T) = 0 + 5 + 0 = 5$$

Only in the first scenario, which is, admittedly, more likely than the other three, the dealer's payoff at the options expiration is precisely equal to zero. This scenario corresponds to the case of equal dollar variance and perfect correlation between diesel and crude oil. In the other three scenarios, which are less likely, the dealer receives additional benefits from options payoffs. In the second scenario, the positive payoff comes from a favorable move in relative realized volatilities when crude oil moves more than diesel. In the last two scenarios, the profit is driven by decorrelation of the two commodities.

This portfolio provides an example of an arbitrage boundary which is independent of any models or assumptions. If one can execute a triangular trade in the market at net zero cost, it would genuinely represent a free option. The only thing to worry about would be the creditworthiness of the option sellers. The dealer can only profit from an OTC trade if neither the refiner, nor the producer, default on their short option obligations. A popular joke, or an interview question for a trading job, is “how much to pay for an OTC option to a counterparty that has an option not to pay you back? The answer is, of course, zero. Some energy volatility dealers, however, learned painful lessons by buying “cheap” options from asset owners with poor credit, who ended up defaulting when their derivatives obligations became particularly large. In one case, the dealer even ended up owning an entire power plant which was placed as collateral against the financial derivative sold by an asset owner. Unfortunately for the dealer, the value of the plant covered only a fraction of the payoff due on the derivatives trade.

A perfect triangular arbitrage is unlikely to be found in competitive markets. However, the pricing boundary created by options on two correlated assets and an option on the spread forms the backbone of spread option and correlation trading. We now discuss how this relationship between the volatility of the spread and the correlation is captured in two alternative modeling paradigms.

### 13.3 The Dichotomy of Spread Option Pricing

Modeling spread options using conventional methodologies developed for financial markets could quickly turn into a quagmire. Petroleum spread options are structurally different from spread options that trade in interest rate and equity markets. A financial spread is typically a derived variable, computed as the difference between the prices of two assets. In contrast, in many energy markets, the spread itself is a primary traded variable that connects a less liquid asset to an industry benchmark. The value of an illiquid asset is then synthetically constructed by adding the traded value of the spread to the price of the liquid leg. For example, nearly all US oil grades trade primarily as the basis to WTI with their prices calculated as the sum of the WTI price and the basis spread.

Certain petroleum spreads exhibit a split personality. This happens when some market participants view the spread as an independently traded asset, while others see it as the difference between two prices. Many refined products, such as diesel or gasoline, trade in such a dual manner. Fundamental traders tend to analyze refined products through the lens of the crack spread, as their trading strategies generally avoid taking directional exposure to petroleum prices. In contrast, financial and systematic traders, who often rely on covariance matrices in managing diversified commodity portfolios, prefer to trade each leg of the spread separately. The way different participants think about the spread dynamics often leads to framing biases which are essential to understand when selecting the model for trading spread options.

Such a trading dichotomy suggests looking at the price of the spread option from two different angles. One approach is to model an option written on an independently traded spread contract that follows a certain stochastic process. An alternative is to specify the stochastic dynamics of each leg of the spread and assume some correlation structure between the two processes. Fortunately, over time the two methodologies have somewhat converged to more practical hybrid solutions, where the volatility of the spread in the former method and the correlation input in the latter are connected to each other.

The first approach of handling the spread as an independent variable is a straightforward extension of methods presented in previous chapters. Let  $S$  be the spread between two futures contracts  $F_1$  and  $F_2$

$$S = F_1 - F_2$$

that follows a diffusion process of the form:

$$dS = \mu(s, t) + \sigma(S, t)dz$$

where  $\sigma(S, t)$  represents the local volatility of the spread.

To apply our previously developed techniques, we only need to replace the futures price  $F$  with the spread  $S$  and acknowledge that the local volatility function refers to the volatility of the spread. Since many petroleum spreads can be either

positive or negative, the local spread volatility must be measured in absolute or dollar terms, and not in percent. As it was shown in Chap. 8, the drift term  $\mu(S, t)$  does not play a direct role in the option valuation. This may sound somewhat counterintuitive, as one would expect relatively fast spread mean-reversion to have an impact on option prices. The mean-reversion does indeed play a role but indirectly, as the speed of mean-reversion affects the steepness of the volatility term structure.<sup>4</sup>

The most straightforward approach for pricing oil spread options assumes that the spread follows an ABM process with constant dollar volatility, for which the option price is given by the Bachelier formula. It is remarkable that a formula originally derived over a century ago to price options on French government bonds retains its relevance a hundred years later in the modern oil market. If spread option prices are available in the market, then they can be easily expressed in terms of implied normal spread volatilities and visualized in the form of the volatility smile for the spread. Such a smile usually has a very pronounced curvature on both sides, reflecting higher normal volatility for out-of-the-money puts and calls.

To adjust the Bachelier model for non-normality and fat tails, which are typical for the behavior of energy spreads, one can use the QN model developed in Chap. 10. The QN model works particularly well for options on the spread between assets that are highly correlated during fundamentally balanced markets, but prone to periodic sharp dislocations. Such dislocations occur when economic linkages between two assets are disrupted by unforeseen events, such as pipeline outages, weather, or geopolitical conflicts. When two assets suddenly diverge, the volatility of the spread spikes, and the value of the spread option explodes.

The parabolic nature of the QN model is very intuitive for many petroleum spreads. The vertex of the parabola that marks the lowest local volatility is often associated with the equilibrium spread level, corresponding to the economics of the physical arbitrage that connects two markets. For example, for CSOs the vertex of the parabola is often located near the spread level that reflects the cost of storage. Likewise, for WTI-Brent options, it is related to the cost of freight. The further away the spread deviates from its normal trading range in either direction, the higher the uncertainty and the resulting dollar volatility. The quadratic local volatility with its increasing curvature on both ends captures such dynamics quite well. Spread options tend to trade under a sticky local volatility smile for specific strikes, which retains its shape regardless of the option moneyness. The ability to capture the tails of the spread behavior is the major advantage of modeling the spread as an independent asset.

Occasionally, one also encounters non-standard spread options, such as spread APOs, early expiry spread options, and even spread swaptions. For example, APOs

---

<sup>4</sup>Formally, this connection follows from reduced-form mean-reverting models, the reference to which are provided in Chap. 3. Less formally and more intuitively, one can see that faster mean-reversion precludes the spread from deviating too far from its mean, which generally results in volatility that declines with time to maturity.

on the spread between gasoil and Brent are particularly popular among European and Asian refineries. Some spread APOs are listed on organized exchanges even though their primary liquidity is still concentrated OTC. All methods developed in Chap. 11 for exotic OTC options apply directly to non-standard spread options, provided that the spread itself is modeled as a stand-alone asset. One can also calibrate the local spread volatility and bootstrap it from market prices of spread options. When the spread is treated as the primary asset, the problem of pricing spread options, as mathematicians often say, reduces to the previous case that has already been solved. This is, by far, the most popular way of handling spread options in petroleum markets.

There are, however, certain spread options for which modeling the spread as a single asset has limitations. Consider a refinery that buys a put option on a spread which approximates its profit margin with a basket of refined product outputs and crude oil inputs. The weights of the basket components are highly specific to individual refineries. For example, one refinery may be producing a 50–50 basket of diesel and gasoline, while another one may have a 60–40 yield split between two of its main products. If one chooses to model such baskets as stand-alone assets, such as basket one, basket two, basket three, etc. with their own independent basket volatilities, then any direct connection between volatilities of closely related baskets will be lost. To manage the portfolio of such refinery margin options, it is better to specify the dynamics for individual components of the spread along with their correlation.

Perhaps a more pressing need to model individual components of the spread arises in other energy markets, such as natural gas and power. For example, a power plant can be modeled as a call option on the spread between the price of electricity that trades in megawatt-hours (MWh) and natural gas that trades in million British thermal units (MMBtu), where the latter is multiplied by the heat rate that measures the efficiency of the plant.<sup>5</sup> Even though this approach of modeling two assets independently is less frequently used for petroleum spread options, it does provide an interesting alternative perspective on valuation.

To keep the exposition simple, we only illustrate the correlation-based framework for two assets as extending it to a multi-asset spread basket is relatively straightforward. We assume that  $F_1$  and  $F_2$  follow two diffusion processes

$$dF_1 = \mu_1(F_1, t)dt + \sigma_1(F_1, t)dz_1$$

$$dF_2 = \mu_2(F_2, t)dt + \sigma_2(F_2, t)dz_2$$

with random components  $dz_1$  and  $dz_2$  that have correlation  $\rho$ . As before, the drift terms are eliminated by delta hedging and the spread option value depends on local volatilities of two futures and the factor correlation which, for simplicity, is assumed to be constant.

---

<sup>5</sup>For more background on spread options that arise in natural gas and power markets, see, for example, Eydeland and Wolyniec (2003) and Swindle (2014).

Consider first the simplest case where the local dollar volatilities of the two legs are constant,  $\sigma_1(F_1, t) = \sigma_{1,A}$  and  $\sigma_2(F_2, t) = \sigma_{2,A}$ . Then both futures are normally distributed, and, therefore, the spread between them is also normally distributed with the arithmetic spread volatility  $\sigma_{S,A}$ , given by

$$\sigma_{S,A} = \sqrt{\sigma_{1,A}^2 - 2\rho\sigma_{1,A}\sigma_{2,A} + \sigma_{2,A}^2}$$

The problem, therefore, again reduces to the previous one, where the spread is an asset by itself, and the option price is given by the Bachelier formula with volatility  $\sigma_{S,A}$ . The normal volatility of the spread is expressed in terms of the normal volatilities of the two legs and the correlation between the two assets.

We next take a short-cut approach similar to the one that was used for pricing APOs and swaptions. We obtain the exact formula for a normal process and then use it to construct an approximation for option prices under other diffusions. For example, consider two conventional lognormal processes, where the volatilities for each leg are proportional to the corresponding futures prices:

$$\sigma_1(F_1, t) = \sigma_{1,G}F_1$$

$$\sigma_2(F_2, t) = \sigma_{2,G}F_2$$

The spread between two lognormal variables is not lognormal. In fact, it is closer to being normal than lognormal. In this case, one can still use the Bachelier formula as an approximation with normal spread volatility given by

$$\sigma_{S,A} = \sqrt{\sigma_{1,G}^2 F_1^2 - 2\rho\sigma_{1,G}\sigma_{2,G}F_1F_2 + \sigma_{2,G}^2 F_2^2} \quad (13.6)$$

Assuming that both futures prices and their implied volatilities are observable in the market, this transformation allows the trader to calculate the volatility of the spread given some assumption about the correlation. Alternatively, if the price for the spread option is observed in the market, then the trader can back out the implied correlation as follows:

$$\rho = \frac{\sigma_{1,G}^2 F_1^2 + \sigma_{2,G}^2 F_2^2 - \sigma_{S,A}^2}{2\sigma_{1,G}\sigma_{2,G}F_1F_2} \quad (13.7)$$

where, for simplicity, we kept the same letter to denote implied and local geometric volatility.

When the spread is modeled as an independent asset, its delta and other Greeks are reported with respect to the spread itself. In other words, the delta with respect to the first leg of the spread is the negative of the delta with respect to the second leg. This makes perfect sense for closely related spreads between assets with similar volatilities, as hedging is typically done by transacting directly in the spread contract. For example, an option on the WTI-Brent spread will always be hedged by trading in

the spread itself. However, when one asset is significantly more volatile than the other, the trader will most likely be looking to hedge the risk of the more volatile asset differently by delta hedging it outright rather than as a spread to something that is moving a lot less. The problem becomes more evident when the two legs of the spread trade in different units. This is more common for cross-product spreads that arise in the natural gas and power markets where the two legs of the spread are more disjoint.

An alternative valuation method attempts to mold spread options into the standard lognormal BSM pricing framework. Its main idea is based on the observation that the ratio of lognormal variables is also lognormal. Furthermore, the variable

$$x = \frac{F_1}{F_2 + K}$$

can be viewed as approximately lognormal if the strike price  $K \ll F_2$ .

More formally, the terminal payoff of the call option on the spread (13.1) can be rewritten in terms of the variable  $x$  as follows:

$$\frac{C_{SP}(F_1, F_2, T)}{F_2 + K} = \frac{\max(0, F_1 - F_2 - K)}{F_2 + K} = \max(0, x - 1)$$

Both the spread option price  $C_{SP}$  and the underlying variable  $x$  are scaled by the same quantity  $F_2 + K$ . They are effectively expressed in terms of the new scaled numeraire. The payoff of the scaled spread option price is then the same as the payoff of the regular call option written on the variable  $x$  with the strike price equal to one.

Unfortunately, the spread option admits an analytic solution in a two-dimensional lognormal setting only in a very special case when the strike price  $K = 0$  and the variable  $x$  becomes lognormal. Such a solution is known as the Margrabe formula, which was initially referred to as an option to exchange one asset for another. For  $K \neq 0$ , there exist several convenient approximations, among which the following Kirk approximation is the most widely used<sup>6</sup>:

$$C_{SP}(F_1, F_2, t) = F_1 N(d_{1,x}) - (F_2 + K) N(d_{2,x}) \quad (13.8)$$

where

$$d_{1,x} = \frac{\ln(x)}{\sigma_{x,G}\sqrt{\tau}} + \frac{\sigma_{x,G}\sqrt{\tau}}{2}$$

---

<sup>6</sup>Margrabe (1978) derived two-dimensional partial differential equation for the price of a spread option and solved it for  $K = 0$ . Kirk (1995) proposed the extension for  $K \neq 0$  but did not publish its derivation. This formula can be formally derived by applying the method of perturbation introduced in Appendix C. The technical details, however, are rather cumbersome, and given the limited usage of correlation-based pricing formulas in the oil market, we chose not to present them.

$$d_{2,x} = \frac{\ln(x)}{\sigma_{x,G}\sqrt{\tau}} - \frac{\sigma_{x,G}\sqrt{\tau}}{2}$$

Here,  $\tau = T - t$  and the effective geometric volatility  $\sigma_{G,x}$  is given by

$$\sigma_{x,G} = \sqrt{\sigma_{1,G}^2 - 2\rho\sigma_{1,G}\bar{\sigma} + \bar{\sigma}^2}, \quad \bar{\sigma} = \sigma_{2,G} \frac{F_2}{F_2 + K}$$

The formula (13.8) resembles the Black formula on the underlying asset  $x$  for the option price scaled by  $F_2 + K$  with the strike price equal to one. If we let  $K = 0$ , then it turns into the Margrabe formula, which represents an exact solution for the two-dimensional lognormal distribution. Many other more sophisticated approaches to the correlation-based pricing of spread options have been proposed, but the accuracy of the approximation (13.8) is usually sufficient for all practical applications in petroleum markets.<sup>7</sup>

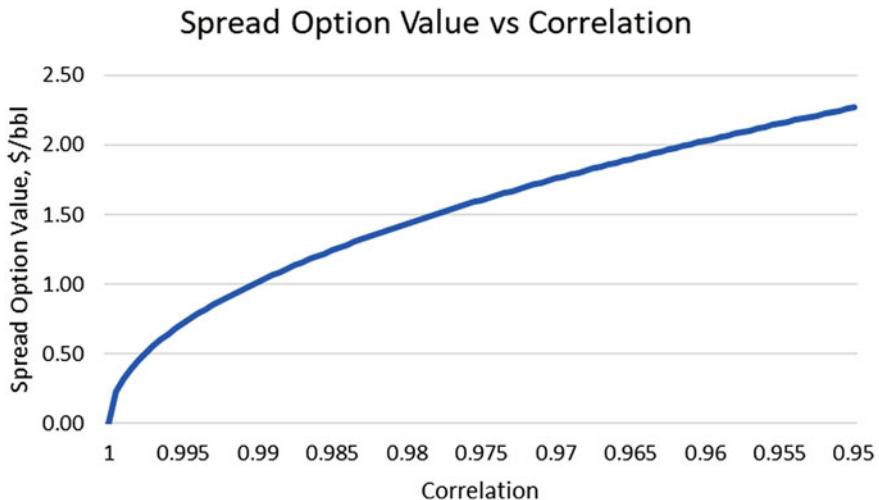
Despite the popularity of the two-dimensional lognormal framework in academic studies, it turns out to be of a limited use for practitioners in the oil market. Using this framework, it is more difficult to incorporate the tails of the distribution, which are critical for pricing energy spread options. In theory, one can extend the idea of the volatility smile for the spread and use (13.7) to construct the implied correlation smile for spread options with different moneyness. In fact, such a correlation smile would appear more like a frown, as lower correlation is needed to generate higher prices for OTM spread options.

However, the process of constructing implied correlations for different moneyness is somewhat ambiguous. For a given strike  $K$  of the spread option, there are infinitely many pairs of strikes for individual assets  $K_1$  and  $K_2$  that match the spread strike defined by  $K = K_1 - K_2$ . Using different pairs of implied volatilities,  $\sigma(K_1)$  and  $\sigma(K_2)$ , inevitably produces different implied correlations. Therefore, the concept of correlation frown is not uniquely defined. Additional constraints must be imposed on the selection of  $K_1$  and  $K_2$ , which further complicates the calibration process.

Having two methodologies for pricing spread options in place, we finish this chapter by highlighting some important practical challenges with empirical estimation of inputs to these models and provide some guidance on approaches favored by professional traders.

---

<sup>7</sup>Similar to (A.10), the solution to the partial differential equation for the price of a spread option can be represented as a double integral of the option payoff multiplied by the risk neutral joint probability density for  $F_1$  and  $F_2$ . Advanced methods for pricing spread options in two-dimensional lognormal setting tend to focus on developing efficient integration techniques and more advanced analytic approximations. See, for example, Shimko (1994), Dempster and Hong (2002), Carmona and Durrelman (2003), and Venkatramanan and Alexander (2011). For petroleum markets, a simple approximation (13.8) is generally sufficient.



**Fig. 13.4** The price sensitivity of a one-year ATM spread option to correlation for  $F_1 = F_2 = 60$ , and  $\sigma_{1,G} = \sigma_{2,G} = 0.30$

### 13.4 Dealing with Unobservables

While correlation is a commonly used metric for describing co-movements between financial assets, applying it to pricing oil spread options is quite dangerous. Many petroleum futures are highly correlated as they are linked to each other by the economics of storage, transportation, and processing. Unfortunately, spread options are extremely sensitive to the correlation input, especially when correlation is particularly high.

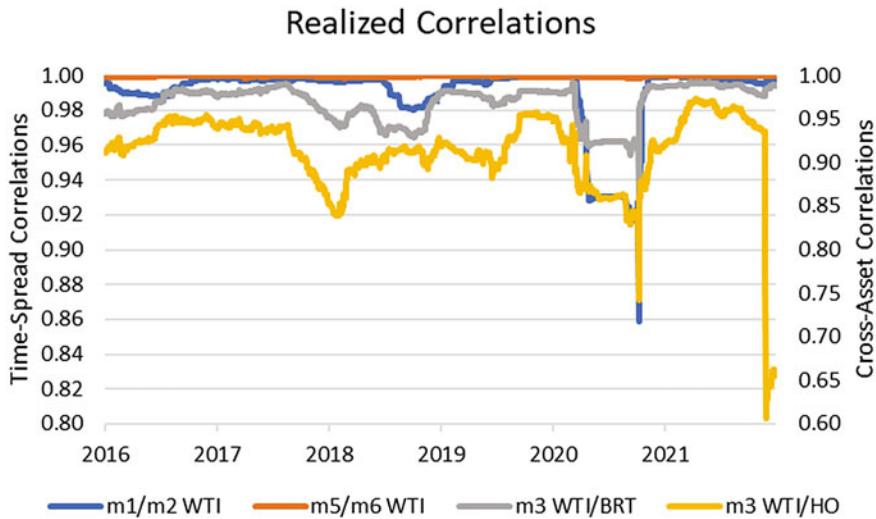
Let us illustrate the challenge with a simplified example of two futures contracts that have the same price  $F_1 = F_2 = F$  and the same percentage volatilities  $\sigma_{1,G} = \sigma_{2,G} = \sigma$ . In this case, the volatility of the spread between the two futures is related to the correlation between them through (13.6):

$$\sigma_{S,A} = \sigma F \sqrt{2(1 - \rho)}$$

The price of an ATM spread option is then given by the Bachelier formula (8.9), where, as usual, we ignore the impact of interest rates:

$$C_{SP,ATM} = \frac{\sigma_{S,A} \sqrt{T-t}}{\sqrt{2\pi}} = \frac{\sigma F \sqrt{T-t}}{\sqrt{\pi}} \sqrt{1-\rho}$$

Figure 13.4 shows how the value of an ATM spread option changes for a wide range of correlations.



**Fig. 13.5** Realized correlations for many petroleum spreads are extremely high and unstable (6-month rolling correlations)

Higher correlation means lower spread option value. It is easy to see that the relationship becomes highly nonlinear for extremely high levels of correlation. In fact, since

$$\frac{\partial C_{SP,ATM}}{\partial \rho} = -\frac{\sigma F}{2} \sqrt{\frac{T-t}{\pi(1-\rho)}}$$

the sensitivity of the spread volatility to correlation becomes infinite when correlation approaches one.

Now let us look at the magnitude of the correlation coefficient for key petroleum spreads that underlie traded spread options. Correlations are especially high for time spreads between futures on the same commodity but with different maturities. The average correlation between the first two nearby WTI futures is approximately 99%. Moreover, the correlation between longer-term futures on the same commodity is basically indistinguishable from one. Imagine now being asked to come up with a bid for a long-term storage asset using the CSO pricing framework, which requires empirically calculated correlation input. Using, for example, 0.996 correlation instead of 0.999 would double the price of a benchmark one-month CSO, or, equivalently, double the value of the physical storage asset. One could easily be relieved from duties as a desk quant for providing such an unacceptable degree of precision. It is unlikely that anyone would be comfortable purchasing an asset based on such an unreliable valuation model.

Figure 13.5 shows other examples of realized correlations for key petroleum spreads. While all correlations generally remain high, one can also observe periodic bouts of rapid decorrelation. These periods of decorrelation are driven by

unexpected short-term disruptions in economic linkages between components of the spread.<sup>8</sup>

Such correlation behavior is typical for petroleum markets, which is also reflected in a parabolic function for the spread volatility versus the spread itself. When the oil market is fundamentally balanced, most futures tend to move together, with spreads exhibiting low volatility. However, when disruptions occur unexpectedly, futures decorrelate and the volatility of spreads rapidly increases. These observations lead us to the following rule of thumb: for spread options on highly correlated assets one should always model the volatility of the spread itself instead of relying on unstable correlation-based models for its two components.

In the case of a CSO, correlation-based pricing models face yet another challenge. Not only is the correlation input hard to estimate, the volatility of the second leg of the time spread is also not observable. Recall from the discussion of EEOs in Chap. 11 that what matters for pricing an option is the local volatility during the lifespan of the option, which is not the same as the implied volatility of a vanilla option. Consider the primary benchmark CSO on the prompt one-month futures spread. The volatility of the front month futures can be considered to be observable.<sup>9</sup> However, for pricing the CSO using a correlation-based model, one also needs to know the local volatility of the second month futures measured only during the life of the CSO, which expires a month earlier. The local volatility of the second leg is not observable, and it can be drastically lower than the implied Black volatility of the second-month futures.

To illustrate, we use as an example a one-month ATM CSO on the spread between the first two futures. Let the implied Black volatilities for these futures be  $v_1 = 0.35$  and  $v_2 = 0.32$ , and the correlation between the two contracts be empirically estimated at 0.995. If we assume again that local volatility is time homogeneous, depending only on the time remaining to maturity, then following the bootstrapping methodology of Chap. 12, the local volatility of the second leg during the first month when the CSO exists is then given by

$$\sigma_2 = \sqrt{2(0.32)^2 - (0.35)^2} = 0.287$$

Using the Bachelier formula, the corresponding price of an ATM CSO call is \$0.49/bbl. If the trader mistakenly uses implied volatility  $v_2$  for the second leg of the spread instead of its local volatility  $\sigma_2$ , then the CSO price drops by more than one-third to \$0.31/bbl. By selling an option at such a lower price, the trader may be making a grave error. This mistake by the seller of the spread option could even violate the boundary set by the triangular arbitrage. An arbitrage would occur if the

---

<sup>8</sup>The day when oil prices went negative is removed as neither percentage returns nor standard correlation metrics can even be computed.

<sup>9</sup>Technically, the volatility of the first leg is also not observable as a vanilla option expires three days prior to futures but the corresponding CSO expires one day before futures. Traders often make an additional ad hoc adjustment to the implied volatility of a vanilla option to incorporate this effect.

CSO and another EEO on the second leg of the spread can be purchased at a total price that does not exceed the price of a vanilla option on the first leg. Surprisingly, such an arbitrage trade has indeed been executed in the oil market on several occasions, rewarding the dealer with a free option.

To summarize, the extreme sensitivity of a spread option price to the correlation input and the need for an additional estimation of the unobservable local volatility of the second leg of the spread makes correlation-based methods practically unusable for pricing many petroleum spread options. As a consequence, the Bachelier model became a de facto industry standard for pricing CSOs and other spread options on highly correlated futures.

The initially obscure market for oil CSOs has gained transparency over the years. At the time of writing this book, market prices for one-month ATM CSOs are regularly quoted by option brokers. One can, therefore, easily calculate CSO implied normal volatility. However, this information is insufficient for backing out implied correlation, as the local volatility of the second leg is still unobservable. There are many combinations of implied correlations and the second local volatility that result in the same volatility of the spread. This is similar to the non-uniqueness problem in model calibration that we have encountered in the previous chapter for the local volatility matrix driven by a multi-factor model. It is also impossible to construct an implied correlation frown for CSOs in a unique way, because it requires the knowledge of the volatility smile for all EEOs.

If the market for CSOs is deemed to be inefficient because of hedging imbalances, then the ability to calculate implied CSO volatility does not necessarily answer the question whether an option is cheap or expensive. Some traders attempt to answer this challenge by using estimates of the past realized volatility to approximate the future implied volatility for the spread. However, for spread options this approach must be taken with the greater degree of caution. The local spread volatility, such as the one shown in Fig. 13.2, is very steep. Therefore, the primary contributor to the total variance of the spread and to the value of a CSO is volatility that is expected to occur in the future near the contract expiration, and not the prior historical volatility of the same spread, which is generally much lower.

For example, if we are pricing a CSO on a January–February spread with, say, six months to maturity, then the recently observed volatility of this spread tells us almost nothing about what volatility to expect in December when the CSO is about to expire. Instead, more meaningful information can be obtained by looking at the realized volatility of previous spreads over the same six-month lifespan prior to their corresponding expirations. Each previous expiration produces a single data point for the realized volatility with given time to expiration. One can then average monthly data to come up with some generic historical term structure of spread volatility that can be compared to the implied volatility of the spread.

This simple approach can only be used as a starting point for modeling volatility term structure for time spreads. The proper pricing of spread options must always be tied to fundamentals, as spread volatility typically depends on the state of local inventories in nearby storage facilities. Many large participants in the CSO market are professional physical traders who use sophisticated technologies to gain access to

up-to-date inventory information and pipeline flows. They tend to price spread options based on anticipated trajectory in forward inventories, specifically based on the likelihood of inventories reaching their boundaries, when the spread volatility is expected to be particularly high. This is akin to modeling the spread dynamics based on the probability of a squeeze, as it was presented in Chap. 3.

One way to link historical volatility estimates to current market conditions is to utilize the concept of analogue periods. Instead of using a simple average of volatility observations over a selected lookback period, the trader can average only observations that are perceived to be relevant to the prevalent fundamental regime. Obviously, establishing what is relevant and what is not is much easier said than done. Some degree of relevance can be quantified using inventory data as a measure of fundamentally similar periods. For example, one could choose analogues by identifying historical periods when inventories were sufficiently close to the current level and average realized volatility only during such periods. One can even assign different weights to each historical volatility data point based on the distance between inventories in the analogue period and their present level.

Numerous other schemes can be developed to weigh historical volatility realizations that can incorporate seasonality or the slope of inventory trends that indicate whether inventories are building or drawing. We should always remember though that for the most part, the value of a spread option is driven by the probability of a squeeze, or the likelihood of an unexpected event. This probability is extremely difficult to extract solely from historical price data. Identifying and comparing the behavior during fundamental analogue periods is probably the best that one can do. Even though history rarely repeats itself, in the case of spread volatilities and correlations it often rhymes with fundamental environments that led to prior squeezes. In the spread options market, quantitative volatility models and fundamental models of supply and demand are deeply intertwined.

These comments provide only a rough guide to estimation of volatility from historical and fundamental data. Volatility forecasting can also involve many other sources of information, such as speculative positioning, open interest, and various measures of liquidity. Any attempt to provide a more rigorous prescription for calculating the fair value of a spread option would have been a disservice to the reader, as volatilities and correlations are impacted by too many constantly evolving factors. The problem of spread option pricing is representative of many other problems that arise in quantitative oil trading. In fact, none of the strategies discussed in this book should be read prescriptively. If one ever claims to have a precise step-by-side guide to making money in the oil market, then my advice would be to stay away from such a guide, as promised money is likely to be an illusion.

Throughout the entire book, we emphasized that quantitative oil trading is a blend of art and science, a blend of a human and a machine. The objective of the book is to provide readers with some science that has been helping the author to paint the picture of the oil market throughout his twenty-five career as a trader. The actual trading strategy is always an art. In this book, I'm happy to share some of mine, both the pieces that worked and the ones that failed, and I'm looking one day to learn about yours.

## References

- Carmona, R., & Durrleman, V. (2003). Pricing and hedging spread options. *SIAM Review*, 45(4), 627–685.
- Considine, J., Galkin, P., & Aldayel, A. (2022). Inventories and the term structure of oil prices: A complex relationship. *Resources Policy*, 77, 1–18.
- Dempster, M. A. H., & Hong, S. S. G. (2002). Spread option valuation and the fast Fourier transform. In H. Geman, D. Madan, S. R. Pliska, & T. Vorst (Eds.), *Mathematical finance, Bachelier congress* (Vol. 1, pp. 203–220). Springer.
- Eydeland, A., & Wolyniec, K. (2003). *Energy and power risk management: New developments in modeling, pricing, and hedging*. Wiley.
- Johnson, O. (2022). *40 classic crude oil trades: Real-life examples of innovative trading*. Routledge.
- Kirk, E. (1995). Correlation in the energy markets. In *Managing energy price risk* (pp. 71–78). Risk Publications.
- Margrabe, W. (1978). The value of an option to exchange one asset for another. *The Journal of Finance*, 33(1), 177–186.
- Shimko, D. C. (1994). Options on futures spreads: Hedging, speculation, and valuation. *The Journal of Futures Markets*, 14(2), 183–213.
- Swindle, G. (2014). *Valuation and risk management in energy markets*. Cambridge University Press.
- Venkatramanan, A., & Alexander, C. (2011). Closed form approximations for spread options. *Applied Mathematical Finance*, 18(5), 447–472.



---

## 14.1 The Roadmap for Energy Transition and Virtual Commodities

As I finish writing this book, the world is actively engaged in the debate about the energy transition. Yet very little attention is paid to the transition of virtual energy in the market for energy derivatives, which, by and large, sets the price for the energy that we all consume. I will conclude the book by outlining the role that virtual barrels could play in the energy transition and its impact on broader markets for virtual commodities.

Regardless of the source of energy, the market will continue to function as a complex dynamic system with multiple feedback loops and bidirectional causality between its components. The entire energy-trading ecosystem will still revolve around real options embedded in physical assets, even if the nature of these assets and their ownership takes a different form. These real options are the foundation for a self-balancing mechanism that allows the entire system to adjust and prevent prices from breaking through the system boundaries.

The nucleus of the energy ecosystem is an option on time, which is provided by storage technology. As we have highlighted before, storage buys time by shifting limited supplies from times of plenty to times of relative scarcity. Without this technology, an energy market is unlikely to exist. For the pioneers of oil trading of the 1860s the technology used for storage was literally a dump. In the new world of the energy transition, storage will take the form of a battery.

Batteries are expensive to build, and they are likely to be financed and leased to professional traders who are in a better position to monetize an embedded optionality. To pay for the lease, battery traders will generate revenues by buying and storing power when the price is low and discharging it when the price is high. In the jargon of virtual barrels this is the strategy of delta hedging the real option. Thus, many core concepts presented in the book, including the theory of storage, the model of the squeeze, and the virtual storage warehouse strategy with time spread options will find new applications in the business of battery storage with only minimal

modifications. The main change is the speed of the market. While the owner of oil storage can afford to trade rather infrequently, the value of the battery option is driven by intraday volatility, and one must participate in the market continuously throughout the day. Trading will inevitably become more data intensive, creating an opportunity that the new generation of quants would most certainly appreciate.

In the energy market, an option on time often goes side by side with an option on location that allows traders to deliver energy from one point to another. New locational spread options are already being created in response to the needs of the physical energy transition. The biggest beneficiary so far is the virtual market for natural gas, which is on its way to become increasingly global, interconnected, and more financialized. An important role in this transition is played by liquefied natural gas (LNG) assets whose owners monetize the transportation option by linking US natural gas with consumers in Europe and Asia. Furthermore, the switching option held by power generators would likely strengthen the link between the prices of power, natural gas, coal, and emission credits. It will further expand the scope of quantitative relative value trading, providing additional diversification to energy stat-arb portfolios.

The growth potential for the energy derivatives market remains highly significant. While this market is already larger than the market for physical commodities, it is still tiny relative to the size of other financial markets. The energy market represents no more than 0.1% of the size of the global equity market, even though the energy share in the world economy is undoubtedly more significant. If every financial investor decides to allocate, for example, 5% of total assets to energy-related strategies, then hypothetically, the market for energy derivatives can grow nearly 50-fold. The larger the energy market, the higher the need for a common benchmark. Even today, WTI and Brent futures behave much more as financial benchmarks, setting the tone for pricing physical barrels.

Does this mean that energy prices will disconnect more often from the underlying physical market? This might be an incorrect question to pose. From the technical pricing perspective, the two markets are unlikely to diverge too far, as the price in the physical market is largely determined by futures. A better question to ask is whether the price is reflective of the prevalent state of physical supply and demand. While the answer is debatable, fundamentals alone will not be able to explain the price. The benchmark futures price will always reflect the combined effect of supply and demand for physical and financial barrels, and more often than not, the latter will dominate the former.

As the energy transition picks up its pace, the strategies of large sovereign producers will change as well. The risk of ending up with stranded oil assets is already forcing OPEC to change its long-term goals. Its main priority has shifted from providing some market stability to maximizing the value of its resources before oil demand starts to wane. OPEC will attempt to exert some control with the goal of setting price floors by collectively cutting production when the price is perceived to be low. This strategy is inherently more volatile, requiring dynamic decision making with higher frequency of OPEC meetings and less predictable outcomes. The perception of such a price floor may change the composition of market participants

and their trading styles. It may trim investor interest in using downside momentum strategies, which may lead to an excess of buyers over sellers in the futures market. This financial imbalance can be further exacerbated by declining interest in hedging by independent producers, many of which remain under investor pressure to divest fossil assets.

The inability to tame energy prices driven by financial markets will undoubtedly frustrate policymakers of consumer nations, and one should expect more frequent market interventions by their governments. There will be more attempts to use strategic reserves, price caps, subsidies, sanctions, tariffs, and other policy tools specifically targeting energy prices. However, the track record of states meddling with energy prices is very poor. Replacing a part of a dynamic ecosystem with an artificial graft is unlikely to make the body of the market any stronger. The system would lose its natural immunity and become more susceptible to a complete collapse in the event of unexpected exogenous shocks. Unfortunately, plenty of such episodes have already been recorded. The way to keep the ecosystem under control is to let it evolve organically under the power of market forces and whenever possible adjust its permissible boundaries. For example, excessive speculation can be better managed with tighter position limits and more transparency around large positions, which will dissuade traders from taking excessive risks.

Any transitional period will inevitably bring more uncertainty and volatility, as changing composition of market participants can distort the supply and demand both for physical energy and for hedging services. Since the entire derivatives ecosystem operates on a highly leveraged basis, increasing volatility raises the cost of trading via more stringent collateral requirements. This could force certain traders out of the market, which, in turn, would adversely impact the market liquidity, and lower liquidity leads to even higher volatility. Thus, volatility begets more volatility.

The need to manage volatility brings us to the final piece in the virtual energy transition: the growth and increasing sophistication of the options market. The presence of real options, either developed by new asset owners or implicitly set by regulators, will leave more footprints in the market. Such real options will certainly be replicated by quantitative traders in the derivatives market. The primary motivation of these traders is to capture an arbitrage between the actuarial value of optionality as it is perceived by end-users and policymakers, and the volatility-based value determined by the cost of dynamic replication.

The dynamics of energy options could start to resemble equity options, where the demand for hedging services exceeds the supply, and buyers are willing to pay up for insurance-like protection. In a less-fossilized energy world, the supply of volatility provided by producer hedgers could decline. As a result, the structural volatility risk premium will likely return to the energy market. The other pillars of quantitative option modeling, such as negative gamma, the volatility smile, its term structure, and the parabolic nature of fat tails, could also play a larger role across more diverse markets for energy options.

Similar quantitative concepts apply to metals and other commodity markets which are widely expected to grow with the energy transition. Industrial metals will play a critical role in the electrification of the global economy. The markets for

copper and aluminum are already highly financialized, being predominantly driven by fluctuations in global demand. Like oil, metals are also valued as an inflation hedge. The markets for precious metals, such as gold and silver, continue to compete with oil for the status of being a safe haven during selloffs across broader financial markets. Finally, the development of corn-based ethanol and biodiesel markets will further tie the agricultural complex to energy. It will provide important diversification and a unique seasonal component to broader portfolios of virtual commodities.

To summarize, whether one likes it or not, commodity prices will be heavily influenced by financial traders. The markets for raw commodities have been studied for more than a century. The supply and demand for physical commodities were much better at explaining price behavior in the past. The present and, most importantly, the future will be more impacted by the supply and demand for virtual commodities traded in the derivatives market. This market requires a different set of skills. While media and policymakers continue to explain market drivers with a simplified narrative for a broader audience, the computers that increasingly drive commodity prices are unlikely to be able to hear them. They will continue to do what we have attempted to describe in this book and a lot more that we cannot pretend to know.

Further quantification and digitalization of commodity markets is on its way. One cannot stop the market evolving from primitive towards more complex ways of doing things. The transition towards more quantitative commodity trading is no different from the technological innovations that we are seeing in many other areas of our lives. The virtual energy transition will inevitably be driven by big data and more advanced quantitative trading methods. Virtual barrels describe only the beginning of this quantitative transition; much more is yet to come in the world of virtual commodities.

---

## Appendix A: Diffusions and Probabilities

We assume that an asset price  $x$  follows the *diffusion* process of the form

$$dx = \mu(x, t)dt + \sigma(x, t)dz \quad (\text{A.1})$$

defined by *drift*  $\mu(x, t)$  and the *local volatility* function  $\sigma(x, t)$ . The increment  $dz$  represents a normally distributed random variable with zero mean and variance equal to  $dt$ .

A function  $G(x, t)$  that depends on a stochastic variable  $x$  also follows the diffusion process which is specified by *Itô's lemma*:

$$dG = \left( \frac{\partial G}{\partial t} + \frac{1}{2} \sigma^2(x, t) \frac{\partial^2 G}{\partial x^2} \right) dt + \frac{\partial G}{\partial x} dx \quad (\text{A.2})$$

Itô's lemma is the stochastic analogue of Taylor's formula used in the ordinary calculus, but it retains the additional second-order term. This is because in the stochastic calculus  $(dx)^2$  is of the same order of magnitude as  $dt$ . Itô's lemma shows that the variable  $x$  and the function  $G(x, t)$  are driven by the same source of uncertainty  $dz$ .

The *Dirac delta function*, or *impulse function*, centered at  $x_0$ , is informally defined as the abstract function that is equal to zero everywhere, except for a single point  $x_0$ , where its value is infinite:

$$\delta(x - x_0) = \begin{cases} 0, & x \neq x_0 \\ \infty, & x = x_0 \end{cases} \quad (\text{A.3})$$

More formally, the Dirac delta function can be understood as the limit of normal probability density with the mean  $x_0$  and the standard deviation  $\sigma \rightarrow 0$ :

$$\delta(x - x_0) = \lim_{\sigma \rightarrow 0} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-x_0}{\sigma}\right)^2}$$

Since the integral of the probability density function over all possible values of  $x$  is equal to one, the integral of the Dirac delta function is also equal to one

$$\int_{-\infty}^{\infty} \delta(x - x_0) dx = 1 \quad (\text{A.4})$$

When the Dirac delta function centered at  $x_0$  is multiplied by a function  $f(x)$  and integrated over all possible values of  $x$ , then the integral isolates the value of the function at the point  $x_0$ :

$$f(x_0) = \int_{-\infty}^{\infty} f(x) \delta(x - x_0) dx \quad (\text{A.5})$$

Since (A.1) has a stochastic component, the behavior of variable  $x$  can only be understood in a probabilistic sense. Let  $p(x, t; x_T, T)$  denote the probability density function generated by the process (A.1). It describes the transition probability of the asset price reaching various levels of  $x_T$  at a future time  $T > t$  given that its price is equal to  $x$  at time  $t$ . To find this probability density, one must solve the following *Fokker-Planck* equation, which is also known as the *Kolmogorov forward equation*:

$$\frac{\partial p}{\partial T} + \frac{\partial}{\partial x_T} (\mu(x_T, T)p) - \frac{1}{2} \frac{\partial^2}{\partial x_T^2} (\sigma^2(x_T, T)p) = 0 \quad (\text{A.6})$$

The equation is written with respect to  $x_T$  and it propagates forward in time for  $T > t$ . The initial condition is given by the Dirac delta function centered at a given asset price  $x$  at time  $T = t$ :

$$p(x, t, x_T, t) = \delta(x_T - x) \quad (\text{A.7})$$

This initial condition means that the asset price at time  $t$  is certain, and the entire probability mass is concentrated in a single point when  $x_T = x$ .

The same probability density function also describes the probability of where the asset  $x$  was at an earlier time  $t$ , if its price at a later time  $T > t$  is known to be  $x_T$ . It satisfies the following *Kolmogorov backward equation*, defined for  $t < T$ :

$$\frac{\partial p}{\partial t} + \mu(x, t) \frac{\partial p}{\partial x} + \frac{1}{2} \sigma^2(x, t) \frac{\partial^2 p}{\partial x^2} = 0 \quad (\text{A.8})$$

This equation is with respect to  $x$  and it moves backward in time for  $t < T$ . The terminal boundary condition at  $t = T$  is given by the Dirac delta function centered at  $x = x_T$ :

$$p(x, T, x_T, T) = \delta(x - x_T) \quad (\text{A.9})$$

In applied mathematics, Eqs. (A.6) and (A.8) are known, respectively, as the *forward and backward parabolic differential equations* that describe the diffusion of heat in a medium, and the function  $p(x, t, x_T, T)$  is called the *fundamental solution* to these equations.

In this book, we are mainly interested in the backward Eq. (A.8), which applies to the pricing of a financial derivative at time  $t$  if its payoff is specified at a later time  $T > t$ . The function  $p(x, t, x_T, T)$  represents only one particular solution to the Eq. (A.8) with a very special boundary condition given by the Dirac delta function. Any other solution  $G(x, t)$  to (A.8) with a different boundary condition specified at time  $T$  by

$$G(x, T) = g(x)$$

can be expressed in terms of its fundamental solution, or the probability density function, as follows:

$$G(x, t) = \int_{-\infty}^{\infty} p(x, t, x_T, T)g(x_T)dx_T \quad (\text{A.10})$$

In other words,  $G(x, t)$  can be understood as the expected value of the function  $g(x)$  when the random behavior of  $x$  is described by the probability density function  $p(x, t, x_T, T)$ .

In Chap. 8, we show that if  $x$  represents the futures contract, then the price of any financial derivative of  $x$  solves (A.8) with  $\mu(x, t) = 0$ . The derivatives price can then be expressed in the form of (A.10) with  $p(x, t, x_T, T)$  representing the so-called *risk-neutral probability density* that corresponds to a modified diffusion process (A.1) for which the drift term is eliminated.

We list three examples of probability density functions used throughout the book. The first one corresponds to *Arithmetic Brownian Motion (ABM)*, which is defined by (A.1) with constant drift and volatility coefficients:

$$dx = \mu_A dt + \sigma_A dt \quad (\text{A.11})$$

The solution to (A.8) that corresponds to ABM is given by the *normal probability density*:

$$p_N(x, t; x_T, T) = \frac{1}{\sqrt{2\pi(T-t)}\sigma_A} e^{-\frac{(x_T - x)^2}{2\sigma_A^2(T-t)}} \quad (\text{A.12})$$

The second important diffusion process is *Geometric Brownian Motion (GBM)*, where the drift and local volatility of the diffusion process are assumed to be proportional to the random variable:

$$dx = \mu_G x dt + \sigma_G x dz \quad (\text{A.13})$$

To find the corresponding probability density function, we define a new variable

$$y = \ln(x)$$

It follows from Itô's lemma (A.2) that

$$dy = \left( \mu_G - \frac{\sigma_G^2}{2} \right) dt + \sigma_G dz$$

This means that the logarithm of the variable  $x$  is normally distributed with mean  $\mu_G - \frac{\sigma_G^2}{2}$  and variance  $\sigma_G^2$ . The corresponding probability function that solves equations (A.6) and (A.8) is given by

$$p_{LN}(x, t; x_T, T) = \frac{1}{\sqrt{2\pi(T-t)\sigma_G x_T}} e^{-\frac{\left(\ln\left(\frac{x_T}{x}\right) - \left(\mu_G - \frac{\sigma_G^2}{2}\right)(T-t)\right)^2}{2\sigma_G^2(T-t)}} \quad (\text{A.14})$$

Since the logarithm of the random variable is normally distributed, (A.14) is called the *lognormal probability density*. It is defined only for positive  $x$  and widely used for modeling financial assets, such as equities, whose prices cannot be negative.

The third important diffusion is the *mean-reverting (MR) or Ornstein-Uhlenbeck process*, which modifies ABM by allowing the drift term to gravitate towards its long-term mean  $\bar{x}$  with speed of mean-reversion  $k > 0$ :

$$dx = k(\bar{x} - x)dt + \sigma_A dz \quad (\text{A.15})$$

Such a mean-reverting process is more suitable for modeling commodity prices where mean-reversion is induced by the cyclicalities of supply and demand.

To derive the probability density for MR process (A.15), consider the following shifted variable

$$y(x, t) = \bar{x} + (x - \bar{x})e^{-k(T-t)}$$

which describes how  $x$  is pulled towards its equilibrium level  $\bar{x}$ , where speed of mean-reversion is characterized by  $k$ .

Applying Itô's lemma (A.2) to  $y(x, t)$ , we obtain that

$$dy = \left( k(x - \bar{x})e^{-k(T-t)} + 0 \right) dt + e^{-k(T-t)} \left( k(\bar{x} - x)dt + \sigma_A dz \right) = e^{-k(T-t)} \sigma_A dz$$

In other words, the shifted variable  $y(t)$  follows the ABM process, where for brevity, we no longer explicitly show its dependence on  $x$ .

The drift of this process is zero, but its variance is reduced by the time-dependent exponential factor. This factor can be eliminated by switching to the new integrated-time variable  $\hat{t}$  defined by

$$T - \hat{t} = \int_t^T e^{-2k(T-s)} ds = \frac{1 - e^{-2k(T-t)}}{2k} \quad (\text{A.16})$$

Following the steps described in Appendix D for pricing options with time-dependent volatility, one can reduce the Eq. (A.8) for MR process (A.15) to the one with constant volatility in variables  $(y, \hat{t})$ :

$$\frac{\partial p}{\partial \hat{t}} + \frac{\sigma_A^2}{2} \frac{\partial^2 p}{\partial y^2} = 0$$

Note that when  $t = T$ , then  $y(T) = x_T$  and  $\hat{t} = T$ , which preserves the boundary condition (A.9) in the new variables  $y$  and  $\hat{t}$ .

The solution to this equation, which is the probability density for MR process (A.15), is then also given by the normal probability density

$$\begin{aligned} p_{MR}(x, t; x_T, T) &= \frac{1}{\sqrt{2\pi(T-\hat{t})\sigma_A^2}} e^{-\frac{(y(T)-y(t))^2}{2\sigma_A^2(T-\hat{t})}} \\ &= \frac{1}{\sqrt{2\pi(T-\hat{t})\sigma_A^2}} e^{-\frac{(x_T - \bar{x} - (x - \bar{x})e^{-k(T-t)})^2}{2\sigma_A^2(T-\hat{t})}} \end{aligned} \quad (\text{A.17})$$

with shifted drift and reduced variance, where  $\hat{t}$  is defined by (A.16).

---

## Appendix B: Option Pricing under Normal and Lognormal Distributions

In Chap. 8, it is shown that the price for a financial derivative that follows a stochastic process (A.1) satisfies the Kolmogorov backward Eq. (A.8) with zero drift  $\mu(F, t) = 0$ . Therefore, its solution is given by the representation (A.10), where  $g(x)$  is the option payoff.

For ABM process (A.11) with constant volatility, the value of the call option is, therefore, obtained by the following integral of the normal probability density (A.12) with  $\mu_A = 0$  and the option payoff:

$$C(x, t) = \int_{-\infty}^{\infty} p_N(x, t, x_T, T) \max(0, x_T - K) dx_T$$

It can be shown that the integration results in the following Bachelier pricing formula with constant normal volatility  $\sigma_A$  and the time remaining to maturity  $\tau = T - t$  is given by the Bachelier formula:

$$C_{BC}(F, t) = e^{-r\tau} \{ (F - K)N(m_A) + \sigma_A \sqrt{\tau} n(m_A) \} \quad (\text{B.1})$$

Here

$$m_A = \frac{F - K}{\sigma_A \sqrt{\tau}}$$

represents the normalized option moneyness,

$$\tau = T - t$$

the time remaining to maturity, and  $N$  is the cumulative normal distribution with zero mean and variance equal to one, whose density is given by

$$\frac{\partial N}{\partial x} = n(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad (\text{B.2})$$

The normal delta is the first derivative of the Bachelier formula with respect to the futures price:

$$\Delta_N = \frac{\partial C_{BC}}{\partial F} = e^{-r\tau} N(m_A) \quad (\text{B.3})$$

The normal gamma is the second derivatives of the Bachelier formula with respect to the futures price:

$$\Gamma_N = \frac{\partial^2 C_{BC}}{\partial F^2} = e^{-r\tau} \frac{n(m_A)}{\sigma_A \sqrt{\tau}} \quad (\text{B.4})$$

The normal vega is the partial derivative of the Bachelier formula with respect to the volatility  $\sigma_A$ :

$$V_N = \frac{\partial C_{BC}}{\partial \sigma_A} = e^{-r\tau} n(m_A) \sqrt{\tau} \quad (\text{B.5})$$

The normal theta is the partial derivative of the Bachelier formula with respect to time:

$$\Theta_N = \frac{\partial C_{BC}}{\partial t} = -e^{-r\tau} \frac{\sigma_A n(m_A)}{2\sqrt{\tau}} + rC_{BC} \quad (\text{B.6})$$

The derivatives pricing Eq. (8.8) is satisfied as

$$\Theta_N + \frac{\sigma_A^2}{2} \Gamma_N - rC_{BC} = 0$$

Similarly, for GBM process (A.13) the pricing formula is given by the integral of the payoff function with lognormal probability density with  $\mu_G = 0$

$$C(x, t) = \int_0^\infty p_{LN}(x, t, x_T, T) \max(0, x_T - K) dx_T$$

The evaluation of this integral results in the following Black formula with constant percentage volatility  $\sigma_G$ :

$$C_{BL}(F, t) = e^{-r\tau} \left\{ FN\left(m_G + \frac{\sigma_G \sqrt{\tau}}{2}\right) - KN\left(m_G - \frac{\sigma_G \sqrt{\tau}}{2}\right) \right\} \quad (\text{B.7})$$

where

$$m_G = \frac{\ln(F/K)}{\sigma_G \sqrt{\tau}}$$

represents the normalized log-moneyness.<sup>1</sup>

---

<sup>1</sup>The integration details are presented in many standard derivatives textbooks, such as Wilmott et al. (1993) and Hull (2018).

The corresponding Black delta, gamma, vega, and theta are:

$$\Delta_{BL} = \frac{\partial C_{BL}}{\partial F} = e^{-r\tau} N\left(m_G + \frac{\sigma_G \sqrt{\tau}}{2}\right) \quad (\text{B.8})$$

$$\Gamma_{BL} = \frac{\partial^2 C_{BL}}{\partial F^2} = e^{-r\tau} \frac{n\left(m_G + \frac{\sigma_G \sqrt{\tau}}{2}\right)}{\sigma_G F \sqrt{\tau}} \quad (\text{B.9})$$

$$V_{BL} = \frac{\partial C_{BL}}{\partial \sigma_G} = e^{-r\tau} n\left(m_G + \frac{\sigma_G \sqrt{\tau}}{2}\right) F \sqrt{\tau} \quad (\text{B.10})$$

$$\Theta_{BL} = \frac{\partial C_{BL}}{\partial t} = -e^{-r\tau} \frac{\sigma_G F n\left(m_G + \frac{\sigma_G \sqrt{\tau}}{2}\right)}{2\sqrt{\tau}} + rC_{BL} \quad (\text{B.11})$$

The following identity

$$Fn\left(m_G + \frac{\sigma_G \sqrt{\tau}}{2}\right) = Kn\left(m_G - \frac{\sigma_G \sqrt{\tau}}{2}\right)$$

was used in the calculation of lognormal partial derivatives. The pricing equation

$$\Theta_{BL} + \frac{\sigma_G^2 F^2}{2} \Gamma_{BL} - rC_{BL} = 0$$

is satisfied. Partial derivatives for put options can be obtained from the Greeks for the call option by differentiating the put-call parity relationship (8.12).

---

## Appendix C: The Perturbation Method and the Quadratic Normal Model

Let us assume that the local volatility function is defined by a small perturbation  $\varepsilon(F)$  of the constant volatility  $\sigma_A$

$$\sigma(F) = \sigma_A + \varepsilon(F) \quad (\text{C.1})$$

Then the partial differential Eq. (8.8) for the option price is

$$\frac{\partial C}{\partial t} + \frac{1}{2}(\sigma_A + \varepsilon(F))^2 \frac{\partial^2 C}{\partial F^2} = 0 \quad (\text{C.2})$$

We seek its solution  $C(F, t)$  as the sum of the Bachelier formula  $C_{BC}(F, t)$  that corresponds to constant normal volatility  $\sigma_A$  and the skew correction function  $U(F, t)$ :

$$C(F, t) = C_{BC}(F, t) + U(F, t) \quad (\text{C.3})$$

To derive the equation for  $U(F, t)$ , we substitute (C.3) into (C.2) and regroup the terms as follows, where, for brevity, we suppress the dependency of function  $\varepsilon$  on futures price  $F$

$$\begin{aligned} & \left( \frac{\partial C_{BC}}{\partial t} + \frac{\sigma_A^2}{2} \frac{\partial^2 C_{BC}}{\partial F^2} \right) + \left( \frac{\partial U}{\partial t} + \frac{\sigma_A^2}{2} \frac{\partial^2 U}{\partial F^2} + \varepsilon \sigma_A \frac{\partial^2 C_{BC}}{\partial F^2} \right) \\ & + \left( \frac{\varepsilon^2}{2} \frac{\partial^2 C_{BC}}{\partial F^2} + \left( \varepsilon \sigma_A + \frac{\varepsilon^2}{2} \right) \frac{\partial^2 U}{\partial F^2} \right) = 0 \end{aligned}$$

The sum of the two terms grouped in the first parenthesis is zero because  $C_{BC}$  solves (C.2) with constant dollar volatility  $\sigma_A$  and  $\varepsilon = 0$ . The terms grouped in the last parenthesis are of a higher order with respect to  $\varepsilon$ . *The method of perturbation or linearization* assumes that for small perturbation  $\varepsilon$ , these terms can be omitted.

For small  $\varepsilon$ , we are left with only the terms grouped in the middle parenthesis, which leads to the following equation for  $U(F, t)$ :

$$\frac{\partial U}{\partial t} + \frac{\sigma_A^2}{2} \frac{\partial^2 U}{\partial F^2} + \varepsilon \sigma_A \frac{\partial^2 C_{BC}}{\partial F^2} = 0$$

To simplify this equation, we let  $\tau = T - t$ , and replace the second derivative of the Bachelier formula, which is simply its gamma, with its explicit formula (B.4). The equation for the skew correction then becomes

$$\frac{\partial U}{\partial \tau} - \frac{\sigma_A^2}{2} \frac{\partial^2 U}{\partial F^2} = \frac{\varepsilon}{\sqrt{\tau}} n \left( \frac{F - K}{\sigma_A \sqrt{\tau}} \right) \quad (\text{C.4})$$

We look for a solution to (C.4) of the following form

$$U(F, \tau) = \sqrt{\tau} p(F, \tau) n \left( \frac{F - K}{\sigma_A \sqrt{\tau}} \right) \quad (\text{C.5})$$

where  $p(F, \tau)$  is a yet to be defined function. We evaluate partial derivatives of  $U(F, \tau)$  using the chain rule, suppressing arguments of functions  $p$  and  $n$ :

$$\begin{aligned} \frac{\partial U}{\partial \tau} &= \frac{pn}{2\sqrt{\tau}} + \sqrt{\tau} \frac{\partial p}{\partial \tau} n + \sqrt{\tau} pn \frac{(F - K)^2}{2\sigma_A^2 \tau^2} \\ \frac{\partial U}{\partial F} &= \sqrt{\tau} \frac{\partial p}{\partial F} n - \sqrt{\tau} pn \frac{(F - K)}{\sigma_A^2 \tau} \\ \frac{\partial^2 U}{\partial F^2} &= \sqrt{\tau} \frac{\partial^2 p}{\partial F^2} n - 2\sqrt{\tau} \frac{\partial p}{\partial F} n \frac{(F - K)}{\sigma_A^2 \tau} + \sqrt{\tau} pn \frac{(F - K)^2}{\sigma_A^4 \tau^2} - \sqrt{\tau} \frac{pn}{\sigma_A^2 \tau} \end{aligned}$$

and substitute them into (C.4). After cancelling the common multiplier  $n$  and multiplying each term by  $\sqrt{\tau}$ , we arrive at the following equation for the function  $p(F, \tau)$

$$p + \tau \frac{\partial p}{\partial \tau} - \frac{\sigma_A^2 \tau}{2} \frac{\partial^2 p}{\partial F^2} + (F - K) \frac{\partial p}{\partial F} = \varepsilon \quad (\text{C.6})$$

We can now choose to specify the perturbation function to be a quadratic of the form

$$\varepsilon(F) = a + bF + cF^2 \quad (\text{C.7})$$

In this case, the Eq. (C.6) can be solved by letting  $p(F, \tau)$  be a polynomial function:

$$p(F, \tau) = w_0 + w_1 F + w_2 F^2 + w_3 \tau$$

where  $w_i$  can be found by the method of undetermined coefficients. Specifically, we equate the coefficients in the left-hand side and the right-hand side of the Eq. (C.6) for polynomial terms of the same order, which leads to the following system of linear equations:

$$w_0 - Kw_1 = a$$

$$2w_1 - 2Kw_2 = b$$

$$3w_2 = c$$

$$2w_3 - \sigma_A^2 w_2 = 0$$

This system is then inverted for coefficients  $w_i$  as follows:

$$w_0 = a + \frac{Kb}{2} + \frac{K^2 c}{3}$$

$$w_1 = \frac{b}{2} + \frac{Kc}{3}$$

$$w_2 = \frac{c}{3}$$

$$w_3 = \frac{\sigma_A^2 c}{6}$$

After some simple algebra, the polynomial function  $p(F, \tau)$  simplifies to

$$\begin{aligned} p(F, \tau) &= \left( a + \frac{Kb}{2} + \frac{K^2 c}{3} \right) + \left( \frac{b}{2} + \frac{Kc}{3} \right) F + \frac{c}{3} F^2 + \frac{c}{6} \sigma_A^2 \tau \\ &= a + \frac{b}{2}(K + F) + \frac{c}{3}(K^2 + KF + F^2) + \frac{c}{6} \sigma_A^2 \tau \end{aligned}$$

and representation (C.5) leads to the solution for the skew correction function

$$U(F, \tau) = \sqrt{\tau} n \left( \frac{F - K}{\sigma_A \sqrt{\tau}} \right) \left( a + \frac{b}{2}(F + K) + \frac{c}{3}(F^2 + FK + K^2) + \frac{c}{6} \sigma_A^2 \tau \right) \quad (\text{C.8})$$

Also, since both  $C(F, T)$  and  $C_{BC}(F, T)$  at expiration are equal to the option payoff,  $\max(0, F - K)$ , the skew correction function at time  $t = T$  must be equal to zero. This boundary condition for the skew correction function at  $\tau = 0$

$$U(F, 0) = 0$$

is clearly satisfied by the formula (C.8).

If necessary, this method can be easily extended to a higher-order polynomial volatility perturbation function, in which case  $p(F, \tau)$  must also be taken to be a polynomial of the same order with coefficients determined by matching corresponding terms of the Eq. (C.6).

---

## Appendix D: Option Pricing with Time-Dependent Volatility

We consider the GBM process for the  $T$ -maturity futures price  $F(t, T)$  with time-dependent lognormal volatility  $\sigma_G(t, T)$  as in (11.1). The option price is then the solution of the following partial differential equation:

$$\frac{\partial C}{\partial t} + \frac{1}{2} \sigma_G^2(t, T) F^2 \frac{\partial^2 C}{\partial F^2} = 0 \quad (\text{D.1})$$

In general, the option expires at time  $T_0 \leq T$ . The payoff of the call option is given by

$$C(F, T_0) = \max(0, F - K)$$

We introduce a new time variable  $\hat{t}$ , defined as

$$\hat{t} = T_0 - \frac{1}{v^2} \int_t^{T_0} \sigma_G^2(s, T) ds \quad (\text{D.2})$$

where  $v$  is constant.

Let  $\hat{C}(F, \hat{t})$  represent the call option price using this new time variable

$$C(F, t) = \hat{C}(F, \hat{t})$$

We calculate the derivative with respect to time using the chain rule

$$\frac{\partial C}{\partial t} = \frac{\partial \hat{C}}{\partial \hat{t}} \frac{\partial \hat{t}}{\partial t} = \frac{\sigma_G^2(t, T)}{v^2} \frac{\partial \hat{C}}{\partial \hat{t}}$$

The substitution of this partial derivative into Eq. (D.1) with time variable volatility  $\sigma_G(t, T)$  replaces it with an identical equation but with constant volatility  $v$

$$\frac{\partial \hat{C}}{\partial \hat{t}} + \frac{1}{2} v^2 F^2 \frac{\partial^2 \hat{C}}{\partial F^2} = 0$$

Note that when  $t = T_0$  then  $\hat{t} = T_0$  so that the boundary condition remains intact

$$\hat{C}(F, T_0) = \max(0, F - K)$$

The price of the call option is simply given by the Black formula with respect to time  $\hat{t}$  and with constant volatility  $v$ :

$$\hat{C}(F, \hat{t}) = C_{BL}(F, \hat{t}; v)$$

It follows from (D.2) that the total variance in the Black formula in the new variables is equal to the integrated local variance in the original variables

$$v^2(T_0 - \hat{t}) = \int_t^{T_0} \sigma_G^2(s, T) ds$$

Therefore, the option price  $C(F, t)$  is given by the same Black formula with implied volatility  $v$  given by the quadratic mean of the local volatility

$$v = \sqrt{\frac{1}{T_0 - t} \int_t^{T_0} \sigma_G^2(s, T) ds}$$

The calculations above are identical for time-dependent normal volatility, in which case the option price is given by the Bachelier formula with volatility  $v_N$  equal to the mean-square of the local volatility

$$v_N = \sqrt{\frac{1}{T_0 - t} \int_t^{T_0} \sigma_A^2(s, T) ds}$$

---

## Appendix E: Average Price Options

Let the futures price follow the general diffusion process (A.1):

$$dF = \mu(F, t)dt + \sigma(F, t)dz \quad (\text{E.1})$$

We consider the problem of pricing an APO within the averaging period, i.e.,  $T_a \leq t \leq T$ . The rolling average is defined as

$$A(t) = \frac{1}{t - T_a} \int_{T_a}^t F(u)du$$

By differentiating  $A(t)$  with respect to  $t$ , we obtain the stochastic differential equation for the rolling average as

$$dA = \left( -\frac{1}{(t - T_a)^2} \int_{T_a}^t F(u)du + \frac{F(t)}{t - T_a} \right) dt = \frac{F - A}{t - T_a} dt$$

The option price  $C_{APO}(F, A, t)$  depends on two spatial variables,  $F$  and  $A$ . The latter, however, does not have the stochastic term that contains  $dz$ . The pricing equation for an APO follows from the two-dimensional Itô's lemma and the BSM hedging argument from Chap. 8<sup>2</sup>:

$$\frac{\partial C_{APO}}{\partial t} + \frac{1}{2}\sigma^2(F, t, T) \frac{\partial^2 C_{APO}}{\partial F^2} + \left( \frac{F - A}{t - T_a} \right) \frac{\partial C_{APO}}{\partial A} = 0 \quad (\text{E.2})$$

Such two-dimensional partial differential equations are difficult to solve analytically. However, the specific form of (E.2) allows for a reduction of dimensionality using the auxiliary variable  $x$

---

<sup>2</sup>See, for example, Kemna and Vorst (1990) and Wilmott et al. (1993) for a more detailed derivation.

$$x(t) = \left( \frac{t - T_a}{T - T_a} \right) A(t) + \left( \frac{T - t}{T - T_a} \right) F(t)$$

We make the change of variables

$$C_{APO}(F, A, t) = U(x, t)$$

and evaluate partial derivatives using the chain rule, as follows:

$$\frac{\partial C_{APO}}{\partial t} = \frac{\partial U}{\partial t} + \frac{\partial U}{\partial x} \frac{\partial x}{\partial t} = \frac{\partial U}{\partial t} + \left( \frac{A - F}{T - T_a} \right) \frac{\partial U}{\partial x}$$

$$\frac{\partial C_{APO}}{\partial F} = \frac{\partial U}{\partial x} \frac{\partial x}{\partial F} = \left( \frac{T - t}{T - T_a} \right) \frac{\partial U}{\partial x}, \quad \frac{\partial^2 C_{APO}}{\partial F^2} = \left( \frac{T - t}{T - T_a} \right)^2 \frac{\partial^2 U}{\partial x^2}$$

$$\frac{\partial C_{APO}}{\partial A} = \frac{\partial U}{\partial x} \frac{\partial x}{\partial A} = \left( \frac{t - T_a}{T - T_a} \right) \frac{\partial U}{\partial x}$$

Then the two-dimensional partial differential Eq. (E.2) for an APO reduces to

$$\frac{\partial U}{\partial t} + \frac{1}{2} \left( \frac{T - t}{T - T_a} \right)^2 \sigma^2(F, t, T) \frac{\partial^2 U}{\partial x^2} = 0, \quad T_a < t < T$$

The local volatility in this equation is adjusted by a linear multiplier that reflects gradually decreasing volatility as the average accumulates throughout the averaging period. Outside of the averaging period for  $t > T_a$ , no adjustment is needed and the standard differential Eq. (8.8) with respect to  $t$  and  $F$  applies with the local volatility function  $\sigma(F, t, T)$ .

---

## Appendix F: The Inverse Diffusion Problem

To derive the Dupire equation, we use the fact that the risk-neutral probability density  $p(F, t; K, T)$  satisfies the Fokker-Planck Eq. (A.6) with zero drift with respect to variables  $K$  and  $T$ :

$$\frac{\partial p}{\partial T} - \frac{1}{2} \frac{\partial^2}{\partial K^2} (\sigma^2(K, T)p) = 0$$

Using representation (12.4), we replace the probability density function with the second derivative of the call option payoff:

$$\frac{\partial}{\partial T} \left( \frac{\partial^2 C}{\partial K^2} \right) - \frac{1}{2} \frac{\partial^2}{\partial K^2} \left( \sigma^2(K, T) \left( \frac{\partial^2 C}{\partial K^2} \right) \right) = 0$$

We then change the order of differentiation in the first term, and integrate both terms twice with respect to  $K$ , which leads to the following equation for the option price with respect to strikes  $K$  and maturities  $T$ <sup>3</sup>:

$$\frac{\partial C}{\partial T} - \frac{1}{2} \sigma^2(K, T) \frac{\partial^2 C}{\partial K^2} = 0 \quad (\text{F.1})$$

We next consider a more practical case, where the local volatility only depends on the spatial variable and option prices are available for the continuum of strikes but only for a fixed maturity  $T$ , which is the case for the oil market.

Let  $\tau = T - t$  represent the time remaining to maturity. Then the Dupire Eq. (F.1) with time-independent volatility is

---

<sup>3</sup>Some technical regularity conditions must be imposed to make sure that boundary terms at  $K \rightarrow 0$  and  $K \rightarrow \infty$  vanish, which for simplicity, we omit here.

$$\frac{\partial C}{\partial \tau} - \frac{1}{2} \sigma^2(K) \frac{\partial^2 C}{\partial K^2} = 0 \quad (\text{F.2})$$

In applied mathematics, this is the equation of heat transfer in a non-homogeneous medium, described by an unknown space-dependent thermal conductivity  $\sigma(K)$ .

The equation is supplemented with an initial condition, which is defined by the payoff of a call option at maturity  $\tau = 0$ :

$$C(K, 0) = \max(0, F - K) \quad (\text{F.3})$$

The futures price  $F$  here plays the role of an exogenous parameter.

Furthermore, we assume that market prices of options with all strikes  $C^*(K)$  are observed today at time  $t = 0$ , or when  $\tau = T$ :

$$C(K, T) = C^*(K) \quad (\text{F.4})$$

We can now formulate the following inverse problem:

*The Inverse Problem with Final Over-Determination*<sup>4</sup>: If the solution to the boundary-value problem (F.2) and (F.3) is measured at a given time by (F.4), then is it possible to uniquely reconstruct the unknown coefficient  $\sigma(K)$ ?

This problem arises in many applications. For example, when applied to the equation of heat transfer in a medium, it raises the question whether it is possible to uniquely reconstruct unknown conductivity properties of the medium from the measurement of the temperature distribution across all points at a given time.

This problem can also be restated using the jargon of probabilities. Let us consider the Eq. (F.2) but with a different initial condition, where the call option payoff is replaced with the Dirac delta function

$$C(K, 0) = \delta(K - F)$$

Then the solution represents the probability density function for a diffusion process with unknown time-independent diffusion coefficient. This means that the inverse problem can be restated in terms of probability densities as follows:

*The Inverse Diffusion Problem*: Let  $p(F, 0; K, \tau)$  represent the diffusion probability of a particle being at the point  $K$  at time  $\tau$  given its starting point  $K = F$  at  $\tau = 0$ . Is it possible to uniquely reconstruct the time-independent diffusion coefficient  $\sigma(K)$  from a single observation of the probability density at time  $T$ ?<sup>5</sup>

---

<sup>4</sup>This problem was formulated in Bouchouev and Isakov (1997, 1999) and solved analytically only under additional assumptions. For numerical solutions, see also references in the footnote in Chap. 12.

<sup>5</sup>The drift of the diffusion process must be fixed, and here it is assumed to be zero. It is well known that two diffusions with different drifts and time-independent volatilities can generate the same probability density. For example, the mean-reverting process and arithmetic Brownian motion represent different diffusion processes, but they can produce the same normal distribution, as shown in Appendix A.

The glossary summarizes key terms used by professional oil derivatives traders. It reflects the jargon and conventions of the oil market, as some terms may be used in different contexts in other markets and in academic studies.

---

## Glossary

**Actuarial Valuation** A method of pricing options based on their average historical payoffs.

**Arbitrage** A strategy that involves buying and selling similar contracts to capture discrepancies in their prices.

**Average Price Option (APO)** An option that settles based on the average price of underlying futures contracts over a given period.

**Backwardation** A situation where the futures curves declines with increasing time to maturity.

**Barrel Counting** Fundamental analysis of supply and demand.

**Basis** The spread between the price of a physical or a less liquid financial contract and the benchmark futures price.

**Bid-Ask Spread** The spread between the price to buy (bid) and the price to sell (ask).

**Black Volatility** Implied volatility computed using the Black (1976) option pricing formula.

**Butterfly Spread** A three-legged option trade that buys options with strikes  $K_1$  and  $K_2$  and sells twice as many options with the strike price set half-way between  $K_1$  and  $K_2$ , all with the same expiration.

**Calendar Spread Option (CSO)** An option on the spread between two futures on the same commodity with different maturities.

**Calibration** A process of fitting volatility models to market prices of traded options.

**Carry** The spread between futures on the same commodity with different maturities, which is used to measure the expected price roll up (down) the curve.

**Cash Price** The price of an over-the-counter contract with physical delivery.

**Cash-Settled Option** An option whose payoff at expiration is calculated financially as the difference between the futures settlement price and the strike price.

**Commodity Currencies** Currencies of commodity-exporting countries.

**Commitments of Traders (CoT) Report** A weekly report on futures and option positions held by different categories of market participants.

**Commodity Trading Advisors (CTAs)** In general, anyone who advises on commodity trading, but in practice, the term is often used to describe hedge funds that trade based on quantitative algorithms.

**Contango** A situation when the futures curve increases with time remaining to maturity.

**Convenience Yield** Consumption benefits accrued to the owner of a physical commodity, but not to the owner of a futures contract.

**Crack Spread** The spread between the price of a refined product and the price of crude oil.

**Cushing** A reference to the storage hub and inventories at the delivery location for WTI futures contract.

**Digital (Binary) Option** An option with a discontinuous payoff that pays a fixed amount or zero depending on the futures price.

**Denomination Effect** An impact on oil price, resulting from oil being denominated in US dollars (USD).

**Excess Return (ER)** A return on buying and rolling futures.

**Fading the Crowded Trade** A contrarian position taken when hedge funds' futures holdings are perceived to be excessive.

**Follow the Flow Strategy** A strategy of mimicking futures position held by hedge funds.

**Fractionation Analysis** Decomposition of profit-and-loss (P&L) of a systematic strategy by the value of a certain explanatory variable.

**Gamma Hedging** The process of option dealers rebalancing their directional price exposure caused by shifts in the futures price.

**Grades** Various types of crude oil that trade as a differential to main futures benchmarks, such as WTI and Brent.

**GSCI Rolls** The process of rolling futures between the fifth and the ninth business days of each month by major commodity indices, such as Goldman Sachs Commodity Index (GSCI).

**Hacienda Hedge** A reference to a large-scale hedging program by the Government of Mexico.

**Hedging Pressure** An imbalance between buyers and sellers of a particular futures or options contract.

**Implied Volatility** An input into an option pricing model that makes the model price of an option to match its market price.

**Local Volatility** A volatility function that depends on futures prices and time and characterizes uncertainty of the random variable in a stochastic diffusion process.

**Market-Implied Probability Distribution** A probability distribution reconstructed from market prices of options with various strikes and the fixed expiration.

**Mean-Reversion** A tendency of prices to revert to a long-term equilibrium.

**Naked Option** An option position (typically a sale) which is not delta hedged.

**Normal Backwardation (Contango)** A situation where the futures price is below (above) the *expected* spot price. The term is equivalent to a risk premium.

**Normal (Dollar) Volatility** Implied volatility computed using the Bachelier option pricing formula.

**Quantamentals** A trading style that combines rule-based decision-making with discretionary overlay.

**Penultimate Expiration** An expiration of derivatives contracts one day prior to the expiration of the futures.

**Physically-Settled Option** An option, which when it is in-the-money, is exercised into futures at expiration.

**Positioning** A reference to positions held by hedge funds and other large market participants.

**Producer Collar (Fence)** A derivatives structure where a producer buys a put option and finances it by selling a call option of the same value.

**Producer Three-Way Collar** A derivatives structure where a producer buys a put spread and finances it by selling a call option of the same value.

**Prompt (Nearby) Futures** A futures contract with the shortest time to expiration.

**Reaction Function** A function that maps the strength of the systematic signal to the position size.

**Realized Volatility** A statistical measure of volatility calculated as the annualized standard deviation of percentage changes in prices over a given period.

**Regime Change** A structural shift in dominant market-driving factors.

**Risk Parity** An asset allocation framework that weighs assets inversely to their volatility and invests in commodity futures as an inflation hedge.

**Roll Return (RR)** The difference between the excess return (ER) and the spot return (SR).

**Signal Blending** A combination of multiple systematic trading signals.

**Skew Delta** A correction to the hedging delta that results from the change in option's vega caused by the move in futures.

**Spot Return (SR)** A return on a hypothetical investment in a spot futures contract with no convenience benefits or storage costs.

**Spread Option** An option on the difference between the prices of two futures.

**Squeeze** A large price move caused by futures traders exiting positions near the expiration of the futures contract.

**Sticky Moneyness** A heuristic rule that assumes that implied volatilities for all options remain unchanged for a given moneyness.

**Sticky Strike** A heuristic rule that assumes that implied volatilities for all options remain unchanged for a given strike.

**Stock-Out** A situation of zero inventories in storage.

**Swap Data Repository (SDR)** A registered entity that collects and disseminates information about over-the-counter transactions on a nearly real-time basis.

**Synthetic Spread Option** The spread between an option on a refined product and an option on crude oil with the same expiration.

**Synthetic Storage** A portfolio of calendar spread options (CSOs) designed to mimic cash flows of physical storage.

**Swaption** An option to enter into a swap.

**Tank-Tops** A situation where inventories reach the maximum storage capacity.

**Trading-at-Settlement (TAS)** A futures contract that allows parties to trade during a trading day at a settlement price which will only be determined later after the market closes.

**Underlying Futures** A futures contract on which the price of an option contract depends.

**Vanilla Options** Exchange-traded options, excluding spread options and APOs.

**Volatility Risk Premium** An investment return of the systematic strategy of selling and delta hedging short term options.

**Volatility Skew (Smile)** A graph of implied volatilities for options versus their strikes or moneyness for the same underlying futures contract.

**Volatility Targeting** A strategy that adjusts a notional position size required to bring expected portfolio volatility to a given target.

**Volatility Term Structure** A graph of implied volatilities versus time to maturity.

---

## References

- Acharya, V. V., Lochstoer, L. A., & Ramadorai, T. (2013). Limits to arbitrage and hedging: Evidence from commodity markets. *Journal of Financial Economics*, 109(2), 441–465.
- Alexander, C. (2001). *Market models*. Wiley.
- Alexander, C. (2008). *Market risk analysis, Vol. III: Pricing, hedging and trading financial instruments*. Wiley.
- Andersen, L. (2011). Option pricing with quadratic volatility: A revisit. *Finance and Stochastics*, 15(2), 191–219.
- Ashton, M., & Greer, R. (2008). History of commodities as the original real return asset class. In *Inflation risk and products* (pp. 85–109). Risk Books.
- Avellaneda, M., Friedman, C., Holmes, R., & Samperi, D. (1997). Calibrating volatility surfaces via relative entropy minimization. *Applied Mathematical Finance*, 4(1), 37–64.
- Bachelier, L. (1900). Théorie de la Spéculation. *Annales scientifiques de l'École Normale Supérieure, Serie 3*, 17, 21–86.
- Backhouse, R. E. (2002). *The ordinary business of life*. Princeton University Press.
- Baker, S. D. (2021). The financialization of storable commodities. *Management Science*, 67(1), 471–499.
- Bakshi, G., Gao, X., & Rossi, A. G. (2019). Understanding the sources of risk underlying the cross section of commodity returns. *Management Science*, 65(2), 619–641.
- Barone-Adesi, G., & Whaley, R. E. (1987). Efficient analytic approximation of American option values. *Journal of Finance*, 42(2), 301–320.
- Black, F. (1976). The pricing of commodity contracts. *Journal of Financial Economics*, 3(1/2), 167–179.
- Black, F., & Scholes, M. (1973). The pricing of options and corporate liabilities. *The Journal of Political Economy*, 81(3), 637–654.
- Blas, J. (2017, April 4). Uncovering the secret history of Wall Street's largest oil trade, *Bloomberg*.
- Bodie, Z., & Rosansky, V. I. (1980, May–June). Risk and return in commodity futures. *Financial Analysts Journal*, 36(3), 27–39.
- Boons, M., & Prado, M. P. (2019). Basis-momentum. *The Journal of Finance*, 74(1), 239–279.
- Bouchouev, I. (1998). Derivatives valuation for general diffusion processes, *Proceedings of the Annual Conference of the International Association of Financial Engineers*, New York, USA, pp. 91–104.
- Bouchouev, I. (2000a, July). *Demystifying Asian options*. Energy and Power Risk Management, pp. 26–27.
- Bouchouev, I. (2000b, August). *Black-Scholes with a smile*. Energy and Power Risk Management, pp. 28–29.
- Bouchouev, I. (2012). Inconvenience yield, or the theory of normal contango. *Quantitative Finance*, 12(12), 1773–1777.
- Bouchouev, I. (2020a, April 30). Negative oil prices put spotlight on investors, *Risk.net*.

- Bouchouev, I. (2020b). From risk bearing to propheteering. *Quantitative Finance*, 20(6), 887–894.
- Bouchouev, I. (2021). A stylized model of the oil squeeze, SSRN.
- Bouchouev, I. (2022). *The Strategic Petroleum Reserve strategies: Risk-free return or return-free risk?* The Oxford Institute for Energy Studies.
- Bouchouev, I. (2023). *Testimony to the US House Subcommittee on Economic Growth, Energy Policy, and Regulatory Affairs of the Committee on Oversight and Accountability*, March 8. Reprinted in *Commodity Insights Digest* (2023), Vol. 1.
- Bouchouev, I., & Isakov, V. (1997). The inverse problem of option pricing. *Inverse Problems*, 13(5), L11–L17.
- Bouchouev, I., & Isakov, V. (1999). Uniqueness, stability and numerical methods for the inverse problem that arises in financial markets. *Inverse Problems*, 15(3), R95–R116.
- Bouchouev, I., Isakov, V., & Valdivia, N. (2002). Recovery of volatility coefficient by linearization. *Quantitative Finance*, 2(4), 257–263.
- Bouchouev, I., & Johnson, B. (2022). The volatility risk premium in the oil market. *Quantitative Finance*, 22(8), 1561–1578.
- Bouchouev, I., & Zuo, L. (2020, Winter). Oil risk premia under changing regimes. *Global Commodities Applied Research Digest*, 5(2), 49–59.
- Boyle, P. C. (1899, September 6). Testimony at the Hearing of the US Industrial Commission on Trusts and Industrial Combinations.
- Breeden, D. T., & Litzenberger, R. H. (1978). Prices of state contingent claims implicit in option prices. *Journal of Business*, 51(4), 621–651.
- Brennan, M. J. (1958). The supply of storage. *The American Economic Review*, 48(1), 50–72.
- Brennan, M. J. (1991). The price of convenience and the valuation of commodity contingent claims. In D. Lund & B. Oksendal (Eds.), *Stochastic models and option values*. North Holland.
- Brennan, M. J., & Schwartz, E. S. (1985). Evaluating natural resource investments. *Journal of Business*, 58(2), 135–157.
- Büyüksahin, B., & Robe, M. A. (2014). Speculators, commodities and cross-market linkages. *Journal of International Money and Finance*, 42, 48–70.
- Carmona, R., & Durrelman, V. (2003). Pricing and hedging spread options. *SIAM Review*, 45(4), 627–685.
- Carmona, R., & Ludkovski, M. (2004). Spot convenience yield models for the energy markets. *Contemporary Mathematics*, 351, 65–79.
- Carr, P., Fisher, T., & Ruf, J. (2013). Why are quadratic normal volatility models analytically tractable? *SIAM Journal on Financial Mathematics*, 4(1), 185–202.
- Carr, P., & Madan, D. (2001). Determining volatility surfaces and option values from an implied volatility smile. In M. Avellaneda (Ed.), *Quantitative analysis in financial markets* (Vol. II, pp. 163–191). World Scientific.
- Casassus, J., & Collin-Dufresne, P. (2005). Stochastic convenience yield implied from commodity futures and interest rates. *The Journal of Finance*, 60(5), 2283–2331.
- Castagna, A., & Mercurio, F. (2007). The vanna-volga method for implied volatilities. *Risk*, 20(1), 106–111.
- Cheng, I.-H., Kirilenko, A., & Xiong, W. (2015). Convective risk flows in commodity futures markets. *Review of Finance*, 19(5), 1733–1781.
- Chiarella, C., Craddock, M., & El-Hassan, N. (2003). An implementation of Bouchouev's method for short time calibration of option pricing models. *Computational Economics*, 22, 113–138.
- Clewlow, L., & Strickland, C. (2000). *Energy derivatives: Pricing and risk management*. Lacima Publications.
- Considine, J., Galkin, P., & Aldayel, A. (2022). Inventories and the term structure of oil prices: A complex relationship. *Resources Policy*, 77, 1–18.
- Daskalaki, C., Kostakis, A., & Skiadopoulos, G. (2014). Are there common factors in individual commodity futures returns? *Journal of Banking and Finance*, 40, 346–363.
- Deaton, A., & Laroque, G. (1992). On the behavior of commodity prices. *The Review of Economic Studies*, 59(1), 1–23.

- Dempster, M. A. H., & Hong, S. S. G. (2002). Spread option valuation and the fast Fourier transform. In H. Geman, D. Madan, S. R. Pliska, & T. Vorst (Eds.), *Mathematical finance, Bachelier congress* (Vol. 1, pp. 203–220). Springer.
- Dempster, M. A. H., Medova, E., & Tang, K. (2012). Determinants of oil futures prices and convenience yields. *Quantitative Finance*, 12(12), 1795–1809.
- Dempster, M. A. H., & Richards, D. G. (2000). Pricing American options fitting the smile. *Mathematical Finance*, 10(2), 157–177.
- Derman, E., & Miller, M. B. (2016). *The volatility smile*. Wiley.
- Doran, J. S., & Ronn, E. I. (2006). The bias in Black-Scholes/Black implied volatility: An analysis of equity and energy markets. *Review of Derivatives Research*, 8(3), 177–198.
- Doran, J. S., & Ronn, E. I. (2008). Computing the market price of volatility risk in the energy commodity markets. *Journal of Banking and Finance*, 32(12), 2541–2552.
- Dupire, B. (1994). Pricing with a smile. *Risk*, 7(1), 18–20.
- Dvir, E., & Rogoff, K. (2009). Three epochs of oil, NBER Working Paper, 14927.
- Ederington, L. H., Fernando, C. S., Holland, K. V., Lee, T. K., & Linn, S. C. (2021). The dynamics of arbitrage. *Journal of Financial and Quantitative Analysis*, 56(4), 1350–1380.
- Ellwanger, R. (2017). On the tail risk premium in the oil market, Bank of Canada Working Paper, 46.
- Erb, C. B., & Harvey, C. R. (2006). The strategic and tactical value of commodity futures. *Financial Analysts Journal*, 62(2), 69–97.
- Eydeland, A., & Wolyniec, K. (2003). *Energy and power risk management: New developments in modeling, pricing, and hedging*. Wiley.
- Fama, E. F., & French, K. R. (1987). Commodity futures prices: Some evidence on forecast power, premiums, and the theory of storage. *Journal of Business*, 60(1), 55–73.
- Fattouh, B. (2011). An anatomy of the crude oil pricing system. *Oxford Institute for Energy Studies*, Working Paper, 40.
- Fattouh, B., & Mahadeva, L. (2014). Causes and implications of shifts in financial participation in commodity markets. *The Journal of Futures Market*, 34(8), 757–787.
- Fernandez-Perez, A., Frijns, B., Fuertes, A.-M., & Miffre, J. (2018). The skewness of commodity futures returns. *The Journal of Banking and Finance*, 86, 143–158.
- Fernandez-Perez, A., Fuertes, A.-M., & Miffre, J. (2021, Summer). On the negative pricing of WTI crude oil futures. *Global Commodities Applied Research Digest*, 6(1), 36–43.
- Fisher, I. (1896). Appreciation and interest. *Publications of the American Economic Association*, 11(4), 331–442.
- Gatheral, J. (2006). *The volatility surface: A Practitioner's guide*. Wiley.
- Geman, H. (2005). *Commodities and commodity derivatives: Modeling and pricing for agriculturals, metals and energy*. Wiley.
- Gibson, R., & Schwartz, E. S. (1990). Stochastic convenience yield and the pricing of oil contingent claims. *The Journal of Finance*, 45(3), 959–976.
- Giddens, P. H. (1947). *Pennsylvania petroleum 1750–1872: A documentary history*. Pennsylvania Historical and Museum Commission.
- Gorton, G. B., Hayashi, F., & Rouwenhorst, K. G. (2012). The fundamentals of commodity futures returns. *Review of Finance*, 17(1), 35–105.
- Gorton, G., & Rouwenhorst, K. G. (2006). Facts and fantasies about commodity futures. *Financial Analysts Journal*, 62(2), 47–68.
- Greer, R. J. (1978, Summer). Conservative commodities: A key inflation hedge. *Journal of Portfolio Management*, 4(4), 26–29.
- Grunspan, C. (2011). A note on the equivalence between the normal and the lognormal implied volatility: A model free approach, *SSRN*.
- Gustafson, R. L. (1958). Carryover levels for grains, *U.S. Department of Agriculture*, Technical Bulletin, 1178.
- Hamilton, J. D. (1983). Oil and the macroeconomy since World War II. *Journal of Political Economy*, 91(2), 228–248.

- Hamilton, J. D. (2003). What is an oil shock? *Journal of Econometrics*, 113(2), 363–398.
- Hamilton, J. D. (2009). Understanding crude oil prices. *The Energy Journal*, 30(2), 179–206.
- Hamilton, J. D., & Wu, J. C. (2014). Risk premia in crude oil futures prices. *Journal of International Money and Finance*, 42, 9–37.
- Hicks, J. R. (1939). *Value and capital: An inquiry into some fundamental principles of economic theory*. Oxford University Press.
- Hirschleifer, D. (1988). Residual risk, trading costs, and commodity futures risk premia. *The Review of Financial Studies*, 1(2), 173–193.
- Hull, J. C. (2018). *Options, futures, and other derivatives* (10th ed.). Pearson.
- Imsirovic, A. (2021). *Trading and price discovery for crude oils: Growth and development of international oil markets*. Palgrave Macmillan.
- Ingersoll, J. E., Jr. (1997). Valuing foreign exchange rate derivatives with a bounded exchange process. *Review of Derivatives Research*, 1, 159–181.
- Interim Staff Report. (2020). *Trading in NYMEX WTI crude oil futures contract leading up to, on, and around April 20, 2020*, Commodity Futures Trading Commission, November 23.
- Jacobs, K., & Li, B. (2023). Option returns, risk premiums, and demand pressure in energy markets. *Journal of Banking and Finance*, 146, 1–26.
- Jawaheri, A. (2005). *Inside volatility arbitrage: The secrets of skewness*. Wiley.
- Johnson, O. (2022). *40 classic crude oil trades: Real-life examples of innovative trading*. Routledge.
- Kaldor, N. (1939). Speculation and economic stability. *The Review of Economic Studies*, 7(1), 1–27.
- Kang, S. B., & Pan, X. (2015). Commodity variance risk premia and expected futures returns: Evidence from the crude oil market, *SSRN*.
- Kang, W., Rouwenhorst, K. G., & Tang, K. (2020). A tale of two premiums: The role of hedgers and speculators in commodity futures markets. *The Journal of Finance*, 75(1), 377–417.
- Kemna, A. G. Z., & Vorst, A. C. F. (1990). A pricing method for options based on average asset values. *Journal of Banking and Finance*, 14(1), 113–129.
- Keynes, J. M. (1923). Some aspects of commodity markets. The Manchester Guardian Commercial, Reconstruction Supplement, March 29.
- Keynes, J. M. (1930). *A treatise on money* (Vol. II). Macmillan.
- Keynes, J. M. (1936). *The general theory of employment, interest, and money*. Macmillan.
- Kilian, L. (2009). Not all oil price shocks are alike: Disentangling demand and supply shocks in the crude oil market. *American Economic Review*, 99(3), 1053–1069.
- Kilian, L. (2020). Understanding the estimation of oil demand and oil supply elasticities, Federal Reserve Bank of Dallas Working Paper, 2027.
- Kilian, L., & Murphy, D. P. (2014). The role of inventories and speculative trading in the global market for crude oil. *Journal of Applied Econometrics*, 29(3), 454–478.
- Kilian, L., & Vigfusson, R. J. (2017). The role of oil price shocks in causing U.S. recessions. *Journal of Money, Credit and Banking*, 49 (8), 1747–1776.
- Kirk, E. (1995). Correlation in the energy markets. In *Managing energy price risk* (pp. 71–78). Risk Publications.
- Koijen, R. S. J., Moskowitz, T. J., Pedersen, L. H., & Vrugt, E. B. (2018). Carry. *Journal of Financial Economics*, 127, 197–225.
- Leoni, P. (2014). *The Greeks and hedging explained*. Palgrave Macmillan.
- Levy, E. (1992). Pricing European average rate currency options. *Journal of International Money and Finance*, 11(5), 474–491.
- Lipton, A. (2001). *Mathematical methods for foreign exchange: A financial engineer's approach*. World Scientific.
- Lipton, A., & Sepp, A. (2011, October). Filling the gaps. *Risk*, 24(10), 78–83.
- Lo, A. W. (2002). The statistics of Sharpe ratios. *Financial Analysts Journal*, 58(4), 36–52.

- Lux, H. (2003, February 1). What becomes a legend? *Institutional Investor*.
- Ma, L. (2022). Negative WTI price: What really happened and what can we learn? *The Journal of Derivatives*, 29(3), 9–29.
- Margrabe, W. (1978). The value of an option to exchange one asset for another. *The Journal of Finance*, 33(1), 177–186.
- Merton, R. C. (1973). Theory of rational option pricing. *The Bell Journal of Economics and Management Science*, 4(1), 141–183.
- Miffre, J. (2016). Long-short commodity investing: A review of the literature. *Journal of Commodity Markets*, 1(1), 3–13.
- Miltersen, K. R. (2003). Commodity price modelling that matches current observables: A new approach. *Quantitative Finance*, 3(1), 51–58.
- Naldi, N. (2015). Sraffa and Keynes on the concept of commodity rates of interest. *Contributions to Political Economy*, 34(1), 17–30.
- Neville, H., Draaisma, T., Funnell, B., Harvey, C. R., & Van Hemert, O. (2021). The best strategies for inflationary times, *SSRN*.
- Pilipović, D. (2007). *Energy risk: Valuing and managing energy derivatives*. McGraw-Hill.
- Pirrong, C. (2012). *Commodity price dynamics: A structural approach*. Cambridge University Press.
- Prokopcuk, M., Symeonidis, L., & Simen, C. W. (2017). Variance risk in commodity markets. *Journal of Banking and Finance*, 81, 136–149.
- Rebonato, R. (2004). *Volatility and correlation: The perfect hedger and the fox*. Wiley.
- Routledge, B. R., Seppi, D. J., & Pratt, C. S. (2000). Equilibrium forward curves for commodities. *The Journal of Finance*, 55(3), 1297–1338.
- Samuelson, P. A. (1965). Proof that properly anticipated prices fluctuate randomly. *Industrial Management Review*, 6(2), 41–49.
- Schachermayer, W., & Teichmann, J. (2008). How close are the option pricing formulas of Bachelier and Black-Merton-Scholes? *Mathematical Finance*, 18(1), 155–170.
- Schwartz, E. S. (1997). The stochastic behavior of commodity prices: Implications for valuation and hedging. *The Journal of Finance*, 52(3), 923–973.
- Shimko, D. C. (1994). Options on futures spreads: Hedging, speculation, and valuation. *The Journal of Futures Markets*, 14(2), 183–213.
- Smiley, A. W. (1907). *A few scraps: Oily and otherwise*. The Derrick Publishing Company.
- Sraffa, P. (1932). Dr. Hayek on money and capital. *The Economic Journal*, 42 (165), 42–53.
- Stoll, H. R. (1979). Commodity futures and spot price determination and hedging in capital market equilibrium. *The Journal of Financial and Quantitative Analysis*, 14(4), 873–894.
- Swindle, G. (2014). *Valuation and risk management in energy markets*. Cambridge University Press.
- Szymanowska, M., De Roon, F., Nijman, T., & Van Den Goorbergh, R. (2014). An anatomy of commodity futures risk premia. *The Journal of Finance*, 69(1), 453–482.
- Tang, K., & Xiong, W. (2012). Index investment and the financialization of commodities. *Financial Analysts Journal*, 68(6), 54–74.
- Till, H. (2022, Winter). Commodities, crude oil, and diversified portfolios. *Global Commodities Applied Research Digest*, 7(2), 65–74.
- Till, H., & Eagleeye, J. (Eds.). (2007). *Intelligent commodity investing*. Risk Books.
- Trolle, A. B., & Schwartz, E. S. (2010, Spring). Variance risk premia in energy commodities. *The Journal of Derivatives*, 17(3), 15–32.
- Tully, S. (1981, February 9). Princeton's rich commodity scholars, *Fortune*.
- Turnbull, S. M., & Wakeman, L. M. (1991). A quick algorithm for pricing European average price options. *Journal of Financial and Quantitative Analysis*, 26(3), 377–389.
- Venkatraman, A., & Alexander, C. (2011). Closed form approximations for spread options. *Applied Mathematical Finance*, 18(5), 447–472.
- Weymar, F. H. (1965). *The dynamics of the world cocoa market*. Ph.D. Thesis, Massachusetts Institute of Technology.

- Whiteshot, C. A. (1905). *The oil-well driller: A history of the world's greatest enterprise, the oil industry*. Acme Publishing Company.
- Williams, J. C., & Wright, B. D. (1991). *Storage and commodity markets*. Cambridge University Press.
- Wilmott, P., Dewynne, J., & Howison, S. (1993). *Option pricing: Mathematical models and computation*. Oxford Financial Press.
- Working, H. (1948). Theory of the inverse carrying charge in futures markets. *Journal of Farm Economics*, 30(1), 1–28.
- Working, H. (1949). The theory of price of storage. *The American Economic Review*, 39(6), 1254–1262.
- Zühlsdorff, C. (2001). The pricing of derivatives on assets with quadratic volatility. *Applied Mathematics Finance*, 8(4), 235–262.

---

# Index

## A

- Actuarial valuation, 188, 331  
Agricultural markets, 30, 56, 64, 65, 129, 155, 308  
Airline hedging, 65, 234  
All Weather strategy, 74  
Aluminum, 308  
American options, 174  
Arbitrage  
    model, 214, 261  
    paper, 118, 120, 121  
    statistical, 5, 109, 122–124, 139, 148, 306  
    triangular correlation, 7, 281, 287–291  
    volatility, 230, 257  
Aristotle, 64, 161  
Arithmetic Brownian motion (ABM), 40, 165, 171, 238, 293, 311  
Asian options, 234  
At-the-money (ATM) options, 162, 163, 172, 183, 188–190, 194, 199, 200, 203–205, 207, 214–225, 289–291, 298, 300  
Autocorrelation, 151  
Availability, 34–37  
Average price options (APOs), 7, 233–237, 241–247, 325–326, 331

## B

- Bachelier formula, 6, 171, 222, 227, 295, 315  
Bachelier, L., 6, 162  
Backtesting, 85  
Backwardation, 15–17, 57, 58, 98, 103, 110–113, 331  
Bank of China, 52  
Barone-Adesi, G., 174  
Barrel counting, 117, 331  
Basis risk, 62  
Basis spread, 2, 126, 292, 331

## Batteries, 282, 305

- Beta hedge ratio, 62, 77, 144, 146  
Biodiesel, 308  
Black, F., 164, 169  
Black formula, 173, 222, 316  
Black-Scholes-Merton (BSM) model, 6, 161, 163, 169, 171, 193, 273  
Black volatility, 181, 192, 245, 331  
Bootstrapping, 7, 257, 261–268  
Breakout strategy, 85  
Brent, 3, 117, 118, 123–125, 130, 132, 145, 174, 236, 287, 293, 295, 306  
Bridgewater Associates, 74  
Brownian motion, 40, 162  
Bureau of Labor Statistics (BLS), 149  
Butane, 127  
Butterfly spread, 271, 272, 331

## C

- Calendar spread options (CSOs), 7, 281–286, 293, 299–301, 331  
Calibration, 7, 257, 259–279  
Call options, 162  
Call spreads, 270, 271, 274  
Canadian dollar (CAD), 143–145, 156, 157  
Capital Asset Pricing Model (CAPM), 69, 74  
Carry, 5, 83, 86, 97–101, 103–106, 110, 111, 114, 118, 128, 331  
Carry-momentum, 101, 102, 106  
Carry trade, 5, 12, 19–21, 30, 35, 36, 42, 43  
Causality, 132, 137, 139, 141, 143, 148, 305  
Chapman-Kolmogorov equation, 163  
Chicago Mercantile Exchange (CME), 4, 125, 130, 131  
Chilean peso (CLP), 145  
Coal, 127, 306  
Coefficient of instability, 163

- Cointegration, 122, 123, 139, 143–146, 148, 156, 157
- Collateral, 24, 65, 67, 170, 174, 211, 234, 255, 291, 307
- Commitments of Traders (CoT) report, 129, 331
- Commodities Corp., 84
- Commodity currencies, 143, 145, 148, 331
- Commodity Futures Trading Commission (CFTC), 131
- Commodity indices, 67, 68
- Commodity own rate of interest, 5, 14, 22, 23, 26
- Commodity terms of trade (CToT), 139–146
- Commodity trading advisors (CTAs), 5, 84, 332
- Concave backwardation, 112, 113
- Concave contango, 103, 104, 110–113
- Consumer hedging, 57, 62–64, 77, 130, 234
- Consumer Price Index (CPI), 71, 73, 149–152, 154, 157
- Contango, 15–17, 21, 26, 58, 75, 76, 98, 103, 107, 110–116, 282, 286, 332
- Convenience yield, 5, 11, 19, 22, 23, 25, 31, 32, 41, 76, 97, 98, 100, 103, 110, 117, 332
- Convex backwardation, 103, 110–112
- Convex contango, 112
- Convexity, 103, 110, 162, 172, 186, 191
- Cootner, P., 84
- Copper, 11, 145, 308
- Corn, 13, 167, 308
- Correlation
- frown, 297, 301
  - implied, 295, 297, 301
  - realized, 299
- Costless collars, 211, 212, 229, 276
- Crack options, 287, 290–292
- Crack spreads, 287, 292, 332
- Credit risks, 212, 214, 234, 255, 291
- Crisis alpha, 85
- Cross-sectional momentum, 87
- Cushing, 46, 114–117, 119, 124, 125, 127, 332
- D**
- Dalio, R., 74
- Dated Brent, 125, 236
- Delta, 169, 172, 183, 184, 189–196, 202–206, 208, 216, 218–220, 225, 226, 229, 259, 271, 279, 295, 315–317
- Delta hedging, 6, 169, 183, 187, 189–196, 202–206, 218, 259, 273, 281, 284–288, 294, 296, 305
- Denomination effect, 140, 143, 332
- Derivatives market, 1, 3, 4
- Diesel, 131, 149, 152, 154, 288, 290–292, 294
- Diffusion process, 6, 7, 161–171, 227, 228, 257–259, 276–278, 309, 328
- Digital options, 270, 271, 332
- Dirac delta function, 41, 43, 44, 268–272, 309, 310, 328
- Direct problems, 257
- Disaggregated reports, 129
- Domestic sweet oil (DSW), 125
- Dubai oil, 125
- Dump men, 20
- Dupire equation, 273–275, 327
- Dynamic systems, 1, 3, 29, 30, 38, 105, 137–139, 148, 305
- E**
- Early exercise premium, 174
- Early expiry options (EEOs), 233, 247–252, 300
- Einstein, A., 163
- Emerging markets, 138
- Emission credits, 127, 306
- Energy Information Administration (EIA), 4, 47
- Energy Policy and Conservation Act (EPCA), 16
- Energy transition, 282, 305–308
- Ethane, 127
- Ethanol, 308
- Euclidean distance, 134
- Euro (EUR), 143
- Eurobob, 126
- European options, 174
- Excess return (ER), 24, 25, 65, 66, 332
- Exchange-traded options, 174, 233
- Expandable options, 255
- Extendable options, 254
- F**
- Fading extreme positioning, 133
- Fading the crowded trade, 133, 332
- Fat tails, 167, 182, 209, 225, 226, 228, 293, 307
- Federal Reserve Economic Data (FRED), 4
- Feedback loops, 1, 3, 5, 29, 30, 35, 52, 137, 138, 305
- Fence, 211, 333
- Financialization, 5, 55, 64–68, 74, 78, 113, 114, 141, 206, 221, 223
- Finite difference methods, 171, 227
- Fisher, I., 5, 12, 97
- Fisher inflation law, 5, 11, 12, 16

- Floating storage, 21, 50  
Fokker-Planck equation, 40, 163, 273, 310, 327  
Follow the flow strategy, 132, 332  
Fourier equation, 163  
Fractionation analysis, 5, 113, 114, 116, 122, 128, 332  
Freight, 120, 121, 282, 293  
Fuel oil, 126  
Fundamental solution, 268, 310  
Futures margin requirements, 66
- G**  
Gamma, 6, 172, 187, 191–196, 199, 200, 207, 215, 218, 225, 228, 307, 316, 317, 320, 332  
Gasoil, 126, 131, 294  
Gasoline, 64, 68, 71, 89, 126–128, 137, 148–154, 292, 294  
Geometric Brownian motion (GBM), 166, 311  
Gold, 11–13, 73, 140, 167, 308  
Goldman Sachs Commodity Index (GSCI), 67, 332
- H**  
Hacienda hedge, 233–237, 332  
Harmonic mean, 64  
Hayek, F., 14  
Heat equation, 161, 171, 258, 310, 328  
Heat rate, 294  
Heaviside function, 269  
Hedging  
airline, 65, 234  
consumer, 57, 62–64, 77, 130, 234  
inflation, 76  
Mexico sovereign, 7, 210, 235–237  
pressure, 4, 5, 55–64, 76–78, 112, 332  
producer, 57, 60–67, 209–214, 234  
Hicks, J.R., 57  
High-Sulfur Fuel Oil (HSFO), 236
- I**  
Ill-posed problems, 260  
Implied Black volatility (IBV), 181, 221–225  
Implied correlation, 295, 297, 301  
Implied normal volatility (INV), 181, 222–224  
Implied volatility, 6, 161, 164, 180–186, 215–226, 239–241, 261–268, 276, 277, 332  
Impulse function, 269, 309  
Incomplete markets, 257, 267  
Inconvenience yield, 76
- Inflation, 11–14, 16, 55, 69–78, 137–139, 148–154, 308  
Inflation base effect, 150  
Inflation breakeven rate, 72  
Inflation hedging, 76  
Inflation pass-through, 71, 138, 141, 148–154  
Inflation swaps, 139, 148–154  
Inflection point, 105, 106, 116, 133, 134, 139  
Information ratio, 91  
Intercontinental Exchange (ICE), 4, 125, 130  
In-the-money (ITM) options, 162  
Inventories, 29–53, 83, 87, 98, 100, 102, 103, 114–117  
Inverse demand function, 33–36, 38  
Inverse diffusion problem, 327–328  
Inverse problem of option pricing, 7, 257–260  
Inverse problems, 258–260  
Itô's lemma, 168, 309
- J**  
Jet fuel, 64, 126, 288
- K**  
Kaldor, N., 22  
Keynes, J.M., 11, 14, 56–60  
K-factor, 236  
Kirk approximation, 296  
Kolmogorov equations, 40, 274, 310  
Kurtosis, 6, 190, 214, 276
- L**  
Law of cosines, 289  
Law of radiation of probability, 163  
Linearization methods, 227, 319  
Liquefied natural gas (LNG), 127, 145, 306  
Liquidity preference theory of money, 23  
Local volatility, 6, 161, 164, 167, 171, 172, 175, 177, 181, 185, 226–230, 237–255, 261, 263–267, 273–279, 285, 292–294, 300, 301, 309, 319, 326, 327, 332  
Lognormal distribution, 166, 167  
Louisiana Light Sweet (LLS), 236
- M**  
Macro fair-value model, 155–158  
Managed money (MM), 131  
Margrabe formula, 296, 297  
Market-implied diffusion, 7  
Market-implied probability distribution, 7, 257, 268–273, 332

Mars oil, 125  
 Maya oil, 235  
 Mean-reversion, 39, 40, 44, 45, 83, 101–104,  
   107, 109, 118, 120–123, 127, 132, 138,  
   247, 254, 293, 312, 332  
 Medium of exchange channel, 140  
 Mental models, 1, 4, 5  
 Merton, R.C., 164, 169, 238  
 Metals markets, 145, 307  
 Model arbitrage, 214, 261  
 Model of the squeeze, 5, 38–46, 305  
 Momentum, 5, 83–97, 132, 133, 138  
 Momentum smile, 95  
 Monte Carlo simulation, 168, 175, 255, 259,  
   276, 278  
 Multi-factor models, 252–255

**N**

Naphtha, 127  
 National Bureau of Economic Research  
   (NBER), 69  
 Natural gas, 68, 99, 127, 131, 148, 168, 219,  
   239, 254, 265, 294, 296, 306  
 Natural gas liquids (NGLs), 123, 127  
 Negative oil prices, 5, 49–53  
 Normal backwardation, 5, 55–60, 63–68, 74,  
   75, 85, 206, 332  
 Normal contango, 5, 55, 75–78, 113, 332  
 Normal distribution, 40, 164–167, 171, 225,  
   311  
 Normal volatility, 165  
 Norwegian krone (NOK), 145  
 Numeraire effect, 140

**O**

**Oil**  
   busts, 177  
   grades, 124, 332  
   heavy, 125  
   light, 125  
   loans, 16–18  
   own rate of interest, 15–19  
   sour, 125  
   swap, 249, 250  
   sweet, 125  
 Open interest, 86, 124, 233, 302  
 Options  
   American, 174  
   Asian, 234  
   at-the-money (ATM), 162, 163, 172,  
     183, 188–190, 194, 199, 200,  
     203–205, 207, 214–225, 289–291,  
     298, 300

average price (APOs), 7, 233–237,  
   241–247, 325–326, 331  
 calendar spread (CSOs), 7, 281–286, 293,  
   299–301, 331  
 call, 162  
 crack, 287, 290–292  
 digital, 270, 271, 332  
 early expiry (EEOs), 233, 247–252, 300  
 European, 174  
 exchange-traded, 174, 233  
 expandable, 255  
 extendable, 254  
 Greeks, 172  
 as insurance, 187–191, 197, 198  
 in-the-money (ITM), 162  
 moneyness, 172, 183, 315, 316  
 out-of-the-money (OTM), 162  
 over-the-counter (OTC), 7, 233, 234, 237,  
   247, 249, 252, 254, 255, 291  
 put, 162  
 real, 2–3, 6, 109, 118, 123, 125, 210–213,  
   282, 287, 305, 307  
 simple, 163, 172  
 spread, 2, 118, 281–302, 333  
 synthetic spread, 288, 333  
 vanilla, 233, 334  
 Organization of Arab Petroleum Exporting  
   Countries (OAPEC), 16, 70, 140  
 Organization of the Petroleum Exporting  
   Countries (OPEC), 17, 47, 100, 157,  
   177, 264, 291, 306  
 Ornstein-Uhlenbeck process, 40, 312  
 Other reportables (OTH), 131  
 Out-of-the-money (OTM) options, 162  
 Over-the-counter (OTC) options, 7, 233, 234,  
   237, 247, 249, 252, 254, 255, 291

**P**

Paper arbitrage, 118, 120, 121  
 Parametrix method, 227  
 Penultimate expiration, 282, 333  
 Perturbation method, 6, 209, 227–229, 319–322  
 Petrodollar recycling, 140  
 Petroleum Administration for Defense Districts  
   (PADDs), 47, 114  
 Pipelines, 2, 21, 23, 46, 117, 119, 123, 125,  
   127, 287, 293, 302  
 Positioning, 128–134  
 Position limits, 307  
 Power markets, 254, 294, 296, 305  
 Premium collars, 214  
 Premium retained ratio, 197, 199  
 Price elasticities of demand and supply, 29, 34,  
   36, 37

Price of storage, 30, 31  
Principle of zero expectations, 163  
Probability density, 40, 309  
Producer hedging, 57, 60–67, 209–214, 234  
Producers, merchants, processors, and users (PMPUs), 129  
Prompt futures, 2, 333  
Propane, 127  
Put-call parity, 173  
Put option, 162

## Q

Quadratic mean, 233, 238–241  
Quadratic normal (QN) model, 6, 228, 229, 276, 277, 293, 319–322  
Quadratic utility function, 60  
Quantamentals, 5, 6, 109, 333

## R

Ratio call spread, 207  
Reaction function, 5, 83, 105–107, 116, 133, 134, 139, 196, 333  
Realized correlation, 299  
Realized volatility, 6, 161, 164, 175–181, 333  
Real options, 2–3, 6, 109, 118, 123, 125, 210–213, 282, 287, 305, 307  
Real rate, 11, 16  
Rebalancing effect, 68  
Recessions, 55, 69–71, 140  
Reduced-form models, 41, 293  
Refined products, 3, 62, 77, 99, 125–128, 219, 239, 292, 294  
Refineries, 3, 23, 46, 51, 112, 119, 123, 287, 288, 290, 294  
Reformulated gasoline blendstock for oxygenate blending (RBOB), 89, 126, 131, 149–154  
Regularization, 260  
Relative vega trading, 210  
Reverse carry trade, 21  
Risk aversion channel, 142  
Risk aversion coefficient, 60, 62, 64, 76  
Risk-neutral pricing, 35, 164, 170, 250  
Risk-neutral probabilities, 43, 268, 269, 311, 327  
Risk parity, 55, 69–76, 105, 148, 333  
Risk premium, 55, 57, 59, 63, 64, 75, 77, 78, 86  
Rockefeller, J.D., 3  
Roll return, 24–26, 66, 75, 333  
Roll yield, 5, 11, 23, 25, 26, 55, 58, 68, 75, 76, 97, 110, 132, 286

## S

Safe haven channel, 142  
Samuelson effect, 185, 240  
Samuelson, P., 84  
Scholes, M., 164, 169  
Seasonality, 88, 89, 99, 111, 117, 151, 153, 239, 254, 265  
Sentiment index, 133  
Shale oil, 47, 114, 116, 119, 120, 123, 125, 140, 177, 206, 210, 211, 221, 223, 236, 284  
Shanghai International Energy Exchange (INE), 124  
Signal blending, 5, 101, 333  
Signal transformation function, 105  
Silver, 11, 12, 308  
Simple options, 163, 172  
Skew correction function, 227, 228, 319, 321  
Skew delta, 215–221, 333  
Skewness, 6, 86, 128, 179, 180, 183, 187, 190, 209, 212–214, 225, 226, 228, 229, 276  
Spot prices, 1  
Spot return, 24–26, 66, 68, 333  
Spread options, 2, 118, 281–302, 333  
Squeeze, 5, 29, 38–46, 48, 52, 116, 117, 286, 302, 333  
Sraffa, P., 14  
Stationarity, 47, 48, 73, 122, 156, 158  
Statistical arbitrage, 5, 109, 122–124, 139, 148, 306  
Sticky moneyness, 6, 209, 217–220, 333  
Sticky strike, 6, 209, 219, 220, 333  
Stochastic dynamic programming, 35  
Stochastic volatility, 279  
Stock-out, 33–36, 45, 333  
Storage  
    boundaries, 29, 32–38, 43, 50, 115–117  
    capacity, 33, 35–37, 51, 76, 115–116  
    capacity utilization, 48, 115  
    costs, 2, 21, 22, 31, 35, 76, 282  
    floating, 21, 50  
    optimization, 29  
    price of, 30, 31  
    synthetic, 281, 283, 286, 333  
    tanks, 21, 123, 125  
    theory of, 5, 29–37, 83, 87, 305  
    virtual, 7, 281, 282, 286, 287, 305  
Straddles, 184, 188–190, 214, 218, 229, 276  
Strangles, 213–214, 229, 276  
Super-backwardation, 112  
Super-contango, 112  
Swap Data Repositories (SDRs), 148, 333  
Swap dealers (SDs), 129  
Swaplets, 249–251, 253, 267

Swaption, 7, 233, 249–254, 266, 267, 333  
 Synthetic spread options, 288, 333  
 Synthetic storage, 281, 283, 286, 333

**T**

Tank-tops, 33, 45, 334  
 Target redemption swap, 255  
 Taylor formula, 168, 309  
 Thales of Miletus, 161  
 Theory of hedging pressure, 4, 5, 55–64, 76–78, 112, 332  
 Theory of storage, 5, 29–37, 83, 87, 305  
 Theta, 172, 193, 283, 286, 287, 316, 317  
 Three-way collars, 212–214, 333  
 Time spreads, 109–117  
 Trading at settlement (TAS) contract, 52, 95, 202, 283, 334  
 Transaction costs, 95, 96, 145, 191, 200–204  
 Trapped option value, 174  
 Treasury Inflation-Protected Securities (TIPS), 72, 148  
 Trend following, 87  
 Triangular correlation arbitrage, 7, 281, 287–291  
 Two-factor models, 42, 253, 267

**U**

Ultra-low-sulfur diesel (ULSD), 126  
 Uncovered interest rate parity, 97  
 US Department of Energy (DOE), 47  
 US dollar (USD), 5, 12, 13, 137–146, 156, 157  
 US Strategic Petroleum Reserve (SPR), 5, 16–18, 21  
 US terms of trade, 140

**V**

Value risk premium, 83, 101, 102  
 Vanilla options, 233, 334  
 Vector autoregressive models (VAR), 71  
 Vega, 6, 172, 181, 208–218, 229, 316, 317  
 Virtual rate of interest in commodities, 14  
 Virtual storage, 7, 281, 282, 286, 287, 305  
 VIX index, 156  
 Volatility  
     arbitrage, 230, 257  
     Black, 181, 192, 245, 331

exponential, 239, 246–247, 251, 254, 266  
 ghost, 176  
 heuristics, 226  
 implied, 6, 161, 164, 180–186, 215–226, 239–241, 261–268, 276, 277, 332  
 implied Black (IBV), 181, 221–225  
 implied normal (INV), 181, 222–224  
 local, 6, 161, 164, 167, 171, 172, 175, 177, 181, 185, 226–230, 237–255, 261, 263–267, 273–279, 285, 292–294, 300, 301, 309, 319, 326, 327, 332  
 matrix, 264  
 normal, 165  
 realized, 6, 161, 164, 175–181, 333  
 risk premium (VRP), 6, 187, 196–208, 286, 307, 334  
 risk premium (VRP) smile, 6, 187, 198  
 risk premium (VRP) term structure, 187, 199  
 skew, 182, 334  
 smile, 6, 161, 182, 183, 209, 214–221, 226, 276–278, 334  
 stochastic, 279  
 targeting, 105, 334  
 term structure, 7, 185, 237–241, 261, 334

**W**

Wealth effect, 143  
 Well-posed problems, 258  
 Western Canadian Select (WCS), 125, 127, 145  
 West Texas Intermediate (WTI), 3, 18, 19, 26, 46, 47, 49–52, 112–115, 117–121, 124, 125, 127, 130, 132  
 West Texas Sour (WTS), 236  
 Weymar, H., 84  
 Whaley, R.E., 174  
 Working, H., 30  
 WTI-Brent accordion, 5, 120  
 WTI Houston, 125, 236  
 WTI Midland, 125

**X**

XOP ETF, 146, 147, 157

**Y**

YuanYouBao, 52