Limits to Arbitrage: An introduction to Behavioral Finance and a Literature Review
Miguel Herschberg

ABSTRACT

This paper is a survey of the developments in the literature of the Limits to Arbitrage. We investigate why investors, even if they know that an asset is not priced correctly, may not be able to profit from an arbitrage opportunity.
1. Definition

Behavioral Finance is the study of the way in which psychology influences the behavior of market practitioners, both at the individual and group level, and the subsequent effect on markets. (Sewell (2010) [31]).

According to Thaler and Barberis (2002) [36], behavioral finance has two building blocks: limits to arbitrage and psychology. Limits to arbitrage seek to explain the existence of arbitrage opportunities which do not quickly disappear. It is associated with arbitrageurs coexisting with not-fully rational investors in the market and themselves not being able to profit from market dislocations.

Understanding the existent of arbitrage opportunities, although theoretically counterintuitive, is not enough to make sharp predictions. Behavioral finance researchers often need to specify the form of the agents’ irrationality. This is related to how they misapply Bayes law or deviate from the Subjective Expected Utility theory. In order to specify the type of irrationality, researchers have turned to experimental evidence compiled by cognitive psychologists on the biases that arise when people form beliefs, and on the people’s preferences, or on how they make decisions, given their beliefs (Thaler and Barberis (2002) [36]).

In this literature review, we will concentrate on the extensive literature on the Limits to Arbitrage. The reader can refer to Camerer (1995) [5], Rabin (1998) [28], Kahneman and Tversky (1982) [21], Kahneman and Tversky (2000) [22] and Gilovitch, Griffin and Kahneman (2002) [18] for an excellent and detailed treatment of the psychology behind people’s irrationality types in behavioral finance.

2. Introduction

It has its origins in the 1970s, a time when academics turned to social psychology, a field neglected by economists, to provide explanations to puzzles that they could not answer. Ever since academics structured Economics as a science itself, they looked after Physics and Mathematics as model sciences upon which it should look alike. The mathematical tradition continued through the 1960s but a revolution led by Professors Tversky, Kahneman, Thaler, Barberis and Shleifer started to foster. They realized that mathematics alone could not explain human interactions and began mixing psychology with economics.

According to Barberis and Thaler, two of the creators of the field, explain that Behavioral Finance has two building blocks: limits to arbitrage, which argues that arbitrageurs may not be able to profit from market dislocations caused by less or not rational traders, and psychology, which catalogues all the possible kinds of deviations that we may see in the financial markets (Thaler and Barberis (2002) [36]).
Modern financial economics relies on the notion of efficient markets which assumes that individuals act in a rational way. In particular, the whole field is based on the assumption that the “representative agent” is rational, which does not imply that all agents are rational. Irrational investors can coexist with rational ones as long as the “marginal investor”, who is defined as the investor who is making the specific decision at hand is rational (See Thaler (1999) [35]).

The traditional financial economic theory is based on Professor Fama’s work. His contribution to Economics is immense, but he is best known for demonstrating that stock prices are close to a random walk (Fama (1965) [15]) and for developing the Efficient Market Hypothesis (EMH hereafter) (Fama (1970) [16]).

Fama’s definition of market efficiency is logically intuitive: it means that asset prices in financial markets fully reflect all the available information (Fama (1970) [16]) and that there are no trading strategies that produce positive, expected, risk adjusted excess returns (Dothan (2008) [13]).

A valid question that we ought to ask ourselves is if there is a connection between prices that fully reflect all available information and the absence of profitable trading strategies that produce positive risk adjusted returns. Some researchers define market efficiency as “circumstances where prices fully reflect all available information” (Lo (1974) [23] and Chordia et al. (2007) [8]) whereas others define it as the non-existence of arbitrage opportunities (Malakiel (2003) [24]). In fact, following the formal proof given in Dothan [13] which we will present later in full generality, all these definitions are equivalent statements. Formally, the Markovian structure of asset prices captures the property that they reflect all the available information and together with the fact that there are no admissible trading strategies which produce excess returns, it follows that the EMH implies that in an efficient market asset prices behave as a martingale.

However intuitively appealing the Efficient Market Hypothesis appears to be, significant empirical evidence has appeared over the years. The weakest assumption of the theory of efficient markets is that it assumes that individual’s are rational. In detail, the EMH does not assume that all investors are ration, but it does assume that markets are rational in the sense that markets make unbiased forecast for the future (for example, in this framework financial bubbles could not exist).

In a market with not-fully rational and rational agents, rational agents will prevent not-fully-rational investors from influencing security prices by trading mispriced securities through a process called arbitrage. Therefore in an efficient market, there should be no arbitrage opportunities because competition will drive prices to their correct values. However, if non-rational investors predominate in the market, it does not follow that prices in the financial markets will fully reflect all the available information and this is equivalent to arguing that there will be arbitrage opportunities.
In the traditional finance paradigm, arbitrage should be riskless and arbitrage opportunities must not exist. However, researchers have found strong evidence to assert the opposite. Arbitrage is generally risky and limited. In fact, there are situations where arbitrage opportunities exist but do not quickly disappear. This is known in the literature as limits to arbitrage and the idea is usually credited to Professors Shleifer and Vishny (Shleifer and Vishny (1997) [33]). Situations in which there is a limit to arbitrage fail to be explained by classical finance theory and can be better understood using behaviour and social psychology.

Behavioural finance emerged as a controversial field to provide an explanation for those puzzles left out from the traditional financial economic theory. Researchers realized that some financial phenomena could be better understood using models where agents are not fully rational. Its purpose has changed over the last fifty years. At the beginning, there was a possibility to argue that it was a reaction to the Efficient Market Hypothesis and that its goal was to study financial economics in situation where humans were not fully rational. Such a definition is now obsolete, as the empirical evidence against the theory of efficient market and theories where individuals are fully rational has been well documented.

The death of Behavioral Finance was predicted in 1999 by one of its founding fathers (Thaler (1999) [35]). The author argues that Behavioral finance, once a controversial revolution against established financial economics theory, has lost its purpose. History has shown that not only Behavioral Finance did not disappear in the first decade of the new millennium, but has found new applications in the financial markets. After the subprime mortgage crisis of 2008, Behavioral Finance and Limits to Arbitrage in particular, has the potential to deliver a more useful method of thought to evaluate the different regulations of the financial sphere.

Among the solutions that the theory of Limits to Arbitrage has provided over the years include: the equity premium puzzle which refers to the empirical fact that stocks have out-performed bonds over the last century (Benartzi and Thaler (1995) [4], Sewell (2005) [31], Thaler and Barberis (2002) [36]), the conservatism principle of share prices, which states that earning reflect bad news more quickly than good news (Basu (1997) [3]), the tendency of investors to sell winning investments too soon and hold investments for too long (Odean (1998) [25], overconfidence of investors (Daniel, Hirshleifer and Subrahmanyam (1998) [9], Camerer and Lovallo (1999) [6], herding behaviour in the financial markets (Wermers (1999) [37]) among others, and is nowadays yielding insight into the effects of social networks (such as the number of Facebook Likes and the number of good/bad tweets) on stock prices of publicly traded companies.

This paper is a reviews the academic literature of the application of Behavioural Finance to Limits to Arbitrage. To begin with, a definition of rationality needs to be developed, as traditional finance relies on it. Then this paper will explain the relationship between the Efficient Market Hypothesis and the No Arbitrage Condition. This will show how relaxing the rationality assumption could lead to the existence of persistent arbitrage opportunities. After this, motivating examples are presented.
3. Rationality

In this section, we present different definitions of what constitutes a rational investor. Professors Barberis and Thaler (Thaler and Barberis (2002) [36]) explain that a agent is rational if two this occur:

First

When individuals receive new information, they update their beliefs correctly, in the manner described by.

Second

Given their beliefs, agents make choices that are normatively acceptable, in the sense that they are consistent with Savage’s notion of Subjective Expected Utility (SEU).

A less technical definition of rationality is given in Thaler(1999) [35]. In this paper the author explains that traditional finance assumes that the “representative agent” in an economy is rational in two ways: The “representative agent”:

First

makes decisions according to the axioms of expected utility theory.

Second

makes unbiased forecasts about the future.

The fact that individuals make unbiased forecasts about the future means that agents’ beliefs are correct. Furthermore, by making decisions according to the axioms of expected utility theory, the rational investors in the economy apply Bayes’s law correctly. Hence, we can see that both definitions are equivalent.

3.1. Subjective Expected Utility

Intuitively, Expected Utility was originally developed to be used with specified or “objective” probabilities (Hammond, Barbera and Seidl (1998) [20]). However, not all uncertainty can be described by a specified probability distribution. The work led by Ramsey (1926) [27] and Bruno de Finetti (1937, 1949) [10], [11] which was further expanded by Savage (1954) [30], Anscombe and Aumann (1963) [1], Arrow (1951) [2], Chernoff (1954) [7], Suppes (1956) [34], Ellsberg (1961), [14] and Rubin (1949) [29]; demonstrated how, under certain assumptions, “subjective” probabilities could still be inferred from behavior in the face of such uncertainty. Properly defining the structure of the utility function is a daunting task and is outside of the scope of this paper. However a perfect explanation can be found in Hammond, Barbera and Seidl (1998) [20].
\[ U(f) = \int_{\Delta S} u \cdot f \, d\pi \]  

(1)

where \( \Pi \in \Delta S \) is a subjective probability over states of the world and 
\( u : R^+ \rightarrow R \) is a cardinal utility index over tokens.

A more formal definition can be given by following the derivations in Hammond, Barbera and Seidl (1998) [20]. Let \( Y \) be a fixed domain of possible consequences and \( S \) be a fixed finite state of possible states of the world. We do not specify a probability distribution over \( S \). An act is a mapping \( a : S \rightarrow Y \) specifying what consequence results in each possible state of the world. The “consequences” are known in the literature as Contingent Consequence Functions or CCFs. Each CCF can be considered as a list \( y^S = (y_s)_{s \in S} \) of contingent consequences in the Cartesian product space \( Y^S := \Pi_{s \in S} Y_s \).

The subjective expected utility framework requires that there exist subjective probabilities \( p_s \), non negative by definition, of different states \( s \in S \) satisfying \( \sum_{s \in S} p_s v(y_s) \). There must also be a von Neumann -Morgenstern utility function \( v : Y \rightarrow \mathbb{R} \), as in the Objective Expected Utility theory. Moreover, we suppose that the agent will choose a CCF \( y^S \) from the relevant feasible set in order to maximize the subjective expected utility function:

\[ U^S(y^S) := \sum_{s \in S} p_s v(y_s) \]  

(2)

Note that this equation (2) is the discrete version of the previous equation (1). The key point in this utility function is that the probabilities \( p_s \) are not objectively specified, but are revealed by the agent’s behavior. It implies that the agents in the market have enough information about the structure of the economy to be able to figure out the correct distribution of the variables in the model.

Behavioral finance departs from Traditional Finance by relaxing these assumptions. If we relax this assumption, investors update their beliefs correctly, applying Bayes’ theorem, but they lack the information required to know the actual distribution variables are drawn from (Thaler and Barberis (2002) [36]). An example will be when we don’t know the growth of an asset, but we learn it from the available data. The problems that arise when we relax the consistent beliefs assumption have been extensively studied. It is commonly referred to as the literature on bounded variation.
4. Efficient Markets and the Existence of Arbitrage Opportunities

Most papers in the literature assume that in an efficient market there will be no arbitrage opportunities because less-than-rational investors coexist with rational investors. Analogously, researchers assume that in an inefficient market, there will be arbitrage opportunities that do not quickly disappear.

The relationship between Efficient Markets and the existence of arbitrage opportunities seems intuitively correct. However, proving that such link exists is not trivial. In this section, we discuss in detail the proof given in [13]. Dothan’s paper starts by explaining the general set up of the model where he defines the characteristics of the market. He then explains why the markovian model of stock prices is useful to capture that prices reflect all available information. Finally, he provides a proof of the non-existence of profitable risk-adjusted trading opportunities.


Essentially a Cox-Ross-Rubinstein model, it includes a probability space \((\Omega, \mathcal{F}, \mathbb{P})\), where \(\mathcal{F} = (\mathcal{F}_n)_{0 \leq t \leq T}\) is a filtration and \(\mathbb{P}\) a probability measure. Additionally, the market has a \(d\)-dimensional process \(X^0\) of prices and a set \(\Theta\) of \(d\)-dimensional trading strategies.

The filtration \((\mathcal{F}_n)_{0 \leq t \leq T}\) is an increasing set of \(\sigma\)-fields such that \((\mathcal{F}_t)\) contains all the decidable events at time \(t\). We can think of it as all the information available at time \(t\). It is increasing because we do not “loose” or forget information at time passes. The process \(X^0\) is assumed to be adapted to the filtration \(\mathcal{F}\), this means that one cannot see into the future, i.e. for every \(n\), the realisation of \(X^n\) is known at time \(n\).

Furthermore, to ensure the gains from trade, modelled by \(\int_0^t \Theta^s dX_s\) exists, we need to assume that the process \(X\) is a \(d\)-dimensional semimartingale. Trading strategies \(\theta\) are restricted to be predicable, \(X\)-integrable, selffinancing and tame.

A self-financing strategy means the purchase of securities must be financed by the sale of existing ones. It rules out injections of capital to purchase assets, in other words, all the cash required to make the strategy work must be internally produced. Predicable means that its value is knowable at a prior time. In my opinion, this is obvious, because the trading strategies must be left continuous, and all left continuous, continuous time processes are predictable. A tame strategy is such that the gain process \(\int_0^t \Theta^s dX_s\) is uniformly bounded from below. This assumption is important because it excludes trading strategies that guarantee an ultimate gain at the cost of unbounded intermediate losses which are not applicable in the real world but theoretically could constitute arbitrage opportunities.
4.2 The Markovian Property of Asset Prices

Professor Dothan explains very clearly why prices must be markovian. If the information $\mathcal{F}_t$ is fully reflected in prices $X_t$, the conditional probability of any relevant future event is the same whether conditioned on the information $\mathcal{F}_t$ or on current prices $X_t$ which implies that the price process $X_t$ is Markovian.

It is worth to note if prices are markovian then this means that new information arrives at $t$ and not an instant after $t$. Mathematically, this means that $\mathcal{F}$ is right continuous. We will use this to prove that arbitrage strategies do not exist.

Furthermore, Dothan argues that the assumption that prices fully reflect all available information is not sufficient to infer that there are no arbitrage opportunities. To deliver the result we need a “no free lunch with vanishing risk” condition. Dothan defines a free lunch with vanishing risk as a self-financing trading strategy that can be approximated by a sequence of tame self-financing trading strategies that converge to an arbitrage strategy.

4.3 No Arbitrage Opportunities

**Theorem 1** If discounted prices $X$ have the Markov property and satisfy the no free lunch with vanishing risk condition, then the discounted value process is a supermartingale.

The proof given by Prof. Dothan [13] is mathematically complex and beyond the scope of this paper. However, he explains the intuition behind this theorem: “The supermartingale property of the discounted value process means that, after adjustment for risk and the time value of money, conditional expected returns are nonpositive. Therefore, after adjustment for risk and the time value of money, it is possible, on average, to break even or have a trading loss, but not a trading gain”.

Hence if prices fully reflect all the available information and satisfy the no free lunch with vanishing risk condition, then there are no admissible strategies that produce positive, expected, risk-adjusted returns.

5. The Theory of The Limits to Arbitrage

The question that remains to be answered is why arbitrage opportunities do not quickly disappear even if investors know how to exploit them?. The idea is that strategies designed to correct the mispricing can be both risky and costly. Professors Thaler and Barberis (Thaler and Barberis [36]) have identified four risks and costs that have been identified.
5.1. Fundamental Risk

Fundamental risk refers to the risk that new bad information arrives to the market after you purchased the security. Theoretically this risk could be perfectly hedged by buying a closely related product. Unfortunately substitute securities are rarely perfect, making it impossible to remove all the fundamental risk.

As an example, suppose you own two shares, one from A, the other one from B. Furthermore assume that both are closely related, share the same industry, and that you buy A and sell B. After the combination of buying and selling, most fundamental risk is removed. Intuitively, we can think of being neutral to the industry where A and B operate. However, the investor is still vulnerable to bad news about the industry as a whole. To sum up, fundamental risk always persists even though there are ways of hedging the portfolio against it.

5.2. Noise Trader Risk

Originally introduced by De Long et al. (1990) [12] and by Shleifer and Vishny (1997) [33], noise trader risk refers to the risk that the mispricing worsens in the short run because there is a possibility that pessimistic traders become even more pessimistic about the future.

Noise trader risk is very important because of its link to other agency problems. It can force people, such as institutional investors and hedge fund managers, to liquidate their positions early, bringing them unwanted and unnecessary steep losses. Hence if investors lack the knowledge to evaluate the managers strategy, they may simply evaluate him based on his returns. If a mispricing worsens in the short run and generates negative returns, investors made decide to withdraw their funds and force him to liquidate his business.

Professors Thaler and Barberis argue that fear of such premature liquidation makes professional managers be less aggressive in combating the mispricing in the first place. Although noise risk most commonly occurs when shorting assets, it may also be present in the collateral of a repurchase agreement (repo) or if the original owner of the stock demands his stock before our strategy corrected the mispricing.

Furthermore Professors Shleiffer and Summers (Shleiffer and Summers (1990) [32]) call this source of risk: “resale risk” because it comes from the unpredictability of future resale prices. This risk will particularly affect investors and professional managers time horizon. In particular, if resale risk is high, investors will have a comparatively shorter time horizon because investors fear having to liquidate because of exogenous reasons.
5.3. Implementation Costs

Implementation costs are well known to any investor. They refer to transaction costs such as commissions, bid-ask spread, premium costs for hot collateral in repurchase agreements, increased commissions for shorting securities among others. Other than monetary costs, there can also be legal constraints and accounting issues.

This category also includes the cost of finding and learning about a mispricing, as well as the cost of the resources needed to exploit it. In particular, finding mispricings can be expensive and time consuming. Learning about them will certainly require highly specialized labor. Finally exploiting mispricing furthermore requires state of the art technology and expensive IT systems that can trade at the high-frequency speed.

6. Evidence of Limits to Arbitrage

As we saw before, there is scope for arbitrage opportunities to exist. In the end, closing price differences in the market can be both costly and risky. We should not forget that behavioral finance started as a reaction to classical financial theory because there were “anomalies” that could not be explained by it.

The relevant anomalies for the limits to arbitrage theory are those concerning the relative prices of assets with closely related payoffs (Gromb and Vayanos (2010) [19]). For example:

First “Siamese twin” stocks, with claims to almost identical dividend streams can trade at significantly different prices.

Second Stocks of a parent and a subsidiary company can trade at prices under which the remainder of the parents company’s assets has negative value.

Third newly issued “on-the-run” bonds can trade at significantly higher prices than older “off-the-run” bonds with almost identical payoffs.

In this section we will present two classic examples of the evidence of the theory of limits to arbitrage. We follow closely the paper by Thaler and Barberis (2002) [36]. Also, interested readers should look into Prast (2004) [26] for a detailed behavioral explanation for six finance puzzles.

6.1. Twin Shares

This example is the most classic example in the literature. We follow closely both the exposition in Thaler and Barberis (2002) [36] which is based on the paper Froot and Dabora (1999) [17].
In 1907, two completely independent companies: Royal Dutch and Shell Transport, agreed to merge their interests on a 60:40 basis, while remaining separate companies. Share of Royal Dutch are a claim to 60% of the total cash flow of the two companies, while Shell is a claim to the remaining 40%.

Since 60% of the new company is comprised by Royal Dutch and 40% by Shell Transport, if prices equal fundamental value, the market value of Royal Dutch equity should always be 1.5 times the market value of Shell equity. However it is not. In the past, Royal Dutch has traded 35% underpriced relative to parity, and sometimes 15% overpriced.

This mispricing is a clear evidence of limits to arbitrage. In this case the shares are good substitutes for each other so fundamental risk is nearly perfectly hedged. In addition there are no major implementation costs because buying and shorting shares of either company is relatively easy.

Professors Thaler and Barberis argue that the one risk that remains is noise trader risk. Whatever investor sentiment is causing one share to be undervalued relative to the other could also cause that share to become even more undervalued in the short term. In fact, if you bought Royal Dutch shares in March 1983 which traded at a price 10% undervalued, you would have seen the stock drop even further in the following six months. This shows that arbitrage in this case was limited because clearly prices were not right, but on the other hand there were no profits to be made for arbitrageurs.

6.2. Index Effects

Is it normal that a stock jumps in price by an average of 3.5% when it is added to an index? Yahoo! for example jumped even further, 24%.

Shares should not change in price upon inclusion in an index. After all, the fundamental values remained intact. Professors Thaler and Barberis argue that this is not only a mispricing but also an evidence of limited arbitrage because if an arbitrageur wished to exploit this mispricing he would have to short a good substitute for the stock being included in the index, and this entails considerable risk. Furthermore, they note that there is also noise trader risk because the share price can continue to increase in the short run, preventing the arbitrageur from closing the mispricing.

On the other hand, in some markets, institutional investors like mutual funds, pension funds and insurance companies for example, can only buy shares of companies that are part of an index like the S&P 500. There is a possibility that share price increase because when they start trading as part of an index, there is more demand.
7. Conclusion

As we mentioned in the introduction, limits to arbitrage is viewed as one of the two building blocks in the field of behavioral finance. This paper reviewed the major developments in this area from its birth around 1970. This paper started by explaining that behavioral finance started as a reaction to the Efficient Markets Hypothesis. We then explained in detail the connection between efficient markets which fully reflect all available information, rationality and the non-existence of profitable risk adjusted opportunities.

Behavioral finance economists who were working on the theory of limits to arbitrage realized that in fact, individuals were not as rational as others previously thought. It was argued that either individuals did not apply Bayes law correctly or they did not behave according to the Subjective Expectations Utility (SEU) theory.

Regardless of the form of investors irrationality, researchers found that it was a real possibility that the market through the process of arbitrage could not prevent mispricings from closing immediately. This meant that the share prices did not reflect all the available information as the Efficient Market Theory had predicted earlier.

Intuitively the idea behind the existence of arbitrage opportunities is that they exist because arbitrage is both risky and costly. Investors may be subject to fundamental risk, noise trader risk, resale risk and implementation costs. Therefore researchers explained that even if investors know that prices are not correct, they may not be able to exploit the mispricing to drive prices to their fundamental values. There is also a strong evidence that arbitrage is limited. In this paper, we explained the two most important cases of limits to arbitrage: the merger between Royal Dutch and Shell Transport, and the index effect puzzle.
References


