Mean-Variance Portfolio Rebalancing with Transaction Costs*

Dybvig Philip, H.† Luca Pezzo‡

November 11, 2019

Abstract

Transaction costs can make it unprofitable to rebalance all the way to the ideal portfolio. A single-period mean-variance theory allows a full solution for many securities with possibly correlated returns, and makes the economics of trading with transaction costs very clear, informing us about theory and practice. As in continuous time models, there is a non-trading region within which trading does not pay. With only variable costs, any trading is to the boundary of the non-trading region, while fixed or mixed costs induce trading to the interior. The exact solution for an arbitrary number of assets and covariance structure is easy to compute (exactly by hand for a small problem). The results complement nicely the continuous-time models for special cases or with approximate numerical solutions. One application shows how to improve on traditional symmetric futures overlay strategies.

Keywords: Transaction Costs, Mean-Variance, Optimization, Asset Allocation

*We are grateful for support from NISA Investment Managers LLC and for helpful comments from Dave Brown, Hong Liu, Bill Marshall, and seminar participants at Washington University. Any errors or opinions are our own.

†Olin Business School, Washington University in St. Louis, +1 314 935-9444, dybvig@wustl.edu.

‡University of New Orleans, department of Economics and Finance, +1 314 422 6083, lpezzo@uno.edu.

1 Introduction

Optimal portfolio rebalancing given transaction costs is a complex problem. Even with only two assets, solving for the optimal strategy in a continuous-time model involves either a primal free boundary problem (see, for example, Davis and Norman (1990), Dumas and Luciano (1991), Liu and Loewenstein (2004), Shreve and Soner (1994) and Taksar et al. (1988)) or its dual formulation (see Goodman and Ostrov (2010) and Schachermayer (2017)). When there are more securities or time is discrete, models have been solved only in the extreme case of uncorrelated returns and constant absolute risk aversion (Liu (2004)) or with numerical or heuristic approximations (Leland (2000), Balduzzi and Lynch (1999), Balduzzi and Lynch (2000), Donohue and Yip (2003), Han (2005), Muthuraman and Kumar (2006), Irle and Prelle (2008), Lynch and Tan (2009), Myers (2009) or Buss and Dumas (2017)). In this paper we study the single-period rebalancing problem in a mean-variance framework that permits exact solutions, providing many interesting insights.

Mean-variance analysis was originated by Markowitz (1952), who described the basic formulations and the quadratic programming tools used to solve them. The theory was further described by Tobin (1958), who focused on macroeconomic implications of the theory. Early discussions of transaction costs often focused on the intuition that small investors who face high costs will choose a smaller and less diversified portfolio than will a large investor with smaller costs. This intuition has been formalized by a constraint on the number of securities in the portfolio (Jacob (1974)), a fixed cost for each security included in the portfolio (Brennan (1975), Goldstein (1979), and Mayshar (1981) or a study of benefits of adding securities without modeling the costs (Mao (1970)). Unfortunately, these assumptions tend to produce a somewhat messy combinatoric problem looking at all possible subsets to include, and their static perspective does not seem suited to questions about rebalancing. The current analysis differs in two important ways from the traditional mean-variance literature on transaction costs. First, we focus on rebalancing from a starting portfolio, not necessarily building a portfolio from scratch. Second, we solve models with variable costs as well fixed costs [1]

A complete solution to a portfolio optimization problem is the rule for rebalancing from initial portfolios to optimal new portfolios. Absent costs and other frictions such as taxes, the optimal portolio after rebalancing depends on the initial portfolio only through the initial wealth, since it does not cost anything to move among positions having the same value. When trading involves transaction costs, we may not trade to this ideal portfolio that would be optimal absent costs. In particular, there exists a set of initial positions which are too close to the ideal portfolio to justify any trade, so that it is optimal not to trade at all. Locally, the cost of moving to the ideal portfolio (constant and/or linear) is a larger order of magnitude compared to the mean-variance loss from having the wrong allocation. The set of portfolios from which it is not optimal to trade is referred to as the non-trading region.

The cost of trading can include a lot of pieces: brokerage fees, bid-ask spread, price pressure (getting a less favorable price the more you trade), time and effort to compute and submit the trade, and the cost of studying the securities in your portfolio. The overall cost of trading is probably complex in practice, for example, the brokerage fee may be constant for small trades with a positive marginal cost kicking in at some level. In this paper we abstract from some of these complexities by modeling costs as either fixed, proportional or both. In particular, we view the variable component of costs as proportional to the size of the trade but possibly different for different securities, and we allow separate fixed components for different subsets, allowing for an overall fixed cost, a per-security fixed cost, or a fixed cost for entering a new exchange. These assumptions are general enough to be useful and simple enough so we can get some sharp results.

In models with only proportional costs, any trading goes only as far as the boundary of the non-trading region, since trading further would incur additional costs that are not justified.[2]

[1] Pogue (1970) formulated models with variable costs and other institutional features, but did not provide any analysis of their solution.

[2] Masters (2003) claims a mean-variance-style analysis with a single risky security and variable trading costs in which it is claimed that it is not optimal to trade to the boundary of the non-trading region. However, this is because the paper computed the non-trading region incorrectly as the set of portfolios from which it would be worth trading to the ideal point that would be chosen absent costs. In fact, if there are only variable costs, there are portfolios from which is not worth trading to the ideal point but from which it is

Furthermore, the cost of trading is additive (if all trades are in the same direction), or less than additive (if the second trade reverses the first trade in some securities). If a candidate trade does not take us to the non-trading region, we could add on the additional trade we would make from that point and be better off. Conversely, if a candidate trade takes us on a line going beyond the boundary of the nontrading region, is better to trade along the line to the boundary because the part of the trade beyond the boundary is not justified. These arguments do not work for fixed costs, because they rely on the cost of a sequence of trades to be no more than additive.

In the models with fixed costs, any trade moves to inside the non-trading region if it is optimal to trade at all: this is because fixed costs, once triggered, become sunk costs. With only an overall fixed cost, any nonzero trade moves to an ideal portfolio that would be held absent costs. This ideal portfolio is in the interior of the non-trading region because the value of trading from nearby is too small to cover the fixed cost. With security-specific fixed costs, any trade will take us to the interior of the non-trading region, and different starting portfolios could cause us to trade to different target portfolios which depend on the subset of securities being traded.

In models featuring both proportional costs and an overall fixed cost, if the starting portfolio is far enough from the ideal portfolio, the optimal trade is exactly the same as the optimal trade with variable costs only. This is because it is obvious it will be optimal to trade something and incur the fixed cost, and therefore we can view the fixed cost as sunk. The argument is slightly more complicated in the presence of asset (or subset) specific fixed costs. Consider an example with independent returns and asset-specific fixed costs for each of two assets. If we are are far away from the ideal position only in one security, the fixed cost for this security can be considered as sunk but the same is not true for the other security. This results in an optimal trade only involving the most displaced asset. However, in the absence of asset-specific fixed costs it would have been optimal to trade in both assets.

The analysis in this article can accommodate multiple risky assets, trading of individual

worth making a smaller trade.

securities or bundles or pairs, and trading futures or swaps as well as stocks. In particular, while applying our model to trading in futures and its underlying, we show how to improve over traditional futures overlay strategies. Perhaps less surprisingly, it can also be shown that inexpensive trading in bundles of assets is beneficial and the non-trading region gets squeezed towards the ideal allocation along the bundles trading directions.

Given proportional costs only, our model is a quadratic program which is solved very quickly by standard software. Adding fixed costs, we have the best solution of a finite set of quadratic programs and the solution is still very fast if there are not too many different fixed costs. The main requirements for the programs to be well-defined with unique solution are (i) the covariance matrix of returns is positive definite, and (ii) the coefficient of risk aversion plus the parameter penalizing tracking error is positive. In the traditional case, there is no penalty for tracking error and the second condition is that the risk aversion parameter is positive.

Although our model is myopic, Maurer et al. (2018) show it is useful for dynamic trading in FX markets: taking into account costs while optimizing over FX carry trades leads to an economically large and statistically significant improvement in the out-of-sample performance with Sharpe ratio increments of at least 25%. They also show that at least 40% of the improvement is due to the proper treatment of correlations. Another attractive empirical feature of our setup is the ease at which it scales up with the number of assets: Pezzo et al. (2019) construct profitable out-of-sample trading strategies in the U.S. stock market involving thousands of stocks.

Our analysis complements nicely Leland (2000) and Liu (2004). They have continuous-time models with many assets and constant investment opportunity sets. Leland (2000) provides an heuristic approach to minimize proportional transaction costs. He assumes the non-trading region has the shape of a parallelogram or its higher dimensional analogs, and our work suggests this to be a good approximation, since that is exactly the form of the non-trading region in our model. Liu (2004) solves proportional and fixed transaction costs problems in the presence of many uncorrelated risky assets. His no-trading regions are rectangles (or higher dimensional analogs) which arise as special cases in our analysis when the covariance matrix of the risky securities is diagonal. Our model helps to clarify the connection between these two papers in the literature.

The rest of this paper is structured as follows: Section 2 introduces the general framework, Section 3 conveys the main economic insights through the graphical analysis of a set of selected examples, ~~and~~ Section 4 provides a formal characterizations of the solutions and their properties, and Section 5 illustrates some extensions. Section 6 concludes.

2 The Mean-Variance Framework

There are N risky securities and a riskless asset (security 0). We assume trading goes through the riskless asset. If we had a matrix of costs for trading each pair of assets, the notation would be more cumbersome, but nothing essential would change. Preferences are given by a linear mean-variance utility function over end-of-period wealth. Transaction costs can be variable, fixed by security or subgroups, or fixed overall, or any combination of these different types. We solve the following problem:

Problem 1 *Given the vector θ_0 of initial risky security positions, choose a nonnegative vector Δ_+ of security purchases and a nonnegative vector Δ_- of security sales to maximize the utility of terminal wealth*

$$U(\Delta_+, \Delta_-) = \theta'\mu - \frac{\lambda}{2}\theta'\mathbf{V}\theta - \frac{\kappa}{2}(\theta - \theta_B)'\mathbf{V}(\theta - \theta_B) - c(\Delta_+, \Delta_-) \tag{1}$$

subject to

$$\theta = \theta_0 + \Delta_+ - \Delta_- \tag{2}$$

and

$$c(\Delta_+, \Delta_-) \equiv \Delta_+' C_+ + \Delta_-' C_- + \sum_{S \subseteq \mathcal{S}} K(S(\Delta_+, \Delta_-)) \tag{3}$$

where we use the following notation:

θ: vector of risky security positions after trades have occurred

c: cost function

μ: vector of risky security expected excess returns net of any liquidation costs

λ: risk aversion parameter (typically positive, but $\lambda + \kappa > 0$ is sufficient for us)

κ: tracking error parameter (typically zero or positive, but $\lambda + \kappa > 0$ is sufficient for us)

$\mathbf{V}$: positive-definite covariance matrix of risky security returns net of any liquidation costs

θ_B: vector of benchmark risky security positions

$C_+ \geq 0$: vector of proportional transaction cost parameters for purchases

$C_- \geq 0$: vector of proportional transaction cost parameters for sales

$\mathcal{S}$: set of all tradable risky securities

$\emptyset$: the empty set $\{\}$

$S(\Delta_+, \Delta_-) = \{i | \Delta_{+,i} > 0 \text{ or } \Delta_{-,i} > 0\} \subseteq \mathcal{S}$: subset of securities tradable paying a fixed cost

$K(S) \geq 0, K(\emptyset) = 0, (\forall S : S_1 \supseteq S_2 \Rightarrow K(S_1) \geq K(S_2))$: fixed cost for trading in subset S

The vector of risky security positions after trade is θ, and the riskless-asset position is implicitly given as a residual, so that the positions in θ do not have to add up to the initial wealth. Utility is given as an excess over what we would get from investing only in the riskless asset without costs. We are suppressing the prices of the securities, which can be thought of as exogenous. In the case of assets such as stocks or bonds, we can think of θ as being the number of dollars invested at the current price, and μ as the vector of expected rates of return in excess of the riskfree rate. In the case of futures, we enter at a price of zero and we can think of θ being measured in terms of the size of the position (either in number of contracts or dollars worth of the underlying), with the excess return and volatility being measured at that scale.

The first two terms in the utility function (1) are standard for mean-variance optimization: $\theta'\mu$ gives the net change in expected return from holding risky assets rather than just the riskless asset, while $\frac{\lambda}{2}\theta'\mathbf{V}\theta$ is the utility penalty for taking on variance $\theta'\mathbf{V}\theta$. The constant $\lambda > 0$ is the coefficient of risk aversion: the larger the value of λ, the more reluctant we are to take on risk in exchange for return.[3]

The third term $\frac{\kappa}{2}(\theta - \theta_B)'\mathbf{V}(\theta - \theta_B)$ is a penalty for tracking error, as in Grinold and Kahn (1995). This term is perhaps controversial (and many academics would like to set $\kappa = 0$) because it depends on the benchmark θ_B and not just on the distribution of returns. The dependence on the benchmark would be unnecessary and damaging in an ideal world, but does arise in practice for good reasons and should be very familiar to practitioners. When plan sponsors allocate funds to different managers the penalty for tracking error gives the managers an incentive not to deviate too much from the plan sponsor's planned allocation reflected in managers' benchmarks. The plan sponsor may care about preserving the intended allocation because it is diversifying or because it reflects the plan sponsor's view of which assets classes are likely to do well. Alternatively, for asset-liability management, the benchmark can be interpreted as the best projection into the space of traded securities of some liability claim that a firm would like to cost-efficiently hedge by trading according to the mean-variance criterion.

The last term is the cost function $c(\Delta_+, \Delta_-)$, which specifies the future value of the transaction costs paid to re-balance the initial position θ_0 to the new position θ according to equation (2). The vector $C_+ \geq 0$ (resp. $C_- \geq 0$) is the security-specific proportional cost of buying (resp. selling) the security. Buying $\Delta_{+,n}$ units (resp. selling $\Delta_{-,n}$ units) incurs a cost of $\Delta_{+,n}C_{+,n}$ (resp. $\Delta_{-,n}C_{-,n}$).[4] For each $s \in \mathcal{S}$, $K(S) \geq 0$ indicates the fixed cost component we have to pay to be able to trade in the subset $S = \{i | \Delta_{+,i} > 0 \text{ or } \Delta_{-,i} > 0\}$ of risky securities. In the absence of trading no cost is incurred, so we set $K(\emptyset) = 0$. Also, it costs no less to trade more securities: $(\forall S_1 \subseteq S_2 \subseteq \mathcal{S})(K(S_1) \leq K(S_2))$. If $K(S)$ is the same positive constant for all nonempty sets S, the constant is an overall fixed cost. When the fixed cost function $K(\cdot)$ has the form $K(S) = \sum_{n \in S} K(\{n\})$, we have pure security-specific fixed costs. As a third example, if $\mathcal{S}$ is partitioned into M different disjoint "markets"

[3] Including 2 in the denominator makes the units the same as absolute risk aversion in a multivariate normal model with exponential utility, and also the 2 cancels when we look at the first-order conditions.

[4] In a single-period model there is no difference in proportional costs per dollar and proportional costs per share.

$S_1,\ S_2,\ \ldots\ \ S_M$ with $\cup_{m=1}^{M} S_m = \mathcal{S}$, suppose we can write $K(S) = \sum_{m \in T} K(S_m)$ where $T \equiv \{m \in M | S \cap S_m \neq \emptyset\}$. Then we can say that we have pure market-specific fixed costs. If costs are proportional only, then $(\forall S)(K(S) = 0)$.

There are a number of ways of making the cost structure more complex but retaining the basic structure that allows us to solve the model using quadratic programming. For example, we could have a fixed cost for shorting any security or more generally the fixed cost could depend both on the set of securities being sold and the securies being bought. One particularly interesting generalization by Maurer et al. (2018) has a separate variable cost for selling an existing long position than for establishing a new short position, emphasizing that the cost may depend on more than the net trade.

Also outside the scope of our analysis are various other market frictions that would maintain the quadratic programming format that keeps our model tractable. Instead, we could add non-negativity constraints for portfolio positions, no-borrowing constraints, margin requirements, or constraints on proportions in individual stocks or industries, and the problem would still be easy to solve, but including such considerations here would only be a distraction from our main message.

3 Examples

In this section we illustrates and motivate the scope of our framework via a set of selected examples.

3.1 Proportional Costs

Figure 1 shows a typical case with proportional costs. If the initial allocation θ_0 is in the non-trading region, labeled "NO TRADE", then there is no trade whose benefit covers the cost and it is best not to trade. The right boundary of the non-trading region is part of the line along which we are just indifferent at the margin about selling security 1, and it is optimal to sell security 1 if we start to the right of this boundary. The left boundary of the non-trading region is part of the line along which we are just indifferent at the margin about purchasing security 1, and it is optimal to purchase security 1 to the left of this boundary. The boundaries for purchasing and selling are different because the costs put a wedge between the effective price with and without costs. Similar to the case of security 1, we sell security 2 if we start above the top boundary and we sell security 2 if we start below the bottom boundary. If we start inside one of the four corner regions, then we trade in both securities. Absent the positive correlations between the returns of the two securities, the non-trading region would have been a square (or a rectangle for non-identical costs) with sides parallel to the axes. With positive correlation (or a positive weight on benchmark deviations and positive correlations with the benchmark), the two securities are substitutes, and over-weighting in one security is more serious if we have the over-weighting in the other security. This is why the non-trading region is larger along the $-45°$ direction in which the over- and under-weightings cancel than along the $45°$ direction in which the over- and under-weightings reinforce each other.

3.1.1 Futures overlays

When security 1 represents equities and security 2 their futures contracts, it might be optimal to deviate from (traditional) symmetric futures overlay strategies to take better advantage of the extra expected return from holding the underlying.

Futures overlays are transaction-cost-aware strategies which use futures as an inexpensive way of keeping effective asset allocations in line with a pre-specified allocation. For example, if we think the ideal weighting in equities is 60%, then as the market rises we become over-weighted and as the market falls we become under-weighted (since the fixed-income part of the portfolio moves less than proportionately with moves in the equity market). Maintaining a weighting near the ideal one by trading equities is very expensive. A futures overlay might correct for minor deviations by trading in futures, which are highly correlated with equities but much cheaper to trade.[5] This is why traditional strategies substitute futures trades for trades in equities. However, the expected returns ("alphas") of these trades are not usually discussed much, but they turn out to be very important.

Figure 2 illustrates a case in which the investor gets similar expected returns on the underlying and on "synthetic equity" composed of futures. In this case, the additional expected return from holding equities is typically too small to justify the additional transaction costs. Consequently, if we are over- or under-weighted in equities initially (with no futures), we correct our position by selling or buying futures. Thus, in the absence of some additional return to holding the underlying equities, a traditional synthetic equity strategy would be optimal, and there would seem to be little reason to hold the underlying equities in the first place (point $(0, 0)$ in Figure 2), or to trade them once we own them (the common initial equity-only exposures represented by the portion of the dotted line in Figure 2 to the right of -15%).

Generally, we might expect the return on the underlying equities to be higher than the return on synthetic equity due to the benefits of active management. Figure 3 illustrates an example in which equities have a significantly higher expected return than the synthetic equity strategy using futures. In this case, there is a trade-off between transaction costs and expected returns and it is optimal to use a futures overlay strategy of using futures to substitute for some trading in the underlying. In practice, most plan sponsors use a "symmetric" futures overlay that uses futures to the same extent for correcting over- and under-exposure to the market: this would indeed be the typical optimal thing to do given the setup of Figure 2. However, the current example shown in Figure 3 prescribes an "asymmetric" strategy that makes good economic sense. For the typical equities-only starting positions on the dotted line, if the market exposure must be reduced, we sell futures, which allows us to keep the extra return on the underlying equities. On the other hand, if the market exposure must be increased, we buy equities, which have the extra return, rather than futures, which don't.

[5]Perhaps futures are used to keep the exposure to equities to within 3% of the ideal allocation with trading in actual shares of stock only when the allocation gets more than 10% out of line.

3.2 Fixed Costs

3.2.1 Overall Fixed Costs

If there is an overall fixed cost, then if we trade the cost is the same whatever trade we make (since it is no more costly to trade to the ideal portfolio than to trade to a less-preferred portfolio). Therefore, as Figure 4 illustrates, if we are going to trade, we always trade to the same ideal portfolio (identified by the green square). Consequently, the choice problem reduces to one of comparing the utility of not trading with the utility of trading to the ideal point but incurring the fixed cost. It is optimal to trade when outside the indifference curve (an ellipse given our mean-variance assumptions) and not trading in the light blue area inside of it. If the asset returns were uncorrelated with identical variances, the non-trading region would be a circle. In this example, everything is symmetric but there is correlation. The correlation means that the two assets are substitutes and it is not so bad if we have too little of one asset if we have too much of the other. As in the case of proportional costs, this is why we are quicker to trade if we are over-weighted in both assets than if we are over-weighted in one and under-weighted in the other.

3.2.2 Asset-specific Fixed Costs

Figure 5 illustrates the various trading regions. Surrounding the ideal point in the middle (identified by the green square), is the non-trading region. The upper left and lower right boundaries are on the same ellipse, and the other boundaries are linear. From the regions in the corners, we trade both securities to the ideal point. From the regions on the right and left, we trade Security 1 but not Security 2 to a line going through the ideal point. This would be a vertical line if we had no correlation, but has negative slope in our case. Similarly, from the regions above and below, we trade only Security 2 to a different line running through the ideal point. The left plot illustrates the trades in both securities, the middle plot illustrates the trades in Security 1 alone, and the right illustrates the trades in Security 2 alone. This example is consistent with Brennan (1975) or Goldstein (1979), in the presence of constraints proportional to the number of securities traded, or implicitly Jacob (1974), where we are given exogenously a small number of securities that can be bought. This analysis is more general mostly in that it does not assume the starting position in cash alone.

3.3 Mixed Costs

Figure 6 shows the various optimal trades under an asset-specific fixed and proportional cost structure. The non-trading region, labeled “NO TRADE”, resembles the one from Figure 5 when costs are only asset-specific fixed and analogous geometrical interpretations apply. The upper left and lower right boundaries are on two different ellipses where we are indifferent between simultaneously trading both securities or staying put, and the other boundaries are straight indifference lines where we are indifferent between trading in one security at a time or do nothing. Optimal trades for initial allocations outside of the non-trading region are very similar to those described in Figure 1 when costs are only proportional and end up in the interior of the non-trading region. This is because the impact of proportional over fixed costs increases with the distance from the ideal point (identified by the green square), and when a position is far enough proportional costs are just what matter. In particular, as the leftmost plot illustrates, from the regions in the corners we trade both securities to the closest corner of the interior non-trading region defined by the intersections of the dashed lines: the region and the optimal allocations are the same as the one described in Figure 1. While, as shown in the middle and right plot respectively, from regions to the left and right and above and below the non-trading region we only trade in security 1 or 2. These trades bring us to the closest dashed straight line inside the non-trading region, part of which representing the border of the non-trading region in Figure 1. The reason why the lines stretch beyond is because in these portions only the position of one security is far enough from the optimum for the fixed cost to be considered sunk, resulting in extra trades involving the most displaced asset.

4 Analytical characterization of the solutions

4.1 Pure proportional costs

When $k = 0$ and $K = 0$, $c(\Delta_+, \Delta_-) = \Delta_+' C_+ + \Delta_-' C_-$ so that the transaction cost for a trade $\Delta_{+,n}$ or $\Delta_{-,n}$ is just proportional to the size of the trade. Then the problem to solve specializes to

Problem 2 *(Proportional Costs) Choose a nonnegative vector Δ_+ of security purchases and a nonnegative vector Δ_- of security sales to maximize the utility of terminal wealth*

$$U^{PC}(\Delta_+, \Delta_-) = \theta'\mu - \frac{\lambda}{2}\theta'\mathbf{V}\theta - \frac{\kappa}{2}(\theta - \theta_B)'\mathbf{V}(\theta - \theta_B) - \Delta_+' C_+ - \Delta_-' C_- \tag{4}$$

subject to $\theta = \theta_0 + \Delta_+ - \Delta_-$

We next give sufficient conditions for existence and uniqueness of solutions in Problem 2.

Proposition 1 *Given the maintained assumptions of a positive definite covariance matrix $\mathbf{V}$ and $\lambda + \kappa > 0$, the optimal solution to Problem 2 exists and the optimal portfolio θ^* we trade to is unique. The optimal trade is of the form $\Delta_{+,n} = \max(\theta_n - \theta_{n,0}, 0) + x_n$ and $\Delta_{-,n} = -\min(\theta_n - \theta_{n,0}, 0) + x_n$ where $x_n \geq 0$. If the round-trip trading cost is positive, $C_{+,n} + C_{-,n} > 0$, then it is suboptimal to simultaneously buy and sell so that $x_n = 0$. In particular, if the round-trip trading cost is positive for all securities, the optimal trades are unique and Problem 2 has a unique solution.*

Proof.

Let $\hat{U}(\theta) \equiv \max\{U^{PC}(\Delta_+, \Delta_-) | \Delta_{+,n}, \Delta_{-,n} \geq 0 \text{ and } \theta = \theta_0 + \Delta_+ - \Delta_-\}$. Define $\hat{\Delta}_+(\theta)$ and $\hat{\Delta}_-(\theta)$ by $\hat{\Delta}_+(\theta) = \max(\theta_n - \theta_0, 0)$ and $\hat{\Delta}_-(\theta) = -\min(\theta_n - \theta_0, 0)$. Then $\hat{U}(\theta) = U^{PC}(\hat{\Delta}_+, \hat{\Delta}_-)$, because any other way of generating θ either has the same value (if trading in at least some securities with round-trip cost of zero) or smaller value. Therefore

$$\hat{U}(\theta) = \theta'\mu - \frac{\lambda}{2}\theta'\mathbf{V}\theta - \frac{\kappa}{2}(\theta - \theta_B)'\mathbf{V}(\theta - \theta_B) - \hat{\Delta}_+(\theta)'C_+ - \hat{\Delta}_-(\theta)'C_- \tag{5}$$

Note that $-\hat{\Delta}_+(\theta)'C_+$ and $-\hat{\Delta}_-(\theta)'C_-$ are concave (since $\hat{\Delta}_+(\theta)'C_+$ and $\hat{\Delta}_-(\theta)'C_-$ are convex) and the quadratic terms are strictly concave. Therefore $\hat{U}(\theta)$ is strictly concave and if an optimal θ exists it is unique. Such an optimum does exists because $\hat{U}(\theta) < \hat{U}(\theta_0)$ outside of the set

$$\theta'\mu - \frac{\lambda}{2}\theta'\mathbf{V}\theta - \frac{\kappa}{2}(\theta - \theta_B)'\mathbf{V}(\theta - \theta_B) \geq \theta_0'\mu - \frac{\lambda}{2}\theta_0'\mathbf{V}\theta_0 - \frac{\kappa}{2}(\theta_0 - \theta_B)'\mathbf{V}(\theta_0 - \theta_B)$$

which is the compact set bounded by a multidimensional ellipse because V is positive-definite and $\kappa + \lambda > 0$.

The final allocation θ can be achieved by any pairs $(\Delta_+, \Delta_-) = (\hat{\Delta}_+ + x, \hat{\Delta}_- + x)$, where $x_n \geq 0$. The utility function for these pairs is given by $\hat{U}(\theta) - C_+'x - C_-'x$. If round-trip trading costs are positive for all securities n, then $x = 0$ is the unique choice, so that the optimal directional trades are unique. ■

The assumption that $\lambda + \kappa > 0$ is important: if $\lambda + \kappa < 0$ the investor is risk-loving and will not have a solution, while if $\lambda + \kappa = 0$ the agent is effectively risk-neutral and will have a solution only in an uninteresting case. Since V is a covariance matrix it must be at least positive semi-definite. If it is positive-semi-definite but not positive definite, we can derive sufficient conditions for a solution but we need need some additional structure to rule out arbitrage, as in the Fundamental Theorem of Asset Pricing (Dybvig and Ross (1987)). HERE: why isn't this citation working?

4.1.1 First Order Conditions

For the rest of the discussion of pure proportional costs, we assume the round-trip cost $C_{+,n} + C_{-,n}$ to be strictly positive for every security n. This makes the optimal trades Δ_+

and $\Delta -$ unique, but is without loss of generality in the sense that Proposition 1 tells us exactly what happens otherwise.

Given θ_0, it is easy and fast to solve Problem 2 numerically for the optimal trades. However, we can learn a lot about the economics by studying the first-order conditions, which also allow us to solve small problems exactly by hand as we show in Section 4.1.3. By Proposition 1, the solution of Problem 2 maximizes the concave non-differentiable function $\hat{U}(\theta)$, so the first-order conditions for an optimum imply[6] $0 \in \nabla_0 \hat{U}(\theta)$, where $\nabla \hat{U}(\theta)$ is the subgradient given by $\{\mu - \lambda \mathbf{V}\theta - \kappa \mathbf{V}(\theta - \theta_B) + y | y_n \in [-C_{-,n}, C_{+,n}](\forall n)\}$ and $\Delta_+ = \hat{\Delta}_+(\theta)$ and $\Delta_- = \hat{\Delta}_-(\theta)$.

We can also write the first-order conditions for not wanting to buy or sell any more of security n as

$$(\mu - \lambda \mathbf{V}\theta - \kappa \mathbf{V}(\theta - \theta_B))_n \leq C_{+,n} \tag{6}$$

and

$$-(\mu - \lambda \mathbf{V}\theta - \kappa \mathbf{V}(\theta - \theta_B))_n \leq C_{-,n} \tag{7}$$

These two inequalities say that the marginal benefit of buying or selling security n is less than the marginal cost. The solution also needs to satisfy the complementary slackness conditions that (6) (resp. (7)) holds with inequality if $(\theta - \theta_0)_n > 0$ (resp < 0). These complementary slackness conditions automatically hold if $\theta = \theta_0$.

4.1.2 No-trading region

Define the No-Trading Region (NTR) to be the set of initial positions θ_0 for which the optimal choice is not to trade ($\theta = \theta_0$).

Absent costs ($C_+ = C_- = 0$) Problem 2 features the standard mean-variance ideal allocation θ_I (adjusted for the presence of the benchmark θ_B)

$$\theta_I = \frac{1}{\lambda + \kappa} \mathbf{V}^{-1}(\kappa \mathbf{V}\theta_B + \mu) \tag{8}$$

[6] See Rockafellar (1981).

and it is optimal to trade directly to this portfolio whatever the initial allocation θ_0. As a result, the NTR is the singletion $\{\theta_0\}$.

With transaction costs, however, there is a non-trivial set of portfolios θ_0 too close to θ_I to justify any trade. This set of portfolios satisfies (6) and (7) and is described by the intersection of half-planes of the form

$$(\mathbf{V}_{,n})'\theta_0 \geq \frac{1}{\kappa+\lambda}(\kappa(\mathbf{V}_{.,n})'\theta_B + (\mu_n - C_{+,n})) \tag{9}$$

$$(\mathbf{V}_{,n})'\theta_0 \leq \frac{1}{\kappa+\lambda}(\kappa(\mathbf{V}_{.,n})'\theta_B + (\mu_n + C_{-,n})) \tag{10}$$

where $\mathbf{V}_{,n}$ is the n-th column of $\mathbf{V}$. These equations with equalities were used to plot the boundaries of the NTRs in Figures 1, 2 and 3. More generally, solving for being on the edge of multiple half-spaces allows us to compute the vertices and faces of the NTR.

In particular, each asset n is associated with two half-spaces (9) and (10), which are parallel to each other and have orientation characterized by the vector $\mathbf{V}_{.,n}$. These half-spaces are separated by the distance $\frac{C_{+,n}+C_{-,n}}{||\mathbf{V}_i||(\kappa+\lambda)}$ which is proportional to the magnitude of the costs, and are parallel to the coordinate axes if and only if $\mathbf{V}$ is a diagonal matrix. The resulting object is a parallelogram for the case of 2 risky securities, or its higher dimensional analog for a generic number $N > 2$

The first-order conditions in the case of proportional costs fully characterizes the NTR. Their linearity suggests that Leland made a good assumption when conjecturing the linearity of the NTR boundaries in the heuristic solution of his continuous time model (Leland (2000)). Our analysis is also consistent with the formal solution in continuous time with uncorrelated assets of Liu (2004): as in our model, if risky returns are uncorrelated, the sides of his NTR are parallel to the coordinate axes.

4.1.3 An easy problem to solve by hand

The solution and the economics of Problem 2 are actually very easy to derive for the case of two risky securities. Let us take a closer look at the example described in Figure 1, for which we have $\lambda = 2, \kappa = 0, \mu_1 = \mu_2 = 0.05$, $\rho = 0.5, V_{1,1} = V_{2,2} = 0.04$ (implying $\mathbf{V} = \begin{bmatrix} 0.04 & 0.02 \\ 0.02 & 0.04 \end{bmatrix}$) and $C_{+,1} = C_{-,1} = C_{+,2} = C_{-,2} = 0.01$. The first order conditions (9) and (10) inform us that the non-trading region is the set of portfolios satisfying

$$0.04\theta_1 + 0.02\theta_2 \begin{cases} \geq \frac{1}{2}0.04 & \text{for buying security 1: } B1 \\ \leq \frac{1}{2}0.06 & \text{for selling security 1: } S1 \end{cases}$$

$$0.02\theta_1 + 0.04\theta_2 \begin{cases} \geq \frac{1}{2}0.04 & \text{for buying security 2: } B2 \\ \leq \frac{1}{2}0.06 & \text{for selling security 2: } S2. \end{cases}$$

A portfolio in the no-trading region is at the boundary if at least one inequality holds with equality and it is a corner if two hold with equality. For example, the bottom-right corner solves $S1$ and $B2$, so it is $(\frac{2}{3}, \frac{1}{6})$. Similarly, we can calculate the corners $(\frac{1}{2}, \frac{1}{2})$, $(\frac{1}{6}, \frac{2}{3})$ and $(\frac{1}{3}, \frac{1}{3})$ solving $S1$ and $S2$, $B1$ and $S2$ and $B1$ and $B2$.

Once we have computed the four corners of the nontrading region, we can write down the entire strategy. In the corner regions, we trade both securities to the corresponding corners. For example, if $\theta_{0,1} > 2/3$ and $\theta_{0,2} < 1/6$, it is optimal to trade to the corner $(2/3, 1/6)$ by selling $\theta_{0,1} - 2/3$ units of security 1 and buying $1/6 - \theta_{0,2}$ units of security 2. In the side and top regions, we buy or sell one security until the corresponding inequality is satisfied with equality. For example, if $\theta_{0,1} > 3/4 - \theta_{0,2}/2$ (so S1 is not satisfied at the starting point) and $1/6 \leq \theta_{0,2} \leq 1/2$ (so B2 and S2 are satisfied if we sell enough of security one so that S1 is just satisfied), then it is optimal to sell $\theta_{0,1} - 2/3 + \theta_2/2$ units of security 1 without trading security 2. Once we determine the nontrading region and the optimal strategies in the corners and on the top, bottom, and sides, we know the entire strategy.

4.2 General Conditions for Fixed and Variable Costs

Proposition 2 *Assume that the covariance matrix* $\mathbf{V}$ *is positive definite, that* $\lambda + \kappa > 0$, *and that the fixed cost structure* $K(S)$ *is nondecreasing (where nondecreasing is defined by* $(\forall S_1, S_2 \subseteq \mathcal{S})(S_1 \supseteq S_2 \Rightarrow K(S_1) \geq K(S_2))$*). Then there always exists a solution to Problem 1, and the optimal solution is the proportional-costs-only solution from Proposition 1 for some subset* $S \subseteq \mathcal{S}$ *of securities.*

Proof. Whenever the fixed cost function $K(S(\Delta_+, \Delta_-))$ is not zero everywhere, Problem 1 can be viewed as a combinatoric problem of computing the optimum given the subset $S \subseteq \mathcal{S}$ of assets traded and then picking the subset with the best optimal value. Given N assets there are finitely many (2^N) different subsets and the solution to our problem takes the maximum over finitely many values. Each problem can be indexed by a different subset $S \subseteq \mathcal{S}$: for a given subset S we pay the fixed cost $K(S)$ which allows us to trade in all the securities inside of it and solve the following optimization

Problem 3 *(Subset S Problem)*

Take as given some subset $S \subseteq \mathcal{S}$ *of the securities, choose a nonnegative vector* Δ_+ *of security purchases and a nonnegative vector* Δ_- *of security sales to maximize the utility of terminal wealth*

$$U^S(\Delta_+, \Delta_-) = \theta'\mu - \frac{\lambda}{2}\theta'\mathbf{V}\theta - \frac{\kappa}{2}(\theta - \theta_B)'\mathbf{V}(\theta - \theta_B) - \Delta_+'C_+ - \Delta_-'C_- - K(S) \tag{11}$$

subject to $\theta = \theta_0 + \Delta_+ - \Delta_-$ *and* $(\forall i \in \mathcal{S} \setminus S)(\Delta_{i,+} = \Delta_{i,-} = 0)$.

Note that we charge a cost of $K(S)$ even though the optimal solution may not trade all securities in S: this avoids a closure problem, as we explain below, and is without loss of generality given our assumption that the cost $K(\cdot)$ is nondecreasing. In particular, the problem is equivalent to a proportional-cost-only optimization since the fixed cost $K(S)$ is a sunk cost and we know from Proposition 1 to admit a unique solution. Let this solution be denoted as (Δ_+^S, Δ_-^S). If it is optimal to trade in all the securities belonging to subset S, the maximum level of utility our investor can achieve is given by $U^S(\Delta_+^S, \Delta_-^S)$. If it is optimal to trade in a strict subset T of S, then (Δ_+^S, Δ_-^S) is also the solution to the Subset T Problem. This is because the objective function is the same up the addition of a constant and the solution to the less constrained Subst S Problem is still feasible for the more constrained Subset T Problem. Furthermore it must be that $U^T(\Delta_+^T, \Delta_-^T) = U^T(\Delta_+^S, \Delta_-^S) \geq U^S(\Delta_+^S, \Delta_-^S)$. This is due to $K(\cdot)$ being nondecreasing and $T \subset S$. Therefore, when it is optimal to trade in a subset T of S the maximum level of utility our investor can achieve is given by $U^T(\Delta_+^S, \Delta_-^S)$ not necessarily $U^S(\Delta_+^S, \Delta_-^S)$.

A solution to Problem 1 is a solution for some Subset S^* Problem where S^* belongs to $argmax_{S \subseteq \mathcal{S}} \left\{U^S(\Delta_+^S, \Delta_-^S)\right\}$. From the previous discussion only two cases are possible: if for every $S \subseteq \mathcal{S}$ the solution to Subset S Problem entails trading in all the securities in the subset, then $(\Delta_+^{S^*}, \Delta_-^{S^*})$ is the solution to Problem 1 and $U^{S^*}(\Delta_+^{S^*}, \Delta_-^{S^*})$ is the optimal utility level. If the solution to Subset S Problem entails trading in a strict subset T of securities, we know that $(\Delta_+^{S^*}, \Delta_-^{S^*}) = (\Delta_+^T, \Delta_-^T)$ and $U^T(\Delta_+^T, \Delta_-^T) = U^T(\Delta_+^{S^*}, \Delta_-^{S^*}) \geq U^{S^*}(\Delta_+^{S^*}, \Delta_-^{S^*})$. Furthermore, because S^* belongs to $argmax_{S \subseteq \mathcal{S}} \left\{U^S(\Delta_+^S, \Delta_-^S)\right\}$, it follows that $U^T(\Delta_+^T, \Delta_-^T) \leq U^{S^*}(\Delta_+^{S^*}, \Delta_-^{S^*})$. ■

The assumptions of a positive definite covariance matrix $\mathbf{V}$ and a positive coefficient of risk aversion λ guarantee existence and uniqueness of solutions in the presence of proportional costs. When costs are required only to (or additionally allowed to) be fixed, we also need $K(\cdot)$ to be nondecreasing. This restriction is mild since it naturally assume investors pay identical or higher costs for trading in bigger subsets of securities. Nonetheless, not imposing it might lead to cases where a solution does not exist.[7]

[7] For example if we only have one risky security and not trading (the null subset) has a higher cost than trading (a set of cardinality 1), in the case in which not trading is the ideal optimum there is no solution. This is because if we pay the lower cost and then trade we might be better off than deciding ex-ante not to trade and pay the higher cost.

5 Extensions

5.1 Price Pressure

Practitioners are concerned with price pressure: buying tends to move prices up and selling tends to move prices down, and either way larger trades receive a less favorable price.[8] This could be due to information as in L.R. and P.R. (1985) or Walter (1971), since the trade could be due to information about the stock. Or, it could be that there is a temporary shortage of liquidity in the market and the price pressure could reflect risk aversion or inventory costs of the market makers. We can think of the impact of price pressure on cash flow as analogous to a trading cost (or as a trading cost of we want to take a broad view), and price pressure can be modelled formally by including an additional term in the transaction cost function. If the increase in price for each stock is proportional to the amount traded (as in the equilibrium in S. (1985) and many papers in the subsequent literature), this introduces a quadratic term in the objective function since the cost of buying the change in position is the price times quantity, where the quantity is linear in the price. A quadratic term fits nicely within mean-variance analysis.

To analyze price pressure, we add the quadratic term $(\theta - \theta_0)'\Pi(\theta - \theta_0)/2$ to the cost function (3). The diagonal elements of Π describe the security-specific price impact parameters while (possibly) non-zero off diagonal elements admits that (as in Back (1993)) trading in one security may impact the price in other securities (because learning about the two is correlated). The equilibrium in Back (1993) has learning about one asset from the order flow of the other because he studied a stock and an option on the stock, which are based on the same underlying value. However, in equilibrium the two are not locally perfectly correlated because the conditional distribution of variance changes based on the order flow. While his model has only two risky assets, this feature is general. In our model, we assume that Π is positive definite, so that the impact of price pressure makes trading at scale more costly no

[8] As usual, matters are somewhat more complicated in practice, and in some markets a trade of intermediate size receives a more favorable price than a small trade, perhaps due to price discrimination against retail customers.

matter in what proportions we trade the assets.

When there are no fixed costs, there is an interesting connection between the problems with and without price pressure. With just proportional costs and price pressure, the choice problem is

Problem 4 *(Proportional Costs with Price Pressure) Choose a nonnegative vector Δ_+ of security purchases and a nonnegative vector Δ_- of security sales to maximize the utility of terminal wealth*

$$\theta'\mu - \frac{\lambda}{2}\theta'\mathbf{V}\theta - \frac{\kappa}{2}(\theta-\theta_B)'\mathbf{V}(\theta-\theta_B) - \Delta_+'C_+ - \Delta_-'C_- - \frac{1}{2}(\theta-\theta_0)'\Pi(\theta-\theta_0) \tag{12}$$

subject to $\theta = \theta_0 + \Delta_+ - \Delta_-$.

For this problem, the first-order conditions (6) and (7) become

$$\left(\mu - \lambda\mathbf{V}\theta - \kappa\mathbf{V}(\theta-\theta_B) - \Pi(\theta-\theta_0)\right)_n \leq C_{+,n} \tag{13}$$

and

$$-\left(\mu - \lambda\mathbf{V}\theta - \kappa\mathbf{V}(\theta-\theta_B) - \Pi(\theta-\theta_0)\right)_n \leq C_{-,n}. \tag{14}$$

These two inequalities say that the marginal benefit of buying or selling security n is less than the marginal cost. The solution also needs to satisfy the complementary slackness conditions that (13) (resp. (14)) holds with inequality if $(\theta-\theta_0)_n > 0$ (resp < 0). The following theorem compares and contrasts the solution with and without price pressure.

Theorem 5 *Consider Problem 4 with price pressure (Π positive definite) and Problem 2 without price pressure, but with all the same parameters except Π. Then the nontrading region is the same in both problems. However, if we start outside the nontrading region, we never trade to a point in the nontrading region if there is price pressure, but we always trade to a point in the nontrading region if there is no price pressure. In particular, this implies that the if we trade at all, the optimal trade with price pressure is never the same as the* *optimal trade without price pressure.*

Proof. When $\theta = \theta_0$, the first-order conditions (6) and (7) are exactly the same as the first-order conditions (13) and (14), since $-\Pi(\theta - \theta_0) = 0$. Also, complementary slackness holds in both cases. Since the objective function is strictly concave and the constraint set is convex, these first-order conditions are necessary and sufficient, so both problems have the same solution. This shows the two problems have the same nontrading region. To show that the optimal solution for Problem 2 is always in the nontrading region, note that the solution to the problems always satisfies the conditions (6) and (7) that characterize the nontrading region (and additional complementary slackness conditions). Now we need to show that if θ_0 is not in the nontrading region, neither is the optimal solution to the Problem 4 with price pressure. We argue by contradiction: assume not, and then for some parameter values, the optimal θ is in the nontrading region but θ_0 is not, so that $\theta - \theta_0$ is nonzero and θ satisfies the first-order conditions (6) and (7) and the complementary slackness conditions that if $(\theta - \theta_0)_n > 0$ then (6) holds with equality, while if $(\theta - \theta_0)_n < 0$ then (7) holds with equality. Now, let

$$C_n^* \equiv \begin{cases} C_{+,n} & \text{if } (\theta - \theta_0)_n > 0 \\ -C_{-,n} & \text{if } (\theta - \theta_0)_n > 0 \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

Then for all n,

$$(\theta - \theta_0)_n \left(\mu - \lambda \mathbf{V}\theta - \kappa \mathbf{V}(\theta - \theta_B) - \Pi(\theta - \theta_0)\right)_n = (\theta - \theta_0)_n C_n^*. \tag{16}$$

When $(\theta - \theta_0)_n > 0$, this follows from complementary slackness of (13); when $(\theta - \theta_0)_n < 0$, this follows from complementary slackness of (14).[9] Otherwise, this follows because $(\theta - \theta_0)_n = 0$. Similarly, since we are given that θ is in the nontrading region, it must be that θ

[9]When when $(\theta - \theta_0)_n < 0$, the inequality in (14) is reversed by multiplication by a negative number.

satisfies (6) and (7), which implies that

$$(\theta - \theta_0)_n \left(\mu - \lambda \mathbf{V}\theta - \kappa \mathbf{V}(\theta - \theta_B)\right)_n \leq (\theta - \theta_0)_n C_n^*. \tag{17}$$

If we sum over n, (16) implies

$$\begin{aligned}
(\theta - \theta_0)' C^* &= (\theta - \theta_0)' \left(\mu - \lambda \mathbf{V}\theta - \kappa \mathbf{V}(\theta - \theta_B) - \Pi(\theta - \theta_0)\right) \\
&= (\theta - \theta_0)' \left(\mu - \lambda \mathbf{V}\theta - \kappa \mathbf{V}(\theta - \theta_B)\right) - (\theta - \theta_0)'\Pi(\theta - \theta_0) \\
&< (\theta - \theta_0)' \left(\mu - \lambda \mathbf{V}\theta - \kappa \mathbf{V}(\theta - \theta_B)\right), \qquad (18)
\end{aligned}$$

since Π is positive definite and $\theta - \theta_0 \neq 0$. However, this is inconsistent with (16) since summing (16) over n gives the opposite ordering (with a weak inequality)

$$(\theta - \theta_0)' C^* \geq (\theta - \theta_0)' \left(\mu - \lambda \mathbf{V}\theta - \kappa \mathbf{V}(\theta - \theta_B)\right). \tag{19}$$

This contradicts our assumption that θ_0 is outside the nontrading region but θ is in it. We have shown that if θ_0 is not in the nontrading region, then the optimal solution to Problem 2 is always there but the solution to Problem 4 is never there. This shows the two solutions cannot be the same and we are done. ■

Intuitively, in the direction we are trading (including both purchases and sales), we must be indifferent at the margin because we are free to go a little further or a little less. However, with price pressure the marginal cost in increasing in how much we trade, if we are indifferent at the margin making our trade, we must strictly prefer to trade at the margin if we start there. Positive definiteness of Π is important for this, because it means the marginal cost of trading in a given direction is increasing in the magnitude of the trade.

It might be tempting to conclude that we trade less if there is price pressure. While this would be true with a single risky asset, substitution of assets that takes place when we impose price pressure means that we could be trading in a different direction and possibly we would be trading more.

Having quadratic costs but no fixed or variable costs (as in Leland (1985)) is simple to analyze, since, as in the mean-variance problem without any costs, the nontrading region is a single point that can be computed directly. Unfortunately, this case is probably not realistic, since transaction fees and bid-ask spread are not very small for small trades.

5.2 Directional costs

Our setup allows variable costs to be different for purchases and sales, however it does not differentiate them depending on where we start the trades from. For example in practice, due to the presence of different margin requirements, it is more costly to open new short positions than to close existing long ones. That is, costs do not just depended on net trades.

Maurer et al. (2018) extend our framework to handle this type of issues in the context of FX markets. Proportional transaction costs are modeled as $\Delta'_{P+}C_{P+} + \Delta'_{P-}C_{P-} + \Delta'_{S+}C_{S+} + \Delta'_{S-}C_{S-}$. Where trades in our framework are purchase(sales), $\Delta_+(\Delta_-)$, they are now decomposed in non-negative increments(decrements) starting from long positions, $\Delta_{P+}(\Delta_{S+})$, and short positions, $\Delta_{P-}(\Delta_{S-})$, which are charged at their corresponding different directional costs C_{P+}, C_{P-}, C_{S+} and C_{S-}.

With the additional requirement of a cost structure such that $0 < C_{P-} \leq C_{P+}$ and $0 < C_{S+} \leq C_{S-}$ and that for every security n, $0 \leq \Delta_{P-,n} \leq -min(\theta_{0,n}, 0)$ and $0 \leq \Delta_{S+,n} \leq max(\theta_{0,n}, 0)$, it follows that whenever it is optimal to buy more of security n (either to decrease open short positions $\Delta_{P-,n} > 0$ or to increase open long positions $\Delta_{P+,n} > 0$), we first close open short positions (if any) and then open new long positions. Similarly, when it is optimal to sell more of the security we first close open long positions (if any) and then open new short positions. Under these assumptions Maurer et al. (2018) show that the optimal solution to Problem 2 exists and the optimal portfolio we trade to is unique and achievable with trades of the form $\Delta_{P+,n} + \Delta_{P-,n} = \max(\theta_n - \theta_{0,n}, 0)$ and $\Delta_{S+,n} + \Delta_{S-,n} = -\min(\theta_n - \theta_{0,n}, 0)$.

5.3 Linear constraints

So far we did not impose any constraints on the purchase and sales of securities. Because Problem 1 is a well-behaved quadratic program, it is possible to add linear constraints. Conceptually, this is nothing new, and Markowitz (1959) included constraints in his formulation. Some examples of the types of constraints commonly imposed are no short sales (long only), no borrowing (portfolio proportions add up to no more than one), holding no more than 5% of your portfolio in one stock, holding no more than 5% of the outstanding shares of any stock, holding nothing outside a target class of shares (e.g. large cap value), holding no more than twice the benchmark weights in any industry, going short no more than the amount of money invested (in a long-short portfolio), and total equity holdings of at least 98% (cash discipline).

6 Conclusion

We have studied transaction costs in a single-period mean-variance setting. This setting allows a full solution for many securities with possibly correlated returns, and makes the economics very clear. As in more complex models in the literature, there is a nontrading region in which the benefits of trading (of second order) do not justify the costs of trading (of zero or first order) towards the ideal point. If there are proportional costs only, any trading goes only to the boundary of the nontrading region, because any additional marginal trade is not justified. On the other hand, if there are only overall fixed costs, we always trade to the same point in the interior of the nontrading region, since once we have decided to trade, the trading cost is sunk and we will always go to the ideal point. More generally, with both fixed and variable costs, we typically trade to a different point inside the nontrading region depending on the portfolio we start from.

The analysis provides new insights into practical problems such as the use of futures overlays. Futures overlays do some portfolio rebalancing in futures instead of the underlying stocks as a way of reducing transaction costs. Our analysis shows that the optimal strategy should typically be asymmetric, buying stocks to increase exposure and shorting futures to reduce exposure. This is in contrast to traditional symmetric overlay strategies.

A variant of our analysis has been used by Maurer et al. (2018) to look at transaction-cost smart futures strategies. The strategy improves a lot on the optimal strategy ignoring costs and more interestingly the optimal strategy ignoring correlations among futures. This shows that our extension to multiple assets that are not correlated is important economically. Another attractive empirical feature of our setup is the ease at which it scales up with the number of assets: Pezzo et al. (2019) construct profitable out-of-sample trading strategies in the U.S. stock market involving thousands of stocks.

One possible use of this analysis is in examining whether asset pricing "anomalies" can be used to generate trading costs. In a world with trading costs, anomalies can exist even if all agents are optimizing and rational, provided the profitability of trading does not cover the costs. With this analysis, we can shine a fresh light on whether trading on anomalies is profitable. Possibly the smarter trading in the face of transaction costs will reveal that trading on anomalies is more profitable than previously thought, or perhaps we will find out that trading on anomalies gets unprofitable exactly when naive analysis (say using a linear forecasting models) suggests that profits are the largest.

The analysis has already provided insights into continuous-time problems that are hard to solve exactly. For example, the high-dimensional parallelogram shape of the nontrading region assumed in the heuristic analysis by Leland (2000) is exactly correct in our model, lending support to this assumption. We hope that insights from our analysis will help scholars to push the continuous model further.

References

Back, Kerry, 1993, Asymmetric Information and Options, *The Review of Financial Studies* 6, 435472.

Balduzzi, Pierluigi, and Anthony W. Lynch, 1999, Transaction costs and predictability: some utility cost calculations, *Journal of Financial Economics* 52, 47–78.

Balduzzi, Pierluigi, and Anthony W. Lynch, 2000, Predictability and Transaction Costs: The Impact on Rebalancing Rules and Behavior, *Journal of Finance* 55, 2285–2309.

Brennan, M., 1975, The Optimal Number of Securities in a Risky Asset Portfolio When There are Fixed Costs of Transacting: Theory and Some Empirical Results, *Journal of Financial and Quantitative Analysis* 10, 483–496.

Buss, Adrian, and Bernard Dumas, 2017, The Dynamic Properties of Financial-Market Equilibrium with Trading Fees, Working paper, INSEAD.

Davis, Mark H.A., and A. R. Norman, 1990, Portfolio Selection With Transaction Costs, *Mathematics of Operations Research* 15, 676–713.

Donohue, C., and K. Yip, 2003, Optimal Portfolio Rebalancing with Transaction Costs: Improving on Calendar- or Volatility-Based Strategies, *Journal of Portfolio Management* 29, 49–63.

Dumas, Bernard, and Elisa Luciano, 1991, An Exact Solution to a Dynamic Portfolio Choice Problem under Transactions Costs, *Journal of Finance* 46, 577–595.

Dybvig, H. Philip, and S. Ross, 1987, Arbitrage, in *The New Palgrave: A Dictionary of Economics* (MacMillan Press Limited).

Goldstein, A., 1979, A Model of Capital Asset Pricing with Transaction Costs, Ph.d dissertation, Yale University.

Goodman, Jonathan, and Daniel N. Ostrov, 2010, Balancing Small Transaction Costs with Loss of Optimal Allocation in Dynamic Stock Trading Strategies, *SIAM Journal on Applied Mathematics* 70, 19771998.

Grinold, Richard C., and Ronald N. Kahn, 1995, Active Portfolio Management, Technical report, Probus: Chicago.

Han, Jiarui, 2005, Dynamic Portfolio Management: An Approximate Linear Programming Approach, Ph.d dissertation, Stanford University.

Irle, A., and C. Prelle, 2008, A renewal theoretic result in porfolio theory under transaction costs with multiple risky assets, Working paper, Kiel Institute for the World Economy.

Jacob, N. L., 1974, A Limited-Diversification Portfolio Selection model for the Small Investor, *Journal of Finance* 29, 874–856.

Leland, Hayne E., 1985, Option Pricing and Replication with Transactions Costs, *The Journal of Finance* 40, 12831301.

Leland, Hayne E., 2000, Optimal porfolio implemenentation with transaction costs and capital gain taxes, Working paper, University of California at Berkeley.

Liu, Hong, 2004, Optimal Consumption and Investment with Transaction Costs and Multiple Risky Assets, *Journal of Finance* 59, 289–338.

Liu, Hong, and M. Loewenstein, 2004, Optimal Porfolio Selection with Transaction Costs and Finite Horizons, *Review of Financial Studies* 15, 805–835.

L.R., Glosten, and Milgrom P.R., 1985, Bid, Ask and Transaction Prices in a Specialist Market with Heterogeneously Informed Traders, *Journal of Financial Economics* 14, 71–100.

Lynch, Anthony W., and Sinan Tan, 2009, Multiple Risky Assets, Transaction Costs, and Return Predictability: Allocation Rules and Implications for the U.S. Investors, *Journal of Financial and Quantitative Analysis* 45, 1015–1053.

Mao, J. C. T., 1970, Essentials of Portfolio Diversification Strategy, *Journal of Finance* 25, 1109–1121.

Markowitz, Harry, 1952, Portfolio Selection, *Journal of Finance* 7, 77–91.

Masters, S. J., 2003, Rebalancing: Establishing a Consistent Framework, *Journal of Portfolio Management* 29, 52–57.

Maurer, Thomas A., Luca Pezzo, and Mark P. Taylor, 2018, Importance of Transaction Costs for Asset Allocations in FX Markets, Working paper, Washington University.

Mayshar, J., 1981, Transaction Costs and the Pricing of Assets, *Journal of Finance* 36, 583–597.

Muthuraman, K., and S. Kumar, 2006, Multidimensional Porfolio Optimization with Proportional Transaction Costs, *Mathematical Finance* 16, 301–335.

Myers, Jeremy D., 2009, Portfolio Optimization with Transaction Costs and Preconceived Portfolio Weights, Ph.d dissertation, Massachusetts Institute of Technology.

Pezzo, Luca, Lei Wang, and Duygu Zirek, 2019, Importance of Transaction Costs for Asset Allocations in the Equity Market, Working paper, University of New Orleans.

Pogue, G. A., 1970, An Extension of the Markowitz Portfolio Selection Problem to Include Variable Transactions' Costs, Short Sales, Leverage Policies and Taxes, *Journal of Finance* 25, 1005–1027.

Rockafellar, R. Tyrrell, 1981, The theory of subgradients and its applications to problems of optimization: convex and nonconvex functions, in *Volume 1 of R E, research and education in mathematics* (Heldermann).

S., Kyle Albert, 1985, Continuous Auctions and Insider Trading, *Econometrica* 6, 1315–1335.

Schachermayer, Walter, 2017, Asymptotic Theory of Transaction Costs , in Jessica James, Ian W. Marsh, and Lucio Sarno, eds., *Asymptotic Theory of Transaction Costs* (European Mathematical Society, Hoboken, NJ).

Shreve, Steven E., and Halil M. Soner, 1994, Optimal Investment and Consumption with Transaction Costs, *The Annals of Applied Probability* 4, 609–692.

Taksar, M., M. J. Klass, and D. Assaf, 1988, A Diffusion Model for Optimal Portfolio Selection in the Presence of Brokerage Fees, *Mathematics of Operations Research* 13, 277–294.

Tobin, J., 1958, Liquidity Preference as Behavior Towards Risk, *Review of Economic Studies* 25, 65–86.

Walter, Bagehot, 1971, The only game in town, *Financial Analysts Journal* 22, 12–14.

Mean-variance problem with proportional costs

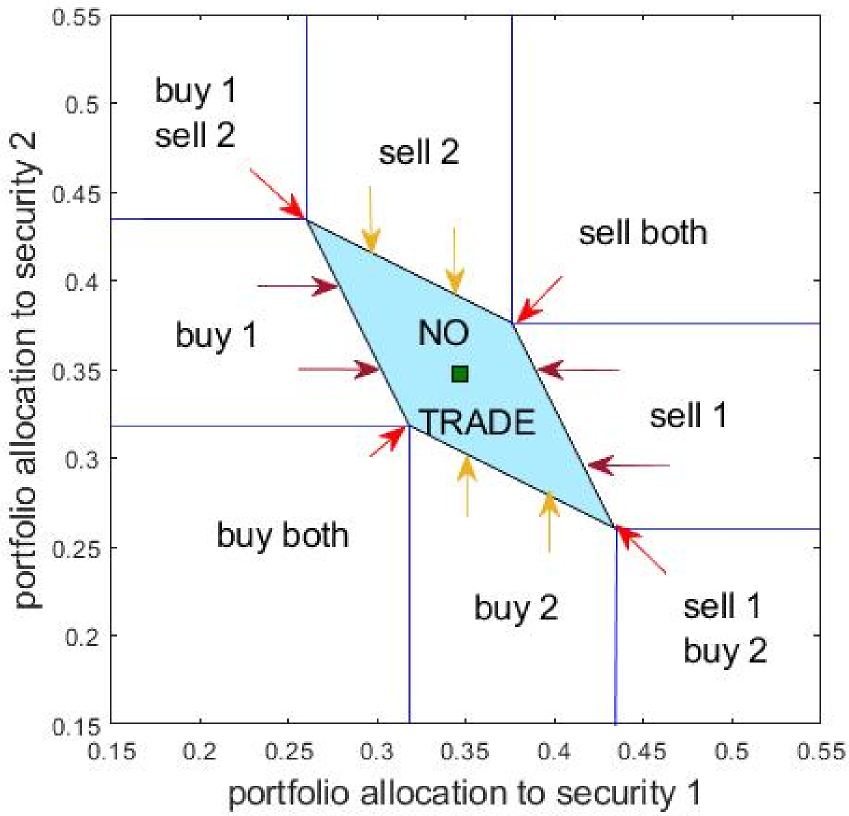

Figure 1: The figure illustrates the optimal strategy for a mean-variance investor with risk aversion of $\lambda = 2$ in the case of two positively correlated identical risky securities ($\rho = 0.5$), with risk premia of 6% ($\mu_1 = \mu_2 = 0.06$), volatilities of 24% ($V_(1,1) = V_(2,2) = 0.24^2$) and symmetric buy and sell proportional costs of 0.5% = 50 basis points ($C_{+,1} = C_{-,1} = C_{+,2} = C_{-,2} = 0.005$). With proportional costs, the non-trading region is the area of a parallelogram. Outside the non-trading region, it is optimal to trade (along the arrows) to the boundary of the non-trading region. If returns were uncorrelated, then the non-trading region would be a square with sides parallel to the axes. In this example, because returns are correlated and the two securities are substitutes, over-weighting in one security is less likely to result in a trade if we are under-weighted in the other security.

.

Traditional futures overlay with proportional costs

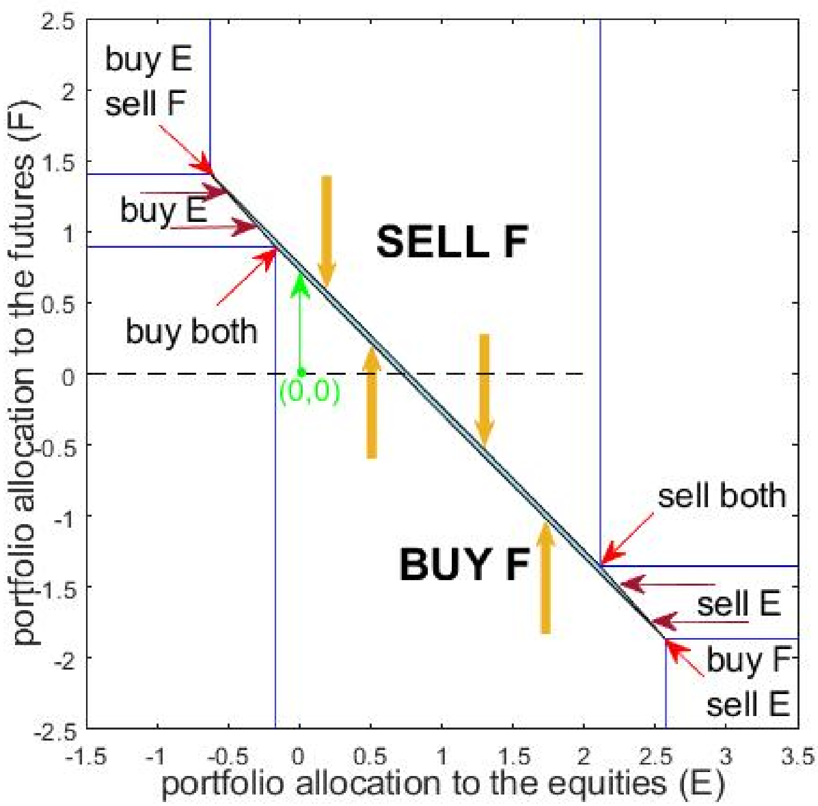

Figure 2: This example considers holdings of equities (E) and futures (F) that are highly correlated ($\rho = 0.95$) in a case in which the equity-equivalent risk premia and volatilities are nearly the same ($\mu_E = 3.4\%, \sigma_E = 21\%$ and $\mu_E = 3\%, \sigma_E = 20\%$), and the mean-variance investor has risk aversion of $\lambda = 1$. In this case, the cheaper trading costs of futures ($C_{+,F} = C_{-,F} = 0.1\%$ versus $C_{+,E} = C_{-,E} = 0.6\%$) are decisive. Starting from the normal range of equities-only positions (the dotted line) the majority of times it is optimal to pursue traditional futures overlay strategies where under or over market exposures are corrected via trading in futures as highlighted in bold. In particular, starting from all cash (highlighted in green), the optimal trade is a "synthetic" strategy consisting in buying futures only.

.

Asymmetric futures overlay with proportional costs

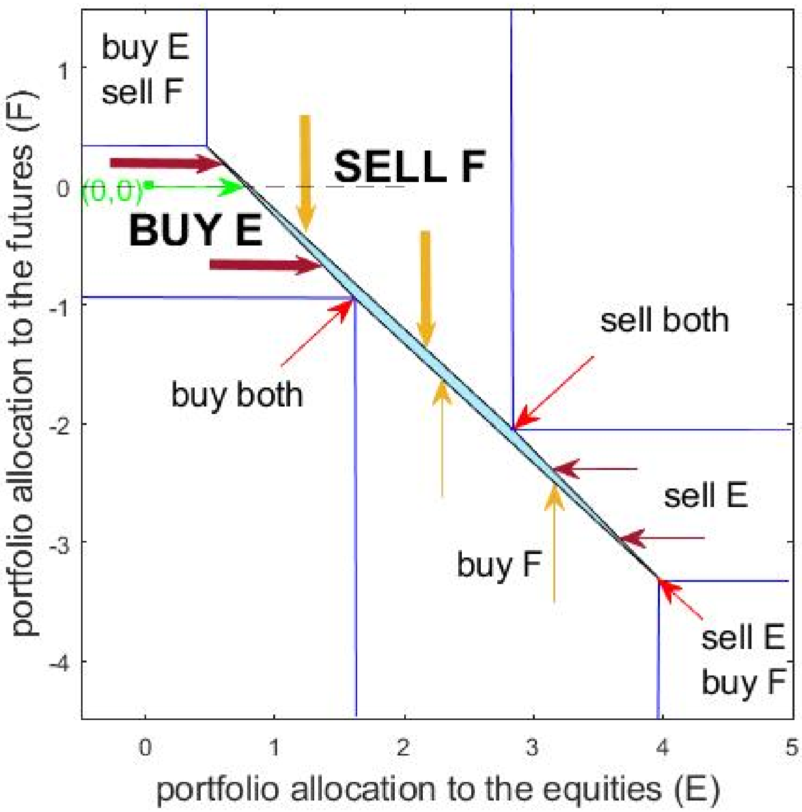

Figure 3: This second example with futures has a significantly higher risk premium ("alpha") on actual equities (E, $\mu_E = 4\%$) than on synthetic equities composed of futures (F, $\mu_F = 3\%$) on top of a stark difference in proportional costs ($C_{+,E} = C_{-,E} = 0.52\%$ versus $C_{+,F} = C_{-,F} = 0.25\%$). Furthermore the futures and equities are highly correlated ($\rho = 0.95$), with similar volatilities ($\sigma_E = 21\%$ and $\sigma_E = 20\%$). The optimal strategy for a mean-variance investor with risk aversion of $\lambda = 1$ is an "asymmetric" futures overlay strategy typically selling futures to correct for overexposure to market risk but buying underlying equities to correct for underexposure to market risk. This asymmetry is due to the fact that selling futures allows the investor to keep the alpha on the exposure she is eliminating, while buying equities allows her to gain alpha on the exposure she is taking on.

.

Mean-variance problem with overall fixed costs

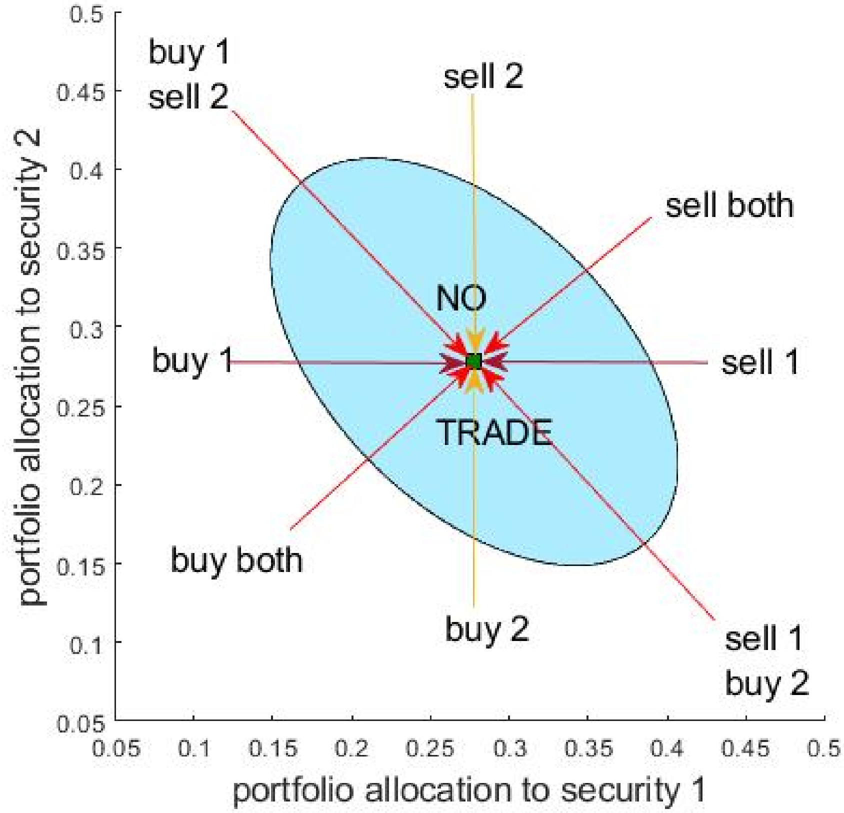

Figure 4: The figure illustrates the optimal strategy for trades in two positively correlated identical risky securities ($\rho = 0.5$), with risk premia of 6% ($\mu_1 = \mu_2 = 0.06$), volatilities of 24% ($V_(1,1) = V_(2,2) = 0.24^2$) and an overall fixed cost of $k = 0.075\%$. With an overall fixed cost, either there is trade immediately to the ideal point (indicated by a green square) or it is not worth trading at all. The non-trading region (shaded light blue) is an ellipse. As with proportional costs, correlation between the assets implies that it is more damaging (and more likely to do trade) when both asset positions are out of line in the same direction.

.

Mean-variance problem with asset-specific fixed costs

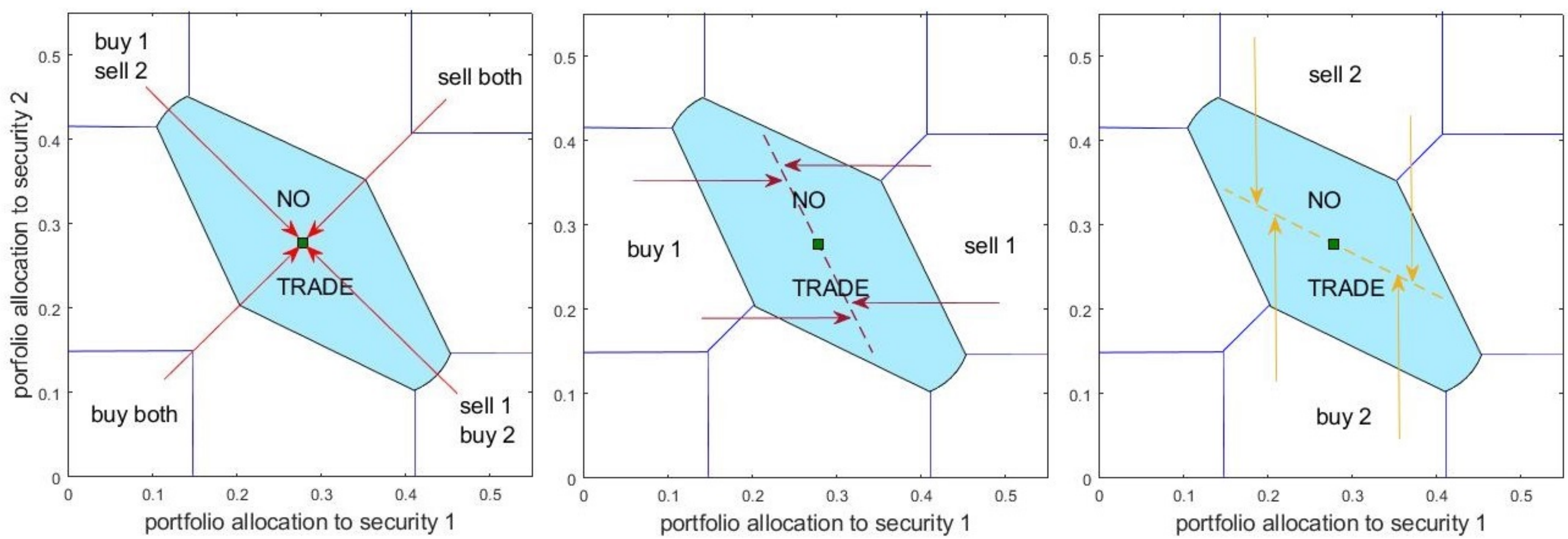

Figure 5: This figure shows the various trading regions for a mean-variance investor with risk aversion of $\lambda = 2$ in an example with security-specific fixed costs ($K_1 = K_2 = 0.075\%$) and identical, positively correlated ($\rho = 0.5$), assets with risk premia of 6% ($\mu_1 = \mu_2 = 0.06$) and volatilities equal to 24% ($V_(1,1) = V_(2,2) = 0.24^2$). With security-specific costs, it only pays to trade whatever security or securities is significantly out of line with the ideal allocation (identified by a green square). From left to right the different arrows indicate the optimal trades to the ideal point when (i) both securities are traded (ii) only security 1 is traded and (iii) only security 2 is traded.

Mean-variance problem with asset-specific fixed and proportional costs

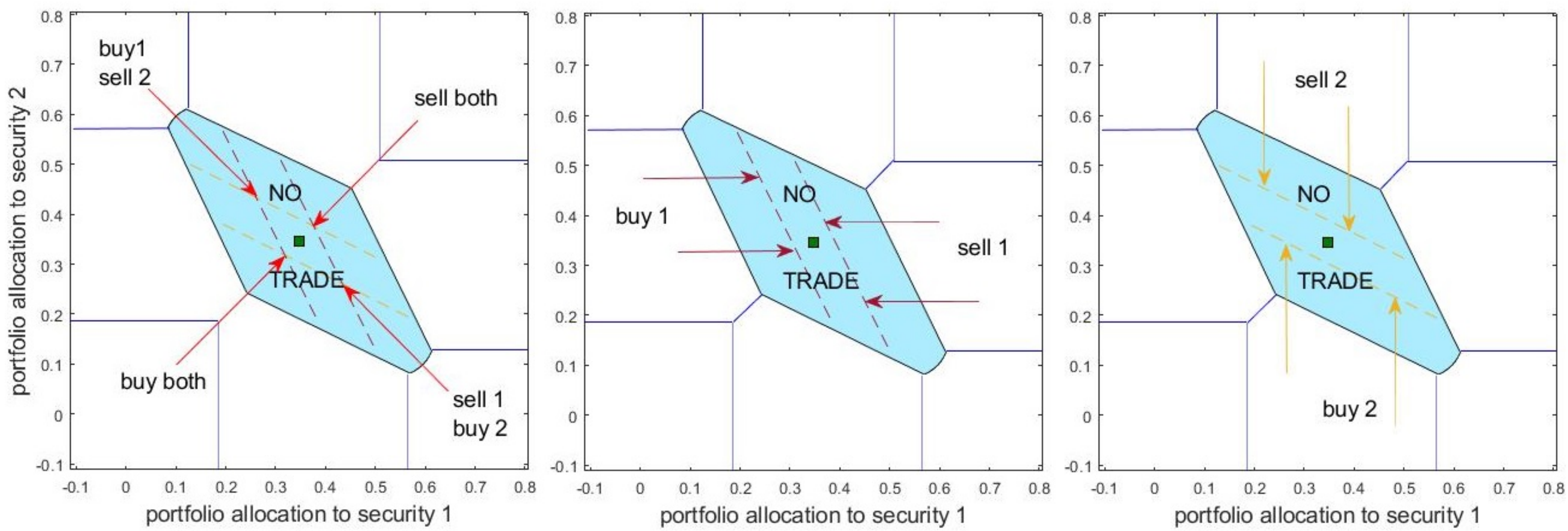

Figure 6: This figure shows the various trading regions for a mean-variance investor with risk aversion of $\lambda = 2$ in an example with security-specific fixed costs ($K_1 = K_2 = 0.075\%$) and symmetric buy and sell proportional costs of $0.5\% = 50$ basis points ($C_{+,1} = C_{-,1} = C_{+,2} = C_{-,2} = 0.0005$). The assets are identical, positively correlated ($\rho = 0.5$), with risk premia of 6% ($\mu_1 = \mu_2 = 0.06$) and volatilities equal to 24% ($V_(1,1) = V_(2,2) = 0.24^2$). With this type of costs the non-trading region is shaped by the fixed components of the costs while the optimal trades (outside of the region) are very similar to the ones prescribed by only considering proportional costs.