CHAPTER 12

# Transaction Costs and Portfolio Implementation

Trading stocks incurs transaction costs. So far, we have not dealt explicitly with the impact of transaction costs on equity portfolio management, with the exception of Chapter 8, where we built optimal alpha models under an aggregate portfolio turnover constraint. However, portfolio turnover is just a proxy for transaction costs, which are often stock specific; trading illiquid stocks would have higher costs than trading liquid stocks even if turnover is the same. Therefore, to fully understand the impact of transaction costs on portfolio management, it is important to incorporate stock-level detail in the analysis.

In this chapter, we study two areas of portfolio management that would benefit from the inclusion of transaction costs. One is portfolio construction or portfolio optimization and the other is portfolio implementation. The processes of portfolio optimization with transaction costs and portfolio implementation should be integrated. Simply put, we cannot know the exact transaction costs without knowing exactly how the portfolio would be implemented. In other words, the transaction costs depend on changes of portfolio (in shares or in portfolio weights), as well as the way the portfolio will be traded. If we denote changes in portfolio by the weight differences, $\Delta \mathbf{w} = \mathbf{w} - \mathbf{w}_0$, where $\mathbf{w}_0$ is the initial weight vector and $\mathbf{w}$ is the optimal weight vector, the transaction costs should be a function $c(\Delta \mathbf{w})$, in which the function form $c(.)$ would be determined by how the trades are executed in addition to the liquidity attributes of stocks. After the function $c(.)$ is determined, the transaction cost $c(\Delta \mathbf{w})$ is incorporated into the portfolio optimization process as another term in the objective function.

In practice, the two processes are often studied separately. As a result, some simple transaction cost functions are used in the portfolio optimization. In this book, we follow this research direction and leave the integrated approach to future research.

## 12.1 COMPONENTS OF TRANSACTION COSTS

To determine a reasonable form for function $c(.)$, we first consider the different components of transaction costs. Broadly speaking, there are two kinds of transaction costs: fixed costs and variable costs. The fixed costs are related to trade commissions and bid/ask spreads. There could be additional service fees but they are often included in the commission. Trade commissions are often quoted at some cost per share whether it is a buy or a sell order. For instance, it could be 2¢ per share. In this case, the cost is a linear function of the traded amount or the number of trade tickets.

The bid/ask spread is another form of fixed cost because it results in investors getting paid less if they were to sell a stock, while paying more if they were to buy a stock. For instance, the spread might be \$10.00/\$10.10, meaning a seller receives \$10.00 per share but a buyer has to pay \$10.10, an extra of 10¢ per share. If nothing changes, a round trip of trading would result in a loss of 10¢ per share for the investor. For this reason, we could model the costs associated with the bid/ask spread as half of the spread between the two prices. The average of the bid and ask is called the *mid-quote*, and hence the cost is the difference between either bid or ask and the mid-quote. Because the cost is on a per-share basis, it is also a linear function of the traded amount.

Hence, we can model the fixed cost as a constant vector times the absolute value of the portfolio weight change,

$$c(\Delta\mathbf{w}) = \boldsymbol{\theta}' \cdot |\Delta\mathbf{w}| = \theta_1|\Delta w_1| + \theta_2|\Delta w_2| + \cdots + \theta_N|\Delta w_N|. \qquad (12.1)$$

- The function (12.1) is always positive with the absolute value function if the coefficients are positive. Also, the proportional constant is different for different stocks. This is a result of different commissions, or different bid/ask spreads for different stocks, or both.

**Example 12.1**

Suppose a stock is originally 10% of a portfolio and we want to reduce it to 5%. The size of the portfolio is \$100 million. This results in a trade of \$5 million worth of stock. Suppose the share price is \$50. We thus need to sell 100,000 shares. Let us say assume a bid/ask spread of

10¢ and a commission of 5¢ per share. The transaction costs would be $c = (0.05 + 0.05) \cdot 100{,}000 = \$10{,}000$, or a loss of 0.01%, or 1 basis point, on the total portfolio. In terms of Equation 12.1, the coefficient equals $\theta = 0.002$, which is cost per share at 10¢ divided by the share price at $50. It can be proved that in terms of percentage loss to the total portfolio, the coefficient $\theta$ equals transaction cost per share divided by the share price (Problem 12.1).

The other component of transaction costs is variable costs, which include market impact and opportunity costs. Market impact refers to the price change due to investors' trading and it occurs when trade size exceeds the quote depth currently available. For instance, we would like to sell 100,000 shares of stock in Example 12.1. However, the bid at $50 is only for 50,000 shares. If we want to sell the additional 50,000 rather quickly, the price is most likely to drop due to the resulting supply and demand imbalance and we might have to accept that lower price to fill the order. The difference between the new price and the bid price prior to the sell order gives rise to the market impact component of total transaction costs.

Thus, the transaction costs associated with market impact are not linear. It is small when the trade size is small but it increases dramatically when the trade size becomes large. For a single stock, one possibility is to model it by a square function

$$c(\Delta w_i) = \psi_i (\Delta w_i)^2, \quad \psi_i \geq 0. \tag{12.2}$$

As we shall see shortly, the simplicity of (12.2) makes portfolio optimization easy.

**Example 12.2**

Continue with Example 12.1. Suppose the quote depth is only 50,000 shares at the selling price of $50 and we have to sell the remaining 50,000 shares at the price of $49.80. The total transaction cost is $c = \$0.05 \cdot 100{,}000 + \$0.05 \cdot 50{,}000 + \$0.25 \cdot 50{,}000 = \$20{,}000$, or twenty thousand dollars. This is equivalent to 20¢ per share, a loss of 0.02%, or 2 basis points, on the total portfolio. If we model the total cost using Equation 12.2, then the coefficient is given by

$$\psi_i = \frac{c}{(\Delta w_i)^2} = \frac{0.02\%}{(5\%)^2} = 0.08.$$

When trading multiple stocks, or a basket of stocks, the market impact on the different stocks can be correlated. Selling two highly correlated stocks would cause a greater market impact on both stocks than selling one stock while buying the other. We can model the transaction costs associated with market impact for a basket of stocks using

$$c(\Delta \mathbf{w}) = \Delta \mathbf{w}' \cdot \boldsymbol{\psi} \cdot \Delta \mathbf{w} \ . \tag{12.3}$$

To ensure that the transaction costs are always positive, the matrix $\boldsymbol{\psi}$ must be positive definite.

Another type of variable cost is the opportunity cost, which is associated with the return impact of trades not getting executed. For instance, investors often use limit orders instead of market orders to buy stocks, in order to reduce market impact. However, if the stock price fails to reach the limit order price, the trade would not be executed. If the stock price continues to rise, then the investor loses the opportunity to participate in the gain on the stock. Compared to the other components of transaction costs, the opportunity cost is the hardest to estimate. We shall not consider it in the book.

## 12.2 OPTIMAL PORTFOLIOS WITH TRANSACTION COSTS: SINGLE ASSET

The problem of incorporating transaction costs into the formation of optimal portfolios is often not analytically tractable. We shall discuss numerical methods to solve it later in the chapter. However, for a single stock or asset, it is possible to analyze and solve the problem analytically, and we can gain valuable insights from it.

### 12.2.1 Single Asset with Quadratic Costs

Mean–variance optimization with the addition of quadratic transaction costs is relatively easy to treat so we shall consider it first. The transaction costs are given in the form of (12.2). The optimization problem in this case can be written as

$$\text{maximize} \ \ U(w) = f \cdot w - \frac{1}{2}\lambda\sigma^2 w^2 - \psi\left(w - w_0\right)^2 . \tag{12.4}$$

The unknown is the optimal weight $w$, and the parameters are: $f$, the return forecast; $\sigma$, the risk of the asset; $\lambda$, the risk-aversion parameter;

$w_0$, the initial weight; and $\psi$, the transaction cost coefficient. We can think of (12.4) as the allocation decision between a single risky asset and cash. The coefficient $\psi$ in this case measures market impact of the cost for a 100% turnover. As opposed to the problem with linear transaction cost, the utility function in (12.4) is well behaved. The cost term is analogous to a variance term, relative to the current position. Taking the derivative with respect to $w$ gives rise to

$$U'(w) = f - \lambda\sigma^2 w - 2\psi(w - w_0). \tag{12.5}$$

The optimal weight is given by $U'(w) = 0$, and we have

$$w^* = \frac{f + 2\psi w_0}{\lambda\sigma^2 + 2\psi}. \tag{12.6}$$

The optimal weight (12.6) is a function of the transaction cost coefficient $\psi$. When $\psi = 0$, then

$$w^* = \tilde{w} \triangleq \frac{f}{\lambda\sigma^2}. \tag{12.7}$$

The weight $\tilde{w}$ is optimal when there are no transaction costs. At the other extreme, when $\psi$ is very large compared to both the forecast and the risk term, then $w^* \to w_0$ slowly.

Let $\Delta w^* = w^* - w_0$ be the optimal trade with transaction costs and $\Delta\tilde{w} = \tilde{w} - w_0$ be the optimal trade without transaction costs. Equation 12.8 shows that $\Delta w^*$ is a fraction of $\Delta\tilde{w}$, and the scaling constant is the ratio of the transaction coefficient to the risk coefficient in the utility function (12.4).

$$w^* - w_0 = \frac{f + 2\psi w_0}{\lambda\sigma^2 + 2\psi} - w_0 = \frac{f - \lambda\sigma^2 w_0}{\lambda\sigma^2 + 2\psi} = \frac{\tilde{w} - w_0}{1 + \left(2\psi/\lambda\sigma^2\right)}. \tag{12.8}$$

**Example 12.3**

Suppose that a single asset has a volatility $\sigma$ is 15%, and we have a return forecast of 15%. The risk-aversion parameter is 10, and the current position is 50%. We can calculate the optimal weight with no transaction costs at

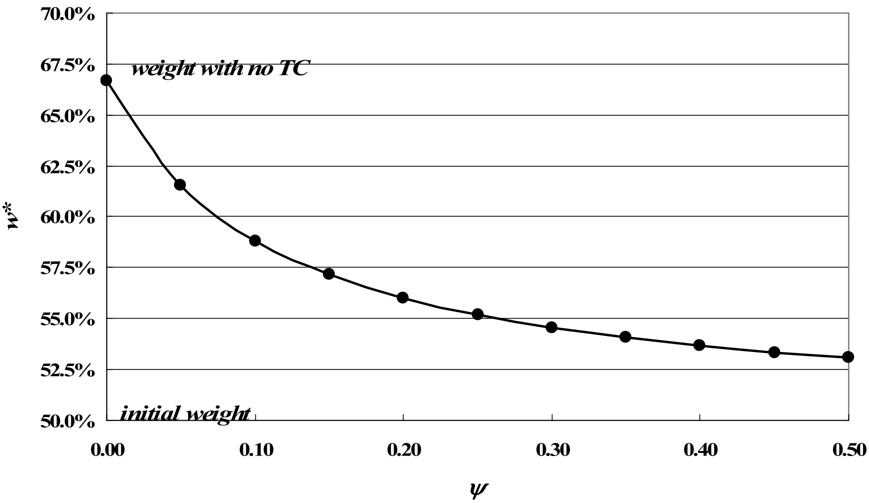


**FIGURE 12.1.** Optimal weight of a single asset with quadratic transaction costs. The initial weight is 50%, and the optimal weight with no transaction costs is 66.7%. Note that the optimal weight is always above the initial weight.

$$\tilde{w} = \frac{f}{\lambda\sigma^2} = \frac{0.15}{10(0.15)^2} = 66.7\% .$$

Therefore, we should be buying more. However, the amount of buying will be tempered by the transaction costs. Suppose $\psi = 0.1$, which corresponds to transaction costs of 10% on 100% turnover. We then have

$$w^* = \frac{f + 2\psi w_0}{\lambda\sigma^2 + 2\psi} = \frac{0.15 + 2(0.1)(0.5)}{10(0.15)^2 + 2(0.1)} = 58.5\% .$$

Figure 12.1 plots the optimal weights for value of $\psi$ from 0 to 0.5. As we can see, the optimal weight declines rather quickly at first, and then the rate of decline slows. When $\psi = 0.5$, the optimal weight is about 53%, a trade of 3%. Note the following remark:

- With quadratic trading costs, there will always be some trading no matter how large $\psi$ is, because the value of the quadratic function of transaction costs will be small when the weight is close to the initial weight. This makes some sense, because the market impact only becomes important when the trade size exceeds the quote depth.

### 12.2.2 Single Asset with Linear Costs

We now consider mean–variance optimization with the addition of transaction costs given in the form of (12.1). The optimization problem in this case can be written as

$$\text{maximize } U(w) = f \cdot w - \frac{1}{2}\lambda\sigma^2 w^2 - \theta\left|w - w_0\right|. \tag{12.9}$$

$\theta$ is the transaction cost coefficient, measuring the cost of 100% turnover. Solving Problem 12.9 poses certain analytical challenges because the absolute value function is not differentiable at the origin.

When there are no transaction costs, i.e., $\theta = 0$, however, the optimal weight is $\tilde{w}$, given by (12.7). When $\theta > 0$, the problem can be formulated in terms of weight change: $\Delta w = w - w_0$. Using $w = w_0 + \Delta w$, we can rewrite the utility function as

$$\begin{aligned} U(\Delta w) &= f \cdot (\Delta w + w_0) - \frac{1}{2}\lambda\sigma^2(\Delta w + w_0)^2 - \theta\left|\Delta w\right| \\ &= U(w_0) + \left[\lambda\sigma^2(\tilde{w} - w_0)\Delta w - \theta\left|\Delta w\right| - \frac{1}{2}\lambda\sigma^2(\Delta w)^2\right] \end{aligned} \tag{12.10}$$

The total utility is a sum of the current utility, a constant, given by

$$U(w_0) = fw_0 - \frac{1}{2}\lambda\sigma^2 w_0^2,$$

and the change in utility caused by the change in weight. The weight $\tilde{w}$ is also a constant given by Equation 12.7.

The change in utility is then

$$\Delta U = U(\Delta w) - U(w_0) = \lambda\sigma^2 \Delta\tilde{w}\Delta w - \theta\left|\Delta w\right| - \frac{1}{2}\lambda\sigma^2(\Delta w)^2, \quad \text{with } \Delta\tilde{w} = \tilde{w} - w_0 \tag{12.11}$$

The optimal weight change must maximize the change in utility, which is zero when $\Delta w = 0$. In other words, at a minimum, we can maintain the current utility with no trading. To find the maximum, we now consider three cases.

The first case is when $\tilde{w} = w_0$, i.e., when the optimal weight disregarding the transaction costs is equal to the initial weight. It is obvious in this case we should not trade at all. Mathematically, $\Delta w = 0$ is the optimal solution for utility (12.10), because any trading would cause the utility to go down.

When $\tilde{w} \neq w_0$, the initial position is not optimal, at least if there were no transaction costs. There is a possibility that we can increase the utility of (12.10) by trading. Because both the second and the third terms, associated with transaction costs and variance, are negative whenever there is trading (either buy or sell), the trading must at least make the first term positive. This implies $\Delta w$ must be of the same sign as $\Delta\tilde{w} = \tilde{w} - w_0$. Therefore, in the second case, we consider $\tilde{w} > w_0$, i.e., the optimal weight in absence of transaction costs is greater than the initial weight, indicating buy. As argued, we should look for solution $\Delta w \geq 0$. In other words, we should look to buy to increase the utility.

If $\Delta w \geq 0$, we have $|\Delta w| = \Delta w$. The utility function becomes differentiable with the derivative

$$U'(\Delta w) = \lambda\sigma^2 \Delta\tilde{w} - \theta - \lambda\sigma^2(\Delta w). \tag{12.12}$$

Setting $U'(\Delta w) = 0$ yields

$$\Delta w^* = w^* - w_0 = \Delta\tilde{w} - \frac{\theta}{\lambda\sigma^2} = \Delta\tilde{w} - w_c. \tag{12.13}$$

We have defined

$$w_c = \frac{\theta}{\lambda\sigma^2}, \tag{12.14}$$

which is an optimal weight associated with the transaction cost as a negative "alpha," or cost weight.

Equation 12.13 is the optimal weight *if* $\Delta w^*$ is greater than or equal to zero, or when

$$\Delta\tilde{w} \geq w_c . \tag{12.15}$$

This condition implies that we would only buy when the costless buying, i.e., $\Delta\tilde{w}$, exceeds the cost weight $w_c$. On the other hand, *when* $\Delta w^*$ is less than zero, the costless buying does not clear the hurdle of cost weight, then (12.13) is certainly not the optimal weight, because it leads to a reduction in utility (12.10). Here, we have a situation in which we would buy if there were no transaction costs, but would not if the transaction cost were factored in. The best course to follow is therefore to stay put: no trade, i.e., $\Delta w^* = 0$.

The analysis applies equally to the last case, in which $\tilde{w} < w_0$. We leave it as an exercise. To summarize the results, we have the optimal trading

$$\Delta w^* = \begin{cases} \Delta\tilde{w} - w_c, & \text{when } \Delta\tilde{w} > w_c \\ 0, & \text{when } \left|\Delta\tilde{w}\right| \leq w_c \\ \Delta\tilde{w} + w_c, & \text{when } \Delta\tilde{w} < w_c \end{cases} \tag{12.16}$$

Figure 12.2 shows the results. Both buys and sells are reduced by the amount, $w_c$, and there is a zone of inaction when the costless trading is less than the cost weight.

Alternatively, we can rewrite the optimal weight as

$$w^* = \frac{f - \theta}{\lambda\sigma^2} \geq w_0 . \tag{12.17}$$

Note that the optimal weight $w^*$ is equivalent to an optimal solution in the case of no transaction costs, but with an adjusted forecast of $f - \theta$. Therefore, we would buy only if the forecast is high enough to offset the transaction costs, such that the optimal weight with the cost-adjusted forecast is still greater than the current weight. Note the following remark:

- The insight from the analysis is that we buy only if the cost-adjusted forecast, $f - \theta$, still leads to a buy decision. In other words, we trim the forecast of a possible buy by the transaction cost, and the adjusted optimal weight must still be higher than the current weight in order for us to trade. In the same vein, we sell only if the cost-adjusted forecast, $f + \theta$, in the case of a sell (see Problem 12.2), still leads to

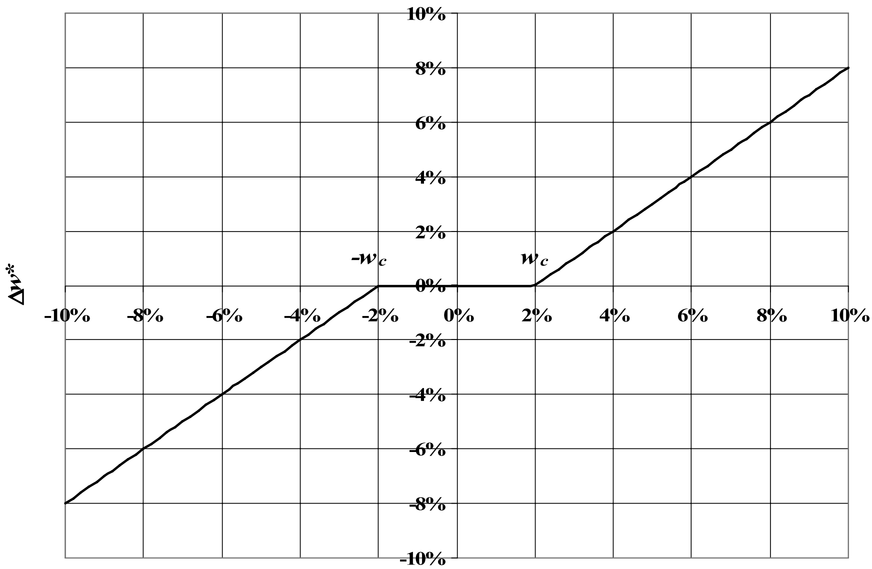


**FIGURE 12.2.** Relationship among the optimal trading $\Delta w^*$, the costless trading $\Delta\tilde{w}$, and the cost weight $w_c$ when transaction cost is a linear function with respect to the size of a trade.

a sell. In other words, we raise the forecast of a possible sell by the transaction cost and the adjusted optimal weight must still be lower than the current weight in order for us to sell. If these conditions are not met, then there is no trade.

**Example 12.4**

We use the same parameters as in Example 12.3: a single asset with volatility $\sigma$ at 15%, and return forecast of 15%. The risk-aversion parameter is 10, and the current position is 50%. The optimal weight with no transaction costs is 66.7%. Therefore, we should be buying more. However, the amount of buying will be tampered by the transaction costs. Suppose $\theta = 0.01$, then the optimal weight is

$$w^* = \frac{f-\theta}{\lambda\sigma^2} = \frac{0.15-0.01}{10(0.15)^2} = 62.2\%.$$

The weight is still above the current weight, by 10.2%. If the transaction cost is increased to $\theta = 0.02$, then the optimal weight decreases to 57.8%.

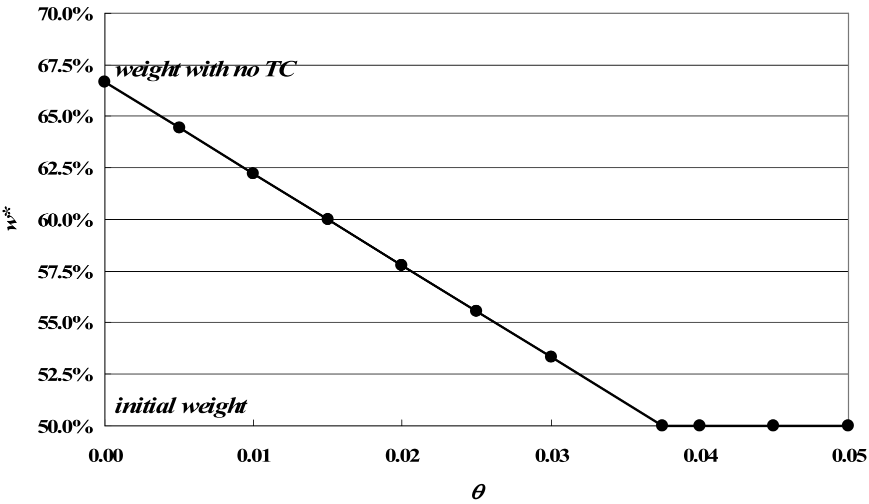


**FIGURE 12.3.** Optimal weight of a single asset with linear transaction costs. The initial weight is 50%, and the optimal weight with no transaction costs is 66.7%. There is no trading when the transaction costs goes beyond a critical value.

Therefore, we are buying less as the costs get higher. The critical value is $\theta = 0.0375$, at which the optimal weight becomes the current weight at 50%.

Figure 12.3 plots the optimal weights for values of $\theta$ from 0 to 0.05. As we can see, the optimal weight declines linearly and it reaches the initial weight when $\theta$ hits the critical value of 0.0375 and stays there.

## 12.3 OPTIMAL PORTFOLIOS WITH TRANSACTION COSTS: MULTIASSETS

Having solved the problem of the optimal weight for a single asset, we now analyze the problem for multiasset portfolios.

### 12.3.1 Multiasset with Quadratic Costs

With a multiasset portfolio, the quadratic transaction cost is given in the form of (12.3), in which $\Delta\mathbf{w} = \mathbf{w} - \mathbf{w}_0$. The optimization problem in this case can be written as

$$\text{maximize } U(\mathbf{w}) = \mathbf{f}' \cdot \mathbf{w} - \frac{1}{2}\lambda \mathbf{w}'\boldsymbol{\Sigma}\mathbf{w} - (\Delta\mathbf{w})' \boldsymbol{\psi}(\Delta\mathbf{w}). \qquad (12.18)$$

Note that for an active portfolio vs. a benchmark, the weight vector is the active weights and for a market-neutral long/short portfolio the weight vector is the absolute weights. We have left out other constraints to isolate the impact of transaction costs.

The solution of (12.18) can be found analytically using the following equation:

$$\frac{\partial U}{\partial \mathbf{w}} = \mathbf{f} - \lambda \mathbf{\Sigma} \mathbf{w} - 2\mathbf{\Psi}\left(\mathbf{w} - \mathbf{w}_0\right) = \mathbf{0}\,. \tag{12.19}$$

We have

$$\mathbf{w}^* = \left(\lambda \mathbf{\Sigma} + 2\boldsymbol{\psi}\right)^{-1}\left(\mathbf{f} + 2\boldsymbol{\psi}\mathbf{w}_0\right). \tag{12.20}$$

In (12.20), both $\mathbf{\Sigma}$ and $\boldsymbol{\psi}$ are square matrices and $\mathbf{f}$ is the forecast vector. Note that it reduces to (12.6) when both matrices are diagonal. In that case, we are simply optimizing uncorrelated individual assets.

### 12.3.2 Portfolio Dynamics

Equation 12.20 gives rise to a dynamic relationship of portfolio weights over time. Applying (12.20) iteratively, we have

$$\mathbf{w}_t = \left(\lambda \mathbf{\Sigma} + 2\boldsymbol{\psi}\right)^{-1}\left(\mathbf{f}_t + 2\boldsymbol{\psi}\mathbf{w}_{t-1}\right) \tag{12.21}$$

and

$$\begin{aligned} \mathbf{w}_t &= \left(\lambda \mathbf{\Sigma} + 2\boldsymbol{\psi}\right)^{-1}\begin{bmatrix} \mathbf{f}_t + 2\boldsymbol{\psi}\left(\lambda \mathbf{\Sigma} + 2\boldsymbol{\psi}\right)^{-1}\mathbf{f}_{t-1} \\ +\left(2\boldsymbol{\psi}\right)\left(\lambda \mathbf{\Sigma} + 2\boldsymbol{\psi}\right)^{-1}\left(2\boldsymbol{\psi}\right)\mathbf{w}_{t-2} \end{bmatrix} \\ &= \left(\lambda \mathbf{\Sigma} + 2\boldsymbol{\psi}\right)^{-1}\left[\mathbf{f}_t + \mathbf{A}\mathbf{f}_{t-1} + \mathbf{A}^2\mathbf{f}_{t-2} + \cdots + \mathbf{A}^{\tau}\mathbf{f}_{t-\tau} + \cdots\right] \end{aligned} \tag{12.22}$$

The matrix $\mathbf{A}$ is defined as

$$\mathbf{A} = \left(\lambda \mathbf{\Sigma} + 2\boldsymbol{\psi}\right)^{-1} 2\boldsymbol{\psi}\,.$$

Based on this relationship, one can build a dynamic model of active portfolios over time, supplemented by a dynamic model of forecasts

$$\mathbf{f}_t = \mathbf{P}_1\mathbf{f}_{t-1} + \mathbf{P}_2\mathbf{f}_{t-2} + \cdots + \mathbf{P}_p\mathbf{f}_{t-p} + \boldsymbol{\varepsilon}_t \tag{12.23}$$

and lagged ICs

$$IC_{t-p,t} = \text{corr}\left(\mathbf{f}_{t-p}, \mathbf{r}_t\right). \tag{12.24}$$

Sneddon (2005) has shown that under simplified assumptions, one can derive the multiperiod information ratio (IR) in a semi-analytical framework that gives valuable insights regarding the combination of forecast signals. His results are consistent with our finding in Chapter 8 (see Grinold 2006 for additional analysis on this topic). For instance, he finds that when incorporating transaction costs, the multiple-period IR can be increased, compared to that of a single-period IR given by the fundamental law of active management, by overweighting the tortoise — signals with lower information coefficient (IC) but slow information decay — and underweighting the hare — signals with higher IC but fast information decay. It remains to be seen if his model can be extended to include more realistic factor and return structures.

### 12.3.3 Multiasset with Linear Costs: Mathematical Formulation

The linear transaction cost of a multiasset portfolio is given previously in (12.1). In terms of a vector of the transaction cost coefficients, $\boldsymbol{\theta}$, and the vector of absolute value of weight changes, $\left|\mathbf{w}-\mathbf{w}_0\right|$, the cost is $\boldsymbol{\theta}'\cdot\left|\mathbf{w}-\mathbf{w}_0\right| = \boldsymbol{\theta}'\cdot\left|\Delta\mathbf{w}\right|$. Thus, the mean–variance cost optimization is

$$\text{maximize } U\left(\mathbf{w}\right) = \mathbf{f}'\cdot\mathbf{w} - \frac{1}{2}\lambda\mathbf{w}'\boldsymbol{\Sigma}\mathbf{w} - \boldsymbol{\theta}'\cdot\left|\Delta\mathbf{w}\right|. \tag{12.25}$$

Unlike the single-asset case, the problem is not analytically tractable unless all assets are uncorrelated: when the covariance matrix is diagonal, because of the presence of the absolute-value function.

The problem can be solved numerically, however, in a number of ways. For example, one can approximate the absolute-value function by some smooth functions. In this chapter we shall present a method that reformulates the transaction cost term in term of two new variables, buys and sells, and solve the reformulated problem with standard quadratic programming.

We define two new vectors, buy vector $\mathbf{w}_B$ and sell vector $\mathbf{w}_S$. Then the new portfolio weights are a combination of the current weights, the buys and the sells

$$\mathbf{w} = \mathbf{w}_0 + \mathbf{w}_B - \mathbf{w}_S. \tag{12.26}$$

Both the buys and the sells are nonnegative, $\mathbf{w}_B \geq 0,\ \mathbf{w}_S \geq 0$, i.e., all elements of the two vectors are either positive or zero. It is also noted that the buys and sells are mutually exclusive: for every stock we either have a buy or sell but never both. These properties enable us to replace the absolute value of weight change by

$$\left|\Delta\mathbf{w}\right| = \mathbf{w}_B + \mathbf{w}_S. \tag{12.27}$$

Substituting both (12.26) and (12.27) into (12.25), we have

$$\begin{aligned}
U\left(\mathbf{w}\right) &= \mathbf{f}'\cdot\left(\mathbf{w}_0+\mathbf{w}_B-\mathbf{w}_S\right)-\frac{1}{2}\lambda\left(\mathbf{w}_0+\mathbf{w}_B-\mathbf{w}_S\right)'\boldsymbol{\Sigma}\left(\mathbf{w}_0+\mathbf{w}_B-\mathbf{w}_S\right)\\
&\quad -\boldsymbol{\theta}'\cdot\left(\mathbf{w}_B+\mathbf{w}_S\right)\\
&= U\left(\mathbf{w}_0\right)+\left(\mathbf{f}-\lambda\boldsymbol{\Sigma}\mathbf{w}_0-\boldsymbol{\theta}\right)'\cdot\mathbf{w}_B+\left(-\mathbf{f}+\lambda\boldsymbol{\Sigma}\mathbf{w}_0-\boldsymbol{\theta}\right)'\cdot\mathbf{w}_S\\
&\quad -\frac{1}{2}\lambda\left(\mathbf{w}_B'\boldsymbol{\Sigma}\mathbf{w}_B-2\mathbf{w}_B'\boldsymbol{\Sigma}\mathbf{w}_S+\mathbf{w}_S'\boldsymbol{\Sigma}\mathbf{w}_S\right)
\end{aligned} \tag{12.28}$$

As before, the initial utility is

$$U\left(\mathbf{w}_0\right) = \mathbf{f}'\cdot\mathbf{w}_0-\frac{1}{2}\lambda\mathbf{w}_0'\boldsymbol{\Sigma}\mathbf{w}_0 .$$

The objective function of (12.28) can be written in terms of a stacked vector, which combines both buys and sells, i.e.,

$$\mathbf{W} = \begin{pmatrix}\mathbf{w}_B\\ \mathbf{w}_S\end{pmatrix}, \tag{12.29}$$

and a stacked forecast vector

$$\mathbf{F} = \begin{pmatrix}\mathbf{f}-\lambda\boldsymbol{\Sigma}\mathbf{w}_0-\boldsymbol{\theta}\\ -\mathbf{f}+\lambda\boldsymbol{\Sigma}\mathbf{w}_0-\boldsymbol{\theta}\end{pmatrix}, \tag{12.30}$$

and an augmented covariance matrix

$$\boldsymbol{\Sigma}_2 = \begin{pmatrix} \boldsymbol{\Sigma} & -\boldsymbol{\Sigma} \\ -\boldsymbol{\Sigma} & \boldsymbol{\Sigma} \end{pmatrix}. \tag{12.31}$$

Combining the equations preceding, we have

$$U(\mathbf{w}) = U(\mathbf{w}_0) + \mathbf{F}' \cdot \mathbf{W} - \frac{1}{2}\lambda \mathbf{W}' \cdot \boldsymbol{\Sigma}_2 \cdot \mathbf{W}. \tag{12.32}$$

The optimization problem with objective function (12.32) can be solved numerically using quadratic programming.

Several constraints can be placed on the augmented weight vector $\mathbf{W}$ to address practical implementation concerns. The first constraint is $\mathbf{W} \geq 0$. Another constraint is related to dollar neutrality; i.e., the total amount of buys and sells should balance. This is a linear equality constraint

$$\mathbf{w}_B' \cdot \mathbf{i} = \mathbf{w}_S' \cdot \mathbf{i}, \text{ or } \mathbf{W}' \cdot \hat{\mathbf{i}} = 0.$$

The vector $\mathbf{i}$ is a vector of ones, of length $N$, and

$$\hat{\mathbf{i}} = \begin{pmatrix} \mathbf{i} \\ -\mathbf{i} \end{pmatrix}.$$

If desired, we can add the turnover constraint as

$$\mathbf{W}' \cdot \mathbf{i}_2 \leq T, \text{ with } \mathbf{i}_2 = \begin{pmatrix} \mathbf{i} \\ \mathbf{i} \end{pmatrix}.$$

$T$ is the maximum turnover allowed and $\mathbf{i}_2$ is a vector of ones, of length $2N$.

Finally, we can require range constraints on the optimal weights

$$\mathbf{l} \leq \mathbf{w} = \mathbf{w}_0 + \mathbf{w}_B - \mathbf{w}_S \leq \mathbf{u}, \tag{12.33}$$

in terms of the augmented weight vector $\mathbf{W}$. This is left as an exercise. Note the following:

- We have not imposed the condition that the buys and the sells are mutually exclusive on the new optimization problem. There is no need to do that because that would certainly result in a suboptimal solution. It is easy to see this in a single-asset case. Suppose both $w_B$ and $w_S$ are positive; then, the new weight defined by the netting of the two would achieve a higher value of utility. For example, let $w_B \geq w_S > 0$, then $w_B' = w_B - w_S$ and $w_S' = 0$ increases the utility, because it has the same mean and variance but less transaction costs.

- The augmented covariance matrix (12.31) is singular, but this is not necessarily an issue for quadratic programming. The matrix can be modified using the fact that the buys and the sells are mutually exclusive, i.e., $w_{B,i} w_{S,i} = 0$ for every stock. Consequently, we can set the diagonal elements of both $\left(-\boldsymbol{\Sigma}\right)$ matrices — upper-right and bottom-left corners in (12.31) — to zeros.

### 12.3.4 Multiasset with Linear Costs: Numerical Example 1

We apply the numerical method to a portfolio of 20 stocks. We start with a market neutral long/short initial portfolio. We then simulate a vector of forecasts and use the forecasts to rebalance the portfolio, incorporating transaction costs. Other inputs are the covariance matrix $\boldsymbol{\Sigma}$ and the transaction cost coefficient $\boldsymbol{\theta}$. For simplicity, we take $\boldsymbol{\Sigma}$ as a diagonal matrix with specific risk of 35% for all stocks. The transaction cost is assumed to be 2% for all stocks. All portfolios, initial and optimized, have a target tracking error of 10%. The forecasts are products of IC, $z$-score, and specific risk. We will let IC = 0.2, and the $z$-scores have 0 mean and standard deviation 1.

Figure 12.4 plots the forecasts vs. the initial portfolio weights (in solid squares) and the optimal portfolio with maximum turnover. As we can see, whereas the initial weights are in general agreement with the forecasts, they are not aligned perfectly. For instance, a stock with a forecast of –3.2% has a weight of 10.3%, whereas another stock with a forecast of 11.2% has a weight of –1.9%. The overall correlation between the forecasts and the initial weights is only 0.48, and the expected return is 4.2%.

The optimal weights are the solution of (12.32) without the turnover constraint. The resulting one-way turnover is about 36%. As we can see, the forecasts and the optimal weights are aligned almost perfectly, with a correlation of 0.97. The only reason that they do not lie on a straight line

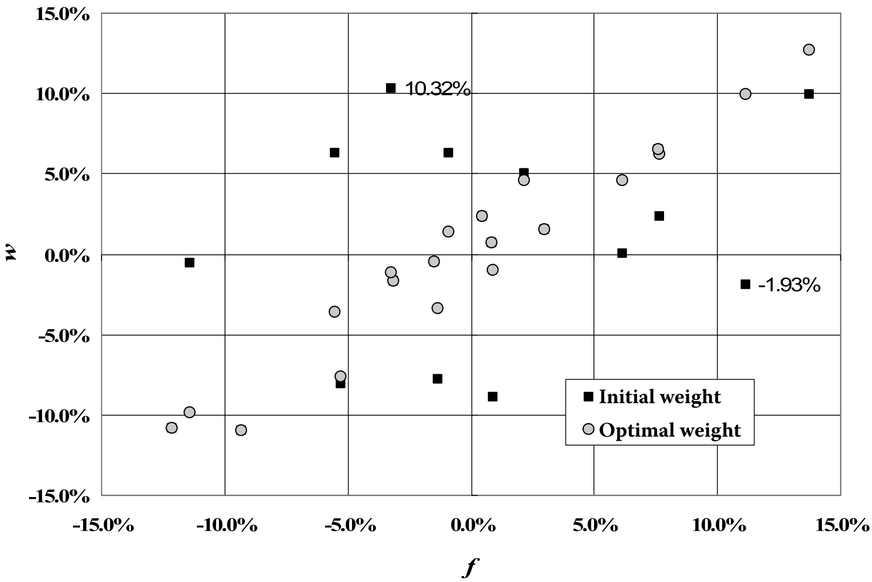


**FIGURE 12.4.** Scatter plot of forecasts vs. initial weights and optimal weights with maximum portfolio turnover.

is due to the $\theta = 2\%$ transaction costs we imposed. The expected return is 8.5% gross of transaction cost and 7.0% net of transaction costs. The gross return is simply the sum of weights times the expected returns and the net return is the gross return minus the transaction costs, $\theta$ times two-way turnover. It is also worth noting that out of the 20 stocks, only 10 stocks, those whose initial weights are too deviated from the optimal weights, show any meaningful weight change. The other 10 stocks are prevented from trading due to the transaction costs.

Imposing additional turnover constraints impacts on optimal weights and expected returns. Figure 12.5 shows the gross and net expected returns as a function of allowed turnover. When no turnover is permitted, both returns are the same as the return of the initial portfolio. As we allow more and more turnover, both returns increase, with the gap between the two widening as the costs increases.

- Note that the rate of increase in the net return slows down as the turnover increases. As a result, when the turnover is 20%, the net return is 6.5%, an increase of 2.3% from the initial 4.2%. This represents a roughly 80% total increase in net return, with about 55% of total turnover.

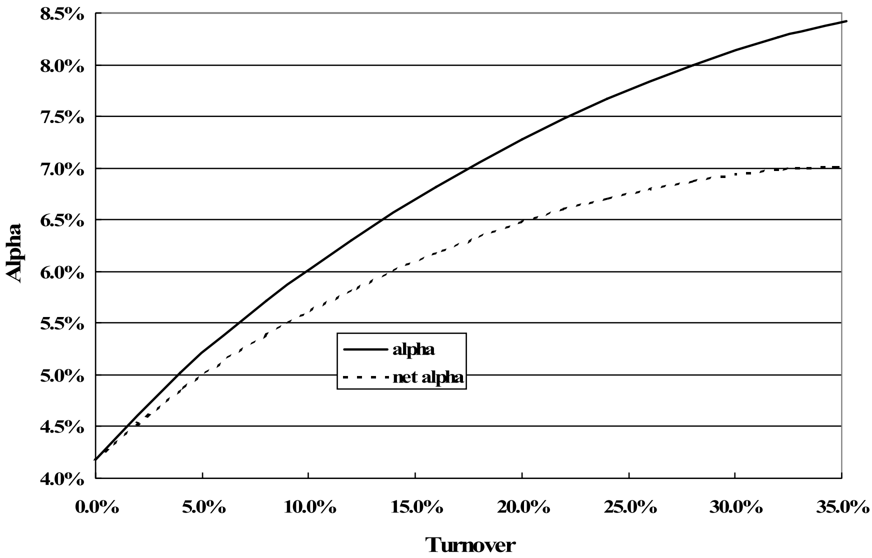


**FIGURE 12.5.** The gross and net expected returns as a function of allowed portfolio turnover.

Figure 12.6 shows the change in portfolio weights from the initial portfolio weights. If $\Delta w > 0$, we buy the stock, whereas if $\Delta w < 0$, we sell the stock. As noted before, only ten stocks show weight changes if maximum turnover is allowed. As we see from Figure 12.6, this number is smaller when the turnover is constrained. For example, at 4% turnover, only the two stocks that are marked in Figure 12.4 are traded. The limited turnover budget is allocated to them, because their positions are most inconsistent with their return projection and trading them increases portfolio alpha the most. As the turnover limit is increased, the trade list expands and the trade sizes expand for stocks that are already on the list.

- We note that the size of buys and sell are monotonic functions of the turnover. If we were to buy a stock, we would buy more if more turnover is allowed up to optimal weight.

### 12.3.5 Multiasset with Linear Costs: Numerical Example 2

In the second example, we study the impact of transaction costs on the optimal weights by varying the level of $\theta$, which is the same for all 20 stocks. For each $\theta$, the optimal portfolio is constructed without additional turnover constraints. Hence, the resulting turnover is the maximum turnover associated with the given transaction costs.

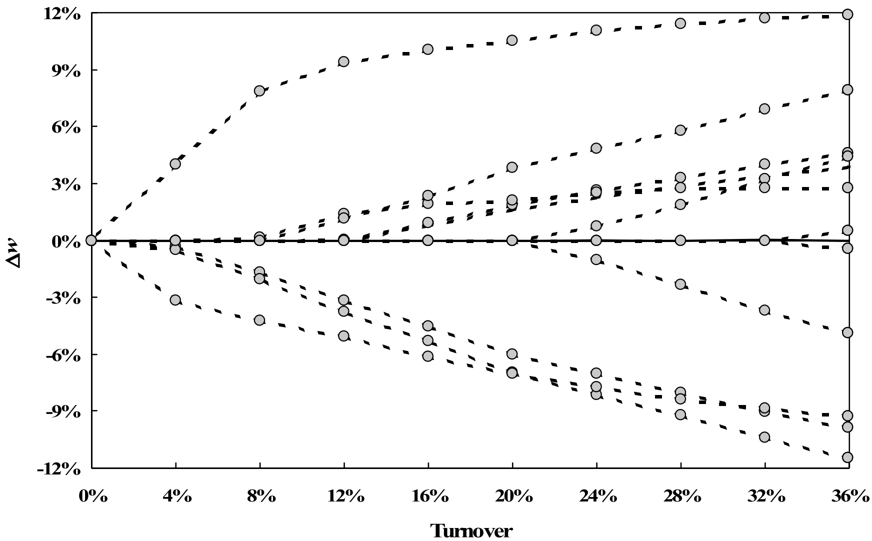


**FIGURE 12.6.** The change of optimal weights from the initial weights as the turnover is increased. Out of 20 stocks, 10 show no weight change; they all lie on the line $\Delta\mathbf{w} = 0$. The remaining 10 stocks show increasing change in weight as more turnovers are permitted.

Figure 12.7 shows the change of the optimal weights from the initial weight, which is the same for all levels of $\theta$, when the transaction costs increase. When $\theta = 0$, i.e., the problem is transaction-cost free, the weight changes are at their maximum for both buys and sells. The difference is just essentially $\Delta\mathbf{w} = \tilde{\mathbf{w}} - \mathbf{w}_0$. As $\theta$ increases, the weight changes for all the stocks shrink toward 0.

- We note that the decline in weight changes follows different patterns for different stocks. Some of them follow a straight line with differing slopes, whereas others are piecewise linear. This feature reflects the nonlinear nature of the objective function and its solution.

Another noteworthy feature of Figure 12.7 is that all weight changes have the same signs as those for $\theta = 0$. In other words, if a stock is a buy (sell) from the optimization with no transaction costs, then it will be a buy (sell) in the optimization with transaction costs. If this is true, it points to an alternative method of constructing an optimal portfolio with transaction costs, using a two-step approach. In the first step, we run an optimization without transaction costs. This is relatively simple as we do not encounter the absolute value function in the objection function (12.25).

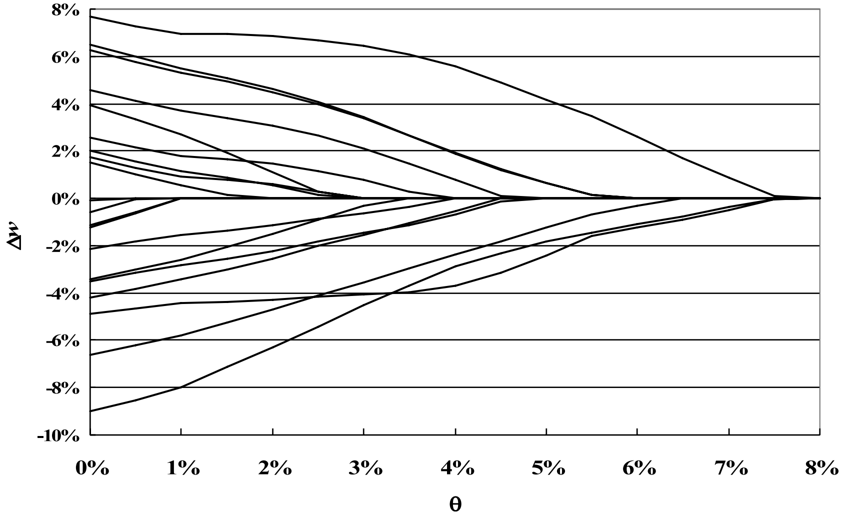


**FIGURE 12.7.** The difference between the optimal weights and the initial weights for varying levels of transaction costs $\boldsymbol{\theta}$.

The solution of this step would provide us a buy list and a sell list. In the second step, we optimize again but with prescribed transaction costs. With the buy and sell lists available, we can now specify the range of optimal weights as $w \geq w_0$ for a buy and $w \leq w_0$ for a sell. The associated transaction costs will be $w - w_0$ for a buy and $w_0 - w$ for a sell. Consequently, we remove the difficulty of dealing with the absolute value function in the objective function. The resulting optimization problem can be solved routinely. However, we caution readers that this may not always be the case.

## 12.4 PORTFOLIO TRADING STRATEGIES

Once optimal portfolio weights are determined, the changes from the initial portfolio weights are the resulting trades that need to be implemented. The goal of portfolio trading strategies is to implement the trades in the most efficient manner. In certain cases, it might be optimal to not implement the full trades, due to either decay in return signals or high transaction costs. In practice, this can also arise due to the use of limit orders, which might not be triggered by price movement resulting in opportunity costs. We shall not consider such cases in our treatment and require all trades to be implemented in the portfolio strategies.

There are at least two conflicting objectives in the portfolio implementation process. On the one hand, one would like to implement the changes as soon as possible to get to the optimal portfolio. The optimal portfolio has the maximum expected return for a specific risk target. Any delay could potentially result in a loss of return, and both the expectation and the variance of that potential loss grow over time. On the other hand, transaction costs from market impact are a direct function of the speed with which the trades are executed. For large trade sizes, immediate execution would cause the greatest market impact. Breaking it in pieces and trading them over an extended period of time would reduce the market impact but at the risk of return loss and tracking error mismatch versus the optimal portfolio, as well as higher fixed costs such as commissions and fees.

For a portfolio of stocks to be traded with both buys and sells, one must consider the trade basket as a whole. For instance, an imbalance between buys and sells might cause an intended net market exposure. The correlation between different stocks is another important issue. For buys and sells that are highly correlated in terms of stock returns, one would like to synchronize the trades, because doing so would reduce systematic exposure. However, if these trades have different market impacts, one would like to execute them at different speeds to minimize the transaction cost. It is therefore necessary to find a balance between the two.

The trading horizon — the length of time we allocate to implement the trades — is another important factor. For trades that are easy to implement based on liquidity, the trading horizon should be short. For difficult trades, the trading horizon can be longer. For a given set of trades, it is better to optimize the trading horizon as well as the actual trade implementation.

## 12.5 OPTIMAL TRADING STRATEGIES: SINGLE STOCK

The problem of optimal trading strategies can be formulated mathematically through an optimization in which the objective function consists of expected return shortfall, return variance, and transaction costs. Grinold and Kahn (2000) considered this problem in continuous time and Almgren and Chriss (2000) used a discrete setting for their analysis. We shall work with the continuous-time case for simplicity in the notations.

We start with the case of a single stock for which the trade is denoted by $\Delta w$. Suppose the trade will be carried out over the horizon $\left[0,T\right]$. We denote the state of the trade at time $t$ in proportion of the total trade: $h(t)\Delta w$, with $h(0)=0$ and $h(T)=1$. The trade shortfall is $h(t)\Delta w-\Delta w=\Delta w\left[h(t)-1\right]$. Suppose the stock's expected return over the horizon is a constant $f$; then the return shortfall is $f\Delta w\left[h(t)-1\right]$. Denoting the stock's risk by $\sigma$, the shortfall variance is $\sigma^2\left(\Delta w\right)^2\left[h(t)-1\right]^2$. We model the transaction costs by two terms, one related to the fixed cost and the other related to the market impact. The fixed cost is assumed to be $-c\left|\Delta w\right|T$ (change in the term), with $c>0$. It is easy to see that the cost is proportional to the trade size. What is new here is that the cost will be proportional to the trading horizon; the longer the horizon, the more often we have to trade (at smaller sizes) and the more we have to pay for fixed costs such as commissions and fees. Finally, we approximate the cost of market impact as being proportional to the square of trading speed, or the derivative of holding: $\left(\Delta w\right)^2\left[\dot{h}\left(t\right)\right]^2$. Combining all four terms and integrating over the time interval $\left[0,T\right]$ gives the objective function

$$
\begin{aligned}
J=&\int_0^T f\Delta w\left[h\left(t\right)-1\right]dt-\frac{1}{2}\lambda\int_0^T \sigma^2\left(\Delta w\right)^2\left[h\left(t\right)-1\right]^2dt-c\left|\Delta w\right|\\
&\int_0^T dt-\psi\int_0^T\left(\Delta w\right)^2\left[\dot{h}\left(t\right)\right]^2dt
\end{aligned}
\tag{12.34}
$$

The additional two parameters are $\lambda$ (the risk-aversion parameter) and $\psi$ (the cost coefficient for market impact). We can simplify (12.34) by scaling it by a positive term $(\Delta w)^2$,

$$
\frac{J}{\left(\Delta w\right)^2}=\int_0^T\left\{f_w\left[h\left(t\right)-1\right]-c_w-\psi\left[\dot{h}\left(t\right)\right]^2-\frac{1}{2}\lambda\sigma^2\left[h\left(t\right)-1\right]^2\right\}dt\,. \tag{12.35}
$$

We have $f_w=f/\left(\Delta w\right)$ and $c_w=c/\left|\Delta w\right|>0$. The goal of optimal trading strategies is to find the solution $h\left(t\right)$ that maximizes (12.35). Note the following:

- Depending on the forecast and the direction of the trade, $f_w=f/\left(\Delta w\right)$ can be zero, positive, or negative. It is zero when the forecast is zero. In this case, the objective function is the same for both buy orders ($\Delta w>0$) and sell orders ($\Delta w<0$). When the forecast is nonzero, the term $f_w=f/\left(\Delta w\right)$ is positive when both have the same sign: buy with a positive forecast or sell with a negative forecast. It is negative when both have opposite signs: buy with a negative forecast or sell with a positive forecast.

- The first three terms of (12.35) are all implementation costs, alpha or transaction costs — whereas the last term is implementation risk. The problem of optimal trading strategies is thus similar to a mean–variance problem of portfolio construction. For a given level of implementation risk, there exists an optimal solution with minimum implementation costs. Similar to the efficient frontier of mean–variance optimization, the optimal trading strategies for varying implementation risks form an efficient risk-cost frontier.
- The *fixed* term has been missing in previous work in optimal trading strategies. Because it is always a cost and it increases with $T$, it has the effect of shortening the optimal trading horizon when we allow $T$ to be free later in the chapter.

### 12.5.1 Optimal Solution with Fixed Trading Horizon

We first treat the trading horizon $T$ as fixed, i.e., the amount of time needed to execute a trade has been determined, maybe by some heuristic estimation or based on traders' experience. We will now solve for the optimal solution $h(t)$ for $t$ in [0,1]. In the next section, we shall also find the optimal trading horizon.

The mathematical technique for solving this type of optimization problem is the calculus of variation. Denote the integrand of (12.35) by

$$L\left(h,\dot{h}\right)=f_w\left[h\left(t\right)-1\right]-c_w-\psi\left[\dot{h}\left(t\right)\right]^2-\frac{1}{2}\lambda\sigma^2\left[h\left(t\right)-1\right]^2. \quad (12.36)$$

Then the solution is given by the following differential equation

$$\frac{d}{dt}\left[\frac{\partial L}{\partial \dot{h}}\right]=\frac{\partial L}{\partial h} \quad (12.37)$$

From (12.36), we have

$$\begin{aligned}\frac{\partial L}{\partial \dot{h}}&=-2\psi\dot{h}\\ \frac{\partial L}{\partial h}&=f_w-\lambda\sigma^2\left(h-1\right)\end{aligned} \quad (12.38)$$

Substituting (12.38) into (12.37) yields

$$2\psi\ddot{h} - \lambda\sigma^2 h = -\left(f_w + \lambda\sigma^2\right). \tag{12.39}$$

Dividing the equation by $2\psi$ leads to the following ordinary differential equation (ODE)

$$\ddot{h} - g^2 h = -s - g^2 \text{ with } s = \frac{f_w}{2\psi},\ g^2 = \frac{\lambda\sigma^2}{2\psi}. \tag{12.40}$$

For the newly defined parameter, we have $g \geq 0$ and $s$ has the same sign as $f_w$. The boundary condition is $h(0) = 0$ and $h(T) = 1$. However, note the following:

- Because the trading horizon $T$ is fixed, the fixed-cost term is then known, and it does not enter the solution. However, it will play a significant role when we have a flexible trading horizon.

We will first consider the solution for the following two special cases:

Case I: $s = g = 0$

This occurs when both forecast and risk-aversion parameter are zero. Now the differential equation reduces to $\ddot{h} = 0$. The solution is therefore

$$h(t) = \frac{t}{T}. \tag{12.41}$$

The optimal solution is linear, implying a constant speed of trading: $\dot{h} = 1/T$. In this case, only the market impact matters. To reduce market impact, the optimal trading strategy is to break the trade evenly during the trade horizon. Furthermore, the total cost would just be

$$\frac{J}{(\Delta w)^2} = c_w T + \int_0^T \left\{ \psi\left(\frac{1}{T}\right)^2 \right\} dt = c_w T + \frac{\psi}{T}. \tag{12.42}$$

Note that the total costs as a function of $T$ go to infinity when $T$ goes to either zero or infinity. It reaches a minimum if $T=\sqrt{\psi/c_w}$ . If $c_w = 0$, the total cost decreases to zero as the trading horizon lengthens to infinity, which is an unrealistic result.

Case II: $g = 0$

In this case, the risk-aversion parameter is zero. Now the differential equation reduces to $\ddot{h}=-s$ . The solution is therefore

$$h(t)=-\frac{s}{2}t^2+at+b\,. \tag{12.43}$$

The constant $a$ and $b$ can be determined by the boundary condition. Therefore, we have

$$h(t)=\frac{t}{T}+\frac{s}{2}t(T-t)\,. \tag{12.44}$$

Equation 12.44 consists of the solution (12.41) and a quadratic term that vanishes at both $t = 0$ and $t = T$. The trading speed is given by

$$\dot{h}(t)=\frac{1}{T}+\frac{sT}{2}-st\,. \tag{12.45}$$

Figure 12.8 plots the solution for three cases, all with $g = 0$ but with three different values of $s$. The solution for the case with $s = 0$ is a straight line. When $s > 0$, by its definition the term $f_w$ is positive, implying either a positive forecast for a buy or a negative forecast for a sell. Hence, there is a need to execute the trade as soon as possible in order to reduce alpha shortfall. This is indeed the case for the optimal solution, the dotted line, which lies above the linear solution. The slope, or the speed of the trade, is higher initially and then slows down as time approaches $T$. On the other hand, when $s < 0$, the term $f_w$ is then negative, implying either a negative forecast for a buy or a positive forecast for a sell. Contrary to the previous case, there is incentive to delay the trade as long as possible, because

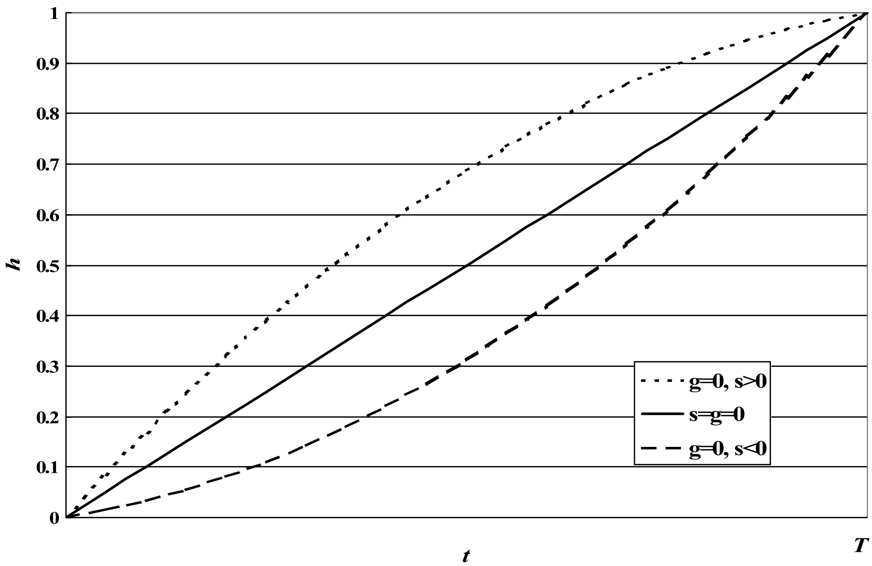


**FIGURE 12.8.** The optimal trading paths for three special cases: the solid line is for the case $s = g = 0$, the dotted line is for $g = 0$, $s > 0$, and the dashed line is for $g = 0$, $s < 0$.

the trade itself leads to lower alpha. Therefore, the optimal solution, the dashed line, lies below the linear solution. The trade fills slowly first and then speeds up as the time approaches $T$.

It is actually possible for the solution (12.44) for $h(t)$ to move out of the range [0,1]. For instance, when $s > 0$, $h(t)$ could be greater than 1. On the other hand, when $s < 0$, $h(t)$ could be less than 0. This implies that the solution may actually switch the direction of the trade during the course of trading! In other words, if the trade were to buy 1000 shares, the optimal strategy could have us buy 1100 shares and later sell the extra 100 shares. This is highly unlikely in practice, because the trading would have stopped once the 1000 shares had been bought. It could happen in the optimal trading solution if the trading horizon is too long, coupled with the fact that we have a strong forecast and a relatively weak market impact. With this combination, the mathematical optimal trading strategy would be to first buy as many shares as possible to generate returns and then later sell them to reach trade size. Because the trading cost is low, this "two-way" strategy would be better than any "one-way" strategy.

Figure 12.9 illustrates this situation. The dotted line is an optimal strategy whose path rises and crosses the line $h = 1$ during the trading horizon. The culprit in this case is the fixed trading horizon $T$, which is too long.

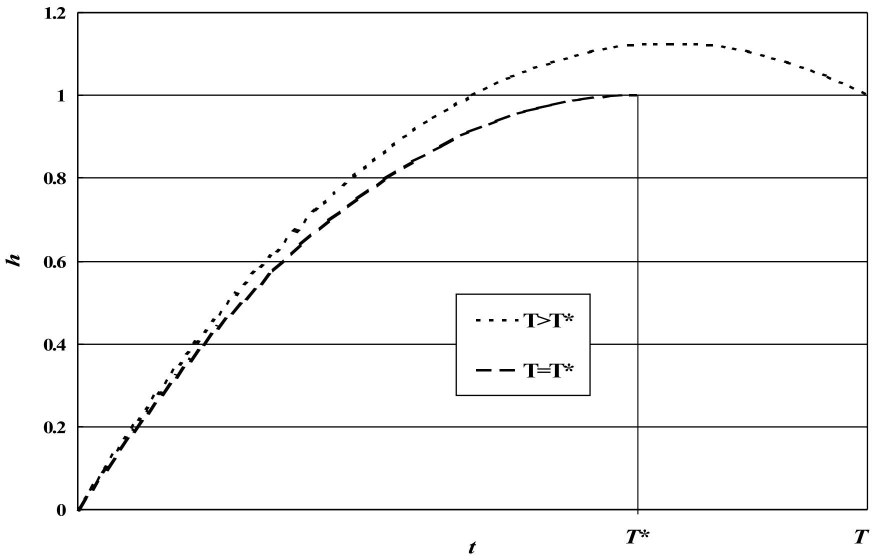


**FIGURE 12.9.** Optimal trading paths for two different trading horizons.

If we allow the trading horizon to be free and optimize it together with the trading path, the horizon will be shortened to $T^*$ and the associated optimal path, the dashed line, will never cross the line $h = 1$. The case of the free trading horizon is solved in the following section.

### *12.5.1.1 The General Case*

When the parameter $g$ is nonzero, the general solution of ODE (12.40) is the exponential functions $\exp(-gt)$ and $\exp(gt)$, which can be combined into hyperbolic functions. The particular solution is given by

$$-g^2 h = -s - g^2 \text{ or } h = 1 + \frac{s}{g^2}.$$

We have (Grinold & Kahn 2000)

$$h(t) = a\sinh(gt) + b\cosh\sinh(gt) + 1 + \frac{s}{g^2}.$$

The constant $a$ and $b$ are determined by the boundary condition; therefore we have

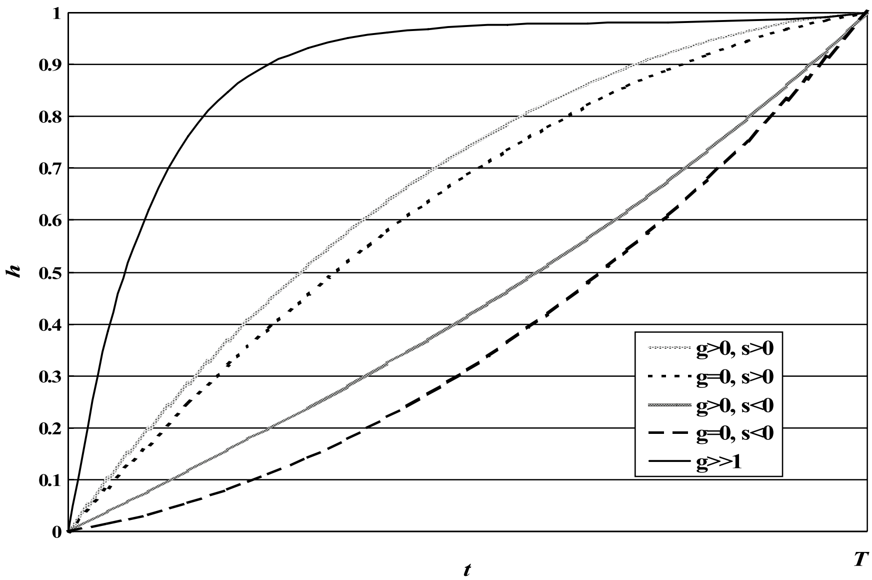


**FIGURE 12.10.** Five different optimal trading paths, two of which are identical to those in Figure 12.7. The other three are for cases with $g > 0$. Two of them have a moderate value of $g$, whereas the steepest path, the thin solid line, has the highest value of $g$, corresponding to extreme risk aversion.

$$h(t)=\frac{\left(1+\frac{s}{g^2}\right)\cosh(gT)-\frac{s}{g^2}}{\sinh(gT)}\sinh(gt)-\left(1+\frac{s}{g^2}\right)\left[\cosh(gt)-1\right]. \quad (12.46)$$

To see the effect of $g$, or variance of shortfall, on the optimal trading strategy, we plot the solution (12.46) in Figure 12.10. There are in all five paths in Figure 12.10, and two of them are identical to those in Figure 12.8 and have zero risk aversion ($g = 0$) but nonzero $s$. The shaded lines next to them are the corresponding trading paths with nonzero g. In both cases, the new trading path is above the previous one, indicating faster execution regardless of the forecast. This makes intuitive sense because higher risk aversion would cause investors to desire speedy execution at the expense of higher transaction costs.

When risk aversion dominates both the return shortfall and market impact, the optimal trading strategy is immediate execution. The thin solid line in Figure 12.10 illustrates this point. It rises rather rapidly and then flattens out. It can be shown mathematically that as $g \to \infty$,

$$\begin{aligned} h(t) &\to 1-\exp(-gt), \quad \text{if } t \text{ is near } 0; \\ h(t) &\to \exp\left[-g(T-t)\right], \quad \text{if } t \text{ is near } T. \end{aligned} \tag{12.47}$$

**Example 12.5**

Consider the case of $s = 0$ in (12.46). Then the solution reduces to

$$h(t) = \coth(gT)\sinh(gt) - \cosh(gt) + 1\,. \tag{12.48}$$

We obtain the implementation costs as

$$\int_0^T \left\{ c_w + \psi\left[\dot{h}(t)\right]^2 \right\} dt = c_w T + \psi g^2 \left[ \frac{1}{2g}\coth(gT) + \frac{T}{2}\operatorname{csch}^2(gT) \right] \tag{12.49}$$

and the implementation risk in terms of variance is

$$\sigma^2 \int_0^T \left[ h(t) - 1 \right]^2 dt = \sigma^2 \left[ \frac{1}{2g}\coth(gT) - \frac{T}{2}\operatorname{csch}^2(gT) \right]. \tag{12.50}$$

Taking the square root of (12.50) gives rise to the implementation risk in standard deviations.

Figure 12.11 plots the implementation costs vs. the risk for varying degrees of risk aversion. The cost is positive in the graph and is a declining function of risk. Each point of the curve corresponds to a different trading strategy, depending on different levels of risk aversion, illustrating the trade-off between risk and cost. When the risk aversion is high, the optimal trading strategy would be to trade fast to reduce implementation risk but incur higher cost. On the other hand, when the risk aversion is low, the optimal trading strategy focuses on lowering cost but incurs higher implementation risk.

### 12.5.2 Optimal Trading Horizon

The analysis so far has assumed a fixed trading horizon. However, in reality, the trading horizon is not precisely known and depends on the trade itself. For instance, for trades that are easy to implement, the trade size is a small fraction of the average daily volume, and the trading horizon can be short; whereas for trades that are difficult to fill, the trading horizon must

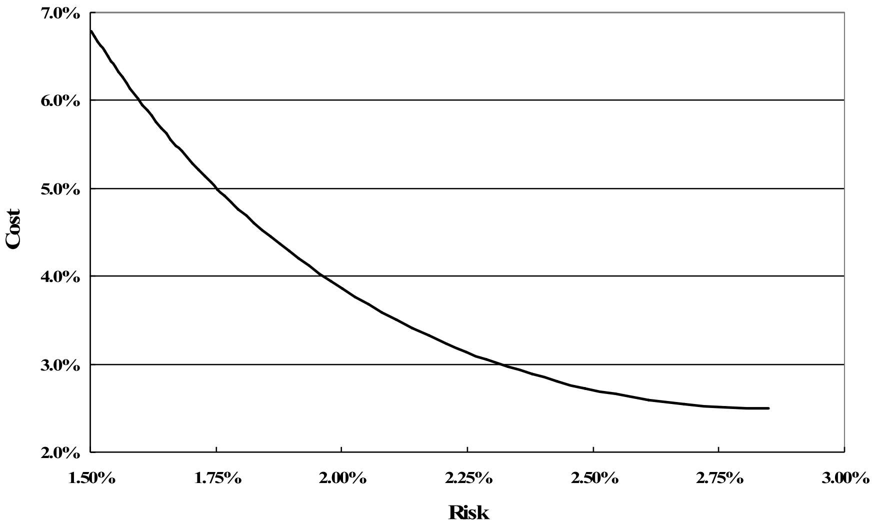


**FIGURE 12.11.** The implementation cost-risk frontier for optimal trading strategies. The parameters are $\psi = 0.05\%$, $\sigma = 35\%$, and $T = 0.02$. We also set $c_w = 0$, which does not affect the shape of the curve, because the fixed cost is a constant for fixed $T$, independent of risk aversion.

be lengthened. The trading horizon may also be dependent on investors' aversion to risks of shortfall. If the risk aversion is high, then the horizon is short; and if the risk aversion is low, then the horizon might be longer.

Mathematically, we can treat the trading horizon as a part of the optimization problem. In other words, we should let $T$ be free or unknown, and we can then solve the optimization problem for both the optimal trading path $h(t)$ and the optimal $T$. In reality, there might be some practical constraints on the trading horizon; for instance, one might want to complete a trade ahead of a long weekend. It is nevertheless useful to compare this with the true optimal.

The mathematical problem is to maximize the objective function (12.35) with both $h(t)$ and free boundary $T$. The problem can similarly be solved with the calculus of variation as follows. The optimal path $h(t)$ must satisfy the same differential equation

$$\frac{d}{dt}\left[\frac{\partial L}{\partial \dot{h}}\right] = \frac{\partial L}{\partial h}.$$

It should also satisfy the same boundary condition $h(0) = 0$ and $h(T) = 1$. In addition, the free boundary condition leads to the following (see Appendix):

$$L(h,\dot{h}) - \frac{\partial L(h,\dot{h})}{\partial \dot{h}} \dot{h} \Bigg|_{t=T} = 0 . \tag{12.51}$$

Because $\frac{\partial L}{\partial \dot{h}} = -2\psi\dot{h}$ and $L = -c_w - \psi(\dot{h})^2$ at $t = T$, equation (12.51) leads to

$$\begin{gathered} -c_w - \psi(\dot{h})^2 + 2\psi(\dot{h})^2 = \psi(\dot{h})^2 - c_w = 0 \\ \dot{h}(T) = \sqrt{\frac{c_w}{\psi}} \triangleq p \end{gathered} \tag{12.52}$$

Hence, the free trading horizon gives rise to a condition on the trading speed at $T$, which allows us to find the optimal trading time as well as the optimal trading path. Note the following:

- We have taken the positive root for $\dot{h}(T)$ because $h(t)$ is a monotonically increasing function if we do not allow the trading strategies to switch the direction of trades. From $h(0) = 0$ and $h(T) = 1$, we conclude $\dot{h}(t) \geq 0$.
- If $c_w = 0$, i.e., the fixed cost of transaction is neglected, then the condition becomes $\dot{h}(T) = 0$. As the trade gets filled, the trading at the end of the trading horizon gets slower and slower, coming to a smooth stop at the end.

**Example 12.6**

Consider the case in which $g = 0$ (zero risk aversion). The solution for $h(t)$ is (12.44) and for the trading speed $\dot{h}(t)$ is (12.45). Hence, (12.52) gives rise to

$$\dot{h}(T) = \frac{1}{T} + \frac{sT}{2} - sT = \frac{1}{T} - \frac{sT}{2} = \sqrt{\frac{c_w}{\psi}} = p .$$

This is a quadratic equation for $T$ and the solution is

$$T = \frac{2}{p + \sqrt{p^2 + 2s}} = \frac{2\sqrt{\psi}}{\sqrt{c_w} + \sqrt{c_w + f_w}} = \frac{2\sqrt{\psi}\,|\Delta w|}{\sqrt{c} + \sqrt{c + f\,\mathrm{sgn}(\Delta w)}} . \tag{12.53}$$

- The optimal trading horizon exists when $f_w$ is positive: a positive forecast for a buy or negative forecast for a sell. In this case, the trading horizon increases with the market impact cost $\psi$ and the trade size $\Delta w$. In other words, if the trade is costly and large, we should allow more time. The trading horizon also decreases with the alpha forecast and the fixed cost. If alpha shortfall is severe or if the fixed cost is large, we should execute the trade sooner.
- The optimal trading horizon does not always exist. If $f_w$ is negative — negative forecast for a buy or positive forecast for a sell — and the magnitude of the forecast exceeds that of the fixed cost $|f|>c$, then there is no optimal trading horizon. In other words, the optimal trading horizon is infinite, because the trade in these circumstances would reduce the return. Coupled with a high forecast, we would gain more if we delayed the trade for as long as possible. These cases might not occur in practice, but one should be aware of the possibilities.

If $c = 0$, i.e., there is no fixed cost, then Equation 12.53 reduces to

$$T = \frac{2\sqrt{\psi}\,|\Delta w|}{\sqrt{f}}. \tag{12.54}$$

**Example 12.7**

Consider the case $s = 0$ (zero forecast) as in Example 12.5. From (12.48), we have

$$\dot{h}(T) = g\left[\coth(gT)\cosh(gT) - \sinh(gT)\right] = \frac{g}{\sinh(gT)}. \tag{12.55}$$

Therefore, the optimal trading horizon is given by

$$\frac{g}{\sinh(gT)} = p \text{ or } T = \frac{1}{g}\sinh^{-1}\left(\frac{g}{p}\right). \tag{12.56}$$

Written in terms of the original parameters, we have

$$T = \sqrt{\frac{2\psi}{\lambda\sigma^2}}\sinh^{-1}\sqrt{\frac{\lambda\sigma^2|\Delta w|}{2c}}. \tag{12.57}$$

In general, the optimal trading horizon lengthens if $\Delta w$ (the trade size) increases, if $\psi$ (market impact) increases, and if $c$ (fixed cost) decreases. It also lengthens if $\lambda\sigma^2$ (risk aversion) decreases, because the function $\sinh^{-1}(x)/x$ is a declining function of $x$.

## 12.6 OPTIMAL TRADING STRATEGIES: PORTFOLIOS OF STOCKS

Much of the analysis of single-stock trading strategies can be extended to multiple stocks, or a portfolio of stocks. We shall formulate the problem first and then find the optimal solution. We shall also allow for the optimal trading horizon $T$. For a portfolio of stock trades, we also discuss additional constraints one might wish to impose during the trading.

### 12.6.1 Formulation

Suppose we have trades in $N$ stocks, and the trade sizes are $(\Delta w_1, \Delta w_2, \cdots, \Delta w_N)$. We denote the trading path by a vector of function $\mathbf{h}(t) = [h_1(t), \cdots, h_N(t)]'$. At any given time $t$, the portfolio position relative to the final position is $[\Delta w_1(h_1 - 1), \Delta w_2(h_2 - 1), \cdots, \Delta w_N(h_N - 1)]$. At the beginning of the trade, we have $h_i(0) = 0,\ i = 1, \cdots, N$ and at the end of the trade $h_i(T) = 1,\ i = 1, \cdots, N$. These are the boundary conditions for $h$'s.

The optimal trading strategy for a portfolio of trades is found by optimizing an objective function similar to that of a single trade. First, the instantaneous return shortfall is given by $f_1 \Delta w_1(h_1 - 1) + f_2 \Delta w_2(h_2 - 1) + \cdots + f_N \Delta w_N(h_N - 1) = \mathbf{f}_w' \cdot (\mathbf{h} - \mathbf{1})$, in which $f$'s are return forecasts and the vector $\mathbf{f}_w = (f_1 \Delta w_1, \cdots, f_N \Delta w_N)'$ and the vector $\mathbf{1} = (1, \cdots, 1)'$. The variance of the return shortfall for a given time $t$ is

$$
\begin{aligned}
&[\Delta w_1(h_1 - 1), \cdots, \Delta w_N(h_N - 1)]\boldsymbol{\Sigma}\begin{pmatrix} \Delta w_1(h_1 - 1) \\ \vdots \\ \Delta w_N(h_N - 1) \end{pmatrix}. \qquad (12.58)\\
&= (\mathbf{h} - \mathbf{1})' \boldsymbol{\Sigma}_w (\mathbf{h} - \mathbf{1})
\end{aligned}
$$

The matrix $\boldsymbol{\Sigma} = (\sigma_{ij})_{i,j=1}^{N}$ is the covariance matrix of returns, and $\boldsymbol{\Sigma}_w = (\sigma_{ij} \Delta w_i \Delta w_j)_{i,j=1}^{N}$ comprises products of the return covariance matrix and the trade size.

Similar to the single-stock trade, there are two components of transaction costs. We model the fixed costs as a multiple of the trading horizon $T$, and the constant is given by $c_w = c_1 \Delta w_1 + c_2 \Delta w_2 + \cdots + c_N \Delta w_N$. The variable costs — the instantaneous market impact — is related to the speeds of the trading in all $N$ stocks

$$\left[\Delta w_1 \dot{h}_1, \cdots, \Delta w_N \dot{h}_N\right] \mathbf{\Psi} \begin{pmatrix} \Delta w_1 \dot{h}_1 \\ \vdots \\ \Delta w_N \dot{h}_N \end{pmatrix} = \dot{\mathbf{h}}' \mathbf{\Psi}_w \dot{\mathbf{h}}, \tag{12.59}$$

where $\mathbf{\Sigma}_w = \left(\Psi_{ij} \Delta w_i \Delta w_j\right)_{i,j=1}^{N}$.

Combining all four terms and integrating them over time gives the objective function of trading strategies

$$J = \int_0^T L\left(\mathbf{h}, \dot{\mathbf{h}}\right) dt, \text{ with}$$

$$L\left(\mathbf{h}, \dot{\mathbf{h}}\right) = \mathbf{f}_w \left[\mathbf{h}(t) - \mathbf{1}\right] - c_w - \dot{\mathbf{h}}(t)' \mathbf{\Psi}_w \dot{\mathbf{h}}(t). \tag{12.60}$$

$$-\frac{1}{2} \lambda \left[\mathbf{h}(t) - \mathbf{1}\right]' \mathbf{\Sigma}_w \left[\mathbf{h}(t) - \mathbf{1}\right]$$

### 12.6.2 Solutions of Optimal Trading Strategies

We derive the differential equation for the optimal trading path with the calculus of variation. We have

$$\frac{\partial L}{\partial \dot{\mathbf{h}}} = -2 \mathbf{\Psi}_w \dot{\mathbf{h}}(t), \ \frac{\partial L}{\partial \mathbf{h}} = \mathbf{f}_w - \lambda \mathbf{\Sigma}_w \left[\mathbf{h}(t) - \mathbf{1}\right] \tag{12.61}$$

and $\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{\mathbf{h}}}\right) = \frac{\partial L}{\partial \mathbf{h}}$ gives rise to

$$2 \mathbf{\Psi}_w \ddot{\mathbf{h}}(t) - \lambda \mathbf{\Sigma}_w \mathbf{h}(t) = -\mathbf{f}_w - \lambda \mathbf{\Sigma}_w. \tag{12.62}$$

Assuming the matrix $\mathbf{\Psi}_w$ is invertible, we can rewrite (12.62) as

$$\ddot{\mathbf{h}}(t) - \frac{\lambda}{2} \mathbf{\Psi}_w^{-1} \mathbf{\Sigma}_w \mathbf{h}(t) = -\frac{1}{2} \mathbf{\Psi}_w^{-1} \mathbf{f}_w - \frac{\lambda}{2} \mathbf{\Psi}_w^{-1} \mathbf{\Sigma}_w. \tag{12.63}$$

The particular solution of (12.63) is obtained by setting $\ddot{\mathbf{h}}=0$

$$\mathbf{h}(t)=-\frac{1}{\lambda}\boldsymbol{\Sigma}_w^{-1}\mathbf{f}_w+\mathbf{1}. \tag{12.64}$$

The general solution is of the form $\mathbf{h}(t)=\mathbf{v}\cdot\exp(pt)$ and

$$\left(p^2\mathbf{I}-\frac{\lambda}{2}\boldsymbol{\Psi}_w^{-1}\boldsymbol{\Sigma}_w\right)\mathbf{v}=0. \tag{12.65}$$

It follows that $p^2$ must be an eigenvalue of the matrix $\frac{\lambda}{2}\boldsymbol{\Psi}_w^{-1}\boldsymbol{\Sigma}_w$ and $\mathbf{v}$ the corresponding eigenvector, both of which can be found by standard numerical routines. Note the following:

- Assuming the matrix $\frac{\lambda}{2}\boldsymbol{\Psi}_w^{-1}\boldsymbol{\Sigma}_w$ is positive definite, there will be $N$ positive eigenvalues and $N$ eigenvectors, and there will be $2N$ general solutions. The weights for these solutions can be found using $2N$ boundary conditions.

### 12.6.3 Optimal Trading Horizon

When the trading horizon is free, we can find the optimal trading horizon using the condition similar to (12.51). In the case of a portfolio trade, we have

$$L(\mathbf{h},\dot{\mathbf{h}})-\dot{\mathbf{h}}'\cdot\left.\frac{\partial L(\mathbf{h},\dot{\mathbf{h}})}{\partial\dot{\mathbf{h}}}\right|_{t=T}=0. \tag{12.66}$$

Using (12.60) and (12.61) gives

$$\left.\dot{\mathbf{h}}'\cdot\boldsymbol{\Psi}\cdot\dot{\mathbf{h}}\right|_{t=T}=c_w. \tag{12.67}$$

The condition is similar to (12.52) and can be combined with the optimal trading solution of the last section to find the optimal $T$.

### 12.6.4 Portfolio Constraints

When trading a portfolio of stocks, one often has to maintain the balance between orders so that the portfolio meets a set of constraints. An example of such constraint is the dollar-neutral constraint: the dollar amount of buys matches that of sells. Other constraints can be risk based. For instance, we might want the portfolio to be beta neutral at all times. These linear constraints can be expressed as

$$\mathbf{h}' \cdot \mathbf{g} = 0 \tag{12.68}$$

where **h** is the trading path for all stocks and **g** a vector of constants.

There are a couple of ways to find the optimal trading strategies with such linear constraints, for example, the method of elimination and the method of the Lagrangian multiplier (Kirk 1970).

## PROBLEMS

12.1 Prove that the coefficient $\theta$ in Equation 12.1 is given by the cost per share divided by the share price.

12.2 Consider the case in which $\tilde{w} < w_0$. Prove that the optimal weight is

$$w^* = \begin{cases} \dfrac{f+\theta}{\lambda\sigma^2}, & \text{if } \dfrac{f+\theta}{\lambda\sigma^2} \le w_0 \\ w_0, & \text{otherwise} \end{cases}. \tag{12.69}$$

12.3 Prove that the critical value of $\theta$, above which there is no trade, is given by

$$\theta_c = \lambda\sigma^2 \left| \tilde{w} - w_0 \right|. \tag{12.70}$$

12.4 Find the optimal position of a single asset when there are both linear and quadratic transaction costs, by maximizing the utility function

$$U(w) = f \cdot w - \frac{1}{2}\lambda\sigma^2 w^2 - \theta \left| w - w_0 \right| - \psi \left( w - w_0 \right)^2. \tag{12.71}$$

12.5 (a) Prove that the utility function in (12.25) can be written as

$$U(\mathbf{w}) = U(\mathbf{w}_0) + \lambda (\Delta\mathbf{w})' \, \boldsymbol{\Sigma} (\tilde{\mathbf{w}} - \mathbf{w}_0) - \frac{1}{2}\lambda (\Delta\mathbf{w})' \, \boldsymbol{\Sigma} (\Delta\mathbf{w}) - \boldsymbol{\theta}' \cdot \left| \Delta\mathbf{w} \right|, \tag{12.72}$$

with $\tilde{\mathbf{w}} = \lambda^{-1} \boldsymbol{\Sigma}^{-1} \mathbf{f}$ as the optimal weights with no transaction costs, and

(b) prove that the optimal weights must satisfy the condition $(\Delta\mathbf{w})'\,\boldsymbol{\Sigma}(\tilde{\mathbf{w}}-\mathbf{w}_0)\geq 0$, i.e., the vector of weight changes must be in the same direction as $(\tilde{\mathbf{w}}-\mathbf{w}_0)$.

12.6 Express the range constraint (12.33) as linear inequality constraints on the augmented vector **W**.

12.7 Verify that solution (12.46) satisfies both the differential equation and the boundary conditions.

12.8 For the optimal trading solution (12.48), prove that the implementation cost is given by (12.49) and the implementation risk is given by (12.50).

12.9 For the general optimal trading solution (12.46) and free $T$, show that the optimal trading horizon $T$ satisfies equation

$$s\cosh(gT)+gp\sinh(gT)=s+g^2.$$

## APPENDIX
## CALCULUS OF VARIATION

We derive the ODE for the optimal trading strategy and the optimal trading horizon using calculus of variation.

Given a functional, a real-valued function of functions

$$J(h,T)=\int_o^T L\left[h(t),\dot{h}(t),t\right]dt,$$

in which $h(0)=0$ and $h(T)=1$, and $T$ is free, then the change in the functional is

$$\begin{aligned}\delta J &= J(h+\delta h,T+\delta T)-J(h,T)\\ &=\int_o^{T+\delta T} L\left[h(t)+\delta h,\dot{h}(t)+\delta\dot{h},t\right]dt-\int_o^T L\left[h(t),\dot{h}(t),t\right]dt\end{aligned}$$

Splitting the first integral in two, we have

$$\delta J = \int_o^T \left\{ L\left[ h(t) + \delta h, \dot{h}(t) + \delta \dot{h}, t \right] - L\left[ h(t), \dot{h}(t), t \right] \right\} dt$$

$$- \int_T^{T+\delta T} L\left[ h(t) + \delta h, \dot{h}(t) + \delta \dot{h}, t \right] dt$$

The second term is approximated by

$$\int_T^{T+\delta T} L\left[ h(t) + \delta h, \dot{h}(t) + \delta \dot{h}, t \right] dt = L\left[ h(t), \dot{h}(t), t \right]\Big|_{t=T} \delta T + o(\delta T). \quad (12.73)$$

The notation $o(.)$ denotes the higher-order term. The first term can be approximated by Taylor expansion

$$\int_o^T \left\{ L\left[ h(t) + \delta h, \dot{h}(t) + \delta \dot{h}, t \right] - L\left[ h(t), \dot{h}(t), t \right] \right\} dt = \int_o^T \left\{ \delta h \frac{\partial L}{\partial h} + \delta \dot{h} \frac{\partial L}{\partial \dot{h}} \right\} dt .$$

Integrating by parts the term containing $\delta \dot{h}$ yields

$$\begin{aligned} &\int_o^T \left\{ L\left[ h(t) + \delta h, \dot{h}(t) + \delta \dot{h}, t \right] - L\left[ h(t), \dot{h}(t), t \right] \right\} dt \\ &= \int_o^T \delta h \left\{ \frac{\partial L}{\partial h} - \frac{d}{dt} \frac{\partial L}{\partial \dot{h}} \right\} dt + \left( \delta h \frac{\partial L}{\partial \dot{h}} \right)\Big|_{t=T} \end{aligned} \quad (12.74)$$

When $T$ is fixed, we have $\delta h = 0$ at $t = T$ . When $T$ is free, we have

$$0 = h(T + \delta T) - h^*(T) \approx h(T) - h^*(T) + \dot{h}^*(T)\delta T = \delta h(T) + \dot{h}^*(T)\delta T .$$

Therefore,

$$\delta h(T) \doteq -\dot{h}^*(T)\delta T . \quad (12.75)$$

Combining (12.73), (12.74), and (12.75) gives

$$0 = \delta J = \int_o^T \delta h \left\{ \frac{\partial L}{\partial h} - \frac{d}{dt}\frac{\partial L}{\partial \dot{h}} \right\} dt + \left( L - \dot{h}\frac{\partial L}{\partial \dot{h}} \right)\Bigg|_{t=T} \delta T \tag{12.76}$$

for optimal path and optimal trading horizon. Because Equation 12.76 is true for the arbitrary function $\delta h$ and arbitrary increment $\delta T$, we must have

$$\frac{d}{dt}\left( \frac{\partial L}{\partial \dot{h}} \right) - \frac{\partial L}{\partial h} = 0,$$

and

$$\left( L - \dot{h}\frac{\partial L}{\partial \dot{h}} \right)\Bigg|_{t=T} = 0.$$

For fixed $T$, only the ODE has to be satisfied.

## REFERENCES

Coppejans, M. and Madhavan, A., Active management and transaction costs, working paper, Barclays Global Investors, 2006.

Grinold, R., A dynamic model of portfolio management, *Journal of Investment Management,* Vol. 4, No. 2, 2006.

Grinold, R.C. and Kahn, R.N., *Active Portfolio Management*, McGraw-Hill, New York, 2000.

Kirk, D.E., *Optimal Control Theory — An Introduction*, Dover, New York, 1970.

Robert, A. and Chriss, N., Optimal execution of portfolio transactions, *Journal of Risk,* Vol. 3, No. 2, 2000.

Sneddon, L., The dynamics of active portfolios, *Proceeding of Northfield Research Conference*, 2005.