# Duality Methods in Portfolio Allocation with Transaction Constraints and Uncertainty 

Christopher W. Miller<br>Department of Mathematics<br>University of California, Berkeley

December 1, 2013


#### Abstract

We examine the application of duality methods to portfolio allocation with transaction constraints. We consider various models of asset returns and transaction constraints. In each model, we propose algorithms to efficiently find optimal solutions to the dual problem and examine when the solution to the dual problem is a minimizer of the primal problem. In particular, we consider mean-variance objective functions, both with correlation and without, as well as uncertainty in return and volatility parameters. We also consider transactions cost models with fixed costs that are different for long and short positions. Finally, we implement a general algorithm for locating optimal solutions to these dual portfolio allocation problem.


## 1 Problem Description

We consider a single-period investment model involving $n$ assets. For each asset, we assume that the risk-return preference is given by a convex function $f_{i}: \mathbb{R} \rightarrow \mathbb{R}$. Our basic goal throughout this paper is to minimize the objective function $f: \mathbb{R}^{n} \rightarrow \mathbb{R}$ given by

$$
f(w)=\sum_{i=1}^{n} f_{i}\left(w_{i}\right)
$$

over the set $\mathcal{W}=\mathcal{W}_{1} \times \cdots \times \mathcal{W}_{n}$, where every $\mathcal{W}_{i}$ is a given interval $\left[\underline{w}_{i}, \bar{w}_{i}\right]$, where $\underline{w}_{i}<\bar{w}_{i}$. This corresponds to choosing a portfolio of uncorrelated assets which minimize the risk-return preference under constraints on sizes of both long and short positions.

We also assume that there are transaction costs. Given an initial position $w^{0} \in \mathbb{R}^{n}$, the total transaction cost for taking a position $w$ from the initial position $w^{0}$ is given by

$$
g(w)=\sum_{i=1}^{n} g_{i}\left(w_{i}-w_{i}^{0}\right),
$$

where, for every $i=1, \ldots, n$, the functions $g_{i}: \mathbb{R} \rightarrow \mathbb{R}$ are given. We do not assume that these cost functions are convex or even continuous.

The crucial assumption which we will rely on in Sections 3 and 4 is that both $f$ and $g$ are decomposable functions. In addition, we will make the assumption that for any scalar $\lambda \geqslant 0$, and every $i=1, \ldots, n$, finding a minimizer for the problem

$$
\min _{\xi \in \mathcal{W}_{i}} f_{i}(\xi)+\lambda g_{i}(\xi)
$$

can be efficiently done, algorithmically or otherwise.
Then in general, we consider an investment problem of the form

$$
p^{*}=\min _{w \in \mathcal{W}}\{f(w): g(w) \leqslant \tau\}
$$

where $\tau \geqslant 0$ is a bound on the total amount of transaction costs. We henceforth refer to this as the primal problem.

## 2 Lagrange Relaxation for the Generic Problem

Rather than considering the primal problem above with inequality constraints, we will consider a dual problem where a Lagrange multiplier $\lambda \geqslant 0$ controls the cost of breaking the transaction bound. For more information on the relationship between primal and dual problems in convex optimization, see [1].

We consider an associated Lagrangian defined by

$$
\mathcal{L}(w, \lambda)=f(w)+\lambda(g(w)-\tau) .
$$

Then we can define a function of $\lambda$ as

$$
D(\lambda)=\min _{w \in \mathcal{W}} \mathcal{L}(w, \lambda) .
$$

We henceforth refer to this minimization as the Lagrangian sub-problem. Note, this is a concave function of $\lambda$ as a minimum over functions which are linear in $\lambda$. Then we can define an associated maximization problem as

$$
d^{*}=\max _{\lambda \geqslant 0} D(\lambda) .
$$

We henceforth refer to this maximization as the dual problem.

### 2.1 General results about dual problem

We can make some immediate observations about the relationship between $p^{*}$ and $d^{*}$ without assuming a particular form for the objective functions or transaction costs.

Proposition 1. With the dual problem defined as above, $p^{*} \geqslant d^{*}$.
Proof. Let $\lambda \geqslant 0$. Then for any $w \in \mathcal{W}$ such that $g(w) \leqslant \tau$, we have

$$
f(w) \geqslant f(w)-\lambda(\tau-g(w))=\mathcal{L}(w, \lambda) .
$$

Now, taking the minimum over $w \in \mathcal{W}$ without the restriction on $g(w)$ can only result in a smaller value, so

$$
p^{*}=\min _{w \in \mathcal{W}}\{f(w): g(w) \leqslant \tau\} \geqslant \min _{w \in \mathcal{W}} \mathcal{L}(w, \lambda)=D(\lambda) .
$$

This inequality holds for all $\lambda \geqslant 0$, so

$$
p^{*} \geqslant \max _{\lambda \geqslant 0} D(\lambda)=d^{*} .
$$

Proposition 2. Let $\lambda^{*}$ be a maximizer of $D(\lambda)$, and $w^{*}$ be a corresponding minimizer of $\mathcal{L}\left(w, \lambda^{*}\right)$. Suppose that $g\left(w^{*}\right)=\tau$. Then $p^{*}=d^{*}$.

Proof. We already showed $p^{*} \geqslant d^{*}$, so it suffices to show that if $g\left(w^{*}\right)=\tau$ then $d^{*} \geqslant p^{*}$. But then

$$
d^{*}=\mathcal{L}\left(w^{*}, \lambda^{*}\right)=f\left(w^{*}\right)-\lambda^{*}\left(\tau-g\left(w^{*}\right)\right)=f\left(w^{*}\right) .
$$

Furthermore, because $g\left(w^{*}\right) \leqslant \tau$, we conclude that

$$
d^{*}=f\left(w^{*}\right) \geqslant \min _{w \in \mathcal{W}}\{f(w): g(w) \leqslant \tau\}=p^{*} .
$$

### 2.2 Algorithm to solve general dual problem

We now consider a general purpose algorithm which may be used to find sub-optimal solutions which approximate the optimal solution of the dual problem described above. We assume that we know an upper bound on the optimal $\lambda^{*}$, given by $\bar{\lambda}$.

Recall that $D(\lambda)$ is a concave function. The key to our algorithm is to note that we can easily compute super-gradients of $D(\lambda)$ and use these to locate a maximum.

Lemma 1. For any $\lambda_{*}$, let $w_{*}$ be a minimizing allocation of $\mathcal{L}\left(w, \lambda_{*}\right)$. Then $g\left(w_{*}\right)-\tau$ is a super-gradient of $D(\lambda)$ at $\lambda_{*}$.

Proof. Fix $\lambda_{*}$ and let $w_{*}$ be a minimizing allocation of $\mathcal{L}\left(w, \lambda_{*}\right)$. Consider any other $\lambda$. Then

$$
\begin{aligned}
D(\lambda) & \leqslant \mathcal{L}\left(w_{*}, \lambda\right)=-\lambda \tau+f\left(w_{*}\right)+\lambda g\left(w_{*}\right) \\
& =\left(g\left(w_{*}\right)-\tau\right)\left(\lambda-\lambda_{*}\right)-\lambda_{*} \tau+f\left(w_{*}\right)+\lambda_{*} g\left(w_{*}\right) \\
& =\left(g\left(w_{*}\right)-\tau\right)\left(\lambda-\lambda_{*}\right)+D\left(\lambda_{*}\right) .
\end{aligned}
$$

Then we conclude that $g\left(w_{*}\right)-\tau$ is a super-gradient of $D(\lambda)$ at $\lambda_{*}$.

Theorem 1. Suppose we are given an upper-bound $\bar{\lambda}$ on $\lambda^{*}$. Furthermore, suppose that we can find minimizing allocations of $\mathcal{L}(w, \lambda)$ for any $\lambda$ in time $O(f(n))$, where $f$ does not depend on $\lambda$. Then for any $\epsilon>0$, we can construct an $\epsilon$-suboptimal solution $\left(\lambda_{*}, w_{*}\right)$ to the dual problem in time $O(-f(n) \log (\epsilon))$ via the algorithm shown in Algorithm 1.

Proof. Let $\lambda^{*}$ be any maximizing value of $D(\lambda)$. By hypothesis, we have that $0 \leqslant \lambda^{*} \leqslant \bar{\lambda}$. We will reduce the size of this closed set by one-half in each step, while guaranteeing an inductive hypothesis that some $\lambda^{*}$ is always contained inside.

For any $\lambda$, we will denote the super-gradient at $\lambda$ determined in the previous lemma by $D^{\prime}(\lambda)=g\left(w_{\lambda}\right)-\tau$. First, note that if $D^{\prime}(\lambda)=0$ then $\lambda$ must be a maximizing value of $D(\lambda)$. However, the converse is not necessarily true. Furthermore, consider any two points $x, y$ such that there is a maximizing point between them, $x<\lambda^{*}<y$. Then $D^{\prime}(x) \geqslant 0$ and $D^{\prime}(y) \leqslant 0$.

Consider a step where we are examining the interval with $\lambda_{l}<\lambda_{r}$. We assume by the inductive hypothesis that there is a maximizing value of $D(\lambda)$ in the interval $\left[\lambda_{l}, \lambda_{r}\right]$. Then if

Input: Lagrangian function $\mathcal{L}$, a cost function $g$, a transaction limit $\tau$, an error tolerance $\epsilon>0$, and an upper bound on the maximizing point $\bar{\lambda}$.
Output: An $\epsilon$-suboptimal solution $\left(\lambda_{*}, w_{*}\right)$ to the dual problem.

```
\(\lambda_{l} \leftarrow 0\)
\(w_{l} \leftarrow \arg \min _{w \in \mathcal{W}} \mathcal{L}\left(w, \lambda_{l}\right)\)
\(m_{l} \leftarrow g\left(w_{l}\right)-\tau\)
\(\lambda_{r} \leftarrow \bar{\lambda}\)
\(w_{r} \leftarrow \arg \min _{w \in \mathcal{W}} \mathcal{L}\left(w, \lambda_{r}\right)\)
\(m_{r} \leftarrow g\left(w_{r}\right)-\tau\)
while \(\lambda_{r}-\lambda_{l}>\epsilon\) or \(\left\|w_{r}-w_{l}\right\|_{\infty}>\epsilon\) do
    if \(m_{l} \leqslant 0\) then
        \(\left(\lambda_{r}, w_{r}, m_{r}\right) \leftarrow\left(\lambda_{l}, w_{l}, m_{l}\right)\)
    else if \(m_{r} \geqslant 0\) then
        \(\left(\lambda_{l}, w_{l}, m_{l}\right) \leftarrow\left(\lambda_{r}, w_{r}, m_{r}\right)\)
    else
        \(\lambda_{m} \leftarrow \frac{1}{2}\left(\lambda_{l}+\lambda_{r}\right)\)
        \(w_{m} \leftarrow \arg \min _{w \in \mathcal{W}} \mathcal{L}\left(w, \lambda_{m}\right)\)
        \(m_{m} \leftarrow g\left(w_{m}\right)-\tau\)
        if \(m_{m} \leqslant 0\) then
            \(\left(\lambda_{r}, w_{r}, m_{r}\right) \leftarrow\left(\lambda_{m}, w_{m}, m_{m}\right)\)
        else
            \(\left(\lambda_{l}, w_{l}, m_{l}\right) \leftarrow\left(\lambda_{m}, w_{m}, m_{m}\right)\)
end
\(\left(\lambda_{*}, w_{*}\right) \leftarrow\left(\lambda_{l}, w_{l}\right)\)
return \(\left(\lambda_{*}, w_{*}\right)\)
```

Algorithm 1: DualProblemSolver
$D^{\prime}\left(\lambda_{l}\right) \leqslant 0$, it must be a maximizing value by the argument in the previous paragraph. Similarly, if $D^{\prime}\left(\lambda_{r}\right) \geqslant 0$, it must be a maximizing value. This case is illustrated in Figure 1a.

Now, assume that $D^{\prime}\left(\lambda_{l}\right)>0$ and $D^{\prime}\left(\lambda_{r}\right)<0$. Let $\lambda_{m}$ be the mid-point of $\lambda_{l}$ and $\lambda_{r}$. If $D^{\prime}\left(\lambda_{m}\right) \leqslant 0$, then the maximizing point in $\left[\lambda_{l}, \lambda_{r}\right]$ cannot be in $\left(\lambda_{m}, \lambda_{r}\right]$, so we conclude that there must be a maximizing point in $\left[\lambda_{l}, \lambda_{m}\right]$. Otherwise, we conclude that there must be a maximizing point in $\left[\lambda_{m}, \lambda_{r}\right]$. This case is illustrated in Figure 1b.

In either case, each step of this process reduces the size of the interval examined by one-half.


Figure 1: Illustrations of different super-gradient cases in the algorithm being described.

By compactness, there is a uniform $\delta>0$, independent of $\lambda$, within which optimal allocations can differ by at most $\epsilon$, so we reach an $\epsilon$-suboptimal solution in $O(-\log (\epsilon))$ steps. Each of these requires one evaluation of a super-gradient, which takes $O(f(n))$ time by assumption, so this algorithm will return an $\epsilon$-suboptimal solution in time $O(-f(n) \log (\epsilon))$.

## 3 Binary Model

In this section we consider a binary model for transaction costs where initial positions $w^{0}$ are all zero, each asset has a fixed transaction cost $b_{i}>0$ for purchases, and short-sales are not possible. In particular, we consider the special case of

$$
g_{i}(\xi)=\left\{\begin{array}{cc}
+\infty & \text { if } \xi<0 \\
0 & \text { if } \xi=0 \\
b_{i} & \text { if } \xi>0
\end{array}\right.
$$

### 3.1 Homogeneous transaction costs

For simplicity, we first assume that all costs $b_{i}$ are equal and, without loss of generality, equal to one. Furthermore, we let $\tau$ be an integer taken in $\{0,1, \ldots, n\}$. In this case we can efficiently find a minimizer of the dual problem and prove that there is no duality gap.

Proposition 3. The solution to the dual problem under the binary model may be written

$$
d^{*}=\max _{\lambda \geqslant 0}\left\{-\tau \lambda+\sum_{i=1}^{n} \min \left(f_{i}^{0}, \lambda+f_{i}^{+}\right)\right\}
$$

for appropriate vectors $f^{0}$ and $f^{+}$.
Proof. If we consider the dual problem from the previous section, immediate simplifications are possible. In particular, we note that for any $\lambda \geqslant 0$ by decomposability,

$$
\begin{aligned}
D(\lambda) & =\min _{w \in \mathcal{W}}\left\{-\lambda \tau+\sum_{i=1}^{n}\left(f_{i}\left(w_{i}\right)+\lambda g_{i}\left(w_{i}\right)\right)\right\} \\
& =-\lambda \tau+\sum_{i=1}^{n} \min _{w_{i} \in \mathcal{W}_{i}}\left\{f_{i}\left(w_{i}\right)+\lambda g_{i}\left(w_{i}\right)\right\}
\end{aligned}
$$

But because $g_{i}(\xi)$ only takes on two values, we can write each minimization problem as

$$
\min _{w_{i} \in \mathcal{W}_{i}}\left\{f_{i}\left(w_{i}\right)+\lambda g_{w}\left(w_{i}\right)\right\}=\min \left(f_{i}(0), \lambda+\min _{0<w_{i} \leqslant \bar{w}_{i}} f\left(w_{i}\right)\right) .
$$

For notational convenience, we define

$$
\left\{\begin{array}{ccc}
f_{i}^{0} & = & f_{i}(0) \\
f_{i}^{+} & = & \min \left\{f_{i}\left(w_{i}\right): 0<w_{i} \leqslant \bar{w}_{i}\right\}
\end{array}\right.
$$

as well as

$$
w_{i}^{*}=\underset{0<w_{i} \leqslant \bar{w}_{i}}{\arg \min } f_{i}\left(w_{i}\right)
$$

Then we can write the dual problem as

$$
d^{*}=\max _{\lambda \geqslant 0}\left\{-\lambda \tau+\sum_{i=1}^{n} \min \left(f_{i}^{0}, \lambda+f_{i}^{+}\right)\right\} .
$$

The usefulness of this result is that $D(\lambda)$ can be easily understood as the sum of a linear function and $n$ kinked piecewise linear functions. This interpretation is useful is finding maximizing values $\lambda^{*}$ of $D(\lambda)$.

Theorem 2. We can construct an optimal solution $\left(\lambda^{*}, w^{*}\right)$ to the dual problem in the binary model in time $O(n)$ via the algorithm shown in Algorithm 2.

```
Input: Objective functions \(f_{i}\), maximal allocations \(\bar{w}_{i}\), and trade cap \(\tau\).
Output: An optimal solution \(\left(\lambda^{*}, w^{*}\right)\) to the dual problem.
for \(i \leftarrow 1\) to \(n\) do
    \(f_{i}^{0} \leftarrow f_{i}(0)\)
    \(w_{i}^{*} \leftarrow \arg \min \left\{f_{i}\left(w_{i}\right): 0<w_{i} \leqslant \bar{w}_{i}\right\}\)
    \(f_{i}^{+} \leftarrow f_{i}\left(w_{i}^{*}\right)\)
    \(\lambda_{i} \leftarrow f_{i}^{0}-f_{i}^{+}\)
end
\(\lambda^{*} \leftarrow(\tau+1)\) th largest value of \(\lambda_{i}\) counting multiplicity
\(\lambda^{*} \leftarrow \max \left\{\lambda^{*}, 0\right\}\)
for \(i \leftarrow 1\) to \(n\) do
        if \(\lambda_{i} \leqslant \lambda^{*}\) then
        \(w_{i}^{*} \leftarrow 0\)
    end
end
return \(\left(\lambda^{*}, w^{*}\right)\)
```


## Algorithm 2: BinaryProblemSolver

Proof. Consider the function which we are maximizing

$$
D(\lambda)=-\lambda \tau+\sum_{i=1}^{n} \min \left(f_{i}^{0}, \lambda+f_{i}^{+}\right) .
$$

Define $\lambda_{i}=f_{i}^{0}-f_{i}^{+}$. Then for each asset, the graph of $\min \left(f_{i}^{0}, \lambda+f_{i}^{+}\right)$is piecewise linear with slope 1 for $\lambda<\lambda_{i}$ and slope 0 for $\lambda>\lambda_{i}$. As a sum of these, $D(\lambda)$ is a piecewise linear function with kinks at each $\lambda_{i}$. Furthermore, for large enough $\lambda, D(\lambda)$ has slope $-\tau$, and for small enough $\lambda, D(\lambda)$ has slope $n-\tau$. An illustration of $D(\lambda)$ with examples values is shown in Figure 2.

Now, consider a permutation $\sigma$ such that

$$
\lambda_{\sigma(1)} \geqslant \lambda_{\sigma(2)} \geqslant \cdots \geqslant \lambda_{\sigma(n)} .
$$

Because the slope changes by one at each $\lambda_{i}$, the slope is zero on the interval $\left[\lambda_{\sigma(\tau+1)}, \lambda_{\sigma_{\tau}}\right]$. Furthermore, $D(\lambda)$ must be non-increasing above $\lambda_{\sigma(\tau+1)}$. Then either $\lambda_{\sigma(\tau+1)}$ is a maximizing value, if it is non-negative, or otherwise a maximizing value is at $\lambda^{*}=0$.

Once a maximizing value $\lambda^{*}$ is chosen, a corresponding optimal allocation can be created by buying all assets with $\lambda_{i}>\lambda^{*}$. This is a feasible allocation because there are not more than $\tau$ assets that satisfy this by our construction of $\lambda^{*}$.

The limiting step in this algorithm is identifying the $(\tau+1)$ th largest value of $\lambda_{i}$. The key to the $O(n)$ time is to recognize that we do not actually have to sort all of the $\lambda_{i}$ values. The

Figure 2: Example of $D(\lambda)$ in the binary model with $n=3$ and $\tau=2$.

$(\tau+1)$ th largest value may be obtained in linear time via the QuickSelect algorithm[4].

Corollary 1. There is no duality gap in the binary model with homogeneous transaction costs and integer-valued $\tau$.

Proof. First, assume there is no multiplicity in the $\lambda_{i}$ in this problem. Then the algorithm above results in an optimal allocation $w^{*}$ such that $g\left(w^{*}\right)=\tau$. Then by Proposition 2 in Section 2, there is no duality gap.

Now, assume that there is multiplicity in the $\lambda_{i}$. If there is no multiplicity on the $\tau$ th largest value, then there are still no problems. However, if there is multiplicity on the $\tau$ th largest value, then this algorithm returns an optimal allocation $\left(w^{*}, \lambda^{*}\right)$ with $g\left(w^{*}\right) \leqslant \tau$. However, if $g\left(w^{*}\right)<\tau$ then there are $\tau-g\left(w^{*}\right)$ assets with $\lambda_{i}=\lambda^{*}$ which are not being purchased. At this value of $\lambda$, the investor is indifferent between purchasing them or not, so by purchasing them, there is a $\tilde{w}^{*}$ such that $g\left(\tilde{w}^{*}\right)=\tau$ and $f\left(\tilde{w}^{*}\right)=f\left(w^{*}\right)$. Then $p^{*}=d^{*}$, even if this particular algorithm does not return an allocation with $g\left(w^{*}\right)=\tau$.

The homogeneous transaction cost binary model algorithm described above was implemented in MATLAB. In Figure 3, we show a graph of the average run-time versus number of assets. It seems that this algorithm successfully runs in approximately linear time with respect to the number of assets. This feature will hold neither in the models to come nor once we add constraints on net allocations of wealth.

### 3.2 Fully general binary model

To begin generalizing the binary model problem, we consider the case when $\tau \in[0, n]$ is not an integer. Intuitively, because all assets have a transaction cost of one, there is no way that the

Figure 3: Runtime of the homogeneous binary model algorithm compared to number of assets.

solution would be to purchase more than the floor of $\tau$ assets. Then we expect that the minimizer of the problem with the floor of $\tau$ will be a minimizer of the problem with non-integral $\tau$.

Proposition 4. If $\tau \in[0, n]$ is not an integer, then if we modify the normal algorithm to find the $\lfloor\tau\rfloor$ th largest $\lambda_{i}$ then it will output an optimal solution to the dual problem.

Proof. Let $\tau \in[0, n]$ not be an integer, so $\lfloor\tau\rfloor<\tau<\lceil\tau]$. Then following the notation of the theorem about the algorithm for the homogeneous binary model, let $\sigma$ be a permutation such that

$$
\lambda_{\sigma(1)} \geqslant \cdots \geqslant \lambda_{\sigma(n)} .
$$

Then $D(\lambda)$ will change sign at $\lambda_{\sigma([\tau])}$. So $\lambda^{*}=\lambda_{\sigma(\lfloor\tau])}$. Then by construction, the algorithm will produce an allocation $w^{*}$ with $g\left(w^{*}\right) \leqslant\lfloor\tau\rfloor<\tau$.

Unfortunately, as soon as we have non-integer transaction cost limits, there may be a duality gap. This is best illustrated by an example.

Proposition 5. When $\tau \in[0, n]$ is not an integer, it is not true that $p^{*}=d^{*}$ in general.
Proof. Consider a counterexample with only one asset. Let $f(\xi)=\frac{1}{2} \xi^{2}-\xi, b=1$, and $\tau=\frac{1}{2}$. Then it is immediately clear that $p^{*}=0$.

On the other hand, we can easily check that

$$
D(\lambda)=-\frac{1}{2} \lambda+\min \left\{0, \lambda-\frac{1}{2}\right\} .
$$

This function is piecewise linear and maximized at $\lambda=\frac{1}{2}$. Here it takes on a value of $d^{*}=-\frac{1}{4}$.
Therefore, we conclude that $p^{*}>d^{*}$.

This example illustrates that in the general case, we cannot expect strong duality to hold. There will be a potential duality gap, even in the binary model with non-integer transaction cost limits.

We next consider the case when not all $b_{i}$ all equal, which we refer to as the inhomogeneous transaction cost binary model. The first thing to notice is not it is natural to ask that $0 \leqslant \tau \leqslant \sum_{i=1}^{n} b_{i}$. Furthermore, there is no reason to ask $\tau$ to be an integer at all.

In this case we can simplify $D(\lambda)$ to

$$
D(\lambda)=-\tau \lambda+\sum_{i=1}^{n} \min \left\{f_{i}^{0}, b_{i} \lambda+f_{i}^{+}\right\} .
$$

Furthermore, we can generalize the algorithm above. In particular, we now compute $\lambda_{i}=$ $b_{i}^{-1}\left(f_{i}^{0}-f_{i}^{+}\right)$and find the smallest $\lambda^{*} \geqslant 0$ such that the sum of all $b_{i}$ corresponding to $\lambda_{i}>\lambda^{*}$ is less than or equal to $\tau$. This can be done in $O(n \log n)$ by sorting all the $\lambda_{i}$.

Once again, however, it is easy to see that as soon as $b_{i}$ are not equal, there is no reason to expect $p^{*}=d^{*}$, even when $\tau$ is an integer. This is illustrated by an example.

Proposition 6. When $b_{i}$ are not all equal, it is not true that $p^{*}=d^{*}$ in general.
Proof. Consider a counterexample with two assets. Let $f_{1}(\xi)=f_{2}(\xi)=\frac{1}{2} \xi^{2}-2 \xi, b_{1}=1, b_{2}=2$, and $\tau=2$.

First, we compute $p^{*}$ directly. Note, buying neither asset or buying exactly one is feasible, but buying both is infeasible. If we buy an asset, it is optimal to purchase 2 units of it and achieve a payoff of -2 . If we purchase neither, we get a payoff of 0 . Then $p^{*}=-2$.

On the other hand, let us compute $d^{*}$ and compare. It is easy to check that

$$
D(\lambda)=-2 \lambda+\min \{0,2 \lambda-2\}+\min \{0, \lambda-2\} .
$$

This function is piecewise linear and maximized at $\lambda=1$. Here it takes on a value of $d^{*}=-3$.

Therefore, we conclude that $p^{*}>d^{*}$.

These two examples illustrate that general models will always have a potential duality gap. In Section 4 we will consider the more general ternary model, and also provide a method for estimating the size of this duality gap.

## 4 Ternary Model

In this section we consider a ternary model for transaction costs where each asset has a fixed transaction cost $b_{i}>0$ for purchases, and a fixed transaction cost $s_{i}>0$ for short-sales. In particular, we consider the special case of

$$
g_{i}(\xi)=\left\{\begin{array}{cc}
s_{i} & \text { if } \xi<0 \\
0 & \text { if } \xi=0 \\
b_{i} & \text { if } \xi>0 .
\end{array}\right.
$$

Furthermore, let us assume that we have the restriction on $\tau$ that

$$
0 \leqslant \tau \leqslant \sum_{i=1}^{n} \max \left(s_{i}, b_{i}\right)
$$

### 4.1 General results in ternary model

Similar to the binary problem, we can simplify the dual problem. Let $\lambda \geqslant 0$, and note by decomposability that

$$
D(\lambda)=\min _{w \in \mathcal{W}}\left\{-\lambda \tau+\sum_{i=1}^{n}\left(f_{i}\left(w_{i}\right)+\lambda g_{i}\left(w_{i}-w_{i}^{0}\right)\right)\right\} .
$$

As in the binary model, we can separate into separate regimes of buying, holding, and selling. However, in order to fully simplify this problem we require the following lemma about when minimums and maximums may be interchanged.

Lemma 2 (Minimax Lemma). Let $K$ be a compact convex subset of a finite-dimensional vector space and $C$ be a convex subset of a vector space. Let $f$ be a real-valued function defined on $K \times C$ such that

1. $x \mapsto f(x, y)$ is convex and lower-semicontinuous for each $y$,
2. $y \mapsto f(x, y)$ is concave for each $x$.

Then

$$
\inf _{x \in K} \sup _{y \in C} f(x, y)=\sup _{y \in C} \inf _{x \in K} f(x, y)
$$

Proof. For a proof of this theorem, see [6].

Proposition 7. The solution of the dual problem under the ternary model may be written

$$
d^{*}=1^{\top} f^{0}+\min _{u^{ \pm}}\left\{\left(h^{+}\right)^{\top} u^{+}+\left(h^{-}\right)^{\top} u^{-}: u^{ \pm} \geqslant 0, u^{+}+u^{-} \leqslant 1, b^{\top} u^{+}+s^{\top} u^{-} \leqslant \tau\right\}
$$

for appropriate vectors $f^{0}, h^{+}$, and $h^{-}$.
Proof. Recall that we can immediately write the dual problem in the ternary model as

$$
\begin{aligned}
D(\lambda) & =\min _{w \in \mathcal{W}}\left\{-\lambda \tau+\sum_{i=1}^{n}\left(f_{i}\left(w_{i}\right)+\lambda g_{i}\left(w_{i}-w_{i}^{0}\right)\right)\right\} \\
& =-\lambda \tau+\sum_{i=1}^{n} \min _{w_{i} \in \mathcal{W}_{i}}\left\{f_{i}\left(w_{i}\right)+\lambda g_{i}\left(w_{i}-w_{i}^{0}\right)\right\}
\end{aligned}
$$

But because $g_{i}(\xi)$ only takes on three values, we can write each minimization problem as

$$
\min _{w_{i} \in \mathcal{W}_{i}}\left\{f_{i}\left(w_{i}\right)+\lambda g_{w}\left(w_{i}-w_{i}^{0}\right)\right\}=\min \left(f_{i}\left(w_{i}^{0}\right), b_{i} \lambda+\min _{w_{i}^{0}<w_{i} \leqslant \bar{w}_{i}} f\left(w_{i}\right), s_{i} \lambda+\min _{\underline{w}_{i} \leqslant w_{i}<w_{i}^{0}} f\left(w_{i}\right)\right) .
$$

For notational convenience, we define

$$
\left\{\begin{array}{ccc}
f_{i}^{0} & = & f_{i}\left(w_{i}^{0}\right) \\
f_{i}^{+} & = & \min \left\{f_{i}\left(w_{i}\right): w_{i}^{0}<w_{i} \leqslant \bar{w}_{i}\right\} \\
f_{i}^{-} & = & \min \left\{f_{i}\left(w_{i}\right): \underline{w}_{i} \leqslant w_{i}<w_{i}^{0}\right\}
\end{array}\right.
$$

as well as

$$
\left\{\begin{array}{c}
w_{i}^{+}=\arg \min _{w_{i}^{0}<w_{i} \leqslant \bar{w}_{i}} f_{i}\left(w_{i}\right) \\
w_{i}^{-}=\arg \min _{\underline{w}_{i} \leqslant w_{i}<w_{i}^{0}} f_{i}\left(w_{i}\right) .
\end{array}\right.
$$

Finally, if we let $h_{i}^{+}=f_{i}^{+}-f_{i}^{0}$ and $h_{i}^{-}=f_{i}^{-}-f_{i}^{0}$, then we can write the dual problem as

$$
\begin{aligned}
d^{*} & =1^{\top} f^{0}+\max _{\lambda \geqslant 0}\left\{-\lambda \tau+\sum_{i=1}^{n} \min \left(0, h_{i}^{+}+\lambda b_{i}, h_{i}^{-}+\lambda s_{i}\right)\right\} \\
& =1^{\top} f^{0}+\max _{\lambda \geqslant 0}\left\{-\lambda \tau+\sum_{i=1}^{n} \min _{u_{i}^{ \pm}}\left\{\left(h_{i}^{+}+\lambda b_{i}\right) u_{i}^{+}+\left(h_{i}^{-}+\lambda s_{i}\right) u_{i}^{-}: u_{i}^{ \pm} \geqslant 0, u_{i}^{+}+u_{i}^{-} \leqslant 1\right\}\right\} \\
& =1^{\top} f^{0}+\max _{\lambda \geqslant 0} \min _{u^{ \pm}}\left\{\left(h^{+}\right)^{\top} u^{+}+\left(h^{-}\right)^{\top} u^{-}+\lambda\left(b^{\top} u^{+}+s^{\top} u^{-}-\tau\right): u^{ \pm} \geqslant 0, u^{+}+u^{-} \leqslant 1\right\}
\end{aligned}
$$

At this point, however, we have an affine objective function being minimized over a convex compact space in $u^{ \pm}$and maximized over a convex space in $\lambda$. Then we can apply the Minimax Lemma to interchange the min and max, and obtain

$$
\begin{aligned}
d^{*} & =1^{\top} f^{0}+\min _{u^{ \pm}}\left\{\left(h^{+}\right)^{\top} u^{+}+\left(h^{-}\right)^{\top} u^{-}+\max _{\lambda \geqslant 0}\left\{\lambda\left(b^{\top} u^{+}+s^{\top} u^{-}-\tau\right)\right\}: u^{ \pm} \geqslant 0, u^{+}+u^{-} \leqslant 1\right\} \\
& =1^{\top} f^{0}+\min _{u^{ \pm}}\left\{\left(h^{+}\right)^{\top} u^{+}+\left(h^{-}\right)^{\top} u^{-}: u^{ \pm} \geqslant 0, u^{+}+u^{-} \leqslant 1, b^{\top} u^{+}+s^{\top} u^{-} \leqslant \tau\right\},
\end{aligned}
$$

where the extra constraint follows because the maximization over $\lambda$ is $+\infty$ when $b^{\top} u^{+}+s^{\top} u^{-}>\tau$ and zero otherwise.

At this point, it is worth noting that we have transformed a general convex optimization problem into a linear programming problem. Then we immediately have major results and algorithms for this problem.

Theorem 3. We can construct an optimal solution $\left(\lambda^{*}, w^{*}\right)$ to the dual problem in the ternary model in polynomial time via the algorithm shown in Algorithm 3.

Proof. Because the objective function is decomposable into quadratics, the computation of all $w_{i}^{+}$and $w_{i}^{-}$can be done in linear time. The limiting step is obtaining an optimal solution $u^{ \pm}$to the equivalent problem from Proposition 7. This can be done in average-case polynomial time with simplex methods, or even in polynomial worst-case time with ellipsoid methods[3].

Because the value of $u_{i}^{ \pm}$corresponds to the minimum of

$$
\min \left(0, h_{i}^{+}+\lambda b_{i}, h_{i}^{-}+\lambda s_{i}\right)
$$

being attained in the holding, buying, or selling regime, once we have obtained an optimal $u^{ \pm}$ from the equivalent linear programming problem, we can convert back to an optimal portfolio by holding at $w_{i}^{0}$ if $u^{+}=u^{-}=0$, buying the optimal amount $w_{i}^{+}$if $u^{+}=1$, or selling the optimal amount $w_{i}^{-}$if $u^{-}=1$. Note, the linear programming problem may have fractional allocations, which correspond to assets which we are indifferent towards buying, selling, or holding. We choose to hold these in the algorithm in order to guarantee that $g\left(w^{*}\right) \leqslant \tau$.

Finally, in order to obtain $\lambda^{*}$, we need to solve the implicit equation

$$
d^{*}=\mathcal{L}\left(w^{*}, \lambda^{*}\right)
$$

Input: Objective functions $f_{i}$, initial allocations $w_{i}^{0}$, allocation limits $\underline{w}_{i}$ and $\bar{w}_{i}$, trading fees $b_{i}$ and $s_{i}$, and the trading fee limit $\tau$.
Output: An allocation $w^{*}$ corresponding to an optimal solution $\left(\lambda^{*}, w^{*}\right)$ to the dual problem.
for $i \leftarrow 1$ to $n$ do
$w_{i}^{+} \leftarrow \arg \min \left\{f_{i}\left(w_{i}\right): w_{i}^{0}<w_{i} \leqslant \bar{w}_{i}\right\}$
$w_{i}^{-} \leftarrow \arg \min \left\{f_{i}\left(w_{i}\right): \underline{w}_{i} \leqslant w_{i}<w_{i}^{0}\right\}$
$h_{i}^{+} \leftarrow f_{i}\left(w_{i}^{+}\right)-f_{i}\left(w_{i}^{0}\right)$
$h_{i}^{-} \leftarrow f_{i}\left(w_{i}^{+}\right)-f_{i}\left(w_{i}^{0}\right)$
end
$\left(d^{*}, u^{ \pm}\right) \leftarrow$ optimal solution to linear programming problem
for $i \leftarrow 1$ to $n$ do
$w_{i}^{*} \leftarrow 0$
if $u_{i}^{+}=1$ then
$w_{i}^{*} \leftarrow w_{i}^{+}$
end
if $u_{i}^{-}=1$ then
$w_{i}^{*} \leftarrow w_{i}^{-}$
end
end
$\lambda^{*} \leftarrow$ solve $\mathcal{L}\left(w^{*}, \lambda\right)=d^{*}$
return $\left(\lambda^{*}, w^{*}\right)$
Algorithm 3: TernaryProblemSolver

But for fixed $w^{*}$, this is linear in $\lambda$, so this can be done in linear time. Therefore, we obtain an optimal solution $\left(\lambda^{*}, w^{*}\right)$ to the dual problem overall in polynomial time.

The ternary model algorithm described above was implemented in MATLAB. In particular, we use the MATLAB linear programming routines with an active-set algorithm and sparse matrices. In Figure 4, we show a graph of the average run-time versus number of assets. The major difference between this and the general Algorithm 1 is that the latter much iteratively determine $\lambda^{*}$. Therefore, we expect this algorithm will be much faster when transaction cost constraints bind (e.g. $\lambda^{*}>0$ ).

### 4.2 Remarks and special cases

We next consider some special cases and remarks concerning the ternary model. In particular, we focus on when we can simplify the solution to the linear programming problem and potentially speed up the algorithm, and when we can determine information about the existence and size of the duality gap.

Consider the case when $b_{i}=s_{i}$ for all assets $i$. It is clear that we can still have a duality gap in this case, by extending the counterexample in Proposition 6 to the Ternary case. However, we can speed up the ternary algorithm above substantially.

Corollary 2. If $b_{i}=s_{i}$ for all assets $i$, then the solution of the dual problem under the ternary

Figure 4: Runtime of the ternary model algorithm compared to number of assets.

model may be written

$$
d^{*}=1^{\top} f^{0}+\min _{u}\left\{h^{\top} u: 0 \leqslant u \leqslant 1, b^{\top} u \leqslant \tau\right\}
$$

for appropriate vectors $f^{0}$ and $h$. In particular, we can still construct an optimal solution ( $\lambda^{*}, w^{*}$ ) to the dual problem in polynomial time.

Proof. We reproduce the proof for the general case from above, but with additional simplifications because $b_{i}=s_{i}$ for all assets. Where details of the argument are omitted, refer to the previous argument.

Because $g_{i}(\xi)$ only takes on two values now, we can write each separated minimization problem as

$$
\min _{w_{i} \in \mathcal{W}_{i}}\left\{f_{i}\left(w_{i}\right)+\lambda g_{w}\left(w_{i}-w_{i}^{0}\right)\right\}=\min \left(f_{i}\left(w_{i}^{0}\right), b_{i} \lambda+\min _{w_{i} \in \mathcal{W}_{i}} f\left(w_{i}\right)\right) .
$$

Then we define

$$
\left\{\begin{array}{ccc}
f_{i}^{0} & = & f_{i}\left(w_{i}^{0}\right) \\
f_{i} & =\min \left\{f_{i}\left(w_{i}\right): w_{i} \in \mathcal{W}_{i}\right\},
\end{array}\right.
$$

as well as

$$
w_{i}=\underset{w_{i} \in \mathcal{W}_{i}}{\arg \min } f_{i}\left(w_{i}\right) .
$$

If let $h_{i}=f_{i}-f_{i}^{0}$, then we can write the dual problem as

$$
\begin{aligned}
d^{*} & =1^{\top} f^{0}+\max _{\lambda \geqslant 0}\left\{-\lambda \tau+\sum_{i=1}^{n} \min \left(0, h_{i}+\lambda b_{i}\right)\right\} \\
& =1^{\top} f^{0}+\max _{\lambda \geqslant 0} \min _{u}\left\{h^{\top} u+\lambda\left(b^{\top} u-\tau\right): 0 \leqslant u \leqslant 1\right\} .
\end{aligned}
$$

Applying the Minimax Lemma as before to switch the minimum and maximum, we obtain

$$
\begin{aligned}
d^{*} & =1^{\top} f^{0}+\min _{u}\left\{h^{\top} u+\max _{\lambda \geqslant 0}\left\{\lambda\left(b^{\top} u-\tau\right)\right\}: 0 \leqslant u \leqslant 1\right\} \\
& =1^{\top} f^{0}+\min _{u}\left\{h^{\top} u: 0 \leqslant u \leqslant 1, b^{\top} u \leqslant \tau\right\},
\end{aligned}
$$

where the extra constraint follows because the maximization over $\lambda$ is $+\infty$ when $b^{\top} u>\tau$ and zero otherwise. As before, we can write an algorithm to solve this linear programming problem and transform the solution $u^{*}$ into an optimal solution $\left(\lambda^{*}, w^{*}\right)$ to the dual problem in polynomial time.

Note, the overall complexity of the algorithm is the same as that when $b_{i} \neq s_{i}$ because it relies on the complexity of the linear programming problem. However, the importance of this result is that we can increase the speed by a constant factor across the board by utilizing this symmetry in the problem.

Because we have shown that there in general may be a duality gap in the ternary model, we would like to be able to estimate the size of this gap. We show below how we can get an upper bound on the duality gap once we have solved the dual problem.

Proposition 8. Let $\left(\lambda^{*}, w^{*}\right)$ be an optimal solution to the dual problem obtained from the algorithm above. Then the duality gap is bounded by

$$
0 \leqslant p^{*}-d^{*} \leqslant \lambda^{*}\left(\tau-g\left(w^{*}\right)\right)
$$

Proof. First, recall from the proof of the algorithm's validity that we will always have $g\left(w^{*}\right) \leqslant \tau$ for the output. Then any allocation is a feasible allocation of the original problem. Therefore, we have the following inequality

$$
d^{*}=f\left(w^{*}\right)-\lambda^{*}\left(\tau-g\left(w^{*}\right)\right) \geqslant p^{*}-\lambda^{*}\left(\tau-g\left(w^{*}\right)\right)
$$

If we rearrange this inequality and also recall from Section 2 that $p^{*} \geqslant d^{*}$ in general, then we obtain the bound

$$
0 \leqslant p^{*}-d^{*} \leqslant \lambda^{*}\left(\tau-g\left(w^{*}\right)\right)
$$

While this result allows estimation of the duality gap, it is important to keep in mind that we can not control the optimal value of $\lambda^{*}$ in the dual problem a priori. Therefore, this result cannot be used to obtain any sort of convergence of the optimal solution to the dual problem to the optimal solution of the primal problem.

## 5 Taking into account uncertainty and correlations

Our model so far takes into account neither any correlation between the assets nor any uncertainty in the model itself. In this section we generalize to an objective function

$$
f(w, \hat{r})=\frac{1}{2} w^{\top} \Sigma w-\hat{r}^{\top} w
$$

where $\hat{r} \in \mathcal{R} \subset \mathbb{R}^{n}$ contains predicted returns and $\Sigma$ is an estimate of the covariance matrix. For simplicity, we assume that $\mathcal{R}=\mathcal{R}_{1} \times \cdots \times \mathcal{R}_{n}$, where $\mathcal{R}_{i}$ is a given interval $\left[\underline{r}_{i}, \bar{r}_{i}\right]$, where $\underline{r}_{i}<\bar{r}_{i}$. We make the same assumptions of the cost function $g$ as in the ternary model. Then we are considering an optimization problem of the form

$$
p^{*}=\min _{w \in \mathcal{W}} \max _{\hat{r} \in \mathcal{R}}\{f(w, \hat{r}): g(w) \leqslant \tau\}
$$

As before, we consider the associated Lagrangian defined by

$$
\mathcal{L}(w, \hat{r}, \lambda)=f(w, \hat{r})+\lambda(g(w)-\tau)
$$

Then we define a function of $\lambda$ as

$$
D(\lambda)=\min _{w \in \mathcal{W}} \max _{\hat{r} \in \mathcal{R}} \mathcal{L}(w, \hat{r}, \lambda) .
$$

Then the corresponding dual problem is given by

$$
d^{*}=\max _{\lambda \geqslant 0} D(\lambda),
$$

and the theory developed in Section 2 for general dual problems still holds.

### 5.1 Uncertainty in returns and no correlation

We consider the case when $\Sigma$ is diagonal. In this case, the assumption of decomposability of $f$ and $g$ still holds. In particular, the dual problem reduces to

$$
\begin{aligned}
d^{*} & =\max _{\lambda \geqslant 0} \min _{w \in \mathcal{W}} \max _{\hat{r} \in \mathcal{R}}\{-\lambda \tau+f(w, \hat{r})+\lambda g(w)\} \\
& =\max _{\lambda \geqslant 0}\left\{-\lambda \tau+\sum_{i=1}^{n} \min _{w_{i} \in \mathcal{W}_{i}} \max _{\hat{r}_{i} \in \mathcal{R}_{i}}\left\{\frac{1}{2} \sigma_{i}^{2} w_{i}^{2}-\hat{r}_{i} w_{i}+\lambda g_{i}\left(w_{i}-w_{i}^{0}\right)\right\}\right\} .
\end{aligned}
$$

Then we can solve this dual problem with the same algorithm as given in Section 2 as long as finding an optimizer for the problem

$$
\min _{w_{i} \in \mathcal{W}_{i}} \max _{\hat{r}_{i} \in \mathcal{R}_{i}}\left\{\frac{1}{2} \sigma_{i}^{2} w_{i}^{2}-\hat{r}_{i} w_{i}+\lambda g_{i}\left(w_{i}-w_{i}^{0}\right)\right\}
$$

can be efficiently done, algorithmically or otherwise. We will consider alternative methods more generally in the case with correlations.

### 5.2 Uncertainty in returns and certain correlation

We next consider the case when $\Sigma=D+u u^{\top}$, where $D$ is a diagonal matrix containing asset variances and $u \in \mathbb{R}^{n}$ is a vector that models how a specific factor affects the entire asset universe. It is important to note that $\Sigma$ is still positive semi-definite because

$$
w^{\top} \Sigma w=w^{\top} D w+w^{\top} u u^{\top} w=w^{\top} D w+\left(w^{\top} u\right)^{2} \geqslant 0
$$

because $D$ is positive semi-definite by assumption. Then the objective function $f$ is still a convex function.

The major issue here, however, is that we no longer can assume decomposability of $f+\lambda g$ into different assets. One might hope to change basis to diagonalize $\Sigma$, but the transaction costs are not basis-invariant, so this direction is fruitless. In this case, we apply the general algorithm from Section 2. Note that in the dual formulation, we now have

$$
D(\lambda)=\min _{w \in \mathcal{W}} \max _{\hat{r} \in \mathcal{R}}\left\{\frac{1}{2} w^{\top} \Sigma w-\hat{r}^{\top} w+\lambda g(w)-\lambda \tau\right\} .
$$

Fortunately, $D(\lambda)$ is still a concave function with a minimizer $\lambda^{*}$. Furthermore, because the objective function is linear in $\hat{r}$, it must be maximized in this variable at a boundary point.

Because $g$ is lower semi-continuous, we can switch the minimum and maximums by the Minimax Lemma and write this as

$$
D(\lambda)=\max _{\hat{r}_{1} \in\left\{\underline{r}_{1}, \bar{r}_{1}\right\}} \cdots \max _{\hat{r}_{n} \in\left\{\underline{r}_{n}, \bar{r}_{n}\right\}} \min _{w \in \mathcal{W}}\left\{\frac{1}{2} w^{\top} \Sigma w-\hat{r}^{\top} w+\lambda g(w)-\lambda \tau\right\} .
$$

Next, we partition $\mathcal{W}$ into $3^{n}$ regions corresponding to holding, buying, or selling each asset. Consider the following notation:

$$
\left\{\begin{array}{c}
\mathcal{W}_{i}^{0}=\left\{w_{i}^{0}\right\} \\
\mathcal{W}_{i}^{+}=\left[w_{i}^{0}, \bar{w}_{i}\right] \\
\mathcal{W}_{i}^{-}=\left[\underline{w}_{i}, w_{i}^{0}\right] .
\end{array}\right.
$$

Then if for any string $\alpha \in\{0,+,-\}^{n}$, we define

$$
\mathcal{W}^{\alpha}={\underset{i=1}{n} \mathcal{W}_{i}^{\alpha_{i}}, ~}_{\text {, }}
$$

we can write the optimization problem as

$$
D(\lambda)=\max _{\hat{r}_{1} \in\left\{\underline{\underline{I}}_{1}, \bar{r}_{1}\right\}} \cdots \max _{\hat{r}_{n} \in\left\{\underline{r}_{n}, \bar{r}_{n}\right\}} \min _{\alpha \in\{0,+,-\}^{n}} \min _{w \in \mathcal{W}^{\alpha}}\left\{\frac{1}{2} w^{\top} \Sigma w-\hat{r}^{\top} w+\lambda g(w)-\lambda \tau\right\} .
$$

It is now clear that, this problem consists of solving $6^{n}$ quadratic programming problems. However, we can cut down on this complexity by employing branch-and-bound techniques to only explore a subset of these combinations.

The idea of our branch-and-bound procedure is to consider a tree generated by choices to hold, buy, or sell each subsequent asset. If we can obtain a lower-bound to an entire sub-tree in constant time, and find that this lower-bound is greater than an upper-bound for the entire tree, then we can ignore that sub-tree. Then the key is to describe a constant time lower-bound of a sub-tree, as well as a procedure for obtaining an upper-bound to an entire tree. For simplicity, we let $\underline{r}=\bar{r}$ and $w^{0}=0$, though all proceeding arguments could be modified for the more general case.

Lemma 3. Let $\alpha \in\{0,+,-\}^{n}$ and $1 \leqslant k<n$. Consider the problem

$$
d_{k}^{\alpha}=\min \left\{\frac{1}{2} w^{\top} \Sigma w-\hat{r}^{\top} w-\lambda \tau+\sum_{i=1}^{k} \lambda g_{i}\left(w_{i}\right): w \in \mathcal{W}_{1}^{\alpha_{1}} \times \cdots \mathcal{W}_{k}^{\alpha_{k}} \times \mathcal{W}_{k+1} \times \cdots \times \mathcal{W}_{n}\right\}
$$

Then we have

$$
d_{k+1}^{\alpha} \geqslant d_{k}^{\alpha}+\lambda b_{k+1} \chi_{\left\{\alpha_{k+1}=+\right\}}+\lambda s_{k+1} \chi_{\left\{\alpha_{k+1}=-\right\}},
$$

where $\chi$ is an indicator function.
Proof. This inequality is nearly trivial from the definition of $d_{k}^{\alpha}$ and the fact that $b_{i}, s_{i} \geqslant 0$. In particular, for $1 \leqslant k<n$, let $w$ be a minimizer of the $(k+1)$-problem. Then this is also a feasible solution of the $k$-problem, so we have

$$
\begin{aligned}
d_{k+1}^{\alpha} & =\frac{1}{2} w \top \Sigma w-\hat{r}^{\top} w-\lambda \tau+\sum_{i=1}^{k+1} \lambda g_{i}\left(w_{i}\right) \\
& =\lambda g_{k+1}\left(w_{k+1}\right)+\frac{1}{2} w \top \Sigma w-\hat{r}^{\top} w-\lambda \tau+\sum_{i=1}^{k} \lambda g_{i}\left(w_{i}\right) \\
& \geqslant d_{k}^{\alpha}+\lambda g_{k+1}\left(w_{k+1}\right) \\
& =d_{k}^{\alpha}+\lambda b_{k+1} \chi_{\left\{\alpha_{k+1}=+\right\}}+\lambda s_{k+1} \chi_{\left\{\alpha_{k+1}=-\right\}} .
\end{aligned}
$$

Figure 5: Illustration of branch-and-bound method for decision to hold, buy, or sell at asset $i$. Sub-trees that have a lower-bound greater than the entire tree's upper-bound are ignored. Otherwise, the process is iterated.


The key to this definition is that each $d_{k}^{\alpha}$ is only a normal quadratic programming problem because the cost function is a constant in the region over which we are minimizing. Furthermore, given a minimal value of $d_{k}^{\alpha}$, we can obtain a lower-bound on the value of $d_{k+1}^{\alpha}$ in constant-time for use in the branch-and-bound method. Furthermore, for at least one $\alpha, d_{n}^{\alpha}$ will correspond to the actual Lagrangian subproblem, $D(\lambda)$.

Lemma 4. There exists an $\alpha \in\{0,+,-\}^{n}$ such that

$$
d_{0}^{\alpha} \leqslant d_{1}^{\alpha} \leqslant \cdots \leqslant d_{n-1}^{\alpha} \leqslant d_{n}^{\alpha}=D(\lambda) .
$$

Proof. Let $w$ be a minimizer of the Lagrangian sub-problem, so $\mathcal{L}(w, \lambda)=D(\lambda)$. Define $\alpha \in$ $\{0,+,-\}^{n}$ via

$$
\alpha_{k}= \begin{cases}0 & \text { if } w_{k}=0 \\ + & \text { if } w_{k}>0 \\ - & \text { if } w_{k}<0\end{cases}
$$

Then it is clear from the definition that $d_{n}^{\alpha}=D(\lambda)$. Furthermore, by the previous lemma, when $\lambda \geqslant 0$, we have

$$
d_{0}^{\alpha} \leqslant d_{1}^{\alpha} \leqslant \cdots \leqslant d_{n-1}^{\alpha} \leqslant d_{n}^{\alpha}=D(\lambda) .
$$

With these results in hand, we can devise an algorithm to solve the Lagrangian sub-problem via the branch-and-bound procedure.

Theorem 4. We can construct an optimal solution $w^{*}$ to the Lagrangian sub-problem in the correlation model via the branch-and-bound algorithm shown in Algorithm 4.

Input: Correlation matrix $\Sigma$, expected returns $\hat{r}$, initial allocations $w_{i}^{0}$, allocation limits $\underline{w}_{i}$ and $\bar{w}_{i}$, trading fees $b_{i}$ and $s_{i}$, a Lagrange multiplier $\lambda$, and the trading fee limit $\tau$.
Output: An allocation $w^{*}$ which minimizes the Lagrangian sub-problem $D(\lambda)$.
Create priority queue $p q$
$\alpha \leftarrow(0,0, \cdots 0)$
$(w, d) \leftarrow$ optimal allocation and value corresponding to $d_{0}^{\alpha}$
$\left(w^{*}, d^{*}\right) \leftarrow(w, \mathcal{L}(w, \lambda))$
Push $(d, w, 1)$ to $p q$ with priority $d$
while $p q$ is not empty do
$(d, w, k) \leftarrow \operatorname{pop}(p q)$
for $\alpha_{k} \leftarrow\{0,+,-\}$ do
$d^{\prime} \leftarrow d+\lambda b_{k} \chi_{\left\{\alpha_{k}=+\right\}}+\lambda s_{k} \chi_{\left\{\alpha_{k}=-\right\}}$
if $d^{\prime} \geqslant d^{*}$ then Continue
$\left(w^{\prime}, d^{\prime}\right) \leftarrow$ optimal allocation and value corresponding to $d_{k}^{\alpha}$
if $d^{\prime} \geqslant d^{*}$ then Continue
if $\mathcal{L}\left(w^{\prime}, \lambda\right) \leqslant d^{*}$ then $\left(w^{*}, d^{*}\right) \leftarrow\left(w^{\prime}, \mathcal{L}\left(w^{\prime}, \lambda\right)\right)$
if $k<n$ then Push $\left(d^{\prime}, w^{\prime}, k+1\right)$ to $p q$ with priority $d^{\prime}$
end
end
return $w^{*}$

## Algorithm 4: CorrelationProblemSolver

Proof. The queue in the algorithm above will only add elements up to $k=n$, so at most there can be $3^{n}$ elements added to the queue. Therefore, this algorithm will terminate eventually. To see that it will return an optimal allocation of the Lagrangian sub-problem, we consider the previous lemmas. We know there exists an $\alpha \in\{0,+,-\}^{n}$ such that

$$
d_{0}^{\alpha} \leqslant d_{1}^{\alpha} \leqslant \cdots \leqslant d_{n}^{\alpha}=D(\lambda)
$$

At any point in the algorithm, $d^{*}=\mathcal{L}(w, \lambda)$ for some $w$. Then for any $k$, we have

$$
d_{k}^{\alpha} \leqslant D(\lambda) \leqslant d^{*}
$$

Therefore, the continues inside the while loop will never be triggered along this sequence corresponding to $\alpha$. Finally, if $w_{n}^{\alpha}$ is the optimal allocation corresponding to $d_{k}^{\alpha}$, then we will have

$$
\mathcal{L}\left(w_{n}^{\alpha}, \lambda\right)=D(\lambda) \leqslant d^{*}
$$

so at some point $w^{*}$ will be set to this optimal allocation. Because $w^{*}$ is a minimizer of the Lagrangian, the condition to reset $w^{*}$ will never we met again unless by another minimizer, so $w^{*}$ will be an optimal allocation of the Lagrangian sub-problem.

For more information on the theory and general results concerning branch-and-bound methods, see [2]. For our purposes, we will note that this allows us to obtain an optimizing vector $w^{*}$ to the Lagrangian sub-problem. We can then compute a super-gradient as $D^{\prime}(\lambda)=g\left(w^{*}\right)-\tau$, and we can proceed with the Algorithm 1 from Section 2.

### 5.3 Uncertainty in both returns and correlation

Lastly, we consider the case when $\Sigma=D+u u^{\top}+\Delta$, with $\Delta$ a positive semi-definite matrix that models error in our measurement of the covariance matrix. Here, $\Delta$ is unknown but bounded: we assume $\|\Delta\| \leqslant \rho$ where $\rho \geqslant 0$ is a known scalar and $\|\cdot\|$ denotes the largest singular value norm. Once again, it is important that $\Sigma$ is still a positive semi-definite operator, as discussed in the previous subsection. Because the results of the previous subsection are a special case of this problem, with $\rho=0$, we expect that the results from the general binary and ternary models will still not hold.

Consider the dual formulation corresponding to this problem. If we hope to apply the general algorithm from Section 2, then we need to be able to solve the optimization problem

$$
D(\lambda)=\min _{w \in \mathcal{W}} \max _{\hat{r} \in \mathcal{R}} \max _{\|\Delta\| \leqslant \rho}\left\{\frac{1}{2} w^{\top} \Sigma w-\hat{r}^{\top} w+\lambda g(w)-\lambda \tau\right\}
$$

The key to simplifying this problem is to note that the maximization over all positive semidefinite matrices is easy to do analytically.

Lemma 5. Let $w \in \mathbb{R}^{n}$. Then if we maximize over all positive semi-definite matrices $\Delta$ with largest singular value norm $\|\Delta\| \leqslant \rho$, we have

$$
\max _{\|\Delta\| \leqslant \rho} w^{\top} \Delta w=\rho w^{\top} w
$$

Proof. If $w=0$, then the result is trivial. Otherwise, the key is to recall that the largest singular value norm of a matrix $\Delta$ is equal to

$$
\max _{w \neq 0} \frac{\|\Delta w\|}{\|w\|}=\|\Delta\| \leqslant \rho
$$

Then for any $w$, we have $\|\Delta w\| \leqslant \rho\|w\|$. By the Cauchy-Schwarz inequality, we have

$$
w^{\top} \Delta w \leqslant\|w\|\|\Delta w\| \leqslant \rho w^{\top} w
$$

Finally, for any $w \neq 0$, we can construct a positive semi-definite matrix $\Delta$ which is $\rho$ times the projection operator onto the subspace spanned by $w$. This matrix has $\|\Delta\|=\rho$ and clearly satisfies $w^{\top} \Delta w=\rho w^{\top} w$.

Therefore, we conclude that $\max _{\|\Delta\| \leqslant \rho} w^{\top} \Delta w=\rho w^{\top} w$.

Using this result, we can simplify the problem above into a problem from the previous subsection.

Theorem 5. Adding uncertainty to asset correlations in the form of a positive semi-definite matrix $\Delta$ with largest singular value norm $\|\Delta\| \leqslant \rho$ is equivalent to adding a constant $\rho$ to the variance of each asset.

Proof. To prove this, we simply consider the $D(\lambda)$ from above. Because the maximization over $\Delta$ comes first, we can immediately use the lemma from above.

$$
\begin{aligned}
D(\lambda) & =\min _{w \in \mathcal{W}} \max _{\hat{r} \in \mathcal{R}} \max _{\|\Delta\| \leqslant \rho}\left\{\frac{1}{2} w^{\top}\left(D+u u^{\top}+\Delta\right) w-\hat{r}^{\top} w+\lambda g(w)-\lambda \tau\right\} \\
& =\min _{w \in \mathcal{W}} \max _{\hat{r} \in \mathcal{R}}\left\{\frac{1}{2} w^{\top}\left(D+\rho+u u^{\top}\right) w-\hat{r}^{\top} w+\lambda g(w)-\lambda \tau\right\}
\end{aligned}
$$

Then we have shifted the uncertainty $\Delta$ onto the diagonal of the covariance matrix. This corresponds to an increase in asset variances in $D$.

This result shows that really the most general problem we are considering in this paper is that from the previous subsection. For an implementation of the correlation model in MATLAB, see the general algorithm in Appendix A.

## 6 Conclusion

In this paper, we have examined the general principle of duality and its applications to portfolio optimization with constraints in the form of transaction costs. We considered three main models of increasing complexity and generality, including the binary model, the ternary model, and the correlation model. The overarching theme is that decomposability of objective functions and transaction costs allows for simplifications of the problem. In the correlation model, the separability of the problem was violated, but we used a branch-and-bound algorithm to dramatically cut down on the overall run-time.

Future work could focus on improving the implementation of these algorithms, and in particular taking advantage of obvious parallelism opportunities. For example, the binary and ternary models both require sequential optimization of individual assets after decomposing the objective function, which could easily be done in parallel. Furthermore, the linear programming algorithm used in the ternary model and quadratic programming algorithm used in the correlation model can both be sped up using parallelism. Finally, and most importantly, the branch-and-bound algorithm used in Section 5 performs several independent minimization problems on different sub-trees of the combinatorial search space. These could each be done in parallel to boost run-time of the algorithm.

Another avenue of future work would be to develop further approximation methods of the dual problem in the correlation method. For example, if it were not for the appearance of the non-polynomial transaction cost terms, one promising strategy would be to change to a basis of eigenvectors of the covariance matrix $\Sigma$. One could hope to approximate the transaction costs with functions that allow this change of basis. This would lead to a variant of the branch-and-bound algorithm where we instead perform relaxation on the cost function. This could substantially increase the speed of obtaining an $\epsilon$-suboptimal solution to the dual problem in practice.

## References

[1] S. Boyd and L. Vandenberghe. Convex Optimization. Cambridge University Press, 2004.
[2] M.J. Brusco and S. Stahl. Branch-and-Bound Applications in Combinatorial Data Analysis. Springer, 2005.
[3] V. Chvatal. Linear Programming. W. H. Freeman, 1983.
[4] T.H. Cormen, C.E. Leiserson, R.L. Rivest, and Stein C. Introduction to Algorithms, Second Edition. The MIT Press, 2001.
[5] Signal Processing Microelectronics. Qpc - quadratic programming in c, 2013.
[6] M. Willem. Minimax Theorems. Birkhauser, 1996.

## Appendix A. Numerical Implementation

In this Appendix, we describe a general solver for all three models which was implemented in MATLAB. In this section we add the extra constraint on overall net wealth being invested,

$$
\sum_{i=1}^{n} w_{i} \leqslant 1
$$

It is worth noting that this condition breaks the separability that allowed for general polynomialtime results in the binary and ternary models.

In the attached MATLAB codes, we provide one general program which computes allocations corresponding to optimizers $\left(w^{*}, \lambda^{*}\right)$ of the dual problem in each model, subject to constraints on transaction costs and the overall net investment constraint above. Because the correlation model generalizes both the ternary and binary model, we use the super-gradient algorithm in Algorithm 1 with the Lagrangian dual optimized via the branch-and-bound algorithm in Algorithm 4. To solve quadratic programming problems, we use the a C-based Quadratic Programming MATLAB interface known as QPC[5]. In particular, we use a primaldual predictor-corrector algorithm for quadratic programming problems constrained by linear inequalities. We use sparse matrices in the binary and ternary models and dense matrices in the correlation model.

To test this algorithm in general cases, we construct random inputs. In order to make sure that the transaction cost and allocation constraints actually bind in general, we generated inputs as follows. Returns are drawn uniformly from $[-1,1]$, standard deviations are set to 1 , upper bounds to 1 , lower bounds to -1 , buy costs to 0.01 and sell costs to 0.02 . The transaction cost limit $\tau$ is set to one-half the sum of the buy costs plus a uniform random number on $[0,0.001]$ so that the inputs are generic in terms of transaction costs. Finally, as applicable, each component of the vector $u$ which controls correlations is drawn uniformly from $[-0.05,0.05]$.

In Figure 6, we show a graph of the average run-time versus number of assets and desired accuracy $\epsilon$ in the full correlation model with a branch-and-bound algorithm. Each line corresponds to a different cut-off for the accuracy of the solution to the dual problem, as described

Figure 6: Timings of full correlation branch-and-bound algorithm with varying number of assets and desired accuracy, $\epsilon$.


Figure 7: Timings of binary, ternary, and full correlation models in the general purpose algorithm with varying number of assets and desired accuracy $\epsilon=0.01$.

in Algorithm 1. In particular, the increase in computation time appears to be linear in $\epsilon$, but exponential in the number of assets. In Figure 7, we show a comparison of the scaling of the binary model, the ternary model, and the full correlation model in this general purpose implementation. In particular, while the correlation model appears to scale exponentially with respect to the number of assets, even with the branch-and-bound algorithm, the binary and ternary models exhibit superior scaling.

