Computing Near Optimal Strategies for Stochastic Investment Planning Problems

Milos Hauskrecht[1], **Gopal Pandurangan**[1,2] **and Eli Upfal**[1,2]
Computer Science Department, Box 1910
Brown University
Providence, RI 02912, U.S.A.
{*milos, gopal, eli*}@*cs.brown.edu*

Abstract

We present efficient techniques for computing near optimal strategies for a class of stochastic commodity trading problems modeled as Markov decision processes (MDPs). The process has a continuous state space and a large action space and cannot be solved efficiently by standard dynamic programming methods. We exploit structural properties of the process, and combine it with Monte-Carlo estimation techniques to obtain novel and efficient algorithms that closely approximate the optimal strategies.

1 Introduction

Investment is an act of incurring immediate cost in the expectation of future rewards. Investments options are typically characterized by three parameters: the initial and accumulated costs, the uncertainty over the future rewards, and the leeway in timing the actions. Investors try to optimize their decision with respect to these parameters. Modern economics theory models the uncertainty of future rewards as a stochastic process defining future price curves. The process is typically an *Ito process* [Dixit and Pindyck, 1994] that is Markovian, thus investment decision can be modeled as a Markov decision process (MDP) where a state of the underlying process needs only to include the current investment portfolio and current prices. While the MDP gives a succinct formalization of the investment decision processes it does not necessarily imply efficient algorithms for computing optimal strategies. The Markov decision process has a continuous state space and a potentially large action space and cannot be solved efficiently by standard dynamic programming methods. A challenging goal in this research area is to characterize special cases of the general investment paradigm that are interesting enough from the application point of view while simple enough to allow efficiently computable analytic solutions.

In this work we focus on stochastic planning in the context of commodity trading. Commodity can be bought, stored and eventually sold. In addition to the initial cost of buying the commodity, the investment decision must take into account the accumulated cost of storage till the commodity is sold. A standard assumption in mathematical economics is that commodity prices (e.g oil and copper) are best modeled as a *mean reverting* stochastic process. We study several versions of the commodity trading problem. In the simplest version we assume that there are no restrictions on the amount of commodity that an investor can buy, store or sell at a given time. We give an efficiently computable optimal analytic solution for that case. The problem becomes significantly harder when we turn to a more realistic setting in which there are constrains on the amount of commodity that a trader can buy store and in particular sell at a given time. We present an efficient Monte-Carlo technique that generates an approximate optimal trading policy with volume and storage constrains.

There has been extensive research in AI in recent years on solving MDPs with large state spaces exploiting specific problem structures, in particular through factoring and decompositions [Boutilier *et al.*, 1995; Dearden and Boutilier, 1997; Dean and Lin, 1995; Meuleau *et al.*, 1998]. However, all these work assume finite or at least discrete state space. Our solution relies on structure analysis of the problem combined with a Monte-Carlo approximation technique. We show that when prices follow a mean-reverting process, the optimal policy for the constrained commodity trading problem has an elegant compact parametric form. This observation allows for a significant reduction in the space of possible optimal policies. Furthermore, we show that individual parameters, which represent a stack of decision thresholds, can be optimized incrementally thus further reducing the complexity of the problem-solving task.

2 The Model

Mathematical economics models commodity price fluctuations as an Ito mean-reverting process (Ornstein-Uhlenbeck process) [Dixit and Pindyck, 1994], where changes in price p satisfy a stochastic differential equation

$$dp = \eta(\mu - p)dt + \sigma dz,$$

such that dz is the (random) increment of a Wiener process (Brownian motion with normally distributed increments), μ is the long term average price of the commodity i.e. a value to which the process reverts, η is the speed of reversion, and σ is the standard deviation of the random component.

[1] Supported by a research grant from Goldman Sachs.
[2] Supported in part by NSF grant CCR – 9731477.

Trading occurs at discrete time steps, and price fluctuation between two consecutive steps is modeled by the discrete version of the mean-reverting process [Dixit and Pindyck, 1994]:

$$p_{t+1} = \mu - e^{-\eta}(\mu - p_t) + \epsilon_t, \qquad (1)$$

where ϵ_t is a sequence of independent random variables with normal distribution $N(0, \sigma_\epsilon)$, $\sigma_\epsilon^2 = \frac{\sigma^2}{2\eta}(1 - e^{-2\eta})$.

The agent (investor) can buy and sell the commodity at different time steps. There is a fixed cost q for storing a unit of community for a unit time interval. The cost of money (the interest rate) is denoted by r. We assume that the volume of commodity traded by the agent is small with respect to the total market of that commodity, so that the agent's strategy does not affect the above parameters. In the more general setting there are limits on the number of units a trader can buy and sell in every step, and on the total number of units that can be stored (denoted $C_{\text{buy}}, C_{\text{sell}}, C_{\text{store}}$ respectively).

Objectives

Our goal is to determine the best, profit maximizing, trading strategy for a given initial state, which is determined by the current price and amount of commodity the agent owns. We use the standard valuation method, *the expected net present value (NPV)* function (see e.g. [Brealey and Myers, 1991; Trigeorgis, 1996])):

$$V^\pi(s) = E(\sum_{t=0}^{T} \gamma^t m_t | \pi, s)$$

where s denotes an initial state, π is a strategy, $\gamma = \frac{1}{1+r}$ is a discount factor, T is the decision horizon, and m_t is cash flow at time t. Thus our goal is to find π such that $V^\pi(s)$ is maximized. The objective function corresponds to discounted versions of standard finite or infinite horizon criteria (see [Bellman, 1957; Puterman, 1994]) where rewards correspond to cash flows.

In the following we focus on the infinite horizon investment problem. The trading strategy for this case is represented as $\pi : R \times N \to N$ mapping the current price p_t and number of units of commodity we own – c_t to the number of units to be held in the next step – c_{t+1}. The results for the infinite trading horizon can be adopted to the finite horizon problem.

Example

Our model fits a variety of applications. A typical example is oil-trading problem. Oil can be bought and sold in the market. The changes of oil prices on the market follow the mean reverting process [Dixit and Pindyck, 1994] and individual trading activities do not affect the market price. Oil can be stored at a fixed cost per unit, and there are obvious constraints on the amounts that can be stored and released in short time intervals.

3 Standard approach and its shortcomings

The value (expected NPV) of a strategy π can be expressed using the Bellman equation [Bellman, 1957]:

$$V^\pi(p, c) = R(p, c, \pi(p, c)) + \gamma \int_{-\infty}^{\infty} V^\pi(p', \tau(c, \pi(p, c))) f(p'|p) dp'$$

where a state is a cross-product of a commodity price p and the amount of commodity held c; $R(p, c, \pi(p, c))$ denotes the expected one-step cash flow resulting from action $\pi(p, c)$; τ maps the current commodity holdings to the next state commodity holdings under the action dictated by the policy π; and $f(p'|p)$ denotes a distribution of next state prices p' (a normal distribution defined by Equation 1). In general it is extremely hard to evaluate a given strategy because of the size of a state space we face (in one step we can reach any price) and there is no closed form solution to the integral part. The task of finding the optimal policy is even worse as the space of policies we need to choose from is even larger. Thus, standard dynamic programming approaches cannot be applied directly and we must resort to methods that exploit the problem structure.

4 Trading without constraints

To solve the case with no limits on buy, sell and store activities, it is sufficient to study the problem of investing one unit of commodity, since the strategy for one unit can be replicated for any number of units.

A strategy for the one-unit problem at any point in time is restricted to only two choices: 0 (do not invest, do not hold the commodity) and 1 (hold, invest in the commodity). Because buy and sell prices are always the same and prices change independently of our trading, the expected net present value of a one-unit strategy π (V^π) can be expressed in terms of step-wise gains:

$$V^\pi(p_0, c_0) = \lim_{T \to \infty} E(p_0 c_0 + \sum_{t=0}^{T} \gamma^t g_t \text{hold}_t | \pi, p_0, c_0)$$

where $g_t = -p_t - q + \gamma p_{t+1}$ is the gain from time step t, q is a fixed storage cost, $\text{hold}_t = 1$ if the strategy π chooses to hold the commodity at time t and $\text{hold}_t = 0$ when not.

Intuitively, we can replicate payoffs from any strategy by always selling the commodity held at the end of the step and buy it back when the strategy requires us to hold the commodity in the next time step. The term $p_0 c_0$ reflects the fact that we always sell (pretend to sell) the initial commodity at price p_0. Thus, we can restrict our search to policies of the form $\pi : R \to \{0, 1\}$, mapping prices to actions. The value function for such a policy π can be rewritten as:

$$V^\pi(p_0, c_0) = p_0 c_0 + V^\pi(p_0, 0),$$

where $V^\pi(p_0, 0)$ is the value function for the policy π starting from a zero commodity state. $V^\pi(p, 0)$ satisfies

$$V^\pi(p, 0) = E(g_1(p))\pi(p) + \gamma \int_{-\infty}^{\infty} V^\pi(p', 0) f(p'|p) dp'. \qquad (2)$$

$E(g_1(p))$ is the *expected gain* for holding the commodity for one step starting from price p

$$E(g_1(p)) = -p - q + \gamma E(p'|p),$$

p' denotes the next step price, $q \geq 0$ is a storage cost and $E(p'|p)$ is the expected next step price. A nice feature of Equation 2 is that the contribution from following the policy π in the future (integral part) is independent of the action choice in the first step. This leads to the fact that the greedy one-step lookahead strategy is optimal.[1]

Theorem 1 *The strategy π maximizing the expected one-step gain is optimal.*
Proof Follows directly from Equation 2 and the independence of future values on current actions. □

The value of the optimal one-unit policy can be written as:

$$V_1^*(p,0) = \max_{j=0,1}\left[\sum_{i=0}^{j} E(g_i(p))\right] + \gamma\int_{-\infty}^{\infty} V^*(p',0)f(p'|p)dp', \tag{3}$$

with $E(g_0(p)) = 0$ denoting the expected gain for holding zero units of commodity. The optimal policy satisfies:

$$\pi^*(p) = \arg\max_{j=0,1}\left[\sum_{i=0}^{j} E(g_i(p))\right].$$

The important thing about this is that we know how to find the optimal policy easily and this despite the fact that its value is hard to compute.

4.1 Decision threshold

As the expected gain from not holding the commodity is 0, the choice to hold the commodity is justified only when $E(g_1(p)) \geq 0$, that is, when the expected gain is positive. Substituting p' using equation 1, we get the condition for choosing the "hold" action:

$$p \leq \frac{\gamma\mu(1-e^{-\eta}) - q}{1-\gamma e^{-\eta}} = p_1^* \tag{4}$$

Thus, we obtain a simple and compact policy: *Hold the commodity when the price p is lower than p_1^*.* Note that the optimal threshold price p_1^* depends solely on the parameters of the mean reverting process and discount γ.

5 Trading with constraints

We consider three types of constraints: buy, sell and store limits. We start with buy and store constraints.

5.1 Buy and store constraints

The buy and store constraints are relatively easy to handle when no restriction on the sell is imposed. Simply, we always want to take the maximum advantage of a positive gain for investing in the next step. Thus, we always hold the maximum amount of commodity allowed by the buy and store limits, whenever the optimal one-unit policy recommends to hold. That is:

$$\pi(p,c) = \begin{cases} \min[C_{\text{store}}; c + C_{\text{buy}}] & \text{if } p \leq p_1^* \\ 0 & \text{otherwise.} \end{cases}$$

C_{buy} and C_{store} denote buy and store constraints.

[1] We note the optimality of a greedy one step lookahead strategy holds not only for the mean reverting process, but also for a more general Ito process.

5.2 Sell limit problem

The sell limit is tricky, since a trader may not be able to immediately realize all gains if he holds more units than the sell limit. A naive solution would be to limit the holding to the sell limit, but such policy is obviously sub-optimal.

In studying this case we can assume w.l.o.g. that the sell limit is 1. The solution for an arbitrary sell constraint $C_{\text{sell}} \geq 1$ is then obtained by replicating the optimal strategy with sell limit 1. In the following we first study a special case in which a trader is allowed to hold at most 2 units of commodity (with sell constraint 1) and find the corresponding optimal policy. Later we generalize the idea to any number of units.

5.3 Trading two units under sell limit

To solve the two units case we restrict our attention to strategies that always try to trade the first unit using the optimal one-unit strategy. Then our goal is to simply find how and when to invest in the second commodity. We show later that this approach indeed leads to the optimal solution.

Let $\tilde{\Pi}_2$ be a set of policies $\tilde{\pi}_2 : R \to \{0,1,2\}$ mapping current prices to commodity choices, such that $\tilde{\pi}_2$ always follows the optimal one-unit strategy for the first unit and chooses to hold the second unit arbitrarily but only when $p \leq p_1^*$. Thus $\tilde{\pi}_2$ is described as:

$$\tilde{\pi}_2(p) = \begin{cases} 1 \text{ or } 2 & p \leq p_1^* \\ 0 & p \geq p_1^* \end{cases}$$

Note that a strategy $\tilde{\pi}_2 \in \tilde{\Pi}_2$ may not be directly applicable to the two-unit case and can violate the sell capacity. This happens when $\tilde{\pi}_2$ recommends to hold two units of the commodity for some price $p \leq p_1^*$ in one step and is forced to recommend 0 units when the next step price p' jumps above p_1^*.

To fix this problem we define a set of policies Π_2, such that for every $\tilde{\pi}_2 \in \tilde{\Pi}_2$ there is a policy $\pi_2 : R \times \{0,1,2\} \to \{0,1,2\}$ in Π_2, mapping current price and current commodity holding to the next step commodity choice and is defined as:

$$\pi_2(p,c) = \max[\tilde{\pi}_2(p), c-1]. \tag{5}$$

The idea behind π_2 is that it replicates $\tilde{\pi}_2$ when it is consistent with sell constraints, otherwise it recommends to reduce the number of units of commodity held by 1. Because $\tilde{\pi}_2$ induces π_2 we call it a *generating policy*.

To find the optimal investment rule for the second unit we try to quantify its added value. To do this, we create a strategy π_2' that recommends to hold two units of commodity when $p \leq p_1^*$ and this only for the first step, otherwise it follows the optimal one-unit policy. The value for such a policy for price $p \leq p_1^*$, and zero units of commodity in terms of gains is:

$$V^{\pi_2'}(p,0) = 2E(g_1(p)) + \gamma\int_{-\infty}^{\infty}\left[E(g_1(p')) + \gamma\int_{-\infty}^{\infty} V_1^*(p'',0)f(p''|p')dp''\right] f(p'|p)dp'.$$

Note that due to the sell limit the second unit of commodity is always held also in the second step. This is captured by the term $E(g_1(p'))$. The key trick now is to rewrite $V^{\pi_2'}(p,0)$ for $p \leq p_1^*$ as:

$$V^{\pi_2'}(p,0) = E(g_2(p)) + V_1^*(p,0),$$

where

$$E(g_2(p)) = E(g_1(p)) + \gamma \int_{p_1^*}^{\infty} E(g_1(p'))f(p'|p)dp' \quad (6)$$

is the *expected added gain* for investing in the second unit of commodity for $p \leq p_1^*$. $E(g_2(p))$ represents the difference from investing in the second unit of commodity at $p \leq p_1^*$ compared to the optimal one-unit strategy. It consists of two terms: the expected gain from holding the commodity for one step plus a correction term (a kind of expected loss or negative gain) for holding it one more step in the case we would like to sell it ($p' \geq p_1^*$), but sell constraint does not allow us to do that. Note that $E(g_2(p))$ depends solely on the price of a commodity and also $E(g_2(p)) \leq E(g_1(p))$.

The other important feature is that by substituting $V_1^*(p,0)$ from Equation 3, the value $V^{\pi_2'}(p,0)$ for $p \leq p_1^*$ equals:

$$V^{\pi_2'}(p,0) = \sum_{i=0}^{2} E(g_i(p)) + \gamma \int_{-\infty}^{\infty} V_1^*(p',0)f(p'|p)dp',$$

which means that the integral part of the expression now becomes independent of the commodity held in the first step and disregards any units we were not able to sell. Intuitively, under expectations, the term $E(g_2(p))$ allows us to pretend that we were able to sell both units at the end of the previous step without restriction, though in reality when the price p' climbs above p_1^* we have to keep one unit and sell it later when the sell capacity is available.

Using the fact that $E(g_2(p))$ allows us to disregard any commodity left from previous steps we can express the value function for an arbitrary policy $\pi_2 \in \Pi_2$ as

$$V_2^{\pi_2}(p,0) = \sum_{i=0}^{\pi_2(p,0)} E(g_i(p)) + \gamma \int_{-\infty}^{\infty} V_2^{\pi_2}(p',0)f(p'|p)dp'$$

For zero units of commodity the policy π_2 always equals its generating $\tilde{\pi}_2$ (see equation 5) and therefore

$$\begin{aligned} V_2^{\pi_2}(p,0) &= V_2^{\tilde{\pi}_2}(p,0) \\ V_2^{\tilde{\pi}_2}(p,0) &= \sum_{i=0}^{\tilde{\pi}_2(p)} E(g_i(p)) + \gamma \int_{-\infty}^{\infty} V_2^{\tilde{\pi}_2}(p',0)f(p'|p)dp' \quad (7) \end{aligned}$$

This means that if we want to find the optimal π_2^* from Π_2 we can do it by simply finding the optimal $\tilde{\pi}_2^*$ from $\tilde{\Pi}_2$.

Theorem 2 *The optimal two-unit strategy $\pi_2^* \in \Pi_2$ maximizing $V_2^{\pi_2}(p,0)$ is defined by a strategy $\tilde{\pi}_2$ maximizing $\sum_{i=0}^{\tilde{\pi}_2(p)} E(g_i(p))$.*

Proof Using Equation 7, and the fact that the integral part becomes independent of the amount of commodity we actually held in the previous step, we can maximize the value of a policy by maximizing the sum of expected gains. That is:

$$V_2^*(p,0) = \max_{j=0,1,2} \left[\sum_{i=0}^{j} E(g_i(p)) \right] + \gamma \int_{-\infty}^{\infty} V^*(p',0)f(p'|p)dp'.$$

□

The theorem shows how to get the optimal policy π_2^* from Π_2 for the zero commodity start state. This policy can be applied also to the non-zero commodity state in a straightforward way.

Up to this point we know how to find the best (optimal) policy from Π_2 ($\tilde{\Pi}_2$). This policy is no worse than π_1^* that trades only one unit (simply because $\pi_1^* \in \Pi_2$). However, this does not necessarily imply that the optimal two-unit strategy $\pi_2^* \in \Pi_2$ is also globally optimal.

Theorem 3 *π_2^* ($\tilde{\pi}_2^*$) is the optimal two-unit strategy.*

Proof To prove this we need to show that the optimal policy always resides in Π_2. The opposite can happen only when: (1) the optimal strategy is not incremental (the optimal choice for the first unit does not imply the optimal choice for two units case) and (2) the optimal strategy depends on c, such that the relation is not captured by Equation 5 ($\tilde{\pi}_2$ and π_2 relation).

The incremental property follows from the fact that the maximum gain we can capture by any unit is $E(g_1(p))$, which equals the expected gain for the first unit. Thus the policy must always trade the first unit optimally. Using the incremental result, the relation in Equation 5 can be violated only when the globally optimal policy at some point recommends to hold 2 units and $\pi_2^* \in \Pi_2$ does not, or vice versa. However, this cannot happen as it would mean that we choose to hold the second unit when $E(g_2(p))$ is negative or not to hold it when it is positive. Thus $\pi_2^* \in \Pi_2$ must be the optimal policy. □

Threshold price for the second unit

The optimal two-unit strategy says that we want to invest into the second unit only when $E(g_2(p)) \geq 0$. The question now is if we can come up with a compact representation of this condition, similarly to the threshold price for the one-unit case. Indeed, we can show that if $E(g_2(p))$ ever becomes positive, there is a unique threshold value p_2^* such that when $p \leq p_2^*$ then $E(g_2(p)) \geq 0$ is guaranteed. This follows directly from the monotonicity of $E(g_2(p))$, which we prove next.

Theorem 4 *For $p \leq p'$ it holds $E(g_2(p)) \geq E(g_2(p'))$*

Proof In order to prove $E(g_2(p)) \geq E(g_2(p'))$ for $p \leq p'$, it is sufficient to show, using Equation 6, that

$$\begin{aligned} E(g_1(p)) &\geq E(g_1(p')), \\ \int_{p_1^*}^{\infty} E(g_1(p''))f(p''|p)dp'' &\geq \int_{p_1^*}^{\infty} E(g_1(p''))f(p''|p')dp'' \end{aligned}$$

hold. This is trivial and follows from the monotonicity of $E(g_1(p))$ and the fact that $f(p''|p)$ and $f(p''|p')$ define normal distributions with means $E(p''|p) \geq E(p''|p')$ and the same standard deviation. □

The main consequence of $E(g_2(p))$ being monotonically decreasing is that there is a unique zero point p_2^*, such that for any $p \leq p_2^*$, $E(g_2(p)) \geq 0$. Therefore, the optimal generating strategy $\tilde{\pi}_2^*$ for trading two units of commodity (with the sell constraint 1) can be defined compactly using a set of threshold prices $\{p_1^*, p_2^*\}$ such that

$$\tilde{\pi}_2^*(p) = \begin{cases} 0 & p_1^* \leq p \\ 1 & p_2^* \leq p \leq p_1^* \\ 2 & p \leq p_2^*. \end{cases}$$

The optimal strategy $\pi_2^*(p,c)$ is then compactly represented as:

$$\pi_2^*(p,c) = \max[\tilde{\pi}_2^*(p), c-1].$$

5.4 Trading k-units under sell limit

In principle the same ideas as used for the two-unit case can be applied to find the optimal strategy π_k^* for trading at most k units of commodity. Such a policy is guaranteed to be no worse than the policy for trading smaller number of units. We summarize the main results and conclusions, the detailed analysis is deferred to the full paper.

The expected added gain from holding k-th unit of commodity – $E(g_k(p))$ is defined recursively as:

$$E(g_k(p)) = E(g_1(p)) + \gamma \int_{p^*_{k-1}}^{\infty} E(g_{k-1}(p'))f(p'|p)dp'. \quad (8)$$

A nice property of $E(g_k(p))$ is that it is monotonically decreasing in p and $E(g_k(p)) \leq E(g_{k-1}(p))$.

The optimal generating policy $\tilde{\pi}_k^* : R \rightarrow \{0,1,2,\cdots,k\}$ equals:

$$\tilde{\pi}_k^*(p) = \arg\max_{j=0,1,2,\cdots,k} \sum_{i=0}^{j} E(g_i(p)).$$

$\tilde{\pi}_k^*$ can be described compactly using a set of threshold values $\{p_1^*, p_2^*, \cdots, p_k^*\}$ such that:

$$\pi_k^*(p) = \begin{cases} 0 & p_1^* \leq p \\ i & p_{i+1}^* \leq p \leq p_i^* \\ k & p \leq p_k^*, \end{cases}$$

where thresholds are unique zero points $E(g_i(p_i^*)) = 0$.

The optimal policy $\pi_k^* : R \times \{0,1,2,\cdots,k\} \rightarrow \{0,1,2,\cdots,k\}$ for the sell limit 1 can be then compactly represented as:

$$\pi_k^*(p,c) = \max[\tilde{\pi}_k^*(p); c-1].$$

5.5 Solution for buy, sell and store limits

The optimal policy for at most k units of commodity and the sell constraint $C_{\text{sell}} \geq 1$, can be derived directly from the solution for the sell limit 1:

$$\pi_k^*(p,c) = \max[C_{\text{sell}}\tilde{\pi}_k^*(p), c - C_{\text{sell}}].$$

Adding buy and store constraints to this result is easy and results in the following policy:

$$\pi_k^*(p,c) = \min[C_{\text{store}}; c+C_{\text{buy}}; \max[C_{\text{sell}}\tilde{\pi}_k^*(p), c-C_{\text{sell}}]].$$

6 Finding optimal thresholds

The optimal policy for k units can be represented compactly using a set of threshold prices $\{p_1^*, p_2^*, \cdots, p_k^*\}$. The threshold price for k-th unit is the zero of the the expected added gain for the k-th unit $E(g_k(p_i^*)) = 0$. As the value of $E(g_k(p))$ depends only on threshold prices $\{p_1^*, p_2^*, \cdots, p_{k-1}^*\}$, the set of threshold prices can be built incrementally.

The main problem in this process is that there is no closed form solution for finding the zero point of $E(g_k(p))$

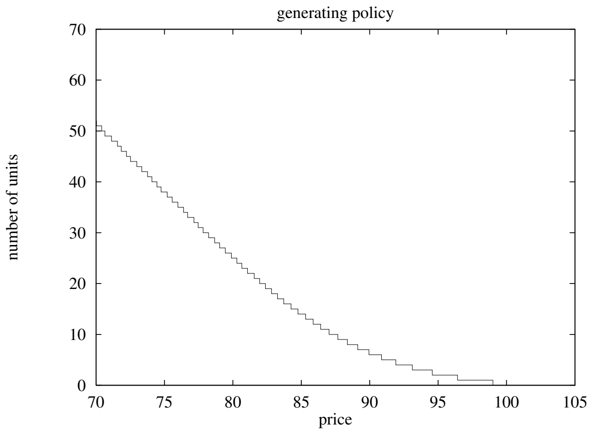

Figure 1: The optimal generating policy with sell limit 1, $\mu = 100, \eta = 0.6, \sigma = 10, \gamma = 0.9975, c = 0.2$. Stepwise changes reflect positions of thresholds.

($E(g_1(p))$ is the only exception). Thus, in order to find the threshold values for k-units we resort to stochastic (Monte Carlo) approximation techniques (see [Kushner and Yin, 1997]). In particular we use the Robbins-Monro scheme, which finds the zero of the function $E(g_k(p))$ iteratively using the update

$$p_{n+1} = p_n + \epsilon_n Y_n$$

where, p_n is the price value at the n-th iteration, Y_n is the "noisy" estimate of the function $E(g_k(p))$ at p_n obtained by Monte Carlo sampling from Equation 8. The sequence ϵ_n is used to "average out" the "noisy" estimates from the Monte-Carlo sampling. We use the standard sequence $\epsilon_n = 1/(n+1)$, which converges (under reasonable assumptions) to the zero value with probability one. Since $E(g_k(p))$ is monotonous there is a unique root p_k^* such that $E(g_k(p_k^*)) = 0$. [Kushner and Yin, 1997] also give more details on the rate of convergence.

The process for finding new thresholds can be applied incrementally to find the solution for an arbitrary number of units. The issue that remains open is that there can be an infinite number of thresholds to define the complete solution. However, thresholds for larger values of k are less likely to be used, as they cover the range of prices with extremely small chance of occurrence. Thus if the initial price is in a reasonable range more complex policies tend to contribute less.

To provide for more robustness, we propose an on-line algorithm that keeps building thresholds on the demand basis. That is, only when price encountered in not covered by a current set of thresholds, we start to work on thresholds for higher values of k. Note that this algorithm can be further refined into an anytime scheme [Dean, 1991], suitable for time critical settings.

7 Experimental results

We have tested our approach on several sets of parameters. Figure 1 illustrates a typical policy. To find the zero points of expected added gain functions $E(g_k(p))$, we used the basic

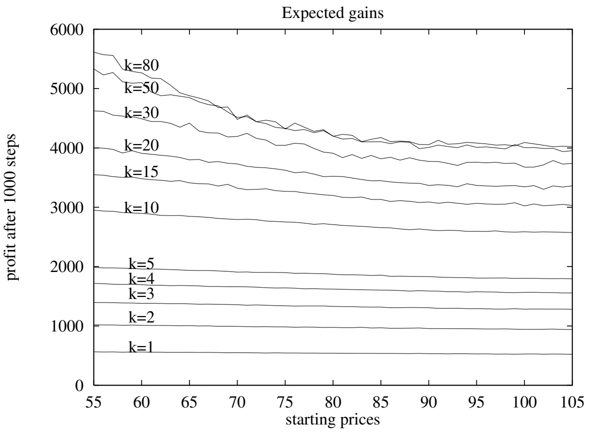

Figure 2: Average NPV for strategies with varying number of thresholds and different initial prices.

Robbins-Monro algorithm for $10,000$ steps. We use the average of 100 runs. The thresholds were computed incrementally starting from p_1^*. The interesting observation is that for larger values of k, thresholds become about equally spaced. In such a situation the policy can be more compactly represented using a linear function, directly mapping prices to recommended commodity holdings.

The effect of strategy refinement (adding more thresholds and potentially investing more units) on the quality of the solution is illustrated in Figure 2. The results show average profit over 1000 runs for different policies and different starting prices. We see that more complex policies with more thresholds tend to improve the profit performance initially. This effect becomes less significant for larger k's, and expected profits converge to the optimal value $V^*(p)$. Since the prices reach lower values very rarely, the probability of using thresholds for larger k's is very small and corresponding more complex policies are only rarely utilized.

8 Conclusion

We have presented an efficient method for finding the near optimal policies for the commodity planning problem with trading and storage constraints. The problem has continuous state space and large action space and cannot be solved efficiently by standard dynamic programming solutions. Our solution relies on the analysis of the domain and takes advantage of a problem structure and Monte Carlo approximation techniques. We showed that the optimal policy for the commodity trading problem with a mean-reverting price model and trading constraints can be represented compactly using a stack of decision (price) thresholds for investing in additional units of commodity. The thresholds correspond to zero points of expected added gain functions, which we also derived. As these functions do not have analytical solution, we apply Monte Carlo approximation techniques to find their zero points. The properties of the functions guarantee uniqueness of the solution and convergence of the approximation scheme.

Interesting questions and opened issues that remain to be addressed include the theoretical bound on the number of thresholds necessary to guarantee the near optimal strategy, and exploration of more compact parametric representations of the policy, mapping prices to the number of units to hold directly (these would eliminate the need to remember all threshold values). Finally, a number of interesting problems will arise if we relax some of the current assumptions of the model. Possible refinements may include price spreads, concurrent trading at multiple interconnected sites, or demand/supply sensitive price models.

Acknowledgement

We wish to thank Oliver Frankel of Goldman Sachs for introducing us to this problem and for valuable technical discussions.

References

[Bellman, 1957] Richard E. Bellman. *Dynamic Programming*. Princeton University Press, Princeton, 1957.

[Boutilier *et al.*, 1995] Craig Boutilier, Richard Dearden, and Moisés Goldszmidt. Exploiting structure in policy construction. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pages 1104–1111, Montreal, 1995.

[Brealey and Myers, 1991] Richard A. Brealey and Stewart C. Myers. *Principles of Corporate Finance*. McGraw-Hill, 1991.

[Dean and Lin, 1995] Thomas Dean and Shieu-Hong Lin. Decomposition techniques for planning in stochastic domains. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pages 1121–1127, Montreal, 1995.

[Dean, 1991] Thomas Dean. Decision-theoretic control of inference for time-critical applications. *International journal of Intelligent Systems*, 6:417–441, 1991.

[Dearden and Boutilier, 1997] Richard Dearden and Craig Boutilier. Abstraction and approximate decision theoretic planning. *Artificial Intelligence*, 89:219–283, 1997.

[Dixit and Pindyck, 1994] Avinash K. Dixit and Robert S. Pindyck. *Investment under Uncertainty*. Princeton University Press, Princeton, 1994.

[Kushner and Yin, 1997] Kushner and Yin. *Stochastic Approximation Algorithms and Applications.* Springer-Verlag, New York, 1997.

[Meuleau *et al.*, 1998] Nicolas Meuleau, Milos Hauskrecht, Kee-Eung Kim, Leonid Peshkin, Leslie Pack Kaelbling, Thomas Dean, and Craig Boutilier. Solving very large weakly coupled markov decision processes. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, pages 165–172, Madison, 1998.

[Puterman, 1994] Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley, New York, 1994.

[Trigeorgis, 1996] Lenos Trigeorgis. *Real Options*. MIT Press, Cambridge, 1996.