Swing Option Pricing
Consistent with Futures Smiles

Roberto Daluiso[*] Emanuele Nastasi[†] Andrea Pallavicini[‡]
Giulio Sartorelli[§]

March 7, 2022

Abstract

In commodity and energy markets swing options allow the buyer to hedge against futures price fluctuations and to select its preferred delivery strategy within daily or periodic constraints, possibly fixed by observing quoted futures contracts. In this paper we focus on the natural gas market and we present a dynamical model for commodity futures prices able to calibrate liquid market quotes and to imply the volatility smile for futures contracts with different delivery periods. We implement the numerical problem by means of a least-square Monte Carlo simulation and we investigate alternative approaches based on reinforcement learning algorithms.

JEL classification codes: C63, G13.
AMS classification codes: 65C05, 91G20, 91G60.
Keywords: Pricing, swing option, volatility smile, least-square Monte Carlo, reinforcement learning.

1 Introduction

In energy markets, a class of commonly traded contracts allows the buyer to select its preferred delivery strategy within daily or periodic constraints,

[*] Intesa Sanpaolo Milan, roberto.daluiso@intesasanpaolo.com
[†] Independent consultant, ...
[‡] Intesa Sanpaolo Milan, a.pallavicini@intesasanpaolo.com
[§] Intesa Sanpaolo Milan, giulio.sartorelli@intesasanpaolo.com

while the purchase price can be fixed at inception or determined before the starting date of the delivery period by observing the prices of quoted futures contracts. These contracts are usually known as swing options since the buyer is allowed to swing between a lower and an upper boundary in the commodity flow.

From the modelling point of view, the daily selection of the delivery strategy along with constraints on the total consumption force us to describe the swing option pricing problem as a specific type of a stochastic control problem for the optimal consumption strategy. Its theoretical aspects are described in Barrera-Esteve et al.[1] and later in Bardou et al.[2]. These papers derive conditions for existence of an optimal consumption strategy of a specific type, named bang-bang, in which only the minimum or maximum consumption allowed by all the constraints is selected on each delivery day. Several other properties of the solution have been proven[3,4] when the set of admissible strategies is subject to such restriction, since the problem then reduces to multiple optimal stopping.

The first numerical works in the literature date back to the nineties and they focus on specific payoffs[5]. Several algorithms have been proposed since then. Tree methods[6] are able to handle fairly general swing option payoffs,[7] different dynamics[8,9,10] and generalizations[11,12] which drop the no-arbitrage pricing setting in favour of a more flexible game-theoretic goal. All of these rely on a Markov chain approximation of the commodity dynamics. A parsimonious but accurate representation by optimal quantization is also discussed in the literature.[13]

Undiscretized continuous state spaces can be handled by least-square Monte Carlo (LSMC) algorithms, first introduced in this context by Barrera-Esteve et al.[1]. LSMC is also considered by other authors[3,14] under the simplified optimal stopping formulation; in this setting, the downward-biased result can be complemented with upper bounds via duality.[15,16,17] Otherwise, if the dynamics of the underlying are driven by Brownian or Lévy factors, continuous time approximations of the payoff lead to Hamilton-Jacobi-Bellman partial (integro) differential equations which can be solved numerically.[18,19] Finally, if the log-underlying itself follows a Lévy or Gaussian mean reverting process, then efficient algorithms based on finite differences[20,21] or Fourier expansions[22,23] can be designed.

Our contribution to the literature is twofold. First, we propose a simple diffusive model for commodity futures prices, which is able to describe the volatility smile quoted by the market for futures contracts with different delivery periods. Our proposal starts from the extension of the local-volatility linear model presented in Nastasi et al.[24]. We stress the importance of modelling futures prices with heterogeneous delivery periods since swing option prices depend both on day-ahead prices through the consumption strategy and on longer period futures contracts (usually one-month contracts) to determine the purchase strike prices. Using this new model, we present a study of the effect on prices of this often overlooked feature of many real swing contracts, by a least-square Monte Carlo implementation which is described in detail.

Second, we investigate reinforcement learning (RL) algorithms as possible alternative tool to solve the stochastic control problem. Reinforcement learning is a modern branch of machine learning, in which an artificial agent automatically learns how to improve the expected cumulated reward of his actions while interacting with a stochastic environment.[25] The number of its applications to finance is rapidly growing.[26,27,28,29] The well studied swing option problem offers an ideal field for benchmarking RL results, both quantitatively against more established algorithms, and qualitatively against the known abstract properties of the optimal solution, in particular the conditions under which it is of bang-bang type. We believe that the positive outcome of such an analysis can be relevant for several reasons.

To begin with, a RL engine does not require any prior knowledge on either the underlying dynamics or the payoff structure; in contrast, most of the existing algorithms cleverly exploit the precise form of the consumption constraints and optimization goal, or properties of the assumed stochastic processes, and need a redesign if any of these is changed. Moreover, RL relies internally on supervised machine learning tools, like neural networks or decision trees, which have been developed specifically to be robust to high-dimensional settings: this means that relevant features for the exercise decision are automatically created or selected even in large problems, where most traditional optimization approaches either break or need hand-crafted dimensionality reductions and simplifications. Finally, the success of this application proves the potential in finance of the recent proximal-policy optimization (PPO) algorithm proposed in Schulman et al.[30], which we adopt in our implementation: differently to Q-learning variants, which have been predominant in applied finance to date, PPO handles continuous actions spaces natively, without the need of artificial discretizations. This implies that also the exercise decision could be easily made multi-dimensional without exponential demands in computational or memory resources.

The paper is organized as follows. In Section 2 we present the model we use to describe the prices of futures contracts on different delivery periods. Then, in Sections 3 we present numerical examples derived by means of least-square Monte Carlo techniques. In particular we check the possibility of optimal bang-bang strategies to test the soundness of the numerical algorithm. We conclude the paper with Section 4 devoted to the application of reinforcement learning to swing option pricing.

Needless to say, the views expressed in this paper are those of the authors and do not necessarily represent the views of their institutions.

2 Modelling Commodity Smiles

The local-volatility linear model presented in Nastasi et al.[24] allows to describe futures prices in a parsimonious way while preserving a perfect fit to plain vanilla options (PVO) quoted in the commodity market. Moreover, mid-curve options and calendar spread options can be calibrated by means of a best-fit procedure. In the original paper some extensions are discussed to introduce multiple risk factors to drive the curve dynamics and to allow for stochastic volatilities. Here, we stick to the one-dimensional specification of the model and we investigate how to extend it to deal with futures contracts on different delivery periods.

2.1 Modelling Futures with Different Delivery Periods

Futures on commodities like natural gas, oil and electricity have as underlying a daily flow for the whole delivery period. Often day-ahead futures are quoted on the market as a close proxy of the spot prices. Moreover, futures on different delivery periods are usually quoted ranging from one day to a whole year. On the other hand, PVO contracts are usually quoted only for the most liquid delivery period.

A common way to model these futures prices consists in introducing a dynamics for futures with the shortest delivery period, and in building longer periods by summing futures prices. Yet, it is difficult to find models which allows to derive closed-form formulae for futures prices with longer periods.[31] Here, we rely on the linear form of the drift coefficient and on using a single risk factor to derive simple closed-form formulae for sums of futures prices.

We start by introducing the instantaneous futures price process $f_t(T)$ with delivery at time T, and we assume that for each T we can model it by the local-volatility linear model,[24] so that we can write

$$f_t(T) = f_0(T) \left(1 - (1 - s_t)e^{-\int_t^T a(u)\,du}\right) \tag{1}$$

where the spot process is given by

$$ds_t = a(t)(1 - s_t)\,dt + \eta(t, s_t)s_t\,dW_t\,, \quad s_0 = 1 \tag{2}$$

where W_t is a standard Brownian motion under the risk neutral measure, a is a non-negative function of time, and η is Lipschitz in the second argument, bounded and greater than a strictly positive constant.

Then, we calculate the futures price for a contract with delivery period $[T + \delta_0, T + \delta_1]$ simply by averaging the instantaneous futures price over the delivery period. We obtain

$$F_t(T, \delta) = \int_0^\infty w(u - T, \delta) f_t(u)\,du\,, \quad w(\tau, \delta) := \frac{1_{\{\delta_0 \leqslant \tau \leqslant \delta_1\}}}{\delta_1 - \delta_0} \tag{3}$$

where we define $\delta := \delta_1 - \delta_0$, and we discard the dependency on δ_0 to lighten the notation. For instance, the futures contracts with a delivery period of one month we mentioned in the introduction are now denoted as $F_t(T, 1\mathrm{m})$.

By straightforward manipulation of the integral we may express $F_t(T, \delta)$ for different delivery period δ as follows:

$$F_t(T, \delta) = F_0(T, \delta) \left(1 - (1 - s_t(\delta))e^{-\int_t^T A(u,\delta)\,du}\right) \tag{4}$$

where we define the fictitious spot process

$$s_t(\delta) := 1 - (1 - s_t)G(t, \delta) \tag{5}$$

which depends on the deterministic functions

$$A(t, \delta) := a(t) - \partial_t \log G(t, \delta) \tag{6}$$

$$G(t, \delta) := \frac{1}{F_0(t, \delta)} \int_0^\infty w(u - t, \delta) f_0(u) e^{-\int_t^u a(v)\,dv}\,du \tag{7}$$

We notice that the relationship between $F_t(T, \delta)$ and $s_t(\delta)$ is the same holding between $f_t(T)$ and s_t up to a change in parameters. Also, we have $f_t(T) = F_t(T, 0)$ and $s_t = s_t(0)$.

Moreover, we can see that the dynamics followed by $s_t(\delta)$ is still a local-volatility linear model. Indeed, we can differentiate the above relationships and we obtain

$$ds_t(\delta) = A(t,\delta)(1 - s_t(\delta))\,dt + \eta(t,\delta,s_t(\delta))s_t(\delta)\,dW_t\,, \quad s_0(\delta) = 1 \tag{8}$$

where we define the local volatility function

$$\eta(t,\delta,k) := \left(1 - \frac{1 - G(t,\delta)}{k}\right)\eta\left(t, 1 - \frac{1-k}{G(t,\delta)}\right) \tag{9}$$

The following bounds hold:

$$s_t(\delta) > 1 - G(t,\delta)\,, \quad 0 < G(t,\delta) \leqslant 1 \tag{10}$$

so that the spot price is always positive.

2.2 Calibration of Futures Option Smile

Usually we can find liquid quotes for PVO only on a single delivery period, we name it $\bar{\delta}$. This is the case for natural gas, the commodity we use as an example in this paper. We focus on this scenario.

In our modelling framework futures prices of each delivery period follow a local-volatility linear model. In particular, this is true for the delivery period $\bar{\delta}$. Thus, we can follow the procedure[24] to calibrate the functions $A(t,\bar{\delta})$ and $\eta(t,\bar{\delta},k)$. We refer to such paper for a discussion of calibration performance and accuracy. Then, we can derive the same functions for any other choice of the delivery period δ by a direct calculation starting from the definitions of these functions. Indeed, we get

$$A(t,\delta) = A(t,\bar{\delta}) + \partial_t \log \frac{G(t,\delta)}{G(t,\bar{\delta})} \tag{11}$$

and

$$\eta(t,\delta,k) = \left(1 - \frac{1}{k}\left(1 - \frac{G(t,\delta)}{G(t,\bar{\delta})}\right)\right)\eta\left(t,\bar{\delta}, 1 - (1-k)\frac{G(t,\bar{\delta})}{G(t,\delta)}\right) \tag{12}$$

which we evaluate for $k > 1 - G(t,\delta)$.

The previous formula allows us to imply volatility smiles for any delivery period. Notice that smiles for different delivery periods can be different only if $a(t) > 0$, since if $a(t) = 0$ we get $G(t,\delta) = 1$.

We show in Figure 1 the volatility smiles implied by the model for different choices of the mean-reversion speed in the case of the TTF natural gas market. The market quotes the volatilities of PVO contracts on futures with one-month delivery period. Since we do not have liquid quotes for mid-curve options or calendar spread options we do not fix the mean-reversion speed to market data. The implied smiles maintain the same shape of the one-month smile because the market smile is almost symmetric in shape. In Figure 2 we plot the volatility backbones, namely the at-the-money implied volatilities as a function of option maturity.

3 Pricing Swing Options

We set up in this numerical section the stochastic control problem required to get swing option prices, and we solve it by means of a least-square Monte Carlo (LSMC) simulation. As a specific example we consider swing options traded in the TTF natural gas market.

The LSMC algorithm is particularly effective when we have to deal only with few risk factors, since the method requires to calculate a linear regression whose size rapidly explodes as the number of risky factors increases. In our case we adopt a parsimonious model with only one risk factor. However, for a better description of curve and smile dynamics we could look at model extensions inclusive of additional risk factors.[24] For this reason we will also investigate in Section 4 solutions which could be applied in higher dimensionality settings.

3.1 Contract Description

A swing option contract guarantees a flexible daily supply of gas with a delivery period of one month. The underlying contracts are the day-ahead futures, namely $F_{T_i}(T_{i+1}, 1\text{d})$ for each fixing date $T_1, \dots, T_{n_f}$ within the delivery period. At each fixing date the owner of the option is allowed to buy a quantity N_{T_i} of gas within a daily range $[N_m, N_M]$ at a strike price K. The total consumption of gas must be within a total range $[C_m, C_M]$. The option price can be written as

$$W_0 := \max_{N \in \mathcal{N}} \sum_{i=1}^{n_f} \mathbb{E}_0[\, N_{T_i}(F_{T_i}(T_{i+1}, 1\text{d}) - K) \,] \, P_0(T_{p,i}; e) \qquad (13)$$

where $P_0(T_{p,i}; e)$ is the price of a zero-coupon bond with yield e_t, and the consumption plan $N := \{N_{T_1}, \dots, N_{T_{n_f}}\}$ can be chosen from a set $\mathcal{N}$ of plans subject to the following constraints:

$$N_m \leqslant N_{T_i} \leqslant N_M \ , \quad C_m \leqslant \sum_{i=1}^{n_f} N_{T_i} \leqslant C_M \tag{14}$$

The strike price of swing options can be either known at inception, or fixed at a forward date. In the latter case it is calculated as the daily average of a specific one-month futures contract over the observation dates $t_1, \dots, t_{n_s}$. For example, the strike price of a swing option with delivery in July 2018 is fixed by averaging the daily observations of the `JUL18` one-month contract, namely we set

$$K_{t_{n_s}} := \frac{1}{n_s} \sum_{j=1}^{n_s} F_{t_j}(\texttt{JUL18}, 1\text{m}) \tag{15}$$

where t_1 is the 1st of June 2018 and t_{n_s} is the 28th of the same month (last trading date).

We show in Figure 3 the `JUL18` swing option on NG TTF day-ahead futures term-sheet data. In this example the strike price is set by observing one-month futures contracts.

3.2 Least-Square Monte Carlo Simulation

We can write the stochastic control problem underlying the pricing of a swing option contract by introducing the consumption strategy N_{T_i} which represents the quantity of gas delivered in T_i, and by defining the total consumption up to time T_i as given by

$$C_{T_i} := \sum_{j=1}^{i} N_{T_j} \tag{16}$$

Thus, on each day T_i in the simulation we have to solve the following control problem

$$W_{T_i} = \max_{N_{T_i}} \left\{ N_{T_i}(F_{T_i}(T_{i+1}, 1\text{d}) - K) + \mathbb{E}\left[\ W_{T_{i+1}}(N_{T_i}) \frac{P_0(T_{p,i+1}; e)}{P_0(T_{p,i}; e)} \ \middle| \ F_{T_i}, C_{T_{i-1}} \right] \right\} \tag{17}$$

In case the option is forward starting we should add to the conditioning factors also the strike price.

We can solve the control problem by means of a LSMC simulation. Here, we describe the details of our implementation, which can be split into three steps: (i) we build consumption grids, (ii) we estimate of the value functions on the grids by means of regressions with a backward procedure, (iii) we compute the swing option price with a standard forward Monte Carlo simulation.

For simplicity, the method is described in the following by considering zero interest rates and fixed strike prices.

3.2.1 Building Consumption Grids

We start by constructing the consumption grid by taking care that the points corresponding to extreme choices of the amount to be consumed are included. We define the global constraint functions at each fixing date T_i as given by

$$U_i := \min\left(C_M - N_m(n_f - i)\,,\ iN_M\right) \tag{18}$$

and

$$D_i := \max\left(C_m - N_M(n_f - i)\,,\ iN_m\right) \tag{19}$$

At each date T_i the total consumption must be within such values: $D_i \leqslant C_{T_i} \leqslant U_i$. We define C^i as the vector representing the consumption grid at fixing date T_i. The consumption grid is built by following Algorithm 1.

On each date T_i we start by adding the points allowed by a bang-bang strategy. We use the term bang-bang as in Jaillet et al.[7] to indicate a strategy which on each date T_i is consuming the minimum or the maximum amount of commodity according to all the constraints. Thus, defining the starting consumption C^0 as a vector with a single component equal to zero, at time T_1 we have only two possible bang-bang states given by N_m and N_M (unless global constraints are tighter). On the following date T_2 the grid has three bang-bang points, obtained by starting from the consumption levels of the previous date and consuming the minimum or maximum allowed quantities, and so on. Then, we refine the grid in between the bang-bang points to allow for intermediate choices (continuous consumption strategy). The resulting grids are depicted in Figure 4.

Algorithm 1 Algorithm to build the consumption grid. Operations from 9 to 11 are performed only when the strategy is continuous.

```
 1: procedure GRID($\{T_i\}_{i=1}^{n_f}, N_m, N_M, C_m, C_M, \Delta$)
 2:     $C^0 := [0]$                                         ▻ Starting consumption
 3:     for $i = 1$ to $n_f$ do
 4:         for $x$ in $C^{i-1}$ do                              ▻ Bang Bang points
 5:             append $\min(U_i\ ,\ x + N_M)$ to $C^i$
 6:             append $\max(D_i\ ,\ x + N_m)$ to $C^i$
 7:         end for
 8:         $C^i$ = unique($C^i$)                 ▻ Sort the grid and erase duplicates
 9:         for $j$ in length($C^i$) - 1 do
10:             append unif($C^i_j, C^i_{j+1}, \Delta$) to $C^i$              ▻ Thicken the grid
11:         end for
12:         $C^i$ = unique($C^i$)                 ▻ Sort the grid and erase duplicates
13:     end for
14: end procedure
```

3.2.2 Calculating the Regression Coefficients

Now, we proceed by describing the simulation algorithm. We start by producing a Monte Carlo simulation on dates T_i for the day-ahead futures prices according to Equations (4) and (8), and we build the consumption grids C^i according to the previous algorithm. We call $F^{(k)}_{T_i}$ the $i-$th fixing of the $k-$th simulation. For each point C^{i-1}_ℓ of the grid at previous time, we introduce the set $N(C^{i-1}_\ell)$ of all the possible consumption levels, whose j-th element can be defined as

$$N_j(C^{i-1}_\ell) := C^i_j - C^{i-1}_\ell \tag{20}$$

and the the set $\mathcal{Q}^{T_i}(C^{i-1}_\ell)$ of admissible consumption levels given global and local constraints relative to the $\ell-$th point of the grid

$$\mathcal{Q}^{T_i}(C^{i-1}_\ell) := \left\{ N \in N(C^{i-1}_\ell) \bigcap [N_m, N_M] \mid C^{i-1}_\ell + N \in [L_i, U_i] \right\} \tag{21}$$

Then, we can write the control problem on the grid as given by

$$W_{T_i}\left(F^{(k)}_{T_i}, C^{i-1}_\ell\right) = \max_{N \in \mathcal{Q}^{T_i}(C^{i-1}_\ell)} \left\{ N\left(F^{(k)}_{T_i} - K\right) + \mathbb{E}\Big[\, W_{T_{i+1}}\left(F_{T_i}, C^{i-1}_\ell + N\right) \,\Big|\, F_{T_i} = F^{(k)}_{T_i} \,\Big] \right\} \tag{22}$$

with terminal condition

$$W_{T_{n_f}}\left(F^{(k)}_{T_{n_f}}, C^{n_f-1}_j\right) = \begin{cases} \min\left(C_M - C^{n_f-1}_j \ , \ N_M\right)\left(F^{(k)}_{T_{n_f}} - K\right) & F^{(k)}_{T_{n_f}} > K \\ \max\left(C^{n_f-1}_j - C_m \ , \ N_m\right)\left(F^{(k)}_{T_{n_f}} - K\right) & F^{(k)}_{T_{n_f}} \leqslant K \end{cases} \tag{23}$$

We can solve the problem backwards in time by starting from the terminal condition in T_{n_f}, and proceeding to the previous steps by numerically evaluating the forward expectation in the right-hand side of Equation (22) by means of the Monte Carlo simulation. We call $f_{T_i}\left(F; C^{i-1}_\ell + N\right)$ the estimate of such forward expectation

$$\mathbb{E}\Big[\, W_{T_{i+1}}\left(F_{T_{i+1}}, C^{i-1}_\ell + N\right) \,\Big|\, F_{T_i} = F^{(k)}_{T_i} \,\Big] \approx f_{T_i}\left(F; C^{i-1}_\ell + N\right) \tag{24}$$

and we suppose that it is quadratic with respect to day-ahead futures prices:

$$f_{T_i}\left(F; C\right) := \alpha^i\left(C\right) + \beta^i\left(C\right)F + \gamma^i\left(C\right)F^2 \tag{25}$$

We note that $C^{i-1}_\ell + N \in C^i$ by construction, hence we can estimate the coefficients by regressing for each j the realizations

$$y^{(k)} := W_{T_{i+1}}\left(F^{(k)}_{T_{i+1}}, C^i_j\right) \tag{26}$$

against

$$x^{(k)} := F^{(k)}_{T_i} \tag{27}$$

The problem given by Equation (22) at each time $T_i < T_{n_f}$ is then solved by replacing the forward expectation with the estimate just performed (25). The procedure is repeated backwards in time until the first fixing.

3.2.3 Pricing with a Forward Simulation

Once the backward procedure is completed we have calculated the coefficients α^i, β^i and γ^i on each grid date T_i, which allows us to approximate the forward expectation given by Equation (25) on any scenario. Then, in order to avoid biases, we proceed by sampling a second Monte Carlo simulation for day-ahead futures prices. On each scenario k of the second simulation and on each date T_i we calculate $F^{(k)}_{T_i}$. Starting from the first fixing, at each step, being at a certain point $C^{(k)}_{i-1}$ on the grid C^{i-1}, we choose the quantity to consume $\hat{N}^{(k)}_i$ solving the problem optimization (22) with the coefficients calculated in the previous simulation. This step takes us to the point $C_i^{(k)} = C_{i-1}^{(k)} + \hat{N}_i^{(k)}$. Repeating the step described until reaching the last fixing we get the reward

$$R_{T_{n_f}}^{(k)} := \sum_{i=1}^{n_f} \hat{N}_i^{(k)} \left(F_{T_i}^{(k)} - K \right) \tag{28}$$

Hence, the swing option price is given by averaging the rewards.

$$W_{T_0}\left(F_{T_0}; 0\right) = \mathbb{E}_{T_0}\Big[\, R_{T_{n_f}} \, \Big] \tag{29}$$

3.3 Numerical Investigations with LSMC

We are now ready to calculate the price of swing options with the local-volatility linear model by using the LSMC algorithm. It is our aim to highlight the impact of the mean-reversion speed in swing option prices. Moreover, we wish to show that Theorem 2.4 in Bardou et al.[13] holds, and the LSMC algorithm is able to select the optimal strategies in agreement with the theorem.

We consider for our numerical analysis futures contracts on the TTF natural gas, quotations are expressed in €/MWh. We calibrate our model to PVO quoted on 29 March 2018 on ICE market. We consider swing option contracts with delivery ranging from May 2018 up to June 2019. We consider both fixed-strike option and floating-strike options with at-the-money strike. The strike price is calculated by considering the one-month futures contract delivering in the same period of delivery of the swing contract. Fixed-strike options evaluate the futures price at contract inception, while floating-strike options make a daily average of the futures prices on a time window starting after contract inception and ending before delivering, as previously depicted in Figure 3. All contracts, if not specified otherwise, have global constraints given by:

$$C_m = 12.5\,\mathrm{MWh}\ , \quad C_M = 20\,\mathrm{MWh} \tag{30}$$

and, following a usual choice in the literature,[13] without loss of generality the daily constraints are set to:

$$N_m = 0\,\mathrm{MWh}\ , \quad N_M = 1\,\mathrm{MWh} \tag{31}$$

We remark that different choices of daily constraints can be obtained by simply shifting and scaling all the relevant quantities.

3.3.1 Fixed vs. Floating-Strike Options

We start by analyzing the impact of the mean-reversion speed on swing option prices. In Figure 5 we show the swing option prices for different delivery periods, each period corresponding to the delivery of a futures contract quoted in the market. The trend of the price of the fixed strike options with respect to the delivery month can be easily understood. As time increases, the option increases its time value, which translates into an increase in price. The increasing trend as a function of the mean-reversion speed is instead explained by the two graphs in Figure 1. This picture shows that the volatility of the day-ahead contract, i.e. one-day delivery period, implied by the model, increases as the mean-reversion speed increases, which translates into an increase in the price of the swing option. On the contrary, if we look at the prices of the forward start options for mean-reversion speed equal to 0, we note that the price trend reproduces the shape of the at the money market volatility at Figure 2. This is due to the fact that the volatility of the monthly Futures observed during the strike period is equal to the volatility of the day-ahead contract. By increasing the mean reversion, instead, we have the two opposite effects: the volatility of the strike decreases, while the volatility of the day-ahead contract, as already said, increases. As a result the forward volatility relative to the fixing period increases, producing increasing prices as the mean-reversion speed increases.

In the following sections we will focus on specific numerical problems, so that we will consider only the case of a fixed-strike option delivering in May 2018. Moreover, we set the mean-reversion speed to 1.

3.3.2 Bang-Bang Strategies

We continue by investigating the strategies selected by the LSMC algorithm. We recall that Theorem 2.4 in Bardou et al.[13] describes the structure of optimal strategies for swing options when the consumption levels have a specific form. In particular, it states that the if the minimum global constraint and the difference between the global constraints can be expressed as an integer multiple of the difference between the daily constraints, then the optimal strategy is consuming the daily minimum or maximum on all dates (we say that the strategy is of bang-bang type). For instance, the swing option contract used for the example of Figure 5 does not satisfy the theorem, while the same contract with integer values for the global constraints is within the theorem since the difference between the local constraints is 1.

We start from one of the cases studied in the previous section (fixed-strike delivering in May 2018 with $a = 1$). Such case does not satisfy the hypotheses of the Theorem since the minimum global constraint is not an integer number. The price of the corresponding swing option contract can be read in the top-left entry of Table 1. Then, we consider three different scenarios.

1. We limit the allowed strategies to be only of bang-bang type. We expect to see a reduced price in this case since we forbid intermediate consumption choices. Indeed, in the top-right entry of the table we obtain a lower price.

2. Without limitations on the strategies we change the minimum global constraint to 12 MWh, so that now we satisfy the hypotheses of the Theorem. In the bottom-left entry of the table we report the price of this scenario. The price is now bigger since the global constraints are wider.

3. With the constraint of the previous scenario we limit the allowed strategies to be only of bang-bang type. We expect not to see a reduced price in this case since we forbid consumption choices which are not selected for the optimal strategy. Indeed, in the bottom-right entry of the table we can see that the price is unchanged.

We support our discussion by showing in Figure 6 the graph of daily consumption N_{T_i} selected by the optimal strategy on a particular simulation path, when we assume that the minimum global constraint is either 12.5 MWh (out-of-theorem hypotheses, left panel) or 12 MWh (within-theorem hypotheses, right panel). When we are out of the theorem hypotheses we can see that exists an optimal strategy (red line) which is not of bang-bang type, and we are able to exploit these strategies.

4 Optimal Strategies via Reinforcement Learning

As we have seen swing option pricing requires to solve a stochastic control problem with a continuous set of actions. Standard techniques rely on regression-based simulations whose performances may degrade when the dimensionality of the problem increases. In particular, if we wish to extend our analysis to long-dated options, we should introduce more driving factors to deal with the curve dynamics and possibly of the volatility dynamics. Indeed, we extend the local-volatility linear model[24] in this direction.

Here, we wish to price swing option with an alternative algorithm based on reinforcement learning (RL) techniques. In our approach we use the recently developed proximal policy optimization (PPO) algorithm proposed in Schulman et al.[30].

4.1 Proximal Policy Optimization Algorithm

RL[25,32] describes how an agent behaves in an environment so to maximize some notion of cumulative reward. The actions of the agent as a function of his observations of the environment are termed the agent policy. In our case the agent can choose the amount of commodity to be delivered within the contract limits, so that the reward at each date is the cash flow generated by the swing option, and the policy is the consumption strategy. Once the agent is trained, and the optimal policy is selected, we can run a further Monte Carlo simulation to calculate the swing option price.

4.1.1 Agent Interaction with the Environment

We consider as before a discrete time-grid of fixing times $T_1, \ldots, T_{n_f}$. The algorithm we chose for the training of the agent belongs to the family of actor-critic algorithms. In particular, in our setting, this means that the agent uses a parametric function with parameters θ to calculate both the quantity $N^\theta_{T_i}$ to consume at time T_i, and its best estimate $V^\theta_{T_i}$ of the value function under the current policy, identified by θ; this value function is the expected value of future rewards, which will match the option price W_{T_i} for optimal N^θ. The agent makes its decision by observing the environment given by: the fixing time T_i, the day-ahead futures price $F_{T_i} := F_{T_i}(T_{i+1}, 1\mathrm{d})$, and the total quantity of gas $C^\theta_{T_{i-1}}$ consumed up to time T_{i-1}. We represent in Figure 7 the relationships between the agent and the environment.

We adopt as learning strategy the PPO algorithm developed in Schulman et al.[30]. This algorithm is well-suited for continuous control problems.[1] The

[1] We use the implementation of the algorithm found in OpenAI Baselines https://github.com/openai/baselines.

PPO algorithm collects a small batch of experiences interacting with the environment to update its decision-making policy. The value function of a new policy is estimated by sampling from the environment. A brief overview of how this is done is provided here below.

4.1.2 Description of the Learning Strategy

In the PPO algorithm the policies are randomized, so they are defined as probability distributions on the set of possible actions. In our case they represent the probability of a specific gas consumption on each fixing date given the value of the day-ahead futures contract and the total level of gas consumption up to the previous fixing date. In case of floating-strike swing options we have to include also the strike level. If the space of controls is a continuum, the algorithm considers the actions to be random variables $\tilde{N}^{\theta}_{T_i}$ distributed according to independent Gaussian distributions ϕ centered on the value $N^{\theta}_{T_i}$, which is determined by a neural network:

$$\pi^{\theta}_{T_i}(n) := \mathbb{Q}\Big\{\, \tilde{N}^{\theta}_{T_i} = n \;\mid\; s_{T_i}, C^{\theta}_{T_{i-1}} \,\Big\} = \phi(n; N^{\theta}_{T_i}, \xi_i) \tag{32}$$

where the variances ξ_i are added to the set θ of parameters which are subject to optimization. Policies are identified by the PPO algorithm with these densities.

If we wish to limit the allowed strategies only to the bang-bang ones, we can simply restrict the action space to a discrete set. Hence, the actions will be binomial random variables $\tilde{N}^{\theta}_{T_i}$, and the neural network will return the probability parameter of such distribution.

Starting from the swing option control problem, we can define the action-value function if a specific action is taken at time T_i as

$$\tilde{Q}^{\theta}_{T_i}(n) := \mathbb{E}\Big[\, r_{T_i}(n) + \tilde{Q}^{\theta}_{T_{i+1}}(\tilde{N}^{\theta}_{T_{i+1}}) D(T_i, T_{i+1}) \,\Big|\, F_{T_i}, C_{T_{i-1}} \,\Big] \tag{33}$$

where $r_{T_i}(n) := n(F_{T_i} - K)$ is the reward at time T_i, and the probability space is extended to incorporate also the uncertainty in the actions.

If the action at time T_i is integrated over its law under the policy, we can write the value function as

$$\tilde{V}^{\theta}_{T_i} := \mathbb{E}\Big[\, r_{T_i}(\tilde{N}^{\theta}_{T_i}) + \tilde{V}^{\theta}_{T_{i+1}} D(T_i, T_{i+1}) \,\Big|\, F_{T_i}, C_{T_{i-1}} \,\Big] \tag{34}$$

The PPO algorithm modifies θ by gradient ascent steps, so to increase the value of an objective function which is made up of two main components,

L^A and L^V. The first component L^A measures the goodness of the policy, and it is a modification of the so-called advantage

$$A^\theta_{T_i} := A^\theta_{T_i}(\tilde{N}^\theta_{T_i}), \qquad \text{where} \qquad A^\theta_{T_i}(n) = \tilde{Q}^\theta_{T_i}(n) - \tilde{V}^\theta_{T_i}. \tag{35}$$

Intuitively,[33] $A^\theta_{T_i}(n)$ measures the average improvement in expected reward given by choosing the action n, over the average reward of the current policy $\tilde{V}^\theta_{T_i}$. Hence, a reasonable idea is to train the agent to favour large values of $A^\theta_{T_i}$. Indeed, one can prove that the gradient of the expected total cumulated reward, which is the ultimate target of optimization, equals

$$\sum_i \mathbb{E}\Big[\, A^\theta_{T_i}(\nabla_\theta \log \pi^\theta_{T_i}(n))|_{n=\tilde{N}^\theta_{T_i}} \Big] . \tag{36}$$

This is beneficial since the realizations of the advantage are typically much less noisy then the realizations of the cumulated rewards[32,34].

In practice one substitutes the unknown A^θ with a pathwise quantity which gives (approximately) the same gradient, namely

$$\hat{A}^\theta_i := \sum_{l=0}^{n_f-i-1} D(T_i, T_{i+l})\lambda^l \left[r_{T_{i+l}}(\tilde{N}^\theta_{T_{i+l}}) + D(T_{i+l}, T_{i+l+1})V^\theta_{t+l+1} - V^\theta_{t+l} \right] \tag{37}$$

Note that if $\lambda = 1$ then $\hat{A}^\theta_{T_i}$ telescopically reduces to the sum of realized discounted rewards, while if $\lambda < 1$ some bias is introduced by reducing the impact on L^A of rewards which are far in the future. This is done to get lower variance.[34]

Instead, the second component L^V of the objective function measures how well V^θ represents the value function $\tilde{V}^\theta$ of the policy π^θ.

Each step of the optimization procedure acts on a batch of so-called "episodes" which are complete trajectories of state and action each resulting from a Monte Carlo simulation of the state up to the swing option maturity. The agent interacting with the environment calculates on each fixing date t the policy density and the advantage for selecting an action $\tilde{N}^\theta_t$ at such time.

After the sampling process a new policy is proposed using a Stochastic Gradient Descent (SGD) with respect to the θ parameters:

$$\theta_{k+1} = \theta_k + \rho \cdot \mathbb{E}\left[\, \nabla_\theta \sum_{i=1}^{n_f} \left(L^A_{T_i}(\theta) - \beta L^V_{T_i}(\theta) \right)\Big|_{\theta=\theta_k} \right] \tag{38}$$

$$L^A_{T_i}(\theta) := \min\left\{ \frac{\pi^\theta_{T_i}(\tilde{N}^{\theta_k}_{T_i})}{\pi^{\theta_k}_{T_i}(\tilde{N}^{\theta_k}_{T_i})} \hat{A}^{\theta_k}_i, \operatorname{clip}\left(1-\varepsilon, \frac{\pi^\theta_{T_i}(\tilde{N}^{\theta_k}_{T_i})}{\pi^{\theta_k}_{T_i}(\tilde{N}^{\theta_k}_{T_i})}, 1+\varepsilon\right) \hat{A}^{\theta_k}_i \right\} \tag{39}$$

$$L^V_{T_i}(\theta) := \left(V^\theta_{T_i} - (\hat{A}^{\theta_k}_i + V^{\theta_k}_{T_i}) \right)^2 \tag{40}$$

where the expected value is estimated over a batch of episodes, ρ is a learning rate, $\operatorname{clip}(a, x, b)$ is the clip function (capped and floored linear function), while ε and β are hyper-parameters.

One of the key ideas of PPO is to ensure that a new policy update is "close" to the previous policy by clipping the advantages. The ratio behind this choice is to keep the new policy $\pi^{\theta_{k+1}}$ within a neighbourhood of the old one π^{θ_k} where one can trust both the first order approximation to the objective function given by the stochastic gradient, and the function V^{θ_k} used in the estimation of the advantage. Once the policy is updated, the experiences are thrown away and a new batch is collected with the updated policy.

4.2 Numerical Investigations with PPO

We focus in this example on a swing option contract with at-the-money fixed strike. The mean reversion is equal to 1 in all experiments.

Several PPO hyper-parameters will be kept fixed to the following values: $\lambda = 0.95$, $\varepsilon = 0.2$, $\rho = 0.0003$. These are general purpose defaults[30] and/or hard-coded in the baselines implementation. The trainable parameters θ are updated once every 2048 training episodes.

In all experiments, two distinct feed-forward neural networks with hyperbolic tangent activation function are used to compute respectively the action and the value function at all times t, where the network inputs are: T_i expressed as a year fraction, the total consumption to-date remapped linearly at each time so that its domain is always $[-0.5, 0.5]$, and $\log(F_{T_i}/F_{T_0})$. In this way all inputs are well normalized, which helps the training of the network. The output layer is linear, i.e. no activation function is applied.

Both when the hypotheses which guarantee the existence of bang-bang optima are satisfied, and when they are not, we can allow for general $[N_m, N_M]$-valued actions; in the former case, the learning algorithm should find out by itself that the best strategy only involves bang-bang consumptions. The continuous-valued consumption is obtained by clipping the network's output to the interval $[0, 1]$ and then remapping the result linearly so that 0 and

1 correspond respectively to the minimum and maximum admissible consumptions given both daily and global constraints. When we want to force bang-bang strategies instead, then only the minimum and maximum are considered as admissible actions, and a softmax layer remaps to probabilities the output units corresponding to these two actions.

4.2.1 Neural Network Fine Tuning

We explored several possible architectures of the neural network to investigate whether it impacts the final price and/or the number of iterations required for convergence. To this aim, we considered an option with a comparatively short delivery period of one week, and constraints

$$\begin{aligned} N_m &= 0\,\mathrm{MWh}\ , \quad N_M = 1\,\mathrm{MWh} \\ C_m &= 3\,\mathrm{MWh}\ , \quad C_M = 5\,\mathrm{MWh} \end{aligned} \tag{41}$$

On this payoff, we tried both wide and deep architectures: 1 hidden layer with 64 units (wide), 5 hidden layers with 4 units each (deep), and 5 hidden layers with 64 neurons each (wide and deep).

We show in Figure 8 the learning curves of the PPO algorithm with β fixed to 0.5 (the default in the baselines implementation). We can see that the shallow architecture finds a slightly suboptimal policy regardless of the high number of units (and hence of parameters). Deeper networks work better, but the very complex 5-by-64 network is prone to over-fitting, and indeed its performance deteriorates if optimized for too long. Hence, in what follows we focus on the 5-by-4 network.

4.2.2 Comparison with the LSMC Algorithm

In this section, we consider the contracts with maturity of one month delivering in May 2018 which were analysed in section 3.3.2 in the context of LSMC pricing.

After training, the option can be priced using either the LSMC or the RL candidate optimal policy. We therefore run a Monte Carlo simulation with 1 million paths to get the price.

We performed a grid search on the hyperparameter β and found that $\beta = 0.01$ was more effective than the default $\beta = 0.5$, corresponding to slower updates of the network which approximates the value function. Moreover, since the objective function is not convex, we run each optimization four times with different random starting guesses for θ, and then choose the optimized network with the best in-sample performance on the last 1,000,000 training episodes. The out-of-sample results of such network are shown in Table 2, and they are compatible with the LSMC results in Table 1 within statistical uncertainty.

We also see that the unconstrained PPO agent successfully identifies a strategy of bang-bang type for the case $C_m = 12\,\mathrm{MWh}$ in which we know that it is optimal to do so. This is exemplified by Figure 9, where we fix a decision time and plot the chosen action as a function of the other two coordinates of the network input (i.e. normalized log-spot and consumption).

5 Conclusion and Further Developments

In this paper we presented a new model to price swing option contracts. The model is able to calibrate liquid market quotes and to imply the volatility smile for futures contracts with different delivery periods. The pricing algorithm is implemented both by using a least-square Monte Carlo approach and by means of a recent reinforcement learning algorithm, namely the proximal policy optimization algorithm. Using the former, we investigate option prices and optimal strategies for different configuration of the model, and we test the impact of constraining the choice of the control problem only to bang-bang strategies. It turns out that the constraint is irrelevant only when the contract anagraphics satisfy certain hypotheses, but not otherwise. The aim of exploring techniques based on reinforcement learning is due to the fact that we wish to investigate calculation tools more suitable in more general high-dimensional settings. We find that these novel techniques also give accurate results. This paper focuses on situations where other techniques are available as a benchmark, to gather evidence on the robustness of the approach; we leave for future developments the exploration of settings where it could be the only possibility.

References

1. Barrera-Esteve C. et al. Numerical methods for the pricing of Swing options: a stochastic control approach. *Methodology and Computing in Applied Probability.* 2006; 8:517–540.

2. Bardou O., Bouthemy S., Pagès G. When are swing options bang-bang? *International Journal of Theoretical and Applied Finance*. 2010; 13:867–899.
3. Carmona R., Touzi N. Optimal multiple stopping and valuation of swing options. *Mathematical Finance*. 2008; 18:239–268.
4. De Angelis Tiziano, Kitapbayev Yerkin. On the Optimal Exercise Boundaries of Swing Put Options. *Mathematics of Operations Research*. 2018; 43:252–274. DOI: 10.1287/moor.2017.0862.
5. Thompson A. C. Valuation of path-dependent contingent claims with multiple exercise decisions over time: the case of Take or Pay. *Journal of Financial and Quantitative Analysis*. 1995; 30:271–293.
6. Lari-Lavassani Ali, Simchi Mohamadreza. A Discrete Valuation Of Swing Options. *Canadian Applied Mathematics Quarterly*. 2001; 9:35–73. DOI: 10.1216/camq/1050519917.
7. Jaillet P., Ronn E. I., Tompaidis S. Valuation of Commodity-Based Swing Options. *Management Science*. 2004; 50.
8. Haarbrücker G., Kuhn D. Valuation of electricity swing options by multistage stochastic programming. *Management Science*. 2009; 45:889–899.
9. Hambly B., Howison S., Kluge T. Modeling Spikes and Pricing Swing Options in Electricity Markets. *Quantitative Finance*. 2009; 9:937–949.
10. Marshall T.J., Reesor R. Mark. Forest of stochastic meshes: A new method for valuing high-dimensional swing options. *Operations Research Letters*. 2011; 39:17–21. ISSN: 0167-6377. DOI: 10.1016/j.orl.2010.11.003.
11. Kovacevic Raimund M., Pflug Georg Ch. Electricity swing option pricing by stochastic bilevel optimization: A survey and new approaches. *European Journal of Operational Research*. 2014; 237:389–403. ISSN: 0377-2217. DOI: 10.1016/j.ejor.2013.12.029.
12. Pflug Georg C., Broussev Nikola. Electricity swing options: Behavioral models and pricing. *European Journal of Operational Research*. 2009; 197:1041–1050. ISSN: 0377-2217. DOI: 10.1016/j.ejor.2007.12.047.
13. Bardou O., Bouthemy S., Pagès G. Optimal quantization for the pricing of swing options. *Applied Mathematical Finance*. 2009; 16:183–217.

14. Kohrs Hendrik et al. Pricing and risk of swing contracts in natural gas markets. *Review of Derivatives Research*. 2019; 22:77–167. DOI: 10.1007/s11147-018-9146-x.

15. Aleksandrov N., Hambly B. A Dual Approach to Multiple Exercise Option Problems under Constraints. *Mathematical Methods of Operations Research*. 2010; 71:503–533. DOI: 10.1007/s00186-010-0310-9.

16. Bender Christian. Dual Pricing of Multi-Exercise Options under Volume Constraints. *Finance and Stochastics*. 2011; 15:1–26. DOI: 10.1007/s00780-010-0134-8.

17. Meinshausen N., Hambly B. M. Monte Carlo Methods for the Valuation of Multiple-Exercise Options. *Mathematical Finance*. 2004; 14:557–583. DOI: 10.1111/j.0960-1627.2004.00205.x.

18. Benth F., Lempa J., Nilssen T. On the optimal exercise of swing options in electricity markets. *Journal of Energy Markets*. 2012; 4:3–28.

19. Eriksson M., Lempa J., Nilssen T. Swing options in commodity markets: A multidimensional Lévy diffusion model. *Mathematical Methods of Operational Research*. 2013; 79:31–67.

20. Kjaer Mats. Pricing of Swing Options in a Mean Reverting Model with Jumps. *Applied Mathematical Finance*. 2008; 15:479–502. DOI: 10.1080/13504860802170556.

21. Kudryavtsev Oleg, Zanette Antonino. Efficient pricing of swing options in Lévy-driven models. *Quantitative Finance*. 2013; 13:627–635. DOI: 10.1080/14697688.2012.717708.

22. Kirkby L., Deng S. Swing option pricing by dynamic programming with B-spline density projection. *International Journal of Theoretical and Applied Finance*. 2020; 22.

23. Zhang B., Oosterlee C. An efficient pricing algorithm for swing options based on Fourier cosine expansions. *Journal of Computational Finance*. 2013; 4:3–34.

24. Nastasi E., Pallavicini A., Sartorelli G. Smile Modelling in Commodity Markets. *Working paper*. 2018. URL: https://arxiv.org/abs/1808.09685.

25. Sutton Richard S., Barto Andrew G. *Reinforcement Learning: An Introduction*. Second. MIT Press, 2018. URL: http://incompleteideas.net/book/the-book-2nd.html.

26. Buehler H. et al. Deep Hedging. *Quantitative Finance*. 2019; 19:1271–1291. DOI: 10.1080/14697688.2019.1571683.

27. Halperin Igor. The QLBS Q-Learner goes NuQLear: fitted Q iteration, inverse RL, and option portfolios. *Quantitative Finance*. 2019; 19:1543–1553. DOI: 10.1080/14697688.2019.1622302.

28. Kolm P., Ritter G. Dynamic replication and hedging: A reinforcement learning approach. *The Journal of Financial Data Science*. 2019; 1:159–171.

29. Ponomarev E.S., Oseledets I. V., Cichocki A. S. Using Reinforcement Learning in the Algorithmic Trading Problem. *Journal of Communications Technology and Electronics*. 2019; 64:1450–1457. DOI: 10.1134/S1064226919120131.

30. Schulman J. et al. Proximal Policy Optimization Algorithms. *Working paper*. 2017. URL: https://arxiv.org/abs/1707.06347.

31. Benth F., Piccirilli M., Vargiolu T. Additive energy forward curves in a Heath-Jarrow-Morton Framework. *Working paper*. 2018. URL: https://arxiv.org/abs/1709.03310.

32. Achiam Josh. *SpinningUP*. https://spinningup.openai.com/en/latest/. OpenAI. 2020. URL: https://spinningup.openai.com/en/latest/ (visited on 05/12/2020).

33. Baird L. C. Reinforcement learning in continuous time: advantage updating. *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*. Vol. 4. 1994:2448–2453.

34. Schulman J. et al. High-Dimensional Continuous Control Using Generalized Advantage Estimation. *Proceedings of ICLR 2016*. 2016. URL: https://arxiv.org/abs/1506.02438.

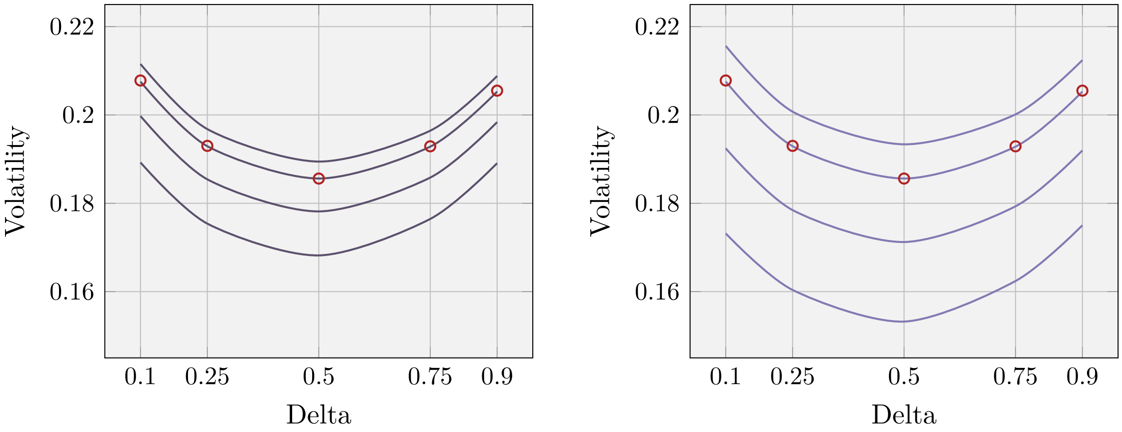

Figure 1: NG TTF PVO on `JUL18` futures quoted on 29 March 2018 on ICE market. Market (red dots) and model (blue lines) implied volatilities. Mean reversion speed equal to 0.5 on left panel, to 1 on right panel. Delivery periods ranging from top to bottom: 1 Day, One-Month, 3 Months, 6 Months.

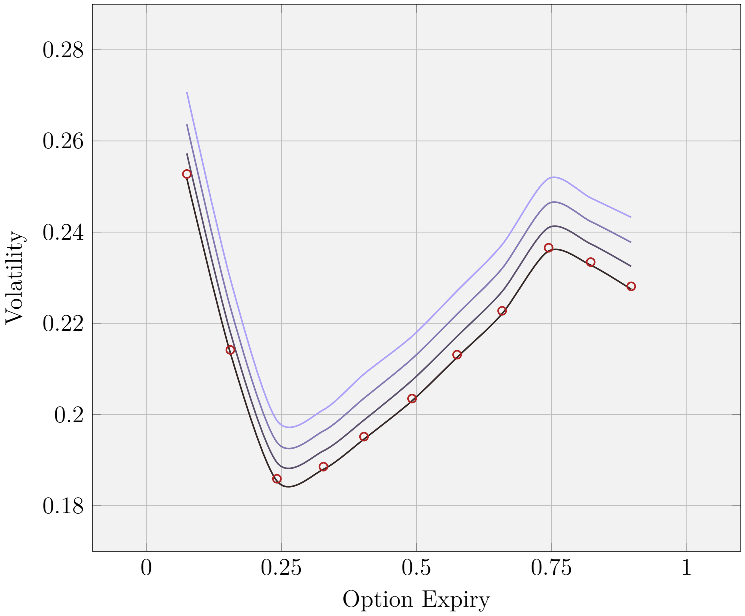

Figure 2: NG TTF PVO quoted on 29 March 2018 on ICE market. Market (red dots) one-month futures PVO at-the-money volatilities. Model (blue lines) day-ahead futures PVO at-the-money implied volatilities. Mean reversion speed ranging from top to bottom from 1.5 to zero with a step of 0.5.

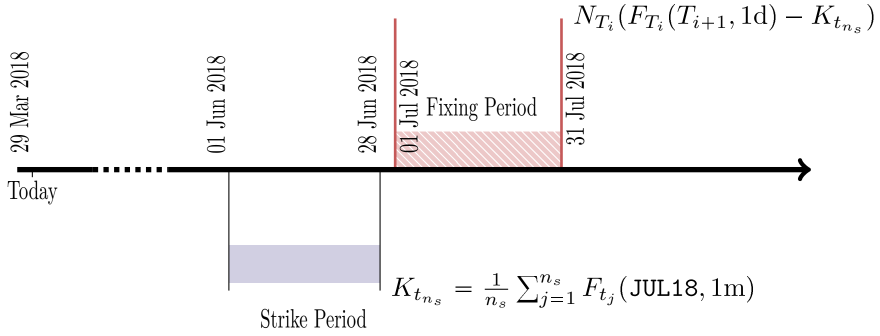
$N_{T_i}(F_{T_i}(T_{i+1}, 1\text{d}) - K_{t_{n_s}})$
29 Mar 2018
01 Jun 2018
28 Jun 2018
01 Jul 2018
Fixing Period
31 Jul 2018
Today
$K_{t_{n_s}} = \frac{1}{n_s}\sum_{j=1}^{n_s} F_{t_j}(\texttt{JUL18}, 1\text{m})$
Strike Period

Figure 3: Term-sheet data for a swing option contract with delivery in July 2018 in TTF natural gas market. The strike is fixed by avering the `JUL18` futures contract observed in the month of June.

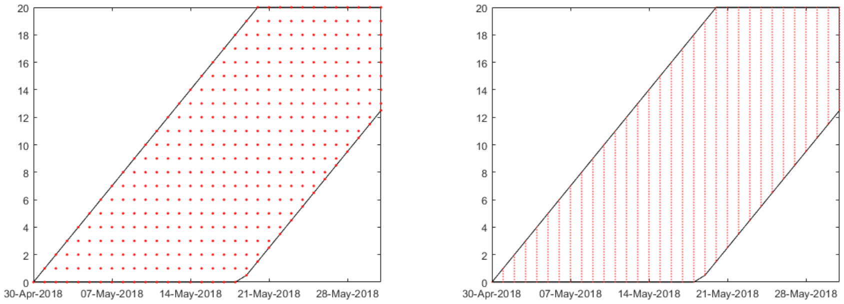

Figure 4: Grids obtained for a swing delivering in the month of April 2019 with $N_M = 1$, $N_m = 0$, $C_M = 15.7$, $C_m = 5.2$ and $\Delta = \frac{1}{6}$. Left and right panel show the bang bang and continuous cases respectively.

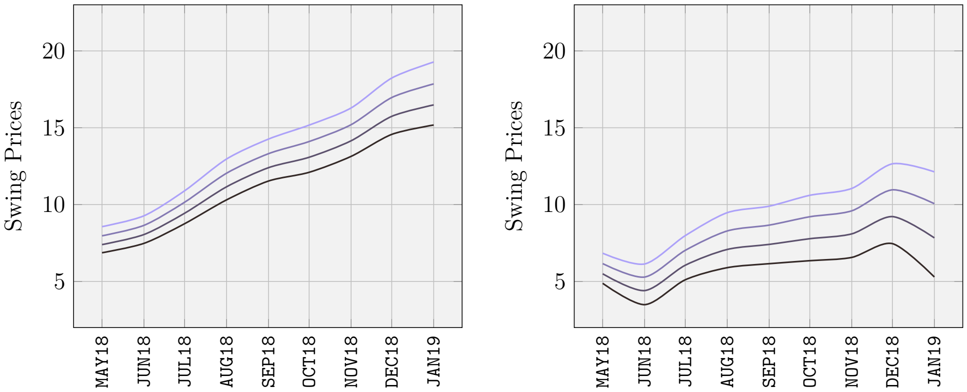

Figure 5: Swing option prices by varying the delivery starting date. Mean reversion ranging from top to bottom from 1.5 to 0 with a step of 0.5. Left panel fixed-strike contracts. Right panel floating-strike contracts.

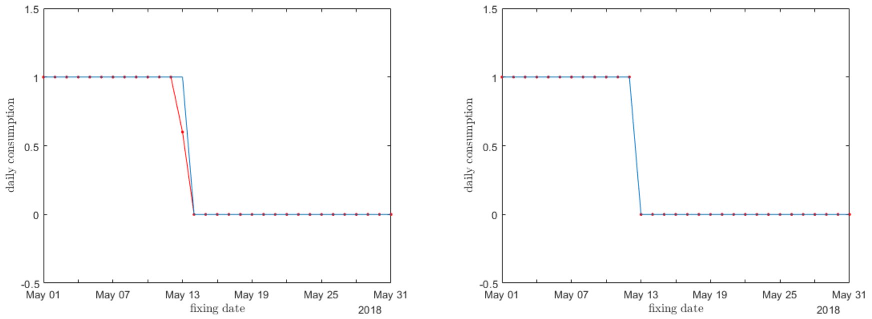

Figure 6: Daily consumption selected by the optimal strategy for a fixed-strike swing option delivering in May 2018. Left panel displays the out-of-theorem scenario, while the right panel the within-theorem. Mean reversion speed $a = 1$. The blue lines refer to the case where only bang-bang strategies are allowed, while red lines to the case without restrictions. In the bang-bang case (right panel) the two lines coincide.

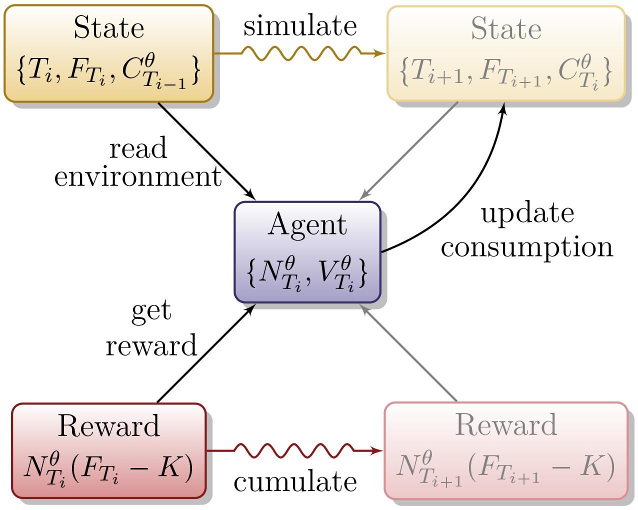
State
$\{T_i, F_{T_i}, C^\theta_{T_{i-1}}\}$
simulate
State
$\{T_{i+1}, F_{T_{i+1}}, C^\theta_{T_i}\}$
read environment
Agent
$\{N^\theta_{T_i}, V^\theta_{T_i}\}$
update consumption
get reward
Reward
$N^\theta_{T_i}(F_{T_i} - K)$
cumulate
Reward
$N^\theta_{T_{i+1}}(F_{T_{i+1}} - K)$

Figure 7: Agent description.

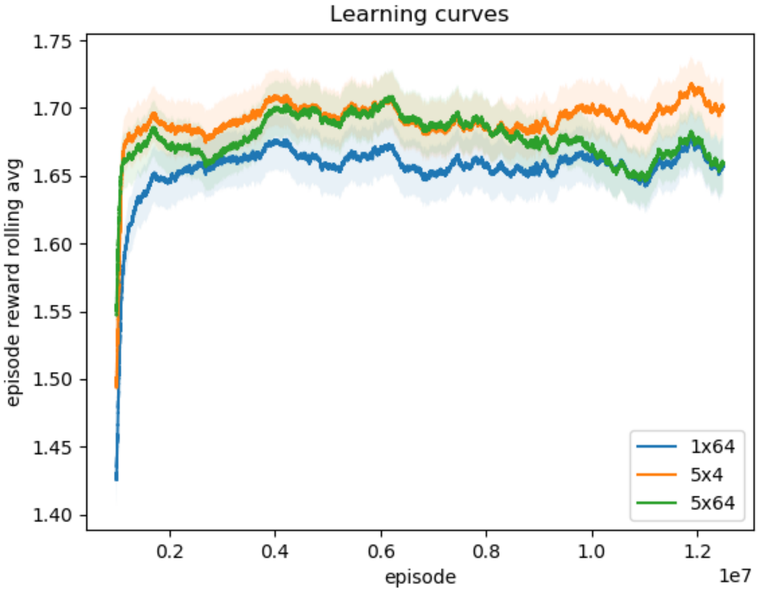

Figure 8: Learning curves. On the horizontal axis the number of training episodes. The solid lines are the moving average of the realized rewards on the last 10^6 episodes. The shadows represent the 98% confidence intervals.

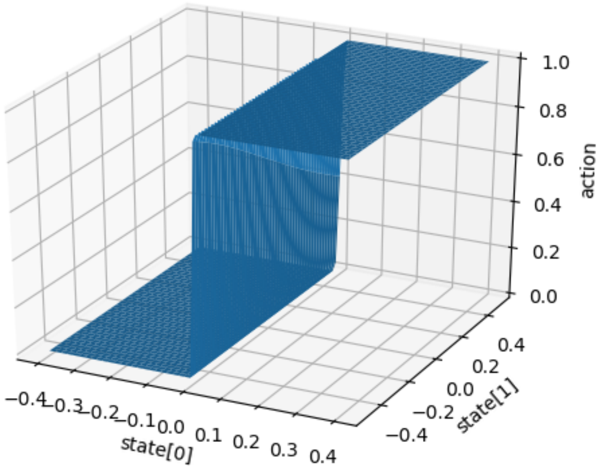

Figure 9: Normalized consumption as a function of normalized log-spot and consumption, on the fourth decision date. The axes are: $\log(F_{T_i}/F_{T_0})$; total consumption to-date remapped linearly so that its domain is $[-0.5, 0.5]$; today's consumption remapped linearly so that its domain is $[0, 1]$.

	All Strategies	Only Bang-Bang
Out of Theorem Hyp.	7.97 ± 0.04	7.76 ± 0.04
Within Theorem Hyp.	8.46 ± 0.04	8.46 ± 0.04

Table 1: Swing option prices in four different cases. We consider the reference case with $C_m = 12.5\,\text{MWh}$ not satisfying the Theorem 2.4 in Bardou et al.[13], and a variant by changing the constraint to $C_m = 12\,\text{MWh}$ so that the Theorem is satisfied. Prices are calculated either allowing all possible strategies or only the bang-bang ones. One-sigma statistical errors are displayed.

	All Strategies	Only Bang-Bang
Out of Theorem Hyp.	7.92 ± 0.04	7.74 ± 0.04
Within Theorem Hyp.	8.40 ± 0.04	8.37 ± 0.04

Table 2: Swing option prices in four different cases, obtained with RL. We consider the reference case with $C_m = 12.5\,\mathrm{MWh}$ not satisfying Theorem 2.4 in Bardou et al.[13], and a variant by changing the constraint to $C_m = 12\,\mathrm{MWh}$ so that the Theorem is satisfied. Prices are calculated either allowing all possible strategies or only the bang-bang ones. One-sigma statistical errors are displayed.