Reinforcement Learning for Finance:

Discrete-time Black-Scholes model

Igor Halperin

NYU Tandon School of Engineering

Brooklyn NY 2017

[1]Opinions presented in these notes are author's only, and not necessarily of his employer. The standard disclaimer applies.

There is nothing more practical than a good theory.

V. Vapnik (?)

In these notes:

- **Can a RL agent learn to trade an option from** *directly from trading data*?
- **Plan:**
  - Take a discrete-time version of the Black-Scholes model for option pricing
  - Convert it to a Markov Decision Process (MPD) model
  - Apply Reinforcement Learning methods to this model, using simulated data

Discrete-time BSM model

We start with a discrete-time version of the BSM model. The problem of option hedging and pricing in this formulation amounts to a sequential risk minimization. To define risk in an option, we follow a local risk minimization approach pioneered in the work of Föllmer and Schweizer, Schweizer (1994). A similar method was developed by physicists Potters and Bouchaud (2001), see also the work by Kapoor et. al. (2010). We use a version of this approach suggested in a Ph.D. thesis by Grau (2007).
In this approach, we take the view of a seller of a European option (e.g. a put option) with maturity T and the terminal payoff of $H_T(S_T)$ at maturity, that depends on the final stock price S_T at that time. To hedge the option, the seller use the proceeds of the sale to set up a replicating (hedge) portfolio Π_t made of the stock S_t and a risk-free bank deposit B_t. The value of hedge portfolio at any time $t \leq T$ is

$$\Pi_t = u_t S_t + B_t \tag{1}$$

where u_t is a stock position at time t, taken to hedge risk in the option.

Hedge portfolio evaluation

The replicating portfolio tries to exactly match the option price in all possible future states of the world. Start at $t = T$: the option position is closed, the hedge u_t is closed ($u_t = 0$), and therefore

$$\Pi_T = B_T = H_T(S_T) \tag{2}$$

This sets a terminal condition for B_T for all future states of the world at T. To find B_t for previous times $t < T$, we impose the *self-financing constraint* which requires that all future changes in the hedge portfolio should be funded from an initially set bank account, without any cash infusions or withdrawals over the lifetime of the option.

$$u_t S_{t+1} + e^{r\Delta t} B_t = u_{t+1} S_{t+1} + B_{t+1} \tag{3}$$

This can be expressed as a recursive relation for B_t at any time $t < T$ using its value at the next time instance:

$$B_t = e^{-r\Delta t} \left[B_{t+1} + (u_{t+1} - u_t) S_{t+1} \right], \quad t = T-1, \ldots, 0 \tag{4}$$

Hedge portfolio evaluation

Plugging this into Eq.(1) produces a recursive relation for Π_t in terms of its values at later times, which can therefore be solved backward in time, starting from $t = T$ with the terminal condition (2), and continued all the way to the current time $t = 0$:

$$\Pi_t = e^{-r\Delta t}\left[\Pi_{t+1} - u_t \Delta S_t\right], \quad \Delta S_t = S_{t+1} - e^{r\Delta t} S_t\,, \quad t = T-1, \ldots, 0 \tag{5}$$

Eqs.(4) and (5) imply that both B_t and Π_t are not measurable at any $t < T$, as they depend on the future. Respectively, their values today B_0 and Π_0 will be random quantities with some distributions. We can compute them using Monte Carlo!

Hedge portfolio evaluation with Monte Carlo

For any given hedging strategy $\{u_t\}_{t=0}^{T}$, these distributions can be estimated using Monte Carlo simulation:

- **Forward pass** : Simulate N paths of the underlying $S_1 \to S_2 \to \ldots \to S_N$,
- **Backward pass**: Evaluate Π_t going backward on each path.

As the choice of a hedge strategy does not affect the evolution of the underlying, such simulation of forward paths should only be performed once. Alternatively, we could use real historical data for stock prices, together with a pre-determined hedging strategy $\{u_t\}_{t=0}^{T}$ and a terminal condition (2).
But first, we need a hedge strategy u_t to implement this Monte Carlo!

Optimal hedging in the discrete-time BSM model

The optimal hedge $u^{\star}(S_t)$ in this model is obtained from the requirement that the variance of Π_t across all simulated MC paths at time t is minimized when conditioned on the currently available *cross-sectional* information $\mathcal{F}_t$, i.e.

$$\begin{aligned} u_t^{\star}(S_t) &= \arg\min_u Var\left[\Pi_t | \mathcal{F}_t\right] \\ &= \arg\min_u Var\left[\Pi_{t+1} - u_t \Delta S_t | \mathcal{F}_t\right] , \quad t = T-1, \ldots, 0 \end{aligned} \tag{6}$$

The first expression in Eq.(7) implies that all uncertainty in Π_t is due to uncertainty regarding the amount B_t needed to be held at the bank account at time t in order to be able to cover the future obligations at the option maturity T. This means that an optimal hedge should minimize the cost of hedge capital at each time step t.

Optimal hedging in the discrete-time BSM model

The optimal hedge $u^{\star}(S_t)$:

$$\begin{aligned} u_t^{\star}(S_t) &= arg\min_u Var\left[\Pi_t|\mathcal{F}_t\right] \\ &= arg\min_u Var\left[\Pi_{t+1} - u_t\Delta S_t|\mathcal{F}_t\right], \quad t = T-1,\ldots,0 \end{aligned} \tag{7}$$

The optimal hedge can be found analytically by setting the derivative of (7) to zero. This gives

$$u_t^{\star}(S_t) = \frac{Cov\left(\Pi_{t+1}, \Delta S_t \middle| \mathcal{F}_t\right)}{Var\left(\Delta S_t \middle| \mathcal{F}_t\right)}, \quad t = T-1,\ldots,0 \tag{8}$$

How one compute one-step expectations depends on whether we deal with a continuous or a discrete state space.
We will use a general notation as in Eq.(8) to denote similar conditional expectations where $\mathcal{F}_t$ stands for cross-sectional information set at time t.

Option pricing in discrete-time BSM

A *mean* option price $\hat{C}_t$ is defined as a time-t expected value of the hedge portfolio Π_t:

$$\hat{C}_t = \mathbb{E}_t \left[\Pi_t \middle| \mathcal{F}_t \right] \tag{9}$$

Using the equation for Π_t and the tower law of conditional expectations, we obtain

$$\begin{aligned} \hat{C}_t &= \mathbb{E}_t \left[e^{-r\Delta t} \Pi_{t+1} \middle| \mathcal{F}_t \right] - u_t(S_t)\, \mathbb{E}_t \left[\Delta S_t \middle| \mathcal{F}_t \right] \\ &= \mathbb{E}_t \left[e^{-r\Delta t} \mathbb{E}_{t+1} \left[\Pi_{t+1} \middle| \mathcal{F}_{t+1} \right] \middle| \mathcal{F}_t \right] - u_t(S_t)\, \mathbb{E}_t \left[\Delta S_t \middle| \mathcal{F}_t \right] \qquad (10) \\ &= \mathbb{E}_t \left[e^{-r\Delta t} \hat{C}_{t+1} \middle| \mathcal{F}_t \right] - u_t(S_t)\, \mathbb{E}_t \left[\Delta S_t \middle| \mathcal{F}_t \right], \quad t = T-1, \ldots, 0 \end{aligned}$$

Fair Option pricing in discrete-time BSM

The dealer cannot just ask the mean price $\hat{C}_0$ when selling the option, as she has to compensate for risk of exhausting the bank account B_t some time in the future, which would require cash infusions into the hedge portfolio, after any fixed amount $\hat{B}_0 = \mathbb{E}_0 [B_0]$ is put in the bank account at time $t = 0$ right after selling the option.
One possible specification of a risk premium is to add the cumulative expected discounted variance of the hedge portfolio along all time steps $t = 0, \ldots, N$, with a risk-aversion parameter λ:

$$C_0^{(ask)}(S, u) = \mathbb{E}_0 \left[\Pi_0 + \lambda \sum_{t=0}^{T} e^{-rt} Var \left[\Pi_t \middle| \mathcal{F}_t \right] \middle| S_0 = S, u_0 = u \right] \quad (11)$$

The option seller should *minimize* this option price to be competitive.

Maximization problem for option pricing

The problem of *minimization* of a fair (to the dealer) option price (11) can be equivalently expresses as the problem of *maximization* of its negative $V_t = -C_t^{(ask)}$, where

$$V_t(S_t) = \mathbb{E}_t \left[-\Pi_t - \lambda \sum_{t'=t}^{T} e^{-r(t'-t)} Var\left[\Pi_{t'} \middle| \mathcal{F}_{t'}\right] \middle| \mathcal{F}_t \right] \tag{12}$$

Hedging and pricing in the BS limit

The framework presented above provides a smooth transition to the strict BS limit $\Delta t \to 0$. In this limit, the BSM model dynamics under the physical measure $\mathbb{P}$ is described by a continuous-time Geometric Brownian motion with a drift μ and volatility σ:

$$\frac{dS_t}{S_t} = \mu dt + \sigma dW_t \tag{13}$$

where W_t is a standard Brownian motion.

Hedging in the BS limit

Consider first the optimal hedge strategy in the BS limit $\Delta t \to 0$.
Using the first-order Taylor expansion

$$\hat{C}_{t+1} = C_t + \frac{\partial C_t}{\partial S_t}\Delta S_t + O\left(\Delta t\right) \tag{14}$$

in (**??**), we obtain

$$u_t^{BS}\left(S_t\right) = \lim_{\Delta t \to 0} u_t^{\star}(S_t) = \frac{\partial C_t}{\partial S_t} \tag{15}$$

which is the correct optimal hedge in the continuous-time BSM model.

Pricing in the BS limit

First compute the limit of the second term in the equation for $\hat{C}_t$:

$$\lim_{\Delta t \to 0} u_t(S_t)\, \mathbb{E}_t \left[\Delta S_t \middle| \mathcal{F}_t \right] = \lim_{dt \to 0} u_t^{BS} S_t (\mu - r) dt = \lim_{dt \to 0} (\mu - r) S_t \frac{\partial C_t}{\partial S_t} dt$$

To evaluate the first term in the equation for $\hat{C}_t$, we use the second-order Taylor expansion:

$$\begin{aligned}
\hat{C}_{t+1} &= C_t + \frac{\partial C_t}{\partial t} dt + \frac{\partial C_t}{\partial S_t} dS_t + \frac{1}{2} \frac{\partial^2 C_t}{\partial S_t^2} (dS_t)^2 + \dots \\
&= C_t + \frac{\partial C_t}{\partial t} dt + \frac{\partial C_t}{\partial S_t} S_t (\mu dt + \sigma dW_t) \\
&+ \frac{1}{2} \frac{\partial^2 C_t}{\partial S_t^2} S_t^2 \left(\sigma^2 dW_t^2 + 2\mu\sigma dW_t dt \right) + O\left(dt^2 \right)
\end{aligned}$$

Pricing in the BS limit

We had the recursive formula for the option price

$$\hat{C}_t = \mathbb{E}_t \left[e^{-r\Delta t} \hat{C}_{t+1} \middle| \mathcal{F}_t \right] - u_t(S_t)\, \mathbb{E}_t \left[\Delta S_t \middle| \mathcal{F}_t \right], \quad t = T-1, \ldots, 0$$

Use Eqs.(15) and (16) for two terms in this expression, use $\mathbb{E}\left[dW_t\right] = 0$ and $\mathbb{E}\left[dW_t^2\right] = dt$, and simplify. We see that the stock drift μ drops out from the problem.
This becomes the celebrated **Black-Scholes equation** in the limit $dt \to 0$:

$$\frac{\partial C_t}{\partial t} + rS_t \frac{\partial C_t}{\partial S_t} + \frac{1}{2}\sigma^2 S_t^2 \frac{\partial^2 C_t}{\partial S_t^2} - rC_t = 0 \tag{16}$$

Therefore, if the world is lognormal, both our hedging and pricing formulae become the original formulae of the Black-Scholes-Merton model in the strict limit $\Delta t \to 0$.

Markov Decision Process model for discrete-time BS model

Now we will re-formulate the discrete-time BSM model as a Markov Decision Processes (MDP) model.
Our discrete-time MDP model works directly in the physical measure $\mathbb{P}$, and addresses the problem of option pricing and hedging as a problem of stochastic optimal control in discrete time, where the system being controlled is a hedge portfolio, and control is a stock position in this hedge portfolio. The problem is then solved by a sequential maximization of "rewards" (negatives of hedge portfolio one-step variances times the risk-aversion λ, plus a drift term).
Two cases are possible:

- The model is *known*
- The model is *unknown*

Respectively, we will have *two ways* to solve a Bellman equation for these two cases.

State variables

We first define a new variable X_t by the following relation:

$$X_t = -\left(\mu - \frac{\sigma^2}{2}\right) t + \log S_t \tag{17}$$

This implies that

$$dX_t = -\left(\mu - \frac{\sigma^2}{2}\right) dt + d\log S_t = \sigma dW_t \tag{18}$$

Therefore, X_t is a standard Brownian motion, scaled by volatility σ. For a given X_t in a MC scenario, the corresponding value of S_t is

$$S_t = e^{X_t + \left(\mu - \frac{\sigma^2}{2}\right) t} \tag{19}$$

As X_t is a martingale, i.e. $\mathbb{E}\left[dX_t\right] = 0$, on average it should not run too far away from X_0 during the lifetime of an option. The state variable X_t is time-uniform, unlike the stock price S_t that has a drift. But Eq.(19) can always be used in order to map the dynamics of S_t into the dynamics of X_t.

Finite-state approximation

We can discretize the set of admissible values of state variables X_t defined in Eq.(17), while keeping the relation (19) that expresses the stock price S_t in terms of the (now discretized) state variable X_t.
Can use a discrete-time, discrete-state Markov Chain model as an approximation to a continuous-state BSM dynamics (Duan and Simonato, 2001).
A state-space discretization is *not* necessary within our framework. Below we assume a general continuous-state case in a Monte Carlo setting.
Simplifications due to discretization:

- Can use simple algorithms for discrete-state MDPs (such as Q-Learning)
- Discrete state formulation simplifies calculation of one-step conditional expectations

Value function

Now we reformulate our risk minimization procedure in a language of MDP problems.
Express the dynamics in terms of variables X_t using Eq.(19). Actions $u_t = u_t(S_t)$ in terms of stock prices are then obtained by the substitution

$$u_t\left(S_t\right) = a_t\left(X_t\left(S_t\right)\right) = a_t\left(\log S_t - \left(\mu - \frac{\sigma^2}{2}\right) t\right) \tag{20}$$

Actual hedging decisions $a_t(x_t)$ are determined by a time-dependent *policy* $\pi\left(t, X_t\right)$. We consider *deterministic policies*, i.e.

$$\pi : \{0, \ldots, T-1\} \times \mathcal{X} \to \mathcal{A} \tag{21}$$

This deterministic policy maps the time t and the current state $X_t = x_t$ into the action $a_t \in \mathcal{A}$:

$$a_t = \pi(t, x_t) \tag{22}$$

Bellman equation

First re-write the value maximization problem of Eq.(12) in terms of a new state variable X_t, and with an upper index to denote its dependence on the policy π:

$$V_t^{\pi}(X_t) = \mathbb{E}_t \left[-\Pi_t(X_t) - \lambda \sum_{t'=t}^{T} e^{-r(t'-t)} Var\left[\Pi_{t'}(X_{t'}) \middle| \mathcal{F}_{t'}\right] \middle| \mathcal{F}_t \right]$$

$$= \mathbb{E}_t \left[-\Pi_t(X_t) - \lambda\, Var\left[\Pi_t\right] - \lambda \sum_{t'=t+1}^{T} e^{-r(t'-t)} Var\left[\Pi_{t'}(X_{t'}) \middle| \mathcal{F}_{t'}\right] \middle| \mathcal{F}_t \right] \quad (23$$

The last term in this expression can be expressed in terms of V_{t+1} using the definition of the value function with a shifted time argument:

$$-\lambda \mathbb{E}_{t+1} \left[\sum_{t'=t+1}^{T} e^{-r(t'-t)} Var\left[\Pi_{t'} \middle| \mathcal{F}_{t'}\right] \right] = \gamma \left(V_{t+1} + \mathbb{E}_{t+1}\left[\Pi_{t+1}\right] \right) \quad (24)$$

Here $\gamma \equiv e^{-r\Delta t}$ is a discrete-time discount factor

Bellman equation

Substitute this into (23), re-arrange terms, and use the portfolio process Eq.(5).
This produces the Bellman equation for our model:

$$V_t^{\pi}(X_t) = \mathbb{E}_t^{\pi}\left[R(X_t, a_t, X_{t+1}) + \gamma V_{t+1}^{\pi}(X_{t+1})\right] \tag{25}$$

Here $R(X_t, a_t, X_{t+1})$ is a one-step time-dependent random reward

$$R_t(X_t, a_t, X_{t+1}) = \gamma a_t \Delta S_t(X_t, X_{t+1}) - \lambda Var\left[\Pi_t \middle| \mathcal{F}_t\right] \tag{26}$$
$$= \gamma a_t \Delta S_t(X_t, X_{t+1}) - \lambda\gamma^2 \mathbb{E}_t\left[\hat{\Pi}_{t+1}^2 - 2a_t\Delta\hat{S}_t\hat{\Pi}_{t+1} + a_t^2\left(\Delta\hat{S}_t\right)^2\right]$$

where we used Eq.(5) in the second line, and $\hat{\Pi}_{t+1} \equiv \Pi_{t+1} - \bar{\Pi}_{t+1}$, where $\bar{\Pi}_{t+1}$ is the sample mean of all values of Π_{t+1}, and similarly for $\Delta\hat{S}_t$. For $t = T$, we have $R_T = -\lambda\, Var\left[\Pi_T\right]$ where Π_T is determined by the terminal condition (2).

Expected rewards

Note that Eq.(26) implies that the expected reward R_t at time step t is *quadratic* in the action variable a_t:

$$\mathbb{E}_t \left[R_t \left(X_t, a_t, X_{t+1} \right) \right] = \gamma a_t \mathbb{E}_t \left[\Delta S_t \right] \tag{27}$$
$$-\lambda \gamma^2 \mathbb{E}_t \left[\hat{\Pi}^2_{t+1} - 2 a_t \Delta \hat{S}_t \hat{\Pi}_{t+1} + a_t^2 \left(\Delta \hat{S}_t \right)^2 \right]$$

As we will see, this simple quadratic dependence on a_t is very useful for a solution of the MDP dynamics in this model.
Note: when $\lambda \to 0$, the expected reward is *linear* in a_t, so it does *not* have a maximum when $\lambda \to 0$ (i.e. there is no risk aversion).
In our framework, quadratic risk is incorporated in a standard (risk-neutral) MDP formulation.

Bellman optimality equation

The *optimal policy* $\pi_t^\star(\cdot|X_t)$ is determined as a policy that maximizes the value function $V_t^\pi(X_t)$:

$$\pi_t^\star(X_t) = arg \max_\pi V_t^\pi(X_t) \tag{28}$$

The optimal value function satisfies the Bellman optimality equation

$$V_t^\star(X_t) = \mathbb{E}_t^{\pi^\star}\left[R_t(X_t, u_t = \pi_t^\star(X_t), X_{t+1}) + \gamma V_{t+1}^\star(X_{t+1})\right] \tag{29}$$

If the system dynamics are *known*, the Bellman optimality equation can be solved using methods of Dynamic Programming such as Value Iteration. If dynamics are *unknown*, the optimal policy should be computed using *samples*. This is a setting of Reinforcement Learning. A formalism based on an action-value function provides a better framework for a RL setting.