Markov Decision Process Model for Discrete-time BS Model

We now re-formulate the discrete-time BSM model as a Markov Decision Processes (MDP) model: 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:

$$X_t = -\left(\mu - \frac{\sigma^2}{2}\right) t + \log S_t \qquad \text{Eq: 22}$$

This implies that

$$dX_t = -\left(\mu - \frac{\sigma^2}{2}\right) dt + d \log S_t = \sigma dW_t \qquad \text{Eq: 23}$$

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

$$S_t = e^{X_t + \left(\mu - \frac{\sigma^2}{2}\right) t} \qquad \text{Eq: 24}$$

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.

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.(24). Actions $u_t = u_t(S_t)$ in terms of stock prices are then obtained by the substitution

$$u_t(S_t) = a_t(X_t(S_t)) = a_t\left(\log S_t - \left(\mu - \frac{\sigma^2}{2}\right)t\right) \qquad \text{Eq: 25}$$

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

$$\pi : \{0, \ldots, T-1\} \times \mathcal{X} \to \mathcal{A} \qquad \text{Eq: 26}$$

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

$$a_t = \pi(t, x_t) \qquad \text{Eq: 27}$$

Bellman Equation

First re-write the value maximization problem of Eq.(14) in terms of a new state variable X_t, and with an upper index to denote its dependence on the policy π:

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

$$= \mathbb{E}_t\left[-\Pi_t(X_t) - \lambda\,\mathrm{Var}_t\left[\Pi_t\right] - \lambda \sum_{t'=t+1}^{T} e^{-r(t'-t)}\mathrm{Var}_{t'}\left[\Pi_{t'}(X_{t'})\right]\right] \qquad \text{Eq: 28}$$

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:

$$-\lambda\mathbb{E}_{t+1}\left[\sum_{t'=t+1}^{T} e^{-r(t'-t)}\mathrm{Var}_{t'}\left[\Pi_{t'}\right]\right] = \gamma\left(V_{t+1} + \mathbb{E}_{t+1}\left[\Pi_{t+1}\right]\right) \qquad \text{Eq: 29}$$

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

Bellman Equation

Substitute this into (28), re-arrange terms, and use the portfolio process Eq.(5).

This produces the Bellman equation for our model:

$$V_t^{\pi}(X_t) = \mathbb{E}_t^{\pi}\left[R(X_t, a_t, X_{t+1}) + \gamma V_{t+1}^{\pi}(X_{t+1})\right] \qquad \text{Eq: 30}$$

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

$$R_t(X_t, a_t, X_{t+1}) = \gamma a_t \Delta S_t(X_t, X_{t+1}) - \lambda \text{Var}_t[\Pi_t] \qquad \text{Eq: 31}$$

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

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 $\Delta\hat{S}_t$.

Expected Rewards

Note that Eq.(31) implies that the expected reward R_t at time step t is quadratic in the action variable a_t :

$$\mathbb{E}_t[R_t(X_t, a_t, X_{t+1})] = \gamma a_t \mathbb{E}_t[\Delta S_t] \qquad \text{Eq: 32}$$

$$-\lambda\gamma^2 \mathbb{E}_t\left[\hat{\Pi}_{t+1}^2 - 2a_t\Delta\hat{S}_t\hat{\Pi}_{t+1} + a_t^2(\Delta\hat{S}_t)^2\right]$$

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 $V_t^\pi(X_t)$:

$$\pi_t^\star(X_t) = arg \max_\pi V_t^\pi(X_t) \qquad \text{Eq: 33}$$

The optimal value function satisfies the Bellman optimality equation

$$V_t^\star(X_t) = \mathbb{E}_t^{\pi^\star}\left[R_t(X_t, u_t = \pi_t^\star(X_t), X_{t+1}) + \gamma V_{t+1}^\star(X_{t+1})\right] \qquad \text{Eq: 34}$$

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.

Action-value Function

The action-value function, or Q-function, is defined by an expectation of the same expression as in the definition of the value function (28), but conditioned on both the current state X_t and the initial action $a = a_t$, while following a policy π afterwards:

$$\begin{aligned} Q_t^\pi(x, a) &= \mathbb{E}_t\left[-\Pi_t(X_t)\,|\,X_t = x, a_t = a\right] \\ &- \lambda\,\mathbb{E}_t^\pi\left[\sum_{t'=t}^{T} e^{-r(t'-t)}\mathrm{Var}_t\left[\Pi_{t'}(X_{t'})\right]\,\middle|\,x, a\right] \end{aligned} \qquad \text{Eq: 35}$$

We can obtain the Bellman equation for the Q- function:

$$Q_t^\pi(x, a) = \mathbb{E}_t\left[R_t(X_t, a_t, X_{t+1})|\,x, a\right] + \gamma\mathbb{E}_t^\pi\left[V_{t+1}^\pi(X_{t+1})|\,x\right] \qquad \text{Eq: 36}$$

An optimal action-value function $Q_T^\star(x, a)$ is obtained when (35) is evaluated with an optimal policy $\pi_t^\star$:

$$\pi_t^\star = \arg\max_\pi Q_t^\pi(x, a) \qquad \text{Eq: 37}$$

Optimal Action-value Function

The optimal value and state-value functions are connected by the following equations

$$Q_t^\star(x,a) = \mathbb{E}_t\left[R_t(x,a,X_{t+1})\right] + \gamma \mathbb{E}_t\left[V_{t+1}^\star(X_{t+1})|\, x\right] \quad \text{Eq: 38}$$

$$V_t^\star(x) = \max_a Q_t^\star(x,a) \quad \text{Eq: 39}$$

The Bellman Optimality equation for the action-value function is obtained by substituting the second of Eq.(38) into the first one:

$$Q_t^\star(x,a) = \mathbb{E}_t\left[R_t\left(X_t,a_t,X_{t+1}\right) + \gamma \max_{a_{t+1}\in\mathcal{A}} Q_{t+1}^\star\left(X_{t+1},a_{t+1}\right)|\, x,a\right] \quad \text{Eq: 40}$$

with a terminal condition at $t = T$ given by

$$Q_T^\star(X_T, a_T = 0) = -\Pi_T\left(X_T\right) - \lambda\, \text{Var}\left[\Pi_T\left(X_T\right)\right] \quad \text{Eq: 41}$$

where Π_T is determined by the terminal condition (2). Recall that Var $[\cdot]$ here means variance with respect to all Monte Carlo paths that end up in a given state.

Greedy Policy

A paired equation defines a "greedy" policy $\pi^\star$ that always seeks an action that maximizes the action-value function in the current state:

$$\pi_t^\star(X_t) = arg \max_{a_t\in\mathcal{A}} Q_t^\star(X_t,a_t) \quad \text{Eq: 42}$$

Backward Recursion for the Q-function

Substitute the expected reward (32) into the Bellman optimality equation

$$Q_t^\star(x,a) = \mathbb{E}_t\left[R_t\left(X_t, a_t, X_{t+1}\right) + \gamma \max_{a_{t+1}\in\mathcal{A}} Q_{t+1}^\star\left(X_{t+1}, a_{t+1}\right) \middle| x, a\right]$$

We see that $Q_t^\star\left(X_t, a_t\right)$ is quadratic in the action variable a_t (here $t = T-1, T-2, \ldots, 0$):

$$\begin{aligned} Q_t^\star(X_t, a_t) &= \gamma \mathbb{E}_t\left[Q_{t+1}^\star\left(X_{t+1}, a_{t+1}^\star\right) + a_t \Delta S_t\right] \\ &- \lambda\gamma^2\, \mathbb{E}_t\left[\hat{\Pi}_{t+1}^2 - 2a_t\hat{\Pi}_{t+1}\Delta\hat{S}_t + a_t^2\left(\Delta\hat{S}_t\right)^2\right] \end{aligned} \qquad \boxed{\text{Eq: 43}}$$

Zero Risk Aversion Limit

Note that in the limit of zero risk aversion $\lambda \to 0$, this formula becomes

$$Q_t^\star(X_t, a_t) = \gamma \mathbb{E}_t\left[Q_{t+1}^\star\left(X_{t+1}, a_{t+1}^\star\right) + a_t \Delta S_t\right] \qquad \boxed{\text{Eq: 44}}$$

As in this limit $Q_t^\star(X_t, a_t) = -\Pi\left(X_t, a_t\right)$, using the the fair option price definition (11), we obtain

$$\hat{C}_t = \gamma \mathbb{E}_t\left[\hat{C}_{t+1} - a_t \Delta S_t\right] \qquad \boxed{\text{Eq: 45}}$$

This coincides with Eq.(12). This means that the recursive formula (43) correctly rolls back the BS fair option price $\hat{C}_t = \mathbb{E}_t\left[\Pi_t\right]$ in the limit $\lambda \to 0$, $\Delta t \to 0$.

Quadratic Q-function

Back to the Bellman optimality equation for the Q-function for a general case $\lambda > 0$:

$$\begin{aligned} Q_t^\star(X_t, a_t) &= \gamma \mathbb{E}_t \left[Q_{t+1}^\star \left(X_{t+1}, a_{t+1}^\star \right) + a_t \Delta S_t \right] \\ &- \lambda \gamma^2 \mathbb{E}_t \left[\hat{\Pi}_{t+1}^2 - 2 a_t \hat{\Pi}_{t+1} \Delta \hat{S}_t + a_t^2 \left(\Delta \hat{S}_t \right)^2 \right] \end{aligned}$$

Note that $Q_t^\star \left(X_t, a_t \right)$ is a quadratic function of a_t.

Optimal Action from the Q-function

In a general case $\lambda > 0$, $Q_t^\star \left(X_t, a_t \right)$ is a quadratic function of a_t:

$$\begin{aligned} Q_t^\star(X_t, a_t) &= \gamma \mathbb{E}_t \left[Q_{t+1}^\star \left(X_{t+1}, a_{t+1}^\star \right) + a_t \Delta S_t \right] \\ &- \lambda \gamma^2 \mathbb{E}_t \left[\hat{\Pi}_{t+1}^2 - 2 a_t \hat{\Pi}_{t+1} \Delta \hat{S}_t + a_t^2 \left(\Delta \hat{S}_t \right)^2 \right] \end{aligned}$$

Set the derivative to zero to find the optimal action (i.e. the hedge) $a_t^\star(S_t)$:

$$0 = \frac{\partial Q_t^\star(X_t, a_t)}{\partial a_t} = \gamma \mathbb{E}_t \left[\Delta S_t \right] + 2 \lambda \gamma^2 \mathbb{E}_t \left[\hat{\Pi}_{t+1} \Delta \hat{S}_t - a_t \left(\Delta \hat{S}_t \right)^2 \right]$$

Optimal Action from the Q-function

Re-arrange this relation to find a_t :

$$0 = \frac{\partial Q_t^\star(X_t, a_t)}{\partial a_t} = \gamma \mathbb{E}_t\left[\Delta S_t\right] + 2\lambda\gamma^2 \mathbb{E}_t\left[\hat{\Pi}_{t+1}\Delta\hat{S}_t - a_t\left(\Delta\hat{S}_t\right)^2\right]$$

The optimal action (i.e. the hedge) $a_t^\star(S_t)$ that maximizes $Q_t^\star(X_t, a_t)$ is thus computed analytically:

$$a_t^\star\left(X_t\right) = \frac{\mathbb{E}_t\left[\Delta\hat{S}_t\hat{\Pi}_{t+1} + \frac{1}{2\gamma\lambda}\Delta S_t\right]}{\mathbb{E}_t\left[\left(\Delta\hat{S}_t\right)^2\right]}$$

Optimal Action: Comparison with Pure Risk-based Hedges

The optimal action $a_t^\star(S_t)$ that maximizes $Q_t^\star(X_t, a_t)$:

$$a_t^\star\left(X_t\right) = \frac{\mathbb{E}_t\left[\Delta\hat{S}_t\hat{\Pi}_{t+1} + \frac{1}{2\gamma\lambda}\Delta S_t\right]}{\mathbb{E}_t\left[\left(\Delta\hat{S}_t\right)^2\right]} \qquad \boxed{\text{Eq: 46}}$$

In a discrete-time BSM model, we had pure risk-based hedges:

$$u_t^\star(S_t) = \frac{\mathrm{Cov}\left(\Pi_{t+1}, \Delta S_t \middle| \mathcal{F}_t\right)}{\mathrm{Var}\left(\Delta S_t \middle| \mathcal{F}_t\right)} = \frac{\mathbb{E}_t\left[\Delta\hat{S}_t\hat{\Pi}_{t+1}\right]}{\mathbb{E}_t\left[\left(\Delta\hat{S}_t\right)^2\right]} \qquad \boxed{\text{Eq: 47}}$$

As $\mathbb{E}_t\left[\Delta S_t\right] \sim \mu - r$, if we set $\mu = r$, or alternatively take the limit $\lambda \to \infty$ in Eq.(46), it coincides with a pure risk-based hedge Eq.(47) of the discrete-time BSM model.

Optimal Action: why the Additional Term?

Why the additional second term in the numerator of the optimal hedge formula?

$$a_t^\star(X_t) = \frac{\mathbb{E}_t\left[\Delta\hat{S}_t\hat{\Pi}_{t+1} + \frac{1}{2\gamma\lambda}\Delta S_t\right]}{\mathbb{E}_t\left[\left(\Delta\hat{S}_t\right)^2\right]} \qquad \text{Eq: 48}$$

The quadratic hedging of the discrete-time BSM model only looks at risk of a hedge portfolio. But here the expected reward has both a drift and variance parts, similar to the Markowitz risk-adjusted portfolio return analysis:

$$\mathbb{E}_t\left[R_t\left(X_t, a_t, X_{t+1}\right)\right] = \gamma a_t \mathbb{E}_t\left[\Delta S_t\right] - \lambda\gamma^2 \text{Var}_t\left[\Pi_t\right]$$

Resulting hedges are therefore different. For a pure risk-focused quadratic hedge, we can set $\mu = r$ or $\lambda \to \infty$ in Eq.(48). In a general case, Eq.(48) gives hedges that can be applied for both hedging and investment with options.

Backward Recursion for Optimal Q-function

As long as we have the analytical formula for optimal action $a_t^\star$, we can do the backward recursion for $t = T-1, \ldots, 0$ directly for the optimal Q-function at the optimal action:

$$\begin{aligned} Q_t^\star(X_t, a_t) &= \gamma\mathbb{E}_t\left[Q_{t+1}^\star\left(X_{t+1}, a_{t+1}^\star\right) + a_t\Delta S_t\right] \\ &- \lambda\gamma^2\,\mathbb{E}_t\left[\hat{\Pi}_{t+1}^2 - 2a_t\hat{\Pi}_{t+1}\Delta\hat{S}_t + a_t^2\left(\Delta\hat{S}_t\right)^2\right] \end{aligned} \qquad \text{Eq: 49}$$

where we use the optimal hedge

$$a_t^\star(X_t) = \frac{\mathbb{E}_t\left[\Delta\hat{S}_t\hat{\Pi}_{t+1} + \frac{1}{2\gamma\lambda}\Delta S_t\right]}{\mathbb{E}_t\left[\left(\Delta\hat{S}_t\right)^2\right]} \qquad \text{Eq: 50}$$

Optimal Price and Hedge from Backward Recursion

The backward recursion given by Eqs. (49) and (50) proceeds all the way backward starting at $t = T - 1$ to the present t = 0.

Note that:

- Optimization is analytical
- As the backward recursion is applied directly to the optimal Q-function Qt $(S_t, a_t^\star)$, neither continuous nor discrete action space representation is required in our setting, as the action in this equation is always just one optimal action.

The ask option price is a negative of the Q-function:

$$C_t^{ask}(S_t) = -Q_t(S_t, a_t^\star) \qquad \text{Eq: 51}$$

Comparison with the Black Scholes Model

- In the DP formulation, both optimal option price and hedge are parts of the same value. $Q_t^\star(X_t, a_t)$. In the BS model, we have two separate formulas for the price and the hedge.
- In the DP formulation, hedging comes ahead of pricing. In the BS model, it is the other way around.
- Vanishing optimization problem in the BS limit:
  - A quadratic objective function in the DP setting with $\lambda > 0$, $\Delta t > 0$. The link between the price and hedge is explicit.
  - If $\lambda = 0$, no quadratic optimization in the DP sense. Can still do risk-minimization hedging, but the link between the price and hedge is lost.
  - If both $\lambda = 0$ and $\Delta t = 0$, then risk is lost two, nothing to optimize any more. Back to the BS formulae

Q-function for a Discrete-state MDP

For a discrete state formulation, there is a finite set of nodes $\{X_n\}_{n=1}^{M}$, with values Q_n of the optimal Q-function at these nodes. We can write the node value using an index-free notation:

$$Q(X) = \sum_{n=1}^{M} Q_n \delta_{X,X_n}$$

where δ_{X,X_n} is the Kronecker symbol:

$$\delta_{X,X_n} = \begin{cases} 1, \text{ if } X = X_n \\ 0, \text{ otherwise} \end{cases}$$

Basis Functions

For a discrete state formulation, we can also write such "index-free" notation as an expansion in basis functions

$$Q(X) = \sum_{n=1}^{M} Q_n \delta_{X,X_n} \equiv \sum_{n=1}^{M} Q_n \Phi_n(X)$$

where $\Phi_n(X)$ are "one-hot" basis functions

$$\Phi_n(X) = \delta_{X,X_n} = \begin{cases} 1, \text{ if } X = X_n \\ 0, \text{ otherwise} \end{cases}$$

This can be generalized to a continuous-state formulation!

Basis Functions for a Continuous-state MDP

For a continuous state formulation, we can write a similar expansion in basis functions

$$Q(X) = \sum_{n=1}^{M} Q_n \delta_{X,X_n} \equiv \sum_{n=1}^{M} Q_n \Phi_n(X)$$

For basis-functions $\Phi_n(X)$, we can take "smoothed-out" one-hot basis functions, e.g. B-splines.

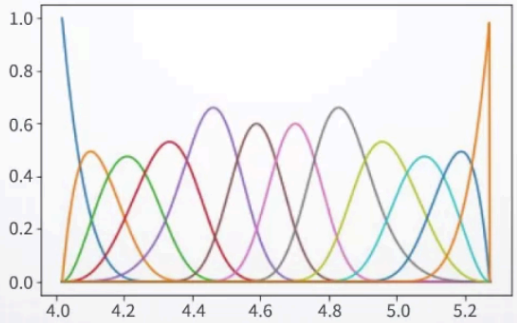

Figure: B-spline basis functions

B-splines as Basis Functions

Interpolate a function between node points $x_0, \ldots, x_M$ with a spline of order n:

Define a basis of B-splines $B_{i,n}(x)$ with $i = -n, \ldots, M-1$:

- Add additional node points $x_{-n}, \ldots, x_{-1}$ to the left of x_0, and $x_{M+1}, \ldots, x_{M+n}$ to the right of x_M
- Set $B_{i,0}(x) = 1$ if $x_i \leq x \leq x_{i+1}$, and zero otherwise.
- Define $B_{i,n}(x)$ for $n > 0$ recursively:

$$B_{i,n}(x) = \frac{x - x_i}{x_{i+n-1} - x_i} B_{i,n-1}(x) + \frac{x_{i+1} - x}{x_{i+n} - x_{i+1}} B_{i+1,n-1}(x)$$

  if $x_i \leq x \leq x_{i+n+1}$, and zero otherwise.

- B-splines are non-negative, integrate to one, and $B_{i,n}(x)$ is only non-zero on the interval $[x_i, x_{i+n+1}]$.

Monte Carlo Implementation of Backward Recursion

Use all MC paths simultaneously to learn optimal actions. Why? Because learning optimal actions for all states simultaneously means learning a policy, which is exactly our objective. Assume we have a set of basis functions $\{\Phi_n(x)\}$. Look the optimal action (hedge) $a_t^\star$ and optimal Q-function $Q_t^\star$ as expansions in basis functions:

$$\begin{aligned} a_t^\star(X_t) &= \sum_n^M \phi_{nt}\Phi_n(X_t) \\ Q_t^\star(X_t, a_t^\star) &= \sum_n^M \omega_{nt}\Phi_n(X_t) \end{aligned} \qquad \boxed{\text{Eq: 52}}$$

Coefficients ϕ_{nt} and ω_{nt} are computed recursively backward in time for $t = T-1, \ldots, 0$.

MC Implementation: Optimal Hedge

We had the backward-recursion formula

$$\begin{aligned} Q_t^\star(X_t, a_t) &= \gamma \mathbb{E}_t\left[Q_{t+1}^\star(X_{t+1}, a_{t+1}^\star) + a_t \Delta S_t\right] \\ &- \lambda\gamma^2 \mathbb{E}_t\left[\hat{\Pi}_{t+1}^2 - 2a_t\hat{\Pi}_{t+1}\Delta\hat{S}_t + a_t^2\left(\Delta\hat{S}_t\right)^2\right] \end{aligned}$$

Find coefficients ϕ_{nt} by replacing the expectation in Eq.(43) by a MC estimate, dropping all a_t-independent terms, substituting the expansion (52) for a_t, and changing the overall sign to convert maximization into minimization:

$$\begin{aligned} G_t(\phi) &= \sum_{k=1}^{N_{MC}}\left(-\sum_n \phi_{nt}\Phi_n\left(X_t^k\right)\Delta S_t^k\right. \\ &+ \left.\gamma\lambda\left(\hat{\Pi}_{t+1}^k - \sum_n \phi_{nt}\Phi_n\left(X_t^k\right)\Delta\hat{S}_t^k\right)^2\right) \end{aligned} \qquad \boxed{\text{Eq: 53}}$$

This is a quadratic function of ϕ_{nt} - can be minimized analytically!

MC Implementation: Optimal Hedge

Minimization of Eq,(53) with respect to coefficients ϕ_{nt} produces a set of linear equations:

$$\sum_{m}^{M} A_{nm}^{(t)} \phi_{mt} = B_n^{(t)}, \quad n = 1, \ldots, M \qquad \text{Eq: 54}$$

where

$$\begin{aligned} A_{nm}^{(t)} &= \sum_{k=1}^{N_{MC}} \Phi_n\left(X_t^k\right) \Phi_m\left(X_t^k\right) \left(\Delta \hat{S}_t^k\right)^2 \\ B_n^{(t)} &= \sum_{k=1}^{N_{MC}} \Phi_n\left(X_t^k\right) \left[\hat{\Pi}_{t+1}^k \Delta \hat{S}_t^k + \frac{1}{2\gamma\lambda} \Delta S_t^k\right] \end{aligned} \qquad \text{Eq: 55}$$

which produces the solution for the coefficients of expansion of the optimal action $a_t^\star(X_t)$ in a vector form:

$$\phi_t^\star = \mathbf{A}_t^{-1} \mathbf{B}_t \qquad \text{Eq: 56}$$

Optimal Hedge: Comparison with the Theoretical Formula

Our resulting expression

$$\phi_t^\star = \mathbf{A}_t^{-1} \mathbf{B}_t$$

with (let's regularize A here)

$$\begin{aligned} A_{nm}^{(t)} &= \sum_{k=1}^{N_{MC}} \Phi_n\left(X_t^k\right) \Phi_m\left(X_t^k\right) \left(\Delta \hat{S}_t^k\right)^2 + \varepsilon \delta_{nm} \\ B_n^{(t)} &= \sum_{k=1}^{N_{MC}} \Phi_n\left(X_t^k\right) \left[\hat{\Pi}_{t+1}^k \Delta \hat{S}_t^k + \frac{1}{2\gamma\lambda} \Delta S_t^k\right] \end{aligned}$$

looks very similar to the analytical expression we had before:

$$a_t^\star(X_t) = \frac{\mathbb{E}_t\left[\Delta \hat{S}_t \hat{\Pi}_{t+1} + \frac{1}{2\gamma\lambda} \Delta S_t\right]}{\mathbb{E}_t\left[\left(\Delta \hat{S}_t\right)^2\right]}$$

MC Recursion for the Optimal Q-function

Once the coefficients $\phi_t^\star$ of the optimal action $a_t^\star$ of the optimal action at time t are found, we turn to the problem of finding coefficients ω_{nt} for the optimal Q-function. To this end, the Bellman optimality equation

$$Q_t^\star(x,a) = \mathbb{E}_t\left[R_t\left(X_t, a_t, X_{t+1}\right) + \gamma \max_{a_{t+1}\in\mathcal{A}} Q_{t+1}^\star\left(X_{t+1}, a_{t+1}\right) \middle| x, a\right]$$

for $a_t = a_t^\star$ is interpreted as regression of the form

$$R_t\left(X_t, a_t^\star, X_{t+1}\right) + \gamma \max_{a_{t+1}\in\mathcal{A}} Q_{t+1}^\star\left(X_{t+1}, a_{t+1}\right) = Q_t^\star(X_t, a_t^\star) + \varepsilon_t \qquad \text{Eq: 57}$$

where ε_t is a random noise at time t with mean zero (check it by taking expectations).

The reward $R_t = \gamma\Pi_{t+1} - \Pi_t - \lambda \text{Var}_t\left[\Pi_t\right]$ is computed from simulated paths.

MC Recursion for the Optimal Q-function

Interpret the Bellman optimality equation as regression

$$R_t\left(X_t, a_t^\star, X_{t+1}\right) + \gamma \max_{a_{t+1}\in\mathcal{A}} Q_{t+1}^\star\left(X_{t+1}, a_{t+1}\right) = Q_t^\star(X_t, a_t^\star) + \varepsilon_t \qquad \text{Eq: 58}$$

Coefficients ω_{nt} are therefore found by solving the following least-square optimization problem:

$$F_t(\omega) = \sum_{k=1}^{N_{MC}} \left(R_t\left(X_t, a_t^\star, X_{t+1}\right) + \gamma \max_{a_{t+1}\in\mathcal{A}} Q_{t+1}^\star\left(X_{t+1}, a_{t+1}\right) - \sum_n^M \omega_{nt}\Phi_n\left(X_t^k\right)\right)^2 \qquad \text{Eq: 59}$$

MC Recursion for the Optimal Q-function

Introducing another pair of a matrix $\mathbf{C_t}$ and a vector $\mathbf{D_t}$ with elements

$$C_{nm}^{(t)} = \sum_{k=1}^{N_{MC}} \Phi_n\left(X_t^k\right) \Phi_m\left(X_t^k\right) \qquad \text{Eq: 60}$$

$$D_n^{(t)} = \sum_{k=1}^{N_{MC}} \Phi_n\left(X_t^k\right) \left(R_t\left(X_t, a_t^\star, X_{t+1}\right) + \gamma \max_{a \in \mathcal{A}} Q_{t+1}^\star\left(X_{t+1}, a\right)\right)$$

we obtain the vector-valued solution for optimal weights ω_t defining the optimal Q-function at time t:

$$\omega_t^\star = \mathbf{C}_t^{-1}\mathbf{D}_t \qquad \text{Eq: 61}$$

For $t = T-1, \ldots, 0$:

- Compute matrix A_t and vector B_t
- Compute the optimal action using $\phi_t^\star = \mathbf{A}_t^{-1}\mathbf{B}_t$
- Compute the optimal reward $R_t^\star$
- Compute matrix C_t and vector D_t
- Compute the optimal Q-function using $\omega_t^\star = \mathbf{C}_t^{-1}\mathbf{D}_t$

This algorithm provides a practical MC implementation of the backward recursion scheme using expansions in basis functions. We can use this approach to find optimal price and optimal hedge when the dynamics are known.