MGTF 413: Computational Finance Methods Lecture Notes

Jiahui Shui

February 28, 2024

Some notes may deviate from what we learned in course. Use at your own risk.

1 Optimization

An optimization problem looks like:

$$\min_{x \in S} f(x) \tag{1}$$

f is called objective function. The components of $x \in \mathbb{R}^n$ are the decision variables. S is the constraint set or feasible set. $x^* = \text{argmin}_{x \in S} f(x)$ is called the minimizer.

Rather than writing in argmax/argmin form, I'll write the optimization into the following form:

$$\begin{aligned} \max_{x} \quad & f(x) \\ \text{s.t.} \quad & g_i(x) = 0, \quad i \in \mathcal{I} \\ & h_j(x) \geq 0, \quad j \in \mathcal{J} \end{aligned} \tag{2}$$

1.1 Linear Programming

A function $l(x)$ for $x \in \mathbb{R}^n$ is called linear if $l(x)$ is a linear combination of the components $x_1, \cdots, x_n$. That is, we can find a vector $c \in \mathbb{R}^n$ such that $l(x) = c^T x$. Property: $l(\alpha x) = \alpha l(x)$ and $l(x + y) = l(x) + l(y)$ for any $x, y \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$.

The graph of a linear function $l(x) = c^T x, x \in \mathbb{R}^n$ is an $n-$dimensional plane living in $\mathbb{R}^{n+1}$. For example, consider $x \in \mathbb{R}$, then $l(x) = cx$ is a line in $\mathbb{R}^2$.

Definition 1 (Level Sets) *We call $\{x | g(x) = \alpha\}$ the $\alpha-$level set of function $g(x)$.*

Definition 2 (Hyperplane) *We call $\{x | c^T x = \alpha\}, c \neq 0$ a hyperplane, which is a $n - 1$ dimensional hyperplanes in $\mathbb{R}^n$.*

Definition 3 (Half-Space) *We call $\{x | c^T x \geq \alpha\}, c \neq 0$ a half space. c is the **outer-norm** of the half-space.*

Standard form of LP:

$$\begin{aligned} \max_{x} \quad & c^T x \\ \text{s.t.} \quad & Ax = b \\ & x \geq 0 \end{aligned} \tag{3}$$

where $A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m, c \in \mathbb{R}^n$. The constraint $x \geq 0$ denotes $x_i \geq 0$ for all $i = 1, \cdots, n$.

Now we might have a question, what if the given problem is not the standard form? For example, consider the following optimization problem:

$$\begin{aligned} \max_{x_1, x_2} \quad & c_1 x_1 + c_2 x_2 \\ \text{s.t.} \quad & 2x_1 + x_2 \leq 12 \end{aligned} \tag{4}$$

Then we can introduce four non-negative variables: y_1, z_1, y_2, z_2, such that

$$x_1 = y_1 - z_1, \quad x_2 = y_2 - z_2$$

Hence, we can rewrite the optimization problem (4) into the following form:

$$\begin{aligned} \max_{y_1, y_2, z_1, z_2} \quad & c_1 y_1 + c_2 y_2 - c_1 z_1 - c_2 z_2 \\ \text{s.t.} \quad & 2y_1 + y_2 - 2z_1 - z_2 \leq 12 \\ & y_1, y_2, z_1, z_2 \geq 0 \end{aligned} \tag{5}$$

Furthermore, introduce $w \geq 0$, then

$$\begin{aligned} \max_{y_1, y_2, z_1, z_2, w} \quad & c_1 y_1 + c_2 y_2 - c_1 z_1 - c_2 z_2 \\ \text{s.t.} \quad & 2y_1 + y_2 - 2z_1 - z_2 + w = 12 \\ & y_1, y_2, z_1, z_2, w \geq 0 \end{aligned} \tag{6}$$

That is, we can add more decision variables into the optimization problem to convert it into standard form. These additional variables are called surplus and slack variables. Summary of procedures:

1. Introduce non-negative variables ($x \geq 0$ in standard form)
2. Convert inequalities into equalities. $Ax \leq b$ can be converted into $Ax + y = b$ for $y \geq 0$. ($Ax \geq b$ can be written as $Ax = b + y$).

Skip this if you want. Simplex Method. For more details, take a look at chapter 2 and chapter 3 of [1]

Definition 4 *A point x in a convex set C is said to be an extreme point of C if there are* ***no*** *two distinct points $x_1, x_2 \in C$ such that $x = \alpha x_1 + (1 - \alpha) x_2$ for some $\alpha \in (0, 1)$.*

Definition 5 (Polytope, Polyhedron) *A set which can be expressed as the intersection of a finite number of closed half spaces is said to be a convex polytope. A nonempty bounded polytope is called a polyhedron.*

1.2 Non-linear Optimization

Example 1 (Markowitz Mean-Variance Optimization) *Let x_i be the proportion of the portfolio invested in asset i, and μ_i be the expected return of asset i. Moreover, let x and μ denote corresponding vector of x_i and μ_i. Σ is the covariance matrix of stocks, i.e.*

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \cdots & \rho_{1n}\sigma_1\sigma_n \\ \rho_{21}\sigma_2\sigma_1 & \sigma_2^2 & \cdots & \rho_{2n}\sigma_2\sigma_n \\ \vdots & \vdots & & \vdots \\ \rho_{n1}\sigma_n\sigma_1 & \rho_{n2}\sigma_n\sigma_2 & \cdots & \sigma_n^2 \end{pmatrix}$$

For simplicity, we denote

$$e = (1, 1, \cdots, 1)^T$$

Therefore, the portfolio has expectation and variance:

$$\mathbb{E}[x] = \mu^T x \quad \mathrm{Var}[x] = x^T \Sigma x$$

The optimization problem is (we allow short-sale here.)

$$\begin{aligned} \min_x \quad & x^T \Sigma x \\ \text{s.t.} \quad & \mu^T x \geq R \\ & e^T x = 1 \end{aligned} \tag{7}$$

Solution: Now we will use matrix calculus and KKT to help us to solve this problem: The Lagrangian is

$$\mathcal{L} = x^T \Sigma x + \lambda(e^T x - 1) + \nu(R - \mu^T x) \tag{8}$$

where λ, ν are Lagrange multipliers. We have

$$\frac{\partial \mathcal{L}}{\partial x} = 2\Sigma x + \lambda e - \nu_1 \mu = 0$$

Hence

$$x^* = \Sigma^{-1} \left(\frac{1}{2}\nu\mu - \frac{1}{2}\lambda e \right)$$

If $\mu^T x > R$, then $\nu = 0$ due to complementary slackness condition. However, since Σ is positive semi-definite, then so does Σ^{-1}. $\nu = 0$ will lead to $x^* \leq 0$, which is not feasible. So $\mu^T x = R$.

Therefore, we will get the following equation system from KKT:

$$\begin{pmatrix} 2\Sigma & e & \mu \\ e^T & 0 & 0 \\ \mu^T & 0 & 0 \end{pmatrix} \begin{pmatrix} x \\ \lambda \\ \nu \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ R \end{pmatrix} \tag{9}$$

Solving this system we get the optimal x^*: ∎

Remark 1 *When non-negative constraint is added, there is no closed-form solution for this problem.*

There are some (numerical) methods to handle non-linear optimization problem:

- (Steepest) descent method (calculus based)
- Newton's method (calculus based)
- Interior point methods
- Sequential quadratic programming

1.2.1 Gradient-Descent

The steepest descent direction for objective function is $-\nabla f(x)$, i.e. negative gradient direction. Steps:

1. Start with location that is a guess of the minimizer: x^0
2. Move a certain distance in $-\nabla f(x^0)$, call it x^1.
$$x^1 = x^0 - \alpha \nabla f(x^0) \tag{10}$$
3. Iterate by
$$x^k = x^{k-1} - \alpha \nabla f(x^{k-1}) \tag{11}$$

actually, α can be varying.

1.2.2 Newton's Method

Newton's method is to solve root for a (nonlinear) function $g(x)$. Recall the first order condition is just $\nabla f(x) = 0$. So we can use Newton's method to find root for gradient, therefore it might be possible minimum/maximum of objective function.

Algorithm for find root for $g(x)$ (univariate):

1. Start with x^0
2. Iterate
$$x^{k+1} = x^k - \frac{g(x^k)}{g'(x^k)} \tag{12}$$
3. Stop iterations if $|g(x^k)|$ is small or $|x^k - x^{k-1}|$ is small.

For univariate optimization problem: $\min_x \quad f(x)$:

1. Start with x^0

2. Iterate
$$x^{k+1} = x^k - \frac{f'(x^k)}{f''(x^k)} \tag{13}$$

3. Stop iterations if $|f'(x^k)|$ is small or $|x^k - x^{k-1}|$ is small.

In multivariate case, first order derivative is gradient, second order derivative is Hessian matrix. Therefore, if we want to find root for $G(x), x \in \mathbb{R}^n$, we can do iteration:

$$x^{k+1} = x^k - [\nabla G(x^k)]^{-1} G(x^k) \tag{14}$$

For $\min_{x \in \mathbb{R}^n} F(x)$, we have

$$x^{k+1} = x^k - [\nabla^2 F(x^k)]^{-1} \nabla F(x^k) \tag{15}$$

where

$$\nabla^2 F(x) = \left(\frac{\partial^2 F}{\partial x_i \partial_j}\right)_{ij} = \begin{pmatrix} \frac{\partial^2 F}{\partial x_1^2} & \frac{\partial^2 F}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 F}{\partial x_1 \partial x_n} \\ \frac{\partial^2 F}{\partial x_2 \partial x_1} & \frac{\partial^2 F}{\partial x_2^2} & \cdots & \frac{\partial^2 F}{\partial x_2 \partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 F}{\partial x_n \partial x_1} & \frac{\partial^2 F}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 F}{\partial x_n^2} \end{pmatrix}$$

I wrote a tutorial for gradient descent and Newton's method. One can find it at Tutorial.

Example 2 (IRR) *(Find definition of IRR at your corporate finance textbook). The IRR of a bond is called its yield. Suppose that a non-callable bond has a maturity of 4 years. The par(face) value is 1000 and the price today is 900. The coupon rate is 10%, annually. Calculate the yield of this bond.*

Solution: We just have to calculate root for following function:

$$g(r) = \frac{100}{1+r} + \frac{100}{(1+r)^2} + \frac{100}{(1+r)^3} + \frac{1100}{(1+r)^4} - 900$$

Choose one programming language to do it! ■

The Quadratic Programming(QP) problem is a simple nonlinear constrained optimization problem. Standard form of QP:

$$\begin{aligned} \min_x \quad & \frac{1}{2} x^T Q x + c^T x \\ \text{s.t.} \quad & Ax = b \\ & x \geq 0 \end{aligned} \tag{16}$$

where $Q \in \mathbb{R}^{n \times n}, A \in \mathbb{R}^{m \times n}$.

1.2.3 Interior Point Method

2 Discrete Time Option Pricing

2.1 Introduction to Derivatives

A derivative is a financial instrument whose value depends on the value of some underlying asset or assets (in the primary market). In other words, the value of the derivative is derived from that of the underlying asset or assets. Derivatives are also called contingent claims because their value is contingent on the value of an underlying asset or assets.

Some examples of derivatives: Options, Forward Contracts, Futures, Swaps.

Forward Contract: holder of the long position promises to pay the forward price F in exchange for one unit of the asset (eg, a ton of oranges) from the holder of the short position at time T. The forward contract does not cost anything at time 0 to either party. The payoff of the forward contract is $S_T - F$.

2.2 Binomial Model

Time is discrete: $t = 0, 1, 2, \cdots$. Two assets:

- Bond: riskless asset. We denote $B_t = (1 + r_0)^t = R^t$ the value of one unit of bond at time t.
- Stock: risky asset. We denote S_t the value of one share of stock at time t. $S_t = S_{t-1}\xi_t$, where $\{\xi_t\}_{t=1}^T$ are i.i.d with

$$\mathbb{P}(\xi_t = u) = p, \quad \mathbb{P}(\xi_t = d) = 1 - p, \quad \text{for } 0 < p < 1$$

No arbitrage condition: $0 < d < R < u$. (Prove that).

Recombining tree has $T + 1$ nodes while non-recombining tree has 2^T nodes.

2.3 European Option: Pricing and Hedging

The final payoff of European Call option is $X = \max(S_T - K, 0)$.

Definition 6 *A **trading strategy** in the binomial model is a sequence of pairs of random variables:*

$$\varphi = \{(\Delta_t, \beta_t) : t = 1, 2, \cdots, T\} \tag{17}$$

where Δ_t = # of shares of stock held over $(t-1, t]$, and β_t = # units of bond held over $(t-1, t]$. Δ_t, β_t are real-valued functions of $S_0, \cdots, S_{t-1}$, which are non-anticipating. The value of strategy at time t is given by

$$V_t(\varphi) = \Delta_t S_t + \beta_t B_t \tag{18}$$

Definition 7 (Self-Financing Trading Strategy) *If* $\{(\Delta_t, \beta_t)\}$ *satisfies*

$$\Delta_t S_t + \beta_t B_t = \Delta_{t+1} S_t + \beta_{t+1} B_t \tag{19}$$

Then we call this trading strategy is self-financing.

For a European derivative with final payoff X, a *replicating strategy*(or hedging strategy) is $\varphi = \{(\Delta_t, \beta_t)\}$ such that $V_T(\varphi) = X$.

Definition 8 (Arbitrage) *An arbitrage opportunity is an opportunity for a risk free profit, i.e., one can start with no money and invest (using borrowing and short selling as needed) in such a way that ones final wealth is always non-negative and is strictly positive with positive probability.*

Theorem 1 *If* φ *is a replicating strategy for a European derivative with final payoff* X*, then the (unique) arbitrage-free initial price for this derivative is* $V_0(\varphi)$*.*

Steps: See the following slide:

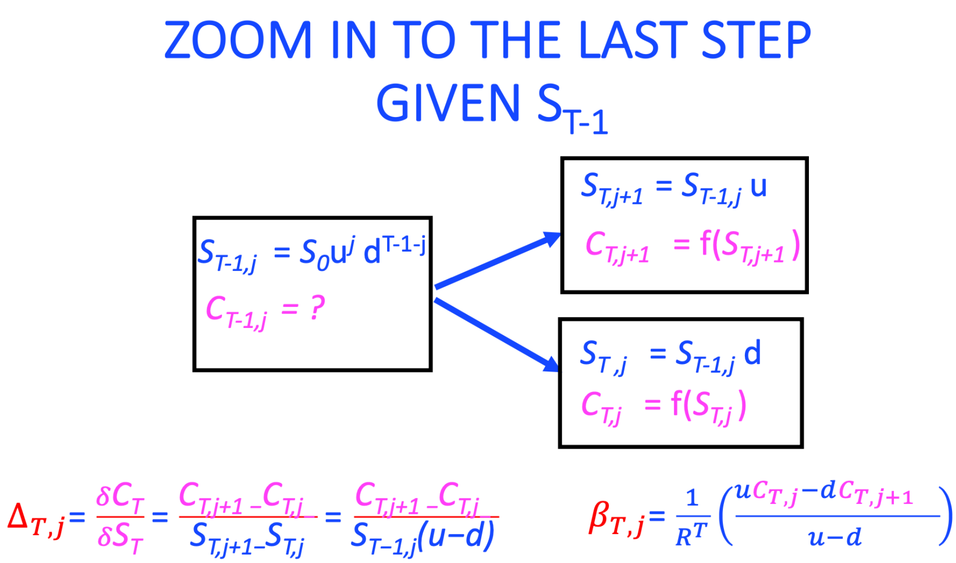

Figure 1: Binomial Model Dynamic Replicating Formulas

2.4 Risk-Neutral Probability

Let

$$p^* = \frac{R-d}{u-d} \tag{20}$$

Then we must have

$$C_{T-1,j} = \frac{1}{R}\mathbb{E}^*[C_T | S_0, \cdots, S_{T-1}] \tag{21}$$

The deep thing behind risk neutral probability is the FTAP(Fundamental Theorem of Asset Pricing). We will take about it in Spring quarter (MGTF411).

2.5 American Option

For American type derivatives, let Z_t be the payoff value at time $t = 0, 1, \cdots, T$. Then $Z_t \in \mathcal{F}_t$, which is non-anticipating(adapted). For example, American put option has payoff $Z_t = (K - S_t)^+$.

Definition 9 (Minimal Superhedging Strategy) *A self-financing strategy φ such that $V_t(\varphi) \geq Z_t$ for $t = 0, \cdots, T$ and $V_0(\varphi)$ is as small as possible.*

Let P_t be the minimal amount needed at time t to cover the payoff value of the option at times $t, \cdots, T$. Good stopping time for buyer is $\tau = \min(t : P_t = Z_t)$.

Note: $V_t(\varphi^*) = P_t$, for $0 \leq t \leq \tau^*$. But we only know $V_t(\varphi^*) \geq P_t$ for $t \geq \tau^*$.

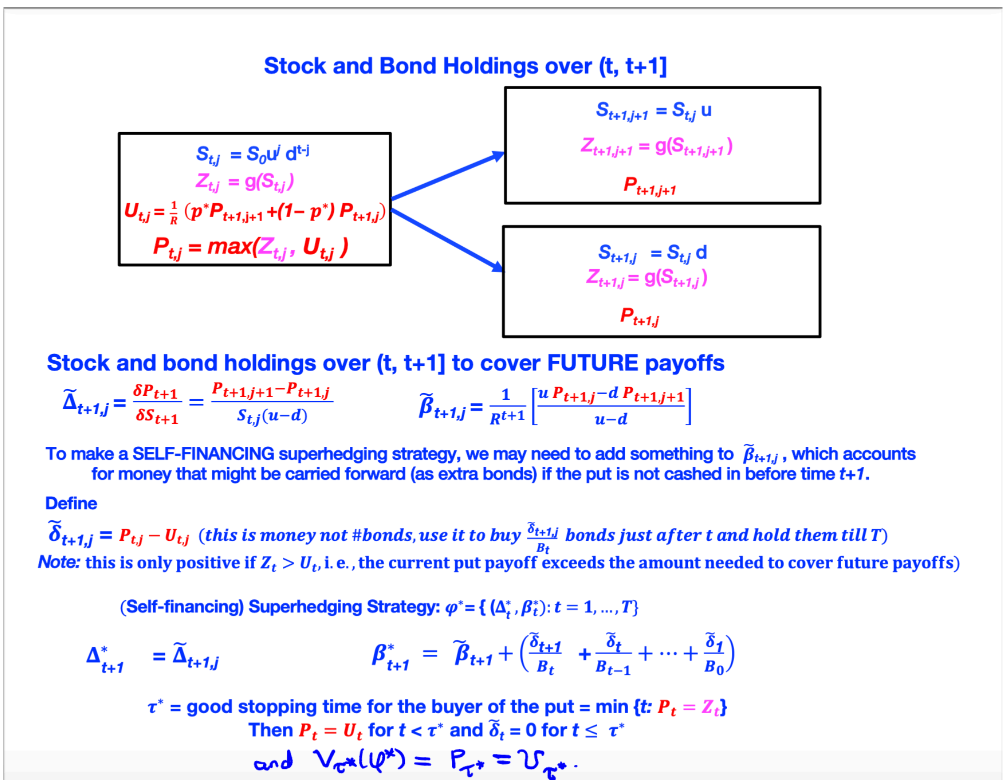

Figure 2: Binomial Model Dynamic Replicating Formulas for American Option

Remark: When we are asked to calculate superhedging portfolio, please don't forget to write down δ.

3 Monte Carlo Method

The European derivative has final payoff:

$$X = F(\{S_t : 0 \leq t \leq T\}) \tag{22}$$

Assume $\mathbb{E}^*[|X|] < \infty$. Then the unique arbitrage free initial price for a European derivative is given by

$$V_0 = e^{-rT}\mathbb{E}^*[X] = e^{-rT}\mathbb{E}^*[F(S_t : 0 \leq t \leq T)] \tag{23}$$

There are two aspects to approximating the value V_0:

1. (Static Monte Carlo): Given independent samples $X^{(1)}, \cdots, X^{(n)}$ of random variable X and then
$$V_0 = e^{-rT}\mathbb{E}^*[X] \approx e^{-rT}\frac{1}{n}\sigma_{i=1}^{n}X^{(i)} \tag{24}$$
2. (Dynamic Monte Carlo): If X is path dependent, then we need to approximate $F(\{S_t : 0 \leq t \leq T\})$ by $F(S_{t_0}, \cdots, S_{t_n})$ for $0 = t_0 < t_1 < \cdots < t_m = T$, $t_j = j\Delta t, \Delta_t = T/m$. Given independent samples $(S_{t_0}^{(i)}, \cdots, S_{t_n}^{(i)})$, $i = 1, \cdots, n$, then
$$V_0 \approx e^{-rT}\frac{1}{n}\sum_{i=1}^{n}\tilde{F}(S_{t_0}^{(i)}, \cdots, S_{t_n}^{(i)}) \tag{25}$$

3.1 Static Monte Carlo

Theorem 2 (SLLN) *Suppose that $X_1, \cdots, X_n$ are i.i.d random variables(provided finite second moments). Then we must have*

$$\frac{1}{n}\sum_{i=1}^{n}X_i \to \mathbb{E}[X_1], \quad a.s. \quad as\ n \to \infty \tag{26}$$

Let$\boldsymbol{X} = (X^{(1)}, \cdots, X^{(n)})$ be a random vector, with independent and identically distributed components $X^{(i)}$ each given by distribution function F_X. Then the sample mean of X is

$$\hat{\theta}_n(\boldsymbol{X}) = \frac{1}{n}\sum_{i=1}^{n}X^{(i)} \tag{27}$$

- Unbiased:
$$\mathbb{E}[\hat{\theta}_n(\boldsymbol{X})] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X^{(i)}] = \mathbb{E}[X] \tag{28}$$
- The variance of estimator is smaller than the variance of X
$$\mathrm{Var}[\hat{\theta}_n(\boldsymbol{X})] = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X^{(i)}) = \frac{\mathrm{Var}(X)}{n} \tag{29}$$

Note that by CLT,

$$Z_n = \frac{\hat{\theta}_n(X) - \theta}{\sqrt{\mathrm{Var}(X)/n}} \to \mathcal{N}(0, 1) \tag{30}$$

3.2 Simulating Random Variables

We have to use **pseudo-random-sample** generated by deterministic computes algorithms as approximate random numbers. The core is to generate uniform distribution on $[0,1]$. Steps:

1. Select a relative larger m.
2. Seed $x_0 \in \{1, \cdots, m\}$
3. $x_{i+1} = ax_i + c \pmod m$, where $0 < a < m, 0 \le c < m$. Often take $c = 0$, m prime, $a^{m-1} - 1$ is a multiple of m but $a^{j-1} - 1$ is not a multiple of m, $j = 1, \cdots, m-2$.
4. $u_{i+1} = x_{i+1}/m$.

3.3 Variance Reduction

Recall: For two random vairables X and Y, the covariance is

$$\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \tag{31}$$

The correlation between them is

$$\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}} \tag{32}$$

3.3.1 Antithetic Variables

Suppose that Y is a random variable and g is aa monotone function (either increasing or decreasing). We are seeking to approximate $\theta = \mathbb{E}[g(Y)]$ (assume $g(Y)$ has finite mean). The standard static Monte Carlo would approxiamate θ by sample mean:

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} g(Y^{(i)})$$

where $g(Y^{(i)})$ are i.i.d with same distribution of $g(Y)$.

Key ideas of antithetic variables: instead of $Y^{(i)}$ i.i.d, we can use pairs of random variables $(Y^{(i)}, \tilde{Y}^{(i)})$ that are independent and i.i.d but $Y^{(i)}$ and $\tilde{Y}^{(i)}$ are usually **negatively** dependent and have same distribution as Y.

Given an antithetic sequence of pairs: $\{(Y^{(i)}, \tilde{Y}^{(i)})\}_{i=1}^{n}$ that are i.i.d with $Y^{(i)} \sim \tilde{Y}^{(i)} \sim Y$ and $\mathrm{Cov}(Y^{(i)}, \tilde{Y}^{(i)}) < 0$. Therefore

$$\hat{\theta}_{av} = \frac{1}{2}\left[\frac{1}{n}\sum_{i=1}^{n} g(Y^{(i)}) + \frac{1}{n}\sum_{i=1}^{n} g(\tilde{Y}^{(i)})\right] \approx \theta \tag{33}$$

It has variance

$$\begin{aligned}\mathrm{Var}(\hat{\theta}_{av}) &= \frac{1}{4n^2}\sum_{i=1}^{n}\mathrm{Var}(g(Y^{(i)}) + g(\tilde{Y}^{(i)})) \\ &= \frac{\mathrm{Var}(Y)}{2n} + \frac{1}{2n}\mathrm{Cov}(g(Y^{(1)}), g(\tilde{Y}^{(1)})) < \mathrm{Var}(\hat{\theta}_{2n}) = \frac{1}{2}\mathrm{Var}(\hat{\theta}_n)\end{aligned} \tag{34}$$

3.3.2 Importance Sampling

The standard Monte Carlo is particularly inefficient whenever the important values of $g(Y)$ occur with low probability (takes a large n to significantly sample these values).

Idea of importance sampling:

- Replace the probability density function for the distribution of Y by a density whose samples are more likely to fall in the region of importance for g.
- Draw Monte Carlo samples from the new density, evaluate theses using g, and weight these values appropriately to compensate for tilting the distribution of Y.

Suppose that Y has probability density function f. Then

$$\theta = \mathbb{E}[g(Y)] = \int_{\mathbb{R}} g(y)f(y)\mathrm{d}y \tag{35}$$

Suppose that h is another probability density function which puts more weight on the important regions for g, and which is positive whenever f is positive.

Then

$$\theta = \mathbb{E}^f[g(Y)] = \int_{\mathbb{R}} g(y)\frac{f(y)}{h(y)}h(y)\mathrm{d}y = \mathbb{E}^h[k(Y)] \tag{36}$$

where $k(y) = g(y)f(y)/h(y)$. And then

$$\hat{\theta}_n^h = \frac{1}{n}\sum_{i=1}^{n} k(Y^{(i)}) \tag{37}$$

The variance of this estimation is

$$\mathrm{Var}(\hat{\theta}_n^h) = \frac{\mathrm{Var}^h(k(Y))}{n} = \frac{1}{n}\left[\int_{\mathbb{R}} g^2(y)\left(\frac{f(y)}{h(y)}\right)^2 h(y)\mathrm{d}y - \left(\int_{\mathbb{R}} g(y)f(y)\mathrm{d}y\right)^2\right] \tag{38}$$

The optimal h is $g(y)f(y)/\int g(z)f(z)\mathrm{d}z$. Not useful, but we can find a h close to $g \times f$.

3.4 Greeks

Suppose that $X = H(S_T)$ with $\mathbb{E}^*[|X|] < \infty$ and H "nice". The value of a hedging strategy at time t when $S_t = x$ is

$$V(t,x) = e^{-r(T-t)}\mathbb{E}^*[H(S_T)|S_t = x] \tag{39}$$

The sensitivity of this value to changes in x, t, σ are given by the Greeks:

$$\Delta_t(X) = \frac{\partial V}{\partial x}, \quad \Gamma_t(x) = \frac{\partial^2 V}{\partial x^2}, \quad \theta_t(x) = \frac{\partial V}{\partial t}, \quad \mathcal{V}_t(x) = \frac{\partial V}{\partial \sigma} \tag{40}$$

3.5 Simulation of Sample Paths

A brief introduction of Stochastic Calculus:

Theorem 3 (Itô's Lemma) *Suppose that $f(t, W_t)$ is a nice function, where W_t is a Brownian Motion. Then*

$$f(t,W_t) = f(0,W_0) + \int_0^t \frac{\partial f}{\partial s}(s,W_s)\mathrm{d}s + \int_0^t \frac{\partial f}{\partial x}(s,W_s)\mathrm{d}W_s + \frac{1}{2}\int_0^t \frac{\partial^2 f}{\partial x^2}(s,W_s)\mathrm{d}(W_s)^2 \quad (41)$$

The solution to SDE:

$$\mathrm{d}S_t = rS_t\mathrm{d}t + \sigma S_t \mathrm{d}W_t \quad (42)$$

is

$$S_t = S_0 \exp\{(r - \frac{1}{2}\sigma^2)t + \sigma W_t\} \quad (43)$$

Let $\Delta t = T/N$, $t_j = j\Delta t, j = 0, \cdots, N$. Then define

$$\tilde{S}_{t_j} = \tilde{S}_{t_{j-1}} \exp\{\sigma\sqrt{\Delta t}Z_j + (r - \frac{1}{2}\sigma^2)\Delta t\} \quad (44)$$

where $Z_1, \cdots, Z_N$ are i.i.d $\mathcal{N}(0,1)$. We can approximate $F(S_t, 0 \le t \le T)$ by $F(\tilde{S}_t, 0 \le t \le T)$. If X is path dependent, need to approximate $X = F(S_t, 0 \le t \le T)$ by $\tilde{F}(\tilde{S}_{t_j}, 0 \le j \le N)$.

3.6 Simulating Solution to a SDE

Itô diffusion:

$$\mathrm{d}X_t = \mu(t,X_t)\mathrm{d}t + \sigma(t,X_t)\mathrm{d}W_t \quad (45)$$

Idea: $\Delta X_t \approx \mu(t,X_t)\Delta t + \sigma(t,X_t)\Delta W_t$.

1. **Euler Scheme**: Fix $N > 0$, let $\Delta_t = T/N$ and $t_j = j\Delta t, j = 0, 1, \cdots, N$. Recursely define

$$\begin{cases} \tilde{X}_{t_j} = \tilde{X}_{t_{j-1}} + \mu(t_{j-1}, \tilde{X}_{t_{j-1}})\Delta t + \sigma(t_{j-1}, \tilde{X}_{t_{j-1}})\sqrt{\Delta t}Z_j \\ \tilde{X}_0 = X_0 \end{cases} \quad (46)$$

   where $Z_j, j = 1, \cdots, N$ i.i.d $\mathcal{N}(0,1)$.

   Error estimates: If μ and σ satisfy for all x, y and $s, t \in [0, T]$

   - $|\mu(t,x) - \mu(t,y)| + |\sigma(t,x) - \sigma(t,y)| \le C_1|x - y|$ (uniform)

4 Partial Differential Equations

References

[1] David G Luenberger, Yinyu Ye, et al. *Linear and nonlinear programming*. Vol. 2. Springer, 1984.

Appendix A: Matrix Calculus

A.1 Scalar Function

Suppose that $f(X)$ is a scalar function of matrix X ($m \times n$). Then the total derivative of f is

$$\mathrm{d}f = \sum_{i=1}^{m}\sum_{j=1}\frac{\partial f}{\partial X_{ij}}\mathrm{d}X_{ij} = \mathrm{tr}\left(\frac{\partial f}{\partial X}^{T}\mathrm{d}X\right) \tag{47}$$

We can use this formula to find the derivative. Here are some properties:

1. $\mathrm{d}(X \pm Y) = \mathrm{d}X \pm \mathrm{d}Y$
2. $\mathrm{d}(XY) = (\mathrm{d}X)Y + X(\mathrm{d}Y)$
3. $\mathrm{d}(X^T) = (\mathrm{d}X)^T$
4. $\mathrm{d}(\mathrm{tr}(X)) = \mathrm{tr}(\mathrm{d}X)$
5. **Inverse**: $\mathrm{d}X^{-1} = -X^{-1}(\mathrm{d}X)X^{-1}$. Sketch of proof: Take differentiation at BHS of $XX^{-1} = I$.
6. **Determinant**: $\mathrm{d}|X| = \mathrm{tr}(X^*\mathrm{d}X)$, where X^* is the adjugate matrix of X. When X is invertible, then $\mathrm{d}|X| = |X|\mathrm{tr}(X^{-1}\mathrm{d}X)$.
7. $\mathrm{d}(X \odot Y) = \mathrm{d}X \odot Y + X \odot \mathrm{d}Y$, where $\odot$ denotes element-wise product, (or Hadamard product, etc.), i.e. $(A \odot B)_{ij} = (A)_{ij}(B)_{ij}$
8. Element-wise Function: suppose that $\sigma(X) := [\sigma(X_{ij})]$. $\sigma'(X) := [\sigma'(X_{ij})]$. Then $\mathrm{d}\sigma(X) = \sigma'(X) \odot \mathrm{d}X$. For example:

$$X = \begin{pmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{pmatrix}, \quad \mathrm{d}\sin(X) = \begin{pmatrix} \cos X_{11}\mathrm{d}X_{11} & \cos X_{12}\mathrm{d}X_{12} \\ \cos X_{21}\mathrm{d}X_{21} & \cos X_{22}\mathrm{d}X_{22} \end{pmatrix} = \cos(X) \odot \mathrm{d}X$$

Some tricks for **trace**:

1. For scalar, $a = \mathrm{tr}(a)$
2. $\mathrm{tr}(A^T) = \mathrm{tr}(A)$
3. Linearity: $\mathrm{tr}(A \pm B) = \mathrm{tr}(A) \pm \mathrm{tr}(B)$
4. **Multiplication**: $\mathrm{tr}(AB) = \mathrm{BA}$, where A has the same size of B^T.
5. $\mathrm{tr}(A^T(B \odot C)) = \mathrm{tr}((A \odot B)^T C)$, where A, B, C has the same dimension.

Ok now let's begin to look at some examples.

Example 3 *Suppose that $f = \boldsymbol{a}^T X \boldsymbol{b}$, where $\boldsymbol{a}$ is a $m \times 1$ vector while $\boldsymbol{b}$ is a $n \times 1$ vector. Find $\frac{\partial f}{\partial X}$*

Appendix B: Lagrange multiplier, KKT

B.1 Gradient