Optimization Theory and Algorithm	**Lecture 3 - 05/07/2021**

Lecture 3

Lecturer:Xiangyu Chang *Scribe: Xiangyu Chang*

Edited by: Xiangyu Chang

Example 1 *(Management Decision Tree Analysis)*

A management decision tree is a branched flowchart showing multiple pathways for potential decisions and outcomes.

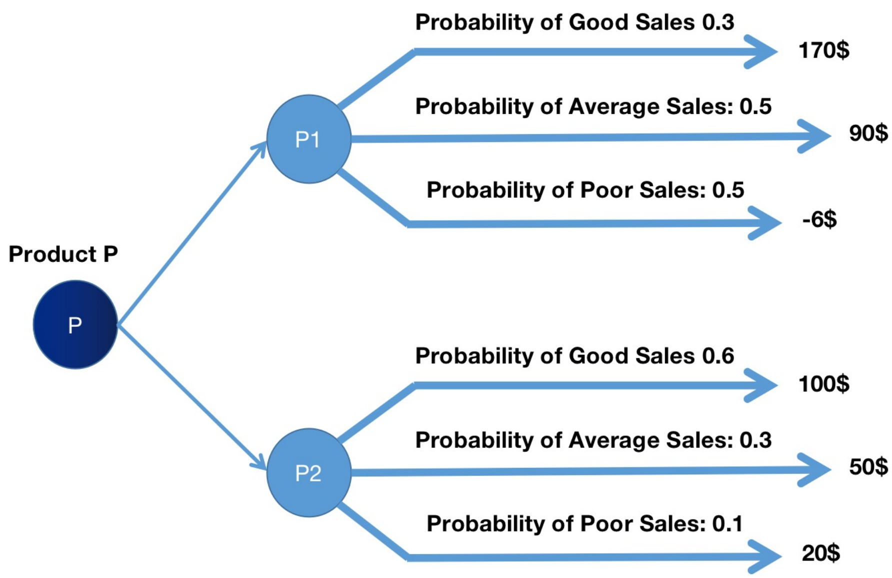

Figure 1: An example of Management Decision Tree

- *Suppose that a company is considering to develop a new product P. The product P includes two different types. The company employ a marketing research institute to study that which type is better?*
- *Based on the study results, the marketing research institute report: (1) If they produce the first type P1, then P1 has 0.3 chance for good sales with profit 170$ per unit; 0.5 chance for average sales with profit 90$ per unit; 0.2 chance for poor sales with -6$ per unit.*
- *Based on the study results, the marketing research institute report: (1) If they produce the first type P2, then P2 has 0.6 chance for good sales with profit 100$ per unit; 0.3 chance for average sales with profit 50$ per unit; 0.1 chance for poor sales with 30$ per unit.*
- **Q:** *which one is better? P1 or P2?*
- *For determining P1 or P2, management decision tree analysis is a commonly used method (see Figure 1).*
- *The main idea is to calculate the so-called* **expected reward** *as follows:*

$$I_1 = 170 \times 0.3 + 90 \times 0.5 - 6 \times 0.2 = 94.8,$$

and

$$I_2 = 100 \times 0.6 + 50 \times 0.3 + 20 \times 0.1 = 77.$$

- *So, $I_1 > I_2$, we need to choice P1.*

Example 1 is a signal step decision making problem. What about multiple step decision making problem?

Example 2 *(Markov Decision Processing and Reinforcement Learning)*

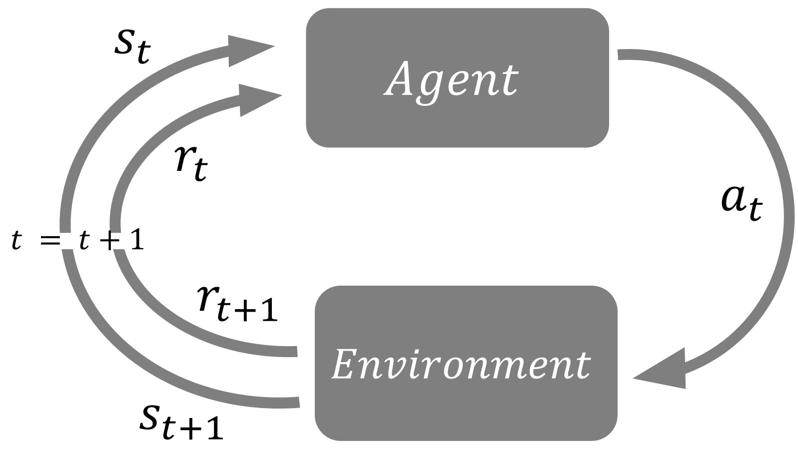

Figure 2: Markov Decision Processing

The above multiple decision making problem (See Figure 2) can be formalized as a Markov Decision Processing (MDP).

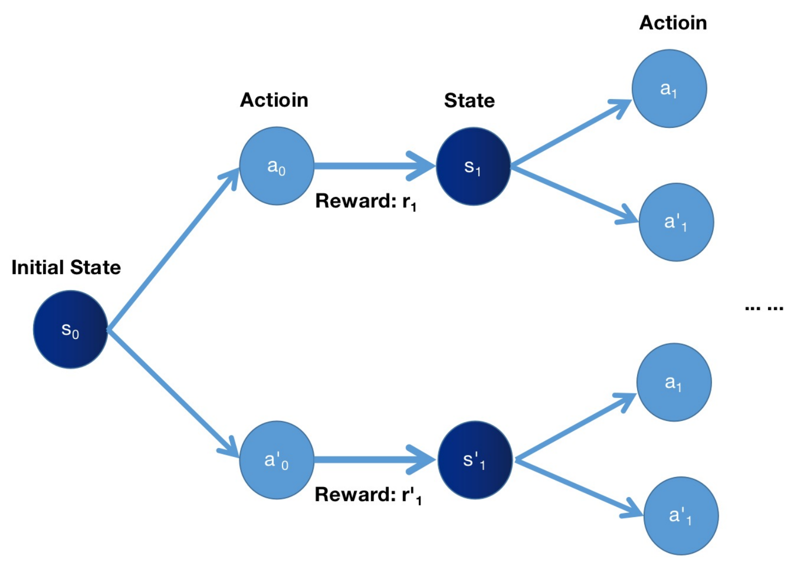

Figure 3: Trajectory of the Markov Decision Processing

- *State Space $\mathcal{S}$ is considered as a finite state space with cardinally $|\mathcal{S}|$.*

- *Action Space $\mathcal{A}$ is considered as a finite action space with cardinally $|\mathcal{A}|$.*
- *Transition Probability:*
$$\mathbb{P}(s_{t+1}|a_t, s_t, a_{t-1}, s_{t-1}, \dots, s_0) = \mathbb{P}(s_{t+1}|a_t, s_t). \tag{1}$$
- *Expected Reward:*
$$\mathbb{E}(r_t|a_t, s_t, a_{t-1}, s_{t-1}, \dots, s_0) = \mathbb{E}(r_{t+1}|a_t, s_t) = r(a_t, s_t). \tag{2}$$
- *Accumulated Reward:*
$$R(\tau) = \sum_{t=0}^{\infty} \gamma^t r(a_t, s_t) \tag{3}$$
where $\tau = (s_0, a_0, s_1, a_1, \dots,)$ is a trajectory (see Figure 3) and $0 < \gamma < 1$ is a discount factor.
- *Policy $\pi : s \in$ $mathcalS \to \Delta(\mathcal{A})$ and $a \sim \pi(a|s)$.*
- *Aim: Finding an optimal policy for maximizing the expected accumulated reward.*

Optimization Formulation:
$$\max_{\pi} \mathbf{E}_{\tau \sim \pi}[R(\tau)]. \tag{4}$$
Reinforcement Learning is commonly used method to solve the above optimization.

Classification of Optimization:

- Linear and Nonlinear Optimization
- Convex and Nonconvex Optimization
- Deterministic and Stochastic Optimization
- Constrained and Non-constrained Optimization
- Integer Program, Mixed Integer Program
- Robust Optimization
- QCQP,...

0.1 Algorithms in Optimization

Let us consider an optimization problem
$$\min_{x} \ f(\mathbf{x}), \tag{5}$$
$$\text{s.t. } \mathbf{x} \in \mathcal{X}, \tag{6}$$
where $f : \mathcal{X} \subset \mathbb{R}^n \to \mathbb{R}$, and further assume that $\mathbf{x}^*$ is the optimal global point or solution for it which is defined by Definition **??**.

An optimization algorithm is to design for pursuing the $\mathbf{x}^*$. However, usually it is not easy.

We consider the least squares problem in Example **??**,
$$\min_{\mathbf{x}} f(\mathbf{x}) = \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2. \tag{7}$$

Q: How to find x^* for the least squares problem?

Generally, I believe that you should known that to compute the derivative to obtain $f'(\mathbf{x})$ and set $f'(\mathbf{x}) = 0$. Then the solution of $f'(\mathbf{x}) = 0$ maybe the optimal solution of Eq.(7). What does it mean $f'(\mathbf{x})$ for a function defined on $\mathbb{R}^n$?

Definition 1 *A function* $f : \mathbb{R}^n \to \mathbb{R}$ *is Fŕechet-differential at* $\mathbf{x}$*, if there exists a vector* $g \in \mathbb{R}^n$ *such that*

$$\lim_{\Delta \mathbf{x} \to 0} \frac{f(\mathbf{x} + \Delta \mathbf{x}) - f(\mathbf{x}) - g^\top \Delta \mathbf{x}}{\|\Delta \mathbf{x}\|} = 0. \tag{8}$$

Then g *is called the gradient of* f *at* $\mathbf{x}$*, denoted as* $g := \nabla f(\mathbf{x})$*. If we further choose that* $\Delta \mathbf{x} = \epsilon e_i$*, and* $e_i = (0, \ldots, \underbrace{1}_{i^{th}\ position}, 0, \ldots, 0)^\top$*, then*

$$\nabla f(\mathbf{x}) = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_n} \right)^\top \in \mathbb{R}^n.$$

Definition 2 *We define the Hessian matrix of function* $f : \mathbb{R}^n \to \mathbb{R}$ *at point* $\mathbf{x}$ *is*

$$\begin{aligned} \nabla^2 f(\mathbf{x}) &= \left(\frac{\partial^2 f}{\partial x_i \partial x_j} \right) \in \mathbb{R}^{n^2} \\ &= \begin{pmatrix} \frac{\partial^2 f}{\partial^2 x_1} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial^2 x_n} \end{pmatrix}. \end{aligned}$$

Commonly, we assume that the Hessian matrix $\nabla^2 f(\mathbf{x})$ *is a symmetric matrix (actually need some regularity conditions).*

Definition 3 *Let* $f : \mathbb{R}^n \to \mathbb{R}^m$*, namely for any* $\mathbf{x} \in \mathbb{R}^n$*,* $f(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_m(\mathbf{x}))^\top \in \mathbb{R}^m$*, the Jacobi matrix is denoted as*

$$J(\mathbf{x}) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix} \in \mathbb{R}^{m \times n}.$$

Q: prove that the Jacobi matrix of gradient is the corresponding Hessian matrix.

Example 3 $f(\mathbf{x}) = \mathbf{a}^\top \mathbf{x}$*, then* $\nabla f(\mathbf{x}) = \mathbf{a}, \nabla^2 f(\mathbf{x}) = \mathbf{0} \in \mathbb{R}^{n^2}$*, why???*

Example 4 $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$.

Let us consider a general case, suppose that $G : \mathbb{R}^n \to \mathbb{R}$ *and* $G(\mathbf{z}) = g(z_1) + g(z_2) + \cdots + g(z_n)$ *and* $z_i = \mathbf{a}_i^\top \mathbf{x}$*, where* $\mathbf{z} = (z_1, \ldots, z_n)^\top$*. Let us derive that*

$$\begin{aligned} \frac{\partial G(A\mathbf{x})}{\partial \mathbf{x}} &= \frac{\partial (g(\mathbf{a}_1^\top \mathbf{x}) + g(\mathbf{a}_2^\top \mathbf{x}) + \cdots + g(\mathbf{a}_n^\top \mathbf{x}))}{\partial \mathbf{x}} && (9) \\ &= \sum_{i=1}^n \frac{\partial g(\mathbf{a}_i^\top \mathbf{x})}{\partial \mathbf{x}} = \sum_{i=1}^n \frac{\partial g(\mathbf{a}_i^\top \mathbf{x})}{\partial \mathbf{a}_i^\top \mathbf{x}} \times \frac{\partial \mathbf{a}_i^\top \mathbf{x}}{\partial \mathbf{x}} && (10) \\ &= \sum_{i=1}^n \frac{\partial g(\mathbf{a}_i^\top \mathbf{x})}{\partial \mathbf{a}_i^\top \mathbf{x}} \mathbf{a}_i && (11) \\ &= A^\top \nabla G(\mathbf{z}). && (12) \end{aligned}$$

Theorem 1 *(First-order Optimality Condition) Consider a non-constrained optimization problem* $\min_{\mathbf{x}} f(\mathbf{x})$*, where* $f : \mathbb{R}^n \to \mathbb{R}$ *and* $f \in C^1$*. If* $\mathbf{x}^*$ *is a local minimum, then*

$$\nabla f(\mathbf{x}^*) = 0.$$

The points which satisfy the equation $\nabla f(\mathbf{x}) = 0$ are called stationary points.

Theorem 2 *(Second-order Optimality Condition) Consider a non-constrained optimization problem $\min_{\mathbf{x}} f(\mathbf{x})$, where $f : \mathbb{R}^n \to \mathbb{R}$ and $f \in C^2$. If $\mathbf{x}^*$ is a local minimum, then*

$$\nabla f(\mathbf{x}^*) = 0 \text{ and } \nabla^2 f(\mathbf{x}^*) \geq 0,$$

where $\nabla^2 f(\mathbf{x}^*) \geq 0$ means the Hessian matrix is a positive semi-definite matrix.

Theorem 3 *(Sufficient Condition) Consider a non-constrained optimization problem $\min_{\mathbf{x}} f(\mathbf{x})$, where $f : \mathbb{R}^n \to \mathbb{R}$ and $f \in C^2$. If*

$$\nabla f(\mathbf{x}^*) = 0 \text{ and } \nabla^2 f(\mathbf{x}^*) > 0,$$

where $\nabla^2 f(\mathbf{x}^*) > 0$ means the Hessian matrix is a positive definite matrix. Then $\mathbf{x}^*$ is a local minimum of f.

These proofs can be found at Page 161-163 of the text book.

We go back to this example and further assume that $G(\mathbf{z}) = \frac{1}{2}\|\mathbf{z} - \mathbf{b}\|^2 = \frac{1}{2}\sum_{i=1}^{n}(z_i - b_i)^2, z_i = \mathbf{a}_i^\top \mathbf{x}$. Thus, $\nabla G(\mathbf{z}) = (z_1 - b_1, \dots, z_n - b_n)^\top$. Finally, based on Eq.(12),

$$\begin{aligned} \nabla f(\mathbf{x}) &= \frac{\partial G(\mathbf{z})}{\partial \mathbf{x}} = A^\top(\mathbf{z} - \mathbf{b}) \\ &= A^\top(A\mathbf{b} - \mathbf{b}) = A^\top A\mathbf{x} - A\mathbf{b} \end{aligned}$$

and

$$\nabla^2 f(\mathbf{x}) = A^\top A.$$

Recall the least squares problem (7), and set $\nabla f(\mathbf{x}) = \nabla \frac{1}{2}\|A\mathbf{x} - \mathbf{b}\|_2^2 = 0$, then we obtain the so-called *normal equation*:

$$A^\top A\mathbf{x} - A^\top \mathbf{b} = 0. \tag{13}$$

If $A^\top A$ is invertible, then $\mathbf{x}^* = (A^\top A)^{-1}A^\top \mathbf{b}$, which is called *a closed form solution.*

According to the definition of stationary point, we know that $\mathbf{x}^* = (A^\top A)^{-1}A^\top \mathbf{b}$ is a stationary point of the least squares problem. Furthermore, if $\nabla^2 f(\mathbf{x}) = A^\top A$ is a positive definite matrix (invertible), then $\mathbf{x}^* = (A^\top A)^{-1}A^\top \mathbf{b}$ is a local minimum according to Theorem 3.

The procedure of obtaining the closed form solution can be seen as an algorithm for solving the linear least squares problem.

References