Optimization Theory and Algorithm **Lecture 2 - 04/30/2021**

Lecture 2

Lecturer:Xiangyu Chang *Scribe: Xiangyu Chang*

Edited by: Xiangyu Chang

Example 1 *(Linear Regression)*

Let us consider a general case.

- *Data:* $\{\mathbf{a}_i, b_i\}_{i=1}^m$, *where* $\mathbf{a}_i \in \mathbb{R}^n, b_i \in \mathbb{R}$.
- *Suppose that*
$$b_i = \mathbf{a}_i^\top \mathbf{x} + \epsilon_i,$$
where $\mathbf{x} = (x_1, \ldots, x_n)^\top$ *is denoted as regression coefficient.*
- *Matrix Form: denote that* $\mathbf{b} = (b_1, \ldots, b_m)^\top \in \mathbb{R}^m$, $\mathbf{A} = (a_{ij}) = (\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m)^\top \in \mathbb{R}^{m\times n}$, *and*
$$\mathbf{b} = \mathbf{A}\mathbf{x} + \boldsymbol{\epsilon},$$
where $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_m)^\top$.

Optimization Formulation:

$$\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2, \tag{1}$$

where $\|\cdot\|_2$ *is the Euclidean norm (vector length), that is* $\|\mathbf{x}\|_2 := \sqrt{\sum_{i=1}^n x_i^2}$.

Example 2 *(Nonlinear Regression)*

- *Data:* $\{\mathbf{a}_i, b_i\}_{i=1}^m$, *where* $\mathbf{a}_i \in \mathbb{R}^n, b_i \in \mathbb{R}$.
- *Suppose that*
$$b_i = \Phi_{\mathbf{x}}(\mathbf{a}_i) + \epsilon_i,$$
where $\Phi_{\mathbf{x}} : \mathbb{R}^n \to \mathbb{R}$ *is a nonlinear function with the model parameter* $\mathbf{x} = (x_1, \ldots, x_n)^\top$.
- *Let us give you a concreted example of* $\Phi_{\mathbf{x}}$*:*
$$\Phi_{\mathbf{x}}(\mathbf{a}) = \frac{\exp(\mathbf{a}^\top \mathbf{x})}{1 + \exp(\mathbf{a}^\top \mathbf{x})}.$$

Optimization Formulation:

$$\min_{\mathbf{x}} \sum_{i=1}^m (b_i - \Phi_{\mathbf{x}}(\mathbf{a}_i))^2. \tag{2}$$

This is the so-called **Nonlinear Regression** or **Nonlinear Least Squares Method**.

Example 3 *(Deep Forward Neural Networks)*

Let us consider a very special nonlinear regression model that is the so-called deep forward neural networks.

- *Data:* $\{\mathbf{a}_i, b_i\}_{i=1}^m$, *where* $\mathbf{a}_i \in \mathbb{R}^n, b_i \in \mathbb{R}$.

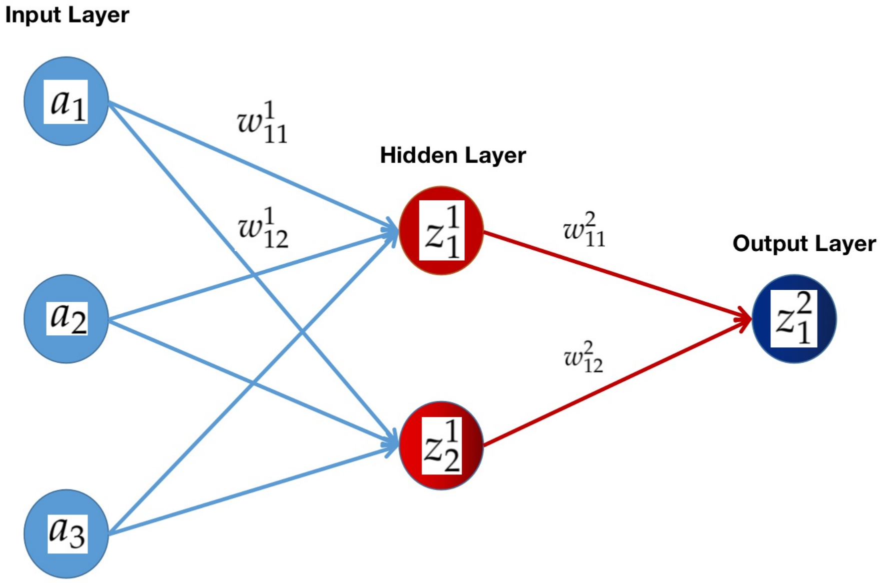

Figure 1: An example of Deep Forward Neural Networks

- *We suppose that any data point* $(\mathbf{a}, b)$ *is inputted into a deep neural network model with the structure in Figure 1.*
- *Then, from the input layer to the first hidden layer, we have*

$$y^1_1 = w^1_{11}a_1 + w^1_{12}a_2 + w^1_{13}a_3, \tag{3}$$
$$z^1_1 = f^1_1(y^1_1), \tag{4}$$
$$y^1_2 = w^1_{21}a_1 + w^1_{22}a_2 + w^1_{23}a_3, \tag{5}$$
$$z^1_2 = f^1_2(y^1_2), \tag{6}$$

where f^1_1 *and* f^1_2 *are the corresponding active function.*

- *From the first hidden layer to the output layer, we have*

$$y^2_1 = w^2_{11}z^1_1 + w^2_{12}z^1_2, \tag{7}$$
$$z^2_1 = f^2_1(y^2_1). \tag{8}$$

- *Integrate them together,*

$$\begin{aligned} z^2_1 &= f^2_1(y^2_1) && (9)\\ &= f^2_1(w^2_{11}z^1_1 + w^2_{12}z^1_2) && (10)\\ &= f^2_1(w^2_{11}f^1_1(y^1_1) + w^2_{12}f^1_2(y^1_2)) && (11)\\ &= F_2 \circ F_1(\mathbf{a}) := FD_W(\mathbf{a}), && (12)\end{aligned}$$

where $FD : \mathbb{R}^3 \to \mathbb{R}$ *is a composite function and* $W = \{w^1_{11}, w^1_{12}, \ldots, \}$ *includes all the parameters we have to estimate.*

Optimization Formulation:

$$\min_W \sum_{i=1}^m (b_i - FD_W(\mathbf{a}_i))^2. \tag{13}$$

This example indicates that the famous deep forward neural networks is a special case of nonlinear regression.

Example 4 *Generalized Linear Model (GLM). Let us consider the following three management problems.*

- $b = \textit{House Price} = F(a_1 = \textit{number of rooms}, a_2 = \textit{school distriction}, a_3, \dots)$
- $b = \textit{Credit Rate} = F(a_1 = \textit{education}, a_2 = \textit{salary}, a_3, \dots)$
- $b = \textit{Number of Visit this month} = F(a_1 = \textit{number of visit last month}, a_2 = \textit{RFM}, a_3, \dots)$

In this example, we introduced three classic regression models, linear regression(house price), Poisson regression (number of visit this month) and logistic regression (credit rate) derived from GLM. We parameterized the parameters in the statistic models as a linear function of covariant variables $\mathbf{a}$*, and formed the optimization problem from the likelihood.*

Consider the input-output pairs $\{\mathbf{a}_i, b_i\}_{i=1}^m$ *as the data. The procedure can be summarized as following recipe,*

1. *write down a probabilistic model for* b_i
2. *link model parameter* $\mathbf{x}$ *with* $\mathbf{a}_i$
3. *formed the optimization problem using maximum likelihood that aim to discover* $\mathbf{x}$ *with all data* $\{\mathbf{a}_i, b_i\}_{i=1}^m$

Next we instantiate this recipe by three examples.

(i) *Linear Regression: Given training data* $\{\mathbf{a}_i, b_i\}_{i=1}^m$ *with* $\mathbf{a}_i \in \mathbb{R}^n$ *and* $b_i \in \mathbb{R}$*. Suppose each* $b_i \overset{i.i.d.}{\sim} N(\mu_i, \sigma^2)$*, that is*

$$\begin{aligned} P(b_i|\mu_i, \sigma^2) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{(b_i - \mu_i)^2}{2\sigma^2}\} \\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{b_i^2}{2\sigma^2}\} \exp\{-\frac{\frac{1}{2}\mu_i^2 - b_i\mu_i}{\sigma^2}\}. \end{aligned}$$

It is convention to choose the parameters that multiply b_i *as the linear function of the variables* $\mathbf{a}_i$ *with the parametric coefficient* $\mathbf{x}$*. Here we make the assumption that*

$$\theta_i = \mu_i = \langle \mathbf{a}_i, \mathbf{x} \rangle.$$

We wish to examine how we find a good $\mathbf{x}$ *to make this work. Our strategy for this is to maximize the* likelihood *of all observations* $\{b_i\}$ *as a function of* $\mathbf{x}$*, i.e.*

$$\max_{\mathbf{x}} \prod_i \exp\{-\frac{1}{\sigma^2}(\frac{1}{2}\mu_i^2 - b_i\mu_i)\} \Rightarrow \max_{\mathbf{x}} \prod_i \exp\{-\frac{1}{2\sigma^2}(\langle \mathbf{a}_i, x \rangle^2 - b_i\langle \mathbf{a}_i, \mathbf{x} \rangle)\}.$$

To maximize this expression, we take the negative log of the expression, i.e. we want to minimize

$$\min_{\mathbf{x}} \frac{1}{\sigma^2} \sum_{i=1}^n (\frac{1}{2}\langle \mathbf{a}_i, \mathbf{x} \rangle^2 - b_i\langle \mathbf{a}_i, \mathbf{x} \rangle).$$

To write it more compactly, we denote,

$$A = \begin{bmatrix} \mathbf{a}_1^\top \\ \vdots \\ \mathbf{a}_m^\top \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix}.$$

And we have,

$$\sum_{i=1}^{m} \langle \mathbf{a}_i, x\rangle^2 = \|A\mathbf{x}\|^2, \qquad \sum_{i=1}^{m} b_i \langle \mathbf{a}_i, \mathbf{x}\rangle = \langle \mathbf{b}, A\mathbf{x}\rangle,$$

we get the minimization problem

$$\arg\min_x \ \frac{1}{2}\|A\mathbf{x}\|^2 - \langle b, A\mathbf{x}\rangle = \arg\min_x \ \frac{1}{2}\|A\mathbf{x} - \mathbf{b}\|^2$$

which is a linear least-squares regression problem.

(ii) Logistic Regression: Let $b_i \in \{0, 1\}$ and p_i be the probability of success, i.e.

$$\begin{aligned} p(b_i|p_i) &= p_i^{b_i}(1-p_i)^{1-b_i} \\ &= \exp\{b_i \ln p_i + (1-b_i)\ln(1-p_i)\} \\ &= \exp\{b_i(\ln p_i - \ln(1-p_i)) + \ln(1-p_i)\} \\ &= \exp\{b_i \ln \tfrac{p_i}{1-p_i} + \ln(1-p_i)\}. \end{aligned}$$

The choice $\theta_i = \ln \frac{p_i}{1-p_i}$ is called the canonical parameter, i.e. $p_i = \frac{\exp(\theta_i)}{1+\exp(\theta_i)} = (1+\exp(-\theta_i))^{-1}$. Letting $\theta_i = \langle a_i, x\rangle$ and noting that $\ln(1-p_i) = -\ln(1+\exp(\theta_i))$ the probability becomes

$$p(b_i|\theta_i) = \exp\{b_i\theta_i - \ln(1+\exp(\theta_i))\}.$$

The fitting problem is found by minimizing the negative log of the above expression,

$$\min_x \sum_{i=1}^{m} [\ln(1+\exp(\theta_i)) - b_i\theta_i] = \min_x \sum_{i=1}^{m} \ln(1+\exp(\langle \mathbf{a}_i, \mathbf{x}\rangle)) - \langle \mathbf{b}, A\mathbf{x}\rangle.$$

(iii) GLM: We find that given a family of distributions for b_i, given μ_i, σ^2 we have

$$f(b_i|\mu_i, \sigma^2) = g_1(b_i, \sigma^2)\exp\{\frac{b_i\mu_i - g_2(\mu_i)}{g_3(\sigma^2)}\}$$

for some functions g_1, g_2, g_3. And g_2 is given by

1. *$g_2(\mu_i) = \frac{1}{2}\mu_i^2$ for linear regression,*
2. *$g_2(\mu_i) = \exp(\mu_i)$ for Poisson regression,*
3. *$g_2(\mu_i) = \ln(1+\exp(\mu_i))$ for logistic regression.*

This gives us the problem

$$\min_{\mathbf{x}} \sum_{i=1}^{n} g_2(\langle \mathbf{a}_i, \mathbf{x}\rangle) - \langle \mathbf{b}, A\mathbf{x}\rangle.$$

The difficulty of this problem depends on properties of g_2. In these three cases g_2 is convex and smooth, but this won't always be the case. This motivates us to look at properties of continuous functions.

We will discuss basic function properties that will determine how good will an optimization algorithm perform on them.

Example 5 *(Portfolio Management)*

Portfolio Management (see Figure 2) is the art and science of making decisions about investment mix and policy, matching investments to objectives and balancing risk against performance.

Modeling:

- *n assets or stocks that are hold over a period of time.*

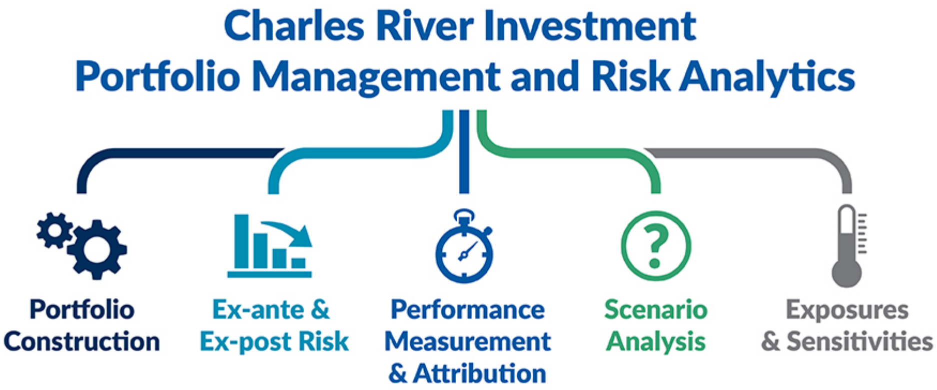
Charles River Investment
Portfolio Management and Risk Analytics
Portfolio Construction
Ex-ante & Ex-post Risk
Performance Measurement & Attribution
Scenario Analysis
Exposures & Sensitivities

Figure 2: An example of Portfolio Management

- *x_i denotes the amount of asset i, the final period.*
- *original price p_{i0} for asset i, the final price p_{it} at time t, then the return on asset i is $r_i = \frac{p_{it}-p_{i0}}{p_{i0}}$.*
- *the overall return is $R = \sum_{i=1}^{n} r_i x_i$.*
- *Suppose that $\mathbf{r} = (r_1, \ldots, r_n)^\top$ is a random vector with expectation of $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top$, and covariance of Σ.*
- *Aim: Finding a set of asset $\mathbf{x} = (x_1, \ldots, x_n)^\top$ to maximize the expected overall return and balancing the risk perform.*

Optimization Formulation:

$$\begin{aligned} \max_{\mathbf{x}} \ & \mathbb{E}(R) - \lambda Var(R), && (14) \\ s.t. \ & x_i \geq 0, i = 1 \ldots, n, && (15) \\ & \sum_{i=1}^{n} x_i = 1, && (16) \end{aligned}$$

where $\mathbb{E}(R)$ is the expectation of R, $Var(R)$ is the variance of R and $\lambda > 0$ is called risk aversion parameter for balancing the investment risk and expected return. Finally, we have that

$$\begin{aligned} \max_{\mathbf{x}} \ & \boldsymbol{\mu}^\top \mathbf{x} - \lambda \mathbf{x}^\top \Sigma \mathbf{x}, && (17) \\ s.t. \ & x_i \geq 0, i = 1 \ldots, n, && (18) \\ & \sum_{i=1}^{n} x_i = 1, && (19) \end{aligned}$$

Q: How to compute $\mathbb{E}(R)$ and $Var(R)$?

Q: Why not $x_i < 0$?

Remark 1 *This example is significantly important. Because*

- *This is called a* **nonlinear program** *due to the nonlinear objective function.*

- *It is also called a* **quadratic program**. *Why???*
- *Harry Markowiz proposed this model called* **Modern Portfolio Theory or Mean-Variance Analysis** *and obtained the* **Nobel Prize in 1990**.

References