Theory sheet 10

Linearity of gradient and Hessian and chain rule

Let $f, g : \mathbb{R}^n \to \mathbb{R}$ be some functions and let $H_f(\boldsymbol{x})$ denote its Hessian.

Since the derivative is linear ($(f+g)' = f' + g'$), we have the following rule, which simplifies finding gradients and Hessians:

$$\boxed{\begin{gathered}\operatorname{grad}(f(\boldsymbol{x}) + g(\boldsymbol{x})) = \operatorname{grad} f(\boldsymbol{x}) + \operatorname{grad} g(\boldsymbol{x}) \\ H_{f+g}(\boldsymbol{x}) = H_f(\boldsymbol{x}) + H_g(\boldsymbol{x})\end{gathered}}$$

If $h : \mathbb{R} \to \mathbb{R}$ is another function and we consider the composition $h(f(\boldsymbol{x}))$, then by the usual chain rule $(h(f(x)))' = h'(f(x))\, f'(x)$ we have that the same holds for the gradient:

$$\boxed{\operatorname{grad} h(f(\boldsymbol{x})) = h'(f(\boldsymbol{x}))\ \operatorname{grad} f(\boldsymbol{x}).}$$

These rules are simple, but very useful.

Example 1. $\operatorname{grad}(f(\boldsymbol{x}))^4 = 4f(\boldsymbol{x})^3\ \operatorname{grad} f(\boldsymbol{x})$.

Gradient and Hessian of a quadratic function

Consider $A \in M_{2,2}$ symmetric, $\boldsymbol{b} \in \mathbb{R}^2$ and the quadratic function

$$f(x_1, x_2) = \boldsymbol{x}^\top A \boldsymbol{x} + \boldsymbol{b}^\top \boldsymbol{x} + c = a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2 + b_1x_1 + b_2x_2 + c.$$

Then

$$\operatorname{grad} f(x_1, x_2) = \begin{pmatrix} f'_{x_1} \\ f'_{x_2} \end{pmatrix} = \begin{pmatrix} 2a_{11}x_1 + 2a_{12}x_2 + b_1 \\ 2a_{12}x_1 + 2a_{22}x_2 + b_2 \end{pmatrix}.$$

Let us note that this may be written as

$$\operatorname{grad} f(x_1, x_2) = 2\begin{pmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = 2A\boldsymbol{x} + \boldsymbol{b}.$$

As it turns out, the exact same formulas hold in higher dimensions:

Theorem 1. *If $A \in M_{n,n}$ and $\boldsymbol{b} \in \mathbb{R}^n$, we have*

$$\operatorname{grad} \boldsymbol{x}^\top A \boldsymbol{x} = 2A\boldsymbol{x} \quad \textit{and} \quad \operatorname{grad} \boldsymbol{b}^\top \boldsymbol{x} = \boldsymbol{b}.$$

Proof. We have

$$\operatorname{grad} \boldsymbol{x}^\top A\boldsymbol{x} = \operatorname{grad} \sum_{i,j=1}^{n} a_{ij} x_i x_j = \sum_{i,j=1}^{n} a_{ij} \operatorname{grad} x_i x_j,$$

so we only need to calculate $\operatorname{grad} x_i x_j$:

$$\operatorname{grad} x_i x_j = x_i\, \boldsymbol{e}_j + x_j\, \boldsymbol{e}_i,$$

where $\boldsymbol{e}_i$ is a vector of zeroes with 1 at i^{th} row. Hence,

$$\operatorname{grad} \boldsymbol{x}^\top A\boldsymbol{x} = \sum_{i,j=1}^{n} a_{ij}(x_i\, \boldsymbol{e}_j + x_j\, \boldsymbol{e}_i) = \sum_{i,j=1}^{n} a_{ij}\, x_i\, \boldsymbol{e}_j + \sum_{i,j=1}^{n} a_{ij}\, x_j\, \boldsymbol{e}_i.$$

Since $a_{ij} = a_{ji}$ by symmetry of A, we have

$$\sum_{i,j=1}^{n} a_{ji}\, x_i\, \boldsymbol{e}_j = \sum_{i,j=1}^{n} a_{ij}\, x_j\, \boldsymbol{e}_i = A\boldsymbol{x},$$

which concludes the proof of the first claim. The second claim follows similarly:

$$\operatorname{grad} \boldsymbol{b}^\top \boldsymbol{x} = \operatorname{grad} \sum_{i=1}^{n} b_i x_i = \sum_{i=1}^{n} b_i \operatorname{grad} x_i = \sum_{i=1}^{n} b_i \boldsymbol{e}_i = \boldsymbol{b},$$

where we used that $\operatorname{grad} x_i = \boldsymbol{e}_i$. □

Therefore,

$$\boxed{\operatorname{grad}(\boldsymbol{x}^\top A\boldsymbol{x} + \boldsymbol{b}^\top \boldsymbol{x} + c) = 2A\boldsymbol{x} + \boldsymbol{b}.}$$

We are going to use this later.

Remark 1. *Compare this with* $\frac{d}{dx}(ax^2 + bx + c) = 2ax + b$.

Next, we want to calculate the Hessian of the quadratic function. Since the Hessian is linear and the second derivatives of $\boldsymbol{b}^\top \boldsymbol{x} + c$ are clearly zero, we only need to find the Hessian of $\boldsymbol{x}^\top A\boldsymbol{x}$. A similar calculation gives

$$\boxed{H_f(\boldsymbol{x}) = 2A.}$$

Therefore, we know how to find both gradient and Hessian of a quadratic function!

Linear regression

Consider the following problem: we have some points x_i and y_i, $i = 1, \ldots, n$ obtained from some data. We plot these points and see that they all lie near some line. How do we find the best line that fits these points?

A line on the plane may be expressed by the following equation:

$$y = ax + b.$$

We cannot just assume that $y_i = ax_i + b$ for all $i = 1, \dots, n$ and try to solve for a, b, because this would be too many equations. That is, unless they all lie exactly on the same line to begin with.

What we can do instead is to assume that $y_i = ax_i + b + \varepsilon_i$, where ε_i are some errors. Now we can find a and b, which minimize the full error.

But what do we mean by full error? There are many ways to answer, the most standard is given by the sum of squared errors:

$$R = \sum_{i=1}^{n} \varepsilon_i^2.$$

Plugging $\varepsilon_i = y_i - ax_i - b$ into this formula, we see that R is a function of a and b:

$$\boxed{R(a, b) = \sum_{i=1}^{n} (y_i - ax_i - b)^2.}$$

We can now minimize this function with respect to a and b and obtain the so-called least square regression line.

Remark 2. *This is by far not the only way to define what "best line" means in this context. One equally interesting alternative is to minimize*

$$L(a, b) = \sum_{i=1}^{n} |\varepsilon_i| = \sum_{i=1}^{n} |y_i - ax_i - b|.$$

This is known as robust linear regression. Unfortunately, $|x|$ is non-differentiable, so minimization in this case becomes more involved.

To minimize R, we first compute its gradient:

$$\operatorname{grad} R(a, b) = \begin{pmatrix} R_a \\ R_b \end{pmatrix} = \begin{pmatrix} -2\sum_{i=1}^{n} (y_i - ax_i - b)x_i \\ -2\sum_{i=1}^{n} (y_i - ax_i - b) \end{pmatrix},$$

set it equal to zero and obtain two equations:

$$\sum_{i=1}^{n} y_i x_i - a\sum_{i=1}^{n} x_i^2 - b\sum_{i=1}^{n} x_i = 0 \quad \text{and} \quad \sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i - bn = 0.$$

Denoting

$$\rho = \frac{1}{n}\sum_{i=1}^{n} y_i x_i, \quad s^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2, \quad \mu_x = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \mu_y = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

we find that the equations above may be rewritten as

$$\begin{cases} \rho - as^2 - b\mu_x = 0, \\ \mu_y - a\mu_x - b = 0. \end{cases}$$

Solving this system, we obtain

$$\boxed{a = \frac{\rho - \mu_x \mu_y}{s^2 - \mu_x^2}, \quad b = \mu_y + a\mu_x.}$$

If we plug in ρ, μ_x, μ_y and s^2, we shall obtain

$$a = \frac{\frac{1}{n}\sum_{i=1}^{n} y_i x_i - \frac{1}{n^2}\sum_{i=1}^{n} x_i \sum_{j=1}^{n} y_j}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2}$$

We can also shuffle this formula a bit and rewrite it in yet another form:

$$a = \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{\sum_{i=1}^{n}(x_i - \mu_x)^2}.$$

Recall, however, that these are candidate points, and we need to check that they indeed minimize R. To this end, we compute the Hessian of R:

$$H_R(\boldsymbol{x}) = \begin{pmatrix} 2\sum_{i=1}^{n} x_i^2 & 2\sum_{i=1} x_i \\ 2\sum_{i=1}^{n} x_i & 2n \end{pmatrix}.$$

Clearly,

$$2\sum_{i=1}^{n} x_i^2 > 0,$$

so we need to check the determinant and apply Sylvester's law of intertia. We have

$$\begin{aligned} \det H_R(\boldsymbol{x}) &= 4n\sum_{i=1}^{n} x_i^2 - 4\left(\sum_{i=1} x_i\right)^2 \\ &= 4n^2\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n^2}\sum_{i=1}^{n} x_i\right)^2\right) \\ &= 4n\sum_{i=1}^{n}(x_i - \mu_x)^2 > 0. \end{aligned}$$

Remark 3. *Above, we used the following calculation:*

$$\begin{aligned} \sum_i (x_i - \mu_x)^2 &= \sum_i (x_i^2 - 2x_i\mu_x + \mu_x^2) \\ &= \sum_i x_i^2 - 2\mu_x \sum_i x_i + n\mu_x^2 \\ &= \sum_i x_i^2 - 2n\mu_x^2 + n\mu_x^2 \\ &= \sum_i x_i^2 - n\mu_x^2. \end{aligned}$$

Linear regression in matrix form

Let us solve the same problem using matrix notation. We will then see that this approach allows to do much more general kinds of regression!

Let y_i and x_i, $i = 1, \dots, n$ be our data points. Denote

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} b \\ a \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \boldsymbol{y} - X\boldsymbol{\beta}.$$

Note that X is a $n \times 2$ matrix. The same function R as above in this notation may be written as

$$R(\boldsymbol{\beta}) = \boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon}.$$

Plugging in the definition of $\boldsymbol{\varepsilon}$, we obtain

$$R(\boldsymbol{\beta}) = (\boldsymbol{y} - X\boldsymbol{\beta})^\top (\boldsymbol{y} - X\boldsymbol{\beta}) = \boldsymbol{y}^\top \boldsymbol{y} - \boldsymbol{\beta}^\top X^\top \boldsymbol{y} - \boldsymbol{y}^\top X\boldsymbol{\beta} + \boldsymbol{\beta}^\top X^\top X\boldsymbol{\beta}.$$

Since $\boldsymbol{y}^\top X\boldsymbol{\beta} = \boldsymbol{\beta}^\top X^\top \boldsymbol{y}$ (we can always transpose a number!), we have

$$R(\boldsymbol{\beta}) = \boldsymbol{\beta}^\top A\boldsymbol{\beta} - 2\,\boldsymbol{b}^\top \boldsymbol{\beta} + c,$$

where $A = X^\top X$, $\boldsymbol{b} = X^\top \boldsymbol{y}$ and $c = \boldsymbol{y}^\top \boldsymbol{y}$.

By the theorem in the beginning of this lecture, the gradient of R is given by

$$\operatorname{grad} R(\boldsymbol{\beta}) = 2A\boldsymbol{\beta} - 2\boldsymbol{b}.$$

Setting it equal zero, we find that $\boldsymbol{\beta}$ satisfies

$$A\boldsymbol{\beta} - \boldsymbol{b} = \boldsymbol{0}.$$

Therefore,

$$\boldsymbol{\beta} = A^{-1}\boldsymbol{b}.$$

It remains to plug in the formulas for A and $\boldsymbol{b}$ to obtain

$$\boxed{\boldsymbol{\beta} = (X^\top X)^{-1} X^\top \boldsymbol{y}.}$$

This formula is called the normal equation of linear regression, it gives the linear regression coefficients directly in terms of the original data points!

Remark 4. *Note that if $\boldsymbol{y} \in \mathbb{R}^n$ and $X \in M_{n,2}$, then $X^\top \in M_{2,n}$, hence $X^\top \boldsymbol{y} \in \mathbb{R}^2$ and $X^\top X \in M_{2,2}$, so all products are well-defined.*

The matrix $X^\top X$ is strictly positive definite, which means that it does not have zero eigenvalues, which means that it is invertible!

Remark 5 (Important)**.** *Note that we cannot simplify $(X^\top X)^{-1} X^\top$ using $(AB)^{-1} = B^{-1}A^{-1}$. Why? Because the last formula is only true if A and B are invertible. In particular, they must be square! Note that although $X^\top X$ is a square matrix (2×2), X and $X^\top$ are not (they are $n \times 2$ and $2 \times n$ correspondingly). Hence, their inverses do not even make sense!*

How to use the normal equation in practice?

Having derived the normal equation, we can forget about its derivation and just use it as follows:

- Given data points $x_i\,,\ y_i\,,\ i=1,\dots,n\,,$ construct a vector $\boldsymbol{y}$ and a matrix X as above.
- Calculate $\boldsymbol{\beta}=(X^\top X)^{-1}X^\top \boldsymbol{y}\,.$
- Now the best line approximating our data points is given by $y=\beta_1+\beta_2 x\,.$

Generalized linear regression

What if we want to find the best curve, instead of line, approximating our dataset $x_i\,,\ y_i\,?$ For example, we can find a polynomial curve approximating our data:

$$y_i \approx \beta_0+\beta_1 x_i+\beta_2 x_i^2+\cdots+\beta_p x_i^p$$

using the same idea! To this end, we build the approximation error

$$R(\boldsymbol{\beta})=\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i-\cdots-\beta_p x_i^p)^2$$

and minimize it with respect to $\boldsymbol{\beta}\,.$ This becomes much easier in the matrix notation. Let

$$\boldsymbol{y}=\begin{pmatrix} y_1\\ y_2\\ \vdots\\ y_n\end{pmatrix},\quad X=\begin{pmatrix} 1 & x_1 & x_1^2 & \dots & x_1^p\\ 1 & x_2 & x_2^2 & \dots & x_2^p\\ \vdots & \vdots & \vdots & \vdots & \\ 1 & x_n & x_n^2 & \dots & x_n^p\end{pmatrix},\quad \boldsymbol{\beta}=\begin{pmatrix}\beta_0\\ \beta_1\\ \vdots\\ \beta_p\end{pmatrix}.$$

With this notation, R becomes

$$R(\boldsymbol{\beta})=(\boldsymbol{y}-X\boldsymbol{\beta})^\top(\boldsymbol{y}-X\boldsymbol{\beta}).$$

Note that the problem is significantly more general (polynomial instead of line), but the function we need to optimize is exactly the same!

Therefore, we can immediately write down the solution of this minimization problem, that is, the normal equation:

$$\boldsymbol{\beta}=(X^\top X)^{-1}X^\top \boldsymbol{y}.$$

This equation now tells us that the best polynomial curve approximating our data is given by

$$y=\beta_0+\beta_1 x+\cdots+\beta_p x^p.$$

Remark 6. *In fact, there is nothing special about polynomials. We can approximate our data by a linear combination of any functions:*

$$y_i \approx \sum_{k=1}^{p} \beta_k \varphi_k(x_i).$$

To this end, we just change the definition of the matrix X to

$$X = \begin{pmatrix} \varphi_0(x_1) & \varphi_1(x_1) & \dots & \varphi_p(x_1) \\ \varphi_0(x_2) & \varphi_1(x_2) & \dots & \varphi_p(x_2) \\ \vdots & \vdots & \vdots & \vdots \\ \varphi_0(x_n) & \varphi_1(x_n) & \dots & \varphi_p(x_n) \end{pmatrix},$$

and use the same normal equation with this new matrix X. One neat example is to approximate seasonal data with periodic functions:

$$y_i \approx \beta_0 + \beta_1 \cos\left(\frac{x_i}{T_1}\right) + \beta_2 \cos\left(\frac{x_i}{T_2}\right),$$

where x_i is interpreted as time and T_1, T_2 are two different time scales.

Another natural extension

Another natural extension of the previous method arises when we want to model the effect of many factors x_i, z_i, t_i, $\dots$ on y_i. Geometrically, this means that we are trying to find the best hyperplane in approximating our points.

Let us assume that $y_i \approx \beta_0 + \beta_1 x_i + \beta_2 z_i$ (only two factors). Then the error of this approximation is

$$R(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i - \beta_2 z_i)^2,$$

which may again be written in the same way with the only difference that X is now

$$X = \begin{pmatrix} 1 & x_1 & z_1 \\ 1 & x_2 & z_2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & z_n \end{pmatrix}.$$

Everything else in the model remains the same, so the resulting $\boldsymbol{\beta}$ is given again by the normal equation.