COMP 7070 Advanced Topics in AI and ML January 22, 2025

Lecture 2: Numerical Analysis

Instructor: Yifan Chen *Scribes: Yifan Xu* *Proof reader: Yifan Chen, Xiong Peng*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*
Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

2.1 Matrix derivation

Matrix derivation refers to the process of computing the derivative of a matrix-valued function with respect to the input matrix, or the derivative of a scalar function to the input matrix. In this section, we study the latter with the matrix $\boldsymbol{X} \in \mathbb{R}^{m\times n}$, and the scalar function $f(\boldsymbol{X}) \in \mathbb{R}$. The derivative $\frac{\partial f}{\partial \boldsymbol{X}}$ can be defined using element-wise derivation:

$$\frac{\partial f}{\partial \boldsymbol{X}} = \left[\frac{\partial f}{\partial \boldsymbol{X}_{ij}}\right]_{m\times n}. \tag{2.1}$$

Computing element-wise derivatives for large matrices is complex and lacks elegance. We aim to interpret $f(\boldsymbol{X})$ as a function of the matrix $\boldsymbol{X}$ rather than as a function of its individual elements $\boldsymbol{X}_{ij}$. Therefore, to establish a matrix-centric perspective on matrix derivation, we first revisit the derivatives we have previously learned.
In univariate calculus, the derivative maps a scalar to a scalar, and the differential is expressed as:

$$\mathrm{d}f = f'(x)\mathrm{d}x,$$

where $\mathrm{d}f$ is the differential, $f'(x)$ is the derivative.
For multivariable calculus, the derivative maps a scalar to a vector, and the differential is defined as:

$$\mathrm{d}f = \sum_{i=1}^{n} \frac{\partial f}{\partial \boldsymbol{x}_i}\mathrm{d}\boldsymbol{x}_i = \frac{\partial f}{\partial \boldsymbol{x}}^T \mathrm{d}\boldsymbol{x},$$

where the total differential $\mathrm{d}f$ is the inner product of the gradient $\frac{\partial f}{\partial \boldsymbol{x}}$ and the differential vector $\mathrm{d}\boldsymbol{x}$.
We generalize the total differential concept from multivariate calculus to matrix calculus as follows:

$$\mathrm{d}f = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{\partial f}{\partial \boldsymbol{X}_{i,j}}\mathrm{d}\boldsymbol{X}_{i,j} = \mathrm{Tr}\left(\frac{\partial f}{\partial \boldsymbol{X}}^T \mathrm{d}\boldsymbol{X}\right) = \left\langle \frac{\partial f}{\partial \boldsymbol{X}}, \mathrm{d}\boldsymbol{X}\right\rangle, \tag{2.2}$$

where $\mathrm{Tr}(\cdot)$ represents matrix trace, which is the sum of the diagonal elements of a square matrix.
Note: Trace $\mathrm{Tr}(\cdot)$ satisfies the property that for matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m\times n}$, $\mathrm{Tr}(\boldsymbol{A}^T\boldsymbol{B}) = \sum_{i,j} \boldsymbol{A}_{ij}\boldsymbol{B}_{ij}$, i.e., $\mathrm{Tr}(\boldsymbol{A}^T\boldsymbol{B})$ is the inner product of matrices $\boldsymbol{A}$ and $\boldsymbol{B}$.
Now we can use differential to compute derivative, we first build rules for basic differential operations.

2.1.1 Differential formulas

1. $\mathrm{d}(\boldsymbol{X}+\boldsymbol{Y}) = \mathrm{d}\boldsymbol{X} + \mathrm{d}\boldsymbol{Y}$ (Addition)
2. $\mathrm{d}(\boldsymbol{X}\boldsymbol{Y}) = \mathrm{d}\boldsymbol{X}\cdot\boldsymbol{Y} + \boldsymbol{X}\cdot\mathrm{d}\boldsymbol{Y}$ (Multiplication)
3. $\mathrm{d}\boldsymbol{X}^{-1} = -\boldsymbol{X}^{-1}\mathrm{d}\boldsymbol{X}\boldsymbol{X}^{-1}$ (Inverse)
   This formula can be proven using $\mathrm{d}\boldsymbol{X}\boldsymbol{X}^{-1} = \mathrm{d}\boldsymbol{I}$.

4. $\mathrm{d}(\boldsymbol{X} \odot \boldsymbol{Y}) = \mathrm{d}\boldsymbol{X} \odot \boldsymbol{Y} + \boldsymbol{X} \odot \mathrm{d}\boldsymbol{Y}$ (Element-wise multiplication),
   where $\odot$ represents element-wise multiplication of matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ of the same size.

5. $\mathrm{d}\sigma(\boldsymbol{X}) = \sigma'(\boldsymbol{X}) \odot \mathrm{d}\boldsymbol{X}, \quad \sigma(\boldsymbol{X}) = [\sigma(\boldsymbol{X}_{ij})]$ (Element-wise function),
   where $\sigma(\boldsymbol{X}) = [\sigma(\boldsymbol{X}_{ij})]$ represents element-wise function, $\sigma'(\boldsymbol{X}) = [\sigma'(\boldsymbol{X}_{ij})]$ represents element-wise derivative.
   E.g., for matrix $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_{11} & \boldsymbol{X}_{12} \\ \boldsymbol{X}_{21} & \boldsymbol{X}_{22} \end{bmatrix}$,

$$\mathrm{d}\sin(\boldsymbol{X}) = \mathrm{d}\begin{bmatrix} \sin \boldsymbol{X}_{11} & \sin \boldsymbol{X}_{12} \\ \sin \boldsymbol{X}_{21} & \sin \boldsymbol{X}_{22} \end{bmatrix} = \begin{bmatrix} \cos \boldsymbol{X}_{11}\mathrm{d}\boldsymbol{X}_{11} & \cos \boldsymbol{X}_{12}\mathrm{d}\boldsymbol{X}_{12} \\ \cos \boldsymbol{X}_{21}\mathrm{d}\boldsymbol{X}_{21} & \cos \boldsymbol{X}_{22}\mathrm{d}\boldsymbol{X}_{22} \end{bmatrix} = \cos(\boldsymbol{X}) \odot \mathrm{d}\boldsymbol{X}.$$

If the scalar function $f(\boldsymbol{X})$ is constructed through operations such as addition, subtraction, multiplication, inversion, or element-wise functions on the matrix $\boldsymbol{X}$, we can apply the above formulas to express the differential $\mathrm{d}f$ in the form presented in Equation (2.2). To achieve this, we need some trace tricks.

2.1.2 Trace tricks

1. If $\boldsymbol{a} \in \mathbb{R}^{n\times 1}, \boldsymbol{B} \in \mathbb{R}^{n\times n}$,
$$\boldsymbol{a}^T\boldsymbol{B}\boldsymbol{a} = \mathrm{Tr}(\boldsymbol{a}^T\boldsymbol{B}\boldsymbol{a}) = \mathrm{Tr}(\boldsymbol{a}\boldsymbol{a}^T\boldsymbol{B}). \tag{2.3}$$
   Sketch of proof: $\boldsymbol{a}^T\boldsymbol{B}\boldsymbol{a} = \sum_{j=1}^n a_{j1}\sum_{i=1}^n a_{i1}b_{ij} = \mathrm{Tr}(\boldsymbol{a}^T\boldsymbol{B}\boldsymbol{a}) = \mathrm{Tr}(\boldsymbol{a}\boldsymbol{a}^T\boldsymbol{B})$.

2. If $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C} \in \mathbb{R}^{m\times m}$,
$$\mathrm{Tr}\left(\boldsymbol{A}^T(\boldsymbol{B} \odot \boldsymbol{C})\right) = \mathrm{Tr}\left[(\boldsymbol{A} \odot \boldsymbol{B})^T\boldsymbol{C}\right]. \tag{2.4}$$
   Sketch of proof: $\mathrm{Tr}\left(\boldsymbol{A}^T(\boldsymbol{B} \odot \boldsymbol{C})\right) = \sum_{i=1}^m\sum_{j=1}^m a_{ij}b_{ij}c_{ij} = \mathrm{Tr}\left[(\boldsymbol{A} \odot \boldsymbol{B})^T\boldsymbol{C}\right]$.

Now the basic operation rules are prepared, to compute complex function derivative, we have one more topic to cover – composite function derivative.

2.1.3 Composite function derivative

If $\boldsymbol{Y}$ is a function of $\boldsymbol{X}$ and $\frac{\partial f}{\partial \boldsymbol{Y}}$ is known, we want to compute $\frac{\partial f}{\partial \boldsymbol{X}}$ using composite function derivative. In univariate calculus, we use the chain rule to compute $\frac{\partial f}{\partial x}$. But in matrix derivation, the derivative between two matrices $\frac{\partial \boldsymbol{Y}}{\partial \boldsymbol{X}}$ is undefined yet. However, we can still use the same differential operation rules to transform $\mathrm{d}\boldsymbol{Y}$ into $\mathrm{d}\boldsymbol{X}$. In this way, it is natural to derive the derivative $\frac{\partial f}{\partial \boldsymbol{X}}$. For example, if $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X}\boldsymbol{B}$, we get for $\mathrm{d}f$,

$$\mathrm{d}f = \mathrm{Tr}\left[\frac{\partial f}{\partial \boldsymbol{Y}}^T \mathrm{d}\boldsymbol{Y}\right] = \mathrm{Tr}\left[\frac{\partial f}{\partial \boldsymbol{Y}}^T \boldsymbol{A}\mathrm{d}\boldsymbol{X}\boldsymbol{B}\right] = \mathrm{Tr}\left[\boldsymbol{B}\frac{\partial f}{\partial \boldsymbol{Y}}^T \boldsymbol{A}\mathrm{d}\boldsymbol{X}\right].$$

According to the form of Equation (2.2), we obtain the derivative of f to $\boldsymbol{X}$ as:

$$\frac{\partial f}{\partial \boldsymbol{X}} = \boldsymbol{A}^T\frac{\partial f}{\partial \boldsymbol{Y}}\boldsymbol{B}^T.$$

Next, we take the above methods into practice.

2.1.4 Example: logistic regression

In logistic regression, $\boldsymbol{y} \in \{0,1\}^k$ is a one-hot vector acting as the label for input $\boldsymbol{x} \in \mathbb{R}^n$, the weight matrix is $\boldsymbol{W} \in \mathbb{R}^{k\times n}$. We define a probability vector $\boldsymbol{p} \in \mathbb{R}^k$, with p_i representing the probability of $\boldsymbol{x}$ belonging to category i. The likelihood of the maximum likelihood estimation in $\boldsymbol{p}$ can be expressed as:

$$\mathcal{L} = \max_{p_i}\prod_{i=1}^k p_i^{y_i},$$

where y_i is the i-th element of $\boldsymbol{y}$, p_i is the i-th element of $\boldsymbol{p}$.
Next, we want to transform $\prod$ into $\sum$ using the log trick:

$$-\log \mathcal{L} = \min_{p_i} \left(-\sum_{i=1}^{k} y_i \log p_i \right),$$

where log represents the natural logarithm.
Therefore, we define the loss function of logistic regression as:

$$l(\boldsymbol{x}; \boldsymbol{W}) = -\boldsymbol{y}^T \log \underbrace{\operatorname{softmax}(\boldsymbol{W}\boldsymbol{x})}_{\boldsymbol{p}}. \tag{2.5}$$

To optimize l, we need to compute the derivative of l to $\boldsymbol{W}$. To simplify notations, we can view $\boldsymbol{W}\boldsymbol{x}$ as a new variable $\boldsymbol{a}$, and Equation (2.5) transforms to:

$$l(\boldsymbol{x}; \boldsymbol{W}) = -\left(\log \operatorname{softmax}(\boldsymbol{x}^T \boldsymbol{W}^T)\right) \boldsymbol{y} = -\left(\log \operatorname{softmax}(\boldsymbol{a}^T)\right) \boldsymbol{y},$$

recall that $\operatorname{softmax}(\boldsymbol{a}) = \frac{\exp(\boldsymbol{a})}{\mathbf{1}_k^T \exp(\boldsymbol{a})}$, where $\mathbf{1}_k$ is a k-dimensional all-ones vector, then we get for $l(\boldsymbol{x}; \boldsymbol{W})$,

$$\begin{aligned}
l(\boldsymbol{x}; \boldsymbol{W}) &= -\log \left[\frac{\exp(\boldsymbol{a}^T)}{\exp(\boldsymbol{a}^T)\mathbf{1}_k} \right] \boldsymbol{y} \\
&= -\log \left[\exp(\boldsymbol{a}^T)\right] \boldsymbol{y} + \log \left[\exp(\boldsymbol{a}^T)\mathbf{1}_k\right] \mathbf{1}_k^T \boldsymbol{y} && \log(\boldsymbol{u}/c) = \log(\boldsymbol{u}) - \mathbf{1}\log(c) \\
&= -\boldsymbol{y}^T \boldsymbol{a} + \log \left[\exp(\boldsymbol{a}^T)\mathbf{1}_k\right]. && \boldsymbol{y}^T \mathbf{1} = 1
\end{aligned}$$

Then, we differentiate both sides of the equation on $\boldsymbol{a}$,

$$\begin{aligned}
\mathrm{d}l &= -\boldsymbol{y}^T \mathrm{d}\boldsymbol{a} + \frac{1}{\exp(\boldsymbol{a}^T)\mathbf{1}_k} \left[\mathrm{d}\exp(\boldsymbol{a}^T)\right] \mathbf{1}_k \\
&= -\boldsymbol{y}^T \mathrm{d}\boldsymbol{a} + \frac{1}{\exp(\boldsymbol{a}^T)\mathbf{1}_k} \left[\exp(\boldsymbol{a}^T) \odot \mathrm{d}\boldsymbol{a}^T\right] \mathbf{1}_k. && \mathrm{d}\sigma(\boldsymbol{a}) = \sigma'(\boldsymbol{a}) \odot \mathrm{d}\boldsymbol{a}
\end{aligned}$$

According to Equation (2.2), we apply the trace operator to both sides of the equation,

$$\begin{aligned}
\mathrm{d}l &= \operatorname{Tr}\left(-\boldsymbol{y}^T \mathrm{d}\boldsymbol{a} + \frac{1}{\exp(\boldsymbol{a}^T)\mathbf{1}_k} \exp(\boldsymbol{a}^T)(\mathrm{d}\boldsymbol{a} \odot \mathbf{1}_k) \right) \\
&= \operatorname{Tr}\left(-\boldsymbol{y}^T \mathrm{d}\boldsymbol{a} + \frac{\exp(\boldsymbol{a}^T)}{\exp(\boldsymbol{a}^T)\mathbf{1}_k} \mathrm{d}\boldsymbol{a} \right) \\
&= \operatorname{Tr}\left(-\left[\boldsymbol{y}^T + \operatorname{softmax}(\boldsymbol{a}^T)\right] \mathrm{d}\boldsymbol{a} \right).
\end{aligned}$$

Therefore,

$$\frac{\partial l}{\partial \boldsymbol{a}} = -\boldsymbol{y} + \operatorname{softmax}(\boldsymbol{a}).$$

Then we apply composite function derivative rules on $\boldsymbol{a}$,

$$\mathrm{d}l = \operatorname{Tr}\left(\frac{\partial l}{\partial \boldsymbol{a}}^T \mathrm{d}\boldsymbol{a} \right) = \operatorname{Tr}\left(\frac{\partial l}{\partial \boldsymbol{a}}^T \mathrm{d}\boldsymbol{W}\boldsymbol{x} \right) = \operatorname{Tr}\left(\boldsymbol{x} \frac{\partial l}{\partial \boldsymbol{a}}^T \mathrm{d}\boldsymbol{W} \right).$$

Therefore,

$$\frac{\partial l}{\partial \boldsymbol{W}} = \frac{\partial l}{\partial \boldsymbol{a}} \boldsymbol{x}^T = -\boldsymbol{y}\boldsymbol{x}^T + \operatorname{softmax}(\boldsymbol{a})\boldsymbol{x}^T.$$

2.2 Numerical analysis

2.2.1 Norm

Norm maps a vector into a scalar "magnitude": $f(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}_0^+$, often written as $\|\boldsymbol{x}\|$. A function $\|\boldsymbol{x}\| : \mathbb{R}^n \to \mathbb{R}_0^+$ is called a norm if and only if it satisfies the following conditions:

1. $\|\boldsymbol{x}\| = 0 \iff \boldsymbol{x} = 0$
2. $\|\alpha\boldsymbol{x}\| = |\alpha|\|\boldsymbol{x}\|$
3. $\|\boldsymbol{x}\| \geq 0$
4. $\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$

A specific norm is determined with a parameter p, referred to as p-norm. If we have $\boldsymbol{x} \in \mathbb{R}^{n\times 1}$, the p-norm of $\boldsymbol{x}$ is defined as:

$$\|\boldsymbol{x}\|_p^p \coloneqq \sum_i^n |x_i|^p, \tag{2.6}$$

when $p = \infty$,

$$\|\boldsymbol{x}\|_\infty = \max_i |x_i|. \tag{2.7}$$

The ∞-norm of a vector is the maximum absolute value of its elements.
when $p = 0$,

$$\|\boldsymbol{x}\|_0 = \sum_{i=1}^n \mathbf{1}\{x_i \neq 0\}, \tag{2.8}$$

where $\mathbf{1}\{\cdot\}$ is an indicator function. The 0-norm counts the number of non-zero elements in the vector.
Further, we discuss **matrix norm**. We begin with the Frobenius norm, if we have $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, the Frobenius norm of $\boldsymbol{A}$ is:

$$\|\boldsymbol{A}\|_F^2 = \sum_{j=1}^n \sum_{i=1}^m a_{ij}^2 = \|\mathrm{Vec}(\boldsymbol{A})\|^2, \tag{2.9}$$

where the $m \times n$ matrix $\boldsymbol{A}$ can be viewed as the vector obtained by concatenating together the columns of $\boldsymbol{A}$, and the Frobenius norm can be viewed as applying the 2-norm on this new vector.
Next we introduce the operator norm. If $\mathbf{X}$ and $\mathbf{Y}$ are two vector spaces with norm $\|\boldsymbol{x}\|_p$ and $\|\boldsymbol{y}\|_q$, respectively. $\boldsymbol{A}$ is the matrix that maps $\mathbf{X}$ to $\mathbf{Y}$, $\boldsymbol{A} : \mathbf{X} \to \mathbf{Y}$. Operator norm $\|\boldsymbol{A}\|_{pq}$ is induced by vector norm:

$$\|\boldsymbol{A}\|_{pq} \coloneqq \inf \{C \geq 0 \mid \|\boldsymbol{A}\boldsymbol{x}\|_q \leq C\|\boldsymbol{x}\|_p, \forall \boldsymbol{x} \in \mathbf{X}\}. \tag{2.10}$$

In this definition, $\|\boldsymbol{A}\|_{pq}$ is the maximum scaling factor that transforms the norm of vector $\boldsymbol{x}$ in space $\mathbf{X}$ to the norm of $\boldsymbol{A}\boldsymbol{x}$ in space $\mathbf{Y}$. The relative scaling effect of $\boldsymbol{A}$ on $\boldsymbol{x}$ is not influenced by the norm of $\boldsymbol{x}$. Therefore, if we simply consider the situation where $\|\boldsymbol{x}\|_p = 1$, we can get for $\|\boldsymbol{A}\|_{pq}$,

$$\|\boldsymbol{A}\|_{pq} = \max_{\|\boldsymbol{x}\|_p=1} \|\boldsymbol{A}\boldsymbol{x}\|_q. \tag{2.11}$$

Taking $p = q = 2$, we have the following inequality,

$$\|\boldsymbol{A}\boldsymbol{x}\|_2 \leq \|\boldsymbol{A}\|_2\|\boldsymbol{x}\|_2. \quad \forall \boldsymbol{x} \in \mathbf{X} \tag{2.12}$$

On the unit sphere in the vector space, the norm of $\boldsymbol{x}$ equals 1,

$$\|\boldsymbol{A}\boldsymbol{x}\|_2 \leq \|\boldsymbol{A}\|_2, \quad \forall \|\boldsymbol{x}\|_2 = 1. \quad \boldsymbol{x} \in \mathbf{X}$$

2.2.2 Conditioning

Conditioning refers to a measure of sensitivity of a function's output to input perturbations, often affecting the numerical stability and accuracy of computations. Relative condition number is defined as the maximum ratio of the relative error in the output of a function to the relative perturbation in the input. If we have an input vector $\boldsymbol{x} \in \mathbb{R}^n$ and a perturbation vector $\boldsymbol{h} \in \mathbb{R}^n$, we give the definition of condition number on function $f(\cdot)$ of $\boldsymbol{x}$ as:

$$\begin{aligned} \kappa(f; \boldsymbol{x}, \boldsymbol{h}) &= \frac{|f(\boldsymbol{x}+\boldsymbol{h}) - f(\boldsymbol{x})| \,/\, |f(\boldsymbol{x})|}{\|\boldsymbol{h}\|/\|\boldsymbol{x}\|} \\ \kappa(f) &\coloneqq \lim_{\epsilon \to 0} \max_{\boldsymbol{x}, \|\boldsymbol{h}\| \leq \epsilon \|\boldsymbol{x}\|} \kappa(f; \boldsymbol{x}, \boldsymbol{h}), \end{aligned} \tag{2.13}$$

where the norm of $\boldsymbol{h}$ is controlled by $\|\boldsymbol{x}\|$. Concisely, we will simply refer to the relative condition number as the condition number in the following analysis. Consider matrix transformation of $\boldsymbol{x}$, if $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{y} = f(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x}$, then we have:

$$\begin{cases} \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} \\ \boldsymbol{y} + \delta\boldsymbol{y} = \boldsymbol{A}(\boldsymbol{x} + \delta\boldsymbol{x}). \end{cases}$$

Taking the norm of $\delta\boldsymbol{y}$, we have,

$$\|\delta\boldsymbol{y}\| = \|\boldsymbol{A}\delta\boldsymbol{x}\| \leq \|\boldsymbol{A}\|\|\delta\boldsymbol{x}\|.$$

We consider three cases,
– If $\boldsymbol{A}$ is a square matrix and the inverse of $\boldsymbol{A}$ exists, we have

$$\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{y} \Rightarrow \|\boldsymbol{x}\| \leq \|\boldsymbol{A}^{-1}\|\|\boldsymbol{y}\| \Rightarrow \frac{1}{\|\boldsymbol{y}\|} \leq \|\boldsymbol{A}^{-1}\|\frac{1}{\|\boldsymbol{x}\|}.$$

Multiplying this inequality with the above inequality of $\|\delta\boldsymbol{y}\|$, we get,

$$\frac{\|\delta\boldsymbol{y}\|}{\|\boldsymbol{y}\|} \leq \|\boldsymbol{A}\|\|\boldsymbol{A}^{-1}\|\frac{\|\delta\boldsymbol{x}\|}{\|\boldsymbol{x}\|}.$$

Based on Equation (2.13), we can compute the condition number of matrix $\boldsymbol{A}$ as:

$$\kappa(f) = \kappa(\boldsymbol{A}) = \lim_{\delta\boldsymbol{x} \to 0} \max_{\boldsymbol{x}, \delta\boldsymbol{x}} \frac{\|\delta\boldsymbol{y}\|/\|\boldsymbol{y}\|}{\|\delta\boldsymbol{x}\|/\|\boldsymbol{x}\|} = \|\boldsymbol{A}\|\|\boldsymbol{A}^{-1}\|. \tag{2.14}$$

– If $m < n$, consider the situation that $\boldsymbol{x} \perp \boldsymbol{A}$, which means that the n-dim vector $\boldsymbol{x}$ is perpendicular to m row vectors in $\boldsymbol{A}$. In this case, $\|\boldsymbol{y}\| = 0$, and the condition number is:

$$\kappa(\boldsymbol{A}) = \lim_{\delta\boldsymbol{x} \to 0} \max_{\boldsymbol{x}, \delta\boldsymbol{x}} \frac{\|\delta\boldsymbol{y}\|/\|\boldsymbol{y}\|}{\|\delta\boldsymbol{x}\|/\|\boldsymbol{x}\|} = \infty. \tag{2.15}$$

– If $m > n$, then

$$\boldsymbol{x} = \boldsymbol{A}^{+}\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}^{+}\boldsymbol{y} \Rightarrow \|\boldsymbol{x}\| = \|\boldsymbol{A}^{+}\boldsymbol{y}\| \leq \|\boldsymbol{A}^{+}\|\|\boldsymbol{y}\|,$$

where $\boldsymbol{A}^{+}\boldsymbol{A} = \boldsymbol{I}$, and $\boldsymbol{A}^{+}$ is the pseudo-inverse matrix of $\boldsymbol{A}$.

$$\kappa(\boldsymbol{A}) = \lim_{\delta\boldsymbol{x} \to 0} \max_{\boldsymbol{x}, \delta\boldsymbol{x}} \frac{\|\delta\boldsymbol{y}\|/\|\boldsymbol{y}\|}{\|\delta\boldsymbol{x}\|/\|\boldsymbol{x}\|} = \|\boldsymbol{A}\|\|\boldsymbol{A}^{+}\|. \tag{2.16}$$

To compute $\boldsymbol{A}^{+}$, we can use singular value decomposition (SVD) on $\boldsymbol{A}$.
Intuitively, if $\boldsymbol{A}$'s rank $r = n$ and $\boldsymbol{A}$ is a square matrix, the equation $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ has only one solution, and the condition number can be expressed using $\boldsymbol{A}^{-1}$. If $r < n$, we refer to the equation $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ as underdetermined, there are infinite solutions for this equation. If $r = n$ and $\boldsymbol{A}$ is not a square matrix, we refer to the equation $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ as overdetermined, there's no solution to the equation, but we can use the least square method to compute the approximate solution.

2.3 Orthogonal matrices

Orthogonal matrices are square matrices whose rows and columns are orthonormal vectors, the transpose of an orthogonal matrix equals its inverse, we define orthogonal matrices as:

$$\boldsymbol{Q}^T \equiv \boldsymbol{Q}^{-1}. \tag{2.17}$$

We can compute the norm of an orthogonal matrix:

$$\begin{aligned} \|\boldsymbol{Q}\|^2 &= \max_{\|\boldsymbol{x}\|=1} \|\boldsymbol{Q}\boldsymbol{x}\|^2 \\ &= \max_{\|\boldsymbol{x}\|=1} \boldsymbol{x}^T\boldsymbol{Q}^T\boldsymbol{Q}\boldsymbol{x} = 1, \end{aligned} \tag{2.18}$$

where $\boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{I}$.
Similarly, we can derive the norm of the inverse of an orthogonal matrix:

$$\|\boldsymbol{Q}^{-1}\| = \max_{\|\boldsymbol{x}\|=1} \|\boldsymbol{Q}^{-1}\boldsymbol{x}\| = \max_{\|\boldsymbol{x}\|=1} \|\boldsymbol{Q}^T\boldsymbol{x}\| = \max \boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{x} = 1. \tag{2.19}$$

2.4 Singular value decomposition

SVD factorizes any matrix into three matrices consisting of two orthogonal matrices and a diagonal matrix of singular values. For matrix $\boldsymbol{A} \in \mathbb{R}^{n\times m}$,

$$\boldsymbol{A} = \boldsymbol{U}\Sigma\boldsymbol{V}^T, \tag{2.20}$$

where $\boldsymbol{U}$ and $\boldsymbol{V}$ are two orthogonal matrices. The columns of $\boldsymbol{U}$ are the left singular vectors of $\boldsymbol{A}$, the columns of $\boldsymbol{V}$ are the right singular vectors of $\boldsymbol{A}$, and Σ is a diagonal matrix whose diagonal elements are the singular values of matrix $\boldsymbol{A}$.
The rank of $\boldsymbol{A}$ satisfies $r \leq \min(n, m)$, then

$$\boldsymbol{A} = \boldsymbol{U}_{n\times r}\Sigma_{r\times r}(\boldsymbol{V}^T)_{r\times m} = \sum_{i=1}^{r} s_i\boldsymbol{u}_i\boldsymbol{v}_i^T, \tag{2.21}$$

where s_i is the i-th element in the diagonal of Σ, also the i-th singular value of $\boldsymbol{A}$, $\boldsymbol{u}_i$ is the i-th column vector in $\boldsymbol{U}$ and $\boldsymbol{v}_i$ is the i-th column vector in $\boldsymbol{V}$.
This equation indicates that a matrix is the summation of the multiplication of its singular values and corresponding singular vectors. In some cases, we only need the first (maximum) k singular values and singular vectors to express $\boldsymbol{A}$, so the truncated SVD can be expressed as:

$$\tilde{\boldsymbol{A}} = \sum_{i=1}^{k} s_i\boldsymbol{u}_i\boldsymbol{v}_i^T. \tag{2.22}$$

Next, we examine the 2-norm of $\boldsymbol{A}$ from SVD perspective (Note: We omit the subscript 2 of the 2-norm in the derivation to keep the proof concise.),

$$\|\boldsymbol{A}\| \leq \|\boldsymbol{U}\|\|\Sigma\|\|\boldsymbol{V}^T\| = \|\Sigma\| = \sigma_{\max},$$

where $\|\boldsymbol{U}\| = \|\boldsymbol{V}\| = 1$.
Similarly, Σ can be expressed using $\boldsymbol{A}$,

$$\Sigma = \boldsymbol{U}^T\boldsymbol{U}\Sigma\boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{U}^T\boldsymbol{A}\boldsymbol{V}.$$

The norm of Σ satisfies the following inequality,

$$\|\Sigma\| \leq \|\boldsymbol{U}^T\|\|\boldsymbol{A}\|\|\boldsymbol{V}\| = \|\boldsymbol{A}\|.$$

Therefore,

$$\|\boldsymbol{A}\| = \|\Sigma\| = \sigma_{\max}. \tag{2.23}$$

This equation indicates that a matrix's norm equals its maximum singular value.
Now we consider the situation of $\boldsymbol{A}^T\boldsymbol{A}$, $\boldsymbol{A}^T\boldsymbol{A}$ can be expressed using the SVD form of $\boldsymbol{A}$:

$$\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{V}\Sigma\boldsymbol{U}^T\boldsymbol{U}\Sigma\boldsymbol{V}^T = \boldsymbol{V}\Sigma^2\boldsymbol{V}^T. \tag{2.24}$$

This equation shows that the diagonal elements in Σ^2 are eigenvalues of $\boldsymbol{A}^T\boldsymbol{A}$.
Using SVD, the pesudo-inverse matrix of $\boldsymbol{A}$ can be defined as:

$$\boldsymbol{A}^+ = \boldsymbol{V}_{m\times r}\Sigma^{-1}_{r\times r}(\boldsymbol{U}^T)_{r\times n}. \tag{2.25}$$

Similar to $\|\boldsymbol{A}\|$, we can derive the norm of $\boldsymbol{A}^+$ as:

$$\|\boldsymbol{A}^+\| = \frac{1}{\sigma_{\min}}. \tag{2.26}$$

2.5 Positive semi-definite

A matrix is positive semi-definite (PSD) if any quadratic form it defines yields no negative values.

$$\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} \geq 0, \quad \forall \boldsymbol{x} \tag{2.27}$$

A symmetric matrix is PSD if all its eigenvalues are non-negative. For a PSD matrix $\boldsymbol{A}$, it can be factorized using eigenvalue decomposition:

$$\boldsymbol{A} = \boldsymbol{U}\Sigma\boldsymbol{U}^T, \tag{2.28}$$

where $\boldsymbol{U}$ is an orthogonal matrix, and Σ is a diagonal matrix with diagonal elements being eigenvalues of $\boldsymbol{A}$.
PSD matrices are very useful in machine learning applications. For example, in the attention mechanism, we have query matrix $\boldsymbol{Q} \in \mathbb{R}^{n\times d}$ and key matrix $\boldsymbol{K} \in \mathbb{R}^{n\times d}$, the attention score is computed using the similarity between $\boldsymbol{Q}$ and $\boldsymbol{K}$, which is the inner product of $\boldsymbol{Q}$ and $\boldsymbol{K}$ through the exponential function, $\exp(\boldsymbol{Q}\boldsymbol{K}^T)$. If we consider a matrix $\boldsymbol{X}$ composed of $\boldsymbol{Q}$ and $\boldsymbol{K}$:

$$\boldsymbol{X} = \begin{bmatrix}\boldsymbol{Q}\\ \boldsymbol{K}\end{bmatrix}$$

Then the matrix $\exp(\boldsymbol{X}\boldsymbol{X}^T)$ is a PSD matrix with $\exp(\boldsymbol{Q}\boldsymbol{K}^T)$ as its right upper component,

$$\exp(\boldsymbol{X}\boldsymbol{X}^T) = \exp\left(\begin{bmatrix}\boldsymbol{Q}\boldsymbol{Q}^T & \boldsymbol{Q}\boldsymbol{K}^T\\ \boldsymbol{K}\boldsymbol{Q}^T & \boldsymbol{K}\boldsymbol{K}^T\end{bmatrix}\right).$$

Further, we normalize the attention score using softmax operation, we first compute the denominator:

$$D = \operatorname{diag}\left(\exp(\boldsymbol{Q}\boldsymbol{K}^T)\cdot \mathbf{1}_n\right),$$

where $\mathbf{1}_n$ is a column vector of all ones with length n, and the diag operation refers to constructing a new matrix by placing all the elements of a given vector along its main diagonal. The attention score matrix is further computed as:

$$A = D^{-1}\exp(\boldsymbol{Q}\boldsymbol{K}^T)$$

Following the above procedure, we can see that the complexity of attention computation is quadratic with sequence length n.

2.6 Revisit linear regression

Recall that the optimization objective of a linear regression model can be described as the equation below:

$$\beta^* = \arg\min_{\beta} < \boldsymbol{X}\beta - \boldsymbol{Y}, \boldsymbol{X}\beta - \boldsymbol{Y} > . \tag{2.29}$$

We make the inner product term as a function $f(\beta)$, then take the derivative of the square loss using matrix derivative rules,

$$\frac{\partial f}{\partial \beta} = 0 \Rightarrow \boldsymbol{X}^T\boldsymbol{X}\hat{\beta} = \boldsymbol{X}^T\boldsymbol{Y}. \tag{2.30}$$

To examine the numerical stability of $\hat{\beta}$, we compute the condition number of $\boldsymbol{X}^T\boldsymbol{X}$ based on Equation (2.14):

$$\kappa(\boldsymbol{X}^T\boldsymbol{X}) = \|\boldsymbol{X}\|^2\|\boldsymbol{X}^+\|^2. \tag{2.31}$$

Using QR factorization $\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{R}$, where $\boldsymbol{Q}$ is an orthogonal matrix and $\boldsymbol{R}$ is an upper triangular matrix. Then we have,

$$\begin{aligned}\boldsymbol{X}^T\boldsymbol{X}\hat{\beta} = \boldsymbol{X}^T\boldsymbol{Y} &\Rightarrow \boldsymbol{R}^T\boldsymbol{Q}^T\boldsymbol{Q}\boldsymbol{R}\hat{\beta} = \boldsymbol{R}^T\boldsymbol{Q}^T\boldsymbol{Y} \\ &\Rightarrow \boldsymbol{R}^T\boldsymbol{R}\hat{\beta} = \boldsymbol{R}^T\boldsymbol{Q}^T\boldsymbol{Y} \\ &\Rightarrow \boldsymbol{R}\hat{\beta} = \boldsymbol{Q}^T\boldsymbol{Y}\end{aligned}$$

The condition number for this linear system is

$$\kappa(\boldsymbol{R}) = \kappa(\boldsymbol{X}) = \|\boldsymbol{X}\|\|\boldsymbol{X}^+\|. \tag{2.32}$$

From the above computation, we can know that using QR factorization significantly reduces condition number, a trick we can use for estimation.
Further, we check the variance for $\hat{\beta}$ from the perspective of conditioning. We represent the condition number using matrix singular value. According to Equation (2.23) and Equation (2.26), the condition number $\kappa(\boldsymbol{X}^T\boldsymbol{X})$ can be further expressed as

$$\kappa(\boldsymbol{X}^T\boldsymbol{X}) = \kappa(\boldsymbol{V}\Sigma^2\boldsymbol{V}^T) = \kappa(\Sigma^2) = \frac{\sigma_{\max}^2}{\sigma_{\min}^2}. \tag{2.33}$$

Revisit the variance of $\hat{\beta}$,

$$\mathrm{Var}(\hat{\beta}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}. \tag{2.34}$$

If we fix the 2-norm of $\boldsymbol{X}$ as 1, then the maximum singular value $\sigma_{\max}$ equals to 1. The condition number of $\boldsymbol{X}^T\boldsymbol{X}$ can be re-written as:

$$\kappa(\boldsymbol{X}^T\boldsymbol{X}) = \frac{1}{\sigma_{\min}^2}.$$

Taking the norm of variance on $\hat{\beta}$, we have:

$$\|\mathrm{Var}(\hat{\beta})\| = \|\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\| = \frac{\sigma^2}{\sigma_{\min}^2}.$$

In this case, when the smallest singular value $\sigma_{\min}$ of $\boldsymbol{X}$ is small, it leads to a large condition number and high variance. This situation indicates the instability of estimation when the collinearity among variables is high.