[Lecture 1] Floating point, vector norms, stability concept

Task: Prove that signed binary fixed-point numbers lay in the range $\pm(2^m - 2^{-n})$, where m, n are sizes of an integer and fractional part, respectively.

> **Solution:** The maximal value in a binary notation is zero sign bit and all ones in the main parts. That gives gives $\sum_{i=0}^{m-1} 2^i = \{\text{geometric sum for finite terms}\} = \frac{1\cdot(2^{(m-1)+1}-1)}{(2-1)} = 2^m$ for the integer part and $\sum_{i=1}^{n} 2^{-i} = \frac{2^{-1}\cdot(2^{-n}-1)}{(2^{-1}-1)} = 2^{-n}$ for the fractional, that in total is equal to $2^m - 2^{-n}$. For the minimal value, we should simply take the sign bit as one, so the value is the maximal value but negative.

Task: Find an angle between $x = [1, 2, 3]$ and $y = [4, 5, 6]$.

> **Solution:**
>
> $$\cos\alpha = \frac{(x, y)}{\|x\|_2 \|y\|_2}$$
>
> $$\begin{cases} (x, y) = 4 + 10 + 18 = 32 \\ \|x\|_2 = \sqrt{1+4+9} = \sqrt{14} \\ \|y\|_2 = \sqrt{16+25+36} = \sqrt{77} \end{cases} \Rightarrow \cos\alpha = \frac{32}{\sqrt{14}\sqrt{77}}$$

Task: Calculate xy^T, x^Ty

> **Solution:**
>
> $$xy^T = 32, \quad x^Ty = \begin{pmatrix} 4 & 5 & 6 \\ 8 & 10 & 12 \\ 12 & 15 & 18 \end{pmatrix}$$

[Lecture 2] Matrix norms, unitary matrices

Task: Compute a Frobenius norm of

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$$

> **Solution:**
>
> $$1^2 + 2^2 + 3^2 + 4^2 = 1 + 4 + 9 + 16 = 30 \Rightarrow \|A\|_F = \sqrt{30}$$

Task: Compute a Frobenius norm of

$$A = [a_{ij}]^{n\times n} \ \text{ where } \ a_{ij} = \begin{cases} 2^{-i/2} & \text{if } \ i = j \\ 0 & \text{otherwise} \end{cases} \ \text{ and } \ n = \infty$$

> **Solution:**
>
> $$\sum_{i=1}^{\infty} (2^{-i/2})^2 = \sum_{i=0}^{\infty} 2^{-i} - 1 = 2 - 1 \Rightarrow \|A\|_F = \sqrt{1} = 1$$

Task: Prove that $\|A\|_F = \sqrt{\text{trace}(AA^*)} = \sqrt{\text{trace}(A^*A)}$

> **Solution:**

$$[AA^*]_{ij} = \sum_{k=1}^{m} A_{ik}A_{kj}^* = \sum_{k=1}^{m} A_{ik}\overline{A_{jk}} \Rightarrow \text{if } i=j \ \text{ (that is trace)} = \sum_{k=1}^{m} |A_{ik}|^2 = \|A\|_F^2$$

Task: Identify the transformation matrix G:

$$G = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix}$$

> **Solution:** This matrix represents a Givens rotation in 2D.

Task: Prove that the matrix G is unitary/orthogonal.

> **Solution:** $G^TG = GG^T = \{\text{trigonometry}\} = I$

Task: Prove that this unitary transformation can zero out the second coordinate of a 2D vector (x_1, x_2).

> **Solution**: For this, we should simply find a suitable α value.
>
> $$x_1\sin\alpha + x_2\cos\alpha = 0 \Rightarrow \begin{cases} \frac{\sin\alpha}{\cos\alpha} = -\frac{x_2}{x_1} \\ \cos\alpha = 0 \end{cases} \Rightarrow \alpha = \begin{cases} \arctan(-\frac{x_2}{x_1}) \\ \pm\pi/2 \end{cases}$$

Task: Compute a Frobenius norm of a Fourier matrix:

$$F = \frac{1}{\sqrt{N}} \begin{bmatrix} 1 & 1 & 1 & 1 & \cdots & 1 \\ 1 & \omega & \omega^2 & \omega^3 & \cdots & \omega^{N-1} \\ 1 & \omega^2 & \omega^4 & \omega^6 & \cdots & \omega^{2(N-1)} \\ 1 & \omega^3 & \omega^6 & \omega^9 & \cdots & \omega^{3(N-1)} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \omega^{N-1} & \omega^{2(N-1)} & \omega^{3(N-1)} & \cdots & \omega^{(N-1)(N-1)} \end{bmatrix} \text{where } \omega = e^{-2\pi i/N}$$

> **Solution:**
>
> Let's show that this matrix is unitary.
>
> $$[FF^*]_{ij} = \frac{1}{\sqrt{N}}\frac{1}{\sqrt{N}}\sum_{k=1}^{N} F_{ik}\overline{F_{jk}} = \frac{1}{N}\sum_{k=1}^{N} \omega^{(i-1)(k-1)}\overline{\omega^{(j-1)(k-1)}} = \frac{1}{N}\sum_{k=1}^{N} \omega^{(i-1)(k-1)}\overline{\omega^{(j-1)(k-1)}}$$
>
> Since
>
> $$\omega^{c_1c_2}\overline{\omega^{c_3c_2}} = e^{-2\pi i c_1c_2/N}\overline{e^{-2\pi i c_3c_2/N}} = \{\overline{e^c} = e^{\overline{c}}\} = e^{-2\pi i c_1c_2/N}e^{2\pi i c_3c_2/N} = e^{-2\pi i(c_1-c_3)c_2/N}$$
>
> Then,
>
> $$[FF^*]_{ij} = \frac{1}{N}\sum_{k=1}^{N} e^{-2\pi i(i-j)(k-1)/N}$$
>
> if $i=j$, $[FF^*]_{ij} = 1$. If not,
>
> $$[FF^*]_{ij} = \frac{1}{N}\sum_{k=1}^{N} e^{-2\pi i(i-j)(k-1)/N} = \{\text{geometric sum}\} = \frac{1}{N}\frac{e^{-2\pi i(i-j)}-1}{\cdots} = \{\text{Euler's identity}\} = 0$$
>
> Then, since $\|A\|_F = \sqrt{\text{trace}(AA^*)}$, $\|F\|_F = \sqrt{N}$.

Task: Compute an infinite norm of a Fourier matrix.

> **Solution:**
>
> $$\|F\|_\infty = \max_i \sum_{j=1}^{N} |F_{ij}|$$
>
> Since
>
> $$\sum_{j=1}^{N} |F_{ij}| = \frac{1}{\sqrt{N}} \sum_{j=1}^{N} |\omega^{(i-1)(j-1)}|$$
>
> and
>
> $$\omega^c = e^{-2\pi i c/N} \Rightarrow |\omega^c| = 1$$
>
> We get
>
> $$\|F\|_\infty = \sqrt{N}$$