Machine Learning

LINEAR ALGEBRA FOR ML

Dr. Panagiotis (Panos) Markopoulos
panos@utsa.edu

A Note on Propositional Logic

Equivalent statements:

A $\Rightarrow$ B

If A, then B

A is **sufficient** for B

B $\Leftarrow$ A

Only if B, then A

B is **necessary** for A

Example: $y = |x|$

$x > 2 \Rightarrow y > 2$; IF $x > 2$ THEN $y > 2$; $x > 2$ is SUFFICIENT for $y > 2$ (but not NECESSARY since $y > 2$ also when $x < -2$ which excludes $x > 2$); $x > 2$ ONLY IF $y > 2$; $y > 2$ is NECESSARY for $x > 2$

Equivalent statements:

A $\Leftrightarrow$ B

Iff A, **then** B

Iff B, **then** A

A is **necessary & sufficient** for B

B is **necessary & sufficient** for A

A and B are **equivalent**

Example: $y = |x|$, $S := (-\infty, -2) \cup (2, +\infty)$

$x \in S \Leftrightarrow y > 2$; IFF $x \in S$ THEN $y > 2$; IFF $y > 2$ THEN $x \in S$; $x \in S$ is NECESSARY & SUFFICIENT for $y > 2$; $y > 2$ is NECESSARY & SUFFICIENT for $x \in S$; $y > 2$ and $x \in S$ are EQUIVALENT

Matrix

- **Consider matrix $\mathbf{X} \in \mathbb{C}^{M \times N}$**. If $M > N$, it is a tall matrix. If $M < N$, it is a wide matrix. If $M = N$, it is a square matrix.
- **Vector $\mathbf{x} \in \mathbb{C}^{M}$** is a matrix with a single column.
- **Scalar** $x \in \mathbb{C}$ is a vector of length 1, or a 1×1 matrix.
- An array with more than 2 ways (sides) is called **tensor**.

Matrix (cont'd)

- **Matrix set:** $\mathcal{X} \subset \mathbb{C}^{M \times N}$ is a set of matrices (not ordered, in general).
- **Cardinality:** $|\mathcal{X}|$ is the number of distinct elements in $\mathcal{X}$.
- **Intersection, union, set-difference:** Consider matrix sets $\mathcal{X}$ and $\mathcal{Y}$. Then, $\mathcal{X} \cap \mathcal{Y}$, $\mathcal{X} \cup \mathcal{Y}$, and $\mathcal{X} \setminus \mathcal{Y}$ are their intersection, union, and set-difference, respectively.

Matrix (cont'd)

❑ Indexing matrix entries

- Consider ordered sets $A \subseteq [N] \coloneqq \{1, \dots, N\}$ and $B \subseteq [M]$. $[\mathbf{X}]_{B,A} \in \mathbb{C}^{|B| \times |A|}$ is the sub-matrix obtained by extracting from $\mathbf{X}$ the rows with index in B and columns with index in A (in the specified order).

- *Special case:* Consider $i \in [M]$ and $j \in [N]$. $[\mathbf{X}]_{i,j} \in \mathbb{C}$ is an entry of $\mathbf{X}$, $[\mathbf{X}]_{i,[N]} \in \mathbb{C}^{1 \times N}$ is the i-th row of $\mathbf{X}$ and $[\mathbf{X}]_{[M],j} \in \mathbb{C}^{M \times 1}$ (or $[\mathbf{X}]_{:,j}$ is the j-th column of $\mathbf{X}$).

Basic Operations

- **Summation:** If $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{M \times N}$, then $\mathbf{Z} = \mathbf{X} + \mathbf{Y}$ is defined, such that $\forall (i,j) \in [M] \times [N]$

$$[\mathbf{Z}]_{i,j} = [\mathbf{X}]_{i,j} + [\mathbf{Y}]_{i,j}.$$

- **Multiplication:** If $\mathbf{X} \in \mathbb{R}^{M \times N}$ and $\mathbf{Y} \in \mathbb{R}^{N \times L}$, then $\mathbf{Z} = \mathbf{XY}$ is defined, such that $\forall (i,j) \in [M] \times [L]$

$$[\mathbf{Z}]_{i,j} = \sum_{n=1}^{N} [\mathbf{X}]_{i,n} [\mathbf{Y}]_{n,j}.$$

- **Hadamard product**: If $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{M \times N}$, then $\mathbf{Z} = \mathbf{X} \odot \mathbf{Y}$ is defined, such that $\forall (i,j) \in [M] \times [N]$

$$[\mathbf{Z}]_{i,j} = [\mathbf{X}]_{i,j} [\mathbf{Y}]_{i,j}.$$

- **Kronecker product**: For any $\mathbf{X} \in \mathbb{R}^{M \times N}$ and $\mathbf{Y} \in \mathbb{R}^{K \times L}$, $\mathbf{Z} = \mathbf{X} \otimes \mathbf{Y}$ is **block-matrix** of $M \times N$ blocks, such that the (i,j)-th block is equal to $[\mathbf{X}]_{i,j}\mathbf{Y}$.

Transpose, Conjugate, Hermitian

- **Transpose of matrix:** For any (i, j), it holds $[\mathbf{X}]_{i,j} = [\mathbf{X}^\top]_{j,i}$. This implies that $(\mathbf{XY})^\top = \mathbf{Y}^\top\mathbf{X}^\top$.
  - $\mathbf{X}$ is called "Symmetric" iff $\mathbf{X}^\top = \mathbf{X}$.
- **Conjugate of matrix:** For any (i, j), it holds $[\mathbf{X}]_{i,j} = [\mathbf{X}^*]_{i,j}$
- **Hermitian of matrix:** $\mathbf{X}^H = (\mathbf{X}^*)^\top \in \mathbb{C}^{N\times M}$.
  - $\mathbf{X}$ is a "Hermitian" matrix iff $\mathbf{X}^H = \mathbf{X}$.

Ortho-gonality/normality

- **Orthogonality:** $\mathbf{X} \in \mathbb{C}^{M \times N}$ is orthogonal iff $[\mathbf{X}^H\mathbf{X}]_{i,j} = 0$ for $i \neq j$ – i.e., the columns of $\mathbf{X}$ are orthogonal vectors.

- **Orthonormality:** $\mathbf{X} \in \mathbb{C}^{M \times N}$ is orthonormal matrix iff $\mathbf{X}^H\mathbf{X} = \mathbf{I}_M$.
  - If $\mathbf{X}$ is square and $\mathbf{X}^H\mathbf{X} = \mathbf{I}_M$, then $\mathbf{X}\mathbf{X}^H = \mathbf{I}_M$. $\mathbf{X}^H\mathbf{X} = \mathbf{I}_N$ only if $N \leq M$.

- **Stiefel Manifold:** $\mathbb{S}_{M,N} = \{\mathbf{X} \in \mathbb{C}^{M \times N} : \mathbf{X}^H\mathbf{X} = \mathbf{I}_N\}$, for any $N \leq M$.
  - Special case is the M-sphere $\mathbb{S}_M = \{\boldsymbol{x} \in \mathbb{R}^{M+1} : \boldsymbol{x}^\top\boldsymbol{x} = 1\}$. Notice that the unit circle is a 1-sphere.

Trace and Entry-wise Norms

- **Trace:** For square $\mathbf{X} \in \mathbb{C}^{M\times M}$, we define $\text{Trace}(\mathbf{X}) := \sum_{i=1}^{M} [\mathbf{X}]_{i,i}$
- **Entry-wise matrix norm**: For any $\mathbf{X} \in \mathbb{C}^{M \in N}$ and $p, q \geq 1$, we define

$$\| \mathbf{X} \|_{p,q} := \left(\sum_{j=1}^{N} \left(\sum_{i=1}^{M} \left| [\mathbf{X}]_{i,j} \right|^{p} \right)^{\frac{q}{p}} \right)^{\frac{1}{q}} .$$

- **Norm properties**: Operator $\|\cdot\|$ is a norm iff, for any $\mathbf{X}, \mathbf{Y} \in \mathbb{C}^{\mathrm{M}\times\mathrm{N}}$ and $\alpha \in \mathbb{C}$:

- $\| \alpha\mathbf{X} \| = |\alpha| \, \| \mathbf{X} \|$ (absolute homogeneity).
- $\| \mathbf{X} + \mathbf{Y} \| \leq \| \mathbf{X} \| + \| \mathbf{Y} \|$ (triangle inequality).
- $\| \mathbf{X} \| \geq 0$. $\| \mathbf{X} \| = 0$ iff $\mathbf{X} = \mathbf{0}_{M,N}$ (non-negativity).

The special case of $p = q = 2$ is also known as **Euclidean** or **Frobenius** norm, denoted as $\| \mathbf{X} \|_F$.

Scalar and Vector Norms

- For a scalar x, all "entry-wise" norms boil down to the absolute value $|x|$.
- For a vector $\mathbf{x} \in \mathbb{R}^{M \times 1}$, the (p, q) entry-wise norm is invariant to q, which can be omitted.

$$\| \mathbf{x} \|_p \coloneqq \| \mathbf{x} \|_{p,q} \;\; (\forall q) = \left(\sum_{j=1}^{1} \left(\sum_{i=1}^{M} \left| [\mathbf{x}]_{i,j} \right|^p \right)^{\frac{q}{p}} \right)^{\frac{1}{q}} = \left(\sum_{i=1}^{M} |[\mathbf{x}]_i|^p \right)^{\frac{1}{p}}$$

Other norm-like notation, but not typical norms:

- $\| \mathbf{x} \|_\infty = \max_{i=1,2,\ldots,M} |[\mathbf{x}]_i|$ (**infinity norm** or **maximum norm**)
- $\| \mathbf{x} \|_0 =$ #non-zero entries in $\mathbf{x}$ (just common notation; not really a norm)

Norm Inequality and Unit-Norm Spheres

- For any $\mathbf{x} \in \mathbb{R}^{M \times 1}$ and $p > r \geq 1$,

$$\| \mathbf{x} \|_p \leq \| \mathbf{x} \|_r \leq M^{\left(\frac{1}{r}-\frac{1}{p}\right)} \| \mathbf{x} \|_p.$$

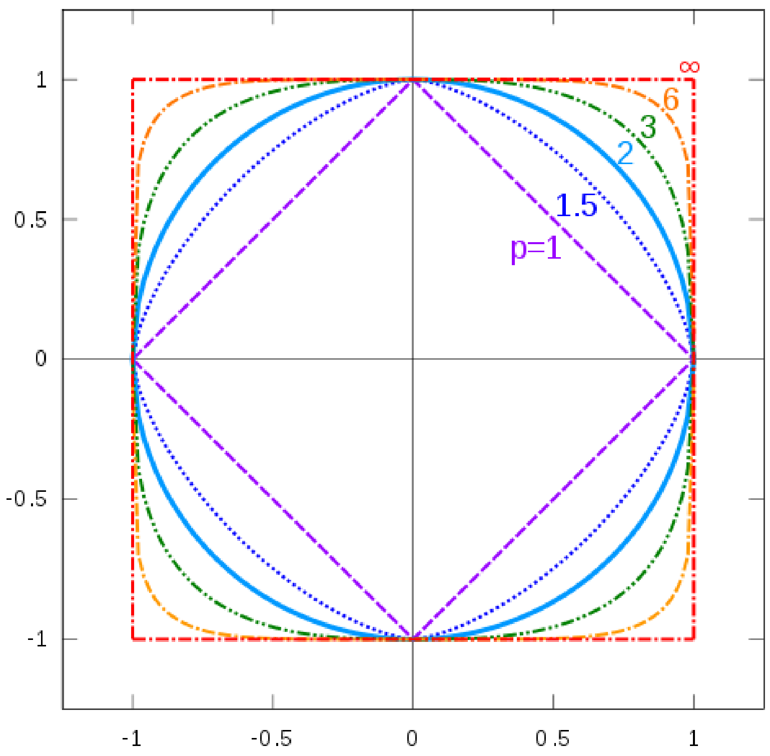
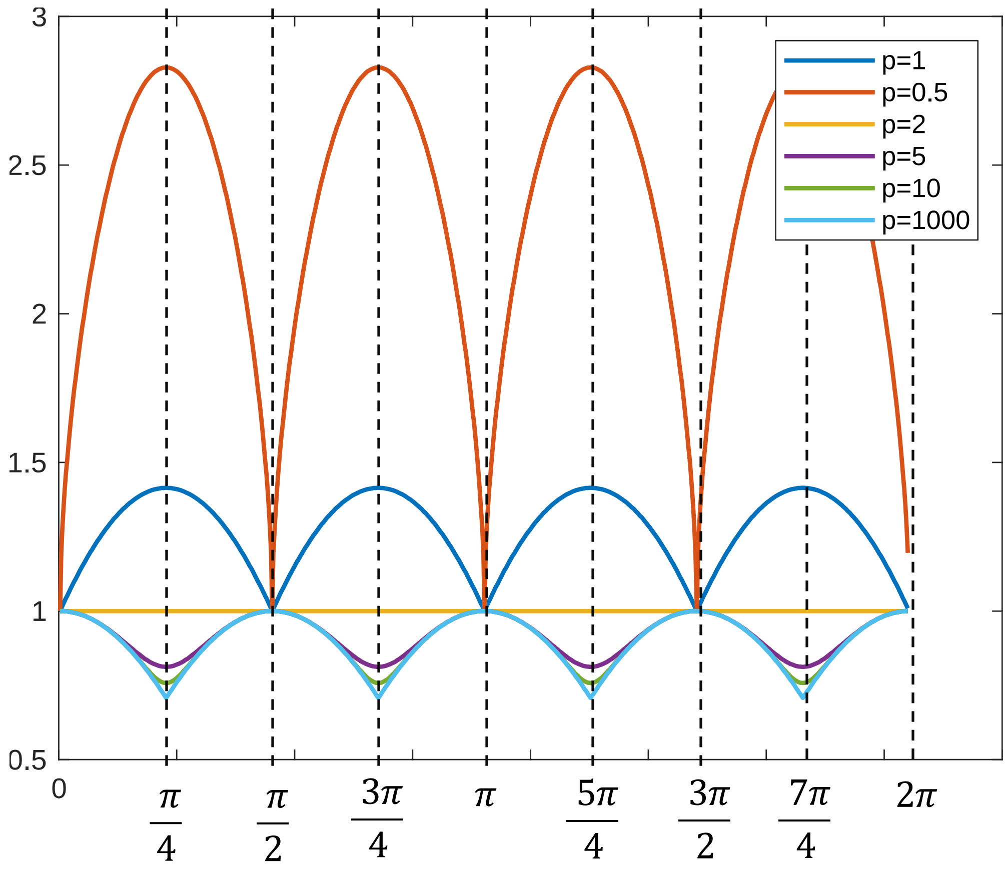


Fig: Unit norm spheres defined upon different norms.

Norm Inequality and Unit-Norm Spheres (cont'd)

- p-norm of all vectors on the 2-norm sphere ($M = 2$ dimensions).

(p>2)-norms are maximized where (p<2) norms are minimized, and vice versa.



Vector Inner Product

- For vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^M$, **inner product** is defined as:

$$\mathbf{x}^\top \mathbf{y} := \sum_{i=1}^{M} [\mathbf{x}]_i \, [\mathbf{y}]_i .$$

  - Algebraicaly, $\| \mathbf{x} + \mathbf{y} \|_2^2 = \| \mathbf{x} \|_2^2 + \| \mathbf{y} \|_2^2 + 2\, \mathbf{x}^\top \mathbf{y}$.
  - Geometrically, $\| \mathbf{x} + \mathbf{y} \|_2^2 = \| \mathbf{x} \|_2^2 + \| \mathbf{y} \|_2^2 + 2\cos(\theta) \, \| \mathbf{x} \|_2 \, \| \mathbf{y} \|_2$.
  - Thus, another expression of inner product is

$$\mathbf{x}^\top \mathbf{y} = \cos(\theta) \, \| \mathbf{x} \|_2 \, \| \mathbf{y} \|_2 \, .$$

Cauchy-Schwarz Inequality

- $\mathbf{x}^\top\mathbf{y} = \cos(\theta) \parallel \mathbf{x} \parallel_2 \parallel \mathbf{y} \parallel_2$
- $|\cos(\theta)| \leq 1$, with equality iff $\theta = 0$.

Cauchy-Schwartz Inequality (CSI):

For any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^M$, it holds

$$|\mathbf{x}^\top\mathbf{y}| \quad \leq \parallel \mathbf{x} \parallel_2 \parallel \mathbf{y} \parallel_2$$

with equality iff $\mathbf{x} = \mathbf{y}c$, for any $c \in \mathbb{R}$.

Hölder's Inequality

CSI also derives from the more general Hölder's Inequality.

Hölder's Inequality (HI):

For any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^M$ and q, p such that $\frac{1}{p} + \frac{1}{q} = 1$, it holds that

$$\sum_{i=1}^{M} |[\mathbf{x}]_i| \, |[\mathbf{y}]_i| \leq \|\mathbf{x}\|_p \|\boldsymbol{y}\|_q$$

with equality iff $\forall i$, $\frac{|[\mathbf{x}]_i \;|^p}{\|\mathbf{x}\|_p^p} = \frac{|[\mathbf{y}]_i \;|^q}{\|\mathbf{y}\|_q^q}$.

Young's Inequality (YI):

If $a, b \geq 0$ and $1 \leq p, q \leq \infty$ such that $\frac{1}{p} + \frac{1}{q} = 1$, then $ab \leq \frac{a^p}{p} + \frac{b^q}{q}$, w/ eq. iff $a^p = b^q$.

- HI derives from YI.
- CSI derives from HI for $p = q = 2$.

HI implies $\left|\sum_{i=1}^{M} [\mathbf{x}]_i \, [\mathbf{y}]_i\right| \leq \|\mathbf{x}\|_p \|\boldsymbol{y}\|_q$, with eq. iff $\forall i$, $\frac{|[\mathbf{x}]_i|^p}{\|\mathbf{x}\|_p^p} = \frac{|[\mathbf{y}]_i|^q}{\|\mathbf{y}\|_q^q}$ and $\mathrm{sgn}([\mathbf{x}]_i \, [\mathbf{y}]_i)$ fixed across i.

Linear Subspaces

- **Set of linearly independent vectors:** A set of vectors $\{\mathbf{x}_i\}_{i=1}^{N} \subset \mathbb{C}^M$ are linearly independent set (LI) iff: for any $i \in [N]$, $\nexists \mathbf{y} \in \mathbb{C}^{N-1}$ such that $\mathbf{x}_i = [\mathbf{X}]_{:,[N]\setminus i}\, \mathbf{y}$.

- **Linear subspace:** $\mathcal{S} \subset \mathbb{C}^M$ is a linear subspace iff for any $\mathbf{x}, \mathbf{y} \in \mathcal{S}$ and $a, b \in \mathbb{C}$, $\mathbf{x}a + \mathbf{y}b \in \mathcal{S}$.

- **Dimensionality:** $\dim(\mathcal{S})$ is the cardinality of the largest linearly independent subset in $\mathcal{S}$. It's a way to measure the "size" of $\mathcal{S}$. $\dim(\emptyset) = \dim(\{\mathbf{0}_M\}) = 0$.

Linear Subspaces (cont'd)

- **Span** or **Range** or **Column Space:** $\text{span}(\mathbf{X}) = \{\mathbf{x} \in \mathbb{C}^M : \mathbf{x} = \mathbf{X}\mathbf{y}, \mathbf{y} \in \mathbb{C}^N\}$.
  - $\text{span}(\mathbf{X})$ is a linear subspace.

- **Basis:** $\mathbf{X}$ is a basis for linear subspace $\mathcal{S}$ iff $\mathcal{S} = \text{span}(\mathbf{X})$.
  - A subspace can be spanned by infinitely many distinct bases.
  - Each matrix spans a unique subspace.
  - If $\mathbf{X}^H\mathbf{X} = \mathbf{I}_N$, then $\mathbf{X}$ is an orthonormal basis for $\text{span}(\mathbf{X})$.

Linear Subspaces (cont'd)

- **Orthogonal subspace:** $\mathcal{S}^{\perp} = \{\mathbf{x} \in \mathbb{C}^{M} : \mathbf{x}^{H}\mathbf{y} = 0 \ \forall \mathbf{y} \in \mathcal{S}\}$.
  - It holds $\dim(\mathcal{S}) = M - \dim(\mathcal{S}^{\perp})$.
  - Consider $\mathbf{X} \in \mathbb{C}^{M \times N}$ and $\mathbf{Y} \in \mathbb{C}^{M \times L}$. Then, $\mathbf{X}^{H}\mathbf{Y} = \mathbf{0}_{N,L} \Leftrightarrow \mathrm{span}(\mathbf{X}) = \mathrm{span}(\mathbf{Y})^{\perp}$.

- **Null-space** or **Kernel:** $\mathcal{N}(\mathbf{X}) = \{\boldsymbol{y} \in \mathbb{C}^{N} : \mathbf{X}\mathbf{y} = \mathbf{0}\}$

Fundamental Theorem of Linear Algebra:

- $\dim(\mathrm{span}(\mathbf{X})) = M - \dim(\mathrm{span}(\mathbf{X})^{\perp})$
- $\mathcal{N}(\mathbf{X}^{H}) = \mathrm{span}(\mathbf{X})^{\perp}$

Matrix Rank

- **Matrix rank:** For $\mathbf{X} = [\mathbf{x}_1, \ldots \mathbf{x}_N] \in \mathbb{C}^{M \times N}$, rank$(\mathbf{X})$ is the size of the largest linearly independent subset among the columns of $\mathbf{X}$, $\{\mathbf{x}_i\}_{i=1}^{N}$.

Remarks:

- $\dim\big(\mathrm{span}(\mathbf{X})\big) = \mathrm{rank}(\mathbf{X})$.
- $\mathrm{rank}(\mathbf{X}) \leq \min\{M, N\}$.
- If $\mathrm{rank}(\mathbf{X}) = M$, $\mathbf{X}$ is full row rank.
- If $\mathrm{rank}(\mathbf{X}) = N$, $\mathbf{X}$ is full column-rank.
- If $\mathrm{rank}(\mathbf{X}) = M = N$, $\mathbf{X}$ is square full-rank.
- If $\mathbf{X} \in \mathbb{S}_{M,N}$, then $\mathrm{rank}(\mathbf{X}) = N$.

Inverse and Pseudo-Inverse

- **Inverse:** If $\mathbf{X}$ is square and full rank, then $\mathbf{X}^{-1}$ exists, such that $\mathbf{X}^{-1}\mathbf{X} = \mathbf{X}\mathbf{X}^{-1} = \mathbf{I}_M$.

- **Moore-Penrose Pseudoinverses:**
  - Iff $\mathbf{X} \in \mathbb{C}^{M \times N}$ is full row rank (thus, wide), then the right-hand pseudoinverse $\mathbf{X}^{\dagger R} = \mathbf{X}^H(\mathbf{X}\mathbf{X}^H)^{-1}$ exists, such that $\mathbf{X}\mathbf{X}^{\dagger R} = \mathbf{I}_M$.
  - Iff $\mathbf{X} \in \mathbb{C}^{M \times N}$ is full column rank (thus, tall), then left-hand MP pseudoinverse $\mathbf{X}^{\dagger L} = (\mathbf{X}^H\mathbf{X})^{-1}\mathbf{X}^H$ exists, such that $\mathbf{X}^{\dagger L}\mathbf{X} = \mathbf{I}_N$.
  - If $\mathbf{X}$ is square full rank, then $\mathbf{X}^{\dagger R} = \mathbf{X}^{\dagger L} = \mathbf{X}^{-1}$.

Low-Rank Subspaces

It is often the case that high dimensional data largely reside on lower-dimensional subspaces. Thus, they can be compressed, denoised, visualized, and ML-processed within those subspaces with significant computational/storage gains and limited information loss.

Projection Matrix

- **Projection matrix:** $\mathbf{P}$ is a projection matrix iff $\mathbf{P} = \mathbf{P}\mathbf{P}$ and $\mathbf{P} = \mathbf{P}^H$.

Remarks:

- The mapping from projection $\mathbf{P}$ to $\text{span}(\mathbf{P})$ is 1-to-1.
- If $\mathbf{P}$ is projection, then $\mathbf{I}_M - \mathbf{P}$ is also projection with $\text{span}\,(\mathbf{I}_M - \mathbf{P}) = \text{span}(\mathbf{P})^{\perp}$.
- $\text{rank}(\mathbf{P}) = M - \text{rank}(\mathbf{I}_M - \mathbf{P})$.
- For any $\mathbf{x} \in \mathbb{C}^M$,

$$\mathbf{P}\mathbf{x} = \underset{\mathbf{y} \in \text{span}(\mathbf{P})}{\text{argmin}} \; \| \mathbf{y} - \mathbf{x} \|_2^2.$$

- If $\mathbf{U} \in \mathbb{S}_{M,K}$, then $\mathbf{U}\mathbf{U}^H$ is a projection matrix on $\text{span}(\mathbf{U}) = \text{span}(\mathbf{U}\mathbf{U}^H)$.