TODO: move this file to folder *signal processing* after migraion.

Orthogonality Principle

Let $\mathbb{F}$ denote either $\mathbb{R}$ or $\mathbb{C}$. Then, $(H, \mathbb{F})$ is a Hilber space iff all of the following are true

1. $(H, \mathbb{F})$ is a vector space.
2. H is equipped with inner product $\langle \cdot, \cdot \rangle : H \times H \to \mathbb{F}$.
3. H is complete w.r.t. the distance induced by $\langle \cdot, \cdot \rangle$.

Remark:

- The axioms of inner product are listed in appendix. Here, we highlight that $\langle \cdot, \cdot \rangle$ is **conjugate** linear w.r.t. the **1st** argument and **linear** w.r.t. the **2nd** argument. This is consistent with physics/engineering convention. NumPy and MATLAB also follows this convention.

$$\langle \lambda \mathbf{x}, \mathbf{y} \rangle = \overline{\lambda} \langle \mathbf{x}, \mathbf{y} \rangle, \quad \langle \mathbf{x}, \lambda \mathbf{y} \rangle = \lambda \langle \mathbf{x}, \mathbf{y} \rangle$$

```python
import numpy as np
a = np.array([1+2j, 3+4j])
b = np.array([5+6j, 7+8j])
print(np.vdot(a, b))  # Output: (70-8j)
```

- If $\mathbb{F} = \mathbb{R}$, then $\langle \cdot, \cdot \rangle$ becomes a bilinear form, i.e. linear w.r.t. both arguments.
- Every inner product induces a norm and thus a distance.

$$\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}, \quad d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\| = \sqrt{\langle \mathbf{x} - \mathbf{y}, \mathbf{x} - \mathbf{y} \rangle}$$

Commonly used Hilbert spaces and their inner products

- Euclidean space $\mathbb{R}^n$.

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{y} = \sum_{i=1}^{n} x_i y_i$$

- Complex coordinate space $\mathbb{C}^n$.

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\mathsf{H} \mathbf{y} = \sum_{i=1}^{n} \overline{x_i} y_i$$

- Space of square integrable functions $L^2(I, \mathbb{F}) = \{ f : I \to \mathbb{F} \mid \int_{t \in I} |f(t)|^2 \, \mathrm{d}t < \infty \}$ where $I \subseteq \mathbb{R}$.

$$\langle f, g \rangle = \int_I \overline{f(t)} g(t) \, \mathrm{d}t$$

- Space of random variables with finite 2nd order moment $L^2(\Omega, \mathbb{F}) = \{ X : \Omega \to \mathbb{F} \mid \mathbb{E}[|X|^2] < \infty \}$.

$$\langle X, Y \rangle = \mathbb{E}[\overline{X}Y] = \int_{\Omega} \overline{X(\omega)} Y(\omega) \, \mathrm{d}\mathbb{P}$$

Subspace Approximation Problem

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{y} \in H$. We woule like to find a vector $\hat{\mathbf{y}} \in \operatorname{span}(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ s.t. $\hat{\mathbf{y}}$ is as close to $\mathbf{y}$ as possible. Formally, we would like to solve

$$\min_{\hat{\mathbf{y}} \in \operatorname{span}(\mathbf{x}_1, \ldots, \mathbf{x}_n)} \|\hat{\mathbf{y}} - \mathbf{y}\|^2 \tag{1}$$

or equivalently

$$\min_{w_1, \ldots, w_n \in \mathbb{F}} \left\| \sum_{k=1}^{n} w_k \mathbf{x}_k - \mathbf{y} \right\|^2 \tag{2}$$

Remark:

- The equivalent formulation is straight forward as $\hat{\mathbf{y}} = \sum_{k=1}^{n} w_k \mathbf{x}_k$.
- The spanning vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are not necessarily linearly independent.

Orthogonality Principle

Let $U = \operatorname{span}(\mathbf{x}_1, \ldots, \mathbf{x}_n)$. The optimal solution of the subspace approximation problem is the orthogonal projection of $\mathbf{y}$ to U. The approximation error $\hat{\mathbf{y}} - \mathbf{y}$ lies in the orthogonal complement of U. In particular,

$$\langle \hat{\mathbf{y}} - \mathbf{y}, \mathbf{x}_k \rangle = 0, \ \forall k = 1, \ldots, n \tag{3}$$

Now, we calculate the optimal coefficients as follows. Reformulate the orthogonality principle into

$$\begin{aligned} \langle \hat{\mathbf{y}}, \mathbf{x}_k \rangle &= \langle \mathbf{y}, \mathbf{x}_k \rangle \\ \left\langle \sum_{\ell=1}^{n} w_\ell \mathbf{x}_\ell, \mathbf{x}_k \right\rangle &= \langle \mathbf{y}, \mathbf{x}_k \rangle \\ \sum_{\ell=1}^{n} \overline{w_\ell} \langle \mathbf{x}_\ell, \mathbf{x}_k \rangle &= \langle \mathbf{y}, \mathbf{x}_k \rangle \end{aligned}$$

Taking the complex conjugate on both sides, we get a linear system of n equations with n unknown coefficients.

$$\sum_{\ell=1}^{n} \langle \mathbf{x}_k, \mathbf{x}_\ell \rangle w_\ell = \langle \mathbf{x}_k, \mathbf{y} \rangle, \ \forall k = 1, \ldots, n \tag{4}$$

Matrix form:

$$\underbrace{\begin{bmatrix} \langle \mathbf{x}_1, \mathbf{x}_1 \rangle & \langle \mathbf{x}_1, \mathbf{x}_2 \rangle & \cdots & \langle \mathbf{x}_1, \mathbf{x}_n \rangle \\ \langle \mathbf{x}_2, \mathbf{x}_1 \rangle & \langle \mathbf{x}_2, \mathbf{x}_2 \rangle & \cdots & \langle \mathbf{x}_2, \mathbf{x}_n \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \mathbf{x}_n, \mathbf{x}_1 \rangle & \langle \mathbf{x}_n, \mathbf{x}_2 \rangle & \cdots & \langle \mathbf{x}_n, \mathbf{x}_n \rangle \end{bmatrix}}_{\mathbf{G} \in \mathbb{F}^{n \times n}} \underbrace{\begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix}}_{\mathbf{w} \in \mathbb{F}^n} = \underbrace{\begin{bmatrix} \langle \mathbf{x}_1, \mathbf{y} \rangle \\ \langle \mathbf{x}_2, \mathbf{y} \rangle \\ \vdots \\ \langle \mathbf{x}_n, \mathbf{y} \rangle \end{bmatrix}}_{\mathbf{r} \in \mathbb{F}^n} \tag{5}$$

Remarks:

- The matrix $\mathbf{G}$ is called ***Gram matrix*** (or ***kernel matrix***) of $\mathbf{x}_1, \dots, \mathbf{x}_n$. The element $G_{ij} = \langle \mathbf{x}_i, \mathbf{x}_j \rangle$ describes the similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$. Likewise, the term $r_i = \langle \mathbf{x}_i, \mathbf{y} \rangle$ on the RHS describes similarity between $\mathbf{x}_i$ and $\mathbf{y}$.
- The Gram matrix $\mathbf{G}$ is Hermitian, i.e. $\mathbf{G}^{\mathsf{H}} = \mathbf{G}$. If $\mathbb{F} = \mathbb{R}$, then $\mathbf{G}$ becomes symmetric.
- The Gram matrix $\mathbf{G}$ is invertible iff $\mathbf{x}_1, \dots, \mathbf{x}_n$ is linearly independent vectors in H. (c.f. Appendix for proof.) Hence, in general, the optimal $\mathbf{w}$ is not unique. However, the optimal approximation $\hat{\mathbf{y}}$ is always unique due to uniqueness of orthogonal projection.

Special case: $\mathbf{x}_1, \dots, \mathbf{x}_n$ are nonzero and orthogonal to each other. The Gram matrix becomes a diagonal matrix and thus invertible

$$\mathbf{G} = \operatorname{diag}(\langle \mathbf{x}_1, \mathbf{x}_1 \rangle, \langle \mathbf{x}_2, \mathbf{x}_2 \rangle, \dots, \langle \mathbf{x}_n, \mathbf{x}_n \rangle)$$

The optimal coefficents becomes

$$w_k = \frac{\langle \mathbf{x}_k, \mathbf{y} \rangle}{\langle \mathbf{x}_k, \mathbf{x}_k \rangle} = \frac{\langle \mathbf{x}_k, \mathbf{y} \rangle}{\|\mathbf{x}_k\|^2}$$

The orthogonal projection becomes

$$\hat{\mathbf{y}} = \sum_{k=1}^{n} \frac{\langle \mathbf{x}_k, \mathbf{y} \rangle}{\|\mathbf{x}_k\|^2} \mathbf{x}_k$$

If $\mathbf{x}_1, \dots, \mathbf{x}_n$ are in addition orthonormal (i.e. $\langle \mathbf{x}_i, \mathbf{x}_j \rangle = \mathbb{I}[i = j]$), the results simplifies further to

$$w_k = \langle \mathbf{x}_k, \mathbf{y} \rangle, \quad \hat{\mathbf{y}} = \sum_{k=1}^{n} \langle \mathbf{x}_k, \mathbf{y} \rangle \mathbf{x}_k$$

Ordinary Least Square

Consider $H = \mathbb{R}^m$ and $\mathbb{F} = \mathbb{R}$. We would like to approximate a vector $\mathbf{y}$ in the span of $\mathbf{x}_1, \dots, \mathbf{x}_n$.

Note that the approximation $\hat{\mathbf{y}}$ can be written as matrix vector product.

$$\hat{\mathbf{y}} = \sum_{k=1}^{n} w_k \mathbf{x}_k = \mathbf{X}\mathbf{w}, \text{ where } \mathbf{X} \triangleq [\mathbf{x}_1, \dots, \mathbf{x}_n] \in \mathbb{R}^{m \times n}$$

The subspace approximation problem

$$\min_{\hat{\mathbf{y}} \in \operatorname{span}(\mathbf{x}_1, \dots, \mathbf{x}_n)} \|\hat{\mathbf{y}} - \mathbf{y}\|^2$$

becomes **ordinary least square**

$$\min_{\mathbf{w} \in \mathbb{R}^n} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 \tag{6}$$

For all $k = 1, \dots, n$, the orthogonality principle $\langle \hat{\mathbf{y}} - \mathbf{y}, \mathbf{x}_k \rangle = 0$ becomes

$$(\hat{\mathbf{y}} - \mathbf{y})^\top \mathbf{x}_k = 0 \tag{7}$$
$$(\mathbf{X}\mathbf{w} - \mathbf{y})^\top \mathbf{x}_k = 0 \tag{8}$$

which can be written more compactly as

$$
\begin{aligned}
(\mathbf{X}\mathbf{w}-\mathbf{y})^\top[\mathbf{x}_1,\dots,\mathbf{x}_n] &= [0,\dots,0] \\
(\mathbf{X}\mathbf{w}-\mathbf{y})^\top\mathbf{X} &= \mathbf{0}^\top \\
\mathbf{X}^\top(\mathbf{X}\mathbf{w}-\mathbf{y}) = \mathbf{0} &\iff \nabla_{\mathbf{w}}\|\mathbf{X}\mathbf{w}-\mathbf{y}\|^2 = 0
\end{aligned} \tag{9}
$$

Hence, we can the optimal weights by solving

$$
\mathbf{X}^\top\mathbf{X}\mathbf{w} = \mathbf{X}^\top\mathbf{y} \tag{10}
$$

Remarks:

- One can verify that $\mathbf{X}^\top\mathbf{X} = \mathbf{G}$ with. $G_{ij} = \mathbf{x}_i^\top\mathbf{x}_j$ and that $\mathbf{X}^\top\mathbf{y} = \mathbf{r}$ with $r_i = \mathbf{x}_i^\top\mathbf{y}$.
- If $\mathbf{X}^\top\mathbf{X}$ is invertible, we have unique solution $\mathbf{w} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$.

Relation to linear regression: Recall that $\mathbf{X} \in \mathbb{R}^{m\times n}$. Consider each row of $\mathbf{X}$ as a data point in $\mathbb{R}^n$. Then, $\mathbf{X}$ represents m data points in $\mathbb{R}^n$ while $\mathbf{y} \in \mathbb{R}^m$ represents their corresponding labels.

- #training samples: m.
- #features per sample: n.
- i-th tranning sample: i-th row of $\mathbf{X}$ and y_i.

LMMSE Estimation

Consider $H = \{X : \Omega \to \mathbb{F} \mid \mathbb{E}[|X|^2] < \infty\}$ where $\mathbb{F}$ is either $\mathbb{R}$ or $\mathbb{C}$. We would like to approximate a random variable Y using a linear combination of $X_1,\dots,X_n$.

$$
\hat{Y} = \sum_{k=1}^n w_k X_k, \quad w_1,\dots,w_n \in \mathbb{F}
$$

Recall for two random variables $U, V \in H$, we have $\langle U, V\rangle = \mathbb{E}[\overline{U}V]$ and $\|U-V\|^2 = \mathbb{E}[|U-V|^2]$.

The subspace approximation problem

$$
\min_{\hat{Y}\in\operatorname{span}(X_1,\dots,X_n)} \|\hat{Y}-Y\|^2
$$

becomes **Linear Minimum Mean Square Error (LMMSE)** estimation:

$$
\min_{\hat{Y}\in\operatorname{span}(X_1,\dots,X_n)} \mathbb{E}\left[\left|\hat{Y}-Y\right|^2\right] \tag{11}
$$

$$
\min_{w_1,\dots,w_n\in\mathbb{F}} \mathbb{E}\left[\left|\sum_{k=1}^n w_k X_k - Y\right|^2\right] \tag{12}
$$

For all $k = 1,\dots,n$, the orthogonality principle $\langle \hat{Y}-Y, X_k\rangle = 0$ becomes

$$
\langle X_k, \hat{Y}-Y\rangle = 0 \tag{13}
$$

$$
\mathbb{E}\left[\overline{X_k}\cdot(\hat{Y}-Y)\right] = 0 \tag{14}
$$

$$
\mathbb{E}\left[\overline{X_k}\cdot\left(\sum_{\ell=1}^n w_\ell X_\ell - Y\right)\right] = 0 \tag{15}
$$

$$
\sum_{\ell=1}^n \mathbb{E}\left[\overline{X_k}X_\ell\right] w_\ell = \mathbb{E}[\overline{X_k}Y] \tag{16}
$$

Matrix form:

$$\begin{bmatrix} \mathbb{E}[\overline{X_1}X_1] & \mathbb{E}[\overline{X_1}X_2] & \cdots & \mathbb{E}[\overline{X_1}X_n] \\ \mathbb{E}[\overline{X_2}X_1] & \mathbb{E}[\overline{X_2}X_2] & \cdots & \mathbb{E}[\overline{X_2}X_n] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}[\overline{X_n}X_1] & \mathbb{E}[\overline{X_n}X_2] & \cdots & \mathbb{E}[\overline{X_n}X_n] \end{bmatrix} \cdot \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix} = \begin{bmatrix} \mathbb{E}[\overline{X_1}Y] \\ \mathbb{E}[\overline{X_2}Y] \\ \vdots \\ \mathbb{E}[\overline{X_n}Y] \end{bmatrix} \tag{$\star$}$$

Define the random vector $\mathbf{X} \triangleq [X_1, \dots, X_n]^\top \in \mathbb{C}^n$. Recall the auto-correlation matrix and cross-correlation matrix are defined as

$$\begin{aligned} \mathbf{R}_{XX} = \mathbb{E}[\mathbf{X}\mathbf{X}^\mathsf{H}] &\iff (\mathbf{R}_{XX})_{ij} = \mathbb{E}[X_i\overline{X_j}] \\ \mathbf{r}_{XY} = \mathbb{E}[\mathbf{X}\overline{Y}] &\iff (\mathbf{r}_{XY})_i = \mathbb{E}[X_i\overline{Y}] \end{aligned}$$

Let $\mathbf{w} \triangleq [w_1, \dots, w_n]^\top$. Equation $(\star)$ becomes

$$\overline{\mathbf{R}_{XX}} \cdot \mathbf{w} = \overline{\mathbf{r}_{XY}} \tag{17}$$

If $\mathbb{F} = \mathbb{R}$, then equation $(\star)$ reduces further to

$$\mathbf{R}_{XX} \cdot \mathbf{w} = \mathbf{r}_{XY} \tag{18}$$

OLS converges to LMMSE estimation

Let $\{(\mathbf{x}_i, y_i)\}_{i=1}^m \subset \mathbb{R}^n \times \mathbb{R}$ be iid from $p(\mathbf{x}, y)$. In OLS, we aim to fit the model $y = \mathbf{w}^\top \mathbf{x}$. Define the data matrix

$$\begin{aligned} \mathbf{X}_{\text{train}} &= [\mathbf{x}_1, \dots, \mathbf{x}_m] \in \mathbb{R}^{n \times m} \\ \mathbf{y}_{\text{train}} &= [y_1, \dots, y_m]^\top \in \mathbb{R}^m \end{aligned}$$

Note: There are two ways to define the data matrix. In OLS section, the data matrix was defined s.t. each row is a data point. Here, $\mathbf{X}_{\text{train}}$ is defined s.t. each column is a data point. The advantage of such definition will be clear later.

By OLS, the optimal weight vector satisfies

$$\mathbf{X}_{\text{train}}\mathbf{X}_{\text{train}}^\top \mathbf{w} = \mathbf{X}_{\text{train}}\mathbf{y}_{\text{train}}$$

Multiplying both sides with sample size m yields

$$\frac{1}{m}\mathbf{X}_{\text{train}}\mathbf{X}_{\text{train}}^\top \mathbf{w} = \frac{1}{m}\mathbf{X}_{\text{train}}\mathbf{y}_{\text{train}}$$

Note that

$$\begin{aligned} \frac{1}{m}\mathbf{X}_{\text{train}}\mathbf{X}_{\text{train}}^\top &= \frac{1}{m}[\mathbf{x}_1, \dots, \mathbf{x}_m] \cdot \begin{bmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_m^\top \end{bmatrix} = \frac{1}{m}\sum_{i=1}^m \mathbf{x}_i\mathbf{x}_i^\top \\ \frac{1}{m}\mathbf{X}_{\text{train}}\mathbf{y}_{\text{train}} &= \frac{1}{m}[\mathbf{x}_1, \dots, \mathbf{x}_m] \cdot \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} = \frac{1}{m}\sum_{i=1}^m \mathbf{x}_i y_i^\top \end{aligned}$$

By the law of large number,

$$\lim_{m\to\infty}\frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_i\mathbf{x}_i^\top=\mathbb{E}[\mathbf{x}\mathbf{x}^\top]=\mathbf{R}_{XX}$$
$$\lim_{m\to\infty}\frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_i y_i^\top=\mathbb{E}[\mathbf{x}y]=\mathbf{r}_{XY}$$

Theorefore, as the sample size $m\to\infty$, OLS becomes LMMSE estimation.

$$\underbrace{\frac{1}{m}\mathbf{X}_{\text{train}}\mathbf{X}_{\text{train}}^\top}_{\to\mathbf{R}_{XX}\text{ as }m\to\infty}\mathbf{w}=\underbrace{\frac{1}{m}\mathbf{X}_{\text{train}}\mathbf{y}_{\text{train}}}_{\to\mathbf{r}_{XY}\text{ as }m\to\infty}$$
$$\mathbf{R}_{XX}\mathbf{w}=\mathbf{r}_{XY}$$

Connection to Fourier Series

Appendix

Axiomic Definition of Inner Product

Let $(V,\mathbb{F})$ be a vector space. Then, $\langle\cdot,\cdot\rangle:V\times V\to\mathbb{F}$ is called a inner product iff all of

1. conjugate symmetry: $\langle\mathbf{x},\mathbf{y}\rangle=\overline{\langle\mathbf{x},\mathbf{y}\rangle}$
2. linear w.r.t. the 2nd argument: $\langle\mathbf{x},\lambda\mathbf{y}\rangle=\lambda\langle\mathbf{x},\mathbf{y}\rangle$ and $\langle\mathbf{x},\mathbf{y}+\mathbf{z}\rangle=\langle\mathbf{x},\mathbf{y}\rangle+\langle\mathbf{x},\mathbf{z}\rangle$
3. positive definite: $\langle\mathbf{x},\mathbf{x}\rangle\geq 0$ with equality iff $\mathbf{x}=\mathbf{0}$

Elementary properties:

- $\langle\lambda\mathbf{x},\mathbf{y}\rangle=\overline{\lambda}\langle\mathbf{x},\mathbf{y}\rangle$
- $\langle\mathbf{x}+\mathbf{y},\mathbf{z}\rangle=\langle\mathbf{x},\mathbf{y}\rangle+\langle\mathbf{x},\mathbf{z}\rangle$
- $\langle\mathbf{x},\mathbf{0}\rangle=0$

Remarks:

- In general, $\langle\mathbf{x},\mathbf{y}\rangle\neq\langle\mathbf{y},\mathbf{x}\rangle$ unless $\mathbb{F}=\mathbb{R}$.
- In math literatures, the inner product is often defined such that it is linear w.r.t the 1st argument.

Invertibility of Gram Matrix

Let $\mathbf{A}\in\mathbb{R}^{m\times n}$, then

1. $\ker(\mathbf{A}^\top)$ and $\operatorname{ran}(\mathbf{A})$ are orthogonal complements. (and thus their intersection is $\{\mathbf{0}\}$)
2. $\ker(\mathbf{A}^\top\mathbf{A})=\ker(\mathbf{A})$.
3. $\mathbf{A}^\top\mathbf{A}$ is invertible $\iff$ $\mathbf{A}$ has linearly independent columns.

Proof 1: We need to show that $\forall\mathbf{x}\in\ker(\mathbf{A}^\top),\forall\mathbf{y}\in\operatorname{ran}(\mathbf{A}),\langle\mathbf{x},\mathbf{y}\rangle=0$. By assumption,

$$\begin{aligned}\mathbf{x}\in\ker(\mathbf{A}^\top)&\implies\mathbf{A}^\top\mathbf{x}=0\\ \mathbf{y}\in\operatorname{ran}(\mathbf{A})&\implies\exists\mathbf{u}\in\mathbb{R}^n,\text{ s.t. }\mathbf{y}=\mathbf{A}\mathbf{u}\end{aligned}$$

Then, we conclude

$$\langle\mathbf{x},\mathbf{y}\rangle=\mathbf{x}^\top\mathbf{y}=\mathbf{x}^\top\mathbf{A}\mathbf{u}=\mathbf{u}^\top\underbrace{\mathbf{A}^\top\mathbf{x}}_{\mathbf{0}}=0$$

■

Proof 2: We only need show that $\ker(\mathbf{A}^\top\mathbf{A}) \subseteq \ker(\mathbf{A})$ since the inclusion in the opposite direction is trivial.

Consider $\mathbf{x} \in \ker(\mathbf{A}^\top\mathbf{A})$. By assumption,

$$\mathbf{A}^\top\mathbf{A}\mathbf{x} = 0 \implies \mathbf{A}\mathbf{x} \in \ker(\mathbf{A}^\top)$$

On the other hand, $\mathbf{A}\mathbf{x} \in \mathrm{ran}(\mathbf{A})$. Hence,

$$\mathbf{A}\mathbf{x} \in \mathrm{ran}(\mathbf{A}) \cap \ker(\mathbf{A}^\top) \implies \mathbf{A}\mathbf{x} = \mathbf{0} \implies \mathbf{x} \in \ker(\mathbf{A}) \qquad \blacksquare$$

Proof 3: This follows directly from $\ker(\mathbf{A}^\top\mathbf{A}) = \ker(\mathbf{A})$ as

$$\mathbf{A}^\top\mathbf{A} \text{ invertible} \iff \ker(\mathbf{A}^\top\mathbf{A}) = \{\mathbf{0}\} \iff \ker(\mathbf{A}) = \{\mathbf{0}\} \iff \mathbf{A} \text{ has lin-indep. col.} \qquad \blacksquare$$