EE499 Cheat Sheet - Midterm 1

Vector space: A vector space V is a set of mathematical objects along with a field of scalars $\mathbb{F}$ that is closed under vector addition & scalar multiplication:

$$\forall x, y \in V,\ x+y \in V$$

$$\forall x \in V, \forall \alpha \in \mathbb{F},\ \alpha x \in V$$

Here these operation satisfy

- Additive commutativity
- Additive associativity
- Existence of additive identity
- Scalar multiplicative associativity
- Distribution laws
- $0\underline{x} = \underline{0}$ & $1\underline{x} = \underline{x}$ (which implies additive inverses)

Subspace: A subset of a vector space that is itself a vector space is called a subspace.

- The same operation is concerned as the parent vector space.
- A subspace must always contain the null vector.

Operations on subspaces: Let S & T be subspaces of a vector space V.

- $S \cap T$ is also a subspace
- $S \cup T$ is not necessarily a subspace

Span: For some vectors $x_i \in V$, $\text{span}(\{x_i\})$ is the set of all linear combinations of x_i

Algebraic Sum: Let $S, T \subset V$. Then the algebraic sum of S & T is

$$S+T = \{ s+t \mid s \in S \ \& \ t \in T \}$$

Direct Sum: The direct sum of two subspaces S, T of V is their algebraic sum if and only if S & T are disjoint subspaces, i.e. $S \cap T = \{\underline{0}\}$.

Linear combination: Consider a vector space V and its subset $S \subset V$. A vector $x \in V$ is a linear combination of elements of S if there exists a finite number of elements $\{s_1, \dots, s_n\} \subset S$ and $\{\alpha_1, \dots, \alpha_n\} \subset \mathbb{F}$ such that

$$\alpha_1 s_1 + \dots + \alpha_n s_n = x$$

Linear Independence: $x_1, \dots, x_n$ are linearly independent vectors if

$$\alpha_1 x_1 + \dots + \alpha_n x_n = 0 \iff \forall k \ \alpha_k = 0$$

Linearly independent vectors enable unique representation of other vectors, i.e.

$$\sum_{k=1}^{n} \alpha_k x_k = \sum_{k=1}^{n} \beta_k x_k \text{ if and only if } \forall k \ \alpha_k = \beta_k$$

Basis: A first definition: For a vector space V, the set $\{x_1, \dots, x_n\}$ are called a basis for V if

1. $x_1, \dots, x_n$ are linearly independent
2. $\text{Span}(x_1, \dots, x_n) = V$.

A more general definition: The set of vectors $\Phi = \{\varphi_k\}_{k \in K} \subset V$ where K is either finite or countably infinite is called a basis for the normed vector space V when

1. $\forall x \in V \ \ x = \sum_{k \in K} \alpha_k \varphi_k, \ \alpha_k \in \mathbb{K} \Rightarrow$ convergence dependent on the norm
2. Any sequence α_k that represents some $x \in V$ is unique.

We are mainly looking for existence & uniqueness of a representation. Linear independence of the elements is equivalent to the uniqueness of the representation.

Change of Basis: Say you have two different bases, e & v, for the same vector space. $[x]_e$ and $[x]_v$ are then the representations of some vector x in those bases. Then, one can pass between these representations through a simple matrix multiplication:

$$e = \{e_1, e_2, \dots, e_n\}$$
$$v = \{v_1, v_2, \dots, v_n\}$$

$$[x]_v = T_{e \to v} [x]_e \quad \text{where}$$

$$T_{e \to v} = \left[[e_1]_v \ [e_2]_v \ \dots \ [e_n]_v \right]$$

$$T_{v \to e} = (T_{e \to v})^{-1} = \left[[v_1]_e \ [v_2]_e \ \dots \ [v_n]_e \right]$$

Functional: A mapping T from a vector space V to the real numbers $\mathbb{R}$ or the complex numbers $\mathbb{C}$.

$$T : V \longrightarrow \mathbb{R}, \mathbb{C}$$

i.e. onto the field.

Norm: A functional is a norm if it satisfies the following axioms:

- Non-negativity: $\|x\| > 0 \quad \forall x \neq \underline{0}, \quad \|x\| = 0 \Leftrightarrow x = \underline{0}$
- Homogeneity: $\|\alpha x\| = |\alpha| \cdot \|x\| \quad \forall x \in V, \forall \alpha \in F$
- Triangle inequality: $\|x+y\| \leq \|x\| + \|y\| \quad \forall x, y \in V$

Distance: A norm induces a metric $d(x,y) = \|x - y\|$

Normed Vector Space: A normed vector space is a pair $(V, \|\cdot\|)$ where V is a vector space and $\|\cdot\|$ is a norm defined on V.

Some typical norms:

- ℓ_p, $p \geq 1$: $\left(\sum |x_i|^p\right)^{1/p}$, ℓ_∞: $\sup_i |x_i|$
- $L_p[a,b]$, $p \geq 1$: $\left(\int_a^b |x|^p dt\right)^{1/p}$, $L_\infty[a,b]$: $\max x(t)$
- On $\mathbb{C}^{N \times N}$, $\|A\| = \max |A_{ij}|$
- On $\mathbb{C}^{N \times N}$, $\|A\|_F = \left(\sum_i \sum_j |A_{ij}|^2\right)^{1/2} = \sqrt{\text{trace}(A^H A)}$

For ℓ_p norms, we have

- (Unit ball)$_p \subset$ (Unit ball)$_q$ if $p < q$
- $\|x\|_p \geq \|x\|_q$ if $p < q$
- $\ell_p \subset \ell_q$ if $p \leq q$

For function spaces, the opposite of the last item is true.

$$\ell_1 \subset \cdots \subset \ell_\infty$$

$$L_1[a,b] \supset \cdots \supset L_\infty[a,b]$$

Inner product: Let V be a vector space, with its field $\mathcal{F}$. An inner product is a mapping $\langle \cdot, \cdot \rangle : V \times V \to \mathcal{F}$ with

1. $\langle x, y \rangle = \langle y, x \rangle^* \quad \forall x, y \in V$
2. $\langle x+y, z \rangle = \langle x, z \rangle + \langle y, z \rangle \quad \forall x, y, z \in V.$
3. $\langle \alpha x, z \rangle = \alpha \langle x, z \rangle$
4. $\langle x, x \rangle > 0 \quad \forall x \in V \ \& \ x \neq \underline{0}, \ \langle x, x \rangle = 0$ iff $x = \underline{0}$

Inner product space: An inner product space is a vector space equipped with an inner product.

Induced norm: An inner product induces a unique norm as

$$\|x\| = \sqrt{\langle x, x \rangle}$$

Cauchy-Schwartz inequality: For all vectors $x, y \in V$,

$$|\langle x, y \rangle| \leq \|x\| \cdot \|y\|$$

Angle: Using the Cauchy-Schwartz inequality, we can define the concept of an angle between vectors:

$$\cos\theta = \frac{\langle x, y \rangle}{\|x\| \cdot \|y\|} \quad \text{for all } x, y \in V.$$

Various angle definitions exist for complex-valued inner products.

Parallelogram law: If V is an inner product space and $\|\cdot\|$ is the induced norm on V, then

$$\|x+y\|^2 + \|x-y\|^2 = 2\|x\|^2 + 2\|y\|^2 \quad \forall x, y \in V$$

If the norm of a normed space satisfies the parallelogram law, then it is induced by an inner product.

$\Rightarrow$ ℓ_2 or L_p spaces are IPS's iff $p = 2$.

Orthogonality: Two vectors $x, y \in V$ are said to be orthogonal if

$$\langle x, y \rangle = 0$$

which is denoted by $x \perp y$.

Similarly, x & y are called parallel if $x = ay$ for some $a \in F$.

Orthogonality of two sets is defined by the orthogonality of their elements:

$$S \perp T \leftrightarrow s \perp t \quad \forall s \in S, \forall t \in T.$$

Orthogonal complement: The orthogonal complement of a set $S \subset V$ is defined as

$$S^{\perp} = \{ x \in V \mid x \perp s \ \ \forall s \in S \}$$

Orthonormal vectors: A set of vectors $\{x_1, \ldots, x_n\}$ is orthonormal if

- $x_i \perp x_j$ for all $i \neq j$
- $\|x_i\| = 1$ for all i.

Gram - Schmidt Orthonormalisation Process: Let $x_1, \ldots, x_n$ be a set of linearly independent vectors. Then one can find an orthogonal sequence inductively as follows:

$$e_1 = x_1$$

$$e_i = x_i - \sum_{j=1}^{i-1} \langle x_i, e_j \rangle e_j \quad \text{for all } 1 < j \leq n$$

If one divides each of these vectors by their norms, one obtains an orthonormal set.

$$u_i = \frac{e_i}{\|e_i\|} \quad \text{for all } 1 \leq i \leq n$$

The benefit of an orthonormal basis is that representation coefficients can be found by a simple inner product operation:

$$x = \sum_{k=1}^{n} \beta_k u_k, \quad \Rightarrow \quad \beta_k = \langle x, u_k \rangle$$

and thus

$$x = \sum_{k=1}^{n} \langle x, u_k \rangle u_k$$

Operator: An operator is a mapping between vector spaces.

Linear operator: An operator $A: X \to Y$ is linear if

$$A(\alpha x + \beta y) = \alpha A(x) + \beta A(y)$$

or equivalently the two following conditions:

- Homogeneity: $A(\alpha x) = \alpha A(x)$
- Additivity: $A(x+y) = A(x) + A(y)$

Range space: The range space of an operator $A: X \to Y$ is defined as

$$R(A) = \{ y \in Y \mid \exists x \in X \; A(x) = y \} \subset Y$$

If $R(A) = Y$ for some operator A, then A is said to be a surjection, or onto/surjective.

Null space: The null space of an operator $A: X \to Y$ is defined as

$$N(A) = \{ x \in X \mid A(x) = 0 \} \subset X$$

If $N(A) = \{0\}$ for some operator A, then A is said to be an injection, or one-to-one/injective.

GUPTA

Both $N(A)$ & $R(A)$ are subspaces.

Inverse problem: Find x, given the operator A & $y = A(x)$.

The inverse problem has at least one solution if A is surjective.

It has at most one solution if A is injective.

It has a unique solution if A is surjective and injective, i.e. bijective.

Linear operators in finite dimensional spaces: An operator A between finite dimensional spaces always has a matrix representation.

Say e is a basis for X, v a basis for Y, and $A: X \to Y$ a linear operator. Then

$$[A(x)]_v = \underline{\underline{A}}\,[x]_e$$

where

$$\underline{\underline{A}} = \left[[A(e_1)]_v \; [A(e_2)]_v \; \cdots \; [A(e_n)]_v \right]$$

If we were to change the basis, from e to e' and v to v', then we can have a new matrix representation $\underline{\underline{A}}'$ as

$$\underline{\underline{A}}'_{e' \to v'} = T_{v \to v'}\, \underline{\underline{A}}_{e \to v}\, T_{e' \to e}$$

so that

$$[A(x)]_{v'} = \underline{\underline{A}}'\,[x]_{e'}$$

Matrix Rank: The rank for a matrix $A \in \mathbb{C}^{M \times N}$ is the number of linearly independent columns of A.

We always have

$$\text{rank}(A) \leq \min\{M, N\}$$

Rank is also defined as the dimensionality of the range space of some matrix A.

Null-space Dimensionality: For a matrix $A \in \mathbb{C}^{M \times N}$

$$\dim(N(A)) = N - \text{rank}(A)$$

Rank-Nullity Theorem: Let $A \in \mathbb{C}^{M \times N}$

$$\dim(X) = \text{rank}(A) + \dim(N(A))$$

Projectors: $\hat{x} \in S \subset V$ is the projection of some $x \in V$ if

$$\hat{x} = \underset{s \in S}{\operatorname{argmin}} \|x - s\|^2$$

Orthonormal bases to Projections: Let $b_1, \ldots, b_n$ be an orthonormal basis for a subspace $S \subset V$. Then the projection matrix $\underline{\underline{P}}$ is given by

$$\underline{\underline{P}} = \underline{\underline{B}}\,\underline{\underline{B}}^H$$

where

$$B = [b_1 \quad b_2 \quad \ldots \quad b_n]$$

If b_i's are a basis for S, although not orthonormal, we have

$$\underline{\underline{P}} = B(B^H B)^{-1} B^H$$

Additional Content:

- Hölder's Inequality: Let $p, q \in (1, \infty)$, $1/p + 1/q = 1$

$$\int |f(t) g(t)|\, dt = \left(\int |f(t)|^p dt \right)^{1/p} \left(\int |g(t)|^q dt \right)^{1/q}$$

- A matrix (or a transform) U is called unitary if the columns of U form an orthonormal basis. In this case,

$$U^H U = I$$

A unitary transformation / matrix preserves the inner product, and hence the norm:

$$\langle Ux, Uy \rangle = \langle x, y \rangle \quad \forall x, y \in V$$

$$\|Ux\| = \|x\|$$

EE499 Cheat Sheet - Midterm 2

Projectors in $\mathbb{C}^N$: Let $x \notin S \subsetneq \mathbb{C}^N$, S a subspace of $\mathbb{C}^N$. Then we call $\hat{x}$ a projection of x onto S if

$$\hat{x} = \underset{s \in S}{\operatorname{argmin}} \|x - s\|^2$$

so $\hat{x}$ is the closest vector in S to x.

If $\underline{b}_1, \ldots, \underline{b}_r$ is a basis (hence linearly independent) for S, we can represent the projection operation as a matrix multiplication:

$$\hat{x} = Px \quad \text{where}$$

$$P = B(B^H B)^{-1} B^H \quad \text{for} \quad B = [\underline{b}_1 \ \ldots \ \underline{b}_r]$$

If further the basis is orthonormal, then

$$P = B\underbrace{(B^H B)^{-1}}_{I} B^H = BB^H$$

Projection theorem in IPS: Let X be an IPS, S a subspace of X, $x \in X$. If there is a vector $\hat{x} \in S$ such that

$$\|x - \hat{x}\| \leq \|x - s\| \quad \forall s \in S$$

then such an $\hat{x}$ is unique. A necessary & sufficient condition that $\hat{x} \in S$ be the unique minimizing vector in S is that the error vector $e = x - \hat{x}$ be orthogonal to S. This is called as the orthogonality principle.

In an arbitrary inner product space, a projection may not exist but if it does, it is unique & fully characterized by the orthogonality principle.

Completeness: A space X is said to be complete if every Cauchy sequence is convergent in X. A vector sequence $\{x_n\}_{n=1}^{\infty}$ in a normed vector space is Cauchy if

$$\lim_{m,n \to \infty} \|x_n - x_m\| = 0$$

Every convergent sequence is Cauchy.

A complete IPS is called a Hilbert space. All finite dimensional spaces, ℓ_2 & $L_2[a,b]$ are Hilbert spaces.

Projection theorem in Hilbert spaces: Let X be a Hilbert space. Then the projection operator P_S (or equivalently a projection $\hat{x}$) exists for a subspace S if and only if S is closed or equivalently complete.

Because $\mathbb{C}^N$ and a subspace $S \subseteq \mathbb{C}^N$ are finite dimensional, we are sure that a projection always exists & is unique and complies with the orthogonality principle. Furthermore, knowing that a projector is a linear operator, it has a matrix representation, as derived earlier.

3 methods of computing projections:

1) Optimization based arguments: Define the cost function J as

$$J: S \longrightarrow \mathbb{R}, \quad s \longrightarrow \|x - s\| \quad \text{or} \quad \|x-s\|^2$$

and minimize it by solving

$$\hat{x} = \underset{s \in S}{\operatorname{argmin}}\, J(s) \iff \nabla J(\hat{x}) = 0$$

2) Find $\hat{x}$ using the orthogonality principle, which provides a system of equations.

3) Find the projection matrix P for S & compute $\hat{x} = Px$.

Solving linear systems of Equation:

$x \in \mathbb{C}^N \rightarrow \boxed{\underline{\underline{A}}_{M \times N}} \rightarrow y \in \mathbb{C}^M$

Given y & $\underline{\underline{A}}$, find x such that $y = Ax$.

- Existence of a solution: For every $y \in \mathbb{C}^M$, there exists at least one solution iff

CASE 2

$\Leftrightarrow R(A) = \mathbb{C}^M$

$\Leftrightarrow \operatorname{rank}(A) = \dim(\overbrace{R(A)}^{\mathbb{C}^M}) = M$ (# of rows)

$\Leftrightarrow$ rows of A are linearly independent (A has full row rank)

One necessary condition is to have $M \leq N$, i.e. A be a square or fat matrix.

- Uniqueness of a solution: If there exists a solution, it is unique iff

CASE 1

$\Leftrightarrow N(A) = \{0\}$

$\Leftrightarrow \operatorname{rank}(A) = N$

$\Leftrightarrow$ columns of A are linearly independent (A has full column rank)

One necessary condition for uniqueness is to have $N \leq M$, i.e. A be a square or a tall matrix.

- Existence & Uniqueness: For every $y \in \mathbb{C}^M$, there exists a unique solution iff

CASE 0

$\Leftrightarrow R(A) = \mathbb{C}^M$ & $N(A) = \{0\}$

$\Leftrightarrow$ rank $(A) = M = N$

$\Leftrightarrow$ rows of A are linearly independent & columns of A are linearly independent

$\Leftrightarrow A$ is invertible.

A necessary condition is to have $M = N$, so A be a square matrix.

- If conditions of both Case 1 & 2 are violated, existence & if so uniqueness are not guaranteed. (CASE 3)

Basics of matrix calculus:

- $\nabla_{\underline{x}} (\underline{a}^T \underline{x}) = \nabla_{\underline{x}} (\underline{x}^T \underline{a}) = \underline{a}$ for any vector or (square) matrix a
- $\nabla_{\underline{x}} (x^T B x) = Bx + B^T x = (B + B^T) x$

Case 0: Least squares solution (derived for real valued case)

- $J(x) = \|y - Ax\|^2 = (y - Ax)^T (y - Ax)$

$= y^T y - y^T A x - x^T A^T y + x^T A^T A x$

$\nabla_x (J(x)) = -(y^T A)^T - A^T y + (A^T A + (A^T A) x = 0$

$-2A^T y + 2A^T A \hat{x}_{LS} = 0$

$A^T A x = A^T y \Rightarrow \hat{x}_{LS} = (A^T A)^{-1} A^T y$ → NORMAL EQUATION

($A^T A$ invertible as A has linearly independent cols (Case 0))

For complex-valued case:

$A^H A \hat{x}_{LS} = A^H y \Rightarrow \hat{x}_{LS} = (A^H A)^{-1} A^H y$

- One can derive the normal equation using projection based arguments as well.

Equivalently define (looking for a projection of y onto $R(A)$)

$$\hat{b}_{LS} = \underset{b \in R(A)}{\operatorname{argmin}} \|y - b\|^2 \quad \leftarrow \quad \hat{b}_{LS} = A\hat{x}_{LS}$$

By the orthogonality principle, $e \perp R(A) = \text{span}\{a_1, \dots, a_n\}$ ($\hookrightarrow A = [a_1 \cdots a_n]$)

$$e \perp a_k \quad \forall k$$

$$\langle e, a_k \rangle = a_k^H e = 0 \quad \forall k$$

$$\begin{bmatrix} a_1^H e \\ a_2^H e \\ \dots \\ a_N^H e \end{bmatrix} = \begin{bmatrix} a_1^H \\ a_2^H \\ \dots \\ a_N^H \end{bmatrix} e = 0 \quad \hookrightarrow A^H$$

$$\Rightarrow A^H e = 0 \quad \leftarrow \text{insert } e = y - \hat{b}_{LS} = y - A\hat{x}_{LS}$$

$$A^H(y - A\hat{x}_{LS}) = 0$$

$$A^H A\hat{x}_{LS} = A^H y \Rightarrow \hat{x}_{LS} = (A^H A)^{-1} A^H y$$

$$\hat{b}_{LS} = \underbrace{A(A^H A)^{-1} A^H}\, y$$

Projection matrix onto $R(A)$!

Case 2: At least one solution exists, the problem is to pick one. If x_p is a solution, elements of $x_p + N(A)$ are all solutions. So, we constrain our solution further. One possible choice is to have a minimum-norm solution.

$$\hat{x}_{MN} = A^H (AA^H)^{-1} y$$

Case 3: We deal with both issues. We can find a least squares solution through the normal equation, but that solution is not unique either, so we may enforce a minimum norm solution as well. So the problem becomes

$$\hat{x}_{MNLS} = \underset{x \text{ s.t. } A^H A x = A^H y}{\operatorname{argmin}} \|x\|_2$$

which turns out to have a closed form solution, which is

$$\hat{x}_{MNLS} = A^+ y$$

where A^+ denotes the pseudo-inverse of A.

Spectral Decomposition Theorem: Let $A \in \mathbb{C}^{n \times n}$ be a Hermitian symmetric matrix. Then there exist a unitary matrix $U \in \mathbb{C}^{n \times n}$ and a real diagonal matrix Λ such that

$$A = U \Lambda U^H = [u_1 \cdots u_n] \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix} \begin{bmatrix} u_1^H \\ \cdots \\ u_n^H \end{bmatrix}$$

$$= \sum_{k=1}^{n} \lambda_k u_k u_k^H$$

λ_k's are the eigenvalues of A & u_k's are the orthonormal eigenvectors of A. We then conclude the following facts:

- $Au_k = \lambda_k u_k$
- $U^H U = I$
- $AU = U\Lambda$
- $\Lambda = U^H A U$
- $\text{rank}(A) = \#$ of non-zero eigenvalues.

The importance is that when multiplying with such a matrix, U^H makes a transition into some other basis of $\mathbb{C}^n$ and the matrix multiplication in that domain becomes elementwise multiplication. We return back to the original basis after the elementwise multiplication using $U = (U^H)^{-1}$.

Singular Value Decomposition Theorem: Let $A \in \mathbb{C}^{m \times n}$ and $\text{rank}(A) = r$, $r \leq m$ & $r \leq n$. Then there exist unitary matrices $U \in \mathbb{C}^{m \times r}$ & $V \in \mathbb{C}^{n \times r}$ and a diagonal matrix $\Sigma = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_r)$ with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ such that

$$A = U \Sigma V^H = [u_1 \cdots u_r] \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{bmatrix} \begin{bmatrix} v_1^H \\ \cdots \\ v_r^H \end{bmatrix}$$

$$= \sum_{i=1}^{r} \sigma_i u_i v_i^H$$

Unitariness here for matrices of arbitrary size is

$$U^H U = I$$
$$V^H V = I$$

so we may indeed have $UU^H \neq I$ & $VV^H \neq I$.

Comparing SVD & Spectral Decomposition:

Spectral Decomposition	Singular Value Decomposition
• $A \in \mathbb{C}^{n\times n}$, $A^H = A$	• $A \in \mathbb{C}^{m\times n}$, no restriction
• $A = U \Lambda U^H = \sum_{k=1}^{n} \lambda_k u_k u_k^H$	• $A = U_{m\times r} \Sigma_{r\times r} (V_{n\times r})^H = \sum_{k=1}^{r} \sigma_k u_k v_k^H$
• $U = [u_1 \; \cdots \; u_n] \in \mathbb{C}^{n\times n}$ ↳ eigenvectors of A	• $U = [u_1 \; \cdots \; u_r] \in \mathbb{C}^{m\times r}$ ↳ left singular vectors of A, eigenvectors of AA^H; $V = [v_1 \; \cdots \; v_r] \in \mathbb{C}^{n\times r}$ ↳ right singular vectors of A, eigenvectors of A^HA.
• $U^HU = I_n = UU^H$	• $U^HU = I$, possibly $UU^H \neq I$; $V^HV = I$, possibly $VV^H \neq I$.
• $A = U\Lambda U^H \Leftrightarrow \Lambda = U^HAU$ ↳ Applying $A \Leftrightarrow$ elt.wise multiplication in U domain	• $R(U) = R(A)$, $P_{R(A)} = UU^H$; $R(V) = R(A^H)$, $P_{R(A^H)} = VV^H$

Extended SVD: In this formulation $A \in \mathbb{C}^{m\times n}$ is expanded such that 0 singular values are also allowed and the singular vector matrices are square and fit the regular unitary matrix property. So,

$$A = \tilde{U}_{m\times m} \tilde{\Sigma}_{m\times n} \left(\tilde{V}_{n\times n}\right)^H$$

$$= \underbrace{[U_{m\times r} \; u_{r+1} \; \cdots \; u_m]}_{\tilde{U}_{m\times m}} \underbrace{\begin{bmatrix} \Sigma_{r\times r} & 0 \\ 0 & 0 \end{bmatrix}}_{\tilde{\Sigma}_{m\times n}} \underbrace{[V_{n\times r} \; v_{r+1} \; \cdots \; v_n]}_{\tilde{V}_{n\times n}}^H$$

$$= \sum_{k=1}^{r} \sigma_k u_k v_k^H$$

with

$$\tilde{U}^H\tilde{U} = \tilde{U}\tilde{U}^H = I_m$$

$$\tilde{V}^H\tilde{V} = \tilde{V}\tilde{V}^H = I_n$$

Pseudo-inverse (Moore-Penrose): Let $A \in \mathbb{C}^{m\times n}$ and $\text{rank}(A) = r$. Then the pseudo inverse is constructed using SVD as

$$A = U\Sigma V^H \longrightarrow A^+ = V\Sigma^{-1}U^H$$

where

$$\Sigma^{-1} = \begin{bmatrix} 1/\sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1/\sigma_r \end{bmatrix}$$

As a closed term, we get

$$A = \sum_{k=1}^{r} \sigma_r u_k v_k^H \longrightarrow A^+ = \sum_{k=1}^{r} \frac{1}{\sigma_r} v_k u_k^H$$

A corollary of this definition is

$$A^+ = A^H (AA^H)^+ = (A^H A)^+ A^H$$

In case A is square and has an inverse, A^+ is an inverse to A. As multiplicative inverses are unique, we obtain

$$A \in \mathbb{C}^{n \times n} \;\&\; \det(A) \neq 0 \longrightarrow A^{-1} = A^+$$

Case-by-case pseudoinverse:

- Case 0: As A is square and nonsingular

$$A^+ = A^{-1} \qquad \Rightarrow \hat{x} = x = A^+ y = A^{-1} y$$

- Case 1: A has full column rank

$$A^+ = (A^H A)^{-1} A^H = A_{LS} \qquad \Rightarrow \hat{x}_{LS} = A^+ y$$

- Case 2: A has full row rank

$$A^+ = A^H (AA^H)^{-1} = A_{MN} \qquad \Rightarrow \hat{x}_{MN} = A^+ y$$

- Case 3: A has neither full row rank nor full column rank

$$A^+ = V \Sigma^{-1} U^H = A_{MNLS} \qquad \Rightarrow \hat{x}_{MNLS} = A^+ y$$

$\Rightarrow$ Therefore no matter in which case we are, one can use the pseudoinverse to find the solution to the inverse problem!

Sensitivity & Conditioning of linear inverse Problems: There may be errors on y and on A, due to

- physical measurement errors (noise)
- modelling errors
- numerical implementation / storage errors (finiteness)

How can we quantify the error in our solution to the inverse problem?

Define the 2-norm condition number as

$$\kappa(A) = \frac{\sigma_{max}}{\sigma_{min}}$$

where σ_{max} & σ_{min} are the maximum & minimum singular values of A, respectively.

We can observe that $K(A) \geq 1$, and $K(A) = 1$ iff all the singular values of A are the same.

- If $K(A) \approx 1$, then A is a well-conditioned matrix
- If $K(A) \gg 1$, then A is an ill-conditioned matrix, meaning it may amplify the error upon inverse problem solution.

- Case 0: We have non-singular & square A, & $x = A^{-1}y$. Then

$$\frac{\|\delta x\|_2}{\|x\|_2} \leq K(A) \frac{\|\delta y\|_2}{\|y\|_2}$$

relative error in the solution x

$$\frac{\|\delta x\|_2}{\|x\|_2} \leq K(A) \left[\frac{\|\delta y\|_2}{\|y\|_2} + \frac{\|\delta A\|_2}{\|A\|_2} \right]$$

- Cases 1-3: $\hat{x} = A^{+}y$. Similar bounds exist for these cases, and they again increase with $K(A)$.

Orthogonality principle in IPS (revisited): Say that

$$y = \sum_{i=1}^{n} c_i a_i + e$$

for some y. Then if we minimize e, $\sum c_i a_i$ becomes the projection of y onto $\text{span}\{a_1, \dots, a_n\}$. As we know in such a case, we require e to be perpendicular to $\text{span}\{a_i\}$, as per the orthogonality principle dictates. This gives us a certain set of equations we may solve to obtain the optimal c_i values, i.e. the coordinates of the projection $\hat{y}$ in the $\{a_i\}_{i=1}^{n}$ basis.

$$\langle y - \sum_{i=1}^{n} c_i a_i, a_k \rangle = 0 \quad \forall k$$

$$\langle y, a_k \rangle = \sum_{i=1}^{n} c_i \langle a_i, a_k \rangle \quad \forall k$$

$$\langle y, a_k \rangle = \sum c_i a_k^{H} a_i \quad \forall k$$

$$\begin{bmatrix} \langle y, a_1 \rangle \\ \langle y, a_2 \rangle \\ \dots \\ \langle y, a_n \rangle \end{bmatrix} = \begin{bmatrix} G_{ij} = \langle a_j, a_i \rangle \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix} \rightarrow \text{optimal } c_i\text{'s}$$

$$d = G \cdot c_{opt}$$

"Gram Matrix"

If a_i's are linearly independent, G is invertible, and so

$$c_{opt} = G^{-1} d$$

Gram Matrix: For a set of vectors $\{a_i\}_{i=1}^n$, the Gram matrix is defined as

$$G_{ij} = \langle a_j, a_i \rangle$$

G is Hermitian symmetric ($G^H = G$) and is positive semi-definite. If further $\{a_i\}_{i=1}^n$ is linearly independent, G is positive definite and hence invertible. Furthermore if $\{a_i\}$'s are orthogonal, G is a diagonal matrix.

The Gram matrix method can be used to derive the projection matrix as well. Note that

$$d = A^H y \quad \& \quad G = A^H A$$

and as c_{opt} are the coordinates in $\{a_i\}_{i=1}^n$ basis, (supposing they are),

$$\hat{y} = A c_{opt} = A G^{-1} d = \underbrace{A(A^H A)^{-1} A^H}_{P} y = Py$$

VECTOR SPACE OF RANDOM VARIABLES

IPS of random variables: Let $\{y_i\}_{i=1}^m$ be a finite collection of real valued random variables, with $E[y_i^2] < \infty \ \forall i$. Then $V = \text{span}\{y_i\}_{i=1}^m$ is an inner product space under the following inner product

$$\langle x, y \rangle = E[xy]$$

$$\|x\| = \sqrt{E[x^2]}$$

We treat $x = y$ iff $P(X(\omega) \neq Y(\omega), \omega: \text{outcome}) = 0$, so $x = y$ a.e.

The usual IPS & RV facts look like the following

- $|E[xy]|^2 \leq E[x^2]\,E[y^2]$ $\Rightarrow$ Cauchy-Schwartz ineq.
- $\sqrt{E[(x+y)^2]} \leq \sqrt{E[x^2]} + \sqrt{E[y^2]}$
- $\rho_{xy} = \dfrac{cov(x,y)}{\sigma_x \sigma_y} = \dfrac{E[(x-\mu_x)(y-\mu_y)]}{\sqrt{E[(x-\mu_x)^2]\,E[(y-\mu_y)^2]}} = \cos\theta$

  θ is the angle between $x-\mu_x$ & $y-\mu_y$
- $x \perp y \iff E[xy] = 0$
- Uncorrelated: $E[(x-\mu_x)(y-\mu_y)] = 0$ i.e. $x-\mu_x \perp y-\mu_y$

For zero mean r.v's, uncorrelated $\iff$ orthogonal.

Decorrelation / Whitening. It is the process of obtaining rv's that are uncorrelated.

1) Let x, y be random variables. Then u, v given by the following equation are orthogonal

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 1/\|x\| & 0 \\ 0 & 1/\|y\| \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

2) Gram-Schmidt orthogonalization is another way of decorrelating a given set of rv's. If they are also zero mean, the process is also whitening.

Gram matrix: For random variables $x_1, \dots, x_N$, define $x = [x_1, \dots, x_N]^T$

$$G = \begin{bmatrix} \langle x_1, x_1 \rangle & \langle x_2, x_1 \rangle & \cdots & \langle x_N, x_1 \rangle \\ \langle x_1, x_2 \rangle & \langle x_2, x_2 \rangle & \cdots & \langle x_N, x_2 \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle x_1, x_N \rangle & \langle x_2, x_N \rangle & \cdots & \langle x_N, x_N \rangle \end{bmatrix}$$

$$= \begin{bmatrix} E[x_1 x_1] & E[x_2 x_1] & \cdots & E[x_N x_1] \\ E[x_1 x_2] & E[x_2 x_2] & \cdots & E[x_N x_2] \\ \vdots & \vdots & \ddots & \vdots \\ E[x_1 x_N] & E[x_2 x_N] & \cdots & E[x_N x_N] \end{bmatrix} = R_x$$

↳ autocorrelation matrix of x

$$= E[x x^T]$$

The Gram matrix is again Hermitian symmetric, & positive semi-definite. If further all vectors are linearly independent, G is positive definite & hence invertible.

LMMSE: Suppose you make $\{y_i\}_{i=1}^N$ observations of a random variable x and want to estimate it, by a linear sum $\alpha_1 y_1 + \cdots + \alpha_N y_N$. Minimize the error by minimizing $\|x - \hat{x}\|^2 = \|e\| = E[(x-\hat{x})^2]$. This amounts to projecting x onto span $\{y_i\}$.

$$e \perp \text{span}\{y_i\} \Rightarrow e \perp y_i \ \forall i$$

By the same derivation with Gram matrix previously,

$$G\alpha = d = \begin{bmatrix} \langle x, y_1 \rangle \\ \langle x, y_2 \rangle \\ \vdots \\ \langle x, y_N \rangle \end{bmatrix}$$

$G \to R_y$, $d \to r_{xy}$

$$\alpha_{opt} = R_y^{-1} r_{xy} \Rightarrow \hat{x}_{opt} = \alpha_{opt}^T y$$

LMMSE estimator

Matrix inverse: $A^{-1} = \frac{1}{\det(A)} ((-1)^{ij} M_{ij})^T$

↳ cofactor

EE499 Cheat Sheet - Final

MSE of LMMSE Estimator:

$$E[(x-\hat{x})^2] = E[e^2] = \|e\|^2 = \|x\|^2 - \|\hat{x}\|^2 = E[x^2] - E[\hat{x}^2] \quad (\text{because } e \perp \hat{x})$$

$$= E[x^2] - E[\alpha^T y y^T \alpha]$$

$$= E[x^2] - \alpha^T E[yy^T]\alpha = E[x^2] - \alpha^T R_y \alpha \qquad (\alpha = R_y^{-1} r_{xy})$$

$$= E[x^2] - r_{xy}^T R_y^{-1} R_y R_y^{-1} r_{xy}$$

$$= E[x^2] - r_{xy}^T R_y^{-1} r_{xy}$$

$$= E[x^2] - r_{xy}^T \alpha = E[x^2] - \alpha^T r_{xy}$$

Vector LMMSE estimation: Say we have $x_1, \dots, x_M$ many r.v.'s, which we observe through $y_1, \dots, y_N$. We know that y_i's may depend on possibly all x_i's. We will estimate x_i's through y_i's, using a linear model as follows:

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_M \end{bmatrix}_{M\times 1}, \quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}_{N\times 1}, \quad k_i = \begin{bmatrix} k_{i1} \\ \vdots \\ k_{iN} \end{bmatrix}_{N\times 1}$$

$$x_i = k_i^T y \quad (i = 1, \dots, M) \Rightarrow \hat{x} = \underbrace{\begin{bmatrix} -k_1^T- \\ \vdots \\ -k_M^T- \end{bmatrix}}_{K} \underbrace{\begin{bmatrix} | \\ y_1 \\ \vdots \\ y_N \\ | \end{bmatrix}}_{y} \qquad \hat{x} = Ky$$

Minimize MSE: $E[e^T e] = E\left[\sum_{i=1}^{M} (x_i - k_i^T y)^2\right] = \sum_{i=1}^{M} E[(x_i - k_i^T y)^2]$

$\Rightarrow$ minimize each $E[(x_i - k_i^T y)]$ equivalently $\forall i$

$\Rightarrow R_y k_i = r_{x_i y} \quad \forall i$

Stack these equations horizontally

$$R_y \underbrace{[k_1 \; k_2 \; \cdots \; k_M]}_{K^T} = \underbrace{[r_{x_1 y} \; r_{x_2 y} \; \cdots \; r_{x_M y}]}_{R_{xy}^T}, \quad R_{xy} = E[x y^T]$$

$$R_y K^T = R_{xy}^T$$

$$\Rightarrow K R_y = R_{xy} \quad (\text{WIENER-HOPF EQUATION})$$

If R_y is invertible,

$$K_{opt} = R_{xy} R_y^{-1}$$

What's the error for this estimator? (MSE = trace(R_e))

$$R_e = E[ee^T] = E[e(x-\hat{x})^T] =$$
$$= E[ex^T] - E[e\hat{x}^T] \quad (E[e\hat{x}^T] \to 0 \text{ due to } \perp \text{ principle})$$
$$= E[(x-Ky)x^T] = E[xx^T] - E[Kyx^T]$$
$$= R_x - K R_{yx} = R_x - R_{xy} R_y^{-1} R_{yx}$$

Now, say our observations y are of the form $y = Ax + n$

$$R_y = E[yy^T] = E[(Ax+n)(Ax+n)^T]$$
$$= A E[xx^T] A^T + E[nn^T] + \cdots \quad (\cdots \to 0)$$
$$= A R_x A^T + R_n$$

$$R_{xy} = E[xy^T] = E[x(Ax+n)^T] = E[xx^T A^T] + E[xn^T] \quad (E[xn^T] \to 0)$$
$$= R_x A^T$$

So the K estimator is, by Wiener-Hopf eqn.

$$K = R_x A^T (A R_x A^T + R_n)^{-1}$$

Then our estimations are

$$\hat{x} = R_x A^T (A R_x A^T + R_n)^{-1} y$$

By the matrix inversion lemma, this is equivalently

$$\hat{x} = (A^T R_n^{-1} A + R_x^{-1})^{-1} A^T R_n^{-1} y$$

Maximum Likelihood Estimation: Treat the unknown x as deterministic, observations as random.

$$\hat{x}_{ML} = \arg\max_x \underbrace{p(y|x)}_{\text{likelihood function}} = \arg\max_x \underbrace{\log p(y|x)}_{\text{log-likelihood function}}$$

ML estimators with different noise statistics:

- iid Gaussian: $\hat{x}_{ML} = \arg\min \|y - Ax\|_2^2 = \hat{x}_{LS}$
- correlated Gaussian: $\hat{x}_{ML} = \arg\min_x (y-Ax)^T \Sigma^{-1} (y-Ax)$
$$= \arg\min_x \|y - Ax\|_{\Sigma^{-1}}$$
↳ "weighted norm"

- independent Poisson:
- iid Generalized Gaussian: $\hat{x}_{ML} = \arg\min_x \|y - Ax\|_p$

Gaussian distribution reminder:

$$p_{\underset{\sim}{x}}(x) = \frac{1}{\sqrt{2\pi\sigma_x^2}} \exp\left(-\frac{1}{2}(x-\mu_x)\frac{1}{\sigma_x^2}(x-\mu_x)\right) = \frac{1}{\sqrt{2\pi\sigma_x^2}} \exp\left(-\frac{(x-\mu_x)^2}{2\sigma_x^2}\right)$$

$$p_{\underset{\sim}{\underline{x}}}(\underline{x}) = \frac{1}{\sqrt{(2\pi)^M |\Sigma|}} \exp\left(-\frac{1}{2}(\underline{x}-\underline{\mu})^T \Sigma^{-1} (\underline{x}-\underline{\mu})\right)$$

$\underline{x} \in \mathbb{C}^M$

Bayesian Estimation: x & y are both random

$$\hat{x}_{Bayesian} = \arg\min_{\hat{x}} E[cost(\hat{x}, x)]$$

↳ MMSE: $cost(\hat{x}-x) = (\hat{x}-x)^2$

MAP: $cost(\hat{x}-x) = \begin{cases} 0, & |\hat{x}-x| \le \Delta \\ 1 & |\hat{x}-x| > \Delta \end{cases}$

MMSE Estimation:

$$\hat{x}_{MMSE} = \arg\min_{\hat{x}} E[(\hat{x}-x)^2] = E[x|y]$$

MAP Estimation:

$$\hat{x}_{MAP} = \arg\max_x p(x|y) = \arg\max_x \frac{p(y|x)p(x)}{p(y)}$$

$$= \arg\max_x \log \frac{p(y|x)p(x)}{p(y)} \quad \text{(discard } p(y)\text{)}$$

$$= \arg\max_x \log p(y|x) + \log p(x)$$

$$= \arg\min_x -\log p(y|x) - \log p(x)$$

General form of MAP estimators

$$\hat{x}_{MAP} = \arg\min_x \Delta(y, Ax) + \lambda R(x)$$

$\Delta(y, Ax)$: model mismatch term / data fidelity term

$R(x)$: ↳ regularization term

λ: → regularization parameter.

Unconstrained optimization: Say $J: \mathbb{R}^N \to \mathbb{R}$ is an objective/cost function. Then this problem is:

$$\hat{x} = \arg\min_x J(x)$$

Convexity: A cost function is called convex if

$$\alpha J(x_1) + (1-\alpha) J(x_2) \geq J(\alpha x_1 + (1-\alpha) x_2)$$

$\forall x_1, x_2 \quad \alpha \in [0,1]$

If a cost function is convex, then it has no local minima other than the global minimum. So we can use any local optimization method at hand.

Local minimization methods: The algorithm is as follows:

$n = 0$
$x^{(n)} = x^{(0)}$ (initial guess)
repeat
  compute $d^{(n)}$, an update direction for x
  compute $\tau^{(n)}$, a step size
  $x^{(n+1)} = x^{(n)} + \tau^{(n)} d^{(n)}$
  $n = n+1$
until convergence
$\hat{x} = x^{(n)}$

Steepest Descent: Pick $d^{(n)}$ in the opposite direction of the gradient:

$$d^{(n)} = -\frac{\nabla J(x^{(n)})}{\| \nabla J(x^{(n)}) \|_2}$$

Newton's method: Use second order approximation of J to obtain faster convergence.

$$J(x + \tau d) \approx J(x) + \tau d^T \nabla J(x) + \frac{\tau^2}{2} d^T \nabla^2 J(x) d \quad \text{(goal} \downarrow \text{)} < J(x)$$

Minimize wrt. d & choose $\tau = 1$

$$\nabla_d \left(J(x) + \tau d^T \nabla J(x) + \frac{\tau^2}{2!} d^T \nabla^2 J(x) d \right) = 0$$

$$\nabla J(x) + \frac{1}{2} \cdot 2 \cdot \nabla^2 J(x) d = 0$$

$$\Rightarrow d = -\left(\nabla^2 J(x^{(n)})\right)^{-1} \nabla J(x^{(n)})$$

$$\Rightarrow x^{(n+1)} = x^{(n)} - \left(\nabla^2 J(x^{(n)})\right)^{-1} \nabla J(x^{(n)})$$

Quasi-Newton methods: Choose this matrix as something else, say B:

$$x^{(n+1)} = x^{(n)} - B \nabla J(x^{(n)})$$

Conjugate Gradient methods: Update the direction as well:

$$d^{(0)} = -\nabla J(x^{(0)})$$

$$x^{(n+1)} = x^{(n)} + \tau d^{(n)}$$

$$d^{(n+1)} = -\nabla J(x^{(n+1)}) + \beta^{(n)} d^{(n)}$$

Choosing the step-size parameter τ Ideally

$$\tau^{(n)} = \underset{\tau}{\operatorname{argmin}}\, J(x^{(n)} + \tau d^{(n)})$$

but this is computationally costly, unnecessary to guarantee convergence and there are approximate methods of choosing τ

Global Optimization methods:

- Deterministic methods
  - ↳ Exhaustive Search,
  - ↳ Branch & Bound,
  - ↳ Dynamic programming, ---
- Stochastic methods
  - ↳ Simulated annealing,
  - ↳ Genetic Algorithms, ---