3. Linear Algebra

Basics of Linear Algebra needed for Optimization Theory

Sun Woo Lim

Jan 6, 2021

Vectors

Physical vectors are objects to represent a **collection of numbers.**

- Column Vector notation:

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

  $x_i, i = 1, 2, 3, ..., n$ are called **"components"** of x and n is called a **"dimension"** of x.

- When $x_i, i = 1, 2, 3, ..., n \in \mathbb{R}$, say $x \in \mathbb{R}^n$.
  More generally, when $x_i \in \mathbb{C}$, say $x \in \mathbb{C}^n$. In this topic, only deal with $x \in \mathbb{R}^n$

- **Transpose**

$$x^T = \begin{bmatrix} x_1 & x_2 & x_3 & \cdots & x_n \end{bmatrix}$$

$$x^{TT} = x$$

Vector Spaces

More richer understanding of vectors comes from the understanding of **"vector spaces"**

A Vector Space is a tuple $\chi = (V, F)$ that is closed under scalar multiplication and vector addition, where V is a set of vectors and F is a field of scalars. In this topic, only deal with $F = \mathbb{R}$.
A representative example of a vector space : $V = \mathbb{R}^n$ with usual definition of vector addition and scalar multiplication.

8 Axioms of Vector Spaces
For any $x, y, z \in V$ and $a, b \in F$,

1. **Associativity of Addition** $x + (y + z) = (x + y) + z$
2. **Commutativity of Addition** $x + y = y + x$
3. **Identity element of addition** $\exists 0 \in V \colon x + 0 = 0 + x = x, \forall x \in V$
4. **Inverse elements of addition** $\forall v \in V, \exists -v : -v + v = 0$
5. **Compatibility of scalar multiplication with field multiplication** $a(bx) = (ab)x$
6. **Identity element of scalar multiplication** $1v = v$
7. **Distribution of scalar multiplication w.r.t. vector addition** $a(x + y) = ax + ay$
8. **Distribution of scalar multiplication w.r.t. field addition** $(a + b)x = ax + bx$

Vector Subspace

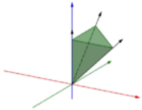
For a vector space (V, F), (W, F) is a **Vector Subspace** when

1. $W \subseteq V$
2. W itself is a **Vector Space** : closed under addition and scalar multiplication

Naturally follows that vector subspace contains 0.

Is this a vector spaces?

1. **Cone** : A "Cone" $W \subseteq \mathbb{R}^n$ is a set if for $w \in W, \alpha \geq 0$, $\alpha w \in W$

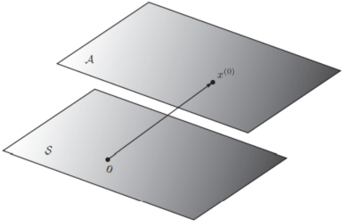


Are they vector subspaces in a given vector space?

1. $\mathbb{S}^n$: A set of symmetric real matrices in $\mathbb{R}^{n,n}$
2. $\mathbb{S}^n_+$: A set of PSD matrices in $\mathbb{R}^{n,n}$
3. $O = \{(x, y) \in \mathbb{R}^2 | x^2 + y^2 = 1\}$ in $\mathbb{R}^2$

Linear Combination, Span, Linear Independence, Basis, Dimension, Affine Set

- A **linear combination** of $\{v_1, v_2, ..., v_n\}$ with $a_1, a_2, ..., a_n$ is $a_1 v_1 + a_2 v_2 + ... + a_n v_n$
- A **span** of $S = \{v_1, v_2, ..., v_n\}$ is **the set of vectors** that can be expressed as linear combination of S. $span(S)$ forms a vector subspace.
- **Linear Independence** $\{v_1, v_2, ..., v_n\}$ is linearly independent if $\sum a_i v_i = 0 \to a_i = 0, \forall i$.
- **Basis** A set of vectors $\{v_1, v_2, ..., v_n\} \in V$ is a basis of V if $\{v_1, v_2, ..., v_n\}$ spans V and are linearly independent.
- **Dimension** is the cardinality of the basis.
- **Affine Set** $A = \{x \in \chi : x = v + x_0, v \in V\}$ where x_0 is a vector and V is a subspace in χ.



It is a **translate** of a subspace V by a vector x_0.

Norm on a vector space

A function $||\cdot|| : V \to \mathbb{R}_+$ is a **norm** if

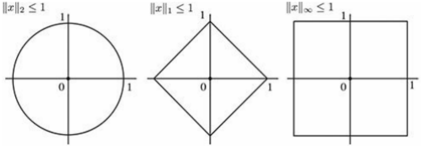
1. $||x|| > 0$ if $x \neq 0$ and $||x|| = 0 \leftrightarrow x = 0$
2. $||x + y|| \leq ||x|| + ||y||$ for $x, y \in V$
3. $||\alpha x|| = |\alpha| ||x||$ for $\alpha \in \mathbb{R}, x \in V$.

The representative example of a norm is the l_p norm : $||x||_p = (\sum_{k=1}^{n} |x_k|^p)^{1/p}, p = 1, 2, 3, ..., \infty$

Useful properties of l_p norms

1. $||x||_\infty := max_i |x_i| = lim_{p\to\infty} ||x||_p$
2. $\forall x \in \mathbb{R}^n$, $||x||_2/\sqrt{n} \leq ||x||_\infty \leq ||x||_2 \leq ||x||_1 \leq \sqrt{n}||x||_2 \leq n||x||_\infty$

Unit balls in the l_p norms
$B_p := \{x \in R^n : ||x||_p \leq 1\}$ is called an unit l_p norm ball.



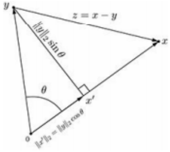
Inner product space, angle, orthogonality

$\forall x, y, z \in V$, $\alpha \in R$, $\langle \cdot \rangle : V \times V \to R$ is a **inner product** if

1. $\langle x, x \rangle > 0$ if $x \neq 0$ and $\langle x, x \rangle = 0$ iff $x = 0$
2. $\langle x, y \rangle = \langle y, x \rangle$
3. $\langle x, y + z \rangle = \langle x, y \rangle + \langle x, z \rangle$
4. $\langle \alpha x, y \rangle = \alpha \langle x, y \rangle$

   In $\mathbb{R}^n$, and only in this case, $\langle x, y \rangle = x^T y = \sum x_i y_i$

In Inner product space, we can discuss the concept of angle and orthogonality.



$cos\theta = \frac{x^T y}{||x||_2 ||y||_2}$. **Cosine Similarity** in natural language processing!

A basis $\{v_1, v_2, ..., v_n\}$ for $\mathbb{R}^n$ is an inner product space of dimension n is called an "orthonormal basis" if $\langle v_i, v_i \rangle = 1, \forall i$ and $\langle v_i, v_j \rangle = 0, \forall i \neq j$

Inner product and Norms

Optimization Examples on l2 norm ball Greet your very first optimization problem.
Let $y \in R^n$ be a GIVEN nonzero vector, $\chi := \{x \in R^n : ||x||_2 \leq r\}$, $r \in R_{++}$

For the following four questions, find the optimal value p^* and the solution of x that achieves the objective function (called the "optimal set").

1. $p^* = min_{x \in \chi} x^T y$
2. $p^* = max_{x \in \chi} x^T y$
3. $p^* = min_{x \in \chi} |x^T y|$
4. $p^* = max_{x \in \chi} |x^T y|$

* Don't mess up with "orthogonal" vs "opposite direction" *

Very Important Inequalities

1. Cauchy Schwarz Inequality
   —$\langle x, y \rangle \leq ||x||_2 ||y||_2$
2. Holder's Inequality in $\mathbb{R}^n$ For $\frac{1}{p} + \frac{1}{q} = 1$, $1 \leq p, q \leq \infty$,
   —$\langle x, y \rangle \leq ||x||_q ||y||_p$

Linear and Affine functions

Linear Functions
A function $f : R^n \to R$ is linear iff $\forall x \in R^n$ and $\alpha \in R$, $f(\alpha x) = \alpha f(x)$ and $x_1, x_2 \in R^n, f(x_1 + x_2) = f(x_1) + f(x_2)$

Can naturally know that $f(0) = 0$.

Linear functions can be represented as $f(x) = a^T x$ where $a \in R^n$

Affine Functions A function $f : R^n \to R$ is linear iff $f(x) - f(0)$ is linear.

Affine functions can be represented as $f(x) = a^T x + b$ where $a \in R^n, b \in R$

Example 2.10 Consider the functions $f_1, f_2, f_3 : \mathbb{R}^2 \to \mathbb{R}$ defined below:

$$\begin{aligned} f_1(x) &= 3.2x_1 + 2x_2, \\ f_2(x) &= 3.2x_1 + 2x_2 + 0.15, \\ f_3(x) &= 0.001x_2^2 + 2.3x_1 + 0.3x_2. \end{aligned}$$

The function f_1 is linear; f_2 is affine; f_3 is neither linear nor affine (f_3 is a quadratic function).

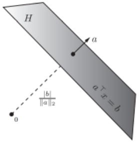
Hyperplane and Half Spaces

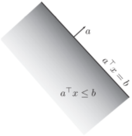
Hyperplane
$H = \{x \in R^n : a^T x = b\}$ where $a \in R^n, a \neq 0, b \in R$ are GIVEN.

Application) When $b = 0$, the hyperplane is the orthogonal complement of $span(a)$. This hyperplane forms a $n-1$ dimensional subspace.

Half Spaces A hyperplane H separates the vector space (usually R^n) into
$H_- := \{x | a^T x \leq b\}$, $H_{++} := \{x | a^T x > b\}$.
Those two regions are called half spaces.

Affine functions can be represented as $f(x) = a^T x + b$ where $a \in R^n, b \in R$





Matrices

$$A_{m,n} = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{pmatrix} \in \mathbb{R}^{m,n}$$

✓ The set of $m \times n$ matrices is a vector space

Matrix multiplication (col, row)

✓ Matrix multiplication on column vectors
For $A \in R^{m,n}$ and $x \in R^n$, $Ax = \sum_{j=1}^{n} x_j a_j$
: a linear combination of column vectors of A. Elements of x are the coefficients.

✓ Matrix multiplication on row vectors
For $A \in R^{m,n}$ and $u \in R^m$, $u^T A = \sum_{i=1}^{m} u_i \alpha_i^T$
: Since we adopt a columnwise notation, u^T is a row vector.
: a linear combination of row vectors of A. Elements of u are the coefficients.

Four Subspaces

$\mathcal{R}(A)$ **: Range Space of** A
For $A \in R^{m,n}$, $\mathcal{R}(A) := \{Ax | x \in R^n\}$ is a subspace of R^m spanned by column vectors of A
: Called as **Column Space = Range Space of** A **= Image of** A
$dim(\mathcal{R}(A))$ is called "rank" of A.

$\mathcal{N}(A)$ **: Null Space of** A **= Kernel of** A
$\mathcal{N}(A) := \{x | Ax = 0\}$ is a subspace of R^n spanned by vectors in the orthogonal complement of A
"A Set of vectors that are 'killed' by A"
$dim(\mathcal{N}(A))$ is called "nullity" of A.

$\mathcal{R}(A^T)$ **: Range Space of** A^T **= Row space of** A
$\mathcal{R}(A^T)$ is a subspace of R^n spanned by Row vectors in A.
In this class, we don't use the term "Row space".
$dim(\mathcal{R}(A^T))$ is called "row rank" of A. Can prove that column rank $=$ row rank.

$\mathcal{N}(A^T)$ **: Null Space of** A^T **= Kernel of** A^T
$\mathcal{N}(A^T)$ is a subspace of R^m spanned by vectors in the orthogonal complement of A^T
"A Set of vectors that are 'killed' by A^T"

Orthogonal Complement, Direct Sum, Projection Theorem

Orthogonal Complement

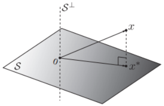
When a vector $x \in \chi$ satisfies $x \perp s, \forall s \in S$ where S is a subset of an inner product space χ, say "x is orthogonal to S".
A set of vectors $x \in \chi$ that are orthogonal to S is called an "orthogonal complement of S, represented as $S^{\perp}$

Direct Sum

If "any" x in a vector space χ can be uniquely represented as $x = a + b$, where $a \in A$ and $b \in B$, for A and B being subspaces of χ, say that
$\chi = A \bigoplus B$, χ is a direct sum of A and B.

Projection Theorem

Let V an inner product space having x as an element (Just think of Euclidean Space) and S is a subspace of V.
There exists unique $x^* \in S$ that satisfies $min_{y \in S}(||x - y||_2)$
: Looks like a complicated theorem! In English, "There exists a unique minimizer x^* in S that minimizes distance between x and y where x is in V and y is in S.

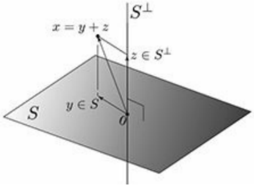


Orthogonal Decomposition Theorem

For any S, the subspace of an inner product space χ, any vector $x \in \chi$ can uniquely be represented as sum of 1) an element in S and 2) an element in $S^{\perp}$.

✓ Orthogonal Decomposition Theorem can be easily followed from the projection theorem.
Why? In the figure below, view y as x^* in the figure above and z as $x - x^*$ in the figure above!

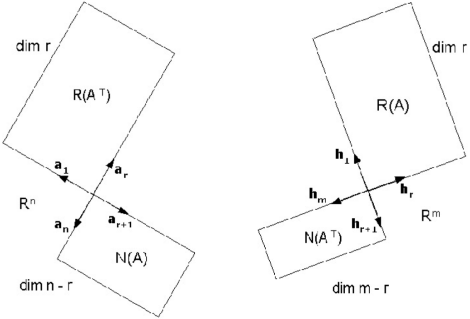
$$\mathcal{X} = \mathcal{S} \oplus \mathcal{S}^{\perp} \textit{ for any subspace } \mathcal{S} \subseteq \mathcal{X}.$$



A figure that explains Orthogonal Complement, Direct Sum, Projection Theorem and Orthogonal Decomposition Theorem
Role of Orthogonal Decomposition Theorem : The treasure derived from the content in the previous slide and is crucial theorem used in the Rank Nullity Theorem.

This theorem is crucial in the Singular Value Decomposition dealt later.
Fundamental Theorem of Linear Algebra

1. $R^n = \mathcal{N}(A) \bigoplus \mathcal{R}(A^T)$ where $\mathcal{R}(A^T) = N(A)^\perp$
2. $R^m = \mathcal{R}(A) \bigoplus \mathcal{N}(A^T)$ where $\mathcal{N}(A^T) = R(A)^\perp$



Square Matrices and Eigenspace
For $A \in \mathbb{R}^{n,n}$, we can discuss concepts of "eigenvalues" and "eigenvectors".

The matrix multiplication is a vector valued function that maps R^n to R^n. It represents a linear transformation.

For $v \neq 0$, if $Av = \lambda v$ or $(\lambda I_n - A)v = 0$, say v is an **eigenvector** and λ is an **eigenvalue** corresponding to that eigenvector.

In English :
1) Eigenvectors are "directions" in C^n invariant in angle under the linear transformation A : "axis of linear transformation".
2) Thus, multiplying A on v only changes the "scale". The eigenvalue represents the scale of linear tranformation".

The eigenspace of λ is $\mathcal{N}(\lambda I - A)$
"One eigenvalue may have distinct eigenvectors. Eigenspace : space of eigenvectors corresponding to the same eigenvalue".

Discussion of trace and determinant through Characteristic Polynomial

The Characteristic Polynomial $p_A(\lambda) = det(\lambda I - A)$ is a function of λ of degree n.
If λ_1, λ_2, ..., λ_n are eigenvalues of A, $p_A(\lambda) = det(\lambda I - A) = \prod_{i=1}^{n}(\lambda_i - \lambda)$.
Use(expand) $det(\lambda I - A) = (-\lambda)^n + tr(A)(-\lambda)^{n-1} + ... + det(A)$ and compare coefficients in the formula above leads to :
1) $tr(A) := \sum_{i=1}^{n} a_{ii} = \sum_{i=1}^{n} \lambda_i$
2) $det(A) = \prod_{i=1}^{n} \lambda_i$

Similar Matrices and Diagonalization

Similar Matrices

Two Matrices A, B are similar if there is an invertible matrix P satisfying $B = P^{-1}AP$

✓ Similar Matrices A, B represents the SAME linear transformation in a different basis.

Consider a linear transformation $y = Ax$.
Since P is an invertible matrix, columns of P form a basis in R^n.
Thus, x and y can be uniquely represented as $x = P\tilde{x}$ and $y = P\tilde{y}$
$\rightarrow P\tilde{y} = AP\tilde{x} \rightarrow \tilde{y} = P^{-1}AP\tilde{x} = B\tilde{x}$
✓ Similar Matrices share the same rank and eigenvalues (so, determinant, trace).

Diagonalizable Matrices

Matrix $A \in R^{n,n}$ is diagonalizable if it is similar to a diagonal matrix.

cf) Orthogonally diagonalizable : $A = U\Lambda U'$ where U : orthogonal

Special Type of Square Matrices

Orthogonal

A square matrix U is **Orthogonal** if $u_i^T u_j = 1$ for $i = j$ and 0 for $i \neq j$.
"Matrices whose columns form an orthonormal basis in R^n."
✓ $U^T U = U U^T = I_n$

1) Orthogonal Matrices preserves **length** : $||Ux||_2^2 = (Ux)^T (Ux) = ||x||_2^2$.
2) Orthogonal Matrices preserves **angles** : $(Ux)^T (Uy) = x^T y \because x' U' U y = x' y$

Diagonal

A square matrix A is diagonal if $a_{ij} = 0$ for $i \neq j$.
Matrix multiplication is easy involving diagonal matrices.

Symmetric

A square matrix A is symmetric if $a_{ij} = a_{ji}$ for $i, j = 1, 2, .., n$.

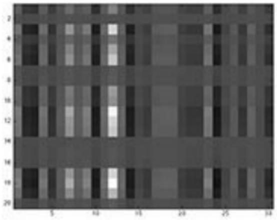
Dyads

Dyads
A matrix $A \in R^{m,n}$ is a dyad if it can be represented as $A = uv^T$ for $u \in R^m$ and $v \in R^n$.
✓ Watch out! Dyads include non-square matrices!

✓ Dyad has rank of 1.

✓ A Matrix $AR^{m,n}$ can be written as a sum of dyads :

$A = U \sum V^T = \sum_{i=1}^{r} \sigma_i u_i v_i^T$ where $r = rank(A)$ and u_i and v_i are column vectors of U and V.



Matrix Norms

Matrix Norm

- $f(A) \geq 0$, and $f(A) = 0$ if and only if $A = 0$;
- $f(\alpha A) = |\alpha| f(A)$;
- $f(A + B) \leq f(A) + f(B)$.

Many of the popular matrix norms also satisfy a fourth condition called *sub-multiplicativity*: for any conformably sized matrices A, B

$$f(AB) \leq f(A) f(B).$$

Three Popular types of Matrix Norm

1. **Frobenius Norm** $||A||_F := (\sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^2)^{\frac{1}{2}} = \sqrt{tr(AA')}$
2. **Nuclear Norm** $||A||_* := \sum \sigma_i$ where σ_i is a singular value in A
3. **Induced Norm = Operator Norm** $||A||_p := max_{x \neq 0} \frac{||Ax||_p}{||x||_p}$: "how much can a matrix scale things up"

   $||A||_2 := max_{x \neq 0} \frac{||Ax||_2}{||x||_2} = \sqrt{\lambda_{max}(A^T A)} = \sigma_1$ where σ_1 is the biggest singular value.

   $||A||_1 := max_{x \neq 0} \frac{||Ax||_1}{||x||_1}$: largest absolute column sum

Symmetric matrices

For $A \in R^{n,n}$, if $A = A^T$, A is symmetric and say $A \in \mathbb{S}^n$

Important Examples of Symmetric matrices

1. Covariance Matrix

$$\begin{aligned}\mathrm{Cov}(\mathbf{Y}) &= \mathrm{E}\left[(\mathbf{Y}-\mathrm{E}(\mathbf{Y}))(\mathbf{Y}-\mathrm{E}(\mathbf{Y}))^T\right] \\ &= \begin{pmatrix} \mathrm{Var}(Y_1) & \mathrm{Cov}(Y_1,Y_2) & \dots & \mathrm{Cov}(Y_1,Y_n) \\ \mathrm{Cov}(Y_2,Y_1) & \mathrm{Var}(Y_2) & \dots & \mathrm{Cov}(Y_2,Y_n) \\ \vdots & \vdots & \vdots & \vdots \\ \mathrm{Cov}(Y_n,Y_1) & \mathrm{Cov}(Y_n,Y_2) & \dots & \mathrm{Var}(Y_n) \end{pmatrix}\end{aligned}$$

2. Hessian Matrix $\nabla^2 f(x_0)$

   Application) Second Order Approximation of f at x_0
   For $f : R \to R, f(x) \approx f(x) + f'(x_0)(x - x_0) + \frac{1}{2}f''(x)(x - x_0)^2$
   For $f : R^n \to R, f(x) \approx f(x_0) + \nabla f(x_0)^T(x - x_0) + \frac{1}{2}(x - x_0)^T\nabla^2 f(x_0)(x - x_0)$

3. $A^T A \in \mathbb{S}^n$ for $A \in R^{m,n}$

Spectral Theorem of $\mathbb{S}^n$

This spectral theorem is the most important theme in Symmetric matrices.

1. All eigenvalues of $A \in \mathbb{S}^n$ are real.

   pf) $Av = \lambda v \to v^* A^* = \lambda^* v^*$ using complex conjugate transpose.
   Note that $A^* = A^T = A$ since A has real entries
   Note that $v^* A v = v^* \lambda v = \lambda v^* v = \lambda^* v^* v$.
   Since $v \neq 0$, $\lambda = \lambda^* \to \lambda \in \mathbb{R}$.

2. For $A \in \mathbb{S}^n$, distinct eigenvectors corresponding to distinct eigenvalues are orthogonal.

   Let $Av = \lambda v$ & $Aw = \tilde{\lambda} w$ for distinct eigenvalues λ and $\tilde{\lambda}$.
   $w' A v = w' \lambda v = \lambda w' v$.
   $w' A v = v' A' w = v' A w = \tilde{v}' w = \tilde{\lambda} w' v$.
   Since $\lambda \neq \tilde{\lambda}, w'v = 0$: w and v are orthogonal.

3. For $A \in \mathbb{S}^n$, can say $A = U \Lambda U^T$, U orthogonal and Λ diagonal

Positive (Semi) Definiteness and Partial Order

For real, symmetric matrices (matrices we can orthogonally diagonalize as $U\Lambda U^T$),

$\mathbb{S}^n_+$ is a set of positive semidefinite real symmetric matrices.
$\leftrightarrow$ All eigenvalues $=$ diagonal entries of Λ are nonnegative

$\mathbb{S}^n_{++}$ is a set of positive definite real symmetric matrices.
$\leftrightarrow$ All eigenvalues $=$ diagonal entries of Λ are positive

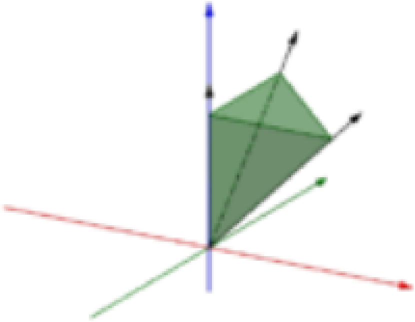
Partial Order of $\mathbb{S}^n$
More generally,

For $A, B \in S^n$, say $A \succeq B$ if $A - B \in \mathbb{S}^n_+ \leftrightarrow x^TAx \geq x^TBx, \forall x \in R^n$
For $A, B \in S^n$, say $A \succ B$ if $A - B \in \mathbb{S}^n_{++} \leftrightarrow x^TAx > x^TBx, \forall x \in R^n$

Why "more generally"?
$A \in \mathbb{S}^n_+$ iff $A \succeq 0$ $A \in \mathbb{S}^n_{++}$ iff $A \succ 0$

Symmetric Matrices and Cone

A "Cone" $W \subseteq \mathbb{R}^n$ is a set if for $w \in W, \alpha \geq 0$, $\alpha w \in W$



PSD Matrices are cones!

S^n_{++} is the interior of the cone S^n_+

S^n is a subspace of $R^{n,n}$ while S^n_+ is not.

Positive Definiteness and Ellipsoids : a good way to represent a PD matrix w.r.t. its eigenstuff

Let $A \in S^n_{++}$. Thoughts to have in mind : A is orthogonally diagonalizable and have all $+$ eigenvalues!

$A = U \Lambda U^T$, $A^{-1} = U \Lambda^{-1} U^T$

Note) PSD matrices don't guarantee inverse (guarantees square root though).

$\xi := \{x \in R^n : x^T A^{-1} x \leq 1\} = \{Uy \in R^n : y^T \Lambda^{-1} y \leq 1\} = \{Uy \in R^n : \sum \frac{1}{\lambda_i} y_i^2 \leq 1\}$

Practice Examples

$X := \{x \in R^n | x^T A x \leq 1\}$. Draw X by considering eigenvalues and eigenvectors of A. Be careful about A and A^{-1}.

A $= \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$

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

Next week Topics

Definition of Singular Values : Eigenvalues and Singular Values

Induced l_2 norm of A and the largest singular value of A : $||A_2|| = \sigma_1$

Rayleigh Quotient

Singular Value Decomposition : Compact Form and Full Form SVD

Principal Component Analysis : A projection (embedding) of the original high dimensional data into a lower dimension data

Least Squares : Some advanced topics such as least squares with constraints

Next week topics are less broad as this week while being harder.