Introduction to Statistical Machine Learning

Christfried Webers

Statistical Machine Learning Group
NICTA
and
College of Engineering and Computer Science
The Australian National University

Canberra
February – June 2013

(Many figures from C. M. Bishop, "Pattern Recognition and Machine Learning")

Part III

Linear Algebra

©2013
Christfried Webers
NICTA
The Australian National University

Intuition

Geometry

- Points and Lines
- Vector addition and scaling
- Humans have experience with 3 dimensions (less with 1, 2 though)

Generalisation to N dimensions (possibly $N \to \infty$)

- Line $\to$ vector space $\mathbb{V}$
- Point $\to$ vector $x \in \mathbb{V}$
- Example : $X \in \mathbb{R}^{n \times m}$
- Space of matrices $\mathbb{R}^{n \times m}$ and the space of vectors $\mathbb{R}^{n \cdot m}$ are isomorphic

©2013

Christfried Webers
NICTA
The Australian National University

Vector Space $\mathcal{V}$, underlying Field $\mathcal{F}$

Given vectors $\mathbf{u}, \mathbf{v}, \mathbf{w} \in \mathcal{V}$ and scalars $\alpha, \beta \in \mathcal{F}$, the following holds:

- Associativity of addition : $\mathbf{u} + (\mathbf{v} + \mathbf{w}) = (\mathbf{u} + \mathbf{v}) + \mathbf{w}$.
- Commutativity of addition : $\mathbf{v} + \mathbf{w} = \mathbf{w} + \mathbf{v}$.
- Identity element of addition : $\exists\, \mathbf{0}$ such that $\mathbf{v} + \mathbf{0} = \mathbf{v}, \quad \forall\, \mathbf{v} \in \mathcal{V}$.
- Inverse of addition : For all $\mathbf{v} \in \mathcal{V}$, $\exists \mathbf{w} \in \mathcal{V}$, such that $\mathbf{v} + \mathbf{w} = \mathbf{0}$. The additive inverse is denoted $-\mathbf{v}$.
- Distributivity of scalar multiplication with respect to vector addition : $\alpha(\mathbf{v} + \mathbf{w}) = \alpha\mathbf{v} + \alpha\mathbf{w}$.
- Distributivity of scalar multiplication with respect to field addition : $(\alpha + \beta)\mathbf{v} = \alpha\mathbf{v} + \beta\mathbf{v}$.
- Compatibility of scalar multiplication with field multiplication : $\alpha(\beta\mathbf{v}) = (\alpha\beta)\mathbf{v}$.
- Identity element of scalar multiplication : $1\mathbf{v} = \mathbf{v}$, where 1 is the identity in $\mathcal{F}$.

Matrix-Vector Multiplication

```
for i in xrange(m):
  R[i] = 0.0;
  for j in xrange(n):
     R[i] = R[i] + A[i,j] * V[j]
```

Listing 1: Code for elementwise matrix multiplication.

©2013
Christfried Webers
NICTA
The Australian National University

Matrix-Vector Multiplication

```
for i in xrange(m):
  R[i] = 0.0;
  for j in xrange(n):
    R[i] = R[i] + A[i,j] * V[j]
```

Listing 2: Code for elementwise matrix multiplication.

$$\begin{array}{cccc} A & V & = & R \\ \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} & \begin{bmatrix} v_1 \\ v_2 \\ \dots \\ v_n \end{bmatrix} & = & \begin{bmatrix} a_{11}v_1 + a_{12}v_2 + \dots + a_{1n}v_n \\ a_{21}v_1 + a_{22}v_2 + \dots + a_{2n}v_n \\ \dots \\ a_{m1}v_1 + a_{m2}v_2 + \dots + a_{mn}v_n \end{bmatrix} \end{array}$$

Matrix-Vector Multiplication

$$\begin{array}{cccc} A & V & = & R \\ \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} & \begin{bmatrix} v_1 \\ v_2 \\ \dots \\ v_n \end{bmatrix} & = & \begin{bmatrix} a_{11}v_1 + a_{12}v_2 + \cdots + a_{1n}v_n \\ a_{21}v_1 + a_{22}v_2 + \cdots + a_{2n}v_n \\ \dots \\ a_{m1}v_1 + a_{m2}v_2 + \cdots + a_{mn}v_n \end{bmatrix} \end{array}$$

©2013 Christfried Webers

NICTA
The Australian National University

Matrix-Vector Multiplication

$$
\begin{array}{cccc}
A & V & = & R \\
\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} &
\begin{bmatrix} v_1 \\ v_2 \\ \dots \\ v_n \end{bmatrix} & = &
\begin{bmatrix} a_{11}v_1 + a_{12}v_2 + \dots + a_{1n}v_n \\ a_{21}v_1 + a_{22}v_2 + \dots + a_{2n}v_n \\ \dots \\ a_{m1}v_1 + a_{m2}v_2 + \dots + a_{mn}v_n \end{bmatrix}
\end{array}
$$

```
R = A[:,0] * V[0];
for j in xrange(1, n):
    R += A[:, j] * V[j];
```

Listing 4: Code for columnwise matrix multiplication.

© 2013

Christfried Webers
NICTA
The Australian National University

Matrix-Vector Multiplication

© 2013 Christfried Webers NICTA The Australian National University

- Denote the n columns of A by $\mathbf{a_1}, \mathbf{a_2}, \dots, \mathbf{a_n}$
- Each $\mathbf{a_i}$ is now a (column) vector

$$\begin{array}{ccc} A & V & = & R \\ \begin{bmatrix} \mathbf{a_1} & \mathbf{a_2} & \dots & \mathbf{a_n} \end{bmatrix} & \begin{bmatrix} v_1 \\ v_2 \\ \dots \\ v_n \end{bmatrix} & = & \begin{bmatrix} \mathbf{a_1} v_1 + \mathbf{a_2} v_2 + \cdots + \mathbf{a_n} v_n \end{bmatrix} \end{array}$$

Transpose of R

- Given $R = AV$,

$$R = \begin{bmatrix} \mathbf{a_1} v_1 + \mathbf{a_2} v_2 + \cdots + \mathbf{a_n} v_n \end{bmatrix}$$

- What is R^T ?

Transpose of R

- Given $R = AV$,

$$R = \begin{bmatrix} \mathbf{a_1} v_1 + \mathbf{a_2} v_2 + \cdots + \mathbf{a_n} v_n \end{bmatrix}$$

- What is R^T ?
- $$R^T = \begin{bmatrix} v_1 \mathbf{a_1^T} + v_2 \mathbf{a_2^T} + \cdots + v_n \mathbf{a_n^T} \end{bmatrix}$$
- NOT equal to $A^T V^T$! (In fact, $A^T V^T$ is not even defined.)

Transpose

- Reverse order rule : $(AV)^T = V^T A^T$

Transpose

- Reverse order rule : $(AV)^T = V^T A^T$

$$\begin{array}{cccc} V^T & A^T & = & R^T \end{array}$$

$$\begin{bmatrix} v_1 & v_2 \dots v_n \end{bmatrix} \begin{bmatrix} \mathbf{a_1^T} \\ \mathbf{a_2^T} \\ \dots \\ \mathbf{a_n^T} \end{bmatrix} = \begin{bmatrix} v_1 \mathbf{a_1^T} + v_2 \mathbf{a_2^T} + \cdots + v_n \mathbf{a_n^T} \end{bmatrix}$$

Trace

- The trace of a square matrix A is the sum of all diagonal elements of A.

$$\operatorname{tr}\{A\} = \sum_{k=1}^{n} A_{kk}$$

- Example

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} \qquad \operatorname{tr}\{A\} = 1 + 5 + 9 = 15$$

- The trace does not exist for a non-square matrix.

©2013 Christfried Webers NICTA The Australian National University

vec(x) *operator*

- Define $\mathrm{vec}(X)$ as the vector which results from stacking all columns of a matrix A on top of each other.
- Example

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} \qquad \mathrm{vec}(A) = \begin{bmatrix} 1 \\ 4 \\ 7 \\ 2 \\ 5 \\ 8 \\ 3 \\ 6 \\ 9 \end{bmatrix}$$

- Given two matrices $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{n \times p}$, the trace of the product $\mathbf{X}^T\mathbf{Y}$ can be written as

$$\operatorname{tr}\left\{\mathbf{X}^T\mathbf{Y}\right\} = \mathrm{vec}(\mathbf{X})^T \, \mathrm{vec}(\mathbf{Y})$$

Inner Product

- Mapping from two vectors $x, y \in \mathbb{V}$ to a field of scalars $\mathbb{F}$ ($\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$):

$$\langle \cdot, \cdot \rangle : \mathbb{V} \times \mathbb{V} \to \mathbb{F}$$

- Conjugate Symmetry :
  $\langle x, y \rangle = \overline{\langle y, x \rangle}$
- Linearity :
  $\langle ax, y \rangle = a \langle x, y \rangle$, and $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$
- Positive-definitness :
  $\langle x, x \rangle \geq 0$, and $\langle x, x \rangle = 0$ for $x = 0$ only.

Canonical Inner Products

- Inner product for real numbers $x, y \in \mathbb{R}$:

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

© 2013 Christfried Webers
NICTA
The Australian National University

Canonical Inner Products

- Inner product for real numbers $x, y \in \mathbb{R}$:

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

- Dot product between two vectors:
  Given $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T \mathbf{y} \equiv \sum_{k=1}^{n} x_k\, y_k = x_1\, y_1 + x_2\, y_2 + \cdots + x_n\, y_n$$

- Canonical inner product for matrices :
  Given $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{n \times p}$

$$\langle \mathbf{X}, \mathbf{Y} \rangle = \operatorname{tr}\left\{\mathbf{X}^T \mathbf{Y}\right\} = \sum_{k=1}^{p} (\mathbf{X}^T \mathbf{Y})_{kk} = \sum_{k=1}^{p} \sum_{l=1}^{n} (\mathbf{X}^T)_{kl} (\mathbf{Y})_{lk}$$

$$= \sum_{k=1}^{p} \sum_{l=1}^{n} \mathbf{X}_{lk} \mathbf{Y}_{lk}$$

©2013
Christfried Webers
NICTA
The Australian National University

Calculations with Matrices

- Denote $(i,j)^{\text{th}}$ element of matrix $\mathbf{X}$ by $\mathbf{X}_{ij}$
- Transpose : $(\mathbf{X}^T)_{ij} = \mathbf{X}_{ji}$
- Product : $(\mathbf{X}\mathbf{Y})_{ij} = \sum_{k=1}^{\dots} \mathbf{X}_{ik}\mathbf{Y}_{kj}$
- Proof of the linearity of the canonical matrix inner product :

$$
\begin{aligned}
\langle \mathbf{X}+\mathbf{Y}, \mathbf{Z} \rangle &= \operatorname{tr}\left\{(\mathbf{X}+\mathbf{Y})^T\mathbf{Z}\right\} = \sum_{k=1}^{p}((\mathbf{X}+\mathbf{Y})^T\mathbf{Z})_{kk} \\
&= \sum_{k=1}^{p}\sum_{l=1}^{n}((\mathbf{X}+\mathbf{Y})^T)_{kl}\mathbf{Z}_{lk} = \sum_{k=1}^{p}\sum_{l=1}^{n}(\mathbf{X}_{lk}+Y_{lk})\mathbf{Z}_{lk} \\
&= \sum_{k=1}^{p}\sum_{l=1}^{n}\mathbf{X}_{lk}\mathbf{Z}_{lk} + \mathbf{Y}_{lk}\mathbf{Z}_{lk} \\
&= \sum_{k=1}^{p}\sum_{l=1}^{n}(\mathbf{X}^T)_{kl}\mathbf{Z}_{lk} + (\mathbf{Y}^T)_{kl}\mathbf{Z}_{lk} \\
&= \sum_{k=1}^{p}(\mathbf{X}^T\mathbf{Z})_{kk} + (\mathbf{Y}^T\mathbf{Z})_{kk} \\
&= \operatorname{tr}\left\{\mathbf{X}^T\mathbf{Z}\right\} + \operatorname{tr}\left\{\mathbf{Y}^T\mathbf{Z}\right\} = \langle \mathbf{X}, \mathbf{Z}\rangle + \langle \mathbf{Y}, \mathbf{Z}\rangle
\end{aligned}
$$

Projection

- In linear algebra and functional analysis, a projection is a linear transformation P from a vector space $\mathbb{V}$ to itself such that

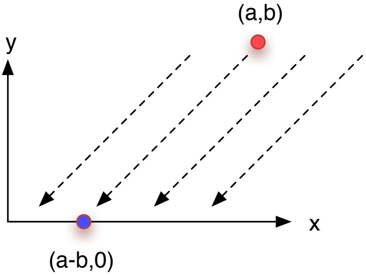
$$P^2 = P$$

Projection transformation can only have effective influence for one time, because after first time projection, we can not move it anymore by further applying that tranformation.

Projection

- In linear algebra and functional analysis, a projection is a linear transformation P from a vector space $\mathbb{V}$ to itself such that

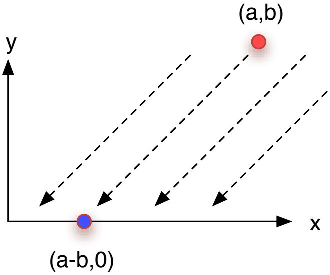
$$P^2 = P$$



Projection

- In linear algebra and functional analysis, a projection is a linear transformation P from a vector space $\mathbb{V}$ to itself such that

$$P^2 = P$$



$$A \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} a - b \\ 0 \end{bmatrix} \qquad A = \begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix}$$

Orthogonal Projection

- Orthogonality : need an inner product $\langle x, y \rangle$
- Choose two arbitrary vectors x and y. Then Px and $y - Py$ are orthogonal.

$$0 = \langle Px, y - Py \rangle = (Px)^T (y - Py) = x^T (P - P^T P) y$$

- Orthogonal Projection

$$P^2 = P \qquad P = P^T$$

- Example : Given some unit vector $u \in \mathbb{R}^n$ characterising a line through the origin in $\mathbb{R}^n$
- project an arbitrary vector $x \in \mathbb{R}^n$ onto this line by

$$P = uu^T$$

- Proof : $P = P^T$, and

$$P^2 x = (uu^T)(uu^T)x = uu^T x = Px$$

©2013
Christfried Webers
NICTA
The Australian National University

ISML 2013

Orthogonal Projection

- Given a matrix $\mathbf{A} \in \mathbb{R}^{n \times p}$ and a vector $\mathbf{x}$. What is the closest point $\widetilde{\mathbf{x}}$ to $\mathbf{x}$ in the column space of $\mathbf{A}$?
- Orthogonal Projection into the column space of $\mathbf{A}$

$$\widetilde{\mathbf{x}} = \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{x}$$

- Projection matrix

$$P = \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$$

- Proof

$$P^2 = \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T = \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T = P$$

- Orthogonal projection ?

$$P^T = (\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T)^T = \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T = P$$

Linear Independence of Vectors

- A set of vectors is linearly independent if none of them can be written as a linear combination of finitely many other vectors in the set.
- Given a vector space $\mathbb{V}$ and k vectors $\mathbf{v}_1, \dots, \mathbf{v}_k \in \mathbb{V}$ and scalars $a_1, \dots, a_k$. The vectors are linearly independent, if and only if

$$a_1 \mathbf{v}_1 + a_2 \mathbf{v}_2 + \cdots + a_k \mathbf{v}_k = \mathbf{0} \equiv \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

has only the trivial solution

$$a_1 = a_2 = a_k = 0$$

©2013 Christfried Webers NICTA The Australian National University

Rank

- The column rank of a matrix A is the maximal number of linearly independent columns of A.
- The row rank of a matrix A is the maximal number of linearly independent rows of A.
- For every matrix A, the row rank and column rank are equal, called the rank of A.

Determinant

Let S_n be the set of all permutation of the numbers $\{1, \dots, n\}$. The determinant of the square matrix A is then given by

$$\det\{A\} = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \prod_{i=1}^{n} A_{i,\sigma(i)}$$

where sgn is the signature of the permutation ($+1$ if the permutation is even, -1 if the permutation is odd).

- $\det\{AB\} = \det\{A\}\det\{B\}$ for A, B square matrices.
- $\det\{A^{-1}\} = \det\{A\}^{-1}$.
- $\det\{A^T\} = \det\{A\}$.

Matrix Inverse

- Identity Matrix I
- $I = AA^{-1} = A^{-1}A$
- The matrix inverse A^{-1} does only exist for square matrices which are NOT singular.
- Singular matrix
  - at least one eigenvalue is zero,
  - determinant $|A| = 0$.
- Inversion 'reverses the order' (like transposition)

$$(AB)^{-1} = B^{-1}A^{-1}$$

(Only then is $(AB)(AB)^{-1} = (AB)B^{-1}A^{-1} = AA^{-1} = I$.)

Matrix Inverse and Transpose

- $(A^{-1})^T$?

©2013 Christfried Webers NICTA The Australian National University

Matrix Inverse and Transpose

- $(A^{-1})^T$?
- need to assume that (A^{-1}) exists
- rule : $(A^{-1})^T = (A^T)^{-1}$

Useful Identity

$$(A^{-1} + B^T C^{-1} B)^{-1} B^T C^{-1} = A B^T (B A B^T + C)^{-1}$$

©2013 Christfried Webers NICTA The Australian National University

Useful Identity

$$(A^{-1} + B^T C^{-1} B)^{-1} B^T C^{-1} = A B^T (B A B^T + C)^{-1}$$

- How to analyse and prove such an equation?
- A^{-1} must exist, therefore A must be square, say $A \in \mathbb{R}^{n \times n}$.
- Same for C^{-1}, so let's assume $C \in \mathbb{R}^{m \times m}$.
- Therefore, $B \in \mathbb{R}^{m \times n}$.

©2013 Christfried Webers NICTA The Australian National University

Prove the Identity

$$(A^{-1} + B^T C^{-1} B)^{-1} B^T C^{-1} = A B^T (B A B^T + C)^{-1}$$

- Multiply by $(BAB^T + C)$

$$(A^{-1} + B^T C^{-1} B)^{-1} B^T C^{-1} (B A B^T + C) = A B^T$$

- Simplify the left-hand side

$$\begin{aligned}
&(A^{-1} + B^T C^{-1} B)^{-1} [B^T C^{-1} B (A B^T) + B^T] \\
&= (A^{-1} + B^T C^{-1} B)^{-1} [B^T C^{-1} B + A^{-1}] A B^T \\
&= A B^T
\end{aligned}$$

Woodbury Identity

$$(A + BD^{-1}C)^{-1} = A^{-1} - A^{-1}B(D + CA^{-1}B)^{-1}CA^{-1}$$

- Useful if A is diagonal and easy to invert, and B is very tall (many rows, but only a few columns) and C is very wide (few rows, many columns).

Manipulating Matrix Equations

- Don't multiply by a matrix which does not have full rank. Why? In the general case, you will loose solutions.
- Don't multiply by nonsquare matrices.
- Don't assume a matrix inverse exists because a matrix is square. The matrix might be singular and you are making the same mistake as if deducing $a = b$ from $a\,0 = b\,0$.

Eigenvectors

- Every square matrix $A \in \mathbb{R}^{n \times n}$ has an Eigenvector decomposition

$$Ax = \lambda x$$

where $x \in \mathbb{R}^n$ and $\lambda \in \mathbb{C}$.

- Example:

$$\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} x = \lambda x$$

$$\lambda = \{-\imath, \imath\}$$

$$x = \left\{ \begin{bmatrix} \imath \\ 1 \end{bmatrix}, \begin{bmatrix} -\imath \\ 1 \end{bmatrix} \right\}$$

Eigenvectors

- How many eigenvalue/eigenvector pairs?
- $$Ax = \lambda x$$
  is equivalent to
  $$(A - \lambda I)x = 0$$
- Has only non-trivial solution for $\det\{A - \lambda I\} = 0$
- polynom of nth order; at most n distinct solutions

Real Eigenvalues

- How can we enforce real eigenvalues?
- Let's look at matrices with complex entries $A \in \mathbb{C}^{n \times n}$.
- Transposition is replaced by Hermitian adjoint, e.g.

$$\begin{bmatrix} 1 + \imath 2 & 3 + \imath 4 \\ 5 + \imath 6 & 7 + \imath 8 \end{bmatrix}^H = \begin{bmatrix} 1 - \imath 2 & 5 - \imath 6 \\ 3 - \imath 4 & 7 - \imath 8 \end{bmatrix}$$

- Denote the complex conjugate of a complex number λ by $\overline{\lambda}$.

Real Eigenvalues

- How can we enforce real eigenvalues?
- Let's assume $A \in \mathbb{C}^{n \times n}$, Hermitian ($A^H = A$).
- Calculate

$$x^H A x = \lambda x^H x$$

  for an eigenvector $x \in \mathbb{C}^n$ of A.
- Another possibility to calculate $x^H A x$

$$\begin{aligned} x^H A x &= x^H A^H x && (A \text{ is Hermitian}) \\ &= (x^H A x)^H && (\text{reverse order}) \\ &= (\lambda x^H x)^H && (\text{eigenvalue}) \\ &= \overline{\lambda} x^H x \end{aligned}$$

- and therefore

$$\lambda = \overline{\lambda} \quad (\lambda \text{ is real}).$$

- If A is Hermitian, then all eigenvalues are real.
- Special case: If A has only real entries and is symmetric, then all eigenvalues are real.

©2013
Christfried Webers
NICTA
The Australian National University

Singular Value Decomposition

© 2013 Christfried Webers NICTA The Australian National University

Every matrix $A \in \mathbb{R}^{n \times p}$ can be decomposed into a product of three matrices

$$A = U \Sigma V^T$$

where $U \in \mathbb{R}^{n \times n}$ and $V \in \mathbb{R}^{p \times p}$ are orthogonal matrices ($U^T U = I$ and $V^T V = I$), and $\Sigma \in \mathbb{R}^{n \times p}$ has nonnegative numbers on the diagonal.

How to calculate a gradient?

- Given a vector space $\mathcal{V}$, and a function $f : \mathcal{V} \to \mathbb{R}$.
- Example: $X \in \mathbb{R}^{n \times p}$, arbitrary $C \in \mathbb{R}^{n \times n}$,

$$f(X) = \operatorname{tr}\left\{X^T C X\right\}.$$

- How to calculate the gradient of f ?

How to calculate a gradient?

- Given a vector space $\mathcal{V}$, and a function $f : \mathcal{V} \to \mathbb{R}$.
- Example: $X \in \mathbb{R}^{n \times p}$, arbitrary $C \in \mathbb{R}^{n \times n}$,

$$f(X) = \operatorname{tr}\left\{X^T C X\right\}.$$

- How to calculate the gradient of f ?
- Ill-defined question. There is no gradient in a vector space.
- Only a Directional Derivative at point $X \in \mathcal{V}$ in direction $\xi \in \mathcal{V}$.

$$\mathcal{D}f(X)(\xi) = \lim_{h \to 0} \frac{f(X + h\xi) - f(X)}{h}$$

- For the example $f(X) = \operatorname{tr}\left\{X^T C X\right\}$

$$\begin{aligned} \mathcal{D}f(X)(\xi) &= \lim_{h \to 0} \frac{\operatorname{tr}\left\{(X + h\xi)^T C (X + h\xi)\right\} - \operatorname{tr}\left\{X^T C X\right\}}{h} \\ &= \operatorname{tr}\left\{\xi^T C X + X^T C \xi\right\} = \operatorname{tr}\left\{X^T (C^T + C) \xi\right\} \end{aligned}$$

How to calculate a gradient?

- Given a vector space $\mathcal{V}$, and a function $f : \mathcal{V} \to \mathbb{R}$.
- How to calculate the gradient of f ?

1. Calculate the Directional Derivative
2. Define an inner product $\langle A, B \rangle$ on the vector space
3. The gradient is now defined as

$$\mathcal{D}f(X)(\xi) = \langle \operatorname{grad} f, \xi \rangle$$

- For the example $f(X) = \operatorname{tr}\left\{X^T C X\right\}$
- Define the inner product $\langle A, B \rangle = \operatorname{tr}\left\{A^T B\right\}$.
- Write the Directional Derivative as an inner product with ξ

$$\begin{aligned} \mathcal{D}f(X)(\xi) &= \operatorname{tr}\left\{X^T (C^T + C) \xi\right\} \\ &= \langle (C + C^T) X, \xi \rangle = \langle \operatorname{grad} f, \xi \rangle \end{aligned}$$

Books

- Gilbert Strang, "Introduction to Linear Algebra", Wellesley Cambridge, 2009.
- David C. Lay, "Linear Algebra and Its Applications", Addison Wesley, 2005.

©2013

Christfried Webers
NICTA
The Australian National University