EC-332 Machine Learning

Nazar Khan

Department of Computer Science
University of the Punjab

Matrix and Vector Calculus

Notation

- Scalars are denoted by lower-case letters like s, a, b.
- Vectors are denoted by lower-case bold letters like $\mathbf{x}, \mathbf{y}, \mathbf{v}$.
- Matrices are denoted by upper-case bold letters like $\mathbf{M}, \mathbf{D}, \mathbf{A}$.
- Any vector $\mathbf{x} \in \mathbb{R}^d$ is by default a column vector.

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

- The corresponding row vector is obtained as $\mathbf{x}^T = \begin{bmatrix} x_1 & x_2 & \dots & x_d \end{bmatrix}$.

Inner Product

For vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$

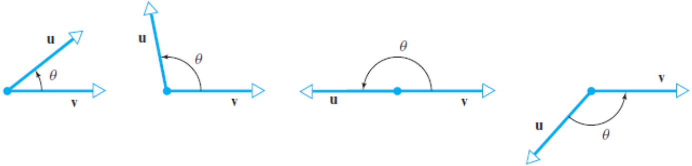
- *Inner product* is a scalar value.

$$\mathbf{x}^T\mathbf{y} = x_1y_1 + x_2y_2 + \cdots + x_dy_d = \|\mathbf{x}\|\|\mathbf{y}\| \cos\theta$$

  where θ is the angle between vectors $\mathbf{x}$ and $\mathbf{y}$.
- Also called *dot product* or *scalar product*. Other representations:

$$\mathbf{x} \cdot \mathbf{y}, (\mathbf{x}, \mathbf{y}) \text{ and } < \mathbf{x}, \mathbf{y} >$$

- Represents similarity of vectors.
  - If $\mathbf{x}^T\mathbf{y} = 0$, then $\mathbf{x}$ and $\mathbf{y}$ are orthogonal vectors (in 2D, this means they are perpendicular).



Euclidean Norm

- *Euclidean norm* of vector

$$\|\mathbf{x}\| = \sqrt{\mathbf{x}^T\mathbf{x}} = \sqrt{x_1x_1 + x_2x_2 + \cdots + x_dx_d}$$

  represents the magnitude of the vector.
- *Euclidean distance* between points $\mathbf{x}$ and $\mathbf{y}$ can be computed as

$$\begin{aligned}\|\mathbf{x}-\mathbf{y}\| &= \sqrt{(\mathbf{x}-\mathbf{y})^T(\mathbf{x}-\mathbf{y})} \\ &= \sqrt{(x_1-y_1)^2 + (x_2-y_2)^2 + \cdots + (x_d-y_d)^2}\end{aligned}$$

- *Unit vector* has norm 1. Also called *normalised vector*.
- If $\|\mathbf{x}\| = 1$ and $\|\mathbf{y}\| = 1$, and $\mathbf{x}^T\mathbf{y} = 0$, then $\mathbf{x}$ and $\mathbf{y}$ are *orthonormal vectors*.

Outer Product

For vectors $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{z} \in \mathbb{R}^k$

- *Outer-product* $\mathbf{x}\mathbf{z}^T$ is a $d \times k$ matrix.

$$\mathbf{x}\mathbf{z}^T = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} \begin{bmatrix} z_1 & z_2 & \dots & z_k \end{bmatrix} = \begin{bmatrix} x_1 z_1 & x_1 z_2 & \dots & x_1 z_k \\ x_2 z_1 & x_2 z_2 & \dots & x_2 z_k \\ \vdots & \vdots & \vdots & \vdots \\ x_d z_1 & x_d z_2 & \dots & x_d z_k \end{bmatrix}$$

Matrix and Vector Calculus

For vector $\mathbf{x} \in \mathbb{R}^d$, scalar function $f(\mathbf{x})$ and vector function $\mathbf{g}(\mathbf{x}) \in \mathbb{R}^k$

- The gradient operator $\frac{d}{d\mathbf{x}}$ is also written as $\nabla_{\mathbf{x}}$ or simply ∇ when the differentiation variable is implied.
- $\nabla_{\mathbf{x}} = \begin{bmatrix} \frac{\partial}{\partial x_1} \\ \frac{\partial}{\partial x_2} \\ \vdots \\ \frac{\partial}{\partial x_d} \end{bmatrix}$ so that $\nabla_{\mathbf{x}}(f(\mathbf{x})) = \frac{d}{d\mathbf{x}}(f(\mathbf{x})) = \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \frac{\partial f(\mathbf{x})}{\partial x_2} \\ \vdots \\ \frac{\partial f(\mathbf{x})}{\partial x_d} \end{bmatrix}$
- $\nabla_{\mathbf{x}}(\mathbf{g}(\mathbf{x})) = \frac{d}{d\mathbf{x}}(\mathbf{g}(\mathbf{x})) = \begin{bmatrix} \frac{\partial g_1(\mathbf{x})}{\partial x_1} & \frac{\partial g_2(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial g_k(\mathbf{x})}{\partial x_1} \\ \frac{\partial g_1(\mathbf{x})}{\partial x_2} & \frac{\partial g_2(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial g_k(\mathbf{x})}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_1(\mathbf{x})}{\partial x_d} & \frac{\partial g_2(\mathbf{x})}{\partial x_d} & \cdots & \frac{\partial g_k(\mathbf{x})}{\partial x_d} \end{bmatrix}$

Matrix and Vector Calculus

For vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and matrices $\mathbf{M} \in \mathbb{R}^{k \times d}$ and $\mathbf{A} \in \mathbb{R}^{d \times d}$

1. $\nabla_{\mathbf{x}}(\mathbf{y}^T\mathbf{x}) = \nabla_{\mathbf{x}}(\mathbf{x}^T\mathbf{y}) = \mathbf{y}$
2. $\nabla_{\mathbf{x}}(\mathbf{M}\mathbf{x}) = \mathbf{M}^T$
3. $\nabla_{\mathbf{x}}(\mathbf{x}^T\mathbf{A}\mathbf{x}) = (\mathbf{A} + \mathbf{A}^T)\mathbf{x}$
4. For symmetric $\mathbf{A}$, $\nabla_{\mathbf{x}}(\mathbf{x}^T\mathbf{A}\mathbf{x}) = 2\mathbf{A}\mathbf{x}$

Matrix and Vector Calculus

Proof of $\nabla_{\mathbf{x}}(\mathbf{y}^T\mathbf{x}) = \nabla_{\mathbf{x}}(\mathbf{x}^T\mathbf{y}) = \mathbf{y}$

First note that

$$\mathbf{y}^T\mathbf{x} = \mathbf{x}^T\mathbf{y} = x_1y_1 + x_2y_2 + \cdots + x_dy_d \tag{1}$$

which is a scalar value.

$$\nabla_{\mathbf{x}}\left(\mathbf{x}^T\mathbf{y}\right) = \nabla_{\mathbf{x}}(x_1y_1 + x_2y_2 + \cdots + x_dy_d) \tag{2}$$

$$= \begin{bmatrix} \frac{d}{dx_1}(x_1y_1 + x_2y_2 + \cdots + x_dy_d) \\ \frac{d}{dx_2}(x_1y_1 + x_2y_2 + \cdots + x_dy_d) \\ \vdots \\ \frac{d}{dx_d}(x_1y_1 + x_2y_2 + \cdots + x_dy_d) \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_d \end{bmatrix} = \mathbf{y} \tag{3}$$

Matrix and Vector Calculus

Proof of $\nabla_{\mathbf{x}}(\mathbf{M}\mathbf{x}) = \mathbf{M}^T$

Let $\mathbf{m}_i^T$ denote the i-th row of matrix $\mathbf{M}$. Then we can write

$$\nabla_{\mathbf{x}}(\mathbf{M}\mathbf{x}) = \nabla_{\mathbf{x}} \begin{bmatrix} \mathbf{m}_1^T\mathbf{x} \\ \mathbf{m}_2^T\mathbf{x} \\ \vdots \\ \mathbf{m}_k^T\mathbf{x} \end{bmatrix} \tag{4}$$

$$= \begin{bmatrix} \nabla_{\mathbf{x}}(\mathbf{m}_1^T\mathbf{x}) & \nabla_{\mathbf{x}}(\mathbf{m}_2^T\mathbf{x}) & \dots & \nabla_{\mathbf{x}}(\mathbf{m}_k^T\mathbf{x}) \end{bmatrix} \tag{5}$$

$$= \begin{bmatrix} \mathbf{m}_1 & \mathbf{m}_2 & \dots & \mathbf{m}_k \end{bmatrix} = \mathbf{M}^T \tag{6}$$

Matrix and Vector Calculus

Proof of $\nabla_{\mathbf{x}}(\mathbf{x}^T\mathbf{A}\mathbf{x}) = (\mathbf{A}+\mathbf{A}^T)\mathbf{x}$

We will use the product rule of differentiation. When applied to vectors, the rule states that

$$\nabla_{\mathbf{x}}\left(\mathbf{u}^T\mathbf{v}\right) = \nabla_{\mathbf{x}}(\mathbf{u})\,\mathbf{v} + \nabla_{\mathbf{x}}(\mathbf{v})\,\mathbf{u} \tag{7}$$

where both $\mathbf{u}$ and $\mathbf{v}$ are functions of $\mathbf{x}$. For our problem, we will take $\mathbf{u}=\mathbf{x}$ and $\mathbf{v}=\mathbf{A}\mathbf{x}$. Then we can write

$$\begin{aligned}
\nabla_{\mathbf{x}}(\mathbf{x}^T\mathbf{A}\mathbf{x}) &= (\nabla_{\mathbf{x}}\mathbf{x})\,\mathbf{A}\mathbf{x} + (\nabla_{\mathbf{x}}\mathbf{A}\mathbf{x})\,\mathbf{x} && (8)\\
&= (\nabla_{\mathbf{x}}\mathbf{I}\mathbf{x})\,\mathbf{A}\mathbf{x} + (\nabla_{\mathbf{x}}\mathbf{A}\mathbf{x})\,\mathbf{x} && (9)\\
&= \mathbf{I}^T\mathbf{A}\mathbf{x} + \mathbf{A}^T\mathbf{x} && (10)\\
&= \mathbf{A}\mathbf{x} + \mathbf{A}^T\mathbf{x} && (11)\\
&= (\mathbf{A}+\mathbf{A}^T)\mathbf{x} && (12)
\end{aligned}$$

Matrix and Vector Calculus

Proof of $\nabla_{\mathbf{x}}(\mathbf{x}^T\mathbf{A}\mathbf{x}) = (\mathbf{A} + \mathbf{A}^T)\mathbf{x}$

When $\mathbf{A}$ is symmetric, $\mathbf{A}^T = \mathbf{A}$, and therefore $(\mathbf{A} + \mathbf{A}^T)\mathbf{x} = 2\mathbf{A}\mathbf{x}$ which proves the last derivative.

Matrices as linear operators

- In a matrix transformation $\mathbf{Mx}$, components of $\mathbf{x}$ are acted upon in a linear fashion.

$$\begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} m_{11}x_1 + m_{12}x_2 \\ m_{21}x_1 + m_{22}x_2 \end{bmatrix}$$

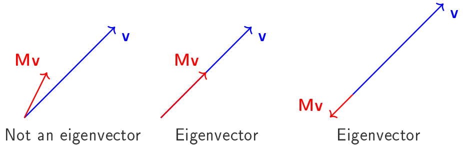
- *Every* matrix multiplication represents a linear transformation.
- *Every* linear transformation can be represented as a matrix multiplication.

Eigenvectors

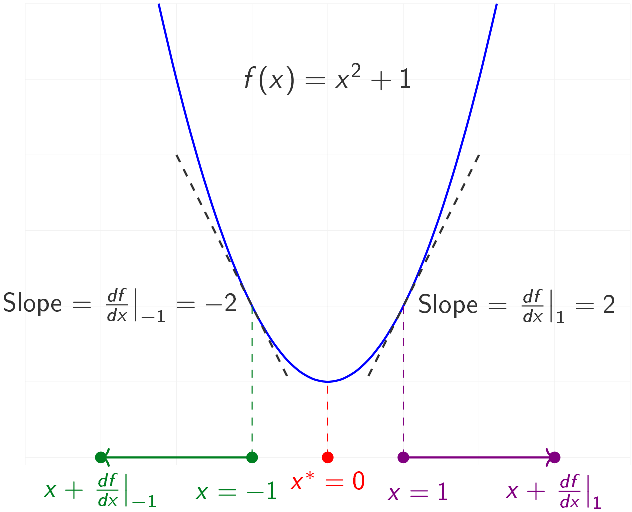
- When a square matrix $\mathbf{M}$ is multiplied with a vector $\mathbf{v}$, the vector is linearly transformed.
  - Rotation/Shearing/Scaling
  - Scaling does not change the direction of the vector.
- If vector $\mathbf{Mv}$ is only a scaled version of $\mathbf{v}$, then $\mathbf{v}$ is called an *eigenvector of* $\mathbf{M}$.
- That is, if $\mathbf{v}$ is an eigenvector of $\mathbf{M}$ then

$$\mathbf{Mv} = \lambda \mathbf{v}$$

where scaling factor λ is also called the *eigenvalue of* $\mathbf{M}$ *corresponding to eigenvector* $\mathbf{v}$.

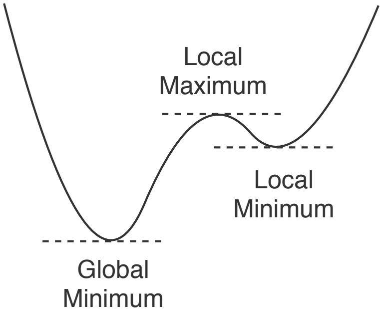


Minimization



What is the slope/derivative/gradient at the minimizer $x^* = 0$?

Minimization
Local vs. Global Minima



- *Stationary point*: where derivative is 0.
- A stationary point can be a minimum or a maximum.
- A minimum can be local or global. Same for maximum.

Constrained Optimization

- For optimizing a function $f(\mathbf{x})$, the gradient of f must vanish at the optimizer $\mathbf{x}^*$.

$$\nabla f|_{\mathbf{x}^*} = \mathbf{0}$$

- For optimizing a function $f(\mathbf{x})$ *subject to some constraint $g(\mathbf{x}) = 0$*, the gradient of the so-called Lagrange function

$$L(\mathbf{x}, \lambda) = f(\mathbf{x}) + \lambda g(\mathbf{x})$$

must vanish at the optimizer $\mathbf{x}^*$. That is,

$$\nabla L(\mathbf{x}, \lambda) = \nabla f|_{\mathbf{x}^*} + \lambda \nabla g|_{\mathbf{x}^*} = \mathbf{0}$$

where λ is the Lagrange (or undetermined) multiplier.

Constrained Optimization

- Quite often, we will need to maximize $\mathbf{x}^T\mathbf{M}\mathbf{x}$ with respect to $\mathbf{x}$ where $\mathbf{M}$ is a symmetric, positive-definite[1] matrix.
  - Trivial solution: $\mathbf{x} = \mathbf{inf}$
- To prevent trivial solution, we must constrain the norm of $\mathbf{x}$. For example, $\mathbf{x}^T\mathbf{x} = 1$.
- This gives us a constrained optimization problem.

Maximize $f(\mathbf{x}) = \mathbf{x}^T\mathbf{M}\mathbf{x}$ *subject to the constraint* $g(\mathbf{x}) = \mathbf{x}^T\mathbf{x} - 1 = 0.$

- Lagrangian becomes $L(\mathbf{x}, \lambda) = \mathbf{x}^T\mathbf{M}\mathbf{x} + \lambda(1 - \mathbf{x}^T\mathbf{x})$
- Use $\nabla_{\mathbf{x}} L|_{\mathbf{x}^*} = \mathbf{0}$ and $\nabla_{\lambda} L|_{\lambda^*} = 0$ to solve for optimal $\mathbf{x}^*$.

[1] $\mathbf{x}^T\mathbf{M}\mathbf{x} > 0$ for all $\mathbf{x} \neq \mathbf{0}$