Linear Algebra Review[1]

Reinforcement Learning (INF8250AE)
Fall 2025

Polytechnique Montréal

[1] Some parts borrowed from Punit Shah's slides.

Basics

- A scalar is a number.

- A vector is a 1-D array of numbers. The set of vectors of length n with real elements is denoted by $\mathbb{R}^n$.
  - Vectos can be multiplied by a scalar.
  - Vector can be added together if dimensions match.

- A matrix is a 2-D array of numbers. The set of $m \times n$ matrices with real elements is denoted by $\mathbb{R}^{m \times n}$.
  - Matrices can be added together or multiplied by a scalar.
  - We can multiply Matrices to a vector if dimensions match.

- In the rest of this tutorial we denote scalars with lowercase letters like a, vectors with bold lowercase $\mathbf{v}$, and matrices with bold uppercase $\mathbf{A}$.

Diagonal Matrix

- A diagonal matrix has all entries equal to zero except the diagonal entries which might or might not be zero, e.g. identity matrix.

- A square diagonal matrix with diagonal enteries given by entries of vector $\mathbf{v}$ is denoted by $\text{diag}(\mathbf{v})$.

- Multiplying vector $\mathbf{x}$ by a diagonal matrix is efficient:

$$\text{diag}(\mathbf{v})\mathbf{x} \;=\; \mathbf{v} \odot \mathbf{x},$$

  where $\odot$ is the entrywise product.

- Inverting a square diagonal matrix is efficient

$$\text{diag}(\mathbf{v})^{-1} \;=\; \text{diag}\Big([\frac{1}{v_1}, \ldots, \frac{1}{v_n}]^\top\Big).$$

Trace

- Trace is the sum of all the diagonal elements of a matrix, i.e.,

$$\mathrm{Tr}(\mathbf{A}) = \sum_i A_{i,i}.$$

- Cyclic property:

$$\mathrm{Tr}(\mathbf{ABC}) = \mathrm{Tr}(\mathbf{CAB}) = \mathrm{Tr}(\mathbf{BCA}).$$

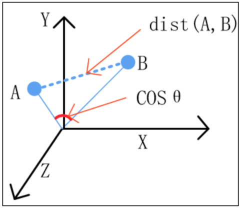
Transposition

- Transposition is an operation on matrices (and vectors) that interchange rows with columns. $(\mathbf{A}^\top)_{i,j} = \mathbf{A}_{j,i}$.
- $(\mathbf{AB})^\top = \mathbf{B}^\top \mathbf{A}^\top$.
- $\mathbf{A}$ is called symmetric when $\mathbf{A} = \mathbf{A}^\top$.
- $\mathbf{A}$ is called orthogonal when $\mathbf{A}\mathbf{A}^\top = \mathbf{A}^\top\mathbf{A} = \mathbf{I}$ or $\mathbf{A}^{-1} = \mathbf{A}^\top$.

Dot Product

- Dot product is defined as $\langle \mathbf{v}, \mathbf{u} \rangle = \mathbf{v} \cdot \mathbf{u} = \mathbf{v}^\top \mathbf{u} = \sum_i u_i v_i$.
- The ℓ_2 norm can be written in terms of dot product: $\|\mathbf{u}\|_2 = \sqrt{\mathbf{u}.\mathbf{u}}$.
- Dot product of two vectors can be written in terms of their ℓ_2 norms and the angle θ between them:

$$\mathbf{a}^\top \mathbf{b} = \|\mathbf{a}\|_2 \|\mathbf{b}\|_2 \cos(\theta).$$



Cosine Similarity

- Cosine between two vectors is a measure of their similarity:

$$\cos(\theta) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\| \, \|\mathbf{b}\|}.$$

- **Orthogonal Vectors**: Two vectors $\mathbf{a}$ and $\mathbf{b}$ are orthogonal to each other if $\mathbf{a} \cdot \mathbf{b} = 0$.

Vector Projection

- Given two vectors $\mathbf{a}$ and $\mathbf{b}$, let $\hat{\mathbf{b}} = \frac{\mathbf{b}}{\|\mathbf{b}\|}$ be the unit vector in the direction of $\mathbf{b}$.
- Then $\mathbf{a}_1 = a_1 \cdot \hat{\mathbf{b}}$ is the orthogonal projection of $\mathbf{a}$ onto a straight line parallel to $\mathbf{b}$, where

$$a_1 = \|\mathbf{a}\| \cos(\theta) = \mathbf{a} \cdot \hat{\mathbf{b}} = \mathbf{a} \cdot \frac{\mathbf{b}}{\|\mathbf{b}\|}$$

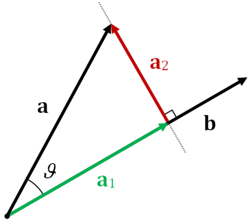


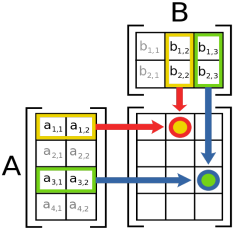
Image taken from wikipedia.

Multiplication

- Matrix-vector multiplication is a linear transformation. In other words,

$$\mathbf{M}(v_1 + av_2) = \mathbf{M}v_1 + a\mathbf{M}v_2 \implies (\mathbf{M}v)_i = \sum_j M_{i,j}v_j.$$

- Matrix-matrix multiplication is the composition of linear transformations, i.e.,
$(\mathbf{AB})v = \mathbf{A}(\mathbf{B}v) \implies (\mathbf{AB})_{i,j} = \sum_k A_{i,k}B_{k,j}.$



Norms

- Norms measure how "large" a vector is. They can be defined for matrices too.

- The ℓ_p-norm for a vector $\mathbf{x}$:

$$\|\mathbf{x}\|_p = \left[\sum_i |x_i|^p\right]^{\frac{1}{p}}.$$

  - The ℓ_2-norm is known as the Euclidean norm.
  - The ℓ_1-norm is known as the Manhattan norm, i.e., $\|\mathbf{x}\|_1 = \sum_i |x_i|$.
  - The ℓ_∞ is the max (or supremum) norm, i.e., $\|\mathbf{x}\|_\infty = \max_i |x_i|$.

Matrix Norms: Basics

- A **matrix norm** $\|A\|$ is a function assigning a nonnegative number to a matrix A.
- Properties (similar to vector norms):
  1. $\|A\| \geq 0$ and $\|A\| = 0 \iff A = 0$.
  2. $\|\alpha A\| = |\alpha| \cdot \|A\|$.
  3. $\|A + B\| \leq \|A\| + \|B\|$.

Operator (Induced) Norms

- Given a vector norm $\|\cdot\|$, the **induced matrix norm** is

$$\|A\| = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|}.$$

- Intuition: measures how much A can stretch a vector.
- Properties:
  1. $\|AB\| \leq \|A\|\|B\|$ (submultiplicative).
  2. $\|Ax\| \leq \|A\| \cdot \|x\|$.

Examples of Matrix Norms

- **Spectral norm:**

$$\|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2.$$

- **Frobenius norm:**

$$\|A\|_F = \sqrt{\sum_{i,j} |a_{ij}|^2}.$$

- ℓ_1**-norm:**
  - **Induced:**

$$\|A\|_1 = \max_j \sum_i |a_{ij}|$$

  - **Entrywise:**

$$\|A\|_{1,\text{entry}} = \sum_{i,j} |a_{ij}|$$

- ℓ_∞**-norm:**
  - **Induced:**

$$\|A\|_\infty = \max_i \sum_j |a_{ij}|$$

  - **Entrywise:**

$$\|A\|_{\infty,\text{entry}} = \max_{i,j} |a_{ij}|$$

Frobenius Norm and Trace

- Frobenius norm can also be written using the trace:

$$\|A\|_F = \sqrt{\sum_{i,j} |A_{i,j}|^2} = \sqrt{\mathrm{Tr}(A^\top A)}$$

- Intuition: it's the Euclidean norm of the flattened matrix.
- Useful in optimization: appears naturally in least-squares and SVD approximations.

Invertibility

- $\mathbf{I}$ denotes the identity matrix which is a square matrix of zeros with ones along the diagonal. It has the property $\mathbf{IA} = \mathbf{A}$ ($\mathbf{BI} = \mathbf{B}$) and $\mathbf{Iv} = \mathbf{v}$

- A square matrix $\mathbf{A}$ is invertible if $\mathbf{A}^{-1}$ exists such that $\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$.

- Not all non-zero matrices are invertible, e.g., the following matrix is not invertible:

$$\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

Inverse of Special Matrices

- **Diagonal Matrix:** $D = \mathrm{diag}(d_1, \ldots, d_n)$

$$D^{-1} = \mathrm{diag}\Big(\frac{1}{d_1}, \ldots, \frac{1}{d_n}\Big)$$

- **Block-Diagonal Matrix:** $A = \mathrm{diag}(A_1, A_2, \ldots, A_k)$

$$A^{-1} = \mathrm{diag}(A_1^{-1}, A_2^{-1}, \ldots, A_k^{-1})$$

- **Orthogonal Matrix:** $Q^\top Q = I$

$$Q^{-1} = Q^\top$$

- **Lower-Triangular Matrix:** L (invertible)

L^{-1} can be computed efficiently via forward substitution.

Lemma A.2 (Golub & Van Loan, 2013)

Lemma: If $A \in \mathbb{R}^{d\times d}$ and $\|A\|_p < 1$, then $I - A$ is non-singular, and

$$\|(I - A)^{-1}\|_p \leq \frac{1}{1 - \|A\|_p}.$$

and

$$(I - A)^{-1} = \sum_{k=0}^{\infty} A^k.$$

Incremental Matrix Inversion (Golub & Van Loan, 2013)

Sherman–Morrison–Woodbury formula: For $A \in \mathbb{R}^{d\times d}$ invertible, and $U, V \in \mathbb{R}^{d\times k}$,

$$(A + UV^\top)^{-1} = A^{-1} - A^{-1}U\big(I + V^\top A^{-1}U\big)^{-1}V^\top A^{-1},$$

assuming A and $(I + V^\top A^{-1}U)$ are invertible.

- $UV^\top$ is a rank-k update of A.
- Inverting the smaller $k \times k$ matrix $\big(I + V^\top A^{-1}U\big)$ is much cheaper than re-inverting the full $d \times d$ matrix.

Special case (Sherman–Morrison): For $u, v \in \mathbb{R}^d$,

$$(A + uv^\top)^{-1} = A^{-1} - \frac{A^{-1}uv^\top A^{-1}}{1 + v^\top A^{-1}u},$$

provided $1 + v^\top A^{-1}u \neq 0$.

Determinant

- Determinant of a square matrix is a mapping to scalars.

$$\det(\mathbf{A}) \quad \text{or} \quad |\mathbf{A}|$$

- Measures how much multiplication by the matrix expands or contracts the space.
- Determinant of product is the product of determinants:

$$\det(\mathbf{AB}) = \det(\mathbf{A})\det(\mathbf{B})$$

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$$

List of Equivalencies

Assuming that $\mathbf{A}$ is a square matrix, the following statements are equivalent

- $\mathbf{Ax} = \mathbf{b}$ has a **unique** solution (for every b with correct dimension).
- $\mathbf{Ax} = \mathbf{m}0$ has a unique, trivial solution: $\mathbf{x} = \mathbf{m}0$.
- Columns of $\mathbf{A}$ are linearly independent.
- $\mathbf{A}$ is invertible, i.e. $\mathbf{A}^{-1}$ exists.
- $\det(\mathbf{A}) \neq 0$

Zero Determinant

If $\det(\mathbf{A}) = 0$, then:

- $\mathbf{A}$ is linearly dependent.
- $\mathbf{Ax} = \mathbf{b}$ has infinitely many solutions or no solution. These cases correspond to when b is in the span of columns of $\mathbf{A}$ or out of it.
- $\mathbf{Ax} = \mathbf{m}0$ has a non-zero solution. (since every scalar multiple of one solution is a solution and there is a non-zero solution we get infinitely many solutions.)

Matrix Decomposition

- We can decompose an integer into its prime factors, e.g., $12 = 2 \times 2 \times 3$.
- Similarly, matrices can be decomposed into product of other matrices.

$$\mathbf{A} \; = \; \mathbf{V}\text{diag}(\boldsymbol{\lambda})\mathbf{V}^{-1}$$

- Examples are Eigendecomposition, SVD, Schur decomposition, LU decomposition, ...

Eigenvectors

- An eigenvector of a square matrix $\mathbf{A}$ is a nonzero vector $\mathbf{v}$ such that multiplication by $\mathbf{A}$ only changes the scale of $\mathbf{v}$.

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

- The scalar λ is known as the **eigenvalue**.
- If $\mathbf{v}$ is an eigenvector of $\mathbf{A}$, so is any rescaled vector $s\mathbf{v}$. Moreover, $s\mathbf{v}$ still has the same eigenvalue. Thus, we constrain the eigenvector to be of unit length:

$$||\mathbf{v}||_2 = 1$$

Characteristic Polynomial(1)

- Eigenvalue equation of matrix **A**.

$$\begin{aligned} \mathbf{A}\mathbf{v} &= \lambda\mathbf{v} \\ \lambda\mathbf{v} - \mathbf{A}\mathbf{v} &= \mathbf{m}0 \\ (\lambda\mathbf{I} - \mathbf{A})\mathbf{v} &= \mathbf{m}0 \end{aligned}$$

- If nonzero solution for **v** exists, then it must be the case that:

$$\det(\lambda\mathbf{I} - \mathbf{A}) = 0$$

- Unpacking the determinant as a function of λ, we get:

$$P_A(\lambda) = \det(\lambda\mathbf{I} - \mathbf{A}) = 1 \times \lambda^n + c_{n-1} \times \lambda^{n-1} + \ldots + c_0$$

- This is called the characterisitc polynomial of A.

Characteristic Polynomial(2)

- If $\lambda_1, \lambda_2, \ldots, \lambda_n$ are roots of the characteristic polynomial, they are eigenvalues of $\mathbf{A}$ and we have $P_A(\lambda) = \prod_{i=1}^{n}(\lambda - \lambda_i)$.
- $c_{n-1} = -\sum_{i=1}^{n} \lambda_i = -tr(A)$. This means that the sum of eigenvalues equals to the trace of the matrix.
- $c_0 = (-1)^n \prod_{i=1}^{n} \lambda_i = (-1)^n det(\mathbf{A})$. The determinant is equal to the product of eigenvalues.
- Roots might be complex. If a root has multiplicity of $r_j > 1$ (This is called the algebraic dimension of eigenvalue), then the geometric dimension of eigenspace for that eigenvalue might be less than r_j (or equal but never more). But for every eigenvalue, one eigenvector is guaranteed.

Example

- Consider the matrix:

$$\mathbf{A} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

- The characteristic polynomial is:

$$\det(\lambda \mathbf{I} - \mathbf{A}) = \det \begin{bmatrix} \lambda - 2 & -1 \\ -1 & \lambda - 2 \end{bmatrix} = 3 - 4\lambda + \lambda^2 = 0$$

- It has roots $\lambda = 1$ and $\lambda = 3$ which are the two eigenvalues of $\mathbf{A}$.

- We can then solve for eigenvectors using $\mathbf{A}\mathbf{v} = \lambda \mathbf{v}$:

$$\mathbf{v}_{\lambda=1} = [1, -1]^\top \quad \text{and} \quad \mathbf{v}_{\lambda=3} = [1, 1]^\top$$

Eigendecomposition

- Suppose that $n \times n$ matrix $\mathbf{A}$ has n linearly independent eigenvectors $\{\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(n)}\}$ with eigenvalues $\{\lambda_1, \ldots, \lambda_n\}$.
- Concatenate eigenvectors (as columns) to form matrix $\mathbf{V}$.
- Concatenate eigenvalues to form vector $\mathbf{m}\lambda = [\lambda_1, \ldots, \lambda_n]^\top$.
- The **eigendecomposition** of $\mathbf{A}$ is given by:

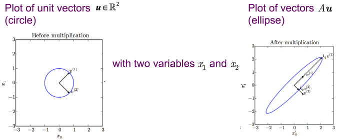
$$\mathbf{AV} \;=\; \mathbf{V}diag(\lambda) \implies \mathbf{A} \;=\; \mathbf{V}\mathrm{diag}(\mathbf{m}\lambda)\mathbf{V}^{-1}$$

Symmetric Matrices

- Every symmetric (hermitian) matrix of dimension n has a set of (not necessarily unique) n orthogonal eigenvectors. Furthermore, all eigenvalues are real.
- Every real symmetric matrix $\mathbf{A}$ can be decomposed into real-valued eigenvectors and eigenvalues:

$$\mathbf{A} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^\top$$

- $\mathbf{Q}$ is an orthogonal matrix of the eigenvectors of $\mathbf{A}$, and $\boldsymbol{\Lambda}$ is a diagonal matrix of eigenvalues.
- We can think of $\mathbf{A}$ as scaling space by λ_i in direction $\mathbf{v}^{(i)}$.



Eigendecomposition is not Unique

- Decomposition is not unique when two eigenvalues are the same.
- By convention, order entries of $\boldsymbol{\Lambda}$ in descending order. Then, eigendecomposition is unique if all eigenvalues have multiplicity equal to one.
- If any eigenvalue is zero, then the matrix is **singular**. Because if $\mathbf{v}$ is the corresponding eigenvector we have: $\mathbf{A}\mathbf{v} = 0\mathbf{v} = 0$.

Positive Definite Matrix

- If a symmetric matrix A has the property:

$$\mathbf{x}^\top \mathbf{A}\mathbf{x} > 0 \quad \text{for any nonzero vector } \mathbf{x}$$

  Then A is called **positive definite**.

- If the above inequality is not strict then A is called **positive semidefinite**.

- For positive (semi)definite matrices all eigenvalues are positive(non negative).

Cholesky Decomposition (Positive Definite Matrices)

- For a symmetric positive definite matrix $A \in \mathbb{R}^{n \times n}$, there exists a lower-triangular matrix L such that

$$A = LL^\top$$

- Useful properties:
  - Solving $Ax = b$ reduces to forward/backward substitution.
  - Numerically stable, used in Gaussian processes, Kalman filters, and RL covariance updates.
  - Efficient way to test positive definiteness: attempt Cholesky decomposition.
- Example:

$$A = \begin{bmatrix} 4 & 2 \\ 2 & 3 \end{bmatrix}, \quad L = \begin{bmatrix} 2 & 0 \\ 1 & \sqrt{2} \end{bmatrix}, \quad A = LL^\top$$

Singular Value Decomposition (SVD)

- If $\mathbf{A}$ is not square, eigendecomposition is undefined.
- **SVD** is a decomposition of the form $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$.
- SVD is more general than eigendecomposition.
- Every real matrix has a SVD.

SVD Definition (1)

- Write $\mathbf{A}$ as a product of three matrices: $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$.
- If $\mathbf{A}$ is $m \times n$, then $\mathbf{U}$ is $m \times m$, $\mathbf{D}$ is $m \times n$, and $\mathbf{V}$ is $n \times n$.
- $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices, and $\mathbf{D}$ is a diagonal matrix (not necessarily square).
- Diagonal entries of $\mathbf{D}$ are called **singular values** of $\mathbf{A}$.
- Columns of $\mathbf{U}$ are the **left singular vectors**, and columns of $\mathbf{V}$ are the **right singular vectors**.

SVD Definition (2)

- SVD can be interpreted in terms of eigendecompostion.
- Left singular vectors of $\mathbf{A}$ are the eigenvectors of $\mathbf{A}\mathbf{A}^\top$.
- Right singular vectors of $\mathbf{A}$ are the eigenvectors of $\mathbf{A}^\top\mathbf{A}$.
- Nonzero singular values of $\mathbf{A}$ are square roots of eigenvalues of $\mathbf{A}^\top\mathbf{A}$ and $\mathbf{A}\mathbf{A}^\top$.
- Numbers on the diagonal of D are sorted largest to smallest and are non-negative ($\mathbf{A}^\top\mathbf{A}$ and $\mathbf{A}\mathbf{A}^\top$ are semipositive definite.).

SVD Optimality

- Given a matrix $\mathbf{A}$, SVD allows us to find its "best" (to be defined) rank-r approximation $\mathbf{A}_r$.
- We can write $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$ as $\mathbf{A} = \sum_{i=1}^{n} d_i \mathbf{u}_i \mathbf{v}_i^\top$.
- For $r \leq n$, construct $\mathbf{A}_r = \sum_{i=1}^{r} d_i \mathbf{u}_i \mathbf{v}_i^\top$.
- The matrix $\mathbf{A}_r$ is a rank-r approximation of A. Moreover, it is the best approximation of rank r by many norms:
  - When considering the operator (or spectral) norm, it is optimal. This means that $\|A - A_r\|_2 \leq \|A - B\|_2$ for any rank r matrix B.
  - When considering Frobenius norm, it is optimal. This means that $\|A - A_r\|_F \leq \|A - B\|_F$ for any rank r matrix B. One way to interpret this inequality is that rows (or columns) of A_r are the projection of rows (or columns) of A on the best r dimensional subspace, in the sense that this projection minimizes the sum of squared distances.

Norms and Singular Values

- Definition (induced from ℓ_2 vector norm):

$$\|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2.$$

- Equivalent formulations:

$$\|A\|_2 = \sqrt{\lambda_{\max}(A^\top A)}.$$

- In terms of singular values:

$$\|A\|_2 = \sigma_{\max}(A),$$

the largest singular value of A.

Norms and Singular Values

Let $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ be the singular values of A.

Norm	**Expression in terms of singular values**
Spectral norm ($\|A\|_2$)	$\|A\|_2 = \sigma_1$
Frobenius norm ($\|A\|_F$)	$\|A\|_F = \sqrt{\sum_{i=1}^r \sigma_i^2}$
Nuclear norm ($\|A\|_*$)	$\|A\|_* = \sum_{i=1}^r \sigma_i$
Schatten-p norm ($\|A\|_p$)	$\|A\|_p = \left(\sum_{i=1}^r \sigma_i^p \right)^{1/p}$

- Special cases: $p = 1$ (nuclear norm), $p = 2$ (Frobenius norm), $p = \infty$ (spectral norm).
- Entrywise norms ($\|A\|_1, \|A\|_\infty$) are not singular-value based.

Matrix Calculus: Basics

- We often need derivatives of scalar, vector, or matrix functions with respect to vectors/matrices.
- Notation:
  - $\frac{\partial f}{\partial \mathbf{x}}$: gradient of scalar f w.r.t vector $\mathbf{x}$ (column vector).
  - $\frac{\partial \mathbf{f}}{\partial \mathbf{x}}$: Jacobian of vector $\mathbf{f}(\mathbf{x})$ w.r.t vector $\mathbf{x}$.
  - $\frac{\partial F}{\partial \mathbf{X}}$: matrix of derivatives of scalar or matrix function F w.r.t matrix $\mathbf{X}$.
- Example:

$$f(\mathbf{x}) = \mathbf{a}^\top \mathbf{x} \implies \frac{\partial f}{\partial \mathbf{x}} = \mathbf{a}$$

Common Vector Derivatives

- Scalar $f = \mathbf{a}^\top \mathbf{x}$: $\frac{\partial f}{\partial \mathbf{x}} = \mathbf{a}$
- Scalar $f = \mathbf{x}^\top \mathbf{A}\mathbf{x}$: $\frac{\partial f}{\partial \mathbf{x}} = (\mathbf{A} + \mathbf{A}^\top)\mathbf{x}$
- Scalar $f = \|\mathbf{x}\|_2^2$: $\frac{\partial f}{\partial \mathbf{x}} = 2\mathbf{x}$
- Scalar $f = \mathrm{Tr}(\mathbf{A}\mathbf{X})$: $\frac{\partial f}{\partial \mathbf{X}} = \mathbf{A}^\top$
- Scalar $f = \mathrm{Tr}(\mathbf{X}^\top \mathbf{A}\mathbf{X})$: $\frac{\partial f}{\partial \mathbf{X}} = (\mathbf{A} + \mathbf{A}^\top)\mathbf{X}$

Useful Matrix Calculus Identities

- $\frac{\partial}{\partial X} \operatorname{Tr}(X) = I$
- $\frac{\partial}{\partial X} \log \det X = (X^{-1})^\top$
- $\frac{\partial}{\partial X} \operatorname{Tr}(AXBX^\top) = A^\top X B^\top + AXB$
- $\frac{\partial}{\partial X} \|X\|_F^2 = 2X$
- $\frac{\partial}{\partial x} x^\top A x = (A + A^\top)x$

Matrix Calculus in RL: Example

Problem: Least-Squares TD error

$$L(\mathbf{w}) = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2$$

where $\mathbf{X} \in \mathbb{R}^{n \times d}$ is a feature matrix, $\mathbf{w} \in \mathbb{R}^d$ are weights, and $\mathbf{y} \in \mathbb{R}^n$ are target values.

Gradient computation:

$$\frac{\partial L}{\partial \mathbf{w}} = 2\mathbf{X}^\top(\mathbf{X}\mathbf{w} - \mathbf{y})$$

Explanation:

- $f(\mathbf{w}) = \mathbf{X}\mathbf{w} - \mathbf{y}$, then $L = f^\top f$.
- Use rule $\frac{\partial}{\partial x}\|Ax - b\|_2^2 = 2A^\top(Ax - b)$
- Gradient points in direction of steepest increase; update rule: $\mathbf{w} \leftarrow \mathbf{w} - \eta \frac{\partial L}{\partial \mathbf{w}}$

Takeaway: Matrix calculus simplifies computing gradients for loss functions common in RL and ML.

Computational LA Tricks: Making Matrices Invertible

- Sometimes matrices are nearly singular, causing numerical issues.
- **Trick: Add a small diagonal (Tikhonov regularization / ridge):**

$$A_{\text{reg}} = A + \epsilon I$$

  where $\epsilon > 0$ small.
- Ensures A_{reg} is invertible and better conditioned.
- Widely used in **ridge regression, TD(λ), policy evaluation**.

Computational LA Tricks: Solving $Ax = b$ with Cholesky Decomposition

- Applicable when A is **symmetric positive definite**.
- Compute the **Cholesky decomposition**:

$$A = LL^\top,$$

  where L is lower triangular.

- Solve in two steps:
  1. **Forward substitution:** solve $Ly = b$ for y.
  2. **Backward substitution:** solve $L^\top x = y$ for x.
- Efficient and numerically stable; avoids explicit matrix inversion.
- Useful in RL for least-squares problems:

$$w = (X^\top X)^{-1} X^\top y \quad \Rightarrow \quad w \text{ via Cholesky of } X^\top X$$

Computational LA Tricks: Incremental / Online Updates

- When a new sample (x_t, y_t) arrives:

$$A_t = A_{t-1} + x_t x_t^\top, \quad b_t = b_{t-1} + x_t y_t$$

- Use Sherman-Morrison or rank-1 updates to update $A^{-1}b$ efficiently.
- Essential for online RL algorithms.

Computational LA Tricks: Eigen Decomposition Tricks

- For SPD or diagonalizable A, $A = V\Lambda V^{-1}$.
- Compute matrix powers or exponentials efficiently:

$$A^k = V\Lambda^k V^{-1}, \quad \exp(A) = V\exp(\Lambda)V^{-1}$$

- Useful in RL for propagating value functions or transition operators.

Computational LA Tricks: Preconditioning

- Idea: Improve convergence of iterative solvers by reducing the condition number of the matrix.
$$\kappa(A) = \|A\| \cdot \|A^{-1}\|$$
- Preconditioner M: Choose $M \approx A^{-1}$ to solve
$$MAx = Mb$$
instead of $Ax = b$. This reduces $\kappa(MA) \ll \kappa(A)$, leading to faster convergence.
- Intuition: Preconditioning "reshapes" the problem so all directions are stretched more evenly.
- Example in RL: Policy evaluation using Least-Squares Temporal Difference (LSTD):
$$w = (X^\top X)^{-1} X^\top y$$
- If $X^\top X$ is ill-conditioned, solve instead:
$$(MX^\top X)w = MX^\top y$$
using a diagonal preconditioner $M = \text{diag}(1/\text{diag}(X^\top X))$
- This improves numerical stability and speeds up iterative solvers.