1 Linear Algebra

Vectors and matrices play a foundational role in machine learning, since they are the main objects we use to represent data and parameters. Much of the emphasis in this section is studying “how big” these objects are.

1. Vector norms: let $v \in \mathbb{R}^d$ be a $d$-dimensional real vector. For each $p \geq 0$, we can measure the $p$-norm of $v$:

$$\|v\|_p = \left( \sum_{i=1}^{d} |v_i|^p \right)^{1/p}.$$  

Important special cases:

- $p = 0$: number of non-zero entries in $v$.
- $p = 1$: used for encouraging sparsity.
- $p = 2$: rotationally invariant (simply written as $\|v\|$)
- $p = \infty$: maximum magnitude of an entry ($\|v\|_\infty = \max_{i=1}^{d} |v_i|$).

2. A vector norm defines a distance metric between two vectors:

$$\|v_1 - v_2\|_p.$$  

3. A vector norm also allows us to define a ball, the set of points with bounded norm:

$$B_p = \{ v \in \mathbb{R}^d : \|v\|_p \leq 1 \}.$$  

4. Relationship between vector norms: As you increase $p$, the norm decreases.

$$\| \cdot \|_\infty \leq \| \cdot \|_2 \leq \| \cdot \|_1.$$  

As you increase $p$, the balls $B_p$ increases in size.

You can also relate the norms in the other direction with an additional factor of $\sqrt{d}$:

$$\|v\|_1 \leq \sqrt{d} \|v\|_2 \quad \|v\|_2 \leq \sqrt{d} \|v\|_\infty$$  

(1)

To remember these relationships, imagine $v = [1, \ldots, 1]$, which has $\|v\|_1 = d, \|v\|_2 = \sqrt{d}$ and $\|v\|_\infty = 1$. 

5. Hölder’s inequality: \(|u \cdot v| \leq \|u\|_p \|v\|_q\) if \(\frac{1}{p} + \frac{1}{q} = 1\). An important special case is the Cauchy-Schwartz inequality \((p = q = 2)\). The main other common case is \(p = 1\) and \(q = \infty\). This is a very important inequality which allows us to decouple things: \(u \cdot v\) could have complex interactions, but we can control the magnitude of \(u\) and \(v\) separately.

6. Now let’s turn from vectors to matrices. Matrices are nominally just a two-dimensional array of numbers, but they have a deeper structure. There are two useful ways to think about matrices.

   First, we can think of an arbitrary matrix \(A \in \mathbb{R}^{m \times n}\) as an linear operator (function) that maps an input vector \(x \in \mathbb{R}^n\) to an output \(Ax \in \mathbb{R}^m\).

   Second, we can think of a positive semidefinite (PSD) matrix as specifying an ellipse or the covariance matrix of a multivariate Gaussian distribution.

7. Singular value decomposition: Let’s consider the linear operator view. Every matrix \(A \in \mathbb{R}^{m \times n}\) of rank \(k\) has a unique thin singular value decomposition (SVD)

   \[ A = U S V^\top, \]

   where the columns of \(U \in \mathbb{R}^{m \times k}\) are the left singular vectors, the columns of \(V \in \mathbb{R}^{n \times k}\) are the right singular vectors, and \(S = \text{diag}(\sigma) \in \mathbb{R}^{k \times k}\) is a diagonal matrix with singular values \(\sigma = (\sigma_1, \ldots, \sigma_k)\); by convention, we assume the entries of \(\sigma\) are non-increasing. The rank of \(A\) is \(k\).

   The right singular vectors \(V\) form an orthogonal basis for the input space \((V^\top V = I)\); the left singular vectors \(U\) form an orthogonal basis for the output space \((U^\top U = I)\).

   The interpretation of \(Ax = U S V^\top x\) is as follows: project \(x\) onto \(V\), scale by \(\sigma\), and then put into the basis of \(U\).

8. Pseudoinverse can be defined effortlessly by passing the inversion operator to the singular values:

   \[ A^\dagger = V S^{-1} U^\top. \]

9. Singular value decompositions expose the singular values, which allows us to think about matrices as vectors for the purposes of computing norms. Recall that the singular values are in some sense how much the input is scaled up. These norms are called Schatten norms:

   \[ \|A\|_p = \|\sigma\|_p \]

   where \(\sigma = (\sigma_1, \ldots, \sigma_k)\) are the singular values of \(A\) defined in 7.

   Important special cases:

   - \(p = 1\): nuclear norm, which is just the sum of the singular values, also written \(\|A\|_*\).
   - \(p = 2\): Frobenius norm, also written \(\|A\|_F\).
   - \(p = \infty\): spectral or operator norm, also written \(\|A\|_2, \|A\|_\text{op}\).

   We inherit the relationship between matrix norms from vectors:

   \[ \|\cdot\|_2 \leq \|\cdot\| \leq \|\cdot\|_* . \]
10. Recall that matrices are linear operators. *Induced norms* allow us to measure how much the matrix “scales up” the input:

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

(Unfortunately, there is a clash in notation with Schatten norms.)

Important special cases:

- \(p = 1\): \(\max_{1 \leq j \leq n} \sum_{i=1}^m |A_{ij}|\), is the maximum \(L_1\) norm over all columns (can skip this).
- \(p = \infty\): \(\max_{1 \leq i \leq m} \sum_{j=1}^n |A_{ij}|\), is the maximum \(L_1\) norm over all rows (can skip this).
- \(p = 2\): spectral or operator norm, also written \(\|A\|\). This is the most important induced norm. Note that it is also a Schatten norm with \(p = \infty\).

For an induced \(p\)-norm:

$$\|Ax\|_p \leq \|A\|_p \|x\|_p.$$

11. Now let us consider the case where \(A \in \mathbb{R}^{n \times n}\) is a symmetric positive semidefinite matrix. In this case, the SVD is an eigendecomposition:

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

so \(S = \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)\) are the eigenvalues (also singular values) and the columns of \(U\) are the eigenvectors (also singular vectors). Note that \(U^\top U = I\). (Eigendecompositions are more general, but let’s not worry about it here.)

An alternative characterization of PSD matrices is

$$v^\top Av \geq 0 \text{ for all } v.$$

If we think about the ellipsoid defined by \(\{v \in \mathbb{R}^n : v^\top Av = 1\}\). Then \(\Lambda\) governs how elongated this ellipsoid is.

12. The trace of a matrix is the sum of its diagonal entries:

$$\text{tr}(A) = \sum_{i=1}^n A_{ii}.$$

13. The cyclic property of traces is pretty useful:

$$\text{tr}(AB) = \text{tr}(BA).$$

14. If \(A\) is symmetric, then the trace is also the sum of its eigenvalues:

$$\text{tr}(A) = \sum_{i=1}^n \lambda_i.$$

$$\text{tr}(A) = \text{tr}(UAU^\top) = \text{tr}(\Sigma U^\top U) = \text{tr}(\Lambda),$$

where we used the cyclic property of traces.
2 Convex Optimization

1. Convex sets: \( x \in V, y \in V \Rightarrow \theta x + (1 - \theta)y \in V \) for any \( \theta \in [0,1] \).

2. Convex function: \( f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta)f(y) \) for any \( x, y \in \text{dom} f \) and \( \theta \in [0,1] \).
   
e.g. Affine function, exponential function, negative entropy \((x \log x)\)

3. Jensen’s inequality: \( f(\mathbb{E}[x]) \leq \mathbb{E}[f(x)] \).
   
   • Exercise: Prove \( D_{\text{KL}}(P||Q) \geq 0 \).
   
   • Exercise: Prove \( \mathbb{E}[\|x\|_2] \leq \sqrt{\mathbb{E}[\|x\|_2^2]} \).

4. Operations that preserve convexity:
   
   (a) Sum: \( f = k_1f_1 + k_2f_2 \) for constants \( k_1, k_2 \geq 0 \).
   
   (b) Affine mapping: \( g(x) = f(Ax + b) \).
   
   (c) Pointwise supremum. [Draw picture]

5. The following functions are convex:
   
   (a) Hinge loss: \( f(w) = \max\{0, 1 - y(w \cdot x)\} \).
   
   (b) L1 norm: \( \|v\|_1 = \sum_i |v_i| \).
   
   (c) Conjugate function of \( f \): \( f^*(y) = \sup_x (y^T x - f(x)) \).
   
   (d) Maximum eigenvalue: \( \lambda_1(A) = \max_{\|v\|_2=1} v^T Av \).

6. Subgradient. For a convex function \( f \) and a point in its domain \( w \), the subdifferential (the elements are the subgradients) of \( f \) at \( w \) is defined as
   \[
   \partial f(w) = \{ z : f(u) \geq f(w) + z \cdot (u - w) \ \forall u \in \text{dom} f \}
   \]
   
   (a) If \( f \) is differentiable at \( w \), then \( \partial f(w) = \{ \nabla f(w) \} \).
   
   (b) Sum: if \( f = k_1f_1 + k_2f_2 \) where \( k_1, k_2 \geq 0 \), then \( \partial f(w) = k_1 \partial f_1(w) + k_2 \partial f_2(w) \).
   
   (c) Affine mapping: \( g(x) = f(Ax + b) \), then \( \partial g(w) = A^T \partial f(Aw + b) \).
   
   (d) Pointwise maximum and supremum: convex hull of the union of the subdifferentials of active functions at the point.

3 Probability

1. Properties of Probability
   
   • if \( A \subseteq B \), \( \Pr(A) \leq \Pr(B) \)
   
   • \( \Pr(A \cap B) \leq \min\{\Pr(A), \Pr(B)\} \)
   
   • (Union Bound) \( \Pr(A \cup B) \leq \Pr(A) + \Pr(B) \)

2. Linearity of Expectation: \( \mathbb{E}[\sum_i X_i] = \sum_i \mathbb{E}[X_i] \) (not necessarily independent)
   
   Linearity of Variance: \( \text{Var}[\sum_i^n X_i] = \sum_i^n \mathbb{E}[X_i^2] \) if \( X_1, \ldots, X_n \) are pair-wise uncorrelated.

- correlation $\Rightarrow$ dependence
- independence $\Rightarrow$ uncorrelated
- uncorrelated $\not\Rightarrow$ independence. (e.g. $X \sim N(0, 1), Y = X^2$)
- uncorrelated $\iff$ independence, if they are Gaussian

4. Conditional Expectation

$$E[Y] = E_X[E[Y|X]] = \int_x E[Y|X = x]dF_X(x)$$

$$E[g(X)Y] = E_X[g(X)E[Y|X]]$$

5. Conditional variance

$$\text{Var}[Y] = E_X[\text{Var}[Y|X]] + \text{Var}_X[\mathbb{E}[Y|X]]$$

6. Convergence. If $\{X_n\}_{i=1}^{\infty}$ are a sequence of random variables,

(a) converge in probability: $X_n \xrightarrow{P} X_\infty$, if

$$\lim_{n \to \infty} \text{Pr}(|X_n - X_\infty| > \epsilon) = 0, \ \forall \epsilon > 0$$

(b) converge almost surely: $X_n \xrightarrow{a.s.} X_\infty$, if

$$\text{Pr}(\{\omega \in \Omega : \lim_{n \to \infty} X_n(\omega) = X_\infty(\omega)\}) = 1.$$

(c) converge in distribution (weak): $X_n \xrightarrow{d} X_\infty$, if

$$\lim_{n \to \infty} F_n(x) = F_\infty(x)$$

(d) converge in order $p$: $X_n \xrightarrow{L^p} X_\infty$, if

$$\lim_{n \to \infty} E[|X_n - X_\infty|^p] = 0.$$

Show the relations between all the convergences.

$$(a) \Rightarrow (c), (b) \Rightarrow (a), (d) \Rightarrow (a)$$