1014 Linear Algebra - Final Notes

Norms

Definition of Norms:

1. Homogeneity: $||\alpha v|| = |\alpha| \times ||v||$ for all $\alpha \in \mathbb{R}^n$ and $v \in V$
2. Positive Definiteness: if $||v|| = 0$ for some $v \in V$ then $v = 0$
3. Triangular Inequality: $||u + v|| \leq ||u|| + ||v||$ for all $u, v \in V$

Inner Products

Definition of Inner Products

1. Symmetry: $\langle u, v\rangle = \langle v, u\rangle$ for all $u, v \in V$
2. Linearity: $\langle u+v, w\rangle = \langle u, w\rangle + \langle v, w\rangle$ and $\langle \alpha v, w\rangle = \alpha\langle v, w\rangle$ for all $v, u, w \in V$ and $\alpha \in \mathbb{R}$
3. Positive Definiteness: $\langle v, v\rangle \geq 0$ with equality if and only if $v = 0$

Proposition: If $\langle \cdot, \cdot \rangle$ is an inner product on V then

$$||v|| = \sqrt{\langle v, v\rangle}$$

is a norm on V. We say that the norm $|| \cdot ||$ is induced by the inner product $\langle \cdot, \cdot \rangle$

L1 Norm

The Euclidean Norm is the sum of the absolute values i^{th} entry in the input vector.

$$||x||_1 = \sum_{i=1}^{n} |x_i|$$

Euclidean Norm:

The Euclidean Norm is the square root of the sum of the products of each i^{th} entry in the input vectors.

$$||x||_2 = \sqrt{\langle x, x\rangle} = \sqrt{x_1^2 + \cdots + x_n^2}$$

Cauchy-Schwartz Inequality Let $|| \cdot ||$ be the norm induced by the inner product $\langle \cdot, \cdot \rangle$ on the vector space V. Then for all $x, y \in V$:

$$|\langle x, y\rangle| \leq ||x|| \times ||y||$$

Moreover, there is equality if and only if x and y are linearly dependent (i.e. $x = \alpha y$ or $y = \alpha x$ for some $\alpha \in \mathbb{R}$)

Orthogonality

Definition of Orthogonality: Let V be a vector space and $\langle \cdot, \cdot \rangle$ be an inner product on V

- We say that vectors x, y are orthogonal if $\langle x, y \rangle = 0$. We write $x \perp y$
- We say that vector x is orthogonal to the set of vectors A if x is orthogonal to all of the vectors in A. We write $x \perp A$

Orthonormal Family of Vectors:
For a Orthonormal Family of Vectors $\{v_1, \ldots, v_n\}$

- The family is orthogonal if $\langle v_i, v_j \rangle = 0$ for all $i \neq j$
- The family is orthonormal if all the vectors are orthogonal and all of the v_i have unit norm $||v_1|| = \cdots = ||v_k|| = 1$
- Orthonormal basis preserve distance and angles: $\langle Ax, Ay \rangle = x^T A^T A y = x^T y = \langle x, y \rangle$
- A vector space of finite dimension admits an orthonormal basis
- Assume that $dim(V) = n$ and let $v_1, \ldots, v_n$ be an orthonormal basis of V. Then the coordinates of a vector $x \in V$ in the basis $v_1, \ldots, v_n$ are

$$x = \langle x, v_1 \rangle v_1 + \cdots + \langle x, v_n \rangle v_n$$

Pythagorean Theorem:
Let $|| \cdot ||$ be the norm induced by $\langle \cdot, \cdot \rangle$ for all $x, y \in V$ we have:

$$x \perp y \longleftrightarrow ||x + y||^2 = ||x||^2 + ||y||^2$$

Proof of Pythagorean Theorem:

$$\begin{aligned} ||x + y||^2 &= \langle x + y, x + y \rangle \\ &= \langle x, x \rangle + \langle y, y \rangle + 2\langle x, y \rangle \\ &= ||x||^2 + ||y||^2 \text{ as } x \perp y \text{ so } \langle x, y \rangle = 0 \quad \square \end{aligned} \tag{1}$$

Orthogonal Projection Let S be a subspace of R^n The orthogonal projection of a vector x onto S is defined as the vector $P_S(x)$ in S that minimizes the distance to x:

$$P_S(x) = argmin||x - y|| \quad for\ y \in S$$

The distance from x to the subspace S is defined by:

$$d(x, S) = min||x - y|| = ||x - P_S(x)||$$

Projection Matrix: $VV^T x$ **Projection Formula:** $P_S(x) = \langle v_i, x \rangle v_i + \cdots + \langle v_n, x \rangle v_n$

Eigenvectors, Square Symmetric Matrices, and SVD

Properties of Eigenvectors:

1. If v_i is an eigenvector with associated eigenvalue λ_i, then: $Av_i = \lambda_i v_i$

2. If you scale a matrix A by some constant $\alpha \in \mathbb{R}$, its eigenvalues get shifted by α as: $\alpha A v_1 = \alpha \lambda_i v_i$ where v_i is an eigenvalue of A.

3. Likewise, if you add to the diagonals of a matrix A using the identity to a matrix, the eigenvalues of A get scaled by α: $(A + \alpha \times Id_n)v_i = \alpha v_i + Av_i = \alpha v_i + \lambda_i v_i = v_i(\alpha + \lambda_i)$

4. If you have two eigenvectors, that share the same eigenspace and eigenvalue, any linear combination is also a eigenvector in that eigenspace with that eigenvalue:

   $Av_1 = \lambda_1 v_1$ and $Av_2 = \lambda_1 v_2$ then $A(v_1+v_2) = A(v_1)+A(v_2) = \lambda_1 v_1+\lambda_1 v_2 = \lambda_1(v_1+v_2)$ $\square$

   Note: that this does not hold if the two eigenvectors do not share the same eigenspace (and therefore, eigenvalue).

5. If you are given a eigenvalue, λ to solve for the eigenspace and thus eigenvector(s) associated with the eigenvalue, you do: $Ker(A - (\lambda \times Id_n))x = 0$ and solve for x via row reduction.

Square Symmetric Matrices:

1. Any square, symmetric matrix S Can be expressed as $S = PDP^T$, where P is the orthonormal eigenvector basis of S. If S is PSD then its eigenvalue are PSD

2. This be interpreted geometrically as a rotation, then a scaling, then a rotation back to canonical coordinates

3. If there is a kernel, the matrix is not invertible, and at least one eigenvector will be associated to the eigenvalue equal to 0

4. The sum of its eigenvalues is equal to its trace $Tr(A) = Tr(PDP^T) = Tr(P^TPD) = Tr(D)$

PCA When doing PCA we pick the largest eigenvalue, as it maximizes the variance. If you take a vector, x where $x = \alpha_1 v_1 + \cdots + \alpha_n v_n$ and $||x|| = 1$, where $v_1, \ldots, v_n$ are eigenvalues of some orthonormal eigenbasis. We then maximize $x^T Ax$
Trace $Tr(AB) = Tr(CB)$ then $Tr(A) = \sum_{i=1}^{n} \lambda_i$ where $A \in \mathbb{R}^{n \times n}$ and A is square symmetric.

SVD: Singular Value Decomposition:

Any matrix $A \in \mathbb{R}^{n \times m}$ (any matrix though even if the dimensions match) can be written in its SVD form as: $A = U\Sigma V^T$ where $\Sigma \in \mathbb{R}^{n \times m}$ is a diagonal matrix with the singular values in the diagonal and 0's everywhere else, $V = Eig(A^TA),\ and\ V \in \mathbb{R}^{m \times m}$ and $U = Eig(AA^T),\ and\ U \in$

$\mathbb{R}^{n\times n}$, where $Eig()$ indicates an orthonormal basis of eigenvectors belonging to some matrix. Some useful identities:

$$A = U\Sigma V^T \qquad A^TA = V\Sigma^T\Sigma V^T = V\Sigma^2V^T \qquad AA^T = U\Sigma V^TV\Sigma U^T = U\Sigma^2U^T$$

The Moore-Penrose psuedo-inverse is defined as:

$$A^\dagger = V\Sigma' U^T \qquad \text{where } \Sigma' \text{ is a diag matrix of}$$

Convexity

Convex Sets:
A set, M is convex if its elements satisfy the following property: that

$$\alpha x + (1-\alpha)y \in M \text{ for } \alpha \in [0,1] \text{ and } x, y \in M$$

Convex Functions:
A function is said to be convex if any line drawn from two points in the function rests above the actual value of the function at that point. Expressed mathematically by:

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y) \text{ with } \alpha \in [0,1]$$

A function is strictly convex if there is strict inequality, which means there is only one minimum point of the function.
Properties of Convexity:

1. A function f is called concave if $-f$ is convex
2. Any linear map is convex and concave: $f(\alpha x + (1-\alpha)y) = \alpha f(x) + (1-\alpha)f(y)$ with $\alpha \in [0,1]$. Also, any norm is convex and the gradient is orthogonal to contour lines.
3. The sum of two convex functions is also a convex function. **If there is a scalar α multiplying one of the two functions, then this does not hold**
4. The graph of a convex function will always be above (or equal) to its tangent

$$f(y) \geq f(x) + \langle \nabla f(x), (y-x) \rangle$$

5. The Hessian is Square Symmetric, and you can tell if the Hessian, and thus function, is PSD/PD/ND/NSD by looking at the eigenvalues.
6. f is only convex if and only if for all $x \in \mathbb{R}^n$ $H_f(x)$ is PSD . If the Hessian is PD then its strictly convex. Likewise if the Hessian is NSD its concave, and ND then strictly concave. Note that if there are variables in the Hessian, you know its not a convex function, though it could be convex in certain ranges.
7. A^TA and AA^T are always PSD, and if A is full rank, then A^TA is PD. Therefore, it's Hessian will be strongly/strictly convex.

Taylor's Formulas:

$$Order\ 1 \to f(x+h) = f(x) + hf'(x) \qquad Order\ 2 \to f(x+h) = f(x) + h\nabla f(x) + \frac{h^2}{2}f''(x)$$

Linear Algebra Form:

$$Order\ 1 \to f(x+h) = f(x) + \langle h, \nabla f(x)\rangle \qquad Order\ 2 \to f(x+h) = f(x) + \langle h, \nabla f(x)\rangle + \frac{1}{2}h^T H_f(x) h$$

Regression

Ordinary Least Squares

1. We want to solve the problem $Ax = b$, but unfortunately, unless our matrix is square and full rank there won't be a solution. A tall matrix risks not having y in its image, and a fat matrix will have a null space, thus there will be infinitely many solutions.

2. We define the problem as Minimize $f(x) = ||Ax - y||^2 = \langle Ax - y, Ax - y\rangle$ with respect to $x \in \mathbb{R}^d$. If the matrix has no kernel, the set of solutions is given by: $x = (A^TA)^{-1}A^Ty$

3. If A^TA is not invertible, which means that A is not full rank and admits a kernel, as $Ker(A^TA) = Ker(A)$, then we will have to use the MP psuedo-inverse, defined as $A^\dagger = V\Sigma' U^T$, then $x^{LS} = A^\dagger y$ and it is a solution to $A^TAx = A^Ty$

$$\begin{aligned} A^Ty &= A^\dagger y \\ A^TAx &= V\Sigma' U^T y \\ x &= (A^TA)^{-1}(V\Sigma' U^T y) \\ x &= (V\Sigma^T U^T U\Sigma V^T)(V\Sigma' U^T y) \\ x &= V\Sigma U^T y = A^T y \quad \square \end{aligned} \tag{2}$$

4. Again, but more detail identities:

$$\begin{aligned} (A^TA)^{-1}A^Ty =& (V\Sigma^T U^T U\Sigma V^T)^{-1}A^Ty \\ =& V\Sigma^{-2}V^T(A^Ty) \\ =& V\Sigma^{-2}V^T(V\Sigma^T U^T y) = V\Sigma' U^T y \\ =& A^\dagger y \end{aligned} \tag{3}$$

5. If there is a kernel, the set of all minimizers is given by $\{x^{LS} + v | v \in Ker(A)\}$

Ridge Regression

1. We define the problem as Minimize $f(x) = ||Ax - y||^2 + \lambda||x||_2^2$ w.r.t. $x \in \mathbb{R}^d$

$$x^{ridge} = (A^TA + \lambda Id_n)A^Ty$$

2. Implies $f(x)$ is a strongly convex function, and therefore strictly convex.

LASSO Regression

1. We define Lasso regression as adding the L1 norm as a penalty to the least squares problem: $f(x) = ||Ax - y||^2 + \lambda||x||_1$

Matrix Norms

NOTE: NORMS (BOTH MATRIX AND VECTOR) ARE NOT LINEAR MAPS!
i.e. $||-x||_2 \neq -||x||_2$
Frobenius Norm: a norm on a matrix $A \in \mathbb{R}^{n \times m}$ is defined as

$$||A||_F = \sqrt{\sum_{i=n}^{n} \sum_{j=1}^{m} A_{i,j}^2} = Tr(A^T A) = \sqrt{\sum_{i=1}^{min(n,m)} \sigma_i(A)^2}$$

Proof: $||A||_F^2 = Tr(AA^T) = Tr(U\Sigma V^T V \Sigma^T U) = Tr(\Sigma\Sigma^T) = \sum_{i=1}^{min(n,m)} \sigma_i(A)^2$
Spectral Norm: $||A||_{sp} = \arg max_{||x||=1} ||Ax|| = \sigma_1(A)$
Nuclear Norm: L1 Norm for matrices $||A||_* = \sum_{i=1}^{min(n,m)} \sigma_i(A)$

Optimality Conditions:

Isotropic $->$ eig vals are in really close ratio
Anistropic they are not

Constrained Optimization

Lagrangian Functions

1. KKT Method: Assume you have a function that is convex (as it is the sum of convex functions) and there exists some point x_0 such that $f_i(x) < 0$ for all i. Then x is a solution if and only if x is feasible and there exists some $\lambda_1, \ldots, \lambda_n \geq 0$ such that:

$$\mathcal{L}(x, \lambda_1, \lambda_2, \ldots) = f(x) + \lambda_1 g(x) + \lambda_2 k(x) + \ldots$$

$$\nabla\mathcal{L}(x, \lambda_1, \lambda_2, \ldots) = \nabla f(x) + \lambda_1 \nabla g(x) + \lambda_2 \nabla k(x) + \ldots$$

Where $f(x)$ is the function you want to minimize and $g(x), k(x), \ldots$ are the constraints

Gradient Descent with Constant Step Size

Goal: minimize a differentiable function f: $\mathbb{R}^n \to \mathbb{R}$. Starting from a point $x_0 \in \mathbb{R}^n$, perform the updates:

$$x_{t+1} = x_t - \alpha_t \nabla f(x_t)$$

For Gradient descent with constant step size, the optimal α is $\frac{u}{L}$
Issues with gradient descent:

1. If the norm $||\nabla f(x)||$ is too small, the step size will be too small

2. The vector $-\nabla f(x)$ does not really point towards the optimal solution. In other words, sometimes the gradient oscillates in big steps around the minimum, but never converges.

Smoothness and Strong Convexity

1. Given $L, \mu > 0$, we say that a twice-differentiable convex function, $f : \mathbb{R}^n \to \mathbb{R}$ is:

   L-Smooth if for all $x \in \mathbb{R}^n$, $\lambda_{max}(H_f(x)) \leq L$

   μ-strongly convex if for all $x \in \mathbb{R}^n$, $\lambda_{min}(H_f(x)) \geq \mu$

2. A L-Smooth μ-strongly convex function verifies the following inequalities:

$$f(x_t) - f(x^*) \leq (1 - \frac{\mu}{L})^t (f(x_0) - f(x^*))$$

   where x^* is the optimal solution, and $||x_0 - x^*||$ is the distance from the start position to the optimal solution

3. Definition: Condition Number, K is equal to $\frac{L}{\mu}$, the speed of convergence decreases if K increases. Also, $K \geq 1$

Backtracking Line Search

1. Start with $\alpha = 1$ and while

$$f(x_t - \alpha \nabla f(x_t)) \geq f(x_t) - \frac{\alpha}{2} ||\nabla f(x_t)||^2$$

   update α by making it smaller, say on the second iteration you set $\alpha = .8\alpha$

Gradient Descent With Momentum

1. Gradient Descent With Momentum is:

$$x_{t+1} = x_t + v_t \text{ where } v_t = -\alpha \nabla f(x_t) + \beta_t v_{t-1}$$

   where $-\alpha \nabla f(x_t)$ is the normal gradient, and $\beta_t v_{t-1}$ is the momentum (some portion of the last step).

2. Pros: Dampens oscillations, promotes direction towards the minimum

3. Error Bound Defined As follows, for some constant, C, that does not depend on t:

$$f(x_t) - f(x^*) \leq C \left(\frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}} \right)^t$$

Newtown's Method, AKA The Virgin Gradient Descent

$$x_{t+1} = x_t - H_f(x)^{-1} \nabla f(x_t)$$

1. Idea: optimizing the learning rate by considering the second order Taylor expansion.

2. Pros: Extremely fast, there exists $C, p > 0$ s.t. $||x_t - x^*||^2 \leq Ce^{-2^t}$

3. Cons: Computationally expensive, if the Hessian is $H_f(x_t) \in \mathbb{R}^{n \times n}$ then it takes n^3 operations to compute its inverse. Also, in non-convex settings, Newton's method gets attracted to any critical point (mins, maxs, saddle points)

Quasi-Newton Methods: try to approximate $H_f(x_t)^{-1}$ by matrices B_t that are easier to compute.

Graphs

Graphs

1. Adjacency Matrix A of Graph G is the square matrix where $A_{ij} = 1$ if nodes i, j connect, else 0.
2. The degree matrix of G is the square diagonal matrix where $D_{i,i} =$ number of neighbors of i
3. Laplacian Matrix of G is defined as:
$$L = D - A$$
then for all $x \in \mathbb{R}^n : x^T L x = \sum_{\text{i connected to j}} (x_i - x_j)^2$ Which is PSD
4. The multiplicity of $eig(L) = 0$ is the number of connected components of graph G. Graph G is connected (all nodes are connected) if the second smallest eig val is greater than 0
5. The adjacency matrix, $A^1_{i,j} = A_{i,j}$ is the number of ways we can go from i to j (or vica versa if we look at $A_{j,i}$ as the adjacency matrix is symmetric) in 1 step.

$$A^{k+1}_{ij} = \sum_{l=1}^{n} A^k_{il} * A_{lj}$$

Proof: For any vector $y \in S$, $\langle x, y\rangle = \langle P_S(x), y\rangle$:

Initialize an orthonormal basis for S as $\{s_1, \ldots, s_d\}$ where d is the dimension of S. The vector y and $P_S(x)$ can each be defined in this basis:

$$y = \langle y, s_1\rangle s_1 + \cdots + \langle y, s_d\rangle s_d \text{ and } P_S(x) = \langle x, s_1\rangle s_1 + \cdots + \langle x, s_d\rangle s_d$$

And let $\alpha_i = \langle y, s_i\rangle$ and $\beta_i = \langle x, s_i\rangle$ then:

$$\langle x, y\rangle = \langle x, \alpha_1 s_1 + \cdots + \alpha_d s_d\rangle \text{ as defined}$$

Split the dot product into many dot products, pull out the scalars α_i

$$\langle x, y\rangle = \sum_{i=1}^{d} \alpha_i \langle x, v_i\rangle = \sum_{i=1}^{d} \langle y, s_i\rangle \langle x, v_i\rangle$$

Follow the same procedure for the right hand side. Note how we will have to do a double sum when foiling the dot product as there are many terms on each side (think a double for loop for intuition):

$$\begin{aligned}
\langle P_S(x), y\rangle &= \langle \langle x, s_1\rangle s_1 + \cdots + \langle x, s_d\rangle s_d, \langle y, s_1\rangle s_1 + \cdots + \langle y, s_d\rangle s_d\rangle \\
&= \langle \beta_1 s_1 + \cdots + \beta_d s_d, \alpha_1 s_1 + \cdots + \alpha_d s_d\rangle \\
&= \sum_{i=1}^{d}\sum_{j=1}^{d} \beta_i s_i \alpha_j s_j \begin{cases} \langle s_i, s_j\rangle = 1 & \text{when } s_i = s_j \\ \langle s_i, s_j\rangle = 0 & \text{when } s_i \neq s_j \end{cases} \\
&= \sum_{i=1}^{d} \beta_i \alpha_i = \sum_{i=1}^{d} \langle y, s_i\rangle \langle x, v_i\rangle
\end{aligned} \tag{4}$$

Therefore:

$$\langle x, y\rangle = \langle P_S(x), y\rangle \rightarrow \sum_{i=1}^{d} \langle y, s_i\rangle\langle x, v_i\rangle = \sum_{i=1}^{d} \langle y, s_i\rangle\langle x, v_i\rangle \quad \square$$

Proof: $x - P_S(x) \perp S$

We can easily show this using what we established in the proof above. Remember that S is defined as $\{s_1, \ldots, s_d\}$, and $P_S(x) = \langle x, s_1\rangle s_1 + \cdots + \langle x, s_d\rangle s_d$. If $x - P_S(x) \perp S$ then $\langle x - P_S(x), s_1 + \cdots + s_d\rangle = 0$. And let $\beta_i = \langle x, s_i\rangle$. We can prove the statement by splitting the dot product into separate dot products, $\langle x, S\rangle$ and $\langle P_S(x), S$

$$\langle x - P_S(x), s_1 + \cdots + s_d\rangle = \langle x, s_1 + \cdots + s_d\rangle - \langle P_S(x), s_1 + \cdots + s_d\rangle$$

LHS can be expressed as follows:

$$\langle x, s_1 + \cdots + s_d\rangle = \sum_{i=1}^{d} \langle x, s_i\rangle$$

RHS can be solved using the same technique as the proof above:

$$\begin{aligned}
\langle P_S(x), s_1 + \cdots + s_d\rangle &= \langle x, s_1\rangle s_1 + \cdots + \langle x, s_d\rangle s_d, s_1 + \cdots + s_d\rangle \\
&= \langle \beta_1 s_1 + \cdots + \beta_d s_d, s_1 + \cdots + s_d\rangle \\
&= \sum_{i=1}^{d}\sum_{j=1}^{d} \beta_i s_i s_j \begin{cases} \langle s_i, s_j\rangle = 1 & \text{when } s_i = s_j \\ \langle s_i, s_j\rangle = 0 & \text{when } s_i \neq s_j \end{cases} \\
&= \sum_{j=1}^{d} \beta_i = \sum_{i=1}^{d} \langle x, s_i\rangle
\end{aligned} \tag{5}$$

Then we have:

$$\langle x - P_S(x), S\rangle = \langle x, S\rangle - \langle P_S(x), S\rangle = \sum_{i=1}^{d} \langle x, s_i\rangle - \sum_{i=1}^{d} \langle x, s_i\rangle = 0 \quad \square$$

Proof:$||P_S(x) \leq ||x||$

Using what we showed above, we know that $x - P_S(x) \perp S$ and since $P_S(x) \in S$. Using the Pythagorean Theorem:

$$\begin{aligned}
||x - P_S(x) + P_S(x)||^2 &= ||x - P_S(x)||^2 + ||P_S(x)||^2 + 2\langle x - P_S(x), P_S(x)\rangle \\
||x||^2 &= ||x - P_S(x)||^2 + ||P_S(x)||^2 + 2\langle x - P_S(x), P_S(x)\rangle \\
||x||^2 &= ||x - P_S(x)||^2 + ||P_S(x)||^2 \quad \square
\end{aligned} \tag{6}$$

As $P_S(x) \in S$ and $x - P_S(x) \perp S$ then $2\langle x - P_S(x), P_S(x)\rangle = 0$. Since you norms are always positive, and you have to add two positive numbers on the RHS to equal to LHS, its clear to see that $||x|| \geq ||P_S(x)||$.

Proof: Square Symmetric Matrices are PSD if their Eigenvalues Are PSD:

Let $A \in \mathbb{R}^n$ be a square symmetric matrix, and $A = PDP^T$. Lets express a vector x as a linear combination of the orthonormal vectors of P: $x = \alpha_1 v_1 + \cdots + \alpha_n v_n$, with associated eigenvalues $\lambda_1, \ldots, \lambda_n$ and $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$. We test positive definiteness with $x^T Ax \geq 0$:

$$\begin{aligned} x^T Ax &= \langle x, Ax \rangle \\ &= \langle \alpha_1 v_1 + \cdots + \alpha_n v_n), A(\alpha_1 v_1 + \cdots + \alpha_n v_n) \rangle \\ &= \langle \alpha_1 v_1 + \cdots + \alpha_n v_n), \lambda_1 \alpha_1 v_1 + \cdots + \lambda_n \alpha_n v_n \rangle \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \lambda_j v_i v_j \begin{cases} \langle v_i, v_j \rangle = 1 & \text{when } v_i = v_j \\ \langle v_i, v_j \rangle = 0 & \text{when } v_i \neq v_j \end{cases} \\ &= \sum_{i=1}^{n} \alpha_i^2 \lambda_i \quad \square \end{aligned} \tag{7}$$

Therefore, the only way for this equation to be negative is if some eigenvector(s) are negative.

GD Proof 1

$$\begin{aligned} &x_{t+1} - x^* = \big(Id - \beta M\big)(x_t - x^*) \text{ definition} \\ &x_{t+1} - x^* = Id_n x_t - Id_n x^* - \beta M x_t + \beta M x^* \text{ foil function} \\ &x_{t+1} - x^* = x_t - x^* + \beta b - \beta \nabla f(x_t) - \beta b + \beta \nabla f(x^*) \text{ gradient plus b equals Mx} \\ &x_{t+1} - x^* = x_t - \beta \nabla f(x_t) - x^* \text{ definition of standard gradient descent} \\ &x_{t+1} - x^* = x_{t+1} - x^* \quad \square \end{aligned} \tag{8}$$

GD Proof 2

$$\begin{aligned} &||x_t - x^*|| = ||\big(Id - \beta M\big)^t (x_0 - x^*)|| \leq ||(Id - \beta M)^t||_{sp} ||x_0 - x^*|| \\ &||x_t - x^*|| \leq ||(Id - \frac{1}{L} M)^t||_{sp} ||x_0 - x^*|| \\ &||x_t - x^*|| \leq (1 - \frac{\mu}{L})^t ||x_0 - x^*|| \quad \square \end{aligned} \tag{9}$$

GD Proof 3

$$\alpha_1(t) v_1 \cdots + \alpha_d(t) v_d = a_1(0)(Id - \frac{\lambda_1}{L})^t v_1 + \cdots + a_d(0)(Id - \frac{\lambda_d}{L})^t v_d$$

$$w_t = \begin{pmatrix} \alpha_1(t) \\ \vdots \\ \alpha_d(t) \end{pmatrix} = \begin{pmatrix} \alpha_1(0)(Id - \frac{\lambda_1}{L})^t \\ \vdots \\ \alpha_d(0)(Id - \frac{\lambda_d}{L})^t \end{pmatrix} \qquad \text{and} \qquad \alpha_i(t) = \alpha_i(0)(Id - \frac{\lambda_i}{L})^t \quad \square$$

Misc.

Convex Sets Misc

1. $S = \{x \in R^n : ||Ax|| = 0\}$ - Yes this is a convex set - use kernel

2. $S = \{x \in R^n : ||Ax|| \leq 1\}$ - Use definition of convex sets, triangle inequality of norms

3. $S = \{x \in R^n : ||Ax|| = 0\}$ - Does not hold, use $n = 2$ identity, and e_1, e_2, set $alpha = .5$ and use convex set definition to show the norms that results is less than 1

4. $S = \{x \in R^2 n : \sum_{k=1}^{n} x_k^2 \leq \sum_{k=n+1}^{2n} x_k^2$ - not a convex set, have one vector that has negative entries and you can find a counter example

If we have some matrix A and a matrix of B that is just the columns of A reordered (permuted) then:

1. $Im(A) = Im(B)$

2. $Im(A^T) \neq Im(B^T)$

3. $Ker(A) \neq Ker(B)$

4. $Ker(A^T) = Ker(B^T)$

5. $||A||_F = ||B_F||$

6. $Rank(A) = Rank(B)$

Derivatives and Matrix/Vector Differentiation

1 Matrix/vector manipulation

You should be comfortable with these rules. They will come in handy when you want to simplify a expression before differentiating. All bold capitals are matrices, bold lowercase are vectors.

Rule	Comments
$(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$	order is reversed, everything is transposed
$(\mathbf{a}^T\mathbf{Bc})^T = \mathbf{c}^T\mathbf{B}^T\mathbf{a}$	as above
$\mathbf{a}^T\mathbf{b} = \mathbf{b}^T\mathbf{a}$	(the result is a scalar, and the transpose of a scalar is itself)
$(\mathbf{A}+\mathbf{B})\mathbf{C} = \mathbf{AC}+\mathbf{BC}$	multiplication is distributive
$(\mathbf{a}+\mathbf{b})^T\mathbf{C} = \mathbf{a}^T\mathbf{C}+\mathbf{b}^T\mathbf{C}$	as above, with vectors
$\mathbf{AB} \neq \mathbf{BA}$	multiplication is **not** commutative

2 Common vector derivatives

You should know these by heart. They are presented alongside similar-looking scalar derivatives to he memory. This doesn't mean matrix derivatives always look just like scalar ones. In these examples, b a constant scalar, and $\mathbf{B}$ is a constant matrix.

Scalar derivative			Vector derivative		
$f(x)$	$\rightarrow$	$\frac{\mathrm{d}f}{\mathrm{d}x}$	$f(\mathbf{x})$	$\rightarrow$	$\frac{\mathrm{d}f}{\mathrm{d}\mathbf{x}}$
bx	$\rightarrow$	b	$\mathbf{x}^T\mathbf{B}$	$\rightarrow$	$\mathbf{B}$
bx	$\rightarrow$	b	$\mathbf{x}^T\mathbf{b}$	$\rightarrow$	$\mathbf{b}$
x^2	$\rightarrow$	$2x$	$\mathbf{x}^T\mathbf{x}$	$\rightarrow$	$2\mathbf{x}$
bx^2	$\rightarrow$	$2bx$	$\mathbf{x}^T\mathbf{Bx}$	$\rightarrow$	$2\mathbf{Bx}$

Derivative	Integral (Antiderivative)
$\frac{d}{dx} n = 0$	$\int 0 \, dx = C$
$\frac{d}{dx} x = 1$	$\int 1 \, dx = x + C$
$\frac{d}{dx} x^n = nx^{n-1}$	$\int x^n \, dx = \frac{x^{n+1}}{n+1} + C$
$\frac{d}{dx} e^x = e^x$	$\int e^x \, dx = e^x + C$
$\frac{d}{dx} \ln x = \frac{1}{x}$	$\int \frac{1}{x} \, dx = \ln x + C$
$\frac{d}{dx} n^x = n^x \ln x$	$\int n^x \, dx = \frac{n^x}{\ln n} + C$
$\frac{d}{dx} \sin x = \cos x$	$\int \cos x \, dx = \sin x + C$
$\frac{d}{dx} \cos x = -\sin x$	$\int \sin x \, dx = -\cos x + C$
$\frac{d}{dx} \tan x = \sec^2 x$	$\int \sec^2 x \, dx = \tan x + C$
$\frac{d}{dx} \cot x = -\csc^2 x$	$\int \csc^2 x \, dx = -\cot x + C$
$\frac{d}{dx} \sec x = \sec x \tan x$	$\int \tan x \sec x \, dx = \sec x + C$
$\frac{d}{dx} \csc x = -\csc x \cot x$	$\int \cot x \csc x \, dx = -\csc x + C$
$\frac{d}{dx} \arcsin x = \frac{1}{\sqrt{1-x^2}}$	$\int \frac{1}{\sqrt{1-x^2}} \, dx = \arcsin x + C$
$\frac{d}{dx} \arccos x = -\frac{1}{\sqrt{1-x^2}}$	$\int -\frac{1}{\sqrt{1-x^2}} \, dx = \arccos x + C$
$\frac{d}{dx} \arctan x = \frac{1}{1+x^2}$	$\int \frac{1}{1+x^2} \, dx = \arctan x + C$
$\frac{d}{dx} \operatorname{arc\,cot} x = -\frac{1}{1+x^2}$	$\int -\frac{1}{1+x^2} \, dx = \operatorname{arc\,cot} x + C$
$\frac{d}{dx} \operatorname{arc\,sec} x = \frac{1}{x\sqrt{x^2-1}}$	$\int \frac{1}{x\sqrt{x^2-1}} \, dx = \operatorname{arc\,sec} x + C$
$\frac{d}{dx} \operatorname{arc\,csc} x = -\frac{1}{x\sqrt{x^2-1}}$	$\int -\frac{1}{x\sqrt{x^2-1}} \, dx = \operatorname{arc\,csc} x + C$

Definitions: Singular values/vectors

$A \in \mathbb{R}^{n \times m}$

For $i = 1, \ldots, m$:

- We define $\sigma_i = \sqrt{\lambda_i}$, called the i^{th} **singular value** of A.

Let $r = \operatorname{rank}(A) =$ number of non-zero λ_i's (exercise!). (λ_i: eigenvalues of A^TA.)

For $i = 1, \ldots, r$:

- We call $u_i = \frac{1}{\sigma_i} A v_i$ the i^{th} **left singular vector** of A. (v_i: eigenvectors of A^TA)
- $u_1, \cdots u_r$ are orhonormal.
- If $r < n$, we add $u_{r+1}, \cdots u_n$ such that $u_1, \cdots u_n$ is an orthonormal basis of $\mathbb{R}^n$.

For $i = 1, \ldots, m$:

- Observe that we have $Av_i = \sigma_i u_i$ and $A^T u_i = \sigma_i v_i$.
- We call v_j the i^{th} **right singular vector** of A.

2.2 Inequalities

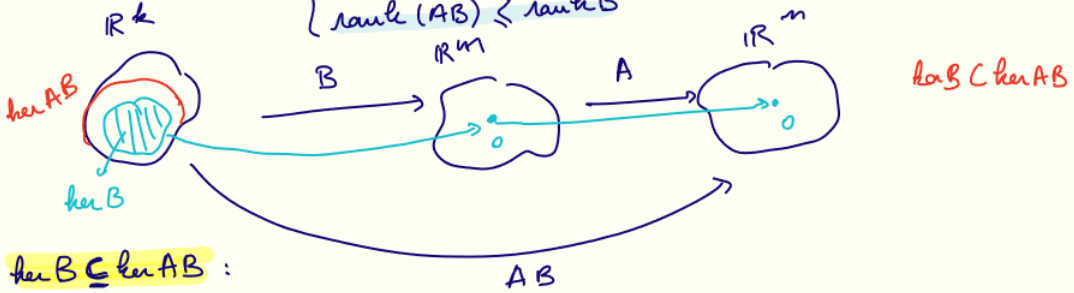
Proposition

$C = \subseteq$

Let $A \in \mathbb{R}^{n \times m}$ and $B \in \mathbb{R}^{m \times k}$. Then the following holds

1. $\operatorname{rank}(A) \leq \min(n, m)$.
2. $\operatorname{rank}(AB) \leq \min(\operatorname{rank}(A), \operatorname{rank}(B))$.

Proof. Show that $\begin{cases} \operatorname{rank}(AB) \leq \operatorname{rank} A \\ \operatorname{rank}(AB) \leq \operatorname{rank} B \end{cases}$ → HW Q°



- $\ker B \subseteq \ker AB$:

for $x \in \ker B$, $Bx = 0 \Rightarrow ABx = 0 \Rightarrow x \in \ker AB$.

- $\dim \ker AB + \operatorname{rank} AB = \dim \ker B + \operatorname{rank} B = k$ (By the rank nullity theorem)

$\dim \ker B \leq \dim \ker AB$

$k - \operatorname{rank} B \leq k - \operatorname{rank} AB$

→ $\operatorname{rank} B \geq \operatorname{rank}(AB)$ □

Problem 10.3 (2 points). *Recall that $\|M\|_{\mathrm{Sp}}$ denotes the spectral norm of a matrix M.*

(a) *Let $A \in \mathbb{R}^{n\times m}$. Show that for all $x \in \mathbb{R}^m$,*

$$\|Ax\| \le \|A\|_{\mathrm{Sp}}\|x\|.$$

(b) *Show that for all $A \in \mathbb{R}^{n\times m}$ and $B \in \mathbb{R}^{m\times k}$:*

$$\|AB\|_{\mathrm{Sp}} \le \|A\|_{\mathrm{Sp}}\|B\|_{\mathrm{Sp}}.$$

PROBLEM 10.4

$\|A\|_{sp} = \max_{\|x\|=1} \|Ax\|$

(a) if $x \neq 0$

$\|Ax\| = \|A \underbrace{\frac{x}{\|x\|}}_{\text{norm 1 vector}} \|\|x\|$ and $\|A\frac{x}{\|x\|}\| \le \|A\|_{sp}$ by def.

$\Rightarrow \|Ax\| \le \|A\|_{sp}\|x\|$.

if $x=0 \rightarrow \|Ax\|=0$ so ok.

(b) $\|AB\|_{sp} = \max_{\|x\|=1} \|ABx\|$

$\|ABx\| \le \|A\|_{sp}\|Bx\|$ for any x

$x^* = \arg\max_{\|x\|=1} \|ABx\|$

$\|ABx^*\| \le \|A\|_{sp}\|Bx^*\|$

$\|ABx^*\| = \|AB\|_{sp}$ $\quad$ $\|Bx^*\| \le \|B\|_{sp}$ ← since this is the max

$\Rightarrow \|AB\|_{sp} \le \|A\|_{sp}\|B\|_{sp}$.

Problem 6. Let $f : \mathbb{R}^2 \to \mathbb{R}$ defined by $f(x) = x_1^2 + 4x_2^2 - 4x_1 - 8x_2 + 1$, for $x = (x_1, x_2) \in \mathbb{R}^2$.

(a) The Hessian matrix of f is given by

$$H_f(x) = \begin{pmatrix} 2 & 0 \\ 0 & 8 \end{pmatrix},$$

for all $x \in \mathbb{R}^2$. Hence the eigenvalues of $H_f(x)$ are 2 and 8 which are non-negative: $H_f(x)$ is positive semi-definite: f is therefore convex.

$$\begin{aligned} x \text{ is a global minimizer of } f &\iff \nabla f(x) = 0 \\ &\iff \begin{pmatrix} 2x_1 - 4 \\ 8x_2 - 8 \end{pmatrix} = 0 \\ &\iff x = (2, 1). \end{aligned}$$

f has therefore one unique minimizer $x^\star = (2, 1)$.

(b)

$$\begin{aligned} w(t+1) &= x(t+1) - x^\star \\ &= x(t) - \alpha \nabla f(x(t)) - x^\star \\ &= w(t) - \alpha \begin{pmatrix} 2x_1(t) - 4 \\ 8x_2(t) - 8 \end{pmatrix}. \end{aligned}$$

Hence

$$\begin{cases} w_1(t+1) = w_1(t) - 2\alpha(x_1(t) - 2) = w_1(t) - 2\alpha w_1(t) = (1 - 2\alpha) w_1(t) \\ w_2(t+1) = w_2(t) - 8\alpha(x_2(t) - 1) = w_2(t) - 8\alpha w_2(t) = (1 - 8\alpha) w_2(t). \end{cases}$$

(c) From the previous question we deduce that

$$w_1(t) = (1 - 2\alpha)^t w_1(0) = (1 - 2\alpha)^t (-2) \quad \text{and} \quad w_2(t) = (1 - 8\alpha)^t w_2(0) = (1 - 8\alpha)^t (-1).$$

- if $0 < \alpha < 1/4$, then $(1 - 2\alpha) \in (0, 1)$ and $(1 - 8\alpha) \in (0, 1)$. Hence $w_1(t)$ and $w_2(t)$ go to zero as $t \to \infty$ which gives that $x(t)$ converge to $x^\star$.

- if $\alpha \geq \frac{1}{4}$, then $1 - 8\alpha \leq -1$ and therefore $w_2(t) = -(1 - 8\alpha)^t$ does not go to zero as $t \to \infty$: $w(t)$ does not go to zero, hence gradient descent does not converge to $x^\star$.