1 Linear Algebra Review

1.1 Linear Systems

Condition on $rank(A)$ for existence of exact solution: System: $Ax = b$. b a weighted sum of the columns of A. Suppose A is full rank. If $rank([A \quad b]) > rank(A)$ (since the number of columns of the matrix increased by 1 and A is assumed full rank, this would imply the rank is $rank(A)+1$), b could not be written as a linear combination of the columns of A. We must have that $rank([A \quad b]) = rank(A)$ in order for the system to have an exact solution. From the def of linear independence applies here, observe that $Ax = b \implies Ax - b = 0$. Therefore $[A \quad b]\begin{bmatrix} x \\ -1 \end{bmatrix} = 0$. If $Ax = b$ has an exact solution, then $[A \quad b]$ does not have linearly independent columns.

Condition on $rank(A)$ for more than one exact solution: If the system of linear equations $Ax = b$ has more than one exact solution, then there is at least one non zero vector w for which $x+w$ is also a solution. That is, $A(x+w) = b$. If x is an exact solution, then $Ax = b$. This implies $Aw = 0$. Therefore, the columns of A are linearly dependent. If $rank(A) < dim(x)$, then there will be more than one exact solution.

Overdetermined: More equations than unknowns. Could either have zero, one, or infinitely many solutions. Underdetermined: Fewer equations than unknowns. Could either have zero or infinitely many solutions.

1.2 Gradients

$x, w \in \mathbb{R}^n$ and $Q \in \mathbb{R}^{n\times n}$

Linear: $\nabla_x x^T w = \nabla_x w^T x = w$

2-norm: $\nabla_x \|x\|_2^2 = 2x$

Quadratic: $\nabla_x x^T Qx = (Q+Q^T)x$. If Q symmetric, then $2Qx$.

1.3 Positive Definite Matrices

1. For any matrix A, $A^TA \succeq 0$ and $AA^T \succeq 0$. Further, if the columns of A are linearly independent, then $A^TA \succ 0$.
2. If $A \succ 0$, then A^{-1} exists.

1.4 Subspace

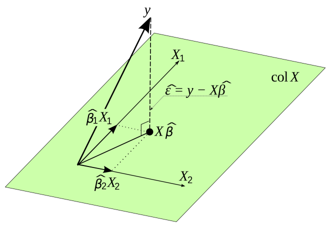
A set of points $S \subseteq \mathbb{R}^n$ is a subspace if 1. $0 \in S$ (S contains the origin) 2. If $x, y \in S$, then $x+y \in S$ 3. If $x \in S$, $\alpha \in \mathbb{R}$, then $\alpha x \in S$.

2 Least Squares

LS *always* has a solution, but does the problem have a unique solution?
1. If X is full rank (its columns are linearly independent), then $\hat{w}_{LS}$ is unique
2. If X is not full rank, then X^TX is not invertible, and $\hat{w}_{LS}$ is not unique

2.1 Geometry

We know $\hat{r} = y - X\hat{w}$ is orthogonal to the span of the columns of X. Thus $x_i^T\hat{r} = 0$, or $X^T\hat{r} = 0$. This implies $X^T(y - X\hat{w}) = 0$. $\hat{w}$ is a solution to the linear system of equations $X^TX\hat{w} = X^Ty$.



• The question we're trying to answer: What is the point in $col(X)$ that has the shortest distance to y? In $\mathbb{R}^2$, what are the weights β_1 and β_2 such that $\beta_1 x_1 + \beta_2 x_2$ has the shortest distance to y? • $colX$ is the space of all vectors that can be written as $\alpha x_1 + \beta x_2$ for some $\alpha, \beta \in \mathbb{R}$, that is the span of the columns of X. y may not lie in this space. • The residual vector will form a right angle with $colX$, because any other angle would correspond to a longer distance.

2.2 Decision Boundary

Given $x, \hat{w} \in \mathbb{R}^p$, the decision boundary is the set of points such that $x^Tw = 0$. Example: $x^T = [x_1 \quad x_2 \quad x_3 \quad 1]$. Find w so that the decision boundary is parallel to the $x_1 - x_2$ plane and includes the point $(0,0,1)$. This implies classification doesn't depend on x_1, x_2, so $w_1 = w_2 = 0$. Further, we require $w_3 + w_4 = 0$, or that $w_3 = -w_4$.

3 Orthogonal Matrices, Projections, and LS

Properties of orthogonal matrix U, V: 1. $U^TU = I$, but in general $UU^T \neq I$. 2. UV orthogonal. 3. Length preserving: $\|Uv\|_2 = \|v\|_2$. Proof: $\|Uv\|_2^2 = (Uv)^T(Uv) = v^TU^TUv = v^Tv = \|v\|_2^2$.

Example: $X \in \mathbb{R}^{n\times p}$ $(n > p)$ with linearly independent columns. U orthonormal basis for the p-dimensional space spanned by the columns of X. Then $X = UT$, where T is $p\times p$ and invertible. T must be invertible. $X = UT$ means any column of X is a weighted combination of the columns of T. Since X and U span the same space, we can also write any column of U as a weighted combination of the columns of X: $U = XB \implies U = UTB \implies TB = I$, or T is invertible.

Application to LS: $\hat{w}$ solution to LS. Then $\hat{y} = X\hat{w} = UU^Ty$. Let $P_x = X(X^TX)^{-1}X^Ty$ be a projection matrix. Since $span(X) = span(U)$, $P_xy = P_uy \implies P_x = P_u = UU^T$. Thus $\hat{y} = P_xy = P_uy = UU^Ty$.

4 Taste Profiles

$X \in \mathbb{R}^{n\times p}$, n movies, p people. $T \in \mathbb{R}^{n\times r}$, and $W \in \mathbb{R}^{r\times p}$. T_k is the kth representative taste profile and w_k^T (the kth row of W) is the affinity of each customer with the kth representative profile.

4.1 Matrix Multiplication

$X = TW \implies X_{ij} = \langle i\text{th row of T}, j\text{th column of W}\rangle$.

1. The jth column of X is a weighted sum of the columns of T, where the jth column of W tells us the weights: $x_j = Tw_j$. Interpretation: the tastes (preferences) of the jth customer.
2. The ith row of X is $x_i^T = t_i^TW$ where t_i^T is the ith row of T. Interpretation: how much each customer likes movie i.

Inner product representation:

$$TW = \begin{bmatrix} —t_1^T— \\ —t_2^T— \\ \vdots \\ —t_n^T— \end{bmatrix} \begin{bmatrix} | & | & & | \\ w_1 & w_2 & \dots & w_p \\ | & | & & | \end{bmatrix} = \begin{bmatrix} t_1^Tw_1 & t_1^Tw_2 & \dots & t_1^Tw_p \\ t_2^Tw_1 & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ t_n^Tw_1 & & & t_n^Tw_p \end{bmatrix} \tag{1}$$

Outer Product Representation:

$$TW = \begin{bmatrix} | & | & & | \\ T_1 & T_2 & \dots & T_r \\ | & | & & | \end{bmatrix} \begin{bmatrix} —w_1^T— \\ —w_2^T— \\ \vdots \\ —w_r^T— \end{bmatrix} = \sum_{k=1}^{r} T_kw_k^T \tag{2}$$

(the sum of rank 1 matrices. TW has rank r if and only if the columns of T are rows of W are linearly independent).

5 Tikhonov Regularization

$X \in \mathbb{R}^{n\times p}$, $y \in \mathbb{R}^n$. Objective: $\hat{w} = \arg\min_w f(w) = \arg\min_w \|y - Xw\|_2^2 + \lambda\|w\|_2^2$, where $\|y - Xw\|_2^2$ measures the fit to the data, $\lambda > 0$ is a regularization parameter or tuning parameter, and $\|w\|_2^2$ is a regularizer. $\|w\|_2^2$ measures the energy in w.

Derivation: $f(w) = y^Ty - 2w^TX^Ty + w^T(X^TX + \lambda I)w$. Then $\nabla_w f(w) = -2X^Ty + 2(X^TX + \lambda I)$. $(X^TX + \lambda I)$ is *always* invertible, since it is pd. Fix $0 \neq a \in \mathbb{R}^n$. Then $a^T(X^TX + \lambda I)a = a^TX^TXa + \lambda a^Ta = \|Xa\|_2^2 + \lambda\|a\|_2^2$. $\|Xa\|_2^2 \geq 0$ (it could be 0 if X is not full rank and a is in the null space of X – this is what causes troubles with LS) but $\lambda\|a\|_2^2 > 0$. Therefore, $(X^TX + \lambda I)$ is positive definite. Pd implies invertible. Therefore the unique solution is $\hat{w} = (X^TX + \lambda I)^{-1}X^Ty$.

Benefits: 1. $\hat{w}$ *always* unique, even when no LS solution exists. 2. Even if X is full rank, X^TX can be badly behaved in LS (the inverse may magnify errors). λ helps us avoid amplifying noise. Example: $y = Xw + \varepsilon$ ($\iff y_i = x_i^Tw + \varepsilon_i$). LS: $\hat{w} = (X^TX)^{-1}X^Ty = (X^TX)^{-1}X^T(Xw + \varepsilon) = w + (X^TX)^{-1}\varepsilon$