Solving Least Squares

Normal Equations

- Let $\tilde{A}$ have full column rank, and be size mxn with $m \geq n$
- Diagonal (nonzero) weighting $A = D\tilde{A}$ does not change the rank/size
  - but changes the answer when $D \neq I$ and $m \neq n$
- Minimizing $\|r\|_2 = \|b - Ac\|_2$ leads to the normal equations $A^T Ac = A^T b$ for the critical point
- Since $A^T A$ is SPD, $A^T Ac = A^T b$ has a unique solution obtainable via fast/efficient SPD solvers
- When b is in the range of A, the unique solution to $A^T Ac = A^T b$ makes $r = 0$, and thus is also the unique solution to $Ac = b$
  - When A is square ($m = n$), and full rank, b is always in the range of A

Condition Number for the Normal Equations

- Compare $A = U\Sigma V^T$ and $A^TA = V\Sigma^T\Sigma V^T = V\Lambda V^T$ where $\Lambda = \Sigma^T\Sigma$ is a diagonal size nxn matrix of singular values squared
- Since the singular values of A^TA are the square of those in A, the condition number $\frac{\sigma_{max}}{\sigma_{min}}$ of A^TA is also squared (compared to A)
  - Thus, solving the normal equations requires twice the precision (e.g. $(10^7)^2 = 10^{14}$)
- It takes twice as much precision to get the same number of significant digits!
- The <u>normal equations are not the preferred approach</u> (unless A is extremely well conditioned)
- However, (like the SVD) it is a great tool for theoretical purposes
  - Can transform any full rank matrix into an SPD system

Understanding Least Squares

- When $A = U\Sigma V^T$ has full column rank, $\Sigma = \begin{pmatrix} \hat{\Sigma} \\ 0 \end{pmatrix}$ with $\hat{\Sigma}$ a size nxn diagonal matrix of (strictly) positive singular values
  - The 0 submatrix is size $(m-n)xn$ and doesn't exist when $m = n$
- Note: $A^T A = V(\hat{\Sigma} \quad 0)\begin{pmatrix} \hat{\Sigma} \\ 0 \end{pmatrix} V^T = V\hat{\Sigma}^2 V^T$ and $(A^T A)^{-1} = V\hat{\Sigma}^{-2} V^T$
- $c = (A^T A)^{-1} A^T b = V\hat{\Sigma}^{-2} V^T V(\hat{\Sigma} \quad 0) U^T b = V(\hat{\Sigma}^{-1} \quad 0) U^T b$
- $Ac = U\begin{pmatrix} \hat{\Sigma} \\ 0 \end{pmatrix} V^T V(\hat{\Sigma}^{-1} \quad 0) U^T b = U\begin{pmatrix} I_{nxn} & 0 \\ 0 & 0 \end{pmatrix} U^T b$
- $r = b - Ac = U I_{mxm} U^T b - U\begin{pmatrix} I_{nxn} & 0 \\ 0 & 0 \end{pmatrix} U^T b = U\begin{pmatrix} 0 & 0 \\ 0 & I_{(m-n)x(m-n)} \end{pmatrix} U^T \text{b}$

Recall: Summary (Unit 3)

- The columns of V that do not correspond to “nonzero” singular values form an orthonormal basis for the null space of A
- The remaining columns of V form an orthonormal basis for the space perpendicular to the null space of A (parameterizing meaningful inputs)

- The columns of U corresponding to “nonzero” singular values form an orthonormal basis for the range of A
- The remaining columns of U form an orthonormal basis for the (unattainable) space perpendicular to the range of A

- One can drop the columns of U and V that do not correspond to “nonzero” singular values and still obtain a valid factorization of A
- One can drop the columns of U and V that correspond to “smaller” singular values and still obtain a reasonable approximation of A

Understanding Least Squares

- A has only n singular values
- So, only the first n columns of U (which has m columns) span the range of A
- Write $\begin{pmatrix} \hat{b}_r \\ \hat{b}_z \end{pmatrix} = U^T b$
- $\hat{b}_r$ (which is size $nx1$) represents the part of b in the range of A
- $\hat{b}_z$ (which is size $(m-n)x1$) represents the part of b which is orthogonal to the range of A
- Then: $c = V\hat{\Sigma}^{-1}\hat{b}_r$, $Ac = U\begin{pmatrix} \hat{b}_r \\ 0 \end{pmatrix}$, and $r = U\begin{pmatrix} 0 \\ \hat{b}_z \end{pmatrix}$

Recall: Singular Value Decomposition (Unit 3)

- Factorization of any size mxn matrix: $A = U\Sigma V^T$
- Σ is mxn diagonal with non-negative diagonal entries (called singular values)
- U is mxm orthogonal, V is nxn orthogonal (their columns are called singular vectors)
  - Orthogonal matrices have orthonormal columns (an orthonormal basis), so their transpose is their inverse. They preserve inner products, and thus are rotations, reflections, and combinations thereof
  - If A has complex entries, then U and V are unitary (conjugate transpose is their inverse)

- Introduced and rediscovered many times: Beltrami 1873, Jordan 1875, Sylvester 1889, Autonne 1913, Eckart and Young 1936. Pearson introduced principal component analysis (PCA) in 1901, which uses SVD. Numerical methods by Chan, Businger, Golub, Kahan, etc.

Orthogonal Matrices and the L2 norm

- An orthogonal $\hat{Q}$ has $\hat{Q}\hat{Q}^T = \hat{Q}^T\hat{Q} = I$

- $\|\hat{Q}r\|_2 = \sqrt{\hat{Q}r \cdot \hat{Q}r} = \sqrt{r^T\hat{Q}^T\hat{Q}r} = \sqrt{r^Tr} = \|r\|_2$
- $\|\hat{Q}^Tr\|_2 = \sqrt{\hat{Q}^Tr \cdot \hat{Q}^Tr} = \sqrt{r^T\hat{Q}\hat{Q}^Tr} = \sqrt{r^Tr} = \|r\|_2$

- That is, orthogonal transformations preserve Euclidean distance

Understanding Least Squares

- Since U is orthogonal, $\|r\|_2 = \left\| U \begin{pmatrix} 0 \\ \hat{b}_z \end{pmatrix} \right\|_2 = \|\hat{b}_z\|_2$

- Consider the diagonalized SVD view of $Ac = b$ (for a full rank A):

$$U\Sigma V^T c = b \;\; \text{or} \;\; \begin{pmatrix} \hat{\Sigma} \\ 0 \end{pmatrix} \hat{c} = \begin{pmatrix} \hat{b}_r \\ \hat{b}_z \end{pmatrix}$$

- The first block row gives $c = V\hat{\Sigma}^{-1}\hat{b}_r$, identical to the least squares solution
- The norm of the residual is $\left\| \begin{pmatrix} \hat{b}_r \\ \hat{b}_z \end{pmatrix} - \begin{pmatrix} \hat{\Sigma} \\ 0 \end{pmatrix} \hat{c} \right\|_2 = \left\| \begin{pmatrix} 0 \\ \hat{b}_z \end{pmatrix} \right\|_2 = \|\hat{b}_z\|_2$, identical to the norm of the residual for the least squares solution

- The SVD approach gives the same (minimum residual) least squares solution

Recall: Gram-Schmidt (Unit 5)

- Orthogonalizes a set of vectors
- For each new vector, subtract its (weighted) dot product overlap with all prior vectors, making it orthogonal to them
- A-orthogonal Gram-Schmidt simply uses an A-weighted dot/inner product
- Given vector $\bar{S}^q$, subtract out the A-overlap with s^1 to s^{q-1} so that the resulting vector s^q has $< s^q, s^{\hat{q}} >_A = 0$ for $\hat{q} \in \{1,2,\cdots,q-1\}$
- That is, $s^q = \bar{S}^q - \sum_{\hat{q}=1}^{q-1} \frac{<\bar{S}^q, s^{\hat{q}}>_A}{<s^{\hat{q}}, s^{\hat{q}}>_A} s^{\hat{q}}$ where the two non-normalized $s^{\hat{q}}$ both require division by their norm (and $< s^{\hat{q}}, s^{\hat{q}} >_A = \left\| s^{\hat{q}} \right\|_A^2$)
- Proof: $< s^q, s^{\tilde{q}} >_A = < \bar{S}^q, s^{\tilde{q}} >_A - \frac{<\bar{S}^q, s^{\tilde{q}}>_A}{<s^{\tilde{q}}, s^{\tilde{q}}>_A} < s^{\tilde{q}}, s^{\tilde{q}} >_A = 0$

Gram-Schmidt for QR Factorization

- From A, create a full rank Q with orthonormal columns
- For each column a_k, subtract the overlap with all prior columns in Q and make the result unit length:

$$q_k = \frac{a_k - \sum_{\hat{k}=1}^{k-1} < a_k, q_{\hat{k}} > q_{\hat{k}}}{\left\| a_k - \sum_{\hat{k}=1}^{k-1} < a_k, q_{\hat{k}} > q_{\hat{k}} \right\|_2}$$

- Define $r_{\hat{k}k} =< a_k, q_{\hat{k}} >$ for $k > \hat{k}$, and $r_{kk} = \left\| a_k - \sum_{\hat{k}=1}^{k-1} < a_k, q_{\hat{k}} > q_{\hat{k}} \right\|_2$
- Then $q_k = \frac{a_k - \sum_{\hat{k}=1}^{k-1} r_{\hat{k}k} q_{\hat{k}}}{r_{kk}}$, and thus $a_k = r_{kk} q_k + \sum_{\hat{k}=1}^{k-1} r_{\hat{k}k} q_{\hat{k}} = \sum_{\hat{k}=1}^{k} r_{\hat{k}k} q_{\hat{k}}$
- This gives $A = QR$ where R is upper triangular and $Q^T Q = I$

Gram-Schmidt for QR (an example)

- Example: $A = QR$ with upper triangular R

$$\begin{pmatrix} 3 & -3 & 3 \\ 2 & -1 & 1 \\ 2 & -1 & -1 \\ 2 & -3 & 3 \\ 2 & -3 & 5 \end{pmatrix} = \begin{pmatrix} 3/5 & 0 & 0 \\ 2/5 & 1/2 & 1/2 \\ 2/5 & 1/2 & -1/2 \\ 2/5 & -1/2 & -1/2 \\ 2/5 & -1/2 & 1/2 \end{pmatrix} \begin{pmatrix} 5 & -5 & 5 \\ 0 & 2 & -4 \\ 0 & 0 & 2 \end{pmatrix}$$

- Note that $Q^T Q = I_{3x3}$ since the columns of Q are orthonormal
- However, $QQ^T \neq I_{5x5}$ since Q is only a subset of an orthogonal matrix

Not Good for Large Matrices

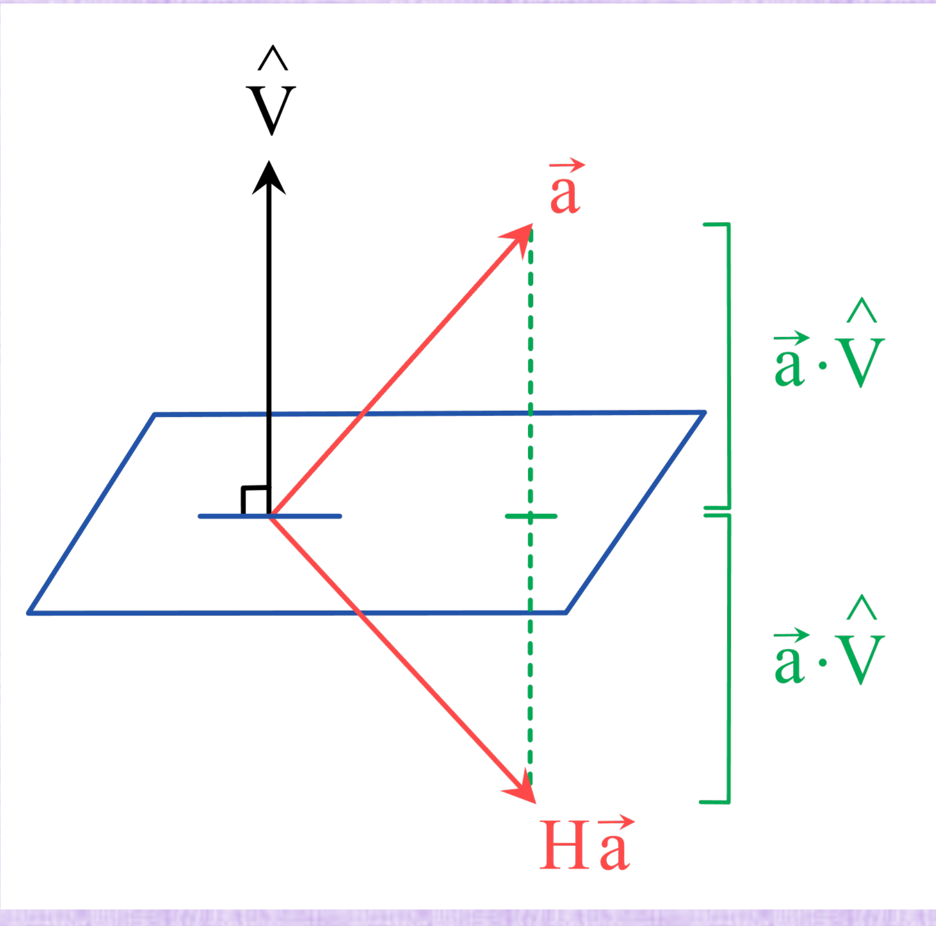
- Gram-Schmidt has too much numerical drift, when used on a large number of vectors
- Don’t use Gram-Schmidt to find $A = QR$

- But it does provide a good conceptual way to think about $A = QR$

QR Factorization

- Consider $A = QR$ with upper triangular R and $Q^T Q = I$
- Q is size mxn (just like A) with n orthonormal columns
- Let $\tilde{Q}$ be the matrix with $m - n$ orthonormal columns that span the space perpendicular to the range of Q
- Then, the size mxm matrix $\hat{Q} = (Q \quad \tilde{Q})$ is orthogonal
- $\|r\|_2 = \|\hat{Q}^T r\|_2 = \left\| \begin{pmatrix} Q^T \\ \tilde{Q}^T \end{pmatrix} (b - QRc) \right\|_2 = \left\| \begin{pmatrix} Q^T b - Rc \\ \tilde{Q}^T b \end{pmatrix} \right\|_2 = \left\| \begin{pmatrix} \hat{b}_Q - Rc \\ \hat{b}_{\tilde{Q}} \end{pmatrix} \right\|_2$
- Only the first (block) row varies with c, so $\|r\|_2$ is minimized by solving $Rc = Q^T b$
  - Then, $\|r\|_2 = \|\tilde{Q}^T b\|_2 = \|\hat{b}_{\tilde{Q}}\|_2$
- Since R is upper triangular, $Rc = \hat{b}_Q$ can be solved via back-substitution

Householder Transform



- A unit normal $\hat{v}$ implicitly defines a plane orthogonal to it
- $H = I - 2\hat{v}\hat{v}^T$ reflects vectors across that plane
- $Ha = a - 2(\hat{v}^T a)\,\hat{v}$

- H is orthogonal with $H = H^T = H^{-1}$
- Don't form the size mxm matrix H
- Instead, apply it using only $\hat{v}$

Householder Transform for QR

- Choose the directions $v_k = a - Ha$ in order to zero out elements
- E.g. $v_k = \begin{pmatrix} a_1 \\ \vdots \\ a_{k-1} \\ a_k \\ a_{k+1} \\ \vdots \\ a_n \end{pmatrix} - \begin{pmatrix} a_1 \\ \vdots \\ a_{k-1} \\ \gamma \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \hat{a}_k - \gamma \hat{e}_k$ where $\hat{a}_k = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ a_k \\ a_{k+1} \\ \vdots \\ a_n \end{pmatrix}$
- Ha (as a reflection) should be the same length as a, so $\gamma = \pm\|\hat{a}_k\|_2$
- Then, $v_k = \hat{a}_k \pm \|\hat{a}_k\|_2 \hat{e}_k$, which is normalized to $\hat{v}_k = \frac{v_k}{\|v_k\|_2}$
- For robustness, $v_k = \hat{a}_k + S(a_k)\|\hat{a}_k\|_2 \hat{e}_k$ where $S(a_k) = \pm 1$ is the sign function

Householder Transform (an example)

- Let $a_1 = \begin{pmatrix} 2 \\ 1 \\ 2 \end{pmatrix}$ and consider $v_k = \hat{a}_k + S(a_k)\|\hat{a}_k\|_2 \hat{e}_k$

- Then, $\hat{a}_1 = \begin{pmatrix} 2 \\ 1 \\ 2 \end{pmatrix}$, $v_1 = \begin{pmatrix} 2 \\ 1 \\ 2 \end{pmatrix} + S(2)\sqrt{9}\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 5 \\ 1 \\ 2 \end{pmatrix}$, $\hat{v}_1 = \frac{1}{\sqrt{30}}\begin{pmatrix} 5 \\ 1 \\ 2 \end{pmatrix}$

- Then, $H_1 a_1 = a_1 - 2\left(\hat{v}_1{}^T a_1\right)\hat{v}_1 = \begin{pmatrix} 2 \\ 1 \\ 2 \end{pmatrix} - 2\,\frac{15}{\sqrt{30}}\frac{1}{\sqrt{30}}\begin{pmatrix} 5 \\ 1 \\ 2 \end{pmatrix} = \begin{pmatrix} -3 \\ 0 \\ 0 \end{pmatrix}$

Householder Transform (another example)

- Let $a_2 = \begin{pmatrix} 6 \\ 3 \\ 4 \end{pmatrix}$ and consider $v_k = \hat{a}_k + S(a_k)\|\hat{a}_k\|_2 \hat{e}_k$

- Then, $\hat{a}_2 = \begin{pmatrix} 0 \\ 3 \\ 4 \end{pmatrix}$, $v_2 = \begin{pmatrix} 0 \\ 3 \\ 4 \end{pmatrix} + S(3)\sqrt{25}\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ -2 \\ 4 \end{pmatrix}$, $\hat{v}_2 = \frac{1}{\sqrt{20}}\begin{pmatrix} 0 \\ -2 \\ 4 \end{pmatrix}$

- Then, $H_2 a_2 = a_2 - 2\left(\hat{v}_2{}^T a_2\right)\hat{v}_2 = \begin{pmatrix} 6 \\ 3 \\ 4 \end{pmatrix} - 2\,\frac{10}{\sqrt{20}}\frac{1}{\sqrt{20}}\begin{pmatrix} 0 \\ -2 \\ 4 \end{pmatrix} = \begin{pmatrix} 6 \\ 5 \\ 0 \end{pmatrix}$

Householder Transform for QR

- For each column of A, construct the Householder transform that zeroes out entries below the diagonal
- Then $H_n H_{n-1} \cdots H_2 H_1 A = \begin{pmatrix} R \\ 0 \end{pmatrix}$ and $H_n H_{n-1} \cdots H_2 H_1 b = \begin{pmatrix} \hat{b}_Q \\ \hat{b}_{\tilde{Q}} \end{pmatrix}$
- Apply each H_k efficiently using $\hat{v}_{\mathrm{k}}$, and remember to apply it to all columns $\geq k$
  - It doesn't affect columns $< k$ (because of all the zeros at the top of $\hat{v}_{\mathrm{k}}$)
- Note: H_n is required to get zeroes at the bottom of the last column
- Letting $\hat{Q}^T = H_n H_{n-1} \cdots H_2 H_1$ gives $\hat{Q}^T A = \begin{pmatrix} R \\ 0 \end{pmatrix}$ or $A = \hat{Q} \begin{pmatrix} R \\ 0 \end{pmatrix}$
- $\|r\|_2 = \left\|\hat{Q}^T r\right\|_2 = \left\|\hat{Q}^T \left(b - \hat{Q} \begin{pmatrix} R \\ 0 \end{pmatrix} c\right)\right\|_2 = \left\|\begin{pmatrix} \hat{b}_Q \\ \hat{b}_{\tilde{Q}} \end{pmatrix} - \begin{pmatrix} Rc \\ 0 \end{pmatrix}\right\|_2$
- Solve $Rc = \hat{b}_Q$ via back-substitution to minimize $\|r\|_2$ to a value of $\left\|\hat{b}_{\tilde{Q}}\right\|_2$