QR factorization

- Gram-Schmidt procedure, QR factorization
- orthogonal decomposition induced by a matrix

Gram-Schmidt procedure

given independent vectors $a_1, \ldots, a_n \in \mathbb{R}^m$, G-S procedure finds orthonormal vectors $q_1, \ldots, q_n$ s.t.

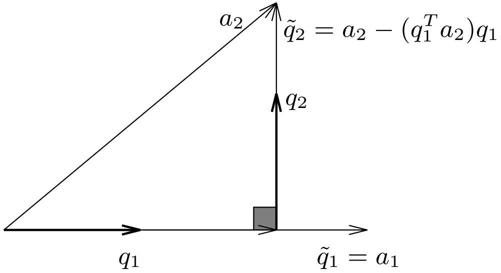
$$\mathbf{span}(a_1, \ldots, a_r) = \mathbf{span}(q_1, \ldots, q_r) \qquad \text{for } r \leq n$$

- thus, $q_1, \ldots, q_r$ is an orthonormal basis for $\mathbf{span}(a_1, \ldots, a_r)$
- rough idea of method: first *orthogonalize* each vector w.r.t. previous ones; then *normalize* result to have norm one

Procedure

$$q_1^T \tilde{q}_3 = q_1^T a_3 - (q_1^T a_3)\underbrace{q_1^T q_1}_{\|q_1\|^2 = 1} - (q_2^T a_3)\underbrace{q_1^T q_2}_{0} = 0$$

- step 1a. $\tilde{q}_1 := a_1$
- step 1b. $q_1 := \tilde{q}_1/||\tilde{q}_1||$ *(normalize)*
- step 2a. $\tilde{q}_2 := a_2 - \underbrace{(q_1^\mathsf{T} a_2)}_{\alpha}\underbrace{q_1}$ *(remove q_1 component from a_2)*
- step 2b. $q_2 := \tilde{q}_2/||\tilde{q}_2||$ *(normalize)*
- step 3a. $\underline{\tilde{q}_3} := a_3 - (q_1^\mathsf{T} a_3)q_1 - (q_2^\mathsf{T} a_3)q_2$ *(remove q_1, q_2 components)*
- step 3b. $\underline{q_3} := \tilde{q}_3/||\tilde{q}_3||$ *(normalize)*
- etc.



for $i = 1, 2, \ldots, n$ we have

$$\begin{aligned} a_i &= (q_1^\mathsf{T} a_i)q_1 + (q_2^\mathsf{T} a_i)q_2 + \cdots + (q_{i-1}^\mathsf{T} a_i)q_{i-1} + ||\tilde{q}_i||q_i \\ &= r_{1i}q_1 + r_{2i}q_2 + \cdots + r_{ii}q_i \end{aligned}$$

(note that the r_{ij}'s come right out of the G-S procedure, and $r_{ii} \neq 0$)

QR decomposition

written in matrix form: $A = QR$, where $A \in \mathbb{R}^{m\times n}$, $Q \in \mathbb{R}^{m\times n}$, $R \in \mathbb{R}^{n\times n}$:

$$\underbrace{\begin{bmatrix} a_1 & a_2 & \cdots & a_n \end{bmatrix}}_{A} = \underbrace{\begin{bmatrix} q_1 & q_2 & \cdots & q_n \end{bmatrix}}_{Q} \underbrace{\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ 0 & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & r_{nn} \end{bmatrix}}_{R}$$

- $Q^{\mathsf{T}}Q = I$, and R is upper triangular & invertible
- called QR **decomposition** (or factorization) of A
- usually computed using a variation on Gram-Schmidt procedure which is less sensitive to numerical (rounding) errors
- columns of Q are orthonormal basis for **range**(A)

General Gram-Schmidt procedure

- in basic G-S we assume $a_1, \ldots, a_n \in \mathbb{R}^m$ are independent
- if $a_1, \ldots, a_n$ are dependent, we find $\tilde{q}_j = 0$ for some j, which means a_j is linearly dependent on $a_1, \ldots, a_{j-1}$
- modified algorithm: when we encounter $\tilde{q}_j = 0$, skip to next vector a_{j+1} and continue:

$$
\begin{array}{l}
r = 0 \\
\textbf{for } i = 1, \ldots, n \\
\quad \tilde{a} = a_i - \sum_{j=1}^{r} q_j q_j^\mathsf{T} a_i \\
\quad \textbf{if } \tilde{a} \neq 0 \\
\quad\quad r = r + 1 \\
\quad\quad q_r = \tilde{a}/||\tilde{a}||
\end{array}
$$

Staircase form

on exit,

- $q_1, \ldots, q_r$ is an orthonormal basis for **range**(A) (hence $r =$ **rank**(A))
- each a_i is linear combination of previously generated q_j's

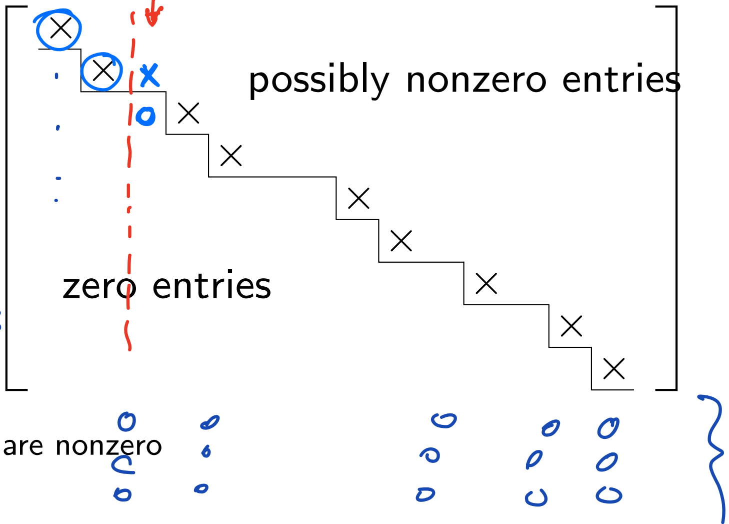
in matrix notation we have $A = \overset{m\times r}{Q}\overset{r\times n}{R}$ with $Q^{\mathsf{T}}Q = I$ and $R \in \mathbb{R}^{r\times n}$ in *upper staircase form*

$a_1 = r_{11} q_1$

$a_2 = r_{12} q_1 + r_{22} q_2$

$a_3 = r_{13} q_1 + r_{23} q_2$

$a_4 = r_{14} q_1 + r_{24} q_2 + r_{34} q_3$



'corner' entries (shown as $\times$) are nonzero

Applications

- directly yields orthonormal basis for **range**(A)
- yields factorization $A = BC$ with $B \in \mathbb{R}^{m\times r}$, $C \in \mathbb{R}^{r\times n}$, $r =$ **rank**(A)
- staircase pattern in R shows which columns of A are dependent on previous ones

works incrementally: one G-S procedure yields QR factorizations of $\begin{bmatrix} a_1 & \cdots & a_p \end{bmatrix}$ for $p = 1, \ldots, n$

$$\begin{bmatrix} a_1 & \cdots & a_p \end{bmatrix} = \begin{bmatrix} q_1 & \cdots & q_s \end{bmatrix} R_p$$

where $s =$ **rank**$(\begin{bmatrix} a_1 & \cdots & a_p \end{bmatrix})$ and R_p is leading $s \times p$ submatrix of R

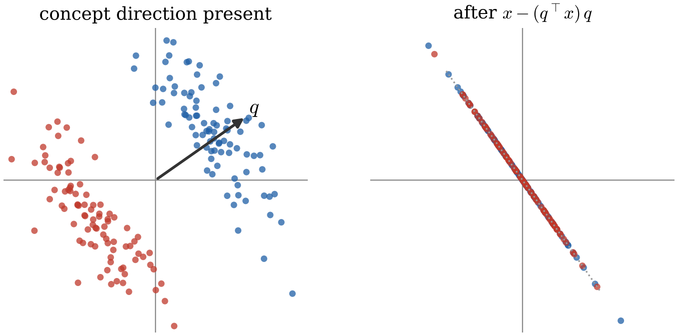
Application: removing a direction

the core Gram-Schmidt step removes a component,

$$x \;\mapsto\; x - (q^\top x)\, q$$

the projection of x onto the subspace orthogonal to q

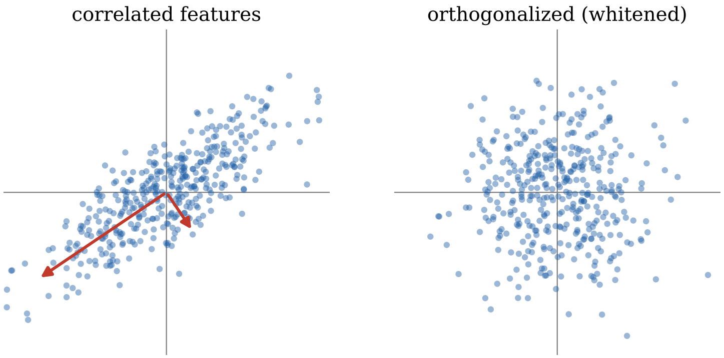
- in language models, subtracting a single direction from the internal activations can *erase a concept* (a topic, a style, or a bias)
- the readout along q is gone, while everything orthogonal to q is left untouched



Application: orthogonalizing features

running Gram-Schmidt on the columns of a data matrix (the Q of $A = QR$) turns *correlated* features into *uncorrelated* ones

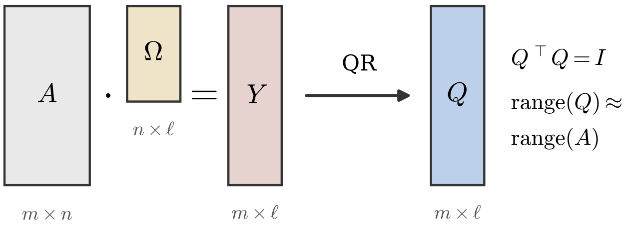
- this is *whitening*, or decorrelation, a standard pre-processing step
- nearly-dependent features make a model unstable to fit; orthogonal features are easier and better conditioned



Application: orthonormal bases at scale

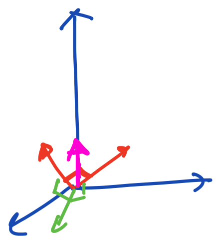
QR is the practical way to *produce* orthonormal vectors

- **randomized SVD:** sketch $Y = A\Omega$ with a random Ω, then take $Y = QR$; the columns of Q span almost all of **range**(A), giving a fast low-rank approximation of a huge matrix
- **orthogonal initialization:** the Q from the QR of a random matrix is a random orthogonal matrix, used to initialize deep networks



$$a_1, \dots, a_k, e_1, e_2, \dots, e_n \xrightarrow{GS} q_1, \dots, q_r \xrightarrow{\text{expand into a basis for } \mathbb{R}^n} q_1, \dots, q_r, q_{r+1}, \dots, q_n$$

$$Q = \left[\begin{array}{c|c} \underset{n\times r}{Q_1} & \underset{n\times(n-r)}{Q_2} \end{array}\right]$$



$$Q^T = Q^{-1}$$

$$A = Q_1 R$$

$$A = \left[\begin{array}{c|c} \underbrace{Q_1}_{n\times r} & \underbrace{Q_2}_{n\times(n-r)} \end{array}\right] \begin{array}{c} \left[\begin{array}{c} R \\ \hline 0 \end{array}\right] \end{array} \begin{array}{l} \} \; r\times n \\ \} \; (n-r)\times n \end{array}$$

$$\begin{bmatrix} q_1 & q_2 \end{bmatrix} \begin{bmatrix} r \\ 0 \end{bmatrix} \to q_1 r + q_2 0 = q_1 r$$

$$= Q_1 R + Q_2 0$$

$$= QR$$

Full QR factorization

with $A = Q_1 R_1$ the QR factorization as above, write

$$A = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} R_1 \\ 0 \end{bmatrix}$$

where $\begin{bmatrix} Q_1 & Q_2 \end{bmatrix}$ is orthogonal, $i.e.$, columns of $Q_2 \in \mathbb{R}^{m \times (m-r)}$ are orthonormal, orthogonal to Q_1

to find Q_2:

- find any matrix $\tilde{A}$ s.t. $\begin{bmatrix} A & \tilde{A} \end{bmatrix}$ has rank m ($e.g.$, $\tilde{A} = I$)
- apply general Gram-Schmidt to $\begin{bmatrix} A & \tilde{A} \end{bmatrix}$
- Q_1 are orthonormal vectors obtained from columns of A
- Q_2 are orthonormal vectors obtained from extra columns ($\tilde{A}$)

$i.e.$, any set of orthonormal vectors can be *extended* to an orthonormal basis for $\mathbb{R}^m$

Example: Full QR factorization

$$A = \begin{bmatrix} 50 & 2 & 37 \\ 54 & -2 & 31 \\ 38 & 41 & 2 \\ 39 & 46 & -16 \end{bmatrix}$$

$$Q = \begin{bmatrix} -0.55 & -0.36 & -0.45 & -0.61 \\ -0.59 & -0.47 & 0.42 & 0.50 \\ -0.42 & 0.52 & -0.57 & 0.48 \\ -0.43 & 0.61 & 0.55 & -0.38 \end{bmatrix}$$

$$R = \begin{bmatrix} -91.55 & -36.53 & -32.51 \\ 0.00 & 49.71 & -36.80 \\ 0.00 & 0.00 & -13.37 \\ 0.00 & 0.00 & 0.00 \end{bmatrix}$$

Pivoted QR

can permute columns with $\times$ to front of matrix:

$$A = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} R_{11} & R_{12} \\ 0 & 0 \end{bmatrix} P^\mathsf{T}$$

where:

- $Q^\mathsf{T} Q = I$
- $R_{11} \in \mathbb{R}^{r \times r}$ is upper triangular and invertible
- $P \in \mathbb{R}^{n \times n}$ is a permutation matrix
  (which moves forward the columns of a which generated a new q)

Example: Pivoted QR

$$A = \begin{bmatrix} -32 & 54 & 70 & -56 \\ 74 & -8 & -6 & 5 \\ 35 & 61 & 38 & -52 \\ 26 & -46 & -20 & 35 \\ -38 & 45 & -61 & -8 \end{bmatrix} \qquad P = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$Q = \begin{bmatrix} -0.52 & 0.18 & 0.56 & -0.03 & 0.62 \\ 0.08 & -0.76 & -0.18 & 0.42 & 0.46 \\ -0.59 & -0.55 & 0.04 & -0.44 & -0.40 \\ 0.44 & -0.14 & -0.04 & -0.79 & 0.40 \\ -0.43 & 0.27 & -0.81 & -0.10 & 0.29 \end{bmatrix}$$

$$R = \begin{bmatrix} -104.12 & 29.69 & -41.50 & 78.81 \\ 0.00 & -94.68 & -17.70 & 8.28 \\ 0.00 & 0.00 & 92.01 & -29.57 \\ 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 \end{bmatrix}$$

Orthogonal decomposition induced by A

$$\textbf{range}(A)^\perp = \textbf{null}(A^T)$$

- from $A^T = \begin{bmatrix} R_1^T & 0 \end{bmatrix} \begin{bmatrix} Q_1^T \\ Q_2^T \end{bmatrix}$ we see that

$$A^T z = 0 \quad \Longleftrightarrow \quad Q_1^T z = 0 \quad \Longleftrightarrow \quad z \in \textbf{range}(Q_2)$$

  so $\textbf{range}(Q_2) = \textbf{null}(A^T)$

- the columns of Q_2 are an orthonormal basis for $\textbf{null}(A^T)$
- called *orthogonal decomposition (of $\mathbb{R}^m$) induced by $A \in \mathbb{R}^{m \times n}$*
- every $y \in \mathbb{R}^n$ can be written uniquely as $y = z + w$, with $z \in \textbf{range}(A)$, $w \in \textbf{null}(A^T)$ (we'll soon see what the vector z is ...)

Proof: Conservation of dimension

$$\boxed{\textbf{dim range}(A) + \textbf{dim null}(A) = n}$$

- using $A = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} R_{11} & R_{12} \\ 0 & 0 \end{bmatrix} P^T$, we have $x \in \textbf{null}(A)$ iff

$$x = P \begin{bmatrix} -R_{11}^{-1} R_{12} \\ I \end{bmatrix} w \qquad \text{for some } w \in \mathbb{R}^{n-r}$$

- the columns of $P \begin{bmatrix} -R_{11}^{-1} R_{12} \\ I \end{bmatrix}$ are independent, and so are a basis for $\textbf{null}(A)$
- so $\textbf{dim null}(A) = n - r$

Proof: Rank of transpose

$$\textbf{rank}(A) = \textbf{rank}(A^\top)$$

$$\begin{aligned}
\textbf{dim range}\, A^\top &= \textbf{dim}(\textbf{null}\, A)^\perp && \text{orthogonal decomposition} \\
&= n - \textbf{dim null}\, A && \text{orthogonal complements} \\
&= \textbf{dim range}\, A && \text{conservation of dimension}
\end{aligned}$$