

QR factorization

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

Gram-Schmidt procedure

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

$$\text{span}(a_1, \dots, a_r) = \text{span}(q_1, \dots, q_r) \quad \text{for } r \leq n$$

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

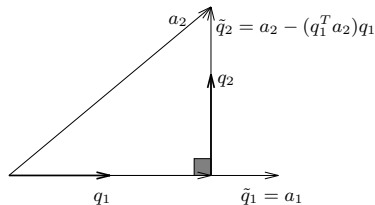
Procedure

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

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

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

(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 $A \in \mathbb{R}^{m \times n}$, $Q \in \mathbb{R}^{m \times n}$, $R \in \mathbb{R}^{n \times n}$:

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

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

$$r = 0$$

for $i = 1, \dots, n$

$$\tilde{a} = a_i - \sum_{j=1}^r q_j q_j^\top a_i$$

if $\tilde{a} \neq 0$

$$r = r + 1$$

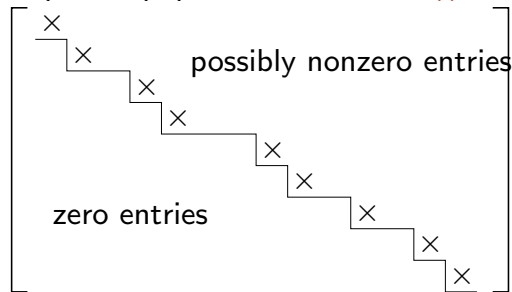
$$q_r = \tilde{a} / \|\tilde{a}\|$$

Staircase form

on exit,

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

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



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

Applications

- ▶ directly yields orthonormal basis for **range**(A)
- ▶ yields factorization $A = BC$ with $B \in \mathbb{R}^{m \times r}$, $C \in \mathbb{R}^{r \times n}$, $r = \mathbf{rank}(A)$
- ▶ 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 $p = 1, \dots, n$

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

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

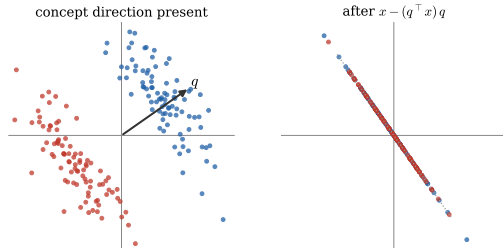
Application: removing a direction

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

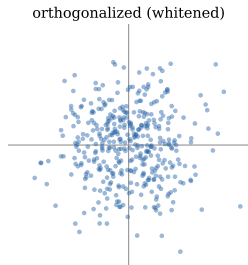
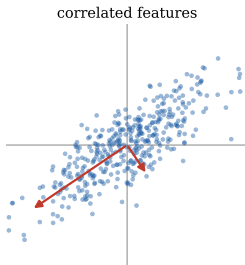
- ▶ in language models, subtracting a single direction from the internal activations can *erase a concept* (a topic, a style, or a bias)
- ▶ the readout along q is gone, while everything orthogonal to q is left untouched



Application: orthogonalizing features

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

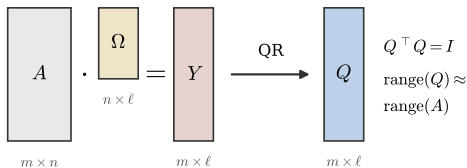
- ▶ this is *whitening*, or decorrelation, a standard pre-processing step
- ▶ nearly-dependent features make a model unstable to fit; orthogonal features are easier and better conditioned



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 $\text{range}(A)$, 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



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:

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

where:

- ▶ $Q^T Q = I$
- ▶ $R_{11} \in \mathbb{R}^{r \times r}$ 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

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

$$P = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

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

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

Orthogonal decomposition induced by A

$$\text{range}(A)^\perp = \text{null}(A^\top)$$

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

$$A^\top z = 0 \iff Q_1^\top z = 0 \iff z \in \text{range}(Q_2)$$

so $\text{range}(Q_2) = \text{null}(A^\top)$

- the columns of Q_2 are an orthonormal basis for $\text{null}(A^\top)$
- 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 $z \in \text{range}(A)$, $w \in \text{null}(A^\top)$ (we'll soon see what the vector z is ...)

Proof: Conservation of dimension

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

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

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

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

► so $\dim \text{null}(A) = n - r$

Proof: Rank of transpose

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

$$\begin{aligned}\mathbf{dim\,range}\,A^{\top} &= \mathbf{dim}(\mathbf{null}\,A)^{\perp} \\ &= n - \mathbf{dim}\,\mathbf{null}\,A \\ &= \mathbf{dim\,range}\,A\end{aligned}$$

orthogonal decomposition
orthogonal complements
conservation of dimension