Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Institute for Operations Research
Department of Mathematics

Robert Weismantel

Week 10: The Gram-Schmidt Process and the pseudoinverse of a matrix

Preparations for the Gram-Schmidt Process

Our task

Construct an orthonormal basis of a given subspace $S \subseteq \mathbb{R}^m$. The subspace is presented by a basis, i.e., vectors $a_1, \ldots, a_n$ such that $S = \mathrm{Span}(a_1, \ldots, a_n)$.

The idea for two vectors

Let a_1, a_2 be linearly independent and $S = \{a_1 x_1 + a_2 x_2 \mid x_1, x_2 \in \mathbb{R}\}$: we first normalize a_1: $q_1 = \frac{a_1}{\|a_1\|}$, then subtract from a_2 a multiple of q_1 so that it becomes orthogonal to q_1, followed by a normalization step:

$$q_2 = \frac{a_2 - (a_2^\top q_1) q_1}{\|a_2 - (a_2^\top q_1) q_1\|}. \qquad \text{Note: } a_2 - (a_2^\top q_1) q_1 \neq 0.$$

Claim: q_1, q_2 are orthogonal.

$$q_1^\top q_2 = q_1^\top \frac{a_2 - (a_2^\top q_1) q_1}{\|a_2 - (a_2^\top q_1) q_1\|} = \frac{q_1^\top a_2 - (a_2^\top q_1) q_1^\top q_1}{\|a_2 - (a_2^\top q_1) q_1\|} = \frac{0}{\|a_2 - (a_2^\top q_1) q_1\|} = 0.$$

The Gram-Schmidt Process

For more vectors:

remove from a vector a_{k+1} the projection of it on the subspace spanned by the k vectors before.

Gram-Schmidt Algorithm

Given n linearly independent vectors $a_1, \ldots, a_n$ that span a subspace S, the Gram-Schmidt process constructs $q_1, \ldots q_n$ in the following way:

- $q_1 = \frac{a_1}{\|a_1\|}$.
- For $k = 2, \ldots, n$ do
  - $q'_k = a_k - \sum_{i=1}^{k-1} (a_k^\top q_i) q_i$
  - $q_k = \frac{q'_k}{\|q'_k\|}$.

Theorem (Correctness of Gram-Schmidt)

Given n linearly independent vectors $a_1, \ldots, a_n$, the Gram-Schmidt process outputs an orthonormal basis for the span of $a_1, \ldots, a_n$.

Proof by induction

Let S_k be the subspace spanned by $a_1, \dots, a_k$. Then $S = S_n$.

Claim: $q_1, \dots, q_k$ are an orthonormal basis for S_k

It is enough to show that $q_1, \dots, q_k \in S_k$ and are orthonormal. (orthonormality implies linearly independence and S_k has dimension k)

The steps

1. Base case: $\|q_1\| = 1$ and q_1 is a multiple of a_1 and so $q_1 \in S_1$.
2. Assume the hypothesis for $i = 1, \dots k-1$:
   - Since a_k is linearly independent from the other original vectors it is not in S_{k-1} and so $q'_k \neq 0$. Thus $\|q_k\| = 1$.
   - By construction $a_k \in S_k$ and so $q_k \in S_k$.
   - Let $1 \le j \le k-1$. Since $q_1, \dots, q_{k-1}$ are orthonormal, we have

$$q_j^\top \left(a_k - \sum_{i=1}^{k-1} (a_k^\top q_i) q_i \right) = q_j^\top a_k - \sum_{i=1}^{k-1} (a_k^\top q_i) q_j^\top q_i = q_j^\top a_k - (a_k^\top q_j) = 0,$$

and $q_j^\top q_k = \frac{1}{\|q'_k\|} q_j^\top q'_k = 0$.

A first application of the Gram-Schmidt Process

Gram-Schmidt actually provides us with a new matrix factorization.

Definition (QR decomposition)

Let A be an $m \times n$ matrix with linearly independent columns. The QR decomposition is given by

$$A = QR,$$

where Q is an $m \times n$ matrix with orthonormal columns returned by the Gram Schmidt Algorithm and R is an upper triangular matrix given by $R = Q^\top A$.

It requires us to show that indeed this is a proper definition.

Lemma

The matrix R defined before is upper triangular. Moreover, R is invertible and $QQ^T A = A$.

Proof of the lemma

R is upper triangular

- We have that $Q^T Q = I$ and hence, $q_k^T q_i = 0$ for all $i = 1, \ldots k-1$.
- $q_1, \ldots, q_{k-1}$ and $a_1, \ldots, a_{k-1}$ span subspace S_{k-1}. Hence,

$$q_k^T a_i = 0 \text{ for all } i = 1, \ldots, k-1.$$

- Hence $R = Q^T A$ is upper triangular.

Moreover, $C(Q) = C(A)$

- Since $Q^T Q = I$ we obtain for the projection matrix onto the subspace $C(Q) = C(A)$ the formula $Q(Q^T Q)^{-1} Q^T = QQ^T$ and notice, for every index i,

$$\text{proj}_{S_n}(a_i) = a_i = QQ^T a_i \quad \Longleftrightarrow \quad QR = QQ^T A = A.$$

- $N(A) = \{0\}$ and since $A = QR$, we must have that $N(R) = \{0\}$. Since R is a matrix of size n by n, we notice that R is invertible.

The QR decomposition is computationally useful.

Recall $C(A) = C(Q)$

Projections on $C(A)$ can be done with Q, i.e., $\text{proj}_{C(A)}(b) = QQ^\top b$.

The least squares solution $\min \|Ax - b\|^2$:

is the point $\hat{x}$ solving the normal equations

$$A^\top A\hat{x} = A^\top b.$$

- Furthermore, $A^\top A = (QR)^\top (QR) = R^\top Q^\top QR = R^\top R$, and so we can write
$$R^\top R\hat{x} = R^\top Q^\top b. \tag{1}$$
- Since R is invertible, R^T is invertible and so we can simplify (1) to
$$R\hat{x} = Q^\top b, \tag{2}$$
which can be solved fast by back-substitution since R is triangular.

The Pseudoinverse or Moore–Penrose Inverse

Our next task

construct an analogue to the inverse of a matrix A for matrices that have no inverse. This is called the pseudoinverse and we will denote it by $A^\dagger$.

The hurdles to overcome

- For some vectors b there might not be a vector x such that $Ax = b$.
- For some vectors b there may be more than one x such that $Ax = b$ and we must pick one.
- Even if we make such choices, it is not clear that such operation will correspond to multiplying by a matrix $A^\dagger$.

Our plan to take the hurdles

- Develop a pseudoinverse for matrices with full column rank.
- Develop a pseudoinverse for matrices with full row rank.
- Write a general matrix as as product of two matrices: one of full column rank and one of full row rank.

Pseudoinverse for matrices with full column rank

The intuition

If the columns of A are linearly independent it makes sense to build $A^{\dagger}$ such that $A^{\dagger}b$ is the Least Squares Solution $\hat{x} = (A^{\top}A)^{-1}A^{\top}b$ (the vector $\hat{x}$ such that $A\hat{x}$ is as close as possible to b).

Definition (Pseudoinverse for matrices with full column rank)

For $A \in \mathbb{R}^{m \times n}$ with $\text{rank}(A) = n$ we define the pseudo-inverse $A^{\dagger} \in \mathbb{R}^{n \times m}$ of A as

$$A^{\dagger} = (A^{\top}A)^{-1}A^{\top}.$$

Proposition

- For $A \in \mathbb{R}^{m \times n}$ with $\text{rank}(A) = n$, the pseudoinverse $A^{\dagger}$ is a left inverse of A, meaning that $A^{\dagger}A = I$.
- Proof. $\text{rank}(A) = n$, $A^{\top}A$ is invertible. Hence, $A^{\dagger}A = (A^{\top}A)^{-1}A^{\top}A = I$.

Pseudoinverse for matrices with full row rank

The intuition

If the rows of A are linearly independent, then A^T has full column rank and we use the pseudo-inverse for A^T to define a pseudo-inverse of A.

Definition (Pseudoinverse for matrices with full row rank)

For $A \in \mathbb{R}^{m \times n}$ with $\mathrm{rank}(A) = m$ we define the pseudo-inverse $A^\dagger \in \mathbb{R}^{n \times m}$ of A as

$$A^\dagger = A^\top (AA^\top)^{-1}.$$

Proposition

- For $A \in \mathbb{R}^{m \times n}$ with $\mathrm{rank}(A) = m$, the pseudoinverse $A^\dagger$ is a right inverse of A, meaning that $AA^\dagger = I$.
- Proof. $\mathrm{rank}(A) = m$, $AA^\top$ is invertible. Hence, $AA^\dagger = AA^\top(AA^\top)^{-1} = I$.

What do we achieve with the pseudo-inverse here?

Since A is full row rank, for all $b \in \mathbb{R}^m$, there exists $x \in \mathbb{R}^n$ such that $Ax = b$. There are many such vectors. Choose one with smallest norm.

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & \|x\|^2 \\ s.t. \quad & Ax = b. \end{aligned} \tag{3}$$

Lemma

For a full row rank matrix A, the (unique) solution to (3) *is given by the vector $\hat{x} \in C(A^\top)$ that satisfies the constraint $A\hat{x} = b$.*

Claim: $\hat{x} = A^\dagger b$ is the solution to (3).

Proof follows from the lemma by noting that

$$A\hat{x} = AA^\dagger b = AA^\top (AA^\top)^{-1} b = b \text{ and hence, } A\hat{x} = b.$$

$\hat{x} \in C(A^\top)$

$$\hat{x} = A^\dagger b = A^\top \left((AA^\top)^{-1} b \right).$$

Proof of the lemma

A solution to (3) is equivalent to

Let x_1 be a vector such that $Ax_1 = b$. The set of solutions to $Ax = b$ are $\{x_1 + y \mid y \in N(A)\}$. Minimize $\|x_1 + y\|$ among all vectors $y \in N(A)$.

$x_1 - \text{proj}_{N(A)}(x_1)$ is the solution to (3).

- $x_1 = \left(x_1 - \text{proj}_{N(A)}(x_1)\right) + \text{proj}_{N(A)}(x_1)$. Since $y \in N(A)$ we have that $\left(x_1 - \text{proj}_{N(A)}(x_1)\right) \perp \left(y + \text{proj}_{N(A)}(x_1)\right)$ and so
- $$\|x_1 + y\|^2 = \left\|\left(x_1 - \text{proj}_{N(A)}(x_1)\right) + \text{proj}_{N(A)}(x_1) + y\right\|^2$$
  $$= \left\|x_1 - \text{proj}_{N(A)}(x_1)\right\|^2 + \left\|\text{proj}_{N(A)}(x_1) + y\right\|^2 \geq \left\|x_1 - \text{proj}_{N(A)}(x_1)\right\|^2.$$

Pseudoinverse for matrices in general

Finally, $x_1 - \text{proj}_{N(A)}(x_1)$ is orthogonal to $N(A)$.

Since $N(A)^\perp = C(A^\top)$, we observe that $x_1 - \text{proj}_{N(A)}(x_1) \in C(A^\top)$.

The idea based on the CR decomposition:

The CR decomposition writes $A = CR$ where $C \in \mathbb{R}^{m \times r}$ has the first r linearly independent columns of A and $R \in \mathbb{R}^{r \times n}$ is upper triangular. Note that C is full column rank and R is full row rank.

Definition (Pseudoinverse for all matrices)

For $A \in \mathbb{R}^{m \times n}$, with $\text{rank}(A) = r$, with CR decomposition $A = CR$ we define the pseudoinverse $A^\dagger$ as

$$A^\dagger = R^\dagger C^\dagger,$$

$$A^\dagger = R^\top \left(RR^\top\right)^{-1} \left(C^\top C\right)^{-1} C^\top = R^\top \left(C^\top CRR^\top\right)^{-1} C^\top = R^\top \left(C^\top AR^\top\right)^{-1} C^\top.$$

What does a pseudoinverse for matrices give us?

Lemma

For $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^n$, the (unique) solution to $(*)$ is given by $\hat{x} = A^\dagger b$.

$$(*) \qquad \min \left\{ \|x\|^2 \ \text{s.t.}\ x \in \mathbb{R}^n, A^\top A x = A^\top b \right\}$$

Proof.

- Let r be the rank of A and $A = CR$ with $C \in \mathbb{R}^{m \times r}$ and $R \in \mathbb{R}^{r \times n}$.
- Then $\hat{x} = A^\dagger b = R^\top \left(C^\top A R^\top\right)^{-1} C^\top b$. Thus,

$$\begin{aligned} A^\top A \hat{x} &= A^\top A R^\top \left(C^\top A R^\top\right)^{-1} C^\top b \\ &= R^\top C^\top A R^\top \left(C^\top A R^\top\right)^{-1} C^\top b = R^\top C^\top b = A^\top b. \end{aligned}$$

- Hence we have verified that $\hat{x}$ satisfies the normal equations.
- $C(A^T A) = C(A^T) = C(R^T)$ and since $\hat{x} = R^\top \left(C^\top A R^\top\right)^{-1} C^\top b$, we have verified that $\hat{x} \in C(A^\top A)$. The result follows with the previous lemma.

A few properties of the pseudo-inverse

Theorem (Let $A \in \mathbb{R}^{m\times n}$.)

1. *$AA^\dagger A = A$ and $A^\dagger AA^\dagger = A^\dagger$.*
2. *$AA^\dagger$ is symmetric. It is the projection matrix for projection on $C(A)$,*
3. *$A^\dagger A$ is symmetric. It is the projection matrix for projection on $C(A^\top)$.*
4. *$\left(A^\top\right)^\dagger = \left(A^\dagger\right)^\top$.*

Proof.

- Let us plug in $A^\dagger = R^\top\left(C^\top AR^\top\right)^{-1}C^\top$ to calculate $AA^\dagger A = CRR^T(C^TCRR^T)^{-1}C^TCR = CRR^T(RR^T)^{-1}(C^TC)^{-1}C^TCR = CR = A$.
- $AA^\dagger$ is symmetric because

$$CRR^T(RR^T)^{-1}(C^TC)^{-1}C^T = C(C^TC)^{-1}C^T = \left(C(C^TC)^{-1}C^T\right)^T = (AA^\dagger)^T$$

- The columns of C are a basis of $C(A)$. Hence, $AA^\dagger = C(C^TC)^{-1}C^T$ is the projection matrix for projecting onto $C(A)$.