Chapter 5: Inner product spaces

1. In Euclidean geometry, we have notions of length of a vector, angle between vectors, projection of a point on a plane.
2. Using the concept of inner product of two vectors which is analogous to the standard dot product of vectors in $\mathbb{R}^n$, we can introduce these geometric concepts.
3. We shall then use these ideas to solve some practical problems related to data and curve fitting.
4. **Notation.** We shall use $\mathbb{F}$ for $\mathbb{R}$ or $\mathbb{C}$. Given $a \in \mathbb{F}$, we write $\overline{a}$ for the complex conjugate of a.
5. Given a matrix A over $\mathbb{F}$ we denote by A^* the **conjugate transpose** of A, i.e., if $A = (a_{ij})$ then $A^* = (\overline{a_{ji}})$.

Inner product of vectors

1. **Definition.** Let V be a vector space over $\mathbb{F}$. An **inner product** on V is a rule which to any ordered pair of elements (u, v) of V associates a scalar, denoted by $\langle u,\ v\rangle$ satisfying the following axioms:

   for all u, v, w in V and c any scalar we have

   i. $\langle u,\ v\rangle = \overline{\langle v,\ u\rangle}$ (Hermitian property or conjugate symmetry)
   ii. $\langle u,\ v+w\rangle = \langle u,\ v\rangle + \langle u,\ w\rangle$ (additivity)
   iii. $\langle u,\ cv\rangle = c\langle u,\ v\rangle$ (homogeneity)
   iv. $\langle v,\ v\rangle \geq 0$ with $\langle v,\ v\rangle = 0 \iff v = 0$ (positive definite).

2. An **inner product space** is a vector space with an inner product.
3. **Remark**. Note that $\langle v,\ v\rangle$ is always real and $\langle cu,\ v\rangle = \bar{c}\langle u,\ v\rangle$. (why?)
4. **Example.** (1) Let $v = (x_1, x_2, \ldots, x_n)^t,\ w = (y_1, y_2, \ldots, y_n)^t \in \mathbb{R}^n$.
5. The **standard inner product** on $\mathbb{R}^n$ is defined as

$$\langle v,\ w\rangle = v^t w = \sum_{i=1}^{n} x_i y_i.$$

Examples of inner products

1. **Example** (2) Let $v = (x_1, x_2, \ldots, x_n)^t,\ w = (y_1, y_2, \ldots, y_n)^t \in \mathbb{C}^n$.
2. The **standard inner product** on $\mathbb{C}^n$ is defined as
$$\langle v,\, w\rangle = v^* w = \sum_{i=1}^{n} \overline{x_i} y_i.$$
3. **Notation.** When we consider $\mathbb{C}^1$ as an inner product space with the standard inner product as defined in the last example, for $z = x + iy \in \mathbb{C}^1$, we write $|z| := \sqrt{\langle z,\, z\rangle} = \sqrt{\overline{z}z} = \sqrt{(x - iy)(x + iy)} = \sqrt{x^2 + y^2}$ as usual.
4. **Example** (3) Let $V = \mathcal{C}[0, 1]$ be the vector space of all real valued continuous functions on the unit interval $[0, 1]$.
5. For $f, g \in V$, define
$$\langle f,\, g\rangle = \int_0^1 f(t)g(t)dt.$$
6. Simple properties of the integral show that $\langle f,\, g\rangle$ is an inner product on $\mathcal{C}[0, 1]$.

Pythagoras Theorem and parallelogram law

1. **Definition.** Given an inner product space V and $v \in V$ we define its **length** or **norm** by $\|v\| = \sqrt{\langle v, v \rangle}$ and v is a **unit vector** if $\|v\| = 1$.
2. Elements v, w of V are said to be **orthogonal** or **perpendicular** if $\langle v, w \rangle = 0$. We write this as $v \perp w$.
3. **Remark.** If $c \in \mathbb{F}, v \in V$ then $\|cv\| = \sqrt{\langle cv, cv \rangle} = \sqrt{\bar{c}c\langle v, v \rangle} = |c|\|v\|$.
4. **Theorem. (Pythagoras)** If $v \perp w$, then $\|v + w\|^2 = \|v\|^2 + \|w\|^2$.
5. **Proof.** We have

$$\|v + w\|^2 = \langle v + w, v + w \rangle = \langle v, v \rangle + \langle w, w \rangle = \|v\|^2 + \|w\|^2. \quad \square$$

6. **Exercise.** *Prove the Parallelogram law:* If $v, w \in V$, then

$$\|v + w\|^2 + \|v - w\|^2 = 2\|v\|^2 + 2\|w\|^2.$$

Projection of a vector onto another vector

1. **Definition.** Let $v, w \in V$ with $w \neq 0$. We define

$$p_w(v) = \frac{\langle w,\ v\rangle}{\langle w,\ w\rangle} w$$

to be the **projection of** v **along** w.

2. Note that the map $p_w : V \to V$ given by $v \mapsto p_w(v)$ is a linear map. (why?)

3. **Proposition.** Let $v, w \in V$ with $w \neq 0$. Then
   (a). $p_w(v) = p_{\frac{w}{\|w\|}}(v)$, i.e., the projection of v along w is same as the projection of v along the unit vector in the direction of w.

   (b). $p_w(v)$ and $v - p_w(v)$ are orthogonal.

   (c). $\|p_w(v)\| \leq \|v\|$ with equality iff $\{v, w\}$ are linearly dependent.

4. **Proof.** (a). We have

$$p_w(v) = \frac{\langle w,\ v\rangle}{\langle w,\ w\rangle} w = \frac{\langle w,\ v\rangle}{\|w\|^2} w = \langle \frac{w}{\|w\|},\ v\rangle \frac{w}{\|w\|} = p_{\frac{w}{\|w\|}}(v).$$

Projection of a vector onto another vector

1. (b). In view of part (a) we may assume that w is a unit vector. So

$$\begin{aligned}
\langle p_w(v),\, v - p_w(v)\rangle &= \langle p_w(v),\, v\rangle - \langle p_w(v),\, p_w(v)\rangle \\
&= \langle\langle w,\, v\rangle w,\, v\rangle - \langle\langle w,\, v\rangle w,\, \langle w,\, v\rangle w\rangle \\
&= \overline{\langle w,\, v\rangle}\langle w,\, v\rangle - \overline{\langle w,\, v\rangle}\langle w,\, v\rangle\langle w,\, w\rangle \\
&= 0 \qquad (\text{since } \|w\| = 1)
\end{aligned}$$

2. (c).

$$\begin{aligned}
\|v\|^2 &= \langle v,\, v\rangle \\
&= \langle p_w(v) + v - p_w(v),\, p_w(v) + v - p_w(v)\rangle \\
&= \|p_w(v)\|^2 + \|v - p_w(v)\|^2 \qquad (\text{since } p_w(v) \perp v - p_w(v)) \\
&\geq \|p_w(v)\|^2.
\end{aligned}$$

3. Clearly, there is equality in the last step $\iff$ $v = p_w(v) = \frac{<w,v>}{<w,w>} w$. $\square$

Cauchy-Schwarz inequality

1. **Theorem** (Cauchy-Schwarz inequality). For $v, w \in V$

$$|\langle w,\ v\rangle| \leq \|w\|\|v\|,$$

   with equality $\iff$ $\{v, w\}$ are linearly dependent.

2. **Proof.** The result is clear if $w = 0$. So we may assume that $w \neq 0$.

3. Case (i): w is a unit vector. In this case the LHS of the C-S inequality is $\|p_w(v)\|$ and the result follows from part (c) of the previous proposition.

4. Case (ii): w is not a unit vector. Set $u = \frac{w}{\|w\|}$.

5. Then we have

$$|\langle w,\ v\rangle| = \|w\|(|\langle \frac{w}{\|w\|},\ v\rangle|) = \|w\||\langle u,\ v\rangle| \text{ and } \|w\|\|v\| = \|w\|(\|u\|\|v\|).$$

6. The result now follows as $|\langle u,\ v\rangle| \leq \|u\|\|v\|$ by Case (i). $\square$

Triangle inequality

1. **Theorem** (Triangle Inequality). For $v, w \in V$
$$\|v + w\| \leq \|v\| + \|w\|.$$

2. **Proof.** We have
$$\begin{aligned}
\|v+w\|^2 &= \langle v+w,\ v+w\rangle \\
&= \langle v,\ v\rangle + \langle v,\ w\rangle + \langle w,\ v\rangle + \langle w,\ w\rangle \\
&= \langle v,\ v\rangle + \langle v,\ w\rangle + \overline{\langle v,\ w\rangle} + \langle w,\ w\rangle \\
&= \langle v,\ v\rangle + 2\mathrm{Re}\langle v,\ w\rangle + \langle w,\ w\rangle \\
&\leq \|v\|^2 + 2|\langle v,\ w\rangle| + \|w\|^2 \qquad (\text{since } x \leq |x+iy| \text{ for } x, y \in \mathbb{R}) \\
&\leq \|v\|^2 + 2\|v\|\|w\| + \|w\|^2 \qquad (\text{using C-S inequality}) \\
&= (\|v\| + \|w\|)^2.
\end{aligned}$$

3. Thus $\|v + w\| \leq \|v\| + \|w\|$. $\square$

Angle and distance between vectors

1. **Definition**. Let V be a real inner product space. Given $v, w \in V$ with $v, w \neq 0$, by C-S inequality
$$-1 \leq \frac{\langle v,\ w\rangle}{\|v\|\|w\|} \leq 1.$$
2. So, there is a unique $0 \leq \theta \leq \pi$ satisfying $cos(\theta) = \frac{\langle v,\, w\rangle}{\|v\|\|w\|}$.
3. This θ is the **angle** between v and w.
4. The **distance** between u and v in V is defined as $d(u, v) = \|u - v\|$.
5. **Proposition.** Let $u, v, w \in V$. Then
   i. $d(u, v) \geq 0$ with equality iff $u = v$
   ii. $d(u, v) = d(v, u)$
   iii. $d(u, v) \leq d(u, w) + d(w, v)$.
6. **Proof**. Exercise.

Orthonormal bases

1. **Definition.** Let V be an n-dimensional inner product space. A basis $\{v_1, v_2, \ldots, v_n\}$ of V is called **orthogonal** if its elements are mutually perpendicular, i.e., if $\langle v_i, v_j \rangle = 0$ for $i \neq j$. If, in addition, $\|v_i\| = 1$, for all i, we say that the basis is **orthonormal**.
2. **Example**. The set $\{e_1, \ldots, e_n\}$ is an orthonormal basis of $\mathbb{F}^n$ with the standard inner product.
3. **Proposition**. Let $U = \{u_1, u_2, \ldots, u_n\}$ be a set of nonzero vectors in an inner product space V. If $\langle u_i, u_j \rangle = 0$ for $i \neq j, 1 \leq i, j \leq n$, then U is linearly independent.
4. **Proof**. Suppose $c_1, c_2, \ldots, c_n$ are scalars with
$$c_1 u_1 + c_2 u_2 + \ldots + c_n u_n = 0.$$
5. Take inner product with u_i on both sides to get $c_i \langle u_i, u_i \rangle = 0$.
6. Since $u_i \neq 0$, we get $c_i = 0$.
7. Thus U is linearly independent. $\square$

Orthonormal bases and Gram-Schmidt process

1. **Theorem**(Gram-Schmidt process). Let V be a finite dimensional inner product space. Let $W \subseteq V$ be a subspace and let $\{w_1, \ldots, w_m\}$ be an orthogonal basis of W.

   If $W \neq V$, then there exist elements $w_{m+1}, \ldots, w_n$ of V such that $\{w_1, \ldots, w_n\}$ is an orthogonal basis of V.
2. **Remark.** Taking $W = L(\{v\})$ for some nonzero $v \in V$, we see that V has an orthogonal, and hence orthonormal, basis.
3. **Proof of the theorem**. The method of proof is as important as the theorem and is called the **Gram-Schmidt orthogonalization process**.
4. Since $W \neq V$, we can find a vector v_{m+1} such that $\{w_1, \ldots, w_m, v_{m+1}\}$ is linearly independent.
5. We take v_{m+1} and subtract from it its projections along $w_1, \ldots, w_m$.
6. Recall that $p_w(v) = \frac{\langle w, v \rangle}{\langle w, w \rangle} w$.
7. Define $w_{m+1} = v_{m+1} - p_{w_1}(v_{m+1}) - p_{w_2}(v_{m+1}) - \cdots - p_{w_m}(v_{m+1})$.
8. Clearly, $w_{m+1} \neq 0$ as otherwise $\{w_1, \ldots, w_m, v_{m+1}\}$ would be linearly dependent.

Gram-Schmidt orthogonalization process

1. We now check that $\{w_1, \ldots, w_{m+1}\}$ is orthogonal.
2. For this, we show that $w_{m+1} \perp w_i$ for $i = 1, 2, \ldots, m$.
3. For $i = 1, 2, \ldots, m$, we have

$$
\begin{aligned}
\langle w_i,\, w_{m+1}\rangle &= \langle w_i,\, v_{m+1} - \sum_{j=1}^{m} p_{w_j}(v_{m+1})\rangle \\
&= \langle w_i,\, v_{m+1}\rangle - \langle w_i,\, \sum_{j=1}^{m} p_{w_j}(v_{m+1})\rangle \\
&= \langle w_i,\, v_{m+1}\rangle - \langle w_i,\, p_{w_i}(v_{m+1})\rangle \text{ (since } \langle w_i,\, w_j\rangle = 0 \text{ for } i \neq j) \\
&= \langle w_i,\, v_{m+1}\rangle - \langle w_i,\, \frac{\langle w_i,\, v_{m+1}\rangle}{\|w_i\|^2} w_i\rangle \\
&= \langle w_i,\, v_{m+1}\rangle - \langle w_i,\, v_{m+1}\rangle = 0.
\end{aligned}
$$

4. **Example**. Let $V = P_3[-1, 1]$ denote the real vector space of polynomials of degree atmost 3 defined on $[-1, 1]$. Note that V is an inner product space under the inner product $\langle f,\, g\rangle = \int_{-1}^{1} f(t)g(t)dt$.

An example for the Gram-Schmidt process

1. We will find an orthogonal basis $\{w_1, w_2, w_3, w_4\}$ of V.
2. For, we begin with the basis $\{1, x, x^2, x^3\}$ of V. Set $w_1 = 1$. Then

$$
\begin{aligned}
w_2 &= x - \frac{\langle x,\, 1\rangle}{\|1\|^2}1 \qquad \text{(what is } ||1||\text{?)} \\
&= x - \frac{1}{2}\int_{-1}^{1} t\,dt = x, \\
w_3 &= x^2 - \langle x^2,\, 1\rangle\frac{1}{2} - \langle x^2,\, x\rangle\frac{x}{(2/3)} \\
&= x^2 - \frac{1}{2}\int_{-1}^{1} t^2\,dt - \frac{3}{2}x\int_{-1}^{1} t^3\,dt \\
&= x^2 - \frac{1}{3}, \\
w_4 &= x^3 - \langle x^3,\, 1\rangle\frac{1}{2} - \langle x^3,\, x\rangle\frac{x}{(2/3)} - \langle x^3,\, x^2 - \frac{1}{3}\rangle\frac{x^2 - \frac{1}{3}}{(2/5)} \\
&= x^3 - \frac{3}{5}x.
\end{aligned}
$$

Subspace and its orthogonal subspace

1. Let V be a finite dimensional inner product space. We have seen how to project a vector onto a nonzero vector.
2. We now discuss the orthogonal projection of a vector onto a subspace.
3. Let W be a subspace of V. Define

$$W^{\perp} = \{u \in V \mid u \perp w \text{ for all } w \in W\}.$$

4. Check that $W^{\perp}$ is a subspace of V and $W \cap W^{\perp} = \{0\}$.
5. The subspace $W^{\perp}$ is called the **orthogonal complement** of W in V.
6. Note that for subspaces W_1 and W_2 of a vector space V, $W_1 \oplus W_2$ is the notation for $W_1 + W_2 = L(W_1 \cup W_2)$ when $W_1 \cap W_2 = \{0\}$.
7. **Theorem**. Every $v \in V$ can be written uniquely as $v = x + y$, where $x \in W$ and $y \in W^{\perp}$ (i.e., $V = W \oplus W^{\perp}$). Moreover $\dim V = \dim W + \dim W^{\perp}$.
8. **Proof**. Let $\{v_1, v_2, \ldots, v_k\}$ be an orthonormal basis of W. For $v \in V$, set

$$x = \langle v_1, v \rangle v_1 + \langle v_2, v \rangle v_2 + \cdots + \langle v_k, v \rangle v_k$$

and put $y = v - x$.

Subspace and its orthogonal subspace

1. Clearly $v = x + y$ and $x \in W$. We now check that $y \in W^\perp$.
2. For $i = 1, 2, \ldots, k$, we have

$$\begin{aligned}
\langle y, v_i \rangle &= \langle v - x, v_i \rangle \\
&= \langle v, v_i \rangle - \langle x, v_i \rangle \\
&= \langle v, v_i \rangle - \langle \sum_{j=1}^{k} \langle v_j, v \rangle v_j, v_i \rangle \\
&= \langle v, v_i \rangle - \sum_{j=1}^{k} \overline{\langle v_j, v \rangle} \langle v_j, v_i \rangle \\
&= \langle v, v_i \rangle - \langle v, v_i \rangle = 0
\end{aligned}$$

3. It follows that $y \in W^\perp$. For uniqueness, let $v = x + y = x' + y'$, where $x, x' \in W$ and $y, y' \in W^\perp$.
4. Then $x - x' = y' - y \in W \cap W^\perp$. But $W \cap W^\perp = \{0\}$. Hence $x = x'$ and $y = y'$. □

Orthogonal projection of a vector onto a subspace

1. **Definition**. For a subspace W, we define a function $p_W : V \to W$ as follows: given $v \in V$, write $v = x + y$, where $x \in W$ and $y \in W^\perp$. The **orthogonal projection** of v onto W is defined to be $p_W(v) = x$.

2. Notice that $v - p_W(v) \in W^\perp$. Notice also that the map p_W is linear (why?).

3. Let W be a subspace of V and let $v \in V$. A **best approximation** to v by vectors in W is a vector w in W such that

$$\| v - w \| \leq \| v - u \|, \ \text{ for all } u \in W.$$

4. The next result shows that the orthogonal projection of v in W gives the unique best approximation to v by vectors in W.

5. **Theorem**. Let $v \in V$ and let W be a subspace of V. Then $p_W(v)$ is the best approximation to v by vectors in W.

Best approximation of a vector in a subspace

1. **Proof**. For any $w \in W$, we have

$$\begin{aligned} \| v - w \|^2 &= \| v - p_W(v) + p_W(v) - w \|^2 \\ &= \| v - p_W(v) \|^2 + \| p_W(v) - w \|^2 \quad \text{(why?)} \\ & \text{(since } v - p_W(v) \in W^{\perp}\text{)} \\ &\geq \| v - p_W(v) \|^2. \end{aligned}$$

2. Therefore $p_W(v)$ is a best approximation to v in W.

3. If $u \in W$ is another best approximation to v, then

$$\begin{aligned} \| v - u \|^2 &= \| v - p_W(v) + p_W(v) - u \|^2 \\ &= \| v - p_W(v) \|^2 + \| p_W(v) - u \|^2 \\ &\geq \| v - p_W(v) \|^2 \\ &\geq \| v - u \|^2 \end{aligned}$$

4. Therefore $\| p_W(v) - u \|^2 = 0$ and hence $u = p_W(v)$. $\square$

Best approximation of a vector in $C(A)$.

1. Consider $\mathbb{R}^n$ with the standard inner product.
2. Let A be an $n \times m$ $(m \leq n)$ matrix and let $b \in \mathbb{R}^n$.
3. We want to project $b \in \mathbb{R}^n$ onto the column space of A.
4. The vector $p = P_{C(A)}(b)$ will be of the form $p = Ax$ for some $x \in \mathbb{R}^m$.
5. We now know that $p = Ax$ is the orthogonal projection of b on $C(A)$ iff $b - Ax$ is orthogonal to every column of A (why?).
6. In other words, x should satisfy the **normal equations:**

$$A^t(b - Ax) = 0 \iff A^t Ax = A^t b.$$

7. Thus, if x is any solution of the normal equations, then $Ax = p_{C(A)}(b)$.
8. **Proposition.** rank $(A) =$ rank $(A^t A)$.
9. **Proof.** We have rank $(A) \geq$ rank $(A^t A)$. (why?)
10. We now find the inequality for the corresponding nullity.

Normal equations for best approximation

1. For, let $z \in \mathcal{N}(A^tA)$. Then $A^tAz = 0$, that is, $A^tw = 0$ for $w = Az$.
2. Hence $w \in C(A) \cap C(A)^{\perp}$. Therefore $w = 0$ and it shows that $z \in \mathcal{N}(A)$ and hence $N(A^tA) \subseteq N(A)$.
3. Therefore nullity $(A) \geq$ nullity (A^tA) and this implies rank $(A) \leq$ rank (A^tA). (why?)
4. It follows from the two above inequalities that rank $(A^tA) =$ rank (A). $\square$
5. **Remark.** If the columns of A are linearly independent, i.e., rank $(A) = m$, the (unique) solution to the normal equations $A^tAx = A^tb$ is $x = (A^tA)^{-1}A^tb$ (why?) and the projection of b onto the column space of A is $A(A^tA)^{-1}A^tb$.
6. Note that the normal equations always have a solution (since rank $A^tA \leq$ rank $[A^tA : A^tb] =$ rank $A^t[A : b] \leq$ rank $A^t =$ rank $A =$ rank A^tA and this implies that rank $A^tA =$ rank $[A^tA : A^tb]$),

   although the solution will not be unique in case the columns of A are linearly dependent (since rank $(A^tA) =$ rank $(A) < m$ and A^tA is an $m \times m$ matrix).
7. However, for any two solutions x_1 and x_2 of the normal equations, as proved earlier the best approximation is unique and hence $Ax_1 = Ax_2$.

Normal equations for best approximation

1. Let $A = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$ and $b = \begin{bmatrix} 1 \\ 0 \\ -5 \end{bmatrix}$.

2. Then $A^t A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$ and $A^t b = \begin{bmatrix} 1 \\ -4 \end{bmatrix}$.

3. The unique solution to the normal equations $A^t Ax = A^t b$ is

$$x = \begin{bmatrix} 2 \\ -3 \end{bmatrix} \text{ and } b - Ax = \begin{bmatrix} 2 \\ -2 \\ -2 \end{bmatrix}.$$

4. Note that this vector is orthogonal to the columns of A.

5. The projection of b onto $C(A)$ is $p = Ax = \begin{bmatrix} -1 \\ 2 \\ -3 \end{bmatrix}$.

Least squares approximation

1. Suppose we have a large number of data points $(x_i, y_i) \quad i = 1, 2, \dots, n$, collected from some experiment.
2. Sometime we believe that these points lie on a straight line.
3. So a linear function $y(x) = s + tx$ may satisfy
$$y(x_i) = y_i,\ i = 1, \dots, n.$$
4. Due to uncertainty in data and experimental error, in practice the points will deviate somewhat from a straight line and so it is impossible to find a linear $y(x)$ that passes through all of them.
5. So we seek a line that fits the data well, in the sense that the errors are made as small as possible.
6. A natural question that arises now is: how do we define the error?

Least squares approximation

1. Consider the following system of linear equations, in the variables s and t, and known coefficients $x_i, y_i,\ i = 1, \ldots, n$:

$$\begin{aligned} s + x_1 t &= y_1 \\ s + x_2 t &= y_2 \\ &\ \ . \\ &\ \ . \\ s + x_n t &= y_n \end{aligned}$$

2. Note that typically n would be much greater than 2. If we can find s and t to satisfy all these equations, then we have solved our problem.

3. However, for reasons mentioned above, this is not always possible.

Least squares approximation

1. For given s and t, the error in the ith equation is $|y_i - s - x_i t|$.
2. There are several ways of combining the errors in the individual equations to get a measure of the total error.
3. The following are three examples:

$$\sqrt{\sum_{i=1}^{n}(y_i - s - x_i t)^2}, \quad \sum_{i=1}^{n}|y_i - s - x_i t|, \quad \max{}_{1\leq i\leq n} |y_i - s - x_i t|.$$

4. Both analytically and computationally, a nice theory exists for the first of these choices and this is what we shall study.
5. The problem of finding s, t so as to minimize $\sqrt{\sum_{i=1}^{n}(y_i - s - x_i t)^2}$

   is called a **least squares problem**.

Least squares approximation

1. Suppose that

$$A = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ . & . \\ 1 & x_n \end{bmatrix}, \ b = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix}, \ x = \begin{bmatrix} s \\ t \end{bmatrix}, \ \text{so } Ax = \begin{bmatrix} s + tx_1 \\ s + tx_2 \\ . \\ . \\ s + tx_n \end{bmatrix}.$$

2. The least squares problem is finding an x such that $||b - Ax||$ is minimized, i.e., find an x such that Ax is the best approximation to b in the column space $C(A)$ of A.

3. This is precisely the problem of finding x such that $b - Ax \in C(A)^{\perp}$.

Least squares approximation

1. **Example.** Find s, t such that the straight line $y = s + tx$ best fits the following data in the least squares sense:

$$y = 1 \text{ at } x = -1, \quad y = 1 \text{ at } x = 1, \quad y = 3 \text{ at } x = 2.$$

2. Project $b = \begin{bmatrix} 1 \\ 1 \\ 3 \end{bmatrix}$ onto the column space of $A = \begin{bmatrix} 1 & -1 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}$.

3. Now $A^t A = \begin{bmatrix} 3 & 2 \\ 2 & 6 \end{bmatrix}$ and $A^t b = \begin{bmatrix} 5 \\ 6 \end{bmatrix}$.

4. The normal equations are

$$\begin{bmatrix} 3 & 2 \\ 2 & 6 \end{bmatrix} \begin{bmatrix} s \\ t \end{bmatrix} = \begin{bmatrix} 5 \\ 6 \end{bmatrix}.$$

5. The solution is $s = 9/7, \ t = 4/7$ and hence the best line is $y = \frac{9}{7} + \frac{4}{7}x$. $\square$

Least squares approximation

1. We can also try to fit an mth degree polynomial

$$y(x) = s_0 + s_1 x + s_2 x^2 + \cdots + s_m x^m$$

to the data points $(x_i, y_i),\ i = 1, \ldots, n$, so as to minimize the error in the least squares sense.

2. In this case $s_0, s_1, \ldots, s_m$ are the variables and we have

$$A = \begin{bmatrix} 1 & x_1 & x_1^2 & . & . & x_1^m \\ 1 & x_2 & x_2^2 & . & . & x_2^m \\ & . & . & . & . & . \\ & . & . & . & . & . \\ 1 & x_n & x_n^2 & . & . & x_n^m \end{bmatrix},\ b = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix},\ x = \begin{bmatrix} s_0 \\ s_1 \\ . \\ . \\ s_m \end{bmatrix}.$$

3. Note that a straight line can be considered as a polynomial of degree 1.