# Least-Squares

Ashek Ahmmed
University of New South Wales

July 31, 2012

Let's consider A$\mathbf{x}$=$\mathbf{y}$ where A$\in \mathbf{R}^{m\times n}$. Previously I learnt about a type of problem: *find a* $\mathbf{x}$ *that gives* $\mathbf{y}$=$\mathbf{y}_{des}$ *in the output but if it is not possible to achieve then find a* $\mathbf{x}$ *for which I can get as close as possible to* $\mathbf{y}_{des}$ *in the output.* The question that arises is **why it is not possible to have y=$\mathbf{y}_{\mathbf{des}}$ at the first place?** The answer is $\mathbf{y}_{\text{des}}\notin$ C(A). Therefore no matter how hard I try, I won't be able to hit $\mathbf{y}_{\text{des}}$ by using the linear combinations of the columns of A.

## 0.1 Overdetermined Linear Equations

If $m > n$ then A is a *skinny* matrix. [If $m \gg n$ then A is called *strictly skinny*].

- this type of system is called **overdetermined** i.e. more equations than unknowns.

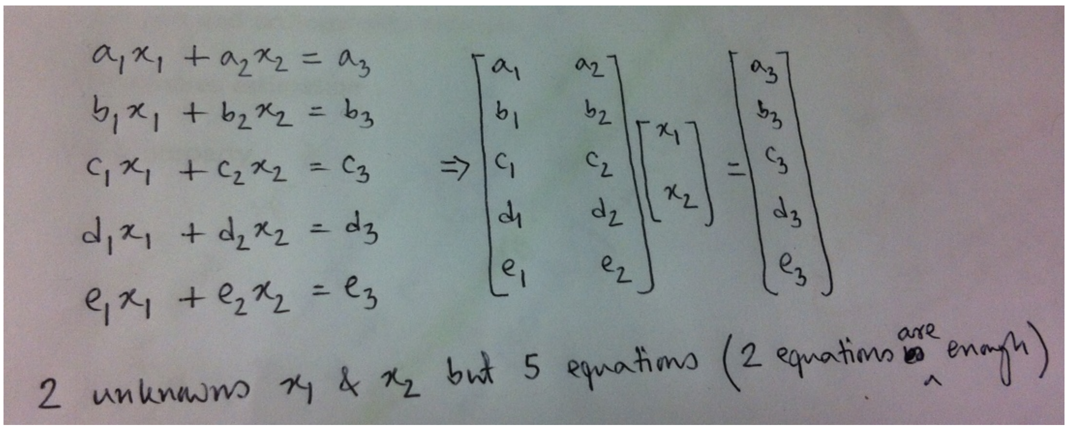


Figure 1: 2 equations would have been sufficient but 5 equations are given.

From Figure 1 it can be stated that to measure/estimate $n$ things I am using $m$ sensors. I have too many sensors in the output i.e. I am producing more than enough output readings.

- for most $\mathbf{y}$, I can not hit/achieve them by using the linear combinations of the columns of A i.e. for most $\mathbf{y}\in\mathbf{R}^m$ I am unable to find a $\mathbf{x}\in\mathbf{R}^n$ such that $\mathbf{y}$=A$\mathbf{x}$. This is because the output space $\mathbf{R}^m$ and the column space C(A) are not the same thing here rather C(A) is a tiny subspace of the output space. Recall that it is possible to hit/achieve only those vectors which are in C(A).

## 0.2 (Approximate) Solution for Overdetermined Systems

Notice that it is not possible to solve an overdetermined system. So whenever I use the term "solution" in this case it means I am talking about an approximate solution. Two things to consider here:

- it is not possible to achieve $\mathbf{y}_{\text{des}}$ so I want to get as close as possible to it.
- it is not possible to achieve any $\mathbf{y}\notin$ C(A).

Combining these two facts I have no other choice other than to set my objective as: *I want such a* $\mathbf{y}$ *that is close to* $\mathbf{y}_{des}$ *and at the same time this* $\mathbf{y}\in$ *C(A).* That means give me the closest point (recall that every vector is a point in the vector space) on C(A) from $\mathbf{y}_{\text{des}}$. To formulate things, I can state that I have to solve the following optimization problem:

$$\begin{array}{ll} \min\limits_{\mathbf{y}} & ||\mathbf{y}_{\text{des}} - \mathbf{y}||_2 \\ s.t. & \mathbf{y} \in C(A) \end{array}$$

or equivalently,

$$
\begin{aligned}
&\min_{\mathbf{x}} && ||\mathbf{y}_{\text{des}} - A\mathbf{x}||_2 \\
\Rightarrow &\min_{\mathbf{x}} && ||\mathbf{y}_{\text{des}} - A\mathbf{x}||_2^2 \\
\Rightarrow &\min_{\mathbf{x}} && \sum_{i=1}^{m}((\mathbf{y}_{\text{des}})_i - \mathbf{a}_i^T\mathbf{x})^2 \\
\Rightarrow &\min_{\mathbf{x}} && \sum_{i=1}^{m}\mathbf{r}_i^2 \\
\Rightarrow &\min_{\mathbf{x}} && ||\mathbf{r}||^2 \qquad (1)
\end{aligned}
$$

The optimization problem (1) is called the *least-squares* problem and note that it is an unconstrained problem. The solution to this optimization problem $\mathbf{x}_{\text{ls}}$ is called the *least-squares* (approximate) solution which suggests that among all $\mathbf{y}\in$ C(A), $\mathbf{y}_{\text{ls}}$=A$\mathbf{x}_{\text{ls}}$ is the closest from $\mathbf{y}_{\text{des}}$ i.e. $||\mathbf{y}_{\text{des}}$-$\mathbf{y}_{\text{ls}}||$ is minimum and $\mathbf{y}_{\text{ls}}\in$C(A) as well :)

## 0.3 Geometric Interpretation

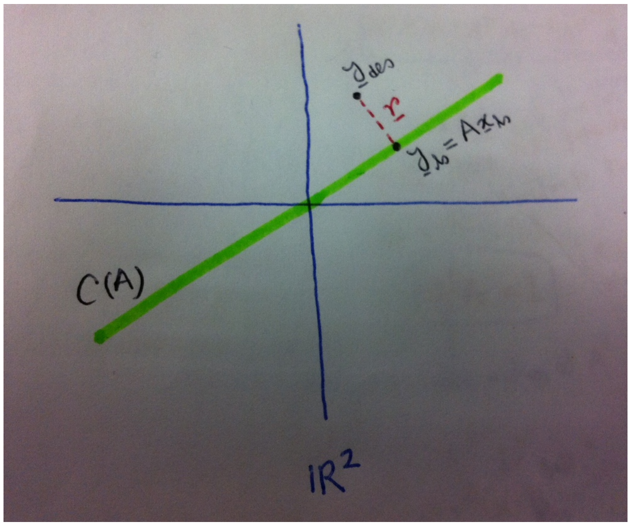


Figure 2: C(A) is a 1D subspace inside $\mathbf{R}^2$ and $\mathbf{y}_{\text{des}}\notin$ C(A).

We have seen that A$\mathbf{x}_{\text{ls}}$ is the point in C(A) closest to $\mathbf{y}_{\text{des}}$. Geometrically, A$\mathbf{x}_{\text{ls}}$ is the projection of $\mathbf{y}_{\text{des}}$ onto C(A) because the distance i.e. the norm of the residual $\mathbf{r}$ is minimum when $\mathbf{r}\perp$ C(A) and this happens only when $\mathbf{y}_{\text{ls}}$ is the projection of $\mathbf{y}_{\text{des}}$ onto C(A).

$$\mathbf{y}_{\text{ls}} = \text{P}\ \mathbf{y}_{\text{des}} \qquad (2)$$

where P is the *projection matrix* which operates on a vector and returns that vector's projection as the output.

## 0.4 Least-squares(Approximate) Solution

Let's assume that A is skinny and full rank ($= n$) i.e. A is *one-to-one* matrix. Later on we will realize the advantage of this assumption. To find $\mathbf{x}_{\text{ls}}$ we need to solve (1) i.e. find the point $\mathbf{x}$ where the optimization function f($\mathbf{x}$) of (1) attains its minimum value.

$$
\begin{aligned}
f(\mathbf{x}) &= ||\mathbf{r}||^2 = \mathbf{r}^T\mathbf{r} = (\mathbf{y}_{\text{des}} - A\mathbf{x})^T(\mathbf{y}_{\text{des}} - A\mathbf{x}) = (\mathbf{y}_{\text{des}}^T - \mathbf{x}^T A^T)(\mathbf{y}_{\text{des}} - A\mathbf{x}) \\
&= \mathbf{y}_{\text{des}}^T\mathbf{y}_{\text{des}} - \mathbf{y}_{\text{des}}^T A\mathbf{x} - \mathbf{x}^T A^T\mathbf{y}_{\text{des}} + \mathbf{x}^T A^T A\mathbf{x} \\
&= \mathbf{x}^T A^T A\mathbf{x} - 2\mathbf{y}_{\text{des}}^T(A\mathbf{x}) + ||\mathbf{y}_{\text{des}}||^2 \qquad (3)
\end{aligned}
$$

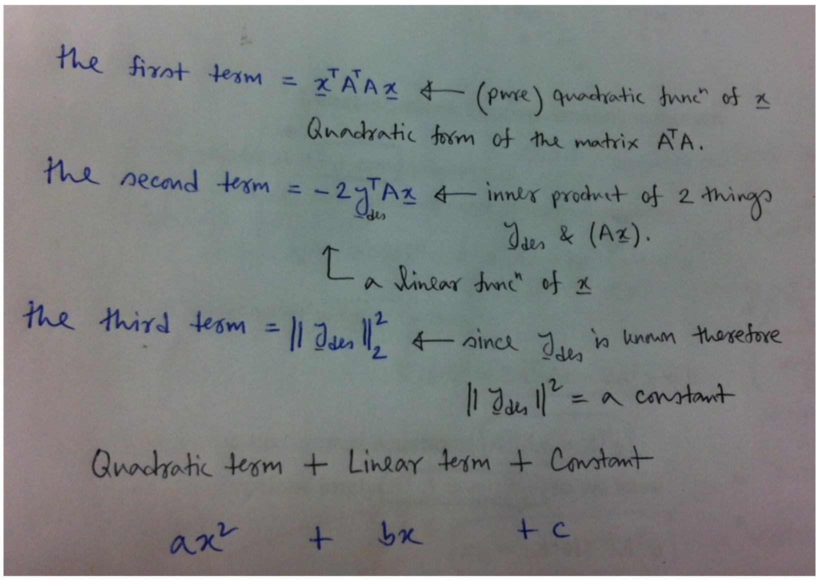


Figure 3: The objective function of (1) is a *quadratic function* in the optimization variable $\mathbf{x}$.

To find the the minimum point of the objective function (3) let's take its gradient w.r.t. $\mathbf{x}$ and set that to zero:

$$
\begin{aligned}
\nabla_{\mathbf{x}} f = 2A^T A\mathbf{x} - 2A^T\mathbf{y}_{\text{des}} = \mathbf{0} \\
\Rightarrow A^T A\mathbf{x} = A^T\mathbf{y}_{\text{des}} \qquad (4)
\end{aligned}
$$

How we ended up at (4) has been shown in Figure 4.

We made the assumption that A is full rank i.e. $rank(A) = n$. As $rank(A^T A) = rank(A) = rank(A^T)$ therefore $A^T A$ is full rank and it is square as well. Hence $A^T A$ *is invertible i.e.* $(A^T A)^{-1}$*exists.* Please note that being square is not enough to be invertible e.g. $AA^T \in \mathbf{R}^{m\times m}$ is also square but it has rank n which is less than m hence $AA^T$ is not full rank therefore not invertible . So it is the rank of a matrix that plays the most important role behind its singularity. From (4) we get our least-squares (approximate) solution $\mathbf{x}_{\text{ls}}$ as:

$$\mathbf{x}_{\text{ls}} = (A^T A)^{-1} A^T\mathbf{y}_{\text{des}} \qquad (5)$$

which is a very famous formula. So famous that in MATLAB we have an (\) operator reserved for it: $\mathbf{x}_{\text{ls}}$=A\\$\mathbf{y}_{\text{des}}$.... so simple! Finally we have found the $\mathbf{x}$ which would take us as close as possible to $\mathbf{y}_{\text{des}}$. Let's now turn our attention to (5) and analyze it.

- **$\mathbf{x}_{\text{ls}}$ is a linear function of $\mathbf{y}_{\text{des}}$.**

  this is because (5) can be re-written as: $\mathbf{x}_{\text{ls}} = B_{\text{ls}}\mathbf{y}_{\text{des}}$ which is a linear operation on $\mathbf{y}_{\text{des}}$ where $B_{\text{ls}} = (A^T A)^{-1} A^T$.

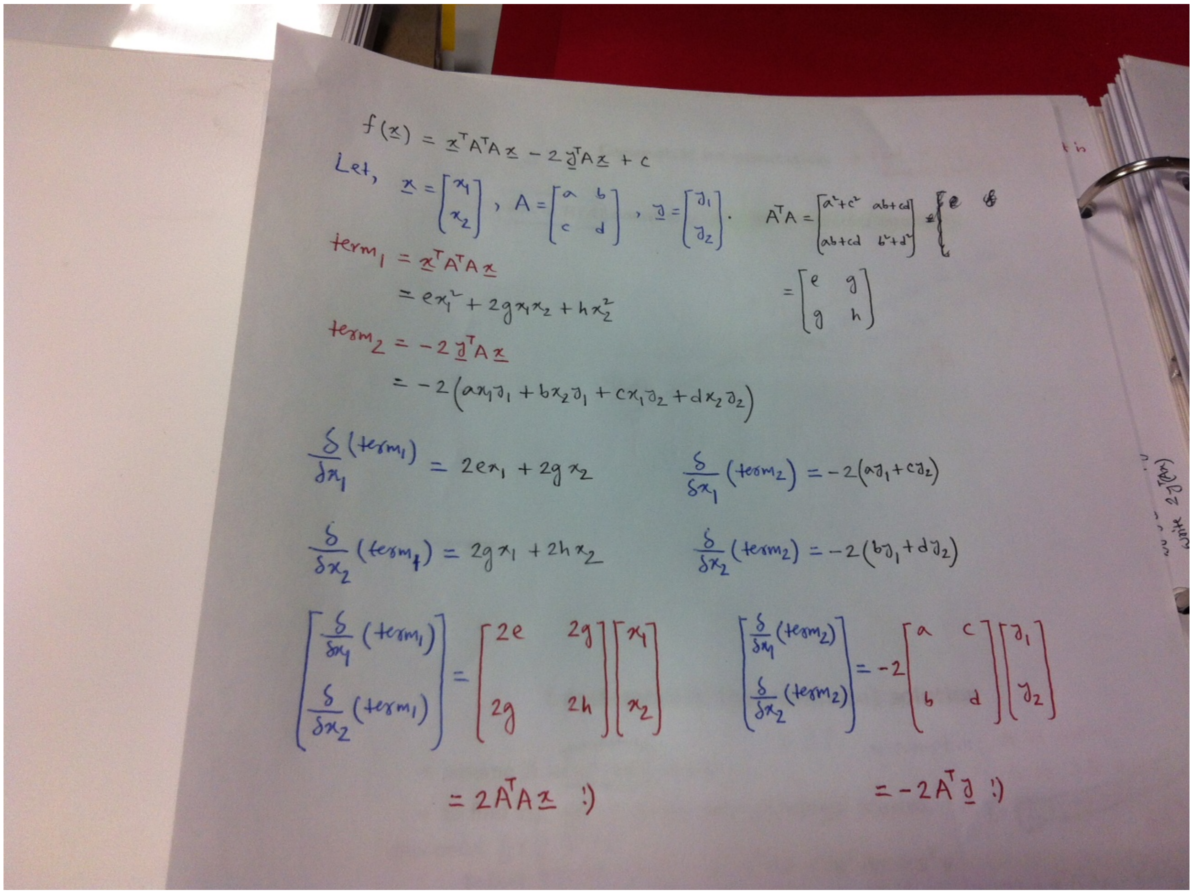


Figure 4: An example showing how we ended up at (4).

- `if A is full rank and square:` $\mathbf{x}_{\mathtt{ls}}$=$A^{-1}\mathbf{y}_{\mathtt{des}}$.

  this is because now A is invertible while for A being full rank,skinny A was singular i.e. typing $A^{-1}$ produced *Syntax Error*.

  $$\mathbf{x}_{\rm ls} = (A^TA)^{-1}A^T\mathbf{y}_{\rm des} = A^{-1}(A^T)^{-1}A^T\mathbf{y}_{\rm des} = A^{-1}\mathbf{y}_{\rm des}$$

  all these tell me that the matrix $(A^TA)^{-1}A^T$ is related to the *inverse* of A and when A is square it is in fact $A^{-1}$. Therefore the matrix $(A^TA)^{-1}A^T$ stands for a more generalized inverse and is known as the *pseudo inverse* or *Moore-Penrose inverse*.

  1. if A is **full rank, skinny** then its inverse is: $A^\dagger = (A^TA)^{-1}A^T$.
  2. if A is **full rank, square** then its inverse is: $A^{-1}$.

  **Rank interpretation:** since A is full rank and square therefore the column space is no longer a tiny subspace of the output space $\mathbf{R}^m$ rather they are the same thing i.e. C(A)=$\mathbf{R}^m$. This tells me that any $\mathbf{y}\in\mathbf{R}^m$ is achievable by using the linear combination of the columns of A. And that linear combination $\mathbf{x}$ is simply given by: $B\mathbf{y}$ where $B = A^{-1}$.

- `if I had` $\mathbf{y}_{\mathtt{des}}$`∈C(A) then I would have had:` A$\mathbf{x}_{\mathtt{ls}}$=$\mathbf{y}_{\mathtt{des}}$.

  $\mathbf{x}_{\rm ls}$ always tries to minimize $||\mathbf{y}_{\rm des}-A\mathbf{x}||$ and for $\mathbf{y}_{\rm des}\in$C(A) this is 0.

- $A^\dagger$ `is a` *`left inverse`* `of the full rank, skinny matrix` $A$`.`

  multiplying in the left of $A$ by $A^\dagger$ gives me the identity matrix.

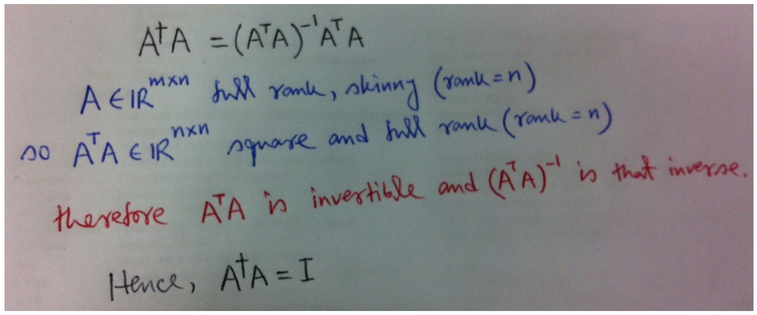


Figure 5: $A^\dagger$ is a *left inverse* of (full rank, skinny) $A$.

1. if A is **full rank, skinny** then it only has *left inverse* and it is: $A^\dagger = (A^TA)^{-1}A^T$.
2. if A is **full rank, square** then it has both *left inverse* and *right inverse* and they are the same thing: $A^{-1}$.

## 0.5 Projection on C(A)

A$\mathbf{x}_{\text{ls}}$=$\mathbf{y}_{\text{ls}}$ is the point in C(A) that is closest to $\mathbf{y}_{\text{des}}$ i.e. $\mathbf{y}_{\text{ls}}$ is the projection of $\mathbf{y}_{\text{des}}$ onto C(A). So what is the projection matrix P in (2) that gives me this?

$$\begin{aligned} \text{P } \mathbf{y}_{\text{des}} = \mathbf{y}_{\text{ls}} = A\mathbf{x}_{\text{ls}} = A(A^TA)^{-1}A^T\mathbf{y}_{\text{des}} \\ \Rightarrow \text{P} = A(A^TA)^{-1}A^T \end{aligned} \tag{6}$$

$A(A^TA)^{-1}A^T$ is the projection matrix associated with the vector space C(A) where A is full rank, skinny.

On the other hand, for a full rank, square matrix the associated projection matrix is: $P = A(A^TA)^{-1}A^T = AA^{-1}(A^T)^{-1}A^T = I \cdot I = I$. This has the implication that P$\mathbf{y}$=I$\mathbf{y}$=$\mathbf{y}$ which suggests that the projection of any vector $\mathbf{y} \in \mathbf{R}^m$ onto C(A) is $\mathbf{y}$ itself. The reason behind this is since C(A) is the output space therefore $\mathbf{y}$is already in C(A) :) Figure 6 points out similar pattern for the projection matrix of (6). This tells me that

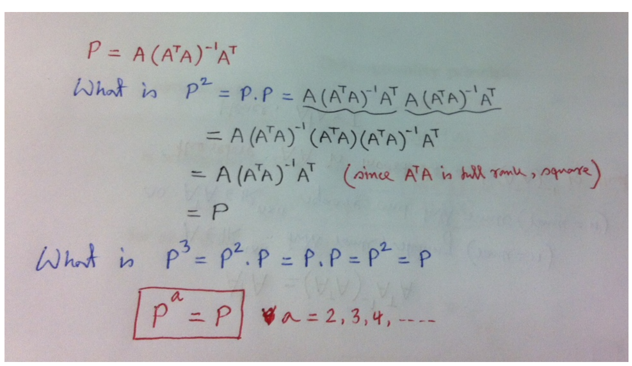


Figure 6: $P^a = P$ where $a = 2, 3, 4, ...$

once I am in C(A) my projection onto C(A) would always be myself. Few more things about the projection matrix of (6).

- `rank(P)=rank(A).`

  In $P = A(A^TA)^{-1}A^T$: rank(A)=n, rank($A^TA)^{-1}$)=n, rank($A^T$)=n. Therefore rank(P)=n=rank(A).

- `C(P)=C(A).`

  With A**x**=**y** I get **y**$in$C(A) therefore P$\mathbf{y}_{\text{des}}$=**y**should be in C(P) but I get **y**$in$C(A). That means C(P) and C(A) are the same thing. They stand for the same vector space but different bases :)

- $P^T = P$ `i.e.  P is symmetric.`

$$P^T = (A(A^TA)^{-1}A^T)^T = (A^T)^T((A^TA)^{-1})^TA^T = A(((A^T))^{-1})^T(A^{-1})^TA^T = A(A^TA)^{-1}A^T = P$$

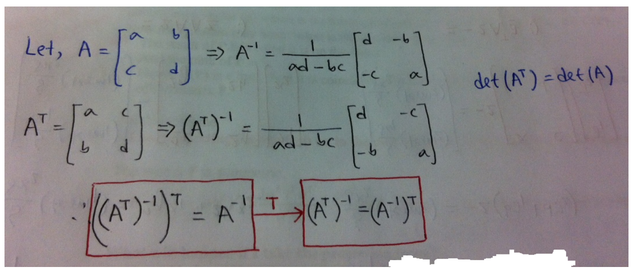


Figure 7: Supporting material for the proof of $P^T = P$.

## 0.6 Orthogonality Principle

The optimal residual $\mathbf{r}_{\text{ls}}$ is given by:

$$\mathbf{r}_{\text{ls}} = \mathbf{y}_{\text{des}} - \mathbf{y}_{\text{ls}} = \mathbf{y}_{\text{des}} - \text{P }\mathbf{y}_{\text{des}}$$

P projects a vector onto C(A). So what is P $\mathbf{r}_{\text{ls}}$ then?

$$\text{P }\mathbf{r}_{\text{ls}} = \text{P}(\mathbf{y}_{\text{des}} - \text{P }\mathbf{y}_{\text{des}}) = \text{P }\mathbf{y}_{\text{des}} - \text{P}^2\ \mathbf{y}_{\text{des}} = \text{P }\mathbf{y}_{\text{des}} - \text{P }\mathbf{y}_{\text{des}} = \mathbf{0}$$

This tells me that the *"C(A) piece"* inside the optimal residual is zero $\Rightarrow$ they are uncorrelated. More on this later. Now since P $\mathbf{r}_{\text{ls}} = \mathbf{0}$ therefore:

$$\mathbf{r}_{\text{ls}} \in \text{N(P)}$$

Since P is symmetric therefore $\mathbf{r}_{\text{ls}} \in N(P^T)$ as well. Then again $N(P^T) \perp C(P)$ and $C(P) = C(A)$ which tells me that the optimal residual $\mathbf{r}_{\text{ls}}$ is *orthogonal* to C(A). This has the following consequences:

- for any vector $\mathbf{y}\in$C(A): $\mathbf{r}_{\text{ls}} \perp \mathbf{y} \Rightarrow \mathbf{r}_{\text{ls}}^T\mathbf{y}=0$.
- $N(A^T) \perp C(A)$ so does $\mathbf{r}_{\text{ls}} \in N(A^T)$?
  well $N(P^T) \perp C(P)$ and $C(P) = C(A)$ so $N(P^T) \perp C(A)$. Hence $N(A^T) = N(P^T)$.
  Since $\mathbf{r}_{\text{ls}} \in N(P^T)$ so we can conclude that: $\mathbf{r}_{\text{ls}} \in N(A^T)$. Figure 8 shows an alternative way of proving this fact.

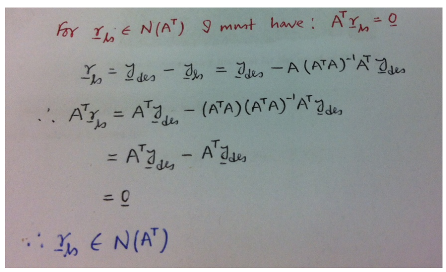


Figure 8: Another way to show that $\mathbf{r}_{\text{ls}} \in N(A^T)$.

**Geometric interpretation:** Let's assume the output space is $\mathbf{R}^3$ and C(A) is a 2D subspace in this output space i.e. a plane in $\mathbf{R}^3$. Then geometrically $\mathbf{r}_{\text{ls}} \perp C(A)$ means that $\mathbf{y}_{\text{ls}}$ is a point on this plane but $\mathbf{y}_{\text{des}}$ is not. And the vector from $\mathbf{y}_{\text{ls}}$ to $\mathbf{y}_{\text{des}}$ i.e. $\mathbf{y}_{\text{des}}-\mathbf{y}_{\text{ls}}$ is *normal* to this plane. This has been shown in Figure 9.

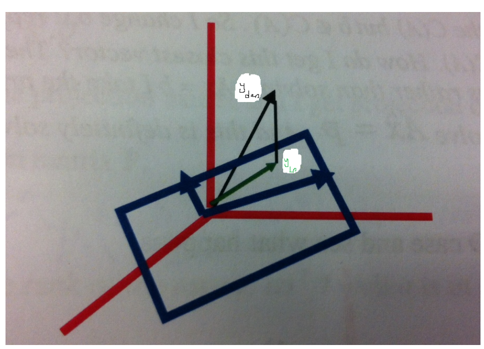

Figure 9: Geometric interpretation of $\mathbf{r}_{\text{ls}} \perp C(A)$.

Let's imagine a scenario where we received a (vector) signal with noise e.g. in a multiple antenna communication system. And somehow we know that the signal should lie in a certain subspace i.e. C(A). But since the signal$\notin$C(A) due to the fact that it has been contaminated with noise so we project the noisy received signal onto C(A) i.e. in the known range. Let,

- received signal: $\mathbf{y}_{\text{rcv}} \in \mathbf{R}^{10}$
- the signal subspace C(A) is a 6D subspace inside $\mathbf{R}^{10}$

So I project $\mathbf{y}_{\text{rcv}}$ onto this 6D subspace whose basis vectors I know i.e. the columns of the matrix A and I get $\mathbf{y}_{\text{proj}}$ as:

$$\mathbf{y}_{\text{proj}} = \mathrm{P}\ \mathbf{y}_{\text{rcv}} = A(A^TA)^{-1}A^T\mathbf{y}_{\text{rcv}}$$

Recall that C(A) and C(P) represents the same vector space with different bases. And the columns of P are a "better" basis than the columns of A. Why? You will understand this momentarily. However if I know A then I can generate P.

What I am claiming that the residual vector $\mathbf{y}_{\text{rcv}} - \mathbf{y}_{\text{proj}}$ is optimal because it is in $N(A^T)$ (the dimension that got crunched to zero) that means $(\mathbf{y}_{\text{rcv}} - \mathbf{y}_{\text{proj}}) \perp$ the signal subspace C(A) which stands for the fact that the de-noised signal $\mathbf{y}_{\text{proj}}$ is "purely" signal and the optimal residual is "purely" noise i.e. they are uncorrelated (in terms of the known range C(A)). This discussion provided me a good insight into how to

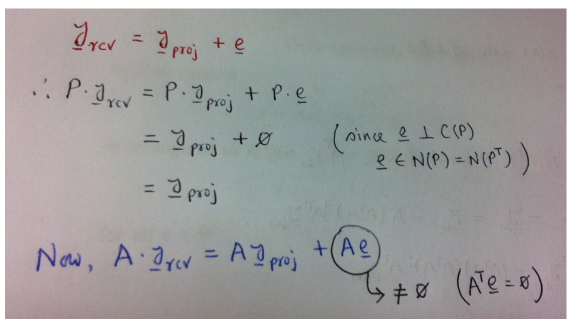


Figure 10: Comparison between the basis $\{P\}$ and the basis $\{A\}$.

design a de-noising transformation matrix.

# Bibliography

[1] Introduction to Linear Algebra, Gilbert Strang.

[2] Introduction to Linear Dynamical Systems, Stephen Boyd.