A New Identity for the Least-square Solution of Overdetermined Set of Linear Equations

Saeid Haghighatshoar*, Mohammad J. Taghizadeh†, and Afsaneh Asaei‡

*Technische Universität, Berlin, Germany

†Huawei European Research Centre, Munich, Germany

‡Idiap Research Institute, Martigny, Switzerland

Emails: saeid.haghighatshoar@tu-berlin.de, mohammad.taghizadeh@huawei.com, afsaneh.asaei@idiap.ch

Abstract

In this paper, we prove a new identity for the least-square solution of an over-determined set of linear equation $Ax = b$, where A is an $m \times n$ full-rank matrix, b is a column-vector of dimension m, and m (the number of equations) is larger than or equal to n (the dimension of the unknown vector x). Generally, the equations are inconsistent and there is no feasible solution for x unless b belongs to the column-span of A. In the least-square approach, a candidate solution is found as the unique x that minimizes the error function $\|Ax - b\|_2$.

We propose a more general approach that consist in considering all the consistent subset of the equations, finding their solutions, and taking a weighted average of them to build a candidate solution. In particular, we show that by weighting the solutions with the squared determinant of their coefficient matrix, the resulting candidate solution coincides with the least square solution.

Index Terms

Over-determined linear equation, Least square solution.

I. Introduction

A. Over-determined Set of Linear Equations

Let A be an $m \times n$ full-rank matrix and let $b \in \mathbb{R}^m$ be a column vector, and consider the linear equation $Ax = b$, to be solved for the unknown vector $x \in \mathbb{R}^n$. Theory and practice of solving these equations play a major role in essentially every part of mathematics such as linear algebra, operational research, optimization, combinatorics, etc. When $m > n$, we call the equations over-determined and there is a solution if and only if b belongs to the column-span of A [1]. Generally, the equations are inconsistent and we need some kind of criteria to build a candidate solution.

One approach for finding a solution is the least-square approach [2], where we find a solution by minimizing the quadratic form $\|Ax-b\|_2^2$. The resulting solution is given by $\hat{x} = A^{\#}b$, where $A^{\#} = (A^tA)^{-1}A^t$ denotes the pseudo-inverse of A. In estimation theory, $\hat{x}$ can be interpreted as the *best linear unbiased estimator* (BLUE) of the signal x observed via a linear channel given by the matrix A and contaminated with an i.i.d. Gaussian noise [3]. Note that in this case, if b is in the column-span of A, the resulting estimation error is zero.

Another approach for building a candidate solution is by some kind of averaging all the possible sub-solutions. To explain this more precisely, we first need to introduce some notations. For $k \in \mathbb{N}$, we define $[k] = \{1, 2, \dots, k\}$ to be the set of all integers from 1 up to k. We denote by $\mathcal{P}$ the set of all subsets of $[m]$ of size n, i.e., $\mathcal{P} = \{p \subset [m] : |p| = n\}$, where $|p|$ denotes the size of the subset p. For a $p \subset \mathcal{P}$, we define A_p to be the $n \times n$ matrix obtained by selecting the rows of the matrix A belonging to p by keeping their order as in A.

Suppose $p \in \mathcal{P}$ is such that $\mathsf{det}(A_p) \neq 0$. By restricting the equations to A_p, we can obtain a sub-solution $x_p = A_p^{-1}b_p$, where b_p is the a sub-vector of b consisting of the components with index in p whose order is the same as in b. Taking the weighted average of all possible sub-solutions with a weighting $\omega_p \geq 0$, $p \in \mathcal{P}$, we can build a candidate solution as follows

$$s^{\omega} = \frac{\sum_{p \in \mathcal{P}} \omega_p x_p}{\sum_{p \in \mathcal{P}} \omega_p}. \tag{1}$$

As the matrix A is full-rank, there is at least one $p \in \mathcal{P}$ with a nonzero $\mathsf{det}(A_p)$, thus s^{ω} is well-defined. By changing the associated weighting ω_p, we obtain a variety of candidate solutions for the over-determined equation $Ax = b$.

Let us consider the weighting function $\omega_p = \mathsf{det}(A_p)^2$, which is equal to the squared determinant of the sub-matrix A_p, and let us define the resulting solution by

$$\hat{x}_{\mathsf{LS}} = \frac{\sum_{p \in \mathcal{P}} \mathsf{det}(A_p)^2 A_p^{-1} b_p}{\sum_{p \in \mathcal{P}} \mathsf{det}(A_p)^2}. \tag{2}$$

If for a specific $p \in \mathcal{P}$, $\mathsf{det}(A_p) = 0$ then A_p^{-1} does not exist but, with some abuse of notation, this term does not play a role because its corresponding weight $\mathsf{det}(A_p)^2$ is equal to 0.

B. Our Contribution

We prove that with the weighting $\omega_p = \mathsf{det}(A_p)^2$, the resulting solution $\hat{x}_{\mathsf{LS}}$ in Eq. (2) coincides with the least-square solution given by $A^{\#}b = (A^tA)^{-1}A^tb$. More importantly, this holds for every full-rank matrix A and for an arbitrary vector b. We have summarized this in the following theorem.

Theorem 1. *Suppose A is a given $m \times n$ full-rank matrix with $m \geq n$ and assume that $b \in \mathbb{R}^m$ is an arbitrary vector. Let $\hat{x}_{\mathsf{LS}}$ be the weighted average solution given by Eq. (2). Then $\hat{x}_{\mathsf{LS}} = (A^tA)^{-1}A^tb$, i.e., $\hat{x}_{\mathsf{LS}}$ coincides* *with the least square solution.*

C. Notation and Auxiliary Results

In this section, we first introduce the required notations for the rest of the paper and prove some auxiliary results that we need to prove Theorem 1. Let B be an arbitrary $n \times n$ matrix and let $p \subset [m]$ of size $|p| = n$. We denote by $\mathsf{embb}(B, p, m)$ the embedding of columns of B inside an $n \times m$ matrix. More precisely, assume that the components of p are sorted with $p_1 < p_2 < \cdots < p_n$. Then $\mathsf{embb}(B, p, m)$ is an $n \times m$ matrix whose p_i-th column, $i \in [n]$, is equal to the i-th column of B, and all the other $m - n$ columns are set to zero.

Let $r, c \in \mathbb{N}$ be arbitrary numbers. We define the linear space of all $r \times c$ real-valued matrices by $M_{\mathbb{R}}(r, c)$ with the traditional matrix addition and scalar-matrix multiplication. For arbitrary matrices $M, N \in M_{\mathbb{R}}(r, c)$, we define the following bilinear form $\langle M, N\rangle = \mathsf{tr}(MN^t) = \sum_{i,j} M_{ij}N_{ij}$. It is not difficult to see that $\langle , \rangle$ defines an inner product on $M_{\mathbb{R}}(r, c)$. We denote the trace and the determinant of a square matrix M by $\mathsf{tr}(M)$ and $\mathsf{det}(M)$ respectively. We need the following auxiliary results from linear algebra. We have included all the proofs in Appendix A.

Lemma 1. *Let $r, c \in \mathbb{N}$ and let $M \in M_{\mathbb{R}}(r, c)$. If $\langle M, N\rangle = 0$ for every $N \in M_{\mathbb{R}}(r, c)$, then $M = 0$.*

Lemma 2. *Let M be an square invertible matrix whose components depend on a parameter u. Then, $\frac{\partial}{\partial u} M^{-1} = -M^{-1}(\frac{\partial}{\partial u} M)M^{-1}$.*

Lemma 3. *Let A be an square matrix whose components depend on a parameter u. Then, $\frac{\partial}{\partial u}\mathsf{det}(A) = \mathsf{det}(A)\mathsf{tr}(A^{-1}\frac{\partial}{\partial u}A)$*

Lemma 4. *Let M and S be $n \times n$ matrices, where S is symmetric. Then $\mathsf{tr}(SM) = \mathsf{tr}(SM^t)$.*

Theorem 2 (Cauchy-Binet)**.** *Let A and B be $m \times n$ matrices with $m \geq n$. Then,*

$$\mathsf{det}(A^t B) = \sum_{p \subset [m], |p| = n} \mathsf{det}(A_p)\mathsf{det}(B_p), \tag{3}$$

where $|p|$ denotes the number of elements of $p \subset [m]$.

II. Proof of the Main Theorem

In the section, we prove Theorem 1. Using Eq. (2), we can write $\hat{x}_{\mathsf{LS}}$ in the following form:

$$\hat{x}_{\mathsf{LS}}(A, b) = \frac{\sum_{p \in \mathcal{P}} \mathsf{det}(A_p)^2 A_p^{-1} b_p}{\mathsf{det}(A^t A)} \tag{4}$$

$$= \frac{\sum_{p \in \mathcal{P}} \mathsf{det}(A_p)^2\, \mathsf{embb}(A_p^{-1}, p, m) b}{\mathsf{det}(A^t A)}, \tag{5}$$

where in the last term we used the definition of $\mathsf{embb}(A_p^{-1}, p, m)$. Recall that for $p \in \mathcal{P}$, with elements $p_1 < p_2 < \cdots < p_n$, we denote by $\mathsf{embb}(A_p^{-1}, p, m)$ an all-zero $n \times m$ matrix except for its p_i-th column witch is equal to the i-th column of A_p^{-1}. Now, we need to prove that for any $b \in \mathbb{R}^m$ and for any $m \times n$ full-rank matrix A, the following identity holds

$$(A^t A)^{-1} A^t b = \frac{\sum_{p \in \mathcal{P}} \mathsf{det}(A_p)^2 \,\mathsf{embb}(A_p^{-1}, p, m)}{\mathsf{det}(A^t A)} b. \tag{6}$$

As this should be true for every $b \in \mathbb{R}^m$, we need to prove the following matrix identity:

$$\mathsf{det}(A^t A)\,(A^t A)^{-1} A^t = \textstyle\sum_{p \in \mathcal{P}} \mathsf{det}(A_p)^2 \,\mathsf{embb}(A_p^{-1}, p, m). \tag{7}$$

As a first step, it is easy to check that both sides are $n \times m$ matrices, thus the dimensions are compatible.

In order to prove the identity (7), let us define the function $f : M_{\mathbb{R}}(m, n) \to \mathbb{R}$ as follows:

$$f(A) = \mathsf{det}(A^t A) - \sum_{p \in \mathcal{P}} \mathsf{det}(A_p)^2. \tag{8}$$

Using the Cauchy-Binet formula as stated in Theorem 2, we obtain

$$\mathsf{det}(A^t A) = \sum_{p \in \mathcal{P}} \mathsf{det}(A_p)\mathsf{det}(A_p^t) = \sum_{p \in \mathcal{P}} \mathsf{det}(A_p)^2, \tag{9}$$

which implies that $f(A) = 0$ for every $A \in M_{\mathbb{R}}(m, n)$. Let $u = A_{ij}$ be a parameter denoting the component of A at row i and column j. As $f(A) = 0$, we have $\frac{\partial}{\partial u} f(A) = 0$, which implies that

$$\begin{aligned}
\frac{\partial}{\partial u}\mathsf{det}(A^t A) &\overset{(a)}{=} \mathsf{det}(A^t A)\mathsf{tr}\Big\{(A^t A)^{-1} \frac{\partial}{\partial u}(A^t A)\Big\} \\
&\overset{(b)}{=} \mathsf{det}(A^t A)\mathsf{tr}\Big\{(A^t A)^{-1}\big((\frac{\partial}{\partial u}A)^t A + A^t \frac{\partial}{\partial u}A\big)\Big\} \\
&\overset{(c)}{=} \mathsf{det}(A^t A)\mathsf{tr}\Big\{(A^t A)^{-1}\big(A^t \frac{\partial}{\partial u}A + A^t \frac{\partial}{\partial u}A\big)\Big\} \\
&= 2\,\mathsf{det}(A^t A)\mathsf{tr}\Big\{(A^t A)^{-1} A^t \frac{\partial}{\partial u}A\Big\} \\
&\overset{(d)}{=} 2\,\mathsf{det}(A^t A)\mathsf{tr}\Big\{(A^t A)^{-1} A^t U_{ij}\Big\} \\
&\overset{(e)}{=} 2\,\mathsf{det}(A^t A)\Big\langle (A^t A)^{-1} A^t, U_{ij}^t \Big\rangle,
\end{aligned}$$

where (a) follows from Lemma 3 applied to the matrix $A^t A$, (b) follows from the chain rule applied to $A^t A$, (c) follows from Lemma 4 applied to the symmetric matrix $(A^t A)^{-1}$ and the matrix $(\frac{\partial}{\partial u}A)^t A$, (d) results by taking the component-wise derivative of A with respect to $u = A_{ij}$ which we denote by U_{ij}, and where (e) results from the definition of the inner product for two matrices. We can simply check that U_{ij} is an $m \times n$ matrix with all-zero components except for ij-th component which is equal to 1.

Now, taking the derivative of the other term in Eq. (8) with respect to $u = A_{ij}$, we obtain

$$
\begin{aligned}
\frac{\partial}{\partial u}\sum_{p\in\mathcal{P}}\mathsf{det}(A_p)^2 &= \sum_{p\in\mathcal{P}} 2\,\mathsf{det}(A_p)\frac{\partial}{\partial u}\mathsf{det}(A_p)\\
&\overset{(a)}{=} \sum_{p\in\mathcal{P}} 2\,\mathsf{det}(A_p)\mathsf{det}(A_p)\mathsf{tr}(A_p^{-1}\frac{\partial}{\partial u}A_p)\\
&\overset{(b)}{=} \sum_{p\in\mathcal{P}} 2\,\mathsf{det}(A_p)^2\mathsf{tr}(\mathsf{embb}(A_p^{-1},p,m)\frac{\partial}{\partial u}A)\\
&\overset{(c)}{=} 2\,\mathsf{tr}\Big\{\sum_{p\in\mathcal{P}}\mathsf{det}(A_p)^2\mathsf{embb}(A_p^{-1},p,m)U_{ij}\Big\}\\
&\overset{(d)}{=} 2\,\Big\langle\sum_{p\in\mathcal{P}}\mathsf{det}(A_p)^2\mathsf{embb}(A_p^{-1},p,m),U_{ij}^t\Big\rangle,
\end{aligned}
$$

where (a) results from Lemma 3 applied to the matrix A_p. We also have (b) from the definition of the embedding n columns of A_p^{-1} in an $m\times n$ matrix. In particular, notice that as the remaining columns of $\mathsf{embb}(A_p^{-1},p,m)$ are all zero, we can replace A_p by A. Finally, (c) results from the linearity of the trace operator tr, and (d) follows from the definition of the inner product. Therefore, we obtain that

$$
\begin{aligned}
0 = \frac{\partial}{\partial u}f(A) = 2\Big\langle\ & U_{ij}^t,\\
&\mathsf{det}(A^tA)(A^tA)^{-1}A^t - \sum_{p\in\mathcal{P}}\mathsf{det}(A_p)^2\mathsf{embb}(A_p^{-1},p,m)\ \Big\rangle.
\end{aligned}
\tag{10}
$$

Notice that equality in Eq. (10) holds for all matrices U_{ij}^t, $i\in[m], j\in[n]$. As, U_{ij}^t form an orthonormal basis for the linear space $M_{\mathbb{R}}(m,n)$, from Lemma 1, it immediately results that

$$\mathsf{det}(A^tA)(A^tA)^{-1}A^t = \textstyle\sum_{p\in\mathcal{P}}\mathsf{det}(A_p)^2\mathsf{embb}(A_p^{-1},p,m).$$

From Eq. (7), this is exactly what we needed to prove.

Appendix A

Proof of the Auxiliary Results

In this section, we provide the proofs of the auxiliary results.

Proof of Lemma 1: Let $i\in[r], j\in[c]$ be arbitrary numbers and let N be an all-zero matrix except for the ij-th element which is set to 1. It results that

$$0 = \langle M,N\rangle = \sum_{k,\ell}M_{k\ell}N_{k\ell} = M_{ij} = 0.$$

As this is true for arbitrary i and j, it results that $M=0$.

Proof of Lemma 2: Let I be the identity matrix of the same order as M. Taking derivative from both sides of the identity $I = MM^{-1}$, and using the chain rule, we obtain that

$$0 = \frac{\partial}{\partial u} M M^{-1} + M \frac{\partial}{\partial u} M^{-1},$$

which implies that $\frac{\partial}{\partial u} M^{-1} = -M^{-1}(\frac{\partial}{\partial u} M) M^{-1}$.

Proof of Lemma 3: Assume that A is a $d \times d$ matrix and let us denote by A_{ij} the component of A in row i and column j. We first find $\frac{\partial}{\partial A_{ij}} \mathsf{det}(A)$ and use the chain rule to obtain

$$\frac{\partial}{\partial u} \mathsf{det}(A) = \sum_{i,j \in [d]} \frac{\partial}{\partial A_{ij}} \mathsf{det}(A) \frac{\partial}{\partial u} A_{ij}. \tag{11}$$

Notice that in order to compute $\mathsf{det}(A)$, we can expand it with respect to the i-th row, where we obtain

$$\mathsf{det}(A) = \sum_{k \in [d]} (-1)^{i+k} \mathsf{det}(\tilde{A}_{ik}), \tag{12}$$

where $\tilde{A}_{ik}$ is a $(d-1) \times (d-1)$ matrix obtained after removing the i-th row and the k-th column of the matrix A. In particular, it can be immediately checked that the only term in the summation (12) that depends on A_{ij} is $(-1)^{i+j} \mathsf{det}(\tilde{A}_{ij})$, thus we obtain

$$\frac{\partial}{\partial A_{ij}} \mathsf{det}(A) = (-1)^{i+j} \mathsf{det}(\tilde{A}_{ij}) = \mathsf{adj}(A)_{ji}, \tag{13}$$

where $\mathsf{adj}(A)$ denotes the *adjoint* of the matrix A. Moreover, from the formula $A^{-1} = \frac{\mathsf{adj}(A)}{\mathsf{det}(A)}$ for the inverse of the matrix A, we immediately obtain that

$$\frac{\partial}{\partial A_{ij}} \mathsf{det}(A) = \mathsf{det}(A) (A^{-1})_{ji}. \tag{14}$$

Using the the chain-rule as in Eq. (11), we have

$$\frac{\partial}{\partial u} \mathsf{det}(A) = \mathsf{det}(A) \sum_{ij} (A^{-1})_{ji} \frac{\partial}{\partial u} A_{ij} = \mathsf{det}(A) \mathsf{tr}(A^{-1} \frac{\partial}{\partial u} A),$$

where tr denotes the trace operator and where $\frac{\partial}{\partial u} A$ denotes the component-wise partial derivative of A with respect to u.

Proof of Lemma 4: The proof simply follows from the properties of the trace operator:

$$\mathsf{tr}(SM) = \mathsf{tr}((SM)^t) = \mathsf{tr}(M^t S^t) = \mathsf{tr}(M^t S) = \mathsf{tr}(S M^t),$$

where we used the symmetry of S and the fact that for arbitrary square matrices K, L of the same dimension, $\mathsf{tr}(KL) = \mathsf{tr}(LK)$.

References

[1] O. Bretscher, *Linear Algebra With Applications (3rd ed.)*, Upper Saddle River, NJ: Prentice Hall, 1995.

[2] http://en.wikipedia.org/wiki/Least_squares

[3] S. M. Stigler, *Gauss and the Invention of Least Squares*, Ann. Statist., 9 (3), 465–474, 1981.