Gradient Descent and Linear Regression with An Appendix on Generalized Inverses: Lectures for SF2930

Timo Koski

March 27, 2024

Department of Mathematics, KTH Royal Institute of Technology

Abstract

The convergence and rate of convergence of gradient descent to the least squares solution is studied. It is additionally shown that gradient descent converges to the unique solution of the minimum norm least squares problem. This gives a new insight in linear regression without any assumptions on the rank of the regressor matrix.

1 Introduction

Gradient descent is a widely used simple but effective method for minimization of functions, when there are no constraints. Applications to training of artificial neural networks and recurrent neural networks are exemplified in [9, Chapter 6 & Chapter 9, respectively]. Gradient decent is also a topic in optimization of convex functions, see, e.g., [2] and [13].

These lecture notes specialize the treatment in [2] and [13] to linear regression models and include a theorem from [6], which connects gradient decent to minimum norm least squares.

2 Gradient Descent Algorithm

2.1 Definition

A great number of tasks in computational statistics and machine learning are as follows.

We have a training set $\mathcal{D}_{tr} = (X, \mathbf{y})$. Here X is a data matrix and $\mathbf{y}$ is a vector of targets. $L(X, \mathbf{y}; \theta)$ is a differentiable cost/loss function, which is to be minimized for a fixed $\mathcal{D}_{tr}$ as a function of $\theta \in \mathbb{R}^d$, i.e., one wants to find a $\widehat{\theta}(X, \mathbf{y})$ satisfying

$$\widehat{\theta}(X, \mathbf{y}) := \text{argmin}_{\theta} L(X, \mathbf{y}; \theta).$$

The Gradient Descent Algorithm (GD) is an iteration for computing $\widehat{\theta}(\mathbf{x})$ as follows, see e.g. [13]:

$$\text{INPUT } \theta_0,\ \mu > 0.$$

1. For $t = 0, \ldots,$ stop
2. $$\theta_{t+1} = \theta_t - \mu \nabla_\theta L\left(X, \mathbf{y}; \theta\right) \tag{2.1}$$
3. Set $\theta_t \leftarrow \theta_{t+1}$, $t + 1 \leftarrow t$.
4. Return to (2.1) with θ_t
5. Continue till $t =$ stop

Here $\nabla_\theta L\left(X, \mathbf{y}; \theta\right)$ is the gradient of the cost function, see (D.1) in the Appendix. One important geometric property of the gradient to be recalled for understanding of (2.1) is that a function $f(\mathbf{x})$ decreases most rapidly in the direction $-\nabla f(\mathbf{x})$, see [1, p. 720].
μ is the *learning rate*. In certain algorithms the learning rate is changing in every iteration. If the learning rate is too high, the GD may overshoot the minimum. If μ is too low, the training will take too long and may never reach the global minimum, or else get stuck in local minima of $L\left(X, \mathbf{y}; \theta\right)$. It is possible to hit saddle points of $L\left(X, \mathbf{y}; \theta\right)$, too.
It holds, under certain assumptions, that if μ is sufficiently small, then

$$L\left(\theta_{t+1}\right) < L\left(\theta_t\right).$$

We shall establish this decrease for minimization of least squares in linear regression, where the assumption of Lipschitz continuity of the gradient $\nabla_\theta L\left(X, \mathbf{y}; \theta\right)$ is shown to hold. We shall also show that the GD Algorithm converges to the minimum norm least squares estimate and provide two convergence rates.

2.2 Stochastic Gradient Descent (SGD)

SGD has been called the *workhorse of statistical/machine learning*. The applications of SGD to training of neural networks by backpropagation are presented in [9, Chapter 6]. In one version SGD computes the gradient of the parameters using a few (=N) examples $X_N, \mathbf{y}_N$ randomly chosen from a training set. The iteration looks then like

$$\theta_{t+1} = \theta_t - \mu \nabla_\theta L\left(X_N, \mathbf{y}_N; \theta_t\right). \tag{2.2}$$

Another version is to pick at random a new item $(\mathbf{x}_i, y_i)$ from the training set for each step of iteration

$$\theta_{t+1} = \theta_t - \mu \nabla_\theta L\left(\mathbf{x}_i, y_i; \theta_t\right). \tag{2.3}$$

2.3 GD derived

GD was invented by Augustin-Louis Cauchy in 1847, [8], as a way of unconstrained minimimization of a function. One makes a second order Taylor expansion

$$f\left(\mathbf{y}\right) \approx f\left(\mathbf{x}\right) + \nabla f\left(\mathbf{x}\right)^{T}\left(\mathbf{y}-\mathbf{x}\right) + \frac{1}{2}(\mathbf{y}-\mathbf{x})^{T}\mathcal{H}\left(\mathbf{x}\right)\left(\mathbf{y}-\mathbf{x}\right),$$

and replaces the Hessian by $\frac{1}{\mu}\mathbb{I}$,

$$f\left(\mathbf{y}\right) \approx f\left(\mathbf{x}\right) + (\mathbf{y}-\mathbf{x})^{T}\nabla f\left(\mathbf{x}\right) + \frac{1}{2\mu}(\mathbf{y}-\mathbf{x})^{T}\mathbb{I}(\mathbf{y}-\mathbf{x}).$$

This means that we go from something elliptical to something spherical. Then one optimizes w.r.t. $\mathbf{y}$ and sets $\nabla_{\mathbf{y}} f\left(\mathbf{y}\right) = \mathbf{0}$, which gives by the rules in Appendix D.1

$$\nabla f\left(\mathbf{x}\right) + \frac{1}{\mu}\left(\mathbf{y}-\mathbf{x}\right) = \mathbf{0}$$

or

$$\mathbf{y} = \mathbf{x} - \mu\nabla f\left(\mathbf{x}\right).$$

Cauchy did not prove the convergence of GD, but he seems to have believed that there was a convergence [8].

2.4 Organization of the Lecture Notes

In section 3 GD is stated for the least squares object function for minimization in linear regression. The crucial Lipschitz continuity of the gradient of the least squares objecti function is established. This requires the Rayleigh principle as found in Appendix B.3. A number of inequalities are established by means of this Lipschitz continuity.

In section 4 these inequalities are used to prove the monotone decrease of the the least squares object function at each iteration and the convergence of GD in linear regression and least squares. Section 5 shoows that the limit of the GD iterations is the unique solution of the minimum norm least squares.

In section 6 rates of convergence statement for GD is established under a sufficiently small learning rate. The general rate is valid for all regressor matrices, but there is another different rate for invertible regressor matrices.

Appendix A recapitulates the proof of Proposition 3.1 on the minimum of the least squares distance. Appendix B.2 contains various standard definitions and rules of linear algebra. Appendix C deals with generalized inverses and their applications to linear regression. The key result is Proposition C.12, which gives a representation of any solution to the normal equations by means of a certain Bjerhammar- Moore-Penrose generalized inverse, a.k.a. pseudoinverse.

In appendix C.4 the theory of minimum norm least squares is presented. This theory is essential for analysis of GD in the cases, where the number of predictors is larger than the number of training samples. Asymptic convergence in mean square of the least squares estimate, as the number of data points in the training set generated by a true multiple regression model increases without bounds, is given in section C.4.2. Appendices D and E contain auxiliary results from multivarible calculus and monotonic convergence of sequences, respectively.

3 Least Squares of Linear Regression

3.1 GD Algorithm for Minimization of the Least Squares Cost Function

We have the least squares distance

$$Q\left(\boldsymbol{\beta}\right) := \frac{1}{2} \parallel \mathbf{y} - X\boldsymbol{\beta} \parallel^2 \tag{3.4}$$

to be minimized as a function of the $k \times 1$ vector $\boldsymbol{\beta}$ of regression coefficients. $\mathbf{y}$ is $n \times 1$ vector of responses and X is $n \times k$ regressor matrix.
We are not assuming anything about the rank of X at this moment, or of whether $k > n$ or $k \leq n$.
By definition of the norm in (B.1) we expand

$$\frac{1}{2} \parallel \mathbf{y} - X\boldsymbol{\beta} \parallel^2 = \frac{1}{2}\left(\mathbf{y} - X\boldsymbol{\beta}\right)^T\left(\mathbf{y} - X\boldsymbol{\beta}\right) = \frac{1}{2}\mathbf{y}^T\mathbf{y} - \boldsymbol{\beta}^T X^T \mathbf{y} + \frac{1}{2}\boldsymbol{\beta}^T X^T X \boldsymbol{\beta}.$$

By the rules in Appendix D.1 we get the gradient

$$\nabla_{\boldsymbol{\beta}} Q\left(\boldsymbol{\beta}\right) = -X^T\mathbf{y} + X^T X\boldsymbol{\beta}, \tag{3.5}$$

where we used the fact that $X^T X$ is symmetric. With respect to (2.1) we have the iteration step

$$\boldsymbol{\beta}_{t+1} = \boldsymbol{\beta}_t + \mu X^T\left(\mathbf{y} - X\boldsymbol{\beta}_t\right). \tag{3.6}$$

The assertion to be proved is that this iteration converges to a minimum of $Q\left(\boldsymbol{\beta}\right)$. Here we need the following result.

Proposition 3.1: *Any $\boldsymbol{\beta}_{\text{sol}} \in \mathbb{R}^k$ that satisfies the equation*

$$\nabla_{\boldsymbol{\beta}} Q\left(\boldsymbol{\beta}_{\text{sol}}\right) = \mathbf{0}_k \tag{3.7}$$

minimizes $Q\left(\boldsymbol{\beta}\right)$, i.e., for all $\boldsymbol{\beta}$

$$Q\left(\boldsymbol{\beta}_{\text{sol}}\right) \leq Q\left(\boldsymbol{\beta}\right).$$

Proof: We are for the moment assuming that there are solutions to (3.7). The proof is found in Appendix A. ■

3.2 Why GD for Linear Regression ?

When $n > k$ and the regression matrix X has full column rank, we know that

$$\widehat{\boldsymbol{\beta}} := \text{argmin}_{\boldsymbol{\beta}} Q\left(\boldsymbol{\beta}\right) \Leftrightarrow \widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{y}.$$

i.e. $\widehat{\boldsymbol{\beta}}$ is the unique solution to the normal equations in (4.24). Hence one might raise the question, as to what is the rationale for discussing GD in this case? A first answer is that forming $X^T X$ is claimed by sources to be unstable for all but the most well-conditioned systems; hence in practice one would avoid forming $X^T X$ directly, c.f. [4, Chapter 2.3].

If $n < k$, then X does not have full column rank, and there are many solutions to the normal equations. Clearly, the gradient descent iteration

$$\boldsymbol{\beta}_{t+1} = \boldsymbol{\beta}_t + \mu X^T \left(\mathbf{y} - X\boldsymbol{\beta}_t\right)$$

can be implemented in both situations, as neither $X^T X$ nor inversion of it is needed. Moreover, it has been pointed out that

- The inversion of a $k \times k$ matrix requires $\mathcal{O}\left(k^3\right)$ operations, c.f., [7, p. 151],
- A GD update requires $\mathcal{O}\left(kn\right)$ operations, as is obvious.

For large k the difference between inversion and GD is more than considerable. Hence, even if $(X^T X)^{-1}$ did exist, it may thus be advantageous to revert to GD. It will turn out that the behavior of GD is critically influenced by a certain matrix norm.

3.3 A Matrix Norm

One of the many possible matrix norms is defined by

$$\|A\| = \sup_{\mathbf{x} \neq 0} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|}, \tag{3.8}$$

where the vector norms $\|A\mathbf{x}\|$ and $\|\mathbf{x}\|$ are defined by (B.1) of Appendix B. One reads sup as 'supremum' and this means the smallest upper bound of a real valued function or a set in $\mathbb{R}^n$, see [17, p. 4]. Here we deal with the smallest upper bound of $\frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|}$. The matrix norm satisfies the properties 1.-4. in Appendix B.1. The norm in (3.8) has by the meaning of supremum the *consistency property*

$$\|A\mathbf{x}\| \leq \|A\| \, \|\mathbf{x}\|. \tag{3.9}$$

The norm in (3.8) is given by the following theorem

Proposition 3.2: *If λ_+ =largest eigenvalue of $A^T A$, then*

$$\|A\| = \sup_{\mathbf{x} \neq 0} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} = \sqrt{\lambda_+}. \tag{3.10}$$

Proof: This is given in Appendix B.3. ■

In fact it holds, see Appendix B.3, that the supremum is assumed by $\mathbf{e}_+$, the eigenvector of $A^T A$ corresponding to λ_+, and thus we have

$$\|A\| = \max_{\mathbf{x}\neq 0} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} = \sqrt{\lambda_+}.$$

3.4 Inequalities for $Q(\boldsymbol{\beta})$

We find next a case of the *Lipschitz continuity of the gradient.*

Lemma 3.3: *Let $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^{'}$ be two vectors of regression coefficients. Then*

$$\parallel \nabla_{\boldsymbol{\beta}} Q\left(\boldsymbol{\beta}^{'}\right) - \nabla_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}) \parallel \leq \lambda_+\left(X^T X\right) \parallel \boldsymbol{\beta}^{'} - \boldsymbol{\beta} \parallel, \tag{3.11}$$

where $\lambda_+\left(X^T X\right)$ is the largest eigenvalue of $X^T X$.

Proof: We get from (3.5) that

$$\nabla_{\boldsymbol{\beta}} Q\left(\boldsymbol{\beta}^{'}\right) - \nabla_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}) = X^T X\left(\boldsymbol{\beta}^{'} - \boldsymbol{\beta}\right).$$

Hence we have

$$\parallel \nabla_{\boldsymbol{\beta}} Q\left(\boldsymbol{\beta}^{'}\right) - \nabla_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}) \parallel = \parallel X^T X\left(\boldsymbol{\beta}^{'} - \boldsymbol{\beta}\right) \parallel .$$

By (3.9) we obtain

$$\parallel X^T X\left(\boldsymbol{\beta}^{'} - \boldsymbol{\beta}\right) \parallel \leq \parallel X^T X \parallel \parallel \left(\boldsymbol{\beta}^{'} - \boldsymbol{\beta}\right) \parallel$$

By Theorem 3.2 $\parallel X^T X \parallel$ equals the square root of the largest eigenvalue of $(X^T X)^T X^T X$. Here $(X^T X)^T X^T X = (X^T X)(X^T X) = (X^T X)^2$. The eigenvectors of $(X^T X)^2$ are eigenvectors of $X^T X$. If λ_i is the eigenvalue of $X^T X$ corresponding to the eigenvector $\mathbf{e}_i$, then λ_i^2 is the eigenvalue of $(X^T X)^2$ corresponding to the eigenvector $\mathbf{e}_i$. By (3.10) we thus obtain, as the eigenvalues of symmetric and positive semidefinite matrices are real and nonnegative, that

$$\parallel X^T X \parallel = \sqrt{\lambda_+((X^T X)^2)} = \sqrt{\lambda_+^2(X^T X)} = |\lambda_+\left(X^T X\right)| = \lambda_+\left(X^T X\right),$$

where $\lambda_+\left((X^T X)^2\right)$ is the largest eigenvalue of $(X^T X)^2$ and $\lambda_+\left(X^T X\right)$ is the largest eigenvalue of $X^T X$. Hence we have established (3.11) as claimed. ■

Proposition 3.4: *For any $\boldsymbol{\beta}^{'}$ and $\boldsymbol{\beta}$ we have*

$$| Q\left(\boldsymbol{\beta}^{'}\right) - Q(\boldsymbol{\beta}) - \nabla Q(\boldsymbol{\beta})^T\left(\boldsymbol{\beta}^{'} - \boldsymbol{\beta}\right) | \leq \frac{\lambda_+\left(X^T X\right)}{2} \parallel \boldsymbol{\beta}^{'} - \boldsymbol{\beta} \parallel^2 . \tag{3.12}$$

Proof: In view of (D.4) with $\mathbf{h} = \left(\boldsymbol{\beta}^{'} - \boldsymbol{\beta}\right)$ and $t \in [0,1]$

$$Q\left(\boldsymbol{\beta}^{'}\right) = Q\left(\boldsymbol{\beta}\right) + \int_0^1 \nabla Q\left(\boldsymbol{\beta} + t\mathbf{h}\right)^T \left(\boldsymbol{\beta}^{'} - \boldsymbol{\beta}\right) dt,$$

and thus

$$\begin{aligned} Q\left(\boldsymbol{\beta}^{'}\right) &= Q\left(\boldsymbol{\beta}\right) + \nabla Q\left(\boldsymbol{\beta}\right)^T \left(\boldsymbol{\beta}^{'} - \boldsymbol{\beta}\right) \\ &\quad + \int_0^1 \left[\nabla Q\left(\boldsymbol{\beta} + t\mathbf{h}\right)^T \left(\boldsymbol{\beta}^{'} - \boldsymbol{\beta}\right) - \nabla Q\left(\boldsymbol{\beta}\right)^T \left(\boldsymbol{\beta}^{'} - \boldsymbol{\beta}\right)\right] dt. \end{aligned}$$

Then

$$\begin{aligned} &\mid Q\left(\boldsymbol{\beta}^{'}\right) - Q\left(\boldsymbol{\beta}\right) - \nabla Q\left(\boldsymbol{\beta}\right)^T \left(\boldsymbol{\beta}^{'} - \boldsymbol{\beta}\right) \mid \\ &= \mid \int_0^1 \left[\nabla Q\left(\boldsymbol{\beta} + t\mathbf{h}\right)^T \left(\boldsymbol{\beta}^{'} - \boldsymbol{\beta}\right) - \nabla Q\left(\boldsymbol{\beta}\right)^T \left(\boldsymbol{\beta}^{'} - \boldsymbol{\beta}\right)\right] dt \mid \\ &\leq \int_0^1 \mid \left[\nabla Q\left(\boldsymbol{\beta} + t\mathbf{h}\right)^T \left(\boldsymbol{\beta}^{'} - \boldsymbol{\beta}\right) - \nabla Q\left(\boldsymbol{\beta}\right)^T \left(\boldsymbol{\beta}^{'} - \boldsymbol{\beta}\right)\right] \mid dt \\ &= \int_0^1 \mid \left[\left(\nabla Q\left(\boldsymbol{\beta} + t\mathbf{h}\right) - \nabla Q\left(\boldsymbol{\beta}\right)\right)^T \left(\boldsymbol{\beta}^{'} - \boldsymbol{\beta}\right)\right] \mid dt. \end{aligned}$$

By the Cauchy-Schwarz inequality (B.2)

$$\leq \int_0^1 \parallel \nabla Q\left(\boldsymbol{\beta} + t\mathbf{h}\right) - \nabla Q\left(\boldsymbol{\beta}\right) \parallel \parallel \boldsymbol{\beta}^{'} - \boldsymbol{\beta} \parallel dt,$$

and (3.11) gives

$$\leq \lambda_+ \left(X^T X\right) \int_0^1 \parallel \boldsymbol{\beta} + t\mathbf{h} - \boldsymbol{\beta} \parallel \parallel \boldsymbol{\beta}^{'} - \boldsymbol{\beta} \parallel dt.$$

and by definition of $\mathbf{h}$

$$= \lambda_+ \left(X^T X\right) \int_0^1 \parallel t(\boldsymbol{\beta}^{'} - \boldsymbol{\beta}) \parallel \parallel \boldsymbol{\beta}^{'} - \boldsymbol{\beta} \parallel dt$$

According to rule 3. for the norm in Appendix B.1 we have, as $0 \leq t \leq 1$, $\parallel t(\boldsymbol{\beta}^{'} - \boldsymbol{\beta}) \parallel = t \parallel (\boldsymbol{\beta}^{'} - \boldsymbol{\beta}) \parallel$. Hence we have

$$= \lambda_+ \left(X^T X\right) \parallel \boldsymbol{\beta}^{'} - \boldsymbol{\beta} \parallel^2 \int_0^1 t dt.$$

Thereby the inequality in (3.12) has been proven. ■

4 On the Convergence of the GD Algorithm

4.1 GD for LSE in Multiple Regression Converges for a Sufficiently Small Learning Rate

Let us next write the GD for LSE as

$$\boldsymbol{\beta}_{t+1} = \boldsymbol{\beta}_t - \mu \nabla_{\boldsymbol{\beta}} Q\left(\boldsymbol{\beta}_t\right). \tag{4.13}$$

We have from (3.12)

$$Q\left(\boldsymbol{\beta}_{t+1}\right) \leq Q\left(\boldsymbol{\beta}_t\right) + \nabla Q\left(\boldsymbol{\beta}_t\right)^T\left(\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}_t\right) + \frac{\lambda_+\left(X^T X\right)}{2} \parallel \boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}_t \parallel^2 \tag{4.14}$$

Remark 1: $|a-b| \leq c$ is equivalent to $-c \leq a-b \leq c$, and the right hand inequality says $a \leq b+c$. Thus we got (4.14) from (3.12).

Next we insert the iteration from (4.13) in (4.14) and we get

$$Q\left(\boldsymbol{\beta}_{t+1}\right) \leq Q\left(\boldsymbol{\beta}_t\right) - \mu \nabla Q\left(\boldsymbol{\beta}_t\right)^T \nabla Q\left(\boldsymbol{\beta}_t\right) + \mu^2 \frac{\lambda_+\left(X^T X\right)}{2} \parallel \nabla Q\left(\boldsymbol{\beta}_t\right) \parallel^2 \tag{4.15}$$

where we used rule 3. for norm in Appendix B. By definitions $\nabla Q\left(\boldsymbol{\beta}_t\right)^T \nabla Q\left(\boldsymbol{\beta}_t\right) = \parallel \nabla Q\left(\boldsymbol{\beta}_t\right) \parallel^2$. Hence we have

$$Q\left(\boldsymbol{\beta}_{t+1}\right) \leq Q\left(\boldsymbol{\beta}_t\right) - \left(1 - \mu \frac{\lambda_+\left(X^T X\right)}{2}\right) \mu \parallel \nabla Q\left(\boldsymbol{\beta}_t\right) \parallel^2 \tag{4.16}$$

Next set $L := \lambda_+\left(X^T X\right)$ and assume that

$$\mu \leq \frac{1}{L}.$$

Then

$$-\left(1 - \mu \frac{\lambda_+\left(X^T X\right)}{2}\right) = \mu \frac{L}{2} - 1 \leq \frac{1}{2} - 1 = -\frac{1}{2}. \tag{4.17}$$

Hence we get in (4.16)

$$Q\left(\boldsymbol{\beta}_{t+1}\right) \leq Q\left(\boldsymbol{\beta}_t\right) - \frac{\mu}{2} \parallel \nabla Q\left(\boldsymbol{\beta}_t\right) \parallel^2 \tag{4.18}$$

But since $\frac{\mu}{2} \parallel \nabla Q\left(\boldsymbol{\beta}_t\right) \parallel^2$ is a positive number, we have

$$Q\left(\boldsymbol{\beta}_{t+1}\right) < Q\left(\boldsymbol{\beta}_t\right). \tag{4.19}$$

Proposition 4.1: *The iterated GD sequence* $\{Q\left(\boldsymbol{\beta}_t\right)\}_{t=0}^{+\infty}$ *is converging for any* $\mu \leq \frac{1}{L}$.

Proof: By (4.19) the positive sequence $\{Q\left(\boldsymbol{\beta}_t\right)\}_{t=0}^{+\infty}$ is monotonically decreasing. It has a lower bound, as all $Q\left(\boldsymbol{\beta}_t\right) > 0$ for all t. By b) in Proposition E.1 the asserted convergence follows. ■

The next question is the limit of $\{\boldsymbol{\beta}_t\}_{t=0}^{+\infty}$. In view of Proposition 3.1 and (4.19) one would expect convergence to some of $\text{argmin}_{\boldsymbol{\beta}} Q\left(\boldsymbol{\beta}\right)$.

4.2 The limit of GD for LSE in Multiple Regression for a Sufficiently Small Learning Rate

Proposition 4.2: *The gradient descent for minimizing the least squares*

$$\boldsymbol{\beta}_{t+1} = \boldsymbol{\beta}_t - \mu \nabla_{\boldsymbol{\beta}} Q\left(\boldsymbol{\beta}_t\right), \tag{4.20}$$

where $\mu \leq \frac{1}{\lambda_+(X^T X)}$, *converges to a* $\boldsymbol{\beta}_c$, *which satisfies*

$$\nabla_{\boldsymbol{\beta}} Q\left(\boldsymbol{\beta}_c\right) = \mathbf{0}_k. \tag{4.21}$$

Proof: Let us recall again that Lemma C.6 shows that solutions to the equations $\nabla_{\boldsymbol{\beta}} Q\left(\boldsymbol{\beta}_c\right) = \mathbf{0}_{k+1}$ exist.
For ease of writing, let us set again $L := \lambda_+\left(X^T X\right)$. Then (4.18) gives

$$Q\left(\boldsymbol{\beta}_t\right) - Q\left(\boldsymbol{\beta}_{t+1}\right) \geq \frac{1}{2L} \parallel \nabla Q\left(\boldsymbol{\beta}_t\right) \parallel^2 \geq \frac{\mu}{2} \parallel \nabla Q\left(\boldsymbol{\beta}_t\right) \parallel^2 \tag{4.22}$$

Let us now sum over the iterations in the both sides of the left hand inequality

$$\sum_{t=0}^{N} \left(Q\left(\boldsymbol{\beta}_t\right) - Q\left(\boldsymbol{\beta}_{t+1}\right)\right) \geq \frac{1}{2L} \sum_{t=0}^{N} \parallel \nabla Q\left(\boldsymbol{\beta}_t\right) \parallel^2 . \tag{4.23}$$

The sum in the left hand side is 'telescoping', i.e.,

$$\begin{aligned}
\sum_{t=0}^{N} \left(Q\left(\boldsymbol{\beta}_t\right) - Q\left(\boldsymbol{\beta}_{t+1}\right)\right) &= Q\left(\boldsymbol{\beta}_0\right) - Q\left(\boldsymbol{\beta}_1\right) \\
&+ Q\left(\boldsymbol{\beta}_1\right) - Q\left(\boldsymbol{\beta}_2\right) \\
&+ Q\left(\boldsymbol{\beta}_2\right) - Q\left(\boldsymbol{\beta}_3\right) \\
&\ldots \\
&+ Q\left(\boldsymbol{\beta}_{N-3}\right) - Q\left(\boldsymbol{\beta}_{N-2}\right) \\
&+ Q\left(\boldsymbol{\beta}_{N-2}\right) - Q\left(\boldsymbol{\beta}_{N-1}\right) \\
&+ Q\left(\boldsymbol{\beta}_{N-1}\right) - Q\left(\boldsymbol{\beta}_N\right) \\
&= Q\left(\boldsymbol{\beta}_0\right) - Q\left(\boldsymbol{\beta}_N\right) \\
&\leq Q\left(\boldsymbol{\beta}_0\right) - Q\left(\boldsymbol{\beta}_{\min}\right),
\end{aligned}$$

where $Q\left(\boldsymbol{\beta}_{\min}\right) = \min_{\boldsymbol{\beta}} Q\left(\boldsymbol{\beta}\right) < Q\left(\boldsymbol{\beta}_N\right)$. It is known that $Q\left(\boldsymbol{\beta}\right)$ has a minimum by Proposition 3.1. We obtain thus via (4.22)

$$Q\left(\boldsymbol{\beta}_0\right) - Q\left(\boldsymbol{\beta}_{\min}\right) \geq \frac{1}{2L} \sum_{t=0}^{N} \parallel \nabla Q\left(\boldsymbol{\beta}_t\right) \parallel^2$$

Here $Q\left(\boldsymbol{\beta}_0\right) - Q\left(\boldsymbol{\beta}_{\min}\right) > 0$, if $\boldsymbol{\beta}_0 \neq \boldsymbol{\beta}_{\min}$. Thus the monotonically increasing series $s_N = \sum_{t=0}^{N} \parallel Q\left(\boldsymbol{\beta}_t\right) \parallel^2$ is bounded. By Theorem E.1 s_N converges as $N \to +\infty$ and by Theorem E.2 this convergence implies that

$$\parallel \nabla Q\left(\boldsymbol{\beta}_t\right) \parallel^2 \to 0, \quad \text{as } t \to +\infty.$$

This means by a property of norms that

$$\nabla Q\left(\boldsymbol{\beta}_t\right) \to \mathbf{0}_{k+1}, \quad \text{as } t \to +\infty.$$

Hence, by (4.20), $\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}_t \to \mathbf{0}_{k+1}$, i.e. there is a $\boldsymbol{\beta}_c$ such that

$$\boldsymbol{\beta}_t \to \boldsymbol{\beta}_c,$$

In view of (3.5), $\nabla_{\boldsymbol{\beta}} Q\left(\boldsymbol{\beta}\right)$ is a continuous function of $\boldsymbol{\beta}$, and thus

$$\nabla Q\left(\boldsymbol{\beta}_c\right) = \mathbf{0}_k.$$

Thus the assertion in the Proposition holds as claimed. ■

In view of (3.5), $\nabla_{\boldsymbol{\beta}} Q\left(\boldsymbol{\beta}_c\right) = \mathbf{0}_k$ gives the *normal equations*

$$X^T X \boldsymbol{\beta}_c = X^T \mathbf{y}. \tag{4.24}$$

We have by Lemma C.6 that there always exist solutions to (4.24). We shall next show that GD converges to the unique solution of (4.24), which has the minimum norm $\parallel \boldsymbol{\beta} \parallel$.

5 The limit of the GD sequence: Minimum Norm Least Squares

Generalized inverses can be used to give (many different) solutions to the normal equations (4.24). The definition and basics of generalized inverses, to the extent required here, are presented in Appendix C. In particular, we need a special case of generalized inverses known as the Bjerhammar-Moore-Penrose (BMP) inverses, also known as pseudoinverses, found in Appendix C.3. The BMP inverse of the regressor matrix X operating on $\mathbf{y}$ is also the unique solution of the minimum norm minimum least squares, c.f. Appendix C.3.3. The next result is [6, Proposition 1,p. 9]. The proof below fills in a number details omitted in [6, loc.cit], where the proof is actually three lines long. The result is dependent on the precise expression of the GD in (3.6).

Proposition 5.1: *Take the starting value of GD in (4.20) as $\boldsymbol{\beta}_0 = \mathbf{0}_k$ and $\mu \leq \frac{1}{\lambda_+(X^TX)}$. Then, as $t \to +\infty$, the $\boldsymbol{\beta}_t$ in (3.6) converges to $\boldsymbol{\beta}^+$, which is the unique minimum norm least squares solution of the normal equations (4.24), and is given by*

$$\boldsymbol{\beta}^+ = X^+ \mathbf{y},$$

where X^+ is the BMP inverse of the regressor matrix X.

Proof: The convergence $\boldsymbol{\beta}_t$, as $t \to +\infty$, is established by Proposition 4.2, as $\mu \leq \frac{1}{\lambda_+(X^TX)}$ is being assumed. The special assumption $\boldsymbol{\beta}_0 = \mathbf{0}$ means that the first iteration of GD in (3.6) is

$$\boldsymbol{\beta}_1 = \mu X^T \mathbf{y}. \tag{5.25}$$

Hence $\boldsymbol{\beta}_1 \in \mathrm{sp}\left(X^T\right)$. Then $\boldsymbol{\beta}_2$ given by $\boldsymbol{\beta}_2 = \boldsymbol{\beta}_1 + \mu X^T\left(\mathbf{y} - X\boldsymbol{\beta}_1\right)$ is a sum of two vectors in $\mathrm{sp}\left(X^T\right)$. Since $\mathrm{sp}\left(X^T\right)$ is a linear subspace (see Lemma C.11), $\boldsymbol{\beta}_2 \in \mathrm{sp}\left(X^T\right)$, too. By this token $\boldsymbol{\beta}_t \in \mathrm{sp}\left(X^T\right)$ for every iteration. Hence the limit satisfies

$$\lim_{t\to+\infty} \boldsymbol{\beta}_t = \boldsymbol{\beta}_c \in \mathrm{sp}\left(X^T\right),$$

since $\mathrm{sp}\left(X^T\right)$ is a closed subspace (see Lemma C.11). Due to Proposition 4.2, $\boldsymbol{\beta}_c$ satisfies the normal equations. When $k > n$, there are many solutions of this linear system of equations. It is shown in Proposition C.12 that every solution of the normal equations satisfies (C.12). Hence it holds also for $\boldsymbol{\beta}_c$ that

$$\boldsymbol{\beta}_c = X^+\mathbf{y} + \left(\mathbb{I}_k - X^+X\right)\mathbf{z}. \tag{5.26}$$

Here $X^+\mathbf{y}$ is a particular solution of the normal equations and $\left(\mathbb{I}_k - X^TX\right)\mathbf{z} \in \mathcal{N}(X^TX)$. $X^+\mathbf{y}$ is characterized in the Lemma C.14. Clearly $X^+\mathbf{y} \in \mathrm{sp}(X^+)$. Proposition C.9 gives that $\mathrm{sp}\left(X^T\right) = \mathrm{sp}\left(X^+\right)$. Thus $X^+\mathbf{y} \in \mathrm{sp}\left(X^T\right)$, too.
X^+X is a projector matrix according to Definition B.2, as is checked in the proof of Proposition C.10. The matrix X^+X projects onto $\mathrm{sp}\left(X^T\right)$ in view of Proposition C.10. By lemma B.3, $\mathrm{sp}\left(X^T\right) = \mathrm{sp}\left(X^TX\right)$. Since $\mathcal{N}(X^TX)$ and $\mathrm{sp}\left(X^TX\right)$ are orthogonal complements, see [11, p. 244], it follows from (5.26) that

$$\boldsymbol{\beta}_c = X^+\mathbf{y} = \boldsymbol{\beta}^+.$$

Proposition B.5 shows that X^+X is a unique projector onto $\mathrm{sp}\left(X^+\right)$. Hence there is no other particular solution defined by some other generalized inverse that could be used in (5.26), if $\boldsymbol{\beta}_c \in \mathrm{sp}\left(X^T\right)$. ■

We consider certain special cases of the preceding Proposition.

Corollary: $\mu \leq \frac{1}{\lambda_+(X^TX)}$. $\boldsymbol{\beta}_0 = \mathbf{0}_k$. If the regressor matrix has full column rank, $k < n$, then it holds for the GD in (3.6) that

$$\lim_{t\to+\infty} \boldsymbol{\beta}_t = \widehat{\boldsymbol{\beta}} := \left(X^TX\right)^{-1} X^T\mathbf{y}. \tag{5.27}$$

■

Proof: In view of (C.3)

$$\boldsymbol{\beta}^+ = X^+\mathbf{y} = \left(X^TX\right)^+ X^T\mathbf{y}.$$

We can check that this solves the normal equations (4.24), as it should, by means of Proposition C.7. If X^TX has full column rank, the BMP inverse $\left(X^TX\right)^+$ is the inverse $\left(X^TX\right)^{-1}$, as checked in section C.1. ■

It should be noted that $\widehat{\boldsymbol{\beta}}$ in (5.27) is in fact a minimum norm least squares solution, since in this case the normal equations (4.24) have a unique solution. We have also that the Hessian of $Q\left(\boldsymbol{\beta}\right)$ is

$$\mathcal{H}\left(\mathbf{x}\right) = X^TX \tag{5.28}$$

by (3.5). If X has full column rank, then $\mathcal{H}(\mathbf{x})$ is positive definite, and this tells also that a critical point is a local minimum of $Q(\boldsymbol{\beta})$, c.f. Proposition D.1.
The rank of an $n \times k$ matrix is the size of the largest invertible square matrix that can be found inside X. When $k > n$, the column rank of X cannot be full, since the row rank equals the column rank. Hence the rank of the $k \times k$ matrix $X^T X$ is smaller than k and the matrix is not invertible and there are many solutions to the normal equations.

Corollary: $\mu \leq \frac{1}{\lambda_+(X^T X)}$. $\boldsymbol{\beta}_0 = \mathbf{0}_k$. $k > n$ and the the regressor matrix X has full row rank. Then it holds for the GD in (3.6) that

$$\lim_{t \to +\infty} \boldsymbol{\beta}_t = \boldsymbol{\beta}^+ = X^T (X X^T)^{-1} \mathbf{y}. \tag{5.29}$$

■

Proof: When X has full row rank, *rank* $X = n$, then by (C.10) the BMP inverse is

$$X^+ = X^T (X X^T)^{-1}$$

and

$$\boldsymbol{\beta}^+ := X^T (X X^T)^{-1} \mathbf{y}.$$

It is checked in section C.3.4 that $\boldsymbol{\beta}^+$ solves the normal equations. ■

By Proposition C.13 GD converges to an interpolation of the training set. The intriguing implications of this are discussed in [6] and [10].

6 Rate of Convergence of GD for Least Squares Regression

6.1 Inequalities

Let $\boldsymbol{\beta}_{\min} = \text{argmin}_{\boldsymbol{\beta}} Q(\boldsymbol{\beta})$. Since the Hessian of $Q(\boldsymbol{\beta})$, $X^T X$, is in view of Lemma B.1, positive definite or positive semidefinite we get from (D.6) for $\boldsymbol{\beta}_{\min}$ and any $\boldsymbol{\beta}$ that

$$Q(\boldsymbol{\beta}_{\min}) \geq Q(\boldsymbol{\beta}) + \nabla Q(\boldsymbol{\beta})^T (\boldsymbol{\beta}_{\min} - \boldsymbol{\beta}). \tag{6.30}$$

This is rearranged as

$$Q(\boldsymbol{\beta}) \leq Q(\boldsymbol{\beta}_{\min}) + \nabla Q(\boldsymbol{\beta})^T (\boldsymbol{\beta} - \boldsymbol{\beta}_{\min}). \tag{6.31}$$

Let us recall (4.14), i.e.,

$$Q(\boldsymbol{\beta}_{t+1}) \leq Q(\boldsymbol{\beta}_t) + \nabla Q(\boldsymbol{\beta}_t)^T (\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}_t) + \frac{\lambda_+(X^T X)}{2} \parallel \boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}_t \parallel^2 .$$

By the GD iteration and rules of norm,

$$\lambda_+(X^T X) \parallel \boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}_t \parallel^2 = \mu^2 \frac{\lambda_+(X^T X)}{2} \parallel \nabla Q(\boldsymbol{\beta}_t) \parallel^2 .$$

When we take $\mu \leq \frac{1}{\lambda_+(X^TX)}$, we get $\mu^2 \frac{\lambda_+\left(X^TX\right)}{2} = \frac{\mu}{2} \cdot \left(\mu\lambda_+\left(X^TX\right)\right) \leq \frac{\mu}{2}$, and

$$Q\left(\boldsymbol{\beta}_{t+1}\right) \leq Q\left(\boldsymbol{\beta}_t\right) + \nabla Q\left(\boldsymbol{\beta}_t\right)^T\left(\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}_t\right) + \frac{1}{2}\mu \parallel \nabla Q\left(\boldsymbol{\beta}_t\right) \parallel^2 .$$

Furthermore the GD iteration gives again

$$\begin{aligned} Q\left(\boldsymbol{\beta}_{t+1}\right) &\leq Q\left(\boldsymbol{\beta}_t\right) - \mu\nabla Q\left(\boldsymbol{\beta}_t\right)^T Q\left(\boldsymbol{\beta}_t\right) + \frac{1}{2}\mu \parallel \nabla Q\left(\boldsymbol{\beta}_t\right) \parallel^2 . \\ &= Q\left(\boldsymbol{\beta}_t\right) - \frac{1}{2}\mu \parallel \nabla Q\left(\boldsymbol{\beta}_t\right) \parallel^2 . \end{aligned}$$

When we use (6.31) in the right hand side of the latest inequality above, we have

$$Q\left(\boldsymbol{\beta}_{t+1}\right) \leq Q\left(\boldsymbol{\beta}_{\min}\right) + \nabla Q\left(\boldsymbol{\beta}\right)^T\left(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{\min}\right) - \frac{1}{2}\mu \parallel \nabla Q\left(\boldsymbol{\beta}_t\right) \parallel^2 .$$

Hence

$$Q\left(\boldsymbol{\beta}_{t+1}\right) - Q\left(\boldsymbol{\beta}_{\min}\right) \leq \frac{1}{2\mu}\left[2\mu\nabla Q\left(\boldsymbol{\beta}\right)^T\left(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{\min}\right) - \mu^2 \parallel \nabla Q\left(\boldsymbol{\beta}_t\right) \parallel^2\right]. \tag{6.32}$$

The right hand side is next rewritten:

$$2\mu\nabla Q\left(\boldsymbol{\beta}\right)^T\left(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{\min}\right) - \mu^2 \parallel \nabla Q\left(\boldsymbol{\beta}_t\right) \parallel^2 =$$

$$2\mu\nabla Q\left(\boldsymbol{\beta}\right)^T\left(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{\min}\right) - \mu^2 \parallel \nabla Q\left(\boldsymbol{\beta}_t\right) \parallel^2 + \parallel \boldsymbol{\beta}_t - \boldsymbol{\beta}_{\min} \parallel^2 - \parallel \boldsymbol{\beta}_t - \boldsymbol{\beta}_{\min} \parallel^2 .$$

Then one observes that

$$\parallel \boldsymbol{\beta}_t - \mu\nabla Q\left(\boldsymbol{\beta}_t\right) - \boldsymbol{\beta}_{\min} \parallel^2 = \parallel \boldsymbol{\beta}_t - \boldsymbol{\beta}_{\min} \parallel^2 - 2\mu\nabla Q\left(\boldsymbol{\beta}_t\right)^T\left(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{\min}\right) + \mu^2 \parallel \nabla Q\left(\boldsymbol{\beta}_t\right) \parallel^2$$

When we insert this in the right hand side of (6.32) we obtain the inequality

$$Q\left(\boldsymbol{\beta}_{t+1}\right) - Q\left(\boldsymbol{\beta}_{\min}\right) \leq \frac{1}{2\mu}\left[\parallel \boldsymbol{\beta}_t - \boldsymbol{\beta}_{\min} \parallel^2 - \parallel \boldsymbol{\beta}_t - \mu\nabla Q\left(\boldsymbol{\beta}_t\right) - \boldsymbol{\beta}_{\min} \parallel^2\right].$$

By the GD iteration $\boldsymbol{\beta}_{t+1} = \boldsymbol{\beta}_t - \mu\nabla Q\left(\boldsymbol{\beta}_t\right)$ and this gives

$$Q\left(\boldsymbol{\beta}_{t+1}\right) - Q\left(\boldsymbol{\beta}_{\min}\right) \leq \frac{1}{2\mu}\left[\parallel \boldsymbol{\beta}_t - \boldsymbol{\beta}_{\min} \parallel^2 - \parallel \boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}_{\min} \parallel^2\right]. \tag{6.33}$$

6.2 Linear Convergence Rate

Now we use the final inequality (6.33) above to prove the following inequality.

Proposition 6.1: *If one runs GD for least squares for linear regression k times with $\mu \leq 1/\lambda_+\left(X^TX\right)$, then it holds that*

$$Q\left(\boldsymbol{\beta}_k\right) - Q\left(\boldsymbol{\beta}_{\min}\right) \leq \frac{1}{2k\mu} \parallel \boldsymbol{\beta}_0 - \boldsymbol{\beta}_{\min} \parallel^2 . \tag{6.34}$$

Proof: Since the inequality (6.33) holds for every iteration of GD,

$$\sum_{t=1}^{k} \left(Q\left(\boldsymbol{\beta}_t\right) - Q\left(\boldsymbol{\beta}_{\min}\right)\right) \leq \frac{1}{2\mu} \sum_{t=1}^{k} \left[\| \boldsymbol{\beta}_{t-1} - \boldsymbol{\beta}_{\min} \|^2 - \| \boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}_{\min} \|^2\right]. \tag{6.35}$$

In view of (4.19) we have

$$k\left(Q\left(\boldsymbol{\beta}_k\right) - Q\left(\boldsymbol{\beta}_{\min}\right)\right) \leq \sum_{t=1}^{k} \left(Q\left(\boldsymbol{\beta}_t\right) - Q\left(\boldsymbol{\beta}_{\min}\right)\right).$$

The sum in the right hand side of (6.35) is again 'telescoping' and this gives

$$\begin{aligned} \sum_{t=1}^{k} \left[\| \boldsymbol{\beta}_{t-1} - \boldsymbol{\beta}_{\min} \|^2 - \| \boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}_{\min} \|^2\right] &= \| \boldsymbol{\beta}_o - \boldsymbol{\beta}_{\min} \|^2 - \| \boldsymbol{\beta}_{k+1} - \boldsymbol{\beta}_{\min} \|^2 \\ &< \| \boldsymbol{\beta}_o - \boldsymbol{\beta}_{\min} \|^2 . \end{aligned}$$

Hence the inequality (6.34) has been established. ■

The convergence rate in (6.34) is called linear and depends on the distance between the initial guess and the minimizer.

6.3 An Alternative Analysis for X with Full Column Rank

When X^TX is invertible, an alternative analysis of convergence of GD is possible. The presentation below is due to [2, Chapter 3]. We have $\widehat{\boldsymbol{\beta}} = (X^TX)^{-1}X^T\mathbf{y}$. From (3.6) we get

$$\boldsymbol{\beta}_{t+1} - \widehat{\boldsymbol{\beta}} = \left(\mathbb{I}_k - \mu X^TX\right)\left(\boldsymbol{\beta}_t - \widehat{\boldsymbol{\beta}}\right). \tag{6.36}$$

To see this note that

$$\left(\mathbb{I}_k - \mu X^TX\right)\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}} - \mu X^TX(X^TX)^{-1}X^T\mathbf{y} = \widehat{\boldsymbol{\beta}} - \mu X^T\mathbf{y},$$

i.e.

$$\widehat{\boldsymbol{\beta}} = \left(\mathbb{I}_k - \mu X^TX\right)\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}} + \mu X^T\mathbf{y} \tag{6.37}$$

With (3.6) and when (6.37) is inserted in right hand side of

$$\boldsymbol{\beta}_{t+1} - \widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}_t + \mu X^T\left(\mathbf{y} - X\boldsymbol{\beta}_t\right) - \widehat{\boldsymbol{\beta}},$$

we get (6.36) as stated. After p iterations of GD in (6.36) we have

$$\boldsymbol{\beta}_p - \widehat{\boldsymbol{\beta}} = \left(\mathbb{I}_k - \mu X^TX\right)^p\left(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}\right). \tag{6.38}$$

Thus by the consistency property (3.9)

$$\| \boldsymbol{\beta}_p - \widehat{\boldsymbol{\beta}} \| \leq \| \left(\mathbb{I}_k - \mu X^TX\right)^p \| \| \boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}} \| . \tag{6.39}$$

Here by we have by Proposition 3.2 that

$$\parallel \left(\mathbb{I}_k - \mu X^T X\right)^p \parallel =$$

the square root of the largest eigenvalue of $((\mathbb{I}_k - \mu X^T X)^p)^T \left(\mathbb{I}_k - \mu X^T X\right)^p$.
But

$$((\mathbb{I}_k - \mu X^T X)^p)^T \left(\mathbb{I}_k - \mu X^T X\right)^p = ((\mathbb{I}_k - \mu X^T X)^T)^p \left(\mathbb{I}_k - \mu X^T X\right)^p .$$

Here $\left(\mathbb{I}_k - \mu X^T X\right)^T = \left(\mathbb{I}_k - \mu X^T X\right)$. Hence $((\mathbb{I}_k - \mu X^T X)^p)^T \left(\mathbb{I}_k - \mu X^T X\right)^p = \left(\mathbb{I}_k - \mu X^T X\right)^{2p}$.
Let $\mathbf{e}_i$ and λ_i be the i:th eigenvector and eigenvalue of $X^T X$. Then

$$\left(\mathbb{I}_k - \mu X^T X\right) \mathbf{e}_i = \mathbf{e}_i - \mu X^T X \mathbf{e}_i = (1 - \mu\lambda_i)\mathbf{e}_i.$$

If λ_1 is the smallest eigenvalue of $X^T X$, then for every i

$$1 - \mu\lambda_i \leq 1 - \mu\lambda_1.$$

Hence $(1-\mu\lambda_1)^{2p} = (1-\mu\lambda_1)^{2p}$ is the largest eigenvalue of $\left(\mathbb{I}_k - \mu X^T X\right)^{2p}$ and

$$\parallel \left(\mathbb{I}_k - \mu X^T X\right)^p \parallel = \sqrt{(1-\mu\lambda_1)^{2p}} = \mid (1-\mu\lambda_1)^p \mid$$

Now we make a theoretically advantageous selection of μ by taking

$$\mu = \frac{2}{\lambda_+ + \lambda_1},$$

where λ_+ is the largest eigenvalue of $X^T X$. This gives

$$1 - \mu\lambda_1 = 1 - \frac{2\lambda_1}{\lambda_+ + \lambda_1} = \frac{\lambda_+ - \lambda_1}{\lambda_+ + \lambda_1} = \frac{\frac{\lambda_+}{\lambda_1} - 1}{\frac{\lambda_+}{\lambda_1} + 1}.$$

Since $X^T X$ is invertible, $\lambda_1 > 0$, and $\frac{\lambda_+}{\lambda_1} - 1 > 0$, hence $\mid (1-\mu\lambda_1)^{p]} \mid = (1-\mu\lambda_1)^p$.
Let us set $c = \frac{\lambda_+}{\lambda_1}$(=condition number of $X^T X$, see [18, pp. 366−367]). Hence we insert in (6.38) to get

$$\parallel \boldsymbol{\beta}_p - \widehat{\boldsymbol{\beta}} \parallel \leq \left(\frac{c-1}{c+1}\right)^p \parallel \boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}} \parallel . \tag{6.40}$$

Here $0 < \frac{c-1}{c+1} < 1$. Hence $\left(\frac{c-1}{c+1}\right)^p \to 0$, as $p \to +\infty$, i.e. GD converges to $\widehat{\boldsymbol{\beta}}$ at an exponential (or linear) rate.
A consequence of (6.40) is that if one desires to have an error $\parallel \boldsymbol{\beta}_p - \widehat{\boldsymbol{\beta}} \parallel < \epsilon$, then this happens, when

$$\left(\frac{c-1}{c+1}\right)^p \parallel \boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}} \parallel < \epsilon,$$

i.e,

$$\left(\frac{c-1}{c+1}\right)^p < \frac{\epsilon}{\parallel \boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}} \parallel}$$

i.e,

$$p \ln\left(\frac{c-1}{c+1}\right) < C_1 \ln \epsilon.$$

where $C_1 = 1/ \parallel \boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}} \parallel$. But $\ln\left(\frac{c-1}{c+1}\right) < 0$, as $0 < \frac{c-1}{c+1} < 1$ and thus

$$p > C_1 \ln \epsilon / \ln\left(\frac{c-1}{c+1}\right),$$

Here we note that if $x < 0$, then $x = (-1)|x|$ and thus with $C_2 = C_1/|\ln\left(\frac{c-1}{c+1}\right)|$

$$p > C_1 \ln \epsilon/(-1)|\ln\left(\frac{c-1}{c+1}\right)| = C_2(-1)\ln \epsilon = C_2 \ln \frac{1}{\epsilon}.$$

Hence it suffices to to make $\approx \ln(1/\epsilon)$ iterations of GD, whereby one ignores the effect of the condition number and the initial error.

7 Discussion

Let us note that by Lemma C.14, the norm $\parallel \widehat{\boldsymbol{\beta}}^+ \parallel$ is the smallest among all solutions of the normal equations. Hence, if k is high, $\boldsymbol{\beta}^+$ should have many small or zero components, it is a regularizing least squares solution. Hence, stopped GD (i.e.. the GD iterations are stopped before convergence) has been investigated for strong connections to ridge regression. For ridge regression see [12, pp. 322−314]. The question is to find rules of early stopping, where one does not wait for full convergence.

In practice there are issues of selection of the learning rate, see e.g. [13]. There are several optimized GD algorithms with names like momentum, adagrad, nesterov accelerated gradient, RMSprop.

There are dozens of tutorials, guised in elementary mathematics, for GD and SGD in the world wide web, see, e.g., the following specialized to linear regression:

https://machinelearningspace.com/a-comprehensive-guide-to-gradient-descent-algorithm/#multi-regression.

Let us finally quote David Donoho [5, p. 17]:

> *The $k > n$ case is not anomalous; it is in some sense (nowadays) the generic case. For many types of event we can think of, we have the potential of a very large number of measurables quantifying that event, and a relatively few instances of that event.*
>
> - *Many genes, relatively few patients with a given genetic disease.*
> - *Many samples of a persons' speech, relatively few speakers sampled.*

As seen above, GD provides a method that works for $k > n$.
Textbooks in linear regression analysis are not responding to the current state of affairs, not even in the recent latest editions, and are harboring the full column rank case, see for example [12]. An exception is [11], where regression models are from the start treated by generalized inverses, GD is not taken up in this text.

Appendices

A Proof of Proposition 3.1

Proof: of Theorem 3.1 follows. We write

$$\mathbf{y} - X\boldsymbol{\beta} = \mathbf{y} - X\boldsymbol{\beta}_{\mathrm{sol}} + X\boldsymbol{\beta}_{\mathrm{sol}} - X\boldsymbol{\beta},$$

where $\boldsymbol{\beta}_{\mathrm{sol}}$ is any solution to $\nabla_{\boldsymbol{\beta}} Q\left(\boldsymbol{\beta}\right) = \mathbf{0}_{k+1}$ and by (3.5)

$$\nabla_{\boldsymbol{\beta}} Q\left(\boldsymbol{\beta}_{\mathrm{sol}}\right) = \mathbf{0}_{k+1} \Leftrightarrow X^T X\boldsymbol{\beta}_{\mathrm{sol}} = X^T\mathbf{y}. \tag{A.1}$$

Let us set for economy of expression $U := \mathbf{y} - X\boldsymbol{\beta}_{\mathrm{sol}}$ and $V := X\left(\boldsymbol{\beta}_{\mathrm{sol}} - \boldsymbol{\beta}\right)$. Then

$$\begin{aligned} Q\left(\boldsymbol{\beta}\right) &= \frac{1}{2} \parallel \mathbf{y} - X\boldsymbol{\beta} \parallel^2 = \frac{1}{2}\left(\mathbf{y} - X\boldsymbol{\beta}\right)^T\left(\mathbf{y} - X\boldsymbol{\beta}\right) \\ &= \frac{1}{2}\left[(U+V)^T(U+V) = (U^T+V^T)(U+V)\right[\\ &= \frac{1}{2}\left[U^TU + U^TV + V^TU + V^TV\right]. \end{aligned}$$

Set $\mathbf{e}_{\mathrm{sol}} := \mathbf{y} - X\boldsymbol{\beta}_{\mathrm{sol}}$. Thus

$$U^TU = \left(\mathbf{y} - X\boldsymbol{\beta}_{\mathrm{sol}}\right)^T\left(\mathbf{y} - X\boldsymbol{\beta}_{\mathrm{sol}}\right) = \mathbf{e}_{\mathrm{sol}}^T\mathbf{e}_{\mathrm{sol}}.$$

Next

$$V^TV = \left(X\boldsymbol{\beta}_{\mathrm{sol}} - \boldsymbol{\beta}\right)^T X\left(\boldsymbol{\beta}_{\mathrm{sol}} - \boldsymbol{\beta}\right) = \left(\boldsymbol{\beta}_{\mathrm{sol}} - \boldsymbol{\beta}\right)^T X^T X\left(\boldsymbol{\beta}_{\mathrm{sol}} - \boldsymbol{\beta}\right).$$

In the above, $U^TV = V^TU$, since these are scalar products. Let us expand U^TV.

$$\left(\mathbf{y} - X\boldsymbol{\beta}_{\mathrm{sol}}\right)^T X\left(\boldsymbol{\beta}_{\mathrm{sol}} - \boldsymbol{\beta}\right) = \left(X^T\mathbf{y} - X^TX\boldsymbol{\beta}_{\mathrm{sol}}\right)^T\left(\boldsymbol{\beta}_{\mathrm{sol}} - \boldsymbol{\beta}\right).$$

By (A.1) we have $X^TX\boldsymbol{\beta}_{\mathrm{sol}} = X^T\mathbf{y}$ and thus the last expression equals zero, since

$$\left(X^T\mathbf{y} - X^TX\boldsymbol{\beta}_{\mathrm{sol}}\right)^T\left(\boldsymbol{\beta}_{\mathrm{sol}} - \boldsymbol{\beta}\right) = \mathbf{0}_k^T\left(\boldsymbol{\beta}_{\mathrm{sol}} - \boldsymbol{\beta}\right) = 0.$$

Hence we have established the following decomposition

$$Q\left(\boldsymbol{\beta}\right) = \frac{1}{2} \parallel \mathbf{y} - X\boldsymbol{\beta} \parallel^2 = \frac{1}{2}\left[\mathbf{e}_{\mathrm{sol}}^T\mathbf{e}_{\mathrm{sol}} + \left(\boldsymbol{\beta}_{\mathrm{sol}} - \boldsymbol{\beta}\right)^T X^TX\left(\boldsymbol{\beta}_{\mathrm{sol}} - \boldsymbol{\beta}\right)\right]. \tag{A.2}$$

Next we apply the lemma B.1 to X, which implies that X^TX is a positive semidefinite matrix, and thus

$$\left(\boldsymbol{\beta}_{\mathrm{sol}} - \boldsymbol{\beta}\right)^T X^TX\left(\boldsymbol{\beta}_{\mathrm{sol}} - \boldsymbol{\beta}\right) \geq 0.$$

Hence

$$Q\left(\boldsymbol{\beta}\right) \geq Q\left(\boldsymbol{\beta}_{\mathrm{sol}}\right) = \mathbf{e}_{\mathrm{sol}}^T\mathbf{e}_{\mathrm{sol}} > 0$$

and the Proposition is proved as asserted. ■

Corollary A.1: *If X has full column rank, then*

$$\widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{y} \tag{A.3}$$

is the unique minimizer of $Q(\boldsymbol{\beta})$.

Proof: $\widehat{\boldsymbol{\beta}}$ is a solution of the normal equations. But if X has full column rank, then Lemma B.1 says that $X^T X$ is positive definite. Hence the quadratic form in the right hand side of (A.2)

$$\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^T X^T X \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) = 0,$$

if and only if $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}$. ■

B Matrix Calculus

B.1 Vector, Scalar product, Norm

$\mathbf{x}$ is an $n \times 1$ vector

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n \quad \mathbf{x}^T = (x_1, x_2, \ldots x_n) \in \mathbb{R}^n$$

$\mathbf{x}^T$ is a $1 \times n$ vector, the transpose of $\mathbf{x}$.

- For $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^n$, the scalar product

$$\mathbf{x}^T \mathbf{y} := \sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n} y_i x_i = \mathbf{y}^T \mathbf{x}$$

$$\mathbf{0}_n = \text{the } n \times 1 \text{ vector with all } n \text{ components} = 0$$

- $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^n$ are called *orthogonal*, if

$$\mathbf{x}^T \mathbf{y} = 0.$$

- The norm $\parallel \mathbf{x} \parallel$ in $\mathbb{R}^n$ is defined by

$$\parallel \mathbf{x} \parallel = \sqrt{\mathbf{x}^T \mathbf{x}} = \sqrt{\sum_{i=1}^{n} x_i^2}. \tag{B.1}$$

  It has the following properties

  1. $\parallel \mathbf{x} \parallel \geq 0$

2. $\parallel \mathbf{x} \parallel = 0$ if and only if $\mathbf{x} = \mathbf{0}_n$.
3. $\parallel a\mathbf{x} \parallel = |a| \parallel \mathbf{x} \parallel$, where a is a scalar.
4. $\parallel \mathbf{x} + \mathbf{y} \parallel \leq \parallel \mathbf{x} \parallel + \parallel \mathbf{y} \parallel$ triangle inequality.

- Distance between $\mathbf{x}$ and $\mathbf{y}$

$$\parallel \mathbf{x} - \mathbf{y} \parallel = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

$$\parallel \mathbf{x} - \mathbf{y} \parallel = 0 \Leftrightarrow \mathbf{x} = \mathbf{y}$$

- $\mathbf{x}$ and $\mathbf{y}$ are two column vectors in $\mathbb{R}^n$. The **Cauchy-Schwartz inequality** is

$$|\mathbf{x}^T\mathbf{y}| \leq \parallel \mathbf{x} \parallel \parallel \mathbf{y} \parallel . \tag{B.2}$$

  Equality holds in (B.2) if and only if $\mathbf{x}$ and $\mathbf{y}$ are linearly dependent.

B.2 Matrices and Matrix Rules

B.2.1 Transpose, Inverse

- A and B conformal, $(AB)^T = B^TA^T$, $(A^T)^T = A$. If A and B are invertible, $(AB)^{-1} = B^{-1}A^{-1}$.

- A and B conformal, $(A+B)^T = A^T + B^T$

- The $n \times n$ identity matrix is

$$\mathbb{I}_n = \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & \ldots & 0 \\ 0 & \ddots & \vdots & \ldots & 0 \\ 0 & 0 & 0 & \ldots & 1 \end{pmatrix} . \tag{B.3}$$

- A is $n \times n$ and invertible.

$$\left(A^T\right)^{-1} = \left(A^{-1}\right)^T . \tag{B.4}$$

  Proof: $A^T\left(A^{-1}\right)^T = \left(A^{-1}A\right)^T = \mathbb{I}_n^T = \mathbb{I}_n$ and $\left(A^{-1}\right)^T A^T = \left(AA^{-1}\right)^T = \mathbb{I}_n^T = \mathbb{I}_n$. Hence,

$$\left((X^TX)^{-1}\right)^T = (X^TX)^{-1} \tag{B.5}$$

■

- $n \times n$ matrix A is *positive semidefinite* if

$$\mathbf{x}^TA\mathbf{x} \geq 0 \quad \text{for all } \mathbf{x}$$

  $n \times n$ matrix A is *positive definite* if

$$\mathbf{x}^TA\mathbf{x} > 0 \quad \text{for all } \mathbf{x} \neq \mathbf{0}_n.$$

B.2.2 Covariance Matrix

Covariance matrix (also denoted by $C_{\mathbf{X}}$)

$$C := E\left[(\mathbf{X}-\boldsymbol{\mu}_{\mathbf{X}})(\mathbf{X}-\boldsymbol{\mu}_{\mathbf{X}})^T\right],$$

where the array in position (i,j) is

$$c_{ij} = E\left[(X_i-\mu_i)(X_j-\mu_j)\right]$$

is the covariance of X_i and X_j. The variances of the components of $\mathbf{X}$ are the elements on the main diagonal, i.e.,

$$c_{ii} = E\left[(X_i-\mu_i)^2\right] = \mathrm{Var}\,(X_i) = \sigma_i^2.$$

$$X \text{ and } Y \text{ are independent } \Rightarrow \mathrm{Cov}(X,Y)=0.$$

B.2.3 Linear Transformations of Covariance Matrices

$$E\left[\mathbf{X}+\mathbf{Y}\right] = \boldsymbol{\mu}_{\mathbf{X}} + \boldsymbol{\mu}_{\mathbf{Y}} \tag{B.6}$$

If $\mathbf{Z} = A\mathbf{X}+\mathbf{b}$, then

$$E\left[\mathbf{Z}\right] = A\boldsymbol{\mu}_{\mathbf{X}} + \mathbf{b}, \tag{B.7}$$

and

$$C_{\mathbf{Z}} = AC_{\mathbf{X}}A^T. \tag{B.8}$$

$$C_{\mathbf{X}} = E\left[\mathbf{X}\mathbf{X}^T\right] - \boldsymbol{\mu}_{\mathbf{X}}\boldsymbol{\mu}_{\mathbf{X}}^T \tag{B.9}$$

$$E\left[\mathbf{X}^T\mathbf{Y}\right] = trE\left[\mathbf{X}\mathbf{Y}^T\right]. \tag{B.10}$$

$$\mathrm{Var}\left[\mathbf{a}^T\mathbf{X}\right] = \mathbf{a}^T C_{\mathbf{X}}\mathbf{a} \tag{B.11}$$

B.2.4 Sherman-Morrison-Woodbury Theorem

A is an invertible square $n\times n$ matrix and $\mathbf{u},\mathbf{v}\in\mathbb{R}^n$ are column vectors. Then $A+\mathbf{u}\mathbf{v}^{\mathsf{T}}$ is invertible iff $1+\mathbf{v}^{\mathsf{T}}A^{-1}\mathbf{u}\neq 0$. In this case,

$$\left(A+\mathbf{u}\mathbf{v}^{\mathsf{T}}\right)^{-1} = A^{-1} - \frac{A^{-1}\mathbf{u}\mathbf{v}^{\mathsf{T}}A^{-1}}{1+\mathbf{v}^{\mathsf{T}}A^{-1}\mathbf{u}}. \tag{B.12}$$

B.2.5 Trace of a Square Matrix

Let A be a square matrix. The **trace** trA of A is the sum of the entries in main diagonal:

$$tr\begin{pmatrix} a_{11} & \cdots & a_{1k} \\ \vdots & \ddots & \vdots \\ a_{k1} & \cdots & a_{kk} \end{pmatrix} = \textstyle\sum_1^k a_{jj}$$

- 1.If A is a $k\times n$-matrix, and B an $n\times k$-matrix, then $tr(AB)=tr(BA)$

- 2. In particular, if a is a column-vector, then $a^T a = tr\left(aa^T\right)$.
- 3. For any real numbers a and b, $tr(aC + bD) = atrC + btrD$
- 4. If A is an $n \times n$ with the eigenvalues $(\lambda_i)_{i=1}^n$, then $trA = \sum_{i=1}^n \lambda_i$, [14, Thm 9.1, p. 280].

  Example: A is $n \times n$ and positive definite. Its eigenvalue λ_i corresponds to the eigenvector $\mathbf{e}_i$. Then

  $$0 < \mathbf{e}_i^T A \mathbf{e}_i = \lambda_i \mathbf{e}_i^T \mathbf{e}_i = \lambda_i,$$

  since eigenvctors are orthonormal. Hence

  $$trA = \sum_{i=1}^n \lambda_i > 0. \tag{B.13}$$

  ■

B.2.6 The Range Space (=Column Space) of a Matrix

Let A be any $n \times p$ matrix. Then the range space of A is defined by

$$\mathrm{sp}\,(A) = \left\{\mathbf{x} \in \mathbb{R}^n | \text{there exists a } \mathbf{b} \in \mathbb{R}^p \text{ such that } \mathbf{x} = A\mathbf{b}\right\}. \tag{B.14}$$

B.3 Rayleigh's Principle

Let us write

$$\|A\mathbf{x}\| = \sqrt{(A\mathbf{x})^T A\mathbf{x}} = \sqrt{\mathbf{x}^T A^T A \mathbf{x}} = \sqrt{\mathbf{x}^T B \mathbf{x}} \quad \text{and} \quad \|\mathbf{x}\| = \sqrt{\mathbf{x}^T \mathbf{x}},$$

where $B = A^T A$. Let us consider the ratio

$$\rho\,(\mathbf{x}) := \frac{\mathbf{x}^T B \mathbf{x}}{\mathbf{x}^T \mathbf{x}}. \tag{B.15}$$

By Lemma B.1, B is a symmetric positive semidefinite square matrix, say $n \times n$. Thus B has n orthonormal eigenvectors $\mathbf{e}_i$ and corresponding real and non-negative eigenvalues ordered as $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$, see [18, pp. 294−296]. An arbitrary vector $\mathbf{x} \in \mathbb{R}^n$ has the expansion

$$\mathbf{x} = \sum_{i=1}^n \alpha_i \mathbf{e}_i.$$

Then

$$B\mathbf{x} = \sum_{i=1}^n \alpha_i B \mathbf{e}_i = \sum_{i=1}^n \alpha_i \lambda_i \mathbf{e}_i$$

and

$$\mathbf{x}^T B \mathbf{x} = \sum_{j=1}^n \sum_{i=1}^n \alpha_j \alpha_i \lambda_i \mathbf{e}_j^T \mathbf{e}_i, \quad \mathbf{x}^T \mathbf{x} = \sum_{j=1}^n \sum_{i=1}^n \alpha_j \alpha_i \mathbf{e}_j^T \mathbf{e}_i.$$

Since $\mathbf{e}_i$ are orthonormal, i.e., $\mathbf{e}_j^T\mathbf{e}_i = 0$ if $i \neq j$ and $= 1$, if $i = j$, we have

$$\rho(\mathbf{x}) := \frac{\sum_{i=1}^n \alpha_i^2 \lambda_i}{\sum_{i=1}^n \alpha_i^2}.$$

Thus

$$\lambda_n - \rho(\mathbf{x}) = \frac{\lambda_n \sum_{i=1}^n \alpha_i^2 - \sum_{i=1}^n \alpha_i^2 \lambda_i}{\sum_{i=1}^n \alpha_i^2} = \frac{\sum_{i=1}^n \alpha_i^2 (\lambda_n - \lambda_i)}{\sum_{i=1}^n \alpha_i^2}.$$

Since $\lambda_n - \lambda_i \geq 0$, we have $\lambda_n - \rho(\mathbf{x}) \geq 0$. Hence for all $\mathbf{x} \in \mathbb{R}^n$

$$\rho(\mathbf{x}) \leq \lambda_n.$$

Hence we have also from (B.15)

$$\rho(\mathbf{e}_n) := \frac{\mathbf{e}_n^T B \mathbf{e}_n}{\mathbf{e}_n^T \mathbf{e}_n} = \lambda_n \tag{B.16}$$

This is a part of a statement known as *Rayleigh's principle.* The proof above follows [14, p. 407].

B.4 Properties of $A^T A$

Definition B.1: An $n \times k$ matrix X has **full column rank** as soon as the k columns of X are linearly independent. ■

Lemma B.1: *Let A be any $n \times p$ matrix. Then*

i) $A^T A$ is symmetric positive semidefinite.

ii) If A has full column rank, then $A^T A$ is symmetric and positive definite.

The positive definiteness *ii)* follows from the fact that A and $A^T A$ have the same rank [14, Thm. 5.15]. The null space of A is

$$\mathcal{N}(A) = \{\mathbf{x} \in \mathbb{R}^p \mid A\mathbf{x} = \mathbf{0}_n\}.$$

The range space of A is

$$\mathrm{sp}(A) = \{\mathbf{x} \in \mathbb{R}^n \mid \exists \mathbf{z} \in \mathbb{R}^p \quad \text{s.t.} \quad A\mathbf{z} = \mathbf{x}\}. \tag{B.17}$$

Lemma B.2:

$$\mathcal{N}(A) = \mathcal{N}(A^T A)$$

Proof: Take $\mathbf{x} \in \mathcal{N}(A)$. Then $A\mathbf{x} = \mathbf{0}_n$ and $A^T(A\mathbf{x}) = A^T\mathbf{0}_n = \mathbf{0}_n$. Thus $\mathcal{N}(A) \subseteq \mathcal{N}(A^T A)$.
Take $\mathbf{x} \in \mathcal{N}(A^T A)$. Then $\mathbf{x}^T A^T A\mathbf{x} = 0$. But $\mathbf{x}^T A^T A\mathbf{x} = \parallel A\mathbf{x} \parallel^2 = 0$, i.e., $A\mathbf{x} = \mathbf{0}_n$, and $\mathcal{N}(A^T A) \subseteq \mathcal{N}(A)$. ■

Lemma B.3:

$$\mathrm{sp}(A^T) = \mathrm{sp}(A^T A)$$

Proof: For any matrix B the spaces $\mathrm{sp}(B)$ and $\mathcal{N}(B^T)$ are orthogonal subspaces [18, p. 138]. This is known as the fundamental theorem of linear algebra [18, loc.cit.]. Hence $\mathrm{sp}(A^T A)$ and $\mathcal{N}(A^T A)$ are orthogonal. By the same token $\mathrm{sp}(A^T)$ and $\mathcal{N}(A)$ are orthogonal. Then we note lemma B.2, $\mathcal{N}(A) = \mathcal{N}(A^T A)$, this implies (details omitted) that $\mathrm{sp}(A^T) = \mathrm{sp}(A^T A)$. ■

B.5 Projector Matrix

We cite a definition from [18, p. 157].

Definition B.2: If H is idempotent, $H^2 = H \cdot H = H$, and symmetric, then H is called a projection matrix. ■

It follows that $\mathbb{I}$ is the only invertible projection matrix, since if H^{-1} exists, then

$$H^2 = H \Leftrightarrow H^{-1}H^2 = H^{-1}H = \mathbb{I} \Leftrightarrow H^{-1}H \cdot H = \mathbb{I} \Leftrightarrow H = \mathbb{I}.$$

Hence $\mathbb{I} - H$ is a projection matrix.

Lemma B.4: *It holds that*

i)

$$\mathcal{N}(H) = \mathrm{sp}(\mathbb{I} - H) \tag{B.18}$$

ii)

$$\mathrm{sp}(H) \cap \mathrm{sp}(\mathbb{I} - H) = \mathbf{0}. \tag{B.19}$$

Proof: We have by definitions of the sets involved:

i) Take $\mathbf{x} \in \mathrm{sp}(\mathbb{I} - H)$. Then there is $\mathbf{z}$ such that $\mathbf{x} = (\mathbb{I} - H)\mathbf{z} = \mathbf{z} - H\mathbf{z}$. Thus $H\mathbf{x} = H\mathbf{z} - H^2\mathbf{z} = H\mathbf{z} - H\mathbf{z} = \mathbf{0}$, i.e., $\mathbf{x} \in \mathcal{N}(H)$. Next, take $\mathbf{x} \in \mathcal{N}(H)$. Then $\mathbf{x} = \mathbf{x} - H\mathbf{x} = (\mathbb{I} - H)\mathbf{x}$, i.e., $\mathbf{x} \in \mathrm{sp}(\mathbb{I} - H)$.

ii) If $x \in \mathrm{sp}(H) \cap \mathrm{sp}(\mathbb{I} - H)$. There is $\mathbf{z}_1$ such that $\mathbf{x} = (\mathbb{I} - H)\mathbf{z}_1$ and $\mathbf{z}_2$ such that $\mathbf{x} = H\mathbf{z}_2$. Thus $H\mathbf{z}_2 = (\mathbb{I} - H)\mathbf{z}_1 \Leftrightarrow H\mathbf{z}_2 = H^2\mathbf{z}_2 = H\mathbf{z}_1 - H^2\mathbf{z}_1 = \mathbf{0}$. I.e., $\mathbf{x} = H\mathbf{z}_2 = \mathbf{0}$. ■

Proposition B.5: *If a projector matrix H maps onto the vector space S, then H is unique.*

Proof: Let H and L be two projector matrices mapping onto the vector space S. For any $\mathbf{x}$ we have $L\mathbf{x} \in S$. Then, since H maps onto S, $H(L\mathbf{x}) = HL\mathbf{x} = L\mathbf{x}$, and we have $HL = L$. In the same way we get $LH = H$. Now

$$(H-L)^T(H-L) = (H-L)(H-L) = H^2 - HL - LH + L^2 = H - L - H + L = \mathbf{O},$$

where $\mathbf{O}$ =the zero matrix. But if we have a matrix $C = (c_{ij})$ such that $C^T C = \mathbf{O}$, then on the main diagonal of $C^T C$ are found the squared norms of all columns in C, $\sum_j c_{ij}^2 = 0$. Thus all $c_{ij} = 0$, all columns in C are zero vectors, and $C = \mathbf{O}$. ■

C Generalized Inverses

C.1 On Linear Systems of Equations & Generalized Inverses

Consider the general equation with an $n \times k$ matrix A

$$A\mathbf{x} = \mathbf{z},$$

where we are seeking a $k \times 1$ vector $\mathbf{x}$ as a solution. There are three possibilities, see, e.g., [16, Chapter 4]:

- There is no solution $\Leftrightarrow$ $\mathbf{z} \notin \mathrm{sp}(A)$= the linear span of the columns in A.
- There is a unique solution $\Leftrightarrow$ A is invertible.
- There are many solutions.

We try to find a $k \times n$ matrix G, which would behave as much like A^{-1} is such that if there are many solutions, then $G\mathbf{z}$ is one of them, i.e.,

$$AG\mathbf{z} = \mathbf{z}$$

Proposition C.1: *A $k \times n$ matrix G is called a* generalized inverse *of an $n \times k$ matrix A if any of the following equivalent conditions hold:*

1. *$G\mathbf{z}$ is a solution to $A\mathbf{x} = \mathbf{z}$ if solutions exist.*
2. *GA is idempotent and* rank GA = rank A $\Leftrightarrow$ *AG is idempotent and* rank AG= rank A
3. *$AGA = A$*

A proof of these equivalences is found in [16, p. 106].
Suppose $n = k$ and that the inverse A^{-1} exists. We leftmultiply in $AGA = A$

$$A^{-1}AGA = A^{-1}A \Leftrightarrow GA = \mathbb{I}.$$

We obtain $AG = \mathbb{I}$ by rightmultiplication in $AGA = A$. Thus $G = A^{-1}$.
It can be shown that a generalized inverse always exists,and that it is unique if and only if A^{-1} exists. With $G \mapsto A^{-}$ the set of all generalized inverses of A is denoted by .

$$\{A\}^{-} = \{A^{-} | AA^{-}A = A\} \tag{C.1}$$

Lemma C.2: *If A^{-} is a generalized inverse of A, then $\left(A^{-}\right)^{T}$ is a generalized inverse of A^{T}, or*

$$\left(A^{T}\right)^{-} = \left(A^{-}\right)^{T} \tag{C.2}$$

Proof:

$$A^{T}\left(A^{-}\right)^{T}A^{T} = \left(A^{-}A\right)^{T}A^{T} = \left(AA^{-}A\right)^{T} = A^{T},$$

where the rule 3. in Proposition C.1 was applied. ■

In order to give a first example of a pseudoinverse we .need an auxiliary matrix result.

Lemma C.3:

$$XA = XB \Leftrightarrow X^T XA = X^T XB.$$

Proof: The statement $XA = XB \Rightarrow X^T XA = X^T XB$ is clear. We are to prove that

$$X^T XA = X^T XB \Rightarrow XA = XB$$

By assumption ($\mathbf{O}$ is a matrix of zeroes)

$$\mathbf{O} = X^T X(A - B).$$

Then

$$\mathbf{O} = (A - B)X^T X(A - \mathbb{I}) = (X(A - B)^T X(A - B).$$

But then it follows as in the proof of Proposition B.5 that $X(A - B) = \mathbf{O}$, i.e., XA=XB. ■

Proposition C.4: *X is $n \times k$. A generalized inverse of X is*

$$X^- = \left(X^T X\right)^- X^T, \tag{C.3}$$

where $\left(X^T X\right)^-$ is a generalized inverse of $X^T X$.

Proof: Set $G = \left(X^T X\right)^- X^T$. We must show that $G \in \{X\}^-$, i.e., by (C.1) that $XGX = X$. Here we compute

$$X^T X \left(X^T X\right)^- X^T X = X^T X, \tag{C.4}$$

where the assumption that $G \in \left\{X^T X\right\}^-$ was used, i.e., (C.1) was used. Set now $A := \left(X^T X\right)^- X^T X$. Hence we have in (C.4)

$$X^T XA = X^T X.$$

But then we apply Lemma C.3 with $B = \mathbb{I}$, which gives $XA = X$, or,

$$X \left(X^T X\right)^- X^T X = X,$$

which shows that $\left(X^T X\right)^- X^T \in \{X\}^-$, which is the Proposition as asserted.■

Lemma C.5: *If $G \in \left\{X^T X\right\}^-$, then $G^T \in \left\{X^T X\right\}^-$*

Proof: Let G be any generalized inverse of $X^T X$. Then $X^T XGX^T X = X^T X$, and hence $(X^T XGX^T X)^T = (X^T X)^T = X^T X$. And $(X^T XGX^T X)^T = X^T XG^T X^T X$, so that

$$X^T XG^T X^T X = X^T X,$$

which shows that $G^T \in \left\{X^T X\right\}^-$. ■

C.2 Generalized Inverses and the Normal Equations

Lemma C.6: *The normal equations.*

$$X^T X \boldsymbol{\beta} = X^T \mathbf{y} \tag{C.5}$$

have solutions.

Proof: By the preceding there is no solution to $A\mathbf{x} = \mathbf{z} \Leftrightarrow \mathbf{z} \notin \mathrm{sp}(A)$. Thus, if $\mathbf{z} \in \mathrm{sp}(A)$, there are solutions. Obviously $X^T\mathbf{y} \in \mathrm{sp}(X^T)$. Hence lemma B.3 implies that $X^T\mathbf{y} \in \mathrm{sp}(X^T X)$. ■

Proposition C.7: *If G is any generalized inverse of $X^T X$, then*

$$\widehat{\boldsymbol{\beta}}^{\dagger} = G X^T \mathbf{y}$$

is one solution to the normal equations (C.5).

Proof: By Lemma C.6, there exists at least one $\boldsymbol{\beta}_o$ such that

$$X^T X \boldsymbol{\beta}_o = X^T \mathbf{y}. \tag{C.6}$$

Then

$$X^T X \widehat{\boldsymbol{\beta}}^{\dagger} = X^T X G X^T \mathbf{y} = X^T X G X^T X \boldsymbol{\beta}_o,$$

where we used (C.6). As $G \in \left\{X^T X\right\}^-$ with the notation in (C.1) we have $X^T X G X^T X = X^T X$,

$$= X^T X \boldsymbol{\beta}_o = X^T \mathbf{y},$$

as was to be proved. ■

It must be observed that $\widehat{\boldsymbol{\beta}}^{\dagger}$ is not a statistical estimate, but just a solution of the normal equations, see [16, pp. 34−35], as it depends on the selection of the generalized inverse.

Let now $\left(X^T X\right)^-$ be any generalized inverse of $X^T X$ and let us define a generalized hat matrix by

$$H^- := X X^- = X \left(X^T X\right)^- X^T, \tag{C.7}$$

where we used (C.3) in Proposition C.4. The following Proposition is from [11, p. 12].

Proposition C.8: *X is $n \times k$ The following properties hold:*

1. *H^- is idempotent.*
2. *H^- is the same independently of which $\left(X^T X\right)^-$ is used.*
3. *The range of H^- is $\mathrm{sp}(X)$.*
4. *H^- is symmetric.*

Proof: The proof proceeds step by step in the given order.

1. By definition of $(X)^{-}$, see (C.1), that we have
$$H^{-}H^{-} = \underbrace{XX^{-}X}_{=X}X^{-} = XX^{-} = H^{-}.$$

2. Let $G_1 \in \left\{X^TX\right\}^{-}$ and $G_2 \in \left\{X^TX\right\}^{-}$. Hence
$$X^TXG_1X^TX = X^TX, \quad X^TXG_2X^TX = X^TX.$$
Take now $A = G_1X^TX$ and $B = G_2X^TX$ and apply Lemma C.3 to obtain
$$XG_1X^TX = XG_2X^TX.$$
If we transpose this result we get
$$X^TXG_1^TX^T = X^TXG_2^TX^T$$
When we exploit Lemma C.3 again with $A = G_1^TX^T$ and $B = G_2^TX^T$, which yields
$$XG_1^TX^T = XG_2^TX^T = H^{-},$$
as in view of Lemma C.5, $G_1^T \in \left(X^TX\right)^{-}$ and $G_2^T \in \left(X^TX\right)^{-}$.

3. By (C.1)
$$X = XX^{-}X \Leftrightarrow X = X\left(X^TX\right)^{-}X^TX \Leftrightarrow X = H^{-}X. \tag{C.8}$$
Hence we get that H^{-} is a projection matrix with range $\mathrm{sp}(X)$.

4. Compute
$$\left(H^{-}\right)^T = \left(X^{-}\right)^TX^T = \left(\left(X^TX\right)^{-}X^T\right)^TX^T = X\left(\left(X^TX\right)^{-}\right)^TX^T.$$
In view of Lemma C.5 $\left(\left(X^TX\right)^{-}\right)^T \in \left\{X^TX\right\}^{-}$. Hence by 2. in this Proposition proved above,
$$X\left(\left(X^TX\right)^{-}\right)^TX^T = H^{-},$$
and the asserted symmetry has been established. ■

Hence we know that H^{-} is a projection matrix with range $\mathrm{sp}(X)$. In Proposition C.8 one considers G, any generalized inverse of XX^T and then defines $\widehat{\boldsymbol{\beta}}^{\dagger} = GX^T\mathbf{y}$, a solution of the normal equations. Next the predictor is
$$\widehat{\mathbf{y}}^{\dagger} = X\widehat{\boldsymbol{\beta}}^{\dagger} = X\left(XX^T\right)^{-}X^T\mathbf{y} = H^{-}\mathbf{y}.$$

By part 2. of Proposition C.8 H^{-} is the same independently of which $\left(X^TX\right)^{-}$ is used. Hence $\widehat{\mathbf{y}}^{\dagger}$ is also independent of of which $\left(X^TX\right)^{-}$ is used.

C.3 Bjerhammar-Moore-Penrose Inverse

C.3.1 Definition and Examples

If a generalized inverse G of A satisfies the four conditions below, then G is most often called the *Moore-Penrose inverse*. The name *pseudoinverse* is also frequently encountered.

MP1 $AGA = A$

MP2 $GAG = G$

MP3 $(AG)^T = AG$

MP4 $(GA)^T = GA$

Moore (1935), Penrose (1955) showed that for a given A there is only one matrix G that satisfies MP1-MP4. The proof of this is quite long. We set $G \mapsto A^+$ to denote the Moore-Penrose inverse of A. Arne Bjerhammar [3] found A^+ independently of Moore and Penrose. We shall talk about the Bjerhammar-Moore-Penrose (BMP) inverse. There are two important special cases of BMP inverses.

Example: Assume that A has full column rank, $rankA = k$. Then the BMP inverse of A is

$$A^+ = (A^T A)^{-1} A^T$$

This holds by a direct check of MP1-MP4 hold. Lemma C.3 is needed. A^+ is the left inverse of A, since

$$A^+ A = (A^T A)^{-1} A^T A = \mathbb{I}_k.$$

Hence, if X has full column rank, then

$$X^+ = (X^T X)^{-1} X^T \tag{C.9}$$

is the BMP inverse of X.

Example: If A has full row rank, $rankA = n$, then

$$A^+ = A^T (AA^T)^{-1} \tag{C.10}$$

is the BMP inverse of A. ■

Again, one checks that MP1-MP4 hold. A^+ is a right inverse of A, since $AA^+ = \mathbb{I}_n$, which also checks with MP4. Addional properties of BMP inverses are

1. $(cA)^+ = \frac{1}{c} A^+$,
2. $(A^+)^+ = A$,
3. if A is $m \times r$ and B is $r \times n$ and both matrices have the same rank, then $(AB)^+ = B^+ A^+$, and
4. there is a counterexample to show that $(AB)^+ \neq B^+ A^+$ can occur.

These are left as exercises for the reader.

C.3.2 On the Range of X^+ and X^+X

The BMP X^+ has the following important property.

Proposition C.9: *X^+ is the BMP inverse of the regressor matrix X. Then*

$$\text{sp}\left(X^+\right) = \text{sp}\left(X^T\right). \tag{C.11}$$

Proof: Take any $\mathbf{y}$. By MP2 and MP4 we get

$$\text{sp}\left(X^+\right) \ni X^+\mathbf{y} = X^+XX^+\mathbf{y} = \left(X^+X\right)X^+\mathbf{y} = \left(X^+X\right)^T X^+\mathbf{y}.$$

Then the rule for transposes of products of matrices

$$= X^T\left(X^+\right)^T X^+\mathbf{y} = X^T\left(\left(X^+\right)^T X^+\right)\mathbf{y} = X^T\mathbf{u} \in \text{sp}\left(X^T\right),$$

with $\mathbf{u} = \left(\left(X^+\right)^T X^+\right)\mathbf{y}$. Hence every $X^+\mathbf{y} \in \text{sp}\left(X^T\right)$. Conversely, take any $\mathbf{u}$.

$$\text{sp}\left(X^T\right) \ni X^T\mathbf{u} = X^+X^TX^+\mathbf{u}$$

$$= X^+(X^TX^+)^T\mathbf{u} = X^+(X^+)^TX\mathbf{u} = X^+\mathbf{w} \in \text{sp}\left(X^+\right),$$

where $\mathbf{w} = (X^+)^TX\mathbf{u}$. This proves the claim. ■

The proof of the next statement is a small modification of [11, Result A.14 p. 251], where it is shown that AA^- projects onto sp $\left(X^T\right)$for any generalized inverse A^- of A .

Proposition C.10: *X^+ is the BMP inverse of the regressor matrix X. Then X^+X projects onto* sp $\left(X^T\right)$.

Proof: X^+X is idempotent, since $X^+XX^+X = X^+X$ by MP2. XX^+ is symmetric by MP4. Hence X^+X is a projector matrix according to Definition B.2. The statement "a projection onto" requires to show that

i) $X^+X\mathbf{z} \in \text{sp}\left(X^T\right)$ for every $\mathbf{z}$

ii) It holds for every $\mathbf{y} \in \text{sp}\left(X^T\right)$ that there is a $\mathbf{z}$ such that $\mathbf{y} = X^+X\mathbf{z}$.

The proof proceeds in the order above.

i) Take any $\mathbf{z}$. By MP4 and the rule for transposes of products of matrices

$$X^+X\mathbf{z} = \left(X^+X\right)^T\mathbf{z} = X^T\left(X^+\right)^T\mathbf{z} = X^T\mathbf{u} \in \text{sp}\left(X^T\right),$$

where $\mathbf{u} = \left(X^+\right)^T\mathbf{z}$.

ii) Take any $\mathbf{y} \in \mathrm{sp}\left(X^T\right)$, that is, there exists a $\mathbf{z}$ such that $\mathbf{y} = X^T\mathbf{z}$. We have by MP4 and (C.2)

$$X^+X = \left(X^+X\right)^T = X^T\left(X^+\right)^T = X^T\left(X^T\right)^+.$$

Hence

$$X^+X\mathbf{y} = X^T\left(X^T\right)^+\mathbf{y} = X^T\left(X^T\right)^+X^T\mathbf{z}$$

and by MP1

$$= X^T\mathbf{z} = \mathbf{y}.$$

Hence the claim about the range of X^+X holds as claimed. ■

We need following observation, too.

Lemma C.11: $\mathrm{sp}\left(X^T\right)$ *is a closed linear subspace of* $\mathbb{R}^k$.

Proof: We observe first that since X^T is $k \times n$, $\mathrm{sp}\left(X^T\right) \subset \mathbb{R}^k$ by (B.17).
For closedness it is required to prove that if $\mathbf{x}_n \in \mathrm{sp}\left(X^T\right)$ for every n, and $\mathbf{x}_n \to \mathbf{x}$, as $n \to +\infty$, then $\mathbf{x} \in \mathrm{sp}\left(X^T\right)$, too. We have $\mathbf{x}_n = X^T\mathbf{y}_n$. If $\mathbf{x}_n \to \mathbf{x}$, then it must be that $\mathbf{y}_n \to \mathbf{y}$. In view of the consistency property (3.9)

$$\parallel X^T\mathbf{y}_n - X^T\mathbf{y} \parallel \leq \parallel X^T \parallel \parallel \mathbf{y}_n - \mathbf{y} \parallel \to 0 \quad \text{as } n \to +\infty.$$

Hence $\mathbf{x}_n = X^T\mathbf{y}_n \to X^T\mathbf{y}$, as $n \to +\infty$. But limits in $\mathbb{R}^k$ with the norm $\parallel \mathbf{x} \parallel$ are unique, and thus $\mathbf{x} = X^T\mathbf{y} \in \mathrm{sp}\left(X^T\right)$.
$\mathrm{sp}\left(X^T\right)$ is also a linear space, since if $\mathbf{x}_i \in \mathrm{sp}\left(X^T\right)$ for $i = 1, 2, \ldots, n$, then

$$\sum_{i=1}^n \lambda_i\mathbf{x}_i = \sum_{i=1}^n \lambda_i X^T\mathbf{y}_i = X^T\left[\sum_{i=1}^n \lambda_i\mathbf{y}_i\right].$$

i.e. $\sum_{i=1}^n \lambda_i\mathbf{x}_i \in \mathrm{sp}\left(X^T\right)$, as was to be proved. ■

C.3.3 BMP Inverses and Normal Equations

Next we find the general solution of the normal equations (C.6).

Proposition C.12: *Let* $\boldsymbol{\beta}_o$ *be an arbitrary solution of the normal equations (C.6).* X^+ *is the BMP inverse of* X*. Then*

$$\boldsymbol{\beta}_o = X^+\mathbf{y} + \left(\mathbb{I}_k - X^+X\right)\mathbf{z}. \tag{C.12}$$

$X^+\mathbf{y}$ *is a particular solution of (C.6) and* $(\mathbb{I}_k - X^+X)\,\mathbf{z} \in \mathcal{N}(X^TX)$.

Proof: Let $\mathbf{t}$ be one solution of $X\mathbf{t} = \mathbf{0}_n$, i.e, $\mathbf{t} \in \mathcal{N}(X)$. Then $\mathbf{t} \in \mathcal{N}(X^+X)$. It was observed in the proof of Proposition C.10 that X^+X is a projector. Lemma B.4 or (B.18) entail that

$$\mathcal{N}(X^+X) = \mathrm{sp}(\mathbb{I}_k - X^+X)$$

Hence there is a $\mathbf{z}$ such that $\mathbf{t} = (\mathbb{I}_k - X^+X)\mathbf{z}$. Next we verify that $\boldsymbol{\beta}_o$ in (C.12) solves (C.6). In this respect we have

$$X^T X\boldsymbol{\beta}_o = X^T XX^+\mathbf{y} + X^T X\left(\mathbb{I}_k - X^+X\right)\mathbf{z},$$

Here

$$X^T XX^+\mathbf{y} = X^T\left(XX^+\right)\mathbf{y} = X^T\left(XX^+\right)^T\mathbf{y},$$

where we used MP3. Rules of transposed product matrices and (C.2) entail

$$= X^T\left(X^+\right)^T X^T\mathbf{y} = X^T\left(X^T\right)^+ X^T\mathbf{y} = X^T\mathbf{y},$$

where the final step follows by MP1. Next

$$X^T X\left(\mathbb{I}_{k+1} - X^+X\right)\mathbf{z} = X^T X\mathbf{z} - X^T XX^+X\mathbf{z}.$$

In this $XX^+ = \left(XX^+\right)^T$ by MP3 and hence $X^T XX^+X = X^T\left(XX^+\right)^T X$ $= X^T(X^T)^+X^T X =(X^T(X^T)^+X^T)X = X^T X$ by Lemma C.2. Hence $X^T X\mathbf{z} - X^T XX^+X\mathbf{z} = \mathbf{0}$, i.e. $\left(\mathbb{I}_{k+1} - X^+X\right)\mathbf{z} \in \mathcal{N}(X^T X)$. Then, by MP3

$$X^T X\boldsymbol{\beta}_o = X^T XX^+\mathbf{y} = X^T(XX^+)\mathbf{y} = X^T(XX^+)^T\mathbf{y}$$

$$= X^T(X^+)^T X^T\mathbf{y} = X^T\mathbf{y},$$

where we used MP1 and (C.2). Hence the expression in (C.12) has been verified. ■

C.3.4 Interpolation Limit

When X has full row rank, *rank* $X = n$, then we define by (C.10)

$$\boldsymbol{\beta}^+ := X^+\mathbf{y}.$$

Then

$$X^T X\boldsymbol{\beta}^+ = X^T\underbrace{XX^T(XX^T)^{-1}}_{=\mathbb{I}_n}\mathbf{y} = X^T\mathbf{y}.$$

Hence $\boldsymbol{\beta}^+$ is a solution to the normal equations (C.5).

Proposition C.13: *X has full row rank n, then the BMP predictor given by $\widehat{\mathbf{y}}^+ = X\boldsymbol{\beta}^+$ is an interpolation of the training set.*

Proof: We have from (C.10)

$$\widehat{\mathbf{y}}^+ = X\boldsymbol{\beta}^+ = XX^+\mathbf{y}$$

$$= XX^T(XX^T)^{-1}\mathbf{y} = \mathbf{y},$$

which is the statement of full interpolation. ■

Here we have the hat matrix $H = \mathbb{I}$, which is a idempotent and symmetric matrix, and the only invertible idempotent and symmetric matrix.

C.4 Minimum Norm Least Squares Estimate

C.4.1 The Minimization Problem

This section is based on [16, pp. 114−115] and one of the steps in [10], c.f., [14, pp. 142−144], too.

Lemma C.14: $\widehat{\boldsymbol{\beta}}^{+} = X^{+}\mathbf{y}$ *is the minimum norm least squares estimate, or,*

$$\widehat{\boldsymbol{\beta}}^{+} = \min \parallel \boldsymbol{\beta} \parallel$$

subject to

$$\boldsymbol{\beta} \quad \textit{minimizes} \quad \parallel \mathbf{y} - X\boldsymbol{\beta} \parallel$$

Proof: By (C.12) it holds for an arbitrary solution $\boldsymbol{\beta}_o$ of the normal equations that

$$\parallel \boldsymbol{\beta}_o \parallel^2 = \parallel X^{+}\mathbf{y} + \left(\mathbb{I}_k - X^{+}X\right)\mathbf{z} \parallel^2$$

$$= \parallel X^{+}\mathbf{y} \parallel^2 + \parallel \left(\mathbb{I}_k - X^{+}X\right)\mathbf{z} \parallel^2 - 2(\mathbf{z}^T \left(\mathbb{I}_k - X^{+}X\right)^T X^{+}\mathbf{y}.$$

Here

$$\mathbf{z}^T \left(\mathbb{I}_k - X^{+}X\right)^T X^{+}\mathbf{y} = \mathbf{z}^T \left(\mathbb{I}_k - X^{+}X\right) X^{+}\mathbf{y},$$

since $\mathbb{I}_k - X^{+}X$ is symmetric. But

$$\left(\mathbb{I}_k - X^{+}X\right) X^{+}\mathbf{y} = X^{+}\mathbf{y} - X^{+}XX^{+}\mathbf{y},$$

and by MP2

$$= X^{+}\mathbf{y} - X^{+}\mathbf{y} = \mathbf{0}.$$

Hence

$$\parallel \boldsymbol{\beta}_o \parallel^2 = \parallel X^{+}\mathbf{y} \parallel^2 + \parallel \left(\mathbb{I}_{k+1} - X^{+}X\right)\mathbf{z} \parallel^2 \geq \parallel X^{+}\mathbf{y} \parallel^2 = \parallel \boldsymbol{\beta}^{+} \parallel^2,$$

which completes the proof. ■

C.4.2 Asymptotic Properties of the Minimum Norm LSE

A training set of observed responses $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$ and the corresponding $n \times (k+1)$ data matrix X are available from some source with the true model

$$\mathbf{Y} = X\boldsymbol{\beta}_* + \boldsymbol{\varepsilon}. \tag{C.13}$$

Here $\boldsymbol{\varepsilon}$ has the mean vector $= \mathbf{0}_n$ and covariance matrix $\sigma^2\mathbb{I}_n$. We use the minimum norm least squares solution of the normal equations

$$\boldsymbol{\beta}^{+} = X^{+}\mathbf{y},$$

where X^{+} is the BMP inverse of the regressor matrix X. The mean squared error is

$$MSE = E \parallel \boldsymbol{\beta}_* - \boldsymbol{\beta}^{+} \parallel^2 .$$

Here

$$\begin{aligned} \boldsymbol{\beta}_* - \boldsymbol{\beta}^+ &= \boldsymbol{\beta}_* - X^+\mathbf{Y} \\ &= \boldsymbol{\beta}_* - X^+X\boldsymbol{\beta}_* - X^+\boldsymbol{\varepsilon} \\ &= \left(\mathbb{I} - X^+X\right)\boldsymbol{\beta}_* - X^+\boldsymbol{\varepsilon}. \end{aligned} \tag{C.14}$$

In view of MP4 and MP2

$$\left(\mathbb{I} - X^+X\right)^T X^+ = \left(\mathbb{I} - (X^+X)^T\right) X^+$$

$$= \left(\mathbb{I} - X^+X\right) X^+ = X^+ - X^+XX^+ = X^+ - X^+ = \mathbf{0}.$$

Hence

$$MSE = \parallel \left(\mathbb{I} - X^+X\right)\boldsymbol{\beta}_* \parallel^2 + E \parallel X^+\boldsymbol{\varepsilon} \parallel^2 .$$

By known rules of covariance matrices, (B.10),

$$E \parallel X^+\boldsymbol{\varepsilon} \parallel^2 = E\left[(X^+\boldsymbol{\varepsilon})^T X^+\boldsymbol{\varepsilon}\right] = \mathrm{tr}E\left[(X^+\boldsymbol{\varepsilon})(X^+\boldsymbol{\varepsilon})^T\right] = \sigma^2\mathrm{tr}(X^+(X^+)^T).$$

From (C.3) we note $X^+ = \left(X^TX\right)^+ X^T$. Hence

$$X^+(X^+)^T = \left(X^TX\right)^+ X^T(\left(X^TX\right)^+ X^T)^T$$

$$= \left(X^TX\right)^+ X^TX((\left(X^TX\right)^+)^T = \left(X^TX\right)^+ X^TX((\left(X^TX\right)^T)^+ = \left(X^TX\right)^+ X^TX \left(X^TX\right)^+$$

where we used (C.2). Then MP2 gives

$$\sigma^2\mathrm{tr}(X^+(X^+)^T) = \sigma^2\mathrm{tr}((X^TX)^+).$$

Hence we have the following formula due to [10].

Proposition C.15:

$$MSE = \parallel \left(\mathbb{I} - X^+X\right)\boldsymbol{\beta}_* \parallel^2 + \sigma^2\mathrm{tr}(X^TX)^+). \tag{C.15}$$

Here we note

$$\boldsymbol{\beta}_* = X^+X\boldsymbol{\beta}_* + \left(\mathbb{I} - X^+X\right)\boldsymbol{\beta}_* \tag{C.16}$$

where X^+X projects orthogonally to $\mathrm{sp}X^T$ by Proposition C.10 and $(\mathbb{I} - X^+X)$ projects oorthogonally to $\mathcal{N}(X)$. Hence by (C.13)

$$\mathbf{Y} = X\boldsymbol{\beta}_* + \boldsymbol{\varepsilon} = XX^+X\boldsymbol{\beta}_* + X\left(\mathbb{I} - X^+X\right)\boldsymbol{\beta}_* + \boldsymbol{\varepsilon}$$

$$= XX^+X\boldsymbol{\beta}_* + \left(X - XX^+X\right)\boldsymbol{\beta}_* + \boldsymbol{\varepsilon} = X\boldsymbol{\beta}_* + \boldsymbol{\varepsilon}.$$

where we used MP1 twice. Hence we see, as pointed out in [10], that the first term in (C.15) is the contribution to the MSE due to the part of the true parameter, which does not influence the training data and thus cannot be estimated from data. Hence $(\mathbb{I} - X^+X)\boldsymbol{\beta}_*$ is called the *unidentifiable* part of the true parameter.

New data $\mathbf{x}_{n+1}, y_{n+1}$ is received and the training set is augmented. The new $(n+1) \times k$ regressor matrix X_{n+1} is

$$X_{n+1} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_i^T \\ \vdots \\ \mathbf{x}_n^T \\ \mathbf{x}_{n+1}^T \end{pmatrix}.$$

The current regressor matrix $X \mapsto X_n$ is found inside the augmented matrix as

$$X_{n+1} = \begin{pmatrix} X_n \\ \mathbf{x}_{n+1}^T \end{pmatrix}.$$

Hence we get

$$\begin{aligned} X_{n+1}^T X_{n+1} &= \begin{pmatrix} X_n \\ \mathbf{x}_{n+1}^T \end{pmatrix}^T \begin{pmatrix} X_n \\ \mathbf{x}_{n+1}^T \end{pmatrix} \\ &= \begin{pmatrix} X^T & \mathbf{x}_{n+1} \end{pmatrix} \begin{pmatrix} X \\ \mathbf{x}_{n+1}^{\mathsf{T}} \end{pmatrix} = X_n^T X_n + \mathbf{x}_{n+1}\mathbf{x}_{n+1}^T. \end{aligned} \tag{C.17}$$

We have the next result from [10]. The proof below fills in a few details not included in loc.cit..

Proposition C.16: *If X has full column rank k, $n > k$, then $MSE_n = \sigma^2 \mathrm{tr}(X_n^T X_n)^1$ and*

$$MSE_{n+1} < MSE_n \tag{C.18}$$

for all $n \geq 1$.

Proof: By (C.9) $X^+ = (X^T X)^{-1} X^T$ and thus

$$\left(\mathbb{I}_k - X^+ X\right) = \left(\mathbb{I}_k - (X^T X)^{-1} X^T X\right) = \mathbf{0}_k$$

and thus the unidentifiable part of the true parameter disappears and $MSE_n = \sigma^2 \mathrm{tr}((X_n^T X_n)^{-1})$.
To establish the inequality (C.18), we apply the Sherman-Morrison-Woodbury inversion rule (B.12) with $A = X^T X$, $\mathbf{u}\mathbf{v}^T = \mathbf{x}_{n+1}\mathbf{x}_{n+1}^T$. We note that $X^T X$ is invertible, since X has full column rank. In addition, $1 + \mathbf{x}_{n+1}^T \left(X^T X\right)^{-1} \mathbf{x}_{n+1} \neq 0$, since $\mathbf{x}_{n+1}^T \left(X^T X\right)^{-1} \mathbf{x}_{n+1} > 0$, because $\left(X^T X\right)^{-1}$ is positive definite. Thus we get

$$\begin{aligned} \left(X_{n+1}^T X_{n+1}\right)^{-1} &= \left(X_n^T X_n + \mathbf{x}_{n+1}\mathbf{x}_{n+1}^T\right)^{-1} \\ &= \left(X_n^T X_n\right)^{-1} - \frac{\left(X_n^T X_n\right)^{-1} \mathbf{x}_{n+1} \mathbf{x}_{n+1}^{\mathsf{T}} \left(X_n^T X_n\right)^{-1}}{1 + \mathbf{x}_{n+1}^{\mathsf{T}} \left(X_n^T X_n\right)^{-1} \mathbf{x}_{n+1}}. \end{aligned}$$

By rules of trace, see section (B.2.5), we get

$$\begin{aligned}\sigma^2\mathrm{tr}\left(X_{n+1}^TX_{n+1}\right)^{-1} &= \sigma^2\mathrm{tr}\left(X_n^TX_n\right)^{-1} \\ &\quad -\sigma^2\mathrm{tr}\left[\frac{\left(X_n^TX_n\right)^{-1}\mathbf{x}_{n+1}\mathbf{x}_{n+1}^T\left(X_n^TX_n\right)^{-1}}{1+\mathbf{x}_{n+1}^T\left(X_n^TX_n\right)^{-1}\mathbf{x}_{n+1}}\right].\end{aligned}$$

We have by trace rules, see section B.2.5, that

$$\mathrm{tr}\left[\frac{\left(X_n^TX_n\right)^{-1}\mathbf{x}_{n+1}\mathbf{x}_{n+1}^T\left(X_n^TX_n\right)^{-1}}{1+\mathbf{x}_{n+1}^T\left(X_n^TX_n\right)^{-1}\mathbf{x}_{n+1}}\right] = \frac{\mathrm{tr}\left[\left(X_n^TX_n\right)^{-1}\mathbf{x}_{n+1}\mathbf{x}_{n+1}^T\left(X_n^TX_n\right)^{-1}\right]}{1+\mathbf{x}_{n+1}^T\left(X_n^TX_n\right)^{-1}\mathbf{x}_{n+1}}.$$

Here for any $\mathbf{x}\neq\mathbf{0}$, and since $\left(X_n^TX_{=n}\right)^{-1}$ is symmetric in view of (B.5)

$$\mathbf{x}^T\left(X_n^TX_n\right)^{-1}\mathbf{x}_{n+1}\mathbf{x}_{n+1}^T\left(X_n^TX_n\right)^{-1}\mathbf{x} = \left(\mathbf{x}_{n+1}^T\left(X_n^TX_n\right)^{-1}\mathbf{x}\right)^T\mathbf{x}_{n+1}^T\left(X_n^TX_n\right)^{-1}\mathbf{x}$$

$$= \left(\mathbf{x}_{n+1}^T\left(X_n^TX_n\right)^{-1}\mathbf{x}\right)^2 > 0,$$

as the null space of $\left(X_n^TX_n\right)^{-1}$ is $=\{\mathbf{0}\}$. Hence, since the trace of a positive definite matrix is positive, (B.13), and since $\sigma^2/(1+\mathbf{x}_{n+1}^T\left(X_n^TX_n\right)^{-1}\mathbf{x}_{n+1}>0)$,

$$\sigma^2\mathrm{tr}\left(X_{n+1}^TX_{n+1}\right)^{-1} < \sigma^2\mathrm{tr}\left(X_n^TX_n\right)^{-1},$$

as was to be proved. ■

The following is not found in [10].

Proposition C.17: *If $\lambda_+(n) := \lambda_+\left((X_n^TX_n)^{-1}\right)\to 0$, as $n\to+\infty$, where $\lambda_+(n)$ is the largest eigenvalue of $(X_n^TX_n)^{-1}$, then*

$$MSE_n\to 0, \quad \textit{as } n\to+\infty.$$

Proof: Since $MSE_n = \sigma^2\mathrm{tr}\left((X_n^TX_n)^{-1}\right) > 0$, it follows from (C.18) that $MSE_n\to MSE_*$, as $n\to+\infty$ in view of Proposition E.1. From section B.2.5 know that

$$\mathrm{tr}\left(X_n^TX_n\right)^{-1} = \sum_{i=1}^k\lambda_i\left((X_n^TX_n)^{-1}\right),$$

where $\lambda_i\left((X_n^TX_n)^{-1}\right)$ are the real positive eigenvalues of $(X_n^TX_n)^{-1}$. Hence

$$\sum_{i=1}^k\lambda_i\left((X_n^TX_n)^{-1}\right) < k\lambda_+(n).$$

Hence $MSE_*=0$, as was to be proved. ■

Proposition C.18: *We have the true model*

$$\mathbf{Y} = X\boldsymbol{\beta}_* + \boldsymbol{\varepsilon}, \tag{C.19}$$

where $n > k$ and X has full column rank, $E[\boldsymbol{\varepsilon}] = \mathbf{0}_n$, $C_{\boldsymbol{\varepsilon}} = \sigma^2\mathbb{I}_n$, and σ^2 is known and does not depend on X. Assume that

$$\lim_{n\to+\infty} \lambda_+(n) = 0,$$

where $\lambda_+(n)$ is the largest eigenvalue of $(X^TX)^{-1}$. $\widehat{\boldsymbol{\beta}}$ is the LSE based on a training set of n samples of (C.23). Then

$$\lim_{n\to+\infty} E\left[\left(\mathbf{l}^T\widehat{\boldsymbol{\beta}} - \mathbf{l}^T\boldsymbol{\beta}_*\right)^2\right] = 0, \tag{C.20}$$

Proof: Under the true model (C.23)

$$\widehat{\boldsymbol{\beta}} = (X^TX)^{-1}X^TX\boldsymbol{\beta}_* + (X^TX)^{-1}X^T\boldsymbol{\varepsilon} = \boldsymbol{\beta}_* + (X^TX)^{-1}X^T\boldsymbol{\varepsilon}. \tag{C.21}$$

Then

$$E\left[\mathbf{l}^T\widehat{\boldsymbol{\beta}}\right] = \mathbf{l}^TE\left[\widehat{\boldsymbol{\beta}}\right] = \mathbf{l}^T\left[(X^TX)^{-1}X^TE[\mathbf{Y}]\right]$$
$$= \mathbf{l}^T(X^TX)^{-1}X^TX\boldsymbol{\beta}_* + \mathbf{l}^T(X^TX)^{-1}X^TE[\varepsilon] = \mathbf{l}^T\boldsymbol{\beta}_*.$$

Hence, by definition of variance,

$$E\left[\left(\mathbf{l}^T\widehat{\boldsymbol{\beta}} - \mathbf{l}^T\boldsymbol{\beta}_*\right)^2\right] = \operatorname{Var}\left[\mathbf{l}^T\widehat{\boldsymbol{\beta}}\right] = \mathbf{l}^TC_{\widehat{\boldsymbol{\beta}}}\mathbf{l},$$

where we used (B.11). From (C.21), (B.8), and since $C_{\boldsymbol{\varepsilon}} = \sigma^2\mathbb{I}_n$ we have

$$C_{\widehat{\boldsymbol{\beta}}} = \sigma^2(X^TX)^{-1}X^T\left((X^TX)^{-1}X^T\right)^T = \sigma^2(X^TX)^{-1}X^T\left(X(X^TX)^{-1}\right),$$

where the symmetry of $(X^TX)^{-1}$ was applied. Thereby

$$= \sigma^2(X^TX)^{-1}(X^TX(X^TX)^{-1}) = \sigma^2(X^TX)^{-1}.$$

Now,

$$\mathbf{l}^TC_{\widehat{\boldsymbol{\beta}}}\mathbf{l} = \sigma^2\mathbf{l}^T(X^TX)^{-1}\mathbf{l} = \sigma^2 \mid \mathbf{l}^T(X^TX)^{-1}\mathbf{l} \mid,$$

since $(X^TX)^{-1}$ is positive definite. Thereafter, by Cauchy-Schwartz inequality (B.2) and consistency property (3.9)

$$\mid \mathbf{l}^T\left((X^TX)^{-1}\mathbf{l}\right) \mid \leq \parallel \mathbf{l}^T \parallel \parallel (X^TX)^{-1}\mathbf{l} \parallel \leq \parallel (X^TX)^{-1} \parallel \parallel \mathbf{l} \parallel^2,$$

since by definition of norm, $\parallel \mathbf{l}^T \parallel = \sqrt{(\mathbf{l}^T)^T\mathbf{l}^T} = \sqrt{\mathbf{l}\mathbf{l}^T} = \sqrt{\mathbf{l}^T\mathbf{l}} = \parallel \mathbf{l} \parallel$.

$$= C \parallel (X^TX)^{-1} \parallel = C\lambda_+(n)$$

as was shown in the proof of Lemma 3.3. When compiling from the above, we obtain

$$E\left[\left(\mathbf{l}^T\widehat{\boldsymbol{\beta}} - \mathbf{l}^T\boldsymbol{\beta}_*\right)^2\right] \leq C\lambda_+(n),$$

and the assertion in (C.24) follows. ■

The Proposition C.16 can be extended to other minimum norm least squares estimators, if there is a suitable Sherman-Morrison-Woodbury generalized inversion rule, of whose existence this author is not aware. However, Proposition 33 can be extended readily to $\beta^+ = X^+\mathbf{y}$, where X^+ is the BMP inverse of X. An additional assumption required is the *estimability* of $\mathbf{l}^T\boldsymbol{\beta}_*$. By this one means that there exists a vector $\mathbf{a}$ so that

$$E\left[\mathbf{a}^T\mathbf{Y}\right] = \mathbf{l}^T\boldsymbol{\beta}_*. \tag{C.22}$$

For the notion of estimability, see [11, pp. 38−41] and [16, pp. 37−38, p. 284].

Proposition C.19: *We have the true model*

$$\mathbf{Y} = X\boldsymbol{\beta}_* + \boldsymbol{\varepsilon}, \tag{C.23}$$

where $n < k$. $E[\boldsymbol{\varepsilon}] = \mathbf{0}_n$, $C_{\boldsymbol{\varepsilon}} = \sigma^2\mathbb{I}_n$, *and* σ^2 *is known and does not depend on* X. *Assume that* $\mathbf{l}^T \in \operatorname{sp}(X^T)$ *and that*

$$\lim_{n\to+\infty} \lambda_+(n) = 0,$$

where $\lambda_+(n)$ *is the largest eigenvalue of* $(X^+X)^2$. $\boldsymbol{\beta}^+ = X^+\mathbf{y}$ *is the minimum norm LSE based on a training set of* n *samples of (C.23). Then*

$$\lim_{n\to+\infty} E\left[\left(\mathbf{l}^T\widehat{\boldsymbol{\beta}} - \mathbf{l}^T\boldsymbol{\beta}_*\right)^2\right] = 0. \tag{C.24}$$

Proof: From (C.14)

$$\boldsymbol{\beta}_* - \boldsymbol{\beta}^+ = \left(\mathbb{I} - X^+X\right)\boldsymbol{\beta}_* - X^+\boldsymbol{\varepsilon}.$$

Let us take $\mathbf{l}^T \in \operatorname{sp}(X^T)$. Since $(\mathbb{I} - X^+X)$ maps to $\mathcal{N}(X^TX) = \mathcal{N}(X^T)$, $\mathbf{l}^T\left(\mathbb{I} - X^+X\right)\boldsymbol{\beta}_* = 0$. Hence

$$E\left[\mathbf{l}^T\boldsymbol{\beta}_* - \mathbf{l}^T\boldsymbol{\beta}^+\right] = \mathbf{0} \Leftrightarrow E\left[\mathbf{l}^T\boldsymbol{\beta}^+\right] = \mathbf{l}^{T\cdot}\boldsymbol{\beta}_*,$$

and hence $\mathbf{l}^T\boldsymbol{\beta}_*$ is estimable, as with respect to (C.22)

$$\mathbf{a}^T\mathbf{Y} = \mathbf{l}^TX^+\mathbf{Y}.$$

Thus again, as in the proof of Proposition 33,

$$E\left[\left(\mathbf{l}^T\widehat{\boldsymbol{\beta}} - \mathbf{l}^T\boldsymbol{\beta}_*\right)^2\right] = \operatorname{Var}\left[\mathbf{l}^T\widehat{\boldsymbol{\beta}}\right] = \mathbf{l}^TC_{\beta^+}\mathbf{l}.$$

By the same rules as in the proof of Proposition 33

$$C_{\beta^+} = \sigma^2X^+(X^+)^T.$$

Then we have as in the proof of Proposition 33

$$E\left[\mathbf{l}^T\boldsymbol{\beta}_* - \mathbf{l}^T\boldsymbol{\beta}^+\right] \leq C \parallel X^+(X^+)^T \parallel = C\lambda_+(n),$$

and the claimed convergence follows. ■

D Matrix Differential Calculus

D.1 Matrix Derivatives

Let $\mathbf{A}$ be a $k \times k$ matrix of constants, $\mathbf{a}$ be a $k \times 1$ vector of constants, and $\mathbf{y}$ be a $k \times 1$ vector of variables.

1. If $\mathbf{z} = \mathbf{a}^\top \mathbf{y}$, then
$$\frac{\partial \mathbf{z}}{\partial \mathbf{y}} = \frac{\partial \mathbf{a}^\top \mathbf{y}}{\partial \mathbf{y}} = \mathbf{a}.$$
2. If $\mathbf{z} = \mathbf{y}^\top \mathbf{y}$, then
$$\frac{\partial \mathbf{z}}{\partial \mathbf{y}} = \frac{\partial \mathbf{y}^\top \mathbf{y}}{\partial \mathbf{y}} = 2\mathbf{y}.$$
3. If $\mathbf{z} = \mathbf{a}^\top \mathbf{A}\mathbf{y}$, then
$$\frac{\partial \mathbf{z}}{\partial \mathbf{y}} = \frac{\partial \mathbf{a}^\top \mathbf{A}\mathbf{y}}{\partial \mathbf{y}} = \mathbf{A}^\top \mathbf{a}.$$
4. If $\mathbf{z} = \mathbf{y}^\top \mathbf{A}\mathbf{y}$, then
$$\frac{\partial \mathbf{z}}{\partial \mathbf{y}} = \frac{\partial \mathbf{y}^\top \mathbf{A}\mathbf{y}}{\partial \mathbf{y}} = \mathbf{A}\mathbf{y} + \mathbf{A}^\top \mathbf{y}.$$
If $\mathbf{A}$ is symmetric, then
$$\frac{\partial \mathbf{y}^\top \mathbf{A}\mathbf{y}}{\partial \mathbf{y}} = 2\mathbf{A}\mathbf{y}.$$

D.2 Gradient and Hessian

$\mathbf{x} \in \mathbb{R}^d$, $f(\mathbf{x})$ is a smooth function, $\mathbf{x} \mapsto f(\mathbf{x}) \in \mathbb{R}$. By smoothness we understand that $f(\mathbf{x})$ has all second order partial derivatives and that these second derivatives are continuous.

1. The gradient vector $\nabla f(\mathbf{x})$ is the $d \times 1$ vector of first order partial derivatives
$$\nabla f(\mathbf{x}) = \left(\frac{\partial}{\partial x_1} f(\mathbf{x}), \frac{\partial}{\partial x_2} f(\mathbf{x}), \ldots \frac{\partial}{\partial x_d} f(\mathbf{x}) \right)^T \tag{D.1}$$
2. The Hessian matrix $\mathcal{H}(\mathbf{x})$ is the $d \times d$ matrix of second order partial derivatives
$$\mathcal{H}(\mathbf{x}) = \left[\frac{\partial^2}{\partial x_i \partial x_j} f(\mathbf{x}) \right]_{i=1,j=1}^{d,d} = \left[\frac{\partial^2}{\partial x_j \partial x_i} f(\mathbf{x}) \right]_{j=1,i=1}^{d,d} \tag{D.2}$$

D.3 Applications of the Chain Rule

$\mathbf{x} \in \mathbb{R}^d$, $f(\mathbf{x})$ is a smooth function, $\mathbf{x} \mapsto f(\mathbf{x}) \in \mathbb{R}$. Set $\mathbf{h} := (\mathbf{y} - \mathbf{x}) \in \mathbb{R}^d$ and
$$F(t) := f\left(\mathbf{x} + t\mathbf{h}\right), 0 \le t \le 1.$$

Next we find two applications of this function.

D.3.1 The Gradient Integral

The general chain rule of calculus, see, e.g., [15, p. 69], yields

$$F^{'}(t) = \frac{d}{dt}F(t) = \frac{\partial}{\partial h_1} f\left(\mathbf{x}+t\mathbf{h}\right) h_1 + \frac{\partial}{\partial h_2} f\left(\mathbf{x}+t\mathbf{h}\right) h_2 + \ldots + \frac{\partial}{\partial h_d} f\left(\mathbf{x}+t\mathbf{h}\right) h_d$$
$$= \nabla f\left(\mathbf{x}+t\mathbf{h}\right)^T \left(\mathbf{y}-\mathbf{x}\right). \tag{D.3}$$

Hence

$$f\left(\mathbf{y}\right) - f\left(\mathbf{x}\right) = F(1) - F(0) = \int_0^1 F^{'}(t)dt$$

i.e.

$$f\left(\mathbf{y}\right) - f\left(\mathbf{x}\right) = \int_0^1 \nabla f\left(\mathbf{x}+t\mathbf{h}\right)^T \left(\mathbf{y}-\mathbf{x}\right) dt. \tag{D.4}$$

D.3.2 The Local Minimum of $f\left(\mathbf{x}\right)$

By another application of the chain rule we get from (D.3)

$$F^{''}(t) = \frac{d^2}{d^2t}F(t) = \sum_{i=1}^{d}\sum_{j=1}^{d} \frac{\partial}{\partial h_i}\frac{\partial}{\partial h_j}\left(\mathbf{x}+t\mathbf{h}\right) h_i h_j = \mathbf{h}^T \mathcal{H}\left(\mathbf{x}+t\mathbf{h}\right)\mathbf{h}, \tag{D.5}$$

where $\mathcal{H}\left(\mathbf{x}+t\mathbf{h}\right)$ is the Hessian of $f\left(\mathbf{x}\right)$ at $\mathbf{x}+t\mathbf{h}$. By a series expansion we have

$$F(1) = F(0) + F^{'}(0) + \frac{1}{2}F^{''}(\eta).$$

i.e.

$$f\left(\mathbf{x}+\mathbf{h}\right) = f\left(\mathbf{x}\right) + \nabla f\left(\mathbf{x}\right)^T \mathbf{h} + \frac{1}{2}\mathbf{h}^T \mathcal{H}\left(\mathbf{x}+\eta\mathbf{h}\right)\mathbf{h}, \tag{D.6}$$

where $\mathcal{H}\left(\mathbf{x}+\eta\mathbf{h}\right)$ the Hessian matrix defined in (D.2).

Proposition D.1: *Let* $\mathbf{a}$ *be a critical point of* $f\left(\mathbf{x}\right)$*, i.e.,*

$$\nabla f\left(\mathbf{a}\right) = \mathbf{0}_d.$$

Then $\mathbf{a}$ *is a local minimum, as soon as the Hessian matrix* $\mathcal{H}\left(\mathbf{a}\right)$ *is a positive definite matrix.*

Proof: By (D.6)

$$\begin{aligned} f\left(\mathbf{a}+\mathbf{h}\right) - f\left(\mathbf{a}\right) &= \underbrace{\nabla f\left(\mathbf{a}\right)^T \mathbf{h}}_{=0} + \frac{1}{2}\mathbf{h}^T \mathcal{H}\left(\mathbf{a}+\eta\mathbf{h}\right)\mathbf{h} \\ &\text{i.e.} \\ f\left(\mathbf{a}+\mathbf{h}\right) - f\left(\mathbf{a}\right) &= \frac{1}{2}\mathbf{h}^T \mathcal{H}\left(\mathbf{a}+\eta\mathbf{h}\right)\mathbf{h} \end{aligned}$$

By the assumed smoothness of $f(\mathbf{x})$, $\mathcal{H}(\mathbf{a}+\eta\mathbf{h})$ is positive definite for $\eta\mathbf{h}$ sufficiently close to $\mathbf{a}$, as $\mathcal{H}(\mathbf{a})$ is a positive definite matrix. Hence for all such $\eta\mathbf{h}$

$$f(\mathbf{a}+\mathbf{h}) - f(\mathbf{a}) > 0,$$

and thus $\mathbf{a}$ is a local minimum. ■

The proof above is a part of the proof of [1, Thm. 3, p. 748].
If the Hessian in (D.6) is positive definite or positive semidefinite, we get

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T (\mathbf{y}-\mathbf{x}). \tag{D.7}$$

E On Convergence of Sequences and Sums

Definition E.1: A sequence $\{s_i\}_{i=0}^{+\infty}$ of real numbers is called

- *monotonically increasing*, if $s_i \leq s_{i+1}$ for all $i = 1, 2, \ldots$
- *monotonically deccreasing*, if $s_i \geq s_{i+1}$ for all $i = 1, 2, \ldots$. ■

Proposition E.1: *A monotonical sequence is convergent if and only if it is bounded.*

For a proof, see, e.g., [17, Thm. 3.14]. Furthermore,

Proposition E.2: *If $s_n = \sum_{i=0}^{n} a_i$ converges, as $n \to +\infty$, then*

$$\lim_{i \to +\infty} a_i \to 0.$$

This is [17, Thm. 3.27].

References

[1] Robert A. Adams & Christopher Essex: *Calculus. A Complete Course*, Eight Edition. Pearson Toronto, 2013.

[2] Sivaranam Balakrishnan: *Lecture Notes on Fundamentals of Convex Optimization*, Carnegie-Mellon University, 2023.

[3] Arne Bjerhammar: Application of calculus of matrices to method of least squares; with special references to geodetic calculations. *Transactions of the Royal Institute of Technology,* Stockholm, 49, 1951.

[4] Åke Björck: *Numerical Methods in Matrix Computations,* Springer 2015.

[5] David L Donoho: High-dimensional data analysis: The curses and blessings of dimensionality, *AMS math challenges lecture*, vol. 1, 2000.

[6] Trevor Hastie & Andrea Montanari & Sharon Rosset & Ryan J. Tibshirani: Surprises in high-dimensional ridgeless least squares interpolation, *Annals of Statistics*, 50, 2, 949−986, 2022.

[7] Lydia I Kronsjö: *Algorithms: Their Complexity and Efficiency*, John Wiley and Sons, 1979.

[8] Claude Lemaréchal: Cauchy and the gradient method. *Documenta Mathematica Extra Volume ISMP* 251−254, 10, 2012.

[9] Bernhard Mehlig: *Machine Learning with Neural Networks. An Introduction for Scientists and Engineers*, Cambridge University Press, 2022.

[10] Per Mattsson & Dave Zachariah & Petre Stoica: Analysis of the Minimum-Norm Least-Squares Estimator and Its Double-Descent Behavior [Lecture Notes], *IEEE Signal Processing Magazine*, 40, 3, 39−75, 2023.

[11] John F. Monahan: *A Primer on Linear Models*, Texts in Statistical Science Series, CRC Press, A Chapman and Hall Book, 2008.

[12] Douglas C Montgomery, & Elizabeth A Peck, & G Geoffrey Vining, *Introduction to linear regression analysis.Sixth Edition*, John Wiley & Sons, 2021.

[13] Yurii Nesterov: *Introductory lectures on convex optimization*, Lecture Notes on Optimization vol 87, Springer Science & Business Media, 2004.

[14] Ben Noble: *Applied Linear Algebra*, Prentice-Hall Inc., 1969.

[15] Arne Persson & Lars-Christer Böiers: *Analys i flera variabler*, Tredje upplagan, Studentlitteratur, 2005.

[16] Simo Puntanen & George P.H. Styan & Jarkko Isotalo: *Matrix Tricks for Linear Statistical Models. Our Personal Top Twenty*, Springer, 2011.

[17] Walter Rudin: *Principles of Mathematical Analysis*, McGraw-Hill Educational, 1976.

[18] Gilbert Strang: *Linear Algebra and Its Applications*, Third Edition, Brooks/Cole, Thomson learning, 1988.