Linear Algebra, Data Science, and Machine Learning

by Jeff Calder and Peter J. Olver

Springer Undergraduate Texts in Mathematics and Technology, 2025

Errata

Last updated December 4, 2025

- Page 99:

  In the displayed formula in the proof of Proposition 4.30, the second equality should be an inequality: $\dim \operatorname{img}(A\,B) \geq \dim(\operatorname{coimg} A \cap \operatorname{img} B)$. Thus, (1.16) proves the lower bound in (4.42), while the upper bound can be easily established directly.

- Page 102: Remark:

  Change "two case" to "two cases".

- Page 141: Equation with $R \circ S$ on left hand side:

  Change $\mathbf{u}_j \mathbf{u}_j^T$ to $\mathbf{u}_i \mathbf{u}_i^T$.

- Page 188: Proof of Theorem 6.9:

  Change $b = \mathbf{y}^T \mathbf{f}$ to $b = -\mathbf{y}^T \mathbf{f}$.

- Page 190: Equation (6.19):

  There is a missing minus sign, the equation should read

  $$P(A\,\mathbf{y}) = \tfrac{1}{2}\,\mathbf{y}^T A^T H A\,\mathbf{y} - \mathbf{y}^T A^T\,\mathbf{f} + c.$$

- Page 190: Equation (6.21):

  In this equation, it should be noted that $\mathbf{b} = A^T \mathbf{f}$.

- Page 191: Remark after Corollary 6.13:

  Insert "in Chapter 4" after "Exercise 2.6".

- Page 195:

  Change "Proposition 6.17" to "Theorem 6.17".

- Page 205: Exercise 4.8:

  Change A to H twice:

  $$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k (H\,\mathbf{x}_k - \mathbf{b}) \qquad \text{and} \qquad \mathbf{r}_k = H\,\mathbf{x}_k - \mathbf{b}.$$

- Page 206: Motivation for the Conjugate Gradient Method:

The formulas for the conjugate gradient method are all correct, but the derivation and motivation has an error. The method should minimize the corresponding quadratic function, and not the residual vector. The text starting right after Equation (6.65) and ending before "Starting with an initial guess..." should be replaced with the text below.

The conjugate gradient algorithm, to be derived below, computes the t_k and $\mathbf{v}_k$ iteratively, so that the k-th approximation to the solution is

$$\mathbf{x}_k = \mathbf{x}_0 + t_1\,\mathbf{v}_1 + \;\cdots\; + t_k\,\mathbf{v}_k, \qquad \text{or, equivalently} \qquad \mathbf{x}_k = \mathbf{x}_{k-1} + t_k\,\mathbf{v}_k,$$

where, once $\mathbf{v}_k$ is specified, t_k is chosen to minimize the quadratic function (6.44). The secret is not to try to specify the conjugate basis vectors in advance, but rather to successively construct them during the course of the algorithm.

We begin with an initial guess $\mathbf{x}_0$ — for example, $\mathbf{x}_0 = \mathbf{0}$. According to (6.45) the residual vector $\mathbf{r}_0 = \mathbf{b} - H\mathbf{x}_0$ is the negative of the Euclidean gradient of P at the point $\mathbf{x}_0$, and hence indicates the direction of steepest decrease. We begin by updating our original guess by moving in this direction, taking $\mathbf{v}_1 = \mathbf{r}_0$ as our first conjugate direction.

The next iterate is $\mathbf{x}_1 = \mathbf{x}_0 + t_1\,\mathbf{v}_1$, and we choose the parameter t_1 so that

$$\begin{aligned} P(\mathbf{x}_1) = P(\mathbf{x}_0 + t_1\,\mathbf{v}_1) &= P(\mathbf{x}_0) + t_1\,\mathbf{v}_1 \cdot (H\mathbf{x}_0 - \mathbf{b}) + \tfrac{1}{2}\,t_1^2\,\mathbf{v}_1^T H\mathbf{v}_1 \\ &= P(\mathbf{x}_0) - t_1\,\mathbf{v}_1 \cdot \mathbf{r}_0 + \tfrac{1}{2}\,t_1^2\,\|\,\mathbf{v}_1\,\|_H^2 \end{aligned} \tag{6.66}$$

is a small as possible, which occurs when

$$t_1 = \frac{\mathbf{v}_1 \cdot \mathbf{r}_0}{\|\,\mathbf{v}_1\,\|_H^2} = \frac{\|\,\mathbf{r}_0\,\|_2^2}{\|\,\mathbf{v}_1\,\|_H^2}. \tag{6.67}$$

We can assume that $t_1 \neq 0$, since otherwise the residual $\mathbf{r}_0 = \mathbf{0}$, which would imply $\mathbf{x}_0 = \mathbf{x}^\star$ is the exact solution of the linear system, and there would be no reason to continue the procedure. We note that the updated residual can be expressed in the form

$$\mathbf{r}_1 = \mathbf{b} - H\mathbf{x}_1 = \mathbf{b} - H\mathbf{x}_0 - t_1 H\mathbf{v}_1 = \mathbf{r}_0 - t_1 H\mathbf{v}_1. \tag{6.68}$$

Observe that, in view of (6.67),

$$\mathbf{r}_0 \cdot \mathbf{r}_1 = \mathbf{v}_1 \cdot \mathbf{r}_1 = \mathbf{v}_1 \cdot \mathbf{r}_0 - t_1 \mathbf{v}_1 \cdot (H\mathbf{v}_1) = \mathbf{v}_1 \cdot \mathbf{r}_0 - t_1\,\|\,\mathbf{v}_1\,\|_H^2 = 0, \tag{6.69}$$

and hence the initial residual — the first conjugate direction — is orthogonal to the first residual under the dot product.

The gradient descent algorithm would tell us to update $\mathbf{x}_1$ by moving in the residual direction $\mathbf{r}_1$. In the conjugate gradient algorithm, we instead choose a direction $\mathbf{v}_2$ which is conjugate, meaning H–orthogonal, to the first direction $\mathbf{v}_1 = \mathbf{r}_0$. Thus, as in the Gram–Schmidt process, we modify the residual direction by setting $\mathbf{v}_2 = \mathbf{r}_1 + s_1\,\mathbf{v}_1$, where the scalar factor s_1 is determined by the imposed orthogonality requirement:

$$0 = \langle\,\mathbf{v}_1, \mathbf{v}_2\,\rangle_H = \langle\,\mathbf{v}_1, \mathbf{r}_1 + s_1\,\mathbf{v}_1\,\rangle_H = \langle\,\mathbf{v}_1, \mathbf{r}_1\,\rangle_H + s_1\,\langle\,\mathbf{v}_1, \mathbf{v}_1\,\rangle_H = \langle\,\mathbf{r}_1, \mathbf{v}_1\,\rangle_H + s_1\,\|\,\mathbf{v}_1\,\|_H^2,$$

and hence we fix

$$s_1 = -\,\frac{\langle\,\mathbf{v}_1, \mathbf{r}_1\,\rangle_H}{\|\,\mathbf{v}_1\,\|_H^2}\,.$$

Now, in view of (6.68) and (6.69),

$$\langle \mathbf{v}_1, \mathbf{r}_1 \rangle_H = \langle \mathbf{r}_1, \mathbf{v}_1 \rangle_H = \mathbf{r}_1^T H \mathbf{v}_1 = \mathbf{r}_1^T \left(\frac{\mathbf{r}_0 - \mathbf{r}_1}{t_1} \right) = -\frac{1}{t_1} \| \mathbf{r}_1 \|_2^2,$$

while, by (6.67),

$$\| \mathbf{v}_1 \|_H^2 = \frac{1}{t_1} \| \mathbf{r}_0 \|_2^2.$$

Therefore, the second conjugate direction is also given by

$$\mathbf{v}_2 = \mathbf{r}_1 + s_1 \mathbf{v}_1, \qquad \text{where} \qquad s_1 = \frac{\| \mathbf{r}_1 \|_2^2}{\| \mathbf{r}_0 \|_2^2}.$$

We then update $\mathbf{x}_2 = \mathbf{x}_1 + t_2 \mathbf{v}_2$, where t_2 is chosen in order to make $P(\mathbf{x}_2)$ as small as possible. By a similar computation as in (6.66), this occurs when

$$t_2 = \frac{\mathbf{v}_2 \cdot \mathbf{r}_1}{\| \mathbf{v}_2 \|_H^2} = \frac{\| \mathbf{r}_1 \|_2^2}{\| \mathbf{v}_2 \|_H^2},$$

where the second expression follows from the formula for $\mathbf{v}_2$ and (6.69). Again, we can assume that $t_2 \neq 0$, as otherwise $\mathbf{r}_1 = \mathbf{0}$ and $\mathbf{x}_1$ would be the exact solution, in which case the algorithm should be terminated.

Continuing in this manner, at the k-th stage for any $k \geq 1$, we have already constructed the conjugate vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$ that are mutually H–orthogonal, and the solution approximation $\mathbf{x}_k$ as a suitable linear combination of them and the initial approximation $\mathbf{x}_0$. We further assume we inductively established the (dot product) orthogonality relations

$$\mathbf{v}_i \cdot \mathbf{r}_k = 0, \qquad 1 \leq i \leq k, \qquad \text{and} \qquad \mathbf{r}_j \cdot \mathbf{r}_k = 0, \qquad 0 \leq j < k.$$

The next conjugate direction is given by

$$\mathbf{v}_{k+1} = \mathbf{r}_k + s_k \mathbf{v}_k, \qquad \text{where} \qquad s_k = \frac{\| \mathbf{r}_k \|_2^2}{\| \mathbf{r}_{k-1} \|_2^2} \tag{6.70}$$

results from the H–orthogonality requirement:

$$\langle \mathbf{v}_i, \mathbf{v}_{k+1} \rangle_H = 0 \qquad \text{for} \qquad 1 \leq i \leq k.$$

The updated solution approximation is

$$\mathbf{x}_{k+1} = \mathbf{x}_k + t_{k+1} \mathbf{v}_{k+1}, \qquad \text{where} \qquad t_{k+1} = \frac{\mathbf{v}_{k+1} \cdot \mathbf{r}_k}{\| \mathbf{v}_{k+1} \|_H^2} = \frac{\| \mathbf{r}_k \|_2^2}{\| \mathbf{v}_{k+1} \|_H^2} \tag{6.71}$$

is then specified so as to make $P(\mathbf{x}_{k+1})$ as small as possible. The second expression for t_{k+1} follows from (6.70) and the orthogonality of $\mathbf{v}_k$ and $\mathbf{r}_k$. The new residual vector is

$$\mathbf{r}_{k+1} = \mathbf{b} - H \mathbf{x}_{k+1} = \mathbf{r}_k - t_{k+1} H \mathbf{v}_{k+1}. \tag{6.72}$$

The induction step is completed by establishing the new orthogonality relations. First, by the already established orthogonality of $\mathbf{v}_i$ and $\mathbf{r}_k$, and the H–orthogonality of the $\mathbf{v}_i$,

$$\mathbf{v}_i \cdot \mathbf{r}_{k+1} = \mathbf{v}_i \cdot \mathbf{r}_k - t_{k+1} \mathbf{v}_i \cdot (H \mathbf{v}_{k+1}) = \mathbf{v}_i \cdot \mathbf{r}_k - t_{k+1} \langle \mathbf{v}_i, \mathbf{v}_{k+1} \rangle_H = 0 \quad \text{for} \quad 1 \leq i \leq k,$$

while, using the first formula for t_{k+1} in (6.71),

$$\mathbf{v}_{k+1} \cdot \mathbf{r}_{k+1} = \mathbf{v}_{k+1} \cdot \mathbf{r}_k - t_{k+1} \mathbf{v}_{k+1} \cdot (H \mathbf{v}_{k+1}) = \mathbf{v}_{k+1} \cdot \mathbf{r}_k - t_{k+1} \| \mathbf{v}_{k+1} \|_H^2 = 0.$$

This also implies $\mathbf{r}_i \cdot \mathbf{r}_{k+1} = 0$ for $0 \leq i \leq k$, since $\mathbf{r}_0 = \mathbf{v}_1$ while $\mathbf{r}_i = \mathbf{v}_{i+1} - s_i \mathbf{v}_i$ for $i \geq 1$.

- Page 213:

  Change x to $\mathbf{x}$ on the line immediately below Eq. (6.79).

- Page 214: Definition 6.32:

  Change the definition to read "is the $n \times n$ self-adjoint matrix defined by the inequality".

- Page 270: Equation 7.53:

  There are missing terms on the right hand side that are independent of w_i (and so do not affect any subsequent computations). The equation and the 2 lines that follow should read as follows:

$$\| X\mathbf{w} - \mathbf{y} \|^2 + \lambda \| \mathbf{w} \|_1 = \| \mathbf{v}_i \|^2 w_i^2 - 2 w_i b_i + \lambda | w_i | + c, \tag{0.1}$$

for any $i = 1, \ldots, n$, where

$$b_i = \mathbf{v}_i \cdot \left(\mathbf{y} - \sum_{j \neq i} w_j \mathbf{v}_j \right) \qquad \text{and} \qquad c = \| \mathbf{y} - \sum_{j \neq i} w_j \mathbf{v}_j \|^2 + \lambda \sum_{j \neq i} |w_j|.$$

- Page 225: Equation 6.109:

  Change μ^2 to μ.

- Page 275: Equation 7.62:

  Change the equation to read

$$\min_{\substack{\mathbf{w} \in \mathbb{R}^n \\ b \in \mathbb{R}}} \left\{ \| \mathbf{w} \|^2 \;\middle|\; y_i(\mathbf{x}_i \cdot \mathbf{w} - b) \geq 1, \quad i = 1, \ldots, m \right\}.$$

- Page 275: Equation 7.63:

  Change the equation to read

$$\min_{\mathbf{w}, b} \left\{ \| \mathbf{w} \|^2 \;\middle|\; \mathbf{z} \cdot \mathbf{w} - b \geq 1, \quad \mathbf{z} \cdot \mathbf{w} + b \geq 1 \right\}.$$

- Page 300: Line directly above Theorem 7.17:

  Remove the text ", as well as soft-margin SVM (7.64)". Technically soft-margin SVM does not fit into this loss function due to the extra parameter b.

- Page 302: Definition 7.18:

  The definition of Mercer kernel should be amended to include a requirement that $\mathcal{K}$ is continuous.

- Page 306: Theorem 7.22:

  The assumptions on convexity/concavity of F should be switched, and D must be nonempty. Also, while the result in Fan [72] holds for nonconvex setes D, Fan has a more general notion of convex/concave functions than we have given in the book, so to be more clear we also assume D is convex. The correct statement of the theorem should read.

  Theorem 7.22. *Let $D \subset \mathbb{R}^m$ be convex, compact, i.e., closed and bounded, and nonempty. Let $F\colon D \times \mathbb{R}^n \to \mathbb{R}$ be continuous, and suppose that $F(\mathbf{x}, \mathbf{y})$ is convex as a function of $\mathbf{y}$ for each fixed $\mathbf{x} \in D$, while $F(\mathbf{x}, \mathbf{y})$ is concave, i.e., $-F(\mathbf{x}, \mathbf{y})$ is convex, as a function of $\mathbf{x}$ for each fixed $\mathbf{y} \in \mathbb{R}^n$. Then,*

  $$\min_{\mathbf{y}\in\mathbb{R}^n} \max_{\mathbf{x}\in D} F(\mathbf{x},\mathbf{y}) = \max_{\mathbf{x}\in D} \min_{\mathbf{y}\in\mathbb{R}^n} F(\mathbf{x},\mathbf{y}).$$

  As a side note, Theorem 7.22 is one place in the book where the minimum is really an infimum.

  For an elementary self-contained proof of Theorem 7.22 in the setting required in the book, we refer to the note [1]. In particular, this note explains saddle points for minimax problems, which are important for understanding the formula for b on page 307 in the book.

- Page 307: Equation (7.112):

  Change $\sum_{i=1}^{n}$ to $\sum_{i=1}^{m}$.

- Page 307: Remark 7.23:

  Replace "exactly those" with "a subset of the".

- Page 309:

  Change $\sum_{i=1}^{n}$ to $\sum_{i=1}^{m}$.

- Page 319: Section 8.8.1. Kernel Principal Component Analysis

  Throughout this section there should be a standing assumption that $d \geq m$ (i.e., the feature dimension is at least as large as the number of data points). Also, we can of course only define the principal components $\mathbf{q}_i$ when $\lambda_i > 0$, and we should also check that $\mathbf{v}_1, \ldots, \mathbf{v}_k$ are orthogonal. Thus, lines -10 through -5 should be replaced with the text below.

Notice that

$$\mathbf{v}_i \cdot \mathbf{v}_j = (\underline{Z}^T \mathbf{p}_i)^T \underline{Z}^T \mathbf{p}_j = \mathbf{p}_i^T \underline{Z}\,\underline{Z}^T \mathbf{p}_j = \lambda_j\, \mathbf{p}_i \cdot \mathbf{p}_j.$$

Thus, if $\lambda_1, \ldots, \lambda_k > 0$, then $\mathbf{v}_1, \ldots, \mathbf{v}_k$ are orthogonal and the top k principal components of the feature vector data $\mathbf{z}_1, \ldots, \mathbf{z}_m$ are given by the corresponding unit vectors

$$\mathbf{q}_i = \frac{\mathbf{v}_i}{\|\mathbf{v}_i\|} = \lambda_i^{-1/2}\, \underline{Z}^T \mathbf{p}_i \qquad \text{for} \qquad i = 1, \ldots, k.$$

- Page 320: Figure 8.8:

Add the line "The colors in the second and third figures in each row indicate which cluster each point is from in the original data depicted on the left." to the figure caption for clarity.

- Page 320: Kernel PCA and Whitening

  The version of kernel PCA described in the book includes whitening. There is also a version without whitening. It is worth making a comment about this. The following paragraph fits nicely at the end of Subsection 8.1.1, after the second to last paragraph in the section.

It is important to note that the method described above is kernel PCA with *whitening*. Indeed, provided $\lambda_k > 0$, so $J\mathbf{p}_i = \mathbf{p}_i$ for $i = 1, \ldots, k$, we have $JP_k = P_k$ and so the covariance matrix of the kernel PCA data is

$$S_{P_k} = P_k^T P_k = \mathrm{I},$$

since the columns of P_k are orthonormal vectors. To produce a non-whitened version of kernel PCA, which is the standard setting for PCA, we define $Q_k = P_k\Sigma_k$, where $\Sigma_k = \operatorname{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_k}) = \operatorname{diag}(\sigma_1, \ldots, \sigma_k)$. Then the covariance matrix of Q_k is given by

$$S_{Q_k} = Q_k^T Q_k = \Sigma_k P_k^T P_k \Sigma_k = \Sigma_k^2,$$

which is the same as for PCA (see (8.7)).

- Page 320: Theorem 8.4:

  The formula for $\mathbf{q}_i$ should read $\mathbf{q}_i = \lambda_i^{-1/2} Z^T \mathbf{p}_i$.

- Page 326: line after (8.18):

  Change "...follows from (2.41)." to "...follows from (2.40)."

- Page 333: Equation (8.33):

  Change Q to Q^T.

- Page 333: Equation (8.34):

  Change Q_k to Q_k^T.

- Page 346, Exercise 4.1:

  Delete "How does the accuracy change with the number of principal components used?"

- Page 352:

  Change "is inversely proportional to similarity" to "decreases as similarity increases".

- Page 541, Eq. (10.94):

  Remove extra left parenthesis, so equation should read

$$\Phi(\mathbf{x}) = \prod_{i=1}^{n} \theta(x_i) = \prod_{i=1}^{n} \big(1 - |\, x_i \,|\big)_+.$$

- Page 543, line -6:

  Remove first “the” in “...it is the only the lower intrinsic...”.

Acknowledgements:

Many thanks to Larry Baker, Gilad Lerman and Markus Schweighofer for sending us their corrections to the printed text.

Bibliography

[1] J. Calder. An Elementary Proof of a Minimax Theorem. *arXiv:2511.19416*, 2025. [arXiv].