“Best” stepsizes for GD of Quadratic Function with known eigenvalues

Consider a quadratic function with $A \succeq 0$, $b \in C(A)$

$$f(x) = \frac{1}{2}x^\top Ax + b^\top x + c$$

with $A \succeq 0$ and $b \in C(A)$. Analysis of GD for this function is equivalent to the analysis of GD for the function

$$g(x) = \frac{1}{2}x^\top Ax$$

Consider one GD step for this case:

$$x_{k+1} = x_k - \alpha_k Ax_k = (I - \alpha_k A)x_k$$

Consider the change of variables $x = H^\top y$, where H is an orthogonal matrix whose columns are eigenvectors of A; then $HAH^\top = D$ is the diagonal matrix with eigenvalues of A on the diagonal:

$$D = diag\{\lambda_1, \lambda_2 \dots, \lambda_m, \lambda_{m+1} \dots \lambda_n\}$$

where

- $\lambda_j > 0$ for $1 \le j \le m$
- Zero otherwise

Under this substitution, we get the following equation for $y_k = H^\top x_k$:

$$y_{k+1} = (I - \alpha_k D)y_k$$

Consider the following stepsizes:

$$\alpha_k = \frac{1}{\lambda_k}, \quad 1 \le k \le m$$

Let's analyze the behaviour of the algorithm on the first step: $y_0 = \begin{pmatrix} c_1 & c_2 & \dots & c_n \end{pmatrix}^\top$

$$y_1 = \left(I - \frac{1}{\lambda_1} D\right) y_0$$

Which will result in $y_1 = \begin{pmatrix} 0 & c_2^{(1)} & \dots & c_n^{(1)} \end{pmatrix}^\top$ - first component will be zero as $I - \frac{1}{\lambda_1} = diag\left\{0, 1 - \frac{\lambda_2}{\lambda_1}, \dots, 1 - \frac{\lambda_n}{\lambda_1}\right\}$.

In general on j-th iteration the $c_j^{(j)}$ will become zero and will stay so in next steps. Algorithm will converge in number of steps which is equal to the number of distinct eigenvalues.

💡 **Correction from Rostyslav**

Assume that $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_k > 0$ are all positive eigenvalues of the matrix A and set

$$q(\lambda) = (\lambda - \lambda_k) \cdots (\lambda - \lambda_2)(\lambda - \lambda_1)$$

According to the Hamilton-Cayley theorem, the matrix A is a zero of its characteristic polynomial $p(\lambda) = \lambda^{n-k} q(\lambda)$, i.e.,

$$p(A) = A^{n-k} q(A) = A^{n-k}(A - \lambda_k I) \cdots (A - \lambda_2 I)(A - \lambda_1 I) = 0$$

Apply the gradient descent with step sizes $\alpha_j = 1/\lambda_j$ for $j = 1, 2, \dots, k$. Since one step of the GD at x produces x^+ of the form

$$x^+ = x - \alpha \nabla f(x) = x - \alpha A x = (I - \alpha A) x,$$

after k steps, we'll see that

$$x_k = (I - \alpha_k A) \cdots (I - \alpha_2 A)(I - \alpha_1 A) x_0 = (-1)^k (\lambda_1 \lambda_2 \cdots \lambda_k)^{-1} q(A) x_0$$

satisfies $A^{n-k} x_k = 0$. Therefore, x_k is in the nullspace of A and thus a minimiser of f.