Consider the problem of minimizing

$$f(\mathbf{x}) = \tfrac{1}{2}\mathbf{x}^T A\mathbf{x} - \mathbf{x}^T\mathbf{b},$$

for $\mathbf{x} \in \mathbb{R}^n$, where A is a given SPD (symmetric positive-definite) matrix, and $\mathbf{b}$ is a given vector. First, we note a few easy-to-prove facts:

1. $\nabla f(\mathbf{x}) = \frac{1}{2}(A^T + A)\mathbf{x} - \mathbf{b} = A\mathbf{x} - \mathbf{b}$ (since A is symmetric, i.e., $A^T = A$).
2. $\nabla\nabla^T f(\mathbf{x}) = \frac{1}{2}(A^T + A) = A$. In particular, f is a convex function, since its Hessian is positive-definite.
3. Since A is SPD, A is invertible, so $A\mathbf{x} = \mathbf{b}$ has a unique solution.
4. The problem of minimizing $f(\mathbf{x})$ and the problem of solving $A\mathbf{x} = \mathbf{b}$ are equivalent, in the sense that they have the same solution.

Let us consider an iteration scheme for the problem, given by

$$\mathbf{x}_{i+1} = \mathbf{x}_i + \alpha_i \mathbf{p}_i \tag{1}$$

where the vector $\mathbf{p}_i$ and the scalar α_i and are to be chosen. (The vector $\mathbf{p}_i$ is called the search direction.)

Let $\mathbf{x}$ be the exact solution, i.e., $\mathbf{x}$ satisfies $A\mathbf{x} = \mathbf{b}$. We define:

$$\begin{aligned} \mathbf{e}_i &= \mathbf{x}_i - \mathbf{x} = \text{the error} \\ \mathbf{r}_i &= \mathbf{b} - A\mathbf{x}_i = \text{the residual (i.e., the error in the output)} \end{aligned}$$

Note that

$$A\mathbf{e}_i = A(\mathbf{x}_i - \mathbf{x}) = A\mathbf{x}_i - A\mathbf{x} = A\mathbf{x}_i - \mathbf{b} = -\mathbf{r}_i \tag{2}$$

We now make a choice:

> Let us decide that the search direction will be in the direction of steepest descent from $\mathbf{x}_i$, that is:
> $\mathbf{p}_i = -\nabla f(\mathbf{x}_i)$

Then,

$$\mathbf{p}_i = -\nabla f(\mathbf{x}_i) = -(A\mathbf{x}_i - \mathbf{b}) = \mathbf{r}_i$$

so that (1) becomes

$$\mathbf{x}_{i+1} = \mathbf{x}_i + \alpha_i \mathbf{r}_i \tag{3}$$

Now that we have decided on $\mathbf{p}_i$, we need to determine α_i. This can be done be choosing the best possible value by letting α vary, and using calculus. Consider the function defined by

$$\varphi(\alpha) := f(\mathbf{x}_i + \alpha\mathbf{r}_i)$$

Since f is convex, φ has a unique global minimum. Since we are trying to minimize f, the minimizer of φ will be the α_i we pick. To find what it is, we set $\varphi'(\alpha) = 0$ and use the multi-variable chain rule to compute:

$$\begin{aligned}
0 = \varphi'(\alpha) &= \mathbf{r}_i^T \nabla f(\mathbf{x}_i + \alpha\mathbf{r}_i) \\
&= \mathbf{r}_i^T (A(\mathbf{x}_i + \alpha\mathbf{r}_i) - \mathbf{b}) \\
&= \mathbf{r}_i^T (A\mathbf{x}_i - \mathbf{b} + \alpha A\mathbf{r}_i) \\
&= \mathbf{r}_i^T (-\mathbf{r}_i + \alpha A\mathbf{r}_i) \\
&= -\mathbf{r}_i^T\mathbf{r}_i + \alpha\mathbf{r}_i^T A\mathbf{r}_i.
\end{aligned}$$

Solving for α (and calling it α_i), we find:

$$\alpha_i = \frac{\mathbf{r}_i^T\mathbf{r}_i}{\mathbf{r}_i^T A\mathbf{r}_i}.$$

Note that, since A is SPD, if $\mathbf{r}_i \neq \mathbf{0}$, then $\mathbf{r}_i^T A\mathbf{r}_i > 0$, to there is no divide-by zero error. On the other hand, if $\mathbf{r}_i = \mathbf{0}$, then the algorithm can stop, since this means we have found an exact solution! Our iteration scheme can be written down as follows.

Steepest Descent Algorithm (naïve form):
Given $\mathbf{x}_i$, $\mathbf{b}$, and an SPD matrix A, set

$$\begin{aligned}
\mathbf{r}_i &= \mathbf{b} - A\mathbf{x}_i \\
\alpha_i &= \frac{\mathbf{r}_i^T\mathbf{r}_i}{\mathbf{r}_i^T A\mathbf{r}_i} \\
\mathbf{x}_{i+1} &= \mathbf{x}_i + \alpha_i\mathbf{r}_i
\end{aligned}$$

Next, we note an important fact about the steepest descent algorithm: successive residuals are orthogonal. To see this, note that, using the above algorithm:

$$\begin{aligned}
\mathbf{r}_i^T\mathbf{r}_{i+1} &= \mathbf{r}_i^T(\mathbf{b} - A\mathbf{x}_{i+1}) = \mathbf{r}_i^T(\mathbf{b} - A(\mathbf{x}_i + \alpha_i\mathbf{r}_i)) \\
&= \mathbf{r}_i^T(\mathbf{b} - A\mathbf{x}_i - \alpha_i A\mathbf{r}_i) \\
&= \mathbf{r}_i^T(\mathbf{r}_i - \alpha_i A\mathbf{r}_i) \\
&= \mathbf{r}_i^T\mathbf{r}_i - \alpha_i\mathbf{r}_i^T A\mathbf{r}_i = \mathbf{r}_i^T\mathbf{r}_i - \left(\frac{\mathbf{r}_i^T\mathbf{r}_i}{\mathbf{r}_i^T A\mathbf{r}_i}\right)\mathbf{r}_i^T A\mathbf{r}_i = 0.
\end{aligned}$$

Thus, $\mathbf{r}_i$ is orthogonal to $\mathbf{r}_{i+1}$.

Another thing to notice is that

$$\mathbf{r}_{i+1} = \mathbf{b} - A\mathbf{x}_{i+1} = \mathbf{b} - A(\mathbf{x}_i + \alpha_i \mathbf{r}_i) = \mathbf{b} - A\mathbf{x}_i - \alpha_i A\mathbf{r}_i = \mathbf{r}_i - \alpha_i A\mathbf{r}_i.$$

Thus, we don't really need to compute $A\mathbf{x}_i$ to find $\mathbf{r}_{i+1}$, so long as we store $\mathbf{r}_i$ and $A\mathbf{r}_i$ that we computed on the previous step. This can reduce the computational cost at the (usually small) cost of storing two additional vectors. The revised algorithms looks like this:

Steepest Descent Algorithm (improved form):
Given $\mathbf{x}_i$, $\mathbf{b}$, and an SPD matrix A, and the vectors $\mathbf{r}_{i-1}$ and $\mathbf{z}_{i-1} := A\mathbf{r}_{i-1}$ from the previous step, compute

$$\begin{aligned} \mathbf{r}_i &= \mathbf{r}_{i-1} - \alpha_i \mathbf{z}_{i-1} \\ \mathbf{z}_i &= A\mathbf{r}_i \\ \alpha_i &= \frac{\mathbf{r}_i^T \mathbf{r}_i}{\mathbf{r}_i^T \mathbf{z}_i} \\ \mathbf{x}_{i+1} &= \mathbf{x}_i + \alpha_i \mathbf{r}_i \end{aligned}$$

The fact that we don't have to compute $A\mathbf{x}_i$ anymore is often a great improvement. The revised algorithm requires only one matrix-vector multiplication per iteration. The algorithm itself is mathematically identical (although round-off errors may make the algorithms computationally different).

Next, let us consider the error. Above, we defined $\mathbf{e}_i = \mathbf{x}_i - \mathbf{x}$. Thus, from the steepest descent algorithm,

$$\begin{aligned} & \mathbf{x}_{i+1} = \mathbf{x}_i + \alpha_i \mathbf{r}_i \\ \Rightarrow \quad & \mathbf{x}_{i+1} - \mathbf{x} = \mathbf{x}_i - \mathbf{x} + \alpha_i \mathbf{r}_i \\ \Rightarrow \quad & \mathbf{e}_{i+1} = \mathbf{e}_i + \alpha_i \mathbf{r}_i \qquad (4) \\ \Rightarrow \quad & A\mathbf{e}_{i+1} = A\mathbf{e}_i + \alpha_i A\mathbf{r}_i \qquad (5) \\ \Rightarrow \quad & -\mathbf{r}_{i+1} = -\mathbf{r}_i + \alpha_i A\mathbf{r}_i \qquad (6) \end{aligned}$$

where we used (2). Now, we can't say very much about the convergence rate from equation (4) directly.

However, using (2) and equation (4), we find

$$\begin{aligned}
\mathbf{e}_{i+1}^T A\mathbf{e}_{i+1} &= (\mathbf{e}_i + \alpha_i \mathbf{r}_i)^T(-\mathbf{r}_{i+1}) \\
&= -\mathbf{e}_i^T \mathbf{r}_{i+1} - \alpha_i \mathbf{r}_i^T \mathbf{r}_{i+1} \\
&= -\mathbf{e}_i^T \mathbf{r}_{i+1} && (\text{since } \mathbf{r}_i^T \mathbf{r}_{i+1} = 0) \\
&= \mathbf{e}_i^T(-\mathbf{r}_i + \alpha_i A\mathbf{r}_i) && (\text{using equation (6)}) \\
&= (-A^{-1}\mathbf{r}_i)^T(-\mathbf{r}_i + \alpha_i A\mathbf{r}_i) && (\text{using (2)}) \\
&= -\mathbf{r}_i^T A^{-1}(-\mathbf{r}_i + \alpha_i A\mathbf{r}_i) && (\text{since } A \text{ is symmetric}) \\
&= \mathbf{r}_i^T(A^{-1}\mathbf{r}_i - \alpha_i \mathbf{r}_i) \\
&= \mathbf{r}_i^T A^{-1}\mathbf{r}_i - \alpha_i \mathbf{r}_i^T \mathbf{r}_i && (\text{using (2)}) \\
&= \mathbf{r}_i^T A^{-1}\mathbf{r}_i - \left(\frac{\mathbf{r}_i^T \mathbf{r}_i}{\mathbf{r}_i^T A\mathbf{r}_i}\right)\mathbf{r}_i^T \mathbf{r}_i \\
&= \mathbf{r}_i^T A^{-1}\mathbf{r}_i \left(1 - \frac{(\mathbf{r}_i^T \mathbf{r}_i)^2}{(\mathbf{r}_i^T A\mathbf{r}_i)(\mathbf{r}_i^T A^{-1}\mathbf{r}_i)}\right) \\
&= (-A\mathbf{e}_i)^T A^{-1}(-A\mathbf{e}_i)\left(1 - \frac{(\mathbf{r}_i^T \mathbf{r}_i)^2}{(\mathbf{r}_i^T A\mathbf{r}_i)(\mathbf{r}_i^T A^{-1}\mathbf{r}_i)}\right) && (\text{using (2)}) \\
&= \mathbf{e}_i^T A^T A^{-1}(A\mathbf{e}_i)\left(1 - \frac{(\mathbf{r}_i^T \mathbf{r}_i)^2}{(\mathbf{r}_i^T A\mathbf{r}_i)(\mathbf{r}_i^T A^{-1}\mathbf{r}_i)}\right) \\
&= \mathbf{e}_i^T A\mathbf{e}_i \left(1 - \frac{(\mathbf{r}_i^T \mathbf{r}_i)^2}{(\mathbf{r}_i^T A\mathbf{r}_i)(\mathbf{r}_i^T A^{-1}\mathbf{r}_i)}\right) && (\text{since } A \text{ is symmetric}).
\end{aligned}$$

Let us introduce the notation $\|\mathbf{x}\|_A^2 = \mathbf{x}^T A\mathbf{x}$. It is straight-forward to show that, so long as A is SPD, $\|\cdot\|_A$ is a norm. Thus, the above relation can be simplified to

$$\|\mathbf{e}_{i+1}\|_A^2 = \|\mathbf{e}_i\|_A^2 \left(1 - \frac{\|\mathbf{r}_i\|^4}{\|\mathbf{r}_i\|_A^2 \|\mathbf{r}_i\|_{A^{-1}}^2}\right).$$

(Note that the above identity shows that if $\mathbf{r}_i$ happens to be an eigenvector of A, then the convergence is immediate).

Let Λ and λ be the largest and smallest eigenvalues of A, respectively (recall that, since A is SPD, all its eigenvalues are positive). It a standard result that

$$\frac{1}{\lambda} = \max_{\mathbf{x}\neq 0} \frac{\mathbf{x}^T A^{-1}\mathbf{x}}{\mathbf{x}^T \mathbf{x}} \qquad \text{and} \qquad \Lambda = \max_{\mathbf{x}\neq 0} \frac{\mathbf{x}^T A\mathbf{x}}{\mathbf{x}^T \mathbf{x}}.$$

Thus,

$$1 - \frac{(\mathbf{r}_i^T \mathbf{r}_i)^2}{(\mathbf{r}_i^T A \mathbf{r}_i)(\mathbf{r}_i^T A^{-1} \mathbf{r}_i)} = 1 - \frac{1}{\frac{(\mathbf{r}_i^T A \mathbf{r}_i)}{(\mathbf{r}_i^T \mathbf{r}_i)} \frac{(\mathbf{r}_i^T A^{-1} \mathbf{r}_i)}{(\mathbf{r}_i^T \mathbf{r}_i)}} \leq 1 - \frac{1}{\Lambda \frac{1}{\lambda}} = 1 - \frac{\lambda}{\Lambda}.$$

Therefore, if $\lambda < \Lambda$ (that is, there is a "gap" between the largest and smallest eigenvalues), then

$$\|\mathbf{e}_{i+1}\|_A^2 \leq \|\mathbf{e}_i\|_A^2 \left(1 - \tfrac{\lambda}{\Lambda}\right) < \|\mathbf{e}_i\|_A^2$$

Thus, the steepest descent method must converge. In fact, a hard lemma called the Kantorovich Lemma shows a sharper bound, namely, it implies that, for any $\mathbf{x} \neq \mathbf{0}$,

$$\left(\frac{\mathbf{x}^T A \mathbf{x}}{\mathbf{x}^T \mathbf{x}}\right) \left(\frac{\mathbf{x}^T A^{-1} \mathbf{x}}{\mathbf{x}^T \mathbf{x}}\right) \leq \frac{(\Lambda + \lambda)^2}{4\lambda\Lambda}$$

so that

$$1 - \frac{1}{\frac{(\mathbf{r}_i^T A \mathbf{r}_i)}{(\mathbf{r}_i^T \mathbf{r}_i)} \frac{(\mathbf{r}_i^T A^{-1} \mathbf{r}_i)}{(\mathbf{r}_i^T \mathbf{r}_i)}} \leq 1 - \frac{4\lambda\Lambda}{(\Lambda + \lambda)^2} = \frac{(\Lambda - \lambda)^2}{(\Lambda + \lambda)^2} = \left(\frac{1 - \frac{\lambda}{\Lambda}}{1 + \frac{\lambda}{\Lambda}}\right)^2$$

Thus, we have the sharper bound

$$\|\mathbf{e}_{i+1}\|_A \leq \frac{1 - \frac{\lambda}{\Lambda}}{1 + \frac{\lambda}{\Lambda}} \|\mathbf{e}_i\|_A.$$

In particular,

$$\lim_{i \to \infty} \frac{\|\mathbf{e}_{i+1}\|_A}{\|\mathbf{e}_i\|_A} \leq \frac{1 - \frac{\lambda}{\Lambda}}{1 + \frac{\lambda}{\Lambda}} = \frac{\Lambda - \lambda}{\Lambda + \lambda},$$

so that the convergence is at least linear. In practice, the convergence is no better than linear for a general SPD matrix. Note also that the larger $\Lambda - \lambda$ is, the slower the convergence. Matrices which have one very large eigenvalue and one very small eigenvalue are often called "ill-conditioned" or "stiff". Solving such problems via a direct approach can often lead to slow convergence. Hence methods such as "preconditioning" are often employed.