Effect of scaling on the method of steepest descent

Let $A \in S^n$ be positive definite, $f : \mathbb{R}^n \to \mathbb{R}$ continuously differentiable. Using the change of variables $x := Ay$, define $h(y) := f(Ay)$.

Compute the gradient of h.

Answer: It holds: $\nabla h(y) = A^\top \nabla f(x) = A\, \nabla f(x)$

Proof:
First we show the following: Let $g : \mathbb{R}^n \to \mathbb{R}^n$, with $g(y) = Ay$ for a matrix $A \in \mathbb{R}^{n\times n}$.
Then, its derivative (Jacobian) at any point y is $J_g(y) = A$.

Since $x := g(y) = Ay$ has components $x_i = \sum_{k=1}^{n} a_{ik} y_k$, it follows for the components:

$$\frac{\partial g_i}{\partial y_j}(x) = \frac{\partial x_i}{\partial y_j} = a_{ij}$$

and thus the derivative (Jacobian) of g at point y is $J_g(y) = \left(\frac{\partial g_i}{\partial y_j}(y)\right)_{\substack{i=1,\dots,n\\ j=1,\dots,n}} = (a_{ij})_{\substack{i=1,\dots,n\\ j=1,\dots,n}} = A$

Recall that the Jacobian of a mapping $f : \mathbb{R}^n \to \mathbb{R}^m$ at a point $x \in \mathbb{R}^n$, is defined as:

$$J_f(x) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(x) & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1}(x) & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial f_m}{\partial x_1}(x) & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix} = \begin{pmatrix} \nabla f_1^\top(x) \\ \nabla f_2^\top(x) \\ \vdots \\ \nabla f_m^\top(x) \end{pmatrix}$$

such that for a scalar-valued function $f : \mathbb{R}^n \to \mathbb{R}$, it holds $\nabla f(x) = J_f(x)^\top$.

From the definition above, we see that $h(y) := f(Ay) = f(g(y)) = (f \circ g)(y)$. Using the Jacobian matrix $J_f(y)$ as representation of the total derivative h' at a point y (see the statement at the bottom), we get:

$$J_h(y) = J_{(f\circ g)}(y) = J_f(g(y)) J_g(y)$$

Since $J_h = \nabla^\top h$, we get: $\nabla^\top h(y) = \nabla^\top f(g(y)) A \implies \nabla h(y) = A^\top \nabla f(g(y))$.
By the symmetry of A and $x = g(y)$, we get the result.

Using a different (but equivalent) formulation of the multivariate chain rule frequently found in textbooks, and due to $x = g(y) = Ay$, and the symmetry of A, we also get the result:

$$\nabla_y h(y) = \nabla_y (f \circ g)(y) = \big(J_g(y)\big)^\top \big(\nabla_x f\big)\underbrace{(g(y))}_{=x} = A^\top \nabla_x f(x) \underset{A^\top = A}{=} A\, \nabla_x f(x)$$

The multivariate chain rule can be stated as follows:

Let $f : \mathbb{R}^p \to \mathbb{R}^m$, and $g : \mathbb{R}^n \to \mathbb{R}^p$, and $a \in \mathbb{R}^n$ be given.
Then, the composition $h = (f \circ g) : \mathbb{R}^n \to \mathbb{R}^m,\ h(a) := f(g(a))$ is well-defined.

If g is differentiable at point $a \in R^n$ with total derivative $g'(a)$
(*represented* by the Jacobian matrix $J_g(a)$),
and if f is differentiable at point $b = g(a)$ with total derivative $f'(b)$
(*represented* by the Jacobian matrix $J_f(b)$),
then the composition h is differentiable at point a with total derivative
$h'(a) = f'(b) \circ g'(a) = f'(g(a)) \circ g'(a)$,
(*represented* by the Jacobian matrix product $J_h(a) = J_f(g(a)) J_g(a)$).