# Optimal Step Size for Gradient Descent on a Quadratic Function

## Problem Setting

Consider the quadratic function:

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

where $A \in \mathbb{R}^{d\times d}$ is a symmetric positive definite matrix, $b \in \mathbb{R}^d$, and $c \in \mathbb{R}$ is a constant.

Its gradient is:

$$\nabla f(x) = Ax - b.$$

## Gradient Descent Update Rule

The update rule of gradient descent is:

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k),$$

where $\alpha_k$ is the step size at iteration $k$.

## Optimal Step Size Derivation

We seek to minimize the function along the gradient direction:

$$\phi(\alpha) = f(x_k - \alpha\nabla f(x_k)).$$

Let $g_k := \nabla f(x_k) = Ax_k - b$. Then:

$$\phi(\alpha) = f(x_k - \alpha g_k) = \frac{1}{2}(x_k - \alpha g_k)^\top A(x_k - \alpha g_k) - b^\top(x_k - \alpha g_k) + c.$$

Expanding:

$$\phi(\alpha) = \frac{1}{2}x_k^\top Ax_k - \alpha x_k^\top Ag_k + \frac{1}{2}\alpha^2 g_k^\top Ag_k - b^\top x_k + \alpha b^\top g_k + c.$$

This is a quadratic function in $\alpha$, so the minimizer is found by taking derivative:

$$\phi'(\alpha) = -x_k^\top Ag_k + \alpha g_k^\top Ag_k + b^\top g_k = 0.$$

Recall $g_k = Ax_k - b$, so $b = Ax_k - g_k$. Substitute:

$$-x_k^\top A g_k + \alpha g_k^\top A g_k + x_k^\top A g_k - g_k^\top g_k = 0.$$

Simplify:

$$\alpha g_k^\top A g_k - g_k^\top g_k = 0,$$

$$\alpha_k = \frac{g_k^\top g_k}{g_k^\top A g_k} = \frac{\|\nabla f(x_k)\|^2}{\nabla f(x_k)^\top A \nabla f(x_k)}.$$

# Conclusion

The optimal step size for each iteration of gradient descent applied to a quadratic function is:

$$\boxed{\alpha_k = \frac{\|\nabla f(x_k)\|^2}{\nabla f(x_k)^\top A \nabla f(x_k)}}.$$

This choice of $\alpha_k$ ensures the fastest decrease of the function value along the gradient direction at each iteration. Since $A$ is positive definite, this step size leads to convergence to the global minimum.

# Remarks

- This is known as the **exact line search** for quadratic functions.
- In one-dimensional case, i.e., $f(x) = \frac{1}{2}ax^2 + bx + c$, the optimal step is $\alpha = \frac{1}{a}$.
- In practice, computing $A\nabla f(x_k)$ may be expensive for large $A$, hence approximations are often used.