Gradient Descent for Solving Least-Squares Problems Proof: Bounds on Step Size for Guaranteed Convergence

Gradient descent minimizes the cost function

$$f(\boldsymbol{w}) = ||\boldsymbol{A}\boldsymbol{w} - \boldsymbol{d}||_2^2$$

using the iterative algorithm

$$\boldsymbol{w}^{(k+1)} = \boldsymbol{w}^{(k)} - \tau \boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{w}^{(k)} - \boldsymbol{d}), \quad k = 0, 1, 2, 3, \ldots \tag{1}$$

where $\tau > 0$ so that we modify the current iterate in the negative gradient direction. Often this algorithm is initialized with $\boldsymbol{w}^{(0)} = \boldsymbol{0}$. The initialization does not affect the convergence behavior because $f(\boldsymbol{w})$ is convex.

The iteration is guaranteed to converge to the minimum of the cost function if the squared error decreases with each iteration, that is, if

$$f(\boldsymbol{w}^{(k+1)}) = ||\boldsymbol{A}\boldsymbol{w}^{(k+1)} - \boldsymbol{d}||_2^2 < f(\boldsymbol{w}^{(k)}) = ||\boldsymbol{A}\boldsymbol{w}^{(k)} - \boldsymbol{d}||_2^2$$

Begin by substituting $\boldsymbol{w}^{(k)} - \tau \boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{w}^{(k)} - \boldsymbol{d})$ for $\boldsymbol{w}^{(k+1)}$ in $f(\boldsymbol{w}^{(k+1)})$ to write

$$\begin{aligned} f(\boldsymbol{w}^{(k+1)}) &= ||\boldsymbol{A}\left(\boldsymbol{w}^{(k)} - \tau \boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{w}^{(k)} - \boldsymbol{d})\right) - \boldsymbol{d}||_2^2 &(2)\\ &= ||(\boldsymbol{A}\boldsymbol{w}^{(k)} - \boldsymbol{d}) - \tau\left(\boldsymbol{A}\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{w}^{(k)} - \boldsymbol{d})\right)||_2^2 &(3)\end{aligned}$$

Now let $\boldsymbol{c} = \boldsymbol{A}\boldsymbol{w}^{(k)} - \boldsymbol{d}$ and $\boldsymbol{e} = \tau\left(\boldsymbol{A}\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{w}^{(k)} - \boldsymbol{d})\right)$ be the first and second terms in parentheses so $f(\boldsymbol{w}^{(k+1)}) = ||\boldsymbol{c} - \boldsymbol{e}||_2^2 = (\boldsymbol{c} - \boldsymbol{e})^T(\boldsymbol{c} - \boldsymbol{e})$. Expand the product to write $f(\boldsymbol{w}^{(k+1)}) = ||\boldsymbol{c}||_2^2 + ||\boldsymbol{e}||_2^2 - 2\boldsymbol{e}^T\boldsymbol{c}$. Substituting for $\boldsymbol{c}$ and $\boldsymbol{e}$ we thus obtain

$$\begin{aligned} f(\boldsymbol{w}^{(k+1)}) &= ||\boldsymbol{A}\boldsymbol{w}^{(k)} - \boldsymbol{d}||_2^2 + \tau^2||\boldsymbol{A}\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{w}^{(k)} - \boldsymbol{d})||_2^2 - 2\tau\left((\boldsymbol{A}\boldsymbol{w}^{(k)} - \boldsymbol{d})^T\boldsymbol{A}\boldsymbol{A}^T\right)(\boldsymbol{A}\boldsymbol{w}^{(k)} - \boldsymbol{d}) &(4)\\ &= f(\boldsymbol{w}^{(k)}) + \tau^2||\boldsymbol{A}\left(\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{w}^{(k)} - \boldsymbol{d})\right)||_2^2 - 2\tau\left((\boldsymbol{A}\boldsymbol{w}^{(k)} - \boldsymbol{d})^T\boldsymbol{A}\right)\left(\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{w}^{(k)} - \boldsymbol{d})\right) &(5)\end{aligned}$$

Define $\boldsymbol{v} = \boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{w}^{(k)} - \boldsymbol{d})$ to simplify the expression and rewrite Eq. 5 as

$$f(\boldsymbol{w}^{(k+1)}) = f(\boldsymbol{w}^{(k)}) + \tau^2||\boldsymbol{A}\boldsymbol{v}||_2^2 - 2\tau\boldsymbol{v}^T\boldsymbol{v}$$

Note that $\boldsymbol{v}$ does not depend on τ. Thus, to prove $f(\boldsymbol{w}^{(k+1)}) < f(\boldsymbol{w}^{(k)})$, we must find the condition for which

$$q(\tau) = \tau^2||\boldsymbol{A}\boldsymbol{v}||_2^2 - 2\tau\boldsymbol{v}^T\boldsymbol{v}$$

is less than zero.

Recall the operator norm of a matrix $\boldsymbol{X}$ satisfies $\max_{\boldsymbol{g}} ||\boldsymbol{X}\boldsymbol{g}||_2 \leq ||\boldsymbol{X}||_{op}||\boldsymbol{g}||_2$, so the first term in $q(\tau)$ may be upper bounded as

$$\tau^2||\boldsymbol{A}\boldsymbol{v}||_2^2 \leq \tau^2||\boldsymbol{A}||_{op}^2||\boldsymbol{v}||_2^2$$

We may rewrite the second term in $q(\tau)$ as

$$-2\tau\boldsymbol{v}^T\boldsymbol{v} = -2\tau||\boldsymbol{v}||_2^2$$

Hence, we obtain an upper bound on $q(\tau)$

$$q(\tau) \leq \tau^2||\boldsymbol{A}||_{op}^2||\boldsymbol{v}||_2^2 - 2\tau||\boldsymbol{v}||_2^2$$

Factoring out the common terms we write

$$q(\tau) \leq \left(\tau||\boldsymbol{A}||_{op}^2 - 2\right)\tau||\boldsymbol{v}||_2^2$$

The second term in $q(\tau)$, $\tau||\boldsymbol{v}||_2^2$, is positive provided $\boldsymbol{v} \neq \boldsymbol{0}$, so we obtain $q(\tau) < 0$ by requiring

$$\left(\tau||\boldsymbol{A}||_{op}^2 - 2\right) < 0$$

which indicates τ must satisfy

$$\tau < \frac{2}{||\boldsymbol{A}||_{op}^2}$$

Note that $\boldsymbol{v} = \boldsymbol{0}$ implies $\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{w}^{(k)} - \boldsymbol{d}) = \boldsymbol{0}$, or $\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{w}^{(k)} = \boldsymbol{A}^T\boldsymbol{d}$, or $\boldsymbol{w}^{(k)} = \left(\boldsymbol{A}^T\boldsymbol{A}\right)^{-1}\boldsymbol{A}^T\boldsymbol{d}$. Thus, if $\boldsymbol{v} = \boldsymbol{0}$, then the iteration has converged to the minimum of the squared error and the update term in Eq. 1 is zero.

Hence, the gradient descent algorithm will converge to the minimum of the squared error cost function provided the step-size τ satisfies $\tau < \frac{2}{||\boldsymbol{A}||_{op}^2}$.