Chapter 8 Gradient Methods

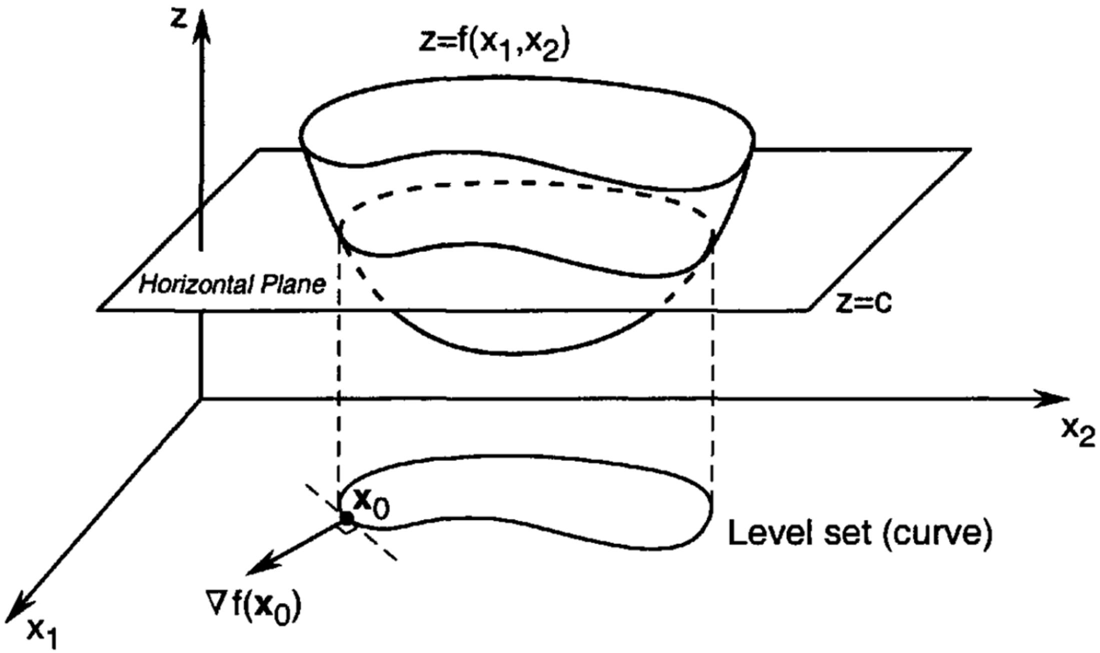
An Introduction to Optimization

Spring, 2014

Wei-Ta Chu

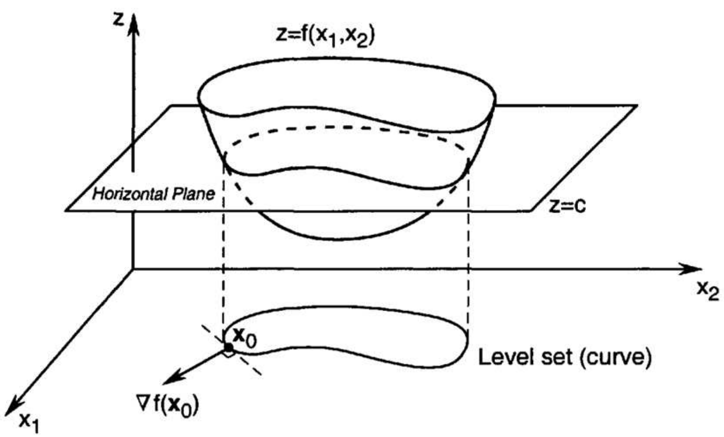
Introduction

- Recall that a *level set* of a function $f : R^n \to R$ is the set of points $\boldsymbol{x}$ satisfying $f(\boldsymbol{x}) = c$ for some constant c. Thus, a point $\boldsymbol{x}_0 \in R^n$ is on the level set corresponding to level c if $f(\boldsymbol{x}_0) = c$
- In the case of functions of two real variables, $f : R^2 \to R$



Introduction

- The gradient of f at $\boldsymbol{x}_0$, denoted by $\nabla f(\boldsymbol{x}_0)$, is orthogonal to the tangent vector to an arbitrary smooth curve passing through $\boldsymbol{x}_0$ on the level set $f(\boldsymbol{x}) = c$
- The direction of maximum rate of increase of a real-valued differentiable function at a point is orthogonal to the level set of the function through that point.
- The gradient acts in such a direction that for a given small displacement, the function f increases more in the direction of the gradient than in any other direction.



$$\langle \nabla f(\boldsymbol{x}), \boldsymbol{d} \rangle \le \|\nabla f(\boldsymbol{x})\| \|\nabla \boldsymbol{d}\|$$

Cauchy-Schwarz inequality

Introduction

- Recall that $\langle \nabla f(\boldsymbol{x}), \boldsymbol{d} \rangle$, $\|\boldsymbol{d}\| = 1$, is the rate of increase of f in the direction $\boldsymbol{d}$ at the point $\boldsymbol{x}$. By the Cauchy-Schwarz inequality,

$$\langle \nabla f(\boldsymbol{x}), \boldsymbol{d} \rangle \le \|\nabla f(\boldsymbol{x})\|$$

because $\|\boldsymbol{d}\| = 1$. But if $\boldsymbol{d} = \nabla f(\boldsymbol{x}) / \|\nabla f(\boldsymbol{x})\|$, then

$$\left\langle \nabla f(\boldsymbol{x}), \frac{\nabla f(\boldsymbol{x})}{\|\nabla f(\boldsymbol{x})\|} \right\rangle = \|\nabla f(\boldsymbol{x})\|$$

- Thus, the direction in which $\nabla f(\boldsymbol{x})$ points is the direction of maximum rate of increase of f at $\boldsymbol{x}$.
- The direction in which $-\nabla f(\boldsymbol{x})$ points is the direction of maximum rate of decrease of f at $\boldsymbol{x}$.
- Hence, the direction of negative gradient is a good direction to search if we want to find a function minimizer.

Introduction

- Let $\boldsymbol{x}^{(0)}$ be a starting point, and consider the point $\boldsymbol{x}^{(0)} - \alpha \bigtriangledown f(\boldsymbol{x}^{(0)})$ Then, by Taylor's theorem, we obtain

$$f(\boldsymbol{x}^{(0)} - \alpha \bigtriangledown f(\boldsymbol{x}^{(0)})) = f(\boldsymbol{x}^{(0)}) - \alpha \|\bigtriangledown f(\boldsymbol{x}^{(0)})\|^2 + o(\alpha)$$

- If $\bigtriangledown f(\boldsymbol{x}^{(0)}) \neq \mathbf{0}$, then for sufficiently small $\alpha > 0$, we have

$$f(\boldsymbol{x}^{(0)} - \alpha \bigtriangledown f(\boldsymbol{x}^{(0)})) < f(\boldsymbol{x}^{(0)})$$

- This means that the point $\boldsymbol{x}^{(0)} - \alpha \bigtriangledown f(\boldsymbol{x}^{(0)})$ is an improvement over the point $\boldsymbol{x}^{(0)}$ if we are searching for a minimizer.

Introduction

- Given a point $\boldsymbol{x}^{(k)}$, to find the next point $\boldsymbol{x}^{(k+1)}$, we move by an amount $-\alpha_k \nabla f(\boldsymbol{x}^{(k)})$, where α_k is a positive scalar called the ***step size***.

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \alpha_k \nabla f(\boldsymbol{x}^{(k)})$$

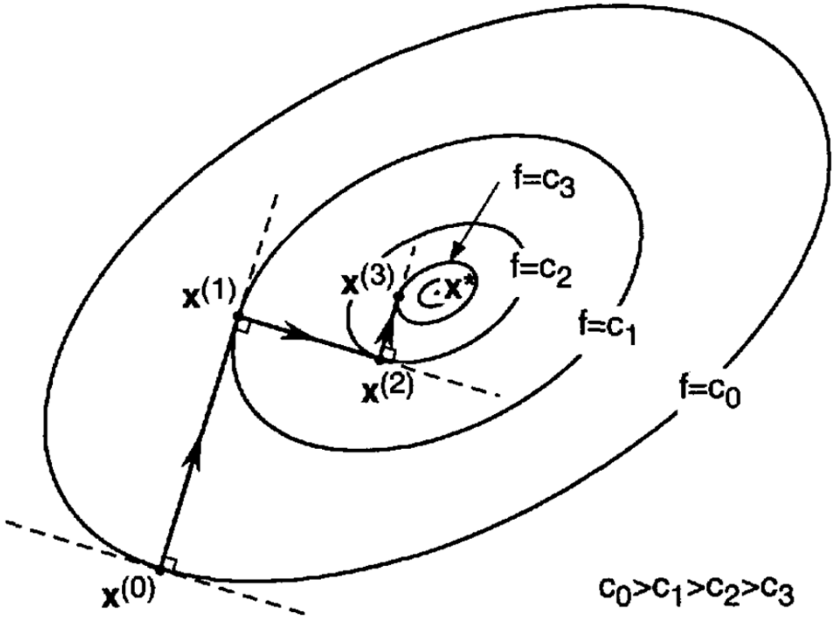
- We refer to this as a ***gradient descent algorithm*** (or ***gradient algorithm***). The gradient varies as the search proceeds, tending to zero as we approach the minimizer.
- We can take very small steps and reevaluate the gradient at every step, or take large steps each time. The former results in a laborious method of reaching the minimizer, whereas the latter may result in a more zigzag path the minimizer.

The Method of Steepest Descent

- Steepest descent is a gradient algorithm where the step size α_k is chosen to achieve the maximum amount of decrease of the objective function at each individual step.

$$\alpha_k = \arg\min_{\alpha \geq 0} f(\boldsymbol{x}^{(k)} - \alpha \bigtriangledown f(\boldsymbol{x}^{(k)}))$$

- At each step, starting from the point $\boldsymbol{x}^{(k)}$, we conduct a line search in the direction $-\bigtriangledown f(\boldsymbol{x}^{(k)})$ until a minimizer, $\boldsymbol{x}^{(k+1)}$, is found.



Proposition 8.1

- Proposition 8.1: If $\{\boldsymbol{x}^{(k)}\}_{k=0}^{\infty}$ is a steepest descent sequence for a given function $f: R^n \to R$, then for each k the vector $\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}$ is orthogonal to the vector $\boldsymbol{x}^{(k+2)} - \boldsymbol{x}^{(k+1)}$
- Proof: From the iterative formula of the method of steepest descent it follows that

$$\langle \boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}, \boldsymbol{x}^{(k+2)} - \boldsymbol{x}^{(k+1)} \rangle = \alpha_k \alpha_{k+1} \langle \bigtriangledown f(\boldsymbol{x}^{(k)}), \bigtriangledown f(\boldsymbol{x}^{(k+1)}) \rangle$$

To complete the proof it is enough to show

$$\langle \bigtriangledown f(\boldsymbol{x}^{(k)}), \bigtriangledown f(\boldsymbol{x}^{(k+1)}) \rangle = 0$$

Observe that α_k is a nonnegative scalar that minimizes $\phi_k(\alpha) \triangleq f(\boldsymbol{x}^{(k)} - \alpha \bigtriangledown f(\boldsymbol{x}^{(k)}))$. Hence, using the FONC and the chain rule gives us

$$0 = \phi_k'(\alpha_k) = \frac{d\phi_k}{d\alpha}(\alpha_k)$$

$$= \bigtriangledown f(\boldsymbol{x}^{(k)} - \alpha_k \bigtriangledown f(\boldsymbol{x}^{(k)}))^T(-\bigtriangledown f(\boldsymbol{x}^{(k)})) = -\langle \bigtriangledown f(\boldsymbol{x}^{(k+1)}), f(\boldsymbol{x}^{(k)}) \rangle$$

Proposition 8.2

- Proposition 8.2: If $\{\boldsymbol{x}^{(k)}\}_{k=0}^{\infty}$ is a steepest descent sequence for a given function $f : R^n \to R$ and if $\bigtriangledown f(\boldsymbol{x}^{(k)}) \neq \mathbf{0}$, then $f(\boldsymbol{x}^{(k+1)}) < f(\boldsymbol{x}^{(k)})$
- Proof: Recall that

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \alpha_k \bigtriangledown f(\boldsymbol{x}^{(k)})$$

  where $\alpha_k \geq 0$ is the minimizer of

$$\phi_k(\alpha) = f(\boldsymbol{x}^{(k)} - \alpha \bigtriangledown f(\boldsymbol{x}^{(k)}))$$

  over all $\alpha \geq 0$. Thus, for $\alpha \geq 0$, we have $\phi_k(\alpha_k) \leq \phi_k(\alpha)$
- By the chain rule,

$$\phi_k'(0) = \frac{d\phi_k}{d\alpha}(0) = -(\bigtriangledown f(\boldsymbol{x}^{(k)} - 0 \bigtriangledown f(\boldsymbol{x}^{(k)})))^T(\bigtriangledown f(\boldsymbol{x}^{(k)})) = -\|\bigtriangledown f(\boldsymbol{x}^{(k)})\|^2 < 0$$

  because $\bigtriangledown f(\boldsymbol{x}^{(k)}) \neq \mathbf{0}$ by assumption. Thus, $\phi_k'(0) < 0$ and this implies that there is an $\bar{\alpha} > 0$ such that $\phi_k(0) > \phi_k(\alpha)$ for all $\alpha \in (0, \bar{\alpha}]$ Hence,

$$f(\boldsymbol{x}^{(k+1)}) = \phi_k(\alpha_k) \leq \phi_k(\bar{\alpha}) < \phi_k(0) = f(\boldsymbol{x}^{(k)})$$

Descent Property

- ***Descent property***: $f(\boldsymbol{x}^{(k+1)}) < f(\boldsymbol{x}^{(k)})$ if $\nabla f(\boldsymbol{x}^{(k)}) \neq \mathbf{0}$
- If for some k, we have $\nabla f(\boldsymbol{x}^{(k)}) = \mathbf{0}$, then the point $\boldsymbol{x}^{(k)}$ satisfies the FONC. In this case, $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)}$. We can use the above as the basis for a stopping criterion for the algorithm.
- The condition $\nabla f(\boldsymbol{x}^{(k)}) = \mathbf{0}$, however, is not directly suitable as a practical stopping criterion, because the numerical computation of the gradient will rarely be identically equal to zero.
- A practical criterion is to check if the norm $\|\nabla f(\boldsymbol{x}^{(k)})\|$ is less than a prespecified threshold.
- Alternatively, we may compute $|f(\boldsymbol{x}^{(k+1)}) - f(\boldsymbol{x}^{(k)})|$, and if the difference is less than some threshold, then we stop.

Descent Property

- Another alternative is to compute the norm $\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\|$, and we stop if the norm is less than a prespecified threshold.
- We may check "relative" values of the quantities above

$$\frac{|f(\boldsymbol{x}^{(k+1)}) - f(\boldsymbol{x}^{(k)})|}{|f(\boldsymbol{x}^{(k)})|} < \epsilon \qquad \frac{\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\|}{\|\boldsymbol{x}^{(k)}\|} < \epsilon$$

  The two relative stopping criteria are preferable because they are "scale-independent." Scaling the objective function does not change the satisfaction of the criterion.
- To avoid dividing by very small numbers, modify as

$$\frac{|f(\boldsymbol{x}^{(k+1)}) - f(\boldsymbol{x}^{(k)})|}{\max\{1, |f(\boldsymbol{x}^{(k)})|\}} < \epsilon \qquad \frac{\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\|}{\max\{1, \|\boldsymbol{x}^{(k)}\|\}} < \epsilon$$

Example

- Use the steepest descent method to find the minimizer of

$$f(x_1, x_2, x_3) = (x_1 - 4)^4 + (x_2 - 3)^2 + 4(x_3 + 5)^4$$

  The initial point is $\boldsymbol{x}^{(0)} = [4, 2, -1]^T$

- We find that

$$\triangledown f(\boldsymbol{x}) = [4(x_1 - 4)^3, 2(x_2 - 3), 16(x_3 + 5)^3]^T$$

  Hence, $\triangledown f(\boldsymbol{x}^{(0)}) = [0, -2, 1024]^T$

- To compute $\boldsymbol{x}^{(1)}$, we need

$$\alpha_0 = \arg\min_{\alpha \geq 0} f(\boldsymbol{x}^{(0)} - \alpha \triangledown f(\boldsymbol{x}^{(0)}))$$

$$= \arg\min_{\alpha \geq 0}(0 + (2 + 2\alpha - 3)^2 + 4(-1 - 1024\alpha + 5)^4)$$

$$= \arg\min_{\alpha \geq 0} \phi_0(\alpha)$$

  Using the secant method from Section 7.4, we obtain

$$\alpha_0 = 3.967 \times 10^{-3}$$

Example

- Plot $\phi_0(\alpha)$ versus α
- We compute

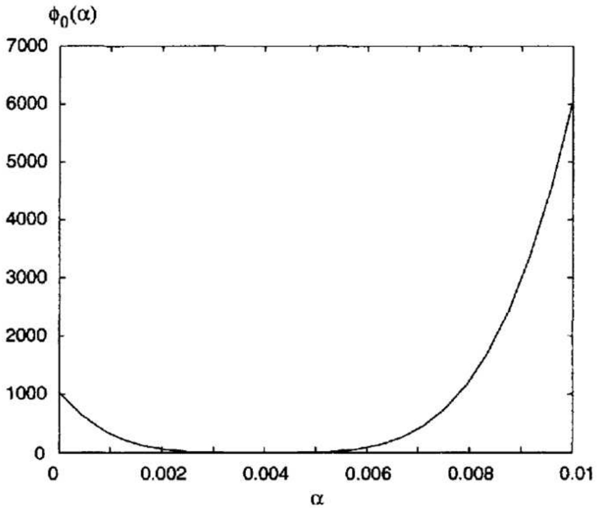
$$\boldsymbol{x}^{(1)} = \boldsymbol{x}^{(0)} - \alpha_0 \bigtriangledown f(\boldsymbol{x}^{(0)}) = [4.000, 2.008, -5.062]^T$$

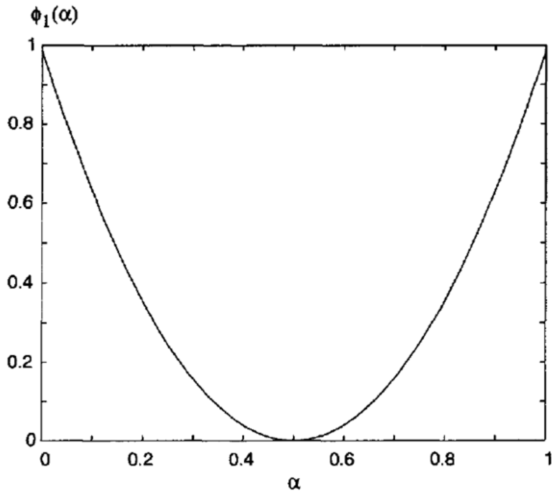
- To find $\boldsymbol{x}^{(2)}$, we first determine $\bigtriangledown f(\boldsymbol{x}^{(1)}) = [0.000, -1.994, -0.003875]^T$
  Next, we find α_1

  $\alpha_1 = \arg\min_{\alpha \geq 0}(0 + (2.008 + 1.984\alpha - 3)^2 + 4(-5.062 + 0.003875\alpha + 5)^4)$

  $= \arg\min_{\alpha \geq 0} \phi_1(\alpha)$

  Using the secant method again, we obtain $\alpha_1 = 0.5000$





Example

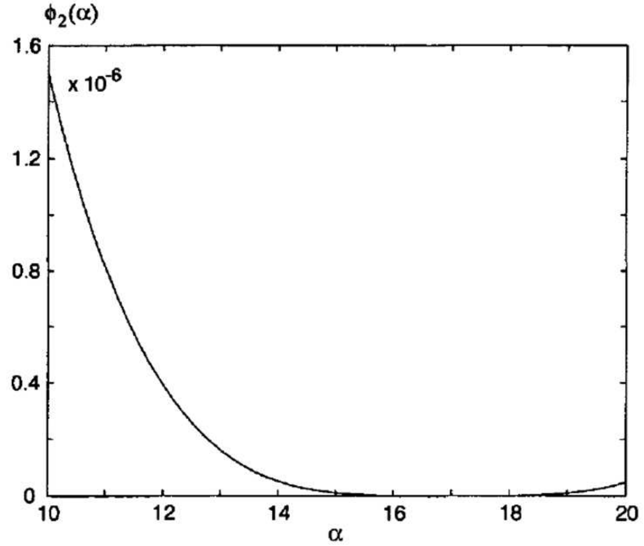
- Thus, $\boldsymbol{x}^{(2)} = \boldsymbol{x}^{(1)} - \alpha_1 \bigtriangledown f(\boldsymbol{x}^{(1)}) = [4.000, 3.000, -5.060]^T$
- To find $\boldsymbol{x}^{(3)}$, we first determine $\bigtriangledown f(\boldsymbol{x}^{(2)}) = [0.000, 0.000, -0.003525]^T$

  $\alpha_2 = \arg\min_{\alpha \geq 0}(0.000 + 0.000 + 4(-5.060 + 0.003525\alpha + 5)^4)$

  $= \arg\min_{\alpha \geq 0} \phi_2(\alpha)$

  $\alpha_1 = 16.29$

- The value $\boldsymbol{x}^{(3)} = [4.000, 3.000, -5.002]^T$
- Note that the minimizer of f is $[4, 3, -5]^T$ and hence it appears that we have arrived at the minimizer in only three iterations.



Steepest Descent for Quadratic Function

- A quadratic function of the form

$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} - \boldsymbol{b}^T\boldsymbol{x}$$

where $\boldsymbol{Q} \in R^{m \times n}$ is a symmetric positive define matrix, $\boldsymbol{b} \in R^n$ and $\boldsymbol{x} \in R^n$. The unique minimizer of f can be found by setting the gradient of f to zero, where

$$\bigtriangledown f(\boldsymbol{x}) = \boldsymbol{Q}\boldsymbol{x} - \boldsymbol{b}$$

because $D(\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x}) = \boldsymbol{x}^T(\boldsymbol{Q} + \boldsymbol{Q}^T) = 2\boldsymbol{x}^T\boldsymbol{Q}$ and $D(\boldsymbol{b}^T\boldsymbol{x}) = \boldsymbol{b}^T$

Steepest Descent for Quadratic Function

- The Hessian of f is $\boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{Q} = \boldsymbol{Q}^T > 0$. To simplify the notation we write $\boldsymbol{g}^{(k)} = \nabla f(\boldsymbol{x}^{(k)})$. Then, for the steepest descent algorithm for the quadratic function can be represented as

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \alpha_k \boldsymbol{g}^{(k)}$$

where

$$\alpha_k = \arg\min_{\alpha \geq 0} f(\boldsymbol{x}^{(k)} - \alpha \boldsymbol{g}^{(k)})$$

$$= \arg\min_{\alpha \geq 0} \left(\frac{1}{2}(\boldsymbol{x}^{(k)} - \alpha \boldsymbol{g}^{(k)})^T \boldsymbol{Q} (\boldsymbol{x}^{(k)} - \alpha \boldsymbol{g}^{(k)}) - (\boldsymbol{x}^{(k)} - \alpha \boldsymbol{g}^{(k)})^T \boldsymbol{b} \right)$$

- In the quadratic case, we can find an explicit formula for α_k. Assume that $\boldsymbol{g}^{(k)} \neq \boldsymbol{0}$, for if $\boldsymbol{g}^{(k)} = \boldsymbol{0}$, then $\boldsymbol{x}^{(k)} = \boldsymbol{x}^*$ and the algorithm stops.

$$\boldsymbol{g}^{(k)} = \nabla f(\boldsymbol{x}^{(k)}) \qquad \nabla f(\boldsymbol{x}) = \boldsymbol{Q}\boldsymbol{x} - \boldsymbol{b}$$

Steepest Descent for Quadratic Function

- Because $\alpha_k \geq 0$ is the minimizer of $\phi_k(\alpha) = f(\boldsymbol{x}^{(k)} - \alpha \boldsymbol{g}^{(k)})$, we apply the FONC to $\phi_k(\alpha)$ to obtain

$$\phi_k'(\alpha) = (\boldsymbol{x}^{(k)} - \alpha \boldsymbol{g}^{(k)})^T \boldsymbol{Q}(-\boldsymbol{g}^{(k)}) - \boldsymbol{b}^T(-\boldsymbol{g}^{(k)})$$

- Therefore, $\phi_k'(\alpha) = 0$ if $\alpha \boldsymbol{g}^{(k)T} \boldsymbol{Q} \boldsymbol{g}^{(k)} = (\boldsymbol{x}^{(k)T}\boldsymbol{Q} - \boldsymbol{b}^T)\boldsymbol{g}^{(k)}$

  But,

$$\boldsymbol{x}^{(k)T}\boldsymbol{Q} - \boldsymbol{b}^T = \boldsymbol{g}^{(k)T}$$

  Hence,

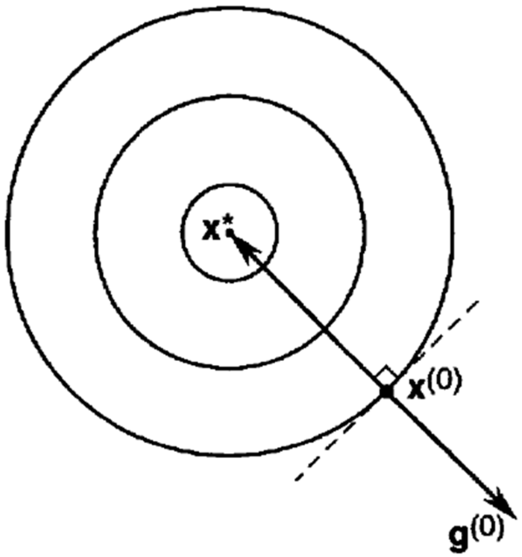
$$\boxed{\alpha_k = \frac{\boldsymbol{g}^{(k)T}\boldsymbol{g}^{(k)}}{\boldsymbol{g}^{(k)T}\boldsymbol{Q}\boldsymbol{g}^{(k)}}}$$

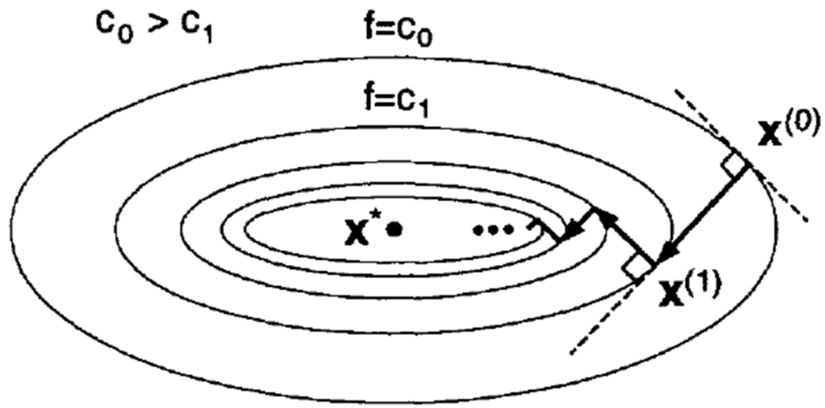
- In summary, the method of steepest descent for the quadratic stakes the form

$$\boxed{\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \frac{\boldsymbol{g}^{(k)T}\boldsymbol{g}^{(k)}}{\boldsymbol{g}^{(k)T}\boldsymbol{Q}\boldsymbol{g}^{(k)}}\boldsymbol{g}^{(k)}} \qquad \boldsymbol{g}^{(k)} = \nabla f(\boldsymbol{x}^{(k)}) = \boldsymbol{Q}\boldsymbol{x}^{(k)} - \boldsymbol{b}$$

Example

- Let $f(x_1, x_2) = x_1^2 + x_2^2$. Then, starting from an arbitrary initial point $\boldsymbol{x}^{(0)} \in R^2$, we arrive at the solution $\boldsymbol{x}^* = \mathbf{0} \in R^2$ at only one step.
- However, if $f(x_1, x_2) = \frac{x_1^2}{5} + x_2^2$, then the method of steepest descent shuffles ineffectively back and forth when searching for the minimizer in a narrow valley. This example illustrates a major drawback in the steepest descent method.





Convergence

- In a ***descent method***, as each new point is generated by the algorithm, the corresponding value of the objective function decreases in value.
- An iterative algorithm is ***globally convergent*** if for any arbitrary starting point the algorithm is guaranteed to generate a sequence of pints converging to a point that satisfies the FONC for a minimizer.
- If not, it may still generate a sequence that converges to a point satisfying the FONC, provided that the initial point is sufficiently close to the point.
  - ***Locally convergent***
- How fast the algorithm converges to a solution point: ***rate of convergence***

$$\boxed{\nabla f(\boldsymbol{x}^*) = \boldsymbol{Q}\boldsymbol{x}^* - \boldsymbol{b} = \boldsymbol{0}}$$

Convergence

- The convergence analysis is more convenient if instead of working with f we deal with

$$V(\boldsymbol{x}) = f(\boldsymbol{x}) + \frac{1}{2}\boldsymbol{x}^{*T}\boldsymbol{Q}\boldsymbol{x}^* = \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}^*)^T\boldsymbol{Q}(\boldsymbol{x} - \boldsymbol{x}^*)$$

where $\boldsymbol{Q} = \boldsymbol{Q}^T > 0$. The solution point $\boldsymbol{x}^*$ is obtained by solving $\boldsymbol{Q}\boldsymbol{x} = \boldsymbol{b}$; that is, $\boldsymbol{x}^* = \boldsymbol{Q}^{-1}\boldsymbol{b}$

- The function V differs from f only by a constant $\frac{1}{2}\boldsymbol{x}^{*T}Q\boldsymbol{x}^*$

$$\boldsymbol{g}^{(k)} = \nabla f(\boldsymbol{x}^{(k)}) \qquad \nabla f(\boldsymbol{x}) = \boldsymbol{Q}\boldsymbol{x} - \boldsymbol{b}$$

Convergence

$$V(\boldsymbol{x}) = f(\boldsymbol{x}) + \frac{1}{2}\boldsymbol{x}^{*T}\boldsymbol{Q}\boldsymbol{x}^* = \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}^*)^T\boldsymbol{Q}(\boldsymbol{x} - \boldsymbol{x}^*)$$

- Lemma 8.1: The iterative algorithm

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \alpha_k \boldsymbol{g}^{(k)}$$

with $\boldsymbol{g}^{(k)} = \boldsymbol{Q}\boldsymbol{x}^{(k)} - \boldsymbol{b}$ satisfies

$$V(\boldsymbol{x}^{(k+1)}) = (1 - \gamma_k)V(\boldsymbol{x}^{(k)})$$

where if $\boldsymbol{g}^{(k)} = \mathbf{0}$, then $\gamma_k = 1$, and if $\boldsymbol{g}^{(k)} \neq \mathbf{0}$, then

$$\gamma_k = \alpha_k \frac{\boldsymbol{g}^{(k)T}\boldsymbol{Q}\boldsymbol{g}^{(k)}}{\boldsymbol{g}^{(k)T}\boldsymbol{Q}^{-1}\boldsymbol{g}^{(k)}}\left(2\frac{\boldsymbol{g}^{(k)T}\boldsymbol{g}^{(k)}}{\boldsymbol{g}^{(k)T}\boldsymbol{Q}\boldsymbol{g}^{(k)}} - \alpha_k\right)$$

Convergence

- Theorem 8.1: Let $\{\boldsymbol{x}^{(k)}\}$ be the sequence resulting from a gradient algorithm $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \alpha_k \boldsymbol{g}^{(k)}$. Let γ_k be as defined in Lemma 8.1, and suppose that $\gamma_k > 0$ for all k. Then, $\{\boldsymbol{x}^{(k)}\}$ converges to $\boldsymbol{x}^*$ for any initial condition $\boldsymbol{x}^{(0)}$ if and only if

$$\sum_{k=0}^{\infty} \gamma_k = \infty$$

- Proof:
- From Lemma 8.1 we have $V(\boldsymbol{x}^{(k+1)}) = (1-\gamma_k)V(\boldsymbol{x}^{(k)})$, from which we obtain

$$V(\boldsymbol{x}^{(k)}) = \left(\prod_{i=0}^{k-1}(1-\gamma_i)\right)V(\boldsymbol{x}^{(0)})$$

- Assume that $\gamma_k < 1$ for all k, for otherwise the result holds trivially.

$$V(\boldsymbol{x}^{(k)}) = \left(\prod_{i=0}^{k-1}(1-\gamma_i)\right)V(\boldsymbol{x}^{(0)})$$

$$V(\boldsymbol{x}) = f(\boldsymbol{x}) + \frac{1}{2}\boldsymbol{x}^{*T}\boldsymbol{Q}\boldsymbol{x}^{*} = \frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}^{*})^{T}\boldsymbol{Q}(\boldsymbol{x}-\boldsymbol{x}^{*})$$

Convergence

- Note that $\boldsymbol{x}^{(k)} \to \boldsymbol{x}^{*}$ if and only if $V(\boldsymbol{x}^{(k)}) \to 0$. We see that this occurs if and only if $\prod_{i=0}^{\infty}(1-\gamma_i) = 0$, which, in turn, holds if and only if $\prod_{i=0}^{\infty} -\log(1-\gamma_i) = \infty$
- Note that by Lemma 8.1, $1-\gamma_i \geq 0$ and $\log(1-\gamma_i)$ is well defined [$\log(0)$ is taken to be $-\infty$]. Therefore, it remains to show that $\prod_{i=0}^{\infty} -\log(1-\gamma_i) = \infty$ if and only if

$$\sum_{i=0}^{\infty}\gamma_i = \infty$$

- We first show that $\sum_{i=0}^{\infty}\gamma_i = \infty$ implies that $\sum_{i=0}^{\infty} -\log(1-\gamma_i) = \infty$. For this, first observe that for any $x \in R, x > 0$, we have $\log(x) \leq x-1$. Therefore, $\log(1-\gamma_i) \leq 1-\gamma_i-1 = -\gamma_i$, and hence $-\log(1-\gamma_i) \geq \gamma_i$. Thus, if $\sum_{i=0}^{\infty}\gamma_i = \infty$, then clearly $\sum_{i=0}^{\infty} -\log(1-\gamma_i) = \infty$

Convergence

- Finally, we show that $\sum_{i=0}^{\infty} -\log(1-\gamma_i) = \infty$ implies that $\sum_{i=0}^{\infty} \gamma_i = \infty$
- By contraposition. Suppose that $\sum_{i=0}^{\infty} \gamma_i < \infty$. Then, it must be that $\gamma_i \to 0$. Observe that for $x \in R, x \leq 1$ and x sufficiently close to 1, we have $\log(x) \geq 2(x-1)$. Therefore, for sufficiently large i, $\log(1-\gamma_i) \geq 2(1-\gamma_i-1) = -2\gamma_i$, which implies that $-\log(1-\gamma_i) \leq 2\gamma_i$. Hence, $\sum_{i=0}^{\infty} \gamma_i < \infty$ implies that $\sum_{i=0}^{\infty} -\log(1-\gamma_i) < \infty$. This completes the proof.

- The assumption in Theorem 8.1 that $\gamma_k > 0$ for all k is significant. Furthermore, the result of the theorem does not hold in general if we do not have this assumption.

$$\boxed{V(\boldsymbol{x}^{(k+1)}) = (1-\gamma_k)V(\boldsymbol{x}^{(k)})}$$

Example

- A counter example to show $\gamma_k > 0$ in Theorem 8.1 is necessary.
- For each $k = 0, 1, 2, \ldots$, choose α_k in such a way that $\gamma_{2k} = -1/2$ and $\gamma_{2k+1} = 1/2$ (we can always do this if, for example, $\boldsymbol{Q} = \boldsymbol{I}_n$). From Lemma 8.1 we have

  $$V(\boldsymbol{x}^{2(k+1)}) = (1-1/2)(1+1/2)V(\boldsymbol{x}^{(2k)}) = (3/4)V(\boldsymbol{x}^{(2k)})$$

  Therefore, $V(\boldsymbol{x}^{(2k)}) \to 0$. Because $V(\boldsymbol{x}^{(2k+1)}) = (3/2)V(\boldsymbol{x}^{(2k)})$, we also have that $V(\boldsymbol{x}^{(2k+1)}) \to 0$.

  Hence, $V(\boldsymbol{x}^{(k)}) \to 0$, which implies that $\boldsymbol{x}^{(k)} \to 0$ (for all $\boldsymbol{x}^{(0)}$). On the other hand, it is clear that

  $$\sum_{i=0}^{k} \gamma_i \leq \frac{1}{2}$$

  for all k. Hence, the result of the theorem does not hold if $\gamma_k \leq 0$ for some k.

Convergence

- Rayleigh's inequality. For any $\boldsymbol{Q} = \boldsymbol{Q}^T > 0$, we have

$$\lambda_{min}(\boldsymbol{Q})\|\boldsymbol{x}\|^2 \leq \boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} \leq \lambda_{max}(\boldsymbol{Q})\|\boldsymbol{x}\|^2$$

We also have

$$\lambda_{min}(\boldsymbol{Q}^{-1}) = \frac{1}{\lambda_{max}(\boldsymbol{Q})}$$

$$\lambda_{max}(\boldsymbol{Q}^{-1}) = \frac{1}{\lambda_{min}(\boldsymbol{Q})}$$

$$\lambda_{min}(\boldsymbol{Q}^{-1})\|\boldsymbol{x}\|^2 \leq \boldsymbol{x}^T\boldsymbol{Q}^{-1}\boldsymbol{x} \leq \lambda_{max}(\boldsymbol{Q}^{-1})\|\boldsymbol{x}\|^2$$

Convergence

- Lemma 8.2: Let $\boldsymbol{Q} = \boldsymbol{Q}^T > 0$ be an $n \times n$ real symmetric positive definite matrix. Then, for any $\boldsymbol{x} \in R^n$, we have

$$\frac{\lambda_{min}(\boldsymbol{Q})}{\lambda_{max}(\boldsymbol{Q})} \le \frac{(\boldsymbol{x}^T\boldsymbol{x})^2}{(\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x})(\boldsymbol{x}^T\boldsymbol{Q}^{-1}\boldsymbol{x})} \le \frac{\lambda_{max}(\boldsymbol{Q})}{\lambda_{min}(\boldsymbol{Q})}$$

- Proof: Appling Rayleigh's inequality and using the properties of symmetric positive definite matrices listed previously, we get

$$\frac{(\boldsymbol{x}^T\boldsymbol{x})^2}{(\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x})(\boldsymbol{x}^T\boldsymbol{Q}^{-1}\boldsymbol{x})} \le \frac{\|\boldsymbol{x}\|^4}{\lambda_{min}(\boldsymbol{Q})\|\boldsymbol{x}\|^2\lambda_{min}(\boldsymbol{Q}^{-1})\|\boldsymbol{x}\|^2} = \frac{\lambda_{max}(\boldsymbol{Q})}{\lambda_{min}(\boldsymbol{Q})}$$

$$\frac{(\boldsymbol{x}^T\boldsymbol{x})^2}{(\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x})(\boldsymbol{x}^T\boldsymbol{Q}^{-1}\boldsymbol{x})} \ge \frac{\|\boldsymbol{x}\|^4}{\lambda_{max}(\boldsymbol{Q})\|\boldsymbol{x}\|^2\lambda_{max}(\boldsymbol{Q}^{-1})\|\boldsymbol{x}\|^2} = \frac{\lambda_{min}(\boldsymbol{Q})}{\lambda_{max}(\boldsymbol{Q})}$$

$$\boxed{\boldsymbol{g}^{(k)} = \nabla f(\boldsymbol{x}^{(k)}) = \boldsymbol{Q}\boldsymbol{x}^{(k)} - \boldsymbol{b}}$$

Convergence

- Theorem 8.2: In the steepest descent algorithm, we have $\boldsymbol{x}^{(k)} \to \boldsymbol{x}^*$ for any $\boldsymbol{x}^{(0)}$
- Proof: If $\boldsymbol{g}^{(k)} = \boldsymbol{0}$ for some k, then $\boldsymbol{x}^{(k)} = \boldsymbol{x}^*$ and the result holds. So assume that $\boldsymbol{g}^{(k)} \neq \boldsymbol{0}$ for all k. Recall that for the steepest descent algorithm,

$$\alpha_k = \frac{\boldsymbol{g}^{(k)T}\boldsymbol{g}^{(k)}}{\boldsymbol{g}^{(k)T}\boldsymbol{Q}\boldsymbol{g}^{(k)}}$$

Substituting this expression for α_k in the formula for γ_k yields

$$\gamma_k = \frac{(\boldsymbol{g}^{(k)T}\boldsymbol{g}^{(k)})^2}{(\boldsymbol{g}^{(k)T}\boldsymbol{Q}\boldsymbol{g}^{(k)})(\boldsymbol{g}^{(k)T}\boldsymbol{Q}^{-1}\boldsymbol{g}^{(k)})}$$

Note that in this case $\gamma_k > 0$ for all k. Furthermore, by Lemma 8.2, we have $\gamma_k \geq (\lambda_{min}(\boldsymbol{Q})/\lambda_{max}(\boldsymbol{Q})) > 0$. Therefore, we have $\sum_{k=0}^{\infty} \gamma_k = \infty$, and hence by Theorem 8.1, we conclude that $\boldsymbol{x}^{(k)} \to \boldsymbol{x}^*$

Convergence

- Consider now a gradient method with fixed step size; that is, $\alpha_k = \alpha \in R$ for all k. The resulting algorithm is of the form $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \alpha g^{(k)}$
- We refer to the algorithm above as a ***fixed-step-size*** gradient algorithm. The algorithm is of practical interest because of its simplicity.
- The algorithm does not require a line search at each step to determine α_k. Clearly, the convergence of the algorithm depends on the choice of α.

Convergence

- Theorem 8.3: For the fixed-step-size gradient algorithm, $\boldsymbol{x}^{(k)} \to \boldsymbol{x}^*$ for any $\boldsymbol{x}^{(0)}$ if and only if

$$0 < \alpha < \frac{2}{\lambda_{max}(\boldsymbol{Q})}$$

- Proof: $\Longleftarrow$: By Rayleigh's inequality we have

$$\lambda_{min}(\boldsymbol{Q})\boldsymbol{g}^{(k)T}\boldsymbol{g}^{(k)} \leq \boldsymbol{g}^{(k)T}\boldsymbol{Q}\boldsymbol{g}^{(k)} \leq \lambda_{max}(\boldsymbol{Q})\boldsymbol{g}^{(k)T}\boldsymbol{g}^{(k)}$$

and

$$\boldsymbol{g}^{(k)T}\boldsymbol{Q}^{-1}\boldsymbol{g}^{(k)} \leq \frac{1}{\lambda_{max}(\boldsymbol{Q})}\boldsymbol{g}^{(k)T}\boldsymbol{g}^{(k)}$$

- Therefore, substituting the above in the formula for γ_k, we have

$$\gamma_k \geq \alpha(\lambda_{min}(\boldsymbol{Q}))^2\left(\frac{2}{\lambda_{max}(\boldsymbol{Q})} - \alpha\right) > 0$$

Therefore, $\gamma_k > 0$ for all k, and $\sum_{k=0}^{\infty} \gamma_k = \infty$. Hence, by Theorem 8.1, we conclude that $\boldsymbol{x}^{(k)} \to \boldsymbol{x}^*$

Convergence

- Proof: $\Longrightarrow$: We use contraposition. Suppose that either $\alpha \leq 0$ or $\alpha \geq 2/\lambda_{max}(\boldsymbol{Q})$. Let $\boldsymbol{x}^{(0)}$ be chosen such that $\boldsymbol{x}^0 - \boldsymbol{x}^*$ is an eigenvector of $\boldsymbol{Q}$ corresponding to the eigenvalue $\lambda_{max}(\boldsymbol{Q})$. Because

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \alpha(\boldsymbol{Q}\boldsymbol{x}^{(k)} - \boldsymbol{b}) = \boldsymbol{x}^{(k)} - \alpha(\boldsymbol{Q}\boldsymbol{x}^{(k)} - \boldsymbol{Q}\boldsymbol{x}^*)$$

we obtain

$$\begin{aligned}\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^* &= \boldsymbol{x}^{(k)} - \boldsymbol{x}^* - \alpha(\boldsymbol{Q}\boldsymbol{x}^{(k)} - \boldsymbol{Q}\boldsymbol{x}^*)\\ &= (\boldsymbol{I}_n - \alpha\boldsymbol{Q})(\boldsymbol{x}^{(k)} - \boldsymbol{x}^*)\\ &= (\boldsymbol{I}_n - \alpha\boldsymbol{Q})^{k+1}(\boldsymbol{x}^{(0)} - \boldsymbol{x}^*)\\ &= (1 - \alpha\lambda_{max}(\boldsymbol{Q}))^{k+1}(\boldsymbol{x}^{(0)} - \boldsymbol{x}^*)\end{aligned}$$

Taking norms on both sides, we get

$$\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^*\| = |1 - \alpha\lambda_{max}(\boldsymbol{Q})|^{k+1}\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\|$$

Because $\alpha \leq 0$ or $\alpha \geq 2/\lambda_{max}(\boldsymbol{Q})$, $|1 - \alpha\lambda_{max}(\boldsymbol{Q})| \geq 1$

Hence, $\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^*|$ cannot converge to 0, and thus the sequence $\{\boldsymbol{x}^{(k)}\}$ does not converge to $\boldsymbol{x}^*$

Example

- Let the function f be given by

$$f(\boldsymbol{x}) = \boldsymbol{x}^T \begin{bmatrix} 4 & 2\sqrt{2} \\ 0 & 5 \end{bmatrix} \boldsymbol{x} + \boldsymbol{x}^T \begin{bmatrix} 3 \\ 6 \end{bmatrix} + 24$$

  We wish to find the minimizer of f using a fixed-step-size gradient algorithm $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \alpha \bigtriangledown f(\boldsymbol{x}^{(k)})$
  where $\alpha \in R$ is a fixed step size.
- Solution: To apply Theorem 8.3, we first symmetrize the matrix in the quadratic term of *f* to get

$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T \begin{bmatrix} 8 & 2\sqrt{2} \\ 2\sqrt{2} & 10 \end{bmatrix} \boldsymbol{x} + \boldsymbol{x}^T \begin{bmatrix} 3 \\ 6 \end{bmatrix} + 24$$

  The eigenvalues of the matrix are 6 and 12. Hence, by Theorem 8.3, the algorithm converges to the minimizer for all $\boldsymbol{x}^{(0)}$ if and only if α lies in the range $0 < \alpha < 2/12$

Convergence Rate

- Theorem 8.4: In the method of steepest descent applied to the quadratic function, at every step we have

$$V(\boldsymbol{x}^{(k+1)}) \leq \frac{\lambda_{max}(\boldsymbol{Q}) - \lambda_{min}(\boldsymbol{Q})}{\lambda_{max}(\boldsymbol{Q})} V(\boldsymbol{x}^{(k)})$$

- Proof: In the proof of Theorem 8.2, we showed that $\gamma_k \geq \lambda_{min}(\boldsymbol{Q})/\lambda_{max}(\boldsymbol{Q})$. Therefore,

$$\frac{V(\boldsymbol{x}^{(k)}) - V(\boldsymbol{x}^{(k+1)})}{V(\boldsymbol{x}^{(k)})} = \gamma_k \geq \frac{\lambda_{min}(\boldsymbol{Q})}{\lambda_{max}(\boldsymbol{Q})}$$

and the result follows.

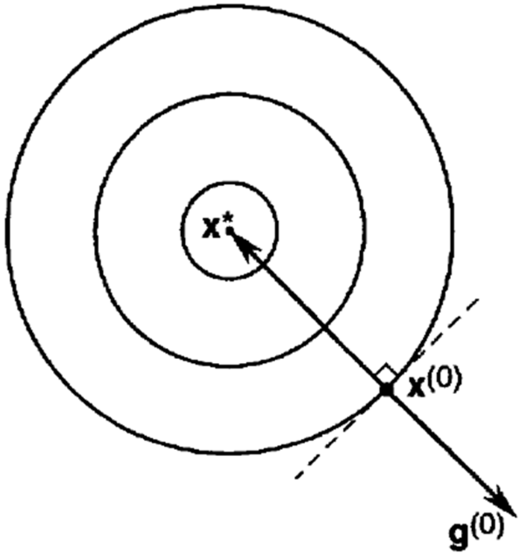
$$\boxed{V(\boldsymbol{x}^{(k+1)}) = (1 - \gamma_k) V(\boldsymbol{x}^{(k)})}$$

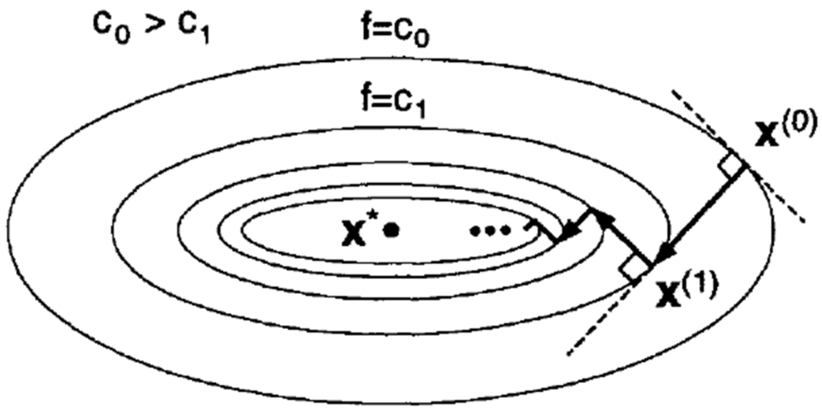
Convergence Rate

- Let $$r = \frac{\lambda_{max}(\boldsymbol{Q})}{\lambda_{min}(\boldsymbol{Q})} = \|\boldsymbol{Q}\|\|\boldsymbol{Q}^{-1}\|$$
  called the ***condition number*** of $\boldsymbol{Q}$. Then, it follows from Theorem 8.4 that
  $$V(\boldsymbol{x}^{(k+1)}) \leq (1 - \tfrac{1}{r})V(\boldsymbol{x}^{(k)})$$
- The term $(1 - 1/r)$ plays an important role in the convergence of $\{V(\boldsymbol{x}^{(k)})\}$ to 0 (and hence of $\{\boldsymbol{x}^{(k)}\}$ to $\boldsymbol{x}^*$). We refer to $(1 - 1/r)$ as the ***convergence ratio***.
- The smaller the value of $(1 - 1/r)$, the smaller $V(\boldsymbol{x}^{(k+1)})$ will be relative to $V(\boldsymbol{x}^{(k)})$, and hence the "faster" $V(\boldsymbol{x}^{(k)})$ converges to 0.

Convergence Rate

- The convergence ratio $(1 - 1/r)$ decreases as r decreases. If $r = 1$ then $\lambda_{max}(\boldsymbol{Q}) = \lambda_{min}(\boldsymbol{Q})$, corresponding to the circular contours of f (Figure 8.6). In this case the algorithm converges in a single step to the minimizer.
- As r increases, the speed of convergence of $\{V(\boldsymbol{x}^{(k)})\}$ (and hence $\{\boldsymbol{x}^{(k)}\}$) decreases. The increase in r reflects that fact that the contours of f are more eccentric.





Convergence Rate

- Definition 8.1: Given a sequence $\{\boldsymbol{x}^{(k)}\}$ that converges to $\boldsymbol{x}^*$, that is, $\lim_{k\to\infty}\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\| = 0$, we say the ***order of convergence*** is p, where $p \in R$, if

$$0 < \lim_{k\to\infty} \frac{\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^*\|}{\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\|^p} < \infty$$

  If for all $p > 0$

$$\lim_{k\to\infty} \frac{\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^*\|}{\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\|^p} = 0$$

  then we say that the order of convergence is ∞

- Note that in the definition above, 0/0 should be understood to be 0.

Convergence Rate

- The order of convergence of a sequence is a measure of its rate of convergence; ***the higher the order, the faster the rate of convergence***.
- The order of convergence is sometimes also called the ***rate of convergence***. If $p = 1$ and $\lim_{k\to\infty} \|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^*\| / \|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\| = 1$ we say that the convergence is ***sublinear***.
- If $p = 1$ and $\lim_{k\to\infty} \|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^*\| / \|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\| < 1$, we say that the convergence is ***linear***.
- If $p > 1$, we say that the convergence is ***superlinear***.
- If $p = 2$, we say that the convergence is ***quadratic***.

Example

- Suppose that $x^{(k)} = 1/k$ and thus $x^{(k)} \to 0$. Then,

$$\frac{|x^{(k+1)}|}{|x^{(k)}|^p} = \frac{1/(k+1)}{1/k^p} = \frac{k^p}{k+1}$$

If $p < 1$, the sequence converges to 0, whereas if $p > 1$, it grows to ∞. If $p = 1$, the sequence converges to 1. Hence, the order of convergence is 1.

- Suppose that $x^{(k)} = \gamma^k$, where $0 < \gamma < 1$, and thus $x^{(k)} \to 0$. Then,

$$\frac{|x^{(k+1)}|}{|x^{(k)}|^p} = \frac{\gamma^{k+1}}{(\gamma^k)^p} = \gamma^{k+1-kp} = \gamma^{k(1-p)+1}$$

If $p < 1$, the sequence converges to 0, whereas if $p > 1$, it grows to ∞. If $p = 1$, the sequence converges to γ. Hence, the order of convergence is 1.

Example

- Suppose that $x^{(k)} = \gamma^{q^k}$, where $q > 1$ and $0 < \gamma < 1$, and thus $x^{(k)} \to 0$
  Then,
  $$\frac{|x^{(k+1)}|}{|x^{(k)}|^p} = \frac{\gamma^{(q^{k+1})}}{(\gamma^{(q^k)})^p} = \gamma^{q^{k+1}-pq^k} = \gamma^{(q-p)q^k}$$

  If $p < q$, the sequence converges to 0, whereas if $p > q$, it grows to ∞ . If $p = q$, the sequence converges to 1. Hence, the order of convergence is q .
- Suppose that $x^{(k)} = 1$ for all k, and thus $x^{(k)} \to 1$. Then,
  $$\frac{|x^{(k+1)} - 1|}{|x^{(k)} - 1|^p} = \frac{0}{0^p} = 0$$

  for all p . Hence, the order of convergence is ∞.

Convergence Rate

- The order of convergence can be interpreted using the notion of the order symbol O. Recall that $a = O(h)$ ("big-oh" of h) if there exists a constant c such that $|a| \leq c|h|$ for sufficiently small h.
- The order of convergence is *at least* p if

$$\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^*\| = O(\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\|^p)$$

Convergence Rate

- Theorem 8.5: Let $\{\boldsymbol{x}^{(k)}\}$ be a sequence that converges to $\boldsymbol{x}^*$. If
$$\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^*\| = O(\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\|^p)$$
then the order of convergence (if it exists) is at least p.
- Proof: Let s be the order of convergence of $\{\boldsymbol{x}^{(k)}\}$. Suppose that
$$\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^*\| = O(\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\|^p)$$
Then, these exists c such that for sufficiently large k,
$$\frac{\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^*\|}{\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\|^p} \leq c$$
Hence,
$$\begin{aligned} &\frac{\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^*\|}{\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\|^s} \\ &= \frac{\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^*\|}{\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\|^p}\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\|^{p-s} \\ &\leq c\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\|^{p-s} \end{aligned}$$

Convergence Rate

- Taking limits yields

$$\lim_{k\to\infty}\frac{\|\boldsymbol{x}^{(k+1)}-\boldsymbol{x}^*\|}{\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^*\|^s}\leq c\lim_{k\to\infty}\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^*\|^{p-s}$$

- Because by definition s is the order of convergence

$$\lim_{k\to\infty}\frac{\|\boldsymbol{x}^{(k+1)}-\boldsymbol{x}^*\|}{\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^*\|^s}>0$$

Combining the two inequalities above, we get

$$c\lim_{k\to\infty}\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^*\|^{p-s}>0$$

Therefore, because $\lim_{k\to\infty}\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^*\|=0$, we conclude that $s\geq p$ that is, the order of convergence is at least p.

Example

- Similarly, we can show that if $\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^*\| = o(\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\|^p)$ then the order of convergence (if it exists) strictly exceeds p.
- Suppose that we are given a scalar sequence $\{x^{(k)}\}$ that converges with order of convergence p and satisfies

$$\lim_{k\to\infty} \frac{|x^{(k+1)} - 2|}{|x^{(k)} - 2|^3} = 0$$

The limit of $\{x^{(k)}\}$ must be 2, because it is clear from the equation that $|x^{(k+1)} - 2| \to 0$. Also, we see that $|x^{(k+1)} - 2| = o(|x^{(k)} - 2|^3)$. Hence, we conclude that $p > 3$

Example

- Consider the problem of finding a minimizer of the function $f : R \to R$ given by $f(x) = x^2 - \frac{x^3}{3}$. Suppose that we use the algorithm $x^{(k+1)} = x^{(k)} - \alpha f'(x^{(k)})$ with step size $\alpha = 1/2$ and initial condition $x^{(0)} = 1$
- We first show that the algorithm converges to a local minimizer of f. We have $f'(x) = 2x - x^2$. The fixed-step-size gradient algorithm is therefore given by

$$x^{(k+1)} = x^{(k)} - \alpha f'(x^{(k)}) = \tfrac{1}{2}(x^{(k)})^2$$

With $x^{(0)} = 1$, we can derive the expression $x^{(k+1)} = (1/2)^{2^k - 1}$
Hence, the algorithm converges to 0, a strict local minimizer of f . Note that

$$\frac{|x^{(k+1)}|}{|x^{(k)}|^2} = \frac{1}{2}$$

Therefore, the order of convergence is 2.

$\boldsymbol{g}^{(k)} = \nabla f(\boldsymbol{x}^{(k)}) = \boldsymbol{Q}\boldsymbol{x}^{(k)} - \boldsymbol{b}$

Convergence Rate

- The steepest descent algorithm has an order of convergence of 1 in the *worst case*.
- Lemma 8.3: In the steepest descent algorithm, if $\boldsymbol{g}^{(k)} \neq \mathbf{0}$ for all k then $\gamma_k = 1$ if and only if $\boldsymbol{g}^{(k)}$ is an eigenvector of $\boldsymbol{Q}$.
- Proof: Suppose that $\boldsymbol{g}^{(k)} \neq \mathbf{0}$ for all k. Recall that for the steepest descent algorithm,

$$\gamma_k = \frac{(\boldsymbol{g}^{(k)T}\boldsymbol{g}^{(k)})^2}{(\boldsymbol{g}^{(k)T}\boldsymbol{Q}\boldsymbol{g}^{(k)})(\boldsymbol{g}^{(k)T}\boldsymbol{Q}^{-1}\boldsymbol{g}^{(k)})}$$

Sufficiency is easy to show by verification. To show necessity, suppose that $\gamma_k = 1$. Then, $V(\boldsymbol{x}^{(k+1)}) = 0$, which implies that $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^*$. Therefore, $\boldsymbol{x}^* = \boldsymbol{x}^{(k)} - \alpha_k \boldsymbol{g}^{(k)}$.

Convergence Rate

$$\boldsymbol{x}^* = \boldsymbol{x}^{(k)} - \alpha_k \boldsymbol{g}^{(k)}$$

- Premultiplying by $\boldsymbol{Q}$ and subtracting $\boldsymbol{b}$ from both sides yields

$$\boldsymbol{0} = \boldsymbol{g}^{(k)} - \alpha_k \boldsymbol{Q}\boldsymbol{g}^{(k)}$$

which can be rewritten as

$$\boldsymbol{Q}\boldsymbol{g}^{(k)} = \frac{1}{\alpha_k}\boldsymbol{g}^{(k)}$$

Hence, $\boldsymbol{g}^{(k)}$ is an eigenvector of $\boldsymbol{Q}$.

- By the lemma, if $\boldsymbol{g}^{(k)}$ is not an eigenvector of $\boldsymbol{Q}$, then $\gamma_k < 1$ (recall that γ_k cannot exceed 1)

Theorem 8.6

- Theorem 8.6: Let $\{\boldsymbol{x}^{(k)}\}$ be a convergent sequence of iterates of the steepest descent algorithm applied to a function f. Then, the order of convergence of $\{\boldsymbol{x}^{(k)}\}$ is 1 in the worst case; that is, there exist a function f and an initial condition $\boldsymbol{x}^{(0)}$ such that the order of convergence of $\{\boldsymbol{x}^{(k)}\}$ is equal to 1.
- Proof: Let $f : R^n \to R$ be a quadratic function with Hessian $\boldsymbol{Q}$. Assume that the maximum and minimum eigenvalues of $\boldsymbol{Q}$ satisfy $\lambda_{max}(\boldsymbol{Q}) > \lambda_{min}(\boldsymbol{Q})$. To show that the order of convergence is 1, it suffices to show that there exists $\boldsymbol{x}^{(0)}$ such that

$$\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^*\| \geq c\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*|$$

for some c.

Theorem 8.6

- By Rayleigh's inequality

$$V(\boldsymbol{x}^{(k+1)}) = \frac{1}{2}(\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^*)^T \boldsymbol{Q}(\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^*) \leq \frac{\lambda_{max}(\boldsymbol{Q})}{2}\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^*\|^2$$

Similarly,
$$V(\boldsymbol{x}^{(k)}) \geq \frac{\lambda_{min}(\boldsymbol{Q})}{2}\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\|^2$$

Combining the inequalities above with Lemma 8.1, we obtain

$$\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^*\| \geq \sqrt{(1-\gamma_k)\frac{\lambda_{min}(\boldsymbol{Q})}{\lambda_{max}(\boldsymbol{Q})}}\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*|$$

Therefore, it suffices to choose $\boldsymbol{x}^{(0)}$ such that $\gamma_k \leq d$ for some $d < 1$

Theorem 8.6

- Recall that for the steepest descent algorithm, assuming that $\boldsymbol{g}^{(k)} \neq \boldsymbol{0}$ for all k,
$$\gamma_k = \frac{(\boldsymbol{g}^{(k)T}\boldsymbol{g}^{(k)})^2}{(\boldsymbol{g}^{(k)T}\boldsymbol{Q}\boldsymbol{g}^{(k)})(\boldsymbol{g}^{(k)T}\boldsymbol{Q}^{-1}\boldsymbol{g}^{(k)})}$$
- First consider the case where $n = 2$. Suppose that $\boldsymbol{x}^{(0)} \neq \boldsymbol{x}^*$ is chosen such that $\boldsymbol{x}^{(0)} - \boldsymbol{x}^*$ is not an eigenvector of $\boldsymbol{Q}$. Then, $\boldsymbol{g}^{(0)} = \boldsymbol{Q}(\boldsymbol{x}^{(0)} - \boldsymbol{x}^*) \neq \boldsymbol{0}$ is also not an eigenvector of $\boldsymbol{Q}$.
- By Proposition 8.1, $\boldsymbol{g}^{(k)} = (\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)})/\alpha_k$ is not an eigenvector of $\boldsymbol{Q}$ for any k [because any two eigenvectors corresponding to $\lambda_{max}(\boldsymbol{Q})$ and $\lambda_{min}(\boldsymbol{Q})$ are mutually orthogonal].
- Also, $\boldsymbol{g}^{(k)}$ lies in one of two mutually orthogonal directions. Therefore, by Lemma 8.3, for each k, the value of γ_k of two numbers, both of which are strictly less than 1. This proves the $n = 2$ case.

Theorem 8.6

- For the general n case, let $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ be mutually orthogonal eigenvectors corresponding to $\lambda_{max}(\boldsymbol{Q})$ and $\lambda_{min}(\boldsymbol{Q})$. Choose $\boldsymbol{x}^{(0)}$ such that $\boldsymbol{x}^{(0)} - \boldsymbol{x}^* \neq \mathbf{0}$ lies in the span of $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ but is not equal to either.
- Note that $\boldsymbol{g}^{(0)} = \boldsymbol{Q}(\boldsymbol{x}^{(0)} - \boldsymbol{x}^*)$ also lies in the span of $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$, but is not equal to either.
- By manipulating $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \alpha_k \boldsymbol{g}^{(k)}$ as before, we can write $\boldsymbol{g}^{(k+1)} = (\boldsymbol{I} - \alpha_k \boldsymbol{Q})\boldsymbol{g}^{(k)}$. Any eigenvector of $\boldsymbol{Q}$ is also an eigenvector of $\boldsymbol{I} - \alpha_k \boldsymbol{Q}$. Therefore, $\boldsymbol{g}^{(k)}$ lies in the span of $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ for all k; that is, the sequence $\{\boldsymbol{g}^{(k)}\}$ is confined within the two-dimensional subspace spanned by $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$. We can now proceed as in the $n = 2$ case.