Recall that a level set of a function $f : \mathbb{R}^n \to \mathbb{R}$ is the set of points $x$ satisfying $f(x) = c$ for some constant $c$. Thus, a point $x_0 \in \mathbb{R}^n$ is on the level set corresponding to level $c$ if $f(x_0) = c$.

In the case of functions of two real variables, $f : \mathbb{R}^2 \to \mathbb{R}$.
Introduction

- The gradient of $f$ at $x_0$, denoted by $\nabla f(x_0)$, is orthogonal to the tangent vector to an arbitrary smooth curve passing through $x_0$ on the level set $f(x) = c$.
- 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.
Recall that \( \langle \nabla f(x), d \rangle, \|d\| = 1 \), is the rate of increase of \( f \) in the direction \( d \) at the point \( x \). By the Cauchy-Schwarz inequality,
\[
\langle \nabla f(x), d \rangle \leq \|\nabla f(x)\|
\]
because \( \|d\| = 1 \). But if \( d = \nabla f(x)/\|\nabla f(x)\| \), then
\[
\langle \nabla f(x), \frac{\nabla f(x)}{\|\nabla f(x)\|} \rangle = \|\nabla f(x)\|
\]
Thus, the direction in which \( \nabla f(x) \) points is the direction of maximum rate of increase of \( f \) at \( x \).

The direction in which \( -\nabla f(x) \) points is the direction of maximum rate of decrease of \( f \) at \( x \).

Hence, the direction of negative gradient is a good direction to search if we want to find a function minimizer.
Introduction

- Let $x^{(0)}$ be a starting point, and consider the point $x^{(0)} - \alpha \nabla f(x^{(0)})$
  Then, by Taylor’s theorem, we obtain
  \[ f(x^{(0)} - \alpha \nabla f(x^{(0)})) = f(x^{(0)}) - \alpha \|\nabla f(x^{(0)})\|^2 + o(\alpha) \]
- If $\nabla f(x^{(0)}) \neq 0$, then for sufficiently small $\alpha > 0$, we have
  \[ f(x^{(0)} - \alpha \nabla f(x^{(0)})) < f(x^{(0)}) \]
- This means that the point $x^{(0)} - \alpha \nabla f(x^{(0)})$ is an improvement over the point $x^{(0)}$ if we are searching for a minimizer.
Given a point $x^{(k)}$, to find the next point $x^{(k+1)}$, we move by an amount $-\alpha_k \nabla f(x^{(k)})$, where $\alpha_k$ is a positive scalar called the \textit{step size}.

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

We refer to this as a \textit{gradient descent algorithm} (or \textit{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 $\alpha_k$ is chosen to achieve the maximum amount of decrease of the objective function at each individual step.

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

- At each step, starting from the point $x^{(k)}$, we conduct a line search in the direction $-\nabla f(x^{(k)})$ until a minimizer, $x^{(k+1)}$, is found.
Proposition 8.1

Proposition 8.1: If \( \{x^{(k)}\}_{k=0}^{\infty} \) is a steepest descent sequence for a given function \( f : \mathbb{R}^n \to \mathbb{R} \), then for each \( k \) the vector \( x^{(k+1)} - x^{(k)} \) is orthogonal to the vector \( x^{(k+2)} - x^{(k+1)} \).

Proof: From the iterative formula of the method of steepest descent it follows that

\[
\langle x^{(k+1)} - x^{(k)}, x^{(k+2)} - x^{(k+1)} \rangle = \alpha_k \alpha_{k+1} \langle \nabla f(x^{(k)}), \nabla f(x^{(k+1)}) \rangle
\]

To complete the proof it is enough to show

\[
\langle \nabla f(x^{(k)}), \nabla f(x^{(k+1)}) \rangle = 0
\]

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

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

\[
= \nabla f(x^{(k)}) - \alpha_k \nabla f(x^{(k)})^T (- \nabla f(x^{(k)})) = -\langle \nabla f(x^{(k+1)}), f(x^{(k)}) \rangle
\]
Proposition 8.2

Proposition 8.2: If \( \{x^{(k)}\}_{k=0}^{\infty} \) is a steepest descent sequence for a given function \( f : \mathbb{R}^n \to \mathbb{R} \) and if \( \nabla f(x^{(k)}) \neq 0 \), then \( f(x^{(k+1)}) < f(x^{(k)}) \)

Proof: Recall that

\[
x^{(k+1)} = x^{(k)} - \alpha_k \nabla f(x^{(k)})
\]

where \( \alpha_k \geq 0 \) is the minimizer of

\[
\phi_k(\alpha) = f(x^{(k)} - \alpha \nabla f(x^{(k)}))
\]

over all \( \alpha \geq 0 \). Thus, for \( \alpha \geq 0 \), we have \( \phi_k(\alpha_k) \leq \phi_k(\alpha) \)

By the chain rule,

\[
\phi_k'(0) = \frac{d\phi_k}{d\alpha}(0) = -((\nabla f(x^{(k)}) - 0 \nabla f(x^{(k)}))^T(\nabla f(x^{(k)})) = -\|\nabla f(x^{(k)})\|^2 < 0
\]

because \( \nabla f(x^{(k)}) \neq 0 \) 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,

\[
f(x^{(k+1)}) = \phi_k(\alpha_k) \leq \phi_k(\bar{\alpha}) < \phi_k(0) = f(x^{(k)})
\]
Descent Property

- **Descent property**: \( f(x^{(k+1)}) < f(x^{(k)}) \) if \( \nabla f(x^{(k)}) \neq 0 \)
- If for some \( k \), we have \( \nabla f(x^{(k)}) = 0 \), then the point \( x^{(k)} \) satisfies the FONC. In this case, \( x^{(k+1)} = x^{(k)} \). We can use the above as the basis for a stopping criterion for the algorithm.
- The condition \( \nabla f(x^{(k)}) = 0 \), 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 \( \| \nabla f(x^{(k)}) \| \) is less than a prespecified threshold.
- Alternatively, we may compute \( |f(x^{(k+1)}) - f(x^{(k)})| \), and if the difference is less than some threshold, then we stop.
Descent Property

- Another alternative is to compute the norm $\|x^{(k+1)} - x^{(k)}\|$, and we stop if the norm is less than a prespecified threshold.

- We may check “relative” values of the quantities above

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

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

$$\frac{|f(x^{(k+1)}) - f(x^{(k)})|}{\max\{1, |f(x^{(k)})|\}} < \epsilon \quad \frac{\|x^{(k+1)} - x^{(k)}\|}{\max\{1, \|x^{(k)}\|\}} < \epsilon$$
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 $x^{(0)} = [4, 2, -1]^T$

We find that

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

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

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

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

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

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

Using the secant method from Section 7.4, we obtain

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

- Plot $\phi_0(\alpha)$ versus $\alpha$
- We compute
  \[ \mathbf{x}^{(1)} = \mathbf{x}^{(0)} - \alpha_0 \nabla f(\mathbf{x}^{(0)}) = [4.000, 2.008, -5.062]^T \]
- To find $\mathbf{x}^{(2)}$, we first determine $\nabla f(\mathbf{x}^{(1)}) = [0.000, -1.994, -0.003875]^T$
  Next, we find $\alpha_1$
  \[ \alpha_1 = \arg \min_{\alpha \geq 0} (0 + (2.008 + 1.984\alpha - 3)^2 + 4(-5.062 + 0.003875\alpha + 5)^4) \]
  \[ = \arg \min_{\alpha \geq 0} \phi_1(\alpha) \]
  Using the secant method again, we obtain $\alpha_1 = 0.5000$
Example

- Thus, $\mathbf{x}^{(2)} = \mathbf{x}^{(1)} - \alpha_1 \nabla f(\mathbf{x}^{(1)}) = [4.000, 3.000, -5.060]^T$

- To find $\mathbf{x}^{(3)}$, we first determine $\nabla f(\mathbf{x}^{(2)}) = [0.000, 0.000, -0.003525]^T$
  
  $\alpha_2 = \arg \min_{\alpha \geq 0} (0.000 + 0.000 + 4(-5.060 + 0.003525\alpha + 5)^4)$
  
  $= \arg \min_{\alpha \geq 0} \phi_2(\alpha)$

  $\alpha_1 = 16.29$

- The value $\mathbf{x}^{(3)} = [4.000, 3.000, -5.002]^T$

- 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.
A quadratic function of the form
\[ f(x) = \frac{1}{2}x^TQx - b^Tx \]
where \( Q \in \mathbb{R}^{m \times n} \) is a symmetric positive define matrix, \( b \in \mathbb{R}^n \) and \( x \in \mathbb{R}^n \). The unique minimizer of \( f \) can be found by setting the gradient of \( f \) to zero, where
\[ \nabla f(x) = Qx - b \]
because \( D(x^TQx) = x^T(Q + Q^T) = 2x^TQ \) and \( D(b^Tx) = b^T \).
Steepest Descent for Quadratic Function

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

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

where

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

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

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

- Because $\alpha_k \geq 0$ is the minimizer of $\phi_k(\alpha) = f(x^{(k)} - \alpha g^{(k)})$, we apply the FONC to $\phi_k(\alpha)$ to obtain
  \[ \phi_k'(\alpha) = (x^{(k)} - \alpha g^{(k)})^T Q(-g^{(k)}) - b^T (-g^{(k)}) \]

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

  Hence,
  \[ \alpha_k = \frac{g^{(k)^T} g^{(k)}}{g^{(k)^T} Q g^{(k)}} \]

- In summary, the method of steepest descent for the quadratic stakes the form
  \[ x^{(k+1)} = x^{(k)} - \frac{g^{(k)^T} g^{(k)}}{g^{(k)^T} Q g^{(k)}} g^{(k)} \]
  \[ g^{(k)} = \nabla f(x^{(k)}) = Qx^{(k)} - b \]
Example

- Let \( f(x_1, x_2) = x_1^2 + x_2^2 \). Then, starting from an arbitrary initial point \( x^{(0)} \in \mathbb{R}^2 \), we arrive at the solution \( x^* = 0 \in \mathbb{R}^2 \) 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 points 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*
The convergence analysis is more convenient if instead of working with $f$ we deal with

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

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

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

\[ V(x) = f(x) + \frac{1}{2}x^TQx^* = \frac{1}{2}(x - x^*)^TQ(x - x^*) \]

- **Lemma 8.1**: The iterative algorithm
  \[ x^{(k+1)} = x^{(k)} - \alpha_k g^{(k)} \]
  with \( g^{(k)} = Qx^{(k)} - b \) satisfies
  \[ V(x^{(k+1)}) = (1 - \gamma_k)V(x^{(k)}) \]
  where if \( g^{(k)} = 0 \), then \( \gamma_k = 1 \), and if \( g^{(k)} \neq 0 \), then
  \[ \gamma_k = \alpha_k \frac{g^{(k)}^TQg^{(k)}}{g^{(k)}^TQ^{-1}g^{(k)}} \left( 2\frac{g^{(k)}^TQg^{(k)}}{g^{(k)}^TQ^{-1}g^{(k)}} - \alpha_k \right) \]
Convergence

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

- Proof:

  - From Lemma 8.1 we have \( V(x^{(k+1)}) = (1 - \gamma_k) V(x^{(k)}) \), from which we obtain
    \[ V(x^{(k)}) = \left( \prod_{i=0}^{k-1} (1 - \gamma_i) \right) V(x^{(0)}) \]

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

\[ V(x^{(k)}) = \left( \prod_{i=0}^{k-1} (1 - \gamma_i) \right) V(x^{(0)}) \]
\[ V(x) = f(x) + \frac{1}{2} x^*^T Q x^* = \frac{1}{2} (x - x^*)^T Q (x - x^*) \]

- Note that \( x^{(k)} \to x^* \) if and only if \( V(x^{(k)}) \to 0 \). 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 \( \sum_{i=0}^{\infty} - \log(1 - \gamma_i) = \infty \).
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.
Example

- A counter example to show $\gamma_k > 0$ in Theorem 8.1 is necessary.
- For each $k = 0, 1, 2, \ldots$, choose $\alpha_k$ in such a way that $\gamma_{2k} = -1/2$ and $\gamma_{2k+1} = 1/2$ (we can always do this if, for example, $Q = I_n$). From Lemma 8.1 we have
  \[ V(x^{(2k+1)}) = (1 - 1/2)(1 + 1/2)V(x^{(2k)}) = (3/4)V(x^{(2k)}) \]
  Therefore, $V(x^{(2k)}) \to 0$. Because $V(x^{(2k+1)}) = (3/2)V(x^{(2k)})$, we also have that $V(x^{(2k+1)}) \to 0$.
  Hence, $V(x^{(k)}) \to 0$, which implies that $x^{(k)} \to 0$ (for all $x^{(0)}$). 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 $k$. 

\[ V(x^{(k+1)}) = (1 - \gamma_k)V(x^{(k)}) \]
Convergence

- Rayleigh’s inequality. For any $Q = Q^T > 0$, we have
  \[ \lambda_{\min}(Q) \|x\|^2 \leq x^T Q x \leq \lambda_{\max}(Q) \|x\|^2 \]
  We also have
  \[ \lambda_{\min}(Q^{-1}) = \frac{1}{\lambda_{\max}(Q)} \]
  \[ \lambda_{\max}(Q^{-1}) = \frac{1}{\lambda_{\min}(Q)} \]
  \[ \lambda_{\min}(Q^{-1}) \|x\|^2 \leq x^T Q^{-1} x \leq \lambda_{\max}(Q^{-1}) \|x\|^2 \]
Lemma 8.2: Let $Q = Q^T > 0$ be an $n \times n$ real symmetric positive definite matrix. Then, for any $x \in \mathbb{R}^n$, we have

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

Proof: Applying Rayleigh’s inequality and using the properties of symmetric positive definite matrices listed previously, we get

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

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

- Theorem 8.2: In the steepest descent algorithm, we have \( x^{(k)} \rightarrow x^* \) for any \( x^{(0)} \)

- Proof: If \( g^{(k)} = 0 \) for some \( k \), then \( x^{(k)} = x^* \) and the result holds. So assume that \( g^{(k)} \neq 0 \) for all \( k \). Recall that for the steepest descent algorithm,

  \[
  \alpha_k = \frac{g^{(k)T}g^{(k)}}{g^{(k)T}Qg^{(k)}}
  \]

  Substituting this expression for \( \alpha_k \) in the formula for \( \gamma_k \) yields

  \[
  \gamma_k = \frac{(g^{(k)T}g^{(k)})^2}{(g^{(k)T}Qg^{(k)})(g^{(k)T}Q^{-1}g^{(k)})}
  \]

  Note that in this case \( \gamma_k > 0 \) for all \( k \). Furthermore, by Lemma 8.2, we have \( \gamma_k \geq (\lambda_{\min}(Q)/\lambda_{\max}(Q)) > 0 \). Therefore, we have \( \sum_{k=0}^{\infty} \gamma_k = \infty \), and hence by Theorem 8.1, we conclude that \( x^{(k)} \rightarrow x^* \)
Consider now a gradient method with fixed step size; that is, $\alpha_k = \alpha \in \mathbb{R}$ for all $k$. The resulting algorithm is of the form

$$x^{(k+1)} = x^{(k)} - \alpha g^{(k)}.$$ 

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 $\alpha_k$. Clearly, the convergence of the algorithm depends on the choice of $\alpha_k$. 

---

---

---
Convergence

- Theorem 8.3: For the fixed-step-size gradient algorithm, \( x^{(k)} \to x^* \) for any \( x^{(0)} \) if and only if
  
  \[
  0 < \alpha < \frac{2}{\lambda_{\text{max}}(Q)}
  \]

Proof: \( \iff \): By Rayleigh’s inequality we have

\[
\lambda_{\text{min}}(Q)g^{(k)^T}g^{(k)} \leq g^{(k)^T}Qg^{(k)} \leq \lambda_{\text{max}}(Q)g^{(k)^T}g^{(k)}
\]

and

\[
g^{(k)^T}Q^{-1}g^{(k)} \leq \frac{1}{\lambda_{\text{max}}(Q)}g^{(k)^T}g^{(k)}
\]

Therefore, substituting the above in the formula for \( \gamma_k \), we have

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

Therefore, \( \gamma_k > 0 \) for all \( k \), and \( \sum_{k=0}^{\infty} \gamma_k = \infty \). Hence, by Theorem 8.1, we conclude that \( x^{(k)} \to x^* \).
Proof: \[\implies\]: We use contraposition. Suppose that either \(\alpha \leq 0\) or \(\alpha \geq 2/\lambda_{\text{max}}(Q)\). Let \(x^{(0)}\) be chosen such that \(x^0 - x^*\) is an eigenvector of \(Q\) corresponding to the eigenvalue \(\lambda_{\text{max}}(Q)\). Because

\[
x^{(k+1)} = x^{(k)} - \alpha(Qx^{(k)} - b) = x^{(k)} - \alpha(Qx^{(k)} - Qx^*)
\]

we obtain

\[
x^{(k+1)} - x^* = x^{(k)} - x^* - \alpha(Qx^{(k)} - Qx^*)
\]
\[
= (I_n - \alpha Q)(x^{(k)} - x^*)
\]
\[
= (I_n - \alpha Q)^{k+1}(x^{(0)} - x^*)
\]
\[
= (1 - \alpha \lambda_{\text{max}}(Q))^{k+1}(x^{(0)} - x^*)
\]

Taking norms on both sides, we get

\[
\|x^{(k+1)} - x^*\| = |1 - \alpha \lambda_{\text{max}}(Q)|^{k+1}\|x^{(0)} - x^*\|
\]

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

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

- Let the function $f$ be given by

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

We wish to find the minimizer of $f$ using a fixed-step-size gradient algorithm $x^{(k+1)} = x^{(k)} - \alpha \nabla f(x^{(k)})$ where $\alpha \in \mathbb{R}$ is a fixed step size.

- Solution: To apply Theorem 8.3, we first symmetrize the matrix in the quadratic term of $f$ to get

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

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

\[ V(x^{(k+1)}) \leq \frac{\lambda_{\text{max}}(Q) - \lambda_{\text{min}}(Q)}{\lambda_{\text{max}}(Q)} V(x^{(k)}) \]

- Proof: In the proof of Theorem 8.2, we showed that 

\[ \gamma_k \geq \frac{\lambda_{\text{min}}(Q)}{\lambda_{\text{max}}(Q)}. \]

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

and the result follows.

\[ V(x^{(k+1)}) = (1 - \gamma_k) V(x^{(k)}) \]
Convergence Rate

- Let \( r = \frac{\lambda_{\max}(Q)}{\lambda_{\min}(Q)} = \|Q\|\|Q^{-1}\| \)
  called the \textit{condition number} of \( Q \). Then, it follows from Theorem 8.4 that
  \[
  V(x^{(k+1)}) \leq (1 - \frac{1}{r})V(x^{(k)})
  \]

- The term \((1 - 1/r)\) plays an important role in the convergence of \( \{V(x^{(k)})\} \) to 0 (and hence of \( \{x^{(k)}\} \) to \( x^* \)). We refer to \((1 - 1/r)\) as the \textit{convergence ratio}.

- The smaller the value of \((1 - 1/r)\), the smaller \( V(x^{(k+1)}) \) will be relative to \( V(x^{(k)}) \), and hence the “faster” \( V(x^{(k)}) \) converges to 0.
Convergence Rate

- The convergence ratio $(1 - 1/r)$ decreases as $r$ decreases. If $r = 1$ then $\lambda_{max}(Q) = \lambda_{min}(Q)$, 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 $\{V(x^{(k)})\}$ (and hence $\{x^{(k)}\}$) decreases. The increase in $r$ reflects that fact that the contours of $f$ are more eccentric.
Convergence Rate

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

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

If for all \( p > 0 \)

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

then we say that the order of convergence is \( \infty \)

- Note that in the definition above, \( 0/0 \) should be understood to be 0.
The order of convergence of a sequence is a measure of its rate of convergence; \textit{the higher the order, the faster the rate of convergence}.

The order of convergence is sometimes also called the \textit{rate of convergence}. If \( p = 1 \) and \( \lim_{k \to \infty} \frac{\| x^{(k+1)} - x^* \|}{\| x^{(k)} - x^* \|} = 1 \) we say that the convergence is \textit{sublinear}.

If \( p = 1 \) and \( \lim_{k \to \infty} \frac{\| x^{(k+1)} - x^* \|}{\| x^{(k)} - x^* \|} < 1 \), we say that the convergence is \textit{linear}.

If \( p > 1 \), we say that the convergence is \textit{superlinear}.

If \( p = 2 \), we say that the convergence is \textit{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 \( \infty \). 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 \( \infty \). If \( p = 1 \), the sequence converges to \( \gamma \). 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,
  \[
  \frac{|x^{(k+1)}|}{|x^{(k)}|^p} = \frac{\gamma^{q^{k+1}}}{(\gamma^q)^p} = \gamma^{q^{k+1} - pq^k} = \gamma^{(q-p)q^k}
  \]

  If \( p < q \), the sequence converges to 0, whereas if \( p > q \), it grows to \( \infty \). 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 \( \infty \).
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 \textit{at least} $p$ if

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

Theorem 8.5: Let \( \{x^{(k)}\} \) be a sequence that converges to \( x^* \). If

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

then the order of convergence (if it exists) is at least \( p \).

Proof: Let \( s \) be the order of convergence of \( \{x^{(k)}\} \). Suppose that

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

Then, there exists \( c \) such that for sufficiently large \( k \),

\[
\frac{\|x^{(k+1)} - x^*\|}{\|x^{(k)} - x^*\|^p} \leq c
\]

Hence,

\[
\frac{\|x^{(k+1)} - x^*\|^p}{\|x^{(k)} - x^*\|^s} = \frac{\|x^{(k+1)} - x^*\|}{\|x^{(k)} - x^*\|^p} \|x^{(k)} - x^*\|^p \|x^{(k)} - x^*\|^{p-s}
\]

\[
\leq c\|x^{(k)} - x^*\|^{p-s}
\]
Convergence Rate

- Taking limits yields

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

- Because by definition \( s \) is the order of convergence

\[
\lim_{k \to \infty} \frac{\|x^{(k+1)} - x^*\|}{\|x^{(k)} - x^*\|_s} > 0
\]

Combining the two inequalities above, we get

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

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

- Similarly, we can show that if \( \|x^{(k+1)} - x^*\| = o(\|x^{(k)} - x^*\|^p) \) 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 \( |x^{(k+1)} - 2| \rightarrow 0 \). Also, we see that

\[
|x^{(k+1)} - 2| = o(|x^{(k)} - 2|^3).
\]

Hence, we conclude that \( p > 3 \).
Consider the problem of finding a minimizer of the function $f : \mathbb{R} \to \mathbb{R}$ 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

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

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.
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 $g^{(k)} \neq 0$ for all $k$ then $\gamma_k = 1$ if and only if $g^{(k)}$ is an eigenvector of $Q$.

Proof: Suppose that $g^{(k)} \neq 0$ for all $k$. Recall that for the steepest descent algorithm,

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

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

$$g^{(k)} = \nabla f(x^{(k)}) = Qx^{(k)} - b$$
Convergence Rate \[ x^* = x^{(k)} - \alpha_k g^{(k)} \]

- Premultiplying by \( Q \) and subtracting \( b \) from both sides yields
  \[ 0 = g^{(k)} - \alpha_k Q g^{(k)} \]
  which can be rewritten as
  \[ Q g^{(k)} = \frac{1}{\alpha_k} g^{(k)} \]
  Hence, \( g^{(k)} \) is an eigenvector of \( Q \).

- By the lemma, if \( g^{(k)} \) is not an eigenvector of \( Q \), then \( \gamma_k < 1 \)
  (recall that \( \gamma_k \) cannot exceed 1)
Theorem 8.6

Theorem 8.6: Let \( \{x^{(k)}\} \) be a convergent sequence of iterates of the steepest descent algorithm applied to a function \( f \). Then, the order of convergence of \( \{x^{(k)}\} \) is 1 in the worst case; that is, there exist a function \( f \) and an initial condition \( x^{(0)} \) such that the order of convergence of \( \{x^{(k)}\} \) is equal to 1.

Proof: Let \( f : \mathbb{R}^n \rightarrow \mathbb{R} \) be a quadratic function with Hessian \( Q \). Assume that the maximum and minimum eigenvalues of \( Q \) satisfy \( \lambda_{\text{max}}(Q) > \lambda_{\text{min}}(Q) \). To show that the order of convergence is 1, it suffices to show that there exists \( x^{(0)} \) such that

\[
\|x^{(k+1)} - x^*\| \geq c \|x^{(k)} - x^*\|
\]

for some \( c \).
Theorem 8.6

By Rayleigh’s inequality

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

Similarly,

\[ V(x^{(k)}) \geq \frac{\lambda_{\min}(Q)}{2} \|x^{(k)} - x^*\|^2 \]

Combining the inequalities above with Lemma 8.1, we obtain

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

Therefore, it suffices to choose \( x^{(0)} \) such that \( \gamma_k \leq d \) for some \( d < 1 \)
Theorem 8.6

- Recall that for the steepest descent algorithm, assuming that \( g^{(k)} \neq 0 \) for all \( k \),
  \[
  \gamma_k = \frac{(g^{(k)T}g^{(k)})^2}{(g^{(k)T}Qg^{(k)})(g^{(k)T}Q^{-1}g^{(k)})}
  \]

- First consider the case where \( n = 2 \). Suppose that \( x^{(0)} \neq x^* \) is chosen such that \( x^{(0)} - x^* \) is not an eigenvector of \( Q \). Then, \( g^{(0)} = Q(x^{(0)} - x^*) \neq 0 \) is also not an eigenvector of \( Q \).

- By Proposition 8.1, \( g^{(k)} = (x^{(k+1)} - x^{(k)})/\alpha_k \) is not an eigenvector of \( Q \) for any \( k \) [because any two eigenvectors corresponding to \( \lambda_{max}(Q) \) and \( \lambda_{min}(Q) \) are mutually orthogonal].

- Also, \( g^{(k)} \) lies in one of two mutually orthogonal directions. Therefore, by Lemma 8.3, for each \( k \), the value of \( \gamma_k \) 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 $v_1$ and $v_2$ be mutually orthogonal eigenvectors corresponding to $\lambda_{\text{max}}(Q)$ and $\lambda_{\text{min}}(Q)$. Choose $x^{(0)}$ such that $x^{(0)} - x^* \neq 0$ lies in the span of $v_1$ and $v_2$ but is not equal to either.
- Note that $g^{(0)} = Q(x^{(0)} - x^*)$ also lies in the span of $v_1$ and $v_2$, but is not equal to either.
- By manipulating $x^{(k+1)} = x^{(k)} - \alpha_k g^{(k)}$ as before, we can write $g^{(k+1)} = (I - \alpha_k Q)g^{(k)}$. Any eigenvector of $Q$ is also an eigenvector of $I - \alpha_k Q$. Therefore, $g^{(k)}$ lies in the span of $v_1$ and $v_2$ for all $k$; that is, the sequence $\{g^{(k)}\}$ is confined within the two-dimensional subspace spanned by $v_1$ and $v_2$. We can now proceed as in the $n = 2$ case.