Analysis of Steepest decent

Oscar Dalmau
dalmau@cimat.mx

Centro de Investigación en Matemáticas
CIMAT A.C. Mexico

February 2018

Outline

1. Steepest descent Method
2. Global Convergence
3. Rate of convergence

Steepest descent Method with exact line search

- The method of *steepest descent with exact step size or with exact line search* is a gradient algorithm where the step size is obtained by solving

$$\alpha_k = \arg\min_{\alpha > 0} \phi(\alpha)$$

  with $\phi(\alpha) = f(\boldsymbol{x}_k + \alpha \boldsymbol{d}_k)$, $\boldsymbol{d}_k = -\boldsymbol{g}_k$.
- The previous method moves in orthogonal steps, see next proposition.

Steepest descent Method with exact line search

Proposition 1.1

(Orthogonality of directions) If $\{\boldsymbol{x}_k\}_{k=1}^{\infty}$ *is a steepest descent sequence for a given function* $f : \mathbb{R}^n \to \mathbb{R}$, *then for each k the vector* $\boldsymbol{x}_{k+1} - \boldsymbol{x}_k$ *is orthogonal to the vector* $\boldsymbol{x}_{k+2} - \boldsymbol{x}_{k+1}$

Remark

1. Note that $\alpha_k \boldsymbol{d}_k = \boldsymbol{x}_{k+1} - \boldsymbol{x}_k$ and $\alpha_{k+1}\boldsymbol{d}_{k+1} = \boldsymbol{x}_{k+2} - \boldsymbol{x}_{k+1}$. Therefore, the proposition states that two consecutive directions are orthogonal, i.e. $\boldsymbol{d}_k \perp \boldsymbol{d}_{k+1}$, where $\boldsymbol{d}_k = -\boldsymbol{g}_k$ and $\boldsymbol{d}_{k+1} = -\boldsymbol{g}_{k+1}$ then $\boldsymbol{g}_k \perp \boldsymbol{g}_{k+1}$.
2. The solution trajectory of the steepest-descent method with exact line search follows a zig-zag pattern.

Steepest descent Method with exact line search

Proposition

If $\{\boldsymbol{x}_k\}_{k=1}^{\infty}$ is a steepest descent sequence for a given function $f : \mathbb{R}^n \to \mathbb{R}$, 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.

As α_k minimizes $\phi(\alpha) = f(\boldsymbol{x}_k + \alpha \boldsymbol{d}_k)$ then $\phi'(\alpha_k) = 0$. Using $\boldsymbol{d}_k = -\boldsymbol{g}_k = -\nabla f(\boldsymbol{x}_k)$,

$$\phi'(\alpha_k) \quad = \quad \nabla f(\boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k)^T \boldsymbol{d}_k = \nabla f(\boldsymbol{x}_{k+1})^T \boldsymbol{d}_k = -\boldsymbol{g}_{k+1}^T \boldsymbol{g}_k$$

then $\boldsymbol{g}_k \perp \boldsymbol{g}_{k+1}$.

□

Steepest descent Method with exact line search

Proposition

If $\{\boldsymbol{x}_k\}_{k=1}^{\infty}$ is a steepest descent sequence for a given function $f : \mathbb{R}^n \to \mathbb{R}$, then for each k the vector $\boldsymbol{x}_{k+1} - \boldsymbol{x}_k$ is orthogonal to the vector $\boldsymbol{x}_{k+2} - \boldsymbol{x}_{k+1}$

Corollary

$\boldsymbol{g}_k$ is parallel to the tangent plane to the level set $\{\boldsymbol{x}|\ f(\boldsymbol{x}) = f(\boldsymbol{x}_{k+1})\}$ at $\boldsymbol{x}_{k+1}$.

Steepest descent Method with exact line search

Proposition 1.2

If $\{\boldsymbol{x}_k\}_{k=1}^{\infty}$ *is a steepest descent sequence for a given function* $f : \mathbb{R}^n \to \mathbb{R}$ *and if* $\nabla f(\boldsymbol{x}_k) \neq \mathbf{0}$ *then* $f(\boldsymbol{x}_{k+1}) < f(\boldsymbol{x}_k)$.

Steepest descent Method with exact line search

Proposition

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

Proof.

As α_k minimizes $\phi(\alpha) = f(\boldsymbol{x}_k + \alpha \boldsymbol{d}_k)$ then $\phi(\alpha_k) \leq \phi(\alpha)$ for all α. On the other hand, $\phi'(0) = \nabla f(\boldsymbol{x}_k)^T \boldsymbol{g}_k = -\|\boldsymbol{g}_k\|^2 < 0$ due to $\boldsymbol{g}_k \neq \mathbf{0}$.

Therefore, there exists $\hat{\alpha}$ (by the sign preserving theorem) such that $\phi'(\alpha) < 0$ for $\alpha \in (0, \hat{\alpha})$. Using Taylor (or the mean value theorem), there exists $\bar{\alpha}$ such that $\phi(\alpha) - \phi(0) = \phi'(\bar{\alpha})\alpha$ with $\bar{\alpha} \in (0, \alpha)$ then $\phi'(\bar{\alpha})\alpha < 0$ and hence $\phi(\alpha) < \phi(0)$ for $\alpha \in (0, \hat{\alpha})$. Then $\phi(\alpha_k) < \phi(0)$ for $\alpha \in (0, \hat{\alpha})$, i.e., $f(\boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k) < f(\boldsymbol{x}_k)$ or $f(\boldsymbol{x}_{k+1}) < f(\boldsymbol{x}_k)$.

□

Stopping criteria

1. The previous proposition states that the steepest descent has the property: $f(\boldsymbol{x}_{k+1}) < f(\boldsymbol{x}_k)$ if $\boldsymbol{g}_k \neq \mathbf{0}$.
2. If for some k, it holds $\nabla f(\boldsymbol{x}_k) = \mathbf{0}$ then $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k$. We can use this as a stopping criterion for the algorithm, however, the gradient will rarely be identically equal to zero.
3. A practical stopping criterion is to check if the norm $\|\nabla f(\boldsymbol{x}_k)\|$ of the gradient is less than a threshold, i.e. $\|\nabla f(\boldsymbol{x}_k)\| \leq \tau$.

Stopping criteria

Other stopping criteria

$$\begin{aligned} |f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}_k)| &\leq \tau \\ \|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\| &\leq \tau \end{aligned}$$

We may check the relative values of the above quantities

$$\begin{aligned} \frac{|f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}_k)|}{|f(\boldsymbol{x}_k)|} &\leq \tau \\ \frac{\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|}{\|\boldsymbol{x}_k\|} &\leq \tau \end{aligned}$$

The above two (relative) stopping criteria are preferable to the previous (absolute) criteria because the relative criteria are scale-independent.

Stopping criteria

To avoid dividing by a small number we may use the following modifications

$$\boxed{\begin{aligned} \frac{|f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}_k)|}{\max\{1, |f(\boldsymbol{x}_k)|\}} &\leq \tau_f \\ \frac{\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|}{\max\{1, \|\boldsymbol{x}_k\|\}} &\leq \tau_x \end{aligned}}$$

Stopping criteria

To avoid dividing by a small number we may use the following modifications

$$\boxed{\begin{aligned}\frac{|f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}_k)|}{\max\{1, |f(\boldsymbol{x}_k)|\}} &\le \tau_f \\ \frac{\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|}{\max\{1, \|\boldsymbol{x}_k\|\}} &\le \tau_x\end{aligned}}$$

$$\boxed{\begin{aligned}\|\nabla f(\boldsymbol{x}_k)\| \le \tau_g \\ k > K_{\max}\end{aligned}}$$

Exact Steepest descent for a quadratic function

Let $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\mathbf{Q}\boldsymbol{x} - \boldsymbol{b}^T\boldsymbol{x}$, with $\mathbf{Q}$ positive definite. As

$$\alpha_k \quad = \quad \arg\max_{\alpha>0} \phi(\alpha) = f(\boldsymbol{x}_k + \alpha \boldsymbol{d}_k)$$

Exact Step for a quadratic function

$$\phi(\alpha) = \frac{1}{2}(\boldsymbol{x}_k + \alpha \boldsymbol{d}_k)^T \mathbf{Q}(\boldsymbol{x}_k + \alpha \boldsymbol{d}_k) - \boldsymbol{b}^T(\boldsymbol{x}_k + \alpha \boldsymbol{d}_k)$$

then, from $\phi'(\alpha) = 0$ and $\boldsymbol{g}_k = \mathbf{Q}\boldsymbol{x}_k - \boldsymbol{b} = -\boldsymbol{d}_k$

$$\begin{aligned}
(\boldsymbol{x}_k + \alpha \boldsymbol{d}_k)^T \mathbf{Q}\boldsymbol{d}_k - \boldsymbol{b}^T \boldsymbol{d}_k &= 0 \\
\boldsymbol{d}_k^T \mathbf{Q}\boldsymbol{d}_k \alpha &= -(\mathbf{Q}\boldsymbol{x}_k - \boldsymbol{b})^T \boldsymbol{d}_k \\
\alpha_k &= \frac{-\boldsymbol{g}_k^T \boldsymbol{d}_k}{\boldsymbol{d}_k^T \mathbf{Q}\boldsymbol{d}_k} = \frac{\boldsymbol{g}_k^T \boldsymbol{g}_k}{\boldsymbol{g}_k^T \mathbf{Q}\boldsymbol{g}_k}
\end{aligned}$$

Exact Steepest descent for a quadratic function

The update formula for the Steepest descent with Exact step size for the quadratic function $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\mathbf{Q}\boldsymbol{x} - \boldsymbol{b}^T\boldsymbol{x}$, with $\mathbf{Q}$ positive definite, is:

$$\boldsymbol{x}_{k+1} \quad = \quad \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k$$

with

$$\boxed{\begin{aligned} \boldsymbol{d}_k &= -\boldsymbol{g}_k \\ \alpha_k &= \frac{\boldsymbol{g}_k^T\boldsymbol{g}_k}{\boldsymbol{g}_k^T\mathbf{Q}\boldsymbol{g}_k} \end{aligned}}$$

Steepest descent: elimination of line search

If $f(x)$ is not quadratic but the **Hessian $\mathbf{H}_k$ is available**: then, we can approximate

$$\phi(\alpha) = f(x_k + \alpha \boldsymbol{d}_k) \quad \approx \quad f(x_k) + \alpha \boldsymbol{g}_k^T \boldsymbol{d}_k + \frac{1}{2}\alpha^2 \boldsymbol{d}_k^T \mathbf{H}_k \boldsymbol{d}_k$$

from $\phi'(\alpha) = 0$ and $\boldsymbol{d}_k = -\boldsymbol{g}_k$

$$\alpha = \alpha_k \quad \approx \quad \frac{\boldsymbol{g}_k^T \boldsymbol{g}_k}{\boldsymbol{g}_k^T \mathbf{H}_k \boldsymbol{g}_k}$$

then

$$\boxed{\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \frac{\boldsymbol{g}_k^T \boldsymbol{g}_k}{\boldsymbol{g}_k^T \mathbf{H}_k \boldsymbol{g}_k} \boldsymbol{g}_k}$$

Steepest descent: elimination of line search

If the Hessian $\mathbf{H}_k$ is not available: suppose we have an estimate $\hat{\alpha}$ of α_k. Then $\hat{f} = f(x_k + \hat{\alpha}\boldsymbol{d}_k)$

$$\phi(\hat{\alpha}) = f(x_k + \hat{\alpha}\boldsymbol{d}_k) \approx f(x_k) + \hat{\alpha}\boldsymbol{g}_k^T\boldsymbol{d}_k + \frac{1}{2}\hat{\alpha}^2\boldsymbol{d}_k^T\mathbf{H}_k\boldsymbol{d}_k$$

$$\boldsymbol{g}_k^T\mathbf{H}_k\boldsymbol{g}_k \approx 2\frac{\hat{f} - f_k + \hat{\alpha}\boldsymbol{g}_k^T\boldsymbol{g}_k}{\hat{\alpha}^2}$$

from $\phi'(\alpha) = 0$ and $\boldsymbol{d}_k = -\boldsymbol{g}_k$

$$\boxed{\alpha_k = \frac{\boldsymbol{g}_k^T\boldsymbol{g}_k}{\boldsymbol{g}_k^T\mathbf{H}_k\boldsymbol{g}_k} \approx \frac{\boldsymbol{g}_k^T\boldsymbol{g}_k\hat{\alpha}^2}{2(\hat{f} - f_k + \hat{\alpha}\boldsymbol{g}_k^T\boldsymbol{g}_k)}}$$

then

$$\boxed{\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \frac{\boldsymbol{g}_k^T\boldsymbol{g}_k\hat{\alpha}^2}{2(\hat{f} - f_k + \hat{\alpha}\boldsymbol{g}_k^T\boldsymbol{g}_k)}\boldsymbol{g}_k}$$

we can use for example: $\hat{\alpha} = \alpha_{k-1}$.

Steepest descent: with fixed step size

$$\boxed{\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha_k \boldsymbol{g}_k}$$

with $\alpha_k = \alpha$ where $\alpha > 0$ is a small value provided by the user, for example, $\alpha = 1e-2$

Global Convergence

1. 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 *first order necessary condition* for a minimizer.
2. When the algorithm is not globally convergent, it may still generate a sequence that converges to a point satisfying the *first order necessary condition*, if the initial point is sufficiently close to the point. In this case, we say that the algorithm is *locally convergent*.
3. Another issue of interest, for both locally or globally convergent algorithms, is the rate of convergence; i.e., how fast the algorithm converges to a solution point.

Quadratic case

Lets start by the Quadratic case.

Let $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\mathbf{Q}\boldsymbol{x} - \boldsymbol{b}^T\boldsymbol{x}$, with $\mathbf{Q}$ is a symmetric positive definite matrix.

Note that there is not loss of generality in considering $\mathbf{Q}$ to be a symmetric matrix, due to, if $\mathbf{A}$ is not symmetric

$$\boldsymbol{x}^T\mathbf{A}\boldsymbol{x} = \frac{1}{2}(\boldsymbol{x}^T\mathbf{A}\boldsymbol{x} + \boldsymbol{x}^T\mathbf{A}^T\boldsymbol{x}) = \boldsymbol{x}^T\frac{\mathbf{A}+\mathbf{A}^T}{2}\boldsymbol{x} := \boldsymbol{x}^T\mathbf{Q}\boldsymbol{x}$$

where $\mathbf{Q} = \frac{\mathbf{A}+\mathbf{A}^T}{2}$ is symmetric!.

Quadratic case

The convergence analysis is more convenient if we consider the following function

$$\begin{aligned} E(\boldsymbol{x}) &= f(\boldsymbol{x}) + \frac{1}{2}(\boldsymbol{x}^*)^T \mathbf{Q} \boldsymbol{x}^* \\ &= \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}^*)^T \mathbf{Q} (\boldsymbol{x} - \boldsymbol{x}^*) \\ &= \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{x}^*\|_{\mathbf{Q}}^2 \end{aligned}$$

which differs from $f(\boldsymbol{x})$ in the constant $\frac{1}{2}(\boldsymbol{x}^*)^T \mathbf{Q} \boldsymbol{x}^* = \frac{1}{2}\|\boldsymbol{x}^*\|_{\mathbf{Q}}^2$, ie

$$E(\boldsymbol{x}) = f(\boldsymbol{x}) + \frac{1}{2}\|\boldsymbol{x}^*\|_{\mathbf{Q}}^2$$

Note: $\|\boldsymbol{x}\|_{\mathbf{Q}}^2 = \boldsymbol{x}^T \mathbf{Q} \boldsymbol{x}$ is the weighted norm.

Quadratic case

Lemma 2.1

The iterates $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha_k \boldsymbol{g}_k$ *with* $\boldsymbol{g}_k = \mathbf{Q}\boldsymbol{x}_k - \boldsymbol{b}$ *satisfies that if* $\boldsymbol{g}_k \neq \mathbf{0}$ *then* $E(\boldsymbol{x}_{k+1}) = (1-\gamma_k)E(\boldsymbol{x}_k)$ *where*

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

if additionally $\alpha_k = \frac{\boldsymbol{g}_k^T \boldsymbol{g}_k}{\boldsymbol{g}_k^T \mathbf{Q} \boldsymbol{g}_k}$ *then*

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

Note: if $\boldsymbol{g}_k = \mathbf{0}$ we consider $\gamma_k = 1$

Quadratic case

$$
\begin{aligned}
E(\boldsymbol{x}_{k+1}) &= \frac{1}{2}(\boldsymbol{x}_k - \boldsymbol{x}^* - \alpha_k \boldsymbol{g}_k)^T \mathbf{Q}(\boldsymbol{x}_k - \boldsymbol{x}^* - \alpha_k \boldsymbol{g}_k) \\
&= \frac{1}{2}(\boldsymbol{x}_k - \boldsymbol{x}^*)^T \mathbf{Q}(\boldsymbol{x}_k - \boldsymbol{x}^*) + \frac{1}{2}\alpha_k^2 \boldsymbol{g}_k^T \mathbf{Q}\boldsymbol{g}_k \\
&- \alpha_k(\boldsymbol{x}_k - \boldsymbol{x}^*)^T \mathbf{Q}\boldsymbol{g}_k
\end{aligned}
$$

Quadratic case

On the other hand $\boldsymbol{g}_k = \mathbf{Q}\boldsymbol{x}_k - \boldsymbol{b} = \mathbf{Q}(\boldsymbol{x}_k - \boldsymbol{x}^*)$ then

$$\boxed{\boldsymbol{x}_k - \boldsymbol{x}^* = \mathbf{Q}^{-1}\boldsymbol{g}_k}$$

and

$$\boxed{E(\boldsymbol{x}_k) = \frac{1}{2}(\boldsymbol{x}_k - \boldsymbol{x}^*)^T\mathbf{Q}(\boldsymbol{x}_k - \boldsymbol{x}^*) = \frac{1}{2}\boldsymbol{g}_k^T\mathbf{Q}^{-1}\boldsymbol{g}_k}$$

$$\begin{aligned} E(\boldsymbol{x}_{k+1}) &= \frac{1}{2}(\boldsymbol{x}_k - \boldsymbol{x}^*)^T\mathbf{Q}(\boldsymbol{x}_k - \boldsymbol{x}^*) + \frac{1}{2}\alpha_k^2\boldsymbol{g}_k^T\mathbf{Q}\boldsymbol{g}_k - \alpha_k\boldsymbol{g}_k^T\boldsymbol{g}_k \\ &= E(\boldsymbol{x}_k) + (\frac{1}{2}\alpha_k^2\boldsymbol{g}_k^T\mathbf{Q}\boldsymbol{g}_k - \alpha_k\boldsymbol{g}_k^T\boldsymbol{g}_k)\frac{E(\boldsymbol{x}_k)}{E(\boldsymbol{x}_k)} \\ &= [1 - \gamma_k]E(\boldsymbol{x}_k) \end{aligned}$$

Quadratic case

$$
\begin{aligned}
E(\boldsymbol{x}_{k+1}) &= (1-\gamma_k)E(\boldsymbol{x}_k) \\
\|\boldsymbol{x}_{k+1}-\boldsymbol{x}^*\|^2_{\mathbf{Q}} &= (1-\gamma_k)\|\boldsymbol{x}_k-\boldsymbol{x}^*\|^2_{\mathbf{Q}}
\end{aligned}
$$

with

$$
\begin{aligned}
\gamma_k &= -\frac{\frac{1}{2}\alpha_k^2\boldsymbol{g}_k^T\mathbf{Q}\boldsymbol{g}_k-\alpha_k\boldsymbol{g}_k^T\boldsymbol{g}_k}{\frac{1}{2}g_k^T\mathbf{Q}^{-1}g_k} \\
&= \alpha_k\frac{\boldsymbol{g}_k^T\mathbf{Q}\boldsymbol{g}_k}{\boldsymbol{g}_k^T\mathbf{Q}^{-1}\boldsymbol{g}_k}(2\frac{\boldsymbol{g}_k^T\boldsymbol{g}_k}{\boldsymbol{g}_k^T\mathbf{Q}\boldsymbol{g}_k}-\alpha_k)
\end{aligned}
$$

Quadratic case

If $\alpha_k = \frac{\boldsymbol{g}_k^T \boldsymbol{g}_k}{\boldsymbol{g}_k^T \mathbf{Q} \boldsymbol{g}_k}$

$$
\begin{aligned}
\gamma_k &= \alpha_k \frac{\boldsymbol{g}_k^T \mathbf{Q} \boldsymbol{g}_k}{\boldsymbol{g}_k^T \mathbf{Q}^{-1} \boldsymbol{g}_k} \left(2 \frac{\boldsymbol{g}_k^T \boldsymbol{g}_k}{\boldsymbol{g}_k^T \mathbf{Q} \boldsymbol{g}_k} - \alpha_k\right) \\
&= \frac{\boldsymbol{g}_k^T \boldsymbol{g}_k}{\boldsymbol{g}_k^T \mathbf{Q} \boldsymbol{g}_k} \frac{\boldsymbol{g}_k^T \mathbf{Q} \boldsymbol{g}_k}{\boldsymbol{g}_k^T \mathbf{Q}^{-1} \boldsymbol{g}_k} \left(2 \frac{\boldsymbol{g}_k^T \boldsymbol{g}_k}{\boldsymbol{g}_k^T \mathbf{Q} \boldsymbol{g}_k} - \frac{\boldsymbol{g}_k^T \boldsymbol{g}_k}{\boldsymbol{g}_k^T \mathbf{Q} \boldsymbol{g}_k}\right) \\
&= \frac{(\boldsymbol{g}_k^T \boldsymbol{g}_k)^2}{\boldsymbol{g}_k^T \mathbf{Q} \boldsymbol{g}_k \boldsymbol{g}_k^T \mathbf{Q}^{-1} \boldsymbol{g}_k}
\end{aligned}
$$

Quadratic case

Remark 2.2

Note that $\gamma_k = 1 - \frac{E(\boldsymbol{x}_{k+1})}{E(\boldsymbol{x}_k)} \geq 0$ due to $E(\boldsymbol{x}_{k+1}) \leq E(\boldsymbol{x}_k)$ and $\gamma_k \leq 1$ due to $E(\boldsymbol{x}) \geq 0$, then $\boxed{0 \leq \gamma_k \leq 1}$.

Quadratic case

Theorem 2.3

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 the previous Lemma, 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

$$\sum_{k=1}^{\infty} \gamma_k = \infty$$

Quadratic case

From $E(\boldsymbol{x}_{k+1}) = (1-\gamma_k)E(\boldsymbol{x}_k)$ we obtain

$$E(\boldsymbol{x}_k) = \prod_{i=0}^{k-1}(1-\gamma_i)E(\boldsymbol{x}_0)$$

1. If $\gamma_i = 1$ the result is trivial.
2. Assume $\gamma_i < 1$ Then, $\boldsymbol{x}_k \to \boldsymbol{x}^*$ if and only if $E(\boldsymbol{x}_k) \to 0$.

Therefore $\prod_{i=0}^{\infty}(1-\gamma_i) \to 0$, iff $-\sum_{i=0}^{\infty} \log(1-\gamma_i) = \infty$

Quadratic case

It remains to proof that $\sum_{i=0}^{\infty} -\log(1-\gamma_i) = \infty$ iff $\sum_{i=0}^{\infty} \gamma_i = \infty$.

($\Leftarrow$) if $\sum_{i=0}^{\infty} \gamma_i = \infty$ and taking into account that

$$\log(x) \leq x - 1$$

then

$$-\log(1-x) \geq x$$

and therefore

$$\sum_{i=0}^{\infty} -\log(1-\gamma_i) \geq \sum_{i=0}^{\infty} \gamma_i = \infty$$

.

Quadratic case

($\Rightarrow$) (by contradiction) Suppose $\sum_{i=0}^{\infty} \gamma_i < \infty$ therefore $\gamma_i \to 0$, ie

$$1 - \gamma_i \approx 1$$

for all $i \geq j$, for sufficiently large j.

As $\log(x) \geq 2(x-1)$ for x close to 1 then $-\log(1-x) \leq 2x$. Therefore

$$\sum_{i=j}^{\infty} -\log(1-\gamma_i) \leq 2 \sum_{i=j}^{\infty} \gamma_i < \infty$$

then

$$\sum_{i=0}^{\infty} -\log(1-\gamma_i) < \infty$$

that contradicts the hypothesis that $\sum_{i=0}^{\infty} -\log(1-\gamma_i) = \infty$.

Quadratic case

Lemma 2.4

Let $\mathbf{Q}$ be an $n \times n$ real symmetric positive definite matrix. Then,for any $\boldsymbol{x} \in \mathbb{R}^n$, we have

$$\frac{a}{A} \leq \frac{(\boldsymbol{x}^T\boldsymbol{x})^2}{\boldsymbol{x}^T\mathbf{Q}\boldsymbol{x}\boldsymbol{x}^T\mathbf{Q}^{-1}\boldsymbol{x}} \leq \frac{A}{a}$$

where a and A are, respectively, the smallest and largest eigenvalues of $\mathbf{Q}$.

Quadratic case

Proof.

Applying Rayleigh's inequality, we get

$$a \leq \frac{\boldsymbol{x}^T \mathbf{Q} \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{x}} \leq A, \text{ and } \frac{1}{A} \leq \frac{\boldsymbol{x}^T \mathbf{Q}^{-1} \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{x}} \leq \frac{1}{a}$$

therefore $\frac{a}{A} \leq \frac{(\boldsymbol{x}^T \boldsymbol{x})^2}{\boldsymbol{x}^T \mathbf{Q} \boldsymbol{x} \boldsymbol{x}^T \mathbf{Q}^{-1} \boldsymbol{x}} \leq \frac{A}{a}$. $\square$

Proposition 2.5

Kantorovich inequality: *Let $\mathbf{Q}$ be a positive definite symmetric $n \times n$ matrix. For any vector $\boldsymbol{x}$ there holds*

$$\frac{(\boldsymbol{x}^T\boldsymbol{x})^2}{\boldsymbol{x}^T\mathbf{Q}\boldsymbol{x}\boldsymbol{x}^T\mathbf{Q}^{-1}\boldsymbol{x}} \geq \frac{4aA}{(a+A)^2}$$

where a and A are, respectively, the smallest and largest eigenvalues of $\mathbf{Q}$.

Remark 2.6

Note that $\frac{4aA}{(a+A)^2} \geq \frac{a}{A}$ therefore

$$\frac{a}{A} \leq \frac{4aA}{(a+A)^2} \leq \frac{(\boldsymbol{x}^T\boldsymbol{x})^2}{\boldsymbol{x}^T\mathbf{Q}\boldsymbol{x}\boldsymbol{x}^T\mathbf{Q}^{-1}\boldsymbol{x}} \leq \frac{A}{a}$$

Quadratic case: Steepest decent with exact line search

Theorem 2.7

In the steepest decent algorithm $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha_k \boldsymbol{g}_k$, *with exact line search, i.e.* $\alpha_k = \frac{\boldsymbol{g}_k^T \boldsymbol{g}_k}{\boldsymbol{g}_k^T \mathbf{Q} \boldsymbol{g}_k}$, *we have that* $\boldsymbol{x}_k \to \boldsymbol{x}^*$ *for any* $\boldsymbol{x}_0$.

Proof.

As $\gamma_k = \frac{(\boldsymbol{g}_k^T \boldsymbol{g}_k)^2}{\boldsymbol{g}_k^T \mathbf{Q} \boldsymbol{g}_k \boldsymbol{g}_k^T \mathbf{Q}^{-1} \boldsymbol{g}_k} \geq \frac{a}{A} > 0$. Therefore $\sum_k^\infty \gamma_k = \infty$ and using Theorem 2.3, the sequence $\{\boldsymbol{x}_k\}$ converges to $\boldsymbol{x}^*$ for any initial condition $\boldsymbol{x}_0$ $\square$

Quadratic case: Gradient method with fixed step size

Theorem 2.8

In the steepest decent algorithm $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha \boldsymbol{g}_k$, with with fixed step size, i.e. $\alpha_k = \alpha$ for all k, we have that $\boldsymbol{x}_k \to \boldsymbol{x}^*$ for any $\boldsymbol{x}_0$ iff $0 < \alpha < \frac{2}{A}$.

Proof.

($\Leftarrow$) By Rayleigh's inequality

$$a\boldsymbol{g}_k^T\boldsymbol{g}_k \leq \boldsymbol{g}_k^T\mathbf{Q}\boldsymbol{g}_k \leq A\boldsymbol{g}_k^T\boldsymbol{g}_k$$

$$\boldsymbol{g}_k^T\mathbf{Q}^{-1}\boldsymbol{g}_k \leq \frac{1}{a}\boldsymbol{g}_k^T\boldsymbol{g}_k$$

As $\gamma_k = \alpha\frac{\boldsymbol{g}_k^T\mathbf{Q}\boldsymbol{g}_k}{\boldsymbol{g}_k^T\mathbf{Q}^{-1}\boldsymbol{g}_k}\left(2\frac{\boldsymbol{g}_k^T\boldsymbol{g}_k}{\boldsymbol{g}_k^T\mathbf{Q}\boldsymbol{g}_k} - \alpha\right) \geq a a\alpha(\frac{2}{A} - \alpha) > 0$ then $\sum_k^\infty \gamma_k = \infty$ and using Theorem 2.3, the sequence $\boldsymbol{x}_k \to \boldsymbol{x}^*$ $\square$

Gradient method with fixed step size

Proof.

($\Rightarrow$) (By contradiction) Suppose that $\alpha \leq 0$ or $\alpha \geq \frac{2}{A}$. On the other hand, we select $\boldsymbol{x}_0$ such that $\boldsymbol{x}_0 - \boldsymbol{x}^*$ is the eigenvector of $\mathbf{Q}$ that corresponds to A.

$$\begin{aligned}
\boldsymbol{x}_{k+1} - \boldsymbol{x}^* &= \boldsymbol{x}_k - \alpha \boldsymbol{g}_k - \boldsymbol{x}^* = \boldsymbol{x}_k - \alpha(\mathbf{Q}\boldsymbol{x}_k - \boldsymbol{b}) - \boldsymbol{x}^* \\
&= \boldsymbol{x}_k - \boldsymbol{x}^* - \alpha(\mathbf{Q}\boldsymbol{x}_k - \mathbf{Q}\boldsymbol{x}^*) = (\mathbf{I} - \alpha\mathbf{Q})(\boldsymbol{x}_k - \boldsymbol{x}^*) \\
&= (\mathbf{I} - \alpha\mathbf{Q})^{k+1}(\boldsymbol{x}_0 - \boldsymbol{x}^*) = (1 - \alpha A)^{k+1}(\boldsymbol{x}_0 - \boldsymbol{x}^*)
\end{aligned}$$

then

$$\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\| = |1 - \alpha A|^{k+1} \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|$$

For $\alpha \leq 0$ or $\alpha \geq \frac{2}{A}$ we have $|1 - \alpha A| \geq 1$ therefore $\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|$ does not converge to 0 and $\{\boldsymbol{x}_k\}$ does not converge to $\boldsymbol{x}^*$. $\square$

Quadratic case: Steepest decent with exact line search

Theorem 3.1

In the steepest decent algorithm with exact line search, i.e. $\alpha_k = \frac{\boldsymbol{g}_k^T \boldsymbol{g}_k}{\boldsymbol{g}_k^T \mathbf{Q} \boldsymbol{g}_k}$, *applied to the quadratic function we have*

$$E(\boldsymbol{x}_{k+1}) \leq \left(1 - \frac{a}{A}\right) E(\boldsymbol{x}_k)$$

Proof.

As $\gamma_k = \frac{(\boldsymbol{g}_k^T \boldsymbol{g}_k)^2}{\boldsymbol{g}_k^T \mathbf{Q} \boldsymbol{g}_k \boldsymbol{g}_k^T \mathbf{Q}^{-1} \boldsymbol{g}_k} \geq \frac{a}{A}$ then $1 - \gamma_k \leq 1 - \frac{a}{A}$. Therefore

$$E(\boldsymbol{x}_{k+1}) = (1 - \gamma_k)\, E(\boldsymbol{x}_k) \leq \left(1 - \frac{a}{A}\right) E(\boldsymbol{x}_k)$$

□

Quadratic case: Summary

- The previous theorem is very important for the *convergence* of the steepest decent algorithm
- The ratio $\kappa := \frac{A}{a} = \|\mathbf{Q}\|_2 \|\mathbf{Q}^{-1}\|_2 = \kappa(\mathbf{Q}) \geq 1$ is the so-called *condition number* of $\mathbf{Q}$
- Recall: $\|\mathbf{Q}\|_2 = \sqrt{\lambda_M(\mathbf{Q}^T\mathbf{Q})} = \sigma_M(\mathbf{Q})$. If $\mathbf{Q} \succ 0$ is symmetric then $\sigma_M(\mathbf{Q}) = \lambda_M(\mathbf{Q})$
- The term $1 - \frac{a}{A} = 1 - \frac{1}{\kappa}$ plays an important role in the convergence of the sequence $\{E(\boldsymbol{x}_k)\}$ (and therefore of the convergence of $\boldsymbol{x}_k$ to $\boldsymbol{x}^*$).

Quadratic case: Summary

- The method of steepest descent converges linearly with a ratio no greater than $1 - \frac{1}{\kappa}$.
- The smaller the value of κ, the smaller the relative value of $E(\boldsymbol{x}_{k+1})$ with respect to $E(\boldsymbol{x}_k)$ and therefore $\{E(\boldsymbol{x}_k)\}$ converges faster to 0.
- If $\kappa = 1$, i.e., the level sets are circulars and $A = a$, the algorithm converges in one iteration to the minimizer. If κ increases then the rate of convergence decreases.

Quadratic case: Steepest descent with exact line search

Lemma 3.2

In the steepest descent algorithm with exact line search, i.e. $\alpha_k = \frac{\boldsymbol{g}_k^T \boldsymbol{g}_k}{\boldsymbol{g}_k^T \mathbf{Q} \boldsymbol{g}_k}$, *if* $\boldsymbol{g}_k \neq \mathbf{0}$ *then* $\gamma_k = 1$ *iff* $\boldsymbol{g}_k$ *is an eigenvector of* $\mathbf{Q}$

Proof.

($\Leftarrow$) if $\boldsymbol{g}_k$ is an eigenvector of $\mathbf{Q}$ then $\mathbf{Q}\boldsymbol{g}_k = \lambda \boldsymbol{g}_k$ and $\mathbf{Q}^{-1}\boldsymbol{g}_k = \lambda^{-1}\boldsymbol{g}_k$ therefore

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

$\square$

Quadratic case: Steepest descent with exact line search

Proof.

($\Rightarrow$) if $\gamma_k = 1$ then $E(\boldsymbol{x}_{k+1}) = \frac{1}{2}(\boldsymbol{x}_{k+1} - \boldsymbol{x}^*)^T \mathbf{Q}(\boldsymbol{x}_{k+1} - \boldsymbol{x}^*) = 0$ therefore $\boldsymbol{x}_{k+1} = \boldsymbol{x}^*$. Hence,

$$\begin{aligned}
\boldsymbol{x}_{k+1} &= \boldsymbol{x}_k - \alpha_k \boldsymbol{g}_k \\
\boldsymbol{x}^* &= \boldsymbol{x}_k - \alpha_k \boldsymbol{g}_k \\
\mathbf{Q}\boldsymbol{x}^* &= \mathbf{Q}\boldsymbol{x}_k - \alpha_k \mathbf{Q}\boldsymbol{g}_k \\
\alpha_k \mathbf{Q}\boldsymbol{g}_k &= \mathbf{Q}\boldsymbol{x}_k - \boldsymbol{b} \\
\mathbf{Q}\boldsymbol{g}_k &= \frac{1}{\alpha_k}\boldsymbol{g}_k
\end{aligned}$$

and then $\boldsymbol{g}_k$ is an eigenvector of $\mathbf{Q}$.

$\square$

Quadratic case: Steepest descent with exact line search

Theorem 3.3

In the steepest decent algorithm with exact line search, the error norm $E(\cdot)$ satisfies $E(\boldsymbol{x}_{k+1}) \leq \left(\frac{A-a}{A+a}\right)^2 E(\boldsymbol{x}_k)$

Proof.

Using the Kantorovich inequality: $\gamma_k = \frac{(\boldsymbol{g}_k^T \boldsymbol{g}_k)^2}{\boldsymbol{g}_k^T \mathbf{Q} \boldsymbol{g}_k \boldsymbol{g}_k^T \mathbf{Q}^{-1} \boldsymbol{g}_k} \geq \frac{4aA}{(a+A)^2}$

then $1 - \gamma_k \leq 1 - \frac{4aA}{(a+A)^2} = \left(\frac{A-a}{A+a}\right)^2$. Therefore

$$E(\boldsymbol{x}_{k+1}) = (1 - \gamma_k)\, E(\boldsymbol{x}_k) \leq \left(\frac{A-a}{A+a}\right)^2 E(\boldsymbol{x}_k)$$

Note also that: $\left(\frac{A-a}{A+a}\right)^2 = \left(\frac{\kappa-1}{\kappa+1}\right)^2$ $\square$

Non-quadratic case

Theorem 3.4

Non-quadratic case Suppose f is defined on $\mathbb{R}^n$, has continuous second partial derivatives, and has a relative minimum at $\boldsymbol{x}^*$. Suppose further that the Hessian matrix of f, $\mathbf{H}(x^*)$, has smallest eigenvalue $a > 0$ and largest eigenvalue $A > 0$. If $\{\boldsymbol{x}_k\}$ is a sequence generated by the method of steepest descent that converges to $\boldsymbol{x}^*$, then the sequence of objective values $\{f(\boldsymbol{x}_k)\}$ converges to $f(\boldsymbol{x}^*)$ linearly with a convergence ratio no greater than $\left(\frac{A-a}{A+a}\right)^2$, i.e., for all k sufficiently large, we have

$$f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}^*) \leq \left(\frac{A-a}{A+a}\right)^2 [f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*)]$$