AARHUS
UNIVERSITET

Optimization
- Gradient Methods

Henrik Karstoft
@
Aarhus University, Department of Engineering

The gradient

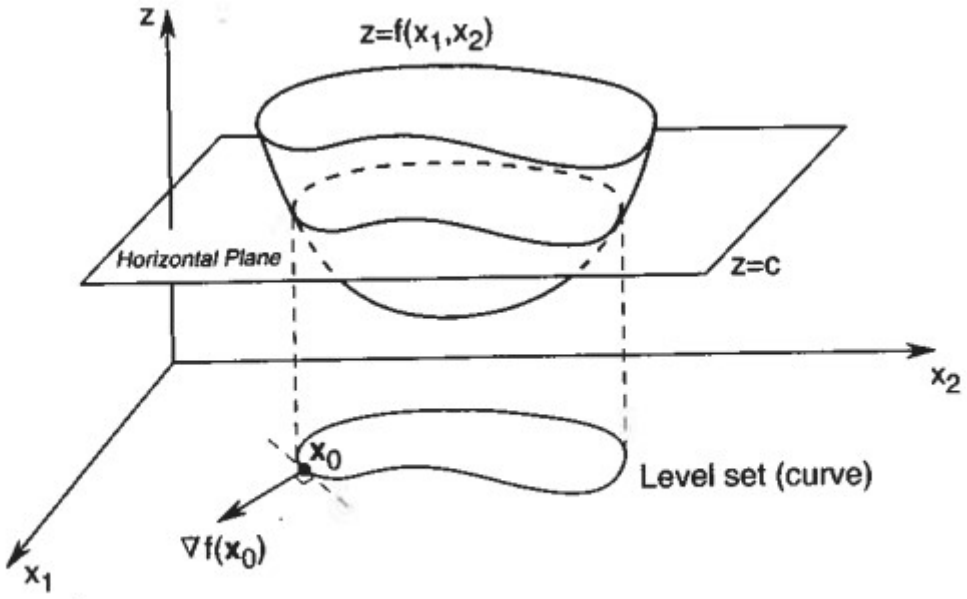

Figure 8.1 Constructing a level set corresponding to level c for f.

Gradient of $f : \mathbb{R}^n \to \mathbb{R}$ is a column vector

$$\nabla f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}$$

First derivative—row vector

$$Df(\boldsymbol{x}) = \nabla f(\boldsymbol{x})^\top = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \cdots & \frac{\partial f}{\partial x_n} \end{bmatrix}$$

The directional derivative of f at $\boldsymbol{x} \in \mathbb{R}^n$ in the direction $\boldsymbol{d}$ is denoted

$$\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{d}}$$

$$\begin{aligned} \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{d}} &= \lim_{\alpha \to 0} \frac{f(\boldsymbol{x} + \alpha \boldsymbol{d}) - f(\boldsymbol{x})}{\alpha} \\ &= \frac{d}{d\alpha} f(\boldsymbol{x} + \alpha \boldsymbol{d}) \Big|_{\alpha = 0} \\ &= \begin{bmatrix} \quad Df(\boldsymbol{x}) \quad \end{bmatrix} \begin{bmatrix} \quad \\ \boldsymbol{d} \\ \quad \end{bmatrix} \end{aligned}$$

The Method og Steepest Decent

Finding best α:

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

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

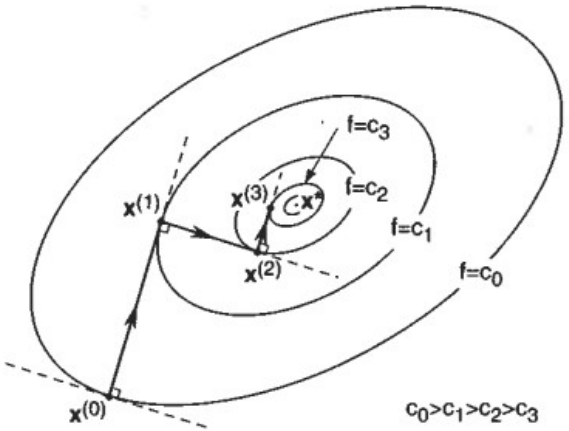

Figure 8.2 Typical sequence resulting from the method of steepest descent.

The Method og Steepest Decent

Example 2.2. This is a 2-dimensional minimization example, illustrated on the front page. A tourist has lost his way in a hilly country. It is a foggy day so he cannot see far and he has no map. He knows that his rescue is at the bottom of a nearby valley. As tools he has a compass and his sense of balance, which can tell him about the local slope of the ground.

In order not to walk in circles he decides to use straight strides, ie with constant compass bearing. From what his feet tell him about the local slope he chooses a direction and walks in that direction until his feet tell him that he is on an uphill slope.

Now he has to decide on a new direction and he starts his next stride. Let us hope that he reaches his goal in the end. ■

Example 2.3. Let us return to our tourist who is lost in the fog in a hilly country. By experimenting with his compass he can find out that "half" the compass bearings give strides that start uphill and that the "other half" gives strides that start downhill. Between the two halves are two strides which start off going neither uphill or downhill. These form the tangent to the level curve corresponding to his position. ■

The Method og Steepest Decent

Proposition 8.2 *If* $\{\boldsymbol{x}^{(k)}\}_{k=0}^{\infty}$ *is the steepest descent sequence for* $f : \mathbb{R}^n \to \mathbb{R}$ *and if* $\nabla f(\boldsymbol{x}^{(k)}) \neq \boldsymbol{0}$, *then* $f(\boldsymbol{x}^{(k+1)}) < f(\boldsymbol{x}^{(k)})$. □

Stopping criteria:

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

$$|f(\boldsymbol{x}^{(k+1)}) - f(\boldsymbol{x}^{(k)})| < \varepsilon,$$

$$\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\| < \varepsilon.$$

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

$$\frac{\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\|}{\|\boldsymbol{x}^{(k)}\|} < \varepsilon.$$

The Method og Steepest Decent

Example 8.1 We use the method of steepest descent 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 $\boldsymbol{x}^{(0)} = [4, 2, -1]^\mathsf{T}$. We perform three iterations.

We find that

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

Hence,

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

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

$$\begin{aligned}\alpha_0 &= \arg\min_{\alpha \geq 0} f(\boldsymbol{x}^{(0)} - \alpha \nabla f(\boldsymbol{x}^{(0)})) \\ &= \arg\min_{\alpha \geq 0} (0 + (2 + 2\alpha - 3)^2 + 4(-1 - 1024\alpha + 5)^4) \\ &= \arg\min_{\alpha \geq 0} \phi_0(\alpha).\end{aligned}$$

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

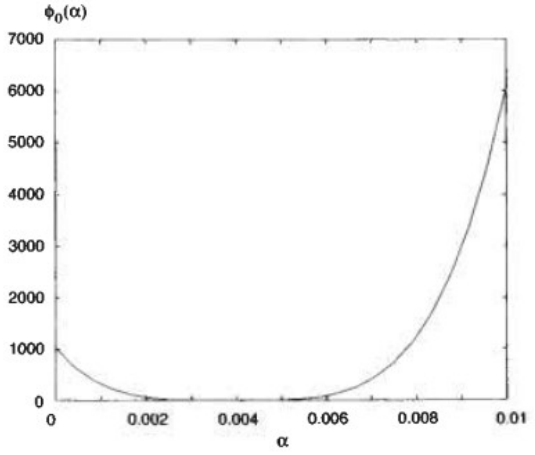

Figure 8.3 Plot of $\phi_0(\alpha)$ versus α.

Review of eigenvalues quadratic forms

Quadratic form

$$Q(x) = x^T A x$$

where $x \in R^n$, A is symmetric $n \times n$ matrix

Eigenvalues for a A $n \times n$ matrix are roots in characteristic the polynomial $\chi(\lambda)$of degree n:

$$\chi(\lambda) = \det(A - \lambda I)$$

where "det" is the determinat of A, and I is the identity matrix

The Method og Steepest Decent, quadratic

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

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

$$\begin{aligned}\alpha_k &= \arg\min_{\alpha \geq 0} f(\boldsymbol{x}^{(k)} - \alpha \boldsymbol{g}^{(k)}) \\ &= \arg\min_{\alpha \geq 0} \left(\frac{1}{2}(\boldsymbol{x}^{(k)} - \alpha \boldsymbol{g}^{(k)})^\mathsf{T} \boldsymbol{Q} (\boldsymbol{x}^{(k)} - \alpha \boldsymbol{g}^{(k)}) - (\boldsymbol{x}^{(k)} - \alpha \boldsymbol{g}^{(k)})^\mathsf{T} \boldsymbol{b} \right).\end{aligned}$$

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \frac{\boldsymbol{g}^{(k)\mathsf{T}} \boldsymbol{g}^{(k)}}{\boldsymbol{g}^{(k)\mathsf{T}} \boldsymbol{Q} \boldsymbol{g}^{(k)}} \boldsymbol{g}^{(k)},$$

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

Analysis of gradient methods, 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*

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

□

$$\boldsymbol{g}^{(k)} = \nabla f(\boldsymbol{x}^{(k)}) = \boldsymbol{Q}\boldsymbol{x}^{(k)} - \boldsymbol{b}. \qquad \gamma_k = \alpha_k \frac{\boldsymbol{g}^{(k)\mathsf{T}} \boldsymbol{Q} \boldsymbol{g}^{(k)}}{\boldsymbol{g}^{(k)\mathsf{T}} \boldsymbol{Q}^{-1} \boldsymbol{g}^{(k)}} \left(2 \frac{\boldsymbol{g}^{(k)\mathsf{T}} \boldsymbol{g}^{(k)}}{\boldsymbol{g}^{(k)\mathsf{T}} \boldsymbol{Q} \boldsymbol{g}^{(k)}} - \alpha_k \right).$$

Theorem 8.2 *In the steepest descent algorithm, we have $\boldsymbol{x}^{(k)} \to \boldsymbol{x}^*$ for any $\boldsymbol{x}^{(0)}$.* □

Analysis of gradient methods, 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*

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

□

Example 8.4 Let the function f be given by

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

We wish to find the minimizer of f using a fixed-step-size gradient algorithm

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

where $\alpha \in \mathbb{R}$ is a fixed step size.

Gradient methods, rate of convergens

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

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

□

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

called the *condition number* of $\boldsymbol{Q}$. Then, it follows from Theorem 8.4 that

$$V(\boldsymbol{x}^{(k+1)}) \leq \left(1 - \frac{1}{r}\right) V(\boldsymbol{x}^{(k)}).$$