Optimization Techniques
Lecture 4
Hale Gonce Köçken

Iterative methods for unconstrained or set-constrained optimization

Gradient methods

The method of steepest descent

The method of steepest descent for a quadratic function

Iterative methods for unconstrained or set-constrained optimization

- In our last lecture, we have seen a theoretical basis for the solution of nonlinear unconstrained problems.
- Suppose that one is confronted with a highly nonlinear function of 20 variables.
- Then the FONC requires the solution of 20 nonlinear simultaneous equations for 20 variables.
- These equations, being nonlinear, will normally have multiple solutions. In addition, we would have to compute 210 second derivatives (provided f in C^2) to use the SOSC.
- Thus, we will now concern with iterative methods of solving such problems.

Gradients Methods

Let $\boldsymbol{x}^{(0)}$ be a starting point, and consider the point $\boldsymbol{x}^{(0)} - \alpha \nabla f(\boldsymbol{x}^{(0)})$.

Then, by Taylor's theorem we obtain

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

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

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

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

To formulate an algorithm that implements the above idea, suppose that we are given a point $\boldsymbol{x}^{(k)}$. To find the next point $\boldsymbol{x}^{(k+1)}$, we start at $\boldsymbol{x}^{(k)}$ and move by an amount $-\alpha_k \nabla f(\boldsymbol{x}^{(k)})$, where α_k is a positive scalar called the *step size*. The above procedure leads to the following iterative algorithm:

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

Remember that the function f increases more in the direction of the gradient than in any other direction.

A numerical example

The gradient vector of a scalar function $f(x_1, x_2, \cdots, x_n)$ is defined as a column vector

$$\nabla f = \left[\frac{\partial f}{\partial x_1} \quad \frac{\partial f}{\partial x_2} \quad \cdots \quad \frac{\partial f}{\partial x_n} \right]^T = \mathbf{c}$$

For example

$$f(x_1, x_2) = 25x_1^2 + x_2^2 \quad \Longrightarrow \quad \mathbf{c} = \nabla f = \begin{bmatrix} 2(25)x_1^* \\ 2x_2^* \end{bmatrix} = \begin{bmatrix} 2(25)(.6) \\ 2(4) \end{bmatrix} = \begin{bmatrix} 30 \\ 8 \end{bmatrix}$$

at the point $x_1^* = .6, x_2^* = 4$

The normalized gradient vector

$$\bar{\mathbf{c}} = \frac{\mathbf{c}}{\sqrt{\mathbf{c}^T \mathbf{c}}}$$

For example, at the point $x_1^* = .6, x_2^* = 4$

$$\Longrightarrow \quad \bar{\mathbf{c}} = \frac{1}{\sqrt{30^2 + 8^2}} \begin{bmatrix} 30 \\ 8 \end{bmatrix} = \begin{bmatrix} .96625 \\ .2577 \end{bmatrix}$$

The gradient vector represents a direction of maximum rate of increase for the function $f(\mathbf{x})$ at $\mathbf{x}^*$. For example,

$f(.6,4) = 25(.6)^2 + 4^2 = 25$ If we increase $\mathbf{x}$ in the direction $\bar{\mathbf{c}}$ by a step size of $\alpha = .5$

$$\mathbf{x}^{(1)} = \mathbf{x}^{(0)} + \alpha\bar{\mathbf{c}} = \begin{bmatrix} .6 \\ 4 \end{bmatrix} + .5\begin{bmatrix} .96625 \\ .2577 \end{bmatrix} = \begin{bmatrix} 1.083125 \\ 4.12885 \end{bmatrix}$$

The function value becomes $f(\mathbf{x}^{(1)}) = 25(1.083125)^2 + (4.122885)^2 = 46.327$

If we move in a direction $\begin{bmatrix} 1 & 0 \end{bmatrix}^T$

$$\mathbf{x}^{(1)} = \mathbf{x}^{(0)} + \alpha\bar{\mathbf{c}} = \begin{bmatrix} .6 \\ 4 \end{bmatrix} + .5\begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1.1 \\ 4 \end{bmatrix}$$

The function value becomes

$f(\mathbf{x}^{(1)}) = 25(1.1)^2 + (4)^2 = 46.25$

If we move in a direction $\begin{bmatrix} 0 & 1 \end{bmatrix}^T$

$$\mathbf{x}^{(1)} = \mathbf{x}^{(0)} + \alpha\bar{\mathbf{c}} = \begin{bmatrix} .6 \\ 4 \end{bmatrix} + .5\begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} .6 \\ 4.5 \end{bmatrix}$$

The function value becomes

$f(\mathbf{x}^{(1)}) = 25(.6)^2 + (4.5)^2 = 29.25$

We can see that moving along the gradient direction results in the maximum increase in the function.

The maximum rate of change of $f(\mathbf{x})$ at any point $\mathbf{x}^*$ is the magnitude of the gradient vector given by $\|\mathbf{c}\| = \sqrt{\mathbf{c}^T\mathbf{c}}$

HW:Find the maximum rate of change of $f(x,y) = xe^{-y} + 3y$ at the point (1,0) and the direction in which it occurs.

Solution: The maximum rate of change occurs in the direction of the gradient vector $\nabla f(x,y)$, and the maximum rate of change is at $|\nabla f(x,y)|$.

$$\nabla f(x,y) = < e^{-y}, -xe^{-y} + 3 >$$

$$\nabla f(1,0) = < 1, 2 >$$

$$\left|\nabla f(1,0)\right| = \left| < 1, 2 > \right| = \sqrt{5}$$

Thus the maximum rate of change is $\sqrt{5}$ in the direction $\nabla f(1,0) = [1 \quad 2]^T$.

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.

The method of steepest descent

The method of 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. Specifically, α_k is chosen to minimize $\phi_k(\alpha) \triangleq f(\boldsymbol{x}^{(k)} - \alpha \nabla f(\boldsymbol{x}^{(k)}))$. In other words,

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

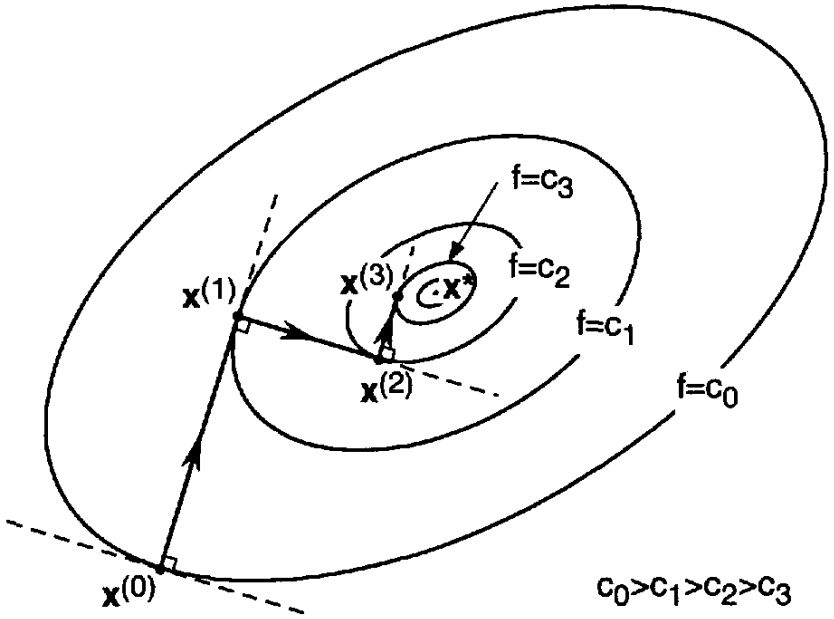


Figure 8.2 Typical sequence resulting from the method of steepest descent

In summary, the steepest descent algorithm can be given as follows:

Step 1: Choose $x^{(0)}$.

Step 2: Calculate $\nabla f\left(x^{(i)}\right)$. If $\nabla f\left(x^{(i)}\right) = 0$, stop.

Step 3: Determine the next point $x^{(i+1)}$ with

$x^{(i+1)} = x^{(i)} - \propto_i \nabla f\left(x^{(i)}\right)$ where $\propto_i$ is choosen to minimize the function $f\left(x^{(i)} - \propto_i \nabla f\left(x^{(i)}\right)\right)$. Set $i = i + 1$ and go to Step 2.

Example : Let $f(x, y) = x^2 + y^2$. Find a minimizer of f with the method of steepest descent assuming the initial point as (1,1).

The method of steepest descent for a quadratic function

Let us now see what the method of steepest descent does with 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 $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$ is a symmetric positive definite matrix, $\boldsymbol{b} \in \mathbb{R}^n$, and $\boldsymbol{x} \in \mathbb{R}^n$. The unique minimizer of f can be found by setting the gradient of f to zero, where

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

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

The Hessian of f is $\boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{Q} = \boldsymbol{Q}^T > 0$.

To simplify the notation we write $g^{(k)} = \nabla f(x^{(k)})$.
Then, the steepest descent algorithm for the quadratic function can be represented as

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

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

In the quadratic case, we can find an explicit formula for α_k. We proceed as follows. Assume $g^{(k)} \neq 0$, for if $g^{(k)} = 0$, then $x^{(k)} = x^*$ and the algorithm stops. Because $\alpha_k \geq 0$ is a 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 $\alpha \boldsymbol{g}^{(k)T}\boldsymbol{Q}\boldsymbol{g}^{(k)} = (\boldsymbol{x}^{(k)T}\boldsymbol{Q} - \boldsymbol{b}^T)\boldsymbol{g}^{(k)}$. But

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

Hence,

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

In summary, the method of steepest descent for the quadratic takes the form

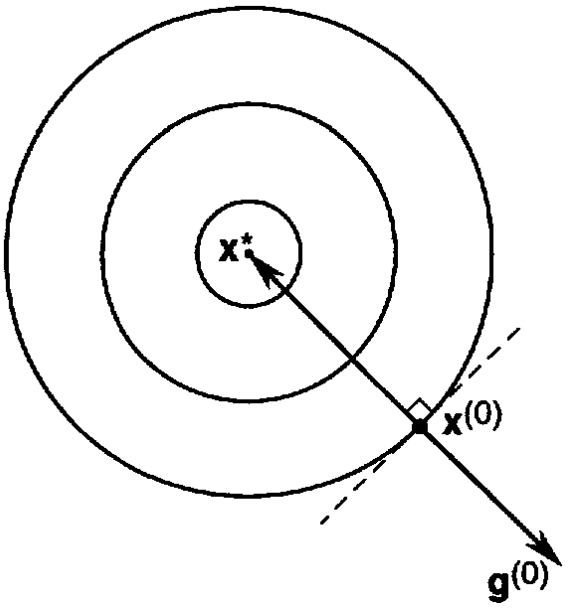
$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \left(\frac{\boldsymbol{g}^{(k)T}\boldsymbol{g}^{(k)}}{\boldsymbol{g}^{(k)T}\boldsymbol{Q}\boldsymbol{g}^{(k)}}\right)\boldsymbol{g}^{(k)},$$

where

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

Example 8.2 Let

$$f(x_1, x_2) = x_1^2 + x_2^2.$$



Then, starting from an arbitrary initial point $\boldsymbol{x}^{(0)} \in \mathbb{R}^2$ we arrive at the solution $\boldsymbol{x}^* = \mathbf{0} \in \mathbb{R}^2$ in only one step. See Figure 8.6.

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 (see Figure 8.7). This example illustrates a major drawback in the steepest descent method. More sophisticated methods that alleviate this problem are discussed in subsequent chapters.

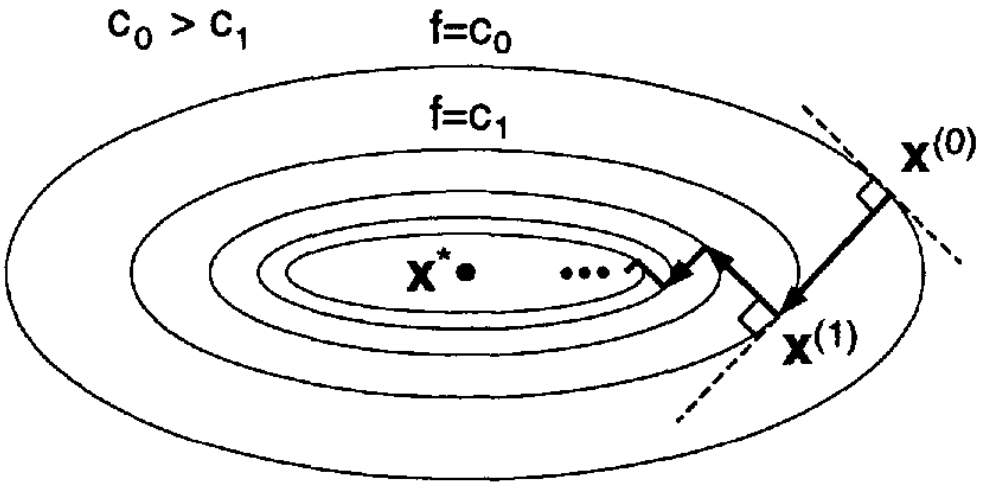


Figure 8.7 Steepest descent method in search for minimizer in a narrow valley

The condition $\nabla f(\boldsymbol{x}^{(k+1)}) = \mathbf{0}$, 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 stopping criterion is to check if the norm $\|\nabla f(\boldsymbol{x}^{(k)})\|$ of the gradient is less than a prespecified threshold, in which case we stop.

Alternatively, we may compute the absolute difference $|f(\boldsymbol{x}^{(k+1)}) - f(\boldsymbol{x}^{(k)})|$ between objective function values for every two successive iterations, and if the difference is less than some prespecified threshold, then we stop; that is, we stop when

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

where $\varepsilon > 0$ is a prespecified threshold.

Yet another alternative is to compute the norm $\|x^{(k+1)} - x^{(k)}\|$ of the difference between two successive iterates, and we stop if the norm is less than a prespecified threshold:

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

Alternatively, we may check "relative" values of the above quantities; for example,

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

or

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

Note that the above stopping criteria are relevant to all the iterative algorithms we discuss in this part.

With the stopping criteria modification, the steepest descent algorithm can be given as follows:

Step 1: Choose $x^{(0)}$.

Step 2: Calculate $\nabla f\left(x^{(i)}\right)$. If $\nabla f\left(x^{(i)}\right) = 0$ or the selected stopping criteria is satisfied, then stop.

Step 3: Determine the next point $x^{(i+1)}$ with $x^{(i+1)} = x^{(i)} - \propto_i \nabla f\left(x^{(i)}\right)$ where $\propto_i$ is choosen to minimize the function $f\left(x^{(i)} - \propto_i \nabla f\left(x^{(i)}\right)\right)$. Set $i = i + 1$ and go to Step 2.

Example : Let $f(x,y) = x^2 + y^2 + xy - 3x$. Find a minimizer of f with

a) Analytical method

b) The method of steepest descent with the following stopping criteria

i. $\|\nabla f(x)\| < 0.8,$

ii. $\left|f\left(x^{(k+1)}\right) - f\left(x^{(k)}\right)\right| < 0.2.$

Point	f value	g_k	α_k
$x^{(0)} = (0,0)^T$	0	$\begin{bmatrix} -3 \\ 0 \end{bmatrix}$	0.5
$x^{(1)} = (3/2\,, 0)^T$	-2.25	$\begin{bmatrix} 0 \\ 3/2 \end{bmatrix}$	0.5
$x^{(2)} = (3/2\,, -3/4)^T$	-2.8116	$\begin{bmatrix} -0.75 \\ 0 \end{bmatrix}$	0.5
$x^{(3)} = (1.875, -0.75)^T$	-2.9531	$\begin{bmatrix} 0 \\ 0.375 \end{bmatrix}$	0.5
$x^{(4)} = (1.875, -0.9375)^T$	-2.9883	$\begin{bmatrix} -0.1875 \\ 0 \end{bmatrix}$	0.5
$x^{(5)} = (1.9688, -0.9375)^T$	-2.9971	$\begin{bmatrix} 0.001 \\ 0.0938 \end{bmatrix}$	0.4995
$x^{(6)} = (1.9688, -0.9844)^T$	-2.9993	$\begin{bmatrix} -0.0468 \\ 0 \end{bmatrix}$	0.5
$x^{(7)} = (1.9922, -0.9844)^T$	-2.9998	$\begin{bmatrix} 0 \\ 0.0234 \end{bmatrix}$	0.5

Point	f value	g_k	α_k
$x^{(8)} = (1.9922, -0.9961)^T$	-3	$\begin{bmatrix} -0.0117 \\ 0 \end{bmatrix}$	0.5
$x^{(9)} = (1.9981, -0.9961)^T$	-3	$\begin{bmatrix} 0.001 \\ 0.0059 \end{bmatrix}$	0.4917
$x^{(10)} = (1.9981, -0.9990)^T$	-3	$\begin{bmatrix} -0.0028 \\ 0.0001 \end{bmatrix}$	0.5185
$x^{(11)} = (1.9996, -0.9991)^T$	-3	$\begin{bmatrix} 0.001 \\ 0.0014 \end{bmatrix}$	0.4668
$x^{(12)} = (1.9996, -0.9998)^T$	-3	$\begin{bmatrix} 0.001 \\ -0.6 \end{bmatrix}$	0.5
$x^{(13)} = (1.9999, -0.9998)^T$	-3	$\begin{bmatrix} 0.001 \\ 0.0003 \end{bmatrix}$	0.5
$x^{(14)} = (1.9999, -1)^T$	-3	$\begin{bmatrix} -0.0002 \\ -0.0001 \end{bmatrix}$	0.3571
$x^{(15)} = (2, -1)^T$	-3	$\begin{bmatrix} 0 \\ 0 \end{bmatrix}$	

Quadratic Function

$\theta \in R^{n\times n}$ is a positive definite matrix and $x \in R^n$;

$$f(x) = \frac{1}{2} x'\theta x - x'b$$

For $n = 2$, $\theta \in R^{2\times 2}$, $x \in R^2$

$$f(x) = \frac{1}{2}\begin{bmatrix} x & y \end{bmatrix}\begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} - \begin{bmatrix} x & y \end{bmatrix}\begin{bmatrix} e \\ f \end{bmatrix}$$

$$= \frac{1}{2}\begin{bmatrix} ax + cy & bx + dy \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} - \begin{bmatrix} ex + fy \end{bmatrix}$$

$$= \frac{1}{2}\left[ax^2 + cxy + bxy + dy^2 \right] - \begin{bmatrix} ex + fy \end{bmatrix}$$

$$f(x) = \frac{1}{2} ax^2 + \frac{1}{2} dy^2 + \frac{1}{2}(b + c)xy - ex + fy$$

$$\theta = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

a : twice the coefficient of x^2

d : twice the coefficient of y^2

$b + c$: twice the coefficient of xy

e : reverse sign of the coefficient of x

f : reverse sign of the coefficient of y

The derivative of a quadratic function

$$\left.\begin{aligned}\frac{\partial f}{\partial x} &= ax+\frac{(b+c)}{2}y-e\\ \frac{\partial f}{\partial y} &= dy+\frac{(b+c)}{2}x-f\end{aligned}\right\}\Rightarrow \nabla f(x)=\begin{bmatrix} a & \frac{(b+c)}{2}\\ \frac{(b+c)}{2} & d\end{bmatrix}\begin{bmatrix}x\\ y\end{bmatrix}-\begin{bmatrix}e\\ f\end{bmatrix}$$

Since θ is a symmetrical matrix, then $b=c$. Thus, we have

$$\nabla f(x)=\begin{bmatrix}a & b\\ b & d\end{bmatrix}\begin{bmatrix}x\\ y\end{bmatrix}-\begin{bmatrix}e\\ f\end{bmatrix}=\theta x-b \quad \text{and also} \quad \nabla^2 f=\begin{bmatrix}a & b\\ b & d\end{bmatrix}=\theta .$$