UP: Solution algorithms

A. J. Conejo, R. Sioshansi, 2016
THE OHIO STATE UNIVERSITY

What

1. Steepest Descent
2. Newton
3. Extended Newton
4. Coordinate Descent
5. Quadratic Case
6. Scaling

Steepest Descent

Steepest Descent

$$\min_{x \in \mathbb{R}^n} \; f(x)$$

$$f(x) : \mathbb{R}^n \rightarrow \mathbb{R}$$

Steepest Descent

Given an unconstrained nonlinear optimization problem and an incumbent point, x^k,

the **steepest-descent** search direction is:

$$d^k = -\nabla f(x^k).$$

Steepest Descent

$$x^{k+1} \leftarrow x^k + \alpha^k d^k$$

$$d^k = -\nabla f(x^k)$$

$$\boxed{x^{k+1} \leftarrow x^k - \alpha^k \nabla f(x^k)}$$

Steepest Descent: Example

Steepest Descent: Example

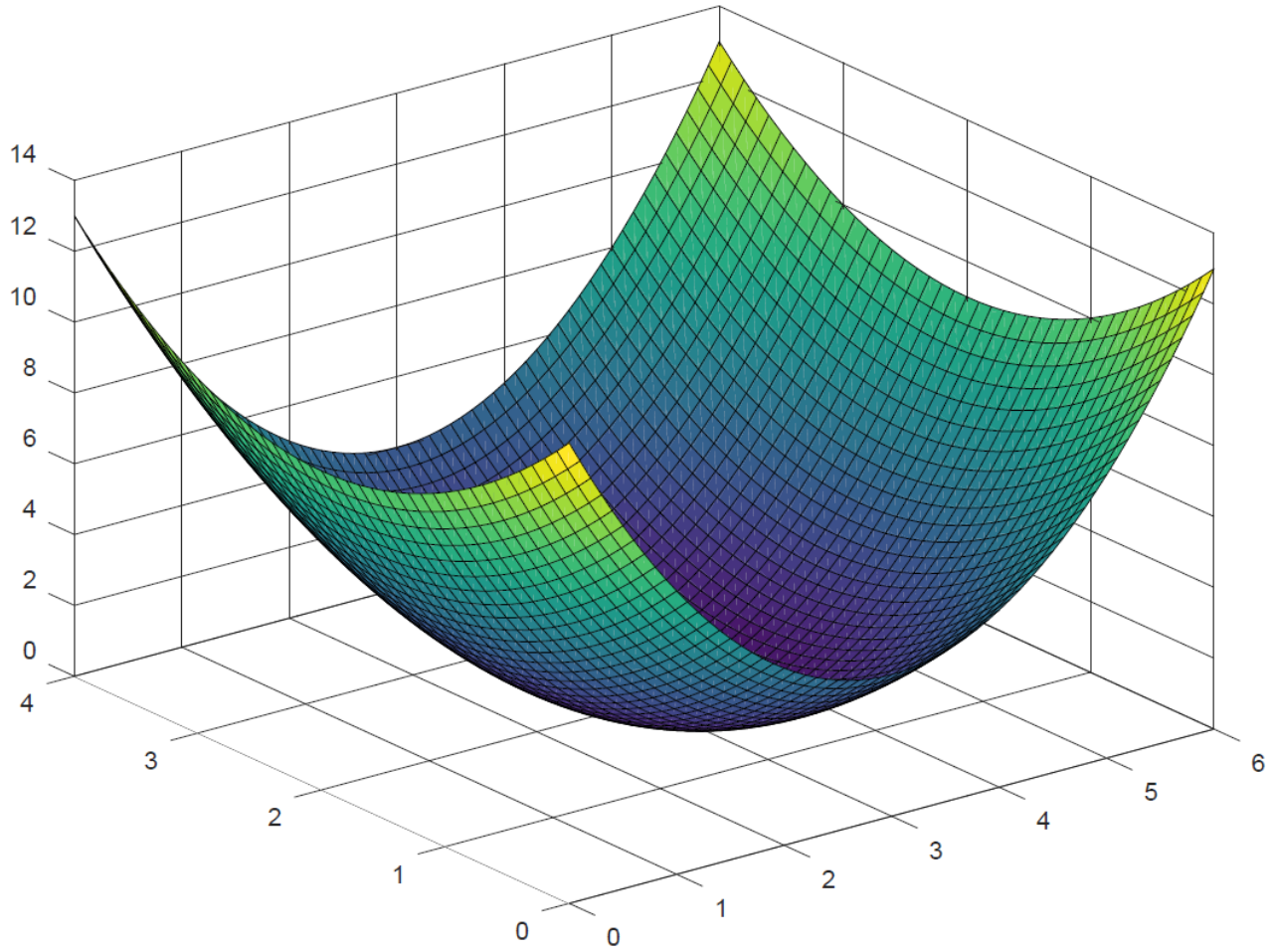
Consider:

$$\min_{x} \; f(x) = (x_1 - 3)^2 + (x_2 - 2)^2$$

Steepest Descent: Example

```
home
x = 0:0.1:6;
y = 0:0.1:4;
[X,Y] = meshgrid(x,y);
Z = (X.-3).^2+(Y.-2).^2;
figure
surf(X,Y,Z)
```

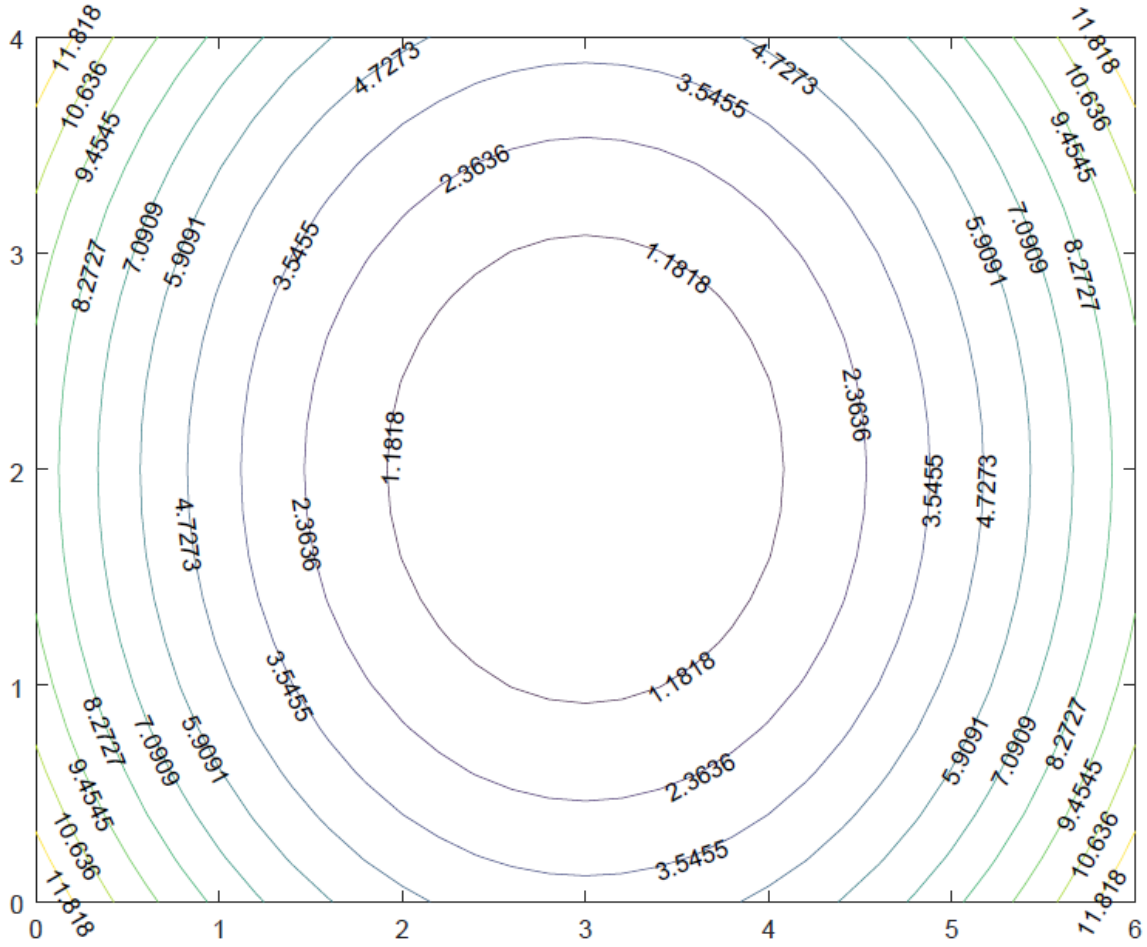
Steepest Descent: Example



Steepest Descent: Example

```
home
x = 0:0.2:6;
y = 0:0.2:4;
[X,Y] = meshgrid(x,y);
Z = (X.-3).^2+(Y.-2).^2;
figure
contour(X,Y,Z,'ShowText','on')
```

Steepest Descent: Example



Steepest Descent: Example

Starting from the point, $x^0 = (1,1)^\top$, we wish to find the steepest-descent direction. To do so, we first compute the gradient of the objective function as:

$$\nabla f(x) = \begin{pmatrix} 2(x_1 - 3) \\ 2(x_2 - 2) \end{pmatrix}.$$

Steepest Descent: Example

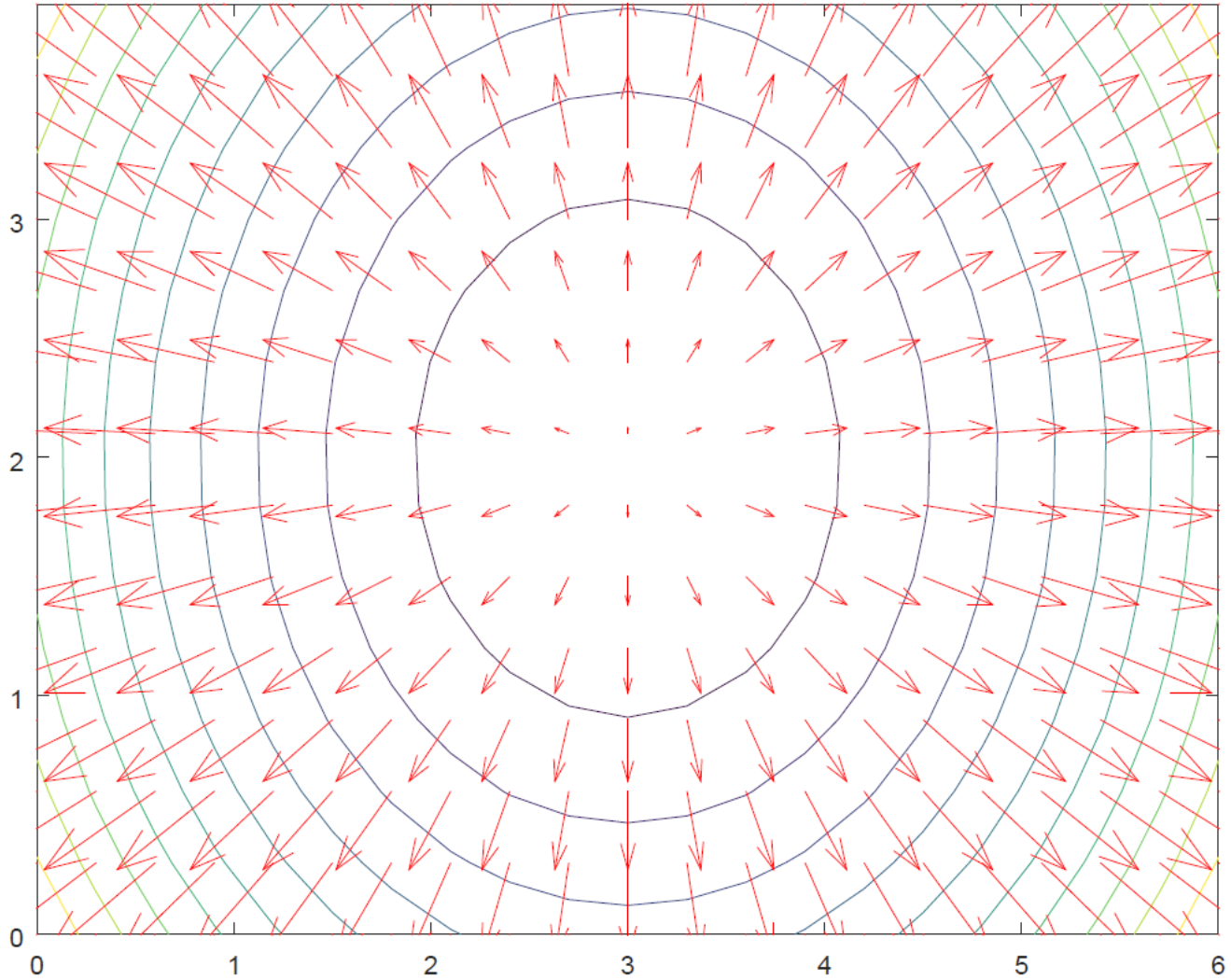
Thus, the steepest-descent direction is:

$$d^0 = -\nabla f(x^0) = \begin{pmatrix} 4 \\ 2 \end{pmatrix}.$$

Steepest Descent: Example

```
home
x = 0:0.2:6;
y = 0:0.2:4;
[X,Y] = meshgrid(x,y);
Z = (X.-3).^2+(Y.-2).^2;
figure
contour(X,Y,Z,'ShowText','off')
[U,V] = gradient(Z);
hold on
quiver(X,Y,U,V,2,'color','red')
hold off
```

Steepest Descent: Example



Steepest Descent: Example

To conduct an exact line search we solve the following minimization problem:

$$\begin{aligned}
\min_{\alpha^0} \; f\left(x^0 + \alpha^0 d^0\right) &= f\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix} + \alpha^0 \begin{pmatrix} 4 \\ 2 \end{pmatrix}\right) \\
&= f\left(\begin{pmatrix} 1 + 4\alpha^0 \\ 1 + 2\alpha^0 \end{pmatrix}\right) \\
&= (4\alpha^0 - 2)^2 + (2\alpha^0 - 1)^2.
\end{aligned}$$

Steepest Descent: Example

To solve this unconstrained minimization, we use the FONC which is:

$$\frac{d}{d\alpha^0}\left[(4\alpha^0-2)^2+(2\alpha^0-1)^2\right]=8(4\alpha^0-2)+4(2\alpha^0-1)=0,$$

which gives $\alpha^0=1/2$.

Steepest Descent: Example

We further have that:

$$\frac{d^2}{d\alpha^{0^2}}\left[(4\alpha^0-2)^2+(2\alpha^0-1)^2\right]=40>0,$$

meaning that this value of α^0 is a global minimum. Thus, our new point is:

$$x^1=x^0+\alpha^0 d^0=(3,2)^\top.$$

Steepest Descent: Example

From this new incumbent point we now conduct another iteration. The steepest descent direction is:

$$d^1 = -\nabla f(x^1) = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

meaning that we are at a stationary point. Thus, we terminate the algorithm at this point.

Steepest Descent: Direction of Descent

Steepest Descent: Direction of Descent

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

$$x^{k+1} = x^k + m^k$$

where

$$m^k = -\alpha^k \nabla f(x^k)$$

Steepest Descent: Direction of Descent

$$f(x^k - \alpha^k \nabla f(x^k)) =$$

$$f(x^k + m^k) \approx \qquad \Leftarrow \quad m^k = -\alpha^k \nabla f(x^k)$$

Taylor $\Rightarrow$

$$f(x^k) + \nabla f(x^k)^\top m^k =$$

$$f(x^k) - \alpha^k \nabla f(x^k)^\top \nabla f(x^k) =$$

$$f(x^k) - \alpha^k [\nabla f(x^k)]^2$$

Descent!

Generalized Steepest Descent

Generalized Steepest Descent

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

$$f(x^k - \alpha^k D^k \nabla f(x^k)) =$$

$$f(x^k + m^k) \approx$$

Taylor

$$f(x^k) + \nabla f(x^k)^\top m^k =$$

$$f(x^k) - \alpha^k \nabla f(x^k)^\top D^k \nabla f(x^k)$$

$$m^k = -\alpha^k D^k \nabla f(x^k)$$

Descent if D^k is definite positive!

Generalized Steepest Descent

To be analyzed later on

$$\text{Steepest Descent: } D^k = I$$

$$\text{Diagonally-Scalled Steepest Descent: } D^k = \begin{bmatrix} d_{11}^k & & \\ & \ddots & \\ & & d_{nn}^k \end{bmatrix}$$

$$\text{Newton: } D^k = [\nabla^2 f(x^k)]^{-1}$$

$$\text{Compute } [\nabla^2 f(x^k)]^{-1} \text{ and use it at } k,\ k+1,\ k+2...$$

$$\text{Discreetely approximate } [\nabla^2 f(x^k)]^{-1}$$

Newton

Newton

Newton's method takes a fundamentally different approach compared to steepest descent in finding a search direction.

Instead of looking for a direction that minimizes the objective, Newton's method uses the knowledge that only stationary points can be minima of unconstrained nonlinear optimization problems.

Newton

Thus, the underlying premise of Newton's method is that we choose a search direction to make

$$\nabla f(x^k + d) = 0.$$

Newton

To derive the Newton direction, we use Taylor's theorem:

$$\nabla f(x^k + d) \approx \nabla f(x^k) + \nabla^2 f(x^k) d.$$

Newton

If we set this approximation equal to zero, we have:

$$\nabla f(x^k) + \nabla^2 f(x^k) d = 0,$$

which gives:

$$d = -\left[\nabla^2 f(x^k)\right]^{-1} \nabla f(x^k),$$

so long as $\nabla^2 f(x^k)$ is invertible.

This is stated in the following Newton's Method Rule.

Newton

Given an unconstrained nonlinear optimization problem and an incumbent point, x^k, the **Newton's method** search direction is:

$$d^k = -\left[\nabla^2 f(x^k)\right]^{-1} \nabla f(x^k),$$

so long as $\nabla^2 f(x^k)$ is invertible.

Otherwise, the Newton's method search direction is not defined.

Newton

$$x^{k+1} \leftarrow x^k - \left[\nabla^2 f(x^k)\right]^{-1} \nabla f(x^k)$$

Newton: Example

Consider the unconstrained nonlinear optimization problem:

$$\min_{x} \; f(x) = (x_1 - 3)^2 + (x_2 - 2)^2,$$

Newton: Example

We already know that the gradient of the objective function is:

$$\nabla f(x) = \begin{pmatrix} 2(x_1 - 3) \\ 2(x_2 - 2) \end{pmatrix}.$$

Newton: Example

We can further compute the Hessian as:

$$\nabla^2 f(x) = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}.$$

Newton: Example

Because the Hessian is invertible, we can compute the Newton direction as:

$$d^0 = -\left[\nabla^2 f(x^0)\right]^{-1} \nabla f(x^0) = \begin{pmatrix} 2 \\ 1 \end{pmatrix}.$$

Newton: Example

We conduct one iteration of Newton's method with an exact line search stating from the point $x^0 = (1, 1)^\top$

Newton: Example

Is this line search needed?

Using this direction, we next conduct an exact line search by solving the following minimization problem:

$$\min_{\alpha^0} \; f\left(x^0 + \alpha^0 d^0\right) = f\left(\begin{pmatrix} 1 + 2\alpha^0 \\ 1 + 1\alpha^0 \end{pmatrix}\right),$$

Newton: Example

which gives $\alpha^0 = 1$ as a solution. We can also easily show that $f(\cdot)$ is convex in α^0, thus we know that $\alpha^0 = 1$ is a global minimum.

This means our new point, when we use Newton's method, is also $x^1 = (3, 2)^\top$.

Newton: not necessarily a decent direction!

Consider the unconstrained nonlinear optimization problem:

$$\min_{x} \; f(x) = -(x_1 - 3)^2 - 2(x_2 - 2)^2.$$

Newton: not necessarily a decent direction!

Starting from the point, $x^0 = (1,1)^\top$ we use Newton's method with a pure step size to solve the problem. To find the Newton direction we first compute the gradient and Hessian of the objective function, which are:

$$\nabla f(x) = \begin{pmatrix} -2(x_1 - 3) \\ -4(x_2 - 2) \end{pmatrix},$$

Newton: not necessarily a decent direction!

and:

$$\nabla^2 f(x) = \begin{bmatrix} -2 & 0 \\ 0 & -4 \end{bmatrix}.$$

Newton: not necessarily a decent direction!

Thus, the Newton direction is $d^0 = (2,1)^\top$. A pure step size means that $\alpha^0 = 1$, meaning that our new point is $x^1 = (3,2)^\top$. If we attempt to conduct another iteration from this incumbent point we have $\nabla f(x^1) = (0,0)^\top$, meaning that we are at a stationary point and the algorithm terminates.

Newton: not necessarily a decent direction!

Note, however, that after conducting this iteration the objective function has gotten worse. We went from an objective function value of $f(x^0) = -6$ to $f(x^1) = 0$. Indeed, it is easy to show that the stationary point we found after a single iteration using the Newton direction is a local and global maximum of $f(\cdot)$.

Steepest Descent vs. Newton: Example

Steepest Descent vs. Newton: Example

Consider the unconstrained nonlinear optimization problem:

$$\min_{x} \; f(x) = 10(x_1 - 3)^2 + 2(x_2 - 2)^2.$$

Steepest Descent vs. Newton: Example

Starting from the point, $x^0 = (1,1)^\top$ we first conduct one iteration of steepest descent with an exact line search. To do so, we compute the gradient of the objective function as:

$$\nabla f(x) = \begin{pmatrix} 20(x_1 - 3) \\ 4(x_2 - 2) \end{pmatrix}.$$

Steepest Descent vs. Newton: Example

Thus, the steepest descent direction is:

$$d^0 = -\nabla f(x^0) = \begin{pmatrix} 40 \\ 4 \end{pmatrix}.$$

Solving the line minimization problem gives $\alpha^0 = 101/2004$, meaning that $x^1 \approx (3.02, 1.20)^\top$.

Note that this is not a stationary point, because:

$$\nabla f(x^1) \approx \begin{pmatrix} 0.4 \\ -3.2 \end{pmatrix}.$$

Steepest Descent vs. Newton: Example

Next, we conduct one iteration of Newton's method with an exact line search, starting from $x^0 = (1, 1)^\top$. To do so, we compute the Hessian of the objective function as:

$$\nabla^2 f(x) = \begin{bmatrix} 20 & 0 \\ 0 & 4 \end{bmatrix}.$$

Steepest Descent vs. Newton: Example

Thus, the Newton direction is:

$$d^0 = -\left[\nabla^2 f(x^0)\right]^{-1} \nabla f(x^0) = \begin{pmatrix} 2 \\ 1 \end{pmatrix}.$$

Solving the line minimization problem gives $\alpha^0 = 1$.
Thus, $x^1 = (3, 2)^\top$.

It is simple to show that this point is stationary and a local and global minimum.

Steepest Descent vs. Newton: Example

In this example, Newton's method performs much better than steepest descent, finding a stationary point and terminating after a single iteration.

Steepest descent is unable to find it in a single iteration with the latter objective function whereas Newton's method is.

Steepest Descent vs. Newton: Example

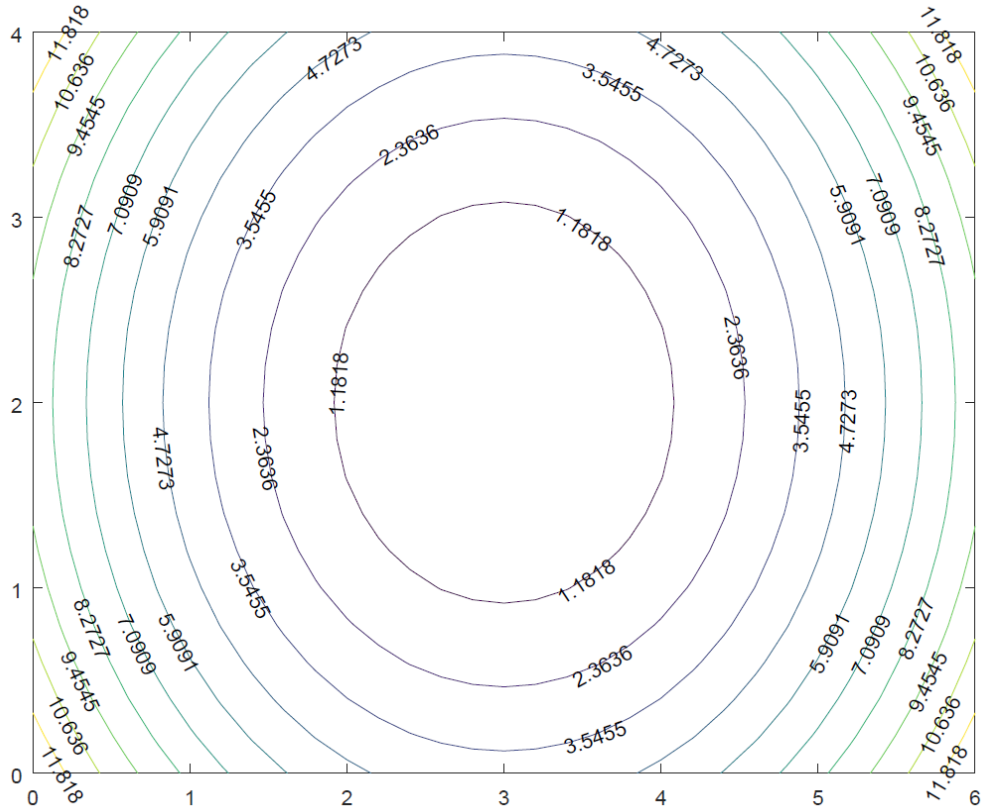
To see why this happens, the figures below show the contour plots of the objective functions in two examples. The contour plot of the objective function in case 1 is a set of concentric circles centered around the stationary point at $x^* = (3, 2)^\top$.

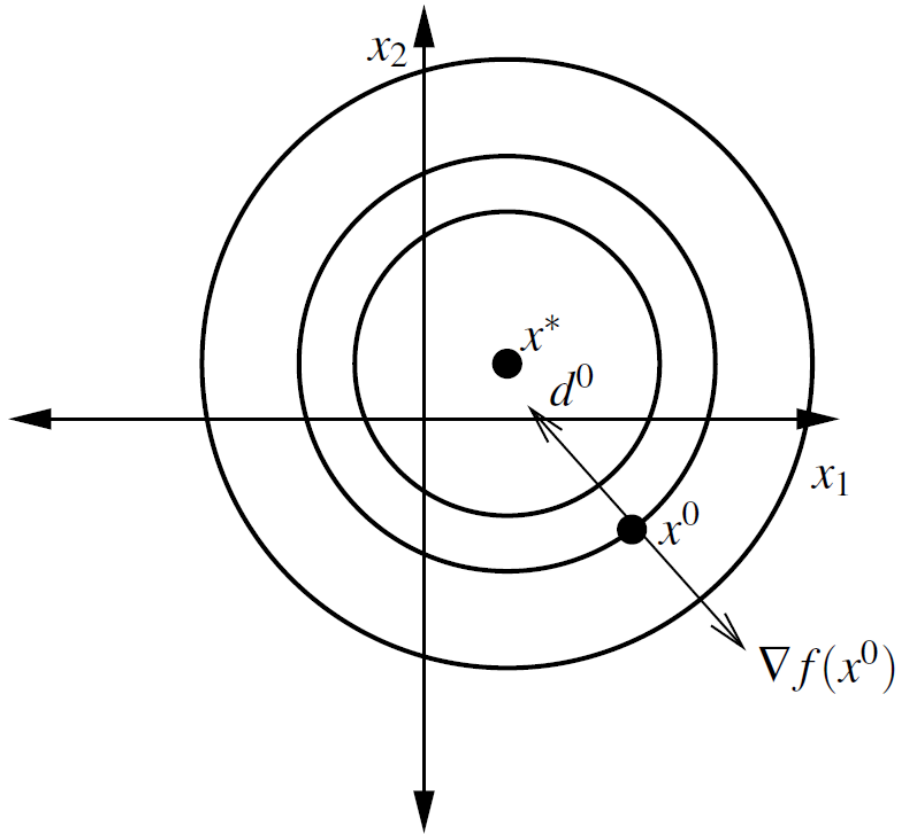
Steepest Descent vs. Newton: Example

This means that starting from any point (not just the choice of $x^0 = (1, 1)$ given in the example), the gradient points directly away from the stationary point and the steepest descent direction points directly at the stationary point. Thus, conducting a single iteration of steepest descent with an exact line search on the problem in this case, starting from any point, gives the stationary point.

Steepest Descent vs. Newton: Example

$$\min_{x} \; f(x) = (x_1 - 3)^2 + (x_2 - 2)^2$$





Steepest Descent vs. Newton: Example

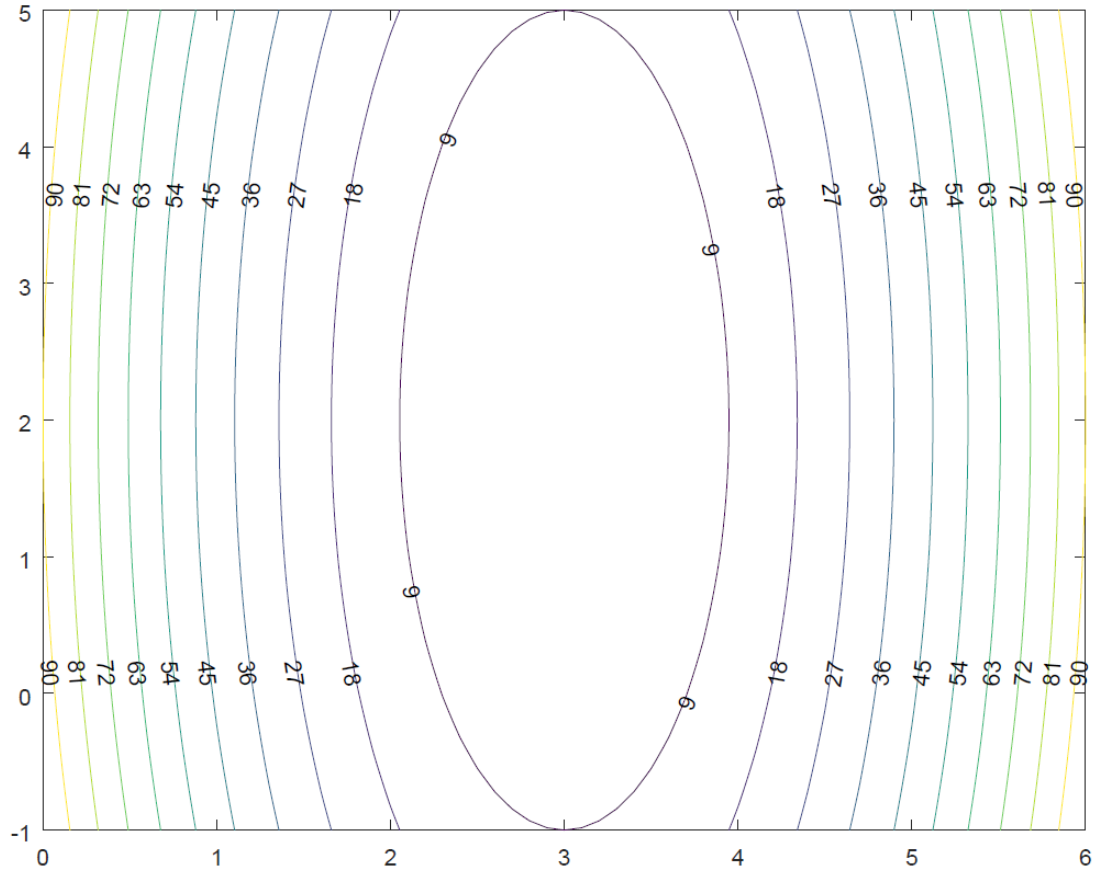
Conversely, the contour plot of the objective function in case 2 is a set of concentric ellipses centered around the stationary point at $x^* = (3, 2)^\top$. This means that starting from almost any point (including the choice of $x^0 = (1, 1)$ given in this case), the gradient does not point directly away from the stationary point. Thus, the steepest descent direction does not point directly at the stationary point.

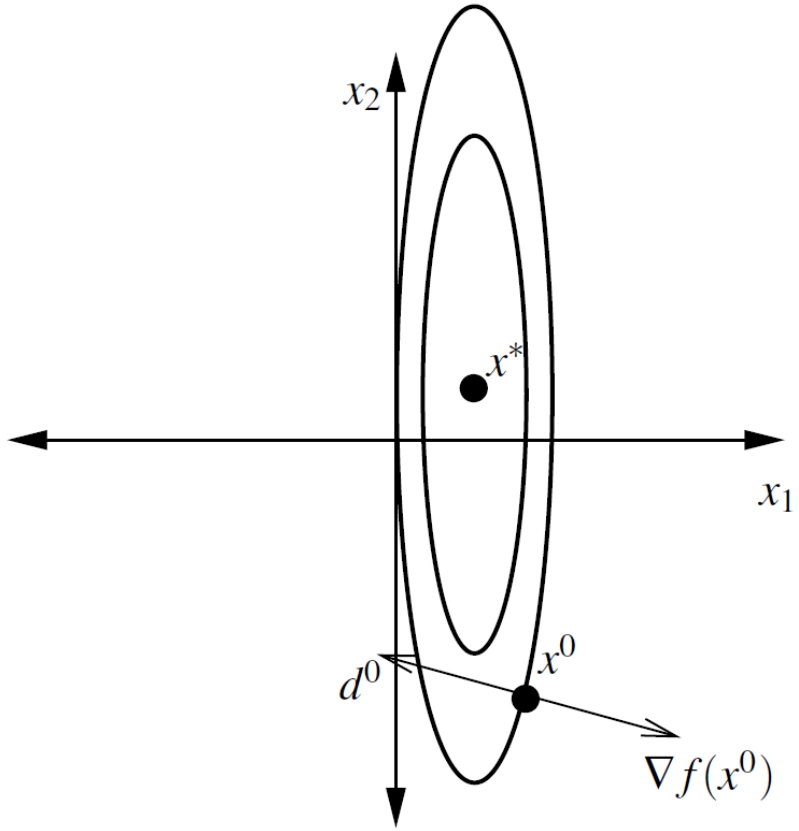
Steepest Descent vs. Newton: Example

This means that an exact line search is not able to find the stationary point in one iteration. Of course, if we are fortunate and choose a starting point which is on the major or minor axis of the concentric ellipses, then the steepest descent direction will point directly toward the stationary point. However, such luck is typically rare.

Steepest Descent vs. Newton: Example

$$\min_x \; f(x) = 10(x_1 - 3)^2 + 2(x_2 - 2)^2$$





Extended Newton

Extended Newton

$$x^{k+1} = x^k - \left[\nabla^2 f(x^k)\right]^{-1} \nabla f(x^k)$$
$$x^{k+1} = x^k - \alpha^k F^k \nabla f(x^k)$$

Options for F^k:

1. Diagonal aproximation of $\left[\nabla^2 f(x^k)\right]^{-1}$
2. Block aproximation of $\left[\nabla^2 f(x^k)\right]^{-1}$
3. Quasi-Newton approximation of $\left[\nabla^2 f(x^k)\right]^{-1}$

Extended Newton

Quasi-Newton approximation of $\left[\nabla^2 f(x^k)\right]^{-1}$

The idea is to build the inverse of the Hessian using gradient information at different iterations.

Extended Newton

$$x^{k+1} = x^k - \left[\nabla^2 f(x^k)\right]^{-1} \nabla f(x^k)$$

$$x^{k+1} = x^k - \alpha^k \left[\nabla^2 f(x^k)\right]^{-1} \nabla f(x^k)$$

$$x^{k+1} = x^k - \alpha^k \left[\beta I + (1-\beta)\left[\nabla^2 f(x^k)\right]^{-1}\right] \nabla f(x^k)$$

Coordinate Descent

Coordinate Descent

$$\min_{x_1,x_2,\ldots,x_n} \quad f(x_1, x_2, \ldots, x_n)$$

$$f(x) : \mathbb{R}^n \rightarrow \mathbb{R}$$

Coordinate Descent

$$\text{Start @ } x^0 = [x_1^0, \ldots, x_r^0, \ldots, x_n^0]$$

$$\text{Solve: } \min_{x_r} \; f(x_1^0, \ldots, x_r, \ldots, x_n^0), \text{ and get } x_r^{0*}$$

$$\text{Set } x^1 = [x_1^0, \ldots, x_r^{0*}, \ldots, x_n^0]$$

$$\text{Repeat for } r+1, \ldots, n, 1, \ldots, r$$

$$\text{Keep repeating until convergence: } |x^{k+1} - x^k| \leq \epsilon \text{ and/or}$$
$$|f(x^{k+1}) - f(x^k)| \leq \epsilon$$

Coordinate Descent

No derivatives!

Convergence guaranteed?

Quadratic case

Quadratic case

A quadratic objective function resembles a convex function!

Quadratic case

$$\min_{x \in \mathbb{R}^n} \; f(x) = \frac{1}{2} x^\top Q x + x^\top b$$

$$\nabla f(x) = Qx + b$$

$$\nabla^2 f(x) = Q$$

Quadratic case

$$\nabla^2 f(x) = Q$$

Relevant information is in Q and in its eigenvalues!

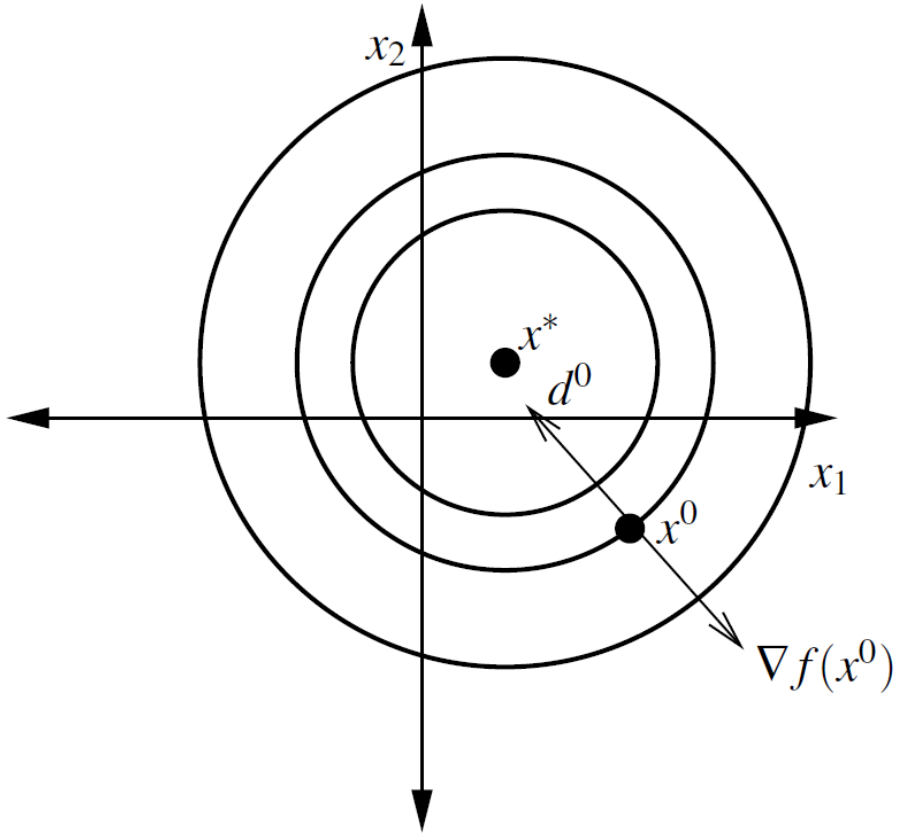
Quadratic case

$$\min_x \; f(x) = (x_1 - 3)^2 + (x_2 - 2)^2$$

$$\nabla^2 f(x) = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

$$\lambda_1 = \lambda_2 = 2$$

$$\frac{\lambda_1}{\lambda_2} = 1$$



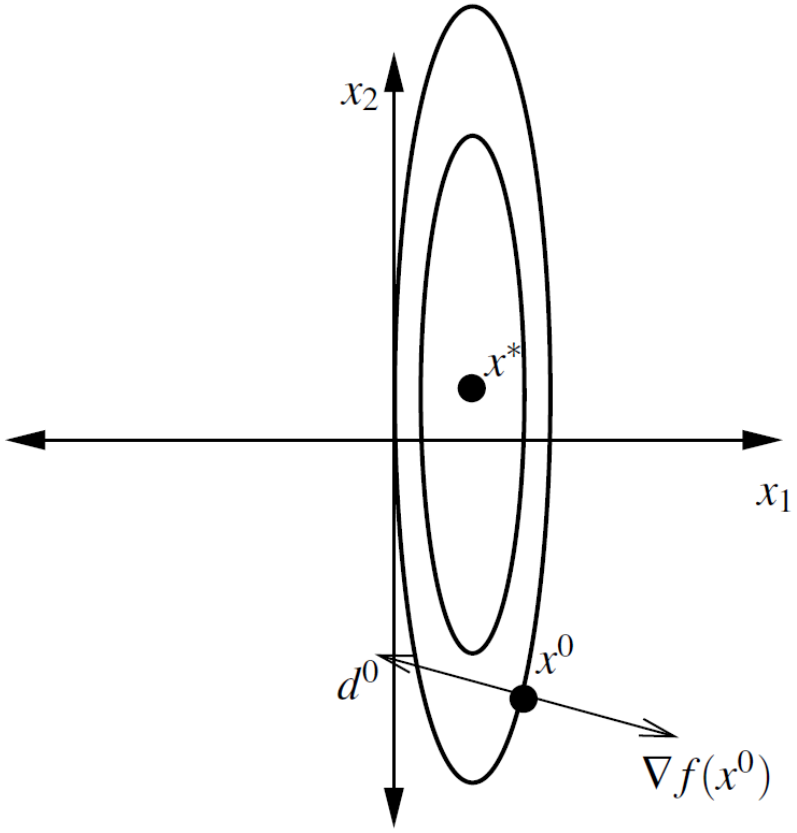
Quadratic case

$$\min_x \; f(x) = 10(x_1 - 3)^2 + 2(x_2 - 2)^2$$

$$\nabla^2 f(x) = \begin{bmatrix} 20 & 0 \\ 0 & 4 \end{bmatrix}$$

$$\lambda_1 = 20, \lambda_2 = 4$$

$$\frac{\lambda_1}{\lambda_2} = 5$$



Scaling

Scaling

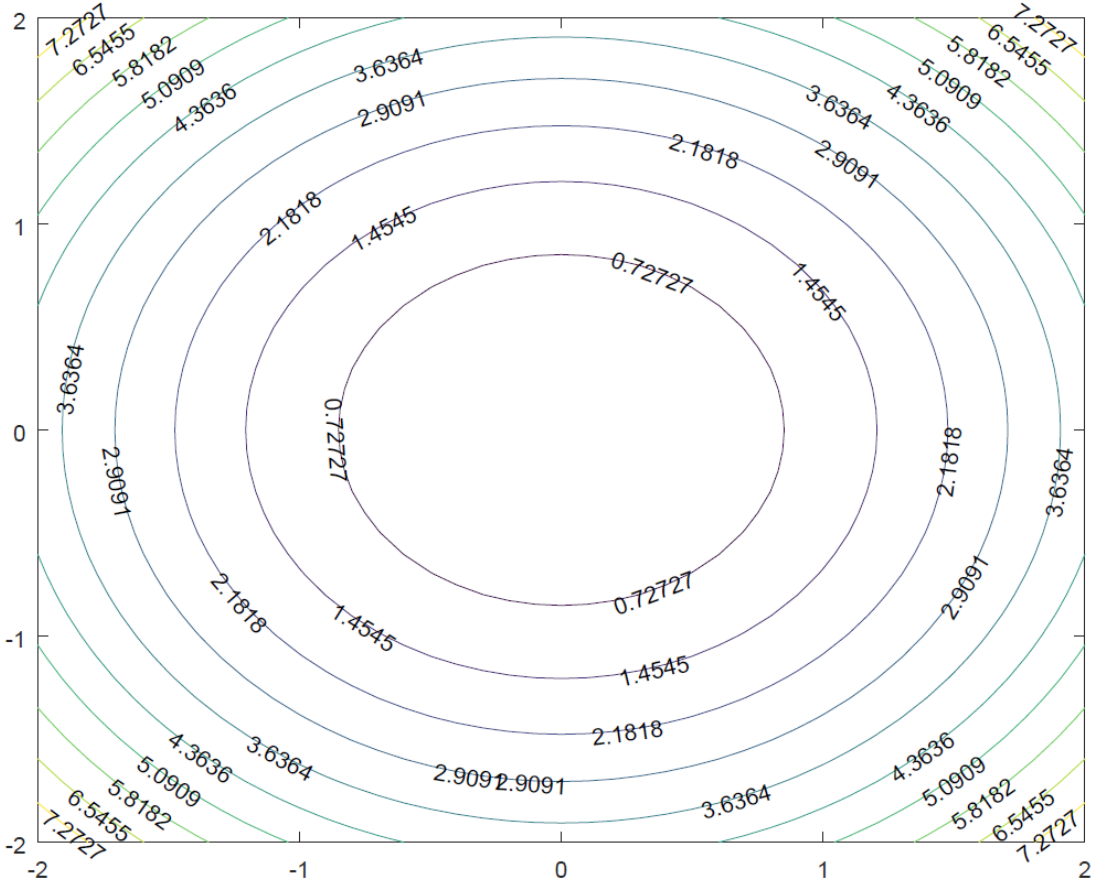
$$\min_{x_1,x_2} f(x_1,x_2) = x_1^2 + x_2^2$$

$$\nabla f(x_1,x_2) = \begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix}$$

$$\nabla^2 f(x_1,x_2) = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

eigenvalues: $\lambda_1 = 2$ and $\lambda_2 = 2$

condition number: $\frac{2}{2} = 1$



Scaling

```
home
x = -2:0.1:2;
y = -2:0.1:2;
[X,Y] = meshgrid(x,y);
Z = (X).^2+(Y).^2;
figure
contour(X,Y,Z,'ShowText','on')
```

Scaling

Try steepest descent!

$$x^{k+1} \leftarrow x^k - \alpha^k \nabla f(x^k)$$

1 step!

Scaling

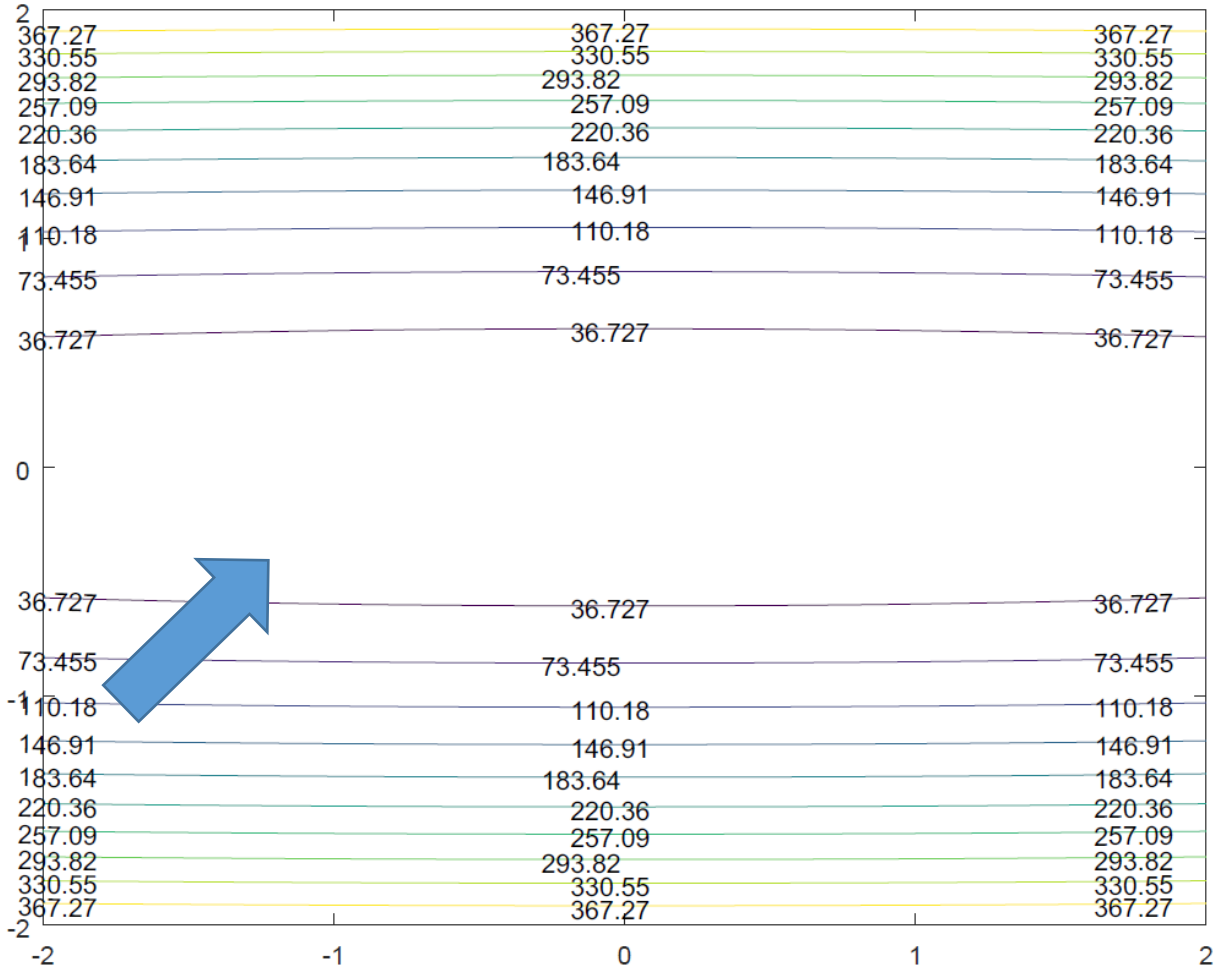
$$\min_{x_1,x_2} f(x_1,x_2) = x_1^2 + 100x_2^2$$

$$\nabla f(x_1,x_2) = \begin{bmatrix} 2x_1 \\ 200x_2 \end{bmatrix}$$

$$\nabla^2 f(x_1,x_2) = \begin{bmatrix} 2 & 0 \\ 0 & 200 \end{bmatrix}$$

eigenvalues: $\lambda_1 = 2$ and $\lambda_2 = 200$

condition number: $\frac{200}{2} = 100$



Scaling

```
home
x = -2:0.1:2;
y = -2:0.1:2;
[X,Y] = meshgrid(x,y);
Z = (X).^2+100.*(Y).^2;
figure
contour(X,Y,Z,'ShowText','on')
```

Scaling

Try steepest descent!

$$x^{k+1} \leftarrow x^k - \alpha^k \nabla f(x^k)$$

Many steps!

Scaling

$$\min_{x_1,x_2} f(x_1, x_2) = x_1^2 + 100x_2^2$$

Change of variable: $x_2 = \frac{1}{10} y_2$

Scaling

Change of variable: $x_2 = \frac{1}{10} y_2$

$$\min_{x_1, y_2} f(x_1, y_2) = x_1^2 + 100(\tfrac{1}{10} y_2)^2 = x_1^2 + y_2^2$$

$$\nabla f(x_1, y_2) = \begin{bmatrix} 2x_1 \\ 2y_2 \end{bmatrix}$$

$$\nabla^2 f(x_1, y_2) = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

eigenvalues: $\lambda_1 = 2$ and $\lambda_2 = 2$

condition number: $\frac{2}{2} = 1$

Scaling

Can we always perform such a change of variable?

This is it!