Numerical Optimization

Unconstrained Optimization

Shirish Shevade

Computer Science and Automation
Indian Institute of Science
Bangalore 560 012, India.

NPTEL Course on Numerical Optimization

Steepest Descent Method

- Uses the steepest descent direction, $\boldsymbol{d}_{SD}^k = -\boldsymbol{g}^k$

Steepest Descent Algorithm

(1) Initialize $\boldsymbol{x}^0$ and ϵ, set $k := 0$.

(2) **while** $\|\boldsymbol{g}^k\| > \epsilon$

  (a) $\boldsymbol{d}^k = -\boldsymbol{g}^k$

  (b) Find $\alpha^k (> 0)$ along $\boldsymbol{d}^k$ such that

    (i) $f(\boldsymbol{x}^k + \alpha^k \boldsymbol{d}^k) < f(\boldsymbol{x}^k)$

    (ii) α^k satisfies Armijo-Wolfe conditions

  (c) $\boldsymbol{x}^{k+1} = \boldsymbol{x}^k + \alpha^k \boldsymbol{d}^k$

  (d) $k := k + 1$

endwhile

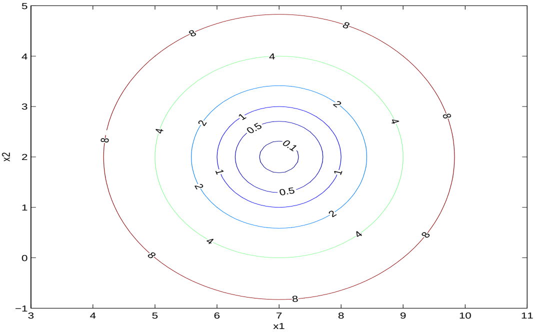
Output : $\boldsymbol{x}^* = \boldsymbol{x}^k$, a stationary point of $f(\boldsymbol{x})$.

- Exact or Backtracking line search can be used in step 2(b)

Example:

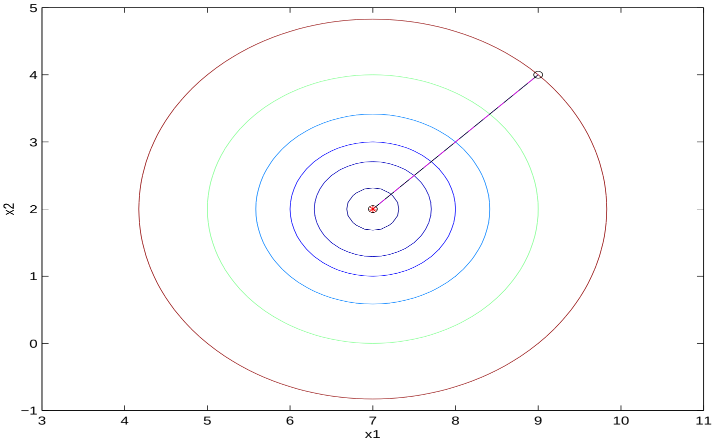
$$\min \;\; f(\boldsymbol{x}) \stackrel{\Delta}{=} (x_1 - 7)^2 + (x_2 - 2)^2$$

- $\boldsymbol{g}(\boldsymbol{x}) = \begin{pmatrix} 2(x_1 - 7) \\ 2(x_2 - 2) \end{pmatrix}$, $\boldsymbol{H}(\boldsymbol{x}) = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$.
- $\boldsymbol{x}^* = \begin{pmatrix} 7 \\ 2 \end{pmatrix}$



Example:

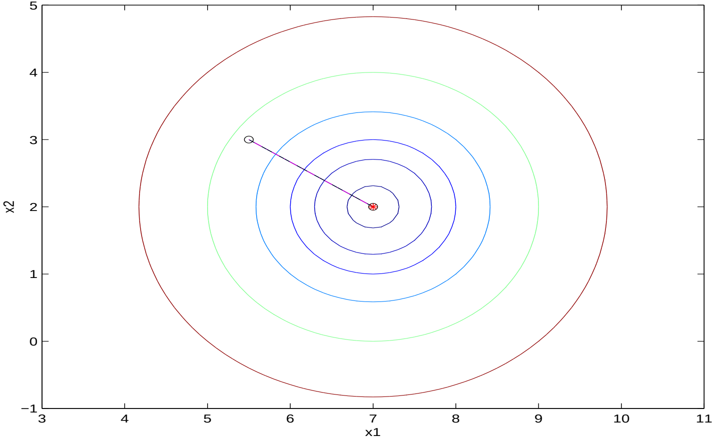
$$\min \quad f(\boldsymbol{x}) \triangleq (x_1 - 7)^2 + (x_2 - 2)^2$$



Behaviour of the steepest descent algorithm (with exact line search) applied to $f(\boldsymbol{x})$ using $\boldsymbol{x}^0 = (9, 4)^T$

Example:

$$\min \quad f(\boldsymbol{x}) \triangleq (x_1 - 7)^2 + (x_2 - 2)^2$$

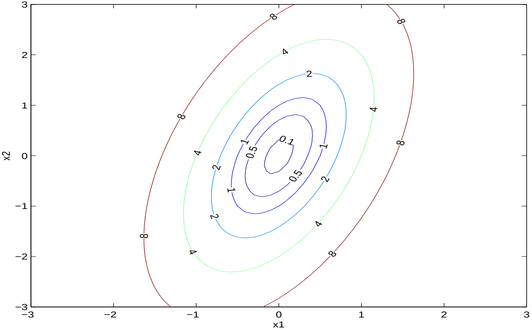


Behaviour of the steepest descent algorithm (with exact line search) applied to $f(\boldsymbol{x})$ using $\boldsymbol{x}^0 = (5.5, 3)^T$

Example:

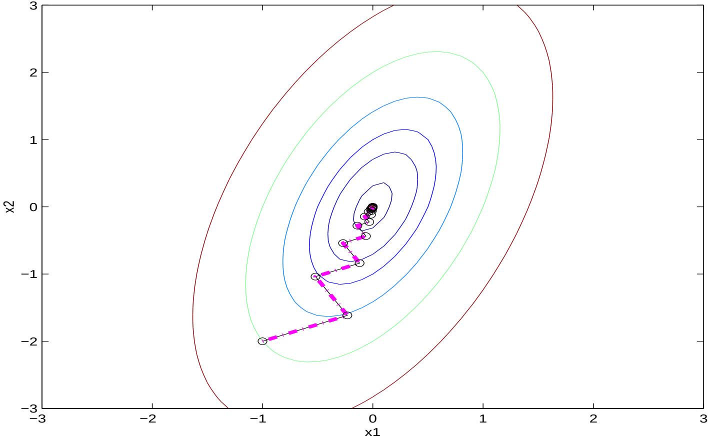
$$\min \; f(\boldsymbol{x}) \triangleq 4x_1^2 + x_2^2 - 2x_1x_2$$

- $\boldsymbol{g}(\boldsymbol{x}) = \begin{pmatrix} 8x_1 - 2x_2 \\ 2x_2 - 2x_1 \end{pmatrix}$, $\boldsymbol{H}(\boldsymbol{x}) = \begin{pmatrix} 8 & -2 \\ -2 & 2 \end{pmatrix}$.
- $\boldsymbol{x}^* = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$



Example:

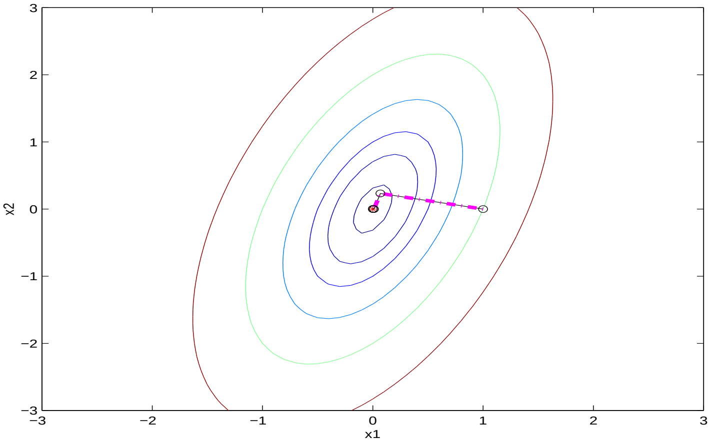
$$\min \;\; f(\boldsymbol{x}) \triangleq 4x_1^2 + x_2^2 - 2x_1x_2$$



Behaviour of the steepest descent algorithm (with exact line search) applied to $f(\boldsymbol{x})$ using $\boldsymbol{x}^0 = (-1, -2)^T$

Example:

$$\min \quad f(\boldsymbol{x}) \triangleq 4x_1^2 + x_2^2 - 2x_1x_2$$

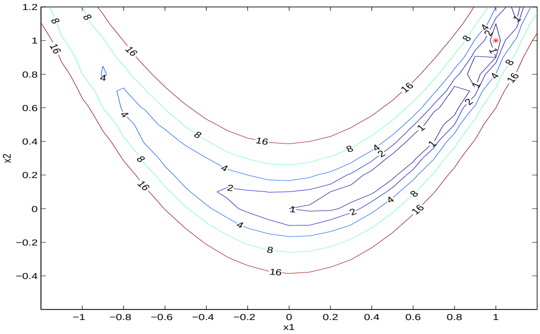


Behaviour of the steepest descent algorithm (with exact line search) applied to $f(\boldsymbol{x})$ using $\boldsymbol{x}^0 = (1, 0)^T$

Example (Rosenbrock function):

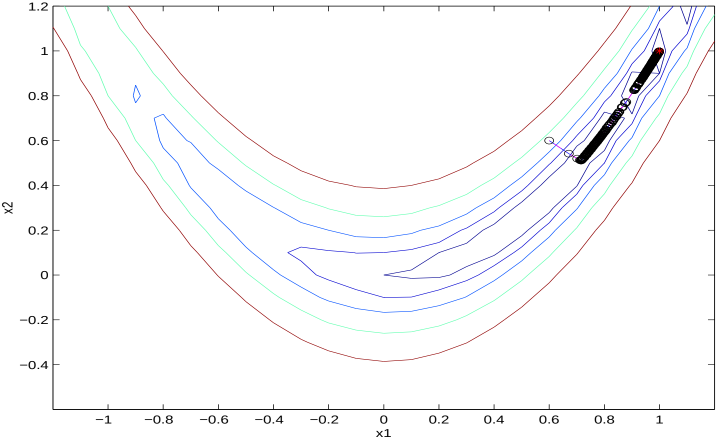
$$\min \quad f(\boldsymbol{x}) \triangleq 100(x_2 - x_1^2)^2 + (1 - x_1)^2$$

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



Example (Rosenbrock function):

$$\min \ f(\boldsymbol{x}) \stackrel{\Delta}{=} 100(x_2 - x_1^2)^2 + (1 - x_1)^2$$



Behaviour of the steepest descent algorithm (with backtracking line search) applied to $f(\boldsymbol{x})$ using $\boldsymbol{x}^0 = (0.6, 0.6)^T$

Example (Rosenbrock function):

$$\min \;\; f(\boldsymbol{x}) \triangleq 100(x_2 - x_1^2)^2 + (1 - x_1)^2$$

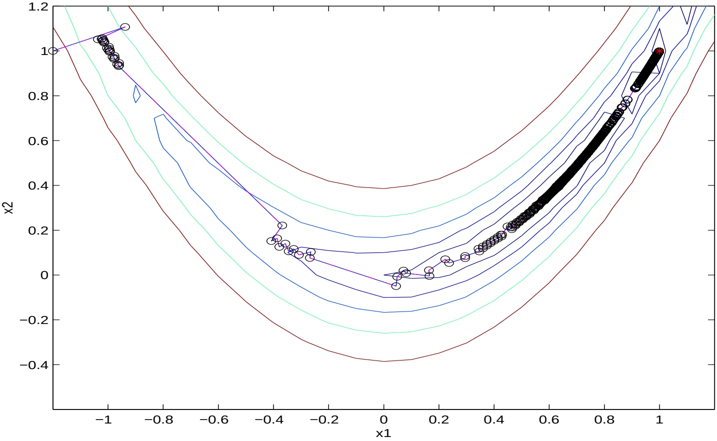
- $\boldsymbol{x}^* = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, $f(\boldsymbol{x}^*) = 0$

k	$\boldsymbol{x}_1^k$	$\boldsymbol{x}_2^k$	$f(\boldsymbol{x}^k)$	$\lVert\boldsymbol{x}^k - \boldsymbol{x}^*\rVert$	$\lVert\boldsymbol{g}^k\rVert$
0	0.6	0.6	5.92	0.5657	75.59
10	0.72	0.52	0.0792	0.5601	0.3938
100	0.78	0.61	0.0465	0.4414	0.2451
1000	0.9914	0.9828	7.45×10^{-5}	0.0192	0.0069
2028	0.9989	0.9978	1.81×10^{-6}	0.0024	9.97×10^{-4}

Table: Steepest descent method (with backtracking line search) applied to Rosenbrock function, using $\boldsymbol{x}^0 = (0.6, 0.6)^T$, $\hat{\alpha} = .5$, $\rho = .3$ and $c_1 = 1.0 \times 10^{-4}$.

Example (Rosenbrock function):

$$\min \quad f(\boldsymbol{x}) \triangleq 100(x_2 - x_1^2)^2 + (1 - x_1)^2$$



Behaviour of the steepest descent algorithm (with backtracking line search) applied to $f(\boldsymbol{x})$ using $\boldsymbol{x}^0 = (-1.2, 1)^T$

Example (Rosenbrock function):

$$\min \quad f(\boldsymbol{x}) \triangleq 100(x_2 - x_1^2)^2 + (1 - x_1)^2$$

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

k	x_1^k	x_2^k	$f(\boldsymbol{x}^k)$	$\|\boldsymbol{x}^k - \boldsymbol{x}^*\|$	$\|\boldsymbol{g}^k\|$
0	-1.2	1.0	24.2	2.2	232.87
10	-1.00	1.01	4.02	2.0042	7.69
100	0.57	0.32	0.1867	0.80	0.84
1000	0.99	0.97	1.99×10^{-4}	0.0314	0.014
2300	0.9989	0.9979	1.11×10^{-6}	0.0024	9.63×10^{-4}

Table: Steepest descent method (with backtracking lines search) applied to Rosenbrock function, using $\boldsymbol{x}^0 = (-1.2, 1)^T$, $\hat{\alpha} = .5$, $\rho = .3$ and $c_1 = 1.0 \times 10^{-4}$.

Convergence of Steepest Descent Method: Quadratic case

Consider the problem:

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} \; f(\boldsymbol{x}) \stackrel{\text{def}}{=} \frac{1}{2}\boldsymbol{x}^T \boldsymbol{H} \boldsymbol{x} - \boldsymbol{c}^T \boldsymbol{x}$$

where $\boldsymbol{H}$ is a symmetric positive-definite matrix.

- $\boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{H}\boldsymbol{x} - \boldsymbol{c}. \quad \therefore \boldsymbol{x}^* = \boldsymbol{H}^{-1}\boldsymbol{c}.$
- How does steepest descent method perform, when applied to $f(\boldsymbol{x})$?
- Assume that *exact line search* is used in each iteration

What is the step length α^k at iteration k?

$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{H}\boldsymbol{x} - \boldsymbol{c}^T\boldsymbol{x}$. $\therefore \boldsymbol{g}^k = \boldsymbol{g}(\boldsymbol{x}^k) = H\boldsymbol{x}^k - \boldsymbol{c}$

Define $\phi(\alpha) = f(\boldsymbol{x}^k + \alpha\boldsymbol{d}^k) = f(\boldsymbol{x}^k - \alpha\boldsymbol{g}^k)$.

Exact line search:

$$\alpha^k = \arg\min_{\alpha>0} \; \phi(\alpha)$$

$$\begin{aligned} \phi'(\alpha) = 0 \quad &\Rightarrow \quad \nabla f(\boldsymbol{x}^k - \alpha\boldsymbol{g}^k)^T(-\boldsymbol{g}^k) = 0 \\ &\Rightarrow \quad (\boldsymbol{H}\boldsymbol{x}^k - \alpha\boldsymbol{H}\boldsymbol{g}^k - \boldsymbol{c})^T\boldsymbol{g}^k = 0 \\ &\Rightarrow \quad (\boldsymbol{g}^k - \alpha\boldsymbol{H}\boldsymbol{g}^k)^T\boldsymbol{g}^k = 0 \end{aligned}$$

Therefore,

$$\alpha^k = \frac{\boldsymbol{g}^{kT}\boldsymbol{g}^k}{\boldsymbol{g}^{kT}\boldsymbol{H}\boldsymbol{g}^k}$$

$$\therefore \boldsymbol{x}^{k+1} = \boldsymbol{x}^k - \left(\frac{\boldsymbol{g}^{kT}\boldsymbol{g}^k}{\boldsymbol{g}^{kT}\boldsymbol{H}\boldsymbol{g}^k}\right)\boldsymbol{g}^k$$

At what rate does $\{\boldsymbol{x}^k\}$ converge?

Define

$E(\boldsymbol{x}^k) = \frac{1}{2}(\boldsymbol{x}^k - \boldsymbol{x}^*)^T \boldsymbol{H}(\boldsymbol{x}^k - \boldsymbol{x}^*)$. $\quad (E(\boldsymbol{x}^k) > 0, \text{ if } \boldsymbol{x}^k \neq \boldsymbol{x}^*)$

Note that $E(\boldsymbol{x}^k) = f(\boldsymbol{x}^k) + \underbrace{\frac{1}{2}\boldsymbol{x}^{*T}\boldsymbol{H}\boldsymbol{x}^*}_{\text{constant}}$.

Define $\boldsymbol{y}^k = \boldsymbol{x}^k - \boldsymbol{x}^*$. $\therefore \boldsymbol{H}\boldsymbol{y}^k = \boldsymbol{g}^k$.

Using

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

Relative decrease in E,

$$\frac{E(\boldsymbol{x}^k) - E(\boldsymbol{x}^{k+1})}{E(\boldsymbol{x}^k)}$$

$$= \frac{(\boldsymbol{x}^k - \boldsymbol{x}^*)^T\boldsymbol{H}(\boldsymbol{x}^k - \boldsymbol{x}^*) - (\boldsymbol{x}^{k+1} - \boldsymbol{x}^*)^T\boldsymbol{H}(\boldsymbol{x}^{k+1} - \boldsymbol{x}^*)}{\boldsymbol{y}^{kT}\boldsymbol{H}\boldsymbol{y}^k}$$

$$
\begin{aligned}
& \frac{E(\boldsymbol{x}^k) - E(\boldsymbol{x}^{k+1})}{E(\boldsymbol{x}^k)} \\
= \;& \frac{(\boldsymbol{x}^k - \boldsymbol{x}^*)^T \boldsymbol{H}(\boldsymbol{x}^k - \boldsymbol{x}^*) - (\boldsymbol{x}^{k+1} - \boldsymbol{x}^*)^T \boldsymbol{H}(\boldsymbol{x}^{k+1} - \boldsymbol{x}^*)}{\boldsymbol{y}^{k^T} \boldsymbol{H} \boldsymbol{y}^k} \\
= \;& \frac{2\alpha^k \boldsymbol{g}^{k^T} \boldsymbol{g}^k - {\alpha^k}^2 \boldsymbol{g}^{k^T} \boldsymbol{H} \boldsymbol{g}^k}{\boldsymbol{y}^{k^T} \boldsymbol{H} \boldsymbol{y}^k}
\end{aligned}
$$

Substituting $\alpha^k = \frac{\boldsymbol{g}^{k^T} \boldsymbol{g}^k}{\boldsymbol{g}^{k^T} \boldsymbol{H} \boldsymbol{g}^k}$, we get

$$
\frac{E(\boldsymbol{x}^k) - E(\boldsymbol{x}^{k+1})}{E(\boldsymbol{x}^k)} = \frac{(\boldsymbol{g}^{k^T} \boldsymbol{g}^k)^2}{(\boldsymbol{g}^{k^T} \boldsymbol{H} \boldsymbol{g}^k)(\boldsymbol{g}^{k^T} \boldsymbol{H}^{-1} \boldsymbol{g}^k)}
$$

Kantorovich inequality

Let $\boldsymbol{H} \in \mathbb{R}^{n \times n}$ be a symmetric positive definite matrix. Let λ_1 and λ_n be respectively the smallest and largest eigenvalues of $\boldsymbol{H}$. Then, for any $\boldsymbol{x} \neq \boldsymbol{0}$,

$$\frac{(\boldsymbol{x}^T\boldsymbol{x})^2}{(\boldsymbol{x}^T\boldsymbol{H}\boldsymbol{x})(\boldsymbol{x}^T\boldsymbol{H}^{-1}\boldsymbol{x})} \geq \frac{4\lambda_1\lambda_n}{(\lambda_1+\lambda_n)^2}$$

Using this inequality,

$$\begin{aligned}\frac{E(\boldsymbol{x}^k)-E(\boldsymbol{x}^{k+1})}{E(\boldsymbol{x}^k)} &= \frac{({\boldsymbol{g}^k}^T\boldsymbol{g}^k)^2}{({\boldsymbol{g}^k}^T\boldsymbol{H}\boldsymbol{g}^k)({\boldsymbol{g}^k}^T\boldsymbol{H}^{-1}\boldsymbol{g}^k)} \\ &\geq \frac{4\lambda_1\lambda_n}{(\lambda_1+\lambda_n)^2}\end{aligned}$$

Therefore,

$$E(\boldsymbol{x}^{k+1}) \leq \left(\frac{\lambda_n-\lambda_1}{\lambda_n+\lambda_1}\right)^2 E(\boldsymbol{x}^k)$$

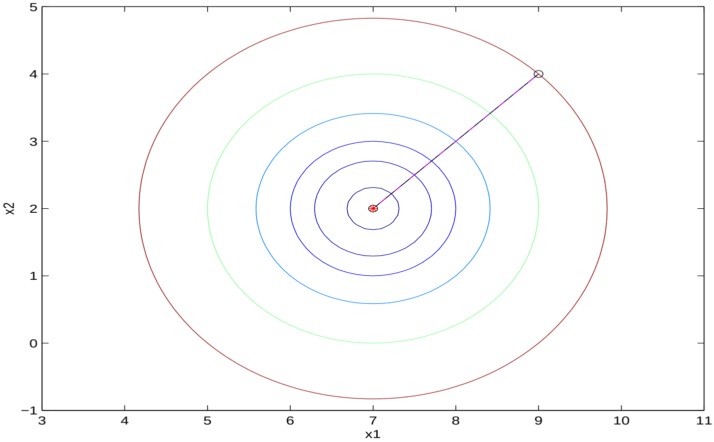
$$E(\boldsymbol{x}^{k+1}) \leq \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}\right)^2 E(\boldsymbol{x}^k)$$

Therefore, $E(\boldsymbol{x}^k) \rightarrow 0$ and $\boldsymbol{x}^k \rightarrow \boldsymbol{x}^*$($\boldsymbol{H}$ is positive definite).
With respect to E, the steepest descent method

- converges linearly with convergence rate no greater than $\left(\frac{\lambda_n-\lambda_1}{\lambda_n+\lambda_1}\right)^2$
- Actual convergence rate depends upon $\boldsymbol{x}^0$
- Define the *condition number* of $\boldsymbol{H}$, $r = \frac{\lambda_n}{\lambda_1}$
- Convergence rate of the steepest descent method depends on the condition number of $\boldsymbol{H}$
  - $r = 1$(circular contours) $\Rightarrow$ convergence in one iteration
  - $r \gg 1$(elliptical contours) $\Rightarrow$ convergence is slow
- For nonquadratic functions, rate of convergence to $\boldsymbol{x}^*$ depends on the condition number of $\boldsymbol{H}(\boldsymbol{x}^*)$

Example:

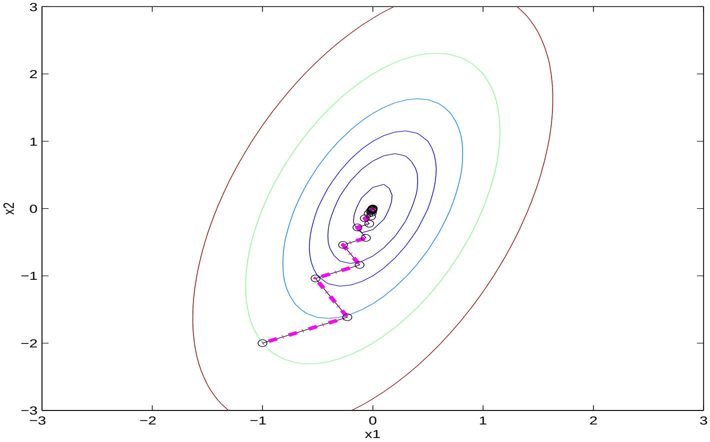
$$\min \quad f(\boldsymbol{x}) \stackrel{\Delta}{=} (x_1 - 7)^2 + (x_2 - 2)^2$$



Steepest descent algorithm (with exact line search) applied to $f(\boldsymbol{x})$ converges in one iteration from any starting point

Example:

$$\min \quad f(\boldsymbol{x}) \triangleq 4x_1^2 + x_2^2 - 2x_1x_2$$



Steepest descent algorithm (with exact line search) applied to $f(\boldsymbol{x})$ requires many iterations before it converges

Consider the problem to minimize

$$\min \;\; f(\boldsymbol{x}) \triangleq \frac{1}{2}\boldsymbol{x}^T\boldsymbol{H}\boldsymbol{x} - \boldsymbol{c}^T\boldsymbol{x}$$

where $\boldsymbol{H}$ is a symmetric positive definite matrix.

- *Condition number* of the Hessian matrix controls the convergence rate of steepest descent method.
- Faster convergence if the Hessian matrix is $\boldsymbol{I}$
- Let $\boldsymbol{H} = \boldsymbol{L}\boldsymbol{L}^T$ be the Cholesky decomposition of $\boldsymbol{H}$
- Define $\boldsymbol{y} = \boldsymbol{L}^T\boldsymbol{x}$. Therefore, the function $f(\boldsymbol{x})$ is transformed to the function $h(\boldsymbol{y})$.

$$h(\boldsymbol{y}) \triangleq f(\boldsymbol{L}^{-T}\boldsymbol{y})$$

$$
\begin{aligned}
h(\boldsymbol{y}) &= f(\boldsymbol{L}^{-T}\boldsymbol{y}) \\
&= \frac{1}{2}\boldsymbol{y}^T\boldsymbol{L}^{-1}\boldsymbol{H}\boldsymbol{L}^{-T}\boldsymbol{y} - \boldsymbol{c}^T\boldsymbol{L}^{-T}\boldsymbol{y} \\
&= \frac{1}{2}\boldsymbol{y}^T\boldsymbol{L}^{-1}\boldsymbol{L}\boldsymbol{L}^T\boldsymbol{L}^{-T}\boldsymbol{y} - \boldsymbol{c}^T\boldsymbol{L}^{-T}\boldsymbol{y} \\
&= \frac{1}{2}\boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{c}^T\boldsymbol{L}^{-T}\boldsymbol{y}
\end{aligned}
$$

- The Hessian matrix of $h(\boldsymbol{y})$ is $\boldsymbol{I}$
- Let us apply steepest descent method in $\boldsymbol{y}$-space

$$
\begin{aligned}
\boldsymbol{y}^{k+1} &= \boldsymbol{y}^k - \nabla h(\boldsymbol{y}^k) \\
&= \boldsymbol{y}^k - \boldsymbol{L}^{-1}\nabla f(\boldsymbol{L}^{-T}\boldsymbol{y}^k) \\
\therefore \boldsymbol{L}^{-T}\boldsymbol{y}^{k+1} &= \boldsymbol{L}^{-T}\boldsymbol{y}^k - \boldsymbol{L}^{-T}\boldsymbol{L}^{-1}\nabla f(\boldsymbol{L}^{-T}\boldsymbol{y}^k) \\
\therefore \boldsymbol{x}^{k+1} &= \boldsymbol{x}^k - \boldsymbol{H}^{-1}\nabla f(\boldsymbol{x}^k)
\end{aligned}
$$