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,

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

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

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,

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

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

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

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,

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

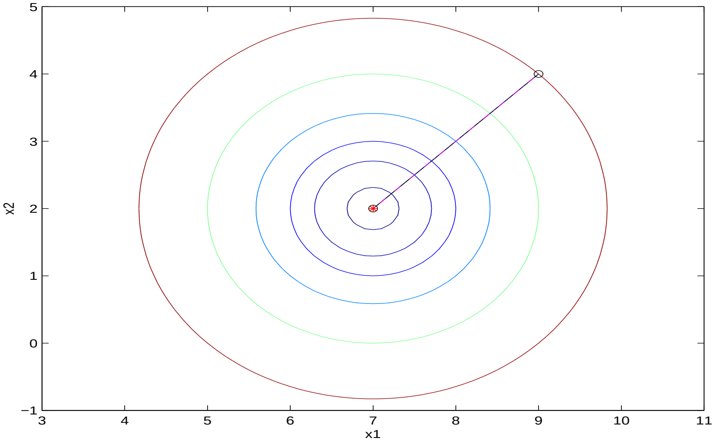
$$E(\boldsymbol{x}^{k+1}) \leq \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}\right)^2 E(\boldsymbol{x}^k)$$

Therefore, $E(\boldsymbol{x}^k) \to 0$ and $\boldsymbol{x}^k \to \boldsymbol{x}^*$($\boldsymbol{H}$ 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:

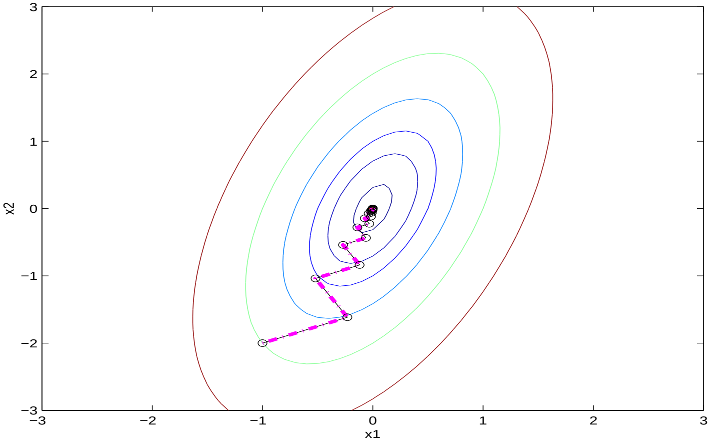
$$\min \; f(\boldsymbol{x}) \triangleq (x_1 - 7)^2 + (x_2 - 2)^2$$



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