Multivariate Linear Regression

- Deterministic method
    - Normal equation, computational complexity and instability issues
- Stochastic method
    - Gradient Descent
    - Steepest Descent
    - Conjugate Gradient

"I am trained to only sleep during national holidays."

Good reference:

Jonathan Richard Shewchuk

An Introduction to Conjugate Gradient Method Without the Agonizing Pain

http://www.cs.cmu.edu/~quake-papers/painless-conjugate-gradient.pdf

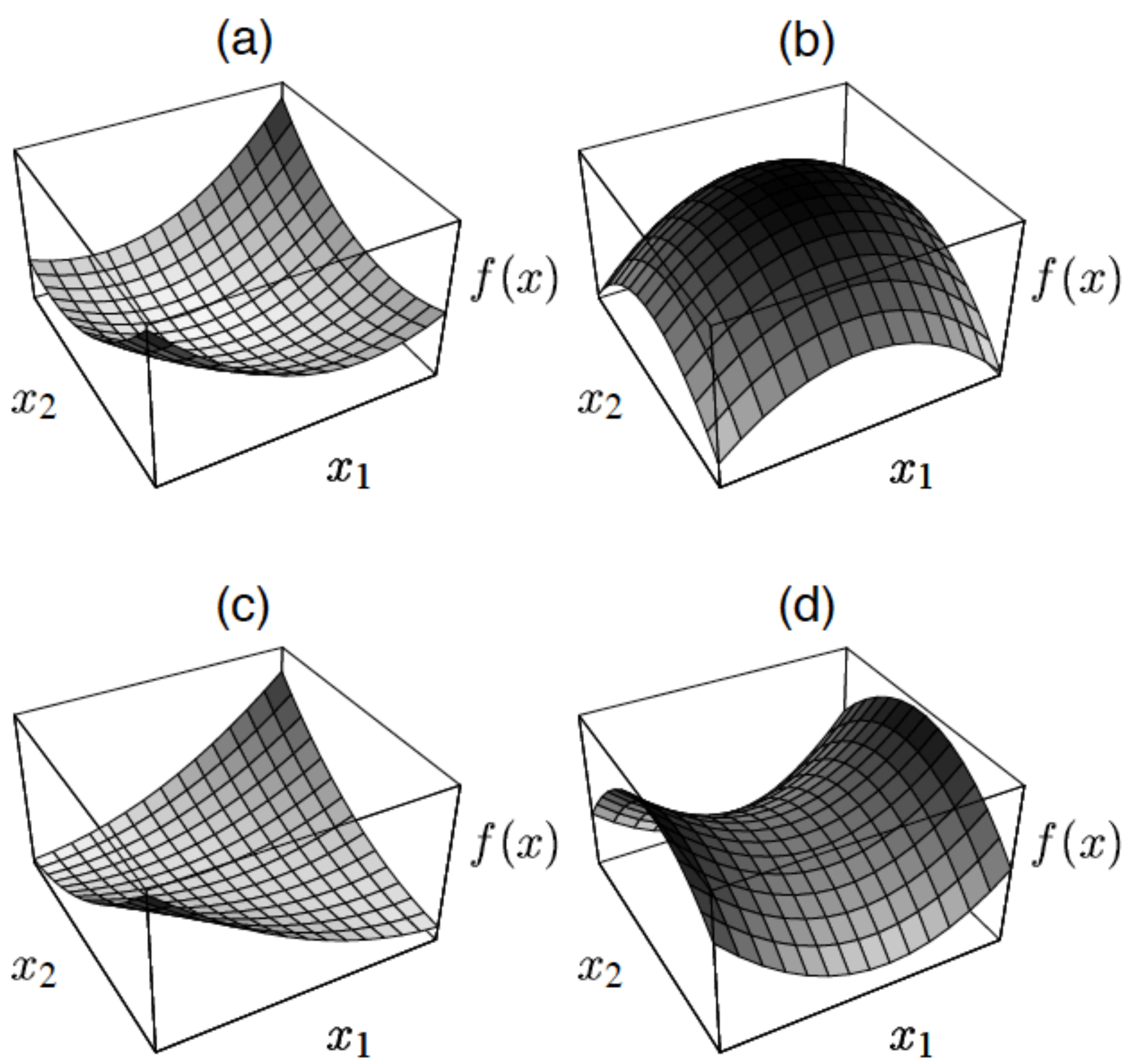
Paraboloid and positive-definite

$$f(x) = \frac{1}{2}x^T Ax - b^T x + c \quad \nabla_x f(x) = Ax - b = 0 \quad Ax = b$$

$$A = \frac{1}{M}X^T X \quad \mathbb{R}^{(N+1)\times(N+1)}$$

$$b = X^T Y \quad \mathbb{R}^{(N+1)\times 1}$$

$$A = \begin{bmatrix} 3 & 2 \\ 2 & 6 \end{bmatrix} \quad b = \begin{bmatrix} 2 \\ -8 \end{bmatrix} \quad c = 0$$

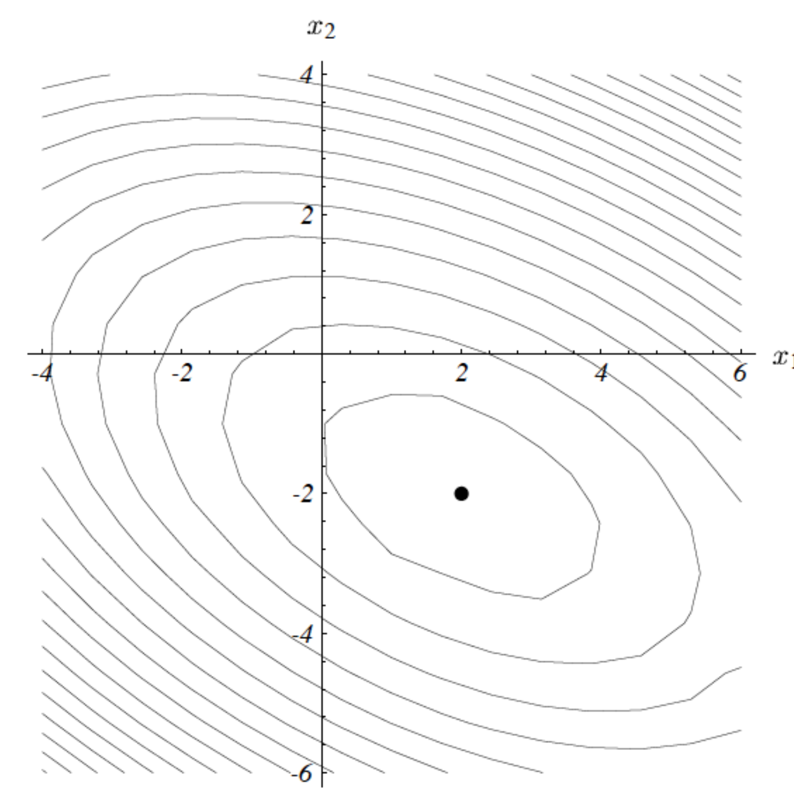


(a) positive-definite, symmetric

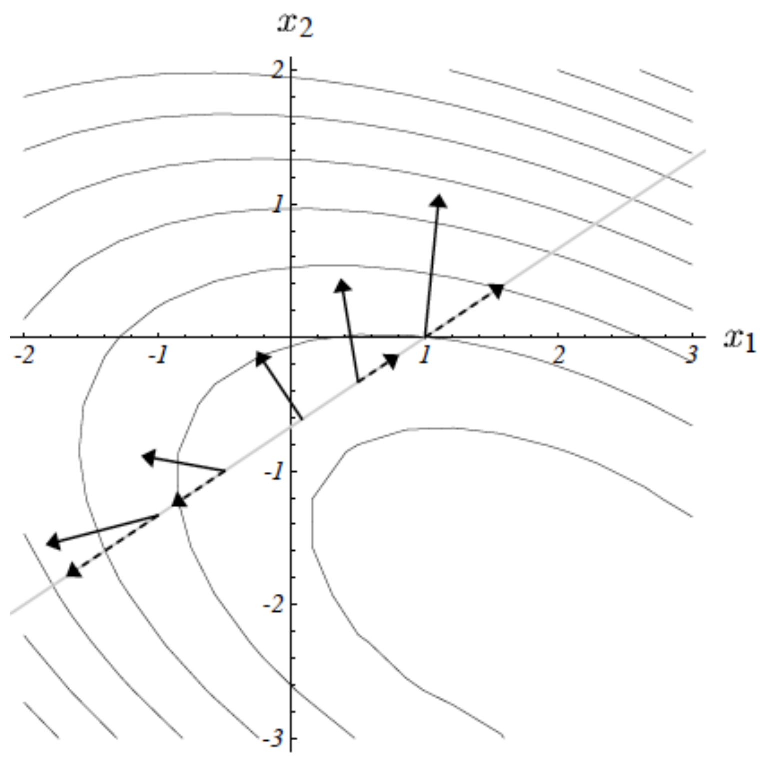
(b) negative-definite

(c) singular, a set of solution

(d) saddle point

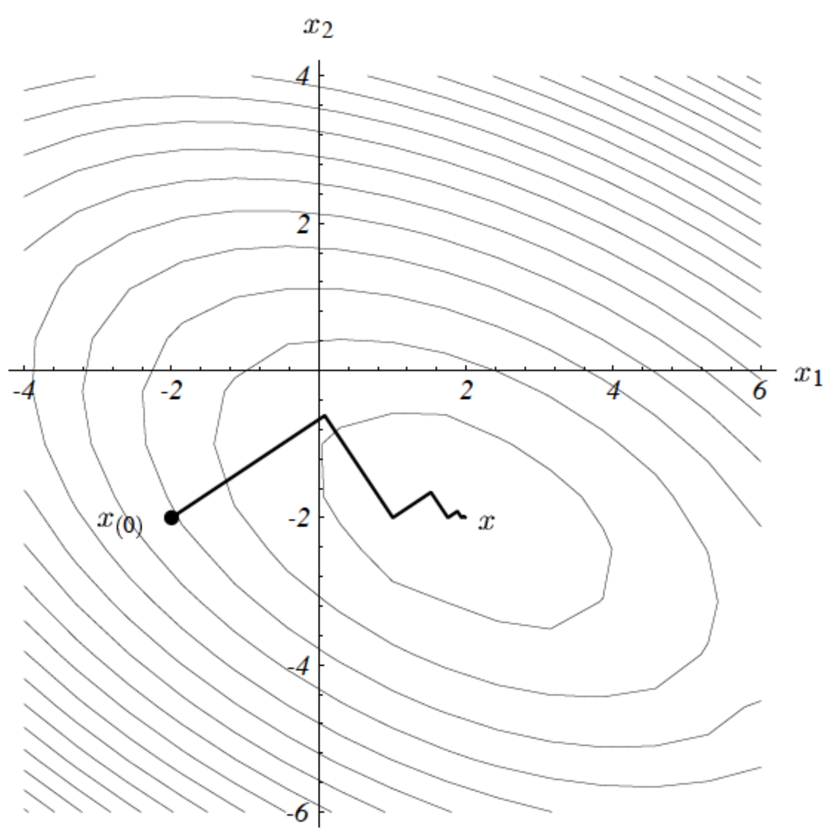


Steepest Descent



gradient

search line,
f(x) is minimised where the gradient is orthogonal to the search line

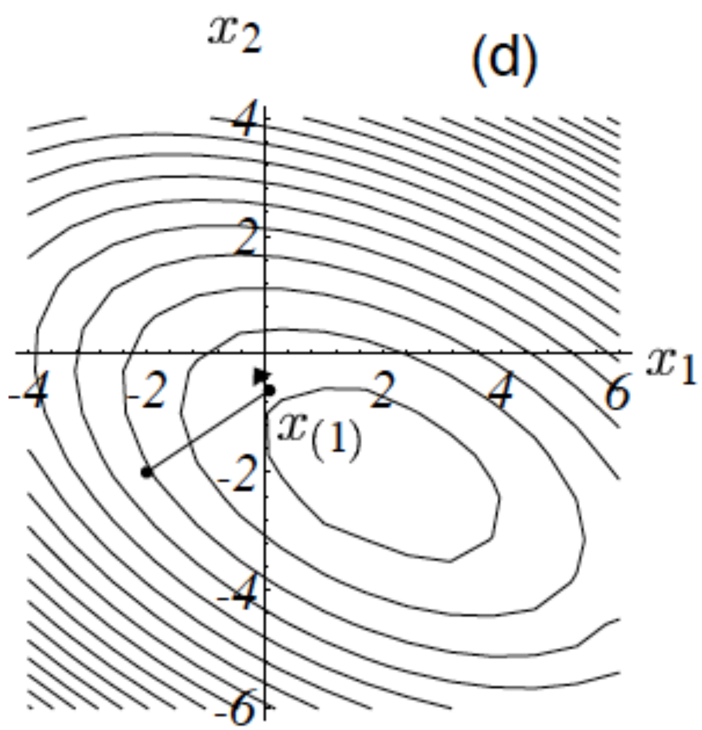


Steepest descent

The search direction is orthogonal, but this is not sufficient,
The searching direction needs to be A-orthogonal

Steepest Descent

$$A = \begin{bmatrix} 3 & 2 \\ 2 & 6 \end{bmatrix} \quad b = \begin{bmatrix} 2 \\ -8 \end{bmatrix} \quad c = 0$$



$$x_{(0)} = [-2, -2]^T$$

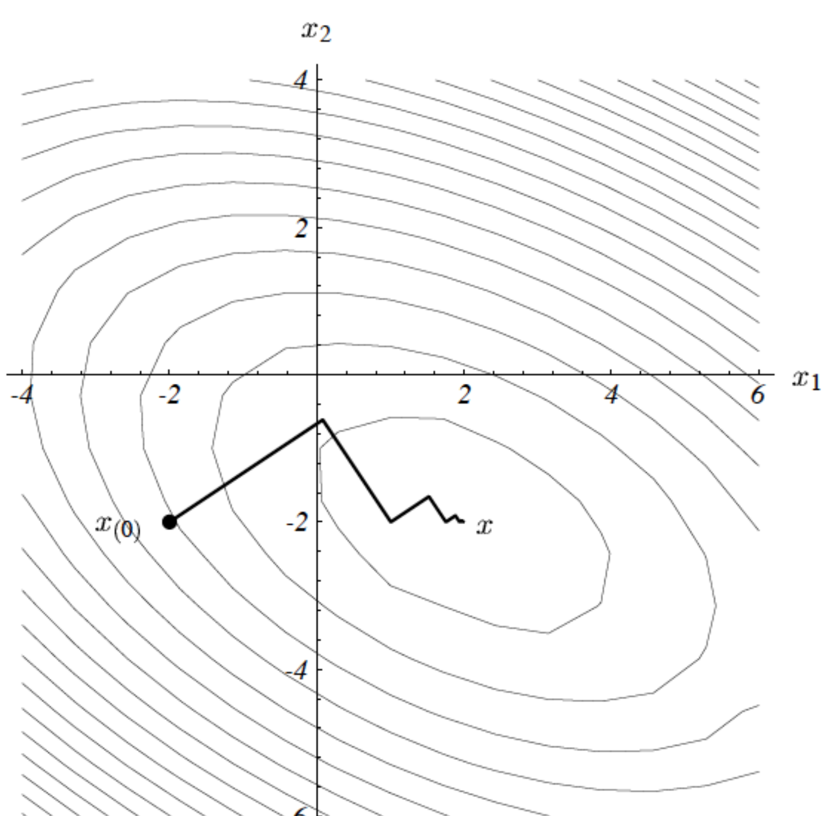
Residual $\quad r_{(i)} = b - Ax_{(i)} \qquad r_i = -\nabla_{x_{(i)}} f(x)$

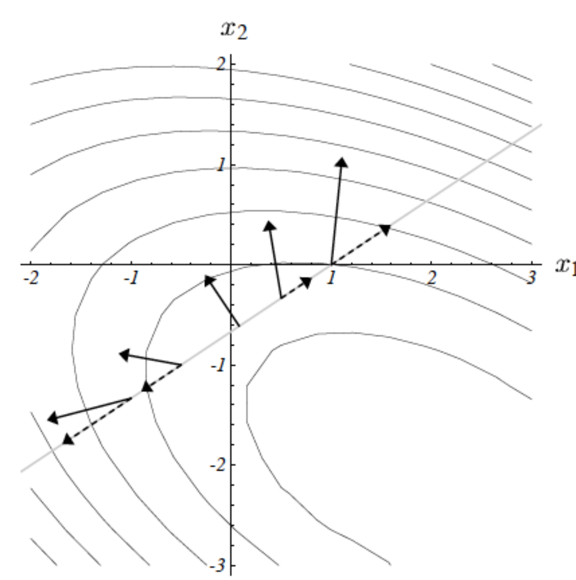
Learning rate $\quad \alpha_{(i)} = \dfrac{r_{(i)}^T r_{(i)}}{r_{(i)}^T A r_{(i)}}$

$$r_{(0)} = [12, 8]^T$$
$$r_{(0)}^T r_{(0)} = 208$$
$$r_{(0)}^T A r_{(0)} = 1200$$

Iteration $\quad x_{(i+1)} = x_{(i)} + \alpha_{(i)} r_{(i)}$

$$x_{(1)} = [-2, -2]^T + \frac{208}{1200}[12, 8]^T = [0.08, -0.62]^T$$

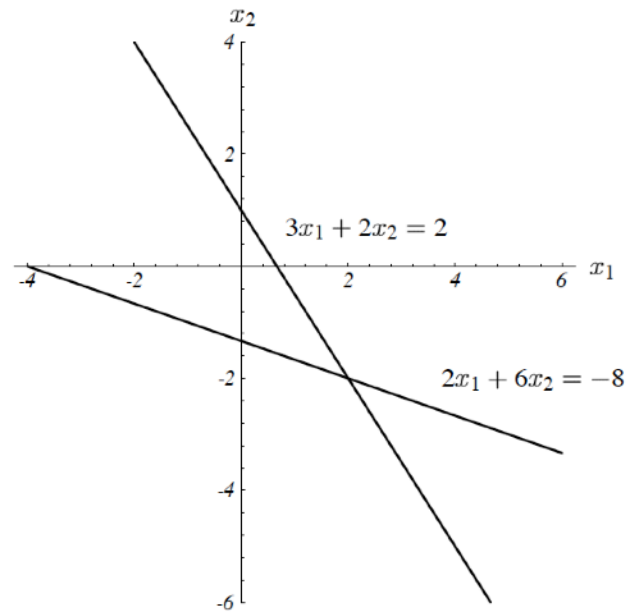


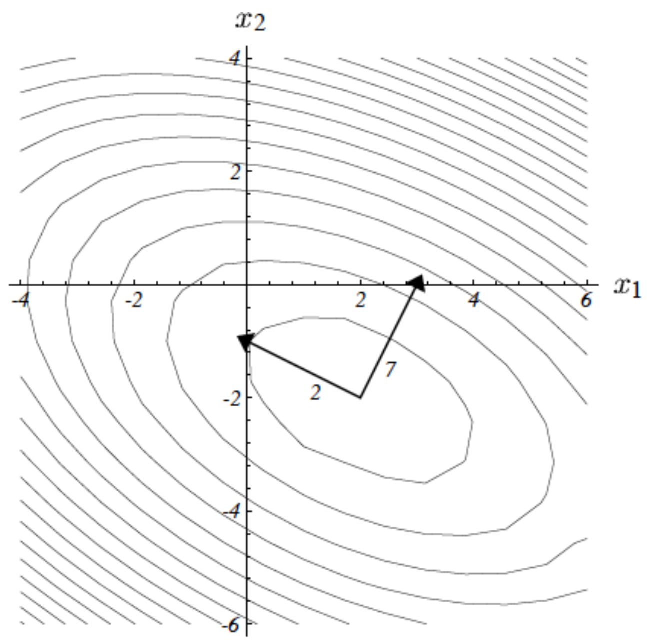


$$r_{(i+1)}^T r_{(i)} = 0$$
$$(b - Ax_{(i+1)})^T r_{(i)} = 0$$
$$(b - A(x_{(i)} + \alpha r_{(i)}))^T r_{(i)} = 0$$
$$(b - AX_{(i)})^T r_{(i)} = \alpha (Ar_{(i)})^T r_{(i)}$$
$$r_{(i)}^T r_{(i)} = \alpha r_{(i)}^T A r_{(i)}$$
$$\alpha = \frac{r_{(i)}^T r_{(i)}}{r_{(i)}^T A r_{(i)}}$$

Each gradient is orthogonal to the previous gradient

Let’s eigen do it





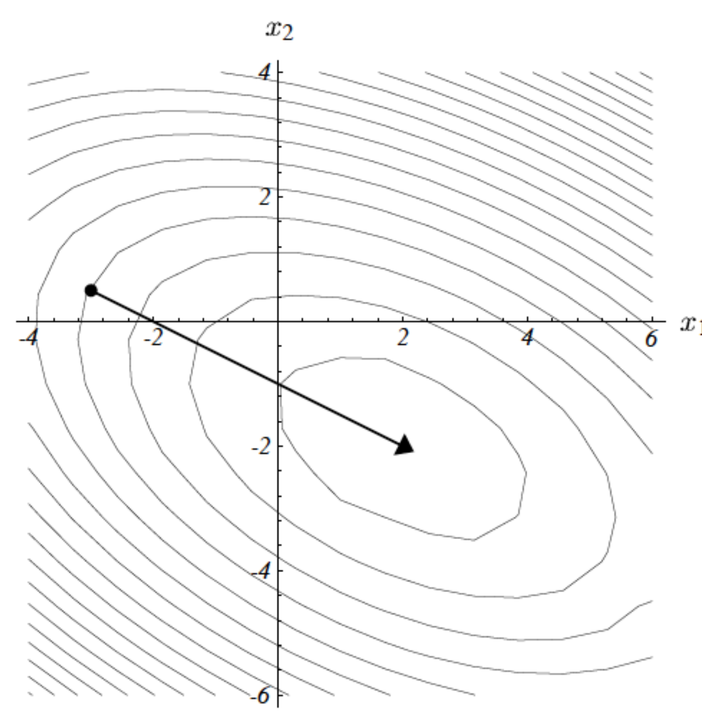
$$f(x) = \frac{1}{2}x^T A x - b^T x + c$$

$$A = \begin{bmatrix} 3 & 2 \\ 2 & 6 \end{bmatrix} \quad b = \begin{bmatrix} 2 \\ -8 \end{bmatrix} \quad c = 0$$

Eigenvalues and eigenvectors:

$$\lambda_1 = 7, \ \ e_{(1)} = [1, 2]^T$$
$$\lambda_2 = 2, \ \ e_{(2)} = [-2, 1]^T$$

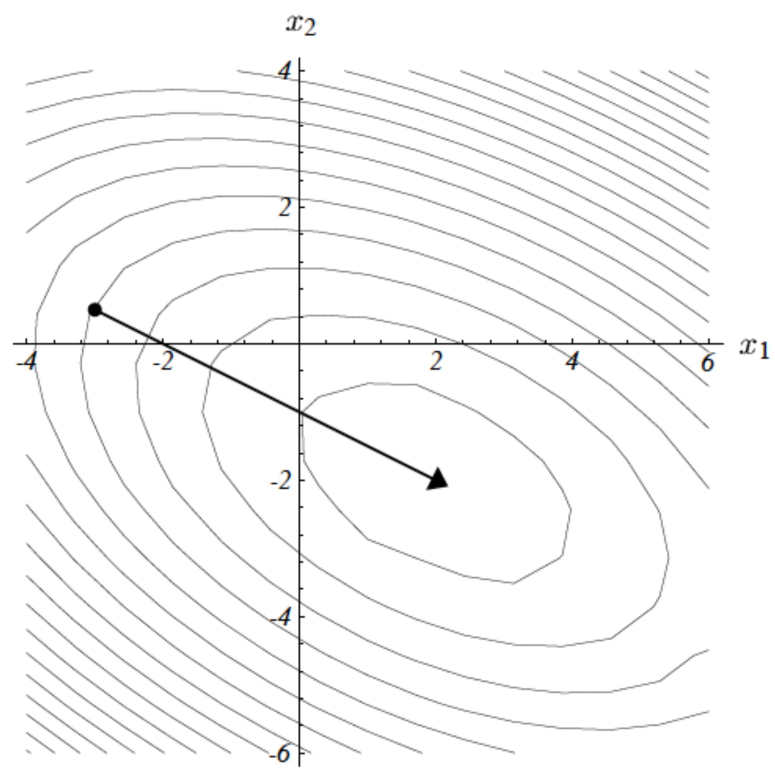
If r_{(i)} is the eigenvector, only one step to converge to the exact solution



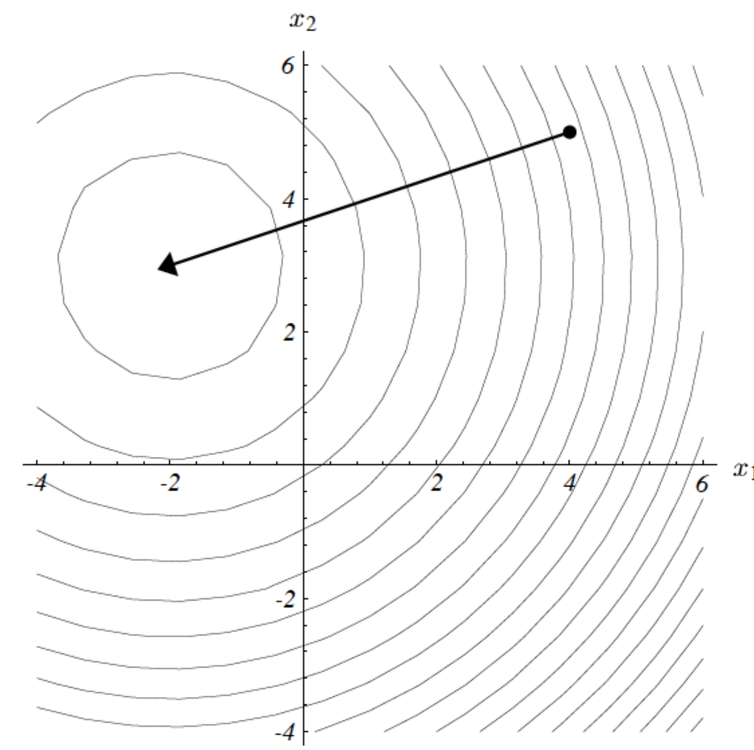
$$Ar_{(i)} = \lambda r_{(i)} \quad \alpha = 1/\lambda$$

$$\begin{aligned} r_{(i+1)} &= b - A(x_{(i)} + \alpha r_{(i)}) \\ &= b - Ax_{(i)} - \alpha A r_{(i)} \\ &= 0 \end{aligned}$$

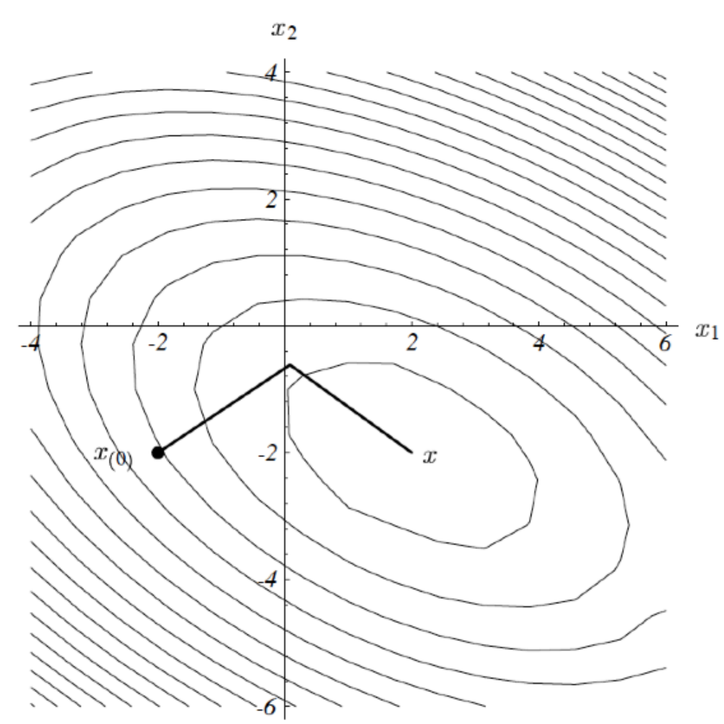
“Eigenvectors are useful tools, and not just bizarre torture devices inflicted on you by your professors for the pleasure of watching you suffer (although the latter is a nice fringe benefit)”



Paraboloid is ellipsoidal, only if r are eigenvectors, can one find the minimal

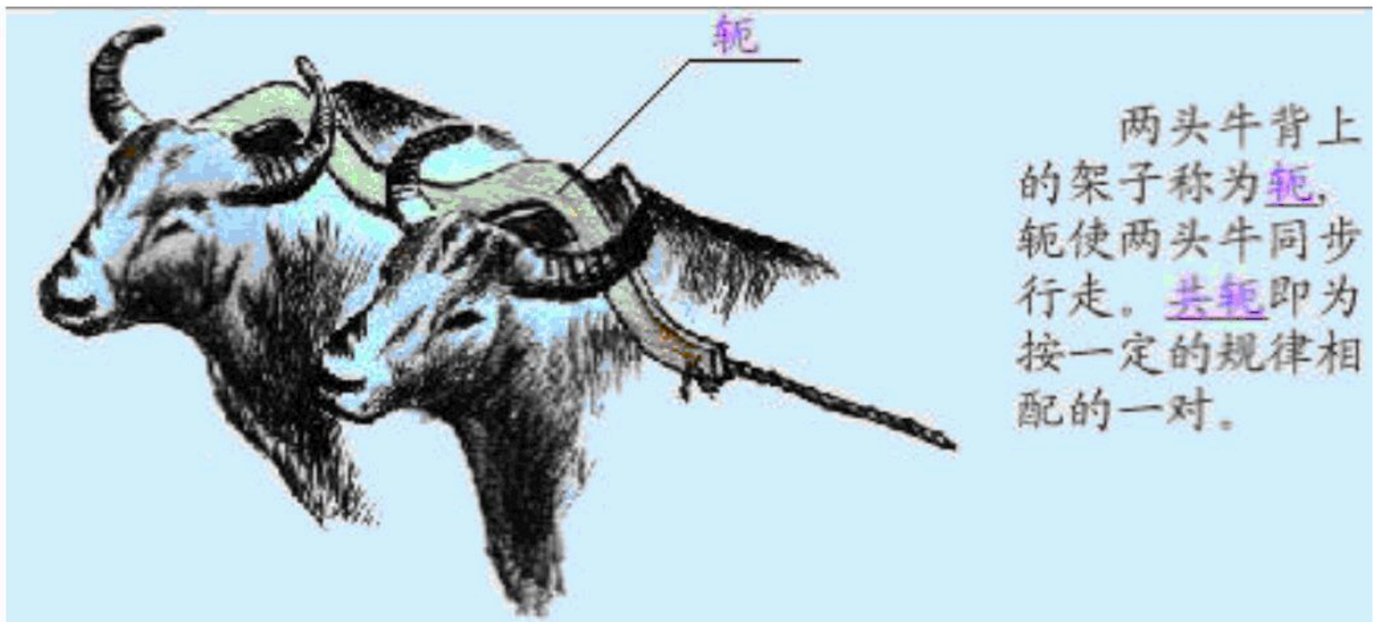


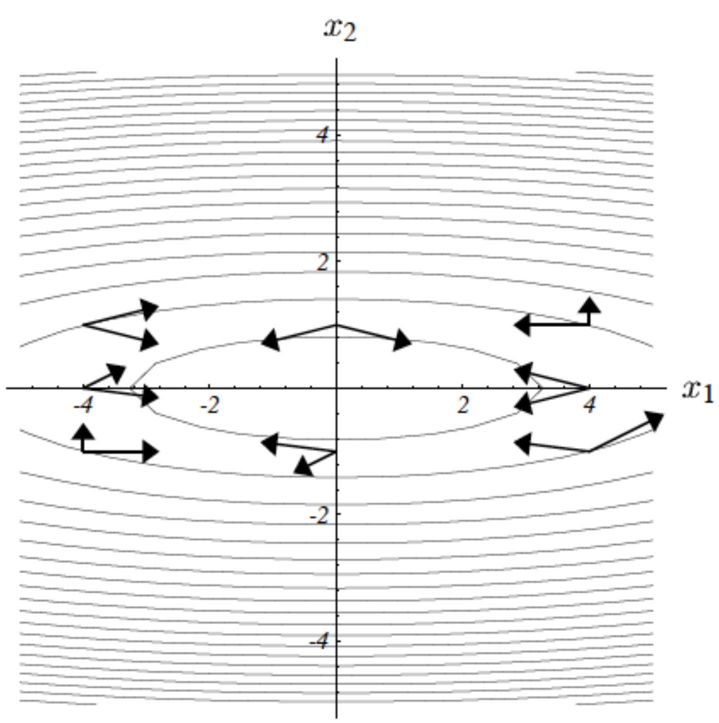
Paraboloid is spherical, no matter what point we start, always find the minimal



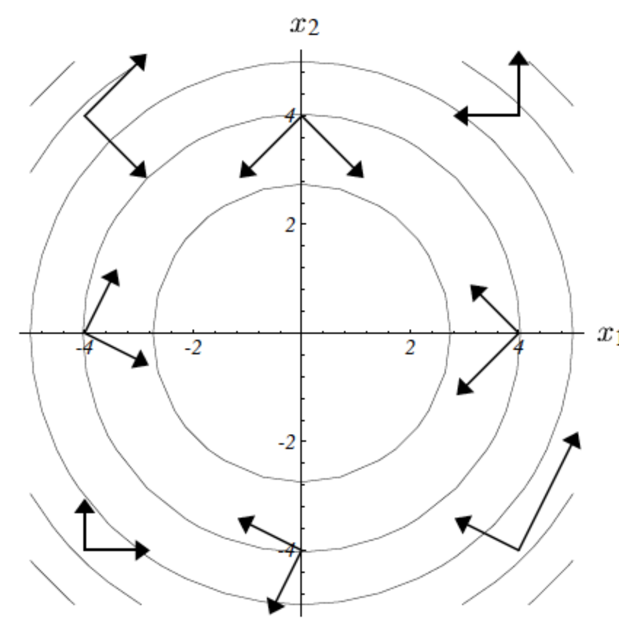
CG is to find the orthogonal directions in a stretched (scaled) space.

Conjugate Gradient Method

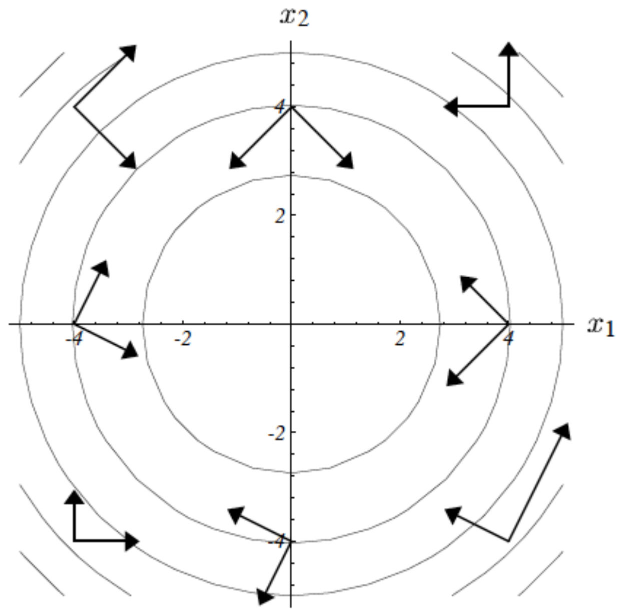




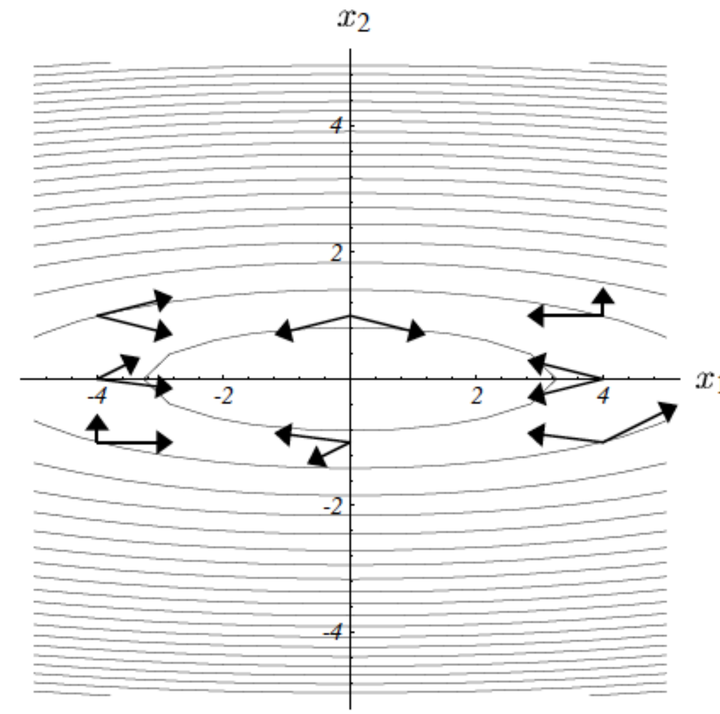
$$d_{(i)}^T A d_{(j)} = 0$$



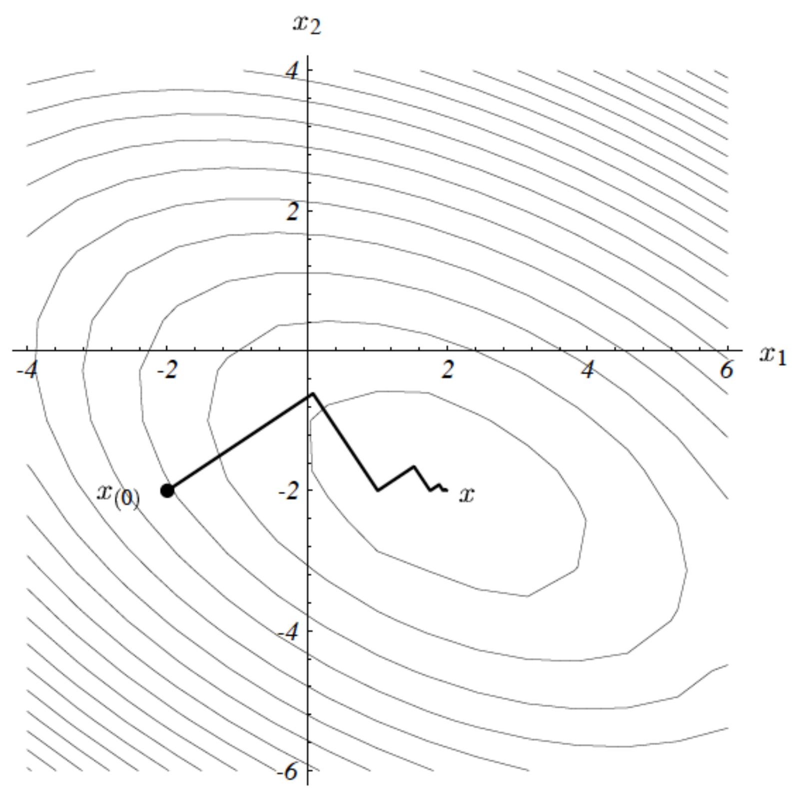
Pairs of vectors that are A-orthogonal, conjugate



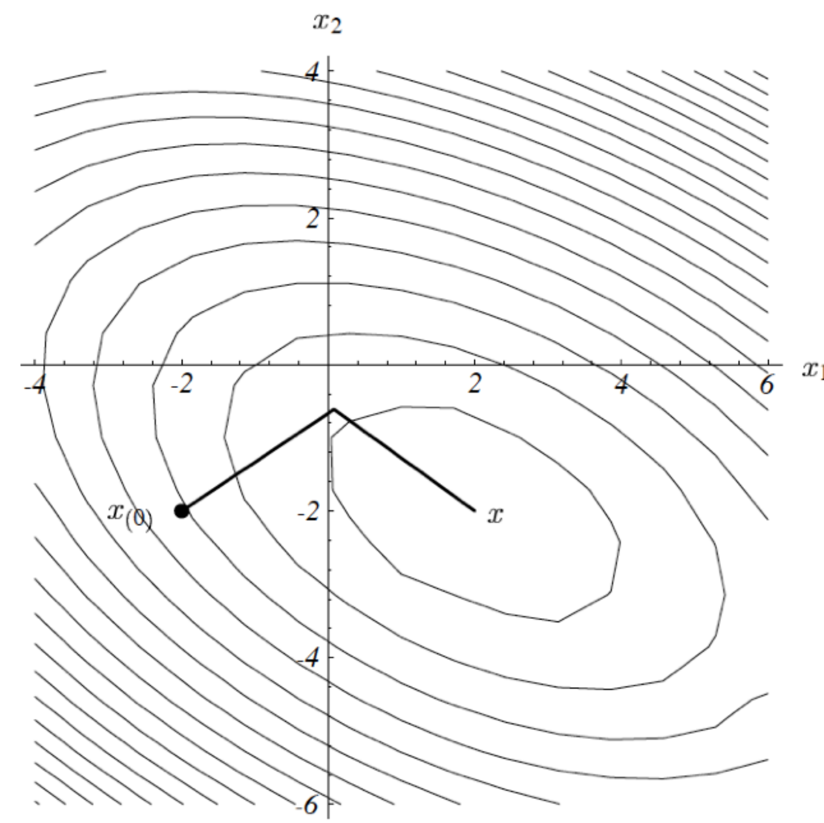
Pairs of vectors that are orthogonal



Pairs of vectors that are A-orthogonal, conjugate



Many steps to find the solution



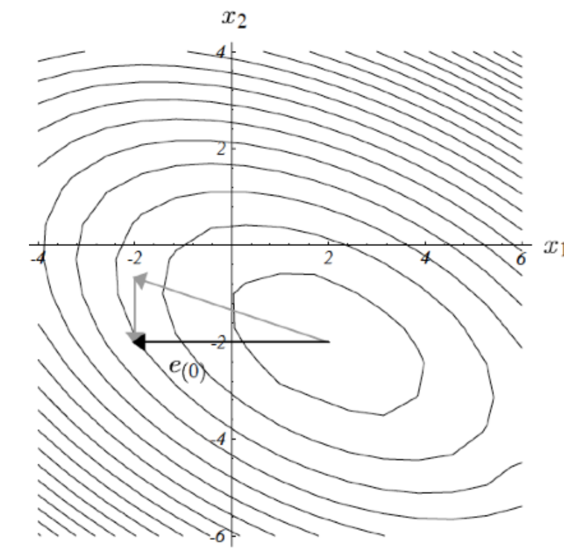
Only takes N-steps to find the solution

$$x^* = \sum_{i=1}^{n} \alpha_i p_i$$

a set of n mutually conjugate vectors (with respect to A) $P = \{P_1, P_2, \cdots, P_n\}$

One can express the solution $x^* = \sum_{i=1}^{n} \alpha_i P_i$

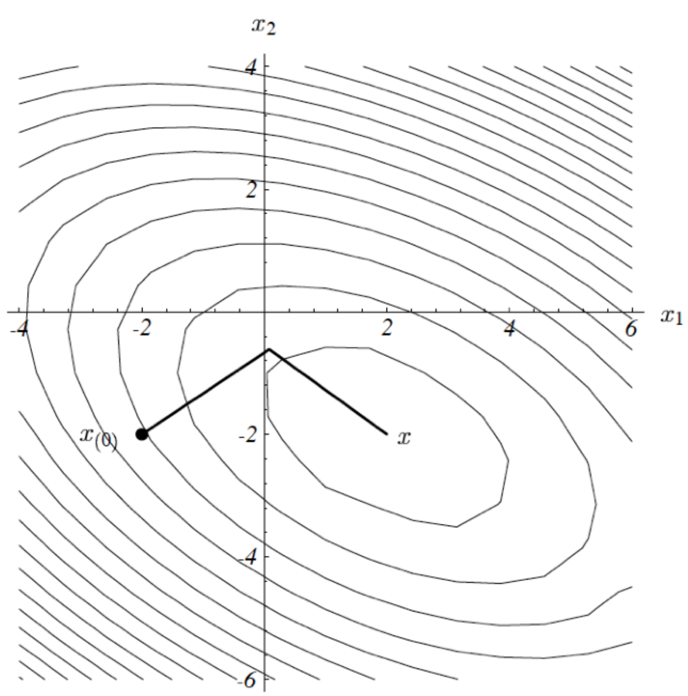
Decompose a vector x* to a sum of A-orthogonal components



$$Ax^* = \sum_{i=1}^{n} \alpha_i A P_i$$

$$P_k^T A x^* = \sum_{i=1}^{n} \alpha_i P_k^T A P_i \qquad \forall i \neq k \quad P_k^T A P_i = 0 \qquad Ax^* = b$$

$$\alpha_k = \frac{P_k^T b}{P_k^T A P_k}$$



$$r_{(0)} = b - Ax_{(0)} \qquad \text{if } r_{(0)} < \epsilon, \text{ return } x_{(0)}$$

$$P_{(0)} = r_{(0)}$$

Repeat from k=0 $\quad \alpha_{(k)} = \dfrac{r_{(k)}^T r_{(k)}}{P_{(k)}^T A P_{(k)}}$

$$x_{(k+1)} = x_{(k)} + \alpha_{(k)} P_{(k)}$$

$$r_{(k+1)} = b - Ax_{(k+1)} = b - A\big(x_{(k)} + \alpha_{(k)} P_{(k)}\big) = r_{(k)} - \alpha_{(k)} A P_{(k)}$$

Gram-Schmidt conjugation $\quad \beta_{(k)} = \dfrac{r_{(k+1)}^T r_{(k+1)}}{r_{(k)}^T r_{(k)}} \qquad \text{if } r_{(k+1)} < \epsilon, \text{ return } x_{(k+1)}$

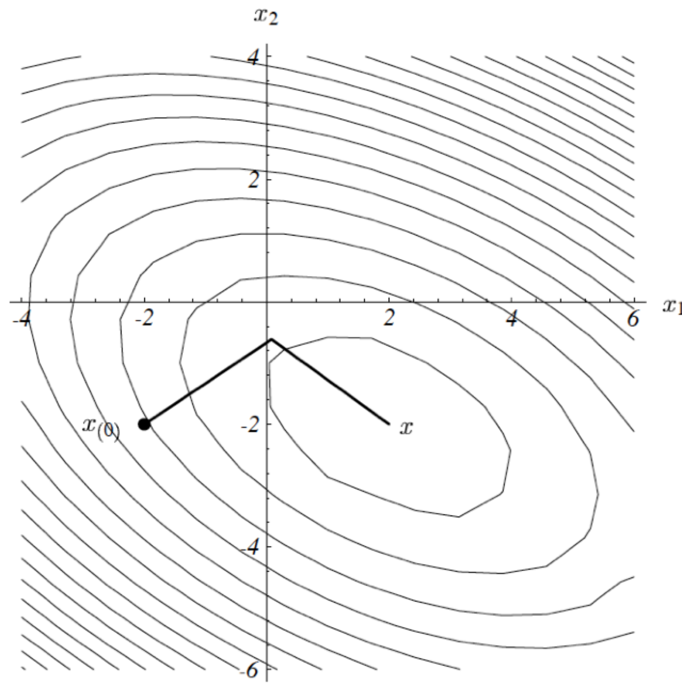
$$P_{(k+1)} = r_{(k+1)} + \beta_{(k)} P_{(k)}$$

$$k = k + 1$$

One example on the fly

$$r_{(0)} = b - Ax_{(0)}$$

$$P_{(0)} = r_{(0)}$$



Repeat from k=0

$$\alpha_{(k)} = \frac{r_{(k)}^T r_{(k)}}{P_{(k)}^T A P_{(k)}}$$

$$x_{(k+1)} = x_{(k)} + \alpha_{(k)} P_{(k)}$$

$$r_{(k+1)} = r_{(k)} - \alpha_{(k)} A P_{(k)}$$

if $r_{(k+1)} < \epsilon$, return $x_{(k+1)}$

$$\beta_{(k)} = \frac{r_{(k+1)}^T r_{(k+1)}}{r_{(k)}^T r_{(k)}}$$

$$P_{(k+1)} = r_{(k+1)} + \beta_{(k)} P_{(k)}$$

$$k = k + 1$$

> $r_{(k)}$ and $r_{(k+1)}$ are orthogonal
>
> $P_{(k)}$ and $P_{(k+1)}$ are A conjugate
>
> N step converge

$A = \begin{bmatrix} 3 & 2 \\ 2 & 6 \end{bmatrix}$ $b = \begin{bmatrix} 2 \\ -8 \end{bmatrix}$ $\lambda_1 = 7 \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ $\lambda_2 = 2 \begin{bmatrix} -2 \\ 1 \end{bmatrix}$ $r_{(1)}^T \cdot r_{(0)} = 0$

$x_{(0)} = \begin{bmatrix} -2 \\ -2 \end{bmatrix}$ $r_{(0)} = b - Ax_{(0)} = \begin{bmatrix} 2 \\ -8 \end{bmatrix} - \begin{bmatrix} 3 & 2 \\ 2 & 6 \end{bmatrix}\begin{bmatrix} -2 \\ -2 \end{bmatrix} = \begin{bmatrix} 12 \\ 8 \end{bmatrix}$

$P_{(0)} = r_{(0)}$

$\alpha_{(0)} = \frac{r_{(0)}^T r_{(0)}}{P_{(0)}^T A P_{(0)}} = \frac{208}{1200} = \frac{13}{75}$

$x_{(1)} = x_{(0)} + \alpha_{(0)} P_{(0)} = \begin{bmatrix} -2 \\ -2 \end{bmatrix} + \frac{13}{75}\begin{bmatrix} 12 \\ 8 \end{bmatrix} = \begin{bmatrix} \frac{2}{25} \\ -\frac{46}{75} \end{bmatrix}$

$r_{(1)} = b - Ax_{(1)} = \begin{bmatrix} 2 \\ -8 \end{bmatrix} - \begin{bmatrix} 3 & 2 \\ 2 & 6 \end{bmatrix}\begin{bmatrix} \frac{2}{25} \\ -\frac{46}{75} \end{bmatrix} = \begin{bmatrix} 2 \\ -8 \end{bmatrix} - \begin{bmatrix} -\frac{74}{75} \\ -\frac{264}{75} \end{bmatrix} = \begin{bmatrix} \frac{224}{75} \\ -\frac{336}{75} \end{bmatrix}$

$\beta_{(0)} = \frac{r_{(1)}^T r_{(1)}}{r_{(0)}^T r_{(0)}} = 0.1394$

$P_{(1)} = r_{(1)} + \beta_{(0)} P_{(0)} = \begin{bmatrix} \frac{224}{75} \\ -\frac{336}{75} \end{bmatrix} + 0.1394\begin{bmatrix} 12 \\ 8 \end{bmatrix} = \begin{bmatrix} 4.6597 \\ -3.3648 \end{bmatrix}$

$\alpha_{(1)} = \frac{r_{(1)}^T r_{(1)}}{P_{(1)}^T A P_{(1)}} = \frac{163072 / 5625}{[4.6597 \; -3.3648]\begin{bmatrix} 3 & 2 \\ 2 & 6 \end{bmatrix}\begin{bmatrix} 4.6597 \\ -3.3648 \end{bmatrix}} = 0.4128$

70.3505

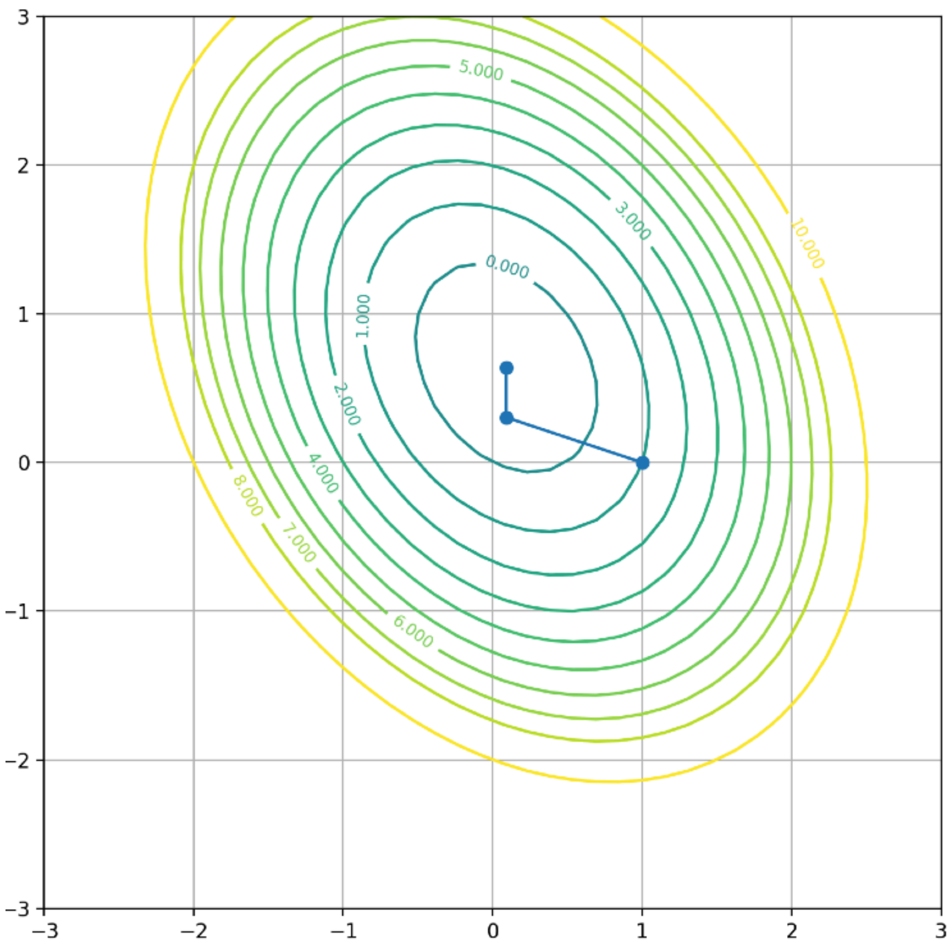
$(7.2481 \; -10.8606)$

$x_{(2)} = x_{(1)} + \alpha_{(1)} P_{(1)} = \begin{bmatrix} \frac{2}{25} \\ -\frac{46}{75} \end{bmatrix} + 0.4128\begin{bmatrix} 4.6597 \\ -3.3648 \end{bmatrix} = \begin{bmatrix} 2 \\ -2 \end{bmatrix}$ ✓

check $r_{(1)}^T \cdot r_{(0)} = [\frac{224}{75} \; -\frac{336}{75}]\begin{bmatrix} 12 \\ 8 \end{bmatrix} = \frac{224 \cdot 12 - 336 \cdot 8}{75} = \frac{2688 - 2688}{75} = 0$ ✓

$P_{(1)}^T A P_{(0)} = [4.6597 \; -3.3648]\begin{bmatrix} 3 & 2 \\ 2 & 6 \end{bmatrix}\begin{bmatrix} 12 \\ 8 \end{bmatrix} = [7.2481 \; -10.8606]\begin{bmatrix} 12 \\ 8 \end{bmatrix} = 0.02684$ ✓

round off errors



$\underline{A}\,\underline{x} = \underline{b}$, $\begin{bmatrix} 4 & 1 \\ 1 & 3 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$, solution $\underline{x} = \begin{pmatrix} 0.0909 \\ 0.6364 \end{pmatrix} = \begin{pmatrix} \frac{1}{11} \\ \frac{7}{11} \end{pmatrix}$

$\underline{x}_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$

$\underline{p}_0 := \underline{r}_0 = \underline{b} - \underline{A}\cdot\underline{x}_0 = \begin{bmatrix} 1 \\ 2 \end{bmatrix} - \begin{bmatrix} 4 & 1 \\ 1 & 3 \end{bmatrix}\begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} -3 \\ 1 \end{bmatrix}$

$\alpha_0 = \dfrac{\underline{r}_0^T \cdot \underline{r}_0}{\underline{p}_0^T \cdot \underline{A} \cdot \underline{p}_0} = \dfrac{[-3 \;\; 1]\cdot\begin{bmatrix} -3 \\ 1 \end{bmatrix}}{[-3 \;\; 1]\begin{bmatrix} 4 & 1 \\ 1 & 3 \end{bmatrix}\begin{bmatrix} -3 \\ 1 \end{bmatrix}} = \dfrac{10}{33}$

$\underline{x}_1 = \underline{x}_0 + \alpha_0 \underline{p}_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \frac{10}{33}\begin{bmatrix} -3 \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{11} \\ \frac{10}{33} \end{bmatrix}$

$\underline{r}_1 = \underline{r}_0 - \alpha_0 \underline{A}\cdot\underline{p}_0 = \begin{bmatrix} -3 \\ 1 \end{bmatrix} - \frac{10}{33}\begin{bmatrix} 4 & 1 \\ 1 & 3 \end{bmatrix}\begin{bmatrix} -3 \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{3} \\ 1 \end{bmatrix}$

$\beta_0 = \dfrac{\underline{r}_1^T \cdot \underline{r}_1}{\underline{r}_0^T \cdot \underline{r}_0} = \dfrac{[\frac{1}{3} \;\; 1]\begin{bmatrix} \frac{1}{3} \\ 1 \end{bmatrix}}{[-3 \;\; 1]\begin{bmatrix} -3 \\ 1 \end{bmatrix}} = \dfrac{1}{9}$

$\underline{p}_1 = \underline{r}_1 + \beta_0 \underline{p}_0 = \begin{bmatrix} \frac{1}{3} \\ 1 \end{bmatrix} + \frac{1}{9}\begin{bmatrix} -3 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ \frac{10}{9} \end{bmatrix}$

$\alpha_1 = \dfrac{\underline{r}_1^T \cdot \underline{r}_1}{\underline{p}_1^T \cdot \underline{A} \cdot \underline{p}_1} = \dfrac{[\frac{1}{3} \;\; 1]\begin{bmatrix} \frac{1}{3} \\ 1 \end{bmatrix}}{[0 \;\; \frac{10}{9}]\begin{bmatrix} 4 & 1 \\ 1 & 3 \end{bmatrix}\begin{bmatrix} 0 \\ \frac{10}{9} \end{bmatrix}} = \dfrac{3}{10}$

$\underline{x}_2 = \underline{x}_1 + \alpha_1 \underline{p}_1 = \begin{bmatrix} \frac{1}{11} \\ \frac{10}{33} \end{bmatrix} + \frac{3}{10}\begin{bmatrix} 0 \\ \frac{10}{9} \end{bmatrix} = \begin{bmatrix} \frac{1}{11} \\ \frac{7}{11} \end{bmatrix}$ ✓

$\underline{r}_2 = \underline{r}_1 - \alpha_1 \underline{A}\cdot\underline{p}_1 = \begin{bmatrix} \frac{1}{3} \\ 1 \end{bmatrix} - \frac{3}{10}\begin{bmatrix} 4 & 1 \\ 1 & 3 \end{bmatrix}\begin{bmatrix} 0 \\ \frac{10}{9} \end{bmatrix} = \begin{bmatrix} \frac{1}{3} \\ 1 \end{bmatrix} - \frac{3}{10}\begin{bmatrix} \frac{10}{9} \\ \frac{10}{3} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ ✓

check: r_k, r_{k+1} orthogonal ✓

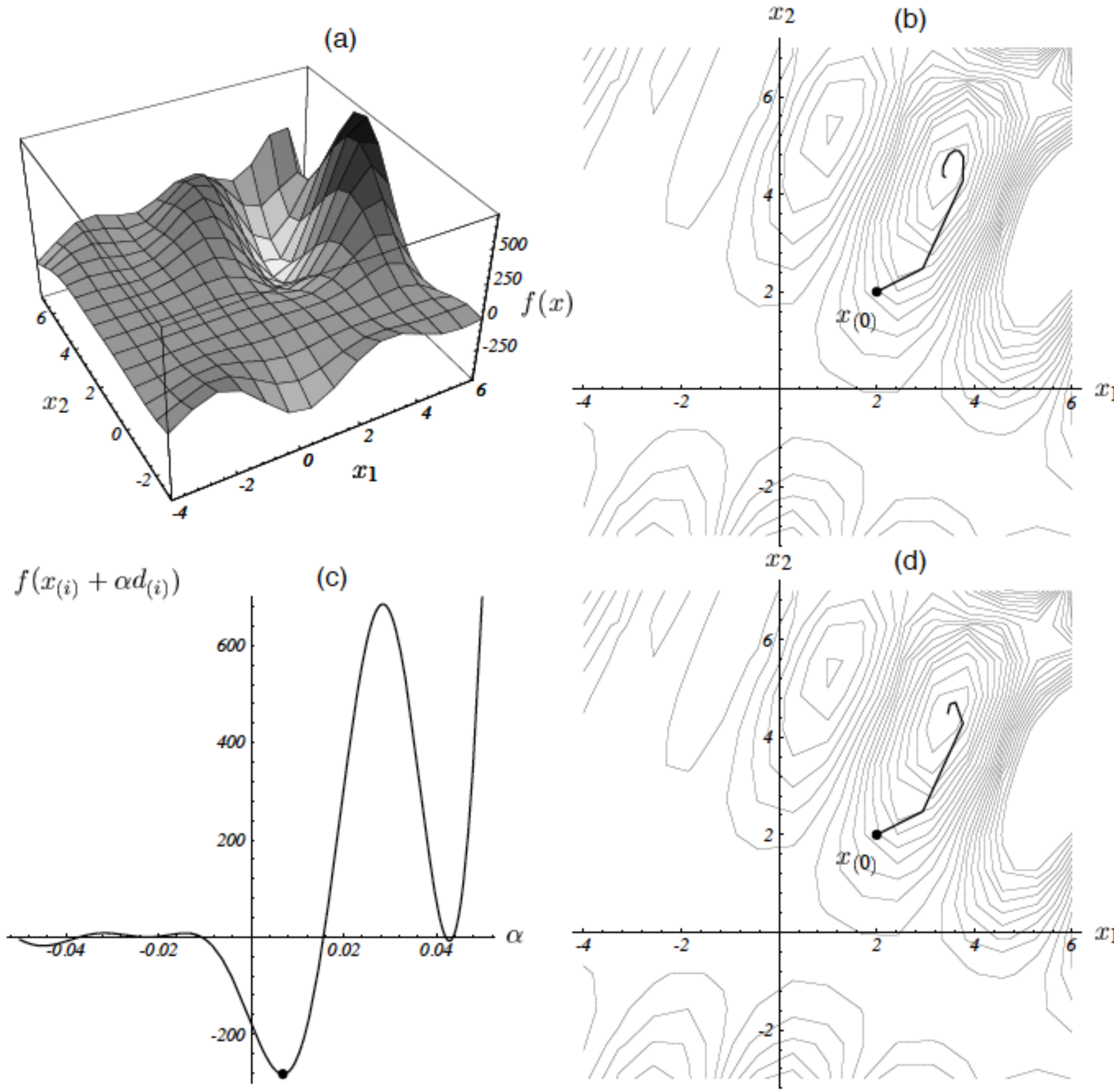
$p_k \cdot A\, p_{k+1}$ orthogonal ✓

n-step converge ✓

$\underline{r}_0^T \cdot \underline{r}_1 = [-3 \;\; 1]\begin{bmatrix} \frac{1}{3} \\ 1 \end{bmatrix} = 0$

$\underline{p}_0^T \cdot \underline{A}\,\underline{p}_1 = [-3 \;\; 1]\begin{bmatrix} 4 & 1 \\ 1 & 3 \end{bmatrix}\begin{bmatrix} 0 \\ \frac{10}{9} \end{bmatrix}$

$= [-11 \;\; 0]\begin{bmatrix} 0 \\ \frac{10}{9} \end{bmatrix}$


(a)
$f(x)$
x_1
x_2
(b)
$x_{(0)}$
(c)
$f(x_{(i)} + \alpha d_{(i)})$
α
(d)