AI & Machine Learning in Physics

PHYS3151 (6 credits)

Time & Place : Tue 13:30-14:20, 14:30-15:20 MB 122
Fri 14:30-15:20 MB 142

Teachers: Zi Yang Meng (zymeng@hku.hk), HOC 231

https://quantummc.xyz/hku-phys3151-machine-learning-in-physics-2023/

Tutor: Ting-Tung Wang (leowdd@connect.hku.hk), HOC 217

AI & Machine Learning in Physics

Teaching Materials:

https://quantummc.xyz/hku-phys3151-machine-learning-in-physics-2023/

Slides / Reading materials

Python notebooks

Assignments

Assessment Methods and Weighting

- Assignments 30%
- Presentation 20%
- Project report 20%
- Exam. 30%

Content

Content

Multivariate Linear Regression

- Ethem Alpaydin, Introduction to Machine Learning, Third Edition, MIT Press 2014

Chap. 4. Parametric Methods
4.6 Regression
4.7 Bias/Variance Dilemma
Chap. 5. Multivariate Methods
5.1 Multivariate data
5.8 Multivariate Regression

- Linear algebra books:

Gilbert Strang, Linear Algebra and its Applications, 1988
David Harville, Matrix Algebra from a Statistician's Perspective, 1997, Springer

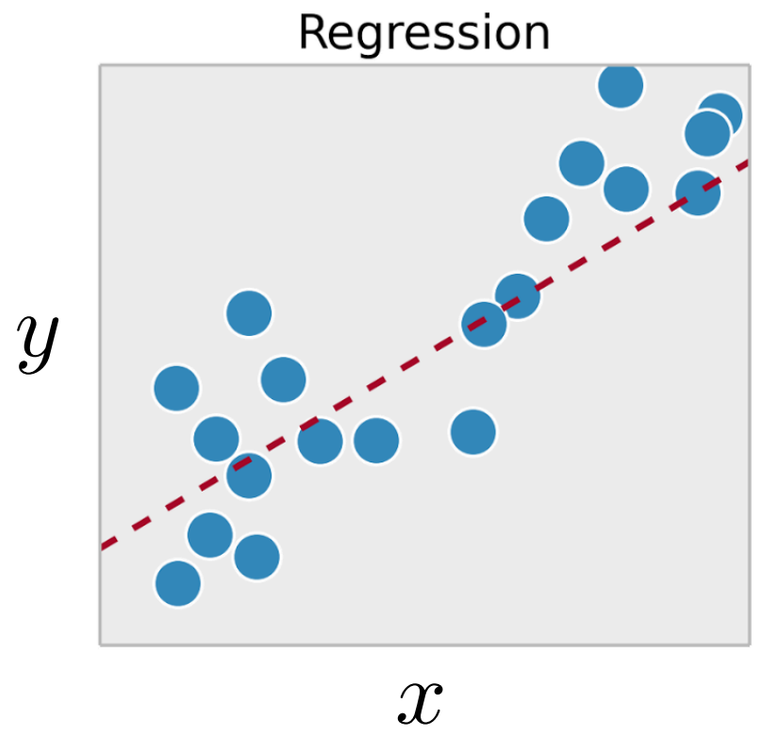
- Most important

https://quantummc.xyz/hku-phys3151-machine-learning-in-physics-2023/

Multivariate Linear Regression

Regression: $y = h_\Theta(x) = \Theta \cdot x$

Statistics	Machine Learning	Notation	Remarks
independent variable	feature	$x_j^{(i)}$	$j = 1, \cdots, N$
dependent variable	outcome	$y^{(i)}$	
sample	example	$(x^{(i)}, y^{(i)})$	$i = 1, \cdots, M$
model	hypothesis	$h_\theta(x)$	
parameter	parameter	θ_j	
intercept	bias	θ_0	



$$\{(x_j^{(i)}, y^{(i)}), \theta_j\};\ j = 1, 2, \cdots, N;\ i = 1, 2, \cdots, M; N < M$$

$$y^{(i)} = \theta_0 + \theta_1 x_1^{(i)} + \theta_2 x_2^{(i)} + \cdots + \theta_N x_N^{(i)}$$

$$\mathbb{R}^{M \times (N+1)} \qquad \begin{bmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \cdots & x_N^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & \cdots & x_N^{(2)} \\ & \vdots & & \vdots & \\ 1 & x_1^{(M)} & x_2^{(M)} & \cdots & x_N^{(M)} \end{bmatrix} \cdot \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_N \end{bmatrix} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(M)} \end{bmatrix}$$

$$\underline{\underline{X}} \cdot \underline{\Theta} = \underline{Y}$$

Multivariate Linear Regression

Basic assumption: examples are independent and identically distributed (i.i.d.).

Likehood function

$$\mathcal{L}(\Theta|X) = \sum_{i=1}^{M} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(y^{(i)} - \theta^T \cdot x^{(i)})^2}{2\sigma^2})$$

$$\Theta^* = \arg\min_{\Theta}\{\sum_{i=1}^{M} \frac{(y^{(i)} - \theta^T \cdot x^{(i)})^2}{2M}\}$$

Cost / Loss function

$$J(\Theta) = \frac{1}{2M} \sum^{M} (h_\theta(x^{(i)}) - y^{(i)})^2$$

Convex optimisation

$$\begin{aligned}
\Delta_\theta J(\theta) &= \nabla_\theta \cdot \nabla_\theta J(\theta) \\
&= \nabla_\theta \cdot \nabla_\theta \left[\frac{1}{2M} \sum_{i=1}^{M} \left(\theta^T x^{(i)} - y^{(i)} \right)^2 \right] \\
&= \nabla_\theta \cdot \left[\frac{1}{M} \sum_{i=1}^{M} \left(\theta^T x^{(i)} - y^{(i)} \right) x^{(i)} \right] \\
&= \frac{1}{M} \sum_{i=1}^{M} \left| x^{(i)} \right|^2 \geq 0
\end{aligned}$$

There exist global minimum, our goal is to find it.

Multivariate Linear Regression

Feature Scaling：

$$x_j^{(i)} := \frac{x_j^{(i)} - \mu_j^x}{\sigma_j^x} \qquad \mu_j^x = \frac{1}{M}\sum_{i=1}^{M} x_j^{(i)} \qquad \sigma_j^x = \sqrt{\frac{\sum_{i=1}^{M}(x_j^{(i)} - \mu_j^x)^2}{M}}$$

Ensure all the features θ_j lie in the same range

$$J(\theta) = \frac{1}{2M}\sum_{i=1}^{M}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2 = \frac{1}{2M}|X\Theta - Y|^2$$

$$= \frac{1}{2M}(\Theta^T X^T - Y^T)(X\Theta - Y)$$

$$= \frac{1}{2}(\Theta^T \frac{1}{M} X^T X\Theta - 2Y^T X\Theta + Y^T Y)$$

$$= \frac{1}{2}\Theta^T Q\Theta - b^T\Theta + c$$

$$Q = \frac{1}{M}X^T X \geq 0$$

$$b = \frac{1}{M}X^T Y$$

$$c = \frac{1}{2M}Y^T Y$$

Paraboloid

Multivariate Linear Regression

Structure of $$Q = \frac{1}{M} X^T X$$

$$\frac{1}{M} \underbrace{\begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1^{(1)} & x_1^{(2)} & \cdots & x_1^{(M)} \\ x_2^{(1)} & x_2^{(2)} & \cdots & x_2^{(M)} \\ \vdots & \vdots & \ddots & \vdots \\ x_N^{(1)} & x_N^{(2)} & \cdots & x_N^{(M)} \end{bmatrix}}_{\mathbb{R}^{(N+1)\times M}} \cdot \underbrace{\begin{bmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \cdots & x_N^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & \cdots & x_N^{(2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_1^{(M)} & x_2^{(M)} & \cdots & x_N^{(M)} \end{bmatrix}}_{\mathbb{R}^{M\times(N+1)}}$$

$$= \underbrace{\begin{bmatrix} \frac{\sum_{i=1}^M 1}{M} & \frac{\sum_{i=1}^M x_1^{(i)}}{M} & \frac{\sum_{i=1}^M x_2^{(i)}}{M} & \cdots & \frac{\sum_{i=1}^M x_N^{(i)}}{M} \\ \frac{\sum_{i=1}^M x_1^{(i)}}{M} & \frac{\sum_{i=1}^M \left(x_1^{(i)}\right)^2}{M} & \frac{\sum_{i=1}^M x_1^{(i)} x_2^{(i)}}{M} & \cdots & \frac{\sum_{i=1}^M x_1^{(i)} x_N^{(i)}}{M} \\ \frac{\sum_{i=1}^M x_2^{(i)}}{M} & \frac{\sum_{i=1}^M x_2^{(i)} x_1^{(i)}}{M} & \frac{\sum_{i=1}^M \left(x_2^{(i)}\right)^2}{M} & \cdots & \frac{\sum_{i=1}^M x_2^{(i)} x_N^{(i)}}{M} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\sum_{i=1}^M x_N^{(i)}}{M} & \frac{\sum_{i=1}^M x_N^{(i)} x_1^{(i)}}{M} & \frac{\sum_{i=1}^M x_N^{(i)} x_2^{(i)}}{M} & \cdots & \frac{\sum_{i=1}^M \left(x_N^{(i)}\right)^2}{M} \end{bmatrix}}_{\mathbb{R}^{(N+1)\times(N+1)}} = \underbrace{\begin{bmatrix} \langle 1 \rangle & \langle x_1 \rangle & \langle x_2 \rangle & \cdots & \langle x_N \rangle \\ \langle x_1 \rangle & \langle \left(x_1\right)^2 \rangle & \langle x_1 x_2 \rangle & \cdots & \langle x_1 x_N \rangle \\ \langle x_2 \rangle & \langle x_2 x_1 \rangle & \langle \left(x_2\right)^2 \rangle & \cdots & \langle x_2 x_N \rangle \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \langle x_N \rangle & \langle x_N x_1 \rangle & \langle x_N x_2 \rangle & \cdots & \langle \left(x_N\right)^2 \rangle \end{bmatrix}}_{\mathbb{R}^{(N+1)\times(N+1)}}$$

Symmetric and positive semi-definite

Multivariate Linear Regression

Positive semi-definite $$v^T Q v = v^T X^T X v = (Xv)^T (Xv) = u^T u \geq 0$$

$$J(\theta) = \frac{1}{2M} \sum_{i=1}^{M} \left(h_\theta \left(x^{(i)} \right) - y^{(i)} \right)^2 = \frac{1}{2M} |X\Theta - Y|^2$$

$$\nabla_\theta J(\theta) = \frac{1}{M} \left(\left(X^T X \right) \theta - \left(X^T Y \right) \right)$$

$$Q = \frac{1}{M} X^T X \qquad b = X^T Y$$

Normal Equation $$Q\theta \equiv \frac{1}{M} \left(X^T X \right) \theta = \frac{1}{M} X^T Y$$

$$\theta = \left(X^T X \right)^{-1} X^T Y = Q^{-1} b$$

$$O((N+1)^3)$$

Multivariate Linear Regression

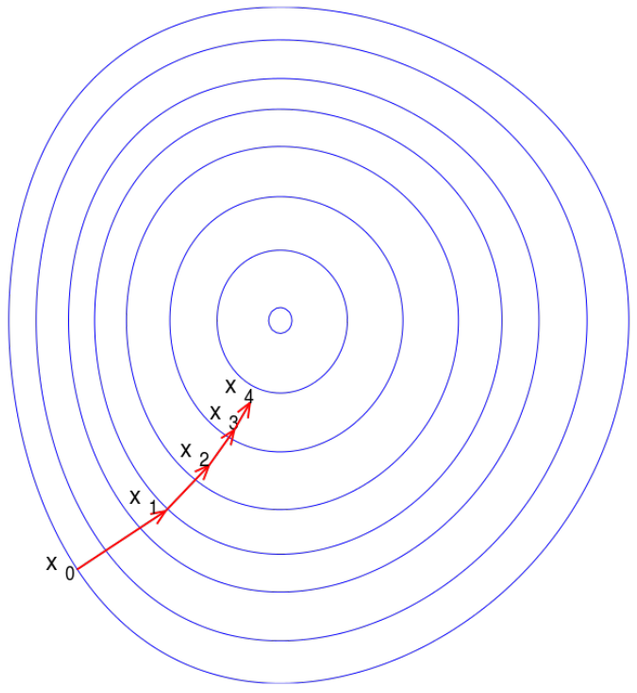
Gradient descent method

Algorithm:

1. Set a precision ϵ, learning rate α, and set the initial guess θ_0
2. Let $\theta_{j+1} := \theta_j - \alpha \nabla_\theta J(\theta_j)$
3. Calculate $J(\theta_{j+1})$
   - If $|\theta_{j+1} - \theta_j| < \epsilon$, stop and return θ_{j+1}
   - Else, return to step 2

The algorithm is based on the Taylor expansion

$$J(\theta_{j+1}) = J(\theta_j - \alpha \nabla_\theta J(\theta_j)) = J(\theta_j) - \alpha (\nabla_\theta J(\theta_j))^2 + O(\alpha^2)$$



Multivariate Linear Regression

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

Good reference:

Jonathan Richard Shewchuk

“I am trained to only sleep during national holidays.”

An Introduction to Conjugate Gradient Method Without the Agonizing Pain

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

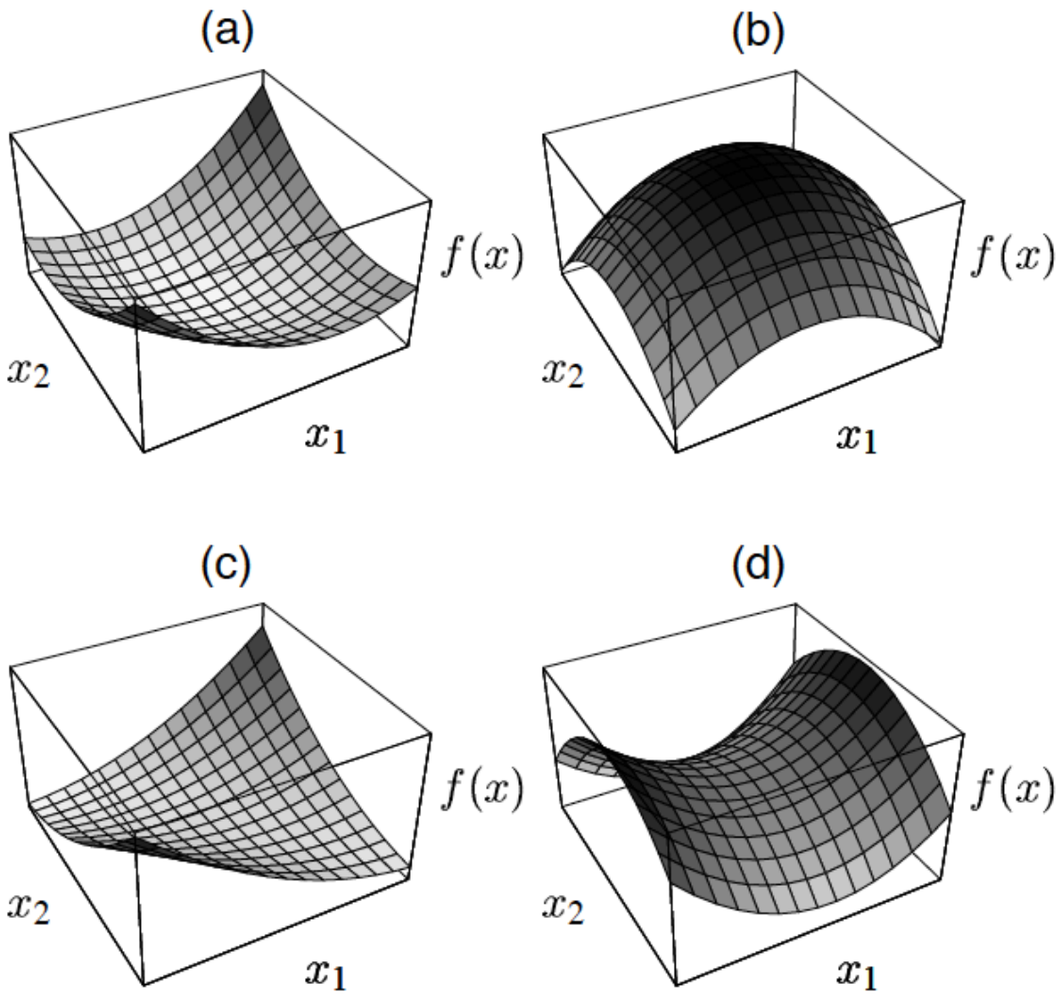
Paraboloid and positive-definite

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

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

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

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

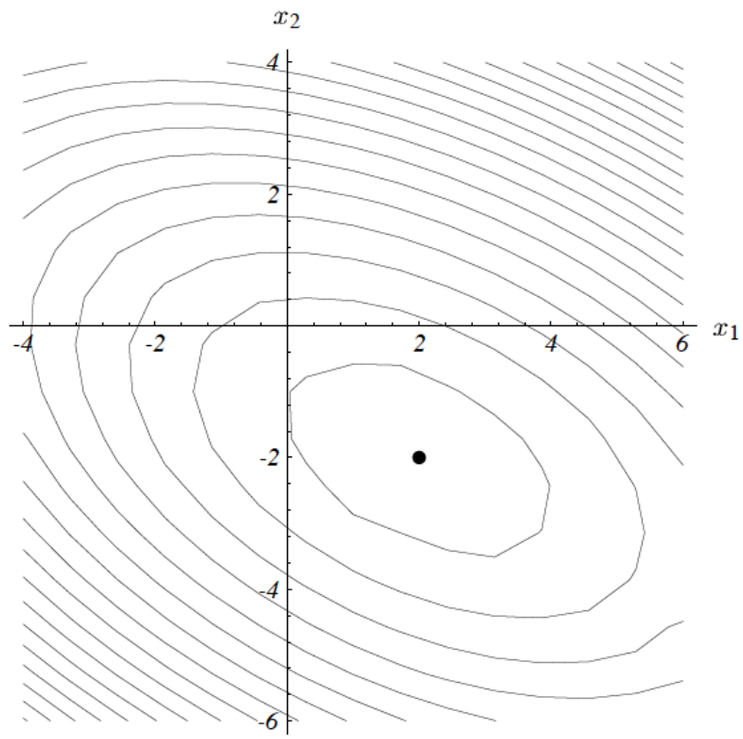


(a) positive-definite, symmetric

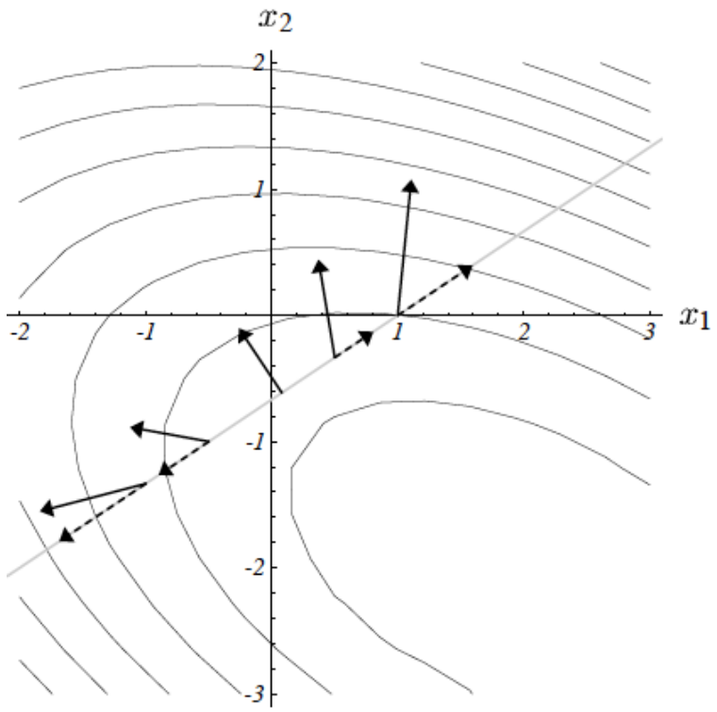
(b) negative-definite

(c) singular, a set of solution

(d) saddle point

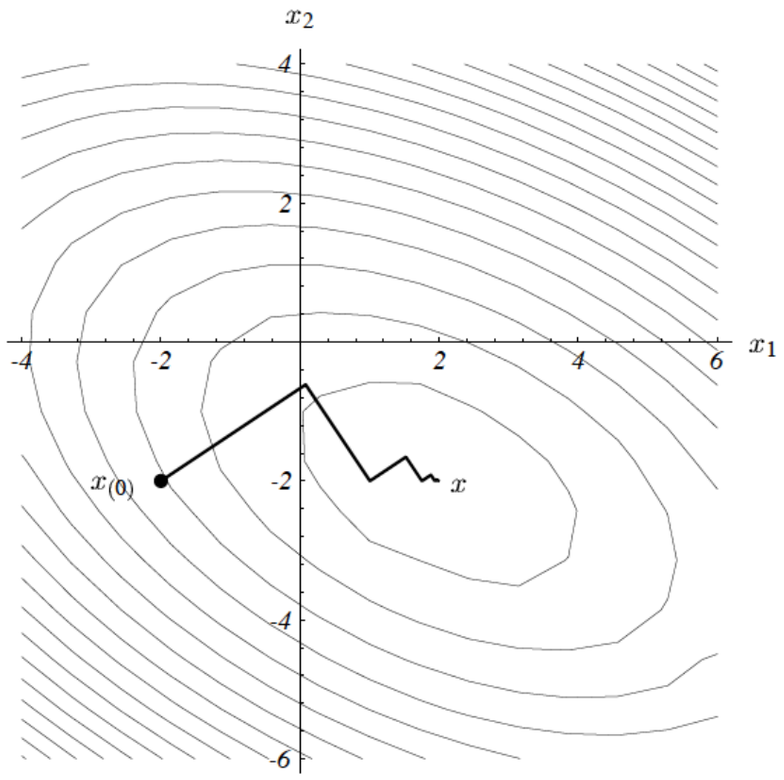


Steepest Descent



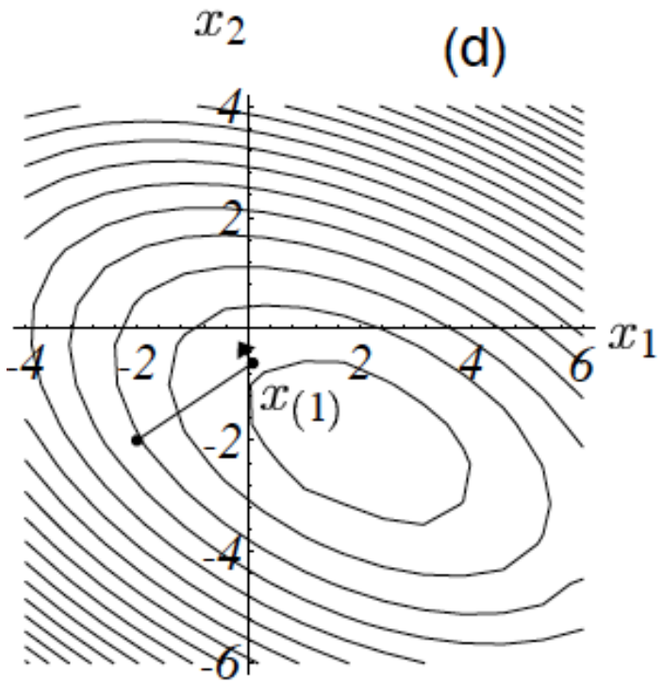
gradient

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

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

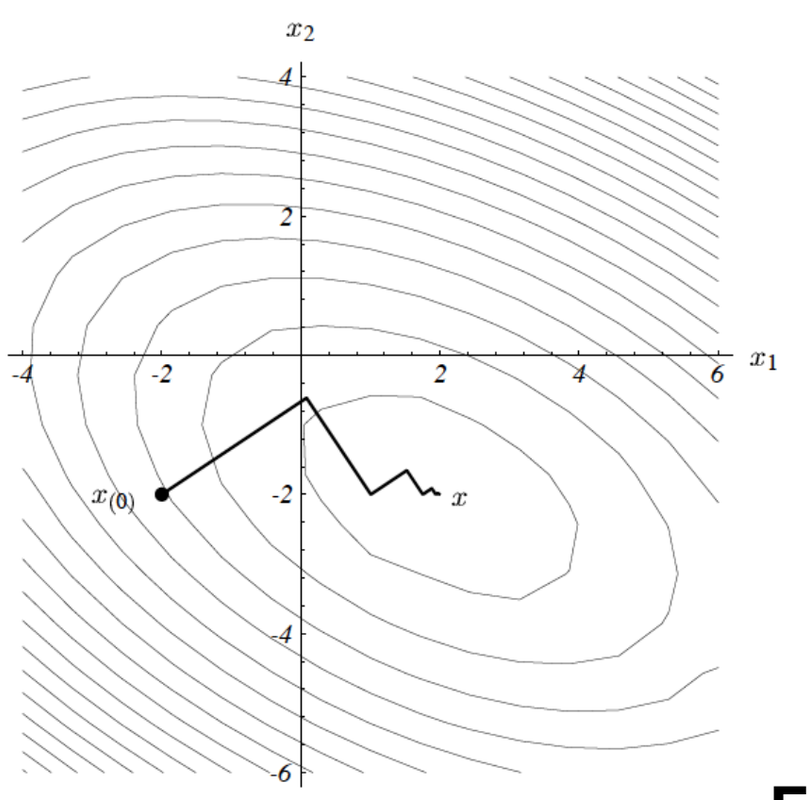
$$r_{(0)} = [12, 8]^T$$

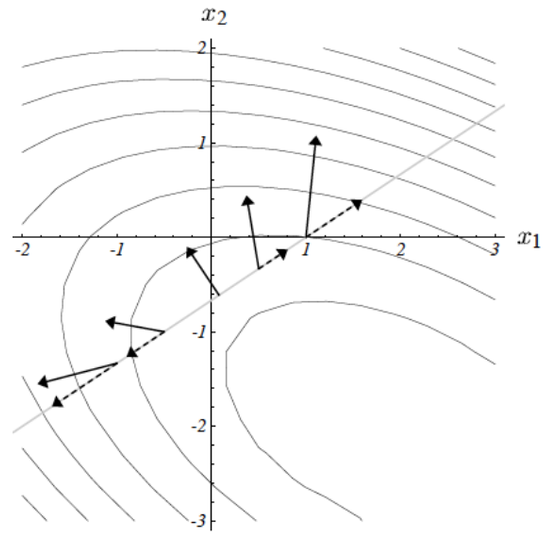
$$r_{(0)}^T r_{(0)} = 208$$

$$r_{(0)}^T A r_{(0)} = 1200$$

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

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





$$r_{(i+1)}^T r_{(i)} = 0$$

$$(b - Ax_{(i+1)})^T r_{(i)} = 0$$

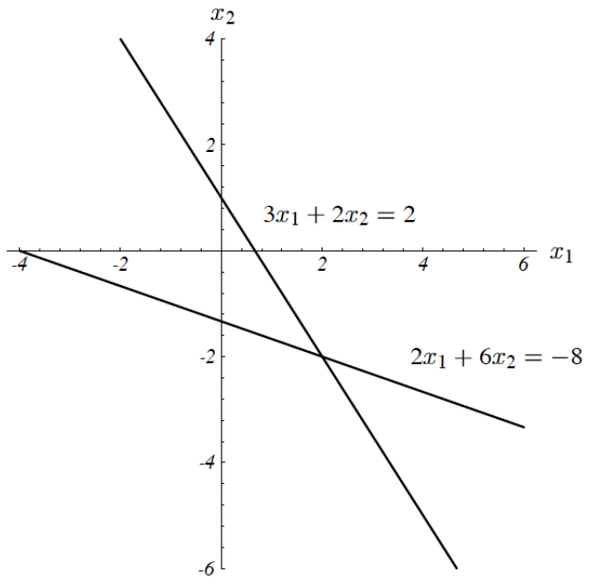
$$(b - A(x_{(i)} + \alpha r_{(i)}))^T r_{(i)} = 0$$

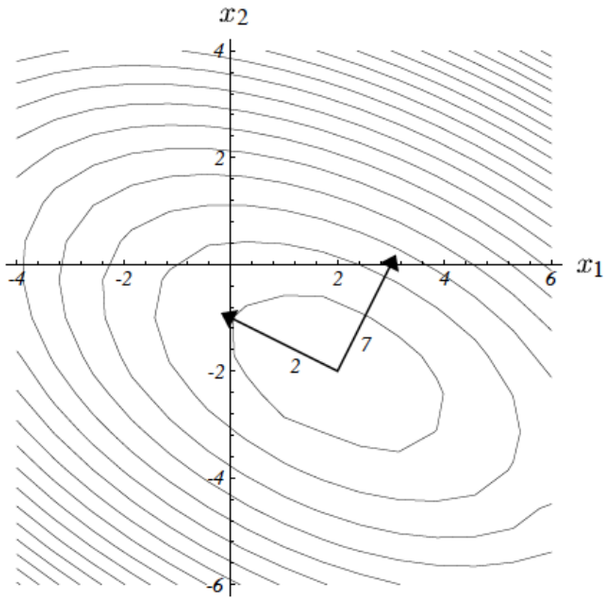
$$(b - AX_{(i)})^T r_{(i)} = \alpha (Ar_{(i)})^T r_{(i)}$$

$$r_{(i)}^T r_{(i)} = \alpha r_{(i)}^T A r_{(i)}$$

$$\alpha = \frac{r_{(i)}^T r_{(i)}}{r_{(i)}^T A r_{(i)}}$$

Each gradient is orthogonal to the previous gradient





Let's eigen do it

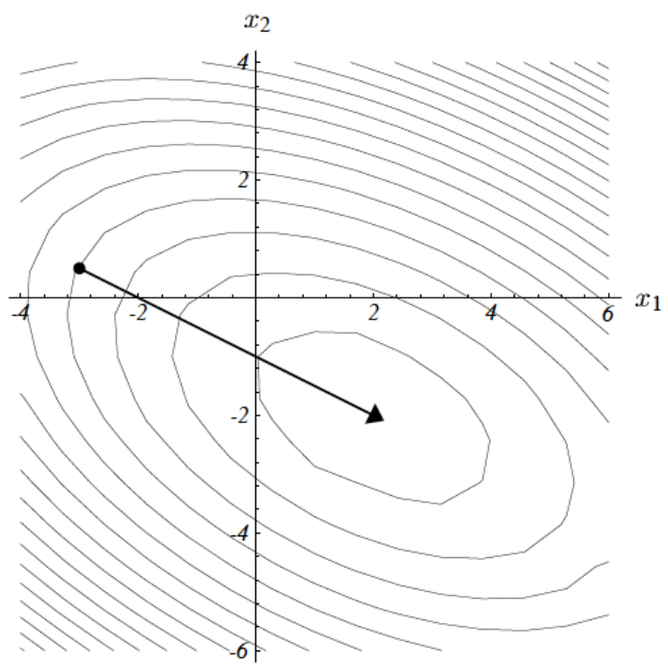
$$f(x) = \frac{1}{2}x^T Ax - b^T x + c$$

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

Eigenvalues and eigenvectors:

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

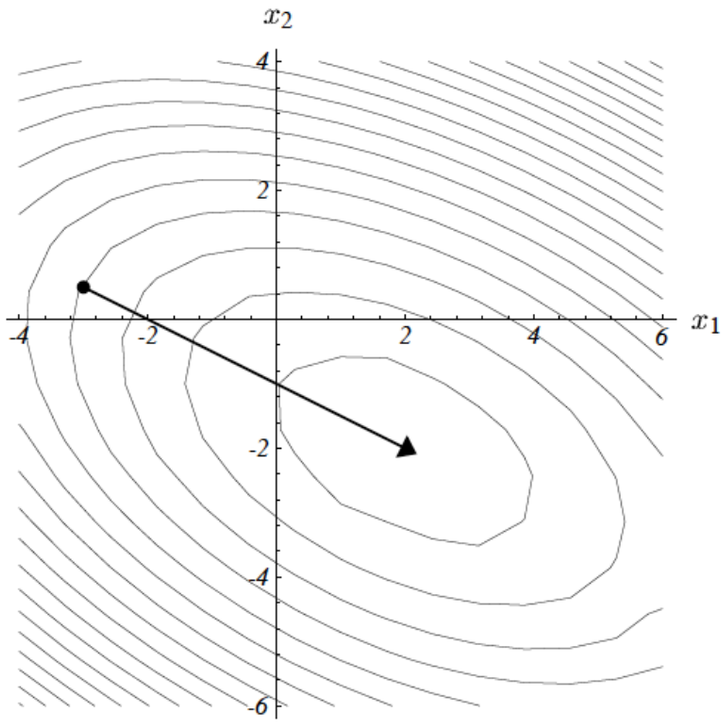
If r_{(i)} is the eigenvector, only one step to converge to the exact solution



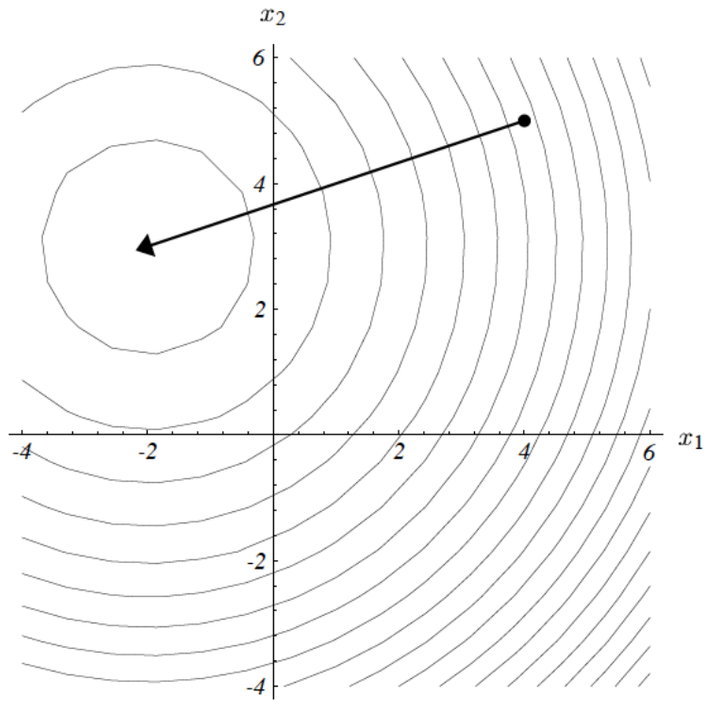
$$Ar_{(i)} = \lambda r_{(i)} \quad \alpha = 1/\lambda$$

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

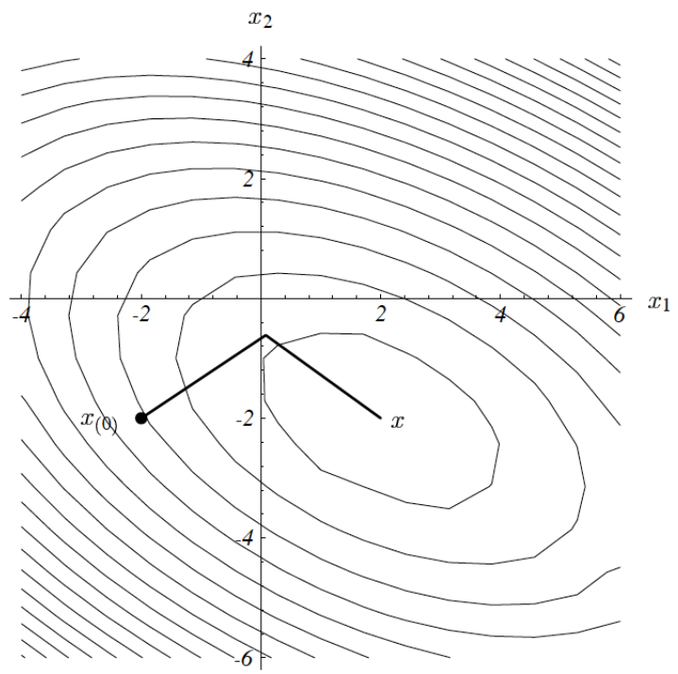
"Eigenvectors are useful tools, and not just bizarre torture devices inflicted on you by your professors for the pleasure of watching you suffer (although the latter is a nice fringe benefit)"



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

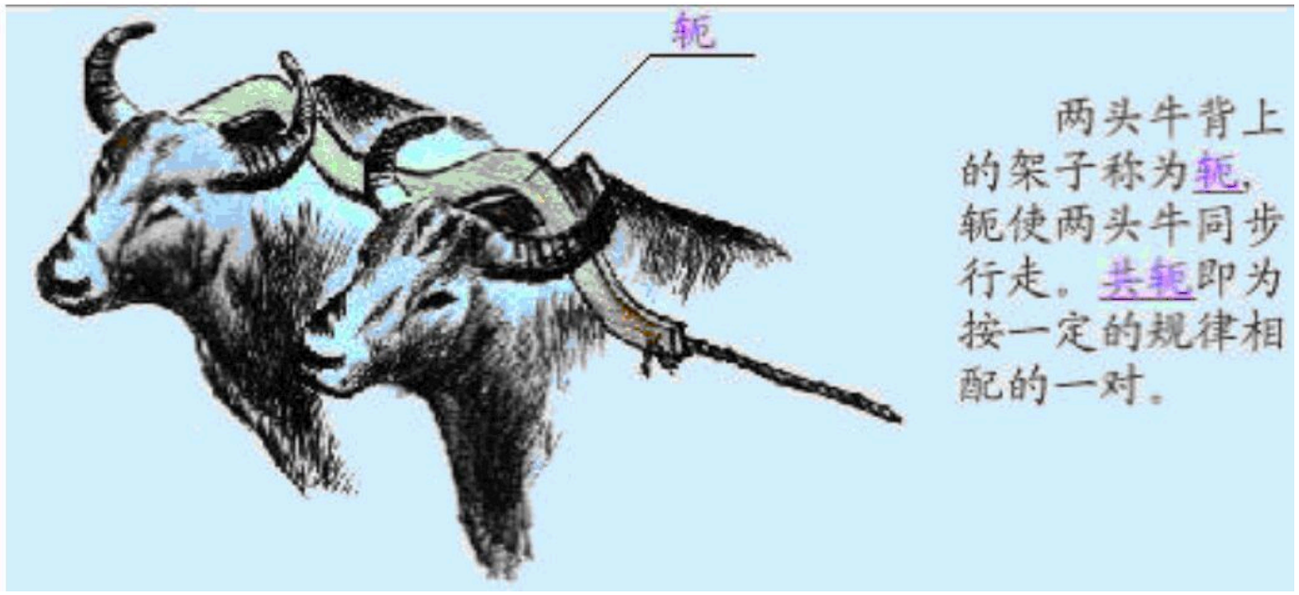


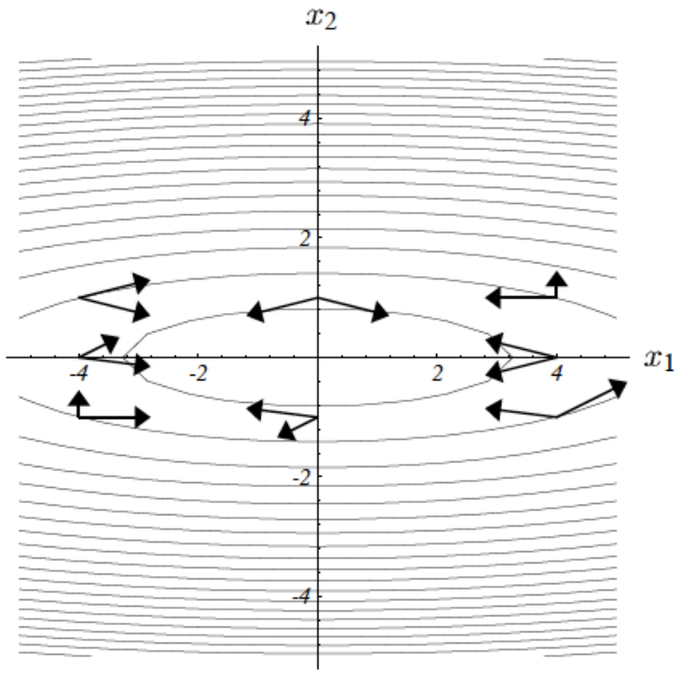
Paraboloid is spherical, no matter what point we start, always find the minimal



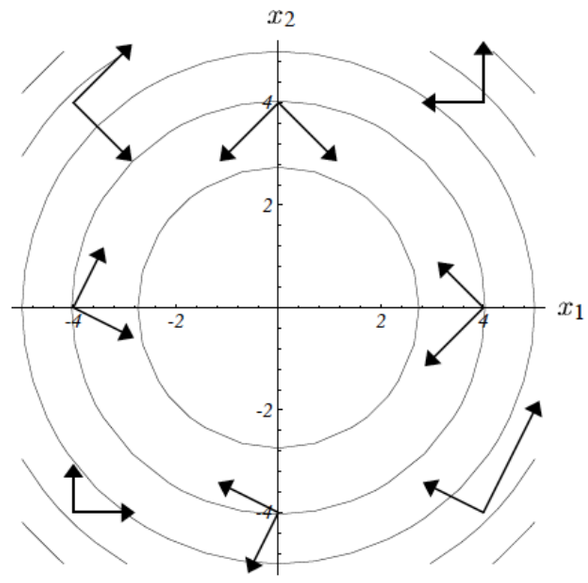
CG is to find the orthogonal directions in a stretched (scaled) space.

Conjugate Gradient Method

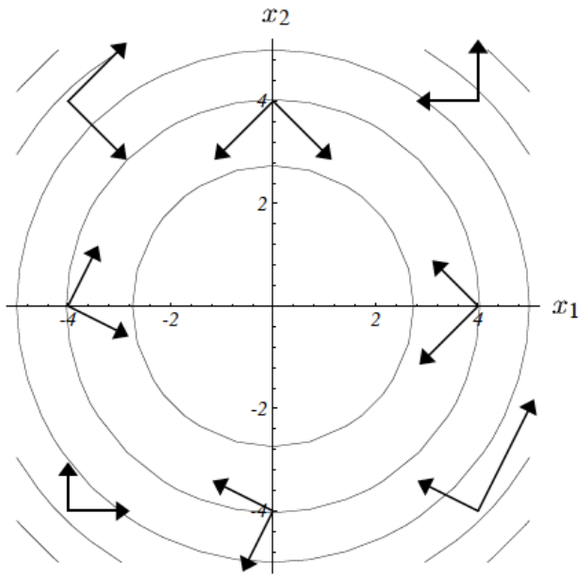




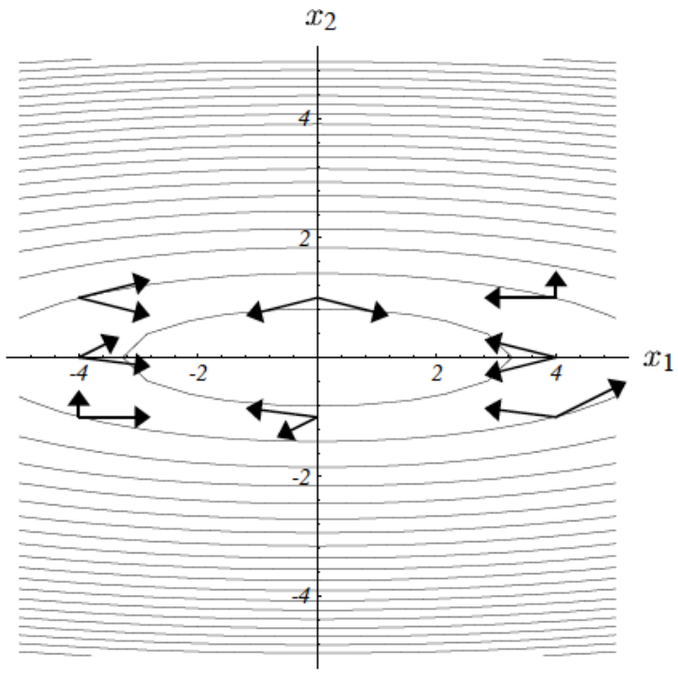
$$d_{(i)}^T A d_{(j)} = 0$$



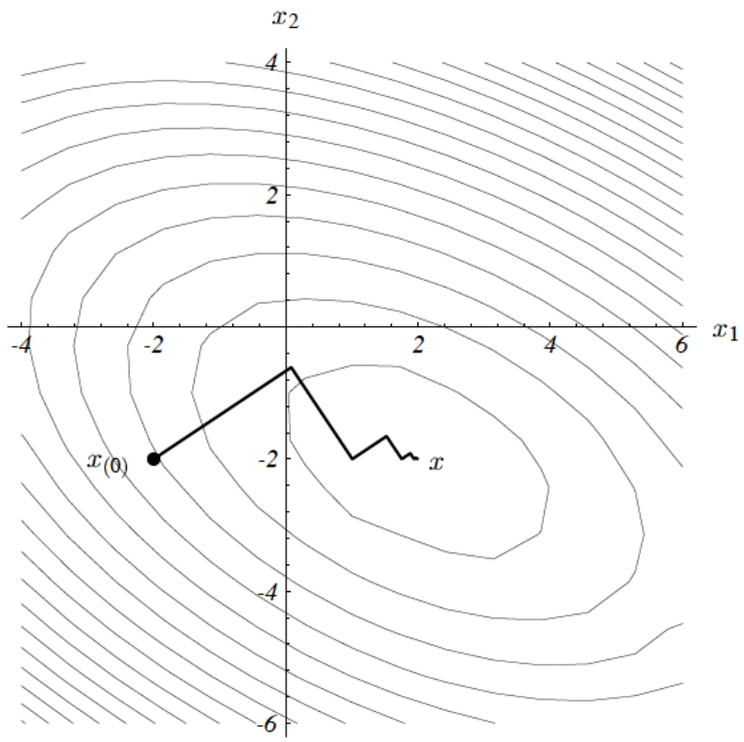
Pairs of vectors that are A-orthogonal, conjugate



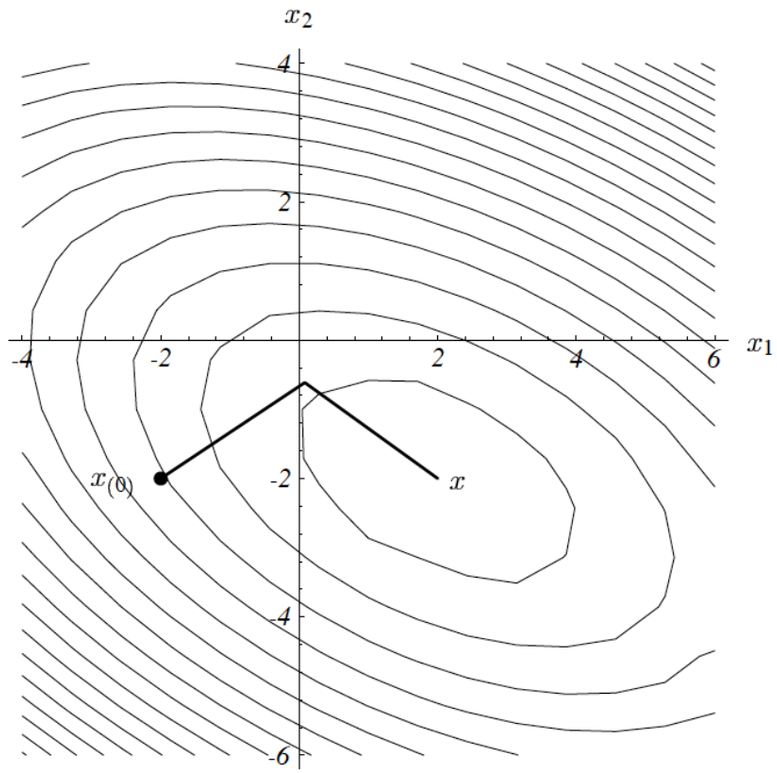
Pairs of vectors that are orthogonal



Pairs of vectors that are A-orthogonal, conjugate



Many steps to find the solution



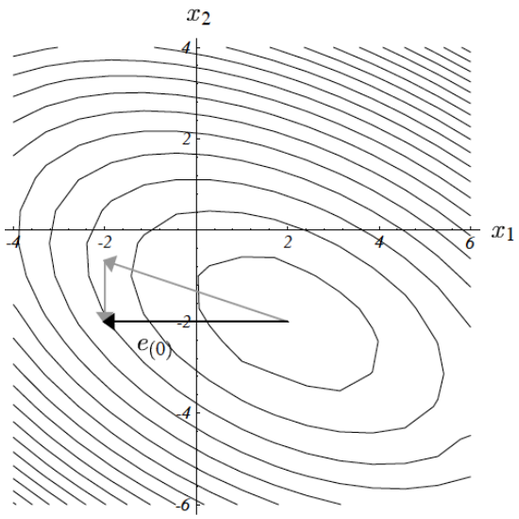
Only takes N-steps to find the solution

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

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

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

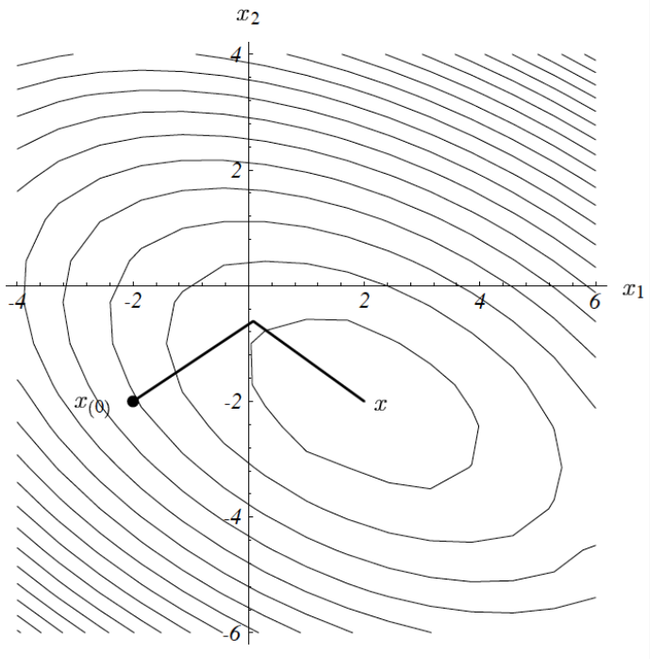
Decompose a vector x* to a sum of A-orthogonal components



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

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

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



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

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

Repeat from k=0 $\qquad \alpha_{(k)} = \frac{r_{(k)}^T r_{(k)}}{P_{(k)}^T AP_{(k)}}$

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

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

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

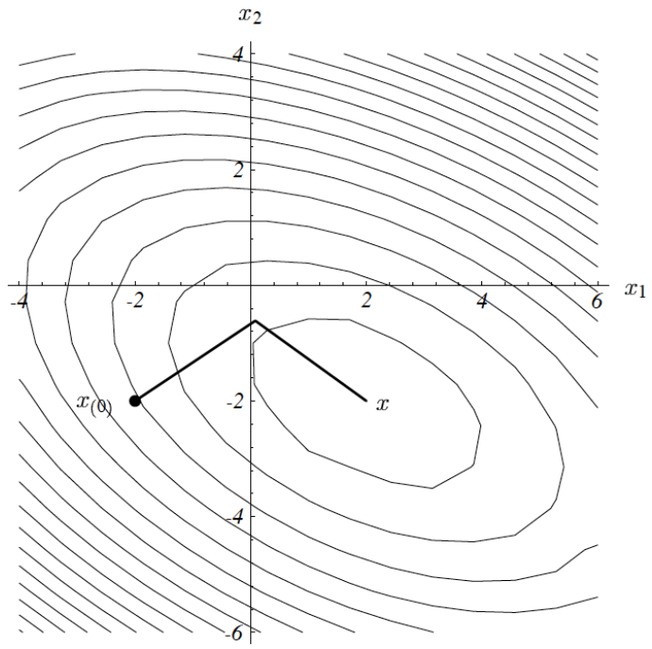
$$P_{(k+1)} = r_{(k+1)} + \beta_{(k)} P_{(k)}$$

$$k = k + 1$$

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

$A = \begin{bmatrix} 3 & 2 \\ 2 & 6 \end{bmatrix}$ $b = \begin{bmatrix} 2 \\ -8 \end{bmatrix}$ $n_1 = 7 \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ $r_{(1)}^T \cdot r_{(0)} = 0$

$n_2 = 2 \begin{bmatrix} -2 \\ 1 \end{bmatrix}$

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

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

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

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

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

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

$P_{(1)} = r_{(1)} + \beta_{(0)} P_{(0)} = \begin{bmatrix} \frac{224}{75} \\ -\frac{336}{75} \end{bmatrix} + 0.1394 \begin{bmatrix} 12 \\ 8 \end{bmatrix} = \begin{bmatrix} 4.6947 \\ -3.3648 \end{bmatrix}$

$\alpha_{(1)} = \frac{r_{(1)}^T r_{(1)}}{P_{(1)}^T A P_{(1)}} = \frac{163072}{5625} = 0.4128$

$\begin{pmatrix} 4.6947 & -3.3648 \end{pmatrix} \begin{pmatrix} 3 & 2 \\ 2 & 6 \end{pmatrix} \begin{pmatrix} 4.6947 \\ -3.3648 \end{pmatrix} = 70.3505$

$X_{(2)} = X_{(1)} + \alpha_{(1)} P_{(1)} = \begin{bmatrix} \frac{2}{25} \\ -\frac{46}{75} \end{bmatrix} + 0.4128 \begin{bmatrix} 4.6947 \\ -3.3648 \end{bmatrix} = \begin{bmatrix} 2 \\ -2 \end{bmatrix}$

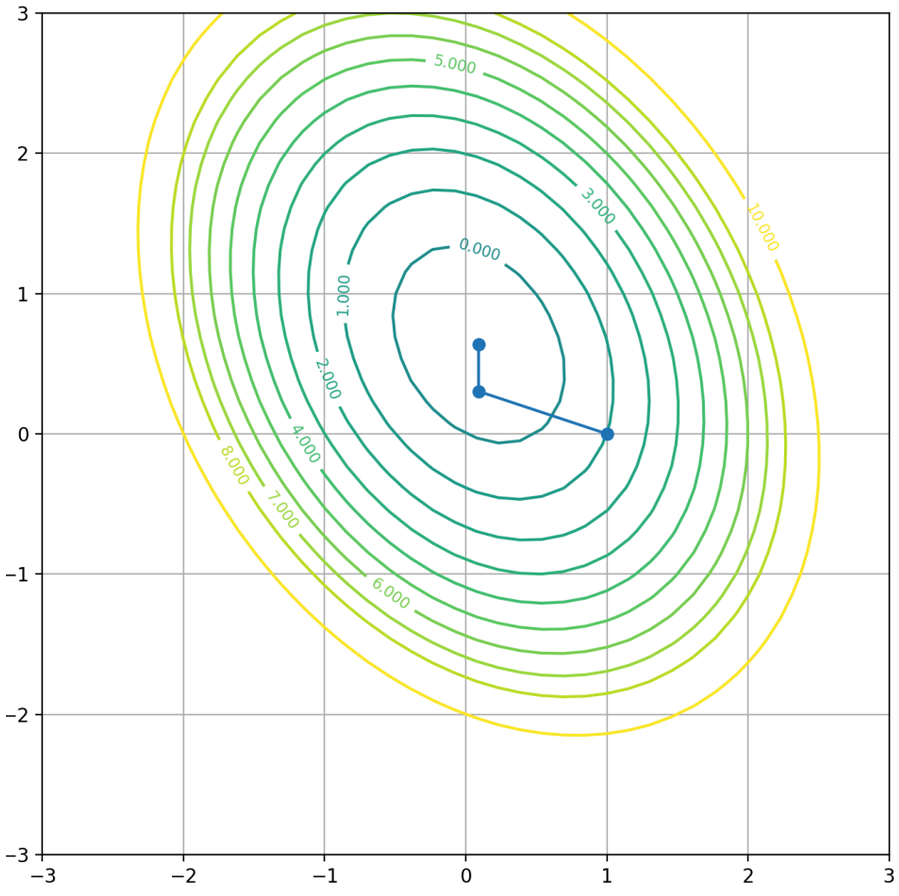
$(7.2481 \quad -10.8626)$

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

$P_{(1)}^T A P_{(0)} = \begin{bmatrix} 4.6947 & -3.3648 \end{bmatrix} \begin{bmatrix} 3 & 2 \\ 2 & 6 \end{bmatrix} \begin{bmatrix} 12 \\ 8 \end{bmatrix} = \begin{bmatrix} 7.2481 & -10.8626 \end{bmatrix} \begin{bmatrix} 12 \\ 8 \end{bmatrix}$

$= 0.02684$ ✓

round off errors



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

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

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

$$\alpha_0 = \frac{\underset{\sim}{r}_0^T \cdot \underset{\sim}{r}_0}{\underset{\sim}{p}_0^T \cdot \underset{=}{A} \cdot \underset{\sim}{p}_0} = \frac{\begin{bmatrix} -3 & 1 \end{bmatrix}\cdot\begin{bmatrix} -3 \\ 1 \end{bmatrix}}{\begin{bmatrix} -3 & 1 \end{bmatrix}\begin{bmatrix} 4 & 1 \\ 1 & 3 \end{bmatrix}\begin{bmatrix} -3 \\ 1 \end{bmatrix}} = \frac{10}{33}$$

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

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

$$\beta_0 = \frac{\underset{\sim}{r}_1^T \cdot \underset{\sim}{r}_1}{\underset{\sim}{r}_0^T \cdot \underset{\sim}{r}_0} = \frac{\begin{bmatrix} \frac{1}{3} & 1 \end{bmatrix}\begin{bmatrix} \frac{1}{3} \\ 1 \end{bmatrix}}{\begin{bmatrix} -3 & 1 \end{bmatrix}\begin{bmatrix} -3 \\ 1 \end{bmatrix}} = \frac{1}{9}$$

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

$$\alpha_1 = \frac{\underset{\sim}{r}_1^T \cdot \underset{\sim}{r}_1}{\underset{\sim}{p}_1^T \cdot \underset{=}{A} \cdot \underset{\sim}{p}_1} = \frac{\begin{bmatrix} \frac{1}{3} & 1 \end{bmatrix}\begin{bmatrix} \frac{1}{3} \\ 1 \end{bmatrix}}{\begin{bmatrix} 0 & \frac{10}{9} \end{bmatrix}\begin{bmatrix} 4 & 1 \\ 1 & 3 \end{bmatrix}\begin{bmatrix} 0 \\ \frac{10}{9} \end{bmatrix}} = \frac{30}{10}$$

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

$$\underset{\sim}{r}_2 = \underset{\sim}{r}_1 - \alpha_1 \underset{=}{A}\cdot p_1 = \begin{bmatrix} \frac{1}{3} \\ 1 \end{bmatrix} - \frac{3}{10}\begin{bmatrix} 4 & 1 \\ 1 & 3 \end{bmatrix}\begin{bmatrix} 0 \\ \frac{10}{9} \end{bmatrix} = \begin{bmatrix} \frac{1}{3} \\ 1 \end{bmatrix} - \frac{3}{10}\begin{bmatrix} \frac{10}{9} \\ \frac{10}{3} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

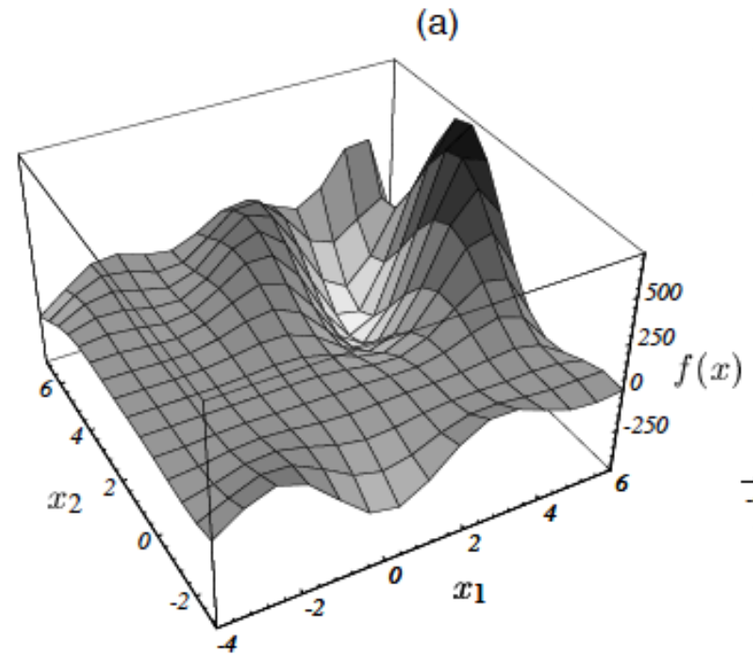
check : r_k, r_{k+1} orthogonal ✓

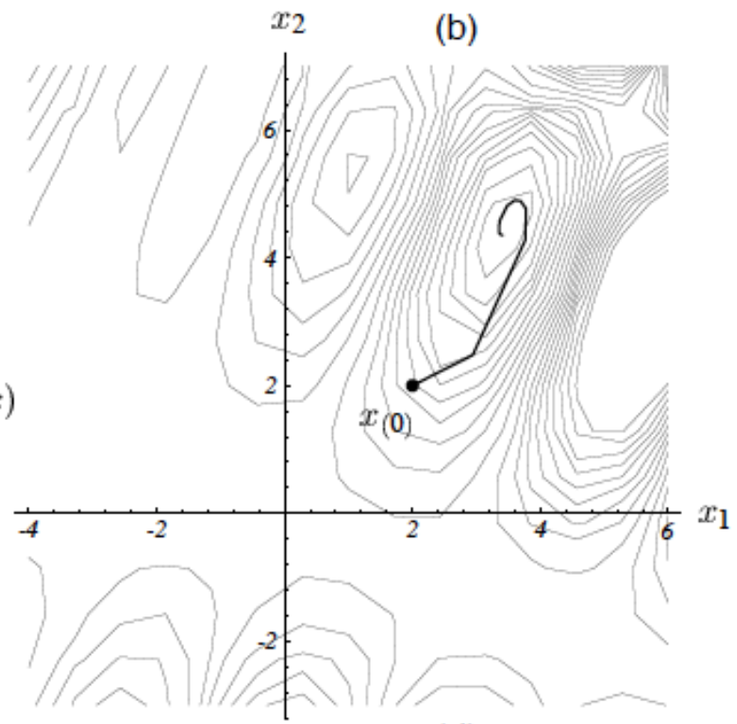
$P_k \cdot A P_{k+1}$ orthogonal ✓

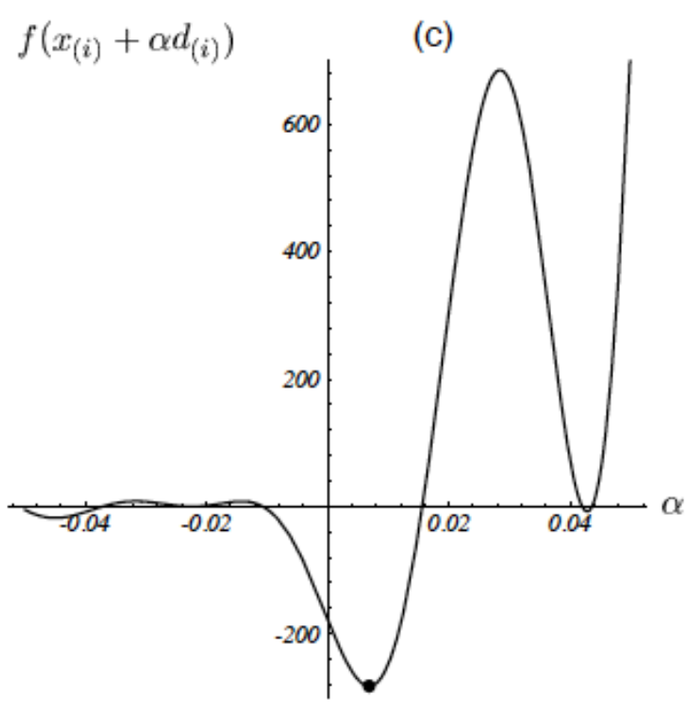
n - stop converge ✓

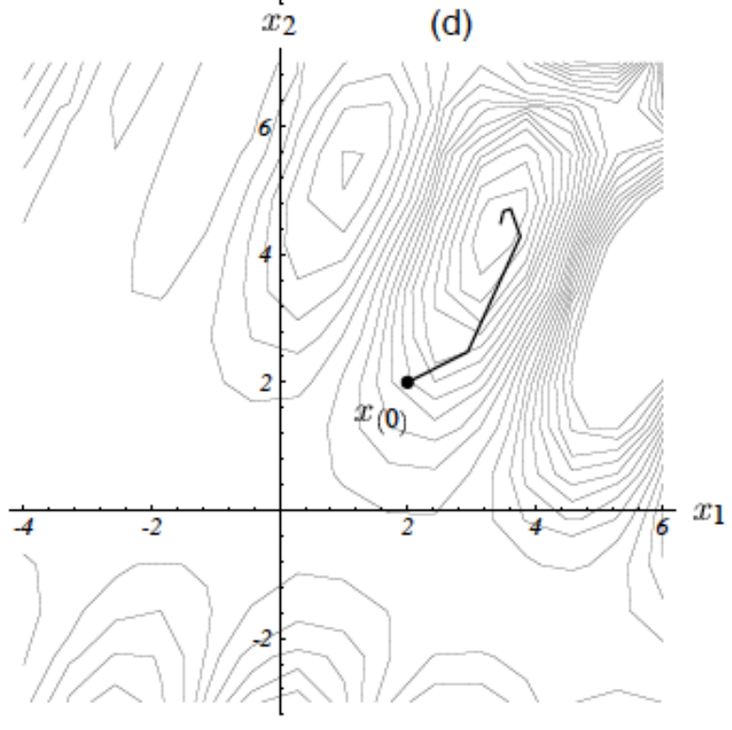
$$\underset{\sim}{r}_0^T \cdot \underset{\sim}{r}_1 = \begin{bmatrix} -3 & 1 \end{bmatrix}\begin{bmatrix} \frac{1}{3} \\ 1 \end{bmatrix} = 0$$

$$\underset{\sim}{p}_0^T \cdot \underset{=}{A}\,\underset{\sim}{p}_1 = \begin{bmatrix} -3 & 1 \end{bmatrix}\begin{bmatrix} 4 & 1 \\ 1 & 3 \end{bmatrix}\begin{bmatrix} 0 \\ \frac{10}{9} \end{bmatrix} = \begin{bmatrix} -11 & 0 \end{bmatrix}\begin{bmatrix} 0 \\ \frac{10}{9} \end{bmatrix}$$


(a)
500
250
0
-250
$f(x)$
6
4
2
0
-2
x_2
-4
-2
0
2
4
6
x_1


(b)
x_2
6
4
2
-2
$x_{(0)}$
-4
-2
2
4
6
x_1

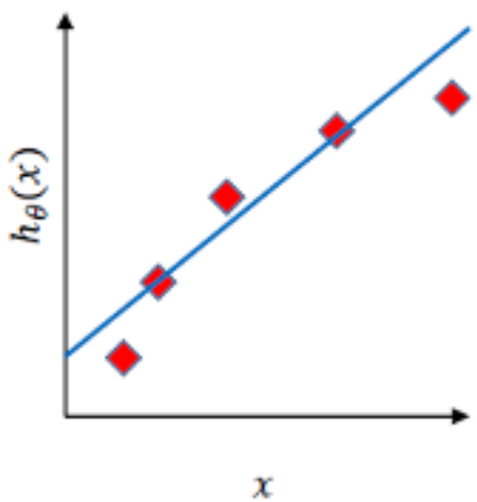

$f(x_{(i)} + \alpha d_{(i)})$
(c)
600
400
200
-200
-0.04
-0.02
0.02
0.04
α


x_2
(d)
6
4
2
-2
$x_{(0)}$
-4
-2
2
4
6
x_1

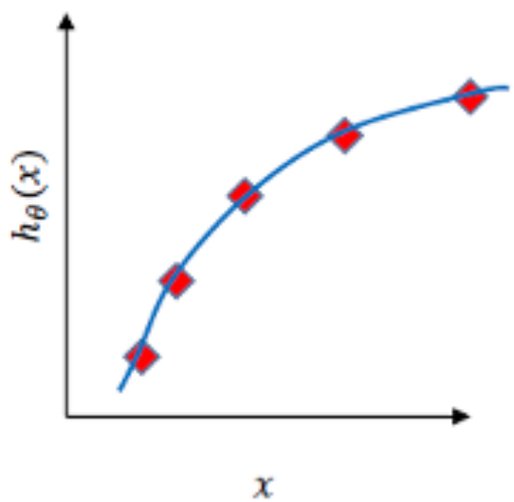
Content

The quality of fitting

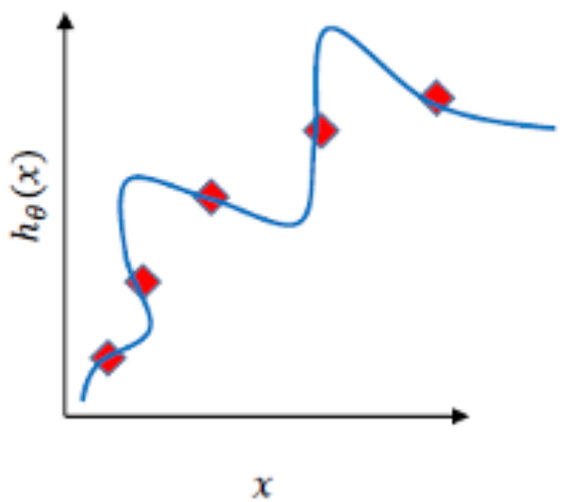
Underfitting (high bias)



High bias



Overfitting (high variance)

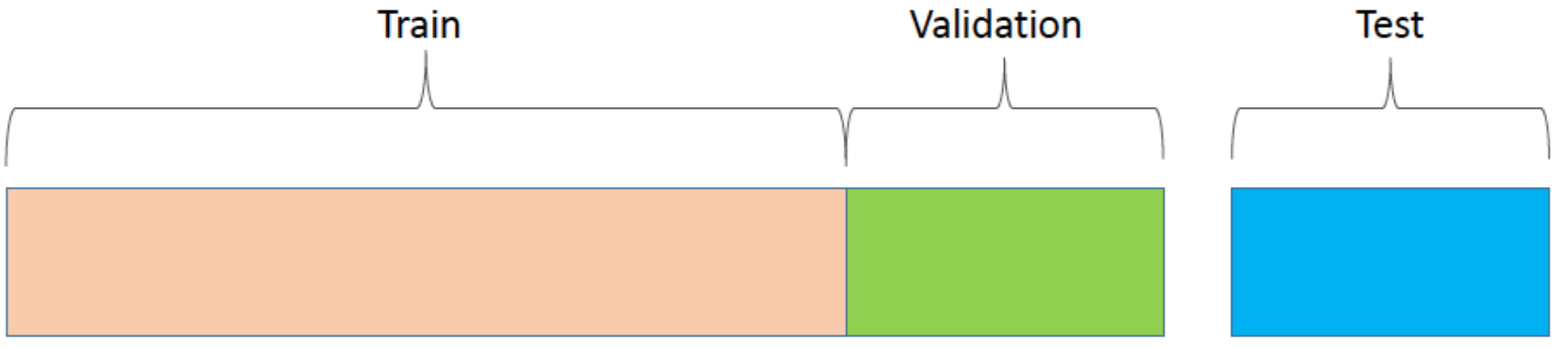


High variance

Polynomial regression

Forecastability

$$h_\theta^{[d]}(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \cdots + \theta_d x^d = \theta^T x$$

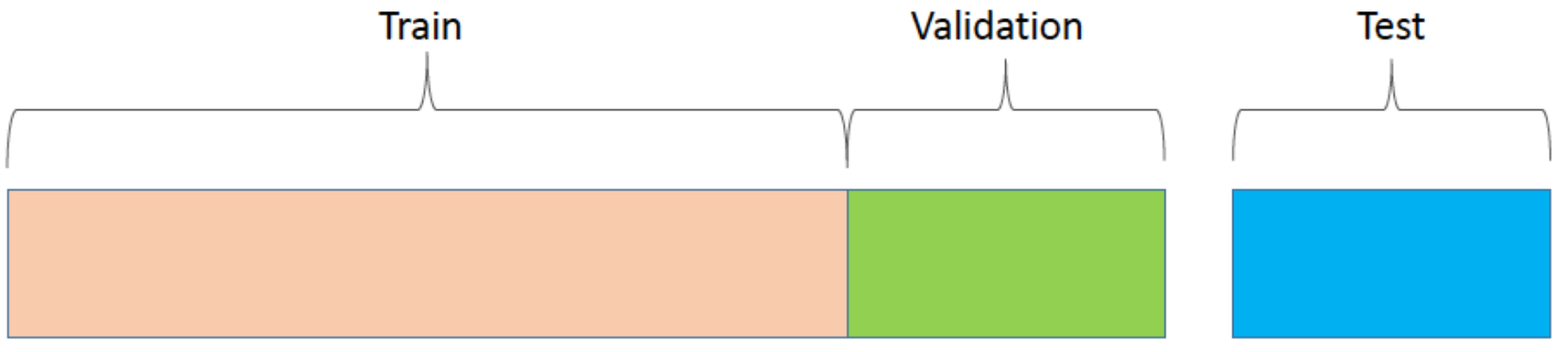


$$J_{\text{train}}(\theta) = \frac{1}{2M_{\text{train}}} \sum_{i_{\text{train}}=1}^{M_{\text{train}}} \left(y^{(i_{\text{train}})} - h_\theta(x)\right)^2$$

$$J_{\text{cv}}(\theta) = \frac{1}{2M_{\text{cv}}} \sum_{i_{\text{cv}}=1}^{M_{\text{cv}}} \left(y^{(i_{\text{cv}})} - h_\theta(x)\right)^2$$

$$J_{\text{test}}(\theta) = \frac{1}{2M_{\text{test}}} \sum_{i_{\text{test}}=1}^{M_{\text{test}}} \left(y^{(i_{\text{test}})} - h_\theta(x)\right)^2$$

Learning Curves and Regularization

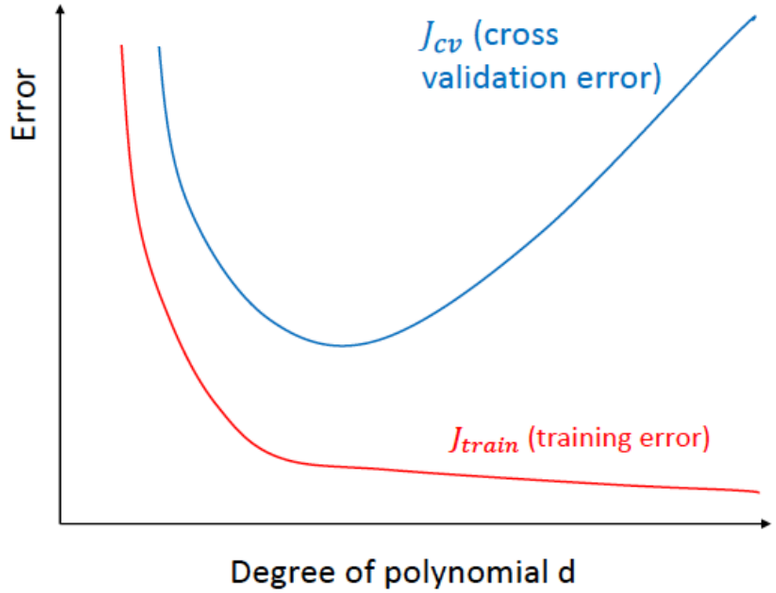


Forecastability

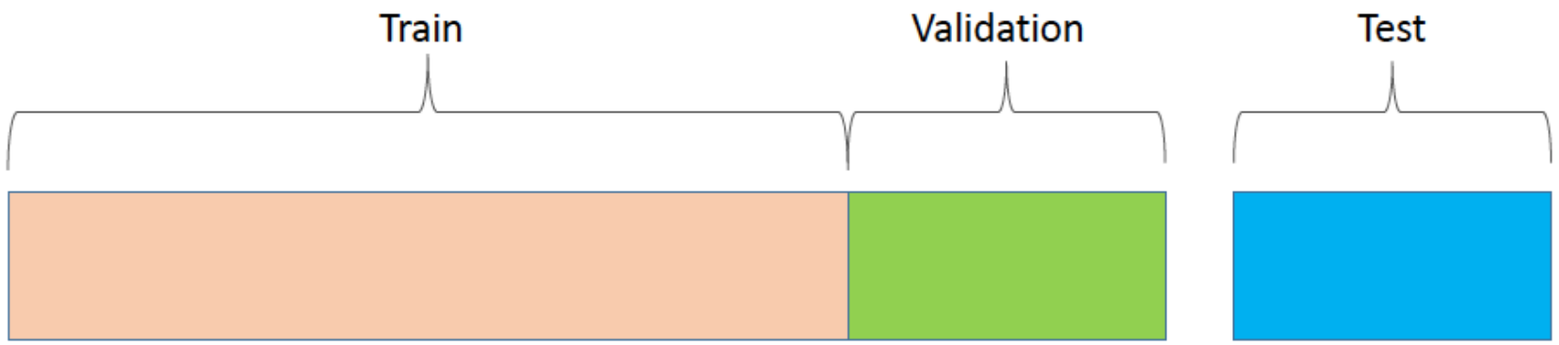
$$h_\theta^{[d]}(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \cdots + \theta_d x^d = \theta^T x$$

Polynomial regression

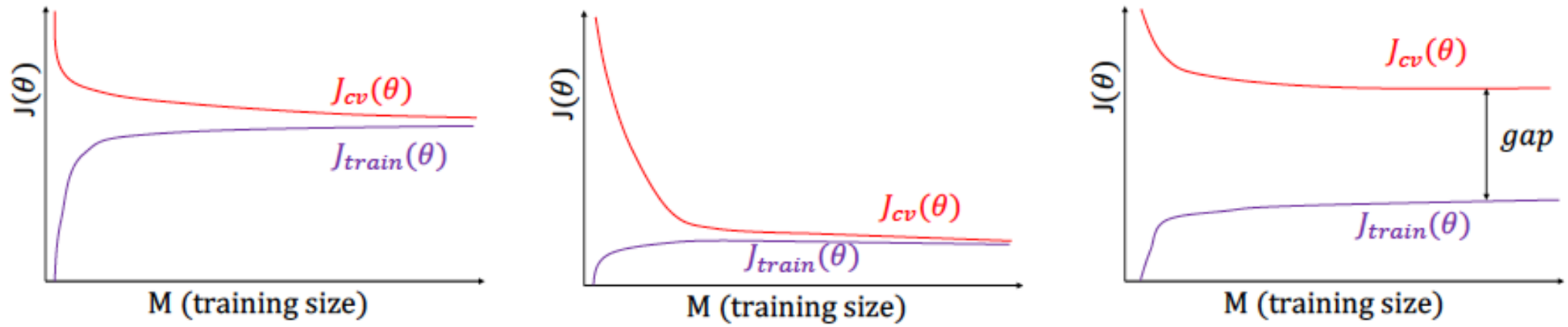
Model Selection



Learning Curves and Regularization



Learning curves



Regularization

$$J(\theta) = \frac{1}{2M}\left[\sum_{i=1}^{M}(h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda\sum_{j=1}^{N}\theta_j^2\right] \qquad \lambda \geq 0$$