Math Review – Background

- Linear Algebra
- Probabilities
- Information Theory
- Optimization

Linear Algebra

- Let $\Re^d$ be the *d*-dimensional Euclidean space
- **Vector:** An ordered tuple

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}$$

- The transpose of **x** is a $1 \times d$ *matrix*: $\mathbf{x}^t = [x_1 \ x_2 \ldots x_d]$
- **Inner (dot) product:** Given **x**, **y** $\in \Re^d$, $\mathbf{x}^t$ **y** given by

  resulting in a scalar that $\in \Re$

$$\mathbf{x}^t \ \mathbf{y} = \sum_{i=1}^{d} x_i y_i$$

Example (Matlab)

```
>> x = [1.2; 3.5; -.8; .3]
y = [.8; 3.6; 2.1; -1.7]
x' *  y
x =
    1.2000
    3.5000
   -0.8000
    0.3000
y =
    0.8000
    3.6000
    2.1000
   -1.7000
ans =
   11.3700
```

Outer product

$$\mathbf{xy}^t = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \begin{bmatrix} y_1 & y_2 & \dots & y_d \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1d} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{nd} \end{bmatrix}$$

where $a_{ij} = x_i y_j$

Example

```
>> x = [2;1]                % Outer product
y = [4;2;3]'
x * y
x =
     2
     1
y =
     4     2     3
ans =
     8     4     6
     4     2     3
```

Euclidean norm

- Also known as L_2-norm obtained as $$\|\mathbf{x}\| = \sqrt{\mathbf{x}^t\ \mathbf{x}} = \left(\sum_{i=1}^{d} x_i^{\ 2}\right)^{\frac{1}{2}}$$
- A vector, **x**, is *normalized* if $$\|\mathbf{x}\| = 1$$
- The angle θ between **x** and **y** is given by:

$$\cos\theta = \frac{\mathbf{x}^t\mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|}$$

- Then, the inner product measures the co-linearity of **x** and **y**
- If **x** and **y** are *orthogonal*, we have $$\mathbf{x}^t\mathbf{y} = 0$$
- If **x** and **y** are *co-linear*, we have $$\left|\mathbf{x}^t\mathbf{y}\right| = \|\mathbf{x}\|\|\mathbf{y}\|$$

```
>> x = [2;1]                % Orthogonal
y = [-1;2]
x' * y
x =
     2
     1
y =
    -1
     2
ans =
     0
```

```
>>  x = [2;1]; y = [4;2]; % Co-linear
x =
     2
     1
y =
     4
     2
abs(x' * y)
ans = 10
norm(x)*norm(y)
ans = 10.0000
```

Basis

- Let the set of vectors $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_d\}$ be *linearly independent*
- Informally, a set of *d* linearly independent vectors "spans" $\Re^d$, and hence it is a *basis* of $\Re^d$.
- A basis $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_d\}$ is *orthogonal*,

  if for all *i* and *j*, $i \neq j$

  $\mathbf{x}_i$ and $\mathbf{x}_j$ are *orthogonal*, i.e. $\mathbf{x}_i^t \mathbf{x}_j = 0$

  if all $\mathbf{x}_i$ are *normalized*,

  the basis is *orthonormal*.

```
>> A = [2,-1;1,2]       % Orthogonal basis
B = orth(A)             % Orthonormal basis
B(1:2) * B(3:4)'
I = [1,0;0,1]
A =
     2    -1
     1     2
B =
   -0.8944   -0.4472
   -0.4472    0.8944
ans =
 -5.5511e-017 ≈ 0
I =
     1     0
     0     1
```

Matrices

- An $n \times d$ matrix, **A**, has the form:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1d} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{nd} \end{bmatrix}$$

- The transpose of **A** is denoted by $\mathbf{A}^t$
- Matrix multiplication: **A B = C**
- Multiplication of a matrix by a vector: **y** = **A x**

- **Rank:** The rank of A is the number of linearly independent rows (columns = "column rank").

- **Inner** (dot) product:

$$\mathbf{A} \cdot \mathbf{B} = \mathrm{tr}\{\mathbf{A}\mathbf{B}^t\}$$

- **Norm** of a matrix (Frobenius norm):

$$\left\|\mathbf{A}\right\|_F = \mathbf{A} \cdot \mathbf{A} = \mathrm{tr}\{\mathbf{A}\mathbf{A}^t\}$$

Square matrix:

- If $d = n$, **A** is called a *square* matrix.
- **A** is symmetric iff $a_{ij} = a_{ji}$

- Let **A** be a square matrix. Then,
- **Determinant:** Denoted by |**A**|, and defined as the product of the eigenvalues λ_i of **A**

$$|\mathbf{A}| = \prod_{i=1}^{d} \lambda_i$$

- **Trace:** Defined as the sum of diagonal elements

$$\mathrm{tr}\{\mathbf{A}\} = \sum_{i=1}^{d} a_{ii}$$

- If **A** is square, rank $\equiv$ # of nonzero eigenvalues of **A**

```
>> A = [2,1,0;1,5,3;0,3,4] % Determinant, trace
det(A)
trace(A)
A =
     2     1     0
     1     5     3
     0     3     4
ans =
    18
ans =
    11
```

Properties

- Let a **A** be square. **A** is not singular iff $|\mathbf{A}| \neq 0$.
- If **A** is not singular, then it has an inverse
- Inverse of **A**, denoted by $\mathbf{A}^{-1}$ satisfies:
  - It is unique
  - $\mathbf{A}\,\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$
- If **A** is not square (or if $\mathbf{A}^{-1}$ does not exist), then…
- **Pseudoinverse**: Denoted by $\mathbf{A}^{\dagger}$ is defined as:
  - $\mathbf{A}^{\dagger} = [\mathbf{A}^{t}\mathbf{A}]^{-1}\,\mathbf{A}^{t}$

  and satisfies:
  - $\mathbf{A}^{\dagger}\,\mathbf{A} = \mathbf{I}$

```
>> B = [1,2;4,5;7,8] % Pseudoinverse
Bpseudo = inv(B' * B) * B'
Bpseudo * B
B =
     1     2
     4     5
     7     8
Bpseudo =
   -1.1667   -0.3333    0.5000
    1.0000    0.3333   -0.3333
ans =
    1.0000   -0.0000
    0.0000    1.0000
```

- **Eigenvectors and eigenvalues:**
- Given a square matrix **A**, a special class of linear equations:

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$$

- for scalar λ, or:

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$$

- where **0** is the *zero* vector.
- The solution vector $\mathbf{x} = \mathbf{e}_i$ and the corresponding scalar λ_i
- are the i^{th} *eigenvector* and the associated *eigenvalue* of **A**

Example

Eigenvectors and Eigenvalues

```
> A := Matrix([[a,b],[c,d]]);
```

$$A := \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

```
> Lambda := DiagonalMatrix([lambda,lambda]);
```

$$\Lambda := \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}$$

```
> eqn  := Determinant(A-Lambda);;
```

$$eqn := a\,d - a\,\lambda - \lambda\,d + \lambda^2 - b\,c$$

```
> solve(eqn,lambda);
```

$$\frac{1}{2}a + \frac{1}{2}d + \frac{1}{2}\sqrt{a^2 - 2\,a\,d + d^2 + 4\,b\,c}, \frac{1}{2}a + \frac{1}{2}d - \frac{1}{2}\sqrt{a^2 - 2\,a\,d + d^2 + 4\,b\,c}$$

Positive Definite Matrices

- **A** is *positive definite* iff for all **x**, $\mathbf{x}^t \mathbf{A} \mathbf{x} > 0$
- If **A** is positive definite, then all $\lambda_i > 0$, and real.
- If **A** is also *symmetric*, all $\mathbf{e}_i$ are orthogonal (orthonormal)
- How to find the eigenvalues and eigenvectors?
- One method: Solve the *characteristic equation*:

$$|\mathbf{A} - \lambda\mathbf{I}| = \lambda^d + a_{d-1}\lambda^{d-1} + \cdots + a_1\lambda^1 + a_0 = 0$$

- If **A** is diagonal, the eigenvectors compose the *canonical* basis in $\Re^d$, i.e. the *identity matrix*
- Other much more efficient methods exist!

Functions of Matrices

Function of a matrix:

$f(\mathbf{A}) = \mathbf{\Phi} f(\mathbf{\Lambda}) \mathbf{\Phi}^{-1}$, where

$\mathbf{\Phi}$: eigenvectors of **A**

$\mathbf{\Lambda}$: eigenvalues of **A**

- $f(\mathbf{\Lambda}) = \mathrm{diag}(f(\lambda_{11}), f(\lambda_{22}), \ldots, f(\lambda_{dd}))$

  f can be any function on $\Re$

- Example: logarithm of **A**, or log(**A**)

  $\log(\mathbf{A}) = \mathbf{\Phi} \log(\mathbf{\Lambda}) \mathbf{\Phi}^{-1}$, where

  $\log(\mathbf{\Lambda}) = \mathrm{diag}(\log(\lambda_{11}), \log(\lambda_{22}), \ldots, \log(\lambda_{dd}))$

  Identity function:

  $\mathbf{A} = \mathbf{\Phi} \mathbf{\Lambda} \mathbf{\Phi}^{-1}$

Derivatives of Matrices

- **Derivatives of matrices** (special cases):
- Let **A** be a matrix, and **x**, **y** be vectors, then:

$$\frac{\partial}{\partial \mathbf{x}} \mathbf{A}\mathbf{x} = \mathbf{A} \qquad \frac{\partial}{\partial \mathbf{x}} \mathbf{y}^t \mathbf{x} = \mathbf{y} \qquad \frac{\partial}{\partial \mathbf{x}} \mathbf{x}^t \mathbf{A}\mathbf{x} = (\mathbf{A} + \mathbf{A}^t)\mathbf{x}$$

- if **A** is symmetric, then

$$\frac{\partial}{\partial \mathbf{x}} \mathbf{x}^t \mathbf{A}\mathbf{x} = 2\mathbf{A}\mathbf{x}$$

Example

Derivatives of Matrices

```
> x := Vector([x1,x2]);
```

$$x := \begin{bmatrix} x1 \\ x2 \end{bmatrix}$$

```
> A . x;
```

$$\begin{bmatrix} a\, x1 + b\, x2 \\ c\, x1 + d\, x2 \end{bmatrix}$$

```
> y := Vector([y1,y2]);
```

$$y := \begin{bmatrix} y1 \\ y2 \end{bmatrix}$$

```
> Transpose(y) . x;
```

$$x1\, y1 + x2\, y2$$

A quadratic equation

```
> eqn2 := expand(Transpose(x) . A . x);
```

$$eqn2 := a\, x1^2 + x1\, x2\, c + x2\, x1\, b + d\, x2^2$$

```
> diff(eqn2,x1); diff(eqn2,x2);
```

$$2\, a\, x1 + x2\, c + b\, x2$$

$$c\, x1 + x1\, b + 2\, d\, x2$$

```
> (A + Transpose(A)) . x;
```

$$\begin{bmatrix} 2\,a\,x1 + (b+c)\,x2 \\ (b+c)\,x1 + 2\,d\,x2 \end{bmatrix}$$

If A is symmetric, then the derivative is 2 A x

```
> A := Matrix([[a,b],[b,c]]);
```

$$A := \begin{bmatrix} a & b \\ b & c \end{bmatrix}$$

```
> eqn2 := expand(Transpose(x) . A . x);
```

$$eqn2 := a\,x1^2 + 2\,x2\,x1\,b + x2^2\,c$$

```
> diff(eqn2,x1); diff(eqn2,x2);
```

$$2\,a\,x1 + 2\,b\,x2$$

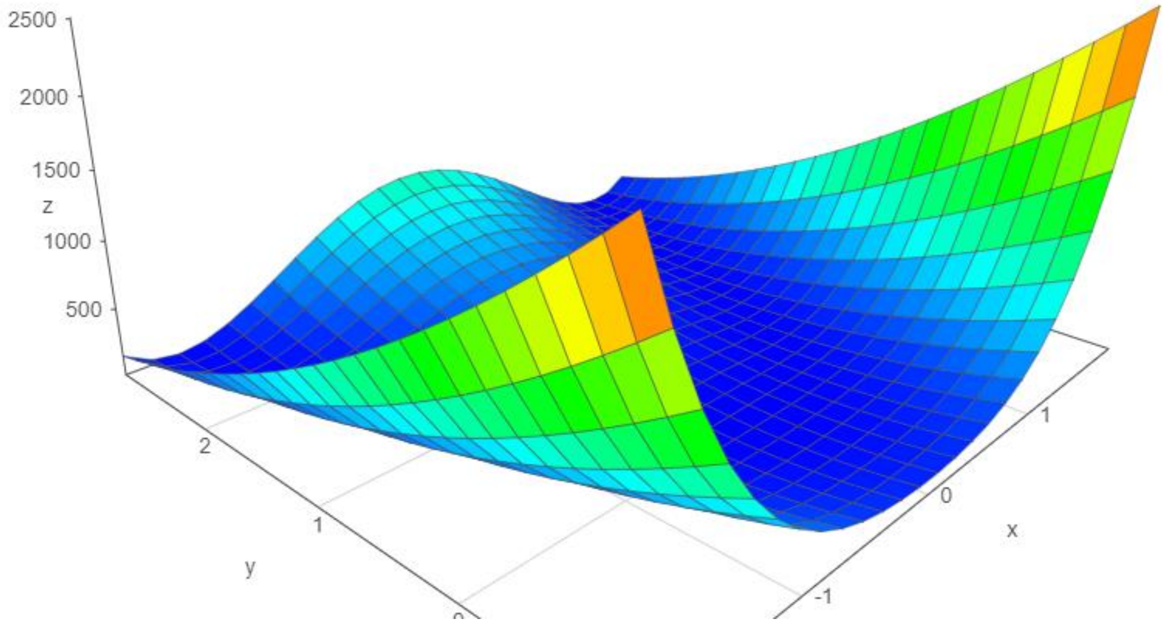
$$2\,x1\,b + 2\,x2\,c$$

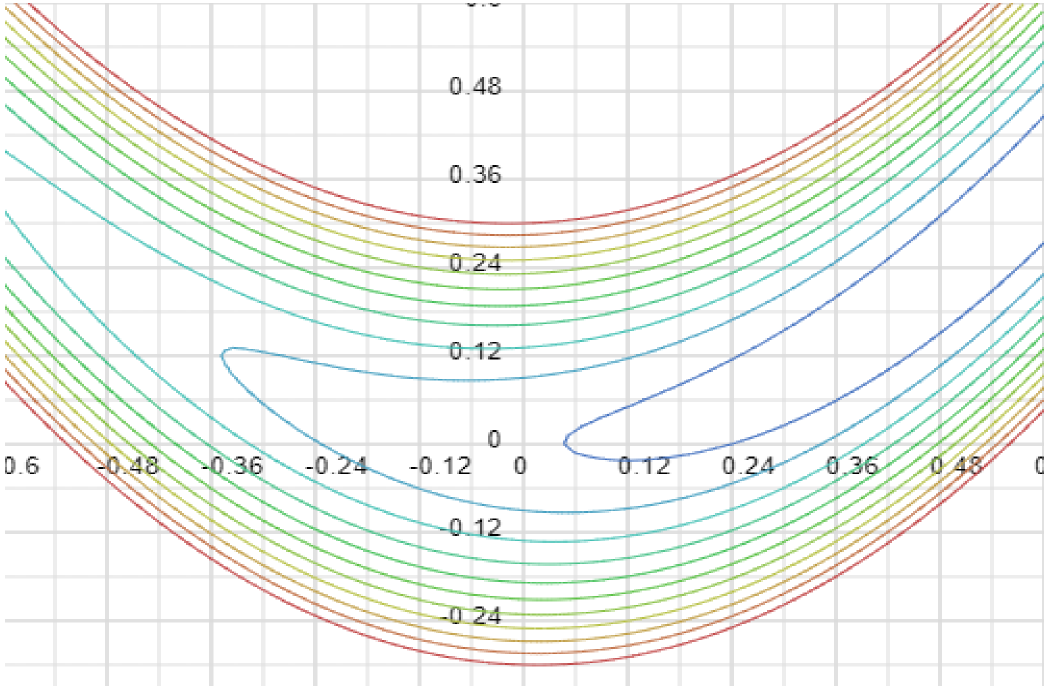
```
> 2 . A . x;
```

$$\begin{bmatrix} 2\,a\,x1 + 2\,b\,x2 \\ 2\,x1\,b + 2\,x2\,c \end{bmatrix}$$

Level sets

- The level set of a function $f: \Re^d \to \Re$ at level c is the set of points
$$S = \{\mathbf{x}: f(\mathbf{x}) = c\}$$
where c is a constant
- If $d = 2$, the level set is a level curve
- Example: Rosenbrock's function (aka "banana" function)
$$f(x) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2$$





Gradient

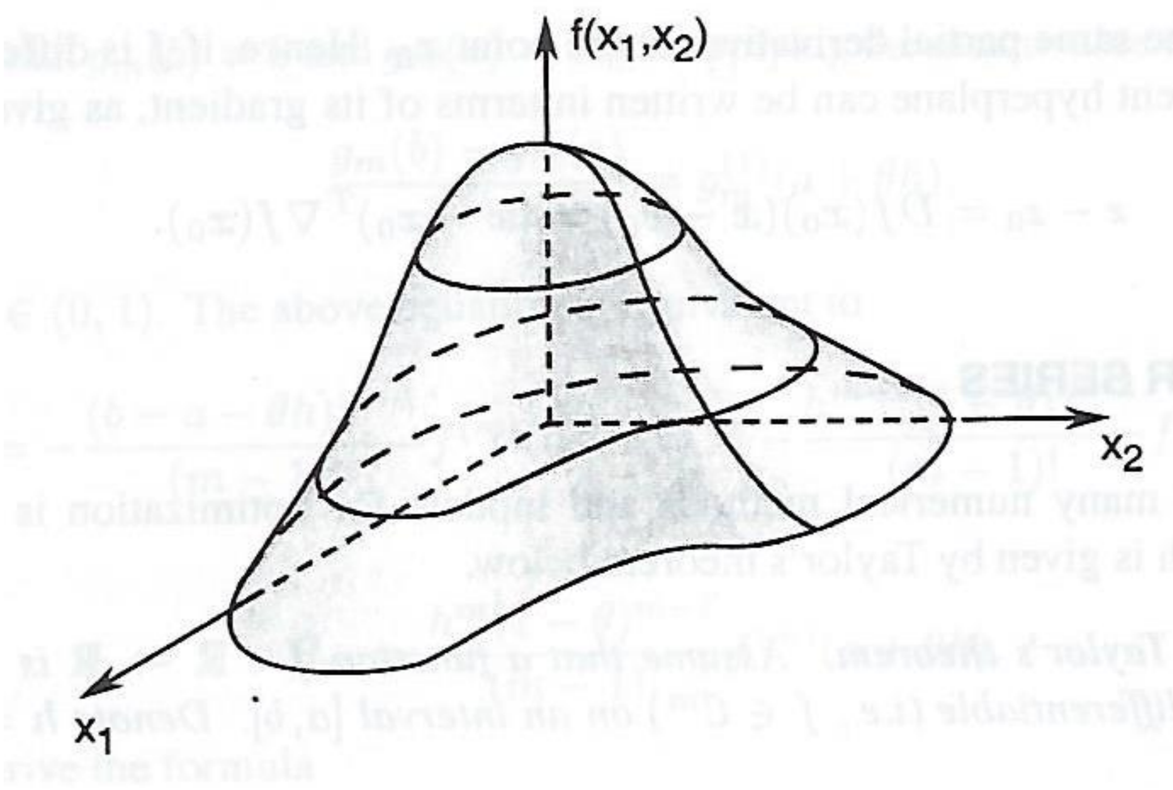
- Given a continuously differentiable function $f\colon \Re^d \rightarrow \Re$
- the gradient ∇ of f is defined as:

$$\nabla f(\mathbf{x}) = \frac{\partial}{\partial \mathbf{x}} f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_d} \end{bmatrix}$$

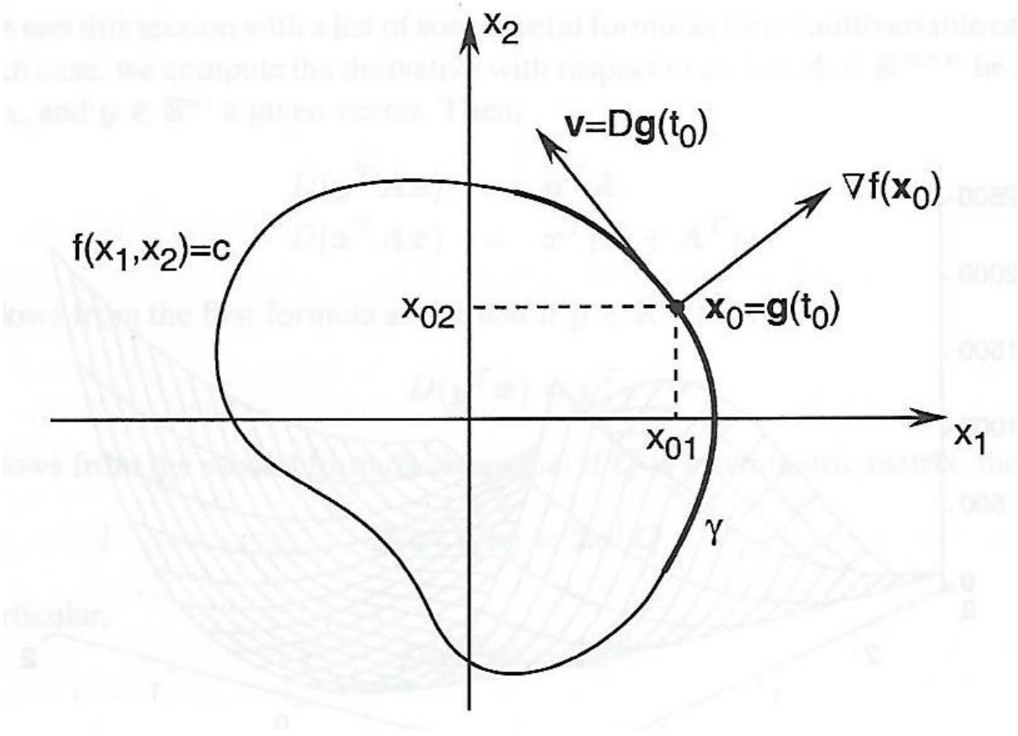
- Suppose a level curve g of level set S at a particular point t_0, and $g(t_0) = \mathbf{x}_0$, where $\mathbf{x}_0 \in S$,
- Then, there is a vector $\mathbf{v}$ that is tangent to g at $\mathbf{x}_0$
- That is, $\mathbf{v} = \partial g(t_0)$
- $\nabla f(\mathbf{x}_0)$ is orthogonal to $\mathbf{v}$

Example

Level sets



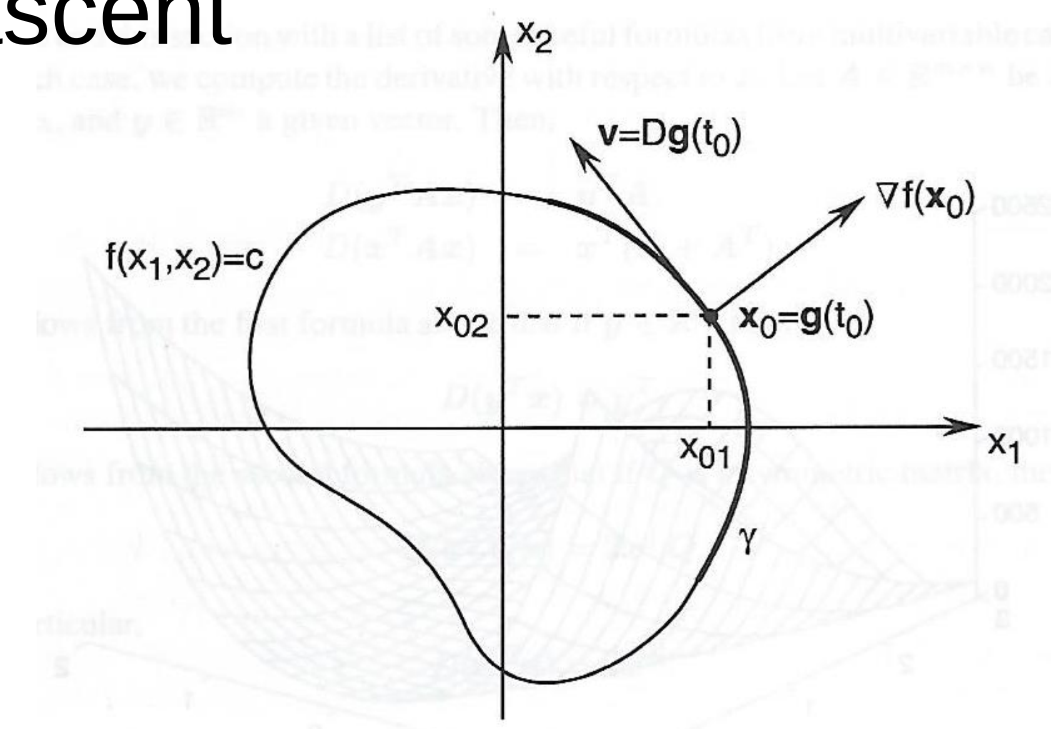
Gradient to a level set

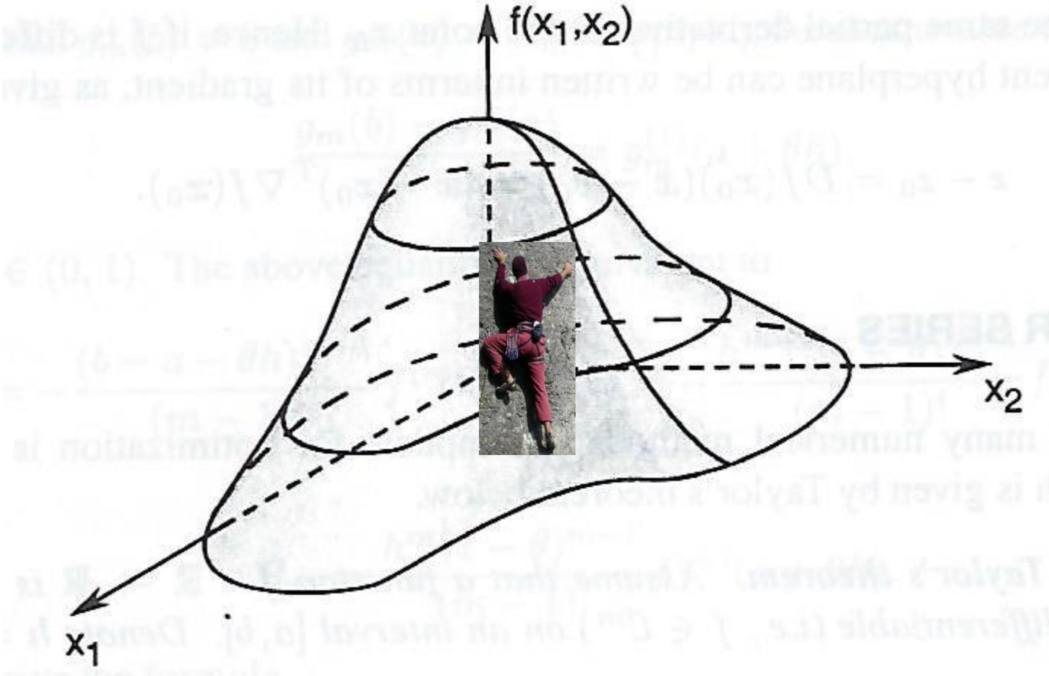


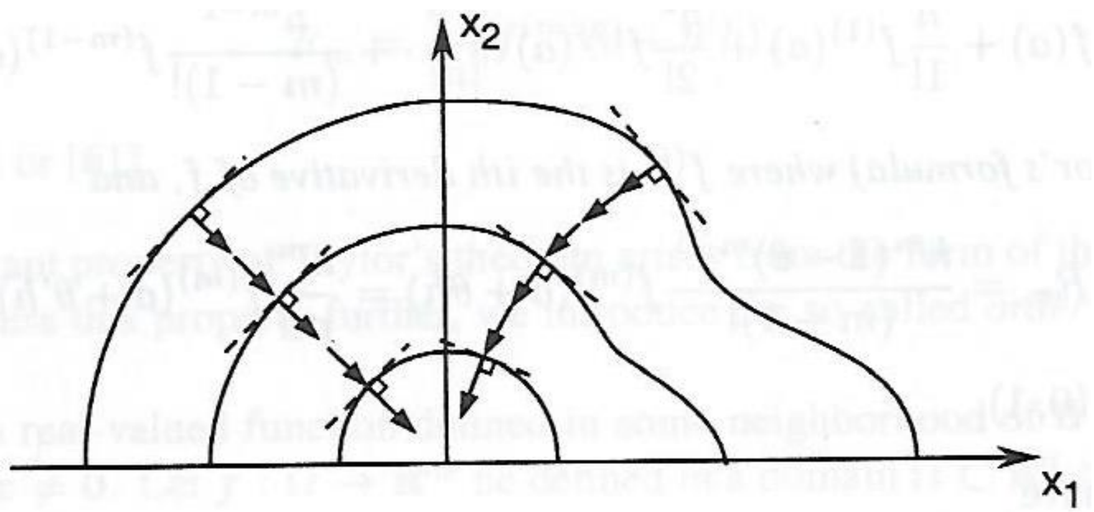
© Luis Rueda, 2019

Steepest ascent

- $\nabla f(\mathbf{x}_0)$ is the maximum rate of increase of f at $\mathbf{x}_0$
- Different level sets will give different directions:
  - of maximum rate of increase
  - called *path of steepest ascent*

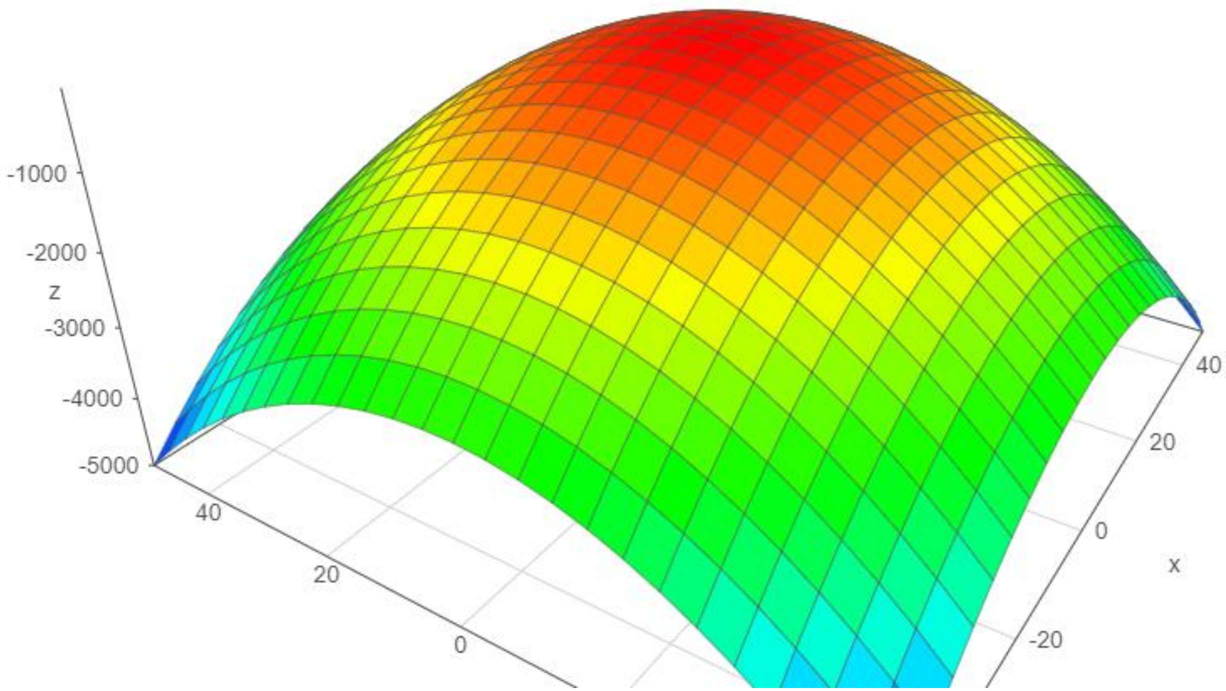


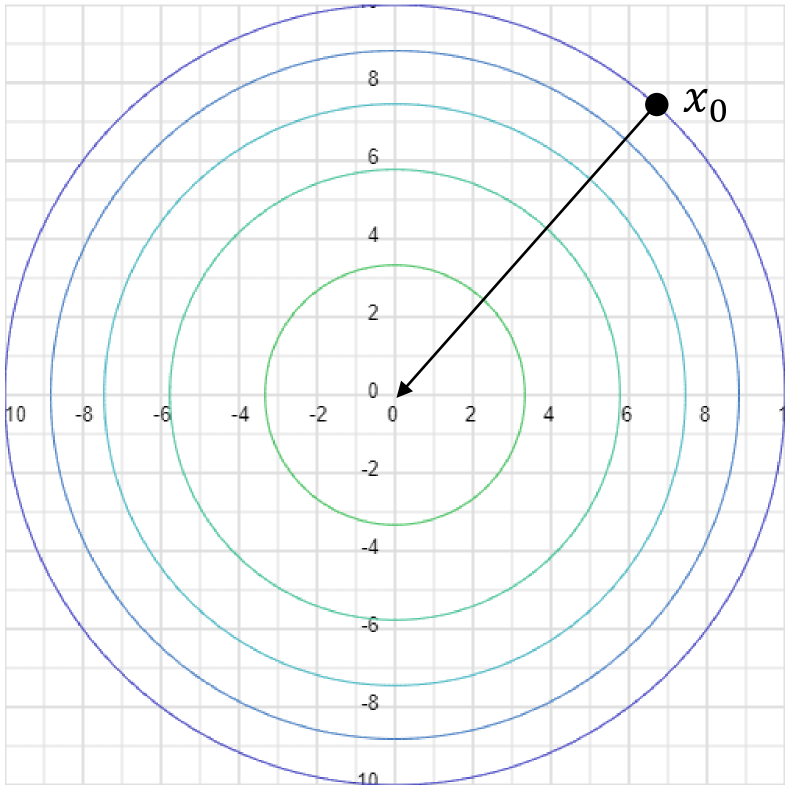




Example 2

- $f(\mathbf{x}) = -x_1^2 - x_2^2$
- In this example, path of steepest ascent always leads to maximum of f in *one* step





© Luis Rueda, 2019

Hessian matrix

- If $\nabla f(\mathbf{x})$ is continuously differentiable, then

$$\mathbf{H} = \frac{\partial}{\partial \mathbf{x}} \nabla f(\mathbf{x}) = \frac{\partial^2}{\partial \mathbf{x}} f(\mathbf{x}) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_2 \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_d \partial x_1} \\ \frac{\partial^2 f}{\partial x_1 \partial x_2} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_d \partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_d} & \frac{\partial^2 f}{\partial x_2 \partial x_d} & \cdots & \frac{\partial^2 f}{\partial x_d^2} \end{bmatrix},$$

where $\mathbf{H}$ is called the *Hessian matrix*

Probability

- If x is a *discrete* random variable, then
- it assumes values from a discrete set $\Omega = \{v_1, v_2, \ldots, v_m\}$, and
- for all i, p_i, which is $\Pr[x = v_i]$, satisfies:

$$p_i \geq 0 \quad \text{and} \quad \sum_{i=1}^{m} p_i = 1$$

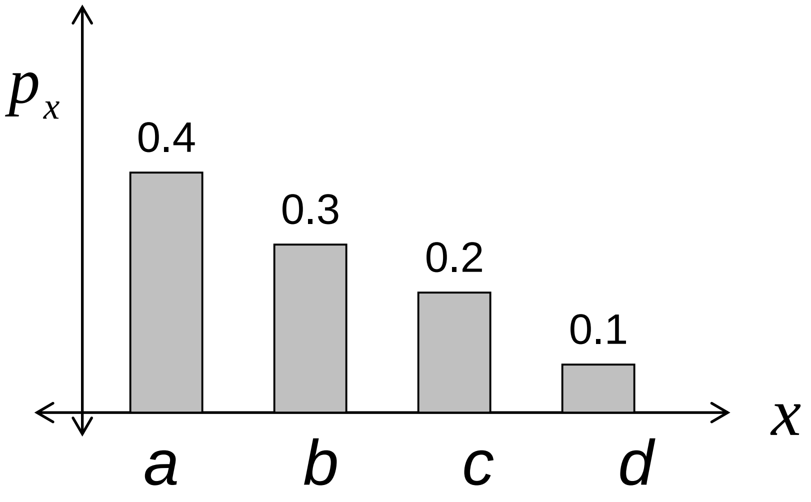
- The set of probabilities, $\{p_1, p_2, \ldots, p_m\}$, can be expressed as a *probability mass function*, $P(x)$, that satisfies:

$$P(x) \geq 0 \quad \text{and} \quad \sum_{i=1}^{m} P(x) = 1$$

Example (discrete r.v.)

$$\Omega = \{a, b, c, d\}$$

$$\Pr[x = a] = p_1 = 0.4$$



$$p_1 = 0.4$$
$$p_2 = 0.3$$
$$p_3 = 0.2$$
$$p_4 = 0.1$$

$$\sum_{i=1}^{4} p_i = 1.0$$

Expected value

- Also, *mean* or *average* of a r.v. *x* :

$$\mathrm{E}[x] = \mu = \sum_{i=1}^{m} xP(x) = \sum_{i=1}^{m} v_i p_i$$

- **Example:**
- Use integers or numeric values:

$$\Omega = \{\mathrm{a}, \mathrm{b}, \mathrm{c}, \mathrm{d}\} \rightarrow \{1,2,3,4\}$$

$$\begin{aligned}\mathrm{E}[x] &= 1(0.4) + 2(0.3) + 3(0.2) + 4(0.1) \\ &= 0.4 + 0.6 + 0.6 + 0.4 = 2\end{aligned}$$

Not necessarily integer, for example if

$$\mathrm{p}_1 = 0.5 \quad \mathrm{p}_2 = 0.25 \quad \mathrm{p}_3 = 0.2 \quad \mathrm{p}_4 = 0.05$$

$$\begin{aligned}\mathrm{E}[x] &= 1(0.5) + 2(0.25) + 3(0.2) + 4(0.05) = \\ &= 0.5 + 0.5 + 0.6 + 0.2 = 1.8\end{aligned}$$

- Ideally, *x* should be on a *measurable* field (e.g., Borel)

Variance

- Also, *second moment* around the mean of *x*

$$\mathrm{Var}[x] = \sigma^2 = \mathrm{E}[(x-\mu)^2] = \sum_{i=1}^{m} (x_i - \mu)^2 P(x_i)$$

- Expanding the quadratic term:

$$\mathrm{Var}[x] = \mathrm{E}[x^2] - \left(\mathrm{E}[x]\right)^2$$

- Example:

$$\mathrm{Var}[x] = \sigma^2 =$$
$$= (1-1.8)^2(0.5) + (2-1.8)^2(0.25) + (3-1.8)^2(0.2) + (4-1.8)^2(0.05)$$
$$= (0.64)^2(0.5) + (0.04)^2(0.25) + (1.44)^2(0.2) + (4.84)^2(0.05)$$
$$= 0.5720$$

Joint Probabilities

- Let x and y be two r.v. taking values from $\{v_1, v_2, \ldots, v_n\}$ and $\{w_1, w_2, \ldots, w_m\}$,
- defined as $p_{ij} = \Pr[x = v_i, y = w_j]$
- The *prob. mass function*, $P(x, y)$, also satisfies:

$$P(x, y) \geq 0 \quad \text{and} \quad \sum_x \sum_y P(x, y) = 1$$

Marginal distribution:

- Like a "separate" distribution for each variable:

$$P_x(x) = \sum_y P(x, y) \quad \text{and} \quad P_y(y) = \sum_x P(x, y)$$

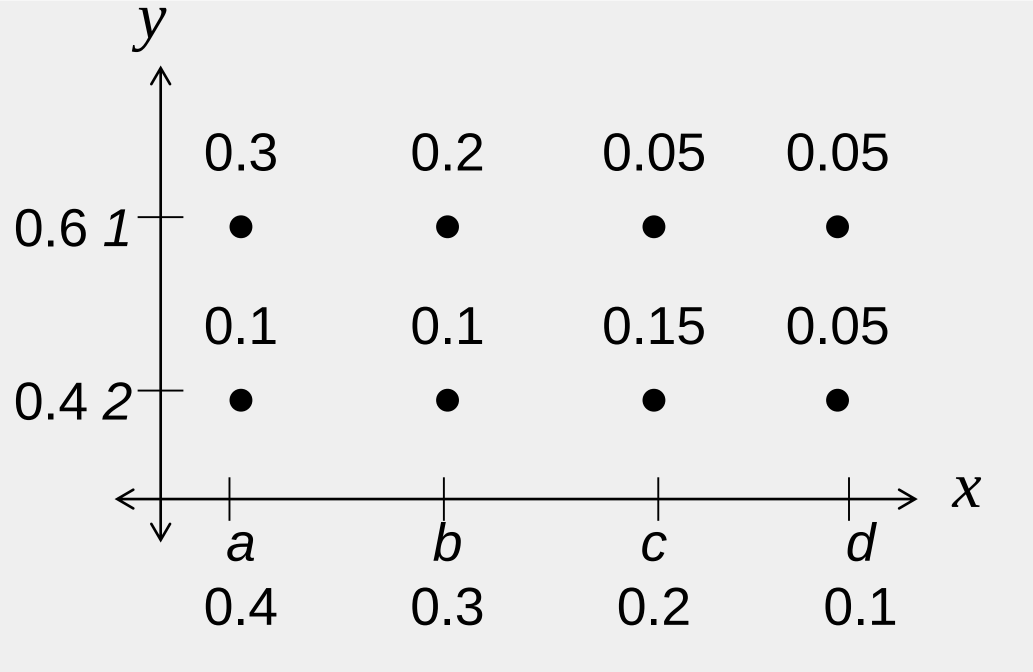
- Notation: P_x is used to denote a function (different from P_y)
- When using only one r.v., we will use P instead

Example

Let x be a r.v. in $\{a,b,c,d\}$

y be a r.v. in $\{1,2\}$

$$\left.\begin{array}{llll} p_{a1} = 0.3 & p_{b1} = 0.2 & p_{c1} = 0.05 & p_{d1} = 0.05 \\ p_{a2} = 0.1 & p_{b2} = 0.1 & p_{c2} = 0.15 & p_{d2} = 0.05 \end{array}\right\} \quad \sum_x \sum_y p_{xy} = 1$$



© Luis Rueda, 2019

$$\mathrm{P_r}[x=a]=\sum_y p(x,y)=0.4$$

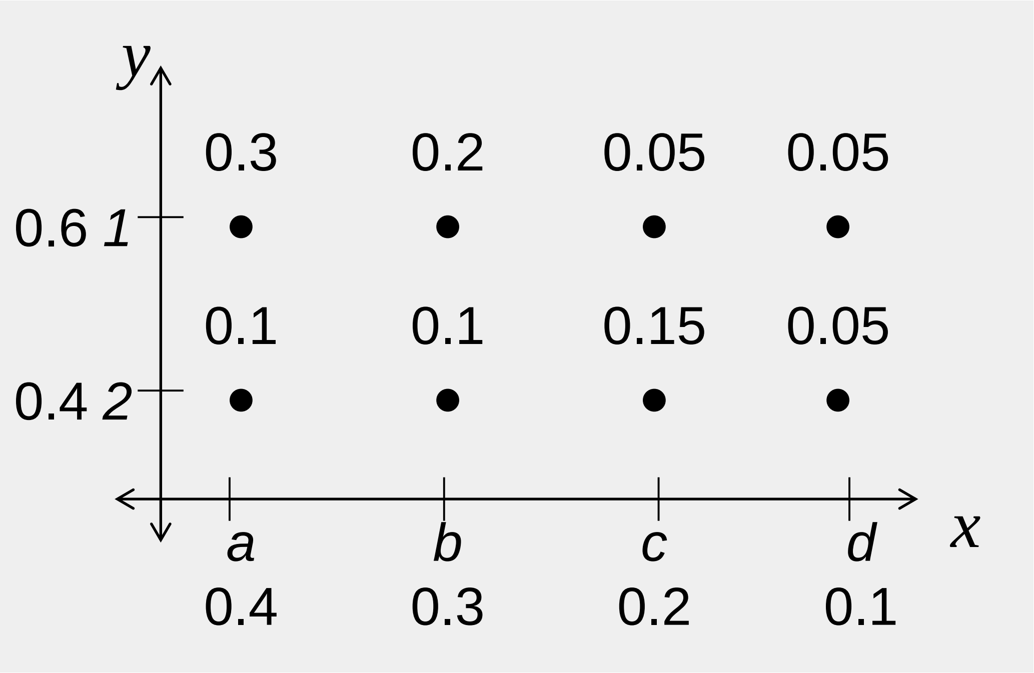
$$\mathrm{P_r}[x=b]=\sum_y p(x,y)=0.3$$

$$\mathrm{P_r}[x=c]=\sum_y p(x,y)=0.2$$

$$\mathrm{P_r}[x=d]=\sum_y p(x,y)=0.1$$

$$\mathrm{P_r}[y=1]=\sum_x p(x,y)=0.6$$

$$\mathrm{P_r}[y=2]=\sum_x p(x,y)=0.4$$



- **Independence:** *x* and *y* are statistically independent iff:

$$P(x, y) = P_x(x) P_y(y)$$

- But, for the example:

$$\mathrm{P_r}[x = a, y = 1] = 0.3$$

$$\neq$$

$$\mathrm{P_r}[x = a]\,\mathrm{P_r}[y = 1] = (0.4)(0.6) = 0.24$$

$$x \text{ and } y \text{ are not independent}$$

- **Mutually exclusive events:**
  - We now talk about "events". E_i and E_j are *mutually exclusive* if

$$\Pr[x = E_i \cap E_j] = 0$$

Conditional Probabilities

Denoted by *P*(*x*|*y*), and defined as:

$$P(x \mid y) = \frac{P(x,y)}{P(y)}$$

- Knowing *y* gives us “some information” about *x*.
- Then, if *x* and *y* are independent, *P*(*x*|*y*) = *P*(*x*).
- **Example:**

$$\mathrm{P_r}[x = a \mid y = 1] = \frac{\mathrm{P_r}[x = a, y = 1]}{\mathrm{P_r}[y = 1]} = \frac{0.3}{0.6} = 0.5$$

$$\mathrm{P_r}[x = a \mid y = 2] = \frac{\mathrm{P_r}[x = a, y = 2]}{\mathrm{P_r}[y = 2]} = \frac{0.1}{0.4} = 0.25$$

- We can also write (Bayes Theorem):

$$P(x,y) = P(x \mid y)P(y) = P(y,x) = P(y \mid x)P(x)$$

- or

$$P(x \mid y) = \frac{P(y \mid x)P(x)}{P(y)}$$

- by the *law of total probabilities*

$$P(y) = \sum_x P(x,y) = \sum_x P(y \mid x)P(x)$$

- then, we write:

$$P(x \mid y) = \frac{P(y \mid x)P(x)}{\sum_x P(y \mid x)P(x)}$$

Example (*law of total probabilities*):

$$P(y) = \sum_x P(x, y) = \sum_x P(y \mid x) P(x)$$

$$\mathrm{P_r}[y = 1 \mid x = a]\,\mathrm{P_r}[x = a] = 0.75 \cdot 0.4 = 0.3$$

$$\mathrm{P_r}[y = 1 \mid x = b]\,\mathrm{P_r}[x = b] = 0.67 \cdot 0.3 = 0.2$$

$$\mathrm{P_r}[y = 1 \mid x = c]\,\mathrm{P_r}[x = c] = 0.25 \cdot 0.2 = 0.05$$

$$\mathrm{P_r}[y = 1 \mid x = d]\,\mathrm{P_r}[x = d] = 0.5 \cdot 0.1 = 0.05$$

$$\mathrm{P_r}[y = 1] = \sum_x P(y \mid x) P(x) = 0.6$$

Example *(Bayes Theorem):*

$$\mathrm{P_r}[x = a \mid y = 1] = \frac{\mathrm{P_r}[y = 1 \mid x = a]\mathrm{P_r}[x = a]}{\mathrm{P_r}[y]} = \frac{0.75 \cdot 0.4}{0.6} = 0.5$$

Chain Rule of Conditional Probabilities

- Given a collection of d random variables:

$$x_1, x_2, \dots, x_d$$

- The joint probability distribution can be computed as:

$$P(x_1, x_2, \dots, x_d) = P(x_d) \prod_{i=1}^{d-1} P(x_i | x_1, x_2, \dots, x_{i-1})$$

- known as the ***chain rule***
- Example:

$$\begin{aligned} P(x, y, z) &= P(x|y, z)P(y, z) \\ &= P(x|y, z)P(y|z)P(z) \end{aligned}$$

Random vectors

A pair of values of two r.v., v_x and v_y, can be considered as a vector **x** in the 2D space, whose prob. is $P(v_x, v_y)$

- Similarly, $\mathbf{x} = [x_1\ x_2\ \ldots\ x_d]^t$ is a *random vector*
- The *joint probability mass* function is now $P(\mathbf{x}) \geq 0$, and
- if $x_1\ x_2\ \ldots\ x_d$ are independent, we have

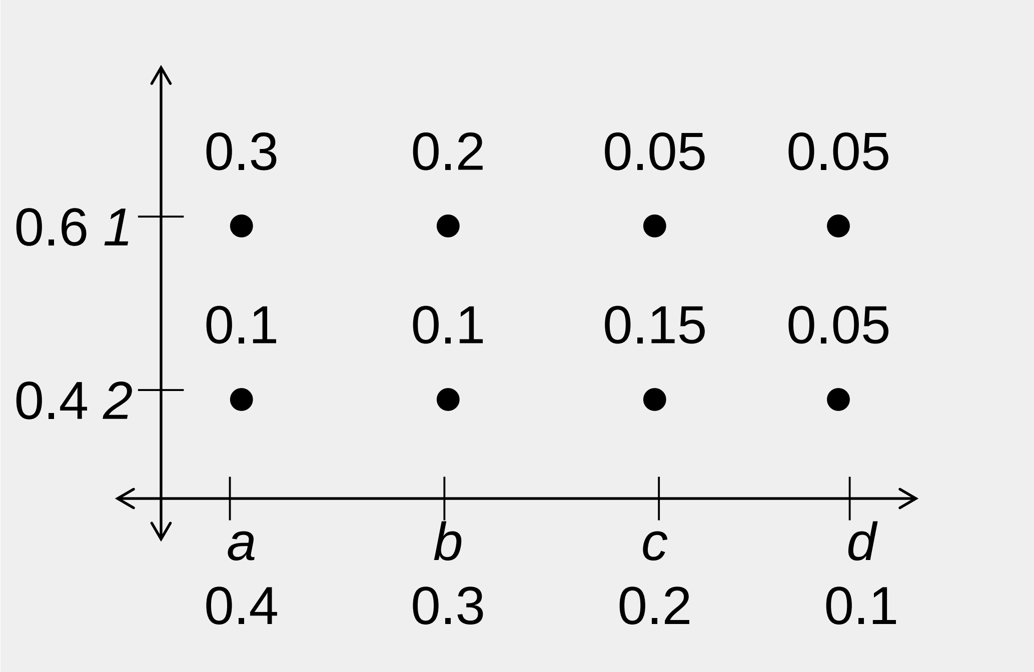
$$P(\mathbf{x}) = \prod_{i=1}^{d} P_{x_i}(x_i)$$

Example

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

x_1 is in $\{1,2,3,4\}$

x_2 is in $\{1,2\}$



$$\mathrm{P}[\mathbf{x}] \text{ for } \mathbf{x} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \text{ its } P_r[x_1 = 1, x_2 = 1] = 0.3$$

$$\mathrm{P}[\mathbf{x}] \text{ for } \mathbf{x} = \begin{bmatrix} 3 \\ 2 \end{bmatrix} \text{ its } P_r[x_1 = 3, x_2 = 2] = 0.15$$

Mean vector: A *d*-dimensional vector, $\boldsymbol{\mu}$, given by

$$\boldsymbol{\mu} = \mathrm{E}[\mathbf{x}] = \sum_{\mathbf{x}} \mathbf{x} P(\mathbf{x})$$

- also known as the *center of the prob. mass*

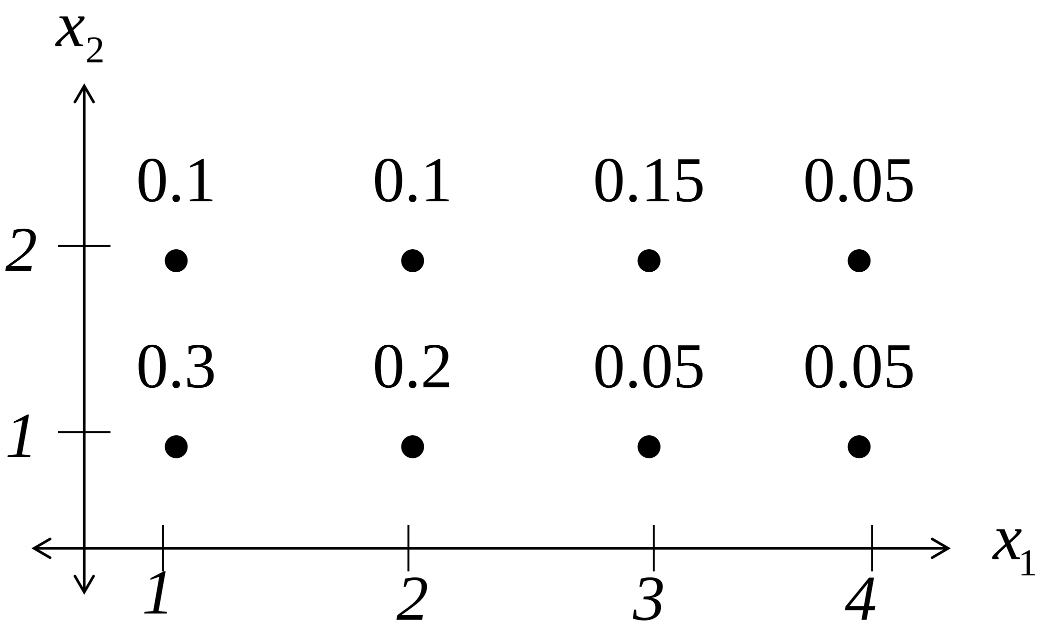
Covariance matrix:

- A $d \times d$ matrix, $\boldsymbol{\Sigma}$, defined using the *outer* product:

$$\boldsymbol{\Sigma} = \sum_{\mathbf{x}} (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^t P(\mathbf{x}) = \mathrm{E}[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^t]$$

- $\boldsymbol{\Sigma}$ is *positive semidefinite* and *symmetric*

Example



$$\boldsymbol{\mu} = \begin{bmatrix}1\\1\end{bmatrix}0.3+\begin{bmatrix}1\\2\end{bmatrix}0.1+\begin{bmatrix}2\\1\end{bmatrix}0.2+\begin{bmatrix}2\\2\end{bmatrix}0.1+\begin{bmatrix}3\\1\end{bmatrix}0.05+\begin{bmatrix}3\\2\end{bmatrix}0.15+\begin{bmatrix}4\\1\end{bmatrix}0.05+\begin{bmatrix}4\\2\end{bmatrix}0.05$$

$$= \begin{bmatrix}2\\1.55\end{bmatrix}$$

$$\mathbf{\Sigma} = \left(\begin{bmatrix}1\\1\end{bmatrix} - \begin{bmatrix}2\\1.55\end{bmatrix}\right)\left(\begin{bmatrix}1\\1\end{bmatrix} - \begin{bmatrix}2\\1.55\end{bmatrix}\right)^t (0.3) + \left(\begin{bmatrix}1\\2\end{bmatrix} - \begin{bmatrix}2\\1.55\end{bmatrix}\right)\left(\begin{bmatrix}1\\2\end{bmatrix} - \begin{bmatrix}2\\1.55\end{bmatrix}\right)^t (0.1) +$$

$$= \left(\begin{bmatrix}2\\1\end{bmatrix} - \begin{bmatrix}2\\1.55\end{bmatrix}\right)\left(\begin{bmatrix}2\\1\end{bmatrix} - \begin{bmatrix}2\\1.55\end{bmatrix}\right)^t (0.2) + \left(\begin{bmatrix}2\\2\end{bmatrix} - \begin{bmatrix}2\\1.55\end{bmatrix}\right)\left(\begin{bmatrix}2\\2\end{bmatrix} - \begin{bmatrix}2\\1.55\end{bmatrix}\right)^t (0.1) +$$

$$= \left(\begin{bmatrix}3\\1\end{bmatrix} - \begin{bmatrix}2\\1.55\end{bmatrix}\right)\left(\begin{bmatrix}3\\1\end{bmatrix} - \begin{bmatrix}2\\1.55\end{bmatrix}\right)^t (0.05) + \left(\begin{bmatrix}3\\2\end{bmatrix} - \begin{bmatrix}2\\1.55\end{bmatrix}\right)\left(\begin{bmatrix}3\\2\end{bmatrix} - \begin{bmatrix}2\\1.55\end{bmatrix}\right)^t (0.15) +$$

$$= \left(\begin{bmatrix}4\\1\end{bmatrix} - \begin{bmatrix}2\\1.55\end{bmatrix}\right)\left(\begin{bmatrix}4\\1\end{bmatrix} - \begin{bmatrix}2\\1.55\end{bmatrix}\right)^t (0.05) + \left(\begin{bmatrix}4\\2\end{bmatrix} - \begin{bmatrix}2\\1.55\end{bmatrix}\right)\left(\begin{bmatrix}4\\2\end{bmatrix} - \begin{bmatrix}2\\1.55\end{bmatrix}\right)^t (0.05)$$

$$= \begin{bmatrix}1.0 & 0.15\\0.15 & 0.2625\end{bmatrix} \qquad \text{(covariance matrix)}$$

Continuous random variables:

- Represented by a *probability* ***density*** *function* (pdf): *p*(*x*)
- But *p*(*x*) is 0 for any value *x*
- Now, if *x* falls in an interval [*a*,*b*]
- *P*(*x*) is the *probability* ***mass*** *function*, and

$$\Pr[x \in [a,b]] = \int_a^b p(x)\,dx$$

- where $p(x) \geq 0$ and $\int_{-\infty}^{\infty} p(x)\,dx = 1$

- Expected values and variances are defined in terms of integrals

The normal distribution

- Fully defined by its two parameters, μ and σ^2.
- Probability density function:

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Notation: A normally distributed r.v. with mean μ and variance σ^2 is denoted by $x \sim N(\mu, \sigma^2)$

Normal distibution (univariate)

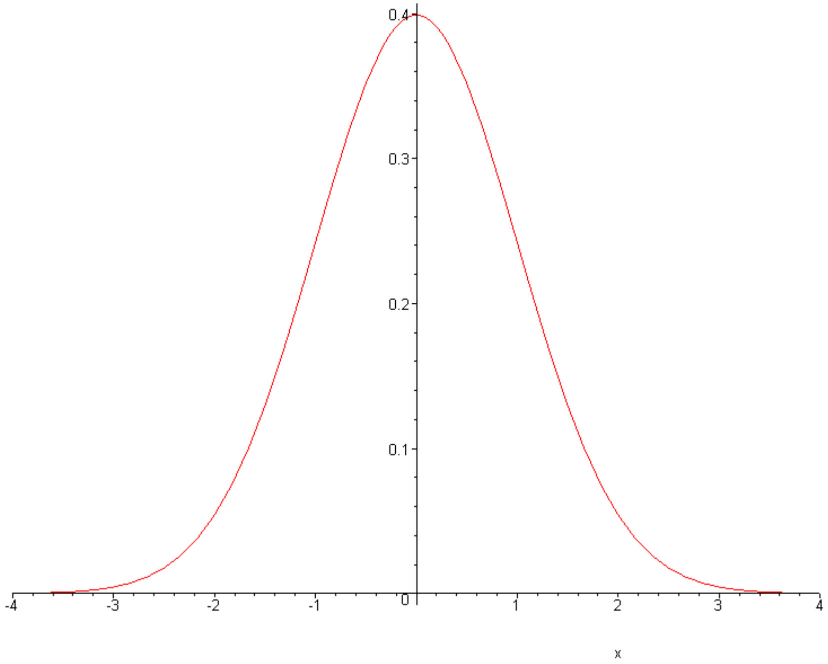
```
> mu := 0;
sigma := 1;
p2 := evalf(1/sqrt(2*Pi)) * 1/sqrt(sigma) * exp(-1/2 * (x-mu)^2/sigma);
```

$$\mu := 0$$

$$\sigma := 1$$

$$p2 := 0.3989422802\mathbf{e}^{(-1/2\, x^2)}$$

```
> plot(p2,x=-4..4);
```



Properties of normal distribution:

- Joint distribution of normal distns is normal
- Generalization: Multivariate of normal distns is normal, i.e., normal random vector
- Linear transformation of normal is normal
- Characterized by just two moments: mean and variance
- Central limit theorem

Central limit theorem:

- If x_1, x_2, …, x_n are n independent r.v. with a common pdf,
- and for all i, E[x_i] = 0 and Var[x_i] = 1, then, as $n \to \infty$

is a $N(0,1)$ r.v.

$$\bar{x} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_i$$

Multivariate normal density

- Similarly, a normally distributed random *vector*, **x**, of dimension *d* is fully defined by
  - a *mean* vector: $\boldsymbol{\mu}$
  - a covariance matrix: $\boldsymbol{\Sigma}$
- The probability density function is given by:

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^t \Sigma^{-1} (\mathbf{x}-\mu)}$$

- The central limit theorem is also valid in *d* > 1 dimensions

Recall, normal r.v. pdf: $p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Normal distribution (multivariate)

```
> X := Vector([x,y]);
Mu := Vector([0,0]);
Sigma := Matrix([[1,0.35],[0.35,0.2625]]);
p3 := evalf(1/(2*Pi) * 1/sqrt(Determinant(Sigma)) * exp(-1/2 * Transpose(X-Mu) . MatrixInverse(Sigma). (X-Mu)));
```

$$X := \begin{bmatrix} x \\ y \end{bmatrix}$$

$$\mathrm{M} := \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

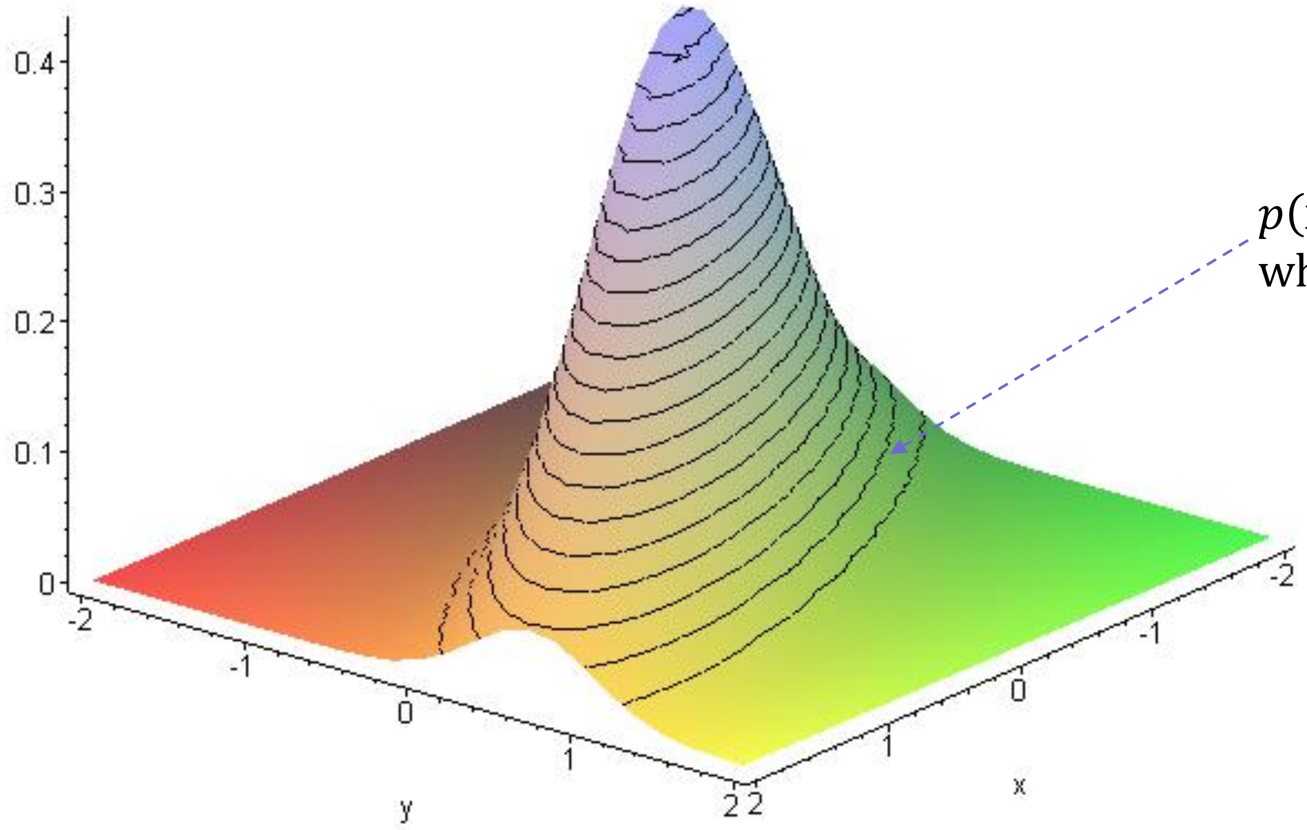
$$\Sigma := \begin{bmatrix} 1 & 0.35 \\ 0.35 & 0.2625 \end{bmatrix}$$

$$p3 := 0.4253594775 \mathbf{e}^{(-1.\, x\,(0.9375000000\ x - 1.250000000\ y) - 1.\, y\,(-1.250000000\ x + 3.571428572\ y))}$$

```
> with(plots):plot3d(p3,x=-2..2,y=-2..2);
```

Graphical Example

- Points with the same probability belong to the same level set



$p(\mathbf{x}) = c$
where c is a constant

Mahalanobis distance

- From a vector $\mathbf{x}$ to the mean $\boldsymbol{\mu}$
- Defined by a positive semidefinite matrix, like $\boldsymbol{\Sigma}$

$$r^2 = (\mathbf{x} - \mu)^t \Sigma^{-1} (\mathbf{x} - \mu)$$

- In the normal distribution, all points with the *same* Mahalanobis distance from μ have the *same* probability
- These points are in the same *ellipsoid*, whose
  - "radii" are the *eigenvalues* of $\boldsymbol{\Sigma}$: $\boldsymbol{\Lambda}$
  - "axes" are in the direction of the *eigenvectors* of $\boldsymbol{\Sigma}$: $\boldsymbol{\Phi}$
- Thus:

$$\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_d \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Phi} = \left[\varphi_1 \middle| \varphi_2 \middle| \cdots \middle| \varphi_d\right]$$

Pictorially

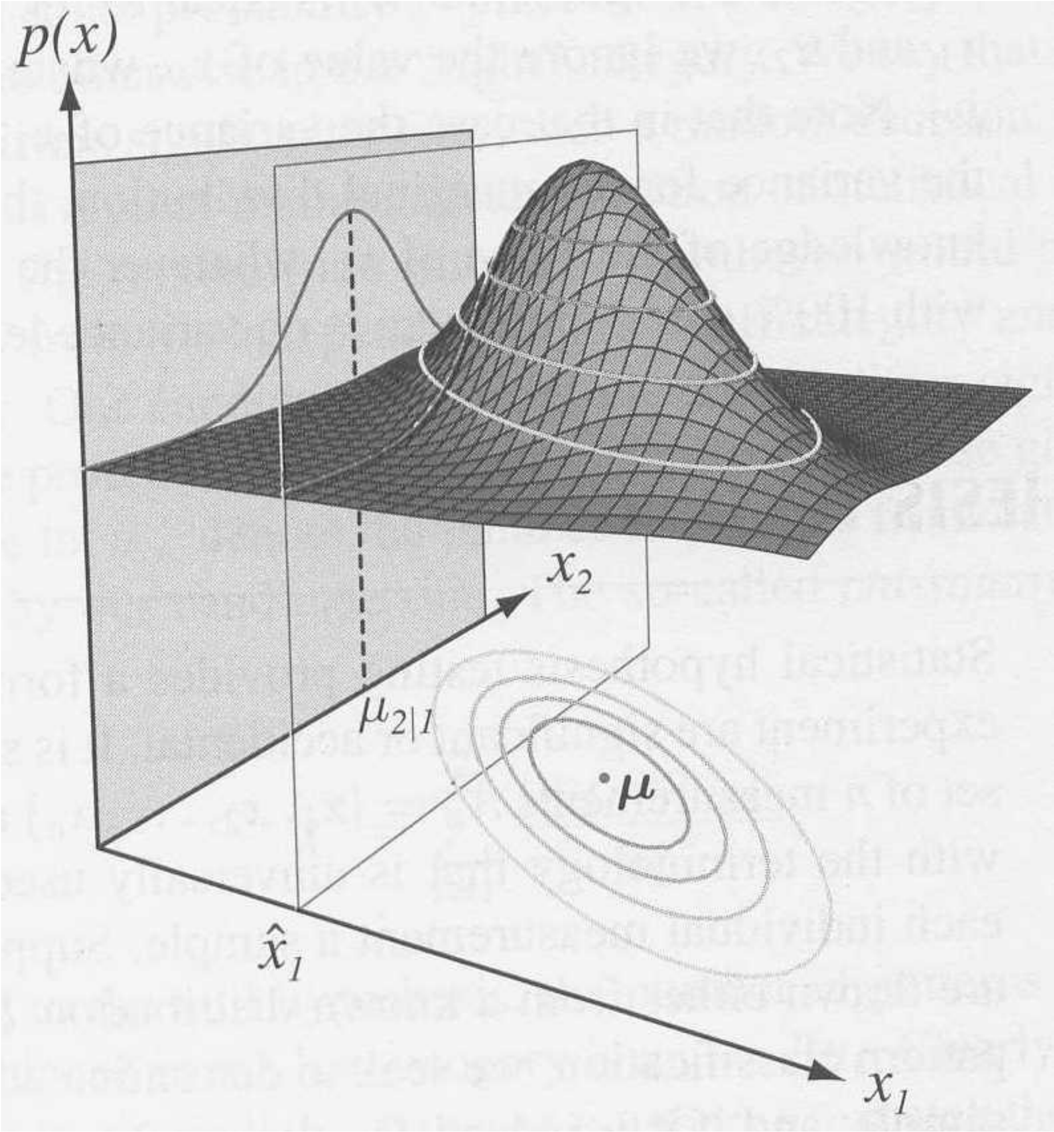


Figure from Duda et al.

Example

Mahalanobis Distance

```
> X := Vector([x,y]);
Mu := Vector([2,1]);
Sigma := Matrix([[1,0.35],[0.35,0.2625]]);
r := expand(Transpose(X-Mu) . MatrixInverse(Sigma) . (X-Mu));
implicitplot(r=0.3, x=0..3,
y=0..10,color=blue,numpoints=20000);
```

$$X := \begin{bmatrix} x \\ y \end{bmatrix}$$

$$\mathrm{M} := \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

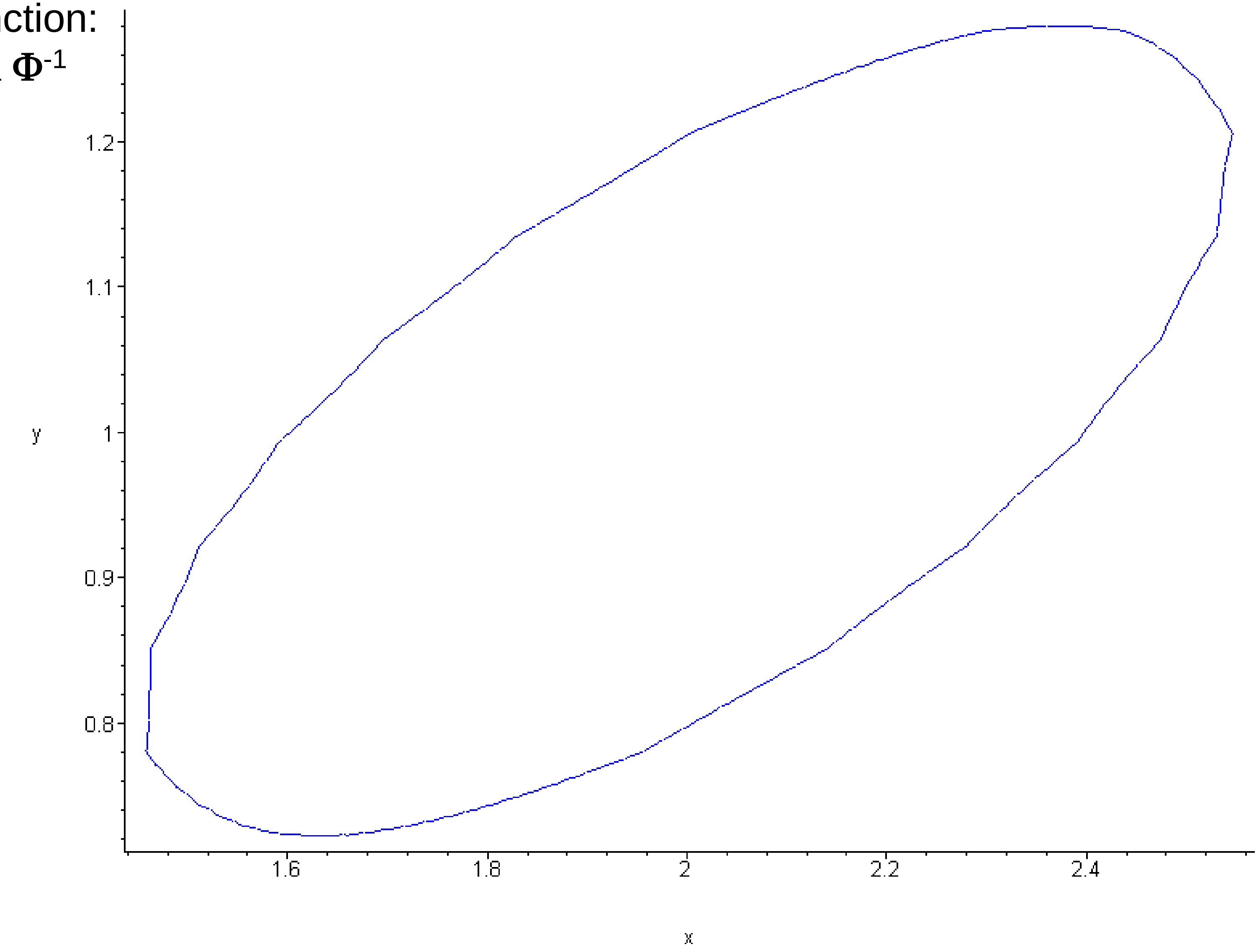
$$\Sigma := \begin{bmatrix} 1 & 0.35 \\ 0.35 & 0.2625 \end{bmatrix}$$

$$r := 1.874999999999978\,x^2 - 2.500000000\,x - 5.000000000\,x\,y + 4.6428571 - 4.285714286\,y + 7.142857142857142\,35\,y^2$$

$$r^2 = (\mathbf{x}-\mu)^t \Sigma^{-1} (\mathbf{x}-\mu)$$

Identity function:

$\Sigma = \Phi \, \Lambda \, \Phi^{-1}$



Correlation coefficient

- Defined as follows: $$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$
- where σ_{xy} is the *covariance* of *x* and *y*, and
- measures the *statistical dependence* of *x* and *y*
- If *x* and *y* are statistically independent, then they are *uncorrelated*, i.e., $\sigma_{xy} = 0$ and $\rho = 0$
- The converse is not always true, but…
- If *x* and *y* are normally dist. r.v.:
- *x* and *y* are independent **if and only if** they are uncorrelated.
- Note: $\sigma_{xy} = E[(x-\mu_x)(x-\mu_y)]$ -- or “cross-moment”

Then, if

- $\mathbf{x} = [x_1\ x_2\ \ldots\ x_d]^t$ is a normal r.v., and
- $x_1\ x_2\ \ldots\ x_d$ are independent,

we have:

- $\mathbf{\Sigma} = \mathbf{\Lambda}$ is a diagonal matrix with the *eigenvalues*.
- The eigenvectors compose the identity matrix, i.e. $\mathbf{\Phi} = \mathbf{I}$
- The ellipsoids containing points with the same Mahalanobis distance have their axes parallel to the system coordinates

```
> Eigenvectors(Sigma);
```

$$\begin{bmatrix} 1.139655903289881 42+0.\,I \\ 0.12284409671011886 6+0.\,I \end{bmatrix},$$

$$\begin{bmatrix} 0.92879122810000003 2+0.\,I & -0.37060336550000000 4+0.\,I \\ 0.37060336550000000 4+0.\,I & 0.92879122810000003 2+0.\,I \end{bmatrix}$$

```
> Mu := Vector([2,1]);
Sigma := Matrix([[1,0],[0,3]]);
r := Transpose(X-Mu) . MatrixInverse(Sigma) . (X-Mu);
implicitplot(r=0.2, x=0..3,
y=0..10,color=blue,numpoints=40000,scaling=constrained);
```

$$M := \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

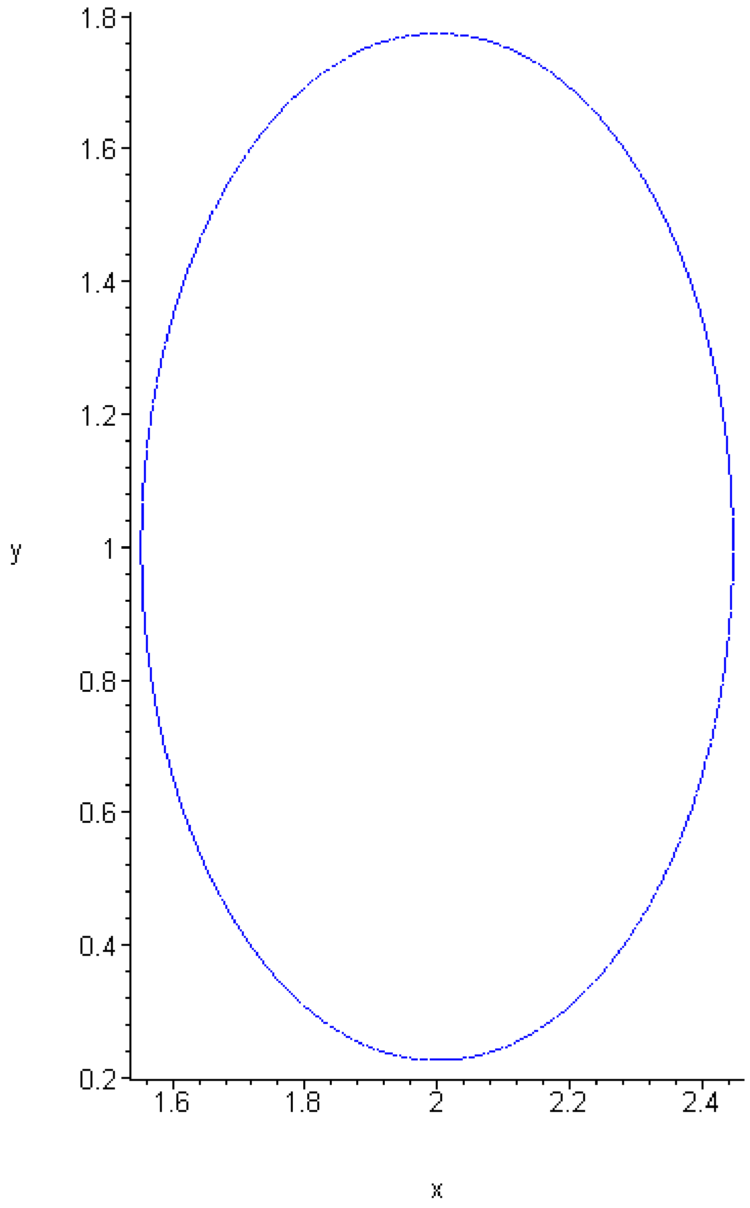
$$\Sigma := \begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix}$$

$$r := (x-2)^2 + (y-1)\left(\frac{1}{3}y - \frac{1}{3}\right)$$

$$r^2 = (\mathbf{x}-\mu)^t \Sigma^{-1} (\mathbf{x}-\mu)$$

Identity function:

$\Sigma = \Phi \Lambda \Phi^{-1}$



Distance Between Distributions

- **Chernoff distance**
  - Between two *normal* random vectors, $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$:

$$k_{ij}(\beta) = \frac{\beta(1-\beta)}{2}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_i)^t[\beta\boldsymbol{\Sigma}_i + (1-\beta)\boldsymbol{\Sigma}_j]^{-1}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_i)$$

$$+\frac{1}{2}\ln\frac{|\beta\boldsymbol{\Sigma}_i + (1-\beta)\boldsymbol{\Sigma}_j|}{|\boldsymbol{\Sigma}_i|^{\beta}|\boldsymbol{\Sigma}_j|^{1-\beta}}$$

  - Among **many** random vectors, $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_c$:
    - Many ways to compute it; one way is to weight pairwise distances:

$$k(\beta) = \sum_{i=1}^{c-1}\sum_{j=i+1}^{c} p_i p_j k_{ij}(\beta)$$

  - Setting $\beta = \frac{1}{2}$ leads to **Battacharyya distance**
  - Chernoff distance can be used even if dist are not normal

Distance Between Distributions

- **Kullback-Leibler** (two *normal* r. v., $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$):

$$d_{ij} = \frac{1}{2} tr\left\{\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\Sigma}_j + \boldsymbol{\Sigma}_j^{-1}\boldsymbol{\Sigma}_i - 2\mathbf{I}\right\} + \frac{1}{2}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^t\left(\boldsymbol{\Sigma}_i^{-1} + \boldsymbol{\Sigma}_j^{-1}\right)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$$

- Multivariate (*normal* r.v.):
  - Many ways… one way is to weight distances for pairs of classes:

$$d = \sum_{i=1}^{c-1} \sum_{j=i+1}^{c} p_i p_j d_{ij}$$

Information and Entropy

- **Information amount:**
  - Measures how surprised we are when we observe an event based on how likely is to occur

  Example:
  - We would not be surprised if we see an “e” in English text (low information amount)
  - But we would be very surprised if we see a “q” (high info amount)

  Given a discrete r.v. x that takes values $\Omega = \{v_1, v_2, \ldots, v_m\}$ whose probs are $\{p_1, p_2, \ldots, p_m\}$

  Definition (information amount):

$$I_i = -p_i \log p_i$$

  log base 2 is commonly used

  related to number of bits that could be used to represent that value

Example: Information amount

- Example:

  x is a symbol from the alphabet $\Omega = \{a, b, c, d\}$

  Probabilities: $\{0.4, 0.3, 0.2, 0.1\}$

  $$I_1 = -\log_2 0.4 \approx 1.32$$
  $$I_2 = -\log_2 0.3 \approx 1.74$$
  $$I_3 = -\log_2 0.2 \approx 2.32$$
  $$I_4 = -\log_2 0.1 \approx 3.32$$

  I_i is the minimum number of bits we could use to encode the symbol

Entropy

- Entropy:
  - Measures the average information amount of a r.v.
  - Summation on discrete r.v.; Integral on continuous r.v.

  Definition (Entropy):

$$H(x) = -\sum_{1}^{m} p_i \log p_i$$

- Continuous r.v.:

$$H(x) = -\int_{-\infty}^{\infty} p(x) \log p(x)$$

- Natural logarithm is common in continuous r.v.
- Normal distribution has maximum entropy

Relative Entropy

- Aka Kullback-Leibler distance or cross entropy
  - Measures the “distance” between two distributions
  - Two distributions over x that takes values $\Omega = \{v_1, v_2, \ldots, v_m\}$:

$$p(x) = \{p_1, p_2, \ldots, p_m\} \text{ and } q(x) = \{q_1, q_2, \ldots, q_m\}$$

Definition (Relative Entropy):

$$D_{KL}(p(x), q(x)) = \sum_{1}^{m} q_i \ln \frac{q_i}{p_i}$$

- Continuous r.v.:

$$D_{KL}(p(x), q(x)) = \int_{-\infty}^{\infty} q(x) \ln \frac{q(x)}{p(x)} dx$$

- $D_{KL}\left(p(x), q(x)\right) = 0$ iff $p(x) = q(x)$

Mutual Information

- Given two discrete r.v., x and y, w.p.:

  $p(x) = \{p_1, p_2, \dots, p_n\}$ and
  $q(y) = \{q_1, q_2, \dots, q_m\}$, and

  $r(x, y)$ is the joint prob of x and y

Definition (Mutual Information):

$$I(p;q) = H(p) - H(p|q)$$
$$= \sum_{x,y} r(x,y) \log_2 \frac{r(x,y)}{p(x)q(x)}$$

- Mutual information measures how much the distributions of the variables differ from statistical independence.

Relationships among entropy, mutual information and conditional entropies

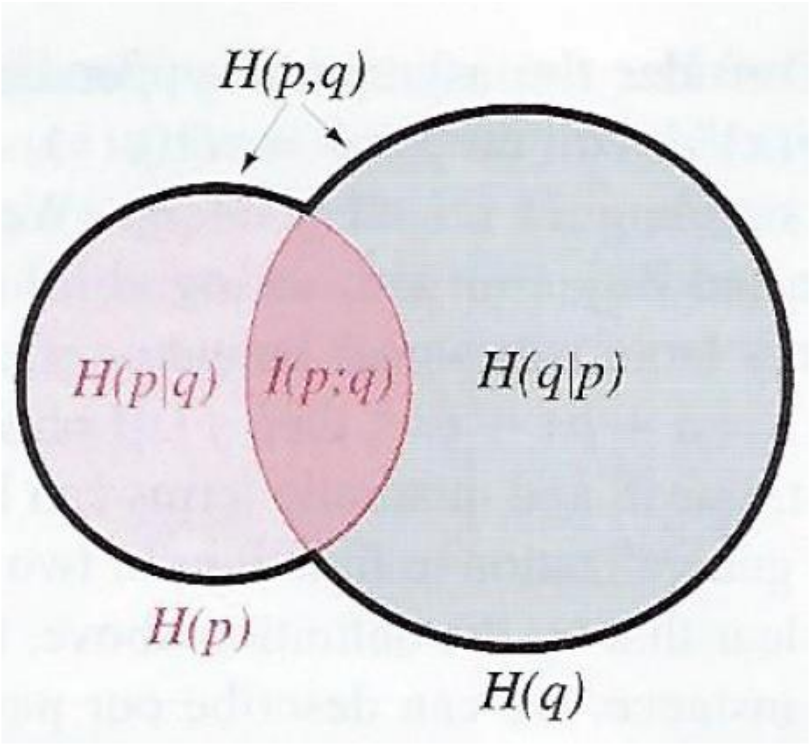


Figure from Duda et al.

Example: Data Encoding

Entropy:

$$\mathcal{H}(\mathcal{S}) = \mathcal{H} = -\sum_{i=1}^{m} p_i \log_r p_i = \sum_{i=1}^{m} p_i \mathcal{I}_i$$

Average code word length of encoding:

$$\bar{\ell} = \sum_{i=1}^{m} \ell_i p_i$$

where ℓ_i is the length of encoding symbol s_i

Example:

$$\mathcal{S} = \{a,b,c,d\}, \ \mathcal{A} = \{0,1\}, \ \mathcal{P} = [0.4, 0.3, 0.2, 0.1]$$

$$\mathcal{C} \to \mathcal{S}: \ a \to 0, \ b \to 10, \ c \to 110, \ d \to 111$$

$$\mathcal{H}(\mathcal{S}) = (0.4)(1.32) + (0.3)(1.74) + (0.2)(2.32) + (0.1)(3.32) \approx 1.846$$

$$\bar{\ell} = 1(0.4) + 2(0.3) + 3(0.2) + 3(0.1) = 1.9$$

Shannon's First Theorem

<u>Shannon's First Theorem:</u>
For any source, there exists at least one encoding scheme (need not be the optimal) such that:

$$\mathcal{H} \leq \bar{\ell} < \mathcal{H} + 1$$

In the previous example:

$$1.846 \leq 1.9 < 2.846$$

Optimization

- Given a function $f: \Re^d \rightarrow \Re$
- Optimization problem

$$\begin{array}{ll} \text{minimize} & f(\mathbf{x}) \\ \text{subject to} & \mathbf{x} \in \Omega \end{array}$$

- Called constrained optimization problem,

  where Ω is the constraint set

- Unconstrained optimization problem:

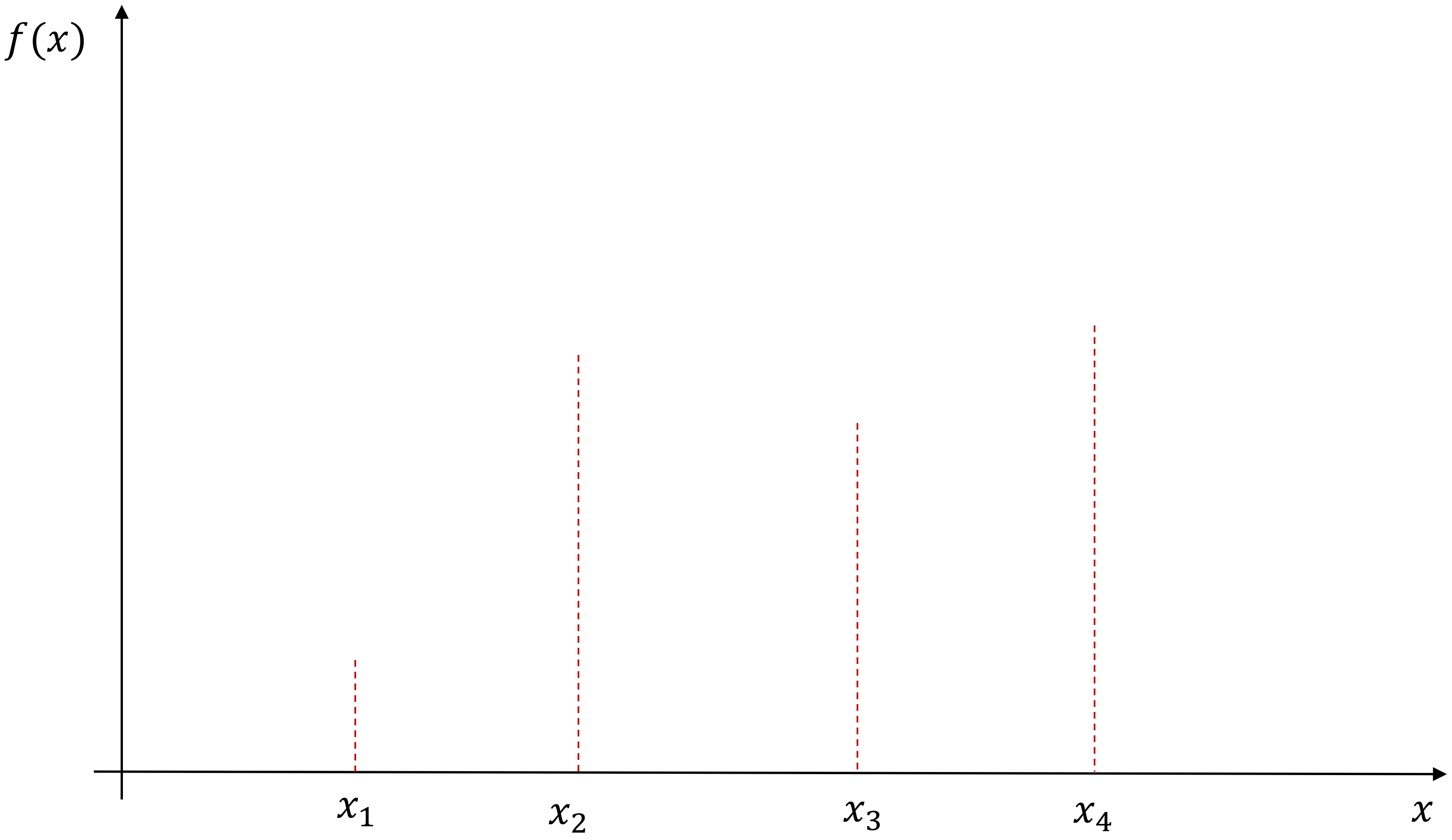
$$\Omega = \Re^d$$

- Here, we want to *minimize* f
- If we want to maximize f,

  reformulate the problem:

$$\begin{array}{ll} \text{minimize} & -f(\mathbf{x}) \\ \text{subject to} & \mathbf{x} \in \Omega \end{array}$$

Minimizers



x_1: strict global minimizer
x_2: inflection (saddle) point
x_3: strict local minimizer
x_4: local (not strict) minimizer

The same principles apply to maximizers, i.e., $-f(x)$

Conditions for minimizers

- First order necessary condition (FONC):

$$\nabla f(\mathbf{x}) = \frac{\partial}{\partial \mathbf{x}} f(\mathbf{x}) = \mathbf{0}$$

- A vector $\mathbf{x}^*$ that satisfies the FONC can be
  - A local minimizer
  - A global minimizer
  - A saddle point

Conditions for minimizers (maximizers)

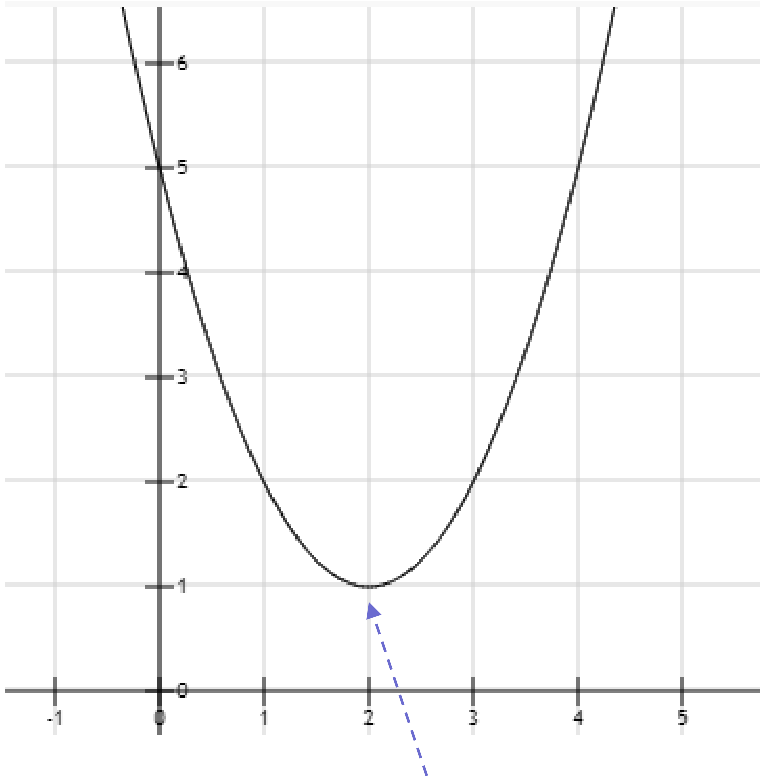
- Given $\mathbf{x}^*$, such that $\nabla f(\mathbf{x}^*) = \mathbf{0}$
- Second order sufficient condition (SOSC):

$$\mathbf{H} = \frac{\partial}{\partial \mathbf{x}} \nabla f(\mathbf{x}) = \frac{\partial^2}{\partial \mathbf{x}} f(\mathbf{x})$$

- One-dimensional optimization:
  - $\mathbf{H}(\mathbf{x}^*) > 0$ yields a minimum
  - $\mathbf{H}(\mathbf{x}^*) < 0$ yields a maximum
  - $\mathbf{H}(\mathbf{x}^*) = 0$ yields a saddle point

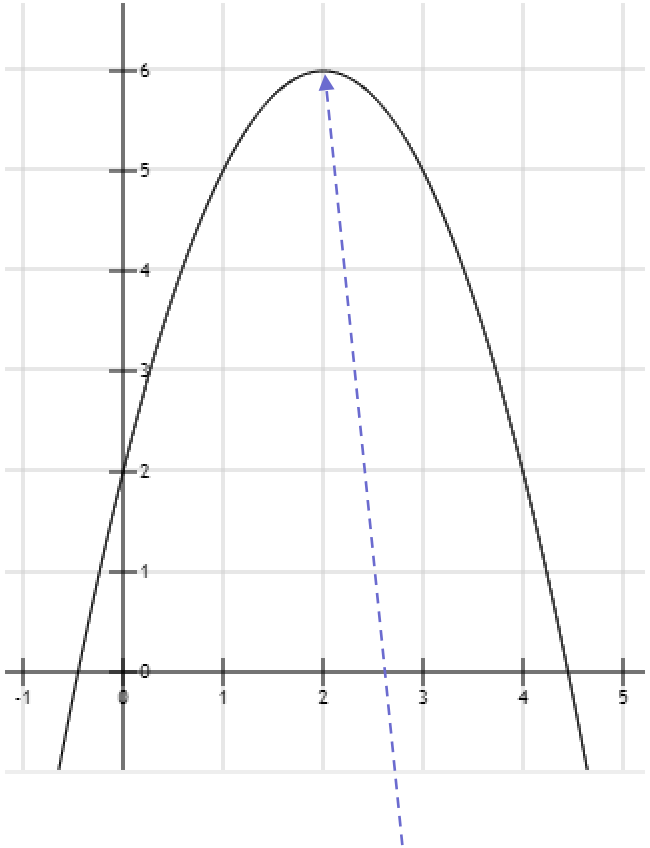
Examples

$(x-2)^2+1$



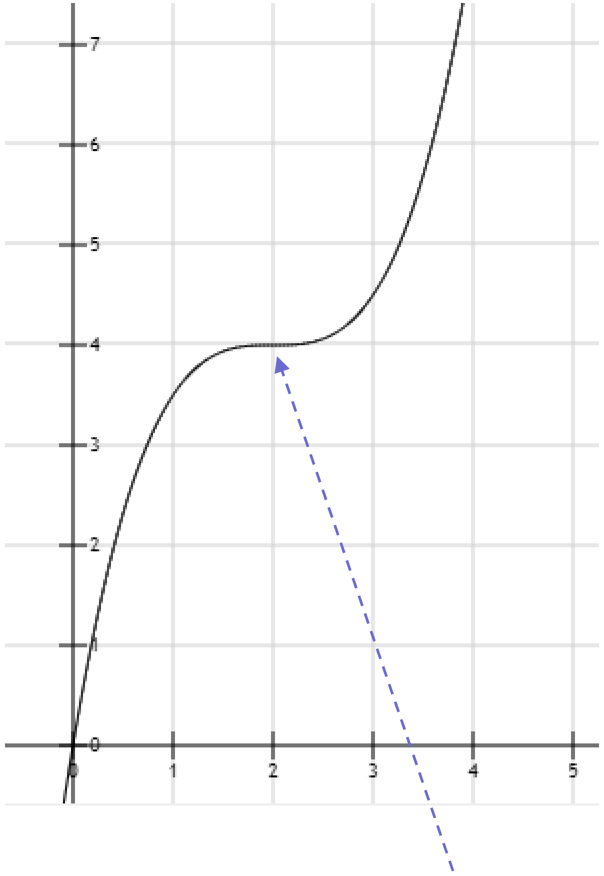
$x^* = 2$ is a minimizer

$-(x-2)^2+6$



$x^* = 2$ is a maximizer

$0.5*(x-2)^3+4$



$x^* = 2$ yields a saddle point

SOSC - Multivariate optimization

- Given $\mathbf{x}^*$, such that $\nabla f(\mathbf{x}^*) = \mathbf{0}$
- Second order sufficient condition (SOSC):

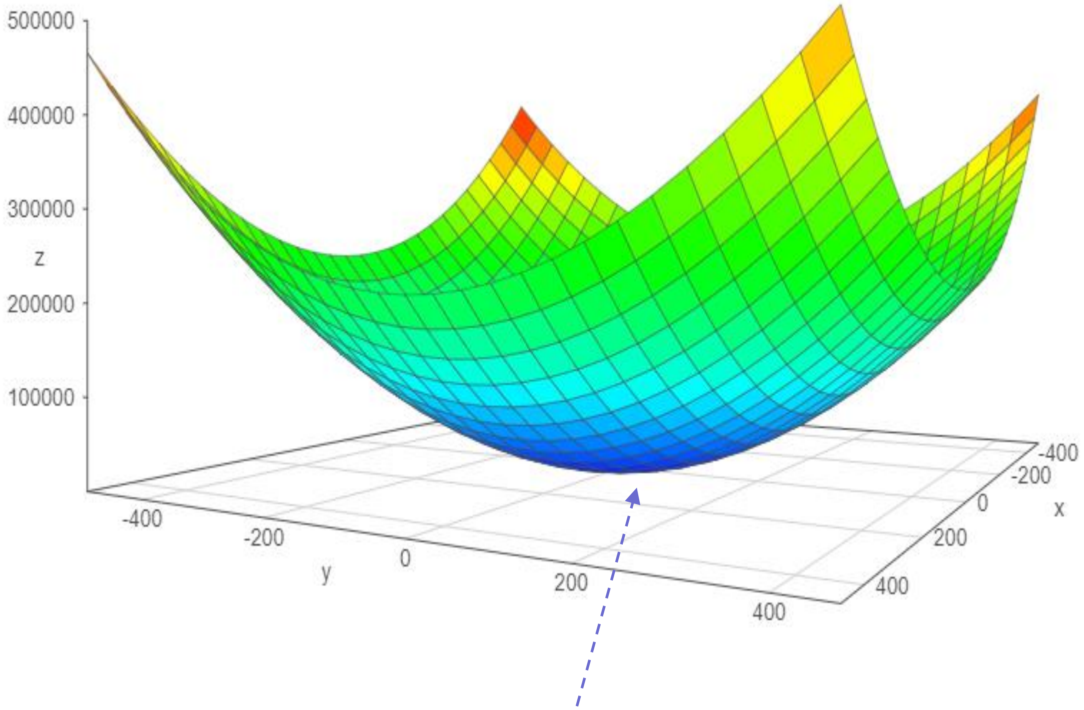
$$\mathbf{H} = \frac{\partial}{\partial \mathbf{x}} \nabla f(\mathbf{x}) = \frac{\partial^2}{\partial \mathbf{x}} f(\mathbf{x})$$

- Then
  - $\mathbf{H}(\mathbf{x}^*) > 0$, i.e., $\mathbf{H}$ is positive definite yields a minimum
  - $\mathbf{H}(\mathbf{x}^*) < 0$, i.e., $\mathbf{H}$ is negative definite yields a maximum
  - $\mathbf{H}(\mathbf{x}^*)$ is indefinite yields a saddle point
    - indefinite means some eigenvalues > 0, some < 0 and/or some = 0

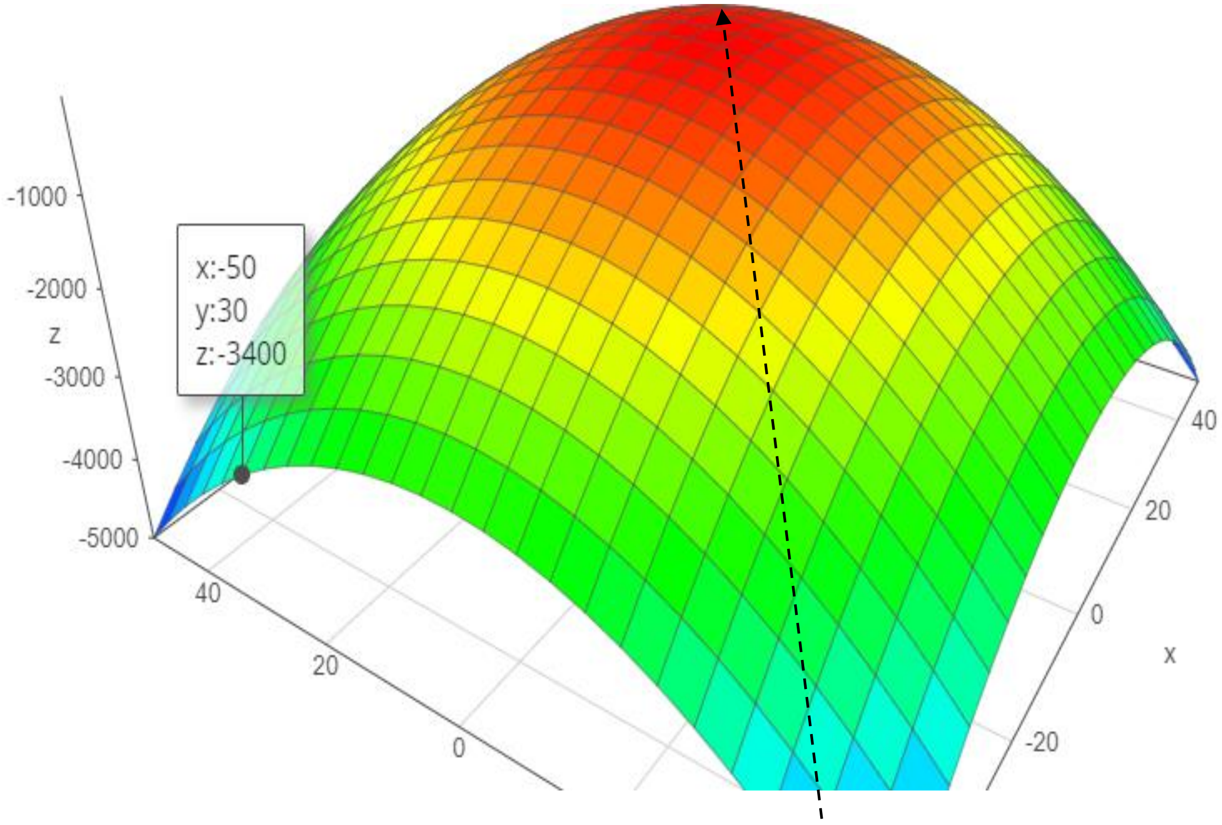
Examples

$$f(\mathbf{x}) = x_1^2 + x_2^2$$
$$\nabla f(\mathbf{x}) = [2x_1, 2x_2]^t$$
$$\mathbf{H}([0,0]^t) = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} > 0$$

$$f(\mathbf{x}) = -x_1^2 - x_2^2$$
$$\nabla f(\mathbf{x}) = [-2x_1, -2x_2]^t$$
$$\mathbf{H}([0,0]^t) = \begin{bmatrix} -2 & 0 \\ 0 & -2 \end{bmatrix} < 0$$



$\mathbf{x}^* = [0,0]^t$ is a minimizer



$\mathbf{x}^* = [0,0]^t$ is a maximizer

Example

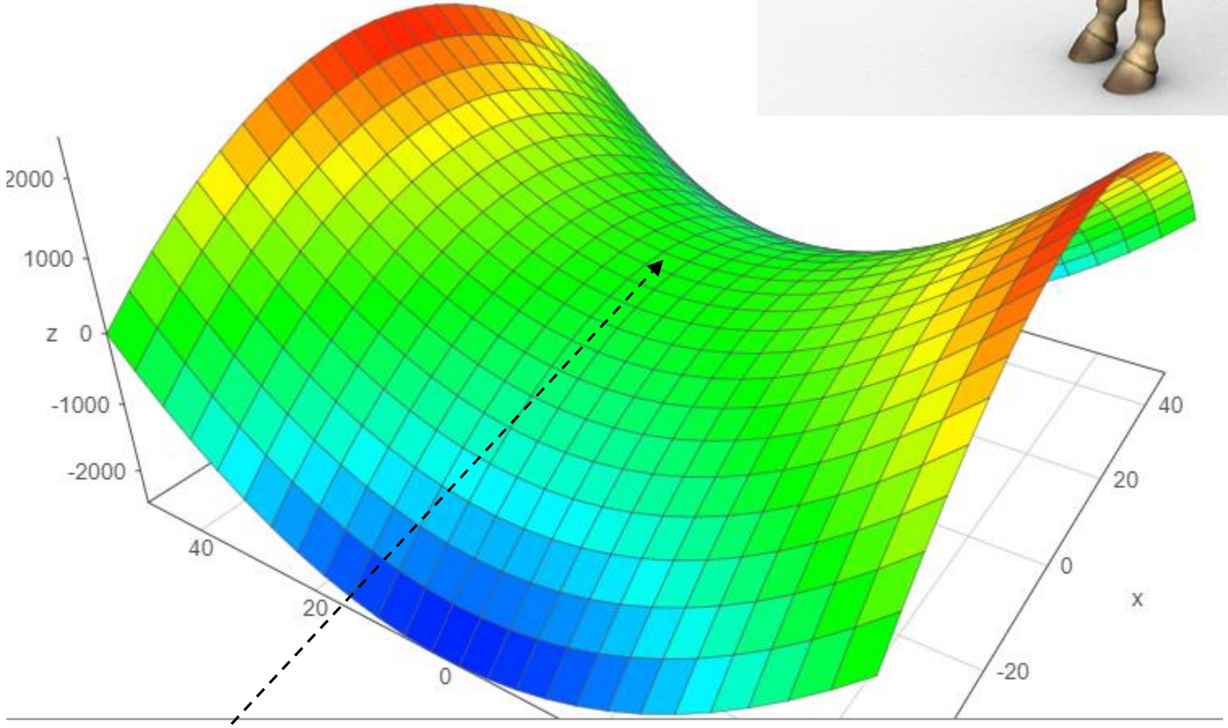
$$f(\mathbf{x}) = -x_1^2 + x_2^2$$

$$\nabla f(\mathbf{x}) = [-2x_1, 2x_2]^t$$

$$\mathbf{H}([0,0]^t) = \begin{bmatrix} -2 & 0 \\ 0 & 2 \end{bmatrix}$$

is indefinite

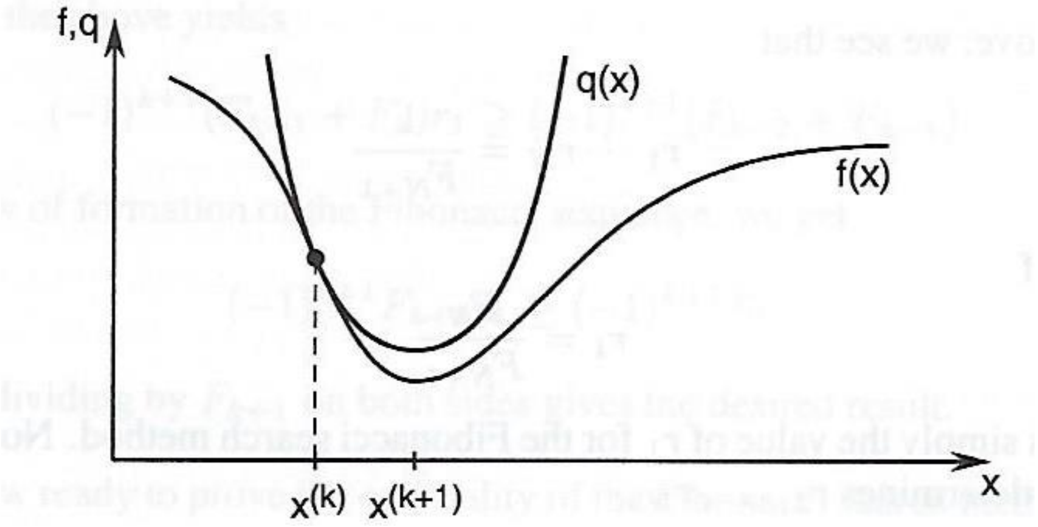
- yields a maximum in direction of eigenvector 1
- yields a minimum in direction of eigenvector 2

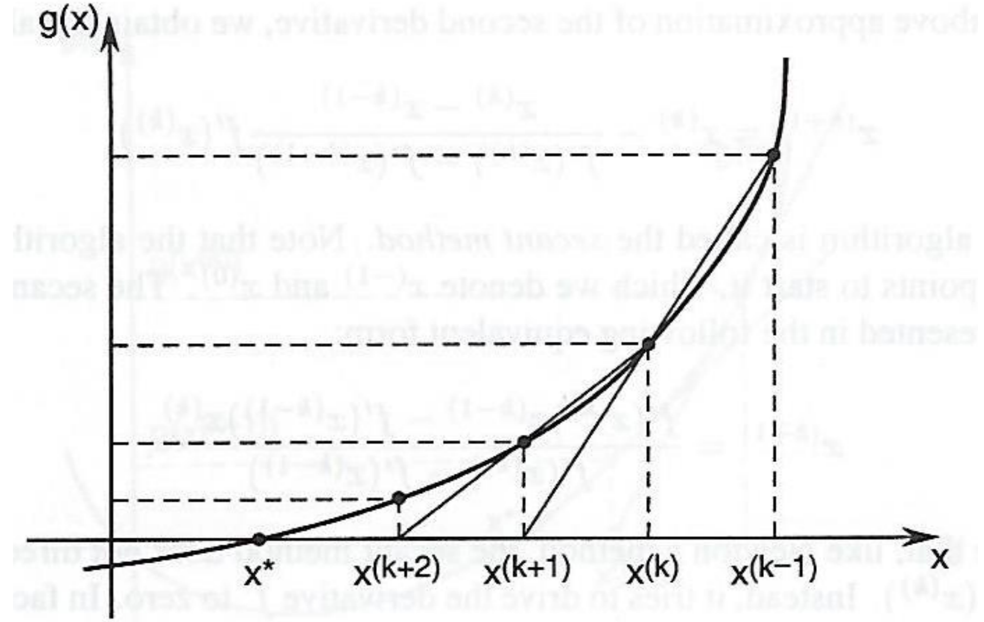


$\mathbf{x}^* = [0,0]^t$ is a saddle point

Iterative Methods for Optimization

- Golden search
- Fibonacci search
- Gradient methods
  - Uses first order derivative
- Newton's method
  - Uses second order derivatives
- Secant method
  - When second order derivative
  - Uses approximation
- Stochastic methods





Gradient Method

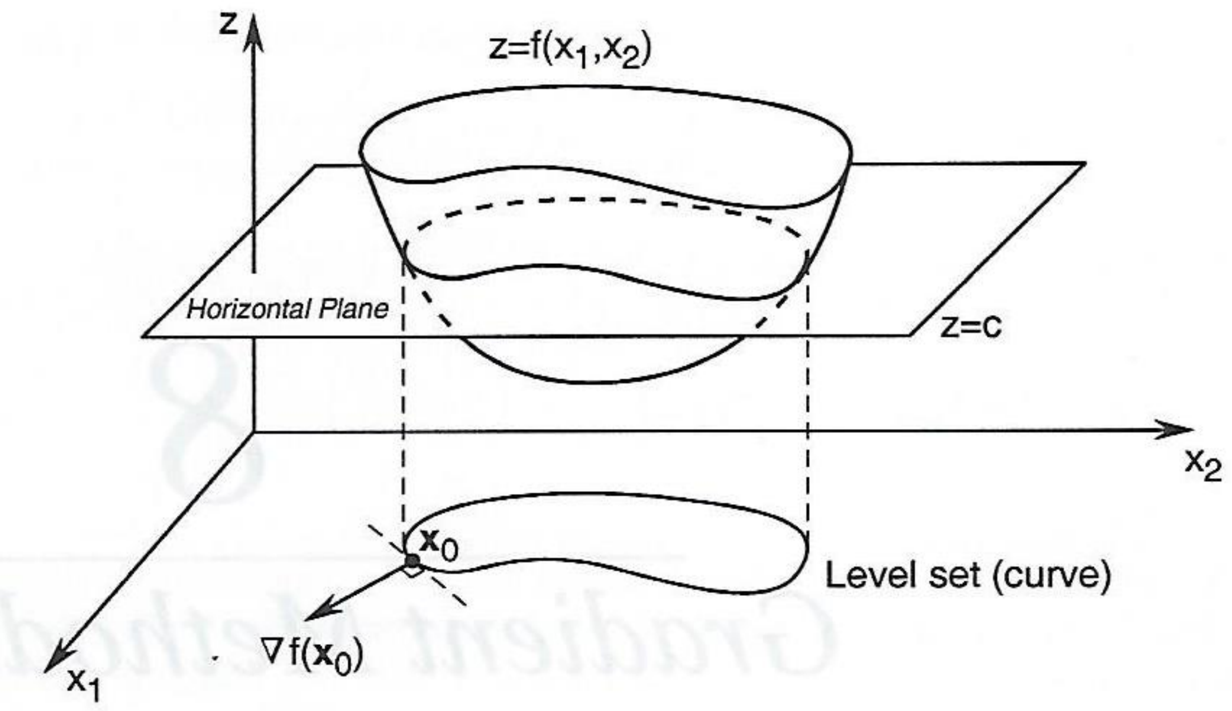
- Uses the gradient operator ∇
- Given a differentiable function

$$f: \Re^d \rightarrow \Re$$

- the gradient ∇ of f :

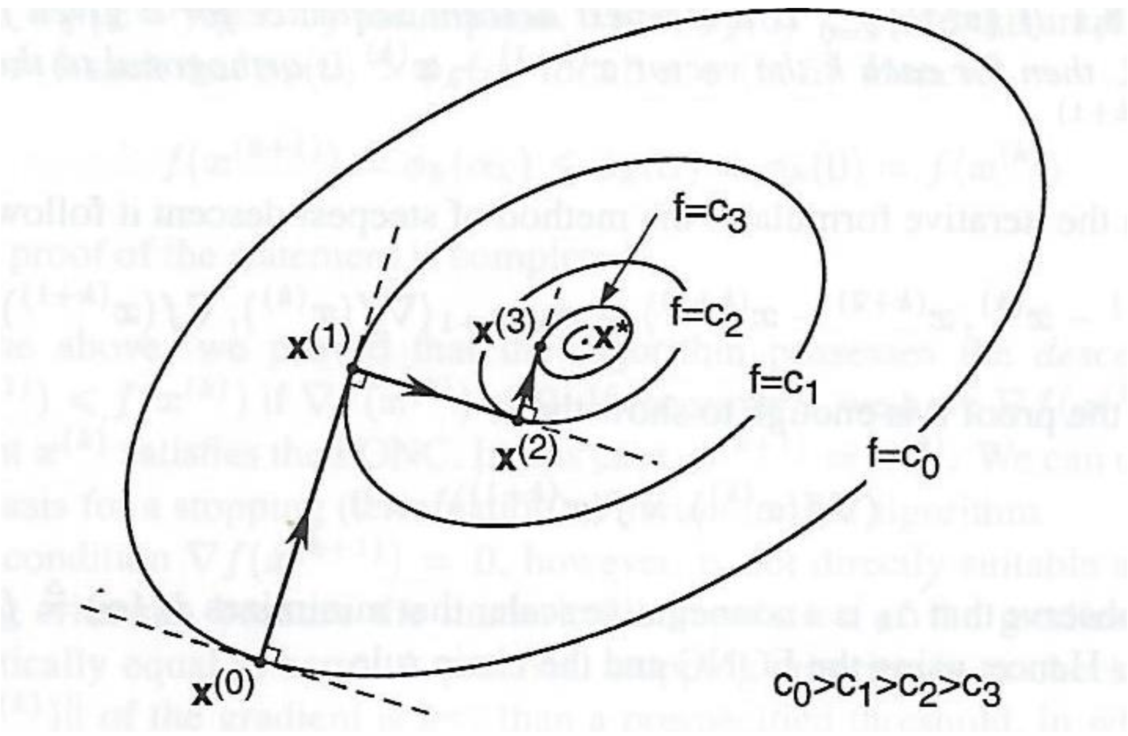
$$\nabla f(\mathbf{x}) = \frac{\partial}{\partial \mathbf{x}} f(\mathbf{x})$$

- On a particular point $\mathbf{x}_0$
- $-\nabla f(\mathbf{x}_0)$ gives the steepest direction of descent



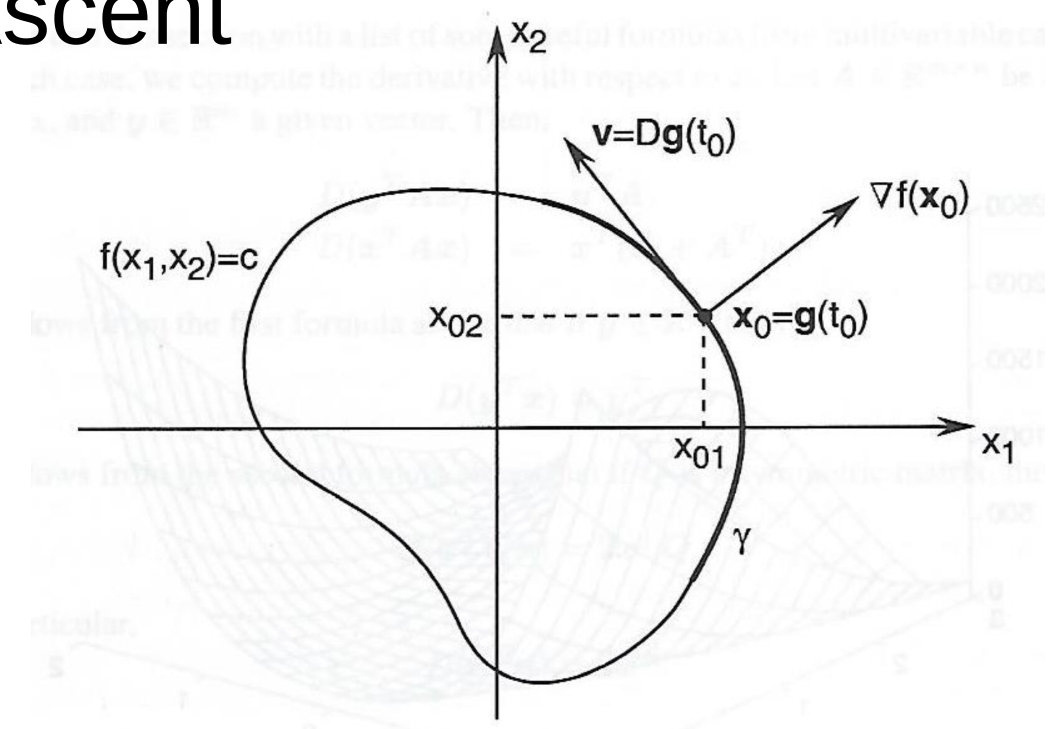
Gradient Descent

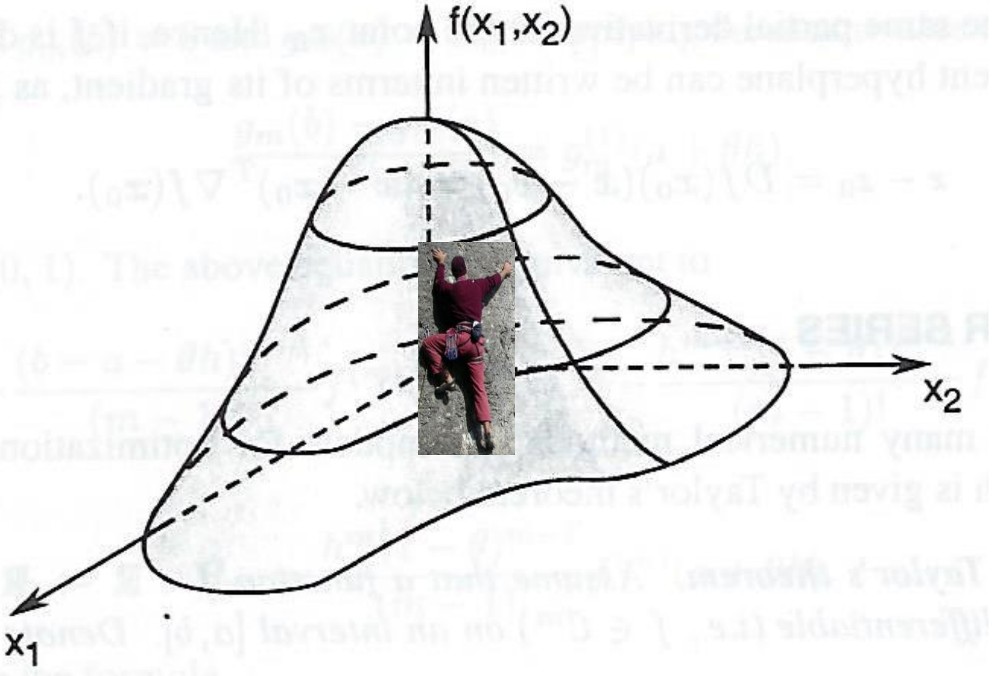
- Iterative optimization
- Start with $\mathbf{x}^{(0)}$
- Repeat
  - $\mathbf{x}^{(k+1)} \leftarrow \mathbf{x}^{(k)} - \eta_k \nabla f\left(\mathbf{x}^{(k)}\right)$
- Until small change
  - $\left|f\left(\mathbf{x}^{(k+1)}\right) - f\left(\mathbf{x}^{(k)}\right)\right| < \theta$
- η_k is called learning rate
- Can be optimized as:
  - $\eta_k \leftarrow \arg\min_{\eta \geq 0} f\left(\mathbf{x}^{(k)}\right) - \nabla f\left(\mathbf{x}^{(k)}\right)$

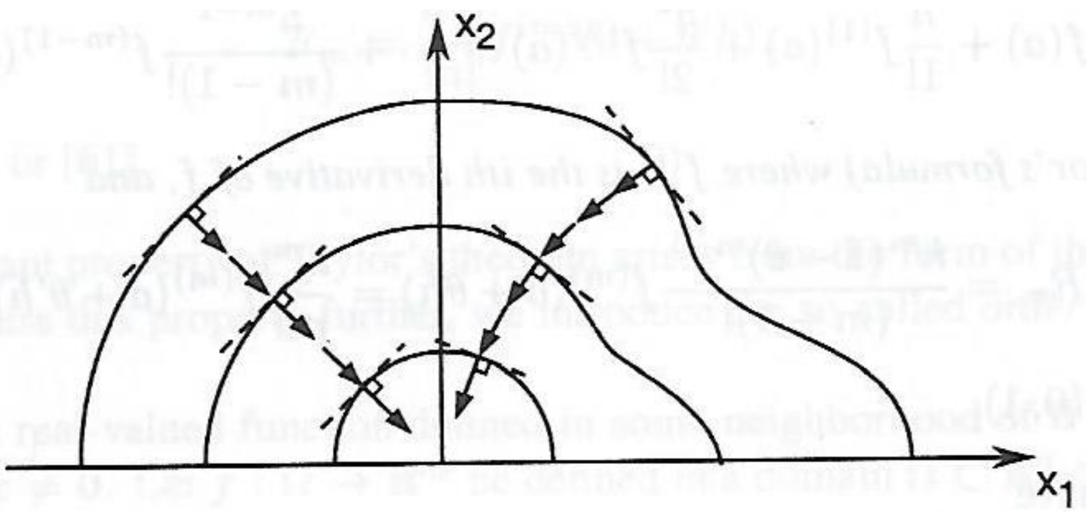


Steepest Ascent

- $\nabla f(\mathbf{x}_0)$ is the maximum rate of increase of f at x_0
- Different level sets will give different directions:
  - of maximum rate of increase
  - called *path of steepest ascent*

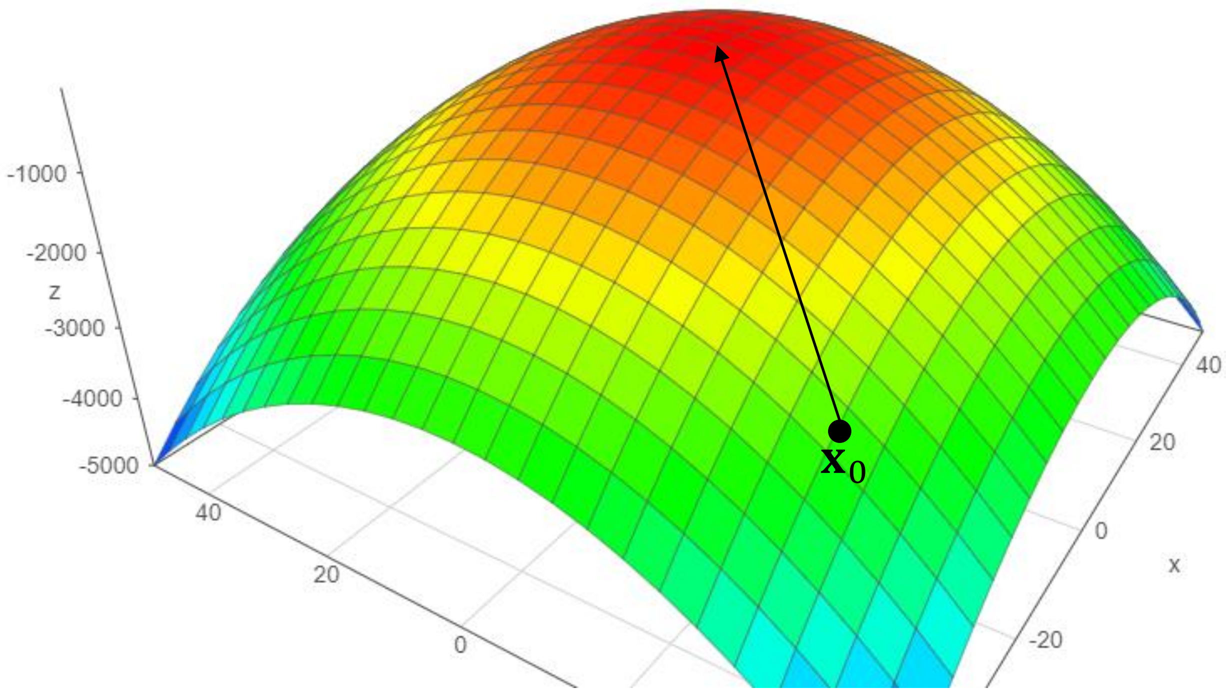


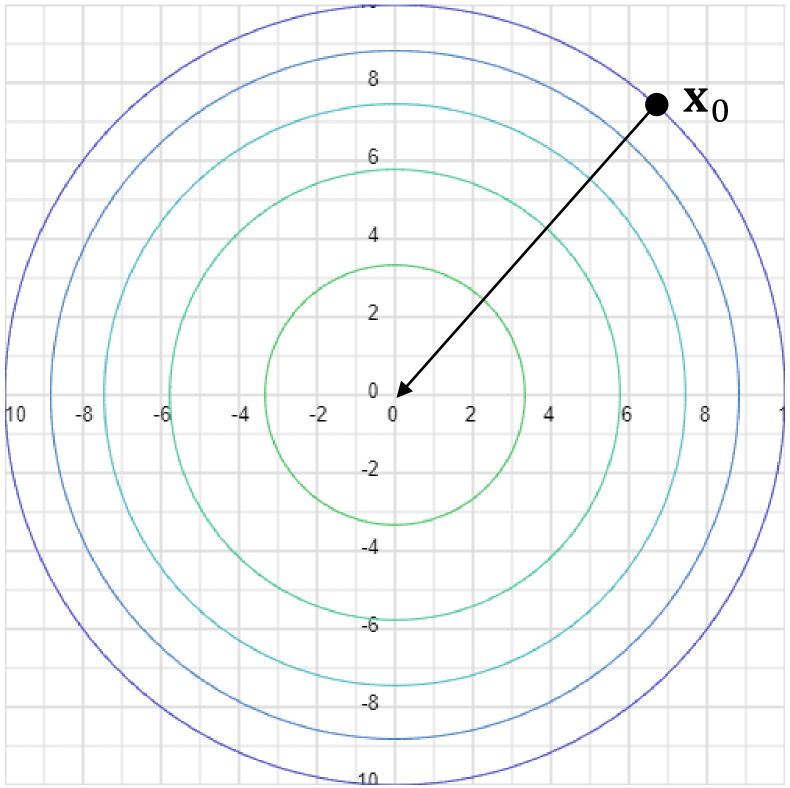




Example 2

- $f(\mathbf{x}) = -x_1^2 - x_2^2$
- In this example, path of steepest ascent always leads to maximum of f in *one* step





Constrained Optimization

- Formulated as follows

$$\begin{array}{ll} \text{minimize} & f(\mathbf{x}) \\ \text{subject to} & h_i(\mathbf{x}) = 0, \;\; i = 1, \dots, m \\ & g_j(\mathbf{x}) \leq 0, \; j = 1, \dots, p \end{array}$$

- $h_i(\mathbf{x}) = 0$ are equality constraints
- $g_j(\mathbf{x}) \leq 0$ are inequality constraints

Equality constraints

- Lagrange multipliers

minimize $\quad f(\mathbf{x})$

subject to $\quad h_i(\mathbf{x}) = 0, \quad i = 1, \ldots, m$

- Transform it into:

minimize $\quad L(\mathbf{x}, \lambda) = f(\mathbf{x}) - \sum_i \lambda_i h_i(\mathbf{x})$

- λ_i are called Lagrange multipliers

Inequality constraints

- Karush-Kuhn-Tucker (KKT) condition (approach)

minimize $f(\mathbf{x})$

subject to $h_i(\mathbf{x}) = 0, \;\; i = 1, \ldots, m$

$g_j(\mathbf{x}) \leq 0, \; j = 1, \ldots, p$

- Transform it into:

minimize $L(\mathbf{x}, \lambda) = f(\mathbf{x}) - \sum_i \lambda_i \, h_i(\mathbf{x}) - \sum_j \alpha_j \, g_j(\mathbf{x})$

subject to $\alpha_j \geq 0$

Convex Optimization

Optimization problem:

minimize $\quad f(\mathbf{x})$

subject to $\quad \mathbf{x} \in \Omega$

- where $\Omega \in \mathcal{R}$ is a convex region

Example:

minimize $\quad f(\mathbf{x}) = x_1^2 + x_2^2$

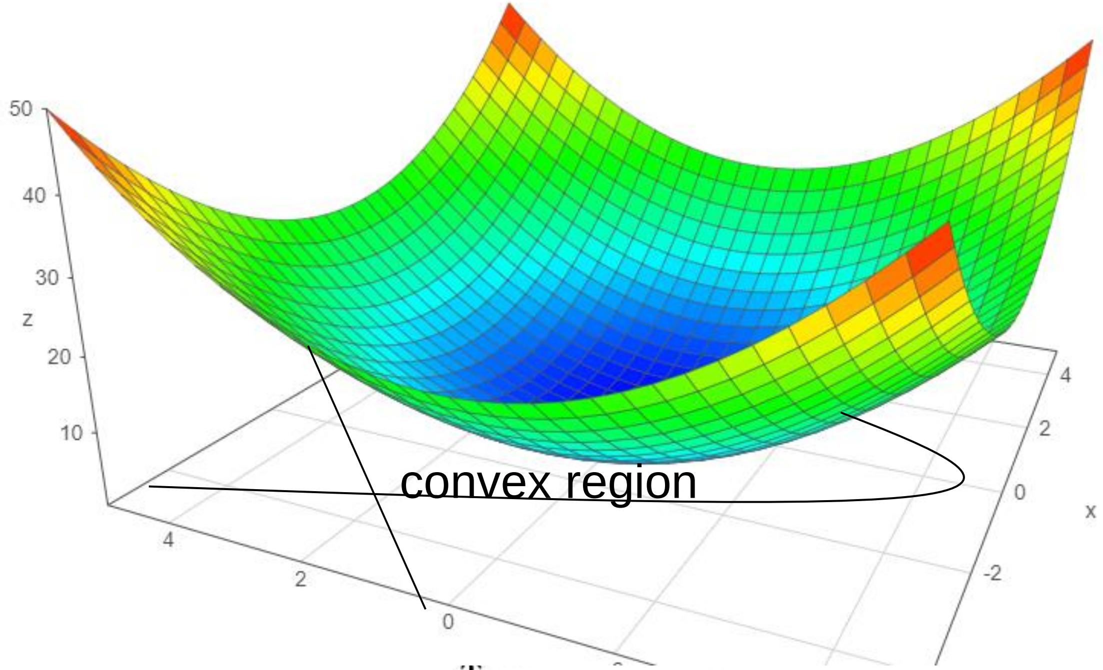
subject to $\quad x_2 - x_1 - 2 \leq 0$

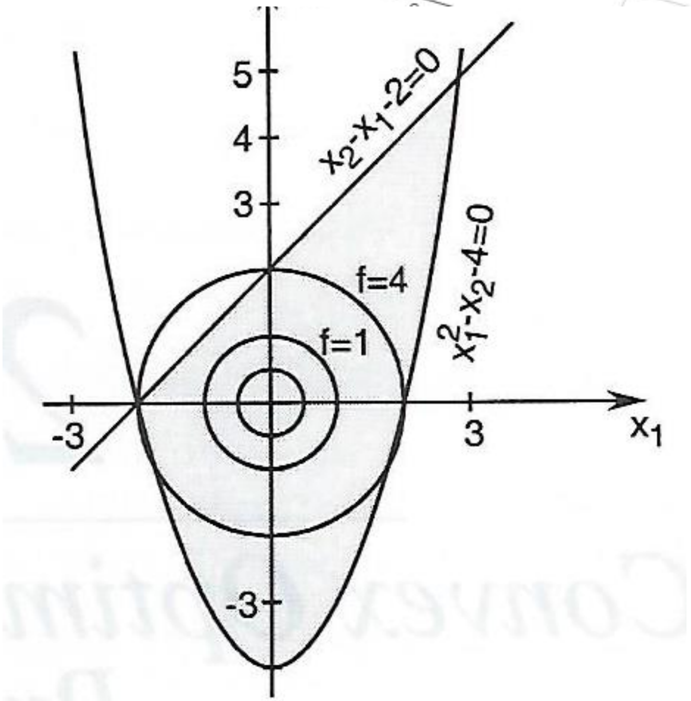
$\quad x_1^2 - x_2 - 4 \leq 0$

A function $f: \Omega \rightarrow \mathcal{R}$ defined on a convex set $\Omega \in \mathcal{R}^d$ has a global minimizer x^* over Ω if f is convex on Ω

That is, Ω and f are both convex

The KKT are satisfied for x^*





© Luis Rueda, 2019

References

1. D. Harville, Matrix Algebra from a Statistician's Perspective, Springer, 1997
2. D. Hankerson et al., Introduction to Information Theory and Data Compression, CRC Press, 1997
3. E. Chong et al., An Introduction to Optimization. 2nd Edition, Wiley, 2001
4. R. Duda et al., Pattern Classification. 2nd Edition, Wiley, 2001