





Sistemi Intelligenti
Relazione tra ottimizzazione e
statistica - IV

Alberto Borghese

Università degli Studi di Milano
Laboratory of Applied Intelligent Systems (AIS-Lab)
Dipartimento di Informatica
borgnese@di.unimi.it



Sommario

- Analisi dell'affidabilità della stima
- Metodo del gradiente
- Linearizzazione e metodo di Gauss-Newton

A.A. 2016-2017 2/31 <http://borgnese.di.unimi.it/>




Valutazione della bontà della stima

$$x = (A' * A)^{-1} A' * b \iff \min_x \sum_k v_k^2 = \min_x (Ax - b)^2$$



Errore di modellizzazione Gaussiano a media nulla $N(0, \sigma^2)$

$$\langle v_k \rangle = 0$$

$$\hat{\sigma}_0^2 = \sum_{k=1}^M (v_k^2) = |v|^2$$

Varianza della stima = varianza dell'errore di misura

A.A. 2016-2017 3/31 http://borghese.di.unimi.it/

Valutazione della bontà della stima del singolo parametro

$$x = (A' * A)^{-1} A' * b$$

$$x = CA' * b \quad \hat{\sigma}_0^2 = \sum_{m=1}^M (v_m^2)$$

Chiamiamo u e v le variabili casuali associate all'errore sui parametri e all'errore di modellizzazione, rispettivamente. Si suppone errore a media nulla e Gaussianamente distribuito.



$$u = \Delta x \quad (x + u) = C A' (b + v)$$

$$\downarrow$$

$$x = CA' b \quad u = C A' * v \quad E[u] = 0$$

C è la matrice di covarianza

A.A. 2016-2017 4/31 http://borghese.di.unimi.it/

Impostazione del calcolo della correlazione tra i parametri

$u = C A' v$

Vogliamo individuare la correlazione tra due parametri i e j . Devo quindi determinare la loro correlazione:



$$\begin{bmatrix} u_1^2 & u_1 u_2 & \dots & u_1 u_W \\ u_2 u_1 & u_2^2 & \dots & u_2 u_W \\ \dots & \dots & \dots & \dots \\ u_W u_1 & u_W u_2 & \dots & u_W^2 \end{bmatrix}$$

$\langle u_i, u_j \rangle$
 $u = C A' v \quad \Rightarrow \quad u' = v' A (C)'$

$uu' = C A' v v' A C' \Rightarrow$ Applicando l'operatore di media, si ottiene:
 $\langle uu' \rangle = C A' \langle v v' \rangle A C'$

Dato che v sono i residui, e sono indipendenti, e tutte i punti di controllo hanno lo stesso tipo di errore di misura, si avrà che $\langle v v' \rangle = I \sigma_0^2$.

A.A. 2016-2017 5/31 http://borghese.di.unimi.it/

Incertezza sulla stima dei parametri

$\langle uu' \rangle = C A' I A C' \sigma_0^2 = C' \sigma_0^2$ $\langle u' u \rangle = C \sigma_0^2$

Segue che: $\sigma^2(u_{ij}) = c_{ij} \sigma_0^2$ Varianza sulla stima del parametro.

Spiegazione intuitiva:



a $x + 3 = y + \text{noise}$

Calcolo x come: $x = y * a^{-1} - 3$

Quanto è sensibile questa stima? Cosa succede se, per effetto del noise, invece di misurare y , misuro $y + v$?

x varierà di $v * a^{-1}$. Il rumore viene cioè moltiplicato per a^{-1} .

A.A. 2016-2017 6/31 http://borghese.di.unimi.it/

Matrice di covarianza



Date N variabili casuali: $x = [x_1, x_2, \dots, x_N]$ si può misurare la correlazione tra coppie di variabili. E' comodo rappresentare la correlazione tra variabili casuali in un'unica matrice detta **matrice di covarianza** come:

$$C = \begin{bmatrix} \sigma_{x_1 x_1} & \sigma_{x_1 x_2} & \cdot & \sigma_{x_1 x_N} \\ \sigma_{x_2 x_1} & \sigma_{x_2 x_2} & \cdot & \sigma_{x_2 x_N} \\ \cdot & \cdot & \cdot & \cdot \\ \sigma_{x_N x_1} & \sigma_{x_N x_2} & \cdot & \sigma_{x_N x_N} \end{bmatrix}$$

Varianza: $\sigma_{x_i x_i} = \sigma^2_{x_i}$ N parametri

Covarianza: $\sigma_{x_i x_j} = \sigma_{x_j x_i}$ $i \neq j$ $(N-1)^2/2$ parametri

A.A. 2016-2017 7/31 <http://borghese.di.unimi.it/>

Correlazione tra coppie di parametri

Date due variabili casuali: x_i, x_j , l'indice di correlazione misura quanto le coppie di variabili estratte: $p(x_i, x_j)$ stanno su una retta:

$$r = \frac{M_{x_i x_j} - M_{x_i} M_{x_j}}{\sigma_{x_i} \sigma_{x_j}} \quad -1 \leq r \leq +1$$



Definendo la covarianza tra x_i ed x_j come:

$$\sigma_{x_i x_j} = \frac{1}{N} \sum_i \sum_j (x_i - M_{x_i})(x_j - M_{x_j})$$

Dalla definizione di deviazione standard risulta:

$$r = \frac{\sigma_{x_i x_j}}{\sigma_{x_i} \sigma_{x_j}}$$

A.A. 2016-2017 8/31 <http://borghese.di.unimi.it/>

Correlazione tra i parametri

$\langle uu' \rangle = CA' IA C' \sigma_0^2 = C' \sigma_0^2$ $\langle u' u \rangle = C \sigma_0^2$

Da cui si giustifica il nome di matrice di covarianza per C.



Segue che: $\sigma^2(u_{ij}) = c_{ij} \sigma_0^2$ Varianza sulla stima del parametro.

$$-1 \leq r_{ij} = \frac{\langle u_i u_j \rangle}{\sqrt{\langle u_i \rangle^2 \langle u_j \rangle^2}} = \frac{c_{ij}}{\sqrt{c_i c_j}} \leq +1$$

Indice di correlazione tra il parametro i ed il parametro j
(empiricamente si scartano parametri quando la correlazione è superiore al 95%)

Vanno rapportati alle dimensioni dei parametri coinvolti.

A.A. 2016-2017 9/31 <http://borghese.di.unimi.it/>

La covarianza: momenti di 2 variabili statistiche

Covarianza = $E[(x - \mu_x)(y - \mu_y)]$

Varianza = $E[(x - \mu_x)(x - \mu_x)]$

Per due variabili indipendenti, la covarianza = 0, non variano assieme (covariano)


$$C = \begin{bmatrix} \sigma_x^2 & \sigma_x \sigma_y \\ \sigma_y \sigma_x & \sigma_y^2 \end{bmatrix}$$

```


>> x = randn(N,1);
>> y = randn(N,1);
>> temp = x.*y;
>> covarianza = mean(temp)

```

A.A. 2016-2017 10/31 <http://borghese.di.unimi.it/>



Misura di correlazione su 2 parametri



Misura la inter-dipendenza tra 2 variabili statistiche:


$$-1 \leq \frac{\sigma_{xy}}{\sigma_x \sigma_y} = c = \lim_{N \rightarrow \infty} \frac{\sum_k (x_k - \mu_x)(y_k - \mu_y)}{\sqrt{\sum_k (x_k - \mu_x)^2} \sqrt{\sum_k (y_k - \mu_y)^2}} \leq +1$$

```


>> x = randn(N,1);
>> y1 = randn(N,1);
>> y2 = x;
>> temp1 = x.*y1;
>> temp2 = x.*y2;
>> covarianza1 = mean(temp1)% Uncorrelated variables(c -> 1)
>> covarianza2 = mean(temp2)% Correlated variables (c = 0)

```

A.A. 2016-2017
11/31
<http://borghese.di.unimi.it/>



Caso 2D



N = 20 punti $\sigma_0^2 = 0.01$
m reale = 1 q reale = 2

C =

0.0494	0.0087
0.0087	0.0515

m stimato = 0.9931
q stimato = 2.0106

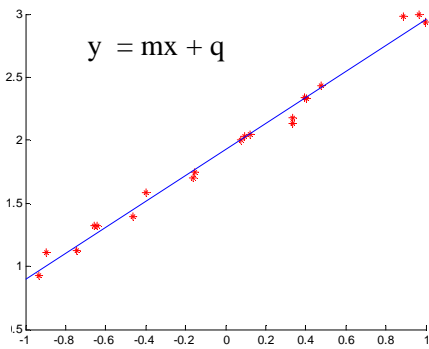
Altra realizzazione del rumore:

C =

0.1702	0.0124
0.0124	0.0509

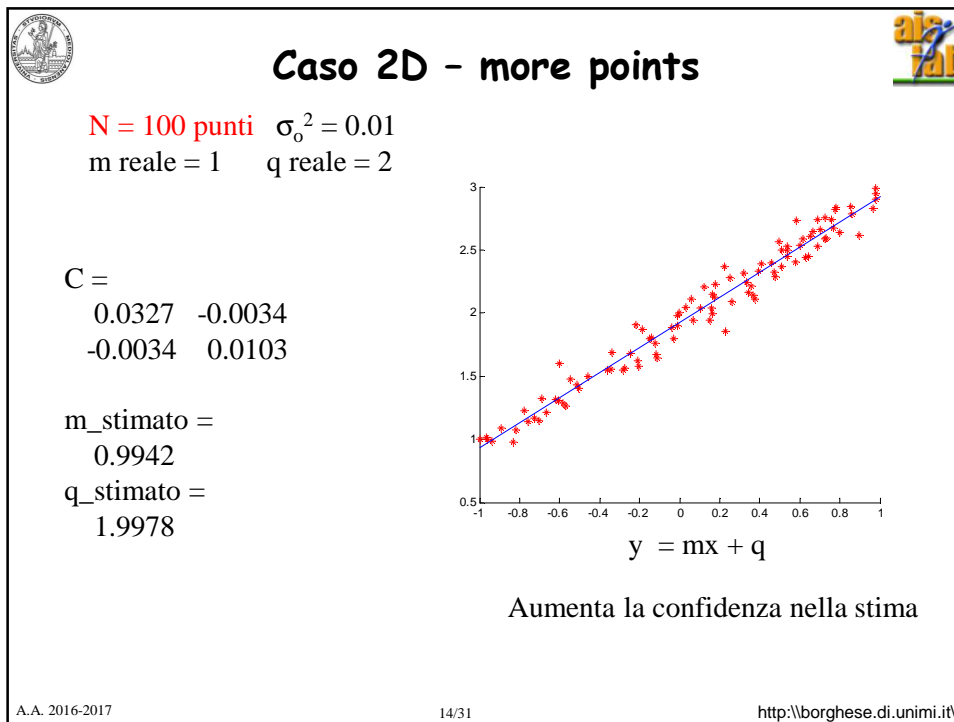
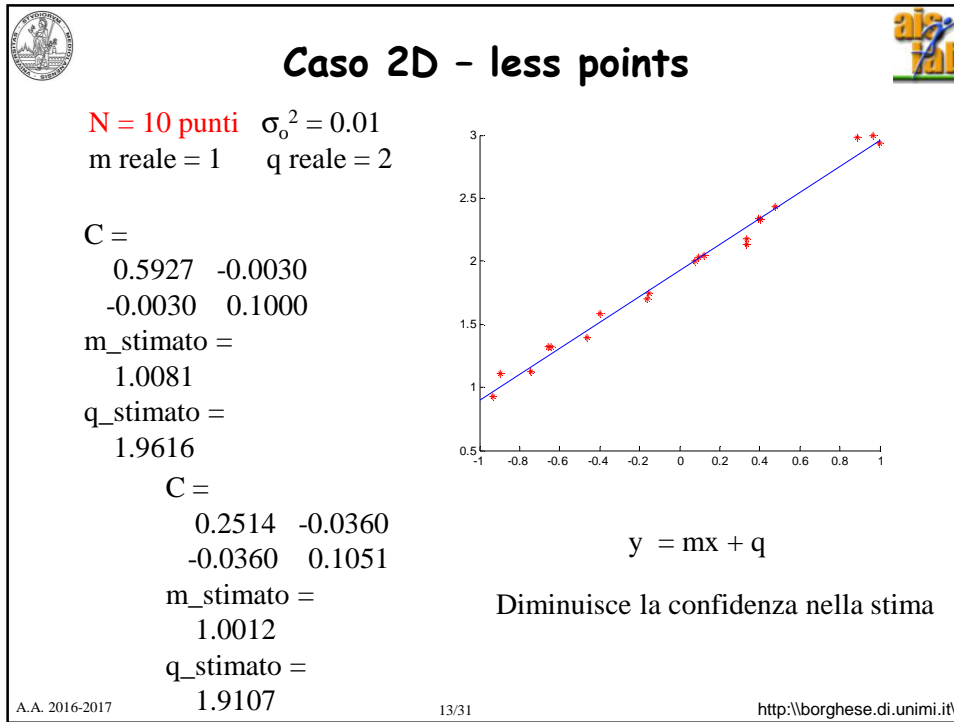
m stimato = 0.9937
q stimato = 1.9522

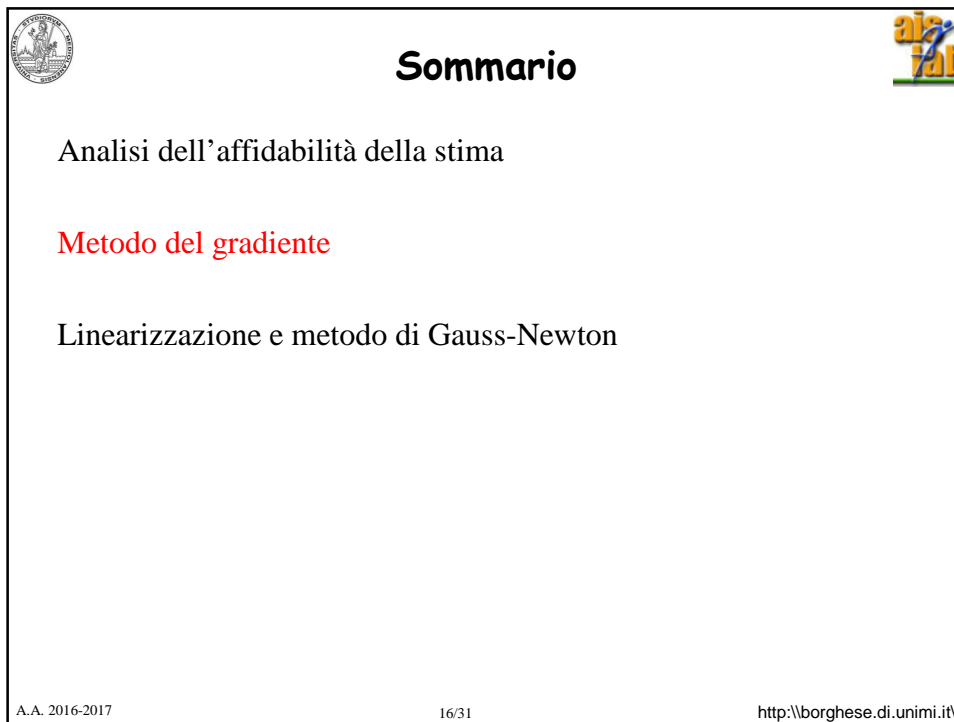
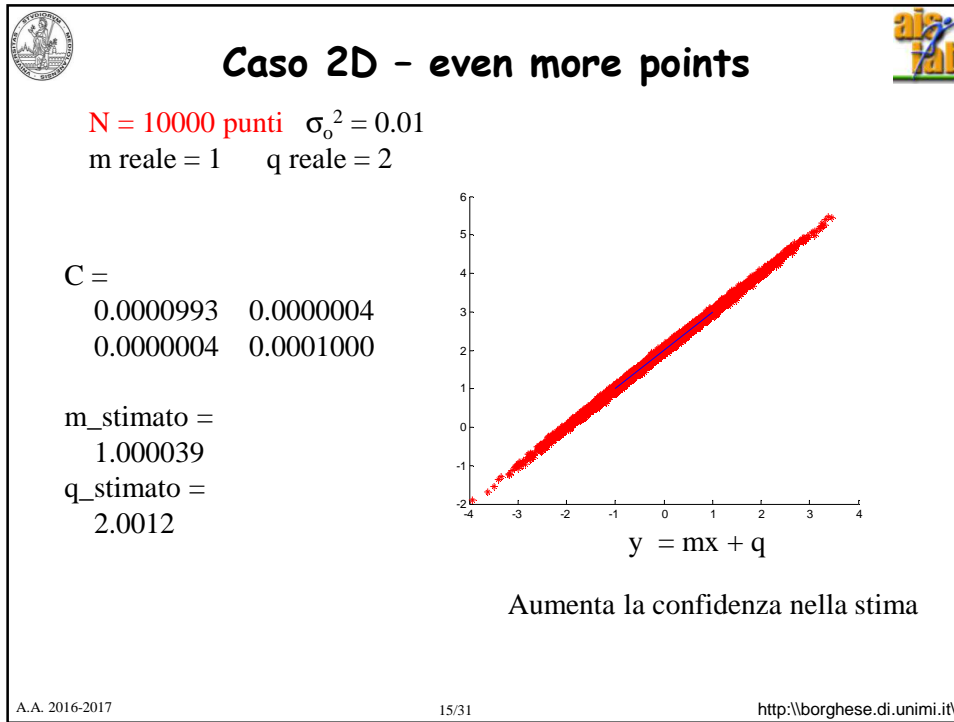
$y = mx + q$




$\sigma^2(m) = c_{11} \sigma_0^2 = 0.1702 * 0.01 = 0.0017 \Rightarrow \sigma = 0.04$
 $\sigma^2(q) = c_{22} \sigma_0^2 = 0.0515 * 0.01 = 0.0005 \Rightarrow \sigma = 0.022$


1
<http://borghese.di.unimi.it/>



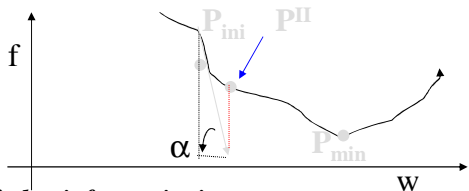




Minimizzazione tramite gradiente (metodo del primo ordine): 1 variabile



Tecnica del gradiente applicata alla minimizzazione di funzioni non-lineari di **una variabile**, x , e di **un parametro**, w : $f = f(x | w)$.



La derivata, mi dà due informazioni:


- 1) In quale direzione di w , la funzione decresce.
- 2) Quanto rapidamente decresce.

Definisco uno spostamento arbitrario lungo la pendenza della funzione f (e.g. errore) nello spazio dei parametri w : maggiore la pendenza maggiore lo spostamento.


$dw \propto -f'(w;P)$ dati P, w . La derivata viene calcolata rispetto a w .

Occorre un'inizializzazione.
Metodo iterativo.

mi.it



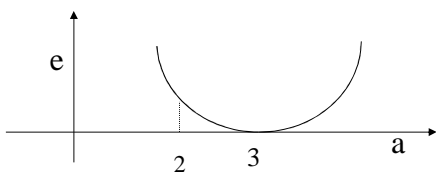
Esempio di applicazione tecnica del gradiente per funzioni di 1 variabile



*Supponiamo che il modello da noi considerato sia semplice: $y = ax^2$
Abbiamo un unico parametro da determinare: a . La funzione è lineare
in a (la soluzione è semplice: $a = y/x^2$!!) ma applichiamo il metodo
del gradiente:*


Misuriamo un punto sulla parabola: $x = 1, y = 3$.
Vogliamo modificare a in modo che la parabola passi per $P(x,y)$.
La funzione costo da minimizzare sarà: $e = f(a | x,y) = (y - ax^2)^2$
La soluzione è $a = 3$

Partiamo da $a_{ini} = 2$.
 $err = (3 - 2 \cdot 1) \cdot 1 = 1$




Utilizziamo il metodo del gradiente:
Calcoliamo la derivata di $f(a | x,y) \rightarrow f'(a) = -2(y - ax^2)x^2$

A.A. 2016-2017 18/31 http://borghese.di.unimi.it/



Minimizzazione - underdamping



Consideriamo $\alpha = 1$
 Calcoliamo la derivata di $f(\cdot) \rightarrow f'(\cdot) = -2 (y - a x^2) x^2$


Utilizziamo il metodo del gradiente:

Passo 1:
 Calcoliamo l'incremento da dare al parametro a:
 $da = -[-2 (3 - 2 \cdot 1) \cdot 1] = -[-6 + 4] = 2 \quad a' = 2 + 2 = 4$


Passo 2:
 Calcoliamo l'incremento da dare al parametro a:
 $da = -[-2 (3 - 4 \cdot 1) \cdot 1] = -[-6 + 8] = -2 \quad a'' = 4 - 2 = 2$
 Oscillazioni!!!

Mi sposto troppo velocemente da una parte all'altra del minimo.

A.A. 2016-2017 19/31 http://borghese.di.unimi.it/



Minimizzazione -2 passi



Consideriamo $\alpha = 0.4$
 Calcoliamo la derivata di $f(\cdot) \rightarrow f'(\cdot) = -2 (y - a x^2) x^2$



Utilizziamo il metodo del gradiente:

Passo 1:
 Calcoliamo l'incremento da dare al parametro a:
 $da = -0.4 [-2 (3 - 2 \cdot 1) \cdot 1] = -[-6 + 4] = 0.8 \quad a' = 2 + 0.8 = 2.8$

Passo 2:
 Calcoliamo l'incremento da dare al parametro a:
 $da = -0.4 [-2 (3 - 2.8 \cdot 1) \cdot 1] = -[-6 + 5.6] = 0.16 \quad a'' = 2.8 + 0.16 = 2.96$
 Converge ad $a = 3$.

Posso correre il rischio di spostarmi troppo lentamente

A.A. 2016-2017 20/31 http://borghese.di.unimi.it/





Osservazioni

- Nel metodo del gradiente mi sposto lungo la tangente alla curva dell'errore per raggiungere il minimo.
- Lo spostamento lungo la tangente non mi porta al minimo direttamente.
- Se mi muovo velocemente lo supero.
- Se mi muovo lentamente arriva lentamente.

- Esistono algoritmi che:
 - ◆ Verificano che il singolo passo di gradiente porta a un miglioramento della soluzione.
 - ◆ Determinano il passo di apprendimento, α .

A.A. 2016-2017 21/31 http://borghese.di.unimi.it/




Sommario

Analisi dell'affidabilità della stima


Metodo del gradiente

Linearizzazione e metodo di Gauss-Newton

A.A. 2016-2017 22/31 http://borghese.di.unimi.it/



Linearizzazione



$y = f(x)$ viene linearizzata utilizzando il differenziale (retta tangente):

$$dy = f(x_0) + \left. \frac{df(x)}{dx} \right|_{x=x_0} dx = y_0 + \left. \frac{df(x)}{dx} \right|_{x=x_0} dx$$


Si può vedere come sviluppo di Taylor arrestato al 1° ordine
E' un'equazione lineare.

Per funzioni di più variabili, $f(\mathbf{P}; \mathbf{W}) = 0$, la linearizzazione nell'intorno di \mathbf{P} , si può scrivere come:


$$F(\mathbf{P}; \mathbf{W}) = F(\mathbf{P}_0; \mathbf{W}_0) + \sum_{j=1}^W \left. \frac{\partial F(\cdot)}{\partial w_j} \right|_{\mathbf{P}_0, \mathbf{W}_0} * dw_j = k - \sum_{j=1}^W a_j * dw_j$$

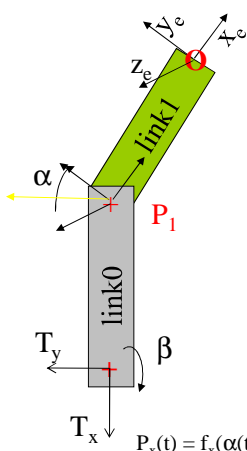
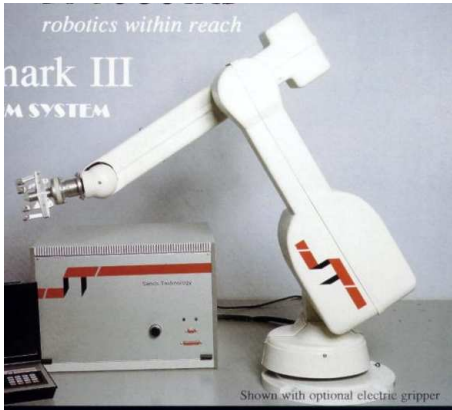
E' un'equazione lineare che descrive il comportamento della funzione $F(\cdot)$ nell'intorno del punto \mathbf{P}_0 con i parametri \mathbf{W}_0 .

A.A. 2016-2017
23/31
<http://borghese.di.unimi.it/>




Esempio di "sistema"







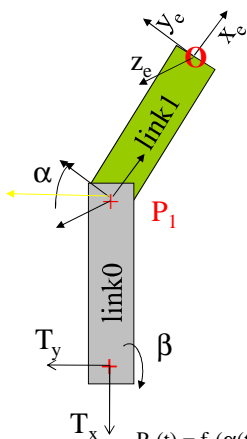
$P_x(t) = f_x(\alpha(t), \beta(t), T_x(t), T_y(t) | l_0, l_1).$
 $P_y(t) = f_y(\alpha(t), \beta(t), T_x(t), T_y(t) | l_0, l_1).$
 $P_z(t) = f_z(\alpha(t), \beta(t), T_x(t), T_y(t) | l_0, l_1).$

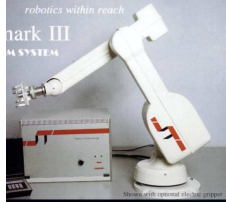
A.A. 2016-2017
24/31
<http://borghese.di.unimi.it/>



Esempio di "sistema"







Le funzioni legano la posizione dell'end point, uscita **P**, alla posizione degli angoli, α e β e della posizione della base, **T**, che rappresentano gli ingressi.

$$P_x(t) = f_x(\alpha(t), \beta(t), T_x(t), T_y(t) | l_0, l_1).$$


$$P_y(t) = f_y(\alpha(t), \beta(t), T_x(t), T_y(t) | l_0, l_1).$$

$$P_z(t) = f_z(\alpha(t), \beta(t), T_x(t), T_y(t) | l_0, l_1).$$


A.A. 2016-2017

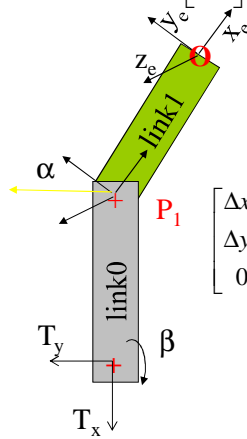
25/31

<http://borghese.di.unimi.it/>



Rappresentazione linearizzata Sistema lineare





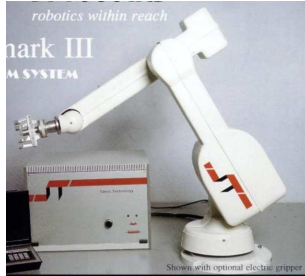
$$\begin{bmatrix} \Delta x_e \\ \Delta y_e \\ 0 \end{bmatrix} = \begin{bmatrix} -l_1 \sin(\alpha + \beta) & -l_1 \sin(\alpha + \beta) - l_0 \sin \beta & 1 & 0 \\ -l_1 \cos(\alpha + \beta) & -l_1 \cos(\alpha + \beta) - l_0 \cos \beta & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \Delta \alpha \\ \Delta \beta \\ \Delta T_x \\ \Delta T_y \end{bmatrix}$$

$$\alpha = 90 \quad l_0 = 2,5$$

$$\beta = 0 \quad l_1 = 2$$

$$\begin{bmatrix} \Delta x_e \\ \Delta y_e \\ 0 \end{bmatrix} = \begin{bmatrix} -2 & -2 & 1 & 0 \\ 0 & -2.5 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \Delta \alpha \\ \Delta \beta \\ \Delta T_x \\ \Delta T_y \end{bmatrix}$$


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



A.A. 2016-2017

26/31

<http://borghese.di.unimi.it/>



Minimizzazione di funzioni di più variabili


$\min(f(\mathbf{x}, \mathbf{w}))$ funzione costo od errore, \mathbf{w} vettore.

Modifico il valore dei pesi di una quantità proporzionale alla pendenza della funzione costo rispetto a quel parametro.
 La pendenza è una direzione nello spazio, non è più solamente destra / sinistra. Devo calcolare la derivata spaziale = **gradiente** della funzione costo, $f(\cdot)$.
 Estensione della tecnica del gradiente a più variabili.

$d\mathbf{w} = -\alpha \nabla f(\mathbf{x}; \mathbf{w})$, dato \mathbf{P}, \mathbf{W} .

Serve un' **approssimazione iniziale** per i pesi $\mathbf{W}_{ini} = \{w_j\}_{ini}$.

A.A. 2016-2017 27/31 http://borghese.di.unimi.it/



Metodo di Gauss-Newton

- L'idea:

Inizializzazione:



- Inizializzo i parametri ad un valore iniziale.

Iterazioni:

- 1) Linearizzazione delle equazioni.
- 2) Stima dell'aggiornamento dei parametri nel modello linearizzato ai minimi quadrati (soluzione ottimale, minimo del problema linearizzato).
- 3) Correzione dei parametri.

Può essere pesante perchè richiede l'inversione della matrice di covarianza. Spesso si preferiscono utilizzare metodi di ottimizzazione del primo ordine.

A.A. 2016-2017 28/31 http://borghese.di.unimi.it/

In pratica

$\mathbf{y} = f(\mathbf{x})$ \mathbf{x}, \mathbf{y} vettori di N ed M elementi rispettivamente



$\mathbf{y}_0 = f(\mathbf{x}_0)$ $\mathbf{x}_0, \mathbf{y}_0$ valore iniziale

Iterazione di (nella prima iterazione $k = 0$):

- $\mathbf{d}\mathbf{y}_k + \mathbf{y}_k = (\sum \delta f(\mathbf{x}) / \delta \mathbf{x})_{\mathbf{x}_k} \mathbf{d}\mathbf{x} + f(\mathbf{x}_k)$ $(\sum \delta f(\mathbf{x}) / \delta \mathbf{x})_{\mathbf{x}_k}$ are numbers!
- Si ottiene un sistema lineare
- Viene risolto come $\mathbf{d}\mathbf{x}_k = (\mathbf{A}\mathbf{A}^T)^{-1} \mathbf{A}^T \mathbf{d}\mathbf{y}_k$
- **Si aggiorna il valore di \mathbf{x} come $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{d}\mathbf{x}_k$**

Fino a convergenza

A.A. 2016-2017 29/31 http://borghese.di.unimi.it/

Evoluzione dei metodi del primo ordine

- α è un parametro critico. Se è troppo piccolo convergenza molto lenta, se è troppo grande overshooting.
- Ottimizzazione di α . Ad ogni passo viene calcolato α ottimale, per cui la funzione è decrescente (line search).

A.A. 2016-2017 30/31 http://borghese.di.unimi.it/



Sommario



Analisi dell'affidabilità della stima

Metodo del gradiente

Linearizzazione e metodo di Gauss-Newton