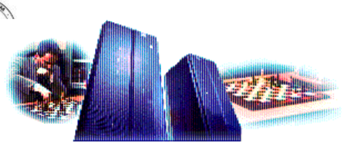
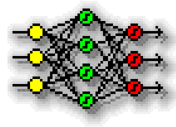
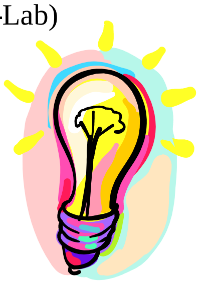




Sistemi Intelligenti Relazione tra ottimizzazione e statistica - IV

Alberto Borghese

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



Sommario

Analisi dell’affidabilità della stima

Metodo del gradiente

Linearizzazione e metodo di Gauss-Newton

Valutazione della bontà della stima

$x = (A'^*A)^{-1}A'^*\,b$ <==> $\min_{\mathbf{x}} \sum_k \nu_k^2 = \min_{\mathbf{x}} (\mathbf{Ax}-\mathbf{b})^2$

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

$<v_k> = 0$

$$\hat{\sigma}_0^2 = \sum_{k=1}^{M} \left(\nu_k^2\right) = |\nu|^2$$

Varianza della stima = varianza dell'errore di misura

.

Valutazione della bontà della stima del singolo parametro, x

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

$x = CA'^*\,b$

$$\hat{\sigma}_0^2 = \sum_{m=1}^{M} \left(\nu_m^2\right)$$

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.

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

$\Downarrow$

$\mathbf{x = CA'b}$ $\qquad \Delta x = C\,A'^*\,v$ $\qquad E[u] = 0$

C è la matrice di covarianza

Impostazione del calcolo della correlazione tra i parametri

$\Delta x = C\ A'\ v$

Abbiamo M parametri

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

$<\Delta x_i, \Delta x_j>$

$$\begin{bmatrix} \Delta x_I^{\ 2} & \Delta x_I \Delta x_{II} & \dots & \Delta x_I \Delta x_M \\ \Delta x_{II} \Delta x_I & \Delta x_{II}^{\ 2} & \dots & \Delta x_{II} \Delta x_M \\ \dots & \dots & \dots & \dots \\ \Delta x_M \Delta x_I & \Delta x_M \Delta x_{II} & \dots & \Delta x_M^{\ 2} \end{bmatrix}$$

$\Delta x = C\ A'\ v \quad => \quad \Delta x' = v' A\ (C)'$

$\Delta x\ \Delta x' = C\ A'\ vv' A\ C'$ => Applicando l'operatore di media, si ottiene:

$<\Delta x\ \Delta x'> = C\ A' <vv'> A\ C'$

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 $<vv'> = I\sigma_0^2$.

Incertezza sulla stima dei parametri

$<\Delta x\ \Delta x'> = CA'\ IA\ C'\sigma_0^2 = C'\ \sigma_0^2$

$\boxed{<\Delta x'\ \Delta x> = C\ \sigma_0^2}$

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

Incertezza su z -> incertezza sui parametri stimati, x

Visione geometrica (1 parametro, m)

$\sigma^2(\Delta x_{ij}) = c_{ij}\ \sigma_0^{\ 2}$ $\boxed{<\Delta x'\ \Delta x> = C\ \sigma_0^{\ 2}}$

u m = z +noise
A x = b + noise Determino la pendenza m della retta

Calcolo m e q ai minimi quadrati.

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

$C = (A'*A)^{-1}$ => $A_{1x1} = u$ => $C = (u * u)^{-1}$

La varianza di m varierà in modo inversamente proporzionale a u^2. Il rumore viene cioè moltiplicato per $1/u^2$. $\sigma^2(m) = c_m\ \sigma_0^{\ 2}$

Tanto più prendo i punti lontani dall'origine tanto meglio riesco a stimare m (tangente angolo).

Matrice di covarianza

Date N variabili casuali: $x = [x_1, x_2, \ldots x_N]$ 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:

$$C = \begin{bmatrix} \sigma_{x_1x_1} & \sigma_{x_1x_2} & . & \sigma_{x_1x_N} \\ \sigma_{x_2x_1} & \sigma_{x_2x_2} & . & \sigma_{x_2x_N} \\ . & . & . & . \\ \sigma_{x_Nx_1} & \sigma_{x_Nx_2} & . & \sigma_{x_Nx_N} \end{bmatrix}$$

Varianza: $\sigma_{x_1x_1} = \sigma^2{}_{x_1}$ N parametri

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

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:

$$r = \frac{M_{x_i x_j} - M_{x_i} M_{x_j}}{\sigma_{x_i} \sigma_{x_j}} \qquad -1 \le r \le +1$$

Definendo la covarianza tra x_i ed x_j come:

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

Dalla definizione di deviazione standard risulta:

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

Correlazione tra i parametri

$<uu'> = CA' \, IA \, C' \sigma_0^2 = C' \, \sigma_0^2$ $\qquad$ $\boxed{<u'u> = C \, \sigma_0^2}$

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

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

$$-1 \le r_{ij} = \frac{< u_i u_j >}{\sqrt{< u_i >^2 < u_j >^2}} = \frac{c_{ij}}{\sqrt{c_i c_j}} \le +1$$

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.

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)
```


Misura di correlazione su 2 parametri

Misura la inter-dipendenza tra 2 variabili statistiche:

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

```
>> 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)
```

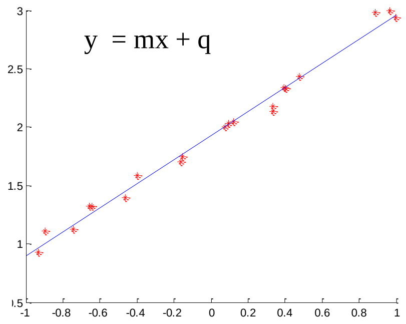
Caso 2D

N = 20 punti $\sigma_o^2 = 0.01$
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

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

Caso 2D – less points

N = 10 punti $\sigma_o^2 = 0.01$
m reale = 1   q reale = 2

C =

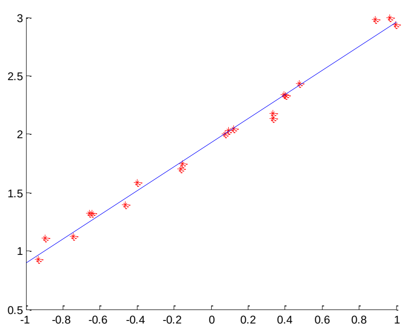
0.5927	-0.0030
-0.0030	0.1000

m_stimato =
1.0081
q_stimato =
1.9616

C =

0.2514	-0.0360
-0.0360	0.1051

m_stimato =
1.0012
q_stimato =
1.9107



y = mx + q

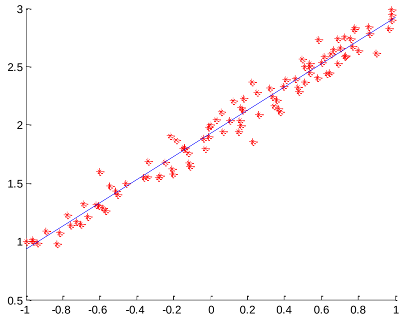
Diminuisce la confidenza nella stima

Caso 2D – more points

N = 100 punti $\sigma_o^2 = 0.01$
m reale = 1   q reale = 2

C =
0.0327 -0.0034
-0.0034 0.0103

m_stimato =
0.9942
q_stimato =
1.9978



y = mx + q

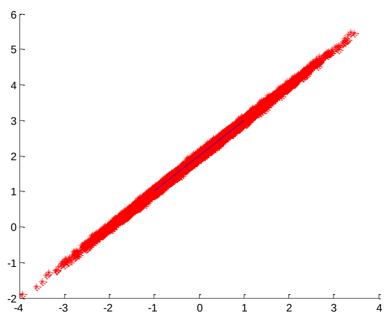
Aumenta la confidenza nella stima

Caso 2D – even more points

N = 10000 punti $\sigma_o^2 = 0.01$
m reale = 1   q reale = 2

C =
0.0000993 0.0000004
0.0000004 0.0001000

m_stimato =
1.000039
q_stimato =
2.0012



y = mx + q

Aumenta la confidenza nella stima

Sommario

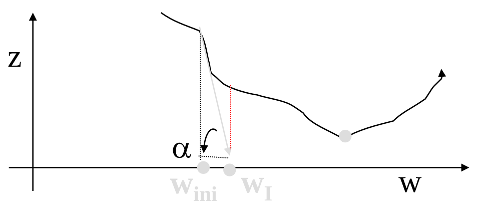
Analisi dell'affidabilità della stima

Metodo del gradiente

Linearizzazione e metodo di Gauss-Newton

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

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



La derivata, mi dà due informazioni:

1) In quale direzione di **w**, la funzione f(.) decresce.
2) Quanto rapidamente decresce.

Definisco uno spostamento arbitrario di w, maggiore la pendenza maggiore la variazione di z. Quindi faccio variare w nella direzione in cui diminuisce z:

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

Occorre un'inizializzazione.

Metodo iterativo.

Esempio di applicazione tecnica del gradiente per funzioni di 1 variabile

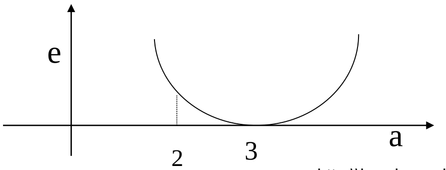
Supponiamo che il modello da noi considerato sia semplice: $z = a\ u^2$

*Abbiamo un unico parametro da determinare: **a**. La funzione f(a,u) è lineare in a (la soluzione è semplice: $a = z / u^2$!!) ma applichiamo il metodo del gradiente:*

Misuriamo un punto sulla parabola, P(u,z) = P(1, 3) La soluzione è $a = 3\ (3 / 1^2)$.

Applichiamo il metodo del gradiente. Partiamo da $a_{ini} = 2$. La parabola $z = 2\ u^2$ non passa per P. Come modificare a per farla passare per P?

La funzione costo da minimizzare sarà: $E = f(a \mid u,z) = (z - a\ u^2)^2$.
Dato il punto P, $E = (3 – 2 * 1^2)^2 = 1$



Minimizzazione - underdamping

Utilizziamo il metodo del gradiente:

Iterazione 1:
Passo 1:
Calcoliamo l'espressione della derivata di f(a | u,z) ➔ $f'(a) = 2\ (z – a\ u^2)\ (-u^2)$
Calcoliamo la derivata nel punto P(1,3), per $a_{ini} = 2$: $f'(a) = 2\ (3 – 2\ 1^2)\ (-1^2) = -2$
Passo 2:
Calcoliamo l'incremento da dare al parametro a:
$a_I = a_{ini} + \Delta a = a_{ini} - f'(a_{ini}; u,z) = 2 –(-2)$ $a_I = 2 + 2 = 4$

Iterazione 2:
Passo 1:
Calcoliamo l'espressione della derivata di f(a | u,z) ➔ $f'(a) = 2\ (z – a\ u^2)\ (-u^2)$
Calcoliamo la derivata nel punto P(1,3), per $a_I = 2$: $f'(a) = 2\ (3 – 4\ 1^2)\ (-1^2) = +2$
Passo 2:
Calcoliamo l'incremento da dare al parametro a:
$a_{II} = a_I + \Delta a = 4 –(+2)$ $a_{II} = 4 - 2 = 2$

La soluzione oscilla tra a= +2 e a = +4 senza trovare il minimo che è in a = 3.

Metodo del gradiente finale

Introduciamo un fattore di damping. **$\Delta w = -\alpha f'(w;u)$ dati P, w. $\alpha < 1$. La derivata f'(.) viene calcolata rispetto a w.** Scegliamo $\alpha = 0.4$.

Iterazione 1:

Passo 1:

Calcoliamo l'espressione della derivata di f(a | u,z) ➔ $f'(a) = 2\,(z - a\,u^2)\,(-u^2)$

Calcoliamo la derivata nel punto P(1,3), per $a_{ini} = 2$: $f'(a) = 2\,(3 - 2\,1^2)\,(-1^2) = -2$

Passo 2:

Calcoliamo l'incremento da dare al parametro a:

$a_I = a_{ini} + \Delta a = a_I = a_{ini} - \alpha\, f'(a_{ini};\, u,z) = 2 - 0.4\,[2\,(3 - 2\,1^2)\,(-1^2)\,] = 2 - 0.4*(-2) = 2.8$

Iterazione 2:

Passo 1:

Calcoliamo l'espressione della derivata di f(a | u,z) ➔ $f'(a) = 2\,(z - a\,u^2)\,(-u^2)$

Calcoliamo la derivata nel punto P(1,3), per $a_I = 2.8$: $f'(a) = 2\,(3 - 2.8\,1^2)\,(-1^2) = -0.4$

Passo 2:

Calcoliamo l'incremento da dare al parametro a:

$a_{II} = a_I + \Delta a = a_I - \alpha\, f'(a_I;\, u,z) = 2.8 - 0.4\,(-0.4) = 2. + 0.16 = 2.96$

E così via fino ad arrivare ad a = 3 che si raggiunge asintoticamente

Osservazioni

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

- Esistono algoritmi che:
  - Verificano che il singolo passo di gradiente porta a un miglioramento della soluzione.
  - Determinano il passo ottimale di apprendimento per ogni passo, $\alpha = \alpha_k$.

Sommario

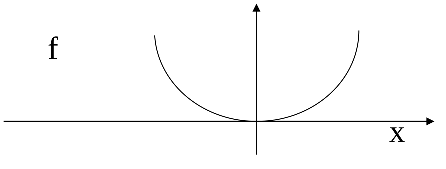
Analisi dell'affidabilità della stima

Metodo del gradiente

Linearizzazione e metodo di Gauss-Newton

Linearizzazione

- Descrizione differenziale locale di una funzione.



$y = ax^2$

Posso descrivere localmente in ogni punto la funzione come:

$y + dy = f(x) + f'(x)dx \Rightarrow dy = f'(x)\, dx \Rightarrow dy = 2ax\, dx$

Linearizzazione

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

$$dy = f(x_o) + \left.\frac{df(x)}{dx}\right|_{x=x_o} dx \qquad = y_o + \left.\frac{df(x)}{dx}\right|_{x=x_o} dx$$

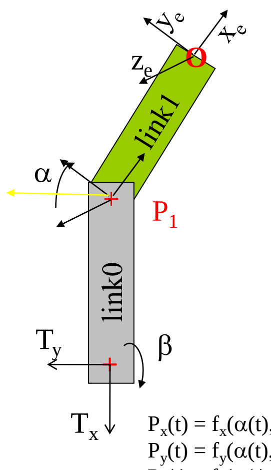
Si può vedere come sviluppo di Taylor arrestato al 1° ordine
E' un'equazione lineare.

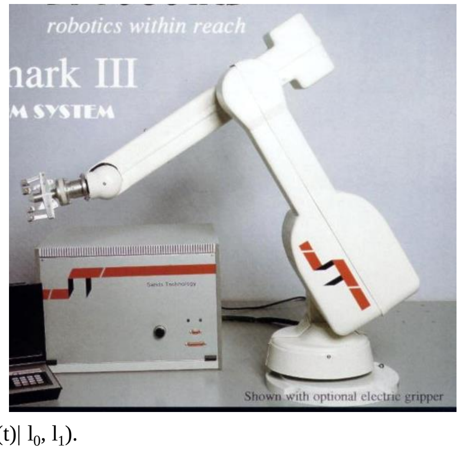
Per funzioni di più variabili, f(**P;W**) = 0, la linearizzazione nell'intorno di **P**, si può scrivere come:

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

E' un'equazione lineare che descrive il comportamento della funzione F(.) nell'intorno del punto **Po** con i parametri **Wo**.

Esempio di sistema

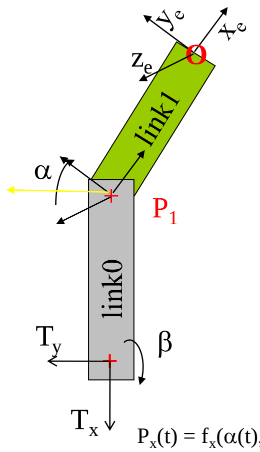




$P_x(t) = f_x(\alpha(t), \beta(t), T_x(t), T_y(t) | l_0, l_1)$.
$P_y(t) = f_y(\alpha(t), \beta(t), T_x(t), T_y(t) | l_0, l_1)$.
$P_z(t) = f_z(\alpha(t), \beta(t), T_x(t), T_y(t) | l_0, l_1)$

Voglio determinare α, β, Tx, Ty per ottenere un certo movimento dell'end-point.

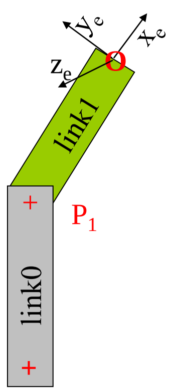
Esempio di "sistema"



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.

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

In forma matriciale



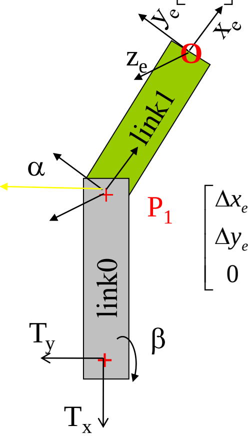
$$\begin{bmatrix} x_e \\ y_e \\ z_e \end{bmatrix} = \begin{bmatrix} l_1 \cos(\alpha+\beta) + l_0 \cos\beta + T_x \\ -l_1 \sin(\alpha+\beta) - l_0 \sin\beta + T_y \\ 0 \\ 1 \end{bmatrix}$$

Sono equazioni non-lineari nei parametri.

Non riesco a calcolare α, β, Tx, Ty per ottenere una certa Posizione dell'end-point

Rappresentazione linearizzata
Sistema lineare

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



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

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

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

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

Minimizzazione di funzioni di più variabili

min(f(**x**, **w**)) funzione costo od errore, **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(.).
Estensione della tecnica del gradiente a più variabili.

dw = - α ∇f(x;w), dato P, W.

Serve un'**approssimazione iniziale** per i parametri $W_{ini} = \{w_j\}_{ini}$.

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.

In pratica

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

$\mathbf{y_0} = f(\mathbf{x_0})$ x_0, y_0 valore iniziale

Iterazione di (nella prima iterazione k = 0):

- $\mathbf{dy_k} + \mathbf{y_k} = (\Sigma\delta f(\mathbf{x})/d\mathbf{x})_{xk}\, \mathbf{dx} + \mathbf{f(x_k)}$ $(\Sigma\delta f(\mathbf{x})/d\mathbf{x})_{xk}$ **are numbers!**
- Si ottiene un sistema lineare
- Viene risolto come $\mathbf{dx_k} = (AA^T)^{-1}A^T\mathbf{dy_k}$
- **Si aggiorna il valore di x come $\mathbf{x_{k+1} = x_k + dx_k}$**

Fino a convergenza

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 a ottimale, per cui la funzione è decrescente (line search).

Sommario

Analisi dell'affidabilità della stima

Metodo del gradiente

Linearizzazione e metodo di Gauss-Newton