GLI ALGORITMI DI OTTIMIZZAZIONE NON VINCOLATA PRESENTANO UNO SCHEMA GENERALE.

A PARTIRE DA UN PUNTO $x^0 \in \mathbb{R}^n$ SI VA A CALCOLARE UNO SPOSTAMENTO S_k FINO A QUANDO $\nabla f(x^k) \neq 0$ ANDANDO AD AGGIORNARE OGNI VOLTA IL VALORE DI x^k:

L'ALGORITMO:

```
Sia x^0 ∈ R^n e k=0
while (∇f(x^k) ≠ 0)
|  - calcolo lo spostamento S_k
|  - x^{k+1} = x^k + S_k
end
return x^k
```

QUESTO ALGORITMO PRODUCE UNA SEQUENZA DI SOLUZIONI $\{x^k\} \subset \mathbb{R}^n$ A CUI SONO ASSOCIATE LE SEQUENZE:

- $\{f(x^k)\} \subset \mathbb{R}$ → SEQUENZA DELLE IMMAGINI
- $\{\nabla f(x^k)\} \subset \mathbb{R}^n$ → SEQUENZA DEI GRADIENTI

$\{\nabla f(x^k)\}$ → SEQUENZA DEI GRADIENTI

SI POSSONO QUINDI AVERE DUE SITUAZIONI DIVERSE:

- $\boxed{\exists \bar{k} \text{ t.c. } \nabla f(x^{\bar{k}}) = 0}$ (CONVERGENZA FINITA)
- $\{x^k\}$ È UNA SEQUENZA INFINITA

VORREMMO QUINDI CHE LA SEQUENZA CONVERGESSE VERSO UN OTTIMO DEL PROBLEMA, OVVERO SI RICHIEDE:

1. ESISTONO DEI PUNTI DI ACCUMULAZIONE PER $\{x^k\}$, OVVERO

   $\exists k_1, k_2, \dots$ TALI CHE: $\boxed{\lim_{j \to \infty} x^{k_j} = \tilde{x}}$

   OVVERO CHE $\{x^k\}$ AMMETTA UNA SOTTOSUCCESSIONE CONVERGENTE

2. CHE $\{x^k\}$ SIA CONVERGENTE ASINTOTICAMENTE VERSO LA STAZIONARIETÀ

3. CHE $\{x^k\}$ CONVERGA RAPIDAMENTE

---

1 ESISTENZA DEI PUNTI DI ACCUMULAZIONE

SIA f CONTINUA E SIA $x^0 \in \mathbb{R}^n$ TALE CHE L'INSIEME DI LIVELLO

$$\boxed{\mathcal{L}_0 = \{x \in \mathbb{R}^n \mid f(x) \leq f(x^0)\}}$$

SIA COMPATTO. SUPPONIAMO DI GENERARE UNA SEQUENZA $\{x^k\}$ DEL TIPO:

$$\boxed{x^{k+1} = x^k + S_k}$$

TALE CHE:

→ LA FUNZIONE DECRESCE AD OGNI PASSO

$$\boxed{f(x^{k+1}) \leq f(x^k)}$$

ALLORA:

1) $\{x^k\}$ HA PUNTI DI ACCUMULAZIONE E OGNI PUNTO DI ACCUMULAZIONE $\bar{x}$ APPARTIENE ALL'INSIEME DI LIVELLO

2) LA SEQUENZA $\boxed{\{f(x^k)\}}$ CONVERGE AD UN VALORE $\bar{f}$ FINITO.

Dim:

1) PER LE IPOTESI SI HA CHE $f(x^{k+1}) \leq f(x^k)$. PER INDUZIONE SI HA CHE $f(x^{k+1}) \leq f(x^0)$ SUPPONENDO CHE $f(x^k) \leq f(x^0)$.

QUINDI $\boxed{x^{k+1} \in \mathcal{L}_0 \quad \forall k}$ PER LA DEFINIZIONE DI $\mathcal{L}_0$, E QUINDI $\boxed{\{x^k\} \subseteq \mathcal{L}_0}$ DOVE $\mathcal{L}_0$ È COMPATTO PER IPOTESI. ALLORA $\{x^k\}$ HA PUNTI DI ACCUMULAZIONE, OGNUNO DEI QUALI APPARTIENE ALL'

INSIEME DI LIVELLO $\mathcal{L}_0$ PER DEFINIZIONE! $\square$

2) ESSENDO $f(x^{k+1}) \leq f(x^k)$ ALLORA $\{f(x^k)\}$ È MONOTONA NON CRESCENTE, QUINDI ESISTE IL LIMITE:

$$\boxed{\lim_{k \to \infty} f(x^k) = \bar{f}}$$

POSSIAMO INOLTRE CONSIDERARE UNA SOTTOSEQUENZA $K \subseteq \{0, 1, \ldots\}$ TALE CHE:

$$\boxed{\lim_{\substack{k \to \infty \\ k \in K}} x^k = \bar{x}}$$

DOVE K ESISTE PER IL PUNTO 1. VALE ALLORA CHE, PER LA CONTINUITÀ DI f:

$$\boxed{\lim_{\substack{k \to \infty \\ k \in K}} f(x^k) = f(\bar{x}) = \bar{f}}$$

$\square$

---

2 CONVERGENZA VERSO LA STAZIONARIETÀ

SI HANNO 3 DIVERSI CASI DI CONVERGENZA:

- $\boxed{\lim_{k \to \infty} x^k = \bar{x}}$ → CONDIZIONE MIGLIORE: SI HA PIENA CONVERGENZA VERSO

$$\lim_{n\to\infty}$$

UN PUNTO STAZIONARIO

- $$\boxed{\lim_{n\to\infty} \|\nabla f(x^n)\| = 0}$$ → IL GRADIENTE TENDE A ZERO, OVVERO TUTTI I PUNTI DI ACCUMULAZIONE SONO STAZIONARI

- $$\boxed{\lim_{n\to\infty} \inf \|\nabla f(x^n)\| = 0}$$ → ALMENO UN PUNTO DI ACCUMULAZIONE È STAZIONARIO
  ↳ CONDIZIONE PIÙ DEBOLE MA LA MINIMA RICHIESTA!

  → INDICA IL VALORE MINIMO A CUI TENDONO I VALORI DEL GRADIENTE CALCOLATI PER CIASCUN PUNTO x^n
  ↳ ALMENO UN VALORE DELLA SUCCESSIONE TENDE AD UN PUNTO STAZIONARIO

ESEMPIO:

SIA $f'(x) = (x-1)(x-3)$, DEFINIAMO:

- $x^n = \{0,9;\ 0,99;\ 0,999;\ 0,9999, \dots\}$ → IN QUESTO CASO SI HA CHE TUTTA LA SEQUENZA CONVERGE AD 1, OVVERO UN PUNTO STAZIONARIO → CASO 1

- $x^n = \{0,9;\ 2,9;\ 0,99;\ 2,99 \dots\}$ → IN QUESTO CASO UNA SOTTOSEQUENZA CONVERGE AD 1 E L'ALTRA A 3 ⇒ $\nabla f(x^n) \to 0$ → CASO 2

- $x^n = \{1;\ 6;\ 1;\ 6;\ 1;\ 6 \dots\}$ → SI HANNO DUE SOTTOSEQUENZE CONVERGENTI DOVE SOLTANTO UNA CONVERGE AD UN PUNTO DI ACCUMULAZIONE → CASO 3

---

3 VELOCITÀ DI CONVERGENZA:

SI DICE CHE UNA SEQUENZA $\{x^n\}$ CONVERGENTE AD x^* HA TASSO DI CONVERGENZA:

- **SUBLINEARE**: $$\lim_{k\to\infty} \frac{\|x^{k+1}-x^*\|}{\|x^k - x^*\|} = 1$$ → PIÙ SI PROCEDE CON L'ITERAZIONE E MENO PROGRESSO SI FA, OVVERO SI TENDE SEMPRE PIÙ A RALLENTARE

- **LINEARE** $$\lim_{k\to\infty} \frac{\|x^{k+1}-x^*\|}{\|x^k - x^*\|} = c\in(0,1)$$ → AD OGNI PASSO DELL'ITERAZIONE CI SI AVVICINA ALLA SOLUZIONE CON RATE COSTANTE

- **SUPERLINEARE** $$\lim_{k\to\infty} \frac{\|x^{k+1}-x^*\|}{\|x^k - x^*\|} = 0$$ → SI VA SEMPRE PIÙ VELOCI A CIASCUNA ITERAZIONE, È L'OPPOSTO DEL SUBLINEARE

- **QUADRATICO** $$\lim_{k\to\infty} \frac{\|x^{k+1}-x^*\|}{\|x^k - x^*\|^2} = c<1$$ → È IL CASO A VELOCITÀ MAGGIORE!

CLASSIFICAZIONE DI ALGORITMI DI OTTIMIZZAZIONE

LA PRIMA SUDDIVISIONE SI BASA SULLE INFORMAZIONI RELATIVE ALLA DISPONIBILITÀ DI $f, \nabla f, \nabla^2 f$

- **ALGORITMI DI PRIMO ORDINE**: SI CALCOLA $f(x), \nabla f(x)$
- **ALGORITMI DI SECONDO ORDINE**: SI CALCOLA $f(x), \nabla f(x), \nabla^2 f(x)$
- **ALGORITMI DI ORDINE ZERO**: SI CALCOLA $f(x)$

GLI ALGORITMI CONVERGENTI SI DIVIDONO IN:

- **GLOBALMENTE CONVERGENTI**: SE $\forall x^0 \in \mathbb{R}$ SI HA CHE $\{x^k\}$ CONVERGE A $\bar{x}$ CON $\nabla f(\bar{x}) = 0$

- **LOCALMENTE CONVERGENTI**: SE $\{x^k\}$ CONVERGE A $\bar{x}$ CON $\nabla f(\bar{x}) = 0$ IN UN INTORNO $B_\rho(\bar{x})$

INFINE SI DISTINGUONO 3 TIPOLOGIE DI ALGORITMO:

1. TIPO LINESEARCH
2. TIPO TRUST-REGION
3. TIPO DIRECT-SEARCH

ALGORITMI LINE-SEARCH

DATI $x^0 \in \mathbb{R}^n$ E $k=0$ IL GENERICO ALGORITMO È:

WHILE ($\nabla f(x^k) \neq 0$):

- SI SCEGLIE UNA DIREZIONE d_k
- SI CALCOLA UN PASSO $\alpha_k > 0$

• SI PONE $\boxed{x^{k+1} = x^k + \alpha_k d_k}$

• $k = k+1$

END WHILE

DEFINIZIONE FUNZIONE DI FORZAMENTO

SIA $\sigma: \mathbb{R}^+ \to \mathbb{R}^+$, σ È UNA FUNZIONE DI FORZAMENTO

SE $\forall \{t^k\}$ TALE CHE:

$$\boxed{\lim_{k \to \infty} \sigma(t^k) = 0}$$

ALLORA VALE CHE:

$$\Downarrow$$

$$\boxed{\lim_{k \to \infty} t^k = 0}$$

OVVERO, SE SI HA CHE $\sigma(t_k) \to 0$, ALLORA SIAMO SICURI CHE ANCHE L'ARGOMENTO $t_k \to 0$.

AD ESEMPIO $\boxed{\sigma(t) = ct}$ CON $c > 0$ È TALE CHE SE

$$\lim_{k \to \infty} \sigma(t^k) = \lim_{k \to \infty} c t^k = 0$$

ALLORA $t_k \to 0$ PER $k \to \infty$

SIA $f \in C^1(\mathbb{R}^n)$ E SIA $x^0 \in \mathbb{R}^n$ t.c.: $\boxed{\mathcal{L}_0 \text{ SIA COMPATTO}}$.
SIA $\{x^k\}$ UNA SEQUENZA TALE CHE $\boxed{\nabla f(x^k) \neq 0 \quad \forall k}$; SI ASSUMANO LE SEGUENTI CONDIZIONI:

1. $\boxed{f(x^{k+1}) \leq f(x^k)}$ $\to$ $\{x^k\}$ MONOTONA NON CRESCENTE

2. $$\boxed{\lim_{k \to \infty} \frac{\nabla f(x^k)^T d_k}{\|d_k\|} = 0}$$

3. $$\boxed{\frac{\nabla f(x^k)^T d_k}{\|d_k\|} \geq \sigma\left(\|\nabla f(x^k)\|\right) \text{ CON } \sigma \text{ DI FORZAMENTO}}$$

ALLORA, SE QUESTE TRE CONDIZIONI RISULTANO VERIFICATE, SI HA:

a. $\boxed{\{x^k\} \subseteq \mathcal{L}_f(x^k)}$

b. $\{x^k\}$ HA TUTTI I PUNTI DI ACCUMULAZIONE APPARTENENTI A $\mathcal{L}_0$

c. $\{f(x^k)\}$ È CONVERGENTE AD UN VALORE FINITO $\bar{f}$

[d. $\boxed{\lim_{k \to \infty} \|\nabla f(x^k)\| = 0}$

[e. CIASCUN PUNTO DI ACCUMULAZIONE DI $\{x^k\}$ È STAZIONARIO

Dim:

I PUNTI a, b, c SEGUONO DALLA CONTINUITÀ DI f, DALLA COMPATTEZZA DI $\mathcal{L}_0$ E DALLA MONOTONIA DI $\{f(x^k)\}$

d: DALLE IPOTESI SAPPIAMO CHE:

$$\frac{|\nabla f(x^k)^T d_k|}{\|d_k\|} \geq \sigma(\|\nabla f(x^k)\|) \geq 0$$

SE SI CONSIDERA IL LIMITE PER $k \to \infty$ SI HA CHE IL MEMBRO DI SINISTRA TENDE A ZERO, ALLORA:

$$\lim_{k \to \infty} \sigma(\|\nabla f(x^k)\|) = 0$$

MA ESSENDO σ DI FORZAMENTO SI HA CHE:

$$\lim_{k \to \infty} \|\nabla f(x^k)\| = 0 \qquad \square$$

(e): CONSIDERO UNA SOTTOSEQUENZA $K \subseteq \{0, 1, \dots\}$ TALE CHE

$$\boxed{\lim_{\substack{k \to \infty \\ k \in K}} x^k = \bar{x}}$$

ALLORA:

$$0 = \underbrace{\lim_{k \to \infty}}_{\text{LIMITE SU TUTTA LA SEQUENZA}} \|\nabla f(x^k)\| \overset{d}{=} \underbrace{\lim_{\substack{k \to \infty \\ k \in K}}}_{\text{LIMITE SULLA SOTTOSEQUENZA}} \|\nabla f(x^k)\| \overset{f \in C^1(\mathbb{R}^n)}{=} \|\nabla f(\bar{x})\|$$

$\Rightarrow \|\nabla f(\bar{x})\| = 0 \Rightarrow$ UN QUALSIASI PUNTO DI ACCUMULAZIONE CONSIDERATO È ANCHE UN PUNTO STAZIONARIO $\square$

OSSERVAZIONE

POSSIAMO RISCRIVERE LA CONDIZIONE (2) NEL SEGUENTE MODO:

$$\boxed{\nabla f(x^k)^T \cdot \frac{d_k}{\|d_k\|} \to 0}$$

E TRAMITE LA NORMALIZZAZIONE DELLA DIREZIONE d_k SI VA A RENDERE TALE LIMITE INDIPENDENTE DAL VALORE DI d_k

PERTANTO DEVE ESSERE LA DERIVATA DIREZIONALE $\nabla f(x^k)^T d_k$ A TENDERE A ZERO E NON LA LUNGHEZZA DELLA DIREZIONE d_k!

INOLTRE, NON DEVE VERIFICARSI CHE, IN MANIERA ASINTOTICA, d_k RISULTI PERPENDICOLARE AL GRADIENTE:

SIA $\boxed{\sigma(t) = ct}$ CON $c > 0$. SAPPIAMO CHE, DALLA 3:

$$\boxed{\frac{|\nabla f(x^k)^T d_k|}{\|d_k\|} \geq c\|\nabla f(x^k)\|}$$

SE $\nabla f(x^k)^T d_k < 0$, ALLORA: → d_k DIREZIONE DI DISCESA

$$\boxed{\frac{\nabla f(x^k)^T d_k}{\|d_k\| \cdot \|\nabla f(x^k)\|} \leq -c < 0}$$

E POSSIAMO SCRIVERE CHE:

IL COS DELL'ANGOLO TRA d_k E $\nabla f(x^k)$ > 90° (?)

$$\frac{\nabla f(x^k)^T d_k}{\ldots} = \cos\left(\theta(d_k, \nabla f(x^k))\right)$$

$$\|d_k\| \cdot \|\nabla f(x^k)\|$$

↳ CONDIZIONE D'ANGOLO

SE PRENDIAMO, COME DIREZIONE, L'ANTIGRADIENTE?

SIA $\boxed{d_k = -\nabla f(x^k)}$

$$\Rightarrow \frac{\nabla f(x^k)^T(-\nabla f(x^k))}{\|\nabla f(x^k)\| \cdot \|-\nabla f(x^k)\|} = \frac{-\|\nabla f(x^k)\|^2}{\|\nabla f(x^k)\|^2} = \boxed{-1}$$

↳ QUINDI L'ANGOLO TRA d_k E $\nabla f(x^k)$ È MAGGIORE DI 90°

⇒ CONDIZIONE SODDISFATTA ✓

SE INVECE SCEGLIAMO $\boxed{d_k = -H_k \nabla f(x^k)}$ CON H_k MATRICE

D.P TALE CHE:

- $\lambda_{min}(H_k) \geq m$
- $\lambda_{max}(H_k) \leq M$

SI OTTIENE CHE:

$$\boxed{\nabla f(x^k)^T d_k = -\nabla f(x^k)^T H_k \nabla f(x^k) < 0}$$

SE $\nabla f(x^k) \neq 0$ OTTENIAMO:

- $|\nabla f(x^k)^T d_k| = |-\nabla f(x^k)^T H_k \nabla f(x^k)| \geq m\|\nabla f(x^k)\|^2$

- $\|d_k\| = \|H_k \nabla f(x^k)\| \leq M\|\nabla f(x^k)\|$

QUINDI SI AVRÀ:

$$\boxed{\frac{|\nabla f(x^k)^T d_k|}{\|d_k\|} \geq \frac{m}{M}\|\nabla f(x^k)\|}$$

E QUINDI SCEGLIENDO $c = \frac{m}{M}$ SI HA CHE LA SCELTA DI d_k VA A SODDISFARE LA CONDIZIONE D'ANGOLO!

LINE-SEARCH: SCELTA DEL PASSO α_k

VEDIAMO ADESSO COME È POSSIBILE SCEGLIERE IL PASSO α_k PER SPOSTARSI LUNGO LA DIREZIONE DI DISCESA d_k.

L'OBIETTIVO È QUINDI QUELLO DI SCEGLIERE α_k DA USARE IN

$$\boxed{x^{k+1} = x^k + \alpha_k d_k}$$

CERCANDO DI MINIMIZZARE $\boxed{f(x^k + \alpha_k d_k)}$.

SI DISTINGUONO DUE SITUAZIONI:

- RICERCA DI LINEA ESATTA: $\boxed{\alpha_k = \arg\min_{\alpha} f(x^k + \alpha d_k)}$

- RICERCA DI LINEA INESATTA: $\boxed{\alpha_k \approx \arg\min_{\alpha} f(x^k + \alpha d_k)}$

RICERCA DI LINEA ESATTA

SE CONSIDERIAMO $\boxed{\ell(\alpha) = f(x^k + \alpha d_k)}$ OTTENIAMO:

- $\boxed{\ell'(\alpha) = \nabla f(x^k + \alpha d_k)^T d_k}$

- $\boxed{\ell(0) = f(x^k)}$

- $\boxed{\ell'(0) = \nabla f(x^k)^T d_k}$

VOGLIAMO FARE IN MODO CHE $\boxed{\alpha_k = \arg\min_{\alpha} \ell(\alpha)}$

OVVERO CERCARE α TALE CHE $\boxed{\ell'(\alpha) = 0}$

UN ESEMPIO DI LINE-SEARCH ESATTA SI HA NEL CASO DEL CLASSICO PROBLEMA QUADRATICO:

$$\min_{x \in \mathbb{R}^n} \frac{1}{2} x^T Q x + c^T x \quad \text{con } Q \text{ D.P}$$

IMPLICA LA STRETTA CONVESSITÀ!

SUPPONIAMO QUINDI DI ESSERE IN UN PUNTO x^k E DI VOLER EFFETTUARE UN CERTO SPOSTAMENTO LUNGO UNA DIREZIONE d_k:

CALCOLIAMO LO SPOSTAMENTO

APPROSSIMAZIONE SECONDO TAYLOR

SAREBBE $\nabla^2 f(\xi)$ MA SAPPIAMO CHE NEL CASO QUADRATICO $\nabla^2 f(\xi) = Q$ cost

$$\varphi(\alpha) = f(x^k + \alpha d_k) = f(x^k) + \alpha \nabla f(x^k)^T d_k + \frac{1}{2} \alpha^2 d_k^T Q d_k$$

NOTIAMO CHE $\varphi(\alpha)$ È UNA FUNZIONE QUADRATICA RISPETTO AD α ED È QUINDI RICONDUCIBILE AD UNA PARABOLA. QUESTO SIGNIFICA CHE PRESENTERÀ UN UNICO MINIMO LOCALE = GLOBALE

PER CALCOLARE QUINDI IL PUNTO DI MINIMO SI CALCOLA $\varphi'(\alpha)$ E SI UGUAGLIA A ZERO:

$$\varphi'(\alpha) = \nabla f(x^k)^T d_k + \alpha d_k^T Q d_k \Rightarrow \boxed{\alpha^* = \frac{-\nabla f(x^k)^T d_k}{d_k^T Q d_k}}$$

QUINDI PER UNA FUNZIONE QUADRATICA È POSSIBILE ESEGUIRE UNA LINE SEARCH ESATTA POICHÈ SIAMO IN GRADO DI CALCOLARE IL PASSO OTTIMO α^*

RICERCA DI LINEA INESATTA

IN QUESTO CASO L'OBIETTIVO È RICERCARE UN PASSO CHE VERIFICHI LE SEGUENTI PROPRIETÀ:

- $\boxed{f(x^k + \alpha_k d_k) \le f(x^k)}$ OPPURE $\boxed{\varphi(\alpha_k) \le \varphi(0)}$

- $f(x^k + \alpha_k d_k) \leq f(x^k) - \underbrace{\varepsilon_k(\alpha_k)}$ → SUFFICIENTE DECREMENTO POSITIVO

$\Downarrow$

- CONDIZIONE DI ARMIJO

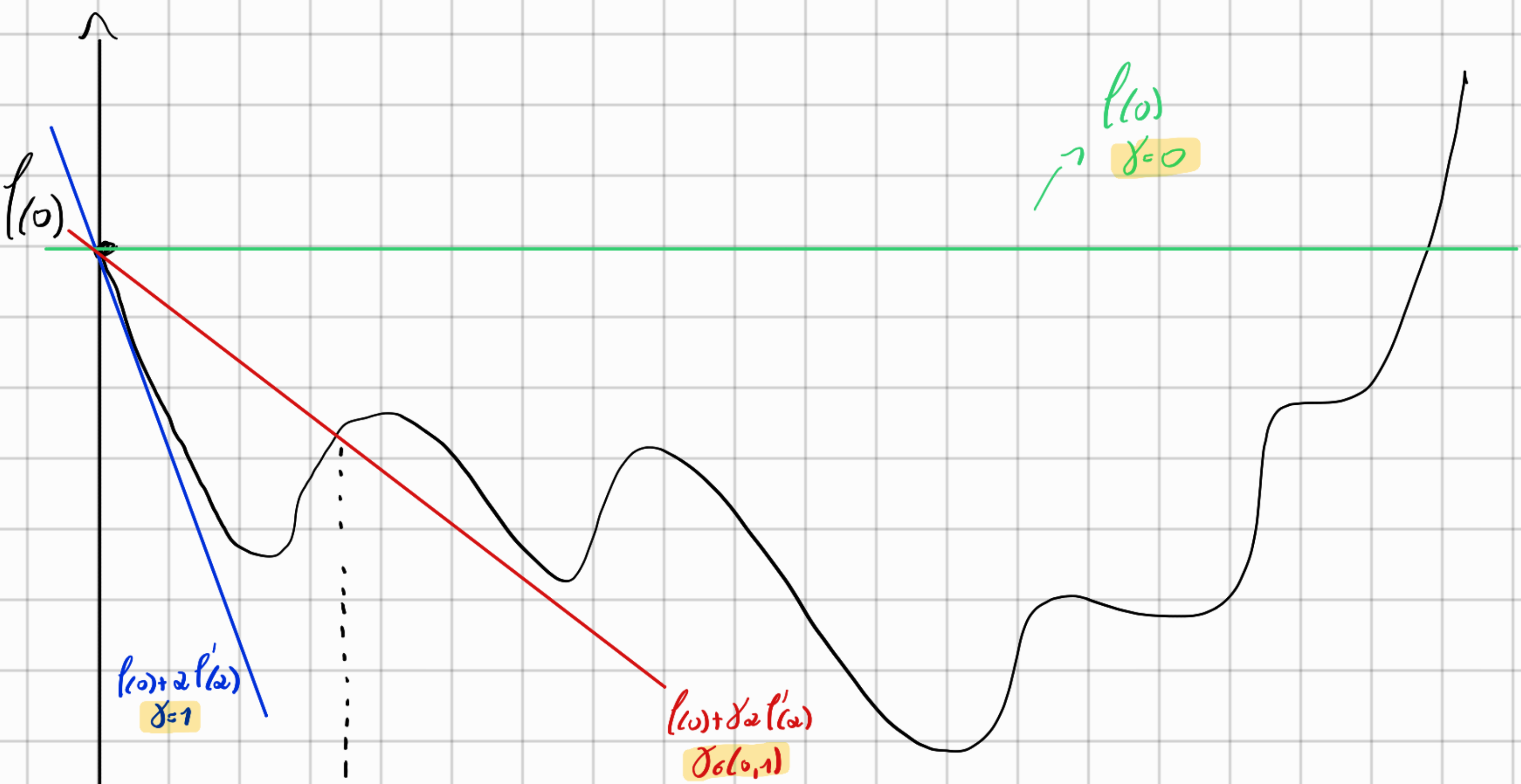
DATO $\gamma \in (0,1)$ E d_k TC: $\nabla f(x^k)^T d_k < 0$ → d_k DIREZIONE DI DISCESA, SI HA CHE:

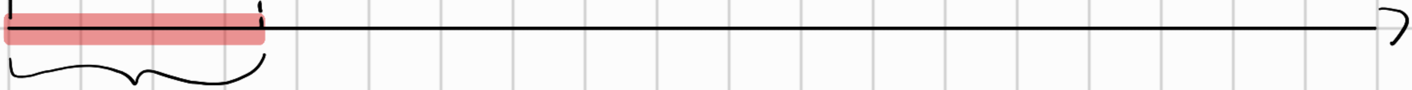
$$\boxed{f(x^k + \alpha_k d_k) \leq f(x^k) + \gamma \alpha_k \nabla f(x^k)^T d_k}$$

POSSIAMO ANCHE RISCRIVERLO COME:

$$\boxed{\varphi(\alpha^k) \leq \varphi(0) + \gamma \alpha_k \varphi'(\alpha)}$$

→ STABILISCE CHE IL PASSO DEVE ESSERE SUFFICIENTEMENTE PICCOLO IN MODO CHE LA FUNZIONE OBIETTIVO DIMINUISCA IN MODO SIGNIFICATIVO MA NON NECESSARIAMENTE FINO AL SUO VALORE MINIMO!





QUESTO È L'INTERVALLO DEI VALORI CHE SODDISFANO LA CONDIZIONE DI ARMIJO

VEDIAMO L'ALGORITMO CHE DESCRIVE IL METODO DI ARMIJO

DATI:
- x^k
- d_k t.c. $\nabla f(x^k)^T d_k < 0$
- $\gamma \in (0,1)$
- $\delta \in (0,1)$
- $\alpha_0 > 0$

OVVERO FINO A QUANDO NON È SODDISFATTA LA CONDIZIONE DI ARMIJO

SIA $t=0$

WHILE $\left(f(x^k + \alpha_t d_k) > f(x^k) + \gamma \alpha_t \nabla f(x^k)^T d_k \right)$

$\alpha_{t+1} = \delta \alpha_t$

$t = t+1$

END WHILE

TERMINAZIONE FINITA DEL METODO DI ARMIJO

SUPPONIAMO CHE $\boxed{\nabla f(x^k)^T d_k < 0}$, $\alpha_0 > 0$, $\underbrace{\delta \in (0,1)}_{\text{SERVE PER DIMINUIRE IL PASSO}}$, $\underbrace{\gamma \in (0,1)}_{\text{PROPRIO DELLA CONDIZIONE ARMIJO}}$, ALLORA:

1. L'ALGORITMO DI ARMIJO PRODUCE, IN UN NUMERO FINITO DI PASSI, UN CERTO VALORE α_k CHE SODDISFA LA CONDIZIONE DI ARMIJO.

② VALE UNA TRA:

CI SI FERMA AL PASSO INIZIALE

- a) $\alpha_K = \alpha_0$
- b) $\alpha_K \leq \gamma\alpha_0$ E $f\left(x^K + \frac{\alpha_K}{\gamma} d_K\right) > f(x^K) + \frac{\alpha_K}{\gamma} \nabla f(x^K)^T d_K$

ABBIAMO FATTO ALMENO UN PASSO

Dim:

① SUPPONIAMO PER ASSURDO CHE LA TESI CHE VOGLIAMO DIMOSTRARE SIA FALSA ALLORA SI HA CHE $\forall t = 0, 1, \dots$ E RICORDANDO CHE $\boxed{\alpha_t = \gamma^t \alpha_0}$:

$$f(x^K + \overbrace{\gamma^t \alpha_0}^{\alpha_t} d_K) > f(x^K) + \overbrace{\gamma\gamma^t \alpha_0}^{\gamma\alpha_t} \nabla f(x^K)^T d_K$$

$$\Rightarrow f(x^K + \gamma^t \alpha_0 d_K) - f(x^K) > \gamma\gamma^t \alpha_0 \nabla f(x^K)^T d_K$$

SI DIVIDE PER $\gamma^t \alpha_0$

$$\Rightarrow \frac{f(x^K + \gamma^t \alpha_0 d_K) - f(x^K)}{\gamma^t \alpha_0} > \gamma \nabla f(x^K)^T d_K$$

PER $t \to \infty$ TENDE ALLA DERIVATA DIREZIONALE $\nabla f(x^K)^T d_K$ POICHÉ $\gamma^t \to 0$ PERCHÉ $\gamma \in (0,1)$

$$\Rightarrow \nabla f(x^K)^T d_K \geq \gamma \nabla f(x^K)^T d_K \to \underbrace{(1-\gamma)}_{<0} \underbrace{\nabla f(x^K)^T d_K}_{<0} \geq 0$$

PERCHÉ, PER ASSURDO, NON SODDISFA ARMIJO! $\Leftarrow$

$\Downarrow$

ASSURDO! $\square$

---

(2)

PER COME È STRUTTURATO L'ALGORITMO SI AVRÀ CHE O SI ACCETTA SUBITO IL PASSO INIZIALE (VALE (a)) OPPURE È AVVENUTA UNA RIDUZIONE DEL PASSO, OVVERO $\alpha_k \le \delta\alpha_0$

INOLTRE, ALLA PENULTIMA ITERAZIONE IL CONTROLLO DELLA CONDIZIONE DI ARMIJO NON VIENE SUPERATO POICHÉ È L'ULTIMO PASSO IN CUI PUÒ EFFETTIVAMENTE AVVENIRE LA RIDUZIONE, QUINDI:

$$\boxed{f\left(x^k+\frac{\alpha_k}{\delta}d_k\right) > f(x^k)+\gamma\frac{\alpha_k}{\delta}\nabla f(x^k)^T d_k}$$

$\square$

CONVERGENZA DEI METODI LINE-SEARCH CON RICERCA DI ARMIJO

SIA $f \in C^1(\mathbb{R}^n)$, $x^0 \in \mathbb{R}^n$ t.c. $\boxed{\mathcal{L}_0 \text{ È COMPATTO}}$ E SIA $\{x^k\}$ LA SEQUENZA PRODOTTA DALL'ALGORITMO, TALE CHE:

$$\boxed{\nabla f(x^k) d_k < 0 \quad \forall k}$$

SE

$$\boxed{\alpha_0(k) \geq \frac{1}{\|d_k\|}\,\sigma\left(\frac{|\nabla f(x^k)^T d_k|}{\|d_k\|}\right)}$$

$\hookrightarrow$ IN PRATICA SIGNIFICA CHE IL PASSO INIZIALE DEVE ESSERE SIGNIFICATIVO

CON σ DI FORZAMENTO, ALLORA:

a. $f(x^{k+1}) < f(x^k)$ → LA SEQUENZA PRODOTTA È MONOTONA NON CRESCENTE

b. $\lim_{k\to\infty} \dfrac{|\nabla f(x^k)^T d_k|}{\|d_k\|} = 0$ → SI HA CONVERGENZA A PUNTI STAZIONARI

Dim:

(a) PER LE PROPRIETÀ DI ARMIJO VALE CHE:

$$f(x^{k+1}) = f(x^k + \alpha_k d_k) \le f(x^k) + \underbrace{\underbrace{\gamma\alpha_k}_{>0}\underbrace{\nabla f(x^k)^T d_k}_{<0}}_{<0}$$

$$< f(x^k) \quad \square$$

ARMIJO CAMBIATO DI SEGNO:

(b) CONSIDERIAMO $\boxed{f(x^k) - f(x^{k+1}) \geq -\gamma\alpha_k \nabla f(x^k)^T d_k}$

DALLA MANIPOLAZIONE DELLA FORMULA DI ARMIJO.

$$\Rightarrow f(x^k) - f(x^{k+1}) \geq -\gamma\alpha_k \overbrace{\nabla f(x^k)^T d_k}^{<0}$$

POICHÉ $\nabla f(x^k)^T d_k$ È NEGATIVA E QUINDI TRAMITE $|\cdot|$ POSSO SEMPLIFICARE IL MENO

$$= \gamma\alpha_k |\nabla f(x^k)^T d_k|$$

$$= \gamma\alpha_k |\nabla f(x^k)^T d_k| \cdot \frac{\|d_k\|}{\|d_k\|}$$

$$= \gamma\alpha_k \|d_k\| \cdot \frac{|\nabla f(x^k)^T d_k|}{\|d_k\|} \geq 0$$

QUINDI OTTENIAMO CHE $\boxed{f(x^k) - f(x^{k+1}) \geq \gamma\alpha_k \|d_k\| \dfrac{|\nabla f(x^k)^T d_k|}{\|d_k\|} \geq 0}$

$\|d_k\|$

SE CONSIDERIAMO LA SEQUENZA $\{f(x^k)\}$ SI HA CHE QUESTA È DECRESCENTE PER IL PUNTO (a).

SAPPIAMO INOLTRE CHE $\mathcal{L}_0$ È COMPATTO PER Hp, QUINDI $\{f(x^k)\}$ AMMETTE LIMITE FINITO.

SE CONSIDERIAMO $\underbrace{\lim_{k\to\infty} f(x^k)-f(x^{k+1})}_{\bar{f}-\bar{f}} \geq \lim_{k\to\infty} \gamma\alpha_k\|d_k\|\frac{|\nabla f(x^k)^T d_k|}{\|d_k\|}$ SI HA CHE:

$$\bar{f}-\bar{f} \geq \lim_{k\to\infty} \gamma\alpha_k\|d_k\|\frac{|\nabla f(x^k)^T d_k|}{\|d_k\|} \geq 0$$

$\Rightarrow$ PER IL TEOREMA DEI CARABINIERI SI HA QUINDI CHE

$$\boxed{\lim_{k\to\infty} \gamma\alpha_k\|d_k\| \cdot \frac{|\nabla f(x^k)^T d_k|}{\|d_k\|} = 0}$$

SE CONSIDERIAMO LA SEQUENZA $\boxed{\left\{\frac{|\nabla f(x^k)^T d_k|}{\|d_k\|}\right\} = \left\{\nabla f(x^k)^T \frac{d_k}{\|d_k\|}\right\}}$

SI HA CHE:

- $\{x^k\} \subseteq \mathcal{L}_0$ È LIMITATA
- $\nabla f(x^k)$ È CONTINUA

$\Rightarrow$ $\boxed{\{|\nabla f(x^k)|\} \text{ È LIMITATA}}$

MA ANCHE $\boxed{\left\{\frac{d_k}{\|d_k\|}\right\} \text{ È LIMITATA}}$ QUINDI $\boxed{\left\{\nabla f(x^k)^T d_k\right\} \text{ È LIMITATA}}$

MA ANCHE $\left\{ \frac{d_k}{\|d_k\|} \right\}$ ..., QUINDI $\left\{ \frac{\nabla f(x^k)^T d_k}{\|d_k\|} \right\}$ È LIMITATA

E QUINDI AMMETTE PUNTI DI ACCUMULAZIONE

ASSUMO ORA CHE (b) SIA FALSA: $\exists\, K \subseteq \{0,1,\dots\}$ TALE CHE:

$$\boxed{\lim_{\substack{k\to\infty \\ k\in K}} \frac{\nabla f(x^k)^T d_k}{\|d_k\|} = -\mu < 0}$$

$\to$ QUINDI NON PIÙ ≤ 0

E POSSIAMO INOLTRE NOTARE CHE $\{x^k\}$ È LIMITATA COSÌ COME $\boxed{\left\{ \frac{d_k}{\|d_k\|} \right\}}$, QUINDI $\exists$ UN'ALTRA SOTTOSEQUENZA $K_1 \subseteq K$ t.c.:

$$\boxed{x^{k} \xrightarrow[K_1]{} \bar{x} \quad \text{E} \quad \frac{d_k}{\|d_k\|} \xrightarrow[K_1]{} \bar{d}}$$

E PER LA CONTINUITÀ DEL GRADIENTE POSSO SCRIVERE:

$$\boxed{\lim_{\substack{k\to\infty \\ k\in K_1}} \frac{\nabla f(x^k)^T d_k}{\|d_k\|} = \nabla f(\bar{x})^T \bar{d} = -\mu < 0}$$

$$\Rightarrow \lim_{\substack{k\to\infty \\ k\in K_1}} \gamma \alpha_k \|d_k\| \frac{\nabla f(x^k)^T d_k}{\|d_k\|} = 0 \Rightarrow \boxed{\lim_{\substack{k\to\infty \\ k\in K_1}} \gamma \alpha_k \|d_k\| = 0}$$

$\hookrightarrow -\mu \neq 0$

SI POSSONO QUINDI DISTINGUERE DUE CASI:

① $\exists \bar{k} \in K_1$ t.c.: $\forall k \in K_1$ E $k \geq \bar{k}$ VALE CHE:

$$\boxed{\alpha_k = \alpha_0(k)}$$

② $\exists K_2 \subseteq K_1$ t.c.: $\boxed{\alpha_k < \alpha_0(k)}$ $\forall k \in K_2$

ANALIZZIAMO IL PRIMO CASO:

1) $$\alpha_k \|d_k\| = \alpha_0(k)\|d_k\| \geq \frac{1}{\cancel{\|d_k\|}} \sigma\left(\frac{|\nabla f(x^k)^T d_k|}{\|d_k\|}\right) \cancel{\|d_k\|}$$

≥ 0 PER DEFINIZIONE DI σ

QUINDI $$\underbrace{\alpha_k \|d_k\|}_{\substack{k\to\infty \\ k\in K_1 \\ \downarrow \\ 0}} \geq \underbrace{\sigma\left(\frac{|\nabla f(x^k)^T d_k|}{\|d_k\|}\right)}_{\hookrightarrow 0 \text{ PER IL TEOREMA DEI CARABINIERI}} \geq 0$$

MA SE $\sigma(\cdot) \to 0$ ALLORA $$\frac{|\nabla f(x^k)^T d_k|}{\|d_k\|} \xrightarrow[k\in K_1]{k\to\infty} 0$$

$\hookrightarrow$ ASSURDO!

$$= \frac{|\nabla f(x^k)^T d_k|}{\ldots} \xrightarrow[k\in K_1]{k\to\infty}$$

MA ANCHE $\frac{\nabla f(x^k)^T d_k}{\|d_k\|} \longrightarrow -\mu < 0$

$\Rightarrow$ QUINDI (1) NON PUÒ ESSERE VERA E QUINDI VALE (2), OVVERO

$$\boxed{\alpha_k < \alpha_0(k) \quad \forall k \in K_2}$$

QUINDI $f\left(x^k + \frac{\alpha_k}{\delta} d_k\right) > f(x^k) + \gamma \frac{\alpha_k}{\delta} \nabla f(x^k)^T d_k$ PER LE PROPRIETÀ DEL METODO DI ARMIJO. CONSIDERIAMO ADESSO IL TEOREMA DELLA MEDIA:

$$\boxed{f\left(x^k + \frac{\alpha_k}{\delta} d_k\right) = f(x^k) + \frac{\alpha_k}{\delta} \nabla f(\xi_k)^T d_k}$$

CON $\xi_k = x^k + t_k \frac{\alpha_k}{\delta} d_k$ PER $t \in (0,1)$

SI HA ALLORA CHE:

$$\frac{\alpha_k}{\delta} \nabla f(\xi_k)^T d_k = f\left(x^k + \frac{\alpha_k}{\delta} d_k\right) - f(x^k) > \gamma \frac{\alpha_k}{\delta} \nabla f(x^k)^T d_k$$

$$\Rightarrow \cancel{\frac{\alpha_k}{\delta}} \nabla f(\xi_k)^T d_k > \gamma \cancel{\frac{\alpha_k}{\delta}} \nabla f(x^k)^T d_k$$

$$\Rightarrow \nabla f(\xi_k)^T \frac{d_k}{\|d_k\|} > \gamma \nabla f(x^k)^T \frac{d_k}{\|d_k\|}$$

PER IL $\lim_{\substack{k\to\infty \\ k\in K_2}}$ SI HA CHE $\frac{d_k}{\|d_k\|} \to \bar{d}$, $x^k \to \bar{x}$

E PER IL $\lim_{\substack{k\to\infty \\ k\in K_2}} \xi_k = \lim_{\substack{k\to\infty \\ k\in K_2}} \overbrace{x^k}^{\to \bar{x}} + \underbrace{t_k}_{\in (0,1)} \overbrace{\frac{\alpha_k}{\delta} d_k}^{\to 0} = \bar{x}$

QUINDI OTTENIAMO CHE $\nabla f(\bar{x})^T \bar{d} \geq \gamma \nabla f(\bar{x})^T \bar{d}$ OVVERO CHE

$$\underbrace{(1-\gamma)}_{>0} \underbrace{\nabla f(\bar{x})^T \bar{d}}_{=-\mu<0} \geq 0 \Rightarrow \text{ASSURDO!} \qquad \square$$

METODO DEL GRADIENTE

DATO $x^0 \in \mathbb{R}^n$ L'ALGORITMO È:

$k = 0$

WHILE $(\nabla f(x^k) \neq 0)$:

- SI PONE $d_k = -\nabla f(x^k)$
- SI CALCOLA α_k LUNGO d_k COL METODO DI ARMIJO
- SI PONE $x^{k+1} = x^k - \alpha_k \nabla f(x^k)$

. $k = k+1$

END WHILE

SIA $f \in C^1(\mathbb{R}^n)$ E SIA $x^0 \in \mathbb{R}^n$ TALE CHE $\mathcal{L}_0$ È COMPATTO, ALLORA IL METODO DEL GRADIENTE GENERA UNA SEQUENZA DI SOLUZIONI $\{x^k\}$ CHE AMMETTE PUNTI DI ACCUMULAZIONE, OGNUNO DEI QUALI RISULTA ESSERE UN PUNTO STAZIONARIO

↳ QUINDI SODDISFA LE IPOTESI CONVERGENZA DEGLI ALGORITMI ITERATIVI

INOLTRE SI HA CHE IL METODO DEL GRADIENTE PRESENTA UN TASSO DI CONVERGENZA SUBLINEARE

↳ $\lim_{k \to \infty} \frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|} = 1$

DEFINIZIONE FUNZIONE FORTEMENTE CONVESSA

SI DICE CHE UNA FUNZIONE $\boxed{f : \mathbb{R}^n \mapsto \mathbb{R}}$ È FORTEMENTE CONVESSA

SE ESISTE UNA FUNZIONE $\boxed{g : \mathbb{R}^n \mapsto \mathbb{R}}$ E $\boxed{\mu > 0}$ TALI CHE:

$$\boxed{f(x) = g(x) + \mu \|x\|^2 \quad \forall x \in \mathbb{R}^n}$$

QUINDI SE f È UNA FUNZIONE FORTEMENTE CONVESSA, ALLORA IL METODO DEL GRADIENTE PRESENTA UN TASSO DI CONVERGENZA LINEARE

LINEARE

$\hookrightarrow \lim_{k \to \infty} \frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|} = c \in (0,1)$

METODO DEL GRADIENTE A PASSO COSTANTE

DEFINIAMO PRIMA LA SEGUENTE IPOTESI:

SIA $f \in C^2(\mathbb{R}^n)$ E SIA ∇f LIEPSCHITZ-CONTINUA DI COSTANTE L, CIOÈ:

$$\boxed{\forall x, y \in \mathbb{R}^n : \|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|}$$

E SE LA FUNZIONE $f \in C^2(\mathbb{R}^n)$ POSSIAMO ANCHE SCRIVERE CHE:

$$\boxed{|\mu^T \nabla^2 f(x) \mu| \leq \|\mu\|^2 \quad \forall x, \mu \in \mathbb{R}^n}$$

QUINDI, CON QUESTE IPOTESI ($f \in C^2(\mathbb{R}^n)$ E ∇f LIEPSCHITZ-CONTINUO) POSSIAMO DEFINIRE IL METODO DEL GRADIENTE A PASSO COSTANTE, DOVE VALE CHE:

$$\boxed{x^{k+1} = x^k - \underbrace{\frac{1}{L}}_{\hookrightarrow\ 2\kappa = \frac{1}{L} \text{ COSTANTE}} \nabla f(x^k)}$$

E VALE CHE, $\forall d, x \in \mathbb{R}^n$ E $\forall t \geq 0$:

$$\boxed{f(x+td) \leq f(x) + t\nabla f(x)^T d + t^2 \frac{L}{2}\|d\|^2}$$

Dim:

CONSIDERANDO LA FORMULA DI TAYLOR:

$$f(x+td) = f(x) + t\nabla f(x)^T d + \frac{t^2}{2} d^T \nabla^2 f(y) d \quad \text{PER QUALCHE } y \in \mathbb{R}^n$$

$$\leq f(x) + t\nabla f(x)^T d + \frac{t^2}{2} \left| d^T \nabla^2 f(y) d \right|$$

$$\leq f(x) + t\nabla f(x)^T d + t^2 \frac{L}{2}\|d\|^2 \quad \square$$

SIA $f : \mathbb{R}^n \mapsto \mathbb{R}$ TALE CHE $\boxed{f \in C^2(\mathbb{R})}$ CON $\boxed{\nabla f \text{ LIEPSCHITZ-CONTINUO}}$ DI COSTANTE L, $x^0 \in \mathbb{R}^n$ TALE CHE $\boxed{\mathcal{L}_0 \text{ E' COMPATTO}}$, ALLORA IL METODO DEL GRADIENTE CON PASSO $\boxed{\alpha_k = \frac{1}{L} \ \forall k}$ GENERA UNA SEQUENZA $\{x^k\}$ t.c.:

1. $\boxed{f(x^{k+1}) \leq f(x^k) - \frac{1}{2L}\|\nabla f(x^k)\|^2 \ \forall k}$

2. $\boxed{\{x^k\} \text{ HA PUNTI DI ACCUMULAZIONE STAZIONARI}}$

Dim:

1) scrivo $\nabla f(x^k) = d_k$:

$$f(x^{k+1}) = f(x) - \alpha_k d_k$$
$$\leq f(x) - \alpha_k d_k^T d_k + \alpha_k^2 \frac{L}{2} \| d_k \|^2$$

ma $\alpha_k = \frac{1}{L}$ e $d_k = -\nabla f(x^k)$

$$\Rightarrow f(x^{k+1}) = f(x^k) - \frac{1}{L} \nabla f(x^k)^T \nabla f(x^k) + \frac{1}{L^2} \frac{L}{2} \| \nabla f(x^k) \|^2$$
$$= f(x^k) - \frac{1}{L} \| \nabla f(x^k) \|^2 + \frac{1}{2L} \| \nabla f(x^k) \|^2$$
$$= f(x^k) - \frac{1}{2L} \| \nabla f(x^k) \|^2 \quad \square$$

---

2) $$f(x^{k+1}) - f(x^k) \leq -\frac{1}{2L} \| \nabla f(x^k) \|^2 < 0$$

- $f(x^k)$ è monotona decrescente
- $\{x^k\} \subseteq \mathcal{L}_0 \Rightarrow \exists$ punti di accumulazione

$\Rightarrow f(x^k) \to f^*$ finito

ALLORA SIA $\bar{x}$ UN PUNTO QUALSIASI TRA QUELLI DI ACCUMULAZIONE.

CIOÈ $\exists K \subseteq \{0,1,\dots\}$ TALE CHE: $\lim_{\substack{k\in K \\ k\to\infty}} x^k = \bar{x}$, ALLORA

$$\underbrace{\lim_{\substack{k\to\infty \\ k\in K}} f(x^{k+1}) - f(x^k)}_{f^* - f^* = 0} \le \underbrace{\lim_{\substack{k\to\infty \\ k\in K}} -\frac{1}{2L}\|\nabla f(x^k)\|^2}_{\le -\frac{1}{2L}\|\nabla f(\bar{x})\|^2 \le 0}$$

$\Rightarrow \|\nabla f(\bar{x})\| = 0$ $\square$

COMPLESSITÀ DEGLI ALGORITMI DI OTTIMIZZAZIONE

SOTTO LE SEGUENTI IPOTESI:

- f LIMITATA INFERIORMENTE DA f^*
- f CONVESSA
- ∇f È ...

- ∇f È LIPSCHITZ-CONTINUO

DEFINIAMO

- UN ALGORITMO HA UN ITERATION ERROR DI $O(h(k))$ SE

$$\boxed{f(x^k) - f^* = O(h(k))}$$

- LA ITERATION COMPLEXITY È $O(\hat{h}(\varepsilon))$ SE:

$$\boxed{\min\{k \mid f(x^k) - f^* \leq \varepsilon\} = O(\hat{h}(\varepsilon))}$$

VEDIAMO IL LEGAME TRA LE DUE QUANTITÀ:

- SE HO $f(x^k) - f^* \approx \frac{1}{k}$ ALLORA:

$$\boxed{f(x^k) - f^* < \varepsilon \Rightarrow k > \frac{1}{\varepsilon} \Rightarrow k = O\left(\frac{1}{\varepsilon}\right)}$$

- SE HO $f(x^k) - f^* \approx \frac{1}{k^2}$ ALLORA:

$$\boxed{f(x^k) - f^* < \varepsilon \Rightarrow k > \frac{1}{\sqrt{\varepsilon}} \Rightarrow k = O\left(\frac{1}{\sqrt{\varepsilon}}\right)}$$

SI HA CHE IL COSTO DI OGNI CIFRA SIGNIFICATIVA DI PRECISIONE CRESCE IN MANIERA ESPONENZIALE CON IL NUMERO DI ITERAZIONI!

ε	# CIFRE SIGNIFICATIVE $= \log\left(\frac{1}{\varepsilon}\right)$	COSTO PER RAGGIUNGERE IL NUMERO DI CIFRE SIGNIFICATIVE SE LA COMPLESSITÀ VALE $O(1/\varepsilon)$
0.1	1	10
0.01	2	100
0.001	3	1000

exp

SE UN CERTO ALGORITMO PRESENTA UN $\varepsilon = O\left(\frac{1}{k}\right)$ OVVERO:

$$\lim_{k\to\infty} \frac{f(x^{k+1})-f^*}{f(x^k)-f^*} \approx \lim_{k\to\infty} O\left(\frac{k}{k+1}\right) = 1$$

SI OTTIENE CHE HA UN TASSO SUBLINEARE. ANALOGAMENTE SI OTTIENE CHE, CON $\delta > 0$:

ITERAZIONI ERRORE

ITERATION ERROR	ITERATION COMPLEXITY	TASSO
$O\left(\frac{1}{k}\right)$	$O\left(1/\varepsilon\right)$	SUBLINEARE ×
$O(\rho^k)$	$O\left(\log \frac{1}{\varepsilon}\right)$	LINEARE ✓
$O(\rho^{2k})$	$O\left(\log\left(\log \frac{1}{\varepsilon}\right)\right)$	SUPERLINEARE ✓✓

PROPOSIZIONE

SIANO LE SEGUENTI IPOTESI:

- $f \in C^2(\mathbb{R}^n)$
- ∇f È LIPSCHITZ-CONTINUO DI COSTANTE L
- $x^0 \in \mathbb{R}^n$ t.c. $\mathcal{L}_0$ È COMPATTO
- $\alpha_k = \frac{1}{L} \ \forall k$
- f CONVESSA CON MINIMO f^* IN x^*

ALLORA POSSIAMO AFFERMARE CHE IL METODO DEL GRADIENTE COSTANTE GENERA UNA SEQUENZA $\{x^k\}$ t.c.:

$$\boxed{f(x^k) - f^* \leq \frac{L}{2} \frac{\|x^* - x^0\|^2}{k+1}}$$

OVVERO LA COMPLESSITÀ DEL METODO DEL GRADIENTE A PASSO COSTANTE È $O(1/k) \Leftrightarrow O(1/\varepsilon)$

Dim:

RICORDIAMO CHE
$$\boxed{f(x^{k+1}) \leq f(x^k) - \frac{\|\nabla f(x^k)\|^2}{2L}} \quad \Delta$$

E INOLTRE PER IPOTESI SI HA CHE f È CONVESSA, QUINDI:

$$\boxed{f(x^*) \geq f(x^k) + \nabla f(x^k)^T(x^* - x^k)} \quad \bigcirc$$

$\hookrightarrow f(x^*)$ STA SOPRA L'IPERPIANO TANGENTE AD f PASSANTE PER x^k E CALCOLATO IN x^*

POSSIAMO QUINDI SCRIVERE PER Δ E $\bigcirc$:

$$f(x^{k+1}) \overset{\Delta}{\leq} f(x^k) - \frac{1}{2L}\|\nabla f(x^k)\|^2 \overset{\bigcirc}{\leq} f(x^*) - \nabla f(x^k)^T(x^* - x^k) - \frac{1}{2L}\|\nabla f(x^k)\|^2$$

$$\Rightarrow f(x^{k+1}) - f(x^*) \leq -\nabla f(x^k)^T(x^* - x^k) - \frac{1}{2L}\|\nabla f(x^k)\|^2$$
$$= \frac{L}{2}\left(-\frac{2}{L}\nabla f(x^k)^T(x^* - x^k) - \frac{1}{L^2}\|\nabla f(x^k)\|^2\right)$$
$$= \frac{L}{2}\left(-\frac{2}{L}\nabla f(x^k)^T(x^* - x^k) - \frac{1}{L^2}\|\nabla f(x^k)\|^2 + \|x^* - x^k\|^2 - \|x^* - x^k\|^2\right.$$

$$= \frac{L}{2}\left(\|x^*-x^k\| - \underbrace{\left(\|x^*-x^k\|^2 + \frac{2}{L}\nabla f(x^k)^T(x^*-x^k) + \frac{1}{L^2}\|\nabla f(x^*)\|^2\right)}_{\Delta}\right) =$$

NOTIAMO CHE SE PONIAMO $(x^*-x^k)=a$ E $\frac{1}{L}\nabla f(x^k)=b$ OTTENIAMO

$$\|a+b\|^2 = a^Ta + b^Tb + 2a^Tb = \|a\|^2 + \|b\|^2 + 2a^Tb$$

QUINDI Δ DIVENTA:

$$\boxed{\|x^*-x^k\|^2 + \frac{1}{L^2}\|\nabla f(x^k)\|^2 + \frac{2}{L}\nabla f(x^k)^T(x^*-x^k)}$$

CONTINUANDO LA CATENA DI ESPRESSIONI SI OTTIENE:

$$= \frac{L}{2}\left(\|x^*-x^k\|^2 - \|\underbrace{x^*-x^k+\frac{1}{2}\nabla f(x^k)}\|^2\right)$$

$$\|$$

$$-\left(x^k - \frac{1}{L}\nabla f(x^k)\right) = -x^{k+1}$$

$$= \frac{L}{2}\left(\|x^*-x^k\|^2 - \|x^*-x^{k+1}\|^2\right)$$

$$\Rightarrow \boxed{f(x^{k+1}) - f(x^*) \le \frac{L}{2}\left(\|x^*-x^k\|^2 - \|x^*-x^{k+1}\|^2\right)}$$

SE CONSIDERIAMO LE SOMME:

$$\boxed{\sum_{t=0}^{K} f(x^{t+1})-f(x^*) \leq \sum_{t=0}^{K} \frac{L}{2}\left(\|x^*-x^t\|^2-\|x^*-x^{t+1}\|^2\right)}$$

OTTENIAMO UNA SERIE TELESCOPICA DOVE SI SEMPLIFICANO TUTTI I TERMINI ADIACENTI:

$$= \frac{L}{2}\left(\|x^*-x^0\|^2-\|x^*-x^{K+1}\|\right) \leq \frac{L}{2}\left(\|x^*-x^0\|^2\right)$$

QUINDI:

$$\boxed{\sum_{t=0}^{K} f(x^{t+1})-f(x^*) \leq \frac{L}{2}\left(\|x^*-x^0\|^2\right)}$$

MA $\{f(x^K)\}$ È MONOTONA DECRESCENTE E QUINDI

$$f(x^{t+1}) \geq f(x^{K+1}) \quad \forall t \leq K$$

QUINDI

$$\boxed{f(x^{t+1})-f(x^t) \geq f(x^{k+1})-f(x^*)}$$

POSSIAMO INFINE SCRIVERE CHE:

$$\sum_{t=0}^{k} f(x^{k+1})-f(x^*) \leq \sum_{t=0}^{k} f(x^{t+1})-f(x^*) \leq \frac{L}{2}\left(\|x^*-x^0\|^2\right)$$

POICHÉ $f(x^{t+1}) \geq f(x^{k+1})$ $\forall t \leq k$

OVVERO:

$$(k+1)\left(f(x^{k+1})-f(x^*)\right) \leq \frac{L}{2}\|x^*-x^0\|^2$$

$$\Rightarrow \boxed{f(x^{k+1})-f(x^*) \leq \frac{L\|x^*-x^0\|^2}{2(k+1)}} \quad \square$$

PROPOSIZIONE

NEL CASO IN CUI f SIA FORTEMENTE CONVESSA SI HA CHE IL METODO DEL GRADIENTE HA COMPLESSITÀ $O(\log\frac{1}{\varepsilon})$ LINEARE

INOLTRE, SE CONSIDERIAMO UN GENERICO PROBLEMA DI APPRENDIMENTO SUPERVISIONATO

$$\boxed{\min_{w} \mathcal{L}(w) + \lambda \Omega(w)}$$

SI HA CHE SE $\mathcal{L}(w)$ È CONVESSA, ALLORA L'AGGIUNTA DEL REGOLARIZZATORE RENDE LA FUNZIONE OBIETTIVO FORTEMENTE CONVESSA E QUINDI MIGLIORA LE PRESTAZIONI!

METODI CON TERMINI DI TIPO MOMENTUM

1) METODO HEAVY-BALL

LA FORMULA DI AGGIORNAMENTO DIVENTA:

$$\boxed{x^{k+1} = x^k - \alpha_k \nabla f(x^k) + \beta \overbrace{(x^k - x^{k-1})}}$$

TERMINE CHE DIPENDE DALLO SPOSTAMENTO ALL'ITERAZIONE PRECEDENTE!

IL METODO SI MUOVE DI UNA COMBINAZIONE TRA L'ANTIGRADIENTE E LA DIREZIONE DELL'ULTIMO SPOSTAMENTO E QUINDI SI PUÒ RISCRIVERE COSÌ:

$$\begin{cases} y^k = x^k - \alpha_k \nabla f(x^k) \\ x^{k+1} = y^k + \beta (x^k - x^{k-1}) \end{cases}$$

$$\begin{cases} v^k = x^k - x^{k-1} \\ x^{k+1} = x^k + v^k - \alpha_k \nabla f(x^k) \end{cases}$$

ESSENDO CHE IL METODO DEL GRADIENTE DIPENDE FORTEMENTE DALLA SCELTA DEL PUNTO INIZIALE. QUINDI QUI IL TERMINE

$$\beta : (x^k - x^{k-1})$$

TENDE AD EVITARE LE VARIE OSCILLAZIONI DOVUTE AD UN CATTIVO PUNTO DI PARTENZA.

- PRO: - SE f È FORTEMENTE CONVESSA ALLORA LA COMPLESSITÀ È MIGLIORE. SE f È QUADRATICA ALLORA α E β POSSONO ESSERE SCELTI IN MODO OTTIMALE E IN FORMA CHIUSA.

- CONTRO: NEL CASO GENERALE NON SAPPIAMO COME SCEGLIERE β ED INOLTRE ESISTONO DEI CASI IN CUI IL METODO NON CONVERGE ANCHE SOTTO IPOTESI DI REGOLARITÀ.

2) METODO DI NESTEROV

LA FORMULA DI AGGIORNAMENTO DIVENTA:

$$x^{k+1} = x^k + \beta(x^k - x^{k-1}) - \alpha_k \nabla f(x^k + \beta(x^k - x^{k-1}))$$

DOVE:

$$\begin{cases} v^k = \beta(x^k - x^{k-1}) \\ x^{k+1} = x^k + v^k - \alpha_k \nabla f(x^k + v^k) \end{cases}$$

$$\begin{cases} y^k = x^k - \alpha_k \nabla f(x^k + \beta(x^k - x^{k-1})) \\ x^{k+1} = y^k + \beta(x^k - x^{k-1}) \end{cases}$$

L'ALGORITMO, CON β SCELTO OPPORTUNAMENTE, HA COMPLESSITÀ $O\left(\frac{1}{k^2}\right) \iff O\left(\frac{1}{\sqrt{\varepsilon}}\right)$ NEL CASO DI f CONVESSA ED È OTTIMO PER GLI ALGORITMI DEL PRIMO ORDINE!

DEFINIZIONE: L'INSIEME DI DIREZIONI MUTUAMENTE CONIUGATE

SIA $Q \in \mathbb{R}^{n\times n}$ SIMMETRICA E D.P., ALLORA UN INSIEME DI DIREZIONI $d_0, d_1, \dots, d_{m-1}$ CON $m \leq n$ È DETTO INSIEME DI DIREZIONI MUTUAMENTE CONIUGATE RISPETTO A Q SE:

$$\boxed{\forall i,j = 0, \dots, m-1 \text{ con } i \neq j: \; d_i^T Q d_j = 0}$$

SIA $d_0, \dots, d_{m-1}$ INSIEME DI DIREZIONI MUTUAMENTE CONIUGATE, ALLORA $d_0, \dots, d_{m-1}$ SONO LINEARMENTE INDIPENDENTI

Dim:

CONSIDERIAMO UNA GENERICA COMBINAZIONE LINEARE NON NULLA DI $d_0, \dots, d_{m-1}$:

$$\boxed{\sum_{j=0}^{m-1} \alpha_j d_j = 0}$$

SIA $i \in \{0, \dots, m-1\}$. MOLTIPLICHIAMO LA COMBINAZIONE LINEARE PER $d_i^T Q$:

$$\boxed{d_i^T Q \sum_{j=0}^{m-1} \alpha_j d_j = 0}$$

MA VALE LA LINEARITÀ:

$$\boxed{\sum_{j=0}^{m-1} \alpha_j d_i^T Q d_j = 0}$$

PER IPOTESI SAPPIAMO CHE $d_i^T Q d_j = 0 \quad \forall i \neq j$ POICHÉ DIREZIONI MUTUAMENTE CONIUGATE. ALLORA:

$$\alpha_i \underbrace{d_i^T Q d_i}_{>0} = 0$$

>0 POICHÉ Q È DEFINITA POSITIVA

ALLORA $\boxed{\alpha_i = 0}$ E POICHÉ i È STATO SCELTO IN MANIERA ARBITRARIA SI HA CHE $\alpha_i = 0 \quad \forall i = 0, \dots, m-1$, OVVERO LE DIREZIONI SONO LINEARMENTE INDIPENDENTI □

CASO PROBLEMA QUADRATICO

SIA $\min_{x \in \mathbb{R}^n} \frac{1}{2} x^T Q x - c^T x$ UN PROBLEMA QUADRATICO. IN QUESTO CASO SAPPIAMO CHE, RAGGIUNTA LA SOLUZIONE OTTIMA, SI HA:

$$\boxed{\nabla f(x^*) = Q x^* - c = 0}$$

SE $d_0, \dots, d_{n-1}$ SONO DIREZIONI MUTUAMENTE CONIUGATE, E QUINDI SONO ANCHE $\underbrace{\text{LINEARMENTE INDIPENDENTI}}$, SI HA:

↳ QUINDI SONO UNA BASE!

$$\boxed{x^* = \sum_{s=0}^{n-1} \alpha_s^* d_s}$$

QUINDI $Qx^* = \sum_{s=0}^{n-1} \alpha_s^* Q d_s$. SE CONSIDERIAMO LA QUANTITÀ:

$$d_i^T Q x^* = \sum_{s=0}^{n-1} \alpha_s^* d_i^T Q d_s$$

$$= \underline{\alpha_i^* d_i^T Q d_i}$$ PER DEFINIZIONE DELLE DIREZIONI

ALLORA, DATO CHE VALE $Qx^* = c$ NELL'OTTIMO:

$$d_i^T c = \alpha_i^* d_i^T Q d_i$$

$$\Rightarrow \boxed{\alpha_i^* = \frac{d_i^T c}{d_i^T Q d_i}}$$

POSSIAMO QUINDI RICAVARE LA SOLUZIONE OTTIMA $x^* = \sum^{n-1} \frac{d_s^T c}{\dots} d_s$

$\boxed{\dots_{s=0} d_s^T Q d_s}$

DIRETTAMENTE, SENZA QUINDI PASSARE DALL'INVERSIONE DI Q PER RISOLVERE IL SISTEMA $Qx^*=c$!

OSSERVAZIONE

DIFFERENZA TRA IL PUNTO DI OTTIMO ED IL PUNTO INIZIALE

CONSIDERIAMO $\boxed{x^*-x^0=\sum_{s=0}^{n-1}\alpha_s d_s}$. SE SCOMPONIAMO LA SOMMATORIA OTTENIAMO:

$$x^*=x^0+\alpha_0 d_0+\alpha_1 d_1+\alpha_2 d_2+\dots+\alpha_{n-1}d_{n-1}$$

CHE RICORDA MOLTO IL PROCESSO DI AGGIORNAMENTO ITERATIVO

$$\boxed{x^k=x^0+\sum_{s=0}^{k-1}\alpha_s d_s}$$

PRENDIAMO ALLORA x^*-x^0 E x^k-x^0. VALGONO:

1) $d_k^T Q(x^k-x^0)=d_k^T Q\sum_{s=0}^{k-1}\alpha_s d_s=\sum_{s=0}^{k-1}\alpha_s d_k^T Q d_s=0$

2) $d_k^T Q(x^*-x^0)=\dots=\sum_{s=0}^{n-1}\alpha_s d_k^T Q d_s=\alpha_k d_k^T Q d_k$

$$d_k^T Q(x^*-x^0) = \ldots = \sum_{s=0} \alpha_s d_k Q d_s = \alpha_k d_k^T Q d_k$$

SEGUE CHE:

$$\alpha_k = \frac{d_k^T Q(x^*-x^0)}{d_k^T Q d_k} = \frac{d_k^T Q(x^* \underbrace{+x^k-x^k}{} -x^0)}{d_k^T Q d_k} = \frac{d_k^T Q(x^*-x^k)}{d_k^T Q d_k} + \cancel{\frac{d_k^T Q(x^k-x^0)}{d_k^T Q d_k}}_{=0 \text{ PER } (1)}$$

$$= \frac{d_k^T (Qx^* - Qx^k)}{d_k^T Q d_k} = \frac{d_k^T (c - Qx^k)}{d_k^T Q d_k} = \boxed{\frac{-d_k^T \nabla f(x^k)}{d_k^T Q d_k}}$$

PASSO OTTIMO PER LA DISCESA NEL CASO QUADRATICO!

E, CON LA NOTAZIONE $\boxed{\nabla f(x^k) = g_k}$ SI HA:

$$\boxed{x^{k+1} = x^k - \frac{g_k^T d_k}{d_k^T Q d_k} d_k}$$

CHE È IL METODO DELLE DIREZIONI CONIUGATE

PROPRIETÀ DI CONVERGENZA FINITA PER IL METODO DELLE DIREZIONI CONIUGATE

SIANO $d_0, \dots, d_{n-1}$ DIREZIONI MUTUAMENTE CONIUGATE E SIA $x^0 \in \mathbb{R}^n$ CON:

$$\boxed{x^{k+1} = x^k + \alpha_k d_k}$$

CON $\boxed{\alpha_k = \frac{-d_k^T g_k}{d_k^T Q d_k}}$, ALLORA:

1) $\forall k$ VALE $\boxed{g_k^T d_i = 0}$ $\forall i = 0, \dots, k-1$

2) $\boxed{\exists\, m \leq n \text{ t.c. } x^m = x^*}$ → OVVERO SI RIESCE A TROVARE LA SOLUZIONE OTTIMA ESATTA IN UN NUMERO FINITO m DI PASSI!

<u>Dim</u>

1) SIA $i \in \{0, \dots, k-1\}$, ALLORA:

$$\boxed{x^k = x^{\overset{\text{i-esimo passo prima di k}}{i}} + \sum_{s=i}^{k-1} \alpha_s d_s}$$

$$\Rightarrow g_k^T d = (Qx^k - c)^T d = \left(Qx^i + \sum_{s=i}^{k-1} \alpha_s Q d_s - c\right)^T d$$

$$= (Qx^i - c)^T d + \sum_{s=i}^{k-1} \alpha_s d_i^T Q d_s$$

$$= g_i^T d_i + \underbrace{\alpha_i}_{\downarrow} d_i^T Q d_i$$

$$= g_i^T d_i - \frac{d_i^T g_i}{\cancel{d_i^T Q d_i}} \cancel{d_i^T Q d_i}$$

$$= g_i^T d_i - d_i^T g_i = 0 \quad \square$$

---

2) SUPPONIAMO DI AVER ESEGUITO n ITERAZIONI. ABBIAMO QUINDI OTTENUTO x^n TALE CHE:

$$\boxed{g_n^T d_i = 0 \quad \forall_{i=0,\dots,n-1}}$$

OVVERO IL GRADIENTE È PERPENDICOLARE AD n DIREZIONI CHE SONO LINEARMENTE INDIPENDENTI, OVVERO CHE COSTITUISCONO UNA BASE, ALLORA DEVE ESSERE $\boxed{g_n = 0}$ OVVERO:

$$\boxed{Q x^n - c = 0} \Rightarrow \boxed{x^n = x^*} \quad \square$$

METODO DEL GRADIENTE CONIUGATO

È IL METODO MIGLIORE NEL CASO DI FUNZIONI QUADRATICHE E SI USA SPESSO PER LA RISOLUZIONE DI SISTEMI LINEARI CON Q SIMMETRICA E DEFINITA POSITIVA.

$x^0 \in \mathbb{R}^n$, $d_0 = -g_0$, $k=0$

WHILE $(\|g_k\| \neq 0)$:

- CALCOLA $\alpha_k = \dfrac{\|g_k\|^2}{d_k^T Q d_k} = -\dfrac{g_k^T d_k}{d_k^T Q d_k}$
- $x^{k+1} = x^k + \alpha_k d_k$
- $\beta_{k+1} = \dfrac{\|g_{k+1}\|^2}{\|g_k\|^2}$
- $d_{k+1} = -g_{k+1} + \beta_{k+1} d_k$
- $k = k+1$

END WHILE

DOVE POSSIAMO NOTARE CHE:

$$g_{k+1} = Qx^{k+1} - c = Q(x^k + \alpha_k d_k) - c$$
$$= Qx^k - c + \alpha_k Q d_k$$
$$= g_k + \alpha_k Q d_k$$

NEL CASO VOLESSIMO APPLICARE QUESTO METODO NEL CASO

GENERALE DOBBIAMO FARE ALCUNE MODIFICHE:

IN PARTICOLARE OCCORRE MODIFICARE LA GUARDIA DEL CICLO POICHÉ NON ABBIAMO PIÙ LA GARANZIA DEL PASSO OTTIMO TIPICA DEI PROBLEMI QUADRATICI.

$$\boxed{\|g_k\| < \varepsilon}$$

PER QUANTO RIGUARDA LA SCELTA DEL PASSO α_k VOGLIAMO GARANTIRE CHE:

$$\boxed{d_{k+1}^T g_{k+1} < 0}$$

DOVE $\boxed{d_{k+1}^T g_{k+1} = -\|g_{k+1}\|^2 + \beta_k d_k^T g_{k+1}}$

QUINDI NON BASTA USARE LA CONDIZIONE DI ARMIJO MA OCCORRE INTRODURRE:

CONDIZIONI DI WOLFE

GARANTISCE CHE IL PASSO SCELTO RIDUCA IN MODO SUFFICIENTE IL VALORE DELLA FUNZIONE OBIETTIVO

- CONDIZIONE DEBOLE:

$$\cdot\ f(x^k + \alpha_k d_k) \le f(x^k) + \gamma \alpha_k \nabla f(x^k)^T d_k$$

$$\cdot\ \nabla f(x^k + \alpha_k d_k)^T d_k \ge \nabla f(x^k)^T d_k\, \sigma \qquad \boxed{\sigma \in (0,1)}$$

$\sigma \in (0, \frac{1}{2})$

$\varphi'(\alpha_k)$ $\varphi'(0)$

- CONDIZIONE FORTE: → GARANTISCE CHE IL PASSO SCELTO NON SCAVALCHI L'OTTIMO

$$\cdot\ f(x^k + \alpha_k d_k) \leq f(x^k) + \gamma \alpha_k \nabla f(x^k)^T d_k$$

$$\cdot\ |\nabla f(x^k + \alpha_k d_k)^T d_k| \leq \sigma |\nabla f(x^k)^T d_k|$$

QUESTE CONDIZIONI GARANTISCONO CHE IL PASSO SCELTO MANTENGA IL PROGRESSO VERSO L'OTTIMO, MA ALLO STESSO TEMPO CHE NON SIA TROPPO LUNGO

METODO DI NEWTON

È UN METODO DEL II° ORDINE, DOVE È QUINDI DISPONIBILE ANCHE IL CALCOLO DELLA MATRICE HESSIANA $\nabla^2 f$:

SE CONSIDERIAMO L'APPROSSIMAZIONE AL SECONDO ORDINE DI x^k:

$$\boxed{m_k(x) = f(x^k) + \nabla f(x^k)^T (x - x^k) + \frac{1}{2}(x - x^k)^T \nabla^2 f(x^k)(x - x^k)}$$

SE L'HESSIANA È DEFINITA POSITIVA, ALLORA $m_k(x)$, ESSENDO UNA FUNZIONE QUADRATICA CON $Q = \nabla^2 f$ DP, È STRETTAMENTE CONVESSA. SI PUÒ QUINDI RICAVARE IL MINIMO PONENDO

$$\boxed{\nabla m_k(\hat{x}) = 0}$$

CALCOLATO RISPETTO AD X

$$\Rightarrow \boxed{\nabla m_k(\hat{x}) = \nabla f(x^k) + \nabla^2 f(x^k)(\hat{x} - x^k) = 0}$$

$$\Rightarrow \nabla^2 f(x^k)(\hat{x} - x^k) = -\nabla f(x^k)$$

$$\Rightarrow \hat{x} - x^k = -\left[\nabla^2 f(x^k)\right]^{-1} \nabla f(x^k)$$

$$\Rightarrow \boxed{\hat{x} = x^k - \left[\nabla^2 f(x^k)\right]^{-1} \nabla f(x^k)}$$

APPROSSIMAZIONE DELL'OTTIMO

SI OTTIENE QUINDI LA FORMULA DI AGGIORNAMENTO DEL METODO DI NEWTON:

$$\boxed{x^{k+1} = x^k - \left[\nabla^2 f(x^k)\right]^{-1} \nabla f(x^k)}$$

CHE PUÒ ESSERE INTERPRETATA COME LA CLASSICA FORMULA:

$$x^{k+1} = x^k + \alpha_k d_k \quad \text{CON } \alpha_k = 1$$

$$d_k = -\left[\nabla^2 f(x^k)\right]^{-1} \nabla f(x^k)$$

PROPRIETÀ DI CONVERGENZA LOCALE DI NEWTON

→ E NON GLOBALE COME IL METODO DEL GRADIENTE!

SIA $f \in C^2(\mathbb{R}^n)$ CON $x^* \in \mathbb{R}^n$ TALE CHE $\boxed{\nabla f(x^*) = 0}$ E SIA $\nabla^2 f(x^*)$ NON SINGOLARE, ALLORA $\exists \varepsilon > 0$ TALE CHE $\forall x^0 \in B_\varepsilon(x^*)$ LA SEQUENZA $\{x^k\}$ OTTENUTA DALL'AGGIORNAMENTO: $x^{k+1} = x^k - [\nabla^2 f(x^k)]^{-1} \nabla f(x^k)$

→ DET ≠ 0

SODDISFA LE SEGUENTI PROPRIETÀ:

1. $\boxed{\{x^k\} \subseteq B_\varepsilon(x^*)}$ → AMMETTE PUNTI DI ACCUMULAZIONE NELL'INTORNO DELL'OTTIMO

2. $\boxed{\lim_{k \to \infty} x^k = x^*}$ → CONVERGE OTTIMAMENTE AD UN PUNTO DI MINIMO!

CONVERGENZA LOCALE!

3. $\boxed{\lim_{k \to \infty} \frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|} = 0}$ → TASSO SUPERLINEARE

↳ MA SE $\nabla^2 f$ È LIPSCHITZ-CONTINUA ALLORA IL TASSO È QUADRATICO

DIM: ①, ②

VOGLIAMO DIMOSTRARE CHE $\{x^k\} \subseteq B_\varepsilon(x^*)$ SE $x^0 \in B_\varepsilon(x^*)$ PER QUALCHE ε CON $x^k \to x^*$

RICHIAMIAMO PRIMA IL TEOREMA DELLA MEDIA INTEGRALE:

$F: \mathbb{R}^n \to \mathbb{R}^m$ allora $\forall x,y$: $\boxed{F(x) = F(y) + \int_0^1 \underbrace{J_F}_{J_F:\mathbb{R}^n \to \mathbb{R}^{m\times n}}(y+t(x-y))(x-y)\,dt}$

Essendo, per Hp, $\nabla^2 f(x^*)$ non singolare, allora $\exists \varepsilon_1 > 0$, $\mu > 0$ tali che:

$$\boxed{\|[\nabla^2 f(x^*)]^{-1}\| \le \mu} \quad \forall x \in B_{\varepsilon_1}(x^*)$$

Inoltre, essendo $f \in C^2(\mathbb{R}^n)$ allora $\exists \varepsilon \le \varepsilon_1$ e un $\sigma > 0$ tali che:

$$\boxed{\|\nabla^2 f(x^k) - \nabla^2 f(y)\| \le \frac{\sigma}{\mu}} \quad \forall x,y \in B_\varepsilon(x^*)$$

Supponiamo allora che $\boxed{x^0 \in B_\varepsilon(x^*)}$ e, per induzione, che $\boxed{x^k \in B_\varepsilon(x^*)}$ con x^k ottenuto dopo x^0.

Dimostriamo allora che $\boxed{x^{k+1} \in B_\varepsilon(x^*)}$:

moltiplico e divido

$$x^{k+1} - x^k = x^k - [\nabla^2 f(x^k)]^{-1} \nabla f(x^k) - x^*$$

$$= - [\nabla^2 f(x^k)]^{-1} \nabla f(x^k) + [\nabla^2 f(x^k)]^{-1} [\nabla^2 f(x^k)] (x^k - x^*)$$

$$= -[\nabla^2 f(x^k)]^{-1} \left(\nabla f(x^k) - \nabla^2 f(x^k)(x^k - x^*) \right)$$

$$= -[\nabla^2 f(x^k)]^{-1} \left(\nabla f(x^k) - \underbrace{\nabla f(x^*)}_{0 \text{ PER Hp}} - \nabla^2 f(x^k)(x^k - x^*) \right)$$

$$\Rightarrow \| x^{k+1} - x^* \| = \| [\nabla^2 f(x^k)]^{-1} \left(\nabla f(x^k) - \nabla f(x^*) - \nabla^2 f(x^k)(x^k - x^*) \right) \|$$

$$\leq \underbrace{\| [\nabla^2 f(x^k)]^{-1} \|}_{\leq \mu} \cdot \| \nabla f(x^k) - \nabla f(x^*) - \nabla^2 f(x^k)(x^k - x^*) \|$$

$$\leq \mu \| \nabla f(x^k) - \nabla f(x^*) - \nabla^2 f(x^k)(x^k - x^*) \| \quad \Delta$$

UTILIZZIAMO ORA IL TEOREMA DELLA MEDIA INTEGRALE:

$$\boxed{\nabla f(x^k) = \nabla f(x^*) + \int_0^1 \nabla^2 f\left(x^* + t(x^k - x^*)\right)(x^k - x^*)\, dt}$$

$$\Rightarrow \Delta = \mu \left\| \int_0^1 \nabla^2 f(x^* + t(x^k - x^*))(x^k - x^*)\, dt - \nabla^2 f(x^k)(x^k - x^*) \right\|$$

$$= \mu \left\| \int_0^1 \nabla^2 f(x^* + t(x^k - x^*))(x^k - x^*) - \nabla^2 f(x^k)\overbrace{(x^k - x^*)}^{\text{RACCOLGO}}\, dt \right\|$$

$$= \mu \left\| \int_0^1 \left(\nabla^2 f(x^* + t(x^k - x^*)) - \nabla^2 f(x^k) \right) (x^k - x^*)\, dt \right\|$$

$$\leq \mu \int_0^1 \left\| \left(\nabla^2 f(x^* + t(x^k - x^*)) - \nabla^2 f(x^k) \right) (x^k - x^*) \right\| dt$$

FACCIO LA NORMA

$$\leq \mu \int_0^1 \| \nabla^2 f(x^* + t(x^k - x^*)) - \nabla^2 f(x^k) \| \cdot \| x^k - x^* \|\, dt \quad \Delta$$

$$\Rightarrow \underline{x^k \in B_\varepsilon(x^*) \text{ per Hp}},\ \underline{x^* \in B_\varepsilon(x^*)},\ x^* + t(x^k - x^*) =$$

$$= t x^k + (1-t) x^*$$

DEFINIZIONE DI CONVESSITÀ

$$\Rightarrow \underline{\text{PER LA CONVESSITÀ DI } B_\varepsilon(x^*)} \Rightarrow \boxed{x^* + t(x^k - x^*) \in B_\varepsilon(x^*)}$$

POSSIAMO ALLORA DIRE CHE:

$$\Delta \le \mu \int_0^1 \frac{\sigma}{\mu} \|x^k - x^*\| dt$$

$$= \sigma \int_0^1 \|x^k - x^*\| dt = \sigma \|x^k - x^*\| \quad \text{con } \sigma \in (0,1)$$

$$\le \|x^k - x^*\|$$

$$\le \varepsilon$$

$$\Rightarrow \underline{x^{k+1} \in B_\varepsilon(x^*) \Rightarrow \{x^k\} \in B_\varepsilon(x^*)} \checkmark$$

INOLTRE $\|x^{k+1} - x^*\| \le \sigma \|x^k - x^*\|$

$$\le \sigma^2 \|x^{k-1} - x^*\|$$

$$\le \sigma^3 \|x^{k-2} - x^*\|$$

$$\vdots$$

$$\Rightarrow \boxed{\|x^{k+1} - x^*\| \leq \sigma^k \|x^0 - x^*\|}$$

OVVERO $\boxed{0 \leq \lim_{k\to\infty} \|x^{k+1} - x^*\| \leq \lim_{k\to\infty} \sigma^k \|x^0 - x^*\| = 0}$ POICHÉ $\sigma < 1$

$$\Rightarrow \boxed{\lim_{k\to\infty} \|x^k - x^*\| = 0 \Rightarrow x^k \to x^*} \quad \square$$

Dim 3

RICORDANDO CHE: $\boxed{\|x^{k+1} - x^*\| \leq \mu \int_0^1 \|\nabla^2 f(x^* + t(x^k - x^*)) - \nabla^2 f(x^k)\| \cdot \|(x^k - x^*)\| dt}$

DIVIDENDO PER $\underline{\|x^k - x^*\|}$:

$$0 \leq \frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|} \leq \mu \int_0^1 \|\nabla^2 f(x^* + t(x^k - x^*)) - \nabla^2 f(x^k)\| dt$$

E SAPENDO CHE $\underline{x^k \to x^*}$ SCRIVIAMO:

$$0 \leq \lim_{k\to\infty} \frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|} \leq \mu \int_0^1 \|\nabla^2 \cancel{f(x^*)} - \nabla^2 \cancel{f(x^*)}\| dt = 0$$

$\Rightarrow x^k$ CONVERGE IN MODO SUPERLINEARE $\square$

$\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq$
$\leq L\|x-y\|$

INOLTRE, SE $\nabla^2 f(x)$ È LIPSCHITZ-CONTINUA, ALLORA:

$$\|x^{k+1} - x^*\| \leq \mu \int_0^1 \|\nabla^2 f(x^* + t(x^k - x^*)) - \nabla^2 f(x^k)\| \cdot \|(x^k - x^*)\| dt$$

$$\leq \mu \int_0^1 L \|(x^* + t(x^k - x^*) - x^k)\| \cdot \|(x^k - x^*) dt$$

$$= L\mu \int_0^1 \|(x^k - x^*)\|^2 (1-t) dt$$

$$= L\mu \|(x^k - x^*)\|^2 \int_0^1 (1-t) dt$$

$$= \frac{1}{2} L\mu \underbrace{\|(x^k - x^*)\|^2}_{\|x^k - x^*\|^2}$$

QUINDI $\dfrac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|^2} \leq \dfrac{\mu L}{2} \Rightarrow \lim_{k\to\infty} \dfrac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|^2} \leq \dfrac{\mu L}{2}$

OVVERO x^k CONVERGE IN MODO QUADRATICO! $\square$

RIASSUNTINO

METODO	f CONVESSA		f FORTEMENTE CONVESSA	
	TASSO	COMPLESSITÀ	TASSO	COMPLESSITÀ
GRADIENTE	SUBLINEARE	$O\left(\frac{1}{\varepsilon}\right)$	LINEARE	$O\left(\log \frac{1}{\varepsilon}\right)$
NESTEROV	SUBLINEARE	$O\left(\frac{1}{\sqrt{\varepsilon}}\right)$	LINEARE	$O\left(\log \frac{1}{\varepsilon}\right)$
NEWTON	SUPERLINEARE	$O\left(\frac{1}{\sqrt{\varepsilon}}\right)$	SUBLINEARE	$O\left(\log\log \frac{1}{\varepsilon}\right)$

OSSERVAZIONI

- NEL CASO DI f CONVESSA IL TASSO È SOLO SUBLINEARE E NON SUPERLINEARE POICHÉ SI RICHIEDE $\nabla^2 f$ NON SINGOLARE, OSSIA DEFINITA-POSITIVA. SIAMO DUNQUE NEL CASO DI f FORTEMENTE CONVESSA. QUINDI LA CONVESSITÀ GARANTISCE SOLO CHE $\nabla^2 f$ SIA SEMI-DEFINITA POSITIVA.

- SE SI GUARDA SOLO IL TASSO IL METODO DI NEWTON È DOMINANTE MA VA CONSIDERATO TUTTAVIA IL COSTO DI OGNI ITERAZIONE. INFATTI ESEGUIRE UNA ITERAZIONE PER IL METODO DEL GRADIENTE COSTA $O(n)$, NEL CASO DEL METODO DI NEWTON IL COSTO È MOLTO MAGGIORE → TRADE-OFF TRA VELOCITÀ DI COMPUTAZIONE E VELOCITÀ DI ESECUZIONE.

VEDIAMO I PROBLEMI CHE POSSONO PRESENTARSI COL METODO DI NEWTON

1) SE $\nabla^2 f(x^k)$ NON È INVERTIBILE ALLORA NON È POSSIBILE ESEGUIRE UN AGGIORNAMENTO CON NEWTON. SI ESEGUE QUINDI UN PASSO DEL METODO DEL GRADIENTE

2) SE $\nabla^2 f(x^k)$ NON È DEFINITA POSITIVA ALLORA $d_k = -[\nabla^2 f(x^k)]^{-1} \nabla f(x^k)$ NON È UNA DIREZIONE DI DISCESA $\Rightarrow$ PASSO DEL GRADIENTE

3) SE $f(x^{k+1}) \geq f(x^k)$, OVVERO $\alpha_k = 1$ NON È BUONO, ALLORA SI VA AD ESEGUIRE UNA LINESEARCH

METODO DI NEWTON GLOBALMENTE CONVERGENTE

UN ALGORITMO ITERATIVO SI DICE GLOBALMENTE CONVERGENTE SE PRODUCE SEQUENZE $\{x^k\}$ CON LE SEGUENTI PROPRIETÀ:

- $\{x^k\}$ HA PUNTI DI ACCUMULAZIONE ED OGNUNO DI QUESTI È UN PUNTO STAZIONARIO
- NESSUNO DEI PUNTI DI ACCUMULAZIONE È UN MASSIMO LOCALE
- SE $\{x^k\}$ CONVERGE AD UN MINIMO LOCALE x^*, ALLORA ESISTE UN

SE ...
PASSO $\bar{k}$ TALE CHE:

$$\boxed{\forall k \geq \bar{k} \text{ si ha } x^{k+1} = x^k - [\nabla^2 f(x^k)]^{-1} \nabla f(x^k)}$$

DA UN CERTO PASSO $\bar{k}$ IN POI, IL PASSO È SEMPRE QUELLO DEL METODO DI NEWTON

METODI QUASI NEWTON

CONSIDERIAMO IL CASO QUADRATICO $\min_x \frac{1}{2} x^T Q x - c^T x$ con $\nabla f(x) = Qx - c$
$\nabla^2 f(x) = Q$

SE CALCOLIAMO $\nabla f(x) - \nabla f(y)$ OTTENIAMO:

$$\nabla f(x) - \nabla f(y) = Qx - \cancel{c} - Qy + \cancel{c} = Q(x-y) = \nabla^2 f(x)(x-y)$$

E, ALL'INTERNO DI UN ALGORITMO ITERATIVO SI AVRÀ:

$$\boxed{\nabla^2 f(x^{k+1}) - \nabla^2 f(x^k) = Q(x^{k+1} - x^k)}$$

PER GLI ALGORITMI ITERATIVI VISTI:

$x^{k+1} - x^k =$

- METODO DEL GRADIENTE: $\boxed{x^{k+1} = x^k - \alpha_k I \nabla f(x^k)}$
- METODO DI NEWTON: $\boxed{x^{k+1} = x^k - \underset{\substack{\shortparallel \\ 1}}{\alpha_k} \left[\nabla^2 f(x^k)\right]^{-1} \nabla f(x^k)}$

IN GENERALE SI HA: $\boxed{x^{k+1} = x^k - \alpha_k \underbrace{B_k}_{dp} \nabla f(x^k)}$

L'IDEA È QUINDI QUELLA DI USARE $\boxed{x^{k+1} = x^k - \alpha_k B_k^{-1} \nabla f(x^k)}$

CON $B_k \approx \nabla^2 f(x^k)$ CHE SODDISFA L'EQUAZIONE QUASI-NEWTON. ▲

SI DISTINGUONO DUE TIPI DI FORMULE DI AGGIORNAMENTO:

- AGGIORNAMENTO DIRETTO: $\boxed{x^{k+1} = x^k - \alpha_k B_k^{-1} \nabla f(x^k) \text{ con } B_k \approx \nabla^2 f(x^k)}$
- AGGIORNAMENTO INVERSO: $\boxed{x^{k+1} = x^k - \alpha_k H_k \nabla f(x^k) \text{ con } H_k \approx \left[\nabla^2 f(x^k)\right]^{-1}}$

SCRIVIAMO CHE $B_{k+1} = B_k + \underbrace{\Delta B_k}_{\text{PERTURBAZIONE}}$ E $H_{k+1} = H_k + \Delta H_k$

SIANO POI $\boxed{s_k = x^k - x^{k-1}}$ E $\boxed{y_k = \nabla f(x^k) - \nabla f(x^{k-1})}$, ALLORA:

▲ → EQUAZIONE QUASI-NEWTON

$$\boxed{\begin{array}{l} B_k s_k = y_k \\ s_k = H_k y_k \end{array}} \quad \begin{array}{l} \Rightarrow B_{k+1} s_{k+1} = y_{k+1} \\ \Rightarrow H_{k+1} y_{k+1} = s_{k+1} \end{array}$$

ALGORITMO GENERICO

DATI: $x^0 \in \mathbb{R}^n$, $k=0$, $B_0 \in \mathbb{R}^{n \times n}$ SIMMETRICA E D.P

WHILE $(\nabla f(x^k) \neq 0)$

- $d_k = -B_k^{-1} \nabla f(x^k)$
- CALCOLO α_k
- $x^{k+1} = x^k + \alpha_k d_k$
- CALCOLO B_{k+1} t.c. $\boxed{B_{k+1}(x^{k+1} - x^k) = \nabla f(x^{k+1}) - \nabla f(x^k)}$

  $\to B_{k+1} s_{k+1} = y_{k+1}$

  $\to$ DEVE SODDISFARE L'EQUAZIONE QUASI-NEWTON
- $k = k+1$

END

AGGIORNAMENTO DI RANGO 1

- VERSIONE DIRETTA: $\boxed{B_{k+1} = B_k + \rho_k \mu_k v_k^T}$ $\rho_k \in \mathbb{R}$, $\mu_k, v_k \in \mathbb{R}^n$
- VERSIONE INVERSA: $\boxed{H_{k+1} = H_k + \rho_k \mu_k v_k^T}$

AGGIORNAMENTO DI RANGO 2

- VERSIONE DIRETTA: $\boxed{B_{k+1} = B_k + a_k u_k u_k^T + b_k v_k v_k^T}$ $a_k, b_k \in \mathbb{R}$, $u_k, v_k \in \mathbb{R}^n$

- VERSIONE INVERSA: $\boxed{H_{k+1} = H_k + a_k u_k u_k^T + b_k v_k v_k^T}$

FORMULE BFGS

- VERSIONE DIRETTA: $$\boxed{B_{k+1} = B_k + \frac{y_k y_k^T}{y_k^T s_k} - \frac{B_k s_k s_k^T B_k^T}{s_k^T B_k s_k}}$$

- VERSIONE INVERSA: $$\boxed{H_{k+1} = H_k + \left(1 + \frac{y_k^T H_k y_k}{s_k^T y_k}\right) \frac{s_k s_k^T}{s_k^T y_k} - \frac{s_k y_k^T H_k + H_k y_k s_k^T}{s_k^T y_k}}$$

MOSTRIAMO CHE LA VERSIONE DIRETTA SODDISFA L'EQUAZIONE QUASI-NEWTON:

$$B_{k+1} s_k = B_k s_k + \frac{y_k \cancel{y_k^T s_k}}{\cancel{y_k^T s_k}} - \frac{B_k \cancel{s_k B_k^T s_k}}{\cancel{s_k^T B_k s_k}} = \cancel{B_k s_k} + y_k - \cancel{B_k s_k} = y_k \checkmark$$

ANALOGAMENTE VALE ANCHE PER LA FORMULA INVERSA.

SE CONSIDERIAMO $\boxed{d_k = -B_k^{-1} \nabla f(x^k)}$ OPPURE $\boxed{d_k = -H_k \nabla f(x^k)}$ SI HA CHE d_k È DIREZIONE DI DISCESA SE LE MATRICI B_k OPPURE H_k SONO DEFINITE POSITIVE

PROPOSIZIONE:

SIA B_k DEFINITA POSITIVA. SE VALE $y_k^T s_k > 0$ ALLORA B_{k+1}, OTTENUTA TRAMITE UN

AGGIORNAMENTO BFGS, È ANCORA UNA MATRICE DEFINITA POSITIVA.

INOLTRE SE $\overbrace{B_0}^{\text{MATRICE INIZIALE}}$ È DEFINITA POSITIVA E VALE $s_k^T y_k > 0$ AD OGNI ITERAZIONE, ALLORA TUTTA LA SEQUENZA $\{B_k\}$ È DEFINITA POSITIVA

MA COME È POSSIBILE GARANTIRE CHE $s_k^T y_k > 0$?

$$\boxed{s_k = x^{k+1} - x^k = x^k + \alpha_k d_k - x^k = \alpha_k d_k}$$

SAPPIAMO CHE $\alpha_k d_k^T y_k > 0 \Rightarrow \alpha_k d_k^T \big(\underbrace{\nabla f(x^{k+1}) - \nabla f(x^k)}_{y_k}\big) > 0$

$\alpha_k > 0$

$$\Rightarrow d_k^T \big(\nabla f(x^{k+1}) - \nabla f(x^k)\big) > 0$$

$$\Rightarrow \boxed{d_k^T \nabla f(x^{k+1}) > d_k^T \nabla f(x^k)}$$

PER LA CONDIZIONE DI WOLFE SI HA $\boxed{\nabla f(x^k + \alpha_k d_k)^T d_k \geq \sigma \nabla f(x^k)^T}$ CON $\sigma \in (0,1)$

TALE CONDIZIONE CI GARANTISCE CHE:

$$\boxed{\nabla f(x^k + \alpha_k d_k)^T d_k \geq \nabla f(x^k)^T d_k}$$

PROPRIETÀ DI CONVERGENZA

- SE f È CONVESSA ALLORA SI HA CONVERGENZA GLOBALE
- SE f È FORTEMENTE CONVESSA ALLORA SI HA UN TASSO DI CONVERGENZA SUPERLINEARE

PROBLEMI SU LARGA SCALA → $\min_{x \in \mathbb{R}^n} f(x)$ con $n \sim 10^4$

IN QUESTO TIPO DI PROBLEMI L'AGGIORNAMENTO BFGS PRESENTA DEI PROBLEMI:

- MEMORIZZAZIONE DI H_k POICHÉ È UNA MATRICE DENSA
- CALCOLARE $d_k = -H_k \nabla f(x^k)$

SI RISOLVE UTILIZZANDO DUE STRATEGIE:

1) SCRIVERE H_{k+1} IN MODO RICORSIVO

$$H_{k+1} = V_k^T H_k V_k + \rho_k s_k s_k^T \quad \text{con} \quad \boxed{\rho_k = \frac{1}{y_k^T s_k}} \quad \text{e} \quad \boxed{V_k = I - \rho_k y_k s_k^T}$$

$$= V_k^T \left(V_{k-1}^T H_{k-1} V_{k-1} + \frac{1}{y_{k-1}^T s_{k-1}} s_{k-1} s_{k-1}^T \right) V_k + \frac{1}{y_k^T s_k} s_k^T s_k \dots$$

$$\vdots$$

$$= \mathcal{H}(H_0, y_0, s_0, y_1, s_1, \dots, y_k, s_k)$$

QUESTO SIGNIFICA CHE H_{k+1} SARÀ FUNZIONE $(\mathcal{H})$ DI H_0 E DELLE SEQUENZE $\{y_k\}$ E $\{s_k\}$

SE PONGO $H_0 = \gamma I$ POSSO RICOSTRUIRE H_{k+1} PARTENDO DA UNO SCALARE E 2K VETTORI IN $\mathbb{R}^n$!

2) TRONCARE IL PROCESSO RICORSIVO

FERMANDO LA RICORSIONE OTTERRÒ QUINDI:

$$\boxed{H_{k+1} = \mathcal{H}(H_0, y_0, s_0, \ldots, y_k, s_k) = \mathcal{H}(\underbrace{H_{k-m}}_{H_0 = \gamma I}, y_{k-m}, s_{k-m}, \ldots, y_k, s_k)}$$

QUESTO CI PERMETTE, AD OGNI ITERAZIONE, DI AVERE IN MEMORIA SOLTANTO $2mn$ VALORI INVECE CHE n^2 VALORI.
SPERIMENTALMENTE SI OSSERVA ANCHE CHE CON $m \sim 10$ SI RIESCE COMUNQUE AD OTTENERE UN'OTTIMA APPROSSIMAZIONE DELLA MATRICE HESSIANA.

ESISTE UN ALGORITMO (METODO HG) CHE CONSENTE, TRAMITE $4mn$ PRODOTTI, DI SVOLGERE IN MODO "IMPLICITO" IL PRODOTTO $H_k \nabla f(x^k)$ USANDO SOLO I VALORI DI $\gamma, s_{k-m}, y_{k-m}, \ldots, s_k, y_k$

LA VARIANTE DI BFGS CHE USA QUESTO METODO È DETTA L-BFGS (L → LIMITED MEMORY) CHE È L'ALGORITMO DI OTTIMIZZAZIONE NON LINEARE E NON VINCOLATA PIÙ PERFORMANTE ANCHE SE NON HA CONVERGENZA GLOBALE!