Optimization Report

Filippo Zaccari

June 26, 2025

1 Obiettivo del progetto

Partendo dall'articolo "Nonmonotone Line Searches Operate at the Edge of Stability" [CF24] l'obiettivo del progetto è il seguente: implementare le classi e le funzioni per caricare dataset e implementare una regressione logistica con regolarizzatore ℓ_2 che calcoli la loss e il gradiente della loss.

Inoltre verranno implementati i seguenti algoritmi:

- Gradient descend con line search di Armijo;
- Gradient descend con line search non monotona;
- Per entrambi gli algoritmi verrano implementate le seguenti strategie di scelta del passo iniziale:
  - Scelta di Polyak;
  - Scelta del passo iniziale tramite euristica. (2.3)

Il codice prodotto può essere trovato nel seguente link: https://github.com/Filo00/optimization_project

Il codice è stato scritto in linguaggio python con l'utilizzo della libreria numpy. I dataset utilizzati sono a4a, a6a, a8a [Ron], la libreria utilizzata per caricare i dataset è sklearn. I grafici generati risultano essere molto simili tra i vari dataset, quindi per migliore leggibilità della relazione sono stati inseriti soltanto i grafici relativi al dataset a4a, il resto dei grafici può essere trovato all'interno della cartella plot della repository git. I valori degli iperparametri utilizzati per gli esperimenti sono:

- Regolarizzatore $\lambda = 0.01$
- Tolleranza $tol = 1 \times 10^{-6}$
- Numero massimo di iterazioni $max\_iter = 10000$

2 Risultati ottenuti

Nella sezione risultati andremo a vedere la stampa dei grafici generati dall'esecuzione degli algoritmi, in particolare per ogni metodo mostreremo l'andamento della funzione loss, la stima del valore LAPPROX calcolato tramite la formula mostrata in [CF24] e la sharpness calcolata come $\lambda_{\max}\left(\nabla^2 f(w)\right)$

2.1 Loss

La funzione utilizzata è la log-loss

$$\ell(w) = \frac{1}{n}\sum_{i=1}^{n} \log\left(1 + e^{-y_i \cdot x_i^\top w}\right) + \frac{\lambda}{2}\|w\|^2 \tag{1}$$

con il rispettivo gradiente

$$\nabla\ell(w) = -\frac{1}{n}\sum_{i=1}^{n} y_i x_i \left(1 - \frac{1}{1 + e^{-y_i x_i^\top w}}\right) + \lambda w \tag{2}$$

ed Hessiana

$$\nabla^2 \ell(w) = \frac{1}{n} X^\top D X + \lambda I \tag{3}$$

dove D è una matrice diagonale con elementi

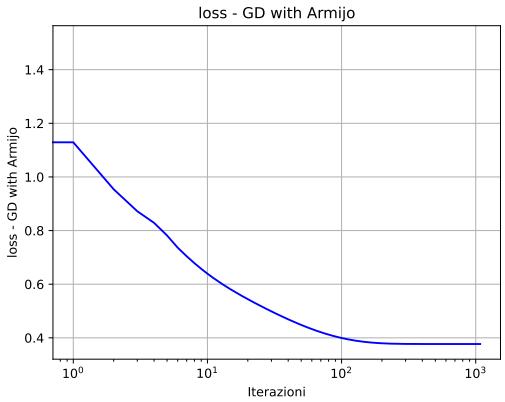
$$D_{ii} = \sigma(z_i)(1 - \sigma(z_i)), \quad \text{con } z_i = x_i^\top w, \quad \sigma(z) = \frac{1}{1 + e^{-z}} \tag{4}$$

2.2 Armijo

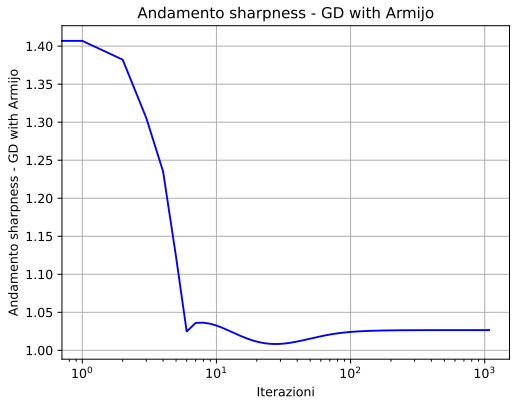
[Criterio di Armijo] Sia $f : R^n \to R$ una funzione differenziabile. Dato un punto $x_k \in R^n$ e una direzione di discesa $d_k \in R^n$ tale che $\nabla f(x_k)^\top d_k < 0$, si cerca un passo $\alpha_k > 0$ che soddisfi la condizione:

$$f(x_k + \alpha_k d_k) \leq f(x_k) + \gamma \alpha_k \nabla f(x_k)^\top d_k$$

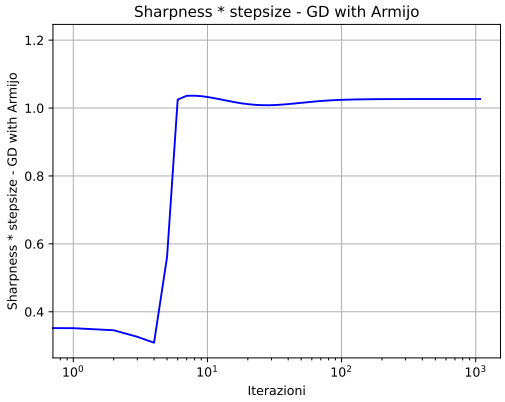
con $\gamma \in (0, 1)$.
I parametri in questo caso sono: δ=0.5, $\gamma = 1 \times 10^{-4}$. Nel caso di Armijo la loss scende in modo monotono, con un valore di LAPPROX che approssima discretamente la sharpness che risulta essere un valore basso con conseguente buona generalizzazione del metodo.



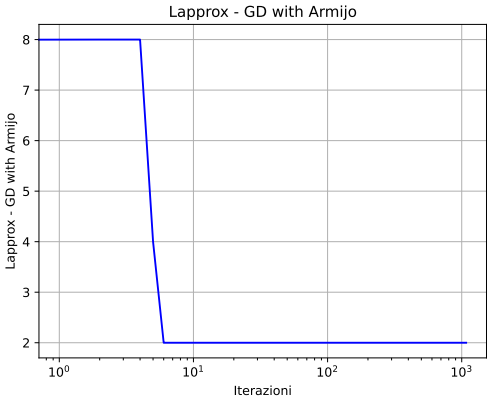
(a) Loss



(b) Sharpness



(c) Sharpness * stepsize



(d) LAPPROX

Figure 1: Grafici Armijo

2.3 Armijo - noTune

Scelta euristica del passo iniziale. Nel caso noTune la scelta del passo iniziale viene effettuata tramite la formula:

$$\alpha_{\text{iniz}} = \begin{cases} \min\left(\frac{\alpha}{\delta}, \alpha_{\max}\right), & \text{se } n = 0 \\ \alpha, & \text{se } n > 0 \end{cases}$$

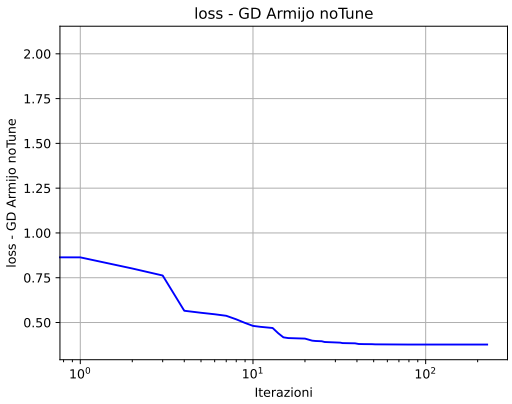
dove:

- $\delta \sim \mathcal{N}(0.5, 1)$ è una variabile casuale con distribuzione normale, troncata nell'intervallo $[0.1, 0.9]$;
- $\alpha_{\max} > 0$ è un valore massimo prefissato per evitare passi troppo grandi altrimenti potrebbe dare overflow, il valore massimo stabilito per il passo è $\alpha = 20$;
- n è il numero di backtrack effettuati.

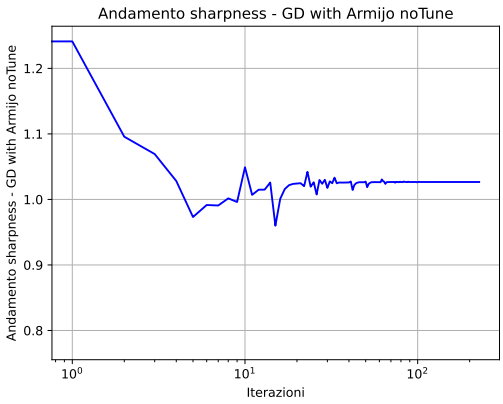
I parametri utilizzati sono δ=0.5, $\gamma = 1 \times 10^{-4}$

Questa scelta permette di introdurre una maggiore aggressività nella prima iterazione, quando non sono ancora stati osservati fallimenti nel soddisfare le condizioni di discesa. In presenza di uno o più tentativi di backtracking ($n > 0$), il passo viene mantenuto costante, favorendo la stabilità del metodo.

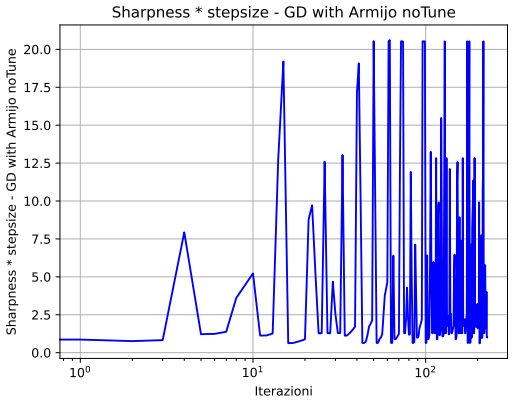
Con questa scelta del passo iniziale la loss parte da un valore più basso rispetto ad Armijo classico e scende più velocemente. In questo caso LAPPROX è molto instabile e non approssima perfettamente la sharpness.



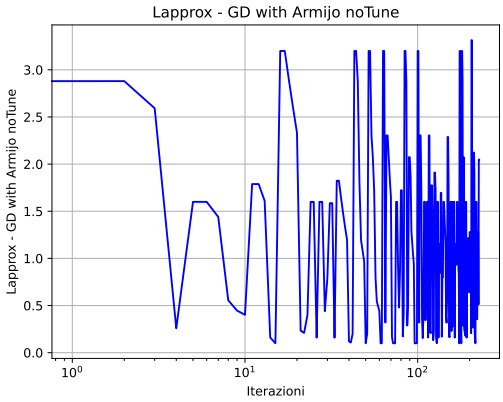
(a) Loss



(b) Sharpness



(c) Sharpness * stepsize



(d) LAPPROX

Figure 2: Grafici Armijo - noTune

2.4 Armijo - Polyak

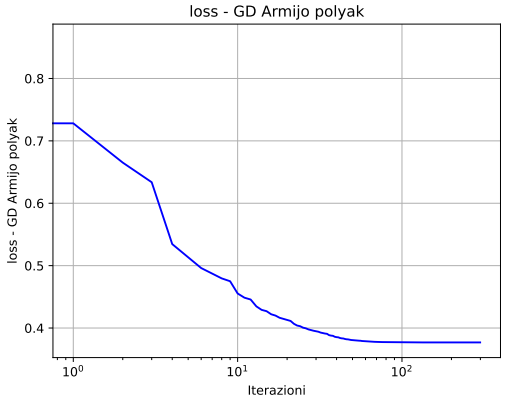
Scelta del passo secondo la regola di Polyak. Un'altra tecnica utilizzata per la selezione del passo α è la *regola di Polyak*, che tiene conto del valore corrente della funzione obiettivo e della norma del gradiente. La regola è data da:

$$\alpha = \min\left(\frac{f(x) - f_{\min}}{\|\nabla f(x)\|^2}, \alpha_{\max}\right)$$

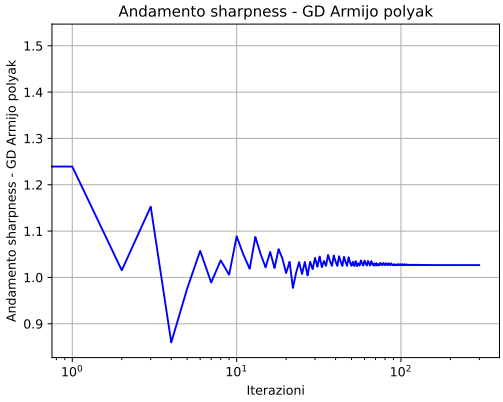
dove:

- $f(x)$ è il valore della funzione obiettivo nel punto x;
- $\nabla f(x)$ è il gradiente della funzione in x;
- $f_{\min}$ è una stima del minimo valore raggiungibile dalla funzione;
- $\alpha_{\max} > 0$ è il limite superiore per il passo, con $\alpha_{\max} = 20$

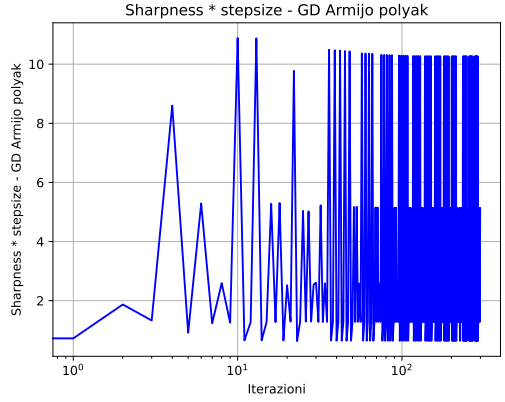
Questa scelta del passo ha il vantaggio di essere adattiva e proporzionale alla distanza dall'ottimo. Quando $f(x)$ si avvicina a $f_{\min}$, il numeratore si riduce e il passo si contrae automaticamente, facilitando la convergenza. In casi in cui la norma del gradiente è così piccola da dare instabilità numerica e risultare come nulla si impone un passo molto piccolo positivo, in questo caso $\alpha = 10^{-6}$, per evitare divisioni per zero.



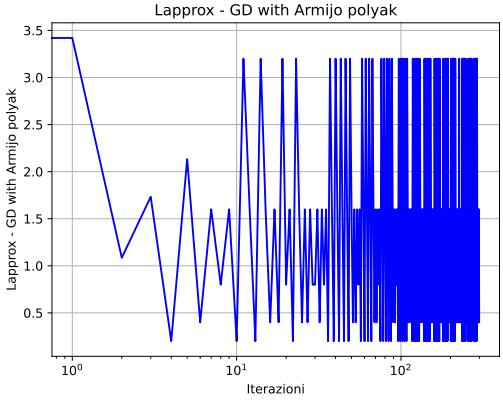
(a) Loss



(b) Sharpness



(c) Sharpness * stepsize



(d) LAPPROX

Figure 3: Grafici Armijo - Polyak

2.5 Nonmonotone

Line search non monotona. Sia $f : R^n \to R$ una funzione obiettivo differenziabile, e $d \in R^n$ una direzione di discesa. Il passo α viene scelto come il primo valore che soddisfa la seguente condizione:

$$f(x + \alpha d) \leq C_k + \gamma \alpha \nabla f(x)^\top d$$

dove:

- $\gamma \in (0, 1)$
- $C_k = \max\{C_k,\, f(x)\}$
- $\nabla f(x)^\top d < 0$, con d una direzione di discesa.

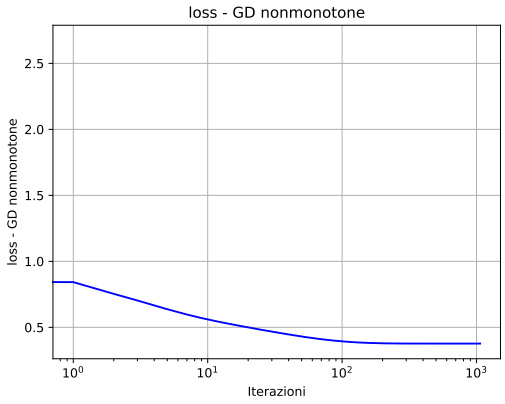
Il valore C_k viene aggiornato ricorsivamente con due variabili ausiliarie Q_k e $\xi \in (0, 1)$ come segue:

$$Q_{k+1} = \xi Q_k + 1$$
$$\widetilde{C}_k = \frac{\xi Q_k C_k + f(x)}{Q_{k+1}}, \quad C_{k+1} = \max\left\{\widetilde{C}_k,\, f(x)\right\}$$

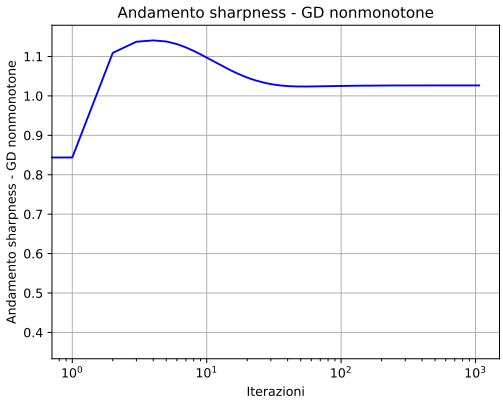
La procedura riduce iterativamente α tramite backtrack:

$$\alpha \leftarrow \delta \cdot \alpha, \quad \text{con } \delta \in (0, 1)$$

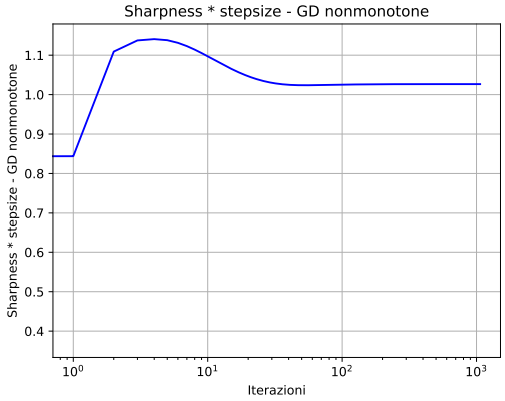
fino a quando la condizione sopra è soddisfatta. Questa strategia consente temporanei incrementi nella funzione obiettivo, evitando un comportamento troppo conservativo e migliorando la convergenza in presenza di irregolarità.



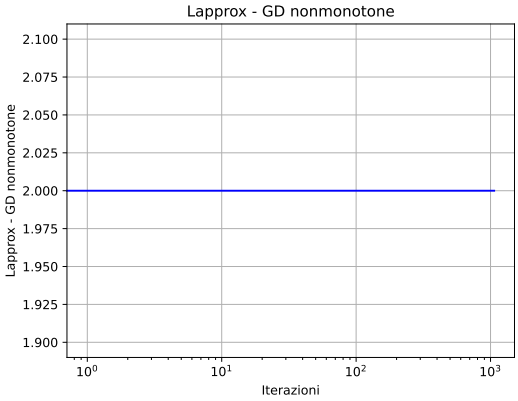
(a) Loss



(b) Sharpness



(c) Sharpness * stepsize

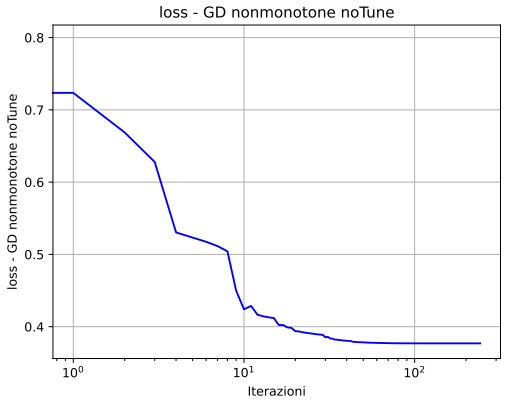


(d) LAPPROX

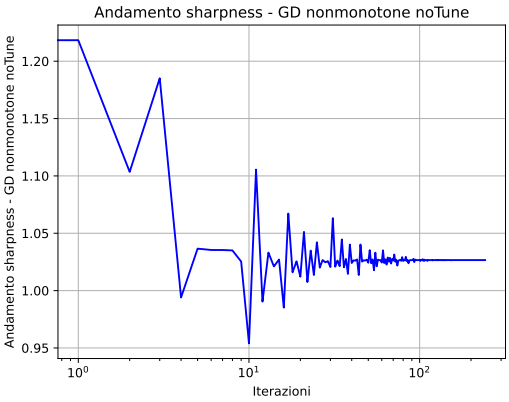
Figure 4: Grafici Nonmonotone

2.6 Nonmonotone - noTune

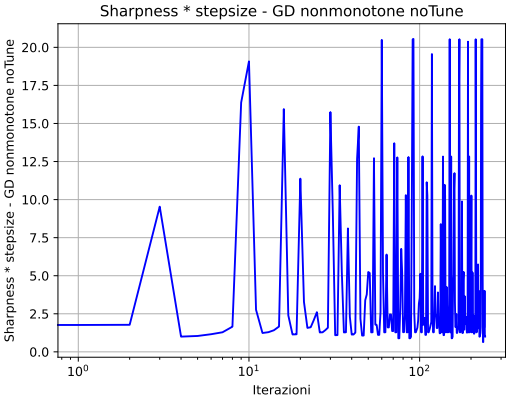
Line search non monotona con passo iniziale euristico. Il seguente metodo combina una strategia di *line search non monotona* con una scelta del *passo iniziale* basata sull'euristica precedentemente descritta(2.3).



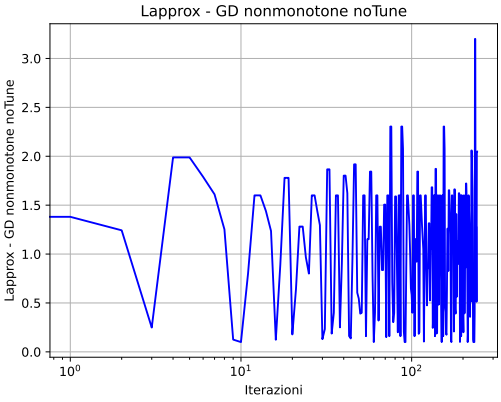
(a) Loss



(b) Sharpness



(c) Sharpness * stepsize

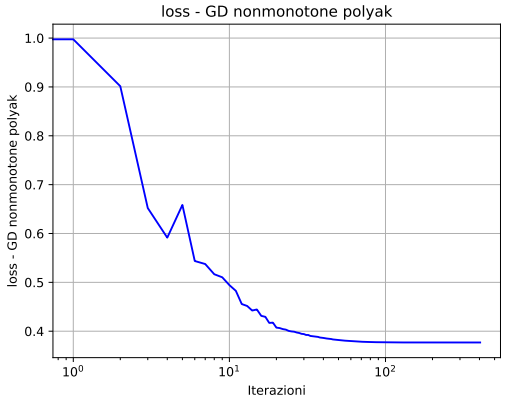


(d) LAPPROX

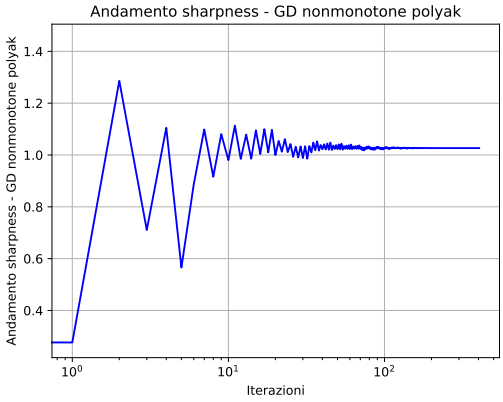
Figure 5: Grafici Nonmonotone - noTune

2.7 Nonmonotone - Polyak

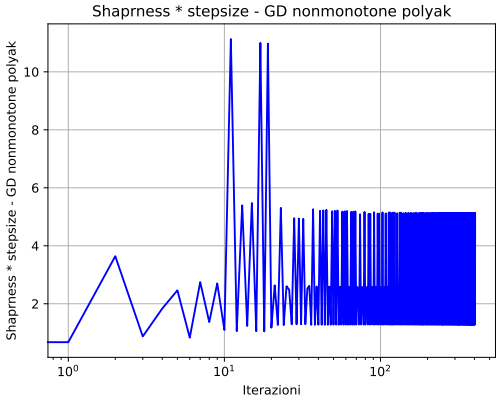
Line search non monotona con passo iniziale secondo Polyak. In questa variante, il metodo nonmonotono viene combinato con la scelta del passo iniziale α secondo la *regola di Polyak* (2.4).



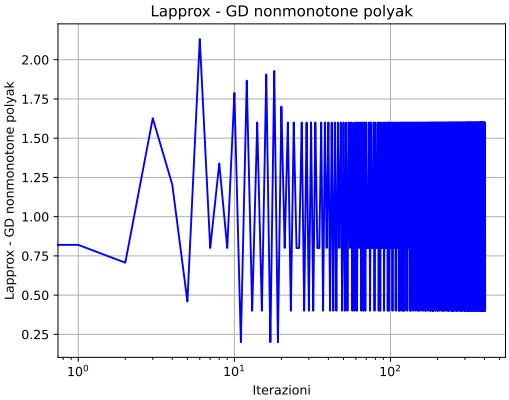
(a) Loss



(b) Sharpness



(c) Sharpness * stepsize



(d) LAPPROX

Figure 6: Grafici Nonmonotone - Polyak

3 Conclusioni

Sulla base dei risultati ottenuti e dei grafici generati possiamo dire che tutti i metodi portano la funzione di loss ad assumere valori che si aggirano intorno a 0.4.

La loss con il metodo non monotono parte più in basso ma ha un andamento più rilassato, Armijo parte più in alto ma scende leggermente più veloce del metodo non monotono. Le varianti noTune e Polyak risultano essere molto simili tra di loro.

La differenza grossa nella loss sembra farla la scelta del passo iniziale più che il metodo stesso, infatti i metodi base sembrano essere leggermente più lenti rispetto alle controparti che implementano una scelta del passo iniziale diversa da quella classica. LAPPROX approssima meglio con il metodo di Armijo e il metodo non monotono. Le versioni noTune e Polyak del metodo non monotono oscillano molto ma complessivamente approssimano discretamente il valore reale. Il metodo di Armijo nelle versioni modificate oscilla molto di più rispetto alla controparte non monotona, approssimando in modo meno preciso il valore della Sharpness. La sharpness reale oscilla leggermente nei metodi con passo iniziale modificato, ma in generale ottiene dei buoni valori in ogni metodo. La metrica $Sharpness \cdot stepsize$ indica quanto il metodo è vicino al'edge of stability, con soglia di edge of stability = 2.

Armijo è ben sotto al valore 2, quindi il metodo è stabile e non rischia di divergere, le sue versioni noTune e Polyak invece oscillano molto tra 1 e 20, la stessa cosa avviene per il metodo non monotono e le sue varianti.

References

[CF24] Mark Schmidt Holger Rauhut Curtis Fox, Leonardo Galli. Nonmonotone line searches operate at the edge of stability. 2024.

[Ron] Rong-En Fan at National Taiwan University. Libsvm - dataset. https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/.