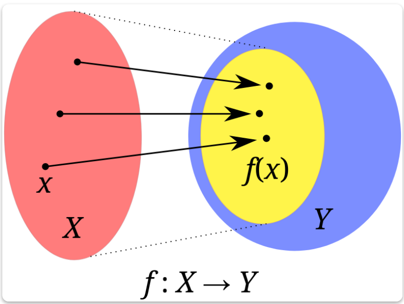
Lezione04

Table of contents

Funzioni reali di più variabili reali

Dominio e codominio



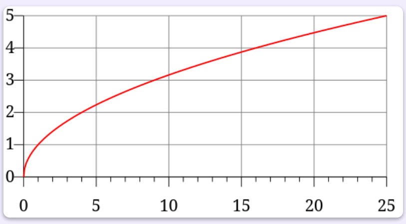
Dati $n, m \in \mathbb{N}_+$, considerando dominio $D \subseteq \mathbb{R}^n$ e codominio $C \subseteq \mathbb{R}^m$, la funzione

$$f : D \to C$$

è una funzione vettoriale con m componenti e n variabili reali.

Il dominio di $\sqrt{x}$

Se non chiaro, visualizziamo il dominio, ragionando nell'insieme dei reali $\mathbb{R}^n$, come l'insieme dei valori che una funzione $f(x)$ può assumere. La funzione $f(x) = \sqrt{x}$ ha come dominio, tutti i numeri reali non negativi.

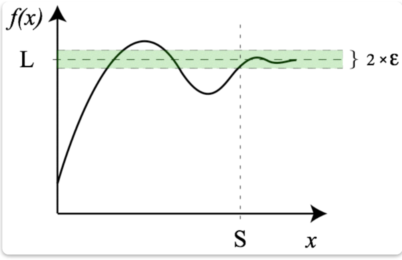


Viene prodotto un vettore come risultato, il che significa che possiamo riscrivere

$$f(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_m(\mathbf{x}))$$

Ci concentreremo sullo studio di più funzioni reali che produrranno una variabile reale: i vettori li potremo spezzare e studiare per ogni singola componente.

Limite



Per lo studio delle funzioni, riprendiamo il concetto di limite di funzione.
Considerando una funzione f su sotto dominio di $\mathbb{R}^n$, e il dominio aperto $D \subseteq \mathbb{R}^n$, diremo che possiamo avvicinarci a un punto contenuto in questa palla da qualsiasi direzione, scrivendo il limite come

$$\lim_{\mathbf{x} \to \mathbf{w}} f(\mathbf{x}) = l$$

e che esiste, se e soltanto se esiste la ϵ delta ("*epsilon delta*")

$$\forall \epsilon \in \mathbb{R}_+ \quad \exists \delta \in \mathbb{R}_+ : \forall \mathbf{x} \in D \quad ||\mathbf{x} - \mathbf{w}|| < \delta \implies |f(\mathbf{x}) - l| < \epsilon$$

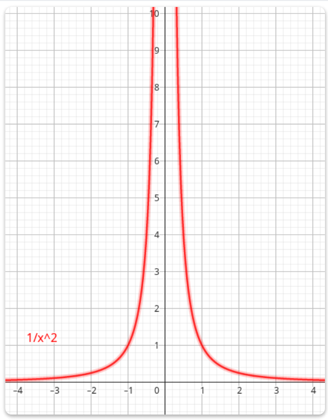
L'affermazione dice: se esiste il limite esiste una tolleranza positiva, che chiamiamo ϵ, la quale può essere contenuta da una palla di raggio D. Possiamo guardare l'immagine sopra per intenderci meglio: notare come la funzione $f(x)$ tenda a questa differenza assoluta, che è a sua volta $< \epsilon$ (la nomenclatura è diversa, ma rende comunque l'idea).

La funzione f ha come limite l per $\mathbf{x}$ che tende a infinito

$$\lim_{\mathbf{x} \to \infty} (\mathbf{x}) = l$$

se allontanandoci sempre di più da m, ci avviciniamo sempre di più a ϵ

$$\forall \epsilon \in \mathbb{R}_+ \quad \exists m \in \mathbb{R}_+ : \forall \mathbf{x} \in D \quad ||\mathbf{x}|| > m \implies |f(\mathbf{x}) - l| < \epsilon$$



La funzione $1/x^2$ tende a ∞

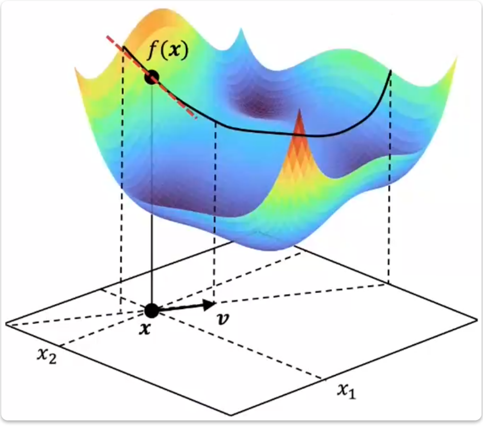
Questa proposizione ci servirà per definire gli integrali impropri e la continuità della funzione; lo studio asimptotico quindi ci serve: il processo d'addestramento che andiamo a fare su una rete neurale, ha lo scopo di minimizzare l'errore sull'approssimazione. Siccome quello che andremo a fare è minimizzare l'errore, capiremo se esiste per davvero questo minimo e se varrà la pena di cercarlo.

Derivata direzionale

Consideriamo la solita funzione f definita per un aperto $D \subseteq \mathbb{R}^n$, consideriamo inoltre un punto che appartenga al dominio $\mathbf{x} \in D$, insieme a un versore $\mathbf{v} \in \mathbb{R}^n$ che ci indica una direzione. Ci poniamo il problema di capire se esiste il limite per h che tende a 0, della quantità che segue in rapporto

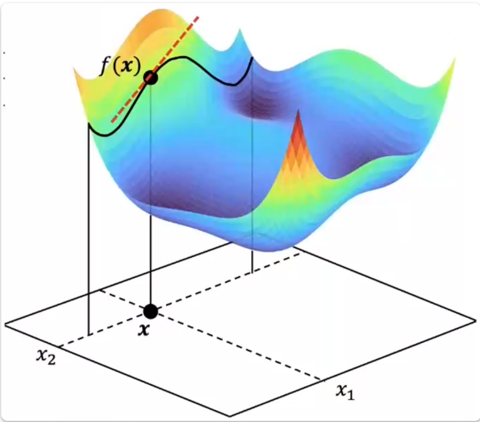
$$\partial_{\mathbf{v}} f(\mathbf{x}) \triangleq \lim_{h \to 0} \frac{f(\mathbf{x} + h\mathbf{v}) - f(\mathbf{x})}{h}$$

Osservando il limite, ci viene da dire che in effetti $\mathbf{x}$ esiste siccome appartenente al dominio; esiste quindi anche il limite di $\mathbf{x}$ che punta nella direzione $\mathbf{v}$? Fintanto che h assuma valori sufficientemente piccoli, allora lo possiamo confermare. Questo limite viene detto **derivata direzionale nella direzione $\mathbf{v}$ della funzione f nel punto $\mathbf{x}$**.



(Osserviamo l'immagine) Abbiamo una funzione di due variabili x_1 e x_2. Fissiamo un punto x su cui la funzione è definita. $f(x)$ è la quota rappresentata sulla superficie. Scegliendo una direzione v, che rimarrà sempre la stessa, ci avviciniamo lungo questa direzione, al punto x, calcolando il rapporto incrementale che nel caso abbia limite (tendono a qualche cosa), confermerà l'esistenza di una derivata direzionale.
Quello che andiamo a ottenere, è in realtà una linea, che corre sulla superficie e si avvicina a due punti: il coefficiente angolare di questa retta è la derivata in quel punto.

Derivata parziale



Le derivate direzionali ci interessano in un caso particolare: se la direzione è parallela a uno degli assi. Anziché prendere un v che è inclinato, ne prendiamo uno parallelo a x_1 oppure uno parallelo a x_2. Nell'insieme $\mathbb{R}^n$ ne esisteranno un numero infinito n di questi *versori ortonormali* (lunghezza 1 e paralleli agli assi).
Per esempio: se prendessimo un punto x lungo il piano, e girassimo il versore v per renderlo parallelo a x_1, allora tutte le componenti del versore saranno a 0 tranne che per la componente la quale descrive la posizione rispetto appunto x_1.

$$\mathbf{v} = (0, 0, \dots, 1, 0, 0, \dots, 0)$$

Se utilizziamo uno di questi per calcolare la nostra derivata, essa prenderà il nome di **derivata parziale**, scritta come

$$\frac{\partial f}{\partial x_i}(\mathbf{x}) \quad \texttt{oppure} \quad \partial_i f(\mathbf{x})$$

Siccome il nostro punto prende ora in considerazione una sola variabile (l'asse che rimane fisso), il calcolo che andremo a eseguire prenderà solo in considerazione questo, mentre le altre variabili saranno costanti e non ci interesseranno (fissate a 0).

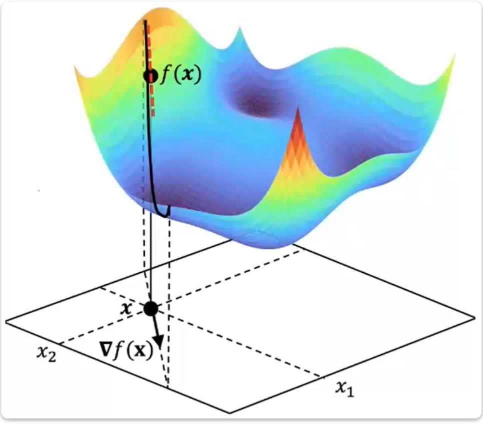
☰ Derivata parziale di una funzione, calcolo del gradiente

$$f(x, y, z) = 3x^2 + 5xy - 7z^3$$

allora la derivata parziale lungo x, y, z usa le normali regole di derivazione

$$\nabla f(x, y, z) = (6x + 5y, 5x, -21z^2)$$

Gradiente della funzione



Se esiste il vettore formato dalle n derivate parziali di una funzione in $\mathbf{x}$, questo viene detto gradiente della funzione e si indica con

$$\nabla f(\mathbf{x}) \triangleq \left(\frac{\partial f}{\partial x_1}(\mathbf{x}), \frac{\partial f}{\partial x_2}(\mathbf{x}), \dots, \frac{\partial f}{\partial x_n}(\mathbf{x}) \right)$$

(Osserviamo l'immagine) Prendiamo la nostra funzione f con tutte le sue depressioni, prendiamo un punto x in cui la funzione esiste, calcoliamo le due derivate parziali una lungo l'asse x_1 e l'altra lungo l'asse x_2, quello che otterremo sarà un vettore $\nabla f(\mathbf{x})$ che prende il nome di gradiente.

Notiamo che:

- se la funzione è esplicitata in formula, questo gradiente sarà semplice da calcolare;
- il nostro gradiente conterrà tutte le informazioni, sull'andamento locale della funzione in prossimità del punto che abbiamo scelto;
- non ha importanza quali derivate vengano calcolate per prime, in quanto se esistenti, allora non cambieranno il risultato finale.

Derivare parzialmente prima una variabile piuttosto che un'altra, non cambia il senso del gradiente, che rimarrà lo stesso

$$f(x, y, z) = 3x^2 + 5xy - 7z^2$$

allora

$$\frac{\partial^2 f}{\partial x \partial y}(x, y, z) = \frac{\partial^2 f}{\partial y \partial x}(x, y, z) = 5$$

Nota la funzione e le sue derivate parziali, ci è possibile calcolare la derivata lungo una direzione arbitraria. Non è più necessario calcolare il limite del rapporto incrementale o girare gli assi, ci basta fare il prodotto scalare del versore normale con il gradiente

$$\partial_{\mathbf{v}} f(\mathbf{x}) = \mathbf{v} \cdot \nabla f(\mathbf{x})$$

Usando la disuguaglianza di Cauchy-Schwarz, possiamo dire che

$$-||\nabla f(\mathbf{x})|| \leq \partial_{\mathbf{v}} f(\mathbf{x}) \leq ||\nabla f(\mathbf{x})||$$

Detto in modo semplice, il gradiente ci dice la direzione da intraprendere per salire o scendere, il più rapidamente possibile: se ci muoviamo lungo la direzione del gradiente in salita, stiamo crescendo il più velocemente possibile $||\nabla f(\mathbf{x})||$, se guardiamo verso il basso, stiamo scendendo il più rapidamente possibile $-||\nabla f(\mathbf{x})||$.
L'uguaglianza ha senso, se esiste un $\lambda \in \mathbb{R}$ tale che

$$\mathbf{v} = \lambda \nabla f(\mathbf{x})$$

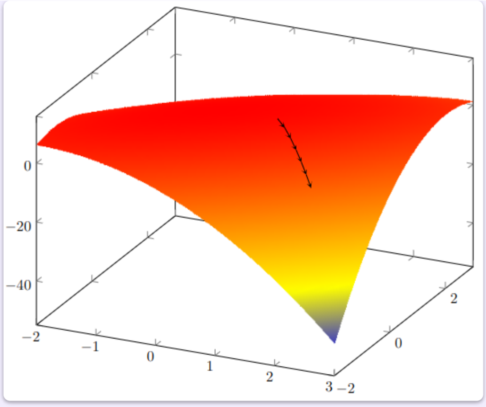
Se il nostro gradiente vale $\nabla f(\mathbf{x}) = 0$, significa che la nostra funzione è stazionaria nel punto identificato da $\mathbf{x}$ (vedere le depressioni in blu scuro dei grafici visti sopra), che prende nome di punto stazionario.
Possono esserci più punti stazionari per una sola funzione e sta a noi capire se questi sono minimo, massimo o un punto di sella, di qui vedremo alcuni esempi fra poco.

> ⚠ **Metodo di Hessian**
>
> Esiste un metodo per la valutazione dei punti stazionari di una funzione, che prende il nome di matrice di Hessian, o semplicemente, Hessian.
>
> $$(\mathbf{H}_f)_{i,j} = \frac{\partial^2 f}{\partial x_i \partial x_j}$$
>
> Il motivo per cui non la usiamo è perché questo metodo non è conclusivo: non è decisionale, perché
>
> - esistono dei casi in qui questo metodo non ci dice nulla sul punto, se sia minimo, massimo o nessuno di questi;
> - finché fatto in $\mathbb{R}^2$ ci è sufficiente guardare il determinante della matrice, ma quando saliamo di una dimensione nello spazio in $\mathbb{R}^3$, non ci basta più.

Massimi e minimi

Se abbiamo una funzione reale f di $n \in \mathbb{N}_+$ variabili reali, definita almeno in un aperto $D \subseteq \mathbb{R}^n$, se il gradiente di f è definito su tutto D allora f può ammettere minimi locali e massimi locali nei punti in cui f è stazionaria in D.

> **Calcolo degli stazionari di una funzione 1**
>
> Cerchiamo il minimo locale della funzione $f : \mathbb{R}^2 \to \mathbb{R}$ per un aperto
>
> $$f(x, y) = -2x^2 - 3y^2 + 4xy + 10$$
>
> 1. calcoliamo il gradiente
>
> $$\nabla f(x, y) = (4y - 4x, 4x - 6y)$$
>
> 2. troviamo il punto stazionario risolvendo il sistema lineare di 2 equazioni
>
> $$f(x, y) = \begin{cases} 4y - 4x = 0 \\ 4x - 6y = 0 \end{cases} \begin{cases} y = x \\ x = 0 \end{cases} \quad \to \quad f(0, 0) = 10$$



Siccome la nostra funzione ha un gradiente, essa si annullerà in un solo punto e quindi il nostro punto è un punto globale

$$f(x,y) = -2(x-y)^2 - y^2 + 10$$

per qualsiasi copia (x,y), $f(0,0) = 10$ ovvero massimo globale

Calcolo degli stazionari di una funzione 2

Consideriamo la funzione definita in $f : D \to \mathbb{R}$

$$f(x,y) = 2x^2 + 5y^2 + 4$$

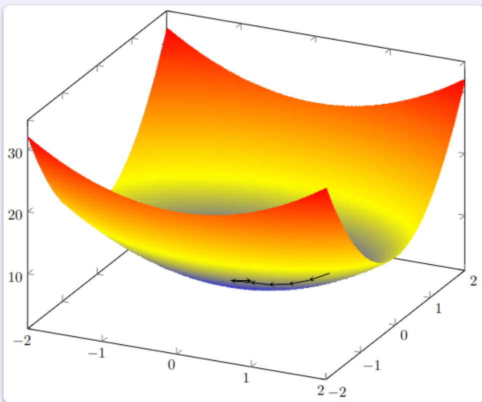
per una palla chiusa $D = \mathcal{B}_2[1,0]$ ($r = 2$, centro $[1,0]$).
Ricordiamo che, per un insieme compatto (chiuso e limitato), se la funzione è continua questa ammette almeno un minimo globale e un massimo globale.

1. calcoliamo il gradiente

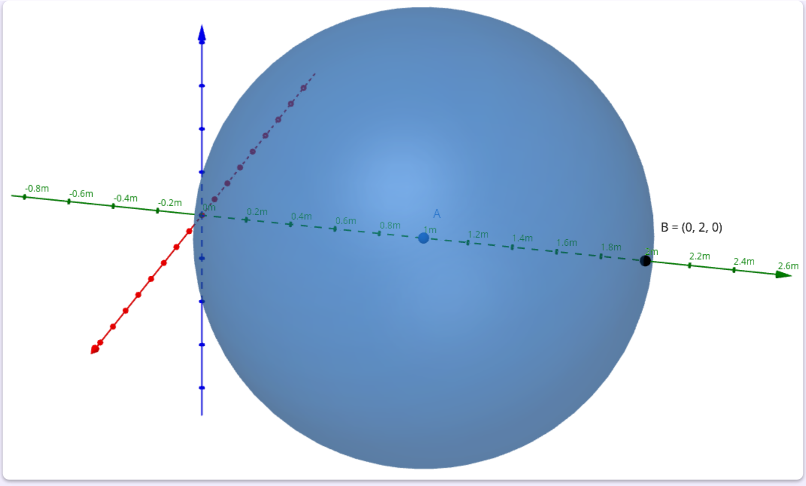
$$\nabla f(x,y) = (4x, 10y)$$

2. troviamo il punto stazionario risolvendo il sistema di 2 equazioni lineari

$$f(x,y) = \begin{cases} 4x = 0 \\ 10y = 0 \end{cases} \begin{cases} x = 0 \\ y = 0 \end{cases} \quad \to \quad f(0,0) = 4$$



Attenzione al fatto che il punto appena trovato non è globale, perché ci basta prendere un altro punto nel dominio, come $f(1,0) = 6$ per cambiare appunto il risultato (vedere il punto nero B nell'immagine sotto).



Ragionando sempre sulla sfera, siccome sappiamo che a prescindere esistono punti globali, questi non possono fare altro che esistere sulla frontiera (o bordo) della sfera

$$\partial D = \{(x, y) : (x-1)^2 + y^2 = 4\}$$

che in funzione di x viene espressa come

$$y^2 = 4 - (x-1)^2$$

che per la funzione $g(x)$ dipendente solo da x, sostituendo ci fa ottenere

$$g(x) = 2x^2 + 5 \cdot [4 - (x-1)^2] + 4 = -3x^2 + 10x + 19$$

Il massimo globale si troverà nel punto di soluzione della nostra parabola con concavità verso il basso

$$P(\frac{5}{3}, \frac{4\sqrt{2}}{3})$$

Algoritmo di discesa del gradiente

L'algoritmo di discesa del gradiente, va alla ricerca di un minimo locale per una funzione definita in tutto $\mathbb{R}^n$, che ammette gradiente nello stesso spazio.
L'algoritmo funziona in modo "greedy", spostandosi sempre nella direzione puntata dal gradiente.

```
function gradient_descent(f, x, α)
    g ← ∇f(x)
    while g ≠ 0 do
        v ← −g/||g||
        x ← x + αv
        g ← ∇f(x)
    end while
    return x
end function
```

- f è la funzione per la quale vogliamo cercare un minimo locale;
- $\mathbf{x}$ è il punto di partenza;
- α la quantità che detta la velocità di discesa;
- $\mathbf{g}$ è il gradiente calcolato per la funzione f;
- $\mathbf{v}$ la direzione.

Calcoliamo il gradiente di f chiamandolo $\mathbf{g}$.

- Se $\mathbf{g} = 0$ siamo arrivati a un punto stazionario e l'algoritmo si ferma, dicendoci che quello trovato è un punto e non procediamo (l'algoritmo non sa nulla di dove siamo, sa solo che ha trovato qualcosa)
- Se $\mathbf{g} \neq 0$ allora del gradiente, quello che a noi interessa, è la direzione contraria a quella di crescita, siccome stiamo cercando i minimi.
    - Prendiamo il gradiente $\mathbf{g}$ e lo dividiamo per la sua norma $||\mathbf{g}||$, ottenendo un versore $\mathbf{v}$ (vettore a norma 1) che cambieremo di segno per rispecchiare il ragionamento appena fatto
    - Prendiamo $\mathbf{x}$ dalla quale siamo partiti e procediamo nella direzione $\mathbf{v}$ per una quantità α
    - Ricalcoliamo il gradiente $\mathbf{g}$

Quello che otteniamo, sarà alla fine almeno un minimo locale $\mathbf{x}$.

L'algoritmo non è infallibile:

- non è detto che termini, perché se prendiamo certe funzioni, tipo un paraboloide, questo comincerà a scendere senza però fermarsi mai;
- non è detto che restituisca minimo o massimo, come nel caso di una sella;
- ha quantità di problemi numerici non banale, siccome dividendo, accumuleremo errore di calcolo a causa della piccolezza del numero a cui cercheremo di avvicinarci.

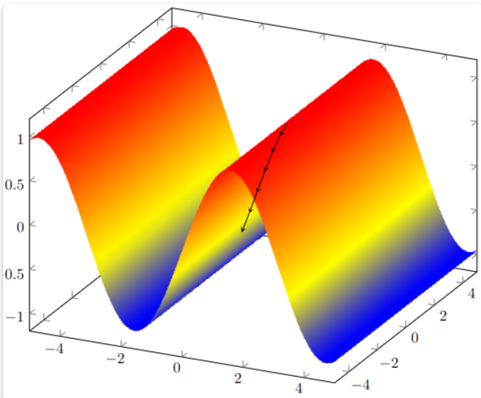
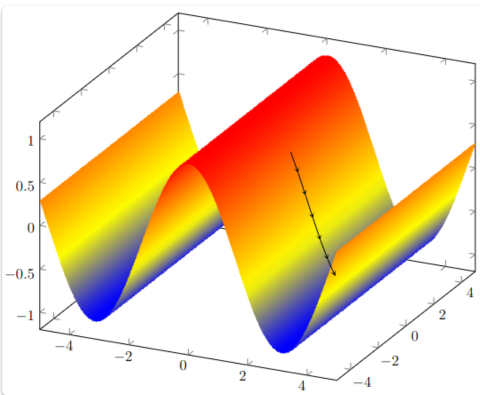
L'algoritmo è tuttavia semplicissimo e lo useremo per addestrare le nostre reti neurali.

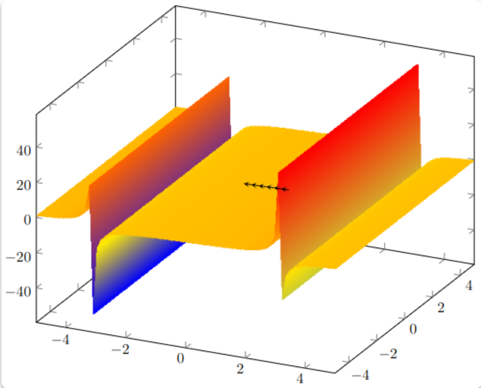
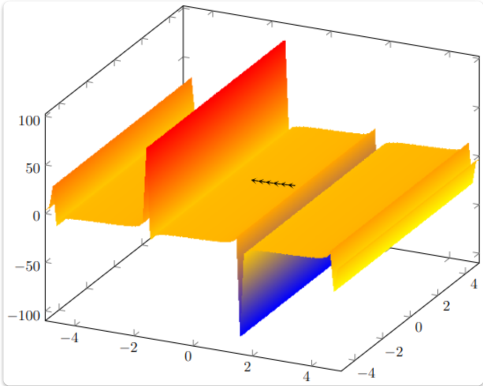
Illustrazioni di alcuni punti stazionari

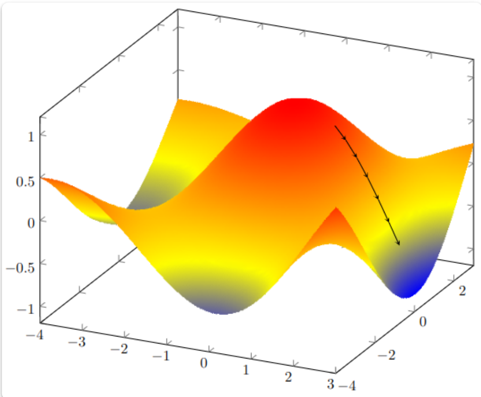
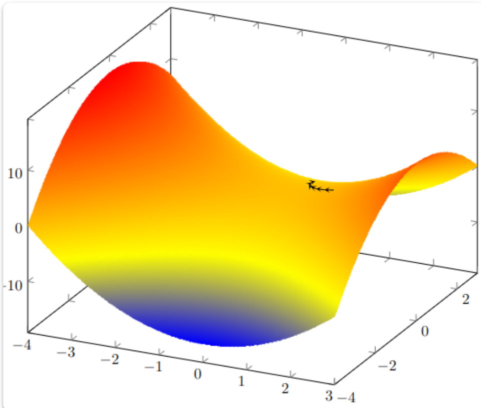
Nelle tabelle in basso, vediamo alcuni esempi curiosi di calcolo del punto stazionario.

L'immagine è stata ottenuta usando il pacchetto TikZ della suite LaTeX, usando un algoritmo per la generazione del grafico in un range definito.

Gli step sono stati normalizzati e aumentati a massimo 80.
(https://tex.stackexchange.com/questions/544796/plot-gradient-descent)

$\sin(x)$	$\cos(x)$
	

$\tan(0.5x)$	$\tan(x)$
Notare come l'algortimo riesca a trovare con successo un minimo locale 	L'algortimo ha trovato effettivamente un minimo locale, ma non il minimo globale (notare come le frecce puntino verso sinistra) 

$\cos(0.8x) \cdot \cos(0.6y) \cdot e^{0.1x}$	$x^2 - y^2$
Esempio interessante di funzione 	La tipica "sella" 

07/03/2023