Università di Pisa

Dipartimento di Matematica
Corso di Laurea Triennale in Matematica

Reti neurali profonde: capacità di approssimazione e convergenza a processi gaussiani

Relatore:
Prof. Dario Trevisan

Presentata da:
Francesco Caporali

Anno Accademico 2020/2021

A Letizia
ed alla mia famiglia.

Sommario

Il presente lavoro è incentrato sullo studio delle reti neurali profonde come modello formale con il quale generare funzioni da $\mathbb{R}^n$ in $\mathbb{R}^m$. Di tali funzioni vengono dapprima indagate le capacità di approssimazione tramite risultati di densità negli spazi L^p e nello spazio delle funzioni continue a supporto compatto; si presentano due distinte versioni del cosiddetto teorema di approssimazione universale, la prima dovuta a K. Hornik e G. Cybenko ed una seconda con stima quantitativa dovuta ad A. R. Barron. In una seconda parte, ripercorrendo quanto svolto da A. Matthews et al., è stata sviluppata la teoria alla base della convergenza di una variabile aleatoria generata da una rete neurale ad un processo gaussiano numerabile, nel limite di architetture di ampiezza infinita, a profondità fissata. La dimostrazione del teorema di convergenza ad un processo gaussiano è stata quindi rielaborata al fine di rendere la notazione chiara ed uniforme. Nel corso della trattazione particolare attenzione viene rivolta verso le possibili architetture di definizione delle reti e, nel dettaglio, si esaminano le condizioni imposte sulle funzioni di attivazione al fine di rendere validi i teoremi in analisi. In questo contesto è inserita una digressione sulla funzione ReLU e sulla possibilità di riformulare i teoremi di densità con questa attivazione, altrimenti esclusa. Al fine di verificare sperimentalmente i risultati teorici discussi sono state sviluppate delle simulazioni in `Python`, sfruttando il pacchetto `PyTorch`, con le quali si sono prodotti alcuni esempi e controesempi notevoli. Per mezzo delle sperimentazioni è stato inoltre possibile osservare direttamente i molteplici comportamenti di reti neurali basate su architetture differenti.

Indice

Notazioni

$[n]$	Insieme dei numeri da 1 ad n, $[n] := \{1, \dots, n\}$.
$\mathbb{N}$	Insieme dei numeri naturali.
$\mathbb{N}_0$	Insieme dei numeri naturali positivi, $\mathbb{N}_0 := \mathbb{N} \setminus \{0\}$.
$\mathbb{R}$	Insieme dei numeri reali.
$\overline{\mathbb{R}}$	Insieme dei numeri reali estesi, $\overline{\mathbb{R}} := \mathbb{R} \cup \{-\infty, +\infty\}$.
$\mathbb{C}$	Insieme dei numeri complessi.
A_i	i-esima riga della matrice $A \in \mathbb{R}^{n \times m}$.
A^i	i-esima colonna della matrice $A \in \mathbb{R}^{n \times m}$.
$v \cdot w$	Prodotto scalare di due vettori $v, w \in \mathbb{R}^n$.
$\delta_{i,j}$	Delta di Kronecker.
$\mathcal{P}(A)$	Insieme delle parti di A.
$\mathcal{B}(X)$	Insieme dei boreliani di uno spazio topologico (X, τ).
B_r	Palla di centro 0 e raggio r k-dimensionale, $B_r := B(0, r)$, per ogni k.
$\operatorname{supp}(f)$	Supporto di $f : X \to Y$, $\operatorname{supp}(f) = \{x \in X \mid f(x) \neq 0\}$.
$C_b(\mathbb{R}^k)$	Classe delle funzioni continue e limitate da $\mathbb{R}^k$ in $\mathbb{R}$ (anche con $k = +\infty$ eventualmente).
dx	Differenziale utilizzato per denotare l'integrazione rispetto alla misura di Lebesgue k-dimensionale, per ogni k.
$\mathcal{L}^k$	Misura di Lebesgue k-dimensionale.
$\widetilde{\forall}\, x \in X$	Per quasi ogni $x \in X$, ovvero per tutti gli $x \in X$ tranne al più per un sottoinsieme di misura nulla.
$\mathcal{N}^k(m, K)$	Distribuzione gaussiana k-dimensionale con vettore delle medie m e matrice delle covarianze K.
$X \sim Y$	Date X e Y v.a., $X \sim Y$ indica che le misure immagine di X e di Y sono coincidenti.

Introduzione

In tempi recenti le reti neurali profonde sono emerse come modelli parametrici molto versatili. Grazie a queste, infatti, si è attualmente in grado di approssimare complessi pattern di dati tramite moderni ed ottimizzati metodi di addestramento. Come conseguenza di questa particolare flessibilità comprendere al meglio le capacità ed i punti deboli di questi oggetti è divenuta una questione di fondamentale importanza in diverse discipline. In questo contesto riveste un ruolo centrale la costruzione di una struttura formale che permetta di indagare in maniera sistematica il comportamento delle reti e che, allo stesso tempo, permetta di costruire dimostrazioni rigorose.
Sebbene tali modelli abbiano avuto molto successo, la teoria sviluppata attorno a questi è ancora debole e, soprattutto, molto distante dalle applicazioni pratiche che ne vengono fatte.
L'obiettivo del presente lavoro è quello di richiamare alcuni dei principali risultati teorici relativi alle reti neurali profonde provando a fornire dimostrazioni dettagliate mantenendo comunque una notazione il più chiara ed uniforme possibile.
La prima delle proprietà ad essere indagata è la densità delle funzioni generabili tramite reti neurali. Si presentano tre distinti teoremi che permettono di affermare che le funzioni continue a supporto compatto, $C(X, \mathbb{R})$ con $X \subset \mathbb{R}^k$ compatto, e le funzioni nello spazio $L^p(\mathbb{R}^k)$ sono approssimabili con funzioni ottenute tramite reti (denotate con $\mathcal{N}_k(\psi)$), per ogni $k \geq 1$.
Dapprima, seguendo quanto fatto da Cybenko [6] ed Hornik [10], si utilizzano degli strumenti analitici per concludere la densità in L^p ed in $C(X)$. Questi primi due risultati vengono dimostrati sotto la condizione di attivazione sigmoidale, ipotesi in realtà abbastanza restrittiva per quel che riguarda le applicazioni pratiche; si fa uso infatti della nozione di funzione discriminatoria, proprietà implicata proprio dalla sigmoidalità.
Dopodiché si prosegue con un teorema di approssimazione universale con stima quantitativa, dovuto a Barron [2]. Questo costituisce una versione del risultato di densità in un particolare spazio di funzioni denotato con $\Gamma_{r,c}$. In più fornisce una stima della capacità di approssimazione delle funzioni implementabili tramite un modello di rete neurale con un numero fissato di neuroni m ($\mathcal{N}_{k,m}(\psi)$). Sebbene il teorema di Barron sia intrinsecamente rilevante, il suo scopo principale è quello di fornire una velocità approssimativa della convergenza ad una $f \in \Gamma_{r,c}$ di una successione di elementi in $\mathcal{N}_{k,m}(\psi)$, $m \in \mathbb{N}$; tuttavia la complessa struttura di tale insieme di funzioni rende molto ardua la verifica sperimentale del risultato.
Per sopperire alle difficoltà di applicazione dei teoremi di densità enunciati ci si è concentrati sulle reti neurali profonde con attivazione di tipo ReLU; questi modelli sono infatti tra i più utilizzati nelle applicazioni sperimentali, tuttavia vengono esclusi dalle ipotesi degli enunciati appena visti. Si è provato che con una rete neurale avente attivazione ReLU e 2 hidden layers si è in grado di generare una qualunque funzione facente parte dell'insieme $\mathcal{N}_k(\text{ReLU1})$ con $\text{ReLU1}(x) = \mathbb{1}_{[0,1]}(x)\, x + \mathbb{1}_{(1,\infty)}(x)$.
Utilizzando l'evidente proprietà di sigmoidalità di quest'ultima funzione si sono, per-

ciò, rielaborati i tre risultati precedentemente descritti al fine di contemplare anche i modelli ReLU.
Nel capitolo che segue si prende in analisi un teorema finalizzato a mostrare la convergenza in legge di variabili aleatorie generate per mezzo di reti neurali ad un processo gaussiano numerabile. Tale comportamento si presenta nel limite dell'ampiezza delle architetture con numero di hidden layers fissato e sotto l'ipotesi di inizializzazione gaussiana dei parametri strutturali delle reti.
Questo teorema, presentato da G. Matthews et al. [8], viene provato con il supporto di diversi risultati noti, il più rilevante dei quali è una versione, adattata a variabili aleatorie scambiabili, del Teorema del limite centrale. La dimostrazione presentata consiste di una rielaborazione di quella riportata nell'articolo principale, nel quale è strutturata in forma più schematica.
In questo caso, come avviene per i teoremi di densità, si assume un'ipotesi sulle funzioni d'attivazione utilizzate: devono essere sub-lineari. Il vincolo imposto questa volta è, in realtà, molto meno restrittivo, ma comunque esclusivo; ad esempio l'ampia classe di *Power Rectified Linear Unit functions* non soddisfa tale condizione.
I teoremi studiati nel corso di tutto il lavoro necessitano di alcune premesse inerenti l'analisi e la probabilità che sono state raccolte, per completezza, in un capitolo preliminare di richiami. Per la stesura di tale capitolo si è fatto riferimento a *Foundations of Modern Analysis* di Friedman [7] per la sezione di analisi funzionale e per il Teorema di decomposizione di Hahn-Jordan, a *Real Analysis and Probability* di Ash [1] per la sezione di teoria della misura ed a *Classical Fourier Analysis* di Grafakos [9] per la parte relativa all'analisi di Fourier. Inoltre alcuni degli enunciati presenti nel medesimo capitolo sono tratti da due distinti testi di Rudin (*Functional Analysis* [13] e *Real and Complex Analysis* [14]), ciò è stato fatto per ripercorrere il più chiaramente possibile il contenuto dell'articolo di Cybenko [6].
Tutti i risultati teorici riportati sono stati anche verificati sperimentalmente nel capitolo conclusivo. Le simulazioni sono state condotte tramite `PyTorch`, una delle più note librerie `Python` per il machine learning. Per mezzo di questo pacchetto è stata sviluppata autonomamente una sottoclasse fortemente parametrizzata della classe `nn.Module`, struttura base per tutti i modelli di rete in `PyTorch`. Avvalendosi di questa sottoclasse sono state condotte delle indagini preliminari sulle funzioni in $\mathcal{N}_k(\psi)$, al variare di ψ, che hanno permesso di procedere in maniera più sistematica alle verifiche dei risultati principali.
Al fine di riprodurre quanto ottenuto teoricamente per l'andamento asintotico si sono fissati parametri grandi per le ampiezze di ciascun hidden layer e si è studiata la densità del vettore aleatorio ottenuto. Andando a violare alcune delle ipotesi del teorema si è poi cercato di esibire dei controesempi.
In ultimo si è cercato di trovare un riscontro ai risultati di densità ottenuti. A tale scopo sono state eseguite delle procedure di training standard su modelli di reti con un neurone in input ed uno in output, così da riuscire ad approssimare delle funzioni continue da $\mathbb{R}$ in $\mathbb{R}$ su un dominio compatto $[a, b]$. Tali addestramenti sono stati effettuati utilizzando come dati un numero prestabilito di campioni della funzione di riferimento. Procedendo in questo modo si è provato a mettere in luce le molteplici capacità di approssimazione di reti neurali con architetture differenti, cercando di mostrare le diversità tra i modelli con attivazione ReLU, tra le più utilizzate, e quelli con attivazioni sigmoidali.

Capitolo 1

Richiami

In questo capitolo iniziale vengono presentati teoremi e proposizioni preliminari necessari per molte delle dimostrazioni presenti nei capitoli a seguire. Vengono omesse le prove di una parte dei risultati più classici, di cui viene soltanto citata la fonte. Per maggiore chiarezza tali enunciati sono suddivisi in sezioni tematiche.

1.1. Analisi funzionale

Definizione 1.1.1. Dato uno spazio vettoriale normato $(X, \|\cdot\|)$ si definisce funzionale lineare continuo un operatore lineare continuo (o equivalentemente limitato) da X in $\mathbb{R}$. Si denota con X^* lo spazio di tali operatori. Tale spazio è a sua volta normato: $\forall\, f^* \in X^*$,

$$\|f^*\| = \sup_{\|x\|=1} |f^*(x)|.$$

Teorema 1.1.1 (Teorema di Hahn-Banach)**.** [7, pp. 150–151] *Dato X spazio vettoriale, $p : X \to \mathbb{R}$,*

$$\forall\, x, y \in X,\ \forall\, \lambda \geq 0, \qquad p(x+y) \leq p(x) + p(y) \quad , \quad p(\lambda x) = \lambda p(x)$$

e $Z \subset X$ sottospazio vettoriale con $f^* \in Z^*$,

$$\forall\, z \in Z, \qquad f^*(x) \leq p(x),$$

allora $\exists\, h^* \in X^*$ tale che

$$\forall\, z \in Z, \quad h^*(z) = f^*(z) \quad e \quad \forall\, x \in X, \quad h^*(x) \leq p(x).$$

Il Teorema 1.1.1 non verrà usato direttamente, ma si farà ampio uso del seguente suo diretto corollario.

Corollario 1.1.1. *Dato uno spazio normato $(X, \|\cdot\|)$ ed un suo sottospazio vettoriale Y, se Y non è denso in X allora esiste un funzionale lineare continuo $h^* \in X^*$, $h^* \neq 0$, tale che $\forall\, y \in Y$, $h^*(y) = 0$.*

La dimostrazione è riportata in Appendice A.1.1.

Definizione 1.1.2. Dato (X, μ) spazio di misura e $f \in L^p(X, \mu)$, si denota

$$\|f\|_p = \left(\int_X |f|^p d\mu \right)^{\frac{1}{p}}.$$

Se invece $f \in L^\infty(X, \mu)$,

$$\|f\|_\infty = \text{ess-sup}|f| = \inf\{M \in \mathbb{R} \mid |f| \leq M, \mu - q.o.\}.$$

Teorema 1.1.2. [7, pp. 176–180] *Dato (X, μ) spazio di misura, con μ finita, siano p, q tali che $1 < p < \infty$ e $\frac{1}{p} + \frac{1}{q} = 1$. Allora per ogni funzionale lineare continuo $h^* \in [L^p(X, \mu)]^*$ si ha una corrispondente funzione $g \in L^q(X, \mu)$ tale che*

$$\forall\, f \in L^p(X, \mu), \qquad h^*(f) = \int_X fg\, d\mu.$$

In particolare

$$\|h^*\| = \|g\|_q.$$

Teorema 1.1.3. [7, pp. 180–181] *Dato (X, μ) spazio di misura, con μ finita si ha che per ogni funzionale lineare continuo, $h^* \in [L^1(X, \mu)]^*$, si ha una corrispondente funzione $g \in L^\infty(X, \mu)$ tale che*

$$\forall\, f \in L^1(X, \mu), \qquad h^*(f) = \int_X fg\, d\mu.$$

In particolare

$$\|h^*\| = \|g\|_\infty.$$

1.2. Teoria della misura

Teorema 1.2.1. [1, pp. 88–90] *Dato (X, μ) spazio di misura, $f \in L^\infty(X, \mu)$, allora $\forall\, \varepsilon > 0\ \exists\, f_\varepsilon$ funzione semplice tale che $\|f - f_\varepsilon\|_\infty < \varepsilon$, ovvero le funzioni semplici sono dense in $L^\infty(X, \mu)$.*

Definizione 1.2.1. Dato uno spazio misurabile $(X, \mathcal{A})$, con $\mathcal{A}$ σ-algebra, si definisce misura segnata una funzione $\mu : X \to \overline{\mathbb{R}}$ che sia σ-additiva; ovvero, data $(A_n)_{n=1}^\infty \in \mathcal{A}$ successione di insiemi a due a due disgiunti, vale

$$\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n).$$

Definizione 1.2.2. Una misura segnata su uno spazio misurabile $(X, \mathcal{A})$ si dice finita se $\forall\, A \in \mathcal{A}$ vale $|\mu(A)| < \infty$. Inoltre si dice boreliana se $\mathcal{B}(X) \subset \mathcal{A}$.

Teorema 1.2.2 (Teorema di decomposizione di Hahn-Jordan)**.** [7, pp. 25–28] *Dato uno spazio misurabile $(X, \mathcal{A})$ con misura segnata μ, esistono due insiemi misurabili P e N tali che:*

- *$P \cup N = X$ e $P \cap N = \emptyset$;*
- *$\forall\, E \in \mathcal{A}$, $E \subset P$ si ha $\mu(E) \geq 0$;*
- *$\forall\, E \in \mathcal{A}$, $E \subset N$ si ha $\mu(E) \leq 0$.*

Inoltre, tale misura segnata μ, può essere decomposta in maniera unica nella differenza di misure

$$\mu = \mu^+ - \mu^-$$

dove

$$\forall\, A \in \mathcal{A}, \qquad \mu^+(A) := \mu(A \cap P) \quad e \quad \mu^-(A) := \mu(A \cap N).$$

μ^+ è detta parte positiva e μ^- è detta parte negativa.

Definizione 1.2.3. Dato uno spazio misurabile $(X, \mathcal{A})$ con misura segnata μ, si può definire la misura $|\mu| = \mu^+ + \mu^-$, detta misura della variazione totale.

Osservazione 1.2.1. Data f integrabile rispetto a $|\mu|$ vale che f lo è anche rispetto a μ^+ e μ^-, dunque ha senso definire

$$\int_X f \, d\mu := \int_X f \, d\mu^+ - \int_X f \, d\mu^-.$$

Teorema 1.2.3 (Teorema di rappresentazione di Riesz)**.** [14, pp. 130–132] *Dato X spazio normato compatto e $f^* : C(X, \mathbb{R}) \to \mathbb{R}$ funzionale lineare continuo, allora $\exists!\, \mu$ misura segnata boreliana tale che*

$$\forall\, g \in C(X), \qquad f^*(g) = \int_X g \, d\mu.$$

In particolare, considerata $|\mu|$ misura della variazione totale di μ, vale anche

$$\|f^*\| = |\mu|(X).$$

Teorema 1.2.4. *Date μ misura finita, ν misura segnata sullo spazio $(X, \mathcal{A})$ tale che $\exists\, f \in L^q(X, \mu)$ con $\nu = f.\mu$ e $\varphi : X \to \mathbb{R}$ funzione misurabile, si ha*

$$\int_X \varphi \, d\nu = \int_X \varphi f \, d\mu.$$

La dimostrazione è riportata in Appendice A.1.2.

Osservazione 1.2.2. Questo teorema risulta essere una versione per misure segnate di un risultato classico per misure positive con densità, sotto ipotesi leggermente più forti. Si richiede infatti la finitezza delle misure, mentre nel caso classico basta la σ-finitezza.

1.3. Analisi di Fourier

Definizione 1.3.1. Data $f : \mathbb{R}^k \to \mathbb{R}$, $f \in L^1(\mathbb{R}^k, \mathcal{L}^k)$, si può definire $\hat{f} : \mathbb{R}^k \to \mathbb{C}$,

$$\hat{f}(\xi) = \int_{\mathbb{R}^k} e^{-2\pi i x \cdot \xi} f(x) \, dx.$$

Tale funzione è chiamata trasformata di Fourier di f.

Teorema 1.3.1 (Teorema di inversione di Fourier)**.** [9, pp. 112–113]
Date $f, \hat{f} \in L^1(\mathbb{R}^k, \mathcal{L}^k)$, vale

$$f(x) = \int_{\mathbb{R}^k} e^{2\pi i \xi \cdot x} \hat{f}(\xi) \, d\xi. \tag{1.1}$$

Osservazione 1.3.1. Assumendo $f, \hat{f} \in L^1(\mathbb{R}^k, \mathcal{L}^k)$ si ha, come conseguenza del Teorema di inversione di Fourier, l'identità

$$f(x) = f(0) + \int_{\mathbb{R}^k} (e^{2\pi i \xi \cdot x} - 1) \hat{f}(\xi) \, d\xi. \tag{1.2}$$

Teorema 1.3.2. [13, p. 176] *Data μ misura segnata su $\mathbb{R}^k$, se*

$$\int_{\mathbb{R}^k} e^{2\pi i \xi \cdot x}\, d\mu(\xi) = 0,$$

allora $\mu = 0$.

Osservazione 1.3.2. Nel caso di $\mu = \hat{f}.\mathcal{L}^k$, con $\hat{f} \in L^1(\mathbb{R}^k, \mathcal{L}^k)$, il Teorema 1.3.2 seguirebbe direttamente dalla Definizione 1.3.1 e dal Teorema 1.3.1.

1.4. Variabili aleatorie discrete su spazi di Hilbert

Dato un sottoinsieme finito $K = \{g_1^*, \dots, g_m^*\}$ di uno spazio di Hilbert $(H, \cdot)$ e uno spazio di probabilità $(\Omega, \mathcal{A}, \mathbb{P})$, si definisca g una variabile aleatoria (v.a.)

$$g : (\Omega, \mathcal{A}, \mathbb{P}) \to (K, \mathcal{P}(K)). \tag{1.3}$$

É immediato notare che la misura immagine $\mathbb{P}_g$, indotta dalla v.a. sullo spazio $(K, \mathcal{P}(K))$ [1], è discreta, essendo K finito:

$$\mathbb{P}_g(g_i^*) = \mathbb{P}(g = g_i^*) = \mathbb{P}(g^{-1}(g_i^*)), \quad \forall\, i \in [m]. \tag{1.4}$$

Fissate dunque f, g v.a., definite come nelle equazioni (1.3) e (1.4), si possono dare le seguenti definizioni.

Definizione 1.4.1 (Valore atteso)**.** $\mathbb{E}[g] = \sum_{i=1}^{m} g_i^* \mathbb{P}_g(g_i^*)$.

Definizione 1.4.2 (Covarianza)**.** $\mathrm{Cov}\,(f, g) = \mathbb{E}\left[(f - \mathbb{E}[f]) \cdot (g - \mathbb{E}[g])\right]$.

Definizione 1.4.3 (Varianza)**.** $\mathrm{Var}\,(g) = \mathrm{Cov}\,(g, g) = \mathbb{E}\left[\|g - \mathbb{E}[g]\|^2\right]$.

Osservazione 1.4.1. Vale anche che

$$\mathrm{Cov}\,(f, g)) = \mathbb{E}[f \cdot g] - \mathbb{E}[f] \cdot \mathbb{E}[g],$$

infatti

$$\begin{aligned}\mathrm{Cov}\,(f, g) &= \mathbb{E}\left[(f - \mathbb{E}[f]) \cdot (g - \mathbb{E}[g]\right] = \\ &= \mathbb{E}[f \cdot g] + \mathbb{E}[f] \cdot \mathbb{E}[g] - \mathbb{E}\left[f \cdot \mathbb{E}[g]\right] - \mathbb{E}\left[\mathbb{E}[f] \cdot g\right] = \\ &= \mathbb{E}[f \cdot g] - \mathbb{E}[f] \cdot \mathbb{E}[g],\end{aligned}$$

dove per l'ultimo uguale si è usata la linearità del valore atteso.
Di conseguenza vale anche

$$\mathrm{Var}\,(g) = \mathbb{E}[\|g\|^2] - \|\mathbb{E}[g]\|^2\,.$$

Per quanto riguarda le definizioni di varianza e covarianza, si può notare che le v.a. $(f - \mathbb{E}[f]) \cdot (g - \mathbb{E}[g])$ e $\|f - \mathbb{E}[f]\|^2$ sono a valori in $\mathbb{R}$, dunque la definizione di valore atteso in questo caso è quella convenzionale[2].

[1] Si può pensare di estendere lo spazio d'arrivo a $(H, \mathcal{P}(H))$ con l'ipotesi che $\forall\, f^* \in H \setminus K$ valga $\mathbb{P}_g(f^*) = 0$.

[2] Anche in questo caso, sebbene le v.a. siano reali, sono comunque discrete.

Proposizione 1.4.1. *Data g v.a. definita come in* (1.3) *e* (1.4) *e $\phi : K \to \mathbb{R}$, la v.a. $\phi(g)$ è tale che*

$$\mathbb{E}[\phi(g)] = \sum_{g^* \in K} \phi(g^*)\mathbb{P}_g(g^*).$$

Proposizione 1.4.2. *Date f, g v.a. definite come in* (1.3) *e* (1.4)*, sono indipendenti se e solo se*

$$\mathbb{P}_{f,g} = \mathbb{P}_f \otimes \mathbb{P}_g.$$

Le dimostrazioni delle Proposizione 1.4.1 e 1.4.2 sono del tutto analoghe al caso reale, sono pertanto omesse.

Proposizione 1.4.3. *Date f, g v.a. indipendenti definite come in* (1.3) *e* (1.4)*, vale*

$$\mathbb{E}[f \cdot g] = \mathbb{E}[f] \cdot \mathbb{E}[g].$$

Proposizione 1.4.4. *Date f, g v.a. indipendenti definite come in* (1.3) *e* (1.4)*, vale*

$$Var(f + g) = Var(f) + Var(g)\,.$$

La dimostrazione delle ultime due proposizioni è riportata in Appendice A.1.3.

Capitolo 2

Densità delle funzioni generabili tramite reti neurali

Al fine di avvicinarsi ai risultati di densità delle funzioni generabili tramite reti neurali si presentano brevemente queste ultime come oggetto matematico formale. Tramite una loro definizione rigorosa è infatti possibile indagarne le proprietà sfruttando nozioni basilari. In questo lavoro ci si concentra su una particolare classe di reti neurali che nella loro rappresentazione sotto forma di grafo risultano completamente connesse ed acicliche (comunemente dette feed-forward).

Definizione 2.0.1 (Rete neurale completamente connessa feed-forward (NN))**.** Una rete neurale completamente connessa feed-forward è definita tramite un'architettura $\alpha = (n, \psi)$ con $l \in \mathbb{N}_{\geq 2}$ numero di layer, $n \in \mathbb{N}^{l+1}$ e $\psi : \mathbb{R} \to \mathbb{R}$ funzione di attivazione. Si indicano con n_0, n_l e n_i per $i = 1, \dots, l-1$ il numero di unità (o neuroni) rispettivamente in input, output e nell'i-esimo hidden layer.
Denotando con $p(n) = \sum_{i=i}^{l} n_i n_{i-1} + n_i$ si definisce una funzione di realizzazione

$$\Phi_\alpha : \mathbb{R}^{n_0} \times \mathbb{R}^{p(n)} \to \mathbb{R}^{n_l}$$

che, fissato un parametro θ,

$$\theta = (\theta_i)_{i=1}^{l} = (A^{(i)}, b^{(i)})_{i=1}^{l} \in \bigtimes_{i=1}^{l} \left(\mathbb{R}^{n_i \times n_{i-1}} \times \mathbb{R}^{n_i}\right) \cong \mathbb{R}^{p(n)}$$

per ogni input $x \in \mathbb{R}^{n_0}$, calcola $\Phi_\alpha(\cdot, \theta) : \mathbb{R}^{n_0} \to \mathbb{R}^{n_l}$ come $\Phi_\alpha(x, \theta) = \Phi_\alpha^{(l)}(x, \theta)$ dove[1]

$$\begin{aligned}
&\Phi_\alpha^{(1)}(x, \theta) = A^{(1)} x + b^{(1)}, \\
&\bar{\Phi}_\alpha^{(1)}(x, \theta) = \psi(\Phi_\alpha^{(1)}(x, \theta)), \\
&\qquad\qquad \vdots \\
&\Phi_\alpha^{(l-1)}(x, \theta) = A^{(l-1)} \bar{\Phi}_\alpha^{(l-2)}(x, \theta) + b^{(l-1)}, \\
&\bar{\Phi}_\alpha^{(l-1)}(x, \theta) = \psi(\Phi_\alpha^{(l-1)}(x, \theta)), \\
&\Phi_\alpha^{(l)}(x, \theta) = A^{(l)} \bar{\Phi}_\alpha^{(l-1)}(x, \theta) + b^{(l)}.
\end{aligned}$$

Nel seguito si farà riferimento a $\bar{\Phi}_\alpha^{(i)}$ e $\Phi_\alpha^{(i)}$, rispettivamente, come attivazione e preattivazione dei neuroni nell'i-esimo layer.
Si definiscono infine ampiezza e profondità di una rete neurale: $\|n\|_\infty$ e l ($l-1$ è il numero di hidden layers).

[1] Si considera ψ applicata componente per componente ogni volta che la si valuta su un vettore.

Il seguente grafo diretto aciclico in fig. 2.1 è un esempio di *NN* ottenuto rappresentando le mappe affini $x \to A^{(i)}x + b^{(i)}$, $\forall\, i \in [l]$ intervallate con l'attivazione ψ. Nel corso del lavoro si mostreranno numerosi esempi di funzioni di attivazione comunemente utilizzate.

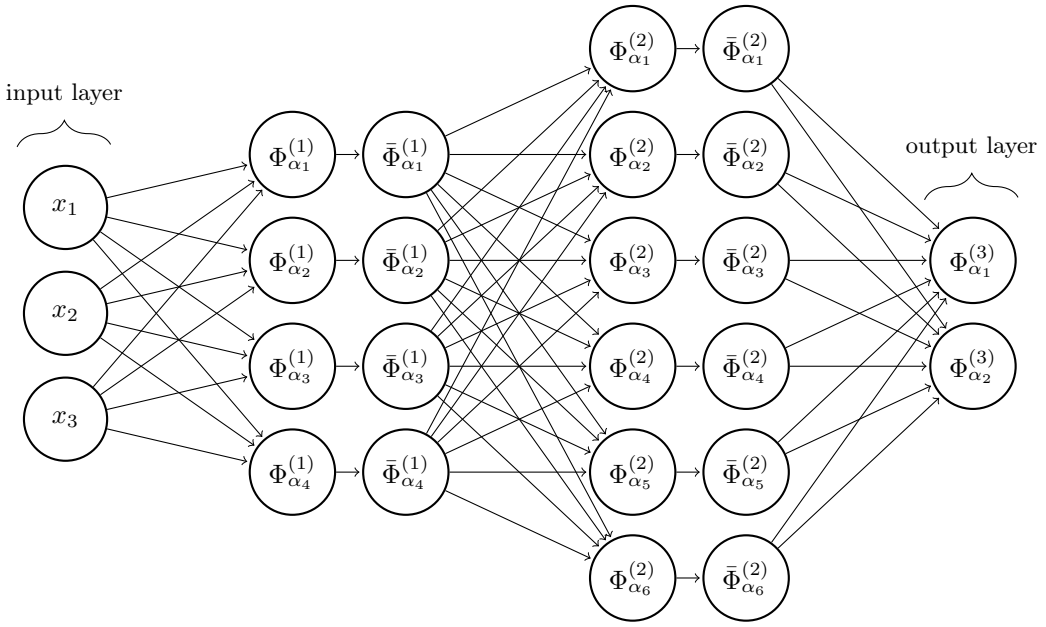

Figura 2.1: Grafo associato ad una rete neurale con 2 hidden layers, architettura $\alpha = ((3, 4, 6, 2), \psi)$ e parametri $\theta = ((A^{(i)}, b^{(i)})_{i=1}^{3}$: $\Phi_\alpha : \mathbb{R}^3 \times \mathbb{R}^{60} \to \mathbb{R}^2$.

Nelle sezioni presentate di seguito si segue il percorso svolto da Cybenko [6] e Hornik [10]. Le dimostrazioni consistono di una versione rivista e dettagliata di quelle presenti nei loro rispettivi articoli. Si andrà pertanto a considerare, nell'enunciare i teoremi di densità, il caso di *NN* con un solo hidden layer, per facilità di notazione. Tuttavia, come osservato da Hornik [10, p. 252], il risultato per reti neurali di tipo *deep* (ovvero con un numero di hidden layers ≥ 2) può essere dedotto dal caso base (Hornik et al. [11, p. 363]).

2.1. Densità in L^p

Data una *NN* con un solo *hidden-layer* composto da m unità, k unità in input, una in output e con funzione di attivazione ψ, l'insieme delle funzioni implementabili[2] al variare di $\theta = (A, b, c) \in \mathbb{R}^{m\times k} \times \mathbb{R}^m \times \mathbb{R}^m$ è

$$\mathcal{N}_{k,m}(\psi) = \left\{ h : \mathbb{R}^k \to \mathbb{R} \;\middle|\; h(x) = \sum_{i=1}^{m} c_i \psi\left(A_i \cdot x - b_i\right) \right\}. \tag{2.1}$$

Se si lascia quindi variare $m \in \mathbb{N}_0$ si può generare

$$\mathcal{N}_k(\psi) = \bigcup_{m=1}^{\infty} \mathcal{N}_{k,m}(\psi),$$

[2]Si suppone che l'ultima funzione di pre-attivazione sia lineare.

ovvero lo spazio che si è finora informalmente definito come insieme delle funzioni generabili tramite reti neurali.
A questo punto si definiscono due proprietà relative a funzioni da $\mathbb{R}$ in $\mathbb{R}$ di fondamentale importanza nel seguito.

Definizione 2.1.1. Data una funzione $\psi : \mathbb{R} \to \mathbb{R}$, si dice sigmoidale se

$$\psi(t) \xrightarrow{t \to +\infty} 1 \quad \text{e} \quad \psi(t) \xrightarrow{t \to -\infty} 0.$$

Definizione 2.1.2. Data una funzione ψ, si dice discriminatoria se data μ misura segnata e finita su $\mathbb{R}^k$, tale che

$$\forall\, a \in \mathbb{R}^k, b \in \mathbb{R}, \qquad \int_{\mathbb{R}^k} \psi(a \cdot x - b)\, d\mu = 0,$$

allora $\mu = 0$.

Come mostra il seguente Lemma 2.1.1 queste definizioni sono in realtà collegate tra loro.

Lemma 2.1.1. *Data ψ misurabile e sigmoidale, ψ è discriminatoria.*

Dimostrazione. Si noti anzitutto che $\forall\, a, x \in \mathbb{R}^k, b, c, \lambda \in \mathbb{R}$

$$\psi(\lambda(a \cdot x - b) + c) \begin{cases} \xrightarrow{\lambda \to +\infty} 1 & \text{se } (a \cdot x - b) > 0 \\ \xrightarrow{\lambda \to +\infty} 0 & \text{se } (a \cdot x - b) < 0 \\ = \psi(c) & \text{se } (a \cdot x - b) = 0 \end{cases}.$$

Definendo

$$\Pi_{a,b} = \left\{ x \in \mathbb{R}^k \,\middle|\, a \cdot x - b = 0 \right\},$$
$$H_{a,b} = \left\{ x \in \mathbb{R}^k \,\middle|\, a \cdot x - b > 0 \right\},$$

rispettivamente l'iperpiano e il semispazio aperto definiti dalla mappa affine $a \cdot x - b$, e $\psi_\lambda(x) = \psi(\lambda(a \cdot x - b) + c)$, si ha che $\psi_\lambda(x) \xrightarrow{\lambda \to +\infty} \gamma(x)$ puntualmente, dove $\gamma(x)$ è definita come segue

$$\gamma(x) = \begin{cases} 1 & \text{se } x \in H_{a,b} \\ \psi(c) & \text{se } x \in \Pi_{a,b} \\ 0 & \text{altrimenti} \end{cases}.$$

Data μ misura segnata, finita, su $\mathbb{R}^k$ tale che

$$\forall\, a \in \mathbb{R}^k, b \in \mathbb{R}, \qquad \int_{\mathbb{R}^k} \psi(a \cdot x - b)\, d\mu(x) = 0, \tag{2.2}$$

si ha, per il Teorema 1.2.2, $\mu = \mu^+ - \mu^-$. In particolare essendo μ finita anche μ^+ e μ^- lo sono.
Si vuole dunque applicare il Teorema di convergenza dominata di Lebesgue a

$$\lim_{\lambda \to \infty} \int_{\mathbb{R}^k} \psi_\lambda \, d\mu = \lim_{\lambda \to \infty} \left(\int_{\mathbb{R}^k} \psi_\lambda \, d\mu^+ - \int_{\mathbb{R}^k} \psi_\lambda \, d\mu^- \right).$$

Essendo ψ sigmoidale, $\forall\, \lambda \in \mathbb{N}$ si ha $|\psi_\lambda| \leq j$ per qualche $j \geq 0$ e anche

$$\int_{\mathbb{R}^k} j\, d\mu^+ = j\mu^+(\mathbb{R}^k \cap P) < \infty \quad \wedge \quad \int_{\mathbb{R}^k} j\, d\mu^- = j\mu^-(\mathbb{R}^k \cap N) < \infty,$$

dunque, per convergenza dominata

$$\lim_{\lambda\to\infty} \int_{\mathbb{R}^k} \psi_\lambda\, d\mu^+ = \int_{\mathbb{R}^k} \lim_{\lambda\to\infty} \psi_\lambda\, d\mu^+ = \int_{\mathbb{R}^k} \gamma\, d\mu^+ < \infty.$$

Analogamente

$$\lim_{\lambda\to\infty} \int_{\mathbb{R}^k} \psi_\lambda\, d\mu^- = \int_{\mathbb{R}^k} \gamma\, d\mu^- < \infty,$$

da cui

$$\lim_{\lambda\to\infty} \int_{\mathbb{R}^k} \psi_\lambda\, d\mu = \int_{\mathbb{R}^k} \gamma\, d\mu.$$

Si noti che per l'equazione (2.2), $\forall\, a, b, c, \lambda$, definiti $a_\lambda := \lambda a$ e $b_\lambda := \lambda b + c$, vale

$$\int_{\mathbb{R}^k} \psi_\lambda\, d\mu = \int_{\mathbb{R}^k} \psi(a_\lambda \cdot x - b_\lambda)\, d\mu(x) = 0,$$

dunque anche

$$0 = \int_{\mathbb{R}^k} \gamma\, d\mu = \int_{H_{a,b}} 1\, d\mu + \int_{\Pi_{a,b}} \psi(c)\, d\mu = \mu(H_{a,b}) + \psi(c)\mu(\Pi_{a,b}). \tag{2.3}$$

Passando al limite per $c \to \infty$, essendo ψ sigmoidale,

$$\forall\, a \in \mathbb{R}^k, b \in \mathbb{R}, \quad \mu(H_{a,b}) + \mu(\Pi_{a,b}) = 0.$$

Si vuole ora concludere che se per una certa misura segnata finita μ, la misura di tutti i semispazi chiusi è nulla, come appena mostrato, allora $\mu = 0$.
A tale scopo, fissando a, si definisce il funzionale lineare F, sullo spazio delle funzioni misurabili e limitate quasi ovunque $L^\infty(\mathbb{R}, \mu)$, come

$$F(h) = \int_{\mathbb{R}^k} h(a \cdot x)\, d\mu(x).$$

Si può notare immediatamente che, per tale funzionale, vale $F \in [L^\infty(\mathbb{R}, \mu)]^*$, poiché è limitato, essendo μ finita, infatti

$$\forall\, h \in L^\infty(\mathbb{R}, \mu) \text{ con } \|h\|_\infty = 1, \qquad F(h) = \int_{\mathbb{R}^k} h(a \cdot x)\, d\mu \leq \mu(\mathbb{R}^k) < \infty.$$

Scelta poi $h(x) = \mathbb{1}_{[b,\infty)}(x) \in L^\infty(\mathbb{R}, \mu)$, vale

$$F(h) = \int_{\mathbb{R}^k} h(a \cdot x)\, d\mu = \int_{\{a\cdot x - b \geq 0\}} d\mu = \mu(H_{a,b}) + \mu(\Pi_{a,b}) = 0.$$

Analogamente se $h(x) = \mathbb{1}_{(b,\infty)}(x)$, $F(h) = 0$, poiché $\mu(H_{a,b}) = 0$ (si ottiene per $c \to 0$ dall'equazione (2.3)).
Per linearità di F si ottiene dunque che per ogni intervallo $I = [d, e]$, $[d, e)$, $(d, e]$ o

(d, e) [3] vale $F(\mathbb{1}_I) = 0$.
Sempre per linearità, $\forall\, h$ funzione semplice, $F(h) = 0$, ed essendo le funzioni semplici dense in $L^\infty(\mathbb{R}^k, \mu)$, per il Teorema 1.2.1 si ha $F = 0$.
In particolare, fissato $y \in \mathbb{R}^k$, usando le funzioni $\cos(2\pi\, y \cdot x)$ e $\sin(2\pi\, y \cdot x)$ si può scrivere

$$\begin{aligned} 0 &= F(\cos(2\pi\, y \cdot x) + i \sin(2\pi\, y \cdot x)) = \\ &= \int_{\mathbb{R}^k} \cos(2\pi\, y \cdot x) + i \sin(2\pi\, y \cdot x)\, d\mu(x) = \\ &= \int_{\mathbb{R}^k} e^{2\pi\, iy \cdot x}\, d\mu(x). \end{aligned}$$

Quindi per il Teorema 1.3.2 $\mu = 0$, da cui segue che ψ è discriminatoria. □

Osservazione 2.1.1. Il Lemma 2.1.1 è valido anche se ψ è misurabile, limitata e non costante. La dimostrazione è più complicata e fa uso di strumenti di analisi di Fourier [10, pp. 255–256].

Grazie all'importante risultato appena dimostrato si è in grado di enunciare e provare un primo teorema di densità dell'insieme $\mathcal{N}_k(\psi)$.

Teorema 2.1.1 (Teorema di densità in L^p)**.** *Fissata ψ funzione di attivazione misurabile e sigmoidale, vale che $\mathcal{N}_k(\psi)$ è denso in $L^p(\mathbb{R}^k, \mu)$ ($\forall\, p \in [1, +\infty)$) per ogni μ misura finita su $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$.*

Dimostrazione. Siccome ψ è limitata, $\mathcal{N}_k(\psi)$ è un sottospazio vettoriale di $L^p(\mathbb{R}^k, \mu)$ per ogni μ misura finita su $\mathbb{R}^k$. Infatti è chiuso per combinazioni lineari e, per ogni $f \in \mathcal{N}_k(\psi)$, si ha

$$\begin{aligned} \|f\|_p &= \left(\int_{\mathbb{R}^k} \left| \sum_{i=1}^n c_i \psi\left(A_i \cdot x - b_i\right) \right|^p d\mu(x) \right)^{1/p} \leq \\ &\leq \left(\int_{\mathbb{R}^k} \left| \sum_{i=1}^n c_i \|\psi\|_\infty \right|^p d\mu(x) \right)^{1/p} = \\ &= \left| \sum_{i=1}^n c_i \|\psi\|_\infty \right| \mu\left(\mathbb{R}^k\right)^{1/p} < \infty. \end{aligned}$$

Supponendo per assurdo che $\overline{\mathcal{N}_k(\psi)} \subsetneq L^p(\mathbb{R}^k, \mu)$, per il Corollario 1.1.1 si avrebbe che $\exists\, h^* \in [L^p(\mathbb{R}^k, \mu)]^*$, $h^* \neq 0$, tale che

$$\forall\, f \in \mathcal{N}_k(\psi) \quad h^*(f) = 0.$$

In particolare essendo $h^* \in [L^p(\mathbb{R}^k, \mu)]^*$, se $p \neq 1$, per il Teorema 1.1.2, definito q in modo che $\frac{1}{p} + \frac{1}{q} = 1$, $\exists\, g \in L^q(\mathbb{R}^k, \mu)$ tale che

$$h^*(f) = \int_{\mathbb{R}^k} f g\, d\mu. \tag{2.4}$$

[3] Per esempio $\mathbb{1}_{[d,e]} = \mathbb{1}_{[d,\infty)} - \mathbb{1}_{(e,\infty)}$.

Se invece $p = 1$, per il Teorema 1.1.3, $\exists\, g \in L^\infty(\mathbb{R}^k, \mu)$ tale che valga nuovamente l'equazione (2.4).
Tale g permette di definire una misura segnata $\sigma = g.\mu$ come segue:

$$\forall\, B \in \mathcal{B}(\mathbb{R}^k), \quad \sigma(B) = \int_B g\, d\mu.$$

Applicando la disuguaglianza di Hölder si può vedere che tale misura segnata è finita, infatti

$$\forall\, B \in \mathcal{B}(\mathbb{R}^k), \quad |\sigma(B)| = \left| \int_{\mathbb{R}^k} \mathbb{1}_B g\, d\mu \right| \leq \|\mathbb{1}_B\|_p \|g\|_q \leq (\mu(\mathbb{R}^k))^{\frac{1}{p}} \|g\|_q < \infty.$$

A questo punto si può affermare che

$$h^*(f) = \int_{\mathbb{R}^k} fg\, d\mu = \int_{\mathbb{R}^k} f\, d\sigma, \tag{2.5}$$

dove l'ultimo passaggio è dato dal Teorema 1.2.4.
In particolare si ha $h^* \neq 0$, dunque $\exists\, f \in L^p(\mathbb{R}^k, \mu)$ tale che

$$h^*(f) = \int_{\mathbb{R}^k} f\, d\sigma \neq 0 \quad \Longrightarrow \quad \sigma \neq 0. \tag{2.6}$$

Sapendo che h^* si annulla su $\mathcal{N}_k(\psi)$ e che $\psi(a \cdot x - b) \in \mathcal{N}_{k,1}(\psi) \subset \mathcal{N}_k(\psi)$, si ha

$$\forall\, a \in \mathbb{R}^k, b \in \mathbb{R}, \qquad \int_{\mathbb{R}^k} \psi(a \cdot x - b)\, d\sigma = 0. \tag{2.7}$$

Usando il Lemma 2.1.1, essendo ψ misurabile e sigmoidale, allora è discriminatoria, dunque per l'equazione (2.7) vale $\sigma = 0$ che crea un assurdo con quanto detto in (2.6). □

2.2. Teorema di approssimazione universale

In maniera del tutto analoga a quanto fatto nella sezione appena conclusa si può enunciare e provare il Teorema di approssimazione universale. In questo caso, infatti, si farà uso del Teorema di rappresentazione di Riesz al posto dei teoremi relativi alla dualità degli spazi L^p, ma ci si appoggerà comunque alla dicriminatorietà della funzione di attivazione per concludere la prova dell'asserto.

Teorema 2.2.1 (Teorema di approssimazione universale)**.** *Fissata ψ funzione di attivazione continua e sigmoidale, vale che $\mathcal{N}_k(\psi)$ è denso in $(C(X), \|\cdot\|_\infty)$ per ogni $X \subset \mathbb{R}^k$ compatto.*

Dimostrazione. In maniera analoga alla dimostrazione del Teorema 2.1.1, usando ψ continua e limitata, $\mathcal{N}_k(\psi)$ è un sottospazio vettoriale di $C(X)$. Infatti, come già detto, è chiuso per combinazioni lineari e per ogni $f \in \mathcal{N}_k(\psi)$ si ha $f \in C(\mathbb{R}^k)$.
Sempre per assurdo, se $\overline{\mathcal{N}_k(\psi)} \subsetneq C(X)$, per il Corollario 1.1.1 si avrebbe che $\exists\, h^* \in [C(X)]^*$, $h^* \neq 0$ tale che

$$\forall\, f \in \mathcal{N}_k(\psi) \quad h^*(f) = 0.$$

Per il Teorema di rappresentazione di Riesz, essendo h^* un funzionale lineare continuo $h^* \in [C(X)]^*$, si ha che $\exists!\, \mu$ misura segnata e boreliana tale che

$$\forall\, f \in C(X), \qquad h^*(f) = \int_X f\, d\mu.$$

Sempre per il Teorema 1.2.3 si ha μ finita, infatti, essendo h^* limitata

$$\forall\, A \in \mathcal{B}(X), \quad |\mu(A)| \leq |\mu|(A) \leq |\mu|(X) = \|h^*\| < \infty.$$

A questo punto si procede con gli stessi passaggi effettuati precedentemente a partire dall'equazione (2.5). Per punti:

- si verifica $\mu \neq 0$;
- utilizzando ψ discriminatoria (Lemma 2.1.1) si ottiene un assurdo.

□

2.3. Approssimazione quantitativa

In ultimo, si presenta una versione quantitativa del Teorema 2.2.1. Tale teorema di densità è tuttavia ristretto ad un particolare sottoinsieme dello spazio delle funzioni continue a supporto compatto ed è perciò una versione indebolita di quanto appena mostrato. Il motivo che porta comunque ad andare in questa direzione è la volontà stimare le capacità di approssimazione di $\mathcal{N}_{k,n}(\psi)$, $\forall\, n \in \mathbb{N}_0$.
Riprendendo la notazione e le definizioni introdotte nella Sezione 1.3 si definisce l'insieme

$$\Gamma_{r,c} = \left\{ f : \mathbb{R}^k \to \mathbb{R} \,\middle|\, \operatorname{supp}(f) \subset B_r \,\wedge\, f, \hat{f} \in L^1(\mathbb{R}^k, \mathcal{L}^k) \,\wedge\, \int_{\mathbb{R}^k} \|\xi\| |\hat{f}(\xi)|\, d\xi \leq \frac{c}{r} \right\}.$$

Si noti che $\forall\, f \in \Gamma_{r,c}$ vale $f, \hat{f} \in L^1(\mathbb{R}^k, \mathcal{L}^k)$, dunque vale lo sviluppo (1.2) dell'Osservazione 1.3.1 per f, ovvero

$$f(x) = f(0) + \int_{\mathbb{R}^k} (e^{2\pi i \xi \cdot x} - 1) \hat{f}(\xi)\, d\xi.$$

Nella definizione di $\Gamma_{r,c}$ si pensa $\mathbb{R}^k$ come uno spazio vettoriale euclideo, dunque $\|\cdot\| = \|\cdot\|_2$, ma quanto segue è indipendente da tale assunzione.

Teorema 2.3.1 (Teorema di approssimazione quantitativa)**.** *Siano ψ funzione di attivazione sigmoidale e μ misura di probabilità su B_r. Allora $\forall\, f \in \Gamma_{r,c}$, fissato $n \in \mathbb{N}_0$, esiste una funzione $f_n \in \mathcal{N}_{k,n}(\psi)$ tale che*

$$\int_{B_r} (f - f_n)^2\, d\mu \leq \frac{(4\pi c)^2}{n}.$$

In particolare i coefficienti c_i usati nella definizione di f_n (si veda l'equazione (2.1)) sono tali che $\sum_{i=1}^n |c_i| \leq 4\pi c$.

Osservazione 2.3.1. Nella dimostrazione si assume $f(0) = 0$; se così non fosse si potrebbe rimuovere dalla costruzione delle funzioni implementabili dalle reti neurali l'ipotesi che l'ultima funzione di pre-attivazione debba essere lineare. Basterebbe, infatti, ridefinire

$$\mathcal{N}_{k,m}(\psi) = \left\{ h : \mathbb{R}^k \to \mathbb{R} \;\middle|\; h(x) = \sum_{i=1}^{m} c_i \psi\left(A_i \cdot x - b_i\right) + c_0 \right\}$$

ed assumere $c_0 = f(0)$ per avere la tesi del teorema.

Per riuscire a provare il Teorema di approssimazione quantitativa è prima necessario dare la definizione di inviluppo convesso ed enunciare il Lemma 2.3.1.

Definizione 2.3.1. Dato V spazio vettoriale reale, si definisce inviluppo convesso di un insieme $G \subset V$, l'insieme di tutte le combinazioni convesse finite di punti di G:

$$co(G) = \left\{ \sum_{i=1}^{n} \lambda_i x_i \;\middle|\; \forall\, i \in [n]\ x_i \in G,\ \lambda_i \geq 0 \ \text{ e } \ \sum_{i=1}^{n} \lambda_i = 1 \right\}.$$

Lemma 2.3.1. *Dato $(H, \cdot)$ spazio di Hilbert e $G \subset H$ tale che $\exists\, b > 0$ $\forall\, g \in G,\ \|g\| \leq b$. Sia $f \in \overline{co(G)}$, allora $\forall n \in \mathbb{N}_0$ e $\forall c' \in \mathbb{R}$ tale che $c' \geq b^2 - \|f\|^2$, esiste $f_n \in co(G)$, combinazione convessa di n elementi di G, tale che*

$$\|f - f_n\|^2 \leq \frac{c'}{n}.$$

Dimostrazione. Siano $\delta > 0$, $n \in \mathbb{N}_0$ e $f^* \in co(G)$ tale che $\|f - f^*\| \leq \delta/n$.
Esiste $m \in \mathbb{N}_0$, sufficientemente grande, tale che f^* è della forma

$$f^* = \sum_{i=1}^{m} \lambda_i g_i^* \quad \text{con} \quad g_i^* \in G,\, \lambda_i \geq 0\ \forall\, i \in [m]\ \wedge\ \sum_{i=1}^{m} \lambda_i = 1.$$

Si definisce la variabile aleatoria g a valori nell'insieme $K = \{g_1^*, \ldots, g_m^*\}$ tale che la sua misura immagine sia la probabilità $\mathbb{P}(g = g_i^*) = \lambda_i$, $\forall i \in [m]$.
Si considerano quindi n copie indipendenti di g: $g_1, \ldots, g_n$. Si costruisce la variabile aleatoria media:

$$f_n = \frac{1}{n} \sum_{i=1}^{n} g_i.$$

Si può notare che

$$\mathbb{E}[f_n] = \mathbb{E}[g] = \sum_{i=1}^{m} g_i^* \lambda_i = f^*.$$

Inoltre

$$\begin{aligned}
\mathrm{Var}\left(f_n\right) &= \mathbb{E}\left[\|f_n - f^*\|^2\right] = \\
&= \mathbb{E}\left[\left\| \frac{1}{n} \sum_{i=1}^{n} g_i - \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[g_i] \right\|^2\right] = \\
&= \frac{1}{n^2} \mathbb{E}\left[\left\| \sum_{i=1}^{n} \left(g_i - \mathbb{E}[g_i]\right) \right\|^2\right] = \\
&= \frac{1}{n} \mathbb{E}\left[\|g - f^*\|^2\right] = \frac{1}{n} \mathrm{Var}\left(g\right),
\end{aligned}$$

ed è anche equivalente, per l'Osservazione 1.4.1,

$$\mathrm{Var}\,(g) = \mathbb{E}\left[\|g\|^2\right] - \|f^*\|^2.$$

Da ciò, usando l'ipotesi $\forall\, g^* \in G,\ \|g^*\| \leq b$, si ha

$$\mathbb{E}\left[\|g\|^2\right] = \sum_{k=1}^{m} \lambda_i \|g_i^*\|^2 \leq b^2 \quad \Longrightarrow \quad \mathbb{E}\left[\|f_n - f^*\|^2\right] \leq \frac{1}{n}\left(b^2 - \|f^*\|^2\right).$$

Essendo valida tale limitazione al valore atteso di $\|f_n - f^*\|^2$, si può facilmente verificare (ad esempio per assurdo) che devono esistere $g_{i_1}^*, \ldots, g_{i_n}^* \in K$ tali che, definita

$$\tilde{f}_n = \frac{1}{n}\sum_{j=1}^{n} g_{i_j}^*, \quad \text{vale} \quad \|\tilde{f}_n - f^*\|^2 \leq \frac{1}{n}\left(b^2 - \|f^*\|^2\right).$$

Si conclude usando la disuguaglianza triangolare e $\|f - f^*\| \leq \delta/n$:

$$\|\tilde{f}_n - f\| \leq \|f - f^*\| + \|\tilde{f}_n - f^*\| \leq \frac{1}{\sqrt{n}}\left(\frac{\delta}{\sqrt{n}} + \sqrt{b^2 - \|f^*\|^2}\right).$$

Elevando poi al quadrato si ottiene

$$\begin{aligned}\|\tilde{f}_n - f\|^2 &\leq \frac{1}{n}\left(b^2 - \|f^*\|^2 + \frac{\delta^2}{n} + 2\sqrt{b^2 - \|f^*\|^2}\,\frac{\delta}{\sqrt{n}}\right) = \\ &= \frac{1}{n}\left(b^2 - \|f^*\|^2 + \delta(\delta c_1 + c_2)\right),\end{aligned}$$

con $c_1, c_2 \in \mathbb{R}$ costanti.
Per arbitrarietà di δ, facendolo tendere a 0, si ha la tesi. □

Si può a questo punto procedere con la dimostrazione del risultato principale.

Dimostrazione del Teorema 2.3.1. Si definiscono anzitutto i seguenti sottoinsiemi di $L^2(B_r, \mu)$:

$$\mathcal{G}_\psi = \left\{\gamma\psi(\alpha\cdot x - \beta) \in \mathcal{N}_{k,1}(\psi) \;\middle|\; \alpha \in \mathbb{R}^k,\ \beta \in \mathbb{R},\ |\gamma| \leq 4\pi c\right\},$$

$$\mathcal{G}_{\cos} = \left\{\frac{\gamma}{r\|\alpha\|}\Big(\cos(2\pi(\alpha\cdot x + \beta)) - \cos(2\pi\beta)\Big) \;\middle|\; \alpha \neq 0 \in \mathbb{R}^k,\ \beta \in \mathbb{R},\ |\gamma| \leq c\right\},$$

$$\mathcal{G}_{\text{step}} = \left\{\gamma\mathbb{1}_{\{x \,\mid\, \alpha\cdot x - \beta \geq 0\}} \;\middle|\; \|\alpha\| = 1/r,\ |\beta| \leq 1,\ |\gamma| \leq 4\pi c\right\}.$$

A partire da $\mathcal{G}_{\text{step}}$ si definisce un suo sottoinsieme imponendo una ulteriore condizione su β, al variare di α.
Si consideri la funzione $z(x) = \alpha\cdot x$, $z : B_r \to \mathbb{R}$, in particolare essendo $x \in B_r$ per Cauchy-Schwarz vale

$$\|\alpha\| = \frac{1}{r} \ \text{ e } \ \|x\| \leq r \quad \Longrightarrow \quad |\alpha\cdot x| \leq 1,$$

dunque $z : (B_r, \mu) \to [-1, 1]$.
Si prende ora in considerazione l'immagine di z e la misura indotta su di essa da μ: $([-1,1], \mathbb{P}_z)$. In particolare si guarda alla funzione di ripartizione (CDF) di tale misura F_z.

Sulla base delle proprietà delle CDF si possono fare le seguenti considerazioni:

- dato $([-1,1],\mathbb{P}_z)$ e la sua CDF F_z, vale che $\beta \in [-1,1]$ è punto di discontinuità di F_z se e solo se $0 < \mathbb{P}_z(\beta) = \mu(\{x \mid \alpha \cdot x - \beta = 0\})$;
- $D_{F_z} = \{\text{insieme dei punti di discontinuità di } F_z\}$ è numerabile, perciò sottraendolo da $[-1,1]$ si ottiene un denso.

Definendo l'insieme $B_\alpha = [-1,1] \setminus D_{F_z}$ si ha dunque che

$$\forall\, \beta \in B_\alpha \quad \mu(\{x \mid \alpha \cdot x - \beta = 0\}) = 0, \tag{2.8}$$

inoltre B_α è denso in $[-1,1]$.
Si definisce quindi il seguente sottoinsieme di $\mathcal{G}_{\text{step}}$,

$$\mathcal{G}^{\mu}_{\text{step}} = \left\{ \gamma \mathbb{1}_{\{x \mid \alpha \cdot x - \beta \geq 0\}} \;\middle|\; \|\alpha\| = 1/r,\ |\beta| \leq 1 \wedge \beta \in B_\alpha,\ |\gamma| \leq 4\pi c \right\}.$$

Da questo punto la dimostrazione viene svolta in più fasi e fa uso del Lemma 2.3.1; per semplificarne la lettura è di seguito suddivisa in sezioni.

Strategia

Si dimostra che $\forall\, f \in \Gamma_{r,c}$ si ha che $f \in \overline{co(\mathcal{G}_\psi)} \subset L^2(B_r,\mu)$.

Fase 1

Si dimostra che $\forall\, f \in \Gamma_{r,c}$ si ha che $f \in \overline{co(\mathcal{G}_{\cos})} \subset L^2(B_r,\mu)$.

Si osserva anzitutto che se $f \in \Gamma_{r,c}$, allora f è una funzione reale per la quale vale lo sviluppo (1.2). Avendo assunto $f(0) = 0$ ed osservando che $\hat{f} : \mathbb{R}^k \to \mathbb{C}$ si può decomporre come

$$\hat{f}(\xi) = e^{2\pi i \theta(\xi)} |\hat{f}(\xi)| \text{ , con } \theta : \mathbb{R}^k \to \mathbb{R},$$

si ha

$$\begin{aligned} f(x) &= \operatorname{Re}\left(\int_{\mathbb{R}^k} (e^{2\pi i \xi \cdot x} - 1) \hat{f}(\xi)\, d\xi \right) = \\ &= \operatorname{Re}\left(\int_{\mathbb{R}^k} (e^{2\pi i \xi \cdot x} - 1) e^{2\pi i \theta(\xi)} |\hat{f}(\xi)|\, d\xi \right) = \\ &= \int_{\mathbb{R}^k} \Big(\cos(2\pi(\xi \cdot x + \theta(\xi))) - \cos(2\pi\theta(\xi)) \Big) |\hat{f}(\xi)|\, d\xi. \end{aligned}$$

Introducendo la funzione $g : \mathbb{R} \to \mathbb{R}$, $g(y) = \cos(2\pi(y + \theta(\xi)))$, si può osservare che parte dell'argomento dell'integrale è $g(\xi \cdot x) - g(0)$, dunque per Lagrange

$$|g(\xi \cdot x) - g(0)| = g'(z) |\xi \cdot x| \text{ , con } z \text{ tra } 0 \text{ e } \xi \cdot x;$$

in particolare $|g(\xi \cdot x) - g(0)| \leq 2\pi |\xi \cdot x| \leq 2\pi r \|\xi\|$.
Si denota inoltre

$$c_{f,r} := \int_{\mathbb{R}^k} r \|\xi\| |\hat{f}(\xi)|\, d\xi;$$

per ipotesi $c_{f,r} \leq c$. Moltiplicando e dividendo $|\hat{f}(\xi)|$ per $\frac{c_{f,r}}{r\|\xi\|}$ ed osservando che

$$\lambda(\xi) := \frac{r\|\xi\||\hat{f}(\xi)|}{c_{f,r}} = \frac{r\|\xi\||\hat{f}(\xi)|}{\int_{\mathbb{R}^k} r\|\xi\||\hat{f}(\xi)|\,d\xi}$$

è una densità di una misura di probabilità su $\mathbb{R}^k$, si può riscrivere $f(x)$ come segue[4]

$$\begin{aligned} f(x) &= \int_{\mathbb{R}^k\setminus\{0\}} \Big(g(\xi\cdot x) - g(0)\Big)\frac{c_{f,r}}{r\|\xi\|}\lambda(\xi)\,d\xi \leq \\ &\leq \int_{\mathbb{R}^k\setminus\{0\}} 2\pi r\|\xi\|\frac{c_{f,r}}{r\|\xi\|}\lambda(\xi)\,d\xi \leq \\ &\leq 2\pi c. \end{aligned}$$

La prima riga della lista di disequazioni precedenti permette di vedere f come combinazione convessa infinita di funzioni della forma

$$h(x,\xi) = \frac{c_{f,r}}{r\|\xi\|}\Big(\cos(2\pi(\xi\cdot x + \theta(\xi))) - \cos(2\pi\theta(\xi))\Big) \quad , \quad |h(x,\xi)| \leq 2\pi c,$$

che, definendo con $\alpha := \xi \neq 0$, $\beta := \theta(\xi) \in \mathbb{R}$ e $\gamma := c_{f,r} \leq c$, si può riscrivere come

$$\frac{\gamma}{r\|\alpha\|}\Big(\cos(2\pi(\alpha\cdot x + \beta)) - \cos(2\pi\beta)\Big) \in \mathcal{G}_{\cos}.$$

Da ciò segue che $f \in \overline{co(\mathcal{G}_{\cos})}$, per una legge dei grandi numeri in L^2.
Infatti date $\xi_1, \ldots, \xi_n$ variabili aleatorie vettoriali (k-dimensionali), indipendenti, con distribuzione $\lambda.\mathcal{L}^k$, si ha che $\forall\, i \in [n]$,

$$\mathbb{E}\left[h(x,\xi_i)\right] = \int_{\mathbb{R}^k\setminus\{0\}} \Big(\cos(2\pi(\xi\cdot x + \theta(\xi))) - \cos(2\pi\theta(\xi))\Big)\frac{c_{f,r}}{r\|\xi\|}\lambda(\xi)\,d\xi = f(x).$$

Dunque, per Fubini-Tonelli,

$$\begin{aligned} &\mathbb{E}\left[\int_{B_r}\left(f(x) - \frac{1}{n}\sum_{i=1}^n h(x,\xi_i)\right)^2 d\mu(x)\right] = \\ &= \int_{B_r}\mathbb{E}\left[\left(f(x) - \frac{1}{n}\sum_{i=1}^n h(x,\xi_i)\right)^2\right] d\mu(x) = \\ &= \int_{B_r}\mathrm{Var}\left(\sum_{i=1}^n h(x,\xi_i)\right) d\mu(x) = \\ &= \frac{1}{n}\int_{B_r}\mathrm{Var}\left(h(x,\xi_1)\right) d\mu(x) = \\ &= \frac{1}{n}\int_{B_r}\mathbb{E}\left[(h(x,\xi_1) - f(x))^2\right] d\mu(x) \leq \frac{16\pi^2c^2}{n}, \end{aligned}$$

[4]Di seguito viene rimosso lo 0 dal dominio di integrazione al fine di poter dividere per $\|\xi\|$.

dove per l'ultima disuguaglianza si è usato che sia h che f sono in modulo $\leq 2\pi c$. Si conclude allora che il valore atteso della norma $L^2(B_r, \mu)$ quadra della differenza tra una combinazione convessa aleatoria di n elementi di $\mathcal{G}_{\cos}$ ed f, converge a 0 per n che tende a $+\infty$. Da ciò si può dedurre, in maniera analoga a quanto fatto nella parte finale della dimostrazione del Lemma 2.3.1, che deve esistere una successione di combinazioni convesse di elementi di $\mathcal{G}_{\cos}$ che converge a f in $L^2(B_r, \mu)$, ovvero $f \in \overline{co(\mathcal{G}_{\cos})} \subset L^2(B_r, \mu)$.

Fase 2

Si dimostra che $\forall\, f \in \mathcal{G}_{\cos}$ si ha che $f \in \overline{co(\mathcal{G}_{\text{step}})} \subset L^2(B_r, \mu)$.

Ogni elemento $f \in \mathcal{G}_{\cos}$ è della forma

$$f(x) = \frac{\gamma}{r\|\alpha\|}\Big(\cos(2\pi(\alpha \cdot x + \beta)) - \cos(2\pi\beta)\Big),$$

dunque, è composizione di una funzione sinusoidale $g : \mathbb{R} \to \mathbb{R}$,

$$g(z) = \frac{\gamma}{r\|\alpha\|}\Big(\cos(2\pi(r\|\alpha\|z + \beta)) - \cos(2\pi\beta)\Big),$$

e di una funzione lineare $h : B_r \to \mathbb{R}$ (essendo $\alpha \neq 0$),

$$h(x) = \frac{1}{r\|\alpha\|}\alpha \cdot x. \tag{2.9}$$

In particolare, essendo $\|x\| \leq r$ e $|\alpha \cdot x| \leq \|x\|\|\alpha\|$, per Cauchy-Schwarz, si ha

$$|h(x)| \leq \frac{1}{r\|\alpha\|}\|x\|\|\alpha\| \leq 1,$$

dunque basta concentrarsi su $g : [-1, 1] \to \mathbb{R}$.
Essendo

$$g'(z) = -2\pi\gamma \sin(2\pi(r\|\alpha\|z + \beta))$$

si ha, poiché $\gamma \leq c$, che g è $2\pi c$-lipschitz su $[-1, 1]$; dunque è approssimabile in norma uniforme tramite funzioni semplici, per ogni successione di partizioni di $[-1, 1]$ con parametro di finezza convergente a 0.
In particolare considerando g ristretta a $[0, 1]$ [5], data $0 = t_0 < t_1 < \cdots < t_n = 1$ partizione di $[0, 1]$, si può definire

$$g_{n,+}(z) := \sum_{i=1}^{n-1}(g(t_i) - g(t_{i-1}))\mathbb{1}_{\{z \geq t_i\}}.$$

Tale funzione semplice interpola g nei punti t_i, $\forall\, i \in [n-1]$. $g_{n,+}$ è inoltre combinazione lineare di indicatrici e la somma dei moduli dei coefficienti moltiplicativi è tale che

$$\sum_{i=1}^{n-1}|g(t_i) - g(t_{i-1})| \leq \sum_{i=1}^{n-1} 2\pi c|t_i - t_{i-1}| = 2\pi c,$$

[5] Si può notare che $g(0) = 0$.

dove la minorazione vale per $2\pi c$-lipschitzianità.
Analogamente si può definire su $[-1,0]$

$$g_{n,-}(z) = \sum_{i=1}^{n-1} (g(-t_i) - g(-t_{i-1}))\mathbb{1}_{\{z \leq -t_i\}}$$

con le medesime proprietà.
Si può a questo punto definire una successione di funzioni semplici $g_{n,+} + g_{n,-}$, convergente in norma uniforme a g su $[-1,1]$ (imponendo che la successione dei parametri di finezza della famiglia di partizioni sia decrescente a 0). Vale che, $\forall\, n \in \mathbb{N}_0$, $g_{n,+} + g_{n,-}$ è combinazione lineare di funzioni indicatrici con la somma dei moduli dei coefficienti moltiplicativi $\leq 4\pi c$.
Si può osservare che ciascuna delle $g_{n,+} + g_{n,-}$ è combinazione di funzioni semplici, rispettivamente, della forma

$$\gamma \mathbb{1}_{\{z \geq \beta\}} \quad \text{e} \quad \gamma \mathbb{1}_{\{-z \geq \beta\}},$$

con $|\gamma| \leq 4\pi c$ e $|\beta| \leq 1$.
Componendo tali $g_{n,+} + g_{n,-}$ con $h(x)$ [6] si ottiene che f è approssimabile in norma uniforme con combinazioni lineari di funzioni della forma

$$\gamma \mathbb{1}_{\{\alpha \cdot x \geq \beta\}} \quad \text{e} \quad \gamma \mathbb{1}_{\{-\alpha \cdot x \geq \beta\}} \text{ , con } \|\alpha\| = \frac{1}{r},\ |\beta| \leq 1 \text{ e } |\gamma| \leq 4\pi c,$$

appartenenti dunque a $\mathcal{G}_{\text{step}}$.
Osservando infine che $\forall p \in L^2(B_r, \mu)$

$$\|p\|_2 = \left(\int_{B_r} |p(x)|^2 d\mu(x) \right)^{\frac{1}{2}} \leq \left(\int_{B_r} \|p\|_\infty^2 d\mu(x) \right)^{\frac{1}{2}} \leq \|p\|_\infty,$$

si ha che l'approssimazione si mantiene anche in $L^2(B_r, \mu)$, dunque, per quanto detto, vale $f \in \overline{co(\mathcal{G}_{\text{step}})} \subset L^2(B_r, \mu)$.

Fase 3

Si dimostra che $\forall\, f \in \mathcal{G}_{\cos}$ si ha che $f \in \overline{co(\mathcal{G}^\mu_{\text{step}})} \subset L^2(B_r, \mu)$.

Ripercorrendo la dimostrazione della fase precedente si nota che l'argomento è ancora valido anche se si impone di prendere le possibili partizioni di $[-1,1]$ nella forma $t_0 < t_1 < \cdots < t_n$ con $t_i \in B_\alpha$ sottoinsieme denso di $[-1,1]$, $\forall i \in [n]$.
Questa particolare restrizione è infatti irrilevante al fine di imporre che la successione dei parametri di finezza della famiglia di partizioni sia decrescente a 0.
Da ciò segue, analogamente alla Fase 2, che $f \in \overline{co(\mathcal{G}^\mu_{\text{step}})} \subset L^2(B_r, \mu)$.

Fase 4

Si dimostra che $\forall\, f \in \mathcal{G}^\mu_{\text{step}}$ si ha che $f \in \overline{co(\mathcal{G}_\psi)} \subset L^2(B_r, \mu)$ (se $|\psi| \leq 1$).

Per fare ciò, fissata $f \in \mathcal{G}^\mu_{\text{step}}$ si costruisce una successione di funzioni

$$\forall\, n \in \mathbb{N}_0, \quad f_n(x) = \gamma\psi(n(\alpha \cdot x - \beta)) \in \mathcal{G}_\psi,$$

[6] Guardando alla definizione di h, (2.9), si può notare che si potrebbe riscrivere come $h(x) = \alpha \cdot x$ con $\|\alpha\| = 1/r$.

che, per $n \to \infty$, converge a $f(x) = \gamma \mathbb{1}_{\{x \mid \alpha \cdot x \geq \beta\}}$, $\forall\, x$ tale che $\alpha \cdot x \neq \beta$. Tuttavia questa limitazione sulla convergenza non è un problema, poiché per quanto visto precedentemente, (2.8), se

$$\gamma \mathbb{1}_{\{x \mid \alpha \cdot x \geq \beta\}} \in \mathcal{G}^{\mu}_{\text{step}} \quad \Longrightarrow \quad \beta \in B_\alpha \ \text{ e } \ \mu(\{x \mid \alpha \cdot x = \beta\}) = 0.$$

Dunque vale che $\widetilde{\forall} x,\ f_n(x) \xrightarrow{n \to \infty} f(x)$. Inoltre, $\forall\, n \in \mathbb{N}_0$ e $\forall x \in B_r$, $|f_n(x)| \leq |\gamma| \leq 4\pi c$, in più ci si trova in uno spazio di probabilità, perciò

$$\left(\int_{B_r} |4\pi c|^2 \, d\mu(x) \right)^{\frac{1}{2}} = 4\pi c < \infty,$$

allora, per la convergenza dominata di Lebesgue generalizzata, vale che f_n converge ad f anche in $L^2(B_r, \mu)$. Segue che $f \in \overline{co(\mathcal{G}_\psi)} \subset L^2(B_r, \mu)$.

Concatenando le 4 fasi precedenti, notando che

$$\mathcal{G}_{\cos} \subset \overline{co(\mathcal{G}^{\mu}_{\text{step}})} \quad \Longrightarrow \quad \overline{co(\mathcal{G}_{\cos})} \subset \overline{co(\mathcal{G}^{\mu}_{\text{step}})}$$

e che, analogamente,

$$\overline{co(\mathcal{G}^{\mu}_{\text{step}})} \subset \overline{co(\mathcal{G}_\psi)},$$

si conclude quanto si voleva dimostrare nella strategia, ossia:

$$\Gamma_{r,c} \subset \overline{co(\mathcal{G}_{\cos})} \subset \overline{co(\mathcal{G}^{\mu}_{\text{step}})} \subset \overline{co(\mathcal{G}_\psi)}.$$

Completamento

Supponendo che $\forall\, x \in \mathbb{R},\ |\psi(x)| \leq 1$ ed osservando che

$$\forall\, h \in \mathcal{G}_\psi, \quad \|h\|_2 = \left(\int_{B_r} |\gamma \psi(\alpha \cdot x - \beta)|^2 \, d\mu \right)^{\frac{1}{2}} \leq 4\pi c,$$

per il Lemma 2.3.1, si ha che $\forall\, n \in \mathbb{N}_0$ e $\forall\, c' > (4\pi c)^2 - \|f\|_2^2$, esiste una $f_n \in co(\mathcal{G}_\psi)$ combinazione convessa di n punti di $\mathcal{G}_\psi$ tale che

$$\int_{B_r} (f - f_n)^2 \, d\mu \leq \frac{c'}{n},$$

dunque in particolare

$$\int_{B_r} (f - f_n)^2 \, d\mu \leq \frac{(4\pi c)^2}{n}.$$

Inoltre f_n è della forma

$$f_n = \sum_{i=1}^{n} \lambda_i g_i, \text{ con } g_i \in \mathcal{G}_\psi,\ \lambda_i \geq 0 \text{ e } \sum_{i=1}^{n} \lambda_i = 1,$$

quindi $f_n \in \mathcal{N}_{k,n}(\psi)$.
In particolare

$$\forall\, i \in [n], \quad g_i(x) = \gamma_i \psi(\alpha_i \cdot x - \beta_i) \text{ , con } |\gamma_i| < 4\pi c,$$

si ha dunque che $\exists A, b, c \in \mathbb{R}^{n\times k} \times \mathbb{R}^n \times \mathbb{R}^n$ tali che

$$f_n = \sum_{i=1}^{n} c_i \psi\left(A_i \cdot x - b_i\right), \text{ con } \sum_{j=1}^{n} |c_i| \leq 4\pi c,$$

ovvero la tesi del Teorema 2.3.1.
Se invece $|\psi(x)| > 1$ per qualche x, si deve utilizzare il Lemma 2.3.1 e, dapprima, solo le prime 3 fasi della strategia: $\Gamma_{r,c} \subset \overline{co(\mathcal{G}^{\mu}_{\text{step}})}$.
Da ciò si può ricavare, con passaggi analoghi a quelli fatti sopra, che esiste una successione di combinazioni convesse di n elementi di $\mathcal{G}^{\mu}_{\text{step}}$, $(f_n)_{n=1}^{\infty}$ tale che, $\forall\, n \in \mathbb{N}_0$,

$$\int_{B_r} (f - f_n)^2 \, d\mu \leq \frac{(4\pi c)^2 - 1/2\|f\|_2^2}{n}.$$

A questo punto si può utilizzare la Fase 4 per rimpiazzare, con delle approssimazioni sufficientemente accurate, le f_n ottenute prima con delle $\overline{f}_n \in \mathcal{G}_{\psi}$ in modo che valga la stima

$$\int_{B_r} (f - \overline{f}_n)^2 \, d\mu \leq \frac{(4\pi c)^2}{n},$$

in questo modo si arriverebbe nuovamente alla tesi del Teorema 2.3.1.

□

Quanto appena dimostrato permette di stimare quantitativamente la capacità delle *NN* di approssimare le funzioni presenti nella classe $\Gamma_{r,c}$. Si può notare che tale classe di funzioni si può semplificare, ad esempio rimuovendo le ipotesi di f integrabile e $\text{supp}\,(f) \subset B_r$ e rimpiazzandole con $f \in C_c(\mathbb{R}^k, \mathbb{R})$; infatti se f ha supporto $K \subset \mathbb{R}^k$ compatto allora $\exists\, r \in \mathbb{R}$ tale che $\text{supp}\,(f) \subset B_r$ ed in particolare

$$\int_{B_r} |f(x)| \, d\mathcal{L}^k(x) \leq \max_{x\in K} |f(x)| \, \mathcal{L}^k(B_r) < \infty. \tag{2.10}$$

Per la minorazione (2.10) si è usato il Teorema di Weierstrass, per affermare che essendo $|f|$ continua ed a supporto compatto ammette massimo, ed anche che la misura di Lebesgue è una misura di Radon, se ristretta ai soli boreliani di $\mathbb{R}^k$, dunque $\mathcal{L}^k(B_r) < \infty$.
Tuttavia anche dopo una simile semplificazione non è del tutto immediato capire quando una f appartenga a

$$\overline{\Gamma}_{r,c} := \left\{ f : \mathbb{R}^k \to \mathbb{R} \;\middle|\; f \in C_c(\mathbb{R}^k, \mathbb{R}) \;\wedge\; \hat{f} \in L^1(\mathbb{R}^k, \mathcal{L}^k) \;\wedge\; \int_{\mathbb{R}^k} \|\xi\| |\hat{f}(\xi)| \, d\xi \leq \frac{c}{r} \right\}.$$

Si presenta quindi una lista non esaustiva di alcune tipologie di funzioni $f \in \overline{\Gamma}_{r,c}$. I seguenti fatti, riportati senza dimostrazione per non appesantire la trattazione, sono stati provati, insieme ad altri risultati analoghi, da Barron [2, pp. 939–942].

- $\forall\, k \in \mathbb{N}_0$, $\forall\, \alpha \in \mathbb{R}^k$, $\forall\, r \in \mathbb{R}$,

$$f(x) = \begin{cases} \alpha \cdot x & \text{se } x \in B_r \\ 0 & \text{altrimenti} \end{cases},$$

  è tale che $f \in \overline{\Gamma}_{r,\|\alpha\| r}$.

- Se $g \in \overline{\Gamma}_{r,c}$ ed in particolare [7]

$$\int_{\mathbb{R}^k} |\hat{g}(\xi)|\, d\xi \leq a,$$

allora $g^k \in \overline{\Gamma}_{r,ka^{k-1}c}$.
Più in generale, dato un polinomio $f(x) = a_n x^n + \cdots + a_0$, denotando con $f_{\text{abs}} := |a_n|x^n + \cdots + |a_0|$, si ha $f \circ g \in \overline{\Gamma}_{r,f'_{\text{abs}}(a)c}$.

- Data una densità gaussiana

$$f(x) = e^{-\frac{\|x\|^2}{2}}, \quad \text{con} \quad f : \mathbb{R}^k \to \mathbb{R},$$

allora $\forall\, r \in \mathbb{R},\ f \in \overline{\Gamma}_{r,k^{\frac{1}{2}}}$.

2.4. Attivazione di tipo ReLU

I precedenti teoremi di approssimazione universale sono stati enunciati tutti per funzioni di attivazione ψ sigmoidali, tuttavia nella pratica non tutte le funzioni di attivazione usate nelle definizioni di reti neurali sono sigmoidali.
Di particolare rilievo è la cosiddetta funzione ReLU (Rectified Linear Unit), ovvero $\text{ReLU} : \mathbb{R} \to \mathbb{R}$ tale che

$$\text{ReLU}(x) = \begin{cases} x & \text{se } x \geq 0 \\ 0 & \text{altrimenti} \end{cases}. \tag{2.11}$$

Tale funzione non rientra nemmeno nella classe delle funzioni misurabili e limitate. Si ricorda che per tale classe la dimostrazione del Teorema 2.2.1 può essere estesa poiché anche queste, come le sigmoidali, sono in realtà discriminatorie (Osservazione 2.1.1).
Si consideri, però, una *NN* con 2 hidden layers (anziché uno, come fatto fin ora):

$$\begin{aligned}
&\Phi^{(1)}(x) = A^{(1)}x + b^{(1)},\\
&\overline{\Phi}^{(1)}(x) = \text{ReLU}(\Phi^{(1)}(x)),\\
&\Phi^{(2)}(x) = A^{(2)}\overline{\Phi}^{(1)}(x) + b^{(2)},\\
&\overline{\Phi}^{(2)}(x) = \text{ReLU}(\Phi^{(2)}(x)),\\
&\Phi^{(3)}(x) = A^{(3)}\overline{\Phi}^{(2)}(x) + b^{(3)}.
\end{aligned}$$

Si può supporre $n_1 = n_2$ e che

$$A^{(2)} = -I, \quad b^{(2)} = e,$$

dove con I si indica la matrice identitaria di taglia $n_1 \times n_1$ e con e si indica il vettore colonna con 1 in ogni componente.
Se, a questo punto, si considerano la prima attivazione, la seconda pre-attivazione e la terza attivazione come una unica funzione di attivazione

$$\psi'(\cdot) = \text{ReLU} \circ(- \cdot +1) \circ \text{ReLU},$$

[7] Essendo $g \in \overline{\Gamma}_{r,c}$, $\hat{f} \in L^1(\mathbb{R}^k, \mathcal{L}^k)$, dunque esiste necessariamente un tale a.

si ottiene la seguente rete:

$$\begin{aligned}
&\Phi'^{(1)}(x) = A^{(1)}x + b^{(1)},\\
&\overline{\Phi'}^{(1)}(x) = \psi'\big(\Phi'^{(1)}(x)\big) = \mathrm{ReLU}\Big(- I\Big[\,\mathrm{ReLU}\left(\Phi'^{(1)}(x)\right)\Big] + e\Big),\\
&\Phi'^{(2)}(x) = A^{(3)}\overline{\Phi'}^{(1)}(x) + b^{(3)}.
\end{aligned}$$

Si noti che per arbitrarietà di $A^{(3)}$ e $b^{(3)}$, considerando l'opposta della matrice ed un traslato del bias nell'ultima funzione di pre-attivazione,

$$\begin{aligned}
&A''^{(1)} = A^{(1)}, \quad A''^{(2)} = -A^{(3)},\\
&b''^{(1)} = b^{(1)}, \quad b''^{(2)} = A^{(3)}e + b^{(3)},
\end{aligned}$$

si può anche riscrivere la rete con la seguente attivazione

$$\psi(x) = -\,\mathrm{ReLU}\Big(- \mathrm{ReLU}(x) + 1\Big) + 1.$$

Solitamente si denomina tale funzione ReLU1 $:= \psi$.
Da ciò segue la Proposizione 2.4.1.

Proposizione 2.4.1. *Una rete neurale con due hidden layers (equivalentemente $2m$) e funzione di attivazione* ReLU*, ammette, come configurazione particolare, una struttura riconducibile ad una NN con un solo hidden layer (equivalentemente m) e funzione di attivazione* ReLU1*.*

Si introduce a questo punto una notazione simile a quella presentata in (2.1) per descrivere l'insieme delle funzioni implementabili per mezzo di reti neurali con due hidden layers aventi architettura fissata $\alpha = ((k, m_1, m_2, 1), \psi)$, al variare di

$$\theta = (A^{(1)}, b^{(1)}, A^{(2)}, b^{(2)}, c), \in \mathbb{R}^{m_1 \times k} \times \mathbb{R}^{m_1} \times \mathbb{R}^{m_2 \times m_1} \times \mathbb{R}^{m_2} \times \mathbb{R}^{m_2}.$$

Si definisce quindi

$$\mathcal{N}^{(2)}_{k,m_1,m_2}(\psi) = \left\{ h : \mathbb{R}^k \to \mathbb{R} \;\middle|\; h(x) = c \cdot \psi\left(A^{(2)}\psi\left(A^{(1)}x + b^{(1)}\right) + b^{(2)}\right)\right\} \quad (2.12)$$

e, in maniera analoga a quanto già fatto, si può denotare l'insieme

$$\mathcal{N}^{(2)}_{k}(\psi) = \bigcup_{m_2=1}^{\infty}\bigcup_{m_1=1}^{\infty} \mathcal{N}^{(2)}_{k,m_1,m_2}(\psi)$$

di tutte le funzioni generabili.
Tramite l'equazione (2.12) si può pertanto riscrivere la Proposizione 2.4.1 con il seguente contenimento

$$\mathcal{N}_{k,m}(\mathrm{ReLU1}) \subset \mathcal{N}^{(2)}_{k,m,m}(\mathrm{ReLU}).$$

Notando inoltre che ReLU1 è una attivazione continua (dunque anche misurabile) e sigmoidale, valgono i seguenti corollari ai teoremi di densità enunciati fino a questo punto.

Corollario 2.4.1 (Teorema di densità in L^p con ReLU)**.** *Fissata* ReLU *funzione di attivazione, vale che* $\mathcal{N}_k^{(2)}(\text{ReLU})$ *è denso in* $L^p(\mathbb{R}^k, \mu)$ $(\forall\, p \in [1, +\infty))$ *per ogni* μ *misura finita su* $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$.

Dimostrazione. Per il Teorema 2.1.1, essendo ReLU1 misurabile e sigmoidale, si ha che $\overline{\mathcal{N}_k(\text{ReLU1})} = L^p(\mathbb{R}^k, \mu)$, $\forall\, p \in [1, +\infty)$ e $\forall\, \mu$ misura finita su $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$. Inoltre, osservando che valgono

$$\mathcal{N}_k(\text{ReLU1}) = \bigcup_{m=1}^{\infty} \mathcal{N}_{k,m}(\text{ReLU1}) \quad \text{e} \quad \mathcal{N}_k^{(2)}(\text{ReLU}) = \bigcup_{m_2=1}^{\infty} \bigcup_{m_1=1}^{\infty} \mathcal{N}_{k,m_1,m_2}^{(2)}(\text{ReLU}),$$

per la Proposizione 2.4.1 si ottiene $\overline{\mathcal{N}_k(\text{ReLU1})} \subset \overline{\mathcal{N}_k^{(2)}(\text{ReLU})}$, dunque anche $\overline{\mathcal{N}_k^{(2)}(\text{ReLU})} = L^p(\mathbb{R}^k, \mu)$. □

Corollario 2.4.2 (Teorema di approssimazione universale con ReLU)**.** *Fissata* ReLU *funzione di attivazione, vale che* $\mathcal{N}_k^{(2)}(\text{ReLU})$ *è denso in* $(C(X), \|\cdot\|_\infty)$ *per ogni* $X \subset \mathbb{R}^k$ *compatto.*

Dimostrazione. La dimostrazione è analoga a quella del Corollario 2.4.1. Si usa ReLU1 continua e sigmoidale per poter applicare il Teorema 2.2.1, si osserva poi che per la Proposizione 2.4.1 si ha $\overline{\mathcal{N}_k(\text{ReLU1})} \subset \overline{\mathcal{N}_k^{(2)}(\text{ReLU})}$ e mettendo insieme i due risultati si ha la tesi. □

Corollario 2.4.3 (Teorema di approssimazione quantitativa con ReLU)**.** *Fissata* ReLU *funzione di attivazione,* μ *misura di probabilità su* B_r*. Allora* $\forall\, f \in \Gamma_{r,c}$*, fissato* $n \in \mathbb{N}_0$*, esiste una funzione* $f_n \in \mathcal{N}_{k,n,n}^{(2)}(\text{ReLU})$ *tale che*

$$\int_{B_r} (f - f_n)^2 \, d\mu \leq \frac{(4\pi c)^2}{n}.$$

Dimostrazione. Per il Teorema 2.3.1, data μ misura di probabilità su B_r allora $\forall\, f \in \Gamma_{r,c}$, fissato $n \in \mathbb{N}_0$, esiste una funzione $f_n' \in \mathcal{N}_{k,n}(\text{ReLU1})$ tale che valga la disequazione nella tesi (essendo ReLU1 sigmoidale). Usando poi la Proposizione 2.4.1, si ottiene che $f_n' \in \mathcal{N}_{k,n}(\text{ReLU1})$ si può pensare come una $f_n \in \mathcal{N}_{k,n,n}^{(2)}(\text{ReLU})$, ovvero la tesi. □

Capitolo 3

Convergenza in legge ad un processo gaussiano

Si consideri una *NN* definita da un'architettura $\alpha = (n, \psi)$, come nella Definizione 2.0.1. Per brevità di notazione si omettono di seguito le dipendenze da α e da θ (parametro della rete), che risultano chiare dal contesto, nella scrittura della funzione di realizzazione; ossia $\Phi := \Phi_\alpha$ [1]. In tal modo si ha che le funzioni denotanti le attivazioni dei layers μ, $\forall\, i \in [l]$, sono così caratterizzabili:

- per $\mu = 1$, in particolare per la prima pre-attivazione, si ha

$$\Phi^{(1)}(x) = A^{(1)}x + b^{(1)}.$$

  Per esteso, $\forall i \in [n_1]$,

$$\Phi_i^{(1)}(x) = A_i^{(1)}x + b_i^{(1)} = \sum_{j=1}^{n_0} A_{i,j}^{(1)} x_j + b_i^{(1)}.$$

- per tutti gli altri layers, per $\mu \in [l-1]$, vale

$$\begin{aligned} \bar{\Phi}^{(\mu)}(x) &= \psi(\Phi^{(\mu)}(x)), \\ \Phi^{(\mu+1)}(x) &= A^{(\mu+1)}\bar{\Phi}^{(\mu)}(x) + b^{(\mu+1)}. \end{aligned}$$

  Analogamente a prima, per esteso, si può scrivere

$$\begin{aligned} &\bar{\Phi}_i^{(\mu)}(x) = \psi(\Phi_i^{(\mu)}(x)), && \forall i \in [n_\mu], \\ &\Phi_i^{(\mu+1)}(x) = A_i^{(\mu+1)}\bar{\Phi}^{(\mu)}(x) + b_i^{(\mu+1)} = \sum_{j=1}^{n_\mu} A_{i,j}^{(\mu+1)}\bar{\Phi}_j^{(\mu)}(x) + b_i^{(\mu+1)}, && \forall i \in [n_{\mu+1}]. \end{aligned}$$

3.1. Parametri delle reti

In questo capitolo viene assunto θ casuale con una particolare distribuzione di probabilità che, assieme al valore dell'input $x \in \mathbb{R}^{n_0}$, indurrà determinate distribuzioni sulle attivazioni e pre-attivazioni. Ricordando che

$$\theta = (\theta_\mu)_{\mu=1}^{l} = (A^{(\mu)}, b^{(\mu)})_{\mu=1}^{l},$$

si assume di scegliere con distribuzioni gaussiane indipendenti ciascun $A^{(\mu)}$ e $b^{(\mu)}$ come segue:

$$\begin{aligned} A_{i,j}^{(\mu)} &\sim \mathcal{N}^1\left(0,\, C_A^{(\mu)}\right) \quad \forall\, \mu \in [l], \\ b_i^{(\mu)} &\sim \mathcal{N}^1\left(0,\, C_b^{(\mu)}\right) \quad \forall\, \mu \in [l]. \end{aligned} \tag{3.1}$$

[1] Si ricorda che Φ_α è definito nella Definizione 2.0.1: $\Phi_\alpha = \Phi_\alpha^{(l)}$.

Nel corso del capitolo viene studiato il comportamento della rete al tendere di n_i, $\forall\, \mu \in [l-1]$ hidden layer, a infinito (si lasceranno invariate le dimensioni di input ed output). Dunque viene assunta una riscalatura sulle varianze $C_A^{(\mu)}$ per evitare che, in tale limite, ci sia una divergenza nelle varianze delle pre-attivazioni[2]:

$$C_A^{(\mu)} = \frac{\hat{C}_A^{(\mu)}}{n_{\mu-1}} \quad \forall\, \mu \in [l].$$

Si assumono $\hat{C}_A^{(\mu)}$ e $C_b^{(\mu)}$ costanti nel limite della taglia degli hidden layers.

Per dare una enunciazione precisa del risultato principale è necessario definire delle reti neurali *dinamiche*, che siano parametrizzate per ampiezza dei layers, ovvero che $\forall\, m$ input abbiano una diversa architettura $\alpha = (n, \psi)$. In particolare la ψ rimarrà invariata per ogni m mentre si avrà $n = (n_0(m), \dots, n_l(m))$.
Per far ciò si ricorre alla seguente definizione.

Definizione 3.1.1 (Funzione di crescita di un layer)**.** Data una *NN* ed un layer μ su tale rete, si definisce funzione di crescita del suddetto layer $h_\mu : \mathbb{N} \to \mathbb{N}$ tale che $\forall\, n$, $h_\mu(n)$ specifica la larghezza del layer al variare del parametro di crescita: $h_\mu(n) = n_\mu(n)$.

Inoltre, a differenza di quanto avviene nelle precedenti sezioni, il seguente enunciato non è valido per le sole reti neurali con funzione d'attivazione sigmoidale, essendo sufficiente la richiesta più debole di attivazioni sub-lineari, definite come segue.

Definizione 3.1.2. Una funzione $\psi : \mathbb{R} \to \mathbb{R}$ si definisce sub-lineare se $\exists\, m, q \in \mathbb{N}$ tali che

$$\forall\, x \in \mathbb{R}, \quad |\psi(x)| \leq m|x| + q.$$

Osservazione 3.1.1. ψ funzione di attivazione continua e sigmoidale è sub-lineare. Infatti, ricordando la Definizione 2.1.1, si ha che, per continuità, ψ è limitata, ovvero $\exists\, q \in \mathbb{N}$ tale che $\forall\, x \in \mathbb{R}$, $|\psi(x)| \leq q$, dunque si ha la sub-linearità con $m = 0$.

Osservazione 3.1.2. La funzione di attivazione ReLU è sub-lineare. Sulla base della definizione (2.11) è immediato notare che $\forall\, x \in \mathbb{R}$, $|\,\mathrm{ReLU}(x)| \leq |x|$, ovvero la tesi per $m = 1$ e $q = 0$.

3.2. Risultato principale

Prima di enunciare il Teorema di convergenza ad un processo gaussiano è necessario dare la definizione di processo gaussiano.

Definizione 3.2.1. Si definisce processo gaussiano (numerabile) un processo stocastico $X = (X_i)_{i \in \mathbb{N}}$ tale che, dato un numero finito di variabili aleatorie nella collezione che forma il processo, esse costituiscono un vettore gaussiano:

$$\forall\, k \in \mathbb{N},\, (i_j)_{j=1}^k \;\text{ vale }\; (X_{i_j})_{j=1}^k \sim \mathcal{N}^k\left(m,\, K\right).$$

[2]Si noti che n_0 è fissato dunque non sarebbe necessario normalizzare $C_A^{(1)}$.

Teorema 3.2.1 (Teorema di convergenza ad un processo gaussiano)**.** *Si consideri una NN definita tramite una funzione di attivazione ψ continua e sub-lineare con parametri scelti con distribuzione gaussiana, come in* (3.1)*. Vale che, per ogni famiglia di funzioni di crescita strettamente crescenti degli hidden layers $(h_\mu(n))_{\mu=1}^{l-1}$ e per ogni insieme di input numerabile $(x[i])_{i=1}^{\infty}$ (con $x[i] \in \mathbb{R}^{n_0}$), la distribuzione dell'output della rete converge in legge ad un processo gaussiano per $n \to \infty$, con n parametro di crescita della rete*

$$(\Phi^{(l)}(x[i])[n])_{i=1}^{\infty} \xrightarrow{n\to\infty} \mathcal{N}^{\infty}\left(0,\, K^{(*)}\right).$$

Osservazione 3.2.1. La convergenza in legge nell'enunciato del Teorema 3.2.1 va intesa in relazione ad una topologia su $\mathbb{R}^{\mathbb{N}}$. Più in generale, su tutti gli spazi astratti X, la convergenza in legge può essere definita soltanto in relazione ad una topologia definita sullo spazio, al fine di dare una precisa definizione di integrazione e misurabilità. Nel caso in questione viene considerata la topologia prodotto di una quantità numerabile di copie di $\mathbb{R}$, che si può mostrare essere indotta dalla metrica ρ:

$$\rho(v,w) = \sum_{n=1}^{\infty} \frac{\min(1, |v_n - w_n|)}{2^n}, \qquad \forall v, w \in \mathbb{R}^{\mathbb{N}}.$$

A partire da questa osservazione, considerando lo spazio $\mathbb{R}^Q$ con Q numerabile e munendolo di una distanza del tutto analoga a quella definita per $\mathbb{R}^{\mathbb{N}}$, denotata anch'essa con ρ, si può dare una condizione equivalente di convergenza di variabili aleatorie su $\mathbb{R}^Q$ che sfrutti tutte le combinazioni lineari finite delle variabili stesse. Per poter procedere in tale direzione è necessario introdurre il concetto di famiglia tesa di misure di probabilità.

Definizione 3.2.2. Una famiglia di misure di probabilità su $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, $(\mu_i)_{i\in I}$, si dice tesa se $\forall\, \varepsilon > 0$, $\exists\, C_\varepsilon \subset \mathbb{R}$ compatto tale che, $\forall\, i \in I$, $\mu_i\,(C_\varepsilon) \geq 1 - \varepsilon$.

Vale, a questo punto, il seguente Lemma 3.2.1.

Lemma 3.2.1. *Dato un insieme numerabile Q e una successione di v.a. $U_n : (X, \mathcal{A}, \mathbb{P}) \to (\mathbb{R}^Q, \rho)$, indicizzate su $n \in \mathbb{N}$ ($(U_n)_{n\in\mathbb{N}}$), allora le seguenti condizioni sono equivalenti:*

- *$U_n \xrightarrow{\mathcal{L}} U_*$ rispetto alla metrica ρ su $\mathbb{R}^Q$;*
- *definendo con $U_n(u)$ "la u-esima componente" di U_n* [3] *(analogamente $U_*(u)$),*

$$\forall\, L \subset Q,\ |L| < \infty \ e\ \forall\, \alpha \in \mathbb{R}^{|L|}, \qquad \sum_{u\in L} \alpha_u U_n(u) \xrightarrow{\mathcal{L}} \sum_{u\in L} \alpha_u U_*(u).$$

Dimostrazione. La dimostrazione si compone di due sezioni principali:

1. la condizione
$$U_n \xrightarrow{\mathcal{L}} U_* \text{ rispetto alla metrica } \rho \text{ su } \mathbb{R}^Q,$$
è equivalente a
$$\forall\, L \subset Q,\ |L| < \infty, \qquad U_n(L) \xrightarrow{\mathcal{L}} U_*(L),$$
dove si definisce $U_n(L) = (U_n(u))_{u\in L}$ (analogamente $U_*(u)$);

[3] Ovvero se $u \in Q$ e $\pi_u : \mathbb{R}^Q \to \mathbb{R}$ è la proiezione di $\mathbb{R}^Q$ sulla componente u, $U_n(u) := \pi_u \circ U_n$.

2. la condizione
$$U_n(L) \xrightarrow{\mathcal{L}} U_*(L),$$
definendo $k = |L|$, è equivalente a
$$\forall\, \alpha \in \mathbb{R}^k, \qquad \sum_{u\in L} \alpha_u U_n(u) \xrightarrow{\mathcal{L}} \sum_{u\in L} \alpha_u U_*(u).$$

Il punto 2 è una diretta conseguenza del Teorema di continuità di Lévy:
$$U_n(L) \xrightarrow{\mathcal{L}} U_*(L) \qquad \Longleftrightarrow \qquad \forall\, t \in \mathbb{R}^k \quad \varphi_{U_n(L)}(t) \to \varphi_{U_*(L)}(t). \tag{3.2}$$

Infatti, si ricorda che data X v.a. vettoriale si ha
$$\varphi_X(t) = \varphi_{t\cdot X}(1)$$
e, valutando in $t = s\alpha$, l'espressione a destra nell'equazione (3.2) si può riscrivere come
$$\forall\, \alpha \in \mathbb{R}^k,\, s \in \mathbb{R}, \qquad \varphi_{\alpha\cdot U_n(L)}(s) \to \varphi_{\alpha\cdot U_*(L)}(s). \tag{3.3}$$
Si usa nuovamente il Teorema di continuità di Lévy per dire che l'equazione (3.3) è equivalente alla condizione
$$\forall\, \alpha \in \mathbb{R}^k, \qquad \alpha \cdot U_n(L) \xrightarrow{\mathcal{L}} \alpha \cdot U_*(L).$$
Scrivendo dunque esplicitamente $U_n(L) = (U_n(u))_{u\in L}$ (analogamente $U_*(u)$) vale
$$\forall\, \alpha \in \mathbb{R}^k, \qquad \alpha \cdot U_n(L) = \sum_{u\in L} \alpha_u U_n(u)$$

Si ha dunque la tesi poiché si hanno tutte equivalenze.
Per quel che riguarda il punto 1 si vuole sfruttare la seguente caratterizzazione di convergenza in legge: data $(X_n)_{n\in\mathbb{N}}$ successione di v.a. a valori in $\mathbb{R}^k$ (eventualmente $k = +\infty$) si ha che $X_n \xrightarrow{\mathcal{L}} X$ se e solo se
$$\forall\, f \in C_b(\mathbb{R}^k), \qquad \mathbb{E}\left[f(X_n)\right] \to \mathbb{E}\left[f(X)\right].$$

Questa caratterizzazione deriva dal fatto che la convergenza in legge è definita come la convergenza stretta della successione di misure immagine associate alle variabili aleatorie[4].
Dunque, per dimostrare 1, è sufficiente mostrare che la condizione
$$\forall\, f \in C_b(\mathbb{R}^{\mathbb{N}}), \qquad \mathbb{E}\left[f(U_n)\right] \to \mathbb{E}\left[f(U_*)\right], \tag{3.4}$$
è equivalente a
$$\forall\, L \subset Q,\ |L| = k < \infty,\ \forall\, f \in C_b(\mathbb{R}^k), \qquad \mathbb{E}\left[f(U_n(L))\right] \to \mathbb{E}\left[f(U_*(L))\right]. \tag{3.5}$$

[4]Si ricorda che essendo la famiglia di misure immagine una famiglia di misure di probabilità, così come la misura limite, in realtà le definizioni di convergenza vaga, debole e stretta coincidono.

Chiaramente (3.4) $\implies$ (3.5). Per l'altra implicazione è necessario tenere presente che ogni famiglia finita di misure di probabilità è tesa. Dunque ciascuna misura immagine di $U_n(l)$, $l \in Q$, è tesa. A questo punto si fa uso della seguente proprietà: se ogni distribuzione marginale è tesa allora tutta la famiglia di misure immagine è tesa. Quanto detto segue dal fatto che un prodotto arbitrario di spazi compatti è compatto. Si procede a questo punto applicando il Teorema di Prokhorov ad ogni sottosuccessione di $(U_n)_{n\in\mathbb{N}}$: $\forall\, (U_{n_k})_{k\in\mathbb{N}}$ sottosuccessione esiste una sotto-sottosuccessione $(U_{n_{k_j}})_{j\in\mathbb{N}}$ convergente in legge. Il limite di tutte le sotto-sottosuccessioni coincide, in quanto univocamente determinato dal limite delle $(U_n(L))_{n\in\mathbb{N}}$, $L \subset Q$, $|L| < \infty$. Infine applicando il criterio sotto-sotto si può concludere $U_n \xrightarrow{\mathcal{L}} U_*$.
Una dimostrazione alternativa di (3.5) $\implies$ (3.4) è presentata da Billingsley [3, p. 19]. □

Dimostrazione del Teorema 3.2.1. Si definisce $Q := X \times \mathbb{N}$ dove $X = (x[i])_{i=1}^{\infty}$ è l'insieme degli input, dunque Q è numerabile.
Fissato $n \in \mathbb{N}$ parametro di crescita della *NN*, si introducono le seguenti definizioni per alleggerire la notazione:

$$\forall\, \mu = 2, \dots, l-1, \qquad U_n^{(\mu)} := (\Phi_i^{(\mu)}(x)[n] - b_i^{(\mu)})_{(x,i)\in Q}, \tag{3.6}$$

dove $\Phi_i^{(\mu)}(x)[n]$ non è altro che $\Phi_i^{(\mu)}(x)$ in corrispondenza del parametro n,

$$\Phi_i^{(\mu)}(x)[n] = \sum_{j=1}^{h_{\mu-1}(n)} A_{i,j}^{(\mu)} \bar{\Phi}_j^{(\mu-1)}(x) + b_i^{(\mu)}.$$

Inoltre, anche per $\mu = l$ si può dare una definizione analoga, anche se la profondità dell'ultimo layer, al variare di n, è costantemente n_l. Si pone perciò in questo caso $Q := X \times [n_l]$ e

$$U_n^{(l)} := (\Phi_i^{(l)}(x)[n] - b_i^{(l)})_{x\in X,\, i\in[n_l]}.$$

Per non dover separare ogni volta il caso $\mu = l$ verrà indicato successivamente $U_n^{(l)}$ con la stessa notazione che si adotta $\forall\, \mu = 2, \dots, l-1$, facendo eventualmente attenzione ai casi in cui la differenza è sostanziale.
Dunque, ricordando che i parametri $(A^{(\mu)}, b^{(\mu)})_{\mu=1}^{l}$ sono v.a. indipendenti con distribuzione come in (3.1), e supponendo $\forall\, \mu \in [l]$

$$A_{i,j}^{(\mu)} : (\Omega, \mathcal{A}, \mathbb{P}) \to \mathbb{R},$$
$$b_i^{(\mu)} : (\Omega, \mathcal{A}, \mathbb{P}) \to \mathbb{R},$$

si può considerare $\mathbb{R}^Q$ munito della metrica introdotta nell'Osservazione 3.2.1.
Di conseguenza anche $\forall\, \mu = 2, \dots, l$ e $\forall n \in \mathbb{N}$, $U_n^{(\mu)}$ è una v.a.

$$U_n^{(\mu)} : (\Omega, \mathcal{A}, \mathbb{P}) \to (\mathbb{R}^Q, \rho).$$

A questo punto l'obiettivo che si cerca di raggiungere è il seguente:

$$\forall\, \mu = 2, \dots, l, \qquad U_n^{(\mu)} \xrightarrow{n\to\infty} \mathcal{N}^{\infty}\left(0,\, (K')^{(\mu)}\right), \tag{3.7}$$

per una certa $(K')^{(\mu)}$.
Si noti che, $\forall\, \mu = 2, \dots, l, \forall\, n \in \mathbb{N}$, la v.a. $b^{(\mu)} = (b_i^{(\mu)})_{(x,i)\in Q}$ è un processo gaussiano indipendente da

$$U_n^{(\mu)} = \left(\sum_{j=1}^{h_{\mu-1}(n)} A_{i,j}^{(\mu)} \bar{\Phi}_j^{(\mu-1)}(x)[n] \right)_{(x,i)\in Q}.$$

Per quel che riguarda l'indipendenza, infatti, le v.a. $A_{i,j}^{(\mu)}$ sono indipendenti dalle $b_i^{(\mu)}$ per ipotesi, $\forall\, i \in \mathbb{N}$, $j \in [h_{\mu-1}(n)]$ ed inoltre lo è anche ogni componente della $(\mu-1)$-attivazione, $\bar{\Phi}_j^{(\mu-1)}(x)[n]$, poiché sono tutte costruite a partire da v.a. a loro volta indipendenti da ciascuna $b_i^{(\mu)}$ [5].
Questa seconda affermazione può essere dedotta dalle definizioni delle v.a. in questione:

$$\forall\, j \in [h_{\mu-1}(n)], \quad \bar{\Phi}_j^{(\mu-1)}(x)[n] = \psi(\Phi_j^{(\mu-1)}(x)[n]) \ \text{ e}$$
$$\Phi_j^{(\mu-1)}(x)[n] = \sum_{k=1}^{h_{\mu-2}(n)} A_{j,k}^{(\mu-1)} \bar{\Phi}_k^{(\mu-2)}(x)[n] + b_j^{(\mu-1)}.$$

Per induzione rispetto a μ: se $\forall\, i \in \mathbb{N}$, $\forall\, k \in [h_{\mu-2}(n)]$, $\bar{\Phi}_k^{(\mu-2)}(x)[n]$ è indipendente da $b_i^{(\mu)}$, è noto che ogni v.a. con la quale si costruisce $\Phi_j^{(\mu-1)}(x)[n]$ è indipendente da $b_i^{(\mu)}$ (per (3.1)), dunque per misurabilità di ψ l'indipendenza si mantiene anche per $\bar{\Phi}_j^{(\mu-1)}(x)[n]$. Inoltre, la base del ragionamento induttivo è vera poiché $\Phi_j^{(1)}(x)[n]$ è combinazione lineare di v.a. indipendenti da $b_i^{(\mu)}$,

$$\Phi_j^{(1)}(x)[n] = \sum_{k=1}^{n_0} A_{j,k}^{(1)} x_k + b_j^{(1)},$$

come sopra l'indipendenza è trasportata, tramite ψ, anche a $\bar{\Phi}_j^{(1)}(x)[n]$.
Per quel che concerne la gaussianità di $b^{(\mu)}$, invece, risulta

$$\forall\, \mu \in [l], \quad b^{(\mu)} \sim \mathcal{N}^{\infty}\left(0,\, (K'')^{\mu}\right),$$

poiché successione di variabili aleatorie gaussiane indipendenti.
Dunque, applicando la Proposizione A.2.1 con $(X_n)_{n\in\mathbb{N}} := (U_n^{(\mu)})_{n\in\mathbb{N}}$ e $Y := b^{(\mu)}$ il limite (3.7) implica anche

$$\forall\, \mu = 2, \dots, l, \qquad U_n^{(\mu)} + b^{(\mu)} \xrightarrow{n\to\infty} \mathcal{N}^{\infty}\left(0,\, K^{(\mu)}\right), \tag{3.8}$$

con $K^{(\mu)} = (K')^{(\mu)} + (K'')^{(\mu)}$.
In particolare l'equazione (3.8) per $\mu = l$ costituisce la tesi del teorema.
Si definisce da qui in avanti la v.a. limite $U_*^{(\mu)}$, avente distribuzione $\mathcal{N}^{\infty}\left(0,\, (K')^{(\mu)}\right)$.

[5] Ciò basta per dichiarare l'indipendenza poiché se $A_{i,j}^{(\mu)}$ e $\bar{\Phi}_j^{(\mu-1)}(x)[n]$ sono indipendenti da $b_i^{(\mu)}$, allora lo è anche il loro prodotto.

Per semplificare la dimostrazione si fa uso del Lemma 3.2.1 che permette di provare l'equazione (3.7) attraverso la seguente condizione equivalente:

$$\begin{gathered}\forall\, \mu = 2, \dots, l,\ \forall\, L \subset Q,\ |L| < \infty \text{ e } \forall\, \alpha \in \mathbb{R}^{|L|}, \\ \sum_{u \in L} \alpha_u U_n^{(\mu)}(u) \xrightarrow{\mathcal{L}} \sum_{u \in L} \alpha_u U_*^{(\mu)}(u) \ \text{ con } \ U_*^{(\mu)} \sim \mathcal{N}^{\infty}\left(0,\, (K')^{(\mu)}\right).\end{gathered} \tag{3.9}$$

Si definisce $\forall\, L \subset Q,\ |L| < \infty$ e $\forall\, \alpha \in \mathbb{R}^{|L|}$,

$$\forall\, \mu = 2, \dots, l, \qquad T^{(\mu)}(L, \alpha)[n] = \sum_{(x,i) \in L} \alpha_{(x,i)} U_n^{(\mu)}(x, i).$$

Riportando la definizione alle sole variabili della *NN* si ha

$$T^{(\mu)}(L, \alpha)[n] = \sum_{(x,i) \in L} \alpha_{(x,i)} (\Phi_i^{(\mu)}(x)[n] - b_i^{(\mu)}). \tag{3.10}$$

Soffermandosi, in particolare, sulla legge del limite nell'equazione (3.9) si può notare che per le proprietà dei processi gaussiani

$$\sum_{u \in L} \alpha_u \mathcal{N}^{\infty}\left(0,\, (K')^{(\mu)}\right)(u) \sim \mathcal{N}^1\left(0,\, \alpha^T (K')^{(\mu)} \alpha\right)^{6}.$$

A questo punto si può affermare che (3.9) è equivalente a

$$\begin{gathered}\forall\, \mu = 2, \dots, l,\ \forall\, L \subset Q,\ |L| < \infty \text{ e } \forall\, \alpha \in \mathbb{R}^{|L|}, \\ T^{(\mu)}(L, \alpha)[n] \xrightarrow{\mathcal{L}} \mathcal{N}^1\left(0,\, \alpha^T (K')^{(\mu)} \alpha\right).\end{gathered} \tag{3.11}$$

Per dare una prova della validità di (3.11) si fa uso della seguente Proposizione 3.2.1.

Proposizione 3.2.1. *$\forall\, \mu = 2, \dots, l$, supposto che valga*

$$(\Phi_i^{(\mu-1)}(x)[n])_{(x,i) \in Q} \xrightarrow{\mathcal{L}} \mathcal{N}^{\infty}\left(0,\, K^{(\mu-1)}\right), \tag{3.12}$$

allora $\exists\, (K')^{(\mu)}$ tale che $\forall\, L \subset Q,\ |L| < \infty$ e $\forall\, \alpha \in \mathbb{R}^{|L|}$,

$$T^{(\mu)}(L, \alpha)[n] \xrightarrow{\mathcal{L}} \mathcal{N}^1\left(0,\, \alpha^T (K')^{(\mu)} \alpha\right).$$

Questa proposizione è espressa sotto forma di induzione.

Base

Per il passo base[7] necessita che

$$(\Phi_i^{(1)}(x))_{(x,i) \in Q} \xrightarrow{\mathcal{L}} \mathcal{N}^{\infty}\left(0,\, K^{(1)}\right).$$

[6] Si deve pensare α immerso in $\mathbb{R}^Q$ con 0 fuori da L: $\alpha[0] = 0$ se $k \in Q \setminus L$.
[7] La prima pre-attivazione non dipende da n, parametro di crescita dei layers.

Questo si può verificare dando una scrittura esplicita di $\Phi_i^{(1)}(x)$ al variare di $(x,i) \in Q$:

$$\Phi_i^{(1)}(x) = \sum_{j=1}^{n_0} A_{i,j}^{(1)} x_j + b_i^{(1)} = \frac{1}{\sqrt{n_0}} \sum_{j=1}^{n_0} \sqrt{\hat{C}_A^{(1)}} \varepsilon_{i,j}^{(1)} x_j + b_i^{(1)},$$

dove $\forall\, i \in \mathbb{N}$, $j \in [n_0]$, si ha $\varepsilon_{i,j}^{(1)} \sim \mathcal{N}^1(0,\,1)$ e $\varepsilon_{i,j}^{(1)}$ indipendente da $b_i^{(1)}$.
Infatti $\forall\,(x,i) \in Q$ si ha

$$\Phi_i^{(1)}(x) \sim \mathcal{N}^1\left(0,\, \sum_{j=1}^{n_0} C_A^{(1)} x_j^2 + C_b^{(1)}\right).$$

Inoltre le famiglie $(\Phi_i^{(1)}(x))_{x\in X}$, $\forall\, i \in \mathbb{N}$, sono processi gaussiani, in quanto ogni combinazione lineare finita di loro elementi è gaussiana: $\forall\, Y \subset X$, $|Y| < \infty$,

$$\sum_{x\in Y} \Phi_i^{(1)}(x) = \sum_{x\in Y} \left(\sum_{j=1}^{n_0} A_{i,j}^{(1)} x_j\right) + |Y| b_i^{(1)} = \sum_{j=1}^{n_0} A_{i,j}^{(1)} \left(\sum_{x\in Y} x_j\right) + |Y| b_i^{(1)}, \quad (3.13)$$

con $A_{i,j}^{(1)}$, $b_i^{(1)}$ gaussiane indipendenti $\forall\, i \in \mathbb{N}$, $j \in [n_0]$ e $\left(\sum_{x\in Y} x_j\right)$, $|Y|$ costanti finite.
Infine, se $i \neq k$, $\Phi_i^{(1)}(x)$ è indipendente da $\Phi_k^{(1)}(y)$, $\forall\, x, y \in X$, dunque, presa una ulteriore combinazione lineare finita di v.a. nel processo $(\Phi_i^{(1)}(x))_{(x,i)\in Q}$, si ha nuovamente la gaussianità: $\forall\, R \subset Q$, $|R| < \infty$,

$$\sum_{(x,i)\in R} \Phi_i^{(1)}(x) = \sum_{i\in\mathbb{N}} \sum_{x\,|\,(x,i)\in R} \Phi_i^{(1)}(x),$$

dove la somma interna costituisce una v.a. gaussiana per (3.13), mentre la somma esterna, che è in realtà finita, è fatta su v.a. indipendenti poiché indicizzata su $i \in \mathbb{N}$.
Dunque $(\Phi_i^{(1)}(x))_{(x,i)\in Q}$ è un processo gaussiano con legge $\mathcal{N}^\infty\left(0,\, K^{(1)}\right)$.
Volendo scrivere esplicitamente $K^{(1)}$ si ha che $\forall\,(x,i),(y,k) \in Q$, usando la bilinearità della covarianza,

$$\begin{aligned}
\left(K^{(1)}\right)_{(x,i),(y,k)} &= \mathrm{Cov}\left(\Phi_i^{(1)}(x), \Phi_k^{(1)}(y)\right) = \\
&= \sum_{j=1}^{n_0} \sqrt{C_A^{(1)}} x_j \mathrm{Cov}\left(\varepsilon_{i,j}^{(1)}, \Phi_k^{(1)}(y)\right) + \mathrm{Cov}\left(b_i^{(1)}, \Phi_k^{(1)}(y)\right) = \\
&= \sum_{j=1}^{n_0}\sum_{l=1}^{n_0} C_A^{(1)} x_j y_l \mathrm{Cov}\left(\varepsilon_{i,j}^{(1)}, \varepsilon_{k,l}^{(1)}\right) + \\
&\quad + \sum_{j=1}^{n_0} \sqrt{C_A^{(1)}} x_j \mathrm{Cov}\left(\varepsilon_{i,j}^{(1)}, b_k^{(1)}\right) + \\
&\quad + \sum_{l=1}^{n_0} \sqrt{C_A^{(1)}} y_l \mathrm{Cov}\left(b_i^{(1)}, \varepsilon_{k,l}^{(1)}\right) + \mathrm{Cov}\left(b_i^{(1)}, b_k^{(1)}\right) =
\end{aligned}$$

$$= \sum_{j=1}^{n_0}\sum_{l=1}^{n_0} C_A^{(1)} x_j y_l \mathrm{Cov}\left(\varepsilon_{i,j}^{(1)}, \varepsilon_{k,l}^{(1)}\right) + \mathrm{Cov}\left(b_i^{(1)}, b_k^{(1)}\right),$$

dove nel penultimo passaggio si è usato che $\forall\, i,k \in \mathbb{N},\, j \in [n_0]$ vale $\varepsilon_{i,j}^{(1)}$ indipendente da $b_k^{(1)}$, dunque $\mathrm{Cov}\left(\varepsilon_{i,j}^{(1)}, b_k^{(1)}\right) = 0$.

In particolare, essendo $\forall\, i \in \mathbb{N},\, j \in [n_0]$, $\varepsilon_{i,j}^{(1)} \sim \mathcal{N}^1(0,\,1)$ e $b_i^{(1)} \sim \mathcal{N}^1\left(0,\, C_b^{(1)}\right)$, valgono le seguenti identità:

$$\mathrm{Cov}\left(\varepsilon_{i,j}^{(1)}, \varepsilon_{k,l}^{(1)}\right) = \begin{cases} 0 & \text{se } (i,j) \neq (k,l) \\ 1 & \text{altrimenti} \end{cases}, \quad \mathrm{Cov}\left(b_i^{(1)}, b_k^{(1)}\right) = \begin{cases} 0 & \text{se } i \neq k \\ C_b^{(1)} & \text{altrimenti} \end{cases}.$$

Dunque si può concludere che

$$\forall\,(x,i),(y,k) \in Q, \quad \left(K^{(1)}\right)_{(x,i),(y,k)} = \delta_{i,k}\, C_A^{(1)} \sum_{j=1}^{n_0} x_j y_j + C_b^{(1)}.$$

Passo induttivo

Al fine di rendere conclusiva la Proposizione 3.2.1 va, inoltre, completato il passo induttivo $\mu - 1 \implies \mu$ con la dimostrazione che

$$T^{(\mu)}(L,\alpha)[n] \xrightarrow{\mathcal{L}} \mathcal{N}^1\left(0,\, \alpha^T (K')^{(\mu)} \alpha\right) \implies (\Phi_i^{(\mu)}(x)[n])_{(x,i)\in Q} \xrightarrow{\mathcal{L}} \mathcal{N}^\infty\left(0,\, K^{(\mu)}\right),$$

risultato ovviamente vero poiché (3.11) $\implies$ (3.8).

□

Dimostrazione della Proposizione 3.2.1. Si parte dalla definizione (3.10) dei $T^{(\mu)}(L,\alpha)[n]$, $\forall\, \mu = 2,\dots,l$ e $n \in \mathbb{N}$. Definendo, poi,

$$\gamma_j^{(\mu)}(L,\alpha)[n] := \sum_{(x,i)\in L} \alpha_{(x,i)} \sqrt{\hat{C}_A^{(\mu)}}\, \varepsilon_{i,j}^{(\mu)}\, \overline{\Phi}_j^{(\mu-1)}(x)[n] \tag{3.14}$$

e dando una scrittura esplicita di $\Phi_i^{(\mu)}(x)[n]$,

$$\Phi_i^{(\mu)}(x)[n] = \frac{1}{\sqrt{h_{\mu-1}(n)}} \sum_{j=1}^{h_{\mu-1}(n)} \sqrt{\hat{C}_A^{(\mu)}}\, \varepsilon_{i,j}^{(\mu)}\, \overline{\Phi}_j^{(\mu-1)}(x)[n] + b_i^{(\mu)},$$

si può facilmente riscrivere la (3.10) come segue:

$$T^{(\mu)}(L,\alpha)[n] = \frac{1}{\sqrt{h_{\mu-1}(n)}} \sum_{j=1}^{h_{\mu-1}(n)} \gamma_j^{(\mu)}(L,\alpha)[n]. \tag{3.15}$$

A questo punto viene applicata una versione adattata del Teorema del limite centrale (CLT) per successioni di v.a. scambiabili, inserito in appendice, per giungere direttamente alla tesi. Infatti, applicando il Teorema A.2.1, usando $h(n) := h_{\mu-1}(n)$

e $X_{n,j} := \gamma_j^{(\mu)}(L,\alpha)[n]$, $\forall\, n, j \in \mathbb{N}_0$, si avrebbe, supponendo di riuscire a soddisfare tutte le ipotesi, la seguente tesi:

$$\begin{aligned} T^{(\mu)}(L,\alpha)[n] &\xrightarrow{\mathcal{L}} \mathcal{N}^1\left(0,\, \sigma^2(\mu,L,\alpha)[*]\right) \\ \text{con } \sigma^2(\mu,L,\alpha)[*] &= \lim_{n\to\infty} \operatorname{Var}\left(\gamma_1^{(\mu)}(L,\alpha)[n]\right). \end{aligned} \tag{3.16}$$

Nel seguito viene data dimostrazione di una parte delle ipotesi, la trattazione comprensiva di tutti i dettagli è stata svolta da G. Matthews et al. [8, pp. 30–36]. Si noti che fino a questo punto non è stato fatto uso dell'ipotesi di sub-linearità di ψ del Teorema 3.2.1, tale assunto viene sfruttato per le dimostrazioni delle Ipotesi 3 e d del Teorema A.2.1.
Si considerino i seguenti enunciati $\forall\, \mu = 2, \ldots, l,\ \forall\, L \subset Q,\ |L| < \infty$ e $\forall\, \alpha \in \mathbb{R}^{|L|}$.

Ipotesi a

$\forall\, n \in \mathbb{N}_0$, $(\gamma_j^{(\mu)}(L,\alpha)[n])_{j\in\mathbb{N}_0}$ è un processo numerabile scambiabile rispetto all'indice j.

Dimostrazione. Per questa dimostrazione si fa uso del Teorema A.2.2 con $(X_j)_{j\in\mathbb{N}} := (\gamma_j^{(\mu)}(L,\alpha)[n])_{j\in\mathbb{N}_0}$. Basta perciò esibire la σ-algebra rispetto alla quale le v.a. in questione sono condizionatamente indipendenti ed identicamente distribuite.
Partendo dall'equazione (3.14) si può effettuare la seguente riscrittura

$$\gamma_j^{(\mu)}(L,\alpha)[n] = \sum_{(x,i)\in L} \alpha_{(x,i)} \sqrt{\hat{C}_A^{(\mu)}}\, \varepsilon_{i,j}^{(\mu)} \overline{\Phi}_j^{(\mu-1)}(x)[n] =$$

$$= \sum_{(x,i)\in L} \alpha_{(x,i)} \sqrt{\hat{C}_A^{(\mu)}}\, \varepsilon_{i,j}^{(\mu)} \psi\left(\frac{1}{\sqrt{h_{\mu-2}(n)}} \sum_{k=1}^{h_{\mu-2}(n)} \sqrt{\hat{C}_A^{(\mu-1)}}\, \varepsilon_{j,k}^{(\mu-1)} \overline{\Phi}_k^{(\mu-2)}(x)[n] + b_j^{(\mu-1)} \right),$$

con la convenzione $h_0(n) = n_0$ e $\forall\, k \in [n_0]$, $\overline{\Phi}_k^{(0)}(x)[n] = x_k$.
Condizionatamente alla σ-algebra $\mathcal{F}$ generata dalle v.a. $\overline{\Phi}_k^{(\mu-2)}(x)[n]$ con $k \in [h_{\mu-2}(n)]$ e x input tale che $(x,i) \in L$ per qualche $i \in \mathbb{N}$, si ha che le v.a. $\gamma_j^{(\mu)}(L,\alpha)[n]$ sono indipendenti ed identicamente distribuite $\forall\, j \in \mathbb{N}_0$ (questo può essere dedotto per ricorsione). Per maggiore precisione si definisce $\mathcal{F}$ come segue

$$\mathcal{F} = \sigma\left(\left(\overline{\Phi}_k^{(\mu-2)}(x)[n] \right)_{(x,k)\in[h_{\mu-2}(n)]\times\pi_1(L)} \right)\ ^{8}.$$

Ipotesi b

$\mathbb{E}\left[\gamma_1^{(\mu)}(L,\alpha)[n]\right] = 0.$

Dimostrazione. La dimostrazione consiste di una verifica diretta:

$$\mathbb{E}\left[\gamma_1^{(\mu)}(L,\alpha)[n]\right] = \mathbb{E}\left[\sum_{(x,i)\in L} \alpha_{(x,i)} \sqrt{\hat{C}_A^{(\mu)}}\, \varepsilon_{i,1}^{(\mu)} \overline{\Phi}_1^{(\mu-1)}(x)[n] \right] =$$

[8] $L \subset Q = X \times \mathbb{N}$, dunque con $\pi_1(L)$ si definisce l'insieme delle $x \in X$ tali che $\exists\, i \in \mathbb{N}$ e $(x,i) \in L$.

$$= \sum_{(x,i)\in L} \text{cost}\ \mathbb{E}\left[\varepsilon_{i,1}^{(\mu)}\ \overline{\Phi}_1^{(\mu-1)}(x)[n]\right] =$$
$$= \sum_{(x,i)\in L} \text{cost}\, 0 = 0.$$

La terza uguaglianza segue dal fatto che, $\forall\,(x,i) \in L$, $\varepsilon_{i,1}^{(\mu)}$ e $\overline{\Phi}_1^{(\mu-1)}(x)[n]$ sono indipendenti. Infatti $\overline{\Phi}_1^{(\mu-1)}(x)[n] = \psi(\Phi_1^{(\mu-1)}(x)[n])$ e $\Phi_1^{(\mu-1)}(x)[n]$ dipende solo dalle v.a. $A_1^{(\nu)}$, $b_1^{(\nu)}$, $\forall\,\nu \in [\mu-1]$ (si può verificare per induzione); a questo punto l'indipendenza segue dalla definizione (3.1).

Ipotesi c

Definendo $\sigma^2(\mu, L, \alpha)[n] := \text{Var}\left(\gamma_1^{(\mu)}(L,\alpha)[n]\right) < \infty$ ed assumendo l'ipotesi (3.12) della Proposizione 3.2.1, vale il seguente limite:

$$\sigma^2(\mu, L, \alpha)[n] \xrightarrow{n\to\infty} \sigma^2(\mu, L, \alpha)[*],$$

con $\sigma^2(\mu, L, \alpha)[*] = \alpha^T (K')^{(\mu)}(L)\alpha$ per una qualche matrice $(K')^{(\mu)}$ [9].

Dimostrazione. L'idea è cercare di riscrivere $\sigma^2(\mu, L, \alpha)[n]$ al fine di semplificare il passaggio al limite:

$$\begin{aligned}
\sigma^2(\mu, L, \alpha)[n] &= \mathbb{E}\left[\left(\gamma_1^{(\mu)}(L,\alpha)[n]\right)^2\right] = \\
&= \mathbb{E}\left[\left(\sum_{(x,i)\in L} \alpha_{(x,i)} \sqrt{\hat{C}_A^{(\mu)}}\, \varepsilon_{i,1}^{(\mu)} \overline{\Phi}_1^{(\mu-1)}(x)[n]\right)^2\right] = \\
&= \hat{C}_A^{(\mu)} \mathbb{E}\left[\left(\alpha \cdot \begin{pmatrix} \varepsilon_{i_1,1}^{(\mu)} \overline{\Phi}_1^{(\mu-1)}(x_1)[n] \\ \vdots \\ \varepsilon_{i_{|L|},1}^{(\mu)} \overline{\Phi}_1^{(\mu-1)}(x_{|L|})[n] \end{pmatrix}\right)^2\right]
\end{aligned}$$

dove $L = (x_j, i_j)_{j=1}^{|L|}$.
Si definisce, per semplificare la scrittura,

$$\tilde{\Phi}_1^{(\mu-1)}(L,x)[n] := \begin{pmatrix} \varepsilon_{i_1,1}^{(\mu)} \overline{\Phi}_1^{(\mu-1)}(x_1)[n] \\ \vdots \\ \varepsilon_{i_{|L|},1}^{(\mu)} \overline{\Phi}_1^{(\mu-1)}(x_{|L|})[n] \end{pmatrix}.$$

Dunque, sostituendo sopra, si ottiene

$$\begin{aligned}
\sigma^2(\mu, L, \alpha)[n] &= \hat{C}_A^{(\mu)} \mathbb{E}\left[\left(\alpha^T \tilde{\Phi}_1^{(\mu-1)}(L,x)[n]\right)^2\right] = \\
&= \hat{C}_A^{(\mu)} \alpha^T \mathbb{E}\left[\tilde{\Phi}_1^{(\mu-1)}(L,x)[n] \left(\tilde{\Phi}_1^{(\mu-1)}(L,x)[n]\right)^T\right] \alpha;
\end{aligned} \tag{3.17}$$

[9] Con $(K')^{(\mu)}(L)$ si indica il minore di $(K')^{(\mu)}$ associato agli indici $L \times L \subset Q \times Q$.

perciò, per il passaggio al limite, è sufficiente analizzare il comportamento di

$$\mathbb{E}\left[\tilde{\Phi}_1^{(\mu-1)}(L,x)[n])\left(\tilde{\Phi}_1^{(\mu-1)}(L,x)[n])\right)^T\right] \text{ per } n\to\infty.$$

Analizzando ciascuna entrata della matrice si ottiene

$$\begin{aligned}
&\mathbb{E}\left[\left(\tilde{\Phi}_1^{(\mu-1)}(L,x)[n]\left(\tilde{\Phi}_1^{(\mu-1)}(L,x)[n])\right)^T\right)_{j,k}\right] = \\
&= \mathbb{E}\left[\left(\tilde{\Phi}_1^{(\mu-1)}(L,x)[n]\right)_j\left(\tilde{\Phi}_1^{(\mu-1)}(L,x)[n]\right)_k\right] = \\
&= \mathbb{E}\left[\left(\varepsilon_{i_j,1}^{(\mu)}\overline{\Phi}_1^{(\mu-1)}(x_j)[n]\right)\left(\varepsilon_{i_k,1}^{(\mu)}\overline{\Phi}_1^{(\mu-1)}(x_k)[n]\right)\right] = \\
&= \delta_{j,k}\,\mathbb{E}\left[\overline{\Phi}_1^{(\mu-1)}(x_j)[n]\overline{\Phi}_1^{(\mu-1)}(x_k)[n]\right] = \\
&= \delta_{j,k}\,\mathbb{E}\left[\psi\left(\Phi_1^{(\mu-1)}(x_j)[n]\right)\psi\left(\Phi_1^{(\mu-1)}(x_k)[n]\right)\right]
\end{aligned}$$

dove per la terza identità si è usato che $\varepsilon_{i_j,1}^{(\mu)}$ ed $\varepsilon_{i_k,1}^{(\mu)}$ sono indipendenti da $\overline{\Phi}_1^{(\mu-1)}(x_j)[n]$ e $\overline{\Phi}_1^{(\mu-1)}(x_k)[n]$ (per (3.1)).
A questo punto, usando il Teorema A.2.3 con

$$X_n := (\Phi_1^{(\mu-1)}(x)[n])_{(x,i)\in Q} \quad \text{e} \quad \phi(\cdot) := \psi\big(\pi_{(x_j,1)}(\cdot)\big)\psi\big(\pi_{(x_k,1)}(\cdot)\big)$$

(con $\pi_{(x_j,1)}(\cdot)$, proiezione sulla componente $(x_j,1)$ di L), si ha

$$\begin{gathered}
\delta_{j,k}\,\mathbb{E}\left[\psi\left(\Phi_1^{(\mu-1)}(x_j)[n]\right)\psi\left(\Phi_1^{(\mu-1)}(x_k)[n]\right)\right] \\
\Big\downarrow\, n\to\infty \\
\delta_{j,k}\,\mathbb{E}\left[\psi\big(\pi_{(x_j,1)}(\cdot)\big)\psi\big(\pi_{(x_k,1)}(\cdot)\big)\,(X)\right] = \\
= \delta_{j,k}\,\mathbb{E}\left[\psi\Big(X_{(x_j,1)}\Big)\psi\Big(X_{(x_k,1)}\Big)\right],
\end{gathered}$$

con $X\sim\mathcal{N}^\infty\left(0,\,K^{(\mu-1)}\right)$.
Dunque, risostituendo nell'equazione (3.17) si conclude

$$\sigma^2(\mu,L,\alpha)[n] \xrightarrow{n\to\infty} \alpha^T(K')^{(\mu)}(L)\alpha,$$

con

$$\begin{gathered}
\Big((K')^{(\mu)}\Big)_{j,k} := \hat{C}_A^{(\mu)}\delta_{j,k}\,\mathbb{E}\left[\psi\left(X_{(x_j,1)}\right)\psi\Big(X_{(x_k,1)}\Big)\right] \\
\text{e } X\sim\mathcal{N}^\infty\left(0,\,K^{(\mu-1)}\right).
\end{gathered} \tag{3.18}$$

Ipotesi 1

$\mathbb{E}\left[\gamma_1^{(\mu)}(L,\alpha)[n]\;\gamma_2^{(\mu)}(L,\alpha)[n]\right] = 0.$

Ipotesi 2

Supposto che valga

$$(\Phi_i^{(\mu-1)}(x)[n])_{(x,i)\in Q} \xrightarrow{\mathcal{L}} \mathcal{N}^\infty \left(0,\, K^{(\mu-1)}\right),$$

allora

$$\mathbb{E}\left[\left|\gamma_1^{(\mu)}(L,\alpha)[n]\ \gamma_2^{(\mu)}(L,\alpha)[n]\right|^2\right] \xrightarrow{n\to\infty} \sigma^4(\mu,L,\alpha)[*].$$

Ipotesi 3 ed ipotesi d

Si ha

$$\mathbb{E}\left[\left|\gamma_1^{(\mu)}(L,\alpha)[n]\right|^3\right] < c < \infty,$$

con c indipendente da n, dunque

$$\mathbb{E}\left[\left|\gamma_1^{(\mu)}(L,\alpha)[n]\right|^3\right] = o\left(\sqrt{h_{\mu-1}(n)}\right).$$

Dunque si può applicare il Teorema A.2.1, usando $X_{n,j} = \gamma_j^{(\mu)}(L,\alpha)[n]$, $\forall\, n, j \in \mathbb{N}_0$; vale perciò l'equazione (3.16), in particolare per l'Ipotesi c si ha

$$\forall\, L \subset Q,\ |L| < \infty \text{ e } \forall\, \alpha \in \mathbb{R}^{|L|} \quad T^{(\mu)}(L,\alpha)[n] \xrightarrow{\mathcal{L}} \mathcal{N}^1\left(0,\, \alpha^T (K')^{(\mu)}(L)\alpha\right),$$

con $(K')^{(\mu)}$ come in (3.18), ovvero la tesi. □

Capitolo 4

Sperimentazioni

Nelle seguenti sezioni sono riportati gli esiti di sperimentazioni atte a riprodurre i due principali risultati dei Capitoli 2 e 3: il Teorema di approssimazione universale ed il Teorema di convergenza ad un processo gaussiano.
Tutte le simulazioni sono state condotte usando il linguaggio di programmazione `Python`; si è in particolare fatto uso del pacchetto `PyTorch`, framework di machine learning, per definire ed utilizzare le reti neurali.

4.1. Test preliminari con PyTorch

La prima fase della sperimentazione è finalizzata alla costruzione di un modello parametrico che ricalchi a pieno la Definizione 2.0.1. In particolare l'obiettivo è quello di creare una sottoclasse della classe `nn.Module` che, a partire dai dati di input

- `nh` := numero di hidden layers,
- `sizeinput` := numero di neuroni in input,
- `sizeoutput` := numero di neuroni in output,
- `act` := funzione di attivazione `act`: $\mathbb{R} \to \mathbb{R}$,
- `growth` := parametro di crescita da cui dipende l'ampiezza degli hidden layers (fa le veci di $n \in \mathbb{N}$ per la funzione $h_\mu(n)$ all'interno delle sperimentazioni),
- `ty` := parametro associato alla definizione di h_μ (ad esempio se `ty = "const"` si ha $h_\mu(n) = n$, o ancora se `ty = "exp"` si ha $h_\mu(n) = n^\mu$),

definisca una *NN* con architettura $\alpha = (n, \psi)$ strutturata come segue:

- $n = (n_\mu)_{\mu=0}^{\texttt{nh}+1}$ con
  1. $n_0 = \texttt{sizeinput}$,
  2. $n_\mu = h_\mu(n),\ \forall\, \mu = 1, \dots, \texttt{nh}$ con $h_\mu(n) :=$ `hmu(`μ`, n, ty)`, ed `n` che assume il valore di `growth` nel metodo `__init__` della sottoclasse,
  3. $n_{\texttt{nh}+1} = \texttt{sizeoutput}$;
- $\psi = \texttt{act}$.

Per completezza si riporta di seguito la struttura della funzione `hmu` usata nelle sperimentazioni:

$$\texttt{hmu(}\mu\texttt{, n, ty)} = \begin{cases} \texttt{n}^\mu & \text{se } \texttt{ty} = \texttt{"exp"} \\ \texttt{n} & \text{se } \texttt{ty} = \texttt{"const"} \end{cases}.$$

Questo risultato è raggiunto facendo uso della struttura dati `nn.ModuleList`, che costituisce una lista di moduli di `PyTorch`. I parametri della *NN* costruita vengono automaticamente inizializzati uniformemente nell'intervallo $[-k_\mu, k_\mu]$, con k_μ dipendente dall'hidden layer μ in cui ci si trova. Entrando più nello specifico: $A_{i,j}^{(\mu)}$ e $b_i^{(\mu)}$ per $\mu = 1, \ldots,$ `nh` $+ 1$ (e $\forall\, i, j$) sono generati con distribuzione $\mathcal{U}(-k_\mu, k_\mu)$ e $k_\mu = \sqrt{1/n_{\mu-1}}$. In fig. 4.1 si riporta il codice utilizzato per dare forma al modello.

```
# 0.1
class NeuralNetwork(nn.Module):
    def __init__(self, nh, sizeinput, sizeoutput, n, ty):
        super(NeuralNetwork, self).__init__()
        self.linears = nn.ModuleList([nn.Linear(sizeinput, hmu(1, n, ty))])
        for i in range(1, nh):
            self.linears.append(nn.Linear(hmu(i, n, ty), hmu(i + 1, n, ty)))
        self.linears.append(nn.Linear(hmu(nh, n, ty), sizeoutput, False))
    def forward(self, x):
        x = self.linears[0](x)
        for i in range(1, inp.nh + 1):
            x = activation(x, inp.act)
            x = self.linears[i](x)
        return x
model = NeuralNetwork(inp.nh, inp.sizeinput, inp.sizeoutput, inp.growth,
    inp.ty).to(device)
```

Figura 4.1: `main/torch_test`, costruzione della rete neurale.

4.1.1. Torch_test

Nel primo test si è cercato di capire quale fosse il comportamento di una rete neurale dopo aver inizializzato $A_{i,j}^{(\mu)}$ e $b_i^{(\mu)}$ con distribuzione gaussiana, $\forall\, \mu, i, j$. In particolare imponendo `sizeinput` $=$ `sizeoutput` $= 1$ ed assegnando ad `act` una specifica funzione di attivazione si è osservato che, completando i parametri restanti in maniera casuale ed eseguendo un'inizializzazione gaussiana dei coefficienti, si ottiene una funzione che assume un andamento assimilabile a quello di `act`.
I coefficienti della rete sono stati generati in accordo con la teoria sviluppata da G. Matthews et al. [8, p. 5], ripresa nel Capitolo 3. Più nel dettaglio si sono usate distribuzioni normali come in (3.1) con $\hat{C}_A^{(\mu)} = 0.8$ e $C_b^{(\mu)} = 0.2$, così come fatto da G. Matthews et al. [8, p. 10]. Tale inizializzazione è riportata in fig. 4.2; si sono sfruttati anche in questo caso i comandi built-in di `PyTorch`.

```
# 1

# 1.1
c = 0.8
ca = torch.ones(inp.nh + 1)
for j in range(inp.nh + 1):
    ca[j] = c/hmu(j - 1, inp.growth, inp.ty)
cb = 0.2
sqrtca = torch.sqrt(ca)
sqrtcb = math.sqrt(cb)
```

```
# 1.2
print("1. NN parameters")
i = 1
bias_flag = 0
for param in model.parameters():
    if (bias_flag == 0):

        # 1.2.1
        print("W_" + str(i) + ":\n- size: " + str(param.size(0)) + "x" +
    str(param.size(1)) + "\n- distribution: N(0, " + str(round(ca[i -
    1].item(), 2)) + ")")
        param.data = torch.normal(mean = torch.zeros(param.size(0),
    param.size(1)), std = sqrtca[i - 1]*torch.ones(param.size(0),
    param.size(1))).to(device)
        bias_flag = 1
    else:

        # 1.2.2
        print("b_" + str(i) + ":\n- size: " + str(param.size(0)) + "\n-
    distribution: N(0, " + str(round(cb, 2)) + ")")
        param.data = torch.normal(mean = torch.zeros(param.size(0)), std =
    sqrtcb*torch.ones(param.size(0))).to(device)
        bias_flag = 0
        i = i + 1
    print("- content: \n", print_tensor(param), "\n\t----")
```

Figura 4.2: `main/torch_test`, inizializzazione dei parametri.

L'output in fig. 4.3 è stato ottenuto con i parametri riportati nella seguente tabella:

act	nh	growth	ty
"ReLU"	4	5	"const"

.

Osservazione 4.1.1. In questo caso la funzione in output è composta di spezzate che si susseguono, esattamente come avviene per l'attivazione ReLU in fig. 4.4. Il risultato rimane analogo per ogni altra scelta degli ultimi tre parametri in tabella.

Utilizzando invece, ad esempio, i parametri

act	nh	growth	ty
"Sigmoid"	3	7	"const"

,

si ottengono le fig. 4.5 e 4.6. Anche in questo caso la funzione generata dalla *NN* risulta molto regolare, così come la funzione `act = "Sigmoid"` [1].

Tali risultati permettono subito di fare le prime fondamentali considerazioni sulla scelta dei parametri da utilizzare per costruire una rete neurale che approssimi una funzione continua da $\mathbb{R}$ in $\mathbb{R}$. Dovendo, ad esempio, approssimare una funzione continua ovunque ma non derivabile in molti punti, sarà preferibile usare un'attivazione che sia a sua volta non derivabile in qualche punto, come la ReLU. Questo ovviamente non preclude l'utilizzo di funzioni d'attivazione più regolari come la Sigmoid, tuttavia appare da subito chiaro che la funzione costruita dalla rete sarà a sua volta abbastanza regolare, quindi meno adatta allo scopo.

[1] Con Sigmoid si denota la seguente funzione: $\text{Sigmoid}(x) = \frac{1}{1+e^{-x}}$.

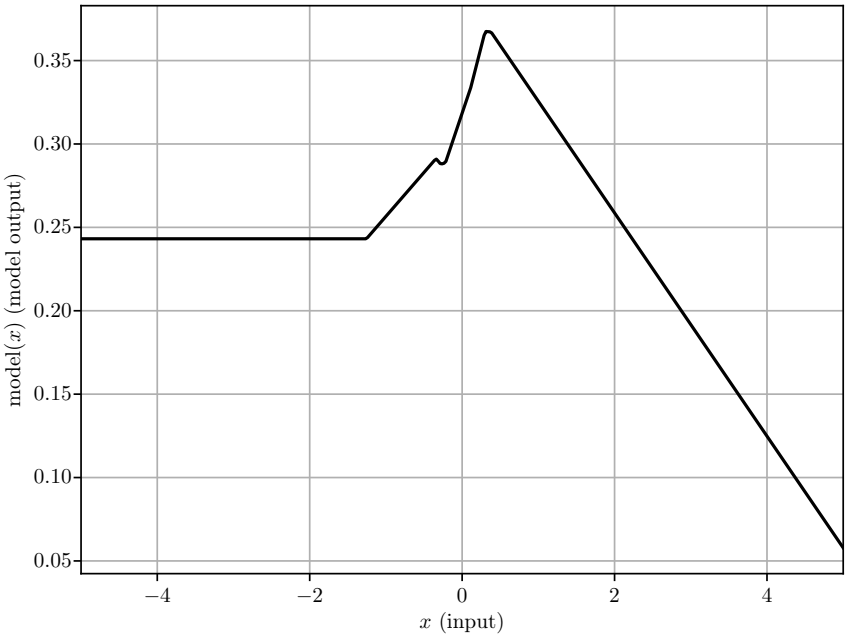

Figura 4.3: `torch_test`, grafico di una *NN* inizializzata con parametri generati con distribuzione normale ed attivazione ReLU.

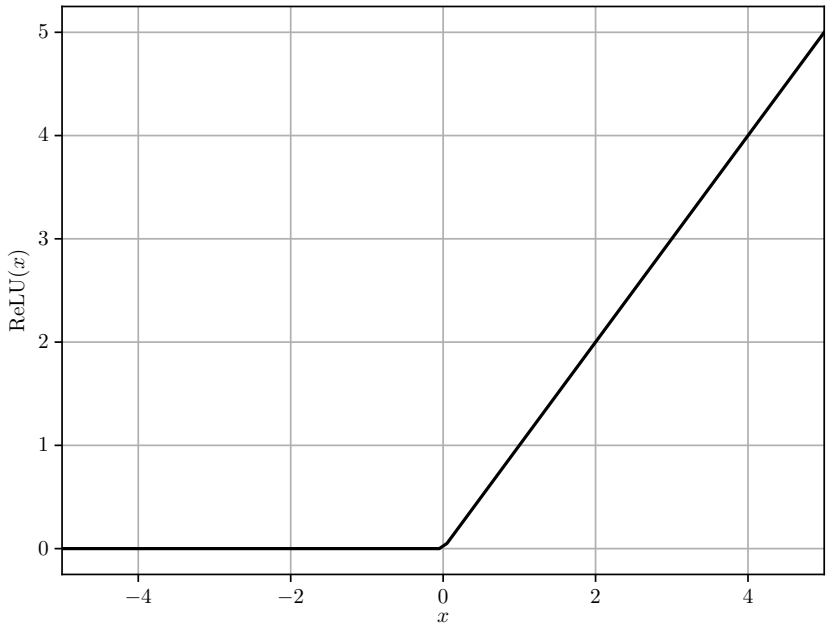

Figura 4.4: `torch_test`, grafico della funzione di attivazione ReLU.

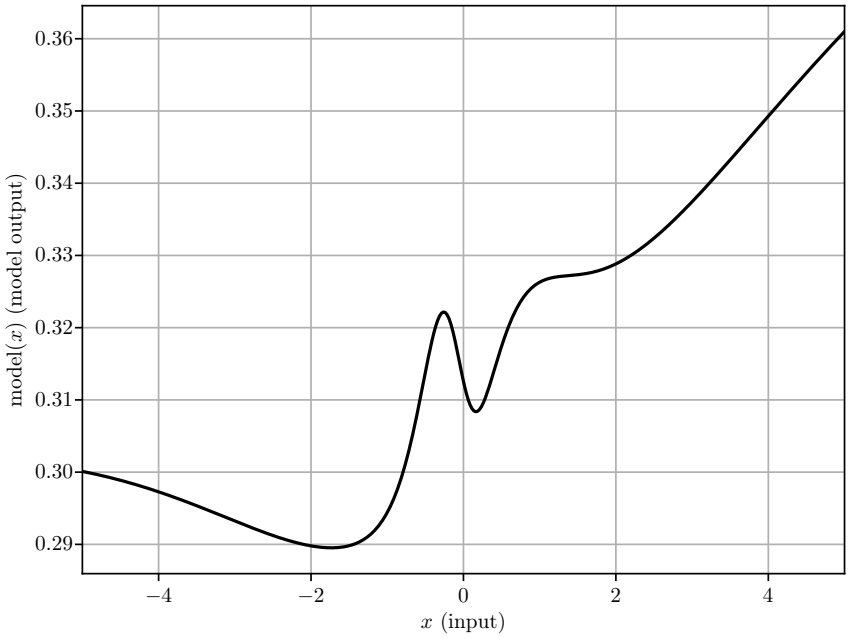

Figura 4.5: `torch_test`, grafico di una *NN* inizializzata con parametri generati con distribuzione normale ed attivazione Sigmoid.

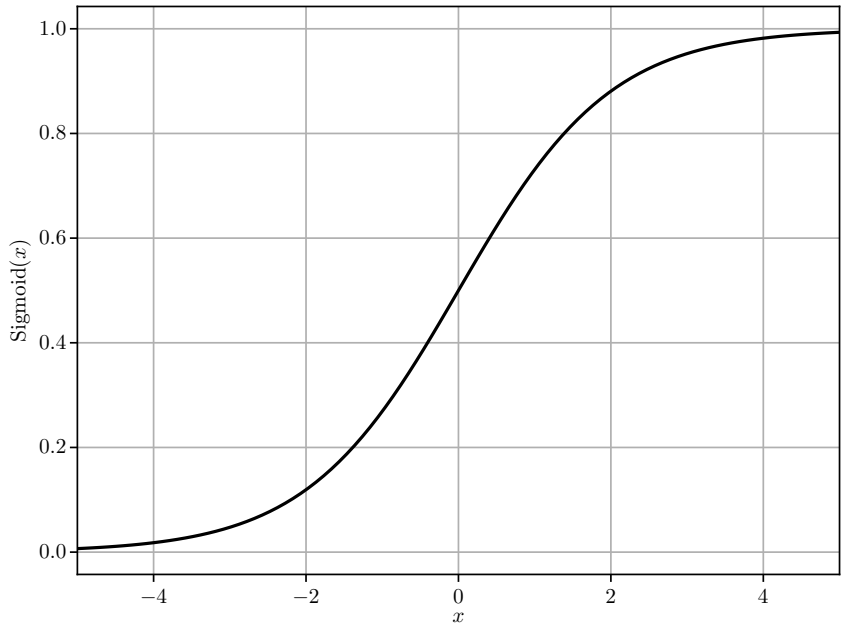

Figura 4.6: `torch_test`, grafico della funzione di attivazione Sigmoid.

4.2. `Exp_1`: convergenza ad un processo gaussiano

In questa sperimentazione si va ad analizzare l'andamento di una rete neurale con le stesse ipotesi presenti nel Teorema di convergenza ad un processo gaussiano. Essendo però quest'ultimo un risultato asintotico, la sperimentazione è volta ad indagare la gaussianità di *NN* con valori fissati di `nh`, `sizeinput`, `sizeoutput` [2] ed `act` e con parametri da cui dipende l'ampiezza degli hidden layers grandi: ad esempio `growth` ≥ 100.
Il comportamento di tali reti è studiato su un ridotto numero di campioni, `nsample`, che vengono dati in input numerose volte, `ntest`, alla medesima *NN*, con parametri estratti ad ogni test con distribuzione $\mathcal{N}^1(0, C_A^{(\mu)})$ se facenti parte della matrice associata al μ-esimo hidden layer e con distribuzione $\mathcal{N}^1(0, C_b^{(\mu)})$ se elementi del μ-esimo bias. Lo scopo è di ottenere un set di `ntest` output che permetta di individuare la *densità della rete neurale* intesa come variabile aleatoria a valori in $\mathbb{R}^{\texttt{ntest}}$. Successivamente i dati raccolti vengono analizzati per testare la correttezza del Teorema 3.2.1. Quello che ci si aspetta è che la v.a. in output si avvicini ad un vettore gaussiano. Per questo controllo si fa riferimento alla definizione di distribuzione normale multivariata: un vettore aleatorio $X : (\Omega, \mathcal{A}, \mathbb{P}) \to (\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$, con $(\Omega, \mathcal{A}, \mathbb{P})$ spazio di probabilità, ha distribuzione normale multivariata se $\forall\, a \in \mathbb{R}^k$, $Y = a^T X$ ha distribuzione gaussiana (con l'usuale convenzione $\mathcal{N}^1(m, 0) := m$).

4.2.1. Funzionamento di `Exp_1`

Nella pratica quel che viene fatto nell'`exp_1` è generare una rete secondo i parametri di input, come fatto in fig. 4.1, ed inizializzare i valori di $C_A^{(\mu)}$ e $C_b^{(\mu)}$ come nella sezione 1.1 della fig. 4.2. Vengono poi generati gli `nsample` array di input e, all'interno di un loop, si ricavano gli `ntest` output applicando la rete generata al passo corrente a tutti i sample, tramite il codice presente nella seguente fig. 4.7.

```
# 0.2.4
for test in range(inp.ntest):
    model = NeuralNetwork(inp.nh, inp.sizeinput, inp.sizeoutput, inp.growth,
    inp.ty).to(device)

    # 0.2.4.1
    i = 1
    bias_flag = 0
    for param in model.parameters():
        if (bias_flag == 0):
            param.data = torch.normal(mean = torch.zeros(param.size(0),
    param.size(1)), std = sqrtca[i - 1]*torch.ones(param.size(0),
    param.size(1))).to(device)
            bias_flag = 1
        else:
            param.data = torch.normal(mean = torch.zeros(param.size(0)), std
    = sqrtcb*torch.ones(param.size(0))).to(device)
            bias_flag = 0
            i = i + 1
```

[2] Il valore di `sizeoutput` è in realtà fissato ad 1 per poter testare la gaussianità della *NN* visivamente.

```
16      # 0.2.4.2
17      for i in range(inp.nsample):
18          outputarray[i, test] = model(sample_input[i,
        :]).cpu().detach().numpy()
19
20      # 0.2.4.3
21      if test%100 == 0:
22          print("\t\t" + str(test) + "...")
```

Figura 4.7: `main/exp_1`, generazione degli `ntest` output.

Il passo successivo consiste nell'eseguire il test di gaussianità sopra descritto sull'output di uno degli `nsample` precedentemente generati (`sample_0`) e su una combinazione lineare di tutti gli output. In entrambi i casi la procedura è la medesima: dapprima si esegue un fit gaussiano dell'array considerato; quest'ultimo è poi utilizzato per generare un grafico della densità di probabilità trovata a confronto con l'istogramma ottenuto a partire dall'array[3]. Si riporta in fig. 4.8, a scopo esemplificativo, il codice usato per eseguire il fit dell'output di `sample_0`.

```
1 # 1.1.1
2 mean = np.zeros(inp.nsample)
3 std = np.zeros(inp.nsample)
4 x_0 = 1.2*max(-min(outputarray[inp.sample_0,:]),
      max(outputarray[inp.sample_0,:]))
5 x = np.linspace(-x_0, x_0, 250)
6 fitted = np.zeros((inp.nsample, 250))
7 for i in range(inp.nsample):
8     mean[i], std[i] = norm.fit(outputarray[i,:])
9     fitted[i, :] = norm.pdf(x, mean[i], std[i])
```

Figura 4.8: `main/exp_1`, fit dell'output relativo all'input `sample_0`.

Per completezza viene infine generata la matrice delle covarianze numeriche calcolate a partire da tutti gli array di output, a tale scopo si ricorre alla funzione `np.cov()` presente all'interno della libreria `numpy`.
Le immagini ed i dati raccolti con un'esecuzione esemplificativa dell'esperimento sono riportati di seguito. I parametri utilizzati per la costruzione della *NN* sono

nh	sizeinput	sizeoutput	act	growth	ty
2	4	1	"ReLU"	100	"const"

;

mentre per il resto della simulazione si sono usati i seguenti valori

ntest	nsample	sample_0
10000	5	0

.

La media e la matrice delle covarianze dell'`outputarray` ottenuto sono trascritte in fig. 4.9. Si noti che sebbene le medie non siano nulle, sono comunque prossime

[3]L'istogramma è normalizzato in modo che l'area sottesa dal grafico sia pari ad 1.

allo 0, in accordo con quanto dichiarato nel Teorema di convergenza ad un processo gaussiano.

```
1.1. mean array and covariance matrix of the NNs randomly generated and
    evaluated on sample_input[i], for i = 0,...,4:
mean:
 [-0.00354243 -0.00665407  0.00248438  0.01066375  0.01299695]
covariance:
 [[17.25171395  9.05261251 15.5282864  21.20272439 13.2272047 ]
 [ 9.05261251  5.63059222  8.3796821  10.59587945  7.12119958]
 [15.5282864   8.3796821  14.70335737 19.13154959 12.5616007 ]
 [21.20272439 10.59587945 19.13154959 29.42393497 20.00177271]
 [13.2272047   7.12119958 12.5616007  20.00177271 17.66254693]]
```

Figura 4.9: `output/exp_1_output`, vettore delle medie e matrice delle covarianze delle *NNs* generate randomicamente e valutate in `sample_input[i]`, $\forall\, i = 0, \ldots,$ `nsample - 1`.

In fig. 4.10 è riportato il grafico della funzione `fitted[inp.sample_0]` (in ascissa si ha il vettore `x` calcolato in fig. 4.8) sovrapposto all'istogramma ottenuto da `outputarray[inp.sample_0, :]`.

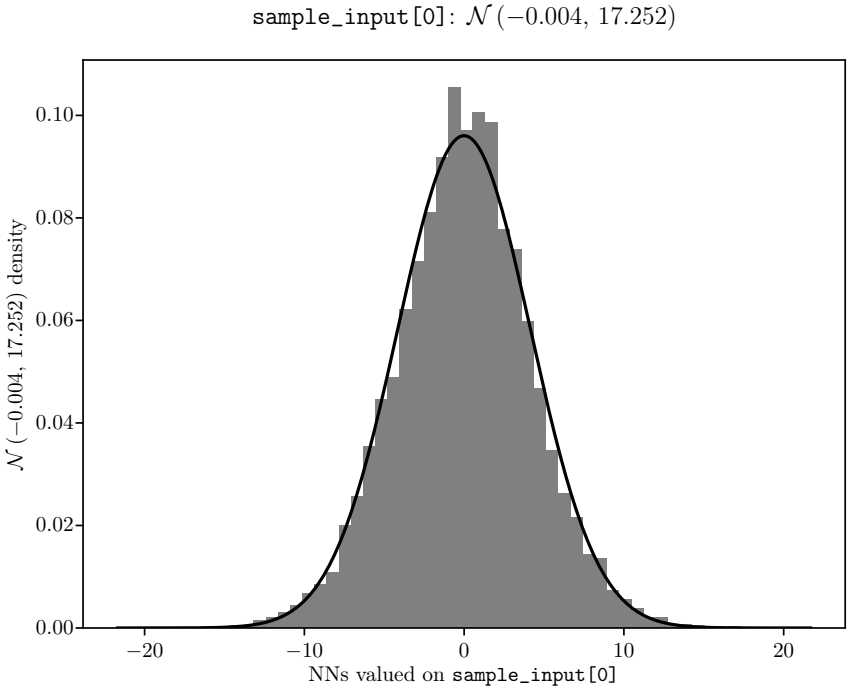

Figura 4.10: `exp_1`, rete neurale valutata in `sample_input[0]`.

L'analogo grafico ottenuto per il vettore `linear_combination` (combinazione lineare di `outputarray[i, :]`, $\forall\, i = 1, \ldots,$ `nsample` $- 1$) è invece riportato in fig. 4.11.

La combinazione lineare presente in figura è realizzata con $a^T = (7, 4, 3, 4, 5)$ e rappresenta a^T`outputarray`.

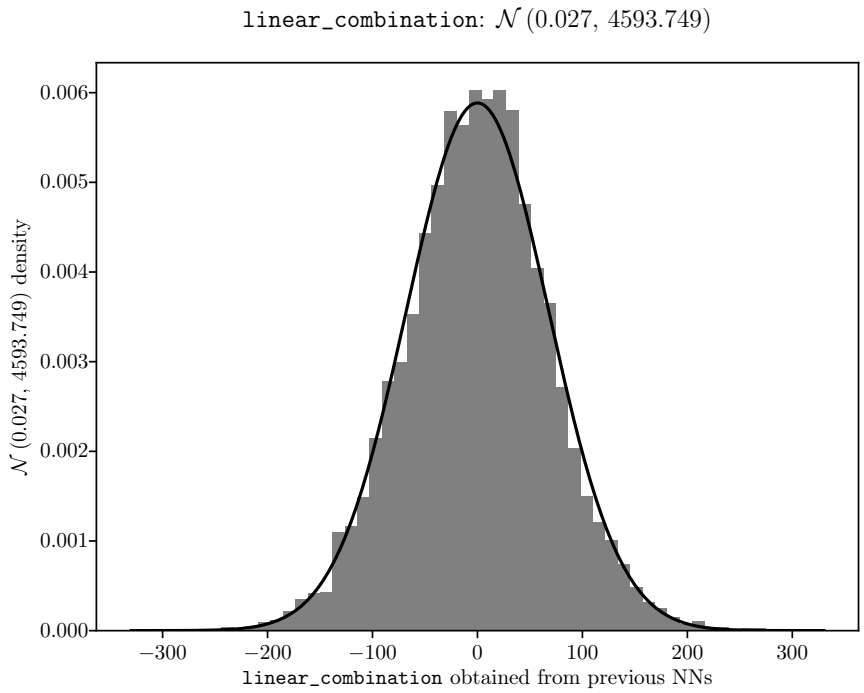

Figura 4.11: `exp_1`, combinazione lineare degli output relativi ai `sample_input[i]` per $i = 1, \ldots, n$.

Questa simulazione ha quindi confermato che una *NN* con parametri inizializzati come nell'equazione (3.1) e con funzione di attivazione sub-lineare ha un comportamento *asintotico* approssimativamente gaussiano e perciò rispetta quanto afferma il Teorema 3.2.1.

Osservazione 4.2.1. Un comportamento del tutto analogo si può ottenere con ogni altra funzione d'attivazione sub-lineare, in accordo con la teoria sviluppata nel Capitolo 3. La scelta di utilizzare la funzione di attivazione ReLU nel condurre il precedente esperimento è dettata unicamente dal fatto che è convenzionalmente la più utilizzata per modelli di reti neurali completamente connesse e feed-forward [5].

4.2.2. Controesempi

Quel che si intende mostrare nella presente sottosezione è che l'andamento normale delle reti neurali appena mostrato non è una proprietà scontata e che, quindi, le ipotesi del Teorema 3.2.1 rappresentano delle condizioni non banali.
La prima verifica che viene effettuata è relativa alla distribuzione con cui sono estratti i parametri della *NN*. In questo caso invece di generare $A_{i,j}^{(\mu)}$ e $b_i^{(\mu)}$ rispettivamente tramite distribuzioni $\mathcal{N}^1(0, C_A^{(\mu)})$ e $\mathcal{N}^1(0, C_b^{(\mu)})$ viene usata una distribuzione esponenziale di parametro $\lambda = 0.5$, $\mathcal{E}(0.5)$. Si conduce un esperimento del tutto analogo

al precedente ma invece di inizializzare i parametri come in fig. 4.2 si usa il codice in fig. 4.12.

```
# 0.2.3
for test in range(inp.ntest):
    model = NeuralNetwork(inp.nh, inp.sizeinput, inp.sizeoutput, inp.growth,
    inp.ty).to(device)

    # 0.2.3.1
    i = 1
    bias_flag = 0

    # 0.2.3.1.1
    if inp.dist == "Exponential":
        lambda_rate = 0.5
        for param in model.parameters():
            if (bias_flag == 0):
                param.data = torch.tensor(np.random.exponential(scale =
    1/lambda_rate, size = (param.size(0), param.size(1)))).float().to(device)
                bias_flag = 1
            else:
                param.data = torch.tensor(np.random.exponential(scale =
    1/lambda_rate, size = (param.size(0)))).float().to(device)
                bias_flag = 0
                i = i + 1
```

Figura 4.12: `main/exp_1_counterexample`, inizializzazione dei parametri con distribuzione $\mathcal{E}(0.5)$

Quello che si ottiene è un andamento evidentemente diverso dalla precedente sperimentazione. La variabile aleatoria risultante ha una densità non simmetrica e quindi non normale.
Il grafico 4.13 è generato con gli stessi parametri usati nella Sottosezione 4.2.1, in aggiunta è stata usata la stringa `dist = "Exponential"` per specificare la distribuzione di inizializzazione dei coefficienti della *NN*.

L'altra condizione necessaria per dare luogo alla gaussianità è imposta sulla funzione di attivazione con la quale si va a definire la *NN*. Tale funzione deve infatti soddisfare la Definizione 3.1.2; è facile notare che ogni funzione sigmoidale ed anche la ReLU soddisfano la suddetta proprietà; tuttavia esiste un'ampia classe di funzioni di attivazione comunemente utilizzate dette *Power Rectified Linear Unit* che non sono sub-lineari. Queste si definiscono come segue: $\forall\, k \in \mathbb{N}$ si definisce k-esima *Power Rectified Linear Unit function* la funzione

$$\text{PRLUk}(x) := \max\{0, x^k\}, \quad \text{ovvero} \quad \text{PRLUk}(x) = \text{ReLU}^k(x).$$

Anche in questo caso la simulazione condotta è simile a quella svolta nella Sottosezione 4.2.1 con l'unica differenza che la rete è generata con una attivazione facente parte della classe appena introdotta; in particolare `act = "PRLU3"`. Il resto dei parametri usati in `exp_1_counterexample` è scelto come nella Sottosezione 4.2.1 (quindi chiaramente `dist = "Normal"`). L'output che ne risulta ha una distribuzione molto concentrata sulla sua media ed, anche in questo caso, evidentemente non gaussiana; il grafico è riportato in fig. 4.14.

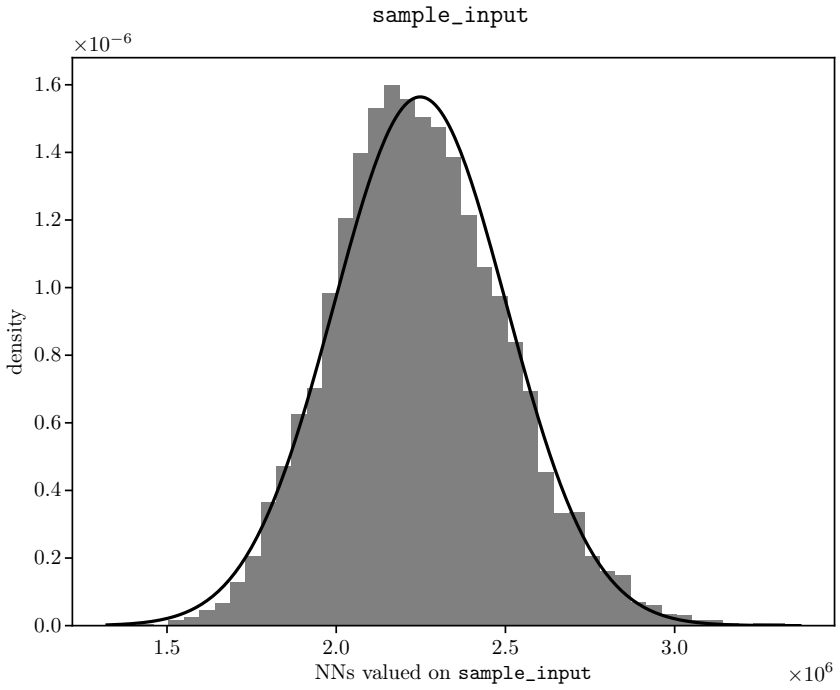

Figura 4.13: `exp_1_counterexample`, controesempio ad andamento gaussiano realizzato con una *NN* avente parametri distribuiti con legge $\mathcal{E}(0.5)$.

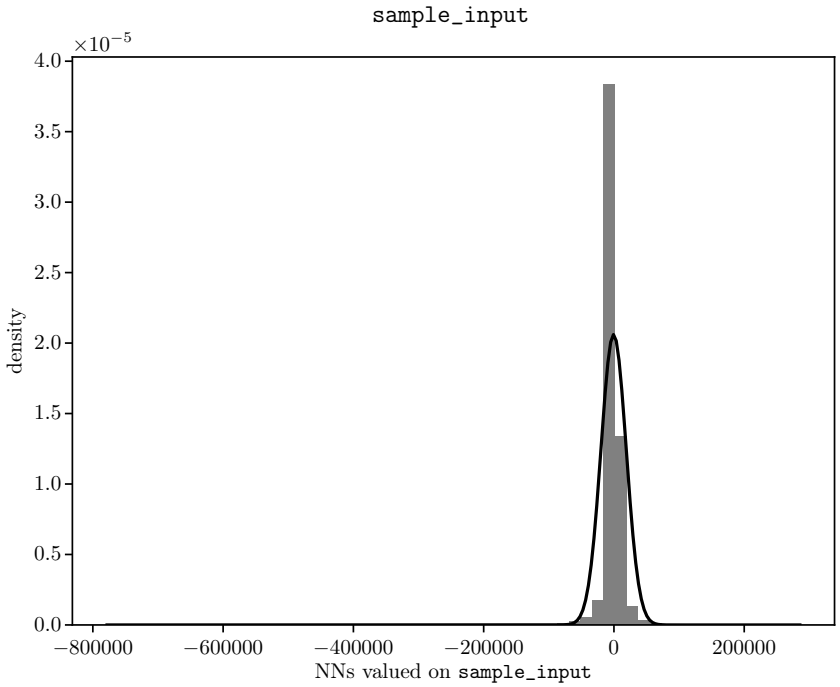

Figura 4.14: `exp_1_counterexample`, controesempio ad andamento gaussiano realizzato con una *NN* avente funzione di attivazione PRLU3.

Osservazione 4.2.2. I due controesempi presentati in questa sezione non rappresentano una dimostrazione della necessarietà delle ipotesi del Teorema di convergenza ad un processo gaussiano tuttavia permettono di affermare che provando ad indebolire queste ultime nella direzione dei controesempi si ha una scarsa probabilità di ottenere un risultato analogo o in qualche modo significativo.

4.2.3. Test di gaussianità in due dimensioni

In ultimo si va ad eseguire un'ulteriore simulazione volta a verificare che l'andamento riscontrato in reti neurali con parametri normalmente distribuiti e numero di neuroni considerevole abbia valenza anche se il numero di neuroni nel layer di output (`sizeoutput`) è maggiore di 1. Nello specifico, per poter visualizzare la densità della v.a. in output, si sceglie `sizeoutput = 2` così da ottenere una funzione $p_{\mathrm{NN}} : \mathbb{R}^2 \to \mathbb{R}$ ancora facilmente visualizzabile.
L'esperimento condotto ripercorre quello delle sottosezioni precedenti; infatti `exp_1_3d` ha una notevole somiglianza con le altre sperimentazioni, quello che lo distingue è la parte necessaria alla computazione della densità della *NN*.
Il risultato generato è quello riportato in fig. 4.15, ovvero un istogramma tridimensionale rappresentante la densità. Il grafico è eseguito su un dominio D contenuto nel piano $\mathbf{0}xy$ (con $\mathbf{0} = (0, 0, 0)$). D è partizionato in quadrati $Q_{i,j}$ a ciascuno dei quali corrisponde una z così determinata:

$$\forall\,(x, y) \in Q_{i,j} \quad \text{si ha} \quad p_{\mathrm{NN}}(x, y) = \frac{\mathbb{P}(\mathrm{NN} \in Q_{i,j})}{\mathrm{area}(Q_{i,j})},$$

dove la probabilità è calcolata come

$$\mathbb{P}(\mathrm{NN} \in Q_{i,j}) = \frac{\#\{(x, y) \in \texttt{output} \mid (x, y) \in Q_{i,j}\}}{\#\texttt{output}}.$$

Al fine di ottenere un risultato il più possibile *vicino* a quello asintotico sono stati utilizzati i seguenti parametri: per la struttura della *NN*

`nh`	`sizeinput`	`sizeoutput`	`act`	`growth`	`ty`
1	3	2	`"ReLU1"`	100000	`"const"`

e per il resto della simulazione `ntest = 100000` [4].
Per ottenere una rappresentazione più uniforme della funzione p_{NN} si è eseguito un ulteriore grafico a partire dai medesimi dati. Questa volta si sono presi come riferimento i centri di ciascuno dei quadrati $Q_{i,j}$ e, tramite il comando `ax.plot_trisurf()`, è stata realizzata la fig. 4.16.

[4] Si noti che il numero di test da effettuare per ottenere un campionamento accurato della funzione di densità cresce molto in fretta, dunque se 1000 test sono più che sufficienti nel caso di un neurone nel layer di output della rete, 100000 è un quantitativo necessario quando i neuroni diventano due.

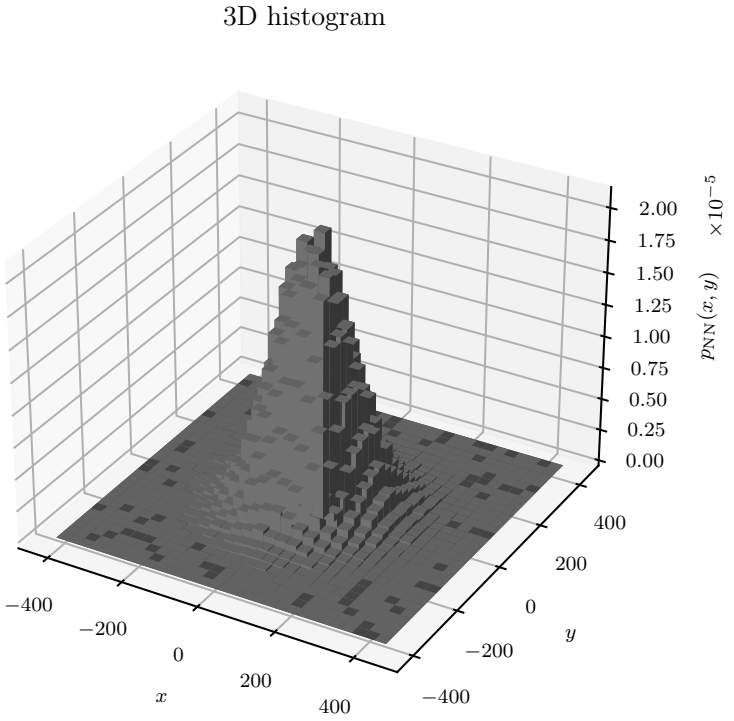

Figura 4.15: `exp_1_3d`, grafico della funzione p_{NN} (istogramma 3D).

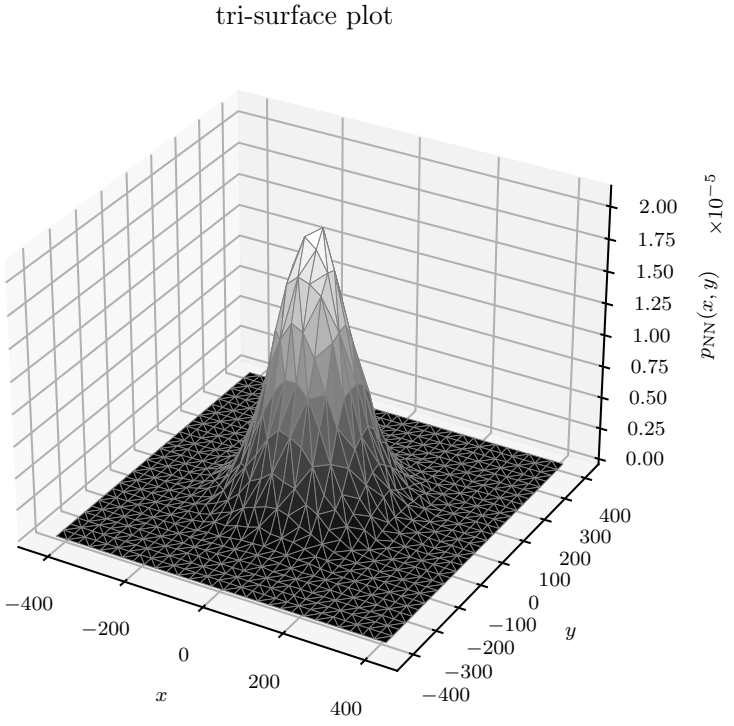

Figura 4.16: `exp_1_3d`, grafico della funzione p_{NN} (tri-surface plot).

Osservazione 4.2.3. Sebbene il risultato teorico dato dal Teorema di convergenza ad un processo gaussiano sia ragionevolmente confermato dalle sperimentazioni svolte, si è notato che al crescere del numero di hidden layers il numero di neuroni che si devono inserire in ciascun layer al fine di ottenere un comportamento gaussiano cresce molto rapidamente.
Una rete neurale di tipo molto *deep* avrà un andamento meno *normale*, a parità di neuroni, rispetto ad una rete con 2 o 3 hidden layers. Questo particolare comportamento fa sì che le simulazioni atte a confermare quanto riportato nel Capitolo 3 diventino computazionalmente proibitive nei casi *deep*, almeno per come sono state implementate nel presente lavoro.

4.3. `Exp_2`: approssimazione universale

Nella seconda, ed ultima, sperimentazione ci si è posti come obiettivo principale quello di riuscire ad ottenere, partendo da una qualsiasi $f \in C(X)$ con $X \subset \mathbb{R}^k$ compatto, una sua approssimazione $g \in \mathcal{N}_k(\psi)$ con ψ funzione di attivazione, continua e sigmoidale, in accordo con le ipotesi dei Teoremi di approssimazione presenti nel Capitolo 2. Non verranno usate esclusivamente funzioni $f \in \Gamma_{r,c}$ in quanto verificare l'appartenenza a tale insieme è un problema non facilmente sistematizzabile[5].
Quello che si andrà a fare concretamente, dopo aver scelto una f, è utilizzare il modello di *NN* introdotto nelle sezioni precedenti ed effettuare un addestramento dei parametri al fine di *avvicinare* la rete alla suddetta funzione. Tutto ciò viene fatto per semplicità con $f : \mathbb{R} \to \mathbb{R}$, ovvero con $k = 1$ per non appesantire il costo computazionale delle sperimentazioni.
Per quanto concerne la strategia di addestramento dei parametri si utilizzano le funzioni già presenti all'interno della libreria `torch`. Gli strumenti utilizzati sono quelli relativi agli algoritmi di training più convenzionali: si utilizza una *Stochastic Gradient Descent (SDG)* per minimizzare una *loss function* denominata in `PyTorch` `MSELoss`, funzione che calcola l'errore quadratico medio tra l'attuale output della rete e quello desiderato.

4.3.1. Funzionamento di `Exp_2`

Per meglio spiegare il modo in cui la sperimentazione è stata condotta è utile analizzare la struttura di `exp_2` per punti. In una prima fase si esegue un semplice campionamento casuale della funzione scelta f (`nsample` campioni) sul dominio compatto $[a, b]$ sul quale è stata definita. Successivamente si costruisce il modello per la *NN* e lo si inizializza come fatto in fig. 4.1 e 4.2. A questo punto inizia la fase del training; in questa sezione vengono richiamate le funzioni della libreria `torch` e si esegue un training standard. Si procede, all'interno di un ciclo `for` per `nepoch` volte[6], con le seguenti istruzioni $\forall\, s = 0, \ldots,$ `nsample` $- 1$:

1. assegnazioni ad `input` ed `exact_output` dei valori corrispondenti alla funzione f nell'`s`-esimo campione (ovvero x ed $f(x)$);

[5] Si ricorda, tuttavia, che i polinomi con dominio ristretto ad un compatto sono appartenenti a $\Gamma_{r,c}$, come osservato alla fine della Sezione 2.3.

[6] In realtà `nepoch` rappresenta un valore massimo del numero di loop eseguiti; se la norma uniforme della funzione $model f - f$ raggiunge il valore minimo di soglia (`err`) si ha un'uscita anticipata.

2. **forward propagation**:

   - computo dell'`output` della rete applicata ad `input`;
   - computo della `loss` tra `output` ed `exact_output` tramite la `loss_function` (ovvero `nn.MSELoss(reduction = "mean")`);

3. **backward propagation**:

   - computo del gradiente della `loss`;
   - esecuzione di un singolo step di ottimizzazione dei parametri della rete sulla base del gradiente appena calcolato, tramite l'`optimizer` scelto (ovvero `optim.SGD(model.parameters(), lr = inp.chosen_lr)`) [7].

Nella parte finale si computano gli `output` della funzione $modelf$ generata dalla rete su tutto il dominio $[a, b]$ e si eseguono i test su tale funzione.
Per capire a pieno le potenzialità delle *NN* sono state condotte 3 distinte sperimentazioni atte ad approssimare delle funzioni continue definite su compatti.

4.3.1.1. Exp_2 su un polinomio

La prima funzione che si è cercato di approssimare è $f(x) = 5x^4 - 2x^2 - x + 7$, sull'intervallo $[a, b] = [-4, 4]$. Per l'addestramento si sono utilizzati `nsample` $= 100$ campioni ed una rete neurale generata a partire dai parametri in tabella:

nh	sizeinput	sizeoutput	act	growth	ty
6	1	1	"ReLU"	100	"const"

.

Il learning rate scelto per eseguire l'algoritmo di discesa al gradiente è il seguente: `chosen_lr = 3.5*1e-7`. Si è infine fissata una soglia massima d'errore sulla norma uniforme pari a `err = 125`, una volta scesa sotto tale valore il training termina. Con tali inizializzazioni è stata prodotta, in 97 `epochs`, l'approssimazione in fig. 4.17.
In fig. 4.18 è riportato il grafico della `loss` media alla fine di ogni `epoch`: ad ogni $i \in \{0, \dots, \texttt{nepoch}\}$, in ascissa, corrisponde un valore sull'ordinata pari alla perdita media su tutti gli `nsample` campioni (media aritmetica).

Osservazione 4.3.1. Si può notare che la funzione in fig. 4.18 non appare come una discesa liscia; tale comportamento dipende dalla scelta del learning rate che influenza notevolmente la velocità di apprendimento della *NN*. Eseguire, infatti, la stessa *SDG* con un learning rate troppo alto può causare un apprendimento *troppo rapido* con conseguente eccessiva oscillazione della perdita. Analogamente se il parametro è troppo basso si ha un apprendimento inutilmente rallentato; in tal caso saranno necessarie moltissime `epochs` per produrre un'approssimazione discreta.
Il risultato ottenuto rappresenta tuttavia una buona approssimazione di f (infatti $\|modelf - f\|_\infty = 100.08 \approx \frac{1}{10}\|f\|_\infty$) raggiunta in una quantità ragionevole di passi.

[7] Per l'esecuzione dell'algoritmo di discesa si deve dichiarare il parametro *learning rate* che va ad influire sulla *velocità* con la quale viene eseguita la discesa al gradiente: tale parametro è denotato come `chosen_lr`.

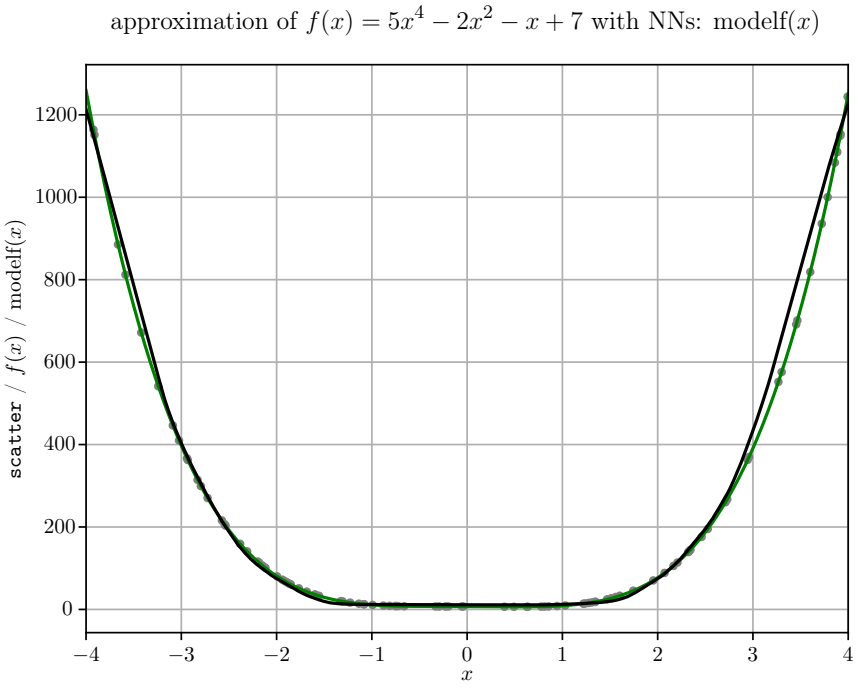

Figura 4.17: `exp_2`, in grigio lo scatter degli `nsample` campioni, in verde la funzione originale f, in nero l'approssimazione di $f(x) = 5x^4 - 2x^2 - x + 7$ tramite la *NN* con attivazione ReLU.

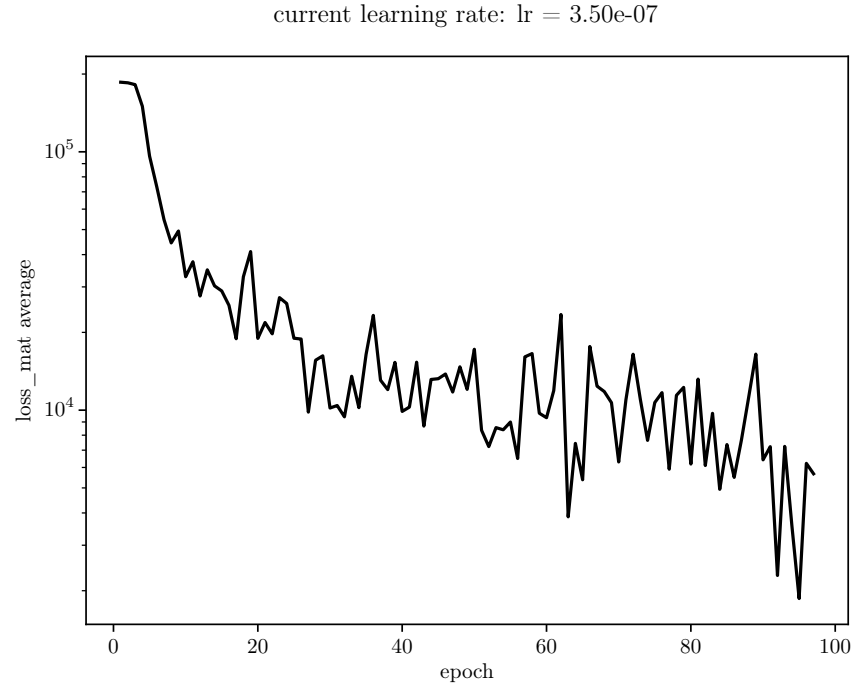

Figura 4.18: `exp_2`, grafico della perdita durante lo scorrimento delle `epochs` nell'approssimazione di $f(x) = 5x^4 - 2x^2 - x + 7$ in scala semilogaritmica.

Osservazione 4.3.2. Di particolare importanza è il fatto che sia stata usata l'attivazione ReLU. La medesima sperimentazione condotta con la funzione sigmoidale ReLU1 conduce, dopo 200 `epochs`, ad una approssimazione che risulta *troncata*, come si vede in fig. 4.19 [8]. Come si vede nel grafico, la *NN* approssima in maniera molto precisa la f solo per le x tali che $f(x) \lesssim 150$. Tutto ciò è dovuto all'applicazione dell'attivazione che causa un appiattimento di $modelf$ risolubile solo con una modifica dell'architettura della *NN*. Quanto osservato nella Sezione 2.4 non è quindi soltanto un risultato teorico ma è anche di fondamentale importanza nelle applicazioni concrete.

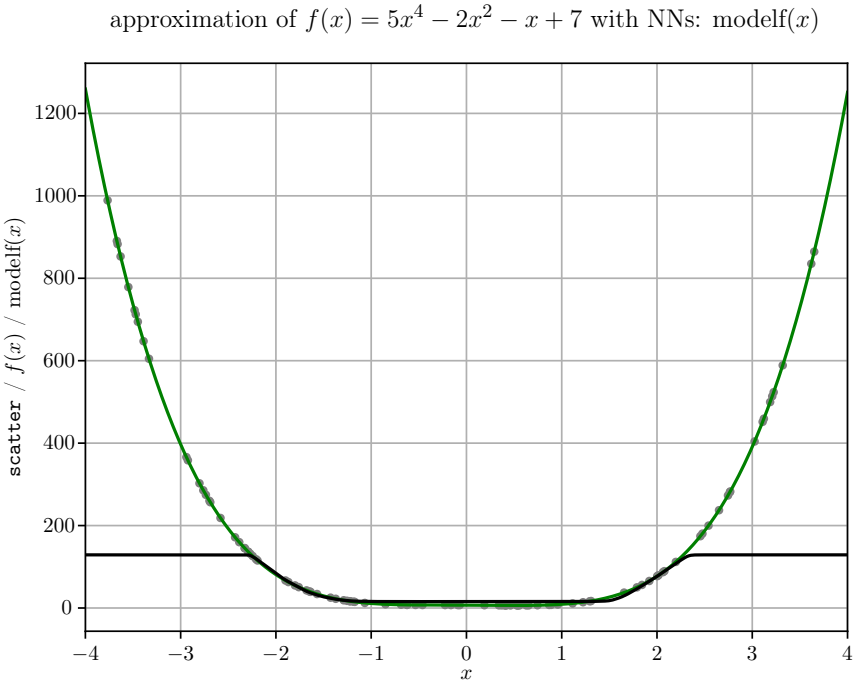

Figura 4.19: `exp_2`, in grigio lo scatter degli `nsample` campioni, in verde la funzione originale f, in nero l'approssimazione di $f(x) = 5x^4 - 2x^2 - x + 7$ tramite la *NN* con attivazione ReLU1.

4.3.1.2. Exp_2 su una funzione goniometrica

Successivamente si è passati ad una funzione goniometrica, con lo scopo di testare le potenzialità della rete nell'approssimare una funzione con delle oscillazioni.
Per raggiungere l'obiettivo è stata campionata, su `nsample = 100` punti aleatori, la funzione $f(x) = \cos(x^2) - \sin(2x + 3)$ sull'intervallo $[a, b] = [-2, 2]$.
Per l'approssimazione di questa funzione sono state testate molte combinazioni dei parametri usuali; in particolare l'approssimazione è riuscita, con risultati via via sempre migliori al crescere di `nh`.

[8] Anche utilizzando gli stessi parametri per la simulazione, cambiando l'attivazione va conseguentemente cambiato anche il learning rate: in questo caso `chosen_lr = 5*1e-7`.

In tutti i casi sono stati fissati i parametri della rete come segue:

`sizeinput`	`sizeoutput`	`act`	`growth`	`ty`
1	1	`"ReLU1"`	100	`"const"`

.

Si riporta, per brevità, la stima più accurata ottenuta con 10 hidden layers dopo 263 `epochs`; per questo test il valore di `err` è stato fissato a 0.15 ($\approx \frac{1}{20}\|f\|_\infty$) ed è stato usato `chosen_lr = 5.5*1e-3`.
I risultati sono riportati nelle fig. 4.20 e 4.21.

Osservazione 4.3.3. In questo caso la funzione $modelf$ approssima quasi alla perfezione la funzione f, discostandosi infatti da quest'ultima solo nelle zone di scarsa densità dei campioni aleatori.
Si nota inoltre che, in questo caso, la scelta del learning rate è stata efficace in quanto la funzione in fig. 4.21 presenta una decrescenza marcata, sebbene siano presenti delle forti oscillazioni.
A differenza dell'esempio precedente, qui l'attivazione ReLU1 non risulta dannosa ai fini dell'approssimazione, nemmeno nel caso di `nh = 1`, poiché il valore di $\|f\|_\infty$ è ridotto e quindi l'appiattimento dell'attivazione non influisce sull'andamento di $modelf$.

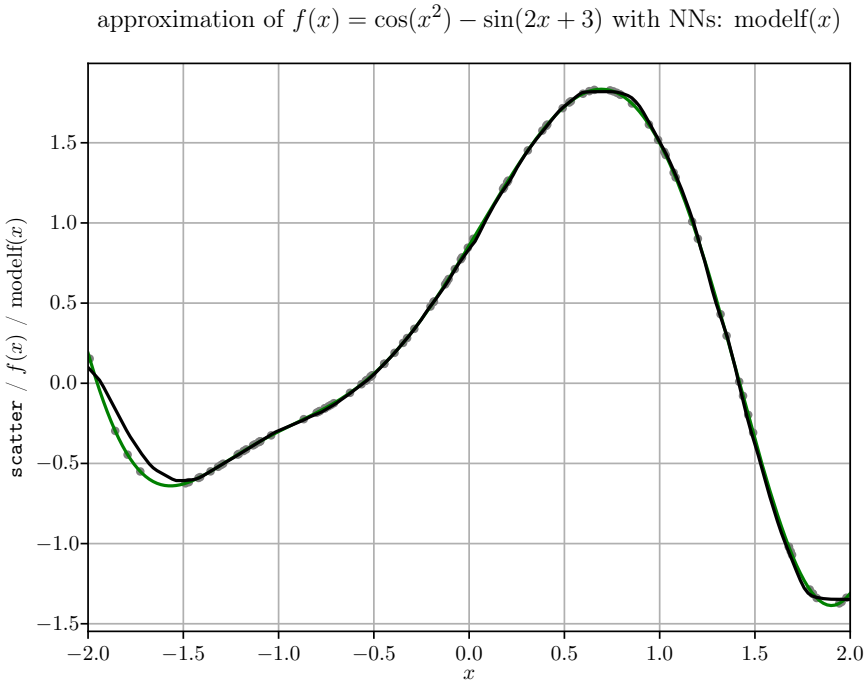

Figura 4.20: `exp_2`, in grigio lo scatter degli `nsample` campioni, in verde la funzione originale f, in nero l'approssimazione di $f(x) = \cos(x^2) - \sin(2x+3)$ tramite la NN.

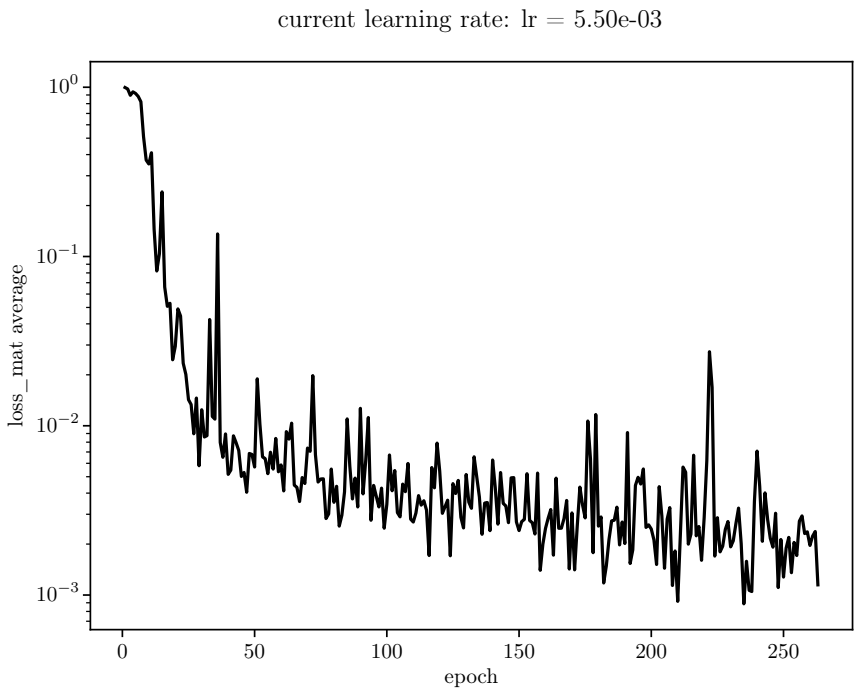

Figura 4.21: `exp_2`, grafico della perdita durante lo scorrimento delle `epochs` nell'approssimazione di $f(x) = \cos(x^2) - \sin(2x + 3)$ in scala semilogaritmica.

4.3.1.3. Exp_2 su una funzione non derivabile

In ultimo si è cercato di approssimare, utilizzando diversi tipi di attivazione, una funzione che fosse continua sul suo dominio, ma non derivabile. Questa sperimentazione è stata sviluppata a partire da quanto detto alla fine della Sottosezione 4.1.1, dove si è osservato che una rete con attivazione regolare risulta poco adatta ad approssimare un punto di non derivabilità.

La funzione presa in considerazione nella presente sottosezione è $f(x) = |x|e^x$ sul dominio $[a, b] = [-1, 1]$ campionata su `nsample` $= 50$ punti aleatori. Tale funzione presenta un punto angoloso in $x = 0$.
Sono state condotte 2 sperimentazioni in parallelo con i seguenti parametri in comune:

`nh`	`sizeinput`	`sizeoutput`	`growth`	`ty`
4	1	1	100	`"const"`

.

Per il primo modello è poi stata usata l'attivazione ReLU1, mentre per il secondo la funzione Sigmoid. Le approssimazioni sono state ottenute rispettivamente con `chosen_lr = 6*1e-2` e `chosen_lr = 1*1e-2` e sono riportate di seguito nelle fig. 4.22 e 4.23.

Osservazione 4.3.4. Come previsto l'attivazione ReLU1 conduce ad una stima più accurata ed in meno passi.

Dopo 200 `epochs` si raggiungerà, infatti, il valore $\|modelf - f\|_\infty = 0.11 \approx \frac{1}{20}\|f\|_\infty$ con la rete che utilizza la funzione ReLU1; invece solamente dopo 1487 `epochs` si raggiungerà il valore $\|modelf - f\|_\infty = 0.2 \approx \frac{1}{10}\|f\|_\infty$ con la rete che utilizza la funzione Sigmoid.
Indipendentemente dalla distanza uniforme raggiunta, comunque, è facile notare che la prima delle due approssimazioni si avvicina al punto angoloso in $x = 0$, senza tuttavia raggiungerlo in quanto 0 non fa parte dell'insieme degli `nsample` campioni. La seconda, invece, è una funzione molto regolare, così come l'attivazione con la quale è stata generata, perciò in $x = 0$ non riesce ad avvicinarsi ad $f(x_0)$.
Va tuttavia sottolineato che aumentando il numero di hidden layers si riesce comunque ad ottenere una stima accurata di f, anche usando l'attivazione Sigmoid.

Osservazione 4.3.5. Quanto osservato riguardo alle differenti capacità di approssimazione delle reti neurali con diverse funzioni di attivazione è già stato notato sperimentalmente in numerose applicazioni pratiche. Comunemente si preferisce utilizzare attivazioni come la Sigmoid nei casi di reti neurali ricorrenti o per modelli di classificazione binaria, piuttosto che per reti neurali completamente connesse feed-forward come quelle prese in analisi nel presente lavoro [5].

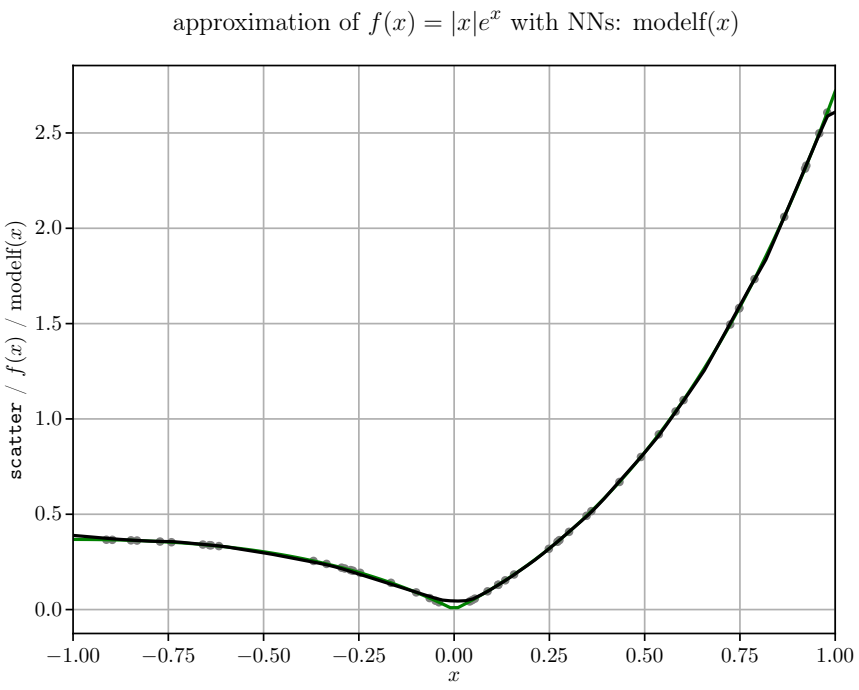

Figura 4.22: `exp_2`, in grigio lo scatter degli `nsample` campioni, in verde la funzione originale f, in nero l'approssimazione di $f(x) = |x|e^x$ tramite la NN con attivazione ReLU1.

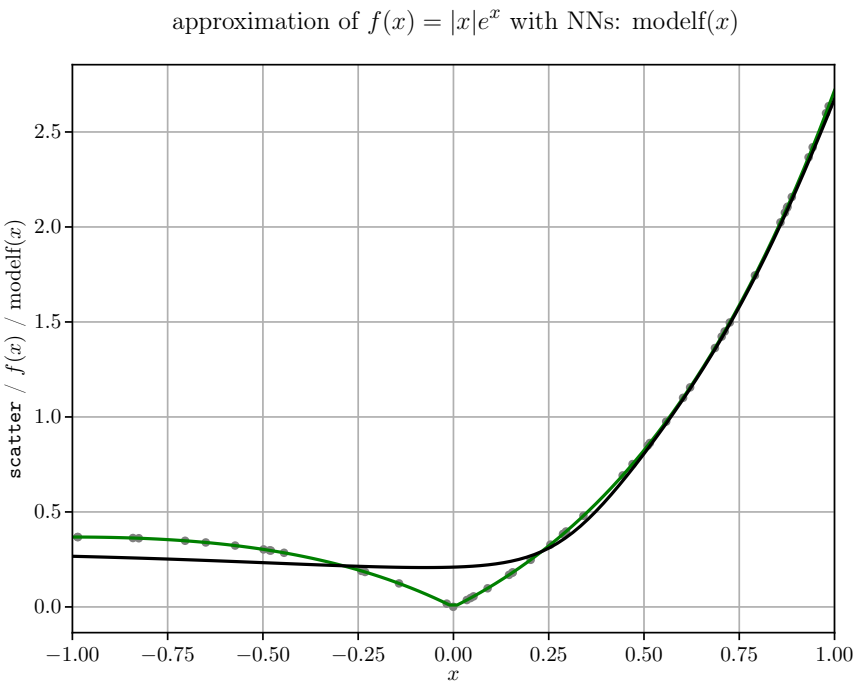

Figura 4.23: `exp_2`, in grigio lo scatter degli `nsample` campioni, in verde la funzione originale f, in nero l'approssimazione di $f(x) = |x|e^x$ tramite la *NN* con attivazione Sigmoid.

Appendice A

Dimostrazioni ed altri risultati

A.1. Richiami

Di seguito sono riportate le dimostrazioni di alcuni dei risultati presentati nel Capitolo 1.

A.1.1. Analisi funzionale

Dimostrazione del Corollario 1.1.1. Essendo $\overline{Y} \subsetneq X$ esiste $x_0 \in X \setminus \overline{Y}$. Denotando con Z lo spazio vettoriale generato da Y e x_0 si ha che, essendo Y un sottospazio vettoriale e $x_0 \notin Y$, $\forall\, z \in Z$ vale $z = y + \lambda x_0$.
Si può quindi considerare il funzionale lineare continuo $g^*(z) = g^*(y + \lambda x_0) = \lambda$.
Vale immediatamente dalla definizione che $g^*(x_0) = 1$ e che $\forall\, y \in Y$, $g^*(y) = 0$.
Dunque esiste un funzionale in Z^* che ha le proprietà cercate, è sufficiente trovarne uno analogo in X^* così da avere la tesi.
Per concludere si applica il Teorema di Hahn-Banach con

$$p(x) = \|g^*\|\|x\| \quad e \quad f^*(x) = g^*(x).$$

Tali funzioni soddisfano le proprietà richieste dal Teorema 1.1.1, infatti $\forall\, x, y \in X,\ \forall\, \lambda \geq 0$,

$$p(x+y) = \|g^*\|\|x+y\| \leq \|g^*\|\|x\| + \|g^*\|\|y\| = p(x) + p(y)$$

$$\text{e} \quad p(\lambda x) = \|g^*\|\lambda\|x\| = \lambda p(x);$$

inoltre $f^* \in Z^*$ e $\forall\, z \in Z$,

$$g^*\left(\frac{z}{\|z\|}\right) \leq \sup_{\|z\|=1} |g^*(z)| = \|g^*\| \implies f^*(z) = g^*(z) \leq \|g^*\|\|z\| = p(z).$$

Dunque $\exists\, h^* \in X^*$ tale che

$$\forall\, z \in Z, \quad h^*(z) = g^*(z) \quad \text{e} \quad \forall\, x \in X, \quad h^*(x) \leq p(x).$$

Quindi $\forall\, y \in Y$ vale $h^*(y) = 0$ ed essendo $x_0 \in Z$ si ha $h^*(x_0) = g^*(x_0) = 1$, che implica $h^* \neq 0$. □

A.1.2. Teoria della misura

Dimostrazione del Teorema 1.2.4. ν è finita, infatti per Hölder, definito p tale che $\frac{1}{p} + \frac{1}{q} = 1$, si ha

$$\forall\, B \in \mathcal{A}, \quad |\nu(B)| = \left| \int_X \mathbb{1}_B f \, d\mu \right| \leq \|\mathbb{1}_B\|_p \|f\|_q \leq (\mu(X))^{\frac{1}{p}} \|f\|_q < \infty.$$

Definendo $P = \{x \in X \mid f(x) \geq 0\}$ e $N = \{x \in X \mid f(x) < 0\}$, valgono le seguenti proprietà:

- $P \cup N = X$ e $P \cap N = \emptyset$;
- $\forall\, E \in \mathcal{A}$, $E \subset P$ si ha $\nu(E) = \int_E f\, d\mu \geq 0$;
- $\forall\, E \in \mathcal{A}$, $E \subset N$ si ha $\nu(E) = \int_E f\, d\mu \leq 0$.

Definendo dunque $\nu = \nu^+ - \nu^-$ come nel Teorema 1.2.2, si nota che

$$\forall\, E \in \mathcal{A}, \quad \nu^+(E) = \nu(E \cap P) = \int_{E\cap P} f\, d\mu = \int_E f^+\, d\mu,$$

ovvero $\nu^+ = f^+.\mu$. Analogamente $\nu^- = f^-.\mu$.
Quindi dall'analogo teorema nel caso di misure positive con densità segue che

$$\int_X \varphi\, d\nu = \int_X \varphi\, d\nu^+ - \int_X \varphi\, d\nu^- = \int_X \varphi f^+\, d\mu - \int_X \varphi f^-\, d\mu = \int_X \varphi f\, d\mu.$$

□

A.1.3. Variabili aleatorie discrete su spazi di Hilbert

Dimostrazione della Proposizione 1.4.3.

$$\begin{aligned}
\mathbb{E}[f \cdot g] &= \sum_{(f^*,g^*)\in K\times K} f^* \cdot g^* \mathbb{P}_{f,g}(f^*, g^*) = \\
&= \sum_{(f^*,g^*)\in K\times K} f^* \cdot g^* \mathbb{P}_f(f^*)\mathbb{P}_g(g^*) = \\
&= \sum_{(f^*,g^*)\in K\times K} (\mathbb{P}_f(f^*)f^*) \cdot (\mathbb{P}_g(g^*)g^*) = \\
&= \left(\sum_{(f^*)\in K} \mathbb{P}_f(f^*)f^*\right) \cdot \left(\sum_{(g^*)\in K} \mathbb{P}_g(g^*)g^*\right) = \\
&= \mathbb{E}[f] \cdot \mathbb{E}[g],
\end{aligned}$$

dove la prima identità è data dalla Proposizione 1.4.1 e, per passare dalla prima alla seconda riga, si è usata la Proposizione 1.4.2. □

Dimostrazione della Proposizione 1.4.4.

$$\begin{aligned}
\mathrm{Var}\,(f+g) &= \mathrm{Cov}\,(f+g, f+g) = \\
&= \mathrm{Var}\,(f) + \mathrm{Var}\,(g) + 2\mathrm{Cov}\,(f,g)\,,
\end{aligned}$$

dove la bilinearità della covarianza segue dalla linearità di prodotto scalare e valore atteso.
Inoltre essendo f e g indipendenti la loro covarianza è nulla, cosa che segue dalla Proposizione 1.4.3; infatti se

$$\mathbb{E}[f \cdot g] = \mathbb{E}[f] \cdot \mathbb{E}[g],$$

usando la definizione alternativa per la covarianza data nell'Osservazione 1.4.1 si ha

$$\mathrm{Cov}\,(f,g) = \mathbb{E}[f \cdot g] - \mathbb{E}[f] \cdot \mathbb{E}[g] = 0.$$

□

A.2. Risultati ausiliari del Capitolo 3

La dimostrazione del Teorema 3.2.1 necessita di diversi enunciati di supporto che vengono riportati a seguire nell'ordine in cui vengono utilizzati nel lavoro. Solo una parte più semplice di questi risultati viene provata nella presente sezione, dei rimanenti vengono comunque riportate le fonti. Di particolare importanza è il contributo di Blum et al. [4, p. 227], questi hanno infatti sviluppato una prima versione del CLT per successioni di v.a. scambiabili rielaborata poi nell'articolo di G. Matthews et al. [8, p. 22].

Nel seguito si assumono le v.a. X_n, X, Y come funzioni da $(\Omega, \mathcal{A}, \mathbb{P})$ in $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$ ($k = +\infty$ eventualmente).

Proposizione A.2.1. *Se $X_n \xrightarrow{\mathcal{L}} X$ e Y indipendente dalla successione $(X_n)_{n\in\mathbb{N}}$ allora $X_n + Y \xrightarrow{\mathcal{L}} X + Y$ (in realtà converge a $X' + Y'$ con X' e Y' copie indipendenti di X e Y).*

Dimostrazione. Per il Teorema di continuità di Lévy vale $\forall\, t \in \mathbb{R}^k,\ \varphi_{X_n}(t) \to \varphi_X(t)$ perciò $\forall\, t \in \mathbb{R}^k$ si ha

$$\varphi_{X_n+Y}(t) = \varphi_{X_n}(t)\varphi_Y(t) \to \varphi_X(t)\varphi_Y(t) = \varphi_{X'}(t)\varphi_{Y'}(t) = \varphi_{X'+Y'}(t),$$

dove per il primo e l'ultimo passaggio si è usata l'indipendenza delle v.a.. Di nuovo, per il Teorema di continuità di Lévy si ha la tesi. □

Teorema A.2.1 (Teorema del limite centrale per successioni di variabili aleatorie scambiabili)**.** [4, p. 227] *$\forall\, n \in \mathbb{N}_0$ sia $(X_{n,j})_{j\in\mathbb{N}_0}$*

a. un processo numerabile scambiabile con

b. $\mathbb{E}[X_{n,1}] = 0$,

c. $Var(X_{n,1}) = \sigma_n^2 < \infty$, $\sigma_n^2 \xrightarrow{n\to\infty} \sigma_*^2$ *e*

d. $\mathbb{E}\left[|X_{n,1}|^3\right] < \infty$.

Si definisce

$$S_n := \frac{1}{\sqrt{h(n)}} \sum_{j=1}^{h(n)} X_{n,j}$$

con $h : \mathbb{N} \to \mathbb{N}$ funzione strettamente crescente. Se valgono le condizioni

1. $\mathbb{E}[X_{n,1}X_{n,2}] = 0$,
2. $\mathbb{E}\left[|X_{n,1}X_{n,2}|^2\right] \xrightarrow{n\to\infty} \sigma_*^4$,
3. $\mathbb{E}\left[|X_{n,1}|^3\right] = o(\sqrt{h(n)})$,

allora

$$S_n \xrightarrow{\mathcal{L}} \mathcal{N}^1\left(0, \sigma_*^2\right)\ ^1.$$

[1] Per convenzione $\mathcal{N}^1(0, 0) := 0$.

Teorema A.2.2 (Teorema di de Finetti)**.** [12] *Una successione di v.a.* $(X_j)_{j\in\mathbb{N}}$ *condizionatamente indipendenti ed identicamente distribuite è scambiabile.*

Teorema A.2.3. *Se* $X_n \xrightarrow{\mathcal{L}} X$ *e* $\phi : \mathbb{R}^k \to \mathbb{R}^k$ *è una funzione continua allora vale* $\phi(X_n) \xrightarrow{\mathcal{L}} \phi(X)$.

Dimostrazione. La convergenza in legge di X_n a X si definisce come convergenza stretta delle misure immagine, dunque si può scrivere:

$$\forall\, f \in C_b(\mathbb{R}^k) \quad \mathbb{E}\left[f(X_n)\right] \to \mathbb{E}\left[f(X)\right]. \tag{A.1}$$

Di conseguenza basta mostrare che

$$\forall\, f \in C_b(\mathbb{R}^k) \quad \mathbb{E}\left[f(\phi(X_n))\right] \to \mathbb{E}\left[f(\phi(X))\right],$$

e ciò è immediato perché $f \circ \phi$ è a sua volta continua e limitata, dunque per l'equazione (A.1) si ha la tesi. □

Bibliografia

[1] Robert B. Ash. *Real Analysis and Probability*. USA: Academic Press, 1972.

[2] Andrew R. Barron. «Universal approximation bounds for superpositions of a sigmoidal function». In: *IEEE Transactions on Information Theory* 39.3 (1993), pp. 930–945. DOI: 10.1109/18.256500.

[3] Patrick Billingsley. *Convergence of Probability Measures*. 2ª ed. USA: John Wiley & Sons, 1999.

[4] Julius R. Blum et al. «Central Limit Theorems for Interchangeable Processes». In: *Canadian Journal of Mathematics* 10 (1958), pp. 222–229. DOI: 10.4153/CJM-1958-026-0.

[5] Jason Brownlee. *How to choose an activation function for deep learning*. URL: https://machinelearningmastery.com/choose-an-activation-function-for-deep-learning/. (accessed: 27.04.2022).

[6] George Cybenko. «Approximation by superposition of a sigmoidal function». In: *Mathematics of Control, Signals and Systems* 2 (1989), pp. 303–314. DOI: 10.1007/BF02551274.

[7] Avner Friedman. *Foundations of Modern Analysis*. USA: Dover Publications, 1982.

[8] Alexander G. de G. Matthews et al. «Gaussian Process Behaviour in Wide Deep Neural Networks». In: (2018), pp. 1–36. DOI: 10.48550/arXiv.1804.11271.

[9] Loukas Grafakos. *Classical Fourier Analysis*. 3ª ed. New York: Springer, 2014. DOI: 10.1007/978-1-4939-1194-3.

[10] Kurt Hornik. «Approximation capabilities of multilayer feedforward networks». In: *Neural Networks* 4.2 (1991), pp. 251–257. DOI: 10.1016/0893-6080(91)90009-T.

[11] Kurt Hornik, Maxwell Stinchcombe e Halbert White. «Multilayer feedforward networks are universal approximators». In: *Neural Networks* 2.5 (1989), pp. 359–366. DOI: 10.1016/0893-6080(89)90020-8.

[12] Paul Ressel. «De Finetti-type Theorems: An Analytical Approach». In: *The Annals of Probability* 13.3 (1985), pp. 898–922. DOI: 10.1214/aop/1176992913.

[13] Walter Rudin. *Functional Analysis*. USA: McGraw-Hill, 1973.

[14] Walter Rudin. *Real and Complex Analysis*. 3ª ed. USA: McGraw-Hill, 1987.