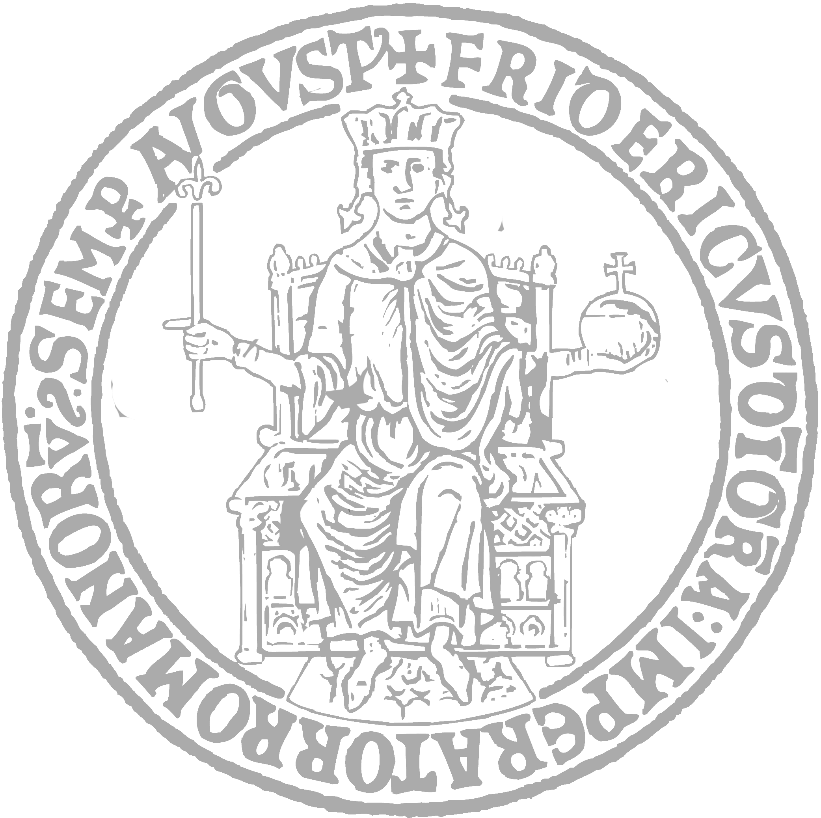
MLeA – Mod. B 2017/2018  
Progetto 5

Prof. Roberto Prevete

Djihad Boukara – N97000275

Emanuele Cioffi – N97000277



# Indice Generale

Sommario

[1. Indice Generale 2](#_Toc527824329)

[2. Indice delle figure 2](#_Toc527824330)

[3. Traccia 3](#_Toc527824331)

[4. Descrizione del problema 3](#_Toc527824332)

[5. Dataset 4](#_Toc527824333)

[6. Implementazione 4](#_Toc527824334)

[A. La rete neurale 4](#_Toc527824335)

[B. L’addestramento 5](#_Toc527824336)

[7. Test Plan 10](#_Toc527824337)

[8. K-Folding 12](#_Toc527824338)

[A. Dati Risultato 12](#_Toc527824339)

[B. Osservazioni sul K-Folding 14](#_Toc527824340)

[9. Test finali e conclusioni 15](#_Toc527824341)

# Indice delle figure

[Figura 1 Esempio di creazione rete neurale 4](#_Toc527822401)

[Figura 2 – Train 5](#_Toc527822402)

[Figura 3 – BackPropagation 6](#_Toc527822403)

[Figura 4 - Forward Propagation 6](#_Toc527822404)

[Figura 5 – SoftMax 7](#_Toc527822405)

[Figura 6 - Cross Entropy 8](#_Toc527822406)

[Figura 7 - Calcolo Derivate dei pesi 8](#_Toc527822407)

[Figura 8 - Discesa del gradiente 9](#_Toc527822408)

[Figura 9 – K-fold 10](#_Toc527822409)

[Figura 10 - SSE: Sigmoide, Identità 12](#_Toc527822410)

[Figura 11 - SSE: TanH, ReLU 13](#_Toc527822411)

[Figura 12 - SSE: TanH, Sigmoide 13](#_Toc527822412)

[Figura 13 - x = Epoche, y = Accuratezza 15](#_Toc527822413)

[Figura 14 - x = epoche, y = errore totale 16](#_Toc527822414)

[Figura 15 - Output finale, 60k esempi d'addestramento, on-line learning, una epoca 17](#_Toc527822415)

# Traccia

**PARTE A.**

“• Progettazione ed implementazione di funzioni per simulare la propagazione in avanti di una rete neurale multi-strato con almeno:

* due strati di pesi, con la sigmoide come funzione di output dei nodi interni e l'identità come funzione di output dei nodi di output.

(FACOLTATIVO: permettere all'utente di implementare reti con più di uno strato di nodi interni e con qualsiasi funzione di output per ciascun strato).

• Progettazione ed implementazione di funzioni per la realizzazione della back-propagation per reti neurali multi-strato con almeno:

* due strati di pesi, con la sigmoide come funzione dioutput dei nodi interni e l'identità come funzione di output dei nodi di output, con la somma dei quadrati come funzione di errore.

(FACOLTATIVO: permettere all'utente di realizzare la back-propagation con più di uno strato di nodi interni, con qualsiasi funzione di output per ciascun strato e con qualsiasi funzione di errore derivabile rispetto all'output).”

**PARTE B.**

“Si consideri come input le immagini raw del dataset mnist. Si ha, allora,un problema di classificazione a C classi, con C=10. Si estragga opportunamente un dataset globale di N coppie (ad esempio, N=200). Si fissi la discesa del gradiente come algoritmo di aggiornamento dei pesi, ed una rete neurale con un unico strato di nodi interni. Si scelgano gli iper-parametri del modello, cioè eta della regola di aggiornamento ed il numerodi nodi interni, sulla base di un approccio di cross-validation k-fold (ad esempio k=5).

Scegliere e mantenere invariati tutti gli altri "parametri" come, ad esempio, le funzioni di output e la funzione di errore. Se è necessario, per questioni di tempi computazionali espazio in memoria, si possono ridurre (ad esempio dimezzarle) le dimensioni delle immagini raw del dataset mnist (ad esempio utilizzando in matlab la funzione imresize).”

# Descrizione del problema

Per risolvere tale problema di classificazione con apprendimento supervisionato, si vuole implementare un modello di rete neurale artificiale feed forward, multistrato e full connected, che permetta all’utente di specificare gli iperparametri e le funzioni di attivazione ed errore desiderate. La rete sarà sviluppata in maniera quanto più generale possibile, nell’ottica di poter accogliere input di diverse tipologie .L’ambiente di sviluppo scelto per implementare tale modello è Matlab. Si vuole inoltre studiare tramite la tecnica di cross validation K-Fold, il miglioramento o il peggioramento delle performance di riconoscimento della rete neurale al variare degli iperparametri, in particolare learning rate, funzioni di attivazione e numero di nodi interni dello strato hidden.

Utilizzando il dataset per caratteri scritti a mano MNIST, dove le classi target discrete consistono dei caratteri da ‘0’ a ‘9’, verranno svolti test in ottica di classificazione (e non regressione), utilizzando come funzione d’errore la Cross Entropy, combinata con la funzione Soft Max per il post processing.

# Dataset

Il dataset impiegato per l’addestramento della rete è il database di caratteri scritti a mano Mnist. Tale dataset è composto da 60000 esempi di caratteri numerici (da 0 a 9) corredati di label compilate a mano e di 10000 esempi ulteriori per il testing. Il dataset è pubblicamente disponibile, e per leggerne i dati ed acquisire in campioni in formato leggibile per Matlab, ci si è serviti delle funzioni sviluppate all’università di Stanford (allegate al progetto, si veda il file *loadMNISTImages.m*). Tali funzioni, una volta letto l’intero dataset, restituiscono una matrice di dimensioni 784\*60000 (per il training set, 784\*10000 per il test set), in cui le colonne rappresentano i singoli campioni, e le righe rappresentano le feature dei campioni. Le 784 feature rappresentano il totale dei pixel per singola immagine grande 28x28, e descrivono tramite un valore numerico compreso tra 0 (il bianco) e 255 (il nero) il contenuto del pixel (in scala di grigi). Un singolo campione del training/test test può quindi essere visto come un vettore colonna della matrice generata. Tale vettore viene poi diviso valore per valore, per 255, per normalizzare i valori di input da consegnare alla rete.

# Implementazione

La rete neurale è stata quindi implementata in un’ottica di flessibilità e facilità di riutilizzo, specie considerando le necessità dettate dalla Cross Validation di poter eseguire l’addestramento su reti sempre diverse (nel senso di scelta di iperparametri). Per ciò, la rete sviluppata permette all’utilizzatore di scegliere un numero arbitrario di strati hidden tramite l’utilizzo di array contenitori, di specificare le funzioni di attivazione per tali strati e per lo strato di output, ed infine di specificare la funzione d’errore da utilizzare. Sebbene gli script corredati per il testing siano pensati per l’utilizzo della rete con il dataset MNIST, nulla vieta di utilizzare la rete con altri tipi di dataset a patto di fornire alla rete i dati nel formato giusto.

## La rete neurale

Il cuore della rete neurale è osservabile nella funzione *neuralNet.m*, che permette di creare la rete neurale specificando dei parametri desiderati:

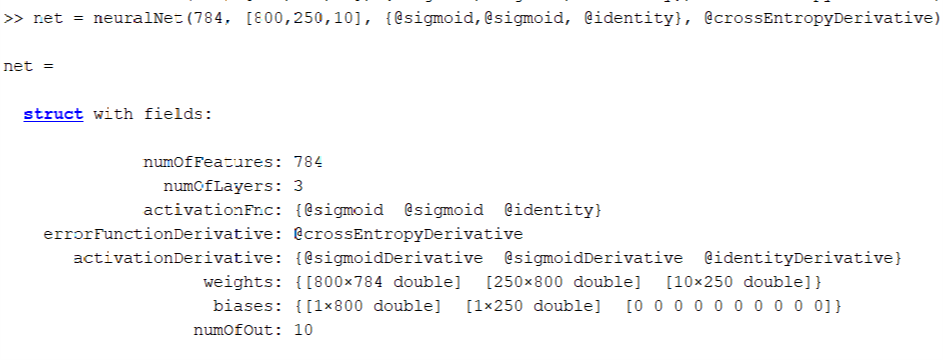


Figura 1 Esempio di creazione rete neurale

In Figura 1 un esempio di creazione di rete neurale con due strati hidden rispettivamente formati da 800 e 250 nodi interni utilizzanti la funzione d’attivazione Sigmoide, e uno strato d’output da 10 nodi con funzione di attivazione Identità (il numero di input e output sono in questo caso quelli necessari al riconoscimento dei dati Mnist). La Cross Entropy è la funzione d’errore utilizzata in questo caso, ma è stata fornita l’implementazione anche per la somma dei quadrati.

Data la necessità di dover parametrizzare il calcolo della derivata (diverso per ogni funzione di attivazione) necessario nella backpropagation, si era pensato in un primo momento di adoperare uno dei toolbox di Matlab per la valutazione delle espressioni e il calcolo delle derivate su queste; tale soluzione si è rivelata però troppo dispendiosa in termini di tempo, in un contesto dove i tempi di esecuzione sono già lunghi di per sé. Pertanto, si è scelto di adottare una soluzione diversa, ossia di fornire in allegato al codice alcune funzioni di attivazione (Sigmoide, Identità, Tangente Iperbolica, ReLU), e le relative implementazioni delle derivate, che l’algoritmo potrà poi chiamare a seconda dei casi. I codici relativi alle funzioni d’attivazione sono stati allegati al progetto (es: *sigmoid.m)*, insieme alle rispettive funzioni per la derivata (es: *sigmoidDerivative.m)*. L’utente dovrà quindi solo specificare le funzioni d’attivazione scelte per ogni strato tramite l’handle matlab (@nomeFunzione), e il sistema provvederà ad accoppiargli le relative funzioni per il calcolo della derivata all’atto della creazione della rete.

Per gestire i pesi della rete, è stato creato un array cell di matrici *weights*, che comprende una matrice di pesi per ogni layer della rete. Ognuna di queste matrici conta una riga per ogni nodo del layer relativo alla matrice, ed una colonna per ogni nodo del layer precedente. Osservando quindi tale matrice riga per riga, si otterrà l’insieme dei pesi/archi entranti nel nodo relativo alla riga, se invece la si osserva per colonna, si otterranno tutti gli archi uscenti del nodo dello strato precedente relativo alla colonna della matrice (ed entranti nei rispettivi nodi dello strato corrente). I bias sono rappresentati tramite pesi fittizi, come matrici con una riga e numero di colonne pari al numero di nodi del layer relativo per facilitare i calcoli.

## L’addestramento

La rete è stata sviluppata in modo da poter essere addestrata tramite l’algoritmo della discesa del gradiente e gli script corredati per l’utilizzo permettono di specificare un learning di tipo Online, MiniBatch e Batch.

Cuore dell’addestramento è visibile nella funzione ’train’ presente nel file *train.m*.



Figura 2 – Train

Come visibile in Figura 2, è possibile partizionare il Training Set in porzioni grandi *batchSize* ed effettuare l’addestramento della rete su tali porzioni. Un singolo training consisterà quindi di una chiamata alla funzione *backPropagation.m,* che dopo aver effettuato forward propagation (*forwardPropagation.m)* e calcolo dei delta, ritornerà le derivate rispetto ai pesi e i bias che verranno poi impiegate dalla funzione *gradientDescent.m* per effettuare l’aggiornamento dei pesi.



Figura 3 – BackPropagation



Figura 4 - Forward Propagation

In Figura 4 è leggibile l’implementare della **Forward Propagation**. Il codice pre-alloca due cell array per un numero di elementi pari a quello degli strati della rete e calcolerà strato per strato i valori di A e Z, rispettivamente l’input e l’output di ogni nodo. Nel dettaglio, le formule implementate risultano essere:

E

Con *l* indice di layer e *g* funzione d’attivazione per lo strato (con z che assumerà valore x per lo strato 1).  
Il calcolo da effettuare in forma matriciale è descritto da:

L’implementazione è quindi stata effettuata tramite moltiplicazioni tra matrici (caso batch/minibatch) o vettori per matrice (caso online) così ottenuti. Per chiarire, ilcell array *a* conterrà in prima posizione, la matrice risultante dal prodotto tra la matrice di BatchSize righe e 784 (l’input) colonne PER la matrice di righe pari al numero di nodi del layer 1 e 784 colonne da trasporre, i cui elementi saranno sommati al bias.

Vi è poi la possibilità di applicare una funzione di post process ai nodi z così ottenuti, in particolare nel nostro caso questa applicherà il softmax per una più chiara classificazione degli esempi rispetto alle label formate da un vettore di dieci componenti, di cui nove impostate al valore 0 ed una al valore 1, in corrispondenza dell’indice corrispondente al valore assegnato.



Figura 5 – SoftMax

Il Softmax viene impiegato quindi come post processing su funzioni di attivazione lineari per normalizzare la somma dei valori (nel nostro caso dei 10 output) portandola a 1, dando così maggiore *peso* in termini di influenza ai valori di grandezza superiore all’interno dei 10 possibili target.

In Figura 3 è possibile osservare l’implementazione della **Back Propagation**. Questa si compone di una chiamata alla forward propagation, del calcolo dei delta dello strato di output e di quelli precedenti a ritroso ed infine del calcolo delle derivate rispetto ai pesi. Il calcolo dei delta dello strato di output è effettuato secondo la formula:

Dove F’ rappresenta la derivata della funzione d’attivazione dello strato i, ai il valore di uscita di un nodo dello strato i-esimo prima della funzione d’attivazione. La derivata dell’errore sull’esempio rispetto all’output Z è stata quindi resa parametrica per permettere di specificare diverse funzioni d’errore. Nel caso d’utilizzo in questione (con Cross Entropy come funzione d’errore su un output a cui è stato applicato softmax), la funzione passata come parametro (*net*.*errorFunctionDerivative*) sarà quindi quella visibile in Figura 5.



Figura 6 - Cross Entropy

Per i restanti delta da calcolare negli strati interni, la formula implementata risulta essere:

Dove g’ rappresenta la derivata della funzione d’attivazione per lo strato i-esimo, applicata sul valore pre-attivazione del nodo i allo strato l, e i termini della sommatoria sono i prodotti dei pesi uscenti dal nodo in questione per i delta associati al nodo dello strato successivo in entrata al peso. I singoli delta vengono calcolati a livello vettore/strato da matlab (*delta{layer*} nel codice) moltiplicando per ognuno la matrice dei pesi dello strato successivo per quella dei delta già calcolati al layer +1. Prima di terminare, la funzione back propagation calcolerà le derivate parziali rispetto ai pesi.



Figura 7 - Calcolo Derivate dei pesi

In Figura 7 è possibile vedere il codice per il calcolo delle derivate rispetto ai pesi. Il calcolo effettuato per ottenere tali derivate è:

Dove Znk assumerà valore Xi (ossia input) nel caso di n=1. Spiegare BIAS



Figura 8 - Discesa del gradiente

In Figura 8 l’implementazione della discesa del gradiente, che aggiorna pesi e bias secondo:

# Test Plan

Per valutare quali iperparametri e quali combinazioni di funzioni di attivazione risultino più performanti in fase di classificazione, si è realizzata una procedura di cross validation su un sottoinsieme ristretto del training set, per testare la capacità di generalizzare le valutazioni da parte della rete. In particolare, si è utilizzata una strategia di K-Folding, visibile nello script *kfold.m*. Tale strategia consiste nel dividere un training set campione in k parti, fissare una delle parti come test set e utilizzare le restanti k-1 come porzione per l’addestramento, ripetendo tale operazione ruotando k volte la k-esima porzione di test. I risultati così ottenuti vengono poi mediati tra le varie rotazioni per fornire un risultato meno sensibile all’overfitting. Il k-fold è stato quindi impiegato per scegliere i migliori iperparametri tra un campione di combinazioni da valutare.

Il training set è stato limitato a 3000 esempi, e per i test effettuati si è scelto *dieci* come valore per il parametro K, formando così 10 raggruppamenti contenenti 300 esempi ciascuno, di cui 9 utilizzati per l’addestramento (in blu in Figura 9), ed uno per il test (in giallo), ruotando a turno su tutte e 10 le posizioni.

Figura 9 – K-fold

Il test plan programmato prevede di far spaziare i seguenti parametri, tra i valori di fianco riportati:

* Numero di **nodi** per lo strato interno:
  + 250
  + 500
  + 800
* Coppia di **funzioni** **d’attivazione** per strato interno ed output:
  + Tangente Iperbolica (tanH) – Unità Lineare Rettificata (ReLU)
  + Sigmoide – ReLU
  + Tangente Iperbolica – Identità
  + Sigmoide - ReLU
* **Eta** (Learning Rate)
  + 0.1
  + 0.05
  + 0.01
  + 0.001

Tali parametri sono stati scelti a seguito di una breve scrematura preventiva, effettuata osservando il funzionamento delle funzioni su determinati parametri nelle fasi iniziali di sviluppo della rete; sono quindi stati scelti i parametri che forniscono un quadro d’insieme più descrittivo e che abbracci una variegata serie di casistiche, escludendo casi limite o poco interessanti (ad esempio learning rate troppo piccoli che risultano sicuramente meno efficaci di vari ordini di grandezza rispetto a tutti gli altri)

Le 9 fasi di addestramento e la singola di test che contraddistinguono un k-folding sono quindi eseguite per un totale di 3x4x4 volte, incrociando le possibili combinazioni.

Una volta ottenuta la tripla di parametri migliore, osservando come misura d’efficacia la deviazione standard dell’errore ottenuto dalla rotazione, si è proceduto ad effettuare dei test di valutazione finali in modalità online e minibatch per confronto.

# K-Folding

## Dati Risultato

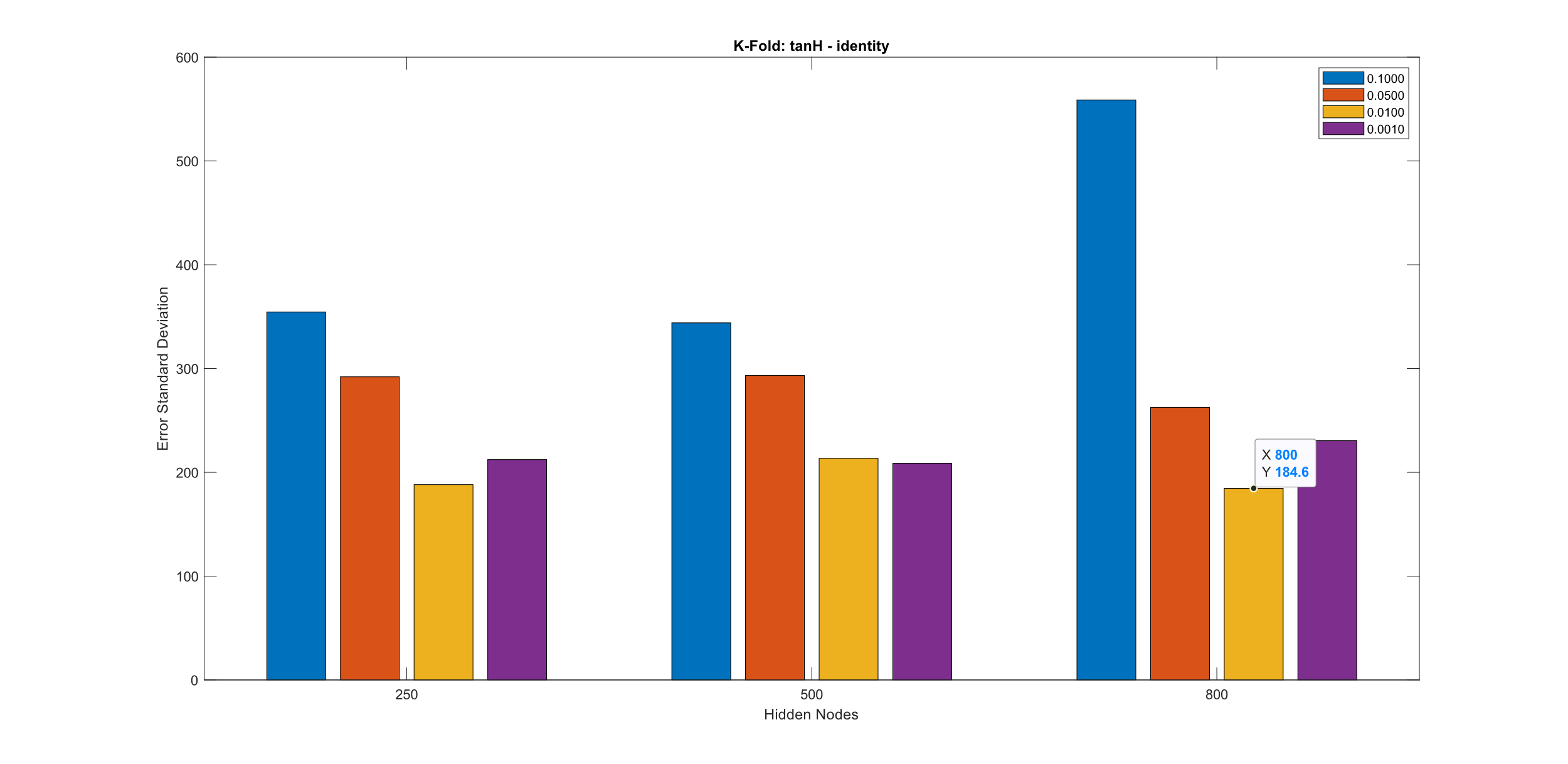
Di seguito vengono mostrati i grafici che descrivono i dati estrapolati dalla cross validation per la ricerca degli iperparametri migliori. I colori delle barre rappresentano il learning rate utilizzato mentre i raggruppamenti sull’asse delle ascisse indicano il numero di hidden nodes utilizzato. Il valore numerico ottenuto rappresenta la deviazione standard dell’errore dato cross entropy. I grafici corrispondono alle 4 coppie di funzioni di output utilizzate come campione d’esame. Per ogni grafico è stata segnato con un’etichetta il valore d’errore minimo rispetto a tutte le combinazioni.

Figura 10 - KFold: tanH-Identity

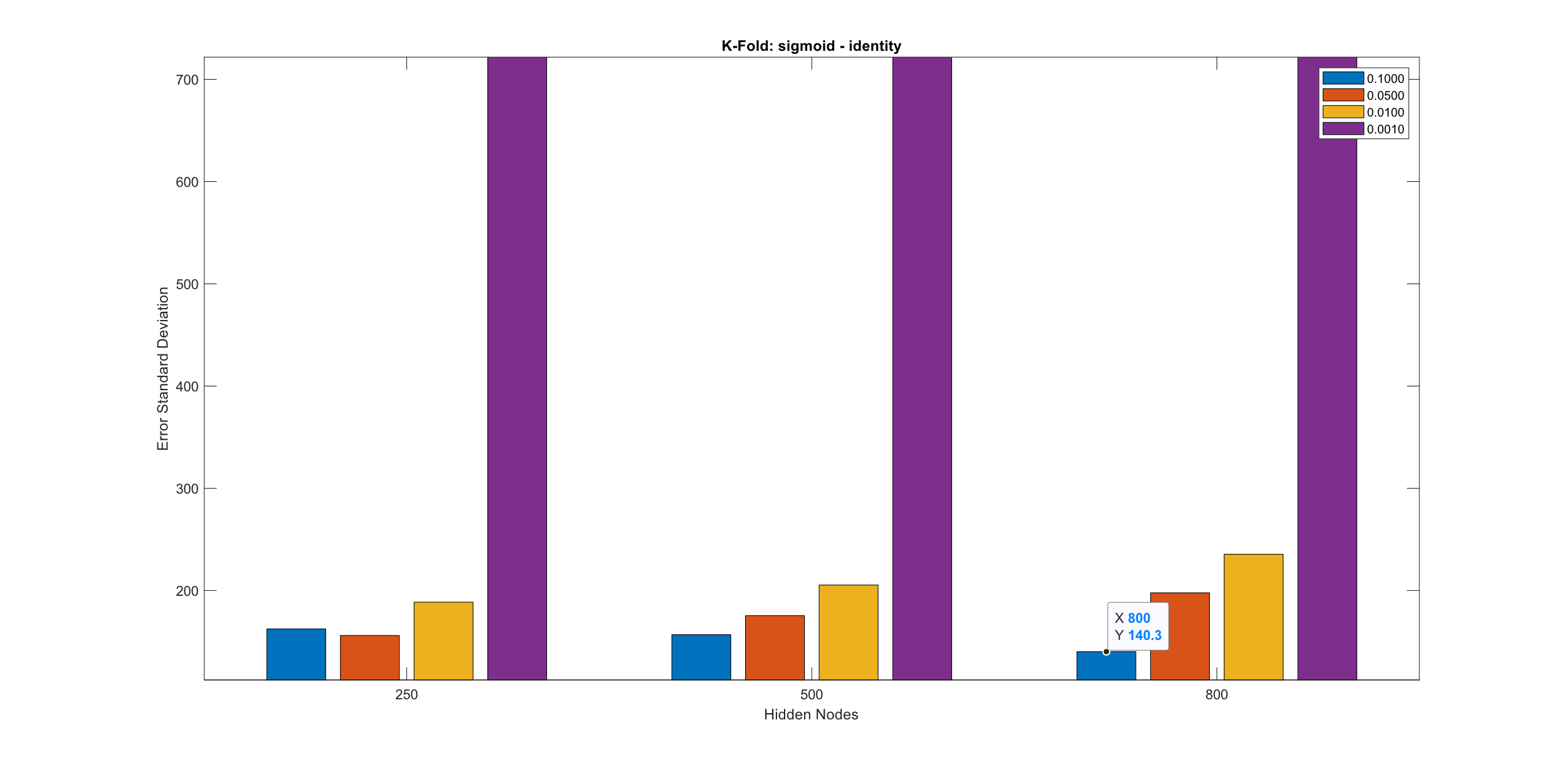
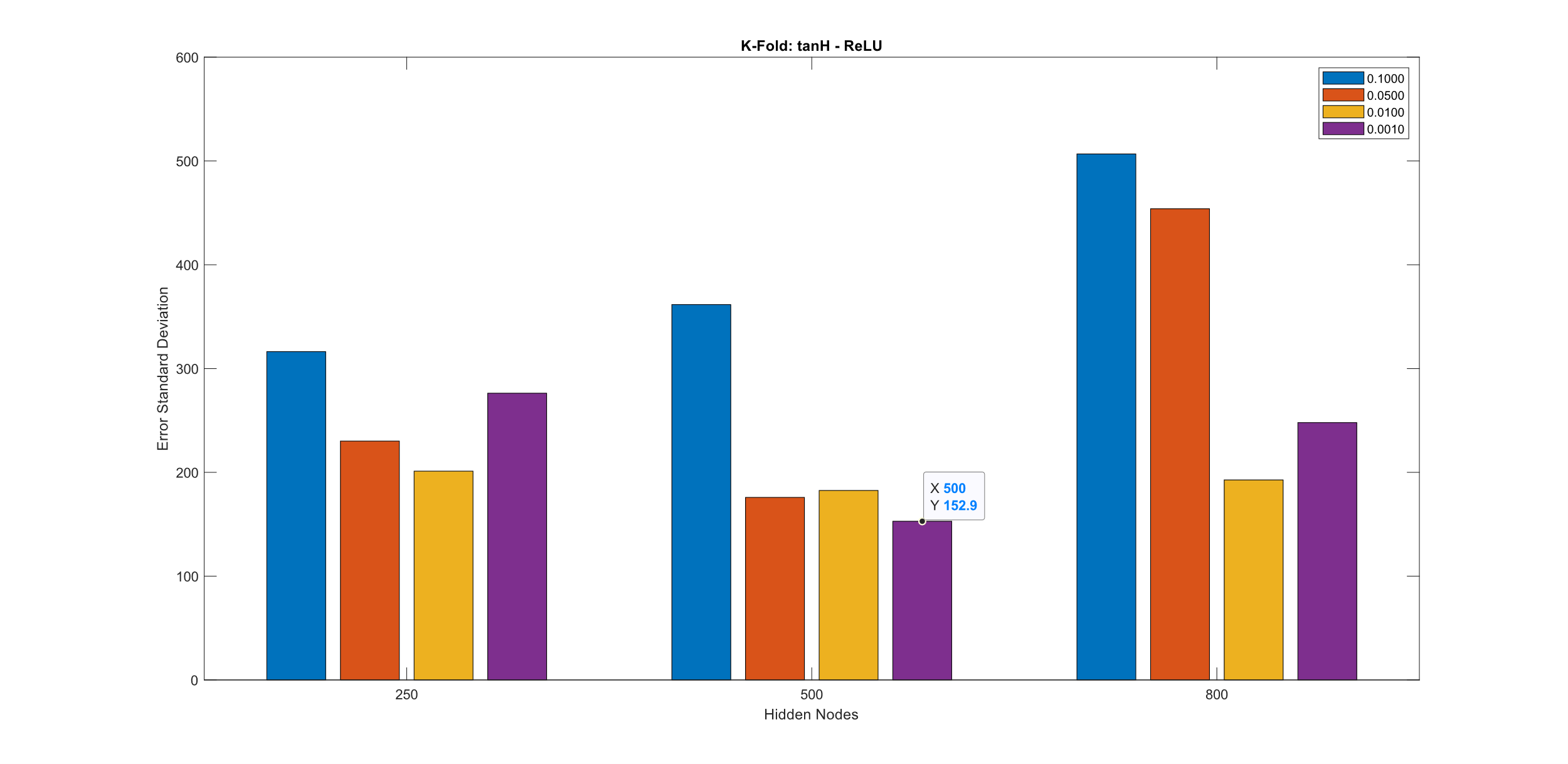


Figura 11- KFold: tanH-ReLU

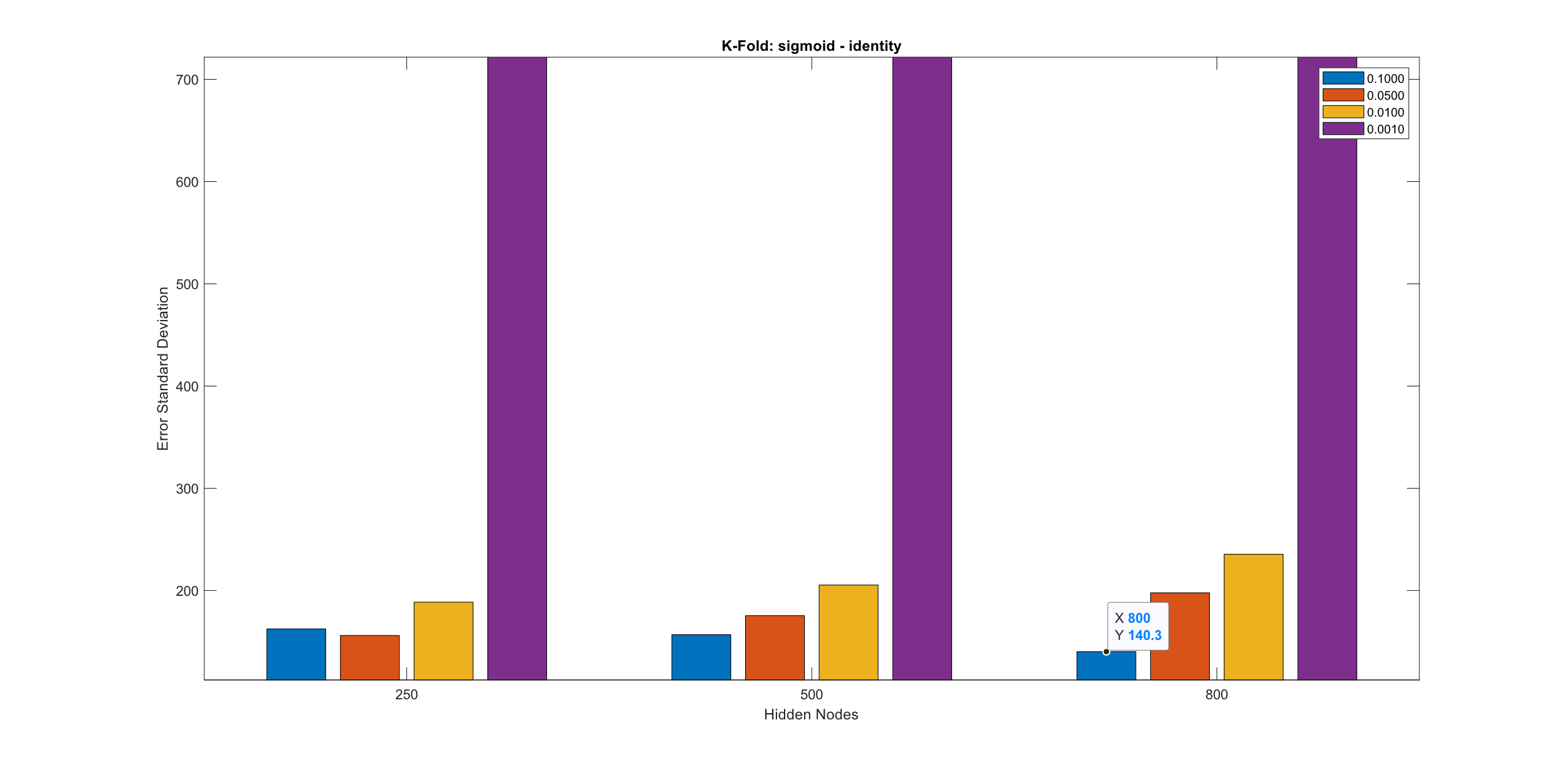


Figura 12 - KFold: sigmoid-Identity

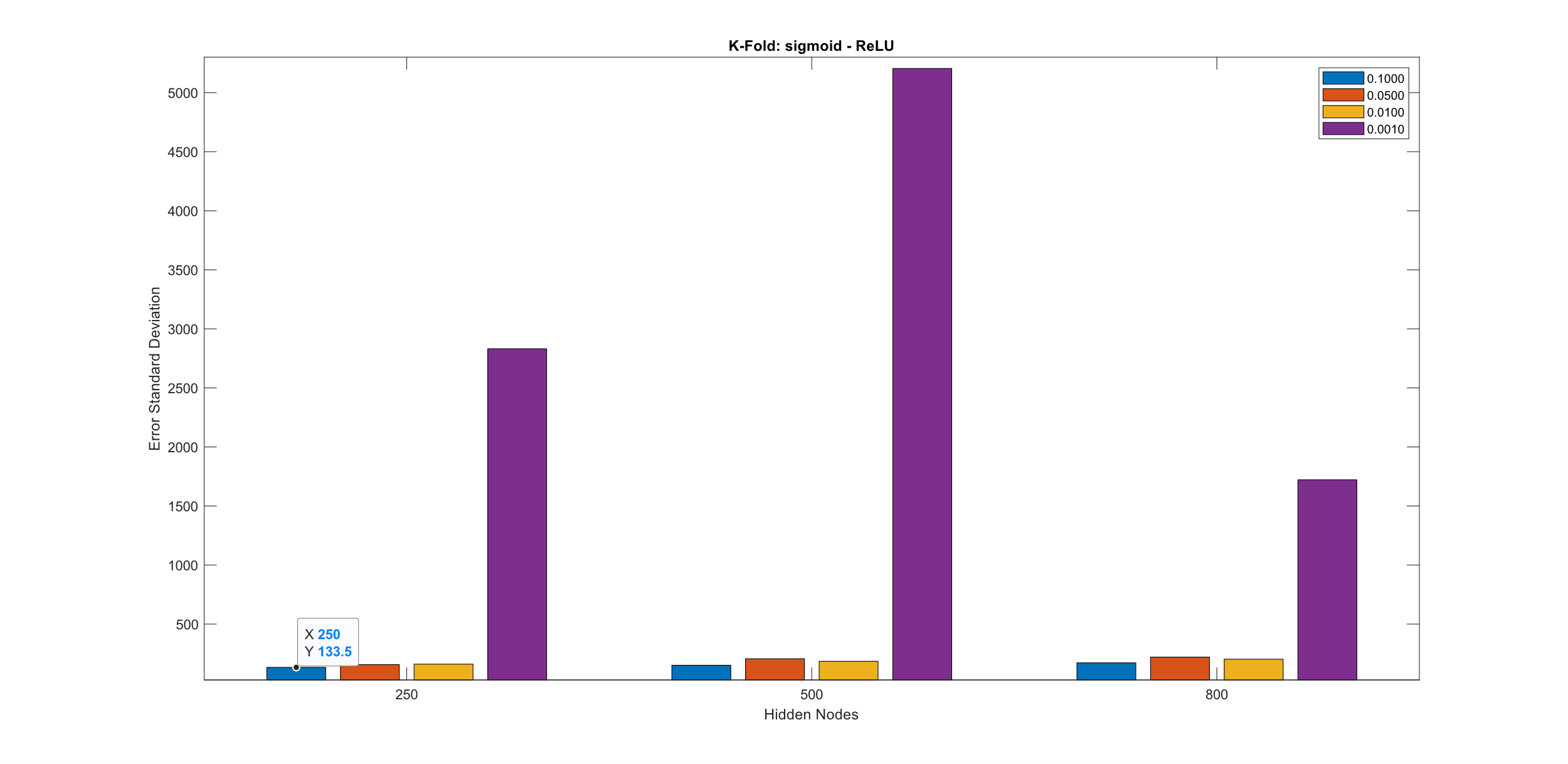


Figura 13 - KFold: sigmoid-ReLU

Il tempo d’esecuzione per l’intero k-folding è ammontato a circa 3 ore e 10 minuti.

## Osservazioni sul K-Folding

Tra le considerazioni che è possibile effettuare osservando i grafici figura la differenza nei risultati d’apprendimento quando si usa la sigmoide rispetto alle altre funzioni. Infatti leggendo i risultati, è possibile vedere come l’utilizzo di tale funzione richieda un learning rate alto per funzionare al meglio, e come visibile nei due casi d’uso (Figura 12,Figura 13), l’uso combinato di sigmoide e l’ultimo learning rate provato (0.001) si traduce in un valore d’errore molto più elevato e fuori scala (e quindi una peggiore classificazione) rispetto a tutte le altre combinazioni. Per quanto riguarda il numero di nodi interni, si evince come l’aumento del numero di nodi dello strato hidden non si traduca in un automatico miglioramento delle performance. Preventivi test effettuati sulla rete hanno inoltre evidenziato nel migliore dei casi, una performance di classificazione scarsa, e nel peggiore il problema dell’esplosione dei gradienti, quando vengono utilizzati learning rate alti e numero di nodi interni elevati. Ciò è visibile nei grafici relativi ai test in cui è stata utilizzata la tangente iperbolica: i valori peggiori sono infatti osservabili in prossimità dell’eta più grande (0.1) e il numero di nodi interni più elevato (800).

Come si evince dall’ultimo grafico (Figura 13), il valore più basso in termini di deviazione standard dell’errore della cross entropy è riscontrato nel test composto da:

* Funzioni: Sigmoide - ReLU:
* Eta: 0.1
* Numero di nodi interni: 250.

Tuttavia, al migliore risultato in termini di errore minimo, non corrisponde un miglior risultato in termini di accuracy. Ciò ci ha quindi spinto a valutare i risultati, mettendo a confronto le migliori combinazioni ottenute in termini di errore minimo e di accuracy massima.



Figura 14 - Kfold: top 9 Accuracy,Loss



Figura 15 - Kfold: top 9 Loss,Accuracy

In Figura 14 sono mostrati i migliori 9 risultati ottenuti dal k-folding, ordinati per migliore accuratezza in ordine decrescente. In Figura 15 invece sono mostrati i migliori 9 questa volta ordinati per deviazione standard del costo della cross entropy in ordine cresente. Analizzando i valori è stato quindi possibile estrapolare la migliore tripletta di iperparametri in termini di errore minimo, di accuratezza massima, ed infine una tripletta di miglior valore medio tra accuratezza ed errore rispetto ai migliori 10 risultati di ciascun parametro.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Hidden layer A.F | Output Layer A.F. | Eta | Hidden nodes | Mean Accuracy | C.E. Standard deviation |
| Best Average Score | sigmoid | identity | 0,05 | 250 | 87,77 | 156,04 |
| Highest Accuracy | tanH | ReLU | 0,01 | 500 | 88,37 | 182,55 |
| Lowest Error | sigmoid | ReLU | 0,1 | 250 | 87,6 | 133,53 |

Figura 16 - Migliori valori k-folding

# Risultati finali

Ottenute le 3 combinazioni di parametri migliori rispetto ai criteri elencati nel paragrafo precedente, si è proceduto ad effettuare dei test finali, utilizzando training set più grandi, con molteplici epoche e confrontando le modalità di apprendimento Online e Minibatch per ognuna di queste. Nella modalità di apprendimento Online, l’aggiornamento dei pesi segue dopo ogni addestramento su singolo esempio, realizzando una rete in grado di convergere ad un minimo in minor numero di epoche impiegando però un tempo d’esecuzione maggiore. La modalità minibatch invece effettua l’addestramento contemporaneamente su porzioni di esempi di grandezza fissata (ad esempio 32 esempi per addestramento). La modalità batch (qui non trattata) può essere vista come un caso particolare della minibatch in cui l’aggiornamento dei pesi viene effettuato una singola volta per epoca sull’intero training set. Per i nostri test è stato scelto un batchSize (grandezza delle partizioni per il minibatch learning) di 32 ed un training set grande complessivamente 3200 esempi. Dopo ogni epoca di apprendimento la rete è stata valutata sull’intero test set di 10000 esempi.

## Combinazione: Errore Minore

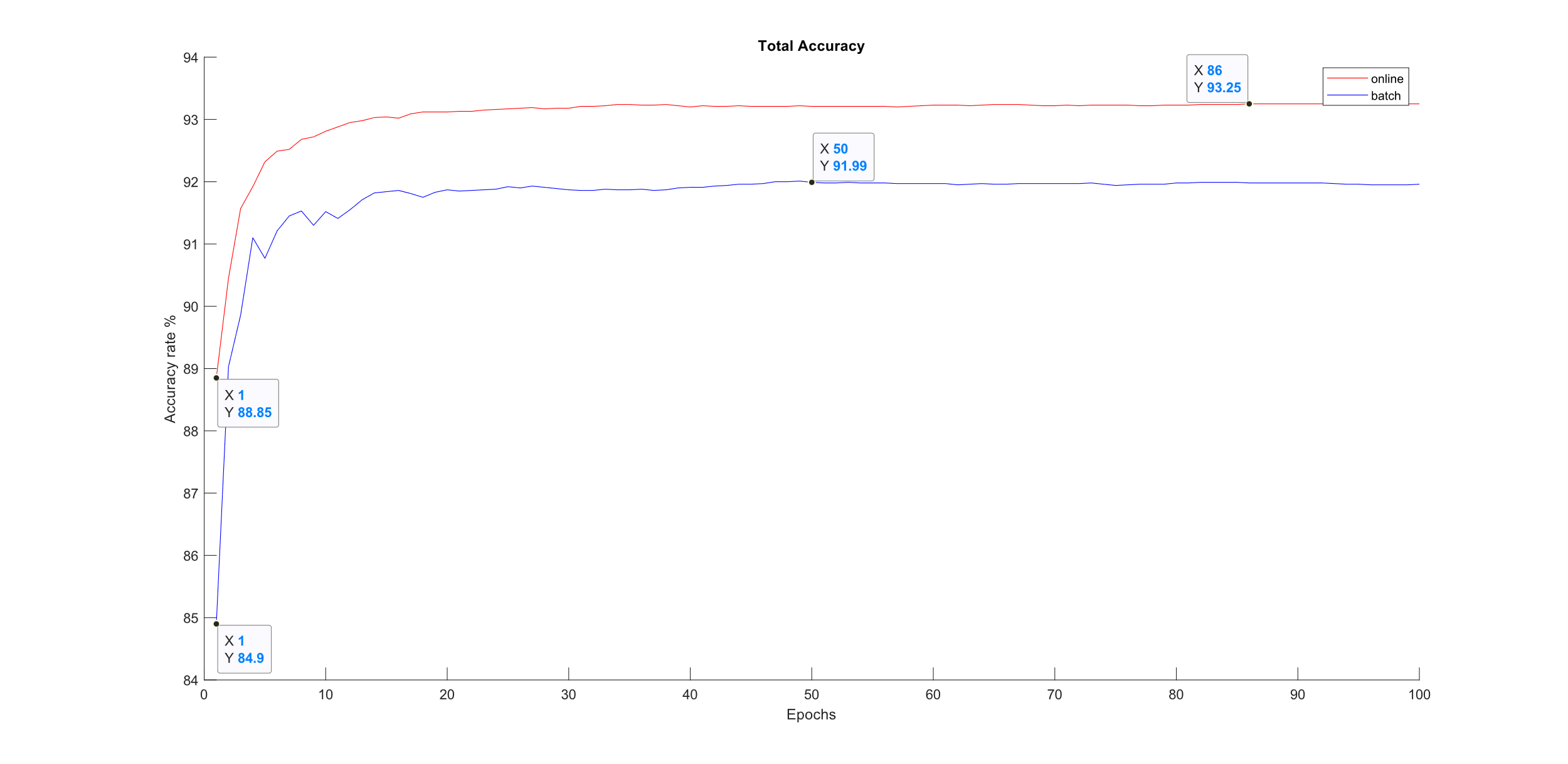


Figura 17 - Combinazione: Errore Minore (sigmoid- ReLU,0.1,250), Misurazione Accuratezza

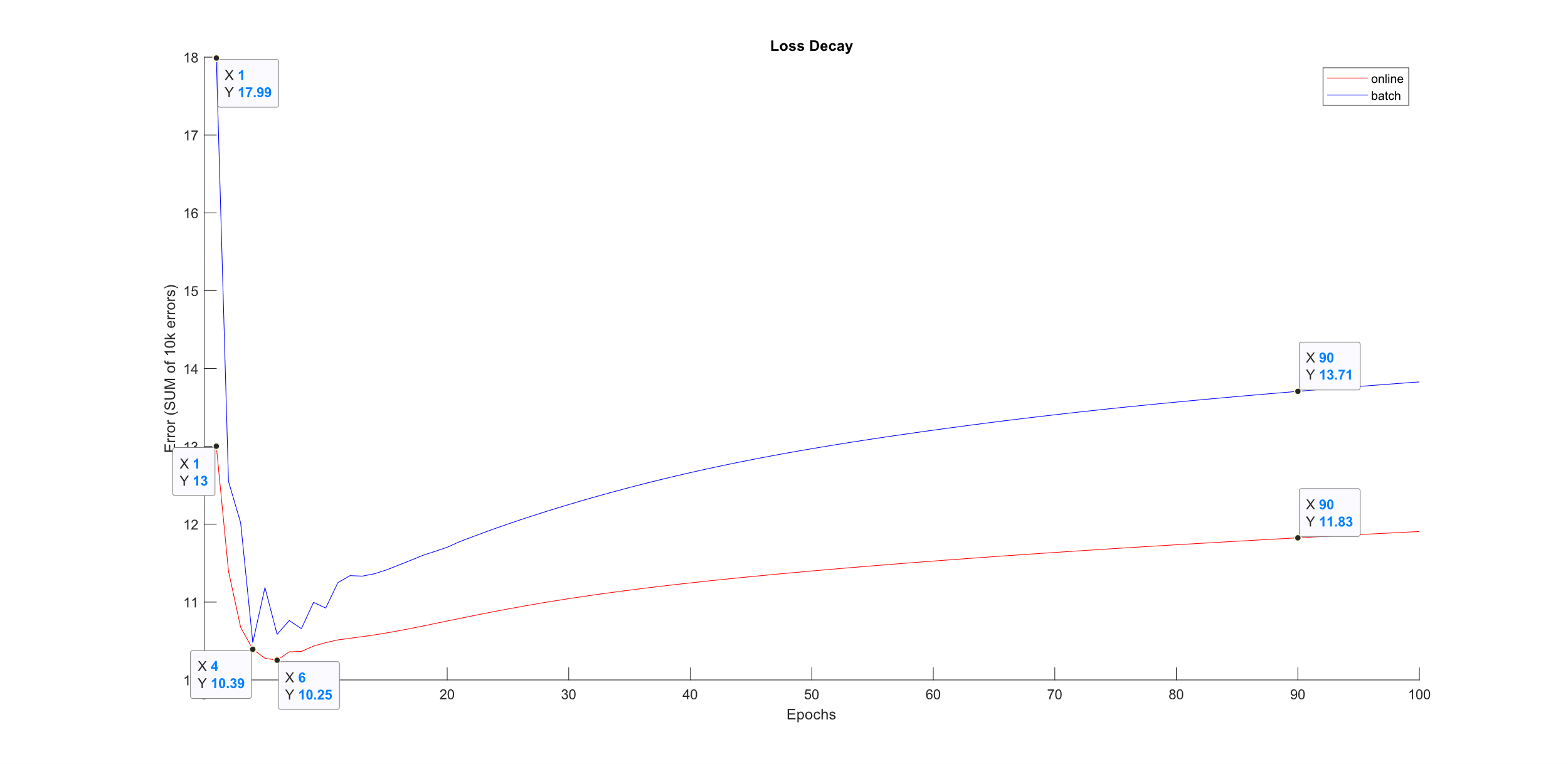


Figura 18 - Combinazione: Errore Minore (sigmoid- ReLU,0.1,250), Misurazione Errore Totale

COMMENTO AI GRAFICI

## Combinazione: Accuratezza Migliore

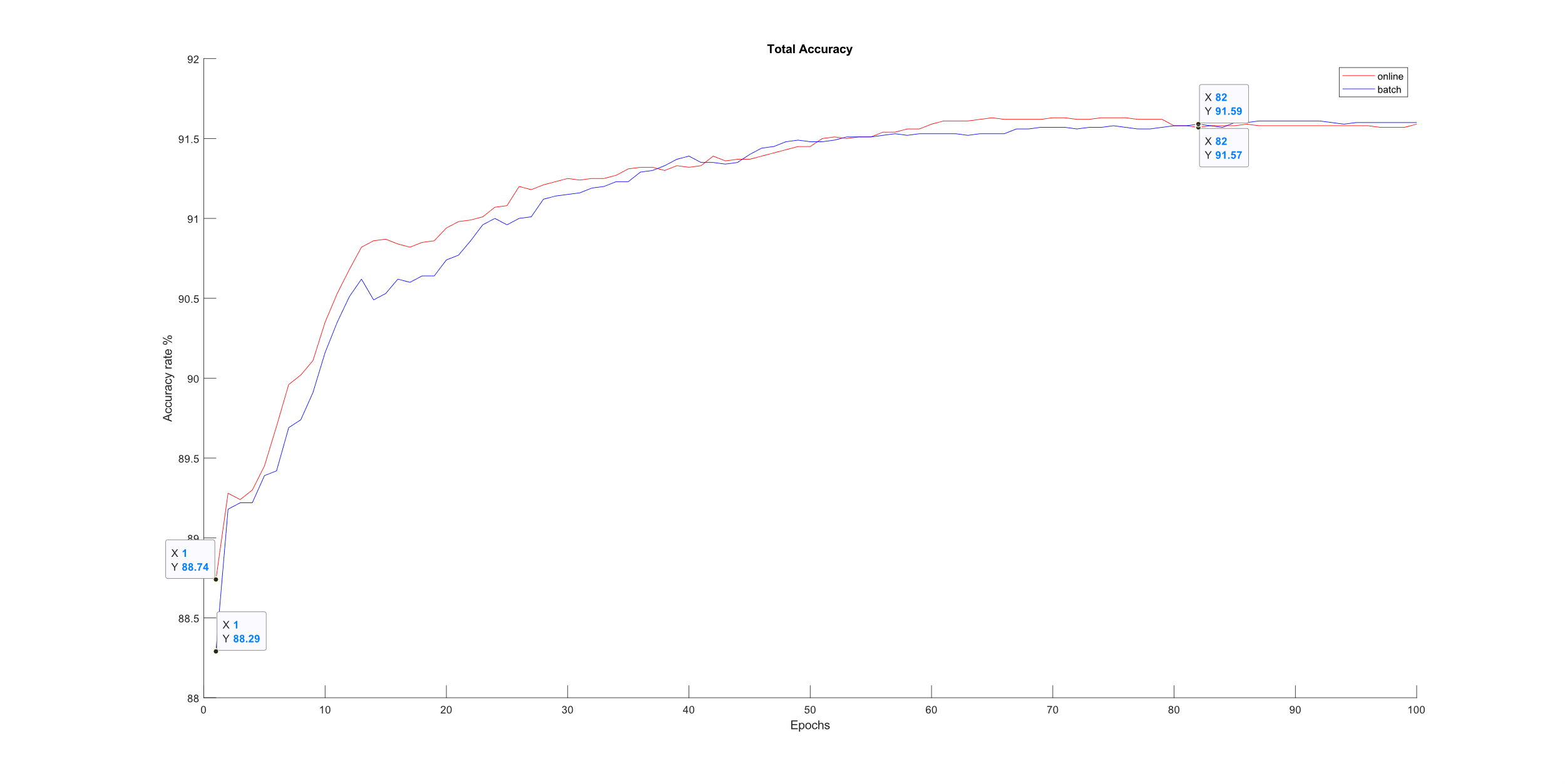


Figura 19 - Combinazione: Accuratezza Migliore (tanH,ReLU,0.01,500), Misurazione Accuratezza

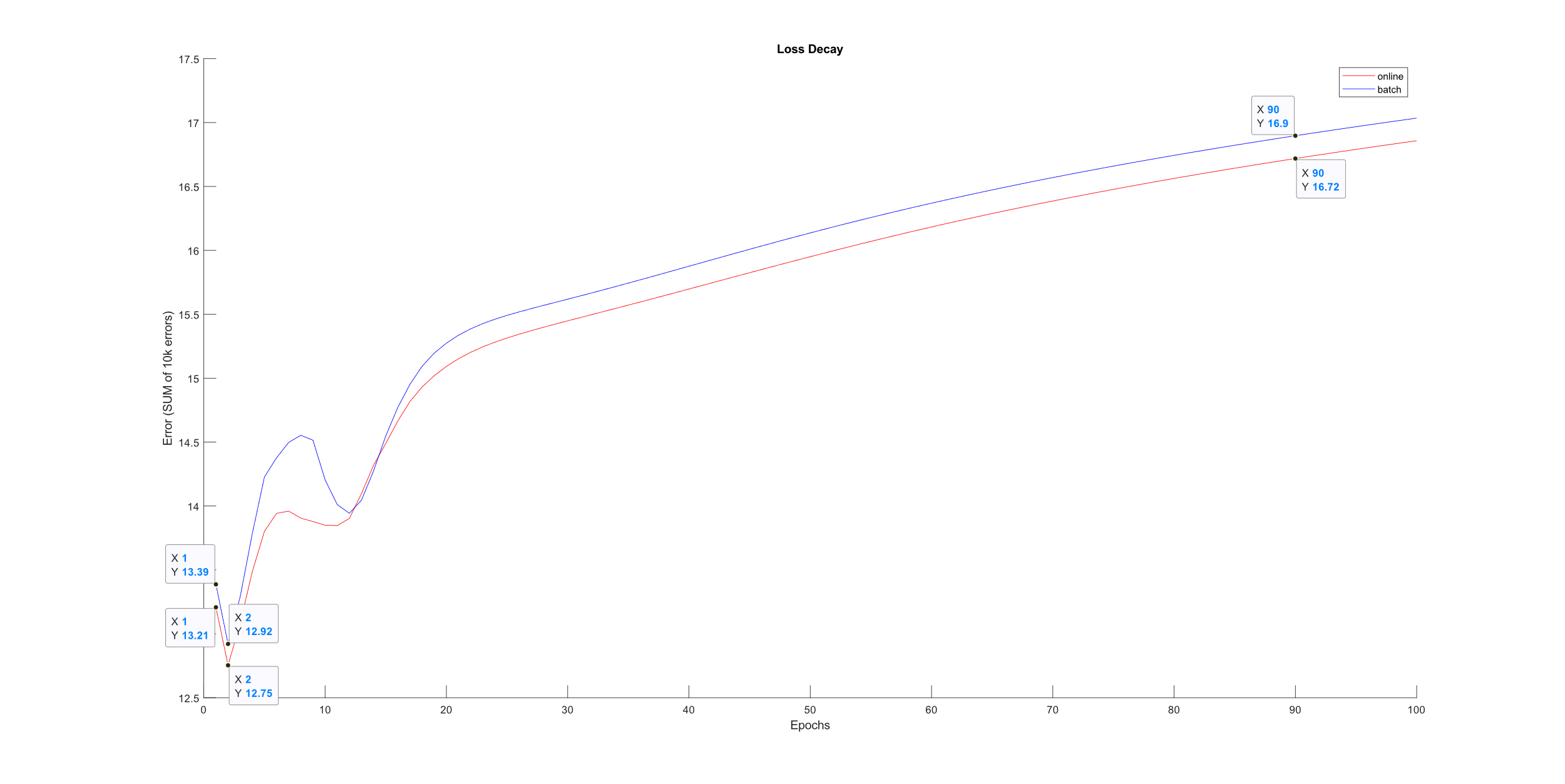


Figura 20 - Combinazione: Accuratezza Migliore (tanH,ReLU,0.01,500), Misurazione Errore Totale

COMMENTO AI GRAFICI

## Combinazione: Migliori valori in media

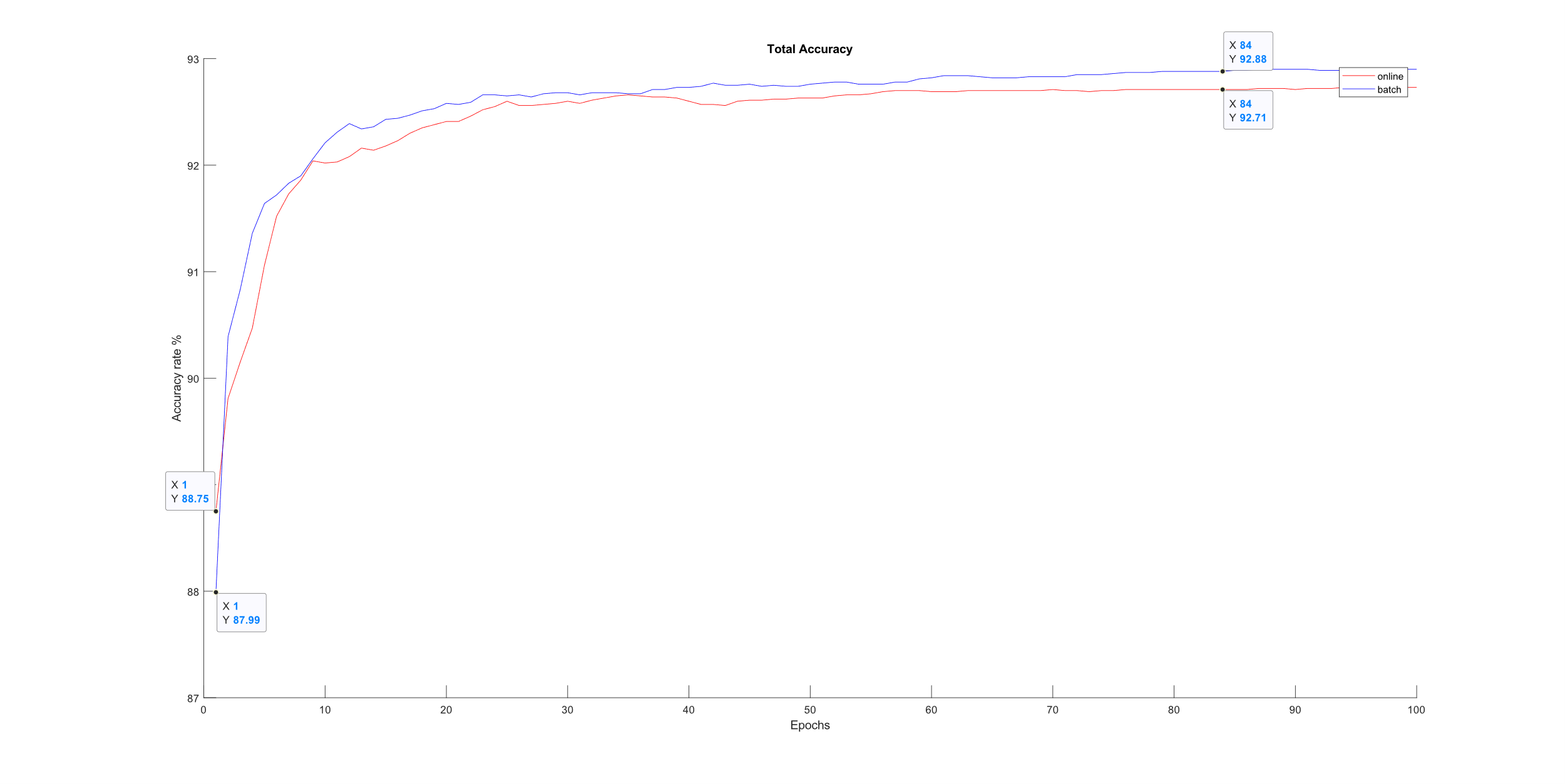


Figura 21 - Combinazione: Migliori valori in media (sigmoid, identity,0.05,250), Misurazione Accuratezza

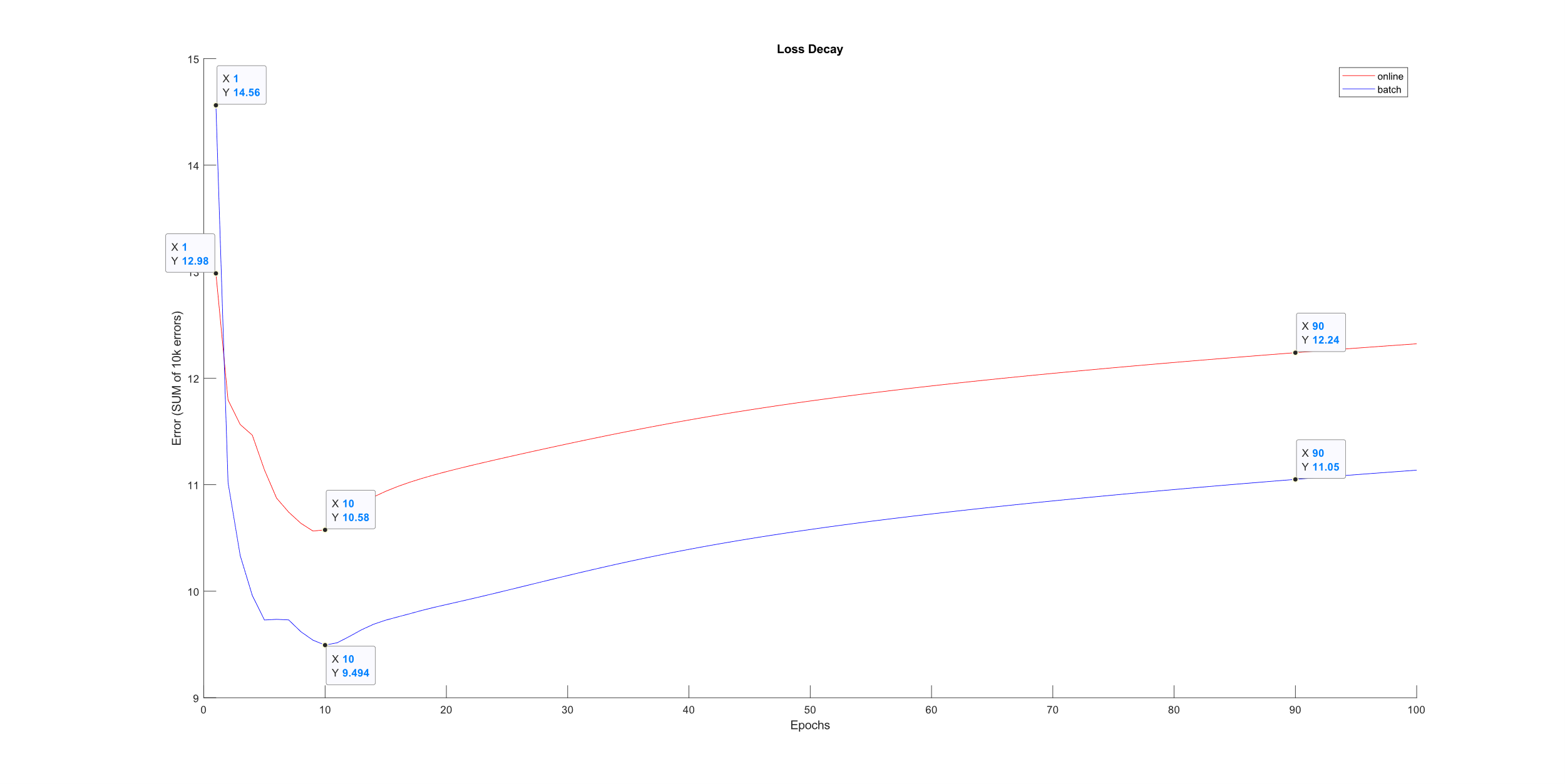


Figura 22 - Combinazione: Migliori valori in media (sigmoid, identity,0.05,250), Misurazione Errore Totale

COMMENTO AI GRAFICI

# Conclusioni

COMMENTO FINALE + eventuale test con best media ma eta ridotto