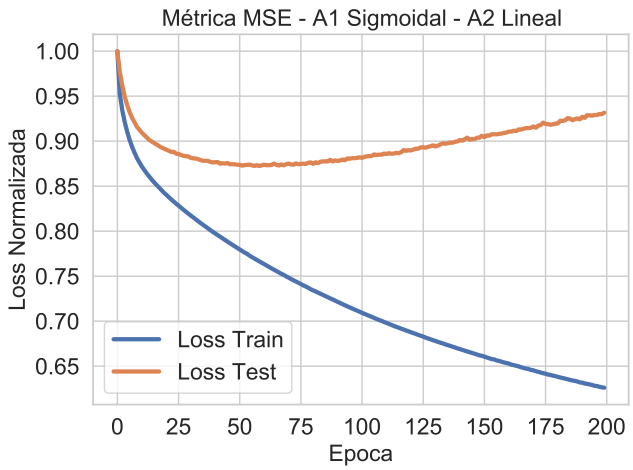
TP2

Tomás Hüttebräucker
Aprendizaje Profundo y Redes neuronales artificiales
(1 de octubre de 2020)

EJ3

Usando una red con una capa oculta de 100 neuronas, se intento resolver la clasificación de datos de CIFAR10. Se eligió como función de costo al error cuadrático medio, como función de activación de la capa oculta a la función sigmoide y como función de activación de la capa de salida a la función lineal. Los datos fueron preprocesados, restandoles la media de los datos de entrenamiento tanto a ellos mismos como a los datos de prueba. El entrenamiento se realizó con el método SGD, y con los siguientes parámetros:

Parámetro	**Valor**
λ	10^{-4}
Learning rate	10^{-4}
Batch size	50
Inicialización de pesos	10^{-3}



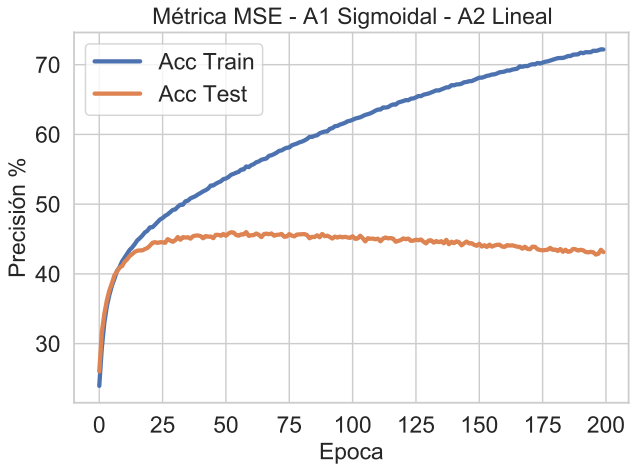


Figura 1: Loss normalizada y precisión del modelo del ejercicio 3 sobre la base de datos CIFAR10 en función de la epoca de entrenamiento.

Como se observa en la figura 1, la función usada con los datos de prueba disminuye primero pero vuelve a aumentar nuevamente, mientras que la función loss con los datos de entrenamiento solo disminuye. Esto es una muestra clara de que hay *overfitting*, por lo que el parámetro λ debería aumentarse.

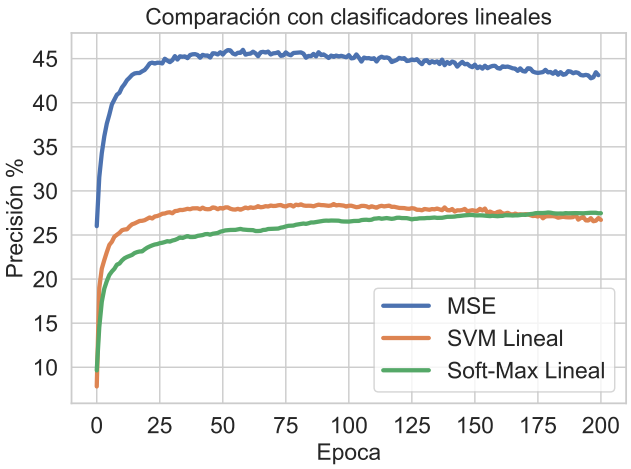


Figura 2: Comparación de la precisión sobre los datos de prueba del modelo planteado con métrica MSE en el ejercicio 3 y los clasificadores lineales implementados en el practico anterior.

A pesar del claro *overfitting* del modelo, este funciona mucho mejor que los clasificadores lineales implementados en el practico anterior. Esto es esperable ya que el nuevo modelo cuenta con mas parámetros.

EJ4

Utilizando el mismo modelo que en el ejercico 3, se cambió la función de costo por la *Categorical Cross Entropy* o CCE. El resto de los parámetros y el pre-procesado de los datos se mantuvieron iguales, con el objetivo de poder comparar la influencia de la función de costo.

Como se observa en la figura 3, el modelo vuelve a mostrar *overfitting*, pero de menor magnitud. También se nota que el modelo tarda mas en converger, con un decrecimiento de la loss y un crecimiento de la precisión menor que al utilizar la métrica MSE. Esta convergencia mas lenta puede ser la que explique por que el modelo presenta menor *overfitting* también. En ambos modelos al principio el *overfitting* no es apreciable, por eso, si se deja entrenar por mas épocas con la métrica CCE, es posible que se llegue a apreciar este fenómeno en mayor magnitud.

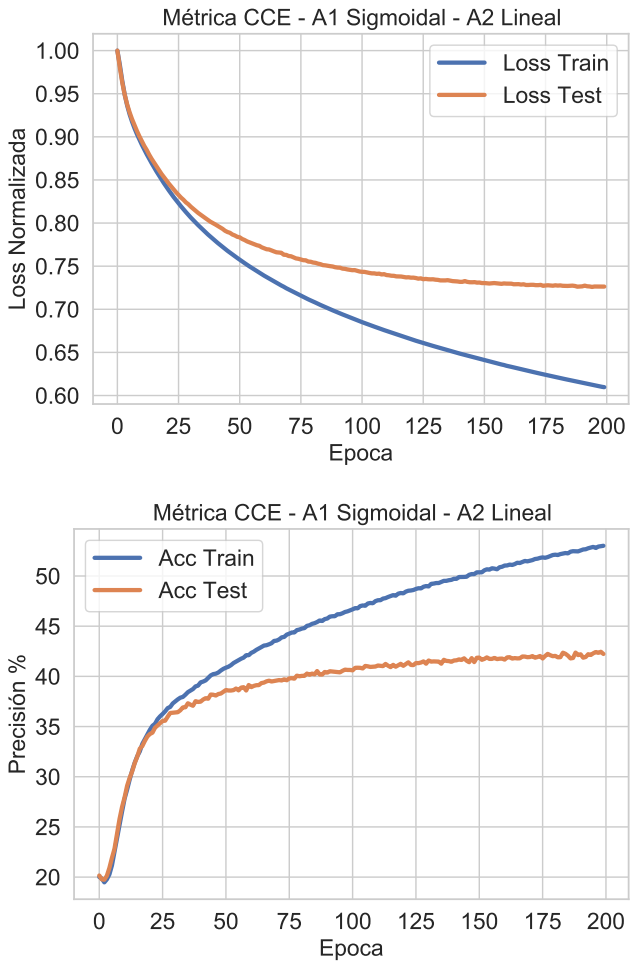


Figura 3: Loss normalizada y precisión del modelo implementado en el ejercicio 4 sobre la base de datos CIFAR10 en función de la época de entrenamiento.

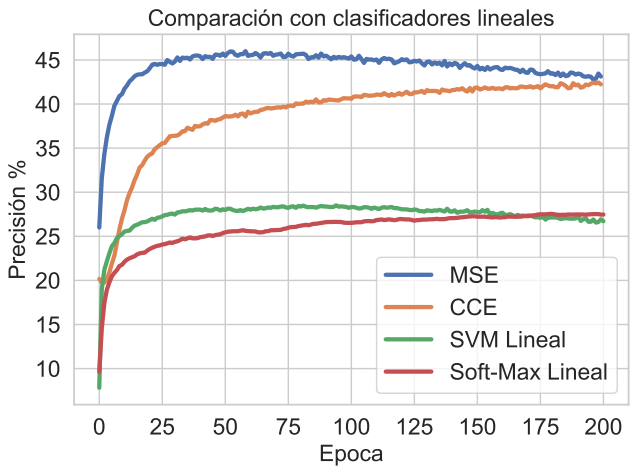


Figura 4: Comparación de la precisión sobre los datos de prueba del modelo planteado en el ejercicio 4 (CCE) y los clasificadores lineales implementados en el practico anterior (SVM-SoftMax) y en el ejercicio 3 (MSE.

Al comparar el modelo con los planteados en el práctico 1 y en el ejercicio 3 se pone en evidencia la convergencia mas lenta de la métrica CCE. Sin embargo, debido al *overfitting* presente en el modelo con métrica MSE, las precisiones de ambos terminan siendo similares. Para hacer una comparación mas interesante debería evitarse el *overfitting* y ver como los dos modelos evolucionan para más épocas de entrenamiento. Ambos modelos son sustancialmente mejores que los clasificadores lineales.

EJ5

Con el objetivo de observar la influencia de las funciones de activación de las diferentes capas, se cambiaron las mismas en los modelos implementados en el ejercicio 3 y 4. Manteniendo el resto de los parámetros, asi como el pre-procesado de los datos para poder comparar apropiadamente. Primero se probó cambiando la función de activación de la primer capa por una ReLU y la de la segunda por una sigmoide.

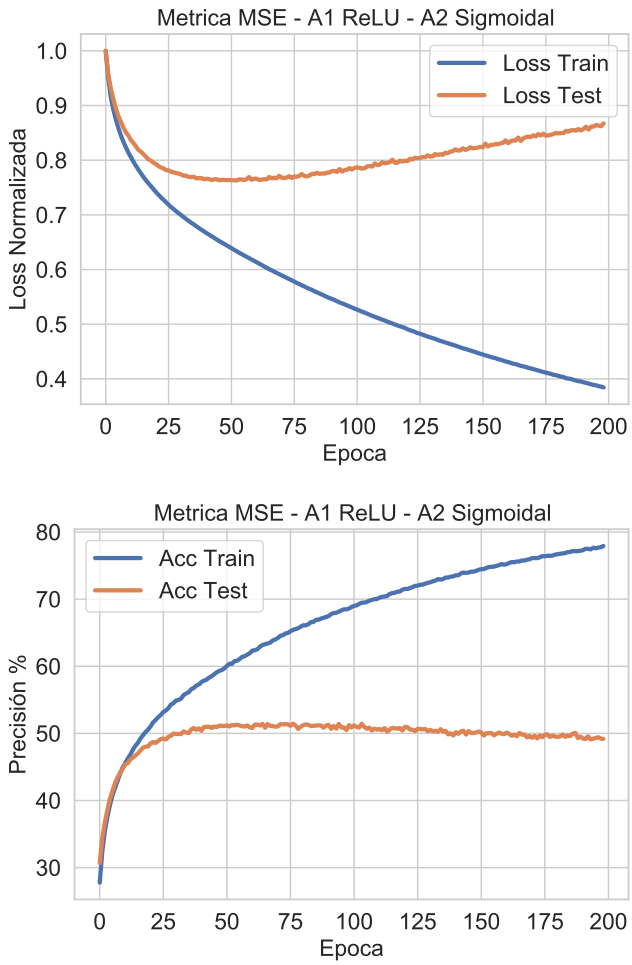


Figura 5: Loss normalizada y precisión del modelo sobre la base de datos CIFAR10 en función de la época de entrenamiento.

En la figura 5 se muestra la evolución de la función de costo y de la precisión del modelo sobre los datos de entrenamiento y evaluación a través de las épocas al usar métrica MSE con funciones de activación ReLU en la primer capa y sigmoidal en la segunda. Se observa nuevamente que el modelo presenta *overfitting*, similar al obtenido en el ejercicio 3, esto podría deberse a que tienen

la misma función de costo. A pesar de esto, en este caso llegó a un nivel de precisión mas alto tanto para los datos de entrenamiento como los de evaluación.

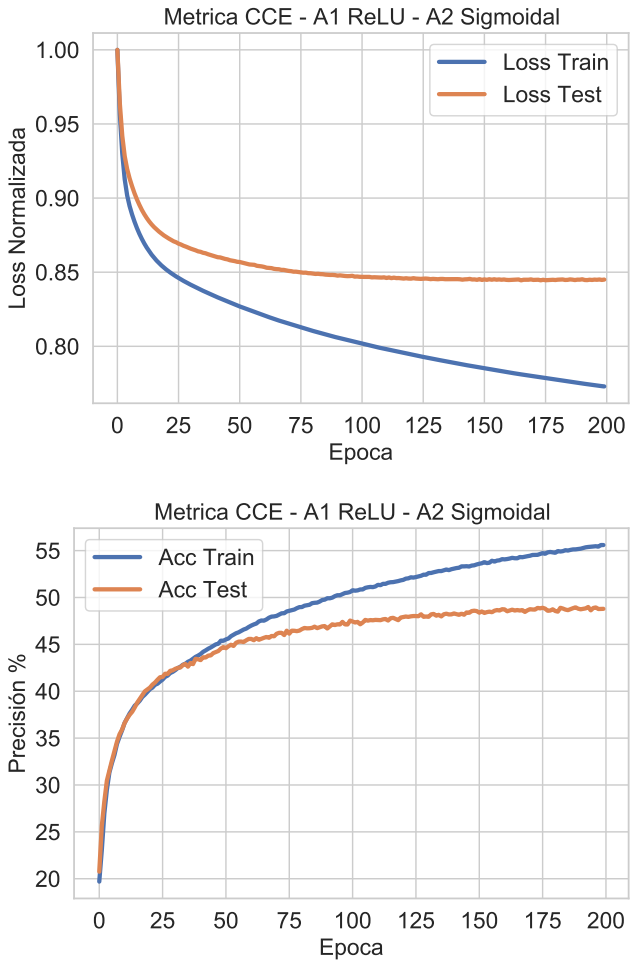


Figura 6: Loss normalizada y precisión del modelo sobre la base de datos CIFAR10 en función de la época de entrenamiento.

En la figura 6 se muestra la evolución de la función de costo y de la precisión del modelo sobre los datos de entrenamiento y evaluación a través de las épocas al usar métrica CCE con funciones de activación ReLU en la primer capa y sigmoidal en la segunda. En este caso el *overfitting* no es tan grave como al usar MSE, lo que de nuevo se condice con el hecho de que el crecimiento de la precisión es mas lento. Nuevamente, las activaciones utilizadas mejoran el resultado respecto al ejercicio 4.

Al usar MSE con activación lineal en la capa de salida, los resultados obtenidos se muestran en la figura 7. En este caso el *overfitting* es realmente problemático, la loss de los datos de evaluación al final del entrenamiento es mayor que al principio. Aunque la precisión en algun momento disminuye, no llega a ser tan baja como al principio, esto puede deberse a que la métrica MSE penaliza valores de clasificación correcta superiores a 1, mientras esto no se ve penalizado en la precisión. El modelo parece converger mas rápido comparado con una activación sigmoidal en la capa de salida, aunque no es una diferencia muy grande y termina con un valor de precisión para los datos de evaluación menor, debido al *overfitting*.

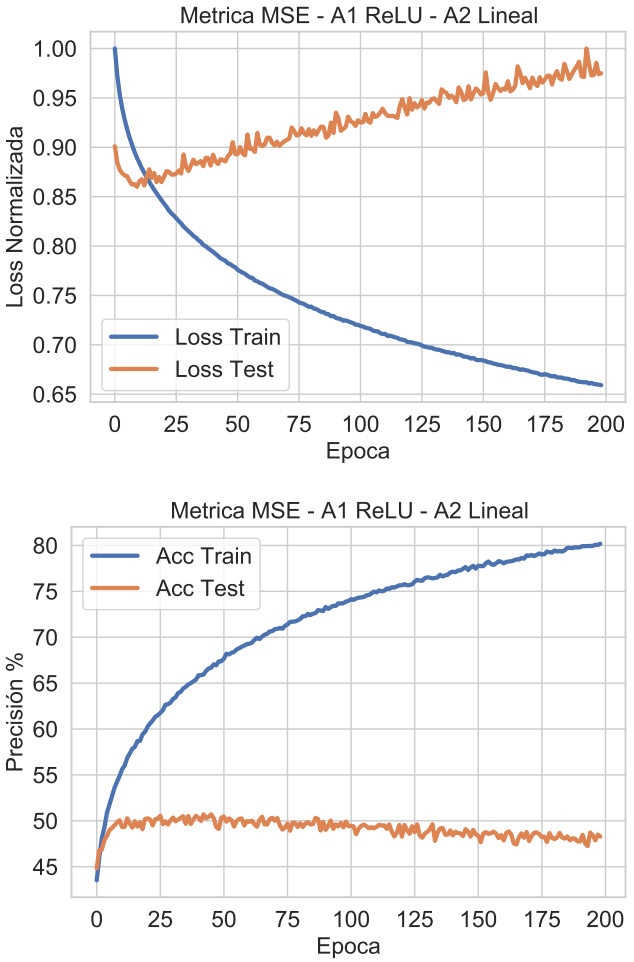


Figura 7: Loss normalizada y precisión del modelo sobre la base de datos CIFAR10 en función de la época de entrenamiento.

En la figura 8 se muestran los resultados al usar métrica CCE con activación ReLU en la primer capa y activación lineal en la capa de salida. Se ve nuevamente que el *overfitting* es muy problemático. Es la primera vez en que es tan notable al usar la métrica CCE. Esto puede deber a que, al utilizar activación lineal en la capa de salida, la convergencia del modelo es más rápida y da lugar a que se magnifique el *overfitting*. También se ve que la precisión de los datos de train llega a un nivel muy alto, el mayor hasta ahora. Esto hace notar que, al usar la métrica CCE, la función sigmoidal es un freno muy importante al avance de la misma, lo que tiene sentido dado que las salidas no tienen que ser acotadas en este caso, y la sigmoidal restringe los posibles valores de las salidas.

Se puede concluir al usar métrica CCE, la función sigmoidal es un freno al aprendizaje de la red, haciendo que avance mas lento debido a su saturación. En el caso de usar la métrica MSE, la red no se ve tan afectada por las activaciones del tipo sigmoide, se comporta de forma bastante similar que al usar activaciones tipo lineales, aunque si es observable una mejora. Para observar bien todos los fenómenos y poder sacar conclusiones mas seguras, sería necesario evitar el sobreajuste de la red, pero por falta de tiempo no se pudo hacer, a pesar de esto, se pueden apreciar a grandes rasgos como afectan las diferentes caracteristicas modificadas.

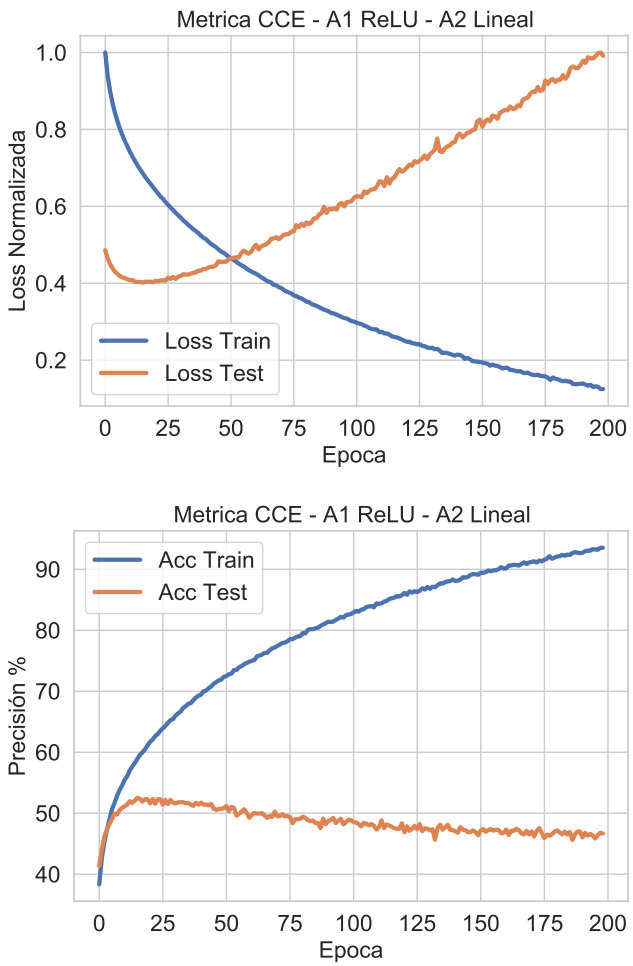


Figura 8: Loss normalizada y precisión del modelo sobre la base de datos CIFAR10 en función de la época de entrenamiento.

EJ6

Se plantearon dos arquitecturas diferentes para resolver el problema del XOR de dos entradas. Las arquitecturas de los modelos usados fueron:

- Modelo 1: Capa oculta de dos neuronas y capa de salida de una neurona
- Modelo 2: Capa oculta de una neurona, capa de concatenación de la entrada + salida de capa oculta a la capa de salida de una neurona

En ambos modelos las funciones de activación de las capas fueron **tanh**. Ambos modelos se entrenaron con todos los datos (todos los números de 2 bits). Los parámetros elegidos para el entrenamiento fueron:

Parametro	**Valor**
Learning rate	10^{-1}
Pesos iniciales	1

En la figura 9 se muestra la evolución de ambos modelos a través de las épocas. Mirando esta imagen se podría decir que el modelo 1 es mejor para el problema, lo que tiene sentido ya que posee mas parámetros para ajustar. Sin embargo no es suficiente evidencia al respecto ya que los modelos son muy dependientes de la inicialización de los pesos de las capas.

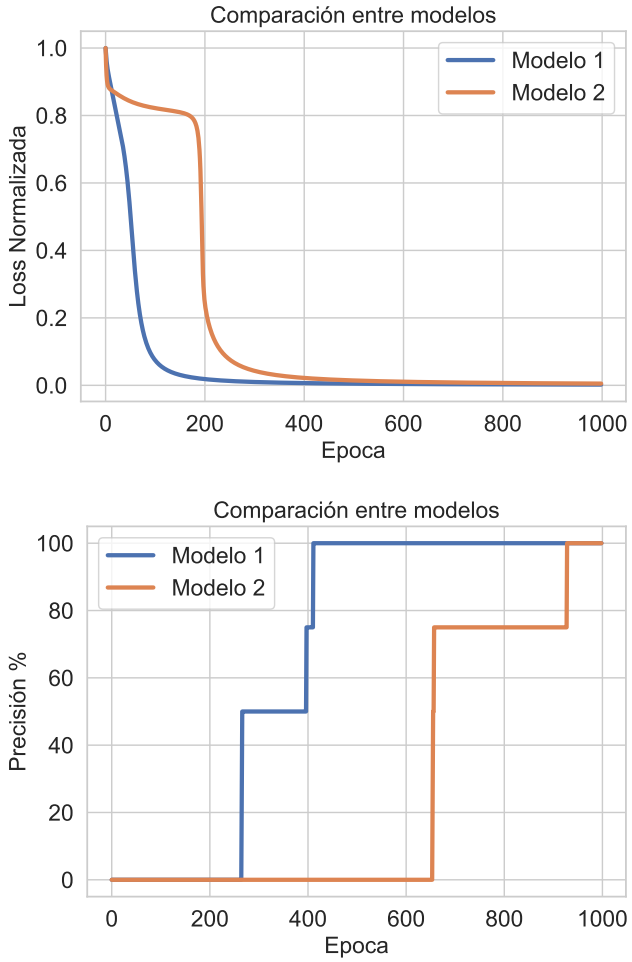


Figura 9: Loss normalizada y precisión en función de la época de entrenamiento para ambos modelos propuesto en el ejericio 6 para resolver la clasificación XOR.

Seria necesario modificar la semilla de la inicialización de los pesos para sacar conclusiones mas acertadas.

EJ7

Usando el modelo 1 del ejercicio 6, se propuso resolver el XOR genérico de N entradas. Eligiendo $N = 10$, se generaron todas las combinaciones posibles de datos y se eligieron un 90 % como datos de entrenamiento y un 10 % como datos de evaluación. Se probaron tres valores de neuronas de la capa intermedia, $N' = N/2, N$ y $2N$. En la figura 10 se muestra la evolución del modelo en el entrenamiento para diferentes valores de N'. Como se ve, para $N' = 2N$ el modelo consigue los mejores resultados y para $N' = N/2$ los peores, esto tiene sentido ya que la cantidad de parémtros es 4 veces mayor para $N' = 2N$. También se el aumento de rendimiento entre $N' = N/2$ y $N' = N$ es mucho mayor que el aumento de rendimiento entre $N' = N$ y $N' = 2N$.

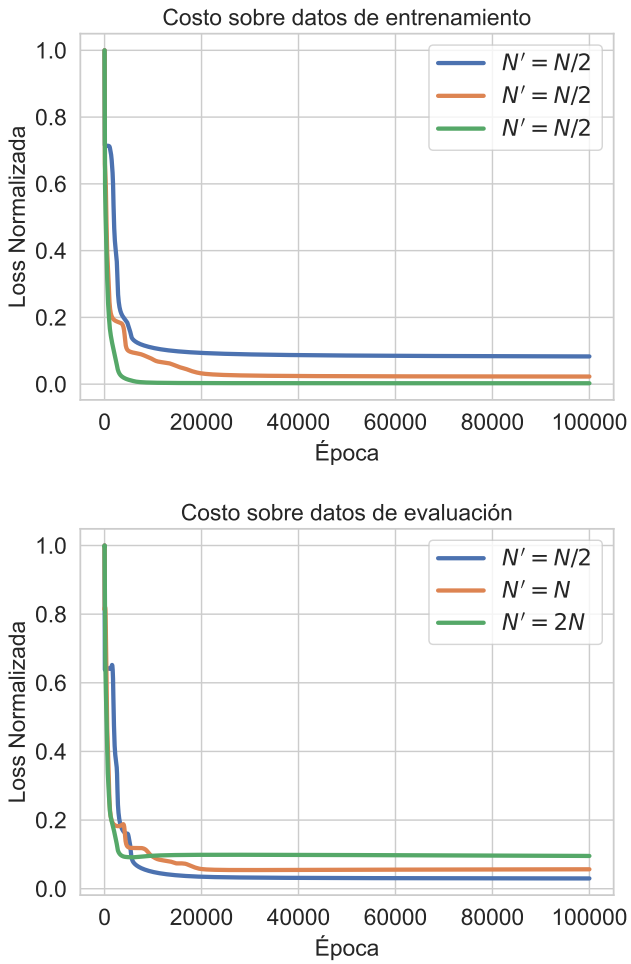


Figura 10: Loss normalizada y precisión del modelo en función de la época de entrenamiento sobre los datos de entrenamiento y evaluación para diversos valores de N' del modelo planteado en el ejercicio 7.

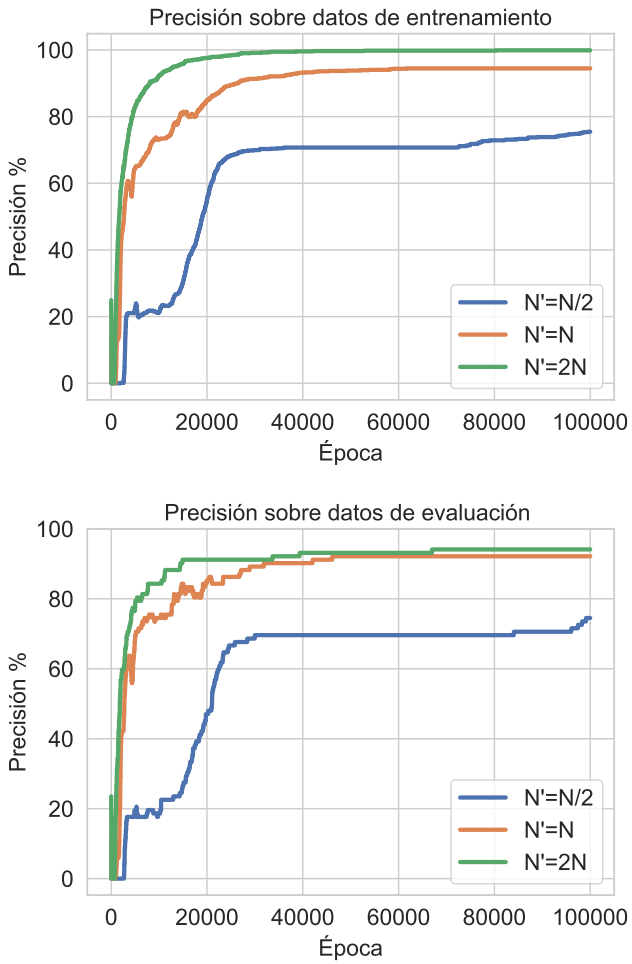


Figura 11: Precisión del modelo en función de la época de entrenamiento sobre los datos de entrenamiento y evaluación para diversos valores de N' del modelo planteado en el ejercicio 7.

Otra característica observable es que a medida que aumenta N', se hace mas notable el *overfitting*, esto puede verse porque la loss sobre los datos de entrenamiento disminuye a medida que aumenta N', pero sobre los datos de evaluación aumenta. Puede tener que ver con la convergencia mas rápida del modelo a mayor N'.

EJ8

Utilizando los módulos diseñados en el ejercicio 6, se busco resolver el problema de CIFAR10, para esto se creo una red con dos capas ocultas densas de 100 neuronas cada una. Se usaron activaciones sigmoidales en ambas capas y activación lineal en la capa de salida. Se usó la métrica MSE y los parámetros de entrenamiento elegidos fueron:

Parametro	**Valor**
λ	10^{-3}
Learning rate	10^{-3}
W	10^{-2}
batch size	50

En la figura 12 se muestra la evolución del modelo en función de las epocas. En este caso el *overtfitting* es mínimo, el modelo se comporta casi igual con los datos de entrenamiento y con los datos de evaluación. A pesar del aumento de los parámetros del problema, no se llega a resultados mejores que a los conseguidos en los ejercicios 3 a 5. Sin embargo esto tiene que ver con la velocidad de aprendizaje de la red, que es claramente menor, a pe-

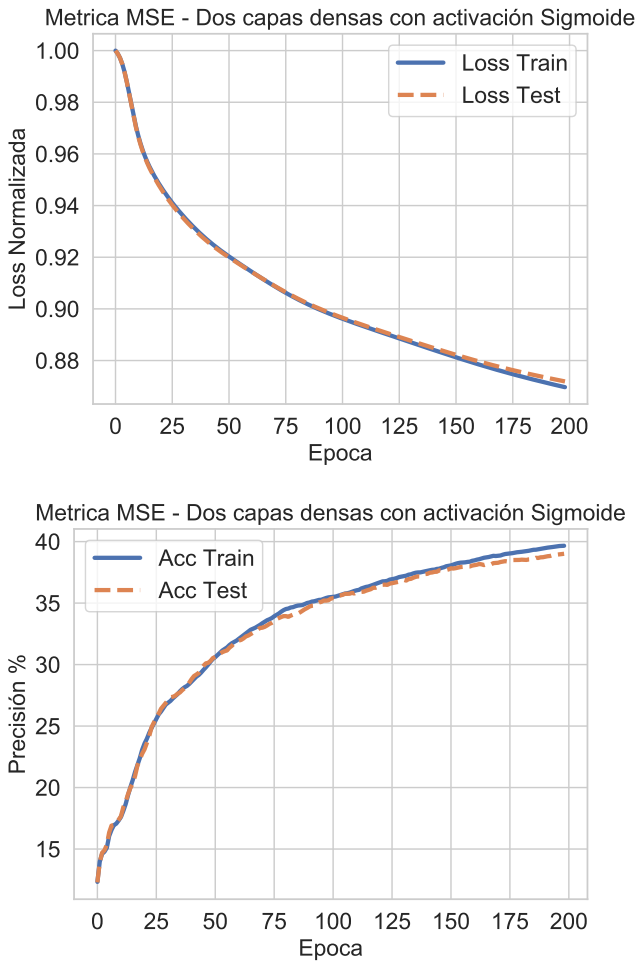


Figura 12: Función de costo y precisión del modelo en función de la época de entrenamiento sobre los datos de entrenamiento y evaluación.

sar de haber aumentado el *learning rate*. Esto puede estar muy relacionado con la utilización de dos funciones de activación sigmoidales. Se ve que el modelo sobre los datos de entrenamiento tiende a seguir creciendo, lo que indica que en mas épocas seria esperable ver un crecimiento de la precisión del modelo.

EJ1-2

Se adjunta una imagen con los ejercicios 1 y 2 resueltos a mano.