Los problemos del back propagation

- Rugosidad de la función: El grado de rugosidad de una función está dado por el número de mínimos locales y puntos de ensilladura. En el caso de las redes neuronales con multiplicidad de capas o *profundas*, se sabe que la cantidad de mínimos y puntos de ensilladura aumentarán exponencialmente mientras crezca el número de acoplamientos sinápticos. Estos, a su vez, dependerán de la arquitectura de la red, de la cantidad de neuronas de entrada, neuronas de salida, cantidad de capas ocultas y el número de neuronas dispuestas en ellas. Entonces cuanto mayor sea la cantidad de neuronas, aumenta el número de acoplamientos y, por lo tanto, aumentarán los mínimos locales en la función (el error en estos puntos sería de un valor distinto de cero), de manera que la función error es muy rugosa (producto de la alternancia de signos y la aleatoriedad). Por consiguiente la función podría o bien quedarse atrapada en mínimos locales o alcanzar puntos de inflexión para los cuales requeriría mucho tiempo salir.

- Valores iniciales de w_{ij}: Cuando los valores asignados a la matriz de acoplamientos sinápticos se eligen inicialmente al azar, la evolución del error (es decir, el descenso que se realiza sobre la superficie en función del gradiente) dependerá de estos valores. En este caso, alcanzar un mínimo local se verá estrechamente afectado por la elección de los valores de los pesos sinápticos iniciales, de forma que la probabilidad de hallar los mínimos locales dependerá de la probabilidad de caer cerca de ellos al seleccionar los valores de los acoplamientos.

- Dependencia de las derivadas: Como es posible observar en la expresión de la función error, para corregir los acoplamientos es preciso conocer las derivadas de las funciones de activación. Esto significa que cuanto mayor sea la complejidad del proceso de derivación de las funciones, mayor será el tiempo necesario para la evolución del sistema. Cuando se analizan los tiempos de análisis, es preciso considerar igualmente lo que ocurriría si la función se encontrara en puntos de ensilladura o si las derivadas tomaran valores próximos a cero dado que la evolución del error dependerá del gradiente. En estos casos, la evolución y aprendizaje del sistema también requeriría mucho tiempo haciendo que el sistema avance cada vez más lentamente a medida que agregan capas y que se aplique la regla de la cadena al cálculo del gradiente. De modo que la anidación de funciones anteriores para cada capa siguiente implican complejidad matemática y computacional dificultando el aprendizaje de la red y la búsqueda de una solución para el problema analizado.

- Dependencia de la razón de aprendizaje: La regla de aprendizaje depende del factor η (razón de aprendizaje) siendo muy sensible a las variaciones en su valor y pudiendo divergir con facilidad (alejándose del valor en el cual $E = 0$) sumado al hecho que conocer con exactitud su valor óptimo (de modo tal que permita alcanzar el mínimo error en el menor número de pasos) resulta una tarea difícil.

- Overfitting[1]: Aunque está más asociado a los problemas propios del machine-learning, el overfitting es otro inconveniente que suele presentarse y que se observa al graficar los valores de la función costo en el tiempo (aumentando la cantidad de épocas). Derivado del objetivo del aprendizaje supervisado (que la respuesta de la red coincida con la respuesta deseada), el overfitting implica una red que predice el conjunto de entrenamiento con una alta precisión de manera que es posible correr el riesgo de incluir el ruido de los datos memorizando peculiaridades y evitando encontrar un conjunto de acoplamiento que permita predecir los resultados. Es decir la red predice muy bien los elementos del conjunto de entrenamiento pero no responde de forma deseada cuando se le presentan los conjuntos de testeos haciendo que la red no generalice correctamente. El exceso de precisión a la hora ajustar el conjunto de entrenamiento se traduce en emplear modelos que incluyen más parámetros o términos de los necesarios o bien modelos más complejos que los necesarios[2], lo que afecta el cálculo del gradiente e incrementa los costos computacionales. Finalmente es importante no perder de vista que el objetivo del aprendizaje supervisado de una red neuronal artificial es poder usarla para resolver problemas de complejidad variable. En otras palabras, no resulta útil contar con una red que ajusta con alta precisión los ejemplos sino es capaz de predecir la respuesta del conjunto de testeos y, por lo tanto no es confiable para ser usada con datos desconocidos.

Fundamentalmente, puede observarse que la complejidad de estas redes está vinculado con la difícil tarea de ajustar una función de $\mathbb{R}^N$a $\mathbb{R}^M$, donde N son las neuronas de entrada y M las neuronas de salida. De manera que es preciso establecer una arquitectura y una matriz de acoplamientos sinápticos que permita ajustar el espacio en que se encuentran los elementos de entrada con el espacio en el que se hallan los estados de las neuronas de salida. Por último y considerando la forma de la regla de aprendizaje, es preciso destacar que la complejidad de su cálculo (y, por lo tanto, los tiempos de procesamiento asociados) irán de la mano con la complejidad de la red neuronal y su arquitectura.