Tentative Solutions

December 2023

- Optimal $\boldsymbol{\epsilon}$

  $$f(\boldsymbol{x}-\boldsymbol{\epsilon}\odot\nabla f(\boldsymbol{x}))\approx f(\boldsymbol{x})-\nabla f(\boldsymbol{x})^\top(\boldsymbol{\epsilon}\odot\nabla f(\boldsymbol{x}))+\frac{1}{2}(\boldsymbol{\epsilon}\odot\nabla f(\boldsymbol{x}))^\top\nabla^2 f(\boldsymbol{x})(\boldsymbol{\epsilon}\odot\nabla f(\boldsymbol{x}))=\mathcal{L}$$

  Here, we let $\boldsymbol{\alpha}=\boldsymbol{\epsilon}\odot\nabla f(\boldsymbol{x})$.

  $$\frac{\partial\mathcal{L}}{\partial\boldsymbol{\alpha}}=-\nabla f(\boldsymbol{x})+\nabla^2 f(\boldsymbol{x})\boldsymbol{\alpha}=\mathbf{0}$$

  Consequently, $\epsilon_i=\frac{\left(\nabla^2 f(\boldsymbol{x})^{-1}\nabla f(\boldsymbol{x})\right)_i}{\nabla f(\boldsymbol{x})_i}$.

- Gradient and evaluation (for simplicity, we have some modifications, considering only one step of updating, and three hidden layers. The final layer is used for linear regression. The initial values have the duplicated structure given in the 2),

  1. Gradient, and let $l_s^{(k)}=\sigma(z_s^{(k)})=\sigma\left(\sum_{b=1}^3 w_{s,b}^{(k)}l_b^{(k-1)}\right)$, and $f=\sum_{s=1}^3 w_s^{(K+1)}l_s^{(K)}$, $K=3$. Note that the last activation function is linear.

  $$\begin{aligned}\sum_{s'=1}^3\frac{\partial L}{\partial l_{s'}^{(K-1)}} &= \frac{\partial L}{\partial f}\sum_{s=1}^3\frac{\partial f}{\partial l_s^{(K)}}\sigma'\left(\sum_{b=1}^3 w_{s,b}^{(K)}l_b^{(K-1)}\right)\sum_{s'=1}^3 w_{s,s'}^{(K)}\\ \frac{\partial L}{\partial l_t^{(K-2)}} &= \frac{\partial L}{\partial f}\sum_{s=1}^3\frac{\partial f}{\partial l_s^{(K)}}\sigma'\left(\sum_{b=1}^3 w_{s,b}^{(K)}l_b^{(K-1)}\right)\sum_{s'=1}^3 w_{s,s'}^{(K)}\\ &\quad\times\sigma'\left(\sum_{b=1}^3 w_{s',b}^{(K-1)}l_b^{(K-2)}\right)w_{s',t}^{(K-1)},\end{aligned}$$

  and

  $$\frac{\partial L}{\partial w_{t,u}^{(K-2)}} = \frac{\partial L}{\partial l_t^{(K-2)}}\frac{\partial l_t^{(K-2)}}{\partial w_{t,u}^{(K-2)}}=\frac{\partial L}{\partial l_t^{(K-2)}}\sigma'(z_t^{(K-2)})l_u^{(K-3)}.$$

2. Evaluation for $x = (0, 1)$ and $y = 1/2$ with $L = (y - f(x))^2$. All initial values are set to $(1/2, -1/2), (1/4, -1/4), (-1, 1)$ for the input layer and the duplicated $(1/3, 2/3, -1/3), (-1/3, -2/3, 1/3), (1/2, -1/2, 1/2)$, and the $(1/3, -1/3, 1/3)$ for intermediate and final layers.

   (a) Forward by $x = (0, 1)$:

$$
\begin{aligned}
&z_1^{(1)} = -1/2, z_2^{(1)} = -1/4, z_3^{(1)} = 1\\
&l_1^{(1)} = \sigma(z_1^{(1)}) = 0, l_2^{(1)} = \sigma(z_2^{(1)}) = 0, l_3^{(1)} = \sigma(z_3^{(1)}) = 1\\
&z_1^{(2)} = -1/3, z_2^{(2)} = 1/3, z_3^{(2)} = 1/2\\
&l_1^{(2)} = \sigma(z_1^{(2)}) = 0, z_2^{(2)} = \sigma(z_2^{(2)}) = 1/3, z_3^{(2)} = \sigma(z_3^{(2)}) = 1/2\\
&z_1^{(3)} = 2/9 - 1/6, z_2^{(3)} = -2/9 + 1/6, z_3^{(3)} = -1/6 + 1/4\\
&l_1^{(3)} = \sigma(z_1^{(3)}) = 1/18, l_2^{(3)} = \sigma(z_2^{(3)}) = 0, l_3^{(3)} = \sigma(z_3^{(3)}) = 1/12\\
&f = (1/18 + 1/12) * (1/3) = 5/108
\end{aligned}
$$

(b) Back Prop.

$$\begin{aligned}
\frac{\partial L}{\partial f} &= 2 \times (5/108 - 54/108) = -0.907 \\
\frac{\partial L}{\partial w^{(4)}} &= -0.907 \times [1/18, 0, 1/12] = [-0.050, 0, -0.076] \\
\frac{\partial L}{\partial l^{(3)}} &= -0.907 \times [1/3, -1/3, 1/3] = [-0.302, 0.302, -0.302] \\
\frac{\partial L}{\partial w^{(3)}} &= [[-0.907 \times [0, 1/3, 1/2]], [0 \times [0, 1/3, 1/2]], [-0.907 \times [0, 1/3, 1/2]]] \\
\frac{\partial L}{\partial l^{(2)}} &= [-0.302, 0.302, -0.302] \cdot [1, 0, 1] \\
&\quad \cdot [[1/3, -1/3, 1/2], [2/3, -2/3, -1/2], [-1/3, 1/3, 1/2]] = [-0.252, -0.050, -0.050] \\
\frac{\partial L}{\partial w^{(2)}} &= [[0 \times [0, 0, 1]], [-0.050 \times [0, 0, 1]], [-0.050 \times [0, 0, 1]]] \\
\frac{\partial L}{\partial l^{(1)}} &= [-0.252, -0.050, -0.050] \cdot [0, 1, 1] \\
&\quad \cdot [[1/3, -1/3, 1/2], [2/3, -2/3, -1/2], [-1/3, 1/3, 1/2]] = [-0.008, 0.058, -0.042] \\
\frac{\partial L}{\partial w^{(1)}} &= [0 \times [0, 1], 0 \times [0, 1], -0.042 \times [0, 1]].
\end{aligned}$$

Learning rate = 0.01

$$\begin{array}{rcl}
\hline
w^{(1)}_{1:} &=& (0.5, 0.5) \\
w^{(1)}_{2:} &=& (0.25, 0.25) \\
w^{(1)}_{3:} &=& (-1, 1 + 0.00042) \\
\hline
w^{(2)}_{1:} &=& (0.33, 0.66, -0.33) \\
w^{(2)}_{2:} &=& (-0.33, -0.66, 0.33 + 0.00050) \\
w^{(2)}_{3:} &=& (0.5, -0.5, 0.5 + 0.00050) \\
\hline
w^{(3)}_{1:} &=& (0.33, 0.66 + 0.000302, -0.33 + 0.000456) \\
w^{(3)}_{2:} &=& (-0.33, -0.66, 0.33) \\
w^{(3)}_{3:} &=& (0.5, -0.5 + 0.000302, -0.33 + 0.000456) \\
\hline
w^{(4)} &=& (0.33 + 0.00050, 0.66, -0.33 + 0.00076) \\
\hline
\end{array}$$

- Effect of layers: Deep layers provide a somewhat larger variation in the predicted function. However, the prediction error does not show a drastic change.