Deep MLP training: stochastic gradient descent.

Deep-MLP-Training ($\mathbf{D}, h, \eta, \texttt{maxiter}, n_1, n_2, \cdots, n_h, f^1, f^2, \cdots, f^{h+1}$):

$n_0 \leftarrow d$ `// input layer size`

$n_{h+1} \leftarrow p$ `// output layer size`

`// Initialize weight matrices and bias vectors`

for $l = 0, 1, 2, \cdots, h$ **do**

    $\boldsymbol{\theta}_l \leftarrow$ random n_{l+1} vector with small values

    $\mathbf{W}_l \leftarrow$ random $n_l \times n_{l+1}$ matrix with small values

$t \leftarrow 0$ `// iteration counter`

repeat

    foreach $(\mathbf{x}_i, \mathbf{y}_i) \in \mathbf{D}$ *in random order* **do**

        `// Feed-Forward Phase`

        $\mathbf{z}^0 \leftarrow \mathbf{x}_i$

        for $l = 0, 1, 2, \ldots, h$ **do**

            $\mathbf{z}^{l+1} \leftarrow f^{l+1}\left(\mathbf{W}_l^T \cdot \mathbf{z}^l\right)$

        $\mathbf{o}_i \leftarrow \mathbf{z}^{h+1}$

        `// Backpropagation Phase`

        $\boldsymbol{\delta}^{h+1} \leftarrow \partial\mathbf{f}^{h+1} \odot \partial\boldsymbol{\mathcal{E}}_{\mathbf{x}}$ `// net gradients at output`

                `// use` $\partial\mathbf{F}^{h+1} \cdot \partial\boldsymbol{\mathcal{E}}_{\mathbf{x}}$ `for softmax`

        for $l = h, h-1, \cdots, 1$ **do**

            $\boldsymbol{\delta}^l \leftarrow \partial\mathbf{f}^l \odot \left(\mathbf{W}_l \cdot \boldsymbol{\delta}^{l+1}\right)$ `// net gradients at layer` l

        `// Gradient Descent Step`

        for $l = 0, 1, \cdots, h$ **do**

            $\nabla_{\mathbf{W}_l} \leftarrow \mathbf{z}^l \cdot \left(\boldsymbol{\delta}^{l+1}\right)^T$ `// weight gradient matrix at layer` l

            $\nabla_{\boldsymbol{\theta}_l} \leftarrow \boldsymbol{\delta}^{l+1}$ `// bias gradient vector at layer` l

        for $l = 0, 1, \cdots, h$ **do**

            $\mathbf{W}_l \leftarrow \mathbf{W}_l - \eta \cdot \nabla_{\mathbf{W}_l}$ `// update` $\mathbf{W}_l$

            $\boldsymbol{\theta}_l \leftarrow \boldsymbol{\theta}_l - \eta \cdot \nabla_{\boldsymbol{\theta}_l}$ `// update` $\boldsymbol{\theta}_l$

    $t \leftarrow t + 1$

until $t \geq \texttt{maxiter}$