We will check techniques to address the difficulty of storing or inverting the Hessian.

But before that let’s derive the mathematical form.
For CNN, the gradient of $f(\theta)$ is

$$\nabla f(\theta) = \frac{1}{C} \theta + \frac{1}{l} \sum_{i=1}^{l} (J^i)^T \nabla_{z^{L+1},i} \xi(z^{L+1,i}; y^i, Z^{1,i}),$$

where

$$J^i = \begin{bmatrix} \frac{\partial z_1^{L+1,i}}{\partial \theta_1} & \cdots & \frac{\partial z_1^{L+1,i}}{\partial \theta_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial z_{n_{L+1}}^{L+1,i}}{\partial \theta_1} & \cdots & \frac{\partial z_{n_{L+1}}^{L+1,i}}{\partial \theta_n} \end{bmatrix}_{n_{L+1} \times n}, \quad i = 1, \ldots, l,$$
Hessian Matrix II

is the Jacobian of $z^{L+1,i}(\theta)$.

The Hessian matrix of $f(\theta)$ is

$$\nabla^2 f(\theta) = \frac{1}{C} I + \frac{1}{l} \sum_{i=1}^{l} (J^i)^T B^i J^i$$

$$+ \frac{1}{l} \sum_{i=1}^{l} \sum_{j=1}^{n_L} \frac{\partial \xi(z^{L+1,i}; y^i, Z^{1,i})}{\partial z^{L+1,i}_j} \frac{\partial^2 z_j^{L+1,i}}{\partial \theta_1 \partial \theta_1} \cdots \frac{\partial^2 z_j^{L+1,i}}{\partial \theta_1 \partial \theta_n} \cdots \frac{\partial^2 z_j^{L+1,i}}{\partial \theta_n \partial \theta_1} \cdots \frac{\partial^2 z_j^{L+1,i}}{\partial \theta_n \partial \theta_n}$$
where $\mathcal{I}$ is the identity matrix and $B^i$ is the Hessian of $\xi(\cdot)$ with respect to $z^{L+1,i}$:

$$B^i = \nabla^2_{z^{L+1,i}, z^{L+1,i}} \xi(z^{L+1,i}; y^i, Z^{1,i})$$

- More precisely,

$$B^i_{ts} = \frac{\partial^2 \xi(z^{L+1,i}; y^i, Z^{1,i})}{\partial z^{L+1,i}_t \partial z^{L+1,i}_s}, \forall t, s = 1, \ldots, n_{L+1}. \quad (3)$$

- Usually $B^i$ is very simple.
Hessian Matrix IV

For example, if the squared loss is used,

\[ \xi(z^{L+1,i}; y^i) = \|z^{L+1,i} - y^i\|^2. \]

then

\[ B^i = \begin{bmatrix} 2 \\ \vdots \\ 2 \end{bmatrix} \]

Usually we consider a convex loss function

\[ \xi(z^{L+1,i}; y^i) \]

with respect to \( z^{L+1,i} \)
Thus $B^i$ is positive semi-definite

The last term of $\nabla^2 f(\theta)$ may not be positive semi-definite

Note that for a twice differentiable function $f(\theta)$

$$f(\theta) \text{ is convex}$$

if and only if

$$\nabla^2 f(\theta) \text{ is positive semi-definite}$$
The Jacobian matrix of $z^{L+1,i}(\theta) \in R^{n_{L+1}}$ is

$$J^i = \begin{bmatrix}
\frac{\partial z_1^{L+1,i}}{\partial \theta_1} & \cdots & \frac{\partial z_1^{L+1,i}}{\partial \theta_n} \\
\vdots & \ddots & \vdots \\
\frac{\partial z_{n_{L}}^{L+1,i}}{\partial \theta_1} & \cdots & \frac{\partial z_{n_{L}}^{L+1,i}}{\partial \theta_n}
\end{bmatrix} \in R^{n_{L+1} \times n}, \ i = 1, \ldots l.$$

- $n_{L+1}$: # of neurons in the output layer
- $n$: number of total variables
- $n_{L+1} \times n$ can be large
The Hessian matrix $\nabla^2 f(\theta)$ is now not positive definite.

We may need a positive definite approximation.

Many existing Newton methods for NN has considered the Gauss-Newton matrix (Schraudolph, 2002)

$$G = \frac{1}{C} I + \frac{1}{l} \sum_{i=1}^{l} (J^i)^T B^i J^i$$

by removing the last term in $\nabla^2 f(\theta)$.
The Gauss-Newton matrix is positive definite if $B^i$ is positive semi-definite. This can be achieved if we use a convex loss function in terms of $z^{L+1,i}(\theta)$. We then solve

$$Gd = -\nabla f(\theta)$$