Supplementary Material for “Relative Fisher Information and Natural Gradient for Learning Large Modular Models”

Ke Sun and Frank Nielsen

Contents

1 Non-linear Activation Functions .............................................. 1

2 Examples of RFIMs .......................................................... 2
   2.1 A Single $\tanh$ Neuron .................................................. 3
   2.2 A Single $\sigma$ Neuron ................................................... 4
   2.3 A Single $\text{relu}$ Neuron .............................................. 5
   2.4 A Single $\text{elu}$ Neuron .............................................. 5
   2.5 RFIM of a Linear Layer ............................................... 5
   2.6 RFIM of a Non-Linear Layer ......................................... 6
   2.7 RFIM of a Softmax Layer ............................................ 7
   2.8 RFIM of Two layers ................................................... 7

3 Proof of Theorem 3 .................................................................. 8

4 Experimental Settings & Zoomed Learning Curves ..................... 8

1 Non-linear Activation Functions

By definition,
\[
\tanh(t) \overset{\text{def}}{=} \frac{\exp(t) - \exp(-t)}{\exp(t) + \exp(-t)},
\]

and
\[
\text{sech}(t) \overset{\text{def}}{=} \frac{2}{\exp(t) + \exp(-t)}.
\]

It is easy to verify that
\[
\text{sech}^2(t) = [1 + \tanh(t)][1 - \tanh(t)] = 1 - \tanh^2(t).
\]
By eq. (1),
\[
\tanh'(t) = \frac{\exp(t) + \exp(-t)}{\exp(t) + \exp(-t)} - \frac{\exp(t) - \exp(-t)}{[\exp(t) + \exp(-t)]^2} [\exp(t) - \exp(-t)]
\]
\[
= \frac{[\exp(t) + \exp(-t)]^2 - [\exp(t) - \exp(-t)]^2}{[\exp(t) + \exp(-t)]^2} = \frac{4}{[\exp(t) + \exp(-t)]^2} = \text{sech}^2(t).
\]

By definition,
\[
\text{sigm}(t) \overset{\text{def}}{=} \frac{1}{1 + \exp(-t)}.
\]

Therefore
\[
\text{sigm}'(t) = -\frac{1}{[1 + \exp(-t)]^2} (-\exp(-t)) = \frac{\exp(-t)}{[1 + \exp(-t)]^2} = \text{sigm}(t)[1 - \text{sigm}(t)].
\]

A smoothed version of the \text{relu} function is given by
\[
\text{relu}_\omega(t) \overset{\text{def}}{=} \omega \ln \left( \exp \left( \frac{\omega}{t} \right) + \exp \left( \frac{t}{\omega} \right) \right),
\]
where \(\omega > 0\) and \(0 \leq \iota < 1\). Then,
\[
\text{relu}_\omega'(t) = \omega \frac{1}{\exp \left( \frac{\omega}{t} \right) + \exp \left( \frac{t}{\omega} \right)} \left( \frac{t}{\omega} \exp \left( \frac{\omega}{t} \right) + \frac{1}{\omega} \exp \left( \frac{t}{\omega} \right) \right)
\]
\[
= \frac{t \exp \left( \frac{\omega}{t} \right) + \exp \left( \frac{t}{\omega} \right)}{\exp \left( \frac{\omega}{t} \right) + \exp \left( \frac{t}{\omega} \right)}
\]
\[
= \frac{t + (1 - \iota) \exp \left( \frac{\iota}{\omega} \right) + \exp \left( \frac{t}{\omega} \right)}{1}
\]
\[
= \iota + (1 - \iota) \text{sigm} \left( \frac{1 - \iota}{\omega} t \right). \quad \text{(2)}
\]

By definition,
\[
\text{elu}(t) = \begin{cases} 
  t & \text{if } t \geq 0 \\
  \alpha (\exp(t) - 1) & \text{if } t < 0.
\end{cases}
\]

Therefore
\[
\text{elu}'(t) = \begin{cases} 
  1 & \text{if } t \geq 0 \\
  \alpha \exp(t) & \text{if } t < 0.
\end{cases} \quad \text{(3)}
\]

2 Examples of RFIMs

Table 1 shows a list of commonly used RFIMs, with detailed derivations given in the following subsections.
Table 1: Commonly used RFIMs

<table>
<thead>
<tr>
<th>Subsystem</th>
<th>the RFIM $g^\beta(w)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A tanh neuron</td>
<td>$\text{sech}^2(w^\top \hat{x})\hat{x}\hat{x}^\top$</td>
</tr>
<tr>
<td>A sign neuron</td>
<td>$\text{sign}(w^\top \hat{x})(1 - \text{sign}(w^\top \hat{x}))\hat{x}\hat{x}^\top$</td>
</tr>
<tr>
<td>A relu neuron</td>
<td>$\left{ \begin{array}{ll} \hat{x}\hat{x}^\top &amp; \text{if } w^\top \hat{x} \geq 0 \ \left( \alpha \exp(w^\top \hat{x}) \right)^2 \hat{x}\hat{x}^\top &amp; \text{if } w^\top \hat{x} &lt; 0 \end{array} \right.$</td>
</tr>
<tr>
<td>A elu neuron</td>
<td>$\left{ \begin{array}{ll} \hat{x}\hat{x}^\top &amp; \text{if } w^\top \hat{x} \geq 0 \ \left{ \nu_f(w_1, \hat{x})\hat{x}\hat{x}^\top, \cdots, \nu_f(w_m, \hat{x})\hat{x}\hat{x}^\top \right} &amp; \text{if } w^\top \hat{x} &lt; 0 \end{array} \right.$</td>
</tr>
<tr>
<td>A linear layer</td>
<td>$\text{diag}[\hat{x}\hat{x}^\top, \cdots, \hat{x}\hat{x}^\top]$</td>
</tr>
<tr>
<td>A non-linear layer</td>
<td>$\text{diag}[\nu_f(w_1, \hat{x})\hat{x}\hat{x}^\top, \cdots, \nu_f(w_m, \hat{x})\hat{x}\hat{x}^\top]$</td>
</tr>
<tr>
<td>A soft-max layer</td>
<td>a dense matrix as shown in eq. (10)</td>
</tr>
<tr>
<td>Two layers</td>
<td>a dense matrix as shown in eq. (12)</td>
</tr>
</tbody>
</table>

2.1 A Single tanh Neuron

Consider a neuron with parameters $w$ and a Bernoulli output $y \in \{+, -\}$, $p(y = +) = p^+$, $p(y = -) = p^-$, and $p^+ + p^- = 1$. By the definition of RFIM, we have

$$g^\beta(y)(w) = p^+ \frac{\partial \ln p^+}{\partial w} \frac{\partial \ln p^+}{\partial w^\top} + p^- \frac{\partial \ln p^-}{\partial w} \frac{\partial \ln p^-}{\partial w^\top} = \frac{1}{p^+} \frac{\partial p^+}{\partial w} \frac{\partial p^+}{\partial w^\top} + \frac{1}{p^-} \frac{\partial p^-}{\partial w} \frac{\partial p^-}{\partial w^\top}.$$ 

Since $p^+ + p^- = 1,$

$$\frac{\partial p^+}{\partial w} + \frac{\partial p^-}{\partial w} = 0.$$ 

Therefore, the RFIM of a Bernoulli neuron has the general form

$$g^\beta(w) = \left( \frac{1}{p^+} + \frac{1}{p^-} \right) \frac{\partial p^+}{\partial w} \frac{\partial p^+}{\partial w^\top} = \frac{1}{p^+p^-} \frac{\partial p^+}{\partial w} \frac{\partial p^+}{\partial w^\top}. \quad (4)$$

A single tanh neuron with stochastic output $y \in \{-1, 1\}$ is given by

$$p(y = 1) = \frac{1 + \mu(x)}{2}, \quad (6)$$
$$p(y = -1) = \frac{1 - \mu(x)}{2}, \quad (5)$$
$$\mu(x) = \tanh(w^\top \hat{x}). \quad (7)$$

By eq. (4),

$$\begin{align*}
g^\beta(w) &= \frac{1}{1 - \mu(x)} \frac{1 + \mu(x)}{2} \left( \frac{1}{2} \frac{\partial \mu}{\partial w} \right) \left( \frac{1}{2} \frac{\partial \mu}{\partial w^\top} \right) \\
&= \frac{1}{(1 - \mu(x))(1 + \mu(x))} \left[ 1 - \mu^2(x) \right]^2 \hat{x}\hat{x}^\top \\
&= \left[ 1 - \tanh^2(w^\top \hat{x}) \right] \hat{x}\hat{x}^\top \\
&= \left[ 1 - \text{sech}^2(w^\top \hat{x}) \right] \hat{x}\hat{x}^\top.
\end{align*}$$
An alternative analysis is given as follows. By eqs. (5) to (7),

\[ p(y = -1) = \frac{\exp(-w^\top \tilde{x})}{\exp(w^\top \tilde{x}) + \exp(-w^\top \tilde{x})}, \]
\[ p(y = 1) = \frac{\exp(w^\top \tilde{x})}{\exp(w^\top \tilde{x}) + \exp(-w^\top \tilde{x})}. \]

Then,

\[ g_y^\mu(w) = \mathbb{E}_{y \sim p(y \mid x)} \left( -\frac{\partial^2 \ln p(y)}{\partial w \partial w^\top} \right) \]
\[ = \frac{\partial^2}{\partial w \partial w^\top} \ln \left( \frac{\exp(w^\top \tilde{x}) + \exp(-w^\top \tilde{x})}{\exp(w^\top \tilde{x}) + \exp(-w^\top \tilde{x})} \right) \]
\[ = \frac{\partial}{\partial \tilde{x}} \ln \left( \frac{\exp(w^\top \tilde{x}) - \exp(-w^\top \tilde{x})}{\exp(w^\top \tilde{x}) + \exp(-w^\top \tilde{x})} \right) \tilde{x} \]
\[ = \frac{\partial}{\partial \tilde{x}} \tanh(w^\top \tilde{x}) \tilde{x} \]
\[ = \text{sech}^2(w^\top \tilde{x}) \tilde{x} \tilde{x}^\top. \]

The intuitive meaning of \( g_y^\mu(w) \) is a weighted covariance to emphasize such “informative” \( x \)'s that

- are in the linear region of \( \tanh \)
- contain “ambiguous” samples

We will need at least \( \dim(w) \) samples to make \( g_y^\mu(w) \) full rank.

### 2.2 A Single sigm Neuron

A single \( \text{sigm} \) neuron is given by

\[ p(y = 0) = 1 - \mu(x), \]
\[ p(y = 1) = \mu(x), \]
\[ \mu(x) = \text{sigm}(w^\top \tilde{x}). \]

By eq. (4),

\[ g_y^\mu(w) = \frac{1}{p(y = 0)p(y = 1)} \frac{\partial p(y = 1)}{\partial w} \frac{\partial p(y = 1)}{\partial w^\top} \]
\[ = \frac{1}{\mu(x)(1 - \mu(x))} \frac{\partial \mu}{\partial w} \frac{\partial \mu}{\partial w^\top} \]
\[ = \frac{1}{\mu(x)(1 - \mu(x))^2} \mu^2(x)(1 - \mu(x))^2 \tilde{x} \tilde{x}^\top \]
\[ = \mu(x)(1 - \mu(x)) \tilde{x} \tilde{x}^\top \]
\[ = \text{sigm}(w^\top \tilde{x}) [1 - \text{sigm}(w^\top \tilde{x})] \tilde{x} \tilde{x}^\top. \]
2.3 A Single relu Neuron

Consider a single neuron with Gaussian output \( p(y \mid w, x) = G(y \mid \mu(w, x), \sigma^2) \). Then

\[
g^y(w \mid x) = E_{p(y \mid w, x)} \left[ \frac{\partial \ln G(y \mid \mu, \sigma^2)}{\partial w} \frac{\partial \ln G(y \mid \mu, \sigma^2)}{\partial w^\top} \right] \\
= E_{p(y \mid w, x)} \left[ \frac{\partial}{\partial w} \left( -\frac{1}{2\sigma^2} (y - \mu)^2 \right) \frac{\partial}{\partial w^\top} \left( -\frac{1}{2\sigma^2} (y - \mu)^2 \right) \right] \\
= E_{p(y \mid w, x)} \left[ \left( -\frac{1}{\sigma^2} (\mu - y) \right)^2 \frac{\partial \mu}{\partial w} \frac{\partial \mu}{\partial w^\top} \right] \\
= \frac{1}{\sigma^4} E_{p(y \mid w, x)} (\mu - y)^2 \frac{\partial \mu}{\partial w} \frac{\partial \mu}{\partial w^\top} \\
= \frac{1}{\sigma^2} \frac{\partial \mu}{\partial w} \frac{\partial \mu}{\partial w^\top}.
\]

We set \( \sigma = 1 \) to get rid of a scale parameter of the RFIM. We get

\[
g^y(w \mid x) = \frac{\partial \mu}{\partial w} \frac{\partial \mu}{\partial w^\top}.
\]

A single relu neuron is given by

\[
\mu(w, x) = \text{relu}_w(w^\top \bar{x}).
\]

By eqs. (2) and (8),

\[
g^y(w) = \left( \ell + (1 - \ell) \text{sign} \left( \frac{1}{\omega^2} \omega^\top w^\top \bar{x} \right) \right)^2 \bar{x} \bar{x}^\top.
\]

2.4 A Single elu Neuron

Similar to the analysis in Subsec. 2.3, a single elu neuron is given by

\[
\mu(w, x) = \text{elu}(w^\top \bar{x}).
\]

By eq. (3),

\[
\frac{\partial \mu}{\partial w} = \begin{cases} 
    \bar{x} & \text{if } w^\top \bar{x} \geq 0 \\
    \alpha \exp(w^\top \bar{x}) \bar{x} & \text{if } w^\top \bar{x} < 0.
\end{cases}
\]

By eq. (8),

\[
g^y(w) = \begin{cases} 
    \bar{x} \bar{x}^\top & \text{if } w^\top \bar{x} \geq 0 \\
    \left( \alpha \exp(w^\top \bar{x}) \right)^2 \bar{x} \bar{x}^\top & \text{if } w^\top \bar{x} < 0.
\end{cases}
\]

2.5 RFIM of a Linear Layer

Consider a linear layer

\[
p(y) = G(y \mid W^\top \bar{x}, \sigma^2 I),
\]

where \( W = (w_1, \ldots, w_D) \). By the definition of the multivariate Gaussian distribution,

\[
\ln p(y) = -\frac{1}{2} \ln 2\pi - \frac{D_y}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{D_y} (y_i - w_i^\top \bar{x})^2.
\]
Therefore,
$$\forall i, \quad \frac{\partial}{\partial w_i} \ln p(y) = -\frac{1}{\sigma^2} (w_i^T \hat{x} - y_i) \hat{x}.$$ 

Therefore,
$$\forall i, \forall j \quad \frac{\partial}{\partial w_i} \ln p(y) \frac{\partial}{\partial w_j} \ln p(y) = \frac{1}{\sigma^2} (y_i - w_i^T \hat{x}) (y_j - w_j^T \hat{x}) \hat{x} \hat{x}^T.$$ 

$W$ is vectorized by stacking its columns $\{w_i\}_{i=1}^{D_y}$. In the following $W$ will be used interchangeably to denote either the matrix or its vector form. Correspondingly, the RFIM $g^y(W)$ has $D_y \times D_y$ blocks, where the off-diagonal blocks are

$$\forall i \neq j, \quad E_{p(y)} \left( \frac{\partial}{\partial w_i} \ln p(y) \frac{\partial}{\partial w_j} \ln p(y) \right) = \frac{1}{\sigma^4} E_{p(y)} [(y_i - w_i^T \hat{x}) (y_j - w_j^T \hat{x})] \hat{x} \hat{x}^T = 0,$$

and the diagonal blocks are

$$\forall i, \quad E_{p(y)} \left( \frac{\partial}{\partial w_i} \ln p(y) \frac{\partial}{\partial w_i} \ln p(y) \right) = \frac{1}{\sigma^2} E_{p(y)} (y_i - w_i^T \hat{x})^2 \hat{x} \hat{x}^T = \frac{1}{\sigma^2} \hat{x} \hat{x}^T.$$

In summary,
$$g^y(W) = \frac{1}{\sigma^2} \text{diag} [\hat{x} \hat{x}^T, \ldots, \hat{x} \hat{x}^T].$$

By setting $\sigma = 1$ we get
$$g^y(W) = \text{diag} [\hat{x} \hat{x}^T, \ldots, \hat{x} \hat{x}^T].$$

### 2.6 RFIM of a Non-Linear Layer

The statistical model of a non-linear layer with independent output units is

$$p(y \mid W, x) = \prod_{i=1}^{D_y} p(y_i \mid w_i, x).$$

Then,
$$\ln p(y \mid W, x) = \sum_{i=1}^{D_y} \ln p(y_i \mid w_i, x).$$

Therefore,
$$\frac{\partial^2}{\partial W \partial W^T} \ln p(y \mid W, x) = \begin{bmatrix}
\frac{\partial^2}{\partial w_i \partial w_i^T} \ln p(y_1 \mid w_1, x) \\
\vdots \\
\frac{\partial^2}{\partial w_{D_y} \partial w_{D_y}^T} \ln p(y_{D_y} \mid w_{D_y}, x)
\end{bmatrix}.$$ 

Therefore the RFIM $g^y(W)$ is a block-diagonal matrix, with the $i$’th block given by

$$-E_{p(y \mid W, x)} \left[ \frac{\partial^2}{\partial w_i \partial w_i^T} \ln p(y_i \mid w_i, x) \right] = -E_{p(y_i \mid w_i, x)} \left[ \frac{\partial^2}{\partial w_i \partial w_i^T} \ln p(y_i \mid w_i, x) \right],$$

which is simply the single neuron RFIM of the $i$’th neuron.
2.7 RFIM of a Softmax Layer

Recall that
\[ \forall i \in \{1, \ldots, m\}, \quad p(y = i) = \frac{\exp(w_i \tilde{x})}{\sum_{i=1}^{m} \exp(w_i \tilde{x})}. \]

Then
\[ \forall i, \quad \ln p(y = i) = w_i \tilde{x} - \ln \sum_{i=1}^{m} \exp(w_i \tilde{x}). \]

Hence
\[ \forall i, \forall j, \quad \frac{\partial \ln p(y = i)}{\partial w_j} = \delta_{ij} \tilde{x} - \frac{\exp(w_j \tilde{x})}{\sum_{i=1}^{m} \exp(w_i \tilde{x})} \tilde{x}, \]

where \( \delta_{ij} = 1 \) if and only if \( i = j \) and \( \delta_{ij} = 0 \) otherwise. Then
\[ \forall i, \forall j, \forall k, \quad \frac{\partial^2 \ln p(y = i)}{\partial w_j \partial w_k} = -\delta_{jk} \sum_{m=1}^{m} \exp(w_m \tilde{x}) \tilde{x} \tilde{x}^T + \frac{\exp(w_j \tilde{x})}{\sum_{i=1}^{m} \exp(w_i \tilde{x})^2} \exp(w_k \tilde{x}) \tilde{x} \tilde{x}^T \]
\[ = (-\delta_{jk} \eta_j + \eta_j \eta_k) \tilde{x} \tilde{x}^T. \]

The right-hand-side of eq. (9) does not depend on \( i \). Therefore
\[ g^y(W) = \begin{bmatrix} (\eta_1 - \eta_1^2) \tilde{x} \tilde{x}^T & -\eta_1 \eta_2 \tilde{x} \tilde{x}^T & \cdots & -\eta_1 \eta_m \tilde{x} \tilde{x}^T \\ -\eta_2 \eta_1 \tilde{x} \tilde{x}^T & (\eta_2 - \eta_2^2) \tilde{x} \tilde{x}^T & \cdots & -\eta_2 \eta_m \tilde{x} \tilde{x}^T \\ \vdots & \vdots & \ddots & \vdots \\ -\eta_m \eta_1 \tilde{x} \tilde{x}^T & -\eta_m \eta_2 \tilde{x} \tilde{x}^T & \cdots & (\eta_m - \eta_m^2) \tilde{x} \tilde{x}^T \end{bmatrix}. \]

2.8 RFIM of Two layers

Consider a two-layer structure, where the output \( y \) satisfies a multivariate Bernoulli distribution with independent dimensions. By a similar analysis to Subsec. 2.1, we have
\[ g^y(W) = \sum_{i=1}^{D_y} \nu_f(c_i, h) \frac{\partial c_i^T h}{\partial W} \frac{\partial c_i^T h}{\partial W^T}. \]

It can be written block by block as \( g^y(W) = G_{ij} \big|_{D_y \times D_y} \), where each block \( G_{ij} \) means the correlation between the \( i \)'th hidden neuron with weights \( \mathbf{w}_i \) and the \( j \)'th hidden neuron with weights \( \mathbf{w}_j \). By eq. (11),
\[ G_{ij} = \sum_{i=1}^{D_y} \nu_f(c_i, h) \frac{\partial c_i^T h}{\partial \mathbf{w}_i} \frac{\partial c_i^T h}{\partial \mathbf{w}_j} = \sum_{i=1}^{D_y} \nu_f(c_i, h) \frac{\partial c_i h_i}{\partial \mathbf{w}_i} \frac{\partial c_j h_j}{\partial \mathbf{w}_j} \]
\[ = \sum_{i=1}^{D_y} \nu_f(c_i, h) c_i c_j \frac{\partial h_i}{\partial \mathbf{w}_i} \frac{\partial h_j}{\partial \mathbf{w}_j} = \sum_{i=1}^{D_y} \nu_f(c_i, h) c_i c_j (\nu_f(w_i, x) \tilde{x}) (\nu_f(w_j, x) \tilde{x}^T) \]
\[ = \sum_{i=1}^{D_y} c_i c_j \nu_f(c_i, h) \nu_f(w_i, x) \nu_f(w_j, x) \tilde{x} \tilde{x}^T. \]
3 Proof of Theorem 3

Proof. By assumption, the joint distribution $p(x, h)$ is in a factorable form. Therefore

$$\log p(x, h) = \sum_{l=1}^{L} \log p(h_l | \theta_l, r_l),$$

where $l = 1, \cdots, L$ is the index of subsystems, $h_l$ is the subsystem output, and $r_l$ is the reference of the subsystem. We have $\bigcup_{l=1}^{L} \{h_l\} = \{x, h\}$ and $\bigcup_{l=1}^{L} \{\theta_l\} = \{\Theta\}$. Therefore

$$E_p \left( - \frac{\partial^2}{\partial \theta_1 \partial \theta_1^\top} \log p(x, h) \right) = E_p \left( - \frac{\partial^2}{\partial \theta_1 \partial \theta_1^\top} \log p(h_1 | \theta_1, r_1) \right)$$

$$= E_{p(r_1)} \left( E_p(h_1 | r_1) \left( - \frac{\partial^2}{\partial \theta_1 \partial \theta_1^\top} \log p(h_1 | \theta_1, r_1) \right) \right)$$

$$= E_p \left( g^{h_1}(\theta_1) \right),$$

and

$$E_p \left( - \frac{\partial^2}{\partial \theta_1 \partial \theta_2^\top} \log p(x, h) \right) = 0 \quad (\forall l_1 \neq l_2).$$

Based on the Hessian expression of RFIM, $J(\Theta)$ is in a block-diagonal form, with each block given by $E_p \left( g^{h_l}(\theta_l) \right)$. \hfill \Box

4 Experimental Settings & Zoomed Learning Curves

The training/validation/testing sets have 50,000/10,000/10,000 images, respectively. Each sample is a gray scale image of size $28 \times 28$ (784 dimensional feature space) and is labeled as one of ten different classes. For all methods, the mini-batch size is fixed to 50 and the $L_2$ regularization strength is fixed to $10^{-3}$. For each optimizer, we try to find the best learning rate in the range $\{\cdots, 10^{-1}, 5 \times 10^{-2}, 10^{-2}, 5 \times 10^{-3}, 10^{-3}, \cdots\}$. On the tested architectures, a good learning rate configuration for RNGD is usually around $10^{-2}$ or $5 \times 10^{-3}$. The optimizers are in their default settings in TensorFlow 1.0. For the Adam optimizer, $\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}$. For RNGD, we set empirically $T = 100, \lambda = 0.005$ and $\omega = 1$. We use the Glorot uniform initializer to set the initial weights.

For each method and each learning rate configuration, we try 40 independent runs with different random seeds. Then, we select the best configuration based on the validation accuracy. Then, we plot the 40 learning curves as well as the average validation curve. The learning curves are obtained by evaluating the training error and validation accuracy after each epoch (one pass over all available training data).

See the following figs. (1–4) for the learning curves on four different architectures with relu activation units and $L_2$ regularization. Only the training curves and validation curves are shown for a clear presentation. The testing accuracy is close to the validation accuracy (run our codes to see the detailed results).
Figure 1: A MLP with shape 784–80–80–80–10.
Figure 2: A MLP with shape 784–80–80–80–10 and batch normalization after each hidden layer.
Figure 3: A MLP with shape 784–100–100–100–10.
Figure 4: A MLP with shape 784–100–100–100–10 and batch normalization after each hidden layer.