Tutorial on Multinomial Logistic Regression

Javier R. Movellan

June 19, 2013

1 General Model

The inputs are n-dimensional vectors the outputs are c-dimensional vectors. The training sample consist of m input output pairs. We organize the example inputs as an $m \times n$ matrix x. The corresponding example outputs are organized as a $m \times c$ matrix y. The models under consideration make predictions

$$\hat{y} = h(u) \tag{1}$$

$$u = x\theta \tag{2}$$

where θ is a $n \times c$ weight matrix. Note θ_{ij} can be seen as the connection strength from input variable X_i to output variable U_j. We evaluate the

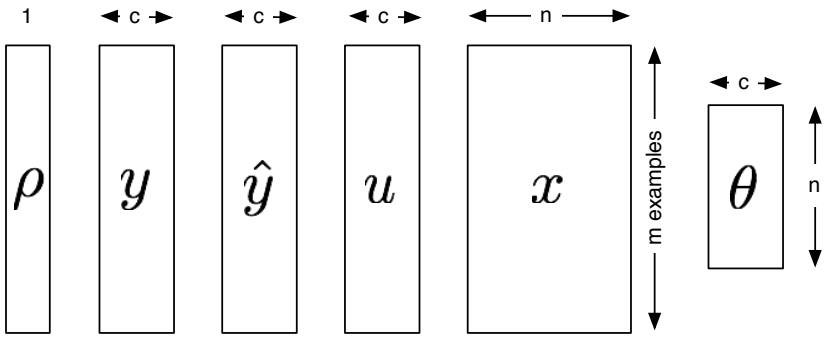

Figure 1: *There are m examples, n input variables and c output variables.*

optimality of $\hat{y}$, and thus of θ, using the following criterion

$$L(\theta) = \Phi(\theta) + \Gamma(\theta) \tag{3}$$

$$\Phi(\theta) = \sum_{j=1}^{m} w_j \rho_j \tag{4}$$

where $\Gamma(\theta)$ is a prior penalty function of the parameters,

$$\rho_j = f(y_{j\cdot}, \hat{y}_{j\cdot}) \tag{5}$$

measures the mismatch between the j^{th} rows of y and $\hat{y}$. The terms $w_1, \ldots, w_m$ are positive weights that capture the relative importance of each of the m input-output pairs.

Our goal is to find a matrix θ that minimizes L. A standard approach for minimizing L is the Newton-Raphson algorithm which calls for computation of the gradient and the Hessian matrix.

1.1 Gradient Vector

Let $\theta_{\cdot i} \in \mathbb{R}^n$ be the i^{th} column of θ, i.e., the set of connection strengths from the n input units to the i^{th} output unit. The overall gradient of Φ with respect to θ is as follows

$$\nabla_{\text{vec}[\theta]}\Phi = \begin{pmatrix} \nabla_{\theta_{\cdot 1}}\Phi \\ \nabla_{\theta_{\cdot 2}}\Phi \\ \vdots \\ \nabla_{\theta_{\cdot c}}\Phi \end{pmatrix} \tag{6}$$

where vec$[\theta]$ is the vectorized version of θ and the gradient of a vector q with respect to a vector p is defined as follows

$$\nabla_p q = \frac{\partial q'}{\partial p} \tag{7}$$

Using the chain rule for gradients we get

$$\frac{\partial \Phi'}{\partial \theta_{\cdot i}} = \frac{\partial u'_{\cdot i}}{\partial \theta_{\cdot i}} \frac{\partial \rho'}{\partial u_{\cdot i}} \frac{\partial \Phi'}{\partial \rho} \tag{8}$$

Note

$$u_{\cdot i} = (x\theta)_{\cdot i} = x\theta_{\cdot i} \tag{9}$$

Thus

$$\frac{\partial u'_{\cdot i}}{\partial \theta_{\cdot i}} = \frac{\partial \theta'_{\cdot i} x'}{\partial \theta_{\cdot i}} = x' \tag{10}$$

Moreover

$$\frac{\partial \rho'}{\partial u_{\cdot i}} = \begin{pmatrix} \frac{\partial \rho_1}{\partial u_{1i}} & \cdots & \frac{\partial \rho_m}{\partial u_{1i}} \\ \vdots & \vdots & \vdots \\ \frac{\partial \rho_1}{\partial u_{mi}} & \cdots & \frac{\partial \rho_m}{\partial u_{mi}} \end{pmatrix} = \begin{pmatrix} \frac{\partial \rho_1}{\partial u_{1i}} & 0 & \cdots & 0 \\ 0 & \frac{\partial \rho_2}{\partial u_{2i}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \frac{\partial \rho_m}{\partial u_{mi}} \end{pmatrix} \tag{11}$$

and

$$\frac{\partial \Phi'}{\partial \rho} = \begin{pmatrix} w_1 \\ \vdots \\ w_m \end{pmatrix} \tag{12}$$

Thus

$$\frac{\partial \Phi'}{\partial \theta_{\cdot i}} = x' \begin{pmatrix} \frac{\partial \rho_1}{\partial u_{1i}} & 0 & \cdots & 0 \\ 0 & \frac{\partial \rho_2}{\partial u_{2i}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \frac{\partial \rho_m}{\partial u_{mi}} \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_m \end{pmatrix} \tag{13}$$

Equivalently

$$\nabla_{\theta_{\cdot i}} \Phi = \frac{\partial \Phi'}{\partial \theta_{\cdot i}} = x' w \Psi_i \tag{14}$$

where

$$w = \begin{pmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & w_m \end{pmatrix} \tag{15}$$

and

$$\Psi_i = \begin{pmatrix} \frac{\partial \rho_1}{\partial u_{1i}} \\ \\ \frac{\partial \rho_2}{\partial u_{2i}} \\ \vdots \\ \frac{\partial \rho_m}{\partial u_{mi}} \end{pmatrix} \tag{16}$$

Thus

$$\nabla_{\mathrm{vec}[\theta]} \Phi = \begin{pmatrix} x' w \Psi_1 \\ \vdots \\ x' w \Psi_c \end{pmatrix} \tag{17}$$

1.2 Hessian Matrix

The Hessian of a scalar v with respect to a vector u is defined as follows

$$\nabla_u^2 v = \nabla_u \Big(\nabla_u v \Big) = \frac{\partial}{\partial u} \Big(\nabla_u v \Big)' \tag{18}$$

Thus

$$\nabla^2_{\text{vec}[\theta]} \Phi = \nabla_{\text{vec}[\theta]} \begin{pmatrix} x' w \Psi_1 \\ \vdots \\ x' w \Psi_c \end{pmatrix} \tag{19}$$

$$= \begin{pmatrix} \nabla_{\text{vec}[\theta]} x' w \Psi_1, & \cdots & , x' w \Psi_c \end{pmatrix} \tag{20}$$

where

$$\nabla_{\text{vec}[\theta]} x' w \Psi_i = \nabla_{\text{vec}[\theta]} \Psi_i \nabla_{\Psi_i} x' w \Psi_i (\nabla_{\text{vec}[\theta]} \Psi_i) wx \tag{21}$$

Thus

$$\nabla^2_{\text{vec}[\theta]} \Phi = \begin{pmatrix} (\nabla_{\text{vec}[\theta]} \Psi_1) wx, & \cdots , & (\nabla_{\text{vec}[\theta]} \Psi_c) wx \end{pmatrix} \tag{22}$$

$$\nabla^2_{\text{vec}[\theta]} \Phi = \begin{pmatrix} \begin{pmatrix} \nabla_{\theta_{\cdot 1}} \Psi_1 \\ \vdots \\ \nabla_{\theta_{\cdot c}} \Psi_1 \end{pmatrix} wx, & \cdots , & \begin{pmatrix} \nabla_{\theta_{\cdot 1}} \Psi_c \\ \vdots \\ \nabla_{\theta_{\cdot c}} \Psi_c \end{pmatrix} wx \end{pmatrix} \tag{23}$$

$$= \begin{pmatrix} (\nabla_{\theta_{\cdot 1}} \Psi_1) wx & \cdots & (\nabla_{\theta_{\cdot 1}} \Psi_c) wx \\ \vdots & \cdots & \vdots \\ (\nabla_{\theta_{\cdot c}} \Psi_1) wx & & (\nabla_{\theta_{\cdot c}} \Psi_c) wx \end{pmatrix} \tag{24}$$

Note

$$\nabla_{\theta_{\cdot i}} \Psi_j = \nabla_{\theta_{\cdot i}} u_{\cdot j} \nabla_{u_{\cdot j}} \Psi_j = x' \Lambda_{ij} \tag{25}$$

where

$$\Lambda_{ij} = \nabla_{u_{\cdot j}} \Psi_j = \begin{pmatrix} \frac{\partial \Psi_{1j}}{\partial u_{1i}} & 0 & \cdots & 0 \\ 0 & \frac{\partial \Psi_{2j}}{\partial u_{2i}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \frac{\partial \Psi_{mj}}{\partial u_{mi}} \end{pmatrix} \tag{26}$$

Thus

$$\nabla_{\theta_{\cdot i}}\Psi_j = \frac{\partial \Psi_j'}{\partial \theta_{\cdot i}} = \begin{pmatrix} x_{11}\frac{\partial \Psi_{1j}}{\partial u_{1i}} & \cdots & x_{m1}\frac{\partial \Psi_{mj}}{\partial u_{mi}} \\ \vdots & \vdots & \vdots \\ x_{1n}\frac{\partial \Psi_{1j}}{\partial u_{ni}} & \cdots & x_{mn}\frac{\partial \Psi_{mj}}{\partial u_{mi}} \end{pmatrix} = x'\Lambda_{ij} \tag{27}$$

where

$$\Lambda_{ij} = \begin{pmatrix} \frac{\partial \Psi_{1j}}{\partial u_{1i}} & 0 & \cdots & 0 \\ 0 & \frac{\partial \Psi_{2j}}{\partial u_{2i}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \frac{\partial \Psi_{mj}}{\partial u_{mi}} \end{pmatrix} \tag{28}$$

Thus

$$\nabla^2_{\mathrm{vec}[\theta]}\Phi = \begin{pmatrix} x'\Lambda_{11}wx & \cdots & x'\Lambda_{1c}wx \\ \vdots & \cdots & \vdots \\ x'\Lambda_{c1}wx & \cdots & x'\Lambda_{cc}wx \end{pmatrix} \tag{29}$$

2 Quadratic Priors

Let

$$\Gamma(\theta) = -\frac{1}{2}(\mathrm{vec}[\theta] - \mathrm{vec}[\mu])'\sigma^{-1}(\mathrm{vec}[\theta] - \mathrm{vec}[\mu])' \tag{30}$$

then

$$\nabla_{\mathrm{vec}[\theta]}\Gamma(\theta) = -\sigma^{-1}(\mathrm{vec}[\theta] - \mathrm{vec}[\mu]) \tag{31}$$

$$\nabla^2_{\mathrm{vec}[\theta]} = \nabla_{\mathrm{vec}[\theta]}\nabla_{\mathrm{vec}[\theta]} = -\sigma^{-1} \tag{32}$$

3 Newton-Raphson Optimization

$$\mathrm{vec}[\theta^{(k+1)}] = \mathrm{vec}[\theta^{(k)}] - \left(\nabla^2_{\mathrm{vec}[\theta]}L(\theta^{(k)}\right)^{-1}\nabla_{\mathrm{vec}[\theta]}L(\theta^{(k)}) \tag{33}$$

where $\theta^{(k)}$ is the value of θ after k iterations of the optimization algorithm.

4 Multivariate Linear Regression

In this case

$$\hat{y}_{i\cdot} = u_{i\cdot} \tag{34}$$

$$\rho_i = \sum_k (\hat{y}_{ik} - y_{ik})^2 \tag{35}$$

Thus

$$\Psi_i = \hat{y}_{\cdot i} - y_{\cdot i} \tag{36}$$

$$\Lambda_{ij} = I_m \tag{37}$$

where I_m is the $m \times m$ identity matrix. Thus

$$\frac{\partial}{\partial \theta_{\cdot i}} \frac{\partial \Phi}{\partial \theta'_{\cdot j}} = x' w x \tag{38}$$

5 Multinomial Logistic Regression

Let

$$\hat{y}_{ij} = f_j(u_{i\cdot}) = \frac{e^{u_{ij}}}{\sum_{k=1}^{c} e^{u_{ik}}} \tag{39}$$

and ρ_i the negative log-likelihood of the output vcector $y_{i\cdot}$ given the input vector $x_{i\cdot}$

$$\rho_i = -\sum_{k=1}^{c} y_{ik} \log \hat{y}_{ik} \tag{40}$$

Note

$$\frac{\partial \hat{y}_{ij}}{\partial u_{ik}} = \hat{y}_{ij} \frac{\partial \log \hat{y}_{ij}}{\partial u_{ik}} = \hat{y}_{ij}(\delta_{jk} - \hat{y}_{ik}) \tag{41}$$

Thus

$$\frac{\partial \rho_i}{\partial u_{ik}} = -\sum_{j=1}^{c} \frac{y_{ij}}{\hat{y}_{ij}} \frac{\partial \hat{y}_{ij}}{u_{ij}} = \sum_{j=1}^{c} y_{ij}(\delta_{jk} - \hat{y}_{ik}) = \hat{y}_{ik} - y_{ik} \tag{42}$$

where we used the fact that $\sum_j y_{ij} = 1$. Thus

$$\Psi_i = \begin{pmatrix} \frac{\partial \rho_1}{\partial u_{1i}} \\ \\ \frac{\partial \rho_2}{\partial u_{2i}} \\ \vdots \\ \frac{\partial \rho_m}{\partial u_{mi}} \end{pmatrix} = \begin{pmatrix} y_{1i} - \hat{y}_{1i} \\ \\ y_{2i} - \hat{y}_{2i} \\ \vdots \\ y_{mi} - \hat{y}_{mi} \end{pmatrix} = \hat{y}_{\cdot i} - y_{\cdot i} \tag{43}$$

Moreover

$$\Lambda_{ij} = \begin{pmatrix} \frac{\partial \Psi_{1j}}{\partial u_{1i}} & 0 & \cdots & 0 \\ 0 & \frac{\partial \Psi_{2j}}{\partial u_{2i}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \frac{\partial \Psi_{mj}}{\partial u_{mi}} \end{pmatrix} \tag{44}$$

where

$$\frac{\partial \Psi_{kj}}{\partial u_{ki}} = \frac{\partial \hat{y}_{kj}}{\partial u_{ki}} = \hat{y}_{kj}(\delta_{ij} - \hat{y}_{kj}) \tag{45}$$

5.1 Relationship to Linear Regression

Note that the gradient in multinomial logistic regression is identical to the gradient in multivariate linear regression.

$$\nabla_{\theta_{\cdot i}} \Phi = \hat{y}_{\cdot i} - y_{\cdot i} \tag{46}$$

The Hessians would are also very simmilar. In linear regression

$$\frac{\partial}{\partial \theta_{\cdot i}} \frac{\partial \Phi}{\partial \theta'_{\cdot j}} = x' w x \tag{47}$$

and in logistic regression

$$\frac{\partial}{\partial \theta_{\cdot i}} \frac{\partial \Phi}{\partial \theta'_{\cdot j}} = x' w \Lambda_{ij} x \tag{48}$$

which can be seen as special case of linear regression where the weight matrix w is substituted by the $w\Lambda_{ij}$ matrix.

5.2 Summary

- **Training Data**
  - **Inputs** $x \in \mathbb{R}^m \times R^n$. Rows are examples, columns are input variables.
  - **Outputs** $y \in \mathbb{R}^m \times R^c$. Rows are examples, columns are labels or label probabilities. All entries are non-negative and each row add up to 1.
  - **Example Weights** $w \in \mathbb{R}^m \times R^r$. Diagonal matrix. Each term is positive and represents the relative importance of each example.
- **Gradients**

$$\nabla_{\text{vec}[\theta]} L = \begin{pmatrix} \nabla_{\theta_{\cdot 1}} \Phi \\ \nabla_{\theta_{\cdot 2}} \Phi \\ \\ \vdots \\ \nabla_{\theta_{\cdot c}} \Phi \end{pmatrix} + \nabla_{\text{vec}[\theta]} \Gamma \tag{49}$$

where

$$\nabla_{\theta_{\cdot i}} \Phi = \frac{\partial \Phi'}{\partial \theta_{\cdot i}} = x' w \Psi_i \tag{50}$$

and

$$\Psi_i = \begin{pmatrix} \frac{\partial \rho_1}{\partial u_{1i}} \\ \\ \frac{\partial \rho_2}{\partial u_{2i}} \\ \vdots \\ \frac{\partial \rho_m}{\partial u_{mi}} \end{pmatrix} = \hat{y}_{\cdot i} - y_{\cdot i} \tag{51}$$

and

$$\nabla_{\text{vec}[\theta]} \Gamma = -\sigma^{-1}(\text{vec}[\theta] - \text{vec}[\mu]) \tag{52}$$

- **Hessian Matrices**

$$\nabla^2_{\text{vec}[\theta]} L = \nabla^2_{\text{vec}[\theta]} \Phi + \nabla^2_{\text{vec}[\theta]} \Gamma \tag{53}$$

$$\nabla^2_{\text{vec}[\theta]}\Phi = \begin{pmatrix} \frac{\partial}{\partial\theta_{\cdot 1}}\frac{\partial\Phi}{\partial\theta'_{\cdot 1}} & \cdots & \frac{\partial}{\partial\theta_{\cdot 1}}\frac{\partial\Phi}{\partial\theta'_{\cdot c}} \\ \vdots & \vdots & \vdots \\ \frac{\partial}{\partial\theta_{\cdot c}}\frac{\partial\Phi}{\partial\theta'_{\cdot 1}} & \cdots & \frac{\partial}{\partial\theta_{\cdot c}}\frac{\partial\Phi}{\partial\theta'_{\cdot c}} \end{pmatrix} \tag{54}$$

where

$$\frac{\partial}{\partial\theta_{\cdot i}}\frac{\partial\Phi}{\partial\theta'_{\cdot j}} = x'\Lambda_{ij}wx \tag{55}$$

where

$$\Lambda_{ij} = \begin{pmatrix} \frac{\partial\Psi_{1j}}{\partial u_{1i}} & 0 & \cdots & 0 \\ 0 & \frac{\partial\Psi_{2j}}{\partial u_{2i}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \frac{\partial\Psi_{mj}}{\partial u_{mi}} \end{pmatrix} \tag{56}$$

and

$$\frac{\partial\Psi_{kj}}{\partial u_{ki}} = \frac{\partial\hat{y}_{kj}}{\partial u_{ki}} = \hat{y}_{kj}(\delta_{ij} - \hat{y}_{kj}) \tag{57}$$

and

$$\nabla^2_{\text{vec}[\theta]}\Gamma = -\sigma^{-1} \tag{58}$$

- **Learning Rule (Newton-Raphson)**

$$\text{vec}[\theta^{(k+1)}] = \text{vec}[\theta^{(k)}] - \left(\nabla^2_{\text{vec}[\theta]}L(\theta^{(k)}\right)^{-1}\nabla_{\text{vec}[\theta]}L(\theta^{(k)}) \tag{59}$$

  where $\theta^{(k)}$ is the value of θ after k iterations of the learning rule.

6 Appendix: L-p priors

From a Bayesian point of view the Φ function can be seen as a log-likelihood term. At times it may be useful to add a log prior term over θ, also known as a regularization term. Most prior terms of interest have the following form

$$\Gamma = \alpha \sum_{i=1}^{n} \sum_{j=1}^{c} \frac{1}{p} \Big(h(\theta_{i,j}) \Big)^p \tag{60}$$

where α is a non-negative constant. If h is the absolute value function then Γ is the *L-p* norm of θ. Two popular options are $p = 2$ and $p = 1$. Note the absolute value function is not differentiable, and thus when p is odd, it useful to approximate it with a differentiable function. Here we will use the log hyperbolic cosine function.

$$h(x) = \frac{1}{\beta} \log(2\cosh(\beta x)) \tag{61}$$

where $\beta \geq 0$ is a gain parameter and

$$\cosh(x) = \frac{e^x + e^{-x}}{2} \tag{62}$$

Note

$$\lim_{\beta \to \infty} h(x) = \text{abs}(x) \tag{63}$$

$$h'(x) = \frac{dh(x)}{dx} = \frac{e^{\beta x} - e^{-\beta x}}{e^{\beta x} + e^{-\beta x}} = \tanh(x) \tag{64}$$

$$\lim_{\beta \to \infty} \frac{dh(x)}{dx} = \text{sign}(x) \tag{65}$$

and

$$h''(x) = \frac{d^2 h(x)}{dx^2} = \frac{d \tanh(\beta x)}{dx} = \beta \, (1 - \tanh^2(\beta x)) \tag{66}$$

6.1 Gradient and Hessian

$$\frac{\partial \Gamma}{\partial \theta_{ij}} = \alpha \Big(h^{p-1}(\theta_{ij}) \; h'(\theta_{ij}) \Big) \tag{67}$$

$$= \alpha \left((\frac{1}{\beta} \log(2\cosh(\beta x)))^{p-1} \; \tanh(x) \right) \tag{68}$$

The $nc \times nc$ Hessian matrix is diagonal. Each term in the diagonal has the following form

$$\frac{\partial^2 \Gamma}{\partial \theta_{ij}^2} = \alpha \Big((p-1) h(\theta_{ij})^{p-2} \; (h'(\theta_{ij}))^2 + h(\theta_{ij})^{p-1} h''(\theta_{ij}) \Big) \tag{69}$$

Thus

$$\frac{\partial^2 \Gamma}{\partial \theta_{ij}^2} = \alpha \Big((p-1) (\frac{1}{\beta} \log(2\cosh(\beta x)))^{p-2} \; (\tanh(\beta x))^2 \tag{70}$$

$$+ (\frac{1}{\beta} \log(2\cosh(\beta x)))^{p-1} \; \beta \; (1 - \tanh^2(\beta x)) \Big) \tag{71}$$

Note for $p = 1$, we get

$$\frac{\partial \Gamma}{\partial \theta_{ij}} = \alpha \frac{1}{\beta} \tanh(\beta \theta_{ij}) \tag{72}$$

$$\frac{\partial^2 \Gamma}{\partial \theta_{ij}^2} = \alpha \; \beta \; (1 - \tanh^2(\beta \theta_{ij})) \tag{73}$$

$$\tag{74}$$

and for $p = 1$, and $\beta \to \infty$ we get

$$\frac{\partial \Gamma}{\partial \theta_{ij}} = \alpha \; \theta_{ij} \tag{75}$$

$$\frac{\partial^2 \Gamma}{\partial \theta_{ij}^2} = \alpha \tag{76}$$

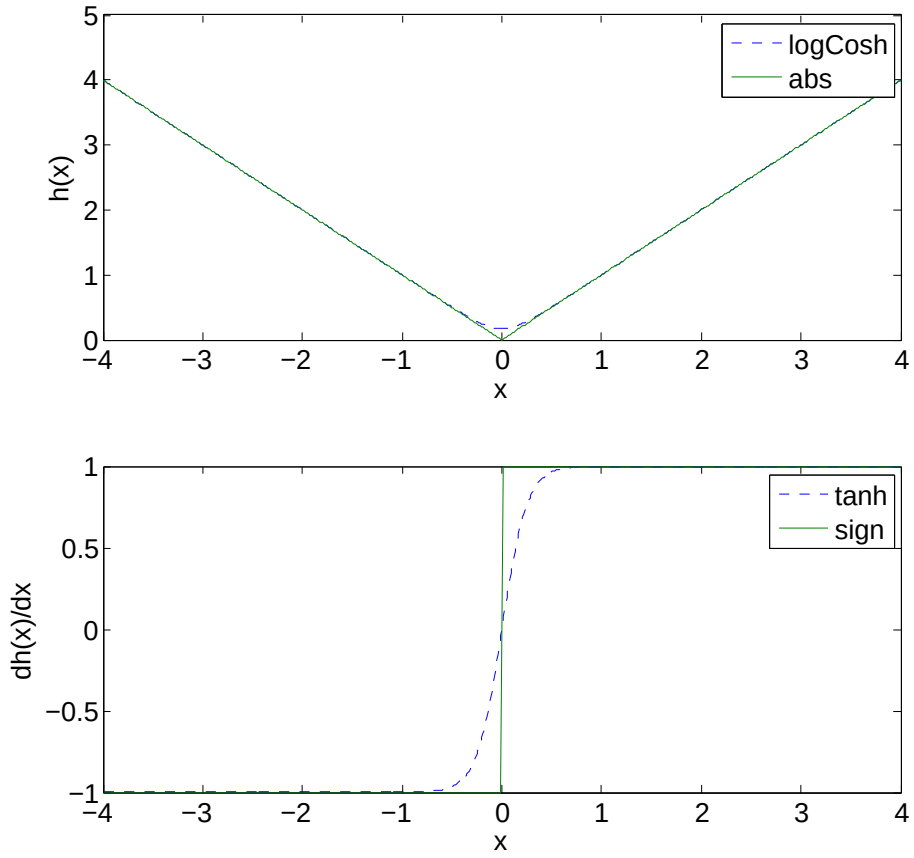

Figure 2: *The absolute value function can be approximated by the log of a hyperbolic cosine. The derivatie of the absolute value is the sign function. It can be approximated by the derivative of the log cosh function, which is the hyperbolic tangent. In this figure we used gain* $\beta = 4$.