The Equivalence of Logistic Regression and Maximum Entropy Modeling

Mingjie Qian*  

University of Illinois at Urbana-Champaign  

(Dated: October 26, 2013)  

Abstract  

In this technical note, we will show the equivalence of maximum entropy modeling and logistic regression. The equivalence is built on the fact that the optimization problem of logistic regression is to maximize the log-likelihood of model parameters knowing the exponential form of posterior probability functions, which is actually the dual problem of maximum entropy modeling. It is the maximum likelihood estimation (MLE) technique that bridges maximum entropy modeling and logistic regression.

*Electronic address: e-mail:mqian2@illinois.edu
I. MAXIMUM ENTROPY MODELING

The idea behind maximum entropy modeling [1] is to find a most uncertain distribution satisfying all known observations or balance equations. These balance equations take the form that the expectation of features under empirical distribution \( \tilde{p}(x, y) \) should be equal to the expectation of features under the modeled distribution \( p(x)p(y|x) \).

A. Primal Problem

\[
p^* (y|x) = \arg \max_{p \in C} - \sum_{x,y} \tilde{p}(x)p(y|x) \log p(y|x). \tag{1}
\]

where

\[
C \equiv \{ p \in \mathcal{P} | p(f_i) = \tilde{p}(f_i), \forall i \in \{1, 2, \ldots, n\} \} \tag{2}
\]

\[
p(f_i) = \sum_{x,y} \tilde{p}(x)p(y|x)f_i(x, y) \tag{3}
\]

\[
\tilde{p}(f_i) = \sum_{x,y} \tilde{p}(x, y)f_i(x, y). \tag{4}
\]

The Lagrange \( L(p, \Lambda, \gamma) \) is

\[
L(p, \Lambda, \gamma) = - \sum_{x,y} \tilde{p}(x)p(y|x) \log p(y|x)
- \sum_i \lambda_i \left[ \sum_{x,y} \tilde{p}(x, y)f_i(x, y) - \sum_{x,y} \tilde{p}(x)p(y|x)f_i(x, y) \right]
- \gamma \left( \sum_y p(y|x) - 1 \right).
\]

Set the derivative of \( L(p, \Lambda, \gamma) \) w.r.t. \( p(y|x) \) to zero, we have

\[
\frac{\partial L(p, \Lambda, \gamma)}{\partial p(y|x)} = -\tilde{p}(x) \left[ 1 + \log p^*(y|x) \right] + \sum_i \lambda_i \tilde{p}(x)f_i(x, y) - \gamma^* = 0.
\]

\[
\tilde{p}(x) \log p^*(y|x) = -\tilde{p}(x) - \gamma^* + \sum_i \lambda_i \tilde{p}(x)f_i(x, y).
\]

We thus have the analytical form of the optimizer

\[
p^*(y|x) = \exp \left\{ \sum_i \lambda_i f_i(x, y) \right\} \exp \left\{ \frac{-\tilde{p}(x) - \gamma^*}{\tilde{p}(x)} \right\}.
\]
The potential or the normalization is
\[
\exp\left\{-\frac{\hat{p}(x) - \gamma^*}{p(x)}\right\} = \frac{1}{Z_\Lambda(x)}
\]

The Lagrange dual function \( \mathcal{L}(p^*, \Lambda, \gamma^*) \) is
\[
\mathcal{L}(p^*, \Lambda, \gamma^*)
= - \sum_{x,y} \hat{p}(x) p^*(y|x) \log p^*(y|x) - \sum_i \lambda_i \left[ \sum_{x,y} \hat{p}(x,y) f_i(x,y) - \sum_{x,y} \hat{p}(x) p^*(y|x) f_i(x,y) \right]
= - \sum_{x,y} \hat{p}(x) p^*(y|x) \log p^*(y|x) + \sum_i \lambda_i \sum_{x,y} \hat{p}(x,y) f_i(x,y) - \sum_i \lambda_i \sum_{x,y} \hat{p}(x,y) f_i(x,y)
= - \sum_{x,y} p^*(y|x) \left[ \hat{p}(x) \log p^*(y|x) - \sum_i \lambda_i \hat{p}(x) f_i(x,y) \right] - \sum_i \lambda_i \sum_{x,y} \hat{p}(x,y) f_i(x,y)
= - \sum_x \left[ -\hat{p}(x) - \gamma^* \right] - \sum_{x,y} \hat{p}(x,y) \sum_i \lambda_i f_i(x,y).
\]

Note that
\[
\exp\left\{-\frac{\hat{p}(x) - \gamma^*}{p(x)}\right\} = \frac{1}{Z_\Lambda(x)} \Rightarrow -\hat{p}(x) - \gamma^* = -\hat{p}(x) \log Z_\Lambda(x),
\]

the Lagrange dual function can be simplified as
\[
\mathcal{L}(p^*, \Lambda, \gamma^*) = - \sum_x -\hat{p}(x) \log Z_\Lambda(x) - \sum_{x,y} \hat{p}(x,y) \sum_i \lambda_i f_i(x,y)
= - \left[ \sum_{x,y} \hat{p}(x,y) \sum_i \lambda_i f_i(x,y) - \sum_x \hat{p}(x) \log Z_\Lambda(x) \right]
= - \left[ \sum_{x,y} \hat{p}(x,y) \log \exp \frac{\sum \lambda_i f_i(x,y)}{Z_\Lambda(x)} - \sum_{x,y} \hat{p}(x,y) \log Z_\Lambda(x) \right]
= - \left[ \sum_{x,y} \hat{p}(x,y) \log \frac{\exp \sum \lambda_i f_i(x,y)}{Z_\Lambda(x)} \right]
= - \sum_{x,y} \hat{p}(x,y) \log p^*(y|x).
\]

Since \(-\sum_{x,y} \hat{p}(x) p(y|x) \log p(y|x)\) is concave then the Lagrange dual function
\[
\mathcal{L}(p^*, \Lambda, \gamma^*) = \sup_{p, \gamma} \mathcal{L}(p, \Lambda, \gamma) \text{ ia bound to be convex w.r.t. } \Lambda.
\]
B. Dual Problem

The dual problem of maximum entropy modeling is an unconstrained optimization problem

\[
\Lambda = \arg \min_\Lambda \mathcal{L}(p^*, \Lambda, \gamma^*) = \arg \min_\Lambda - \sum_{x,y} \tilde{p}(x, y) \log p^*(y|x). \tag{5}
\]

C. Relation to Maximum Likelihood

The log-likelihood of probabilistic function \( p_\Lambda (y|x) \) can be written as

\[
L_\tilde{p}(p_\Lambda (y|x)) = \sum_{x,y} \tilde{p}(x, y) \log p_\Lambda (y|x).
\]

Knowing \( p_\Lambda (y|x) = \frac{\exp^i \sum \lambda_i f_i(x,y)}{Z_\Lambda (x)} \), we can use ML (maximum likelihood) to estimate the model parameters:

\[
\Lambda = \arg \max_\Lambda L_\tilde{p}(p_\Lambda (y|x)) = \arg \max_\Lambda \sum_{x,y} \tilde{p}(x, y) \log \frac{\exp^i \sum \lambda_i f_i(x,y)}{Z_\Lambda (x)}
\]

\[
= \arg \max_\Lambda \sum_{x,y} \tilde{p}(x, y) \sum_i \lambda_i f_i(x,y) - \sum_{x,y} \tilde{p}(x, y) \log \sum_{y'} \exp^i \sum \lambda_i f_i(x,y')
\]

\[
= \arg \max_\Lambda \sum_{x,y} \tilde{p}(x, y) \sum_i \lambda_i f_i(x,y) - \sum_{x} \tilde{p}(x) \log \sum_{y'} \exp^i \sum \lambda_i f_i(x,y').
\]

D. Concavity of the Log-likelihood

Now we show the concavity of the log-likelihood of probabilistic function \( p(y|x) \).

\[
L_\tilde{p}(p_\Lambda (y|x)) = \sum_{x,y} \tilde{p}(x, y) \log \frac{\exp^i \sum \lambda_i f_i(x,y)}{Z_\Lambda (x)}
\]

\[
= \sum_{x,y} \tilde{p}(x, y) \sum_i \lambda_i f_i(x,y) - \sum_{x,y} \tilde{p}(x, y) \log \sum_{y'} \exp^i \sum \lambda_i f_i(x,y')
\]

\[
= \sum_{x,y} \tilde{p}(x, y) \sum_i \lambda_i f_i(x,y) - \sum_{x} \tilde{p}(x) \log \sum_{y'} \exp^i \sum \lambda_i f_i(x,y')
\]

\[
= \sum_{n,k} \tilde{p}(x_n, k) \sum_i \lambda_i f_i(x_n, k) - \sum_{n} \tilde{p}(x_n) \log \sum_{k} \exp^i \sum \lambda_i f_i(x_n, k).
\]
Define the feature matrix for \( x_n \) by \( F_{ik} (x_n) \triangleq f_i (x_n, k) \),

\[
L\bar{p} (p_{\lambda} (y|x)) = \sum_{n,k} \bar{p} (x_n, k) \sum_i \lambda_i F_{ik} (x_n) - \sum_n \bar{p} (x_n) \log \sum_k \exp \sum_i \lambda_i F_{ik} (x_n)
\]

\[
= \sum_{n,k} \bar{p} (x_n, k) [F_{ik} (x_n)]^T \lambda - \sum_n \bar{p} (x_n) \log \sum_k \exp [F_{ik} (x_n)]^T \lambda
\]

Let \( g(y^{x_n}) = \log \sum_k \exp (y^{x_n}) \) and \( y^{x_n} (\lambda) = F^T (x_n) \lambda \), then

\[
\log \sum_k \exp [F_{ik} (x_n)]^T \lambda = g(F^T (x_n) \lambda).
\]

since \( g(y^{x_n}) \) is convex, so is its composition with affine function \( g(F^T (x_n) \lambda) \).

In order to get the optimal parameter \( \lambda \), we need to calculate the gradient or sometimes the Hessian matrix of the log-likelihood function (i.e., Newton-Raphson method is utilized). We first derive the gradient and Hessian of the basic log-sum-exp function.

Let \( z(x) = \exp [x] \in \mathcal{R}^{d \times 1} \), we can write the log-sum-exp function as

\[
g(x) = \log \sum_{i=1}^d e^{x(i)} = \log 1^T z.
\]

The first derivative of \( g(x) \) w.r.t. \( x \) is

\[
Dg(x) = D \left[ \log (1^T z (x)) \right] D_x \left[ 1^T z (x) \right] D_x [z (x)] = \frac{1}{1^T z (x)} 1^T diag(z (x)) = \frac{z(x)^T}{1^T z (x)},
\]

by which we obtain the gradient of \( g(x) \) w.r.t. \( x \) as

\[
\nabla g(x) = \frac{z (x)}{1^T z (x)}.
\]

The Hessian of \( g(x) \) w.r.t. \( x \) is

\[
Hg(x) = D \nabla g(x)
\]

\[
= D_x \left[ \frac{z (x)}{1^T z (x)} \right]
\]

\[
= \frac{D_x [z (x)] \left( 1^T z (x) \right) - z (x) D_x \left[ 1^T z (x) \right]}{(1^T z (x))^2}
\]

\[
= \frac{\text{diag}(z (x)) 1^T z (x) - z (x) z (x)^T}{(1^T z (x))^2}.
\]

Thus we have the gradient and Hessian of \( g(F^T (x_n) \lambda) \):

\[
D \left[ g \circ F^T (x_n) \left( \lambda \right) \right] = D \left[ g \left( F^T (x_n) \lambda \right) \right] D \left[ F^T (x_n) \lambda \right] = \frac{z(F^T (x_n) \lambda)^T}{1^T z (F^T (x_n) \lambda)} F^T (x_n),
\]
i.e.,
\[
\nabla g \circ F^T (x_n) (\lambda) = F (x_n) \frac{z (F^T (x_n) \lambda)}{1^T z (F^T (x_n) \lambda)}.
\]

\[
Hg \circ F^T (x_n) (\lambda) = D\nabla g \circ F^T (x_n) (\lambda)
\]
\[
= D_{\lambda} \left[ F (x_n) \frac{z (F^T (x_n) \lambda)}{1^T z (F^T (x_n) \lambda)} \right]
\]
\[
= F (x_n) D_{\lambda} \left[ \frac{z (F^T (x_n) \lambda)}{1^T z (F^T (x_n) \lambda)} \right] D_{\lambda} [F^T (x_n) \lambda]
\]
\[
= F (x_n) Hg (\mu = F^T (x_n) \lambda) F^T (x_n)
\]
\[
= F (x_n) \text{diag} (\nu (\lambda)) \left[ \frac{z (\mu)}{1^T z (\mu)} \right] \frac{1^T z (\mu)}{(1^T z (\mu))^2} |_{\mu = F^T (x_n) \lambda} F^T (x_n).\]

We can also derive the Hessian of \( g (F^T (x_n) \lambda) \) by applying the chain rule for composition with an affine function:
\[
Hg (Ax) = D_x \nabla g (Ax) = D_x D_x g (Ax)^T = D_x [D_{\mu = Ax} g (\mu) D_x [Ax]]^T = D_x [A^T \nabla g (\mu)]
\]
\[
= A^T D_x \nabla g (\mu) = A^T D_{\mu = Ax} \nabla g (\mu) D_x [Ax] = A^T Hg (\mu) |_{\mu = Ax A}.
\]

Define \( z_n = \exp [F^T (x_n) \lambda] \in \mathcal{R}^{K \times 1} \) and \( \frac{z_n}{1^T z_n} = s_n (x_n) \), the gradient of \( L_\tilde{p} (p_\lambda (y|x)) \) can be written as
\[
\frac{\partial L_\tilde{p} (p_\lambda (y|x))}{\partial \lambda} = \sum_{n,k} \tilde{p} (x_n, k) F_{jk} (x_n) - \sum_{n} \tilde{p} (x_n) \nabla g \circ F^T (x_n) (\lambda)
\]
\[
= \sum_{n,k} \tilde{p} (x_n, k) F_{jk} (x_n) - \sum_{n} \tilde{p} (x_n) F (x_n) \frac{z (F^T (x_n) \lambda)}{1^T z (F^T (x_n) \lambda)}
\]
\[
= \sum_{n,k} \tilde{p} (x_n, k) F_{jk} (x_n) - \sum_{n} \tilde{p} (x_n) F (x_n) \frac{z_n}{1^T z_n}
\]
\[
= \sum_{n,k} \tilde{p} (x_n, k) F_{jk} (x_n) - \sum_{n} \tilde{p} (x_n) F (x_n) s_n (x_n)
\]
\[
= \frac{1}{N} \sum_{n=1}^N F_{\tilde{y} n} (x_n) - \frac{1}{N} \sum_{n=1}^N E_{p_\lambda (y|x_n)} F_{\tilde{y} n (x_n)},
\]

6
and the Hessian matrix of $L_\tilde{p} (p_\Lambda (y|x))$ can be derived as

$$H (\lambda) = -\sum_{n=1}^{N} \tilde{p} (x_n) \frac{F (x_n)}{(1^T z_n)^2} \left[ 1^T z_n \text{diag} (z_n) - z_n z_n^T \right] F^T (x_n)$$

$$= -\sum_{n=1}^{N} \tilde{p} (x_n) F (x_n) \left[ \text{diag} \left( \frac{z_n}{1^T z_n} \right) - \frac{z_n z_n^T}{1^T z_n 1^T z_n} \right] F^T (x_n)$$

$$= -\sum_{n=1}^{N} \tilde{p} (x_n) F (x_n) \left[ \text{diag} (s_\lambda (x_n)) - s_\lambda (x_n) s_\lambda^T (x_n) \right] F^T (x_n)$$

$$= -\frac{1}{N} \sum_{n=1}^{N} \left\{ \sum_{k=1}^{K} s_\lambda (x_n) F_{:k} (x_n) F_{:k} (x_n)^T \right\} - F (x_n) s_\lambda (x_n) \left[ F (x_n) s_\lambda (x_n) \right]^T$$

$$= -\frac{1}{N} \sum_{n=1}^{N} \left\{ E_{p_\lambda (y|x_n)} \left[ F_{:y} (x_n) F_{:y} (x_n)^T \right] - E_{p_\lambda (y|x_n)} F_{:y} (x_n) E_{p_\lambda (y|x_n)} F_{:y} (x_n)^T \right\}$$

$$= -\frac{1}{N} \sum_{n=1}^{N} \left\{ E_{p_\lambda (y|x_n)} \left[ F_{:y} (x_n) \times F_{:y} (x_n) \right] - E_{p_\lambda (y|x_n)} F_{:y} (x_n) \times E_{p_\lambda (y|x_n)} F_{:y} (x_n) \right\},$$

which can be easily shown to be negative semi-definite.

Set the derivative of the log-likelihood function w.r.t. $\Lambda$ to be zero,

$$\frac{\partial L_\tilde{p} (p_\Lambda (y|x))}{\partial \Lambda} = \sum_{x,y} \tilde{p} (x, y) f_i (x, y) - \sum_{x} \tilde{p} (x) \frac{1}{\sum_{y'} \exp \sum_{i} \lambda_i f_i (x, y')} \sum_{y'} \exp \sum_{i} \lambda_i f_i (x, y') f_i (x, y') = 0,$$

we thus have the balance equations

$$\sum_{x, y} \tilde{p} (x, y) f_i (x, y) = \sum_{x} \sum_{y'} \tilde{p} (x) p (y'|x) f_i (x, y').$$

According to Eq.(5), the Lagrange dual function of maximum entropy model is the negative log-likelihood

$$\mathcal{L} (p^*, \Lambda, \gamma^*) = -L_\tilde{p} (p_\Lambda (y|x)).$$

Some comments: We usually have two systems. In the first system, we know the analytical form of the solution function, and we aim to optimize the objective function. For the second system, we didn’t know the form of the solution, but instead we have a collection of constraints (e.g., the balance equations). We seek to optimize the objective function under the set of constraints. System 2 provides the primal problems, and system 1 plays the role of dual problems. Since system 1 is an unconstrained optimization problem, it is easier to solve.

In the next section, we will show that maximum likelihood given $p_\Lambda (y|x) = \frac{\exp \sum_{i} \lambda_i f_i (x, y)}{Z_\Lambda (x)}$ is actually the basic idea of multi-class logistic regression.
II. LOGISTIC REGRESSION

For multi-class logistic regression [2], the posterior probabilities are given by a softmax transformation of linear functions of the feature variables

\[ p(C_k | \phi) = y_k(\phi) = \frac{\exp(a_k)}{\sum_j \exp(a_j)} \]

where the activations \( a_k \) are given by

\[ a_k = w_k^T \phi. \]

The derivatives of \( y_k \) w.r.t. all of the activations \( a_j \), which are given by

\[ \frac{\partial y_k}{\partial a_k} = \frac{\exp(a_k) \sum_j \exp(a_j) - \exp(a_k) \exp(a_k)}{\left( \sum_j \exp(a_j) \right)^2} = \frac{\exp(a_k) \left( \sum_j \exp(a_j) - \exp(a_k) \right)}{\left( \sum_j \exp(a_j) \right)^2} = y_k (1 - y_k) \]

\[ \frac{\partial y_k}{\partial a_{j \neq k}} = -\frac{\exp(a_k) \exp(a_j)}{\left( \sum_j \exp(a_j) \right)^2} = -y_k y_j \]

or

\[ \frac{\partial y_k}{\partial a_j} = y_k (I_{kj} - y_j). \]

Define the target vector \( t_n \) for a feature vector \( \phi_n = \phi(x_n) \) belonging to class \( C_k \) is a binary vector with all elements zero except for element \( k \), which equals one. The likelihood function is then given by

\[ p(T | w_1, w_2, \cdots, w_K) = \prod_{n=1}^{N} \prod_{k=1}^{K} p(C_k | \phi_n)^{t_{nk}} = \prod_{n=1}^{N} \prod_{k=1}^{K} y_{nk}^{t_{nk}}, \]

where \( y_{nk} = y_k(\phi_n) \) and

\[ t_{nk} = \begin{cases} 1, & \text{if } x_n \in C_k; \sum_{k=1}^{K} t_{nk} = 1. \\ 0, & \text{otherwise} \end{cases} \]

Logistic regression aims to minimize the cross-entropy error function defined by

\[ E(W) = -\ln p(T | w_1, w_2, \cdots, w_K) = -\sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \ln y_{nk}. \]
The gradient of $E(W)$ w.r.t. $w_i$ is

$$\frac{\partial E(W)}{\partial w_i} = -\sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \frac{1}{y_{nk}} (I_{ki} - y_{ni}) \phi(x_n) = -\sum_{n=1}^{N} t_{nk} (I_{ki} - y_{ni}) \phi(x_n)$$

$$= -\sum_{n=1}^{N} t_{nk} I_{ki} \phi(x_n) + \sum_{n=1}^{N} t_{nk} y_{ni} \phi(x_n) = -\sum_{n=1}^{N} t_{ni} \phi(x_n) + \sum_{n=1}^{N} y_{ni} \phi(x_n)$$

$$= \sum_{n=1}^{N} (y_{ni} - t_{ni}) \phi(x_n) = [X(Y - T)]_{1i},$$

$$\therefore \frac{\partial E(W)}{\partial W} = X(Y - T), \text{ where } X = \left[ \phi(x_1) \phi(x_2) \cdots \phi(x_N) \right].$$

Set the gradient of $E(W)$ w.r.t. $w_i$ to be zero,

$$\frac{\partial E(W)}{\partial w_i} = \sum_{n=1}^{N} (y_{ni} - t_{ni}) \phi(x_n) = 0,$$

we thus obtain the set of balance equations which have already been shown in Eq.(6)

$$\sum_{n=1}^{N} y_{ni} \phi(x_n) = \sum_{n=1}^{N} t_{ni} \phi(x_n), \quad i = 1, 2, \cdots, K. \quad (7)$$

The Hessian matrix that comprises blocks of size $K \times K$ in which block $i, j$ is given by

$$\frac{\partial^2 E(W)}{\partial w_j \partial w_i} = \sum_{n=1}^{N} \frac{\partial \phi(x_n)}{\partial w_j} \frac{\partial y_{ni}}{\partial w_i}$$

$$= \sum_{n=1}^{N} \frac{\partial \phi(x_n)}{\partial y_{ni}} \frac{\partial y_{ni}}{\partial w_i} \frac{\partial a_{nj}}{\partial w_j}$$

$$= \sum_{n=1}^{N} \phi(x_n) y_{ni} (I_{ij} - y_{nj}) \phi^T(x_n)$$

$$= \sum_{n=1}^{N} y_{ni} (I_{ij} - y_{nj}) \phi(x_n) \phi^T(x_n)$$

$$= \sum_{n=1}^{N} y_{ni} I_{ij} \phi(x_n) \phi^T(x_n) - \sum_{n=1}^{N} y_{ni} y_{nj} \phi(x_n) \phi^T(x_n).$$

Define the $n$-th part of the Hessian matrix

$$H_n = \left[ \begin{array}{c} y_{n1} \\ y_{n2} \\ \vdots \\ y_{nK} \end{array} \right] \otimes \phi(x_n) \phi^T(x_n),$$

9
and

\[ Y_n = \begin{bmatrix} y_{n1} \\ y_{n2} \\ \vdots \\ y_{nK} \end{bmatrix} - \begin{bmatrix} y_{n1} \\ y_{n2} \\ \vdots \\ y_{nK} \end{bmatrix}, \]

we have

\[ H_n = Y_n \otimes \phi (x_n) \phi^T (x_n). \]

Leveraging the equation \( \sum_{i=1}^{K} y_{ni} = 1 \), for \( \forall v \in \mathcal{R}^K \),

\[ v^T Y_n v = \sum_{i=1}^{K} v_i^2 y_{ni} - \left( \sum_{i=1}^{K} v_i y_{ni} \right)^2 = \sum_{i=1}^{K} y_{ni} \sum_{i=1}^{K} v_i^2 y_{ni} - \left( \sum_{i=1}^{K} v_i y_{ni} \right)^2 \geq 0, \]

we thus have,

\[ Y_n \succeq 0 \]

and

\[ H_n \succeq 0, \]

therefore,

\[ H (W) = \sum_{n=1}^{N} H_n \succeq 0. \]

Thus the cross-entropy error function \( E (W) \) is convex w.r.t. \( W \) and therefore has a unique minimum (but may have multiple optimizers).

A. The equivalence of logistic regression and maximum entropy modeling

The log-likelihood function for logistic regression is

\[ L (W) = \ln p (T | w_1, w_2, \ldots, w_K) = -E (W) = \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \ln y_{nk}. \]

We rewrite it as

\[ L (W) = \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \ln y_{nk} \]

\[ \propto \sum_{n=1}^{N} \sum_{k=1}^{K} \tilde{p} (X = x_n, Y = k) \ln p (Y = k | X = x_n) \]

\[ = \sum_{x} \sum_{y} \tilde{p} (x, y) \log p (y | x) \]
Thus the multi-class logistic regression is equivalent to using ML (Maximum Likelihood) to estimate the model parameters of \( p(y = k|x) \) given its form as

\[
p(y = k|x) = y_k(\phi(x)) = \frac{\exp(w_k^T\phi(x))}{\sum_j \exp(w_j^T\phi(x))},
\]

which is actually the dual problem of maximum entropy modeling.

For maximum entropy modeling,

\[
p_{\Lambda}(y|x) = \frac{\exp{\sum_i \lambda_i f_i(x, y)}}{Z_{\Lambda}(x)}.
\]

Define \( F^n_{ik} \triangleq f_i(x_n, k) \), we have

\[
p_{\Lambda}^{\text{ME}}(y = k|x_n) = \frac{1}{Z_{\Lambda}(x_n)} \exp{\left(\lambda^T F^n_{ik}\right)}.
\]

For logistic regression,

\[
p_{\text{LR}}^W(y = k|x_n) = \frac{\exp\left(w_k^T\phi(x_n)\right)}{\sum_j \exp\left(w_j^T\phi(x_n)\right)}.
\]

If we introduce \( \tilde{\lambda} = \text{vec}(W) \), i.e.,

\[
\tilde{\lambda} = \begin{bmatrix} w_1^T, w_2^T, \ldots, w_K^T \end{bmatrix}^T,
\]

and \( \tilde{F}^n_{ik} = e_k \otimes \phi(x_n) \), i.e.,

\[
\tilde{F}^n = \begin{bmatrix} \phi(x_n) \\ \phi(x_n) \\ \vdots \\ \phi(x_n) \end{bmatrix} = I_K \otimes \phi(x_n),
\]

then

\[
p_{\text{LR}}^W(y = k|x_n) = \frac{1}{\sum_j \exp{\left(\tilde{\lambda}^T \tilde{F}^n_{jk}\right)}} \exp{\left(\tilde{\lambda}^T \tilde{F}^n_{ik}\right)} = \frac{1}{Z_{\Lambda}(x_n)} \exp{\left(\tilde{\lambda}^T \tilde{F}^n_{ik}\right)},
\]

which shares the same form of maximum entropy modeling.

The above suggests that logistic regression can be viewed as a special form of maximum entropy modeling by setting the feature matrix w.r.t. \( x_n \) to be \( \tilde{F}^n = I_K \otimes \phi(x_n) \). It also tells us that there are two formulations for multi-class logistic regression, one is the original form, the other is the special form of maximum entropy modeling with \( F^n = I_K \otimes \phi(x_n) \). Therefore
multi-class logistic regression can be solved by either using its original form or casting its parameter estimation problem to that of maximum entropy modeling with \( \hat{F}^n = I_K \otimes \phi (x_n) \).

Now let’s see how a general maximum entropy modeling can be used to derive the balance equation for multi-class logistic regression. Recall that \( \hat{F}^n_k = e_k \otimes \phi (x_n) \),

\[
\frac{\partial L_{\hat{p}} (p^{LR}_{W} (y|x))}{\partial \lambda} = \sum_{n,k} \hat{p} (x_n, k) \hat{F}_{jk} (x_n) - \sum_n \hat{p} (x_n) \nabla g \circ \hat{F}^T (x_n) (\hat{\lambda})
\]

\[
= \sum_{n,k} \hat{p} (x_n, k) \hat{F}_{jk} (x_n) - \sum_n \hat{p} (x_n) \hat{F} (x_n) \frac{z (\hat{F}^T (x_n) \hat{\lambda})}{1^T z (\hat{F}^T (x_n) \hat{\lambda})}
\]

\[
= \sum_{n,k} \hat{p} (x_n, k) \hat{F}_{jk} (x_n) - \sum_n \hat{p} (x_n) \hat{F} (x_n) s_{\hat{\lambda}} (x_n)
\]

\[
= \frac{1}{N} \sum_{n=1}^N \sum_{k=1}^K t_{nk} \hat{F}^n_{jk} - \frac{1}{N} \sum_{n=1}^N \sum_{k=1}^K y_{nk} \hat{F}^n_{jk}
\]

\[
= \frac{1}{N} \sum_{n=1}^N \sum_{k=1}^K t_{nk} e_k \otimes \phi (x_n) - \frac{1}{N} \sum_{n=1}^N \sum_{k=1}^K y_{nk} e_k \otimes \phi (x_n)
\]

\[
= \frac{1}{N} \sum_{n=1}^N \sum_{k=1}^K e_k \otimes t_{nk} \phi (x_n) - \frac{1}{N} \sum_{n=1}^N \sum_{k=1}^K e_k \otimes y_{nk} \phi (x_n)
\]

\[
= \frac{1}{N} \sum_{k=1}^K e_k \otimes \left[ \sum_{n=1}^N t_{nk} \phi (x_n) - \sum_{n=1}^N y_{nk} \phi (x_n) \right],
\]

we thus have

\[
\frac{\partial L_{\hat{p}} (p^{LR}_{W} (y|x))}{\partial \lambda} = 0 \Rightarrow \sum_{n=1}^N t_{nk} \phi (x_n) = \sum_{n=1}^N y_{nk} \phi (x_n), \quad k = 1, 2, \ldots, K,
\]

which is exactly the balance equations for logistic regression in Eq.(7).

Even without deriving the first order derivative of \( E (W) \), we know that Eq.(7) must be held. Since the dual problem of maximum entropy modeling with balance equations is MLE given that \( p (y = k|x) \) has an exponential form, logistic regression must satisfy these balance equations.
III. CONCLUSION

The equivalence between maximum entropy modeling and logistic regression can help us find alternatives to solve maximum entropy modeling by techniques solving logistic regression such as iterative re-weighted least squares. It also helps us understand more deeply on the essence of logistic regression.
