The Geometric Structure of Generalized Softmax Learning

Xiangxiang Xu*, Shao-Lun Huang†, Lizhong Zheng‡ and Lin Zhang†
* Dept. of Electronic Engineering, Tsinghua University, Beijing 100084, China
† Tsinghua-Berkeley Shenzhen Institute, Tsinghua University, Shenzhen 518055, China
‡ Dept. of EECS, Massachusetts Institute of Technology, Cambridge, MA 02139-4307, USA
Email: *xuxx14@mails.tsinghua.edu.cn, †{shaolun.huang, linzhang}@sz.tsinghua.edu.cn, ‡lizhong@mit.edu

***Abstract*—In this paper, we formulate the generalized softmax learning (GSL) problem, as a symmetric extension of the softmax regression problem. We further study the geometric structure of GSL and demonstrate the equivalence of GSL and the original softmax regression problem. Besides, this geometric structure indicates the symmetry between a neural network and its reverse network, and the symmetric roles of the weights and feature in a neural network. Finally, we present a numerical simulation to verify these symmetry properties in neural networks.**

I. Introduction

Given a pair of discrete random variables X, Y with the joint distribution $P_{X,Y}$, we study the *generalized softmax learning* (GSL) problem of approximating the distribution $P_{X,Y}$ by the exponential family of the form:

$$Q_{X,Y}(x,y) = \frac{e^{f^{\mathrm{T}}(x)g(y)+a(x)+b(y)}}{\sum_{(x',y')\in\mathcal{X}\times\mathcal{Y}} e^{f^{\mathrm{T}}(x')g(y')+a(x')+b(y')}}, \quad (1)$$

where $f\colon \mathcal{X}\to\mathbb{R}^k$, $g\colon \mathcal{Y}\to\mathbb{R}^k$ are k-dimensional functions, and $a\colon \mathcal{X}\to\mathbb{R}$, $b\colon \mathcal{Y}\to\mathbb{R}$ are scalar functions of X, Y, respectively. These functions are the parameters to be designed to minimize the K-L divergence between $P_{X,Y}$ and $Q_{X,Y}$, which can be expressed by an M-projection problem:

$$\mathcal{M}_{\mathcal{E}_k}(P_{X,Y}) \triangleq \operatorname*{arg\,min}_{Q_{X,Y}\in\mathcal{E}_k} D(P_{X,Y}\|Q_{X,Y}), \quad (2)$$

where $\mathcal{E}_k$ is the family of distributions of the form (1). Our goal in this paper is to investigate the properties of $\mathcal{M}_{\mathcal{E}_k}(P_{X,Y})$ and its applications in machine learning.

To see how (2) can be applied to machine learning, recall that in the original softmax regression, the discriminant model[1]

$$\tilde{P}_{Y|X}(y|x) = \frac{e^{f^{\mathrm{T}}(x)g(y)+b(y)}}{\sum_{y'\in\mathcal{Y}} e^{f^{\mathrm{T}}(x)g(y')+b(y')}} \quad (3)$$

is used to predict y from x, where the feature function $f(x)\in\mathbb{R}^k$, weights $g(y)\in\mathbb{R}^k$, and bias $b(y)\in\mathbb{R}$ are the parameters for training, so that the K-L divergence between $P_{X,Y}$ and $P_X\tilde{P}_{Y|X}$ is minimized. Note that the softmax function (3) is not symmetric to X, Y. Therefore, it is in general unclear whether or not the trained parameters for predicting Y from X can also be used in the symmetric problem of predicting X from Y. In this paper, we reveal the symmetry implied in softmax regression with the developed geometric structure of GSL. In particular, we show that the optimal solutions of the original softmax regression problem coincide with the solutions of GSL (2), and such optimal solutions, due to the symmetry form of (1), are symmetric to both X and Y. As a result, the optimal feature function $f(x)$ for predicting Y from X is precisely the optimal weights for predicting X from Y in the symmetric softmax regression problem.

The rest of this paper is organized as follows. We first introduce the notations in Section II, then explore the existence and non-uniqueness of the solutions of GSL (2) in Section III. In Section IV, we develop a geometric structure of (2), which leads to a Pythagorean theorem. With this structure, we show that the solutions of (2) coincide with the solutions of the original softmax regression problem and further demonstrate the symmetric roles of feature and weights in a neural network. Furthermore, Section V illustrates the theoretical results through a numerical simulation. Finally, Section VI presents some proof details, and Section VII concludes the paper.

II. Problem Formulation

In this paper, the alphabets $\mathcal{X}$ and $\mathcal{Y}$ are both finite sets with $\mathcal{X}=\{1,2,\ldots,|\mathcal{X}|\}$ and $\mathcal{Y}=\{1,2,\ldots,|\mathcal{Y}|\}$. We use subscripts to distinguish the joint distribution and marginal distributions, e.g., the marginal distributions of joint distribution $Q_{X,Y}$ are denoted by Q_X and Q_Y. We use $\mathcal{P}^{\mathcal{X}\times\mathcal{Y}}$ to denote the probability simplex supported on $\mathcal{X}\times\mathcal{Y}$, and $\mathrm{relint}(\mathcal{P}^{\mathcal{X}\times\mathcal{Y}})$ to denote the relative interior of $\mathcal{P}^{\mathcal{X}\times\mathcal{Y}}$. Besides, we will focus on the case where $P_{X,Y}\in\mathrm{relint}(\mathcal{P}^{\mathcal{X}\times\mathcal{Y}})$.

In GSL, the exponential family $\mathcal{E}_k$ can be equivalently expressed as

$$\left\{Q_{X,Y}\in\mathcal{P}^{\mathcal{X}\times\mathcal{Y}}\colon Q_{X,Y}(x,y)=e^{f^{\mathrm{T}}(x)g(y)+a(x)+b(y)}\right\}. \quad (4)$$

To obtain this equivalence, first, note that the set (4) is a subset of $\mathcal{E}_k$. Then, with b' defined as $b'(y)=b(y)-\log\left(\sum_{(x',y')\in\mathcal{X}\times\mathcal{Y}} e^{f^{\mathrm{T}}(x')g(y')+a(x')+b(y')}\right)$ for $Q_{X,Y}$ in (1), we have $Q_{X,Y}(x,y)=e^{f^{\mathrm{T}}(x)g(y)+a(x)+b'(y)}$.

In our development, it is more convenient to use functions α,β to replace the original a,b in (4), with definitions

[1]The function $f(x)$ here models the case that in neural networks, the input to the softmax regression layer is a feature of data x generated from the hidden layers, and the design of $f(x)$ corresponds to the design of parameters in hidden layers. Also note that the ordinary softmax regression corresponds to the case $f(x)=x$.

IEEE Xplore link to the published article: **DOI** 10.1109/ITW.2018.8613303.

$\alpha(x) \triangleq \log P_X(x) - a(x)$ and $\beta(y) \triangleq \log P_Y(y) - b(y)$. Thus distributions in $\mathcal{E}_k$ can be expressed as

$$Q_{X,Y}(x,y) = P_X(x)P_Y(y)e^{f^{\mathrm{T}}(x)g(y)-\alpha(x)-\beta(y)}, \quad (5)$$

and we use $Q_{X,Y} = Q[f,g,\alpha,\beta]$ to denote this parameterization. Similarly, we use $\tilde{P}_{X,Y} = \tilde{P}[f,g,b]$ to denote the joint distribution $\tilde{P}_{X,Y} \triangleq P_X\tilde{P}_{Y|X}$ in the original softmax regression problem, where $\tilde{P}_{Y|X}$ is the estimated conditional probability defined in (3).

Moreover, we use notations $\mathbf{F}, \mathbf{G}, \boldsymbol{\alpha}, \boldsymbol{\beta}$ to characterize functions f, g, α, β, with $\mathbf{F} = [f(1), \ldots, f(|\mathcal{X}|)]^{\mathrm{T}} \in \mathbb{R}^{|\mathcal{X}|\times k}$, $\mathbf{G} = [g(1), \ldots, g(|\mathcal{Y}|)]^{\mathrm{T}} \in \mathbb{R}^{|\mathcal{Y}|\times k}$, $\boldsymbol{\alpha} = [\alpha(1), \ldots, \alpha(|\mathcal{X}|)]^{\mathrm{T}} \in \mathbb{R}^{|\mathcal{X}|}$, and $\boldsymbol{\beta} = [\beta(1), \ldots, \beta(|\mathcal{Y}|)]^{\mathrm{T}} \in \mathbb{R}^{|\mathcal{Y}|}$.

III. Existence and Non-uniqueness

The following lemma is useful for illustrating the existence of (2), with its proof presented in Section VI.

Lemma 1. *For all $Q_{X,Y} \in \mathcal{E}_k$ with $D(P_{X,Y}\|Q_{X,Y}) \le D(P_{X,Y}\|P_XP_Y)$, there exist parameters f, g, α, β and a constant $M(P_{X,Y})$ independent of $Q_{X,Y}$, such that $Q_{X,Y} = Q[f,g,\alpha,\beta]$ and*

$$\max_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \{\|f(x)\|, \|g(y)\|, |\alpha(x)|, |\beta(y)|\} \le M(P_{X,Y}).$$

Then we have the following theorem.

Theorem 1. *The solutions of the GSL problem* (2) *exist.*

Proof: Since $P_XP_Y = Q[0,0,0,0] \in \mathcal{E}_k$, to find $Q_{X,Y} \in \mathcal{E}_k$ that minimizes $D(P_{X,Y}\|Q_{X,Y})$, it suffices to consider the $Q_{X,Y} \in \mathcal{E}_k$ that satisfies $D(P_{X,Y}\|Q_{X,Y}) \le D(P_{X,Y}\|P_XP_Y)$. Such $Q_{X,Y}$, from Lemma 1, belongs to the set

$$\Big\{Q_{X,Y} \in \mathcal{P}^{\mathcal{X}\times\mathcal{Y}} \colon Q_{X,Y} = Q[f,g,\alpha,\beta], \max_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \{\|f(x)\|, \|g(y)\|, |\alpha(x)|, |\beta(y)|\} \le M\Big\}, \quad (6)$$

where M is a constant independent of $Q_{X,Y}$.

Note that $D(P_{X,Y}\|Q_{X,Y})$ is a continuous function of $Q_{X,Y}$ on the compact set (6), thus can attain its minima. ∎

In the following, we construct an example to show that the optimal solutions of (2) can be non-unique. First, we introduce a useful lemma, of which the proof is presented in Section VI.

Lemma 2. *For a distribution $R_{X,Y} \in \mathcal{P}^{\mathcal{X}\times\mathcal{Y}}$ with corresponding PMI (Pointwise Mutual Information) matrix $\mathbf{\Gamma} = [\Gamma_{x,y}] \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{Y}|}$ defined by*

$$\Gamma_{x,y} \triangleq \log \frac{R_{X,Y}(x,y)}{R_X(x)R_Y(y)}, \quad \forall\,(x,y) \in \mathcal{X}\times\mathcal{Y}, \quad (7)$$

we have $R_{X,Y} \in \mathcal{E}_k$ if $\mathrm{rank}(\mathbf{\Gamma}) \le k$, and $R_{X,Y} \notin \mathcal{E}_k$ if $\mathrm{rank}(\mathbf{\Gamma}) > k+2$.

Consider the GSL example with parameters $k = 1, |\mathcal{X}| = |\mathcal{Y}| = 4$, and $P_{X,Y}$ given by

$$P_{X,Y}(x,y) = u\,\delta_{x,y} + v(1-\delta_{x,y}), \quad (8)$$

where $u > v > 0$, and δ is the Kronecker delta.

Suppose $Q_{X,Y}$ is the unique element of $\mathcal{M}_{\mathcal{E}_1}(P_{X,Y})$, then $Q_{X,Y}$ must have the same form as $P_{X,Y}$, i.e., $\exists\, u', v'$ such that $Q_{X,Y}(x,y) = u'\delta_{x,y} + v'(1-\delta_{x,y})$. Otherwise, we can construct $Q'_{X,Y} \in \mathcal{E}_1$ via permuting the elements in $Q_{X,Y}$, such that $Q_{X,Y} \ne Q'_{X,Y}$ and $D(P_{X,Y}\|Q_{X,Y}) = D(P_{X,Y}\|Q'_{X,Y})$, which contradicts the uniqueness of $Q_{X,Y}$.

If $u' \ne v'$, then the PMI matrix of $Q_{X,Y}$ has full rank $4 > k+2 = 3$. From Lemma 2, we have $Q_{X,Y} \notin \mathcal{E}_1$. On the other hand, $u' = v'$ implies $Q_{X,Y} = P_XP_Y$. This is impossible, however, as we can find $Q''_{X,Y} \in \mathcal{E}_1$ such that $D(P_{X,Y}\|Q''_{X,Y}) < D(P_{X,Y}\|P_XP_Y)$. An example of such $Q''_{X,Y}$ is

$$Q''_{X,Y}(x,y) = \begin{cases} u, & \text{if } x = y = 1, \\ \frac{1-u}{15}, & \text{otherwise.} \end{cases}$$

As a consequence, the M-projection of this $P_{X,Y}$ onto $\mathcal{E}_1$ is not unique.

IV. The Geometric Structure

In this section, we consider the geometric structure of GSL and present its applications in machine learning.

A. Stationary Distributions of GSL

Suppose $Q_{X,Y} = Q[f,g,\alpha,\beta] \in \mathcal{M}_{\mathcal{E}_k}(P_{X,Y})$, then (f,g,α,β) is a stationary point of the Lagrange function $\mathcal{L}(f,g,\alpha,\beta,\lambda)$ corresponding to the GSL problem (2), where $\mathcal{L}$ is defined as

$$\mathcal{L} = D(P_{X,Y}\|Q_{X,Y}) + \lambda\left[\sum_{x',y'} Q_{X,Y}(x',y') - 1\right]. \quad (9)$$

The independent variables of $\mathcal{L}$ are all possible values of functions f, g, α, β, i.e., $\{f(x), g(y), \alpha(x), \beta(y)\}_{(x,y)\in\mathcal{X}\times\mathcal{Y}}$, together with the Lagrange multiplier λ. The stationary points of $\mathcal{L}$ satisfy that,

$$\frac{\partial\mathcal{L}}{\partial f(x)} = \frac{\partial\mathcal{L}}{\partial g(y)} = 0, \quad \forall\,(x,y) \in \mathcal{X}\times\mathcal{Y}, \quad (10a)$$

$$\frac{\partial\mathcal{L}}{\partial \alpha(x)} = \frac{\partial\mathcal{L}}{\partial \beta(y)} = 0, \quad \forall\,(x,y) \in \mathcal{X}\times\mathcal{Y}, \quad (10b)$$

$$\frac{\partial\mathcal{L}}{\partial \lambda} = \sum_{(x',y')\in\mathcal{X}\times\mathcal{Y}} Q_{X,Y}(x',y') - 1 = 0. \quad (10c)$$

To reduce these conditions, note that

$$\begin{aligned} &D(P_{X,Y}\|Q_{X,Y}) \\ &= \sum_{x,y} P_{X,Y}(x,y)\log\frac{P_{X,Y}(x,y)}{Q_{X,Y}(x,y)} \\ &= D(P_{X,Y}\|P_XP_Y) - \mathbb{E}_{P_{X,Y}}\left[f^{\mathrm{T}}(X)g(Y)\right] \\ &\quad + \mathbb{E}_{P_X}\left[\alpha(X)\right] + \mathbb{E}_{P_Y}\left[\beta(Y)\right], \end{aligned} \quad (11)$$

then (10b) implies $P_X = \lambda Q_X, P_Y = \lambda Q_Y$, thus $\lambda = 1$. The conditions (10) for stationary points then become

$$Q_X = P_X,\ Q_Y = P_Y, \quad (12a)$$

$$\mathbb{E}_{Q_{X|Y}}\left[f(X) \mid Y\right] = \mathbb{E}_{P_{X|Y}}\left[f(X) \mid Y\right], \quad (12b)$$

$$\mathbb{E}_{Q_{Y|X}}\left[g(Y) \mid X\right] = \mathbb{E}_{P_{Y|X}}\left[g(Y) \mid X\right]. \quad (12c)$$

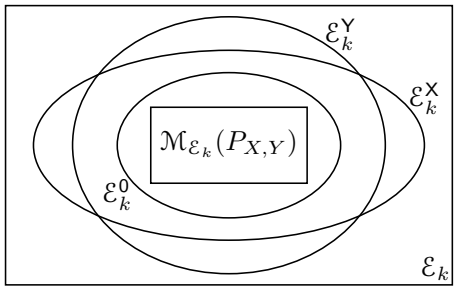


Fig. 1. Relationship between different distribution families

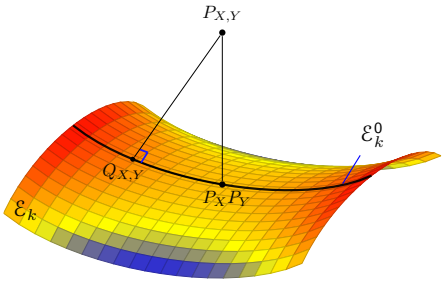


Fig. 2. Pythagorean theorem for stationary distributions $Q_{X,Y} \in \mathcal{E}_k^0$: $D(P_{X,Y}\|P_XP_Y) = D(P_{X,Y}\|Q_{X,Y}) + D(Q_{X,Y}\|P_XP_Y)$

A distribution $Q_{X,Y} = Q[f,g,\alpha,\beta] \in \mathcal{E}_k$ is called a *stationary distribution* (of GSL) if it satisfies (12). Let $\mathcal{E}_k^0$ denote the set of all stationary distributions, then we have $\mathcal{M}_{\mathcal{E}_k}(P_{X,Y}) \subset \mathcal{E}_k^0$. Furthermore, from (12) we have $\mathcal{E}_k^0 \subset \mathcal{E}_k^{\mathsf{X}} \cap \mathcal{E}_k^{\mathsf{Y}}$, where $\mathcal{E}_k^{\mathsf{X}}$ and $\mathcal{E}_k^{\mathsf{Y}}$ are the subsets of $\mathcal{E}_k$ with definitions $\mathcal{E}_k^{\mathsf{X}} \triangleq \{Q_{X,Y} \in \mathcal{E}_k \colon Q_X = P_X\}, \mathcal{E}_k^{\mathsf{Y}} \triangleq \{Q_{X,Y} \in \mathcal{E}_k \colon Q_Y = P_Y\}$. The relationship between different distribution families is shown in Fig. 1.

The set $\mathcal{E}_k^0$ has the following properties.

Property 1. $\forall k \in \mathbb{N}_+, \mathcal{E}_k^0 \subset \mathcal{E}_{k+1}^0$.

Proof: For each $Q_{X,Y} = Q[f_k, g_k, \alpha, \beta] \in \mathcal{E}_k^0$, we can construct $f_{k+1}(x) = [f_k^{\mathrm{T}}(x), 0]^{\mathrm{T}} \in \mathbb{R}^{k+1}, g_{k+1}(y) = [g_k^{\mathrm{T}}(y), 0]^{\mathrm{T}} \in \mathbb{R}^{k+1}$, such that $Q_{X,Y} = Q[f_{k+1}, g_{k+1}, \alpha, \beta] \in \mathcal{E}_{k+1}^0$. ■

Property 2 (Pythagorean theorem)**.** $\forall Q_{X,Y} \in \mathcal{E}_k^0$,

$$D(P_{X,Y}\|Q_{X,Y}) + D(Q_{X,Y}\|P_XP_Y) = D(P_{X,Y}\|P_XP_Y).$$

Proof: For each $Q_{X,Y} = Q[f,g,\alpha,\beta] \in \mathcal{E}_k^0$, we have $D(Q_{X,Y}\|P_XP_Y) = \mathbb{E}_{Q_{X,Y}}\left[f^{\mathrm{T}}(X)g(Y)\right] - \mathbb{E}_{Q_X}\left[\alpha(X)\right] - \mathbb{E}_{Q_Y}\left[\beta(Y)\right]$. From the definition of $\mathcal{E}_k^0$ in (12), we have $\mathbb{E}_{Q_X}\left[\alpha(X)\right] = \mathbb{E}_{P_X}\left[\alpha(X)\right], \mathbb{E}_{Q_Y}\left[\beta(Y)\right] = \mathbb{E}_{P_Y}\left[\beta(Y)\right]$, and

$$\begin{aligned}\mathbb{E}_{Q_{X,Y}}\left[f^{\mathrm{T}}(X)g(Y)\right] &= \mathbb{E}_{Q_X}\left[f^{\mathrm{T}}(X)\,\mathbb{E}_{Q_{Y|X}}\left[g(Y)|X\right]\right] \\ &= \mathbb{E}_{P_X}\left[f^{\mathrm{T}}(X)\,\mathbb{E}_{P_{Y|X}}\left[g(Y)|X\right]\right] \\ &= \mathbb{E}_{P_{X,Y}}\left[f^{\mathrm{T}}(X)g(Y)\right].\end{aligned}$$

The above relationships together with (11) lead to the theorem. ■

Property 3. $\forall\, k \in \mathbb{N}_+, \mathcal{M}_{\mathcal{E}_k}(P_{X,Y}) = \mathcal{E}_k^0$ *if and only if* $P_{X,Y} = P_XP_Y$.

Proof: If $P_{X,Y} = P_XP_Y \in \mathcal{E}_k^0$, then $\mathcal{M}_{\mathcal{E}_k}(P_{X,Y}) = \{P_{X,Y}\}$ by definition. In addition, from Property 2, we have $D(P_{X,Y}\|Q_{X,Y}) + D(Q_{X,Y}\|P_XP_Y) = D(P_{X,Y}\|P_XP_Y) = 0$ for all $Q_{X,Y} \in \mathcal{E}_k^0$, which implies $Q_{X,Y} = P_{X,Y} = P_XP_Y$. As a result, $\mathcal{E}_k^0 = \{P_{X,Y}\} = \mathcal{M}_{\mathcal{E}_k}(P_{X,Y})$.

If $\mathcal{M}_{\mathcal{E}_k}(P_{X,Y}) = \mathcal{E}_k^0$, then $P_XP_Y \in \mathcal{M}_{\mathcal{E}_k}(P_{X,Y})$. Since $P_XP_Y = Q[f,0,0,0]$ for all choices of f, $(f,0,0,0)$ should satisfy the stationary point conditions (12). For all $\hat{x} \in \mathcal{X}$, let $f(x) = [\mathbb{1}_{x=\hat{x}}, 0, \ldots, 0]^{\mathrm{T}} \in \mathbb{R}^k$, then (12b) implies that $\forall y \in \mathcal{Y}, P_{X|Y}(\hat{x}|y) = P_X(\hat{x})$, thus $P_{X,Y} = P_XP_Y$. ■

Property 1 demonstrates that $\{\mathcal{E}_k^0\}$ forms a non-decreasing sequence of sets. Property 2 gives a Pythagorean theorem on $\mathcal{E}_k^0$, as shown in Fig. 2. Property 3 illustrates that, $\mathcal{M}_{\mathcal{E}_k}(P_{X,Y})$ is a proper subset of stationary distributions $\mathcal{E}_k^0$ unless X and Y are independent.

B. Equivalence of Softmax Learning Problems

To establish the equivalence of GSL and the original softmax regression problem, we first show that softmax regression is equivalent to the M-projection problem $\mathcal{M}_{\mathcal{E}_k^{\mathsf{X}}}(P_{X,Y})$.

In softmax regression, with a series of data samples $\{(x_i, y_i)\}_{i=1}^N$, the conditional distribution $\tilde{P}_{Y|X}$ (3) is estimated by the softmax function based on the feature $f(x) \in \mathbb{R}^k$, weights $g(y) \in \mathbb{R}^k$ and bias $b(y) \in \mathbb{R}$. The parameters (f,g,b) are then chosen to maximize the empirical expectation of log-likelihood function $\log \tilde{P}_{Y|X}$:

$$\frac{1}{N}\sum_{i=1}^N \log \tilde{P}_{Y|X}(y_i|x_i) = \mathbb{E}_{P_{X,Y}}\left[\log \tilde{P}_{Y|X}(Y|X)\right],$$

which is equivalent to solving the optimization problem

$$\underset{f,g,b}{\text{minimize}}\ D(P_{X,Y}\|\tilde{P}[f,g,b]), \tag{13}$$

where $P_{X,Y}$ is the empirical distribution of data samples.

To demonstrate the equivalence of (13) and the M-projection problem $\mathcal{M}_{\mathcal{E}_k^{\mathsf{X}}}(P_{X,Y})$, it suffices to show that $\mathcal{E}_k^{\mathsf{X}}$ is the family of distributions of the form $\tilde{P}[f,g,b]$, as will be clarified in the following lemma.

Lemma 3. *Given parameters f, g, the following conditions are equivalent for distribution $R_{X,Y}$:*

1) *$\exists\, b$, such that $R_{X,Y} = \tilde{P}[f,g,b]$.*
2) *$\exists\, \alpha, \beta$, such that $R_{X,Y} = Q[f,g,\alpha,\beta] \in \mathcal{E}_k^{\mathsf{X}}$.*

Proof: 1) $\Longrightarrow$ 2) For given f,g,b, suppose $\alpha(x) = \log \sum_{y' \in \mathcal{Y}} e^{f^{\mathrm{T}}(x)g(y')+b(y')}, \beta(y) = -b(y) + \log P_Y(y)$, then we have $R_{X,Y} = \tilde{P}[f,g,b] = Q[f,g,\alpha,\beta] \in \mathcal{E}_k$. Besides, the definition of $\tilde{P}$ implies that $R_X = P_X$. As a result, $R_{X,Y} = Q[f,g,\alpha,\beta] \in \mathcal{E}_k^{\mathsf{X}}$.

2) $\Longrightarrow$ 1) The definition of $\mathcal{E}_k^{\mathsf{X}}$ indicates that

$$\sum_{y' \in \mathcal{Y}} P_X(x)P_Y(y')e^{f^{\mathrm{T}}(x)g(y')-\alpha(x)-\beta(y')} = P_X(x),$$

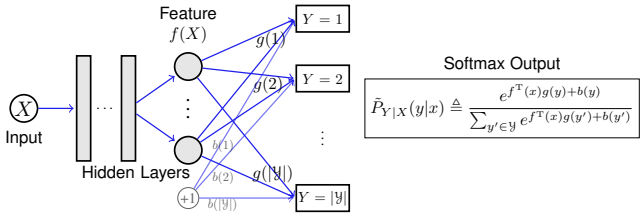


Fig. 3. A forward neural network for classification where f is the feature generated by the last hidden layer, g and b are the weights and bias in the last layer.

which implies $\alpha(x) = \log \sum_{y'\in\mathcal{Y}} P_Y(y') e^{f^{\mathrm{T}}(x)g(y')-\beta(y')}$. Then we have $Q[f,g,\alpha,\beta] = \tilde{P}[f,g,b]$ with $b(y) = -\beta(y) + \log P_Y(y)$. ■

Then, the following theorem shows the equivalence of the original softmax regression problem (13) and the GSL problem (2).

Theorem 2. *The M-projections of $P_{X,Y}$ onto $\mathcal{E}_k$, $\mathcal{E}_k^{\mathsf{X}}$, and $\mathcal{E}_k^{\mathsf{Y}}$ are the same, i.e.,*

$$\mathcal{M}_{\mathcal{E}_k}(P_{X,Y}) = \mathcal{M}_{\mathcal{E}_k^{\mathsf{X}}}(P_{X,Y}) = \mathcal{M}_{\mathcal{E}_k^{\mathsf{Y}}}(P_{X,Y}).$$

Proof: For all $Q_{X,Y} \in \mathcal{M}_{\mathcal{E}_k}(P_{X,Y}) \subset \mathcal{E}_k^0 \subset \mathcal{E}_k^{\mathsf{X}}$, by the definition of $\mathcal{M}_{\mathcal{E}_k}(P_{X,Y})$, we have

$$D(P_{X,Y}\|Q_{X,Y}) \le D(P_{X,Y}\|Q'_{X,Y}),\ \forall\, Q'_{X,Y} \in \mathcal{E}_k^{\mathsf{X}} \subset \mathcal{E}_k,$$

which implies $Q_{X,Y} \in \mathcal{M}_{\mathcal{E}_k^{\mathsf{X}}}(P_{X,Y})$. As a result, we have $\mathcal{M}_{\mathcal{E}_k}(P_{X,Y}) \subset \mathcal{M}_{\mathcal{E}_k^{\mathsf{X}}}(P_{X,Y})$. Suppose $\mathcal{M}_{\mathcal{E}_k}(P_{X,Y}) \neq \mathcal{M}_{\mathcal{E}_k^{\mathsf{X}}}(P_{X,Y})$, then for $Q_{X,Y} \in \mathcal{M}_{\mathcal{E}_k^{\mathsf{X}}}(P_{X,Y}) \setminus \mathcal{M}_{\mathcal{E}_k}(P_{X,Y})$, $\exists\, Q'_{X,Y} \in \mathcal{M}_{\mathcal{E}_k}(P_{X,Y}) \subset \mathcal{E}_k^{\mathsf{X}}$ such that

$$D(P_{X,Y}\|Q_{X,Y}) > D(P_{X,Y}\|Q'_{X,Y}),$$

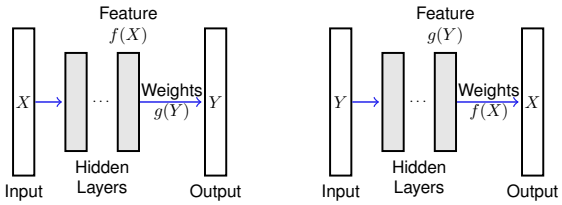
which implies $Q_{X,Y} \notin \mathcal{M}_{\mathcal{E}_k^{\mathsf{X}}}(P_{X,Y})$ and contradicts the assumption. Thus, we have $\mathcal{M}_{\mathcal{E}_k}(P_{X,Y}) = \mathcal{M}_{\mathcal{E}_k^{\mathsf{X}}}(P_{X,Y})$. Similarly, we can prove that $\mathcal{M}_{\mathcal{E}_k}(P_{X,Y}) = \mathcal{M}_{\mathcal{E}_k^{\mathsf{Y}}}(P_{X,Y})$. ■

Remark 1. It can be verified that the stationary distributions of the original softmax regression problem (13) also coincide with the stationary distributions of the GSL problem (2).

C. Applications in Neural Networks

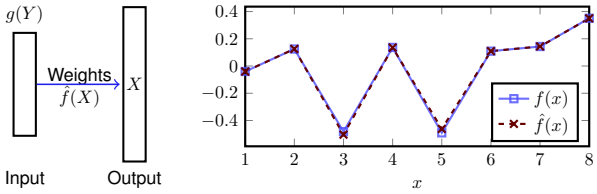
Consider the forward neural network that uses data X to predict label Y shown in Fig. 3, where f is the feature of X extracted by the last hidden layer, g and b correspond to the weights and bias in the last layer, respectively. When there are enough hidden neurons in the neural network, f can express any desired function [1]. Then training this neural network is equivalent to solving the original softmax regression problem (13). We will focus on such neural networks of which the hidden layers have ideal expressive power.

From Theorem 2, the original softmax regression problem $\mathcal{M}_{\mathcal{E}_k^{\mathsf{X}}}(P_{X,Y})$ and the GSL problem $\mathcal{M}_{\mathcal{E}_k}(P_{X,Y})$ have the same solution set. Since f and g are symmetric in the formulation of the GSL problem (2), feature f and weights g extracted by this neural network have symmetric roles in training.



(a) X-Y network (b) Y-X network

Fig. 4. Symmetric (feature, weights) pairs generated by the X-Y network and the Y-X network



(a) The reverse network (b) Comparison between $f(x)$ and $\hat{f}(x)$

Fig. 5. The simulation to verify the symmetry of neural networks

Furthermore, the following proposition as a straightforward corollary of Theorem 2, demonstrates the symmetry between the original network which uses X to predict Y and its reverse network that uses Y to predict X.

Proposition 1. *A neural network that uses X to predict Y (X-Y network) and its reverse network that uses Y to predict X (Y-X network) generate symmetric (feature, weights) pairs, as illustrated in Fig. 4.*

V. Simulation Results

We design a numerical simulation to verify the symmetry property of neural networks presented in Proposition 1. In particular, we set $|\mathcal{X}| = 8$, $|\mathcal{Y}| = 6$, with feature dimension $k = 1$. For a given distribution $P_{X,Y}$, we generate $N = 100\,000$ samples of (x_i, y_i) then feed them to the X-Y network shown in Fig. 4a. The input x is one-hot encoded, and the only hidden layer is a fully connected layer used to generate $f(x)$. With proper weights in this hidden layer, f can express all possible features of x. After training this network, we can obtain a (feature, weights) pair (f, g).

Then, we use $g(y_i)$ as the input feature of the reverse neural network shown in Fig. 5a[2] to predict the label x_i, and obtain the trained weights $\hat{f}$. As shown in Fig. 5b, $\hat{f}(x)$ matches the original feature $f(x)$ precisely, thus verifying the symmetry of these two neural networks.

VI. Proofs

A. Proof of Lemma 1

Proof: As $P_{X,Y} \in \mathrm{relint}(\mathcal{P}^{\mathcal{X}\times\mathcal{Y}})$, $\exists\, \delta_P > 0$ such that $\forall\, (x,y) \in \mathcal{X}\times\mathcal{Y}$, $P_{X,Y}(x,y) > \delta_P$. It can be shown that

[2]Due to the possibly non-uniqueness of (2), we use this simplified reverse network instead of the Y-X network shown in Fig. 4b, to force the feature in the reverse network to be $g(Y)$.

$\exists\, \delta_Q > 0$, such that

$$\min_{(x,y)\in\mathcal{X}\times\mathcal{Y}} Q_{X,Y}(x,y) > \delta_Q$$

for all $Q_{X,Y} \in \mathcal{E}_k$ with $D(P_{X,Y}\|Q_{X,Y}) \le D(P_{X,Y}\|P_XP_Y)$.

Indeed, $\forall\,(x,y) \in \mathcal{X}\times\mathcal{Y}$, suppose $p = P_{X,Y}(x,y), q = Q_{X,Y}(x,y)$, then

$$\begin{aligned} D(P_{X,Y}\|P_XP_Y) &\ge D(P_{X,Y}\|Q_{X,Y}) \\ &\ge D(\mathrm{Bern}(p)\|\,\mathrm{Bern}(q)) \\ &= -H(p) - p\log q - (1-p)\log(1-q) \\ &> -1 - \delta_P \log q, \end{aligned}$$

where the second inequality follows from the data processing inequality. As a result, a valid choice of δ_Q is given by

$$\delta_Q \triangleq \exp\left(-\frac{1}{\delta_P}[D(P_{X,Y}\|P_XP_Y)+1]\right), \tag{14}$$

and we have $0 \le \delta_Q < Q_{X,Y}(x,y) < 1$. Suppose $Q_{X,Y} = Q[f,g,\alpha,\beta]$, then $f^{\mathrm{T}}(x)g(y) - \alpha(x) - \beta(y)$ is bounded, i.e., there exists $M_1 > 0$ such that $\forall\,(x,y)\in\mathcal{X}\times\mathcal{Y}$,

$$\left|f^{\mathrm{T}}(x)g(y) - \alpha(x) - \beta(y)\right| \le M_1. \tag{15}$$

Note that the parameterization of $Q_{X,Y}$ is non-unique, e.g., we have $Q[f,g,\alpha,\beta] = Q[f+c,g,\alpha+c^{\mathrm{T}}g,\beta]$ for any constant vector c. Without loss of generality, we can reparameterize $Q_{X,Y}$ and assume that $\mathbb{E}_{P_X}[f(X)] = \mathbb{E}_{P_Y}[g(Y)] = 0$ and $\mathbb{E}_{P_X}[\alpha(X)] = 0$. Then from Jensen's inequality, $\forall\,(x,y)\in\mathcal{X}\times\mathcal{Y}$,

$$\begin{aligned} |\beta(y)| &= \left|\mathbb{E}_{P_X}\left[f^{\mathrm{T}}(X)g(y) - \alpha(X) - \beta(y)\right]\right| \\ &\le \mathbb{E}_{P_X}\left[\left|f^{\mathrm{T}}(X)g(y) - \alpha(X) - \beta(y)\right|\right] \\ &\le M_1, \end{aligned} \tag{16}$$

$$\begin{aligned} |\alpha(x)| &\le |\alpha(x) + \mathbb{E}_{P_Y}[\beta(Y)]| + |\mathbb{E}_{P_Y}[\beta(Y)]| \\ &\le \left|\mathbb{E}_{P_Y}\left[f^{\mathrm{T}}(x)g(Y) - \alpha(x) - \beta(Y)\right]\right| + M_1 \\ &\le \mathbb{E}_{P_Y}\left[\left|f^{\mathrm{T}}(x)g(Y) - \alpha(x) - \beta(Y)\right|\right] + M_1 \\ &\le 2M_1. \end{aligned} \tag{17}$$

As a result,

$$\begin{aligned} \left|f^{\mathrm{T}}(x)g(y)\right| &\le \left|f^{\mathrm{T}}(x)g(y) - \alpha(x) - \beta(y)\right| + |\alpha(x)| + |\beta(y)| \\ &\le 4M_1, \end{aligned}$$

i.e., all elements in $\mathbf{F}\mathbf{G}^{\mathrm{T}}$ are bounded by $4M_1$, then the norm equivalence [2] implies that

$$\left\|\mathbf{F}\mathbf{G}^{\mathrm{T}}\right\|_2 \le \sqrt{|\mathcal{X}||\mathcal{Y}|}\left\|\mathbf{F}\mathbf{G}^{\mathrm{T}}\right\|_{\max} \le 4\sqrt{|\mathcal{X}||\mathcal{Y}|}M_1. \tag{18}$$

Suppose $\mathbf{F}\mathbf{G}^{\mathrm{T}}$ has the compact SVD (Singular Value Decomposition) $\mathbf{F}\mathbf{G}^{\mathrm{T}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}$, where $\mathbf{\Sigma}$ is a square matrix containing all positive singular values of $\mathbf{F}\mathbf{G}^{\mathrm{T}}$, then we can construct $\hat{\mathbf{F}} = \mathbf{U}\mathbf{\Sigma}^{1/2}$ and $\hat{\mathbf{G}} = \mathbf{V}\mathbf{\Sigma}^{1/2}$, such that

$$\hat{\mathbf{F}}\hat{\mathbf{G}}^{\mathrm{T}} = \mathbf{F}\mathbf{G}^{\mathrm{T}}, \left\|\hat{\mathbf{F}}\right\|_2 = \left\|\hat{\mathbf{G}}\right\|_2 = \left\|\mathbf{F}\mathbf{G}^{\mathrm{T}}\right\|_2^{1/2}. \tag{19}$$

Let $\hat{f},\hat{g}$ be the functions corresponding to $\hat{\mathbf{F}},\hat{\mathbf{G}}$, respectively, then $Q[f,g,\alpha,\beta] = Q[\hat{f},\hat{g},\alpha,\beta]$. Since $\|\hat{\mathbf{F}}\|_2$ and $\|\hat{\mathbf{G}}\|_2$ are bounded, from norm equivalence, we have $\|\hat{\mathbf{F}}\|_{\mathrm{F}} \le M_2, \|\hat{\mathbf{G}}\|_{\mathrm{F}} \le M_2$ with $M_2 \triangleq 2\sqrt{|\mathcal{X}||\mathcal{Y}|M_1}$. Thus $\forall\,(x,y)\in\mathcal{X}\times\mathcal{Y}$,

$$\|\hat{f}(x)\| \le \|\hat{\mathbf{F}}\|_{\mathrm{F}} \le M_2, \|\hat{g}(y)\| \le \|\hat{\mathbf{G}}\|_{\mathrm{F}} \le M_2. \tag{20}$$

Let $M \triangleq \max\{2M_1, M_2\}$, then M only depends on $P_{X,Y}$. As a result, we have $Q_{X,Y} = Q[\hat{f},\hat{g},\alpha,\beta]$ with

$$\max_{(x,y)\in\mathcal{X}\times\mathcal{Y}}\left\{\|\hat{f}(x)\|, \|\hat{g}(y)\|, |\alpha(x)|, |\beta(y)|\right\} \le M(P_{X,Y}).$$

■

B. Proof of Lemma 2

Proof: Suppose $R_{X,Y} \in \mathcal{E}_k$, then there exist parameters f,g,α,β such that $R_{X,Y} = Q[f,g,\alpha,\beta]$. Thus we have, $\forall\,(x,y)\in\mathcal{X}\times\mathcal{Y}$,

$$\log\frac{R_{X,Y}(x,y)}{R_X(x)R_Y(y)} = f^{\mathrm{T}}(x)g(y) - \alpha(x) - \beta(y). \tag{21}$$

The above condition can be equivalently expressed as

$$\mathbf{\Gamma} = \mathbf{F}\mathbf{G}^{\mathrm{T}} - \mathbf{1}_{|\mathcal{X}|}\boldsymbol{\beta}^{\mathrm{T}} - \boldsymbol{\alpha}\mathbf{1}_{|\mathcal{Y}|}^{\mathrm{T}}, \tag{22}$$

where $\mathbf{1}_m$ is an m-dimensional column vector with all elements equaling 1. If $\mathrm{rank}(\mathbf{\Gamma}) \le k$, we can find an $(\mathbf{F},\mathbf{G})$ pair such that $\mathbf{\Gamma} = \mathbf{F}\mathbf{G}^{\mathrm{T}}$ and $R_{X,Y} = Q[f,g,0,0]$. Besides, (22) implies

$$\mathrm{rank}(\mathbf{\Gamma}) \le \mathrm{rank}(\mathbf{F}\mathbf{G}^{\mathrm{T}}) + 2 \le k + 2. \tag{23}$$

As a consequence, $R_{X,Y} \notin \mathcal{E}_k$ if $\mathrm{rank}(\mathbf{\Gamma}) > k+2$. ■

VII. Conclusion

In this work, we studied the geometric structure of the generalized softmax learning (GSL) problem. This geometric structure established the equivalence of GSL and the original softmax regression problem. Moreover, using the connection between neural networks and softmax learning problems, we presented the symmetric roles of the weights and feature in a neural network and the symmetry between a neural network and its reverse network.

Acknowledgment

The research of Shao-Lun Huang was funded by the Shenzhen Municipal Scientific Program JCYJ20170818094022586.

References

[1] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Mathematics of control, signals and systems*, vol. 2, no. 4, pp. 303–314, 1989.

[2] G. H. Golub and C. F. Van Loan, *Matrix computations*. JHU Press, 2012, vol. 3.

[3] S.-i. Amari, *Information geometry and its applications*. Springer, 2016.

[4] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*. MIT press Cambridge, 2016, vol. 1.

[5] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

[6] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[7] S.-L. Huang, A. Makur, L. Zheng, and G. W. Wornell, "An information-theoretic approach to universal feature selection in high-dimensional inference," in *Information Theory (ISIT), 2017 IEEE International Symposium on*. IEEE, 2017, pp. 1336–1340.

[8] S.-L. Huang, L. Zhang, and L. Zheng, "An information-theoretic approach to unsupervised feature selection for high-dimensional data," in *Information Theory Workshop (ITW), 2017 IEEE*. IEEE, 2017, pp. 434–438.