The Geometric Structure of Generalized Softmax Learning

Xiangxiang Xu[1]

Collaborators: Shao-Lun Huang[2] Lizhong Zheng[3] Lin Zhang[2]

ITW 2018

[1]Dept. of Electronic Engineering, Tsinghua University

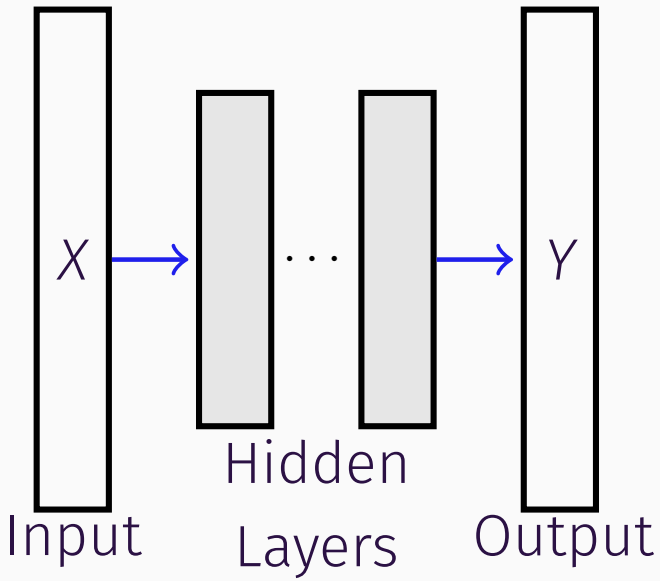
[2]Tsinghua-Berkeley Shenzhen Institute, Tsinghua University

[3]Dept. of EECS, Massachusetts Institute of Technology

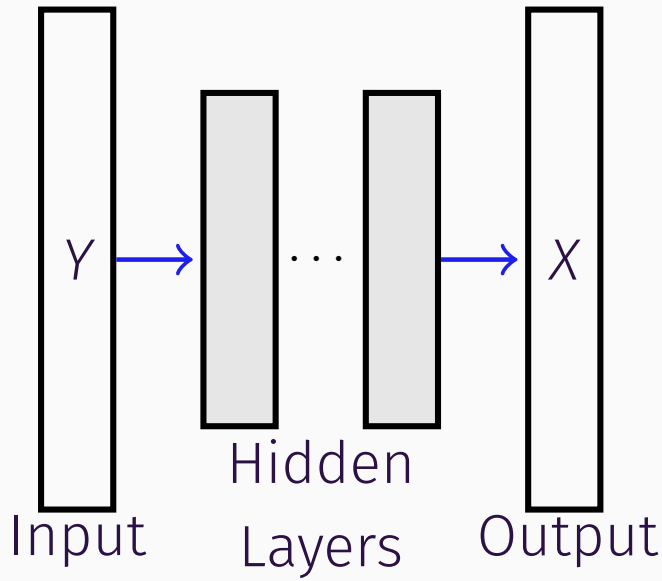
Introduction

Reverse a neural network

Use X to predict Y

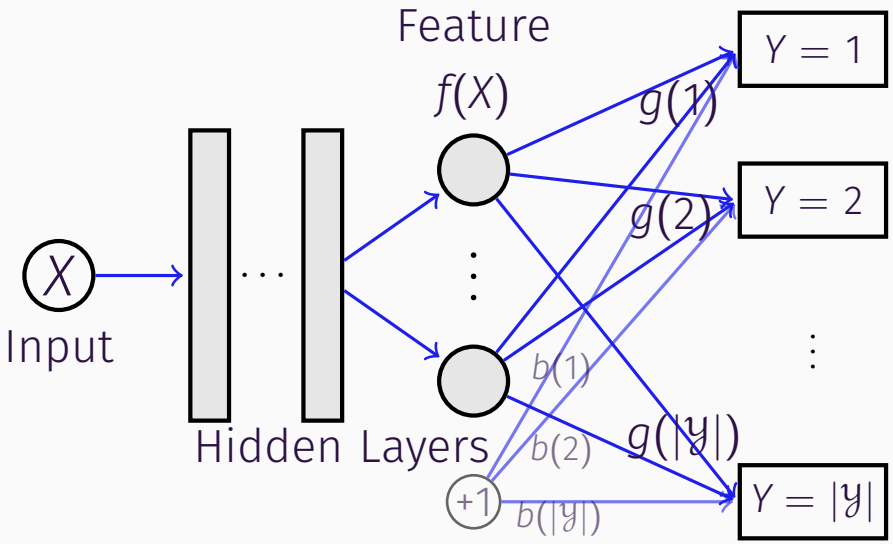


Use Y to predict X



What do they have in common?

Neural network and M-projection



Softmax Output

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

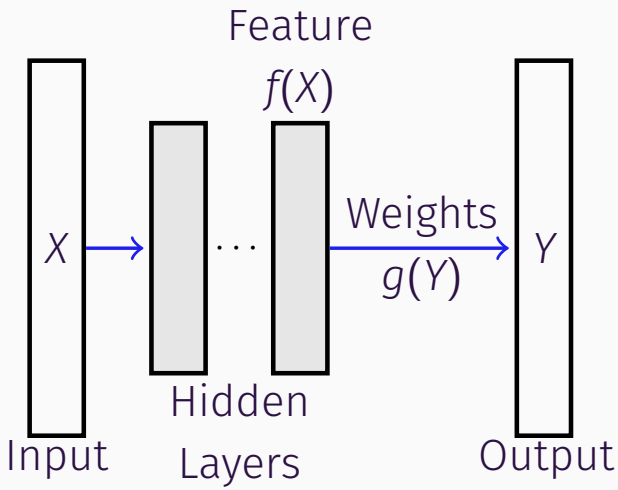
Maximum likelihood estimation

$$(f, g, b)^\star = \arg\max_{f,g,b} \mathbb{E}_{P_{X,Y}} \left[\log \tilde{P}_{Y|X}(Y|X) \right]$$

$$= \arg\min_{f,g,b} D(P_{X,Y} \| P_X \tilde{P}_{Y|X})$$

Generalized softmax learning

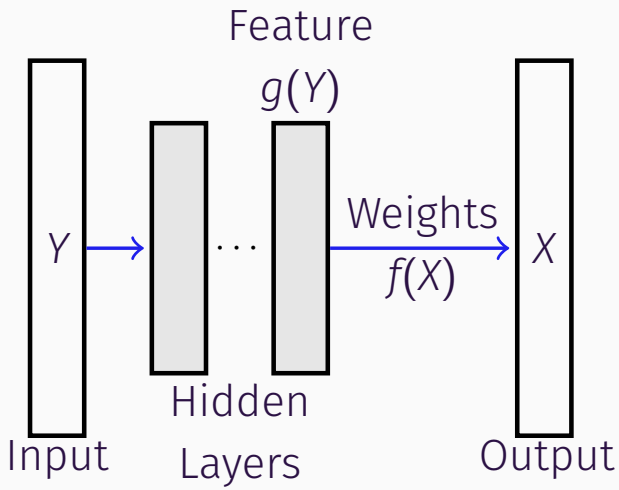
Use X to predict Y



$$\tilde{P}_{Y|X}(y|x) \propto \exp(f^{\mathrm{T}}(x)g(y) + b(y))$$

minimize $D(P_{X,Y} \| P_X \tilde{P}_{Y|X})$

Use Y to predict X



$$\tilde{P}_{X|Y}(x|y) \propto \exp(g^{\mathrm{T}}(y)f(x) + a(x))$$

minimize $D(P_{X,Y} \| P_Y \tilde{P}_{X|Y})$

Generalized softmax learning

minimize $D(P_{X,Y} \| Q_{X,Y})$

$$Q_{X,Y}(x, y) \propto \exp(f^{\mathrm{T}}(x)g(y) + a(x) + b(y))$$

Generalized Softmax Learning

Generalized softmax learning (GSL)

The M-projection of $P_{X,Y}$ onto

$$\mathcal{E}_k \triangleq \left\{Q_{X,Y} \colon Q_{X,Y}(x,y) \propto \exp(f^{\mathrm{T}}(x)g(y) + a(x) + b(y))\right\},$$

where $f \colon \mathcal{X} \to \mathbb{R}^k, g \colon \mathcal{Y} \to \mathbb{R}^k, a \colon \mathcal{X} \to \mathbb{R}, b \colon \mathcal{Y} \to \mathbb{R}$.

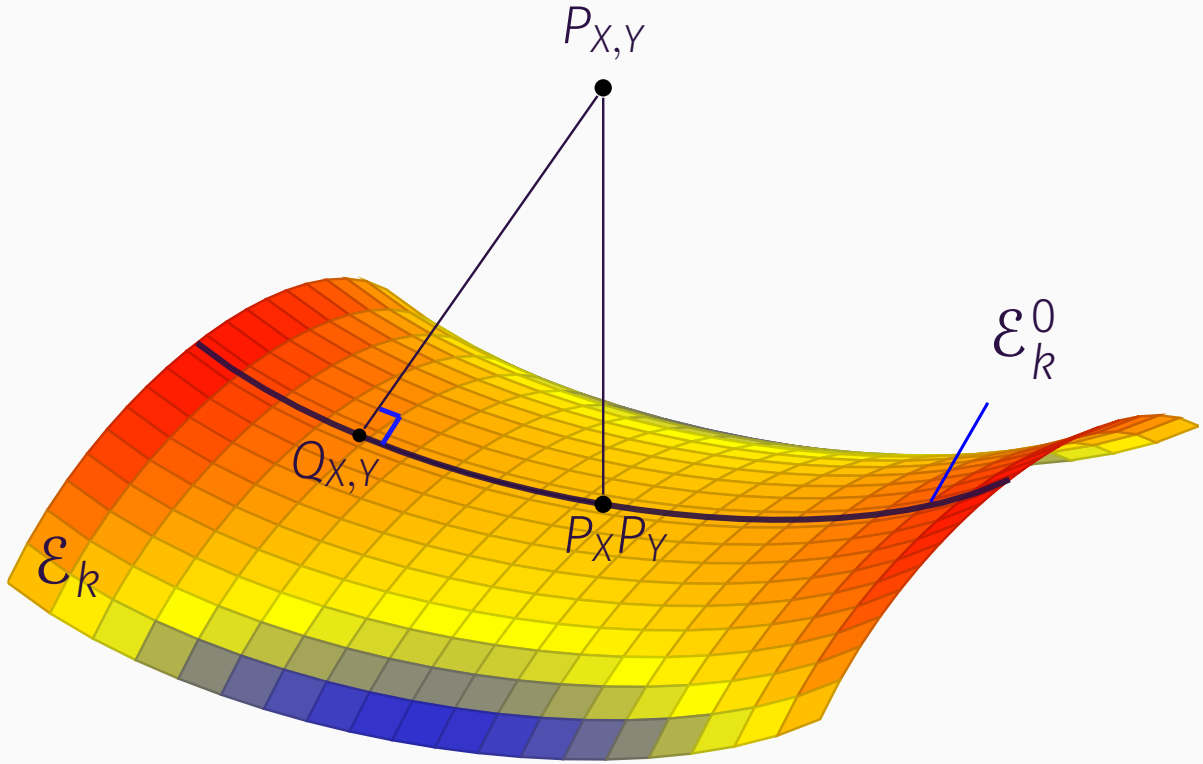
$$\mathcal{M}_{\mathcal{E}_k}(P_{X,Y}) \triangleq \operatorname*{arg\,min}_{Q_{X,Y} \in \mathcal{E}_k} D(P_{X,Y} \| Q_{X,Y}),$$

Stationary distributions $\mathcal{E}_k^0$: $\nabla = 0$

$$Q_X = P_X,\ Q_Y = P_Y,$$
$$\mathbb{E}_{Q_{X|Y}}[f(X) \mid Y] = \mathbb{E}_{P_{X|Y}}[f(X) \mid Y],$$
$$\mathbb{E}_{Q_{Y|X}}[g(Y) \mid X] = \mathbb{E}_{P_{Y|X}}[g(Y) \mid X].$$

$$P_X P_Y \in \mathcal{E}_k^0$$

Pythagorean theorem



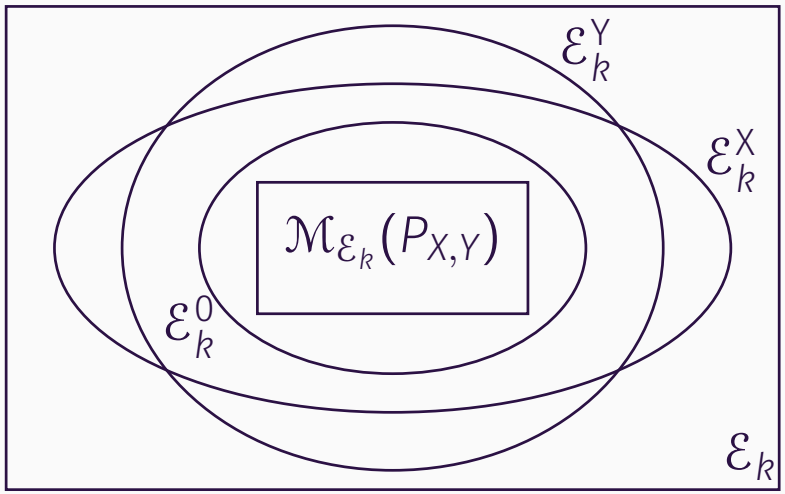
$$\forall\, Q_{X,Y} \in \mathcal{E}_k^0,\, D(P_{X,Y} \| Q_{X,Y}) + D(Q_{X,Y} \| P_X P_Y) = D(P_{X,Y} \| P_X P_Y).$$

$$P_X P_Y \in \mathcal{M}_{\mathcal{E}_k}(P_{X,Y}) \iff X \perp\!\!\!\perp Y$$

Stationary distributions $\mathcal{E}_k^0$

$$Q_{X,Y} \in \mathcal{E}_k^0 \implies Q_X = P_X, Q_Y = P_Y$$

Different distribution families

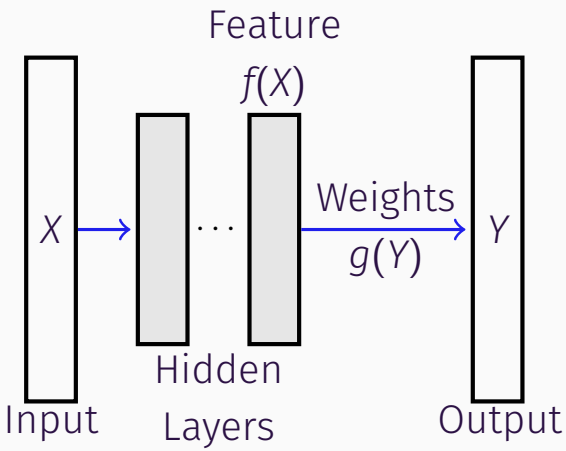


- $\mathcal{E}_k^0$: stationary distributions
- $\mathcal{M}_{\mathcal{E}_k}(P_{X,Y})$: GSL solutions
- $\mathcal{E}_k^X \leftrightarrow Q_{X,Y} \in \mathcal{E}_k, Q_X = P_X$
- $\mathcal{E}_k^Y \leftrightarrow Q_{X,Y} \in \mathcal{E}_k, Q_Y = P_Y$

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

Equivalence of softmax learning problems

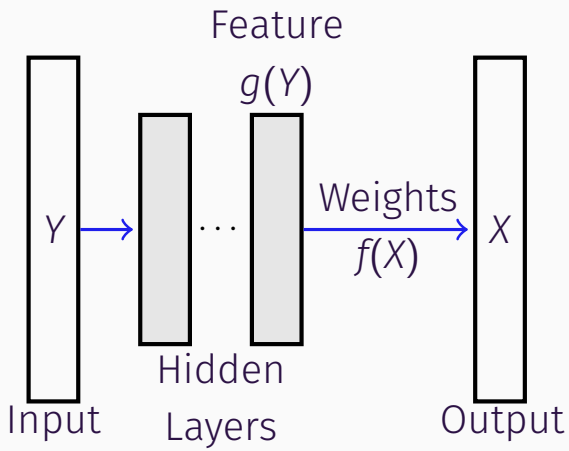
Use X to predict $Y \leftrightarrow \mathcal{M}_{\mathcal{E}_k^X}(P_{X,Y})$



$$\tilde{P}_{Y|X}(y|x) \propto \exp(f^{\mathrm{T}}(x)g(y) + b(y))$$

minimize $D(P_{X,Y} \| P_X \tilde{P}_{Y|X})$

Use Y to predict $X \leftrightarrow \mathcal{M}_{\mathcal{E}_k^Y}(P_{X,Y})$



$$\tilde{P}_{X|Y}(x|y) \propto \exp(g^{\mathrm{T}}(y)f(x) + a(x))$$

minimize $D(P_{X,Y} \| P_Y \tilde{P}_{X|Y})$

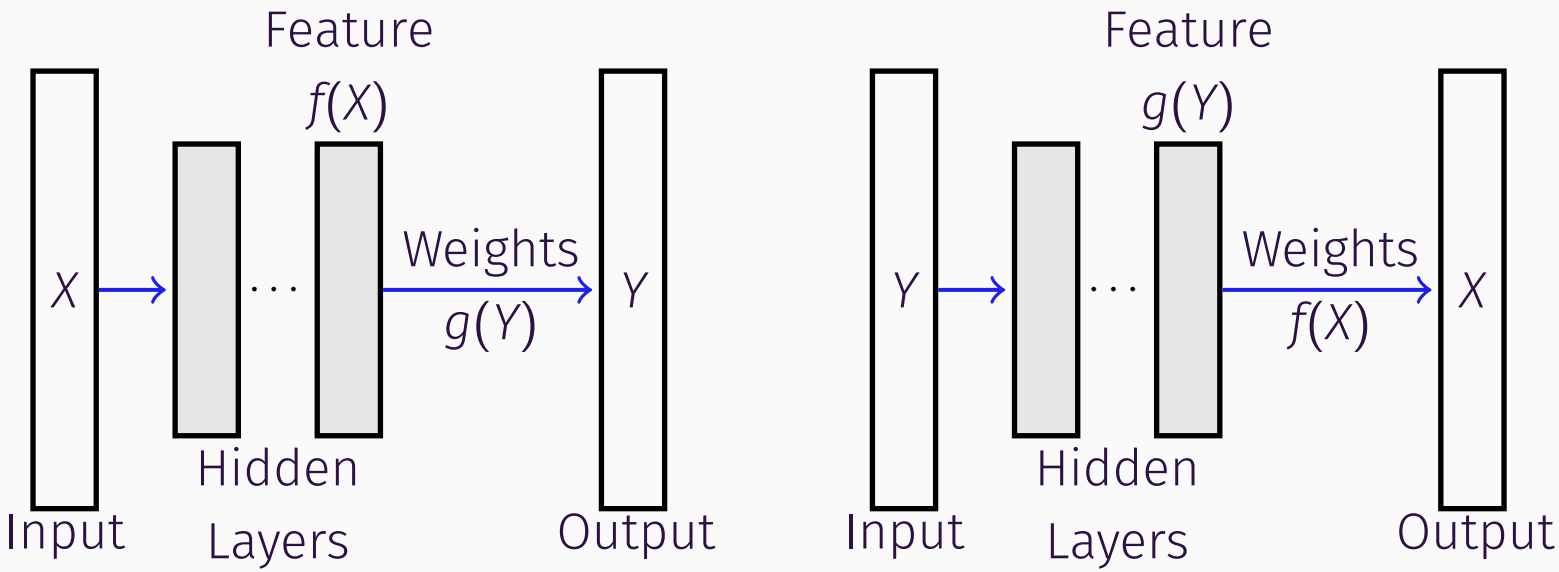
Generalized softmax learning $\leftrightarrow \mathcal{M}_{\mathcal{E}_k}(P_{X,Y})$

minimize $D(P_{X,Y} \| Q_{X,Y})$

$$Q_{X,Y}(x,y) \propto \exp(f^{\mathrm{T}}(x)g(y) + a(x) + b(y))$$

Symmetry in neural networks

Symmetry between a network and its reverse network



Solutions in local analysis regime

Local analysis regime

$P_{X,Y} \approx P_X P_Y$

The B matrix

$$B_{X,Y}(x,y) \triangleq \frac{P_{X,Y}(x,y)}{\sqrt{P_X(x)}\sqrt{P_Y(y)}}$$

Solutions

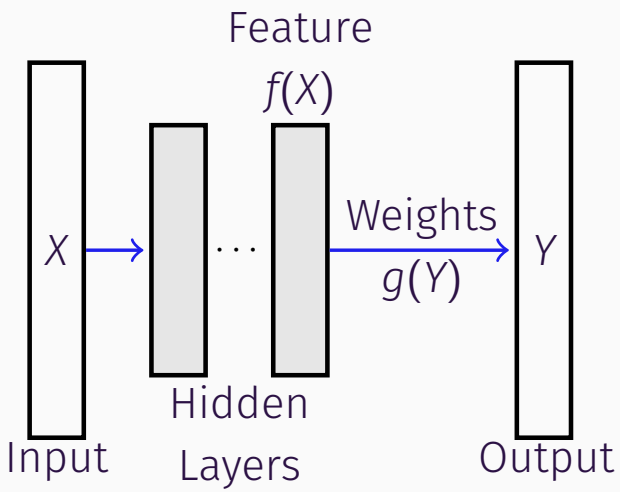
$f^\star(x) \leftrightarrow$ Left singular vectors of $B_{X,Y}$

$g^\star(y) \leftrightarrow$ Right singular vectors of $B_{X,Y}$

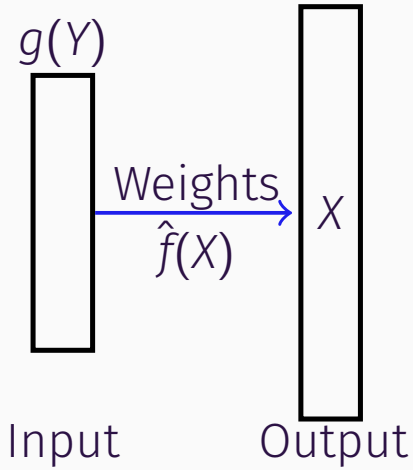
Simulation Results

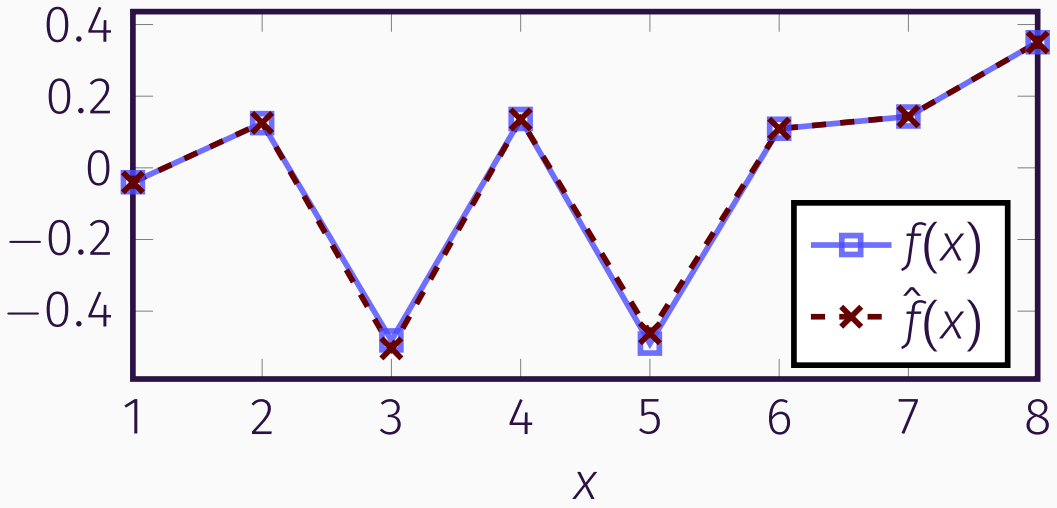
Simulation

Original network



Reverse network*





Summary

Summary

- Equivalence of softmax learning problems
- Discriminative $\leftrightarrow$ Generative
- Asymmetric $\leftrightarrow$ Symmetric
- Symmetry in neural networks