ON THE USEFULNESS OF STATISTICAL NORMALISATION OF BOTTLENECK FEATURES FOR SPEECH RECOGNITION

Erfan Loweimi, Peter Bell and Steve Renals

The Centre for Speech Technology Research (CSTR), University of Edinburgh

{e.loweimi, peter.bell, s.renals}@ed.ac.uk

Abstract

GOAL: An attempt to understand the DNNs from a statistical perspective

HOW: Statistical properties of bottleneck (BN) layer pre-activations (Z) and activations (Y) are studied

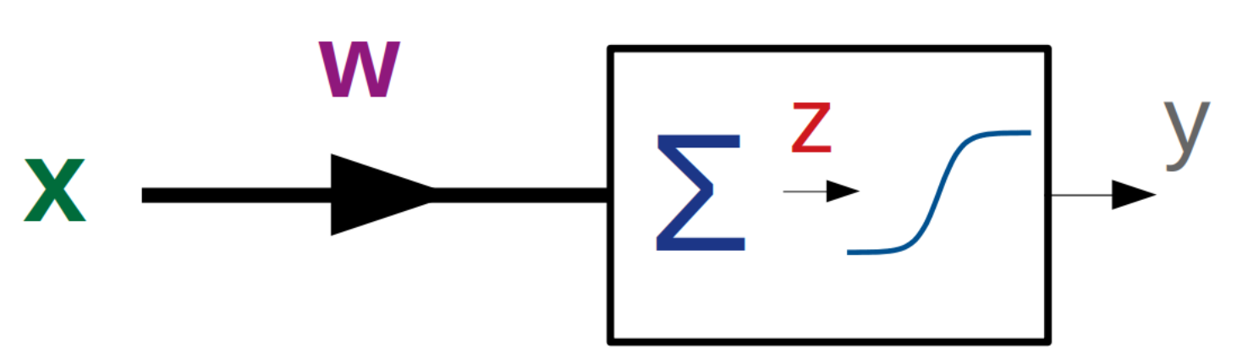
CONTRIBUTIONS:

1. Distribution of the NN activation in the BN layer was analytically derived
2. Statistical properties of the BN features were empirically studied and compared with analytic pdf
3. Sparsity of ReLU was (re-)explained
4. Post-processing of the BN features through statistical normalisation for ASR were investigated

EXPERIMENTS: Aurora-4, train by clean/additive

RESULTS: Up to 2% absolute (9% relative) performance gain (WER reduction) was achieved in mismatch condition

STATISTICAL DISTRIBUTION OF BOTTLENECK FEATURES



$$\mathrm{y} = \mathrm{f}(\mathbf{w}^T\mathbf{x}) = \mathbf{f}(\mathbf{z}) \Rightarrow \mathbf{z} = \mathbf{f}^{-1}(\mathbf{y})$$

$$\mathrm{P}_Y(y) = \left|\frac{d}{dy} f^{-1}(y)\right| P_Z(f^{-1}(y))$$

$$P_Y^{\tanh}(y) = \frac{1}{1-y^2} P_Z(\frac{1}{2}\log\frac{1+y}{1-y})$$

Assumptions for Approximating $P_Z(z)$

1. Central Limit Theorem (CLT)

$$\mathrm{z} \,\dot{\sim}\, \mathcal{N}(z; \mu_z, \sigma_z^2)$$

2. Prob(z > 0) ≈ Prob(z < 0)

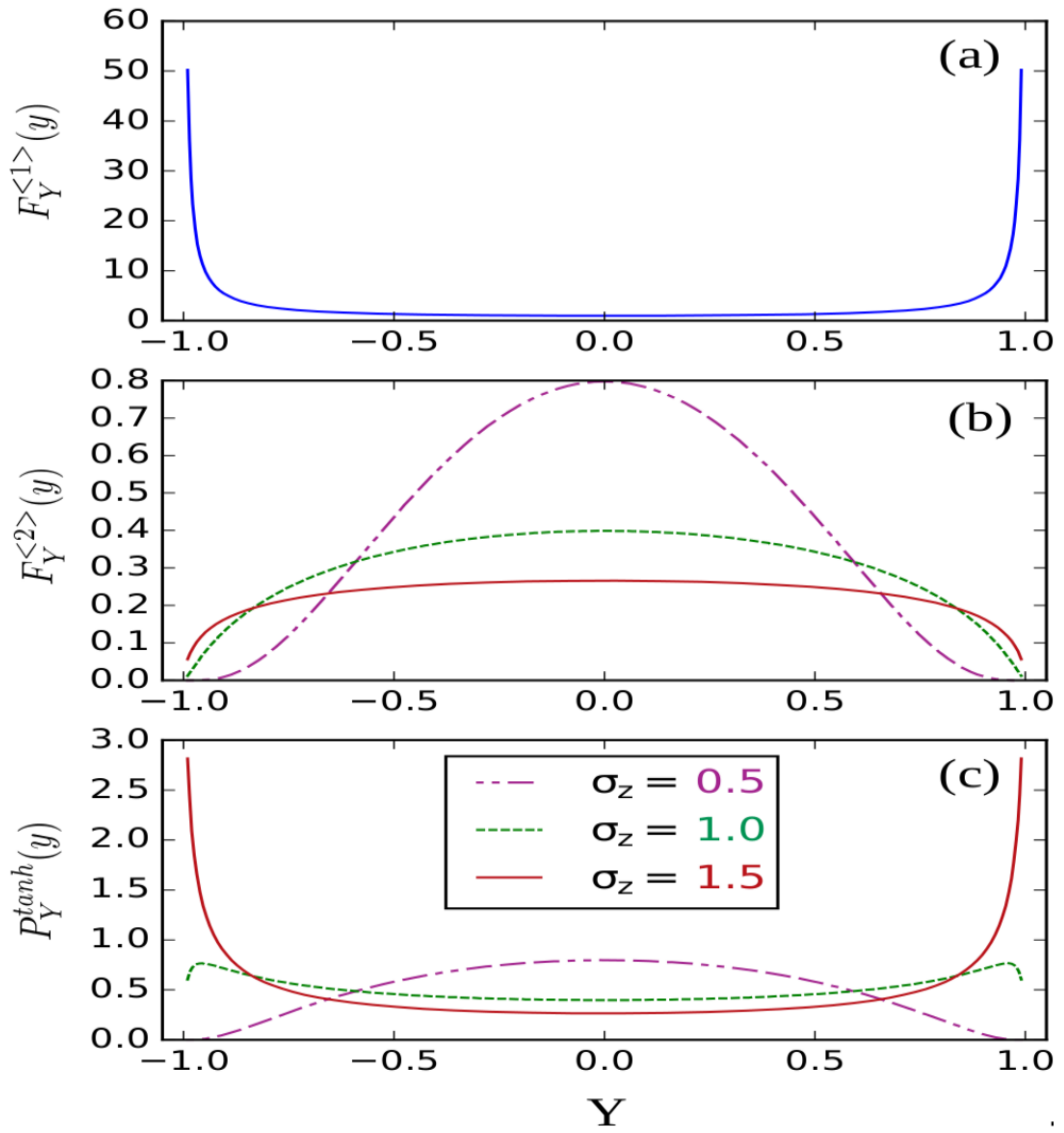
$$\mu_z \to 0 \qquad \mathrm{z} \,\dot{\sim}\, \mathcal{N}(z; 0, \sigma_z^2)$$

Density Estimating for Nodes with tanh Activation

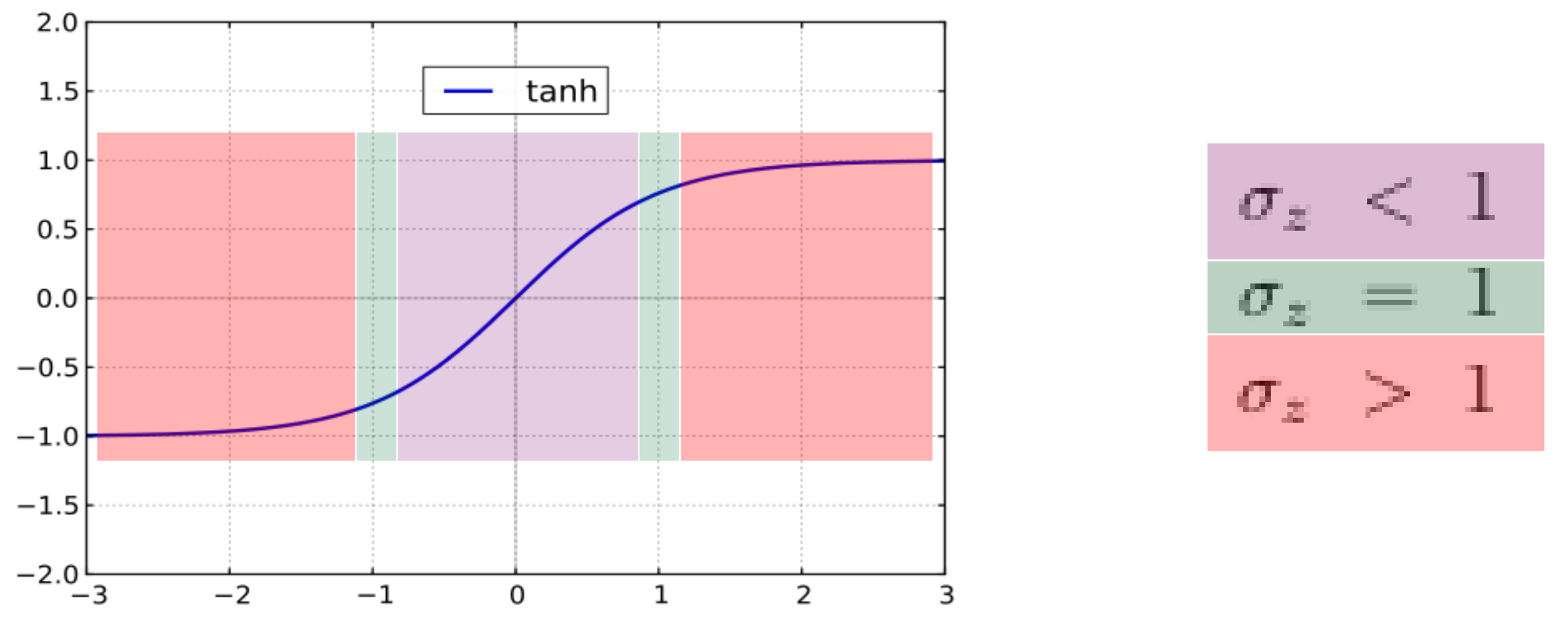
$$P_Y^{\tanh}(y) = \frac{1}{1-y^2}\mathcal{N}(\frac{1}{2}\log\frac{1+y}{1-y}; 0, \sigma_z^2))$$

$$= \underbrace{\frac{1}{1-y^2}}_{F_Y^{<1>}(y)} \underbrace{\frac{1}{\sqrt{2\pi}\sigma_z}(\frac{1+y}{1-y})^{-\frac{1}{8\sigma_z^2}\log\frac{1+y}{1-y}}}_{F_Y^{<2>}(y,\sigma_z)}$$

Factor 1 ← → ← Factor 2 → a function of σ_z →



Non-linearity of NNs and Density Shape Parameter

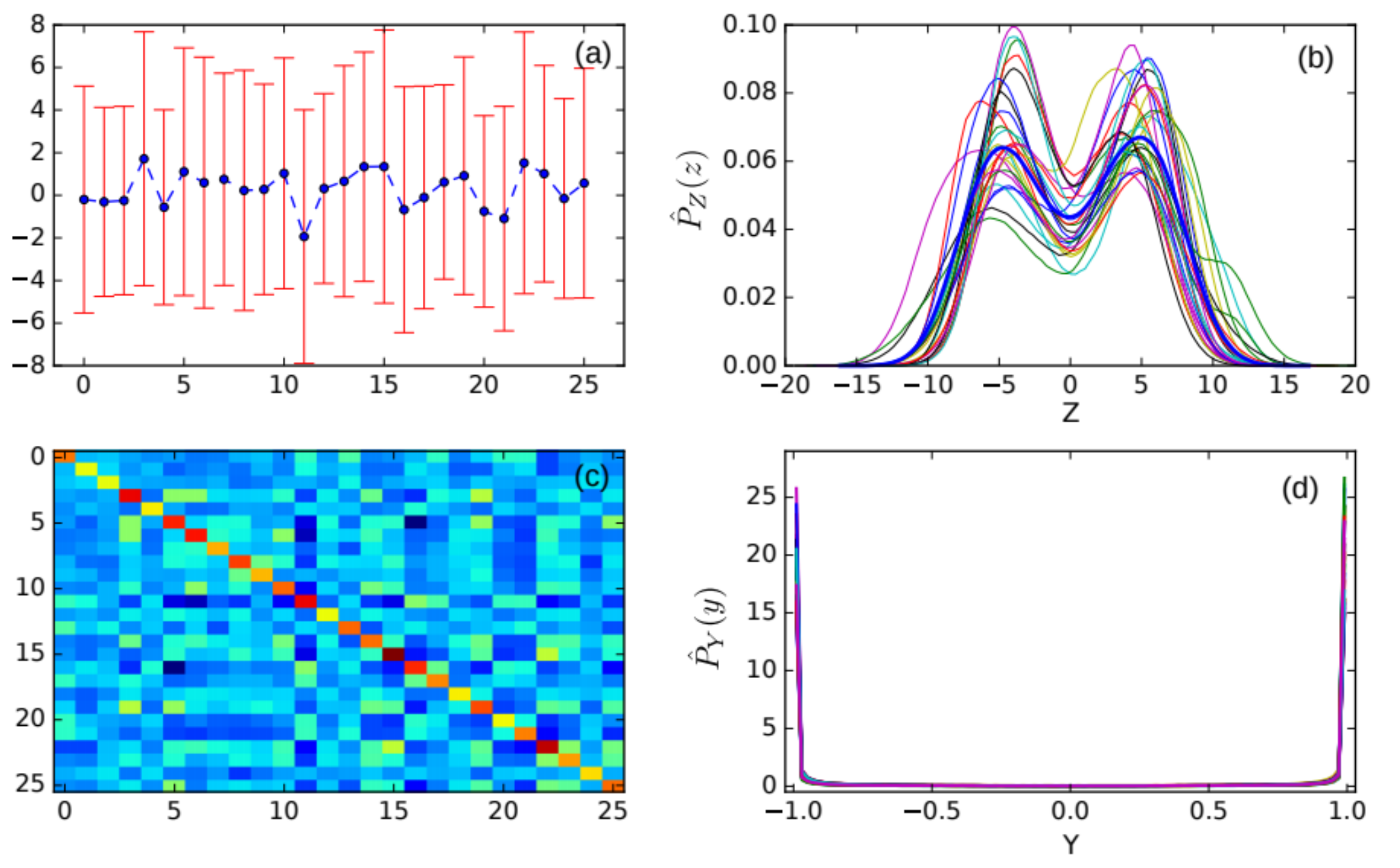


$\sigma_z < 1$ → Nodes/NN operates in linear mode

$\sigma_z > 1$ → Nodes/NN operates in non-linear mode

EMPIRICAL STUDIES

Density of Tanh



(1) Zero mean approximation for Z is reasonable

(2) σ_z > 1 ==>> DNN operates in the non-linear mode

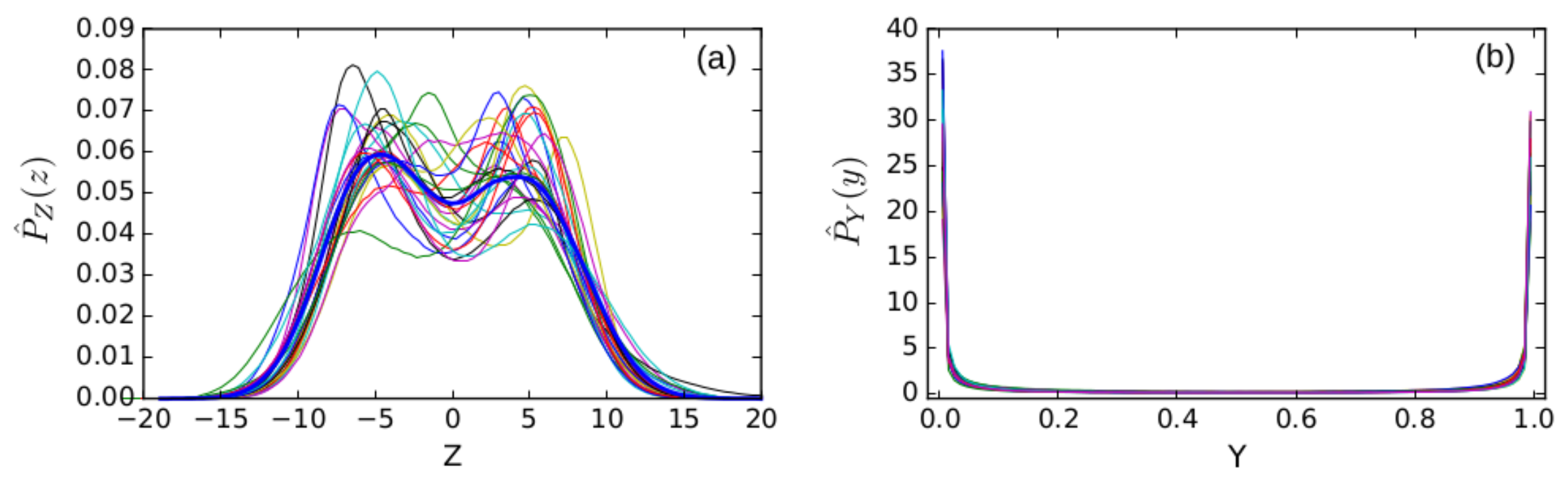
(3) Distribution of Z can be easily fitted by a GMM

(4) Distribution of Y mayNOT be fitted by a GMM

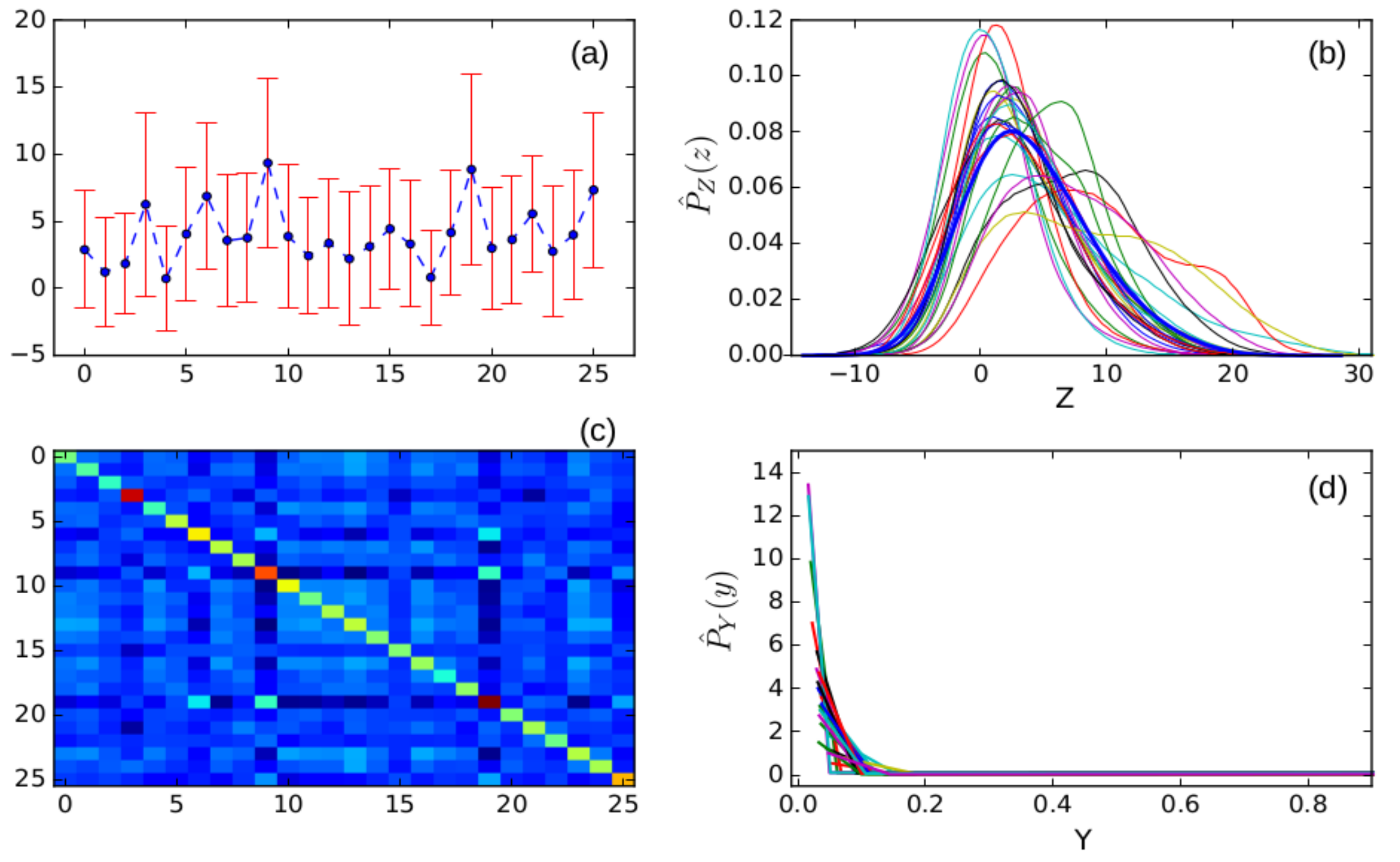
(5) DNN decorrelates the features in the BN layer

(6) Distribution of Y matches the derived equation

Density of Sigmoid



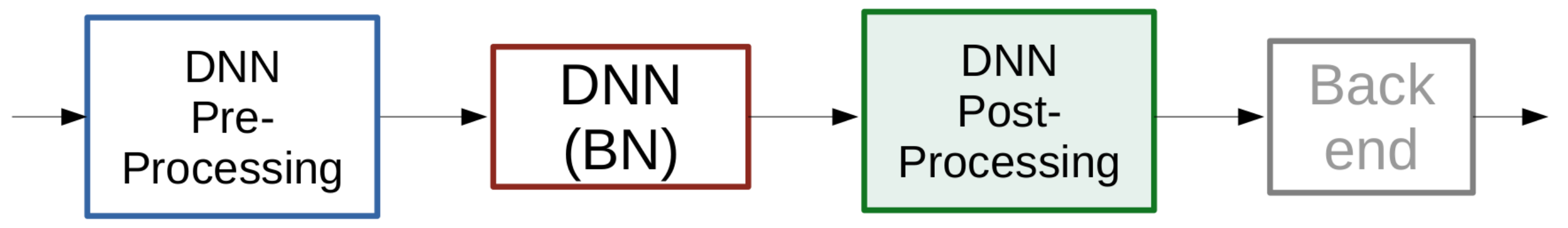
Sparsity of ReLU



* Glorot, et al, "Deep Sparse Rectifier Neural Networks", 2011
– 50% negative preacitivaitons → 50% of activations are 0

* Our argument: Coincidence of the positive zero (0^+) activation with the non-linear operating mode regions
– Before zero → Blocked; After zero → Linear

STATISTICAL NORMALISATION OF THE BOTTLENECK (BN) FEATURES



Experimental Results

Table 1: *WER for Aurora-4 (Kaldi-LDA-MLLT).*

Feature	A	B	C	D	*Ave4*
BN (baseline)	3.87	7.96	21.80	32.72	16.58
BN+MN	3.64	7.66	21.02	32.20	16.13
BN+MVN	4.07	8.31	20.34	33.04	16.44
BN+Gauss	4.15	8.12	20.18	32.67	16.28
BN+HEQ	3.96	7.43	19.76	30.87	15.50
BN+PCA	3.75	7.88	21.56	32.46	16.41
BN+DCT	3.77	7.77	21.76	32.49	16.44