# Learning Sparse Neural Networks Through Mixture-Distributed Regularization

Chang-Ti Huang[1], Jun-Cheng Chen[2], Ja-Ling Wu[1]
1. National Taiwan University, Taipei, Taiwan
2. Academia Sinica, Taipei, Taiwan
cthuang@cmlab.csie.ntu.edu.tw,pullpull@citi.sinica.edu.tw,wjl@cmlab.csie.ntu.edu.tw

## A. The gradient variance of the MDR

In the following subsections, we show both the qualitative and quantitative analysis for the proposed MDR method.

### A.1. Qualitative visualization

As shown in the third column of Fig. (3), we can qualitatively verify that the gradient variances of both the power-law function distribution and the exponential-uniform distribution are lower than the gradient variance of the exponential distribution.

### A.2. Quantitative analysis

As for the three MDRs compared with the Concrete distribution w.r.t. the gradient variance, it is hard to distinguish the pros and cons from Fig. (3). Hence, we quantitatively validate on the gradient variances of both the MDR and the Concrete distribution using a toy experiment. We generate $10^6$ samples drawn from the PDF of the MDRs, i.e.:

$$Q(\zeta) = (1-q)\, r\left(\zeta | z=0\right) + q r\left(\zeta | z=1\right), \tag{1}$$

where $q$ is the mixture weight associated with 1-component, setting $q = 0.5$. Besides, we also generate $10^6$ samples from the PDF of the Concrete distribution, setting $\log \alpha = 0$. The experiment consists of two measurements, one of which is the gradient variance of drawn samples $\boldsymbol{\zeta}$ and the other is the average $L_1$ distance $\|\boldsymbol{\zeta} - \boldsymbol{z}\|_1$, where

$$z_i = \begin{cases} 0 & \text{if } \zeta_i \leq 0.5 \\ 1 & \text{if } \zeta_i > 0.5 \end{cases}. \tag{2}$$

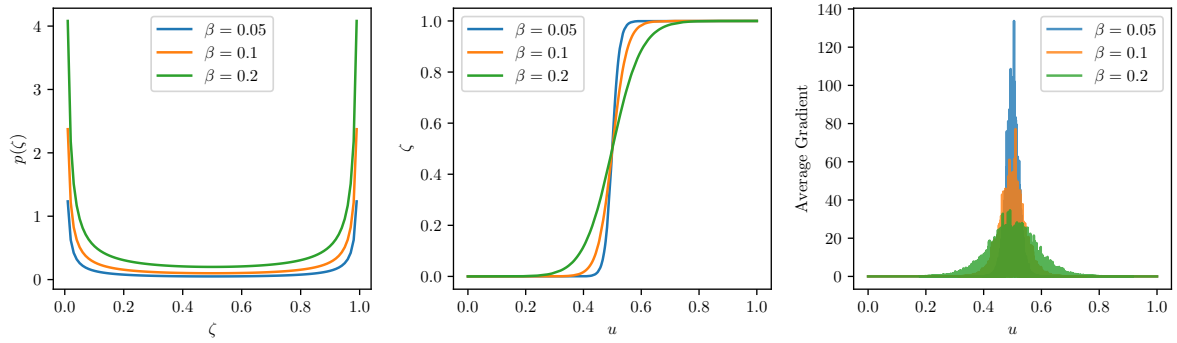
Each curve in Fig. (4) is plotted with 8 dots, which indicate different temperatures $\beta \in \{40, 34, 28, 22, 16, 10, 7, 4\}$ for the Concrete distribution, $\beta \in \{8, 9, 10, ..., 15\}$ for the exponential distribution, $\beta \in \{9, 10, 11, ..., 16\}$ for the exponential-uniform distribution, and $\beta \in \{10, 20, 30, ..., 80\}$ for the power-law function distribution, respectively.

There is a trade-off between the gradient variance and the estimated binary gate $\zeta$, as shown in Fig. (4). As the gradient variance becomes smaller (i.e. decreasing $\beta$ for the MDRs), $\zeta$ tends to be biased toward the original binary gate $z$. Hence, the choice of $\beta$ is still an empirical study. However, from Fig. (4) we observe that for a given gradient variance, the power-law function distribution provides more accurate estimates to the binary gate $z$.

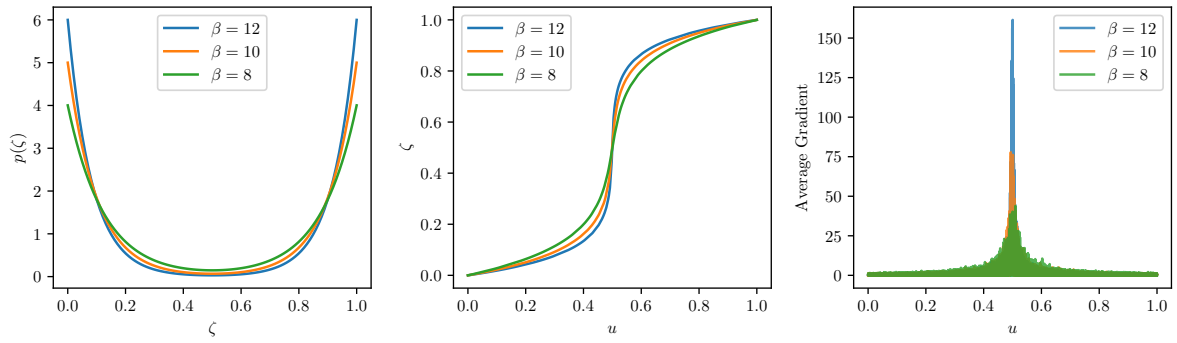
## B. Deriving the reparameterization method for the mixture distribution

Let $r(\zeta|z)$ be the PDF of the mixture component and $q$ be the mixture weight where $q \equiv q(z=1|\theta)$. Applying the definition of a univariate CDF to the components $z = \{0, 1\}$ yields

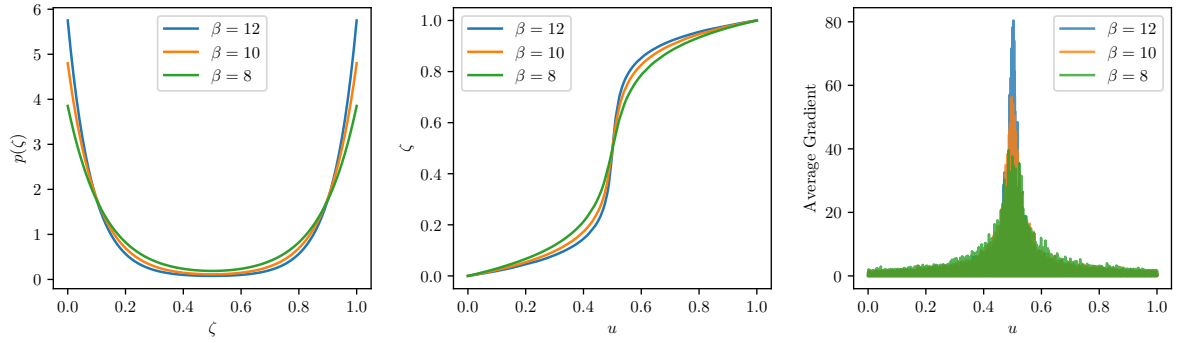
$$\text{CDF}^{\text{mix}}(\zeta) = \underbrace{(1-q)\int_{t=-\infty}^{\zeta} r(t|z=0)dt}_{\text{0-component}} + \underbrace{q\int_{t=-\infty}^{\zeta} r(t|z=1)dt}_{\text{1-component}} = u, \tag{3}$$



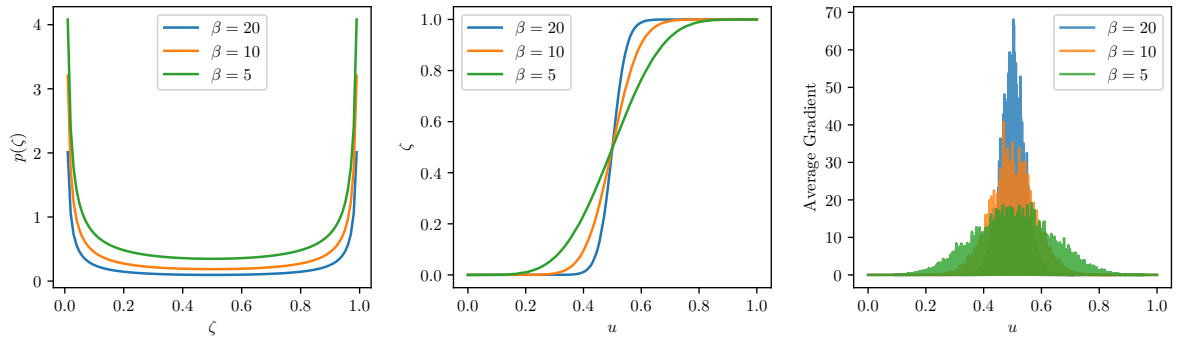

(a) Concrete




(b) Exponential




(c) Exponential-Uniform




(d) Power-Law

Figure 3: (a) indicating the Concrete distribution and (b)(c)(d) corresponding to different mixture distributions. We visualize the PDF (first column) and the inverse CDF (second column) of the mixture distribution $q(\zeta|\theta) = \sum_{z_i} q(z_i|\theta) r(\zeta|z_i)$ where $q(z = 1|\theta) = 0.5$. In the third column, we take the gradient of $\zeta$ w.r.t. $q$ from the inverse CDF. Observing the third column, we see that the variance of the gradient increases with increasing $\beta$. Further, (c) and (d) generally have lower gradient variance estimates than that of (b).

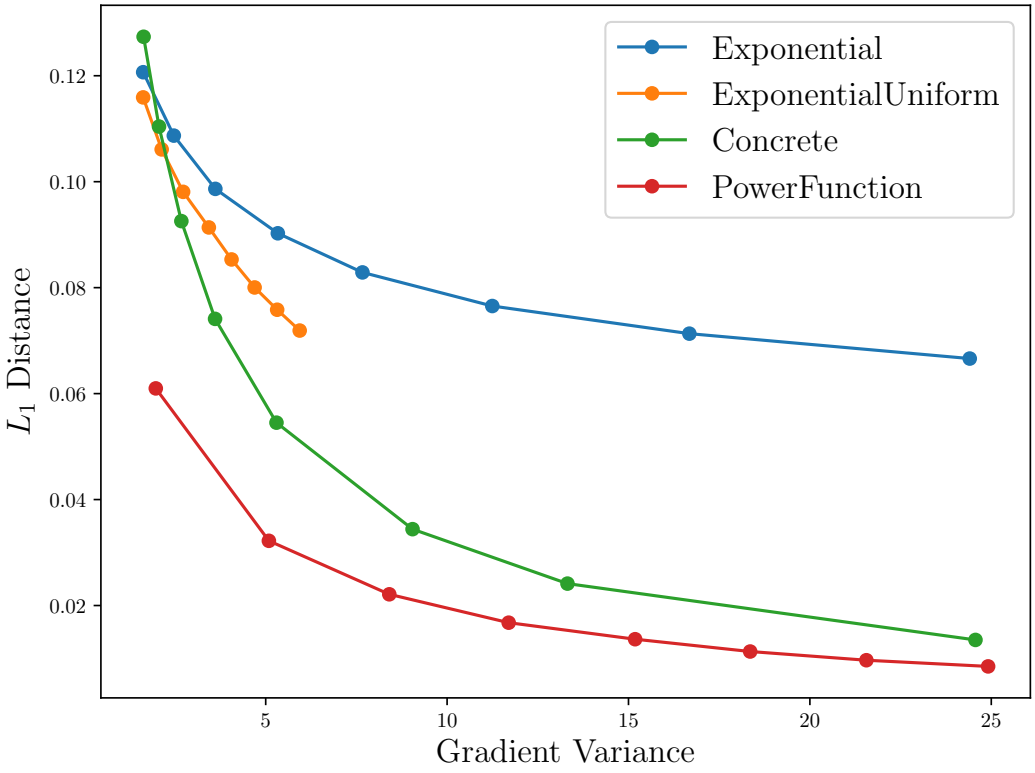



Figure 4: Visualization on the gradient variance (horizontal axis) and the average $L_1$ distance (vertical axis) of the MDRs and the Concrete. Dots per MDR curve indicate different $\beta$ values increasing from left to right, whereas dots in the Concrete curve refer to different $\beta$ values decreasing from left to right.

where $u \sim \text{Uniform}(0, 1)$. Here, we take the gradient w.r.t. $q$ from both sides of Eq. (3) giving

$$\frac{\partial \text{CDF}^{\text{mix}}(\zeta)}{\partial q} = \frac{\partial u}{\partial q} \tag{4}$$

$$\Rightarrow \frac{\partial(1-q)}{\partial q} \int_{t=-\infty}^{\zeta} r\left(t|z=0\right) dt + (1-q) \frac{\partial}{\partial q} \left( \int_{t=-\infty}^{\zeta} r\left(t|z=0\right) dt \right) \tag{5}$$

$$+ \frac{\partial q}{\partial q} \int_{t=-\infty}^{\zeta} r\left(t|z=1\right) dt + \quad q \quad \frac{\partial}{\partial q} \left( \int_{t=-\infty}^{\zeta} r\left(t|z=1\right) dt \right) \tag{6}$$

$$= 0, \tag{7}$$

where Eq. (5) and Eq. (6) are given by respectively differentiating the 0- and the 1- components in Eq. (3). Differentiating the mixture weights and computing the integration over the PDFs, we can rewrite the first term in Eq. (5) as

$$\frac{\partial(1-q)}{\partial q} \int_{t=-\infty}^{\zeta} r\left(t|z=0\right) dt = -R\left(\zeta|z=0\right) \tag{8}$$

and the first term in Eq. (6) as

$$\frac{\partial q}{\partial q} \int_{t=-\infty}^{\zeta} r\left(t|z=1\right) dt = R\left(\zeta|z=1\right), \tag{9}$$

where $R$ are the CDFs of the mixture components. As for the second term in Eq. (5), we apply the general form of Leibniz integral rule to it:

$$\frac{\partial}{\partial q}\left(\int_{t=-\infty}^{\zeta} r\left(t|z=0\right)dt\right) = r\left(\zeta|z=0\right)\frac{\partial \zeta}{\partial q} - r\left(-\infty|z=0\right)\frac{\partial(-\infty)}{\partial q} + \int_{t=-\infty}^{\zeta}\frac{\partial}{\partial q} r\left(t|z=0\right)dt. \tag{10}$$

Observing that the derivatives of the second and the third terms on the right-hand side of Eq. (10) equal to zero, we can summarize the above as

$$\frac{\partial}{\partial q}\left(\int_{t=-\infty}^{\zeta} r\left(t|z=0\right)dt\right) = r\left(\zeta|z=0\right)\frac{\partial \zeta}{\partial q}, \tag{11}$$

and hence the second term of Eq. (6) becomes

$$\frac{\partial}{\partial q}\left(\int_{t=-\infty}^{\zeta} r\left(t|z=1\right)dt\right) = r\left(\zeta|z=1\right)\frac{\partial \zeta}{\partial q}. \tag{12}$$

Now, we are ready to substituting Eqs. (8), (9), (11) and (12) into Eqs. (5) and (6), giving

$$\frac{\partial \text{CDF}^{\text{mix}}(\zeta)}{\partial q} = -R\left(\zeta|z=0\right) + (1-q)r\left(\zeta|z=0\right)\frac{\partial \zeta}{\partial q} \tag{13}$$

$$+R\left(\zeta|z=1\right) + \quad q \quad r\left(\zeta|z=1\right)\frac{\partial \zeta}{\partial q} = 0. \tag{14}$$

After some rearrangements, we have

$$\frac{\partial \zeta}{\partial q}\left[(1-q)r\left(\zeta|z=0\right) + qr\left(\zeta|z=1\right)\right] + R\left(\zeta|z=1\right) - R\left(\zeta|z=0\right) = 0, \tag{15}$$

and the gradient w.r.t. $q$ can be obtained as

$$\frac{\partial \zeta}{\partial q} = \frac{R\left(\zeta|z=0\right) - R\left(\zeta|z=1\right)}{(1-q)r\left(\zeta|z=0\right) + qr\left(\zeta|z=1\right)}. \tag{16}$$

As shown in Eq. (16), given the PDF and the CDF of the mixture components, we can take gradients of $\zeta$ w.r.t. $q$. Reparameterization for the MDR is highly required when the inverse CDF of the mixture distribution is intractable. In that case, we can not sample $\zeta$ through inverse transform sampling, instead we optimize the MDR through backpropagating $\frac{\partial \zeta}{\partial q}$.

With this reparameterization method, we can also take the gradient w.r.t. other model parameters (e.g. $\beta$) in a similar way, i.e.:

$$\frac{\partial \text{CDF}^{\text{mix}}(\zeta)}{\partial \beta} = \frac{\partial u}{\partial \beta} \tag{17}$$

$$\Rightarrow \frac{\partial(1-q)}{\partial \beta}\int_{t=-\infty}^{\zeta} r\left(t|z=0\right)dt + (1-q)\frac{\partial}{\partial \beta}\left(\int_{t=-\infty}^{\zeta} r\left(t|z=0\right)dt\right) \tag{18}$$

$$+\frac{\partial q}{\partial \beta}\int_{t=-\infty}^{\zeta} r\left(t|z=1\right)dt + \quad q \quad \frac{\partial}{\partial \beta}\left(\int_{t=-\infty}^{\zeta} r\left(t|z=1\right)dt\right) \tag{19}$$

$$= 0. \tag{20}$$

Observing that $\frac{\partial(1-q)}{\partial\beta} = 0$ and $\frac{\partial q}{\partial\beta} = 0$, we can rewrite the first term in Eq. (18) as

$$\frac{\partial(1-q)}{\partial\beta}\int_{t=-\infty}^{\zeta} r\left(t|z=0\right)dt = 0, \tag{21}$$

and the first term in Eq. (6) as

$$\frac{\partial q}{\partial\beta}\int_{t=-\infty}^{\zeta} r\left(t|z=1\right)dt = 0. \tag{22}$$

Considering the second term in Eq. (18), we also apply the general form of Leibniz integral rule giving

$$\begin{aligned}\frac{\partial}{\partial\beta}\left(\int_{t=-\infty}^{\zeta} r\left(t|z=0\right)dt\right) = r\left(\zeta|z=0\right)\frac{\partial\zeta}{\partial\beta} - r\left(-\infty|z=0\right)\frac{\partial(-\infty)}{\partial\beta} \\ + \int_{t=-\infty}^{\zeta}\frac{\partial}{\partial\beta}r\left(t|z=0\right)dt.\end{aligned} \tag{23}$$

It is trivial that $\frac{\partial(-\infty)}{\partial\beta} = 0$. However, the third term on the right-hand side of Eq. (10) equals to zero since $r\left(t|z=0\right)$ does not depend on $q$, whereas the third term on the right-hand side of Eq. (23) contains the dependency between $r\left(t|z=0\right)$ and $\beta$. To sum up, Eq. (23) can be rewritten as

$$\frac{\partial}{\partial\beta}\left(\int_{t=-\infty}^{\zeta} r\left(t|z=0\right)dt\right) = r\left(\zeta|z=0\right)\frac{\partial\zeta}{\partial\beta} + \int_{t=-\infty}^{\zeta}\frac{\partial}{\partial\beta}r\left(t|z=0\right)dt. \tag{24}$$

The same rule can be applied to the second term of Eq. (19) giving

$$\frac{\partial}{\partial\beta}\left(\int_{t=-\infty}^{\zeta} r\left(t|z=1\right)dt\right) = r\left(\zeta|z=1\right)\frac{\partial\zeta}{\partial\beta} + \int_{t=-\infty}^{\zeta}\frac{\partial}{\partial\beta}r\left(t|z=1\right)dt. \tag{25}$$

Finally, after some rearrangements, we have

$$\frac{\partial\zeta}{\partial\beta}\left[(1-q)r\left(\zeta|z=0\right) + qr\left(\zeta|z=1\right)\right] + \frac{\partial}{\partial\beta}R\left(\zeta|z=0\right) + \frac{\partial}{\partial\beta}R\left(\zeta|z=1\right) = 0, \tag{26}$$

and the gradient w.r.t. $\beta$ can now be written as

$$\frac{\partial\zeta}{\partial\beta} = -\frac{(1-q)\frac{\partial}{\partial\beta}R\left(\zeta|z=0\right) + q\frac{\partial}{\partial\beta}R\left(\zeta|z=1\right)}{(1-q)r\left(\zeta|z=0\right) + qr\left(\zeta|z=1\right)}. \tag{27}$$

## C. Visualization of the emulated binary gates $\zeta$

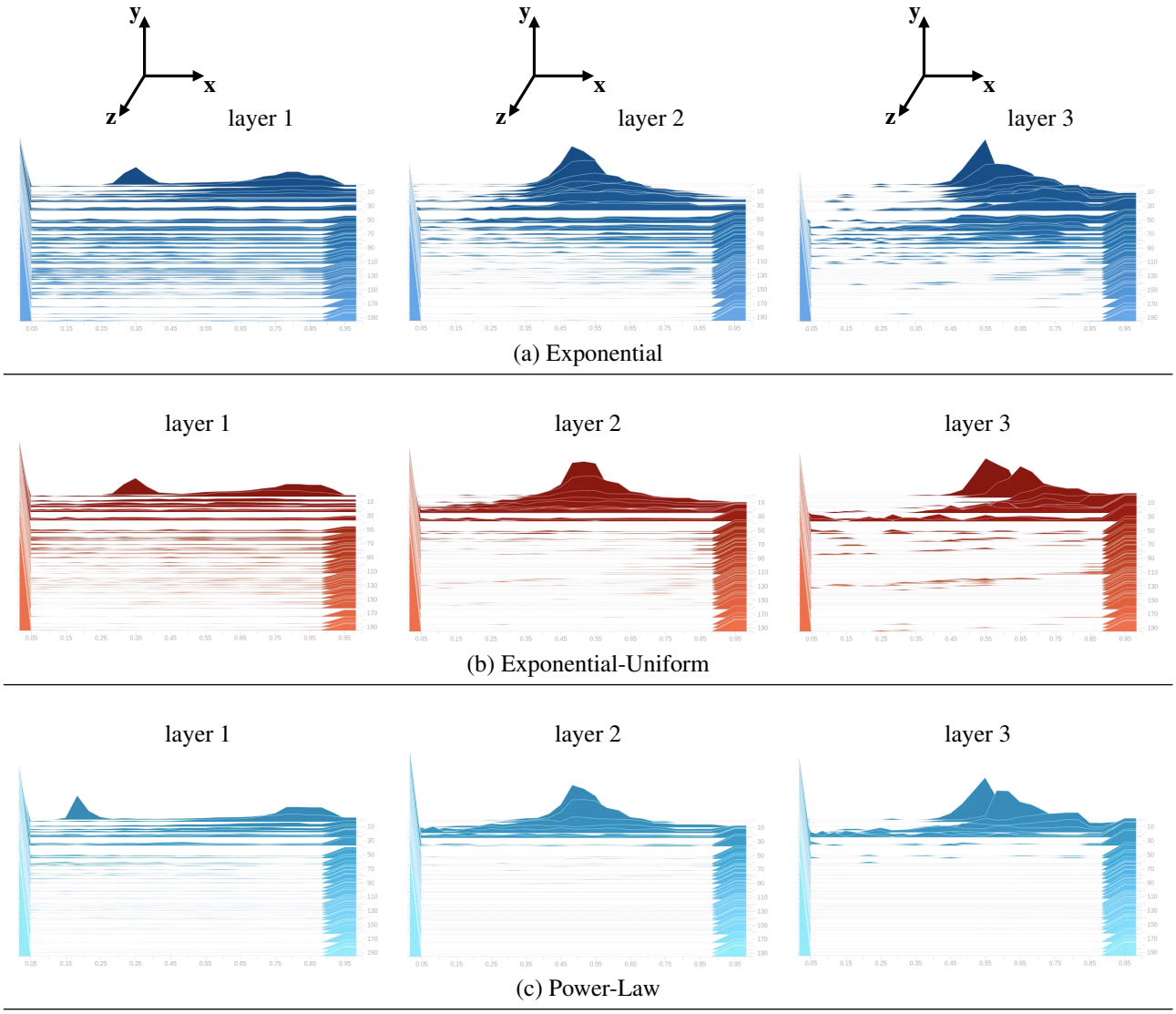



Figure 5: Visualization of the emulated binary gates $\zeta$ using LeNet-300-100 adapted on MNIST. Each column indicates the corresponding layer of LeNet-300-100. Each row refers to the mixture distribution used. All of the mixture distributions converge to Bernoulli-like distributions yet still allow for reparameterization. The x- and y- axes in each subfigure represent the histogram of the drawn $\zeta$, while the z-axis indicates the training epochs. As we can see, $\zeta$ drawn from the MDR converges to $\delta(\zeta = 0)$ and $\delta(\zeta = 1)$ with extremely high mixing proportion at the 0-component.

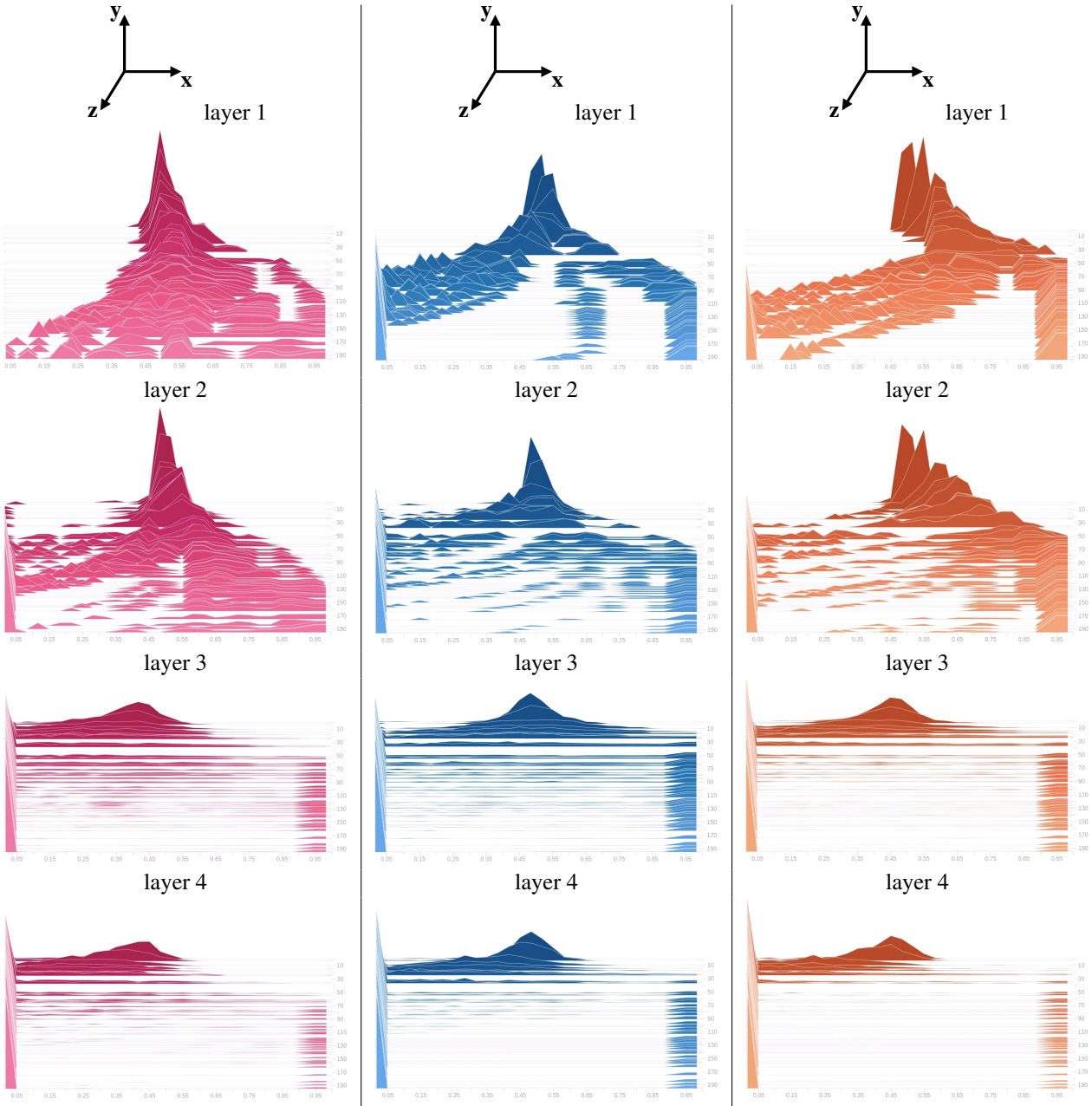



Figure 6: Visualization of the emulated binary gates $\zeta$ using LeNet-5-Caffe adapted on MNIST. Each row indicates the corresponding layer of LeNet-5-Caffe. Each column refers to the mixture distribution used. The x- and y- axes in each subfigure represent the histogram of the drawn $\zeta$, while the z-axis indicates the training epochs. Note that the reported pruned-architecture of MDR-Exp is inferred by the trained architecture, not the estimate $\zeta^*$. As we can observe, the convolutional layers converge much slower than the fc layers; however, $\zeta$ in general still converges to $\delta(\zeta = 0)$ and $\delta(\zeta = 1)$ at the end of training, with high mixing proportion of zero.