Analytic Expressions for Probabilistic Moments of PL-DNN with Gaussian Input (Supplementary Material)

Adel Bibi*, Modar Alfadly,* and Bernard Ghanem

King Abdullah University of Science and Technology (KAUST), Saudi Arabia

We start first by proving lemma 2.

Lemma 1. *Let* $\mathbf{x} \in \mathbb{R}^n \sim \mathcal{N}(\mu_x, \Sigma_x)$ *for any even* p*, where* $\sigma_{ij} = \Sigma_x(i,j) \forall i \neq j$*. Then for any arbitrary nonlinear map* $\Psi : \mathbb{R}^n \to \mathbb{R}$ *the following holds* $\frac{\partial^{\frac{p}{2}} \mathbb{E}[\Psi(\mathbf{x})]}{\prod_{\forall odd i} \partial \sigma_{ii+1}} = \mathbb{E}[\frac{\partial^p \Psi(\mathbf{x})}{\partial x_1 \dots \partial x_p}]$.

Proof. First we define the characteristic function and it's inverse Fourier Transform of the joint Gaussian as follows ($f(x_1, \dots, x_n)$):

$$\Phi(w_1, \dots, w_n) = \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{n} f(x_1, \dots, x_n) e^{j(w_1 x_1 + \cdots + w_n x_n)} dx_1 \dots dx_n$$

$$f(x_1, \dots, x_n) = \left(\frac{1}{2\pi}\right)^n \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{n} \Phi(w_1, \dots, w_n) e^{-j(w_1 x_1 + \cdots + w_n x_n)} dw_1 \dots dw_n$$

$$\frac{\partial^n f(x_1, \dots, x_n)}{\partial x_1 \dots \partial x_n} = \left(\frac{1}{2\pi}\right)^n \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{n} \Phi(w_1, \dots, w_n)(-j)^n \prod_i^n (w_i)$$

$$e^{-j(w_1 x_1 + \cdots + w_n x_n)} dw_1 \dots dw_n \tag{1}$$

Note that

$$\Phi(w_1, \dots, w_n) = e^{\left(\mathbf{w}^\top \mu_x\right) i - \frac{1}{2} \mathbf{w}^\top \Sigma_x \mathbf{w}} = e^{\left(\mathbf{w}^\top \mu_x\right) - \frac{1}{2} \sum_i^n \sum_j^n w_i w_j \Sigma_x(i,j)} \tag{2}$$

Then

$$\frac{\partial^{\frac{n}{2}} \Phi}{\prod_{\forall \text{odd} i} \partial \rho_{ii+1}} = \prod_{\forall \text{odd} i}^{\frac{n}{2}} -\frac{1}{2}(2 w_i w_{i+1}) \Phi(w_1, \dots, w_n) = (-1)^{\frac{n}{2}} \prod_i^n w_i \Phi(w_1, \dots, w_n) \tag{3}$$

Then:

$$\begin{aligned}\mathbb{E}[\Psi(x_1,\dots,x_n)] &= \underbrace{\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}}_{n}\Psi(x_1,\dots,x_n)f(x_1,\dots,x_n)dx_1\dots dx_n \\ &= \left(\frac{1}{2\pi}\right)^n \underbrace{\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}}_{2n}\Phi(w_1,\dots,w_n) \\ &\quad e^{-j(w_1x_1+\cdots+w_nx_n)}\Psi(x_1,\dots,x_n)dw_1\dots dw_ndx_1\dots dx_n\end{aligned} \tag{4}$$

Now by applying the theorem we get:

$$\begin{aligned}\frac{\partial^{\frac{n}{2}}\mathbb{E}[\Psi(x_1,\dots,x_n)]}{\prod_{\forall \text{odd} i}\partial\rho_{ii+1}} &= \left(\frac{1}{2\pi}\right)^n \underbrace{\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}}_{2n}\frac{\partial^{\frac{n}{2}}\Phi(w_1,\dots,w_n)}{\prod_{\forall \text{odd} i}\partial\rho_{ii+1}} \\ &\quad e^{-j(w_1x_1+\cdots+w_nx_n)}\Psi(x_1,\dots,x_n)dw_1\dots dw_ndx_1\dots dx_n \\ &= \left(\frac{1}{2\pi}\right)^n \underbrace{\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}}_{2n}(-1)^{\frac{n}{2}}\prod_{i}^{p}w_i\Phi(w_1,\dots,w_n) \\ &\quad e^{-j(w_1x_1+\cdots+w_nx_n)}\Psi(x_1,\dots,x_n)dw_1\dots dw_pdx_1\dots dx_n \\ &\overset{(i)}{=} \underbrace{\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}}_{n}\Psi(x_1,\dots,x_n)\frac{\partial^n f(x_1,\dots,x_n)}{\partial x_1\dots\partial x_n}dx_1\dots dx_n \\ &= \underbrace{\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}}_{n}\frac{\partial^n\Psi(x_1,\dots,x_n)}{\partial x_1\dots\partial x_n}f(x_1,\dots,x_n)dx_1\dots dx_n\end{aligned} \tag{5}$$

The equality (i) is only true for when $(-1)^{\frac{n}{2}} = (-j)^n$ (n is even). As for the last equality it holds since the Gaussian probability density function $f(x_1,\dots,x_n)$ is in Schwarz class, then the last equality holds by integrating by parts n times where n is even.■

For clarity, we will demonstrate the lemma with an example. This example following comments and other examples by Price [1], will help in understanding why PL-DNNs are best suited for such an application to the Lemma.

Example: 1 Consider the following nonsmooth function $g(x, y, z) = xy\text{sign}(z)$, where the random variables (x,y,z) are jointly Gaussian with unit variance and we wish to find the analytic expression for $\mathbb{E}[g(x, y, z)]$. By applying lemma 1 (note one can apply price theorem since the variances are identities and since the function is written in a product form [1] while with lemma 1 we don't maintain tan of these assumptions in general) and taking the variables (x, z) under differeitation we have

$$\begin{aligned}\mathbb{E}\Big[\frac{\partial^2 g}{\partial x \partial z}\Big] &= \mathbb{E}\Big[2y\delta(z)\Big] = \int_{-\infty}^{\infty} 2y f_{y,z}(y,0)dy \\ &= \frac{2}{2\pi\sigma_y\sigma_z\sqrt{(1-\rho_{yz}^2)}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}\frac{1}{(1-\rho_{yz}^2)}(y-(\mu_y-\rho_{yz}\mu_z))^2} e^{-\frac{1}{2}\mu_z^2} dy \\ &= \frac{2}{2\pi\sqrt{(1-\rho_{yz}^2)}} e^{-\frac{1}{2}\mu_z^2}\sqrt{2\pi}\sqrt{1-\rho_{yz}^2}(\mu_y-\rho_{yz}\mu_z) \\ &= \sqrt{\frac{2}{\pi}} e^{-\frac{1}{2}\mu_z^2}(\mu_y-\rho_{yz}\mu_z) = \frac{\partial \mathbb{E}[g]}{\partial \rho_{xz}}\end{aligned} \tag{6}$$

Then $\mathbb{E}[g] = \sqrt{\frac{2}{\pi}} e^{-\frac{1}{2}\mu_z^2}(\mu_y - \rho_{yz}\mu_z)\rho_{xz} + C$. When x, y, z are independent, the $\rho_{ij} = 0 \forall i, j$. $\mathbb{E}[g] = \mathbb{E}[x]\mathbb{E}[y]\mathbb{E}[\text{sign}(z)] = \mu_x\mu_y\text{erf}(\frac{\mu_z}{\sqrt{2}}) = C$ Therefore.

$$\mathbb{E}[g] = \sqrt{\frac{2}{\pi}} e^{-\frac{1}{2}\mu_z^2}(\mu_y - \rho_{yz}\mu_z)\rho_{xz} + \mu_x\mu_y\text{erf}(\frac{\mu_z}{\sqrt{2}}) \tag{7}$$

As evident from the previous result, in PL-DNNs setting the RHS of lemma 1 would reduce to dirac functions just like in the previous example with $sign(\mathbf{z})$. This results in an easier computation of the RHS of lemma 2.

Tightness on Synthetic Networks. The following table lists the details of the network architectures for both the fully connected and convolutional networks. The input to the fully connected layers is $\in \mathbb{R}^{100}$ (FC-A, FC-B and FC-C) while it is $\in \mathbb{R}^{20\times 20}$ for the convolutional networks (Conv-A, Conv-B and Conv-C). Note that all networks have a fully connected layer at the end converting all outputs to a single function.

Table 1. Shows the synthetic fully connected (on the left) and convolutional (on the right) network configurations used in the experiments. Note that the convolutional network parameters are denoted as "conv(filter size - number of channels)".

FC-A	FC-B	FC-C
FC-80	FC-80	FC-80
FC-1	FC-60	FC-60
-	FC-1	FC-40
-	-	FC-1

Conv-A	Conv-B	Conv-C
conv1x5	conv1x5	conv1x5
FC-1	conv1x10	conv1x10
-	FC-1	conv1x15
-	-	FC-1

Fully Connected. First we show the tightness of the analytic expressions on all fully connected networks under 3 different levels of input variance.

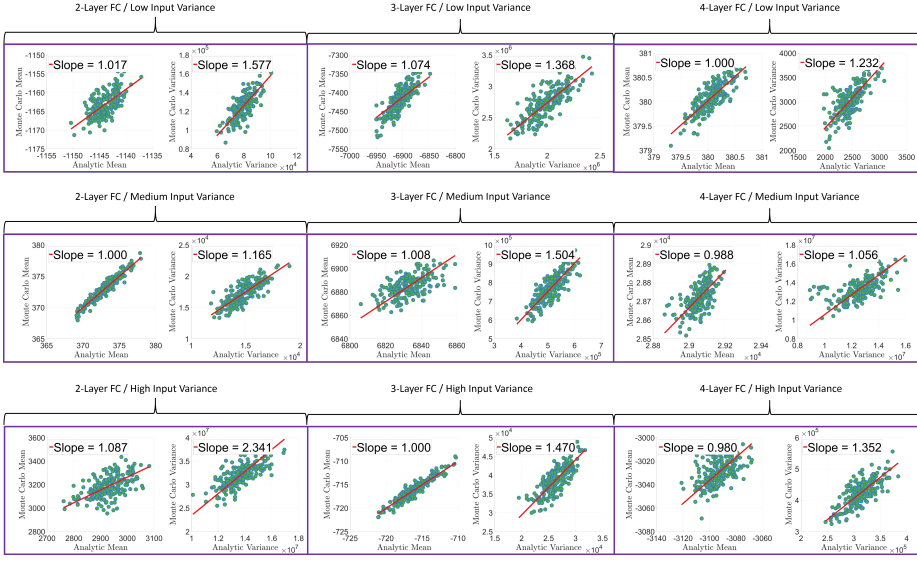

Fig. 1. Shows the tightness between the analytic expressions against the Monte Carlo estimates with random Gaussian inputs on Fully connected networks. The experiment is conducted across different types of networks, different depths and different input variance level. To compare the tightness, we report the slope of the fitted line in the legend. The closer the slope to 1.0 the tighter the expressions to the Monte Carlo estimates.

Convolutional. In here, we show the tightness of the analytic expressions on all convolutional networks under 3 different levels of input variance.

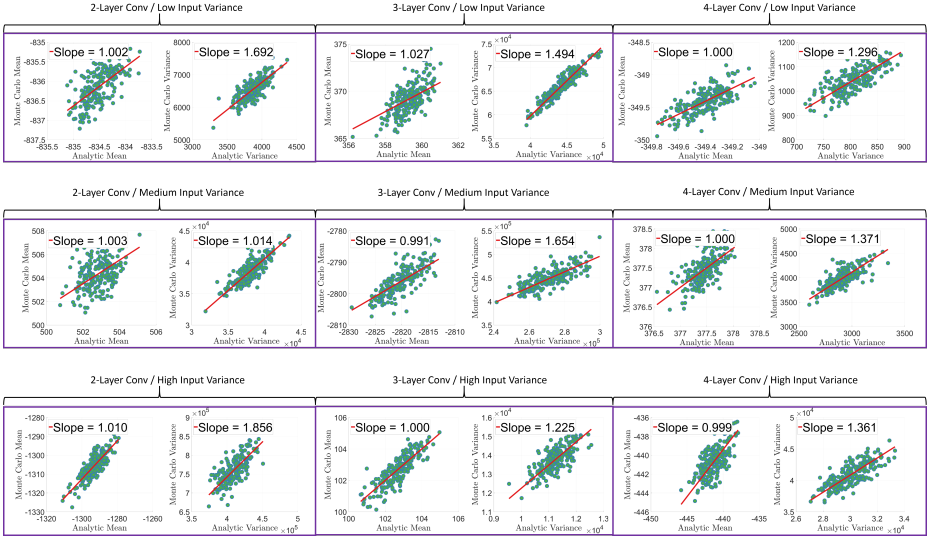

Fig. 2. Shows the tightness between the analytic expressions against the Monte Carlo estimates with random Gaussian inputs on convolutional networks. The experiment is conducted across different types of networks, different depths and different input variance level. To compare the tightness, we report the slope of the fitted line in the legend. The closer the slope to 1.0 the tighter the expressions to the Monte Carlo estimates.

Tightness on LeNet. In here, we show the tightness of the analytic expressions on LeNet under 2 different levels of input variance.

Table 2. Tightness under Σ_1 input variance.

Logits	$\mathbb{E}[\mathbb{E}_{ratio}]$	$\sigma(\mathbb{E}_{ratio})$	$\mathbb{E}[\text{var}_{ratio}]$	$\sigma(\text{var}_{ratio})$
$\mathbf{g}_o(\mathbf{x})$	0.999	0.016	0.619	0.024
$\mathbf{g}_1(\mathbf{x})$	1.002	0.038	0.690	0.027
$\mathbf{g}_2(\mathbf{x})$	0.995	0.081	0.416	0.015
$\mathbf{g}_3(\mathbf{x})$	1.005	0.060	0.592	0.023
$\mathbf{g}_4(\mathbf{x})$	0.972	0.305	0.537	0.020
$\mathbf{g}_5(\mathbf{x})$	0.983	0.220	0.487	0.016
$\mathbf{g}_6(\mathbf{x})$	1.000	0.008	0.546	0.021
$\mathbf{g}_7(\mathbf{x})$	1.000	0.008	0.483	0.016
$\mathbf{g}_8(\mathbf{x})$	1.000	0.008	0.517	0.017
$\mathbf{g}_9(\mathbf{x})$	1.011	0.126	0.716	0.026

Table 3. Tightness under Σ_2 input variance.

Logits	$\mathbb{E}[\mathbb{E}_{ratio}]$	$\sigma(\mathbb{E}_{ratio})$	$\mathbb{E}[\text{var}_{ratio}]$	$\sigma(\text{var}_{ratio})$
$\mathbf{g}_o(\mathbf{x})$	1.007	0	0.5431	0.021
$\mathbf{g}_1(\mathbf{x})$	1.043	0.001	0.6192	0.021
$\mathbf{g}_2(\mathbf{x})$	0.916	0.001	0.3681	0.014
$\mathbf{g}_3(\mathbf{x})$	1.061	0.001	0.517	0.019
$\mathbf{g}_4(\mathbf{x})$	0.581	0.006	0.4878	0.016
$\mathbf{g}_5(\mathbf{x})$	0.699	0.004	0.4267	0.015
$\mathbf{g}_6(\mathbf{x})$	0.999	0	0.4795	0.017
$\mathbf{g}_7(\mathbf{x})$	0.990	0	0.446	0.014
$\mathbf{g}_8(\mathbf{x})$	1.001	0	0.4476	0.014
$\mathbf{g}_9(\mathbf{x})$	1.164	0.002	0.5931	0.020

0.1 Analyzing LeNet's Logits' Behaviour.

To do so, we consider the simplest form of Gaussian noise where $\Sigma_x = \sigma^2 \mathbf{I}_n$. This noise model treats all pixels independently and gives them all the same variance level σ^2. We use the same LeNet network here with a similar linearization as the previous LeNet experiments. Under this noise model assumption, we can now plot the analytic mean of the output as a function of the input noise level σ for each label, where the input to the network is $\mathbf{x} \sim \mathcal{N}(\mu_x, \sigma^2 \mathbf{I}_n)$ and μ_x is taken to be a sampled image from each of the 10 MNIST labels (refer to Figure 3). This plot illustrates how the logit outputs for each label interact and how interclass confusion among labels take place with an increasing input noise level.

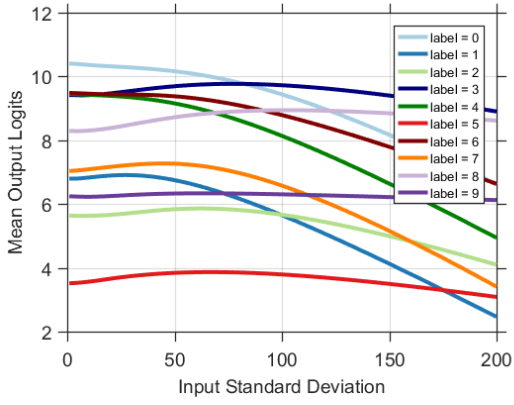

Fig. 3. The effect of varying the the input noise standard deviation σ on the change of the mean output logits.

As for the variance and under the previously stated conditions, the output-input variance has a linearly related ship with the following slope:

$$\frac{\text{var}[\mathbf{g}_i(\mathbf{x})]}{\sigma^2} = \frac{\pi - 1}{2\pi} \sum_{v=1}^{k} \mathbf{B}(i,v)^2 \|\mathbf{A}(v,:)\|_2^2 \tag{8}$$
$$- \frac{1}{2\pi} \sum_{r=1}^{k} \sum_{t=1}^{r-1} \mathbf{B}(i,r)\mathbf{B}(i,t) \|\mathbf{A}(r,:)\|_2 \|\mathbf{A}(t,:)\|_2$$

References

1. Price, R.: A useful theorem for nonlinear devices having gaussian inputs. IRE Transactions on Information Theory **4**(2) (1958) 69–72

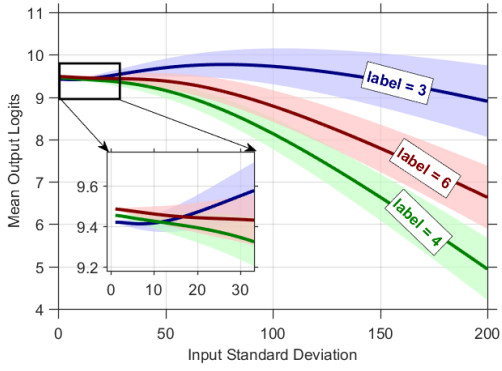

Fig. 4. The effect of varying the the input noise standard deviation σ on the change of the mean output logits for labels 3,6 and 4 and their corresponding variance. Note that the variance linearly increases for all labels with the increase of the input variance.