Neural Networks as a Gaussian Process

Richard Xu

August 15, 2020

1 Preamble

This talk is heavily referenced to the following:

- `https://www.uv.es/gonmagar/blog/2019/01/21/DeepNetworksAsGPs`
- J. H. Lee, Y. Bahri, R. Novak, S. S. Schoenholz, J. Pennington, and J. Sohl-Dickstein. Deep neural networks as gaussian processes. ICLR, 2018
- Radford M. Neal. Priors for infinite networks (tech. rep. no. crg-tr-94-1). University of Toronto, 1994

I tries to unify notations of the above references

2 Gaussian Process

- if one is to perform a predictive distribution $p(y^\star|y, X, x^\star)$ through GP:

$$
\begin{aligned}
p\left(\begin{bmatrix} y \\ y^\star \end{bmatrix} \Bigg| \begin{bmatrix} X \\ x^{\star\top} \end{bmatrix}\right) &= \int p\left(\begin{bmatrix} y \\ y^\star \end{bmatrix} \Bigg| \begin{bmatrix} X \\ x^{\star\top} \end{bmatrix}, f\right) p(f\big|X)\mathrm{d}f \\
&= \int \mathcal{N}\left(\begin{bmatrix} y \\ y^\star \end{bmatrix} \Big| \begin{bmatrix} f(X) \\ f(x^{\star\top}) \end{bmatrix}, \sigma_\epsilon^2 I\right) p(f\big|X, x^\star)\mathrm{d}f
\end{aligned}
$$

- This is the **key**: prior $p(f\big|X, x^\star)$ is defined over function $f(X)$ instead of X
- Imagine, if instead, prior is defined over X, i.e., $p(X)$ is the prior:

$$
\int \mathcal{N}\left(\begin{bmatrix} y \\ y^\star \end{bmatrix} \Big| \begin{bmatrix} f(X) \\ f(x^{\star\top}) \end{bmatrix}, \sigma_\epsilon^2 I\right) p(X)\mathrm{d}X
$$

  Then, non-linear f is **not** making integral tractable!

3 GP for Neural Network: Directly computation

3.1 neural network function

using parameters:

$$\omega \equiv \{W^L, b^L, \dots W^1, b^1\}$$

Deep neural network function $f_\omega(X)$ is defined as:

$$\begin{aligned} f_\omega(X) &= W^L \phi^L(X) + b^L \\ &= W^L\big(\phi^{L-1}(X)W^{L-1} + b^{L-1}\big) + b^L \\ \dots \\ &= W^L \cdots (W^1\phi^1(X) + b^1) + \dots) + b^L \end{aligned}$$

it should be noted that non-linear output $\phi^l(.)$:

$$\begin{aligned} \phi^L(X) &\equiv \phi^L(X\,|\omega^1, \dots, \omega^{L-1}) \\ &\equiv \phi^L(X\,|W^1, b^1, \dots, W^{L-1}, b^{L-1}) \end{aligned}$$

3.2 Apply NN function in predictive distribution

- However, applying NN function in predictive distribution: prior is defined over ω instead of over f. i.e., i.i.d noises are injected to each element of ω. The predictive distribution:

$$p\left(\begin{bmatrix} y \\ y^\star \end{bmatrix} \middle| \begin{bmatrix} X \\ x^{\star\top} \end{bmatrix}\right) = \int \mathcal{N}\left(\begin{bmatrix} y \\ y^\star \end{bmatrix} \middle| \begin{bmatrix} f_\omega(X) \\ f_\omega(x^\star) \end{bmatrix}, \sigma_\epsilon^2 I\right) \mathcal{N}(\omega\big|0, \sigma_\omega^2 I)\mathrm{d}\omega$$

- The integral is **not** analytic!!

3.3 what is the predictive distribution

- eventually, we will need to ask an even harder question on, i.e., suppose we let $N^l \equiv |W^l|$, i.e., the "width" of the neural network at each layer l, and we would like to study the effect of:

$$p\left(\begin{bmatrix} y \\ y^\star \end{bmatrix} \middle| \begin{bmatrix} X \\ x^{*\top} \end{bmatrix}\right) \xrightarrow[N^1, \dots, N^L \to \infty]{d} ?$$

- however, firstly, we ask the question on, what is:

$$p\left(\begin{bmatrix} y \\ y^\star \end{bmatrix} \middle| \begin{bmatrix} X \\ x^{*\top} \end{bmatrix}\right) = ?$$

- attempt to compute it **directly**, by looking the **mean** and **variance**:
  - $\mathbb{E}\left[\begin{bmatrix} y \\ y^\star \end{bmatrix} \middle| \begin{bmatrix} X \\ x^{*\top} \end{bmatrix}\right]$
  - $\mathbb{E}\left[\begin{bmatrix} y \\ y^\star \end{bmatrix} \begin{bmatrix} y^\top & y^\star \end{bmatrix} \middle| \begin{bmatrix} X \\ x^{\star\top} \end{bmatrix}\right]$

3.3.1 look at the mean:

$$\begin{aligned}
&\mathbb{E}\left[\begin{bmatrix} y \\ y^\star \end{bmatrix} \middle| \begin{bmatrix} X \\ x^{*\top} \end{bmatrix}\right] \\
&= \int_y \int_{y^\star} \begin{bmatrix} y \\ y^\star \end{bmatrix} p\left(\begin{bmatrix} y \\ y^\star \end{bmatrix} \middle| \begin{bmatrix} X \\ x^{*\top} \end{bmatrix}\right) \mathrm{d}y \, \mathrm{d}y^\star \\
&= \int_y \int_{y^\star} \begin{bmatrix} y \\ y^\star \end{bmatrix} \int_\omega p\left(\begin{bmatrix} y \\ y^\star \end{bmatrix} \middle| \underline{\omega, \begin{bmatrix} X \\ x^{*\top} \end{bmatrix}}\right) p(\omega|\sigma^2_\omega) \, \mathrm{d}\omega \, \mathrm{d}y \, \mathrm{d}y^\star \\
&= \int_\omega \underbrace{\int_y \int_{y^\star} \begin{bmatrix} y \\ y^\star \end{bmatrix} \mathcal{N}\left(\begin{bmatrix} y \\ y^\star \end{bmatrix} \middle| \underline{\begin{bmatrix} f_\omega(X) \\ f_\omega(x^\star) \end{bmatrix}}, \sigma^2_\epsilon I\right) \mathrm{d}y \, \mathrm{d}y^\star}_{\mathbb{E}\left[\begin{bmatrix} y \\ y^\star \end{bmatrix}\right] = \begin{bmatrix} f_\omega(X) \\ f_\omega(x^\star) \end{bmatrix}} \mathcal{N}(\omega \mid 0, \sigma^2_\omega I) \mathrm{d}\omega \\
&= \int \begin{bmatrix} f_\omega(X) \\ f_\omega(x^\star) \end{bmatrix} \mathcal{N}(\omega \mid 0, \sigma^2_\omega I) \, \mathrm{d}\omega \quad \text{to expand one layer :} \\
&= \int \begin{bmatrix} \phi^L(X) W^L + b^L \\ \phi^L(x^{*\top}) W^L + b^L \end{bmatrix} \mathcal{N}(W^L \mid 0, \sigma^2_w I) \mathcal{N}(b^L \mid 0, \sigma^2_b I) \mathcal{N}(\omega^{1,\dots,L-1} \mid 0, \sigma^2_\omega I) \mathrm{d}\omega^{1,\dots,L-1} \mathrm{d}W^L \, \mathrm{d}b^L \\
&= \int \begin{bmatrix} \phi^L(X) \underbrace{\int W^L \mathcal{N}(W^L \mid 0, \sigma^2_w I) \mathrm{d}W^L}_{=0} + \underbrace{\int b^L \mathcal{N}(b^L \mid 0, \sigma^2_b I) \mathrm{d}b^L}_{=0} \\ \phi^L(x^{*\top}) \underbrace{\int W^L \mathcal{N}(W^L \mid 0, \sigma^2_w I) \mathrm{d}W^L}_{=0} + \underbrace{\int b^L \mathcal{N}(b^L \mid 0, \sigma^2_b I) \mathrm{d}b^L}_{=0} \end{bmatrix} \mathcal{N}(\omega^{1,\dots,L-1} \mid 0, \sigma^2_\omega I) d\omega^{1,\dots,L-1} \\
&= \begin{bmatrix} 0 \\ 0 \end{bmatrix}
\end{aligned}$$

note we are not dealing with infinity at the moment

3.3.2 look at co-variance

$$\mathbb{E}\left[\begin{bmatrix} y \\ y^\star \end{bmatrix}\begin{bmatrix} y^\top & y^\star \end{bmatrix} \middle| \begin{bmatrix} X \\ x^{\star\top} \end{bmatrix}\right]$$

Apply same trick as calculating mean, i.e., introducing ω and then integrate it out:

$$= \int_y \int_{y^\star} \int_\omega p\left(\begin{bmatrix} y \\ y^\star \end{bmatrix}\begin{bmatrix} y^\top & y^\star \end{bmatrix} \middle| \omega, \begin{bmatrix} X \\ x^{\star\top} \end{bmatrix}\right) p(\omega|\sigma_\omega^2)\,\mathrm{d}\omega\,\mathrm{d}y\,\mathrm{d}y^\star$$
$$= \int_\omega \underbrace{\int_y \int_{y^\star} \begin{bmatrix} y \\ y^\star \end{bmatrix}\begin{bmatrix} y^\top & y^\star \end{bmatrix} \mathcal{N}\left(\begin{bmatrix} y \\ y^\star \end{bmatrix} \middle| \begin{bmatrix} f_\omega(X) \\ f_\omega(x^\star) \end{bmatrix}, \sigma_\epsilon^2 I\right) \mathrm{d}y\,\mathrm{d}y^\star}_{\mathbb{E}[Z^2] \quad Z \text{ is } \textbf{not} \text{ mean-subtracted}} \mathcal{N}(\omega \mid 0, \sigma_\omega^2 I)\mathrm{d}\omega$$

Let $Z = \begin{bmatrix} y \\ y^\star \end{bmatrix}$:

$$\mathrm{Var}[Z] = \mathbb{E}[Z^2] - (\mathbb{E}[Z])^2 \quad \Longrightarrow \quad \mathbb{E}[Z^2] = \mathrm{Var}[Z] + (\mathbb{E}[Z])^2$$
$$= \int_\omega \underbrace{\sigma_\epsilon^2 I}_{\mathrm{Var}[Z]} + \underbrace{\begin{bmatrix} f_\omega(X) \\ f_\omega(x^\star) \end{bmatrix}\begin{bmatrix} f_\omega(X)^\top & f_\omega(x^\star) \end{bmatrix}}_{(\mathbb{E}[Z])^2} \mathcal{N}(\omega \mid 0, \sigma_\omega^2 I)\mathrm{d}\omega$$

$$= \sigma_\epsilon^2 I + \int_\omega \begin{bmatrix} (\phi^L(X)W^L + b^L)(W^{L\top}x^L(X)^\top + b^{L\top}) & (\phi^L(X)W^L + b^L)(W^{L\top}\phi^L(x^{\star\top})^\top + b^{L\top}) \\ (\phi^L(\phi^{\star\top})W^L + b^L)(W^{L\top}\phi^L(X)^\top + b^{L\top}) & (\phi^L(x^{\star\top})W^L + b^L)(W^{L\top}\phi^L(x^{\star\top})^\top + b^{L\top}) \end{bmatrix} \mathcal{N}(\omega \mid 0, \sigma_\omega^2 I)d\omega$$

realize $\mathbf{Cov}(x^L(X)W^L, b^L) = 0$:

$$= \sigma_\epsilon^2 I + \int_\omega \begin{bmatrix} \phi^L(X)W^L W^{L\top}x^L(X)^\top + b^L b^{L\top} & \phi^L(X)W^L W^{L\top}x^L(x^{\star\top})^\top + b^L b^{L\top} \\ \phi^L(x^{\star\top})W^L W^{L\top}\phi^L(X)^\top + b^L b^{L\top} & \phi^L(x^{\star\top})W^L W^{L\top}\phi^L(x^{\star\top})^\top + b^L b^{L\top} \end{bmatrix} \mathcal{N}(\omega \mid 0, \sigma_\omega^2 I)d\omega$$

factorize $\mathcal{N}(\omega)$ as each element of ω is independent:

$$\mathcal{N}(\omega \mid 0, \sigma_\omega^2 I)d\omega = \mathcal{N}(\omega^L \mid 0, \sigma_\omega^2 I)\,\mathcal{N}(\omega^{1,\dots,L-1} \mid 0, \sigma_\omega^2 I)d\omega^{1,\dots,L-1}$$

$$= \int \begin{bmatrix} \sigma_w^2 \phi^L(X)x^L(X)^\top + \sigma_b^2 & \sigma_w^2 \phi^L(X)\phi^L(x^{\star\top})^\top + \sigma_b^2 \\ \sigma_w^2 \phi^L(x^{\star\top})\phi^L(X)^\top + \sigma_b^2 & \sigma_w^2 \phi^L(x^{\star\top})\phi^L(x^{\star\top})^\top + \sigma_b^2 \end{bmatrix} \mathcal{N}(\omega^{1,\dots,L-1} \mid 0, \sigma_\omega^2 I)d\omega^{1,\dots,L-1}$$

let's taking the left corner element, and expand ω by one:

$$\int \sigma_w^2 \phi^L(X)\phi^L(X)^\top \mathcal{N}(\omega^{1,\dots,L-1} \mid 0, \sigma_\omega^2 I)\,\mathrm{d}\omega^{1,\dots,L-1} + \int \sigma_b^2 \mathcal{N}(\omega^{1,\dots,L-1} \mid 0, \sigma_\omega^2 I)\,\mathrm{d}\omega^{1,\dots,L-1}$$
$$= \sigma_w^2 \int \phi^L(X)\phi^L(X)^\top \mathcal{N}(\omega^{1,\dots,L-1} \mid 0, \sigma_\omega^2 I)\,\mathrm{d}\omega^{1,\dots,L-1} + \sigma_b^2$$

as we know $\quad \phi^L(X)\phi^L(X)^\top \mathcal{N}(\omega^{1,\dots,L-1} \mid 0, \sigma_\omega^2 I)\,\mathrm{d}\omega^{1,\dots,L-1} + \sigma_b^2$:

$$= \sigma_b^2 + \sigma_w^2 \int \left[\phi(W^{L-1}\phi^{L-1}(X) + b^{L-1})\phi(W^{L-1}\phi^{L-1}(X) + b^{L-1})^\top\right] \mathcal{N}(\omega^{1,\dots,L-1} \mid 0, \sigma_\omega^2 I)\,\mathrm{d}\omega^{1,\dots,L-1}$$

it's difficult to see what is this distribution is.

4 Single layer neural network

$$f_k(x) = b_k + \sum_{j=1}^{H} v_{jk} h_j(x)$$
$$h_j(x) = \tanh\left(a_j + \sum_{i=1}^{I} u_{ij} x_i\right)$$

this is very strange way to define neural network, and it defines it to part of the second layer:

$$\underbrace{f_k(x)}_{z_k^l} = \underbrace{b_k}_{b_k^l} + \sum_{j=1}^{\overbrace{H}^{N_l}} \underbrace{v_{jk}}_{W_{k,j}^l} \times \underbrace{\tanh}_{\phi} \underbrace{\left(\underbrace{a_j}_{b_j^{l-1}} + \underbrace{u_{:,j}^\top}_{W_{:,j}^{l-1\top}} \; x \right)}_{z_j^{l-1}(x)}$$

$$\implies z_k^l(x) = b_k^l + \sum_{j=1}^{N_l} W_{k,j}^l \times \phi\big(z_j^{l-1}(x)\big) \quad \text{modern notation}$$

4.1 $p\big(z_k^l(x)\big)$ for single input x

We need CLT for computing this probability.

4.1.1 Central Limit Theorem:

$$X^{(1)}, X^{(2)}, \ldots, X^{(n)} \quad \text{are i.i.d samples}$$

- note any **arbitrary** distribution with *bounded variance* for $X^{(i)}$ will do
- let $\overline{X}$ be sample mean, and let: $\sigma^2 = \text{Var}[X^{(1)}]$
- Limiting form of the distribution:

$$\sqrt{n}\big(\overline{X} - \mathbb{E}[X^{(1)}]\big) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$
$$\big(\overline{X} - \mathbb{E}[X^{(1)}]\big) \xrightarrow{d} \mathcal{N}(0, \frac{\sigma^2}{n})$$
$$\frac{1}{\sigma}\sqrt{n}\big(\overline{X} - \mathbb{E}[X^{(1)}]\big) \xrightarrow{d} \mathcal{N}(0, 1)$$

- Similarly, instead of "**sample mean**", it can be also be applied to "**sample sum**" of i.i.d random variables:

$$\sqrt{n}\big(\overline{X} - \mathbb{E}[X^{(1)}]\big) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$
$$\implies \sqrt{n}\sqrt{n}\big(\overline{X} - \mathbb{E}[X^{(1)}]\big) \xrightarrow{d} \mathcal{N}(0, \sqrt{n}^2\sigma^2) = \mathcal{N}(0, n\sigma^2)$$
$$\implies n\big(\overline{X} - \mathbb{E}[X^{(1)}]\big] \xrightarrow{d} \mathcal{N}(0, n\sigma^2)$$
$$\implies \left(\sum_{i=1}^{n} X_i - n\mathbb{E}[X^{(1)}]\right) \xrightarrow{d} \mathcal{N}(0, n\sigma^2)$$

choose one of these conditions to suit the situation

4.1.2 Apply CLT to compute $p\big(z_k^l(x)\big)$

- let's pick any arbitrary x, since we only pick a single x, so the index is **not** important, there is no need to use $x^{(1)}$ like in the literature:
- computing $p\big(z_k^l(x)\big)$ directly is hard!
- however, $z_k^l(x)$ is $b_k^l +$ sum of i.i.d elements using CLT notations:

$$z_k^l(x) = b_k^l + \underbrace{\sum_{j=1}^{N_l} \underbrace{W_{k,j}^l \phi\big(z_j^{l-1}(x)\big)}_{X_j}}_{\sum_{j=1}^{N_l} X_j}, \quad \text{note we are not taking average}$$

- therefore, we can just compute mean and variance of its individual element, i.e., an arbitrary $j = 1$ and then apply CLT!

$$X_j \equiv W_{k,j}^l \phi\big(z_j^{l-1}(x)\big)$$

4.1.3 mean and variance of $W_{k,j}^l \phi\big(z_j^{l-1}(x)\big)$

- **Expectation**

$$\begin{aligned}
\mathbb{E}\big[W_{k,j}^l\, \phi\big(z_j^{l-1}(x)\big)\big] &= \mathbb{E}[W_{k,j}^l]\, \mathbb{E}\big[\phi\big(z_j^{l-1}(x)\big)\big] \qquad \text{since } W_{k,j}^l \text{ and } \phi\big(z_j^{l-1}(x)\big) \text{ are independent} \\
&\qquad\qquad \text{as } z_j^{l-1}(x) \text{ depends on } (W^{l-1}, b^{l-1}) \\
&= 0 \times \mathbb{E}[\phi\big(z_j^{l-1}(x)\big)] \qquad \text{because we choose} \quad W_{k,j}^l \sim \mathcal{N}(0, \sigma_w) \\
&= 0
\end{aligned}$$

- **Variance**

$$\begin{aligned}
&\mathrm{Var}\big[W_{k,j}^l \phi\big(z_j^{l-1}(x)\big)\big] \\
&= \mathbb{E}\Big[\Big(W_{k,j}^l \phi\big(z_j^{l-1}(x)\big)\Big)^2\Big] \\
&= \mathbb{E}\big[\big(W_{k,j}^l\big)^2\big]\, \mathbb{E}\big[\phi\big(z_j^{l-1}(x)\big)^2\big] \quad \text{since } W_{k,j}^l \text{ and } \phi\big(z_j^{l-1}(x)\big) \text{ are independent} \\
&= \sigma_w^2 \mathbb{E}\big[\underbrace{\phi\big(z_j^{l-1}(x)\big)\big)^2}_{\text{bounded}}\big] \qquad \Longrightarrow \mathrm{Var}\big[W_{k,j}^l \phi\big(z_j^{l-1}(x)\big)\big] \text{ to be bounded} \\
&= \sigma_w^2\, \mathbb{E}\big[\phi\big(z_j^{l-1}(x)\big)^2\big]
\end{aligned}$$

we leave in this form, as

$$\mathbb{E}\big[\phi\big(z_j^{l-1}(x)\big)^2\big] \equiv \mathbb{E}_{W^{l-1},\ldots,b^{l-1},\ldots}\big[\phi\big(z_j^{l-1}(x)\big)^2\big]$$

4.1.4 apply CLT:

However, we can apply CLT: making $p(z^l(x))$ distributed as Gaussian where its variance is dependent on variance of previous layer, a recursion.

$$\text{using} \quad \left(\sum_{i=1}^{n} X_i - n\mathbb{E}[X_1]\right) \xrightarrow{d} \mathcal{N}(0, n\sigma^2)$$
$$\implies \left(\sum_{j=1}^{N_l} W_{k,j}^l \phi\big(z_j^{l-1}(x)\big) - 0\right) \sim \mathcal{N}\Big(0, N_l\ \sigma_w^2\ \mathbb{E}\big[\phi\big(z_1^{l-1}(x)\big)^2\big]\Big) \quad N_l \to \infty$$

- However, variance under this expression $N_l\ \sigma_w^2\ \big[\phi\big(z_1^{l-1}(x)\big)^2\big]$ is divergent because of N_l!
- luckily, we can take control the choice of σ_w^2, if we let:

$$\sigma_w = \frac{C_w}{\sqrt{N_l}} \quad \implies \sigma_w^2 = \frac{C_w^2}{N_l}$$

- the above is the key, implication is:

$$\implies \left(\sum_{j=1}^{N_l} W_{k,j}^l \phi\big(z_j^{l-1}(x)\big) - 0\right) \sim \mathcal{N}\Big(0, N_l \frac{C_w^2}{N_l}\ \mathbb{E}\big[\phi\big(z_1^{l-1}(x)\big)^2\big]\Big)$$
$$= \mathcal{N}\Big(0, C_w^2\, \underbrace{\mathbb{E}\big[\phi\big(z_1^{l-1}(x)\big)^2\big]}_{\text{bounded}}\Big)$$

- finally adding the bias b_k^l:

  Note that sum of two **independent** Gaussian random variables is also Gaussian: (not to confuse with GMM!)

$$X \sim \mathcal{N}(\mu_X, \sigma_X^2)$$
$$Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$$
$$Z = X + Y \quad Z = X + Y$$
$$\implies Z \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$$

  Therefore:

$$\left(z_k^l(x) = b_k^l + \sum_{j=1}^{N_l} W_{k,j}^l \phi\big(z_j^{l-1}(x)\big)\right) \xrightarrow{d} \mathcal{N}\Big(0, \underbrace{\sigma_b^2}_{\sigma_X^2} + \underbrace{T_w^2\ \mathbb{E}\big[\phi\big(z_1^{l-1}(x)\big)^2\big]}_{\sigma_Y^2}\Big) \quad \text{as } N_l \to \infty$$

- appreciate the recursion here

4.2 given two inputs $x^{(p)}$, $x^{(q)}$: compute Cov$\big[z_k^l(x^{(p)})\ z_k^l(x^{(q)})\big]$

To do so, we need to used **Multidimensional CLT**

4.2.1 Multidimensional CLT:

$$\sum_{i=1}^{n} \mathbf{X}_i = \underbrace{\begin{bmatrix} X_1^{(1)} \\ \vdots \\ X_1^{(p)} \\ \vdots \\ X_1^{(q)} \\ \vdots \\ X_1^{(k)} \end{bmatrix}}_{\mathbf{X}_1} + \underbrace{\begin{bmatrix} X_2^{(1)} \\ \vdots \\ X_2^{(p)} \\ \vdots \\ X_2^{(q)} \\ \vdots \\ X_2^{(k)} \end{bmatrix}}_{\mathbf{X}_2} + \cdots + \underbrace{\begin{bmatrix} X_n^{(1)} \\ \vdots \\ X_n^{(p)} \\ \vdots \\ X_n^{(q)} \\ \vdots \\ X_n^{(k)} \end{bmatrix}}_{\mathbf{X}_n} = \underbrace{\begin{bmatrix} \sum_{i=1}^n X_i^{(1)} \\ \vdots \\ \sum_{i=1}^n X_i^{(p)} \\ \vdots \\ \sum_{i=1}^n X_i^{(q)} \\ \vdots \\ \sum_{i=1}^n X_i^{(k)} \end{bmatrix}}_{\sum_{i=1}^n \mathbf{X}_i}$$

$$\implies \overline{\mathbf{X}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i = \begin{bmatrix} \frac{1}{n} \sum_{i=1}^n X_i^{(1)} \\ \vdots \\ \frac{1}{n} \sum_{i=1}^n X_i^{(p)} \\ \vdots \\ \frac{1}{n} \sum_{i=1}^n X_i^{(q)} \\ \vdots \\ \frac{1}{n} \sum_{i=1}^n X_i^{(k)} \end{bmatrix} = \begin{bmatrix} \overline{\mathbf{X}}^{(1)} \\ \vdots \\ \overline{\mathbf{X}}^{(p)} \\ \vdots \\ \overline{\mathbf{X}}^{(q)} \\ \vdots \\ \overline{\mathbf{X}}^{(k)} \end{bmatrix}$$

Therefore:

$$\begin{aligned} &\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\mathbf{X}_i - \mathbb{E}[\mathbf{X}_i]\right] \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\mathbf{X}_i - \mathbb{E}[\mathbf{X}_1]) = \frac{\sqrt{n}}{\sqrt{n}} \frac{1}{\sqrt{n}} \left(\sum_{i=1}^{n} \mathbf{X}_i \right) - n\mathbb{E}[\mathbf{X}_1] \\ &= \sqrt{n} \left(\overline{\mathbf{X}} - \mathbb{E}[\mathbf{X}_1] \right) \end{aligned}$$

- **Sample mean version**:

$$\sqrt{n} \left(\overline{\mathbf{X}} - \mathbb{E}[\mathbf{X}_1] \right) \xrightarrow{d} \mathcal{N}_k(0, \mathbf{\Sigma})$$

$$\implies \sqrt{n}\, \mathbb{E}\Big[\Big(\underbrace{\overline{\mathbf{X}}^{(p)} - \mathbb{E}[\overline{\mathbf{X}}_1^{(p)}]}_{\text{scalar}} \Big) \Big(\underbrace{\overline{\mathbf{X}}^{(q)} - \mathbb{E}[\, \overline{\mathbf{X}}_1^{(q)}]}_{\text{scalar}} \Big) \Big] = \mathbf{\Sigma}_{(p),(q)}$$

for each co-variance/non-diagonal elements $(p, q) \in \{1, \ldots, k\}$:

- **Sample sum version**:

$$\left(\Big[\sum_{i}^{n}\mathbf{X}_i\Big]-n\mathbb{E}[\mathbf{X}_1]\right)\xrightarrow{d}\mathcal{N}_k(0,n\mathbf{\Sigma})$$

$$\implies \mathbb{E}\Bigg[\bigg(\Big[\sum_{i}^{n}\mathbf{X}_i\Big]^{(p)}-n\mathbb{E}[\mathbf{X}_1]^{(p)}\bigg)\bigg(\Big[\sum_{i}^{n}\mathbf{X}_i\Big]^{(q)}-n\mathbb{E}[\mathbf{X}_1]^{(q)}\bigg)\Bigg]=n\mathbf{\Sigma}_{(p),(q)}$$

$$\implies \mathbb{E}\Big[\Big(n\overline{\mathbf{X}}^{(p)}-n\mathbb{E}[X_1^{(p)}]\Big)\Big(n\overline{\mathbf{X}}^{(q)}-n\mathbb{E}[X_1^{(q)}]\Big)\Big]=n\mathbf{\Sigma}_{(p),(q)}$$

$$\implies \mathbb{E}\Bigg[\bigg(\Big[\sum_{i}^{n}\mathbf{X}_i\Big]^{(p)}-n\mathbb{E}[X_1^{(p)}]\bigg)\bigg(\Big[\sum_{i}^{n}\mathbf{X}_i\Big]^{(p)}-n\mathbb{E}[X_1^{(q)}]\bigg)\Bigg]=n\mathbf{\Sigma}_{(p),(q)}$$

where $\mathbf{\Sigma}_{(p),(q)}=\mathrm{Cov}\big(X_1^{(p)},X_1^{(q)}\big)$

4.2.2 put in Multidimensional CLT structure:

$$\begin{bmatrix}\vdots\\ W^l_{k,1}\phi\big(z_1^{l-1}(x^p)\big)\\ \vdots\\ W^l_{k,1}\phi\big(z_j^{l-1}(x^q)\big)\\ \vdots\end{bmatrix}+\cdots+\begin{bmatrix}\vdots\\ W^l_{k,N_l}\phi\big(z_j^{l-1}(x^p)\big)\\ \vdots\\ W^l_{k,N_l}\phi\big(z_j^{l-1}(x^q)\big)\\ \vdots\end{bmatrix}=\underbrace{\begin{bmatrix}\vdots\\ \sum_{j=1}^{N_l}W^l_{k,j}\phi\big(z_j^{l-1}(x^p)\big)\\ \vdots\\ \sum_{j=1}^{N_l}W^l_{k,j}\phi\big(z_j^{l-1}(x^q)\big)\\ \vdots\end{bmatrix}}_{\begin{bmatrix}\sum_{i=1}^{n}X_i^{(1)}\\ \vdots\\ \sum_{i=1}^{n}X_i^{(p)}\\ \vdots\\ \sum_{i=1}^{n}X_i^{(q)}\\ \vdots\\ \sum_{i=1}^{n}X_i^{(k)}\end{bmatrix}}=\underbrace{\begin{bmatrix}\vdots\\ z_k^l(x^{(p)})\\ \vdots\\ z_k^l(x^{(q)})\\ \vdots\end{bmatrix}}_{\sum_{i=1}^{n}\mathbf{X}_i}$$

Since we already know that:

$$\mathbb{E}\Bigg[\bigg(\Big[\sum_{i}^{n}\mathbf{X}_i\Big]^{(p)}-n\mathbb{E}[X_1^{(p)}]\bigg)\bigg(\Big[\sum_{i}^{n}\mathbf{X}_i\Big]^{(q)}-n\mathbb{E}[X_1^{(q)}]\bigg)\Bigg]=n\mathbf{\Sigma}_{(p),(q)}$$

$$\implies \mathbb{E}\Bigg[\bigg(\sum_{j=1}^{N_l}W^l_{k,j}\phi(z_j^{l-1}(x^{(p)}))-N_l\underbrace{\mathbb{E}\big[W^l_{k,1}\phi(z_1^{l-1}(x^{(p)}))\big]}_{=0}\bigg)\times$$

$$\bigg(\sum_{j=1}^{N_l}W^l_{k,j}\phi(z_j^{l-1}(x^{(q)}))-N_l\underbrace{\mathbb{E}\big[W^l_{k,1}\phi(z_1^{l-1}(x^{(q)}))\big]}_{=0}\bigg)\Bigg]=N_l\mathbf{\Sigma}_{(p),(q)}$$

for any arbitrary $j = 1$, and then:

$$
\begin{aligned}
&\mathbb{E}\Big[\Big(\sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(x^{(p)}))\Big)\Big(\sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(x^{(q)}))\Big)\Big] \\
&= N_l\ \mathbf{\Sigma}_{(p),(q)} \\
&= N_l \operatorname{Cov}\Big(W_{k,1}^l \phi(z_1^{l-1}(x^{(p)})), W_{k,1}^l \phi(z_1^{l-1}(x^{(q)}))\Big) \\
&= N_l\ \mathbb{E}\Big[W_{k,1}^l \phi(z_1^{l-1}(x^{(p)})) \times W_{k,1}^l \phi(z_1^{l-1}(x^{(q)}))\Big]
\end{aligned}
$$

add b_k^l into, and look at $z_k^l(x)$:

$$
\begin{aligned}
\mathbb{E}\big[z_k^l(x^{(p)}) z_k^l(x^{(q)})\big] &= \sigma_b^2 + \mathbb{E}\Big[\Big(\sum_{j=1}^{N_l} W_{k,j}^l \phi\big(z_j^{l-1}(x^{(p)})\big)\Big)\Big(\sum_{j=1}^{N_l} W_{k,j}^l \phi\big(z_j^{l-1}(x^{(q)})\big)\Big)\Big] \\
&= \sigma_b^2 + N_l \operatorname{Cov}\big(W_{k,1}^l \phi\big(z_1^{l-1}(x^{(p)})\big), W_{k,1}^l \phi\big(z_1^{l-1}(x^{(q)})\big)\big) \quad \text{use CLT result above} \\
&= \sigma_b^2 + N_l \sigma_w^2 \operatorname{Cov}\big(\phi\big(z_1^{l-1}(x^{(p)})\big), \phi\big(z_1^{l-1}(x^{(q)})\big)\big) \\
&= \sigma_b^2 + N_l \frac{C_w^2}{N_l} \operatorname{Cov}\big(\phi\big(z_1^{l-1}(x^{(p)})\big), \phi\big(z_1^{l-1}(x^{(q)})\big)\big) \\
&= \sigma_b^2 + C_w^2 \operatorname{Cov}\big(\phi\big(z_1^{l-1}(x^{(p)})\big), \phi\big(z_1^{l-1}(x^{(q)})\big)\big) \\
&= \sigma_b^2 + C_w^2\ \mathbb{E}\big[\phi\big(z_1^{l-1}(x^{(p)})\big) \times \phi\big(z_1^{l-1}(x^{(q)})\big)\big]
\end{aligned}
$$

- **note 1**: this co-variance is same $\forall k$ in $z_k^l(x)$, so right hand side does not need to keep k index

  because in this particular setting, since b_k, $b_{k'}$, $W_{k,j}$ and $W_{k',j'}$ are independent variables, co-variance between any of them are zero:

$$
\begin{aligned}
&z_k^l(x) = b_k + \sum_{j=1}^{N_l} W_{k,j}^l \phi\big(z_j^{l-1}(x)\big) \\
&z_{k'}^l(x) = b_{k'} + \sum_{j=1}^{N_l} W_{k',j}^l \phi\big(z_j^{l-1}(x)\big) \\
&\implies \mathbb{E}\Big[W_{k,j}^l \phi\big(z_j^{l-1}(x)\big) \times W_{k',j'}^l \phi\big(z_{j'}^{l-1}(x)\big)\Big] = 0 \quad \forall\{k, k', j, j'\}
\end{aligned}
$$

- **note 2**: in literature, it is written:

$$
\begin{aligned}
\mathbb{E}\big[z_k^l(x^{(p)}) z_k^l(x^{(q)})\big] &= \sigma_b^2 + \sigma_w^2\ \mathbb{E}\Big[\sum_{j=1}^{N_l} \phi\big(z_j^{l-1}(x^p)\big)\phi\big(z_j^{l-1}(x^q)\big)\Big] \\
\text{instead of } &= \sigma_b^2 + \mathbb{E}\Big[\Big(\sum_{j=1}^{N_l} W_{k,j}^l \phi\big(z_j^{l-1}(x^p)\big)\Big)\Big(\sum_{j=1}^{N_l} W_{k,j}^l \phi\big(z_j^{l-1}(x^q)\big)\Big)\Big]
\end{aligned}
$$

  This is because of $\textbf{note1}$ above

- regardless of this special property CLT still apply.

4.2.3 Relationship with Gaussian Process (GP):

let $f_k(x) \equiv z_k^l(x)$ be some function, and since for every arbitrary point pair, $x^{(p)}$ and $x^{(q)}$, we have:

$$\begin{aligned}
\mathbb{E}\big[f(x)\big] &= 0 \\
\mathbb{E}\big[f(x^{(p)}, f(x^{(q)})\big] &= \mathbf{\Sigma}_{(p),(q)} \\
&\implies f \sim \mathcal{GP}(0, \mathbf{\Sigma})
\end{aligned}$$

- looking at mean and co-variance as $N_l \to \infty$

$$\begin{aligned}
\mathrm{Cov}\Big[z_k^l(x^{(p)}), z_k^l(x^{(q)})\Big] &= \sigma_b^2 + C_w^2\ \mathbb{E}\big[\phi\big(z_1^{l-1}(x^{(p)})\big) \times \phi\big(z_1^{l-1}(x^{(q)})\big)\big] \quad \text{as } N_l \to \infty \\
z_k^l(x) &\xrightarrow{d} \mathcal{N}\Big(0, \sigma_b^2 + C_w^2\ \mathbb{E}\big[\phi\big(z_1^{l-1}(x)\big)^2\big]\Big) \quad \text{as } N_l \to \infty
\end{aligned}$$

- putting it in layer specific GP:

$$\begin{aligned}
&\implies z_k^l(x) \sim \mathcal{GP}(0, \mathbf{\Sigma}) \\
&\quad \text{where} \quad \mathbf{\Sigma}_{p,q} = \sigma_b^2 + C_w^2\ \mathbb{E}\big[\phi\big(z_1^{l-1}(x^{(p)})\big) \times \phi\big(z_1^{l-1}(x^{(q)})\big)\big] \quad \text{as } N_l \to \infty
\end{aligned}$$

4.3 more on GP

first define $K^l(x^{(p)}, x^{(q)})$ in terms of pre-activation $z_k^l(x)$ in this section, it will be changed later to post-activation

$$\begin{aligned}
K^l(x^{(p)}, x^{(q)}) &= \mathbb{E}\big[z_k^l(x^{(p)}) z_k^l(x^{(q)}) \big|\ z^{l-1}\big] \\
&= \mathbb{E}\Bigg[\Bigg(b_k^l + \sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(x^{(p)}))\Bigg) \times \Bigg(b_k^l + \sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(x^{(q)}))\Bigg)\Bigg] \\
&= \sigma_b^2 + \sigma_w^2 \mathbb{E}\Bigg[\sum_{j=1}^{N_l} \phi(z_j^{l-1}(x^{(p)})) \ \times \ \sum_{j=1}^{N_l} \phi(z_j^{l-1}(x^{(q)}))\Bigg] \\
&= \sigma_b^2 + \sigma_w^2 \mathbb{E}\big[\phi(z_1^{l-1}(x^{(p)})) \ \times \ \phi(z_1^{l-1}(x^{(q)}))\big] \quad \text{apply CLT} \quad N_l \to \infty \\
&= \sigma_b^2 + \sigma_w^2 \underbrace{\mathbb{E}_{z_1^{l-1} \sim \mathcal{GP}(0, K^{l-1})}\Big[\phi(z_1^{l-1}(x^{(p)}))\phi(z_1^{l-1}(x^{(q)}))\Big]}_{\text{since } \mathbb{E}[\phi(z)] = \mathbb{E}_{z \sim p(z)}[\phi(z)]} \\
&= \underbrace{\sigma_b^2 + \sigma_w^2 \underbrace{F_\phi\big(K^{l-1}(x^{(p)}, x^{(q)}), K^{l-1}(x^{(p)}, x^{(p)}), K^{l-1}(x^{(q)}, x^{(q)})\big)}}_{G(F_\phi(K^{l-1}))} \\
&= G \circ F_\phi(K^{l-1}(x^{(p)}, x^{(q)}))
\end{aligned}$$

using properties of point Marginals of Gaussian Process:

$$
\begin{aligned}
F_\phi(K^{l-1}(x^{(p)}, x^{(q)})) &= \mathbb{E}_{z_j^{l-1} \sim \mathcal{GP}(0, K^{l-1})}\Big[\phi(z_j^{l-1}(x^{(p)}))\phi(z_j^{l-1}(x^{(q)}))\Big] \\
&= \mathbb{E}_{\underbrace{(z_j^{l-1}(x^{(p)}), z_j^{l-1}(x^{(q)}))}_{\text{2 points on function } z_j^{l-1}} \sim \underbrace{\mathcal{N}(0, K^{l-1}(x^{(p)}, x^{(q)}))}_{\text{2D Gaussian}}}\Big[\phi(z_j^{l-1}(x^{(p)}))\phi(z_j^{l-1}(x^{(q)}))\Big]
\end{aligned}
$$

$$
\begin{bmatrix} z_j^{l-1}(x^{(p)}) \\ z_j^{l-1}(x^{(q)}) \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}\,, \begin{bmatrix} K^{l-1}(x^{(p)}, x^{(p)}) & K^{l-1}(x^{(p)}, x^{(q)}) \\ K^{l-1}(x^{(p)}, x^{(q)}) & K^{l-1}(x^{(q)}, x^{(q)}) \end{bmatrix}\right)
$$

assume z^{l-1} can be integrated out:

$$
= F_\phi\big(K^{l-1}(x^{(p)}, x^{(q)}), K^{l-1}(x^{(p)}, x^{(p)}), K^{l-1}(x^{(q)}, x^{(q)})\big)
$$

5 Expand GP across all layers

5.1 Overall objective

Looking the probability of the final layer output z^L depending on input x:

$$
\begin{aligned}
p(z^L|x) &= \int p(z^L, K^0, K^1, \ldots, K^L|x)\,\mathrm{d}K^{0,\ldots,L} \\
&= \int p(z^L|K^L)\Big(\prod_{l=1}^{L} p(K^l|K^{l-1})\Big)p(K^0|x)\,\mathrm{d}K^{0,\ldots,L}
\end{aligned}
$$

5.2 $p(z^L|K^L)$: conditions on $K^l \equiv \Big\{\phi\big(z^{l-1}\big)\big(x^{(p)}\big)\phi\big(z^{l-1}\big)\big(x^{(q)}\big)\Big\}_{p,q}$

(J. H. Lee et. all 2018) presents an **alternative** definition of K^l, where no longer define K from pre-activation:

$$
K^l(x^{(p)}, x^{(q)}) = \mathbb{E}\big[z_k^l(x^{(p)}) z_k^l(x^{(q)})\big|\, z^{l-1}\big]
$$

instead it define K^l in terms of post-activation of previous later $\phi(z^{l-1})$ for reason illustrated later

- look at Neural Network function:

$$
z_k^l(x) = b_k^l + \sum_{j=1}^{N_l} W_{k,j}^l \phi(z_j^{l-1}(x))
$$

let's make it dependent on $\big\{\phi(z_j^{l-1}(x))\big\}_j^{N_l}$, i.e.:

- Conditional Marginal

$$
\begin{aligned}
z_k^l(x) \big| \left\{\phi(z_j^{l-1}(x))\right\}_j^{N_l} &= b_k^l + \sum_{j=1}^{N_l} W_{k,j}^l \underbrace{\phi(z_j^{l-1}(x))}_{\text{constant}} \\
\implies z_k^l(x) \big| \left\{\phi(z_j^{l-1}(x))\right\}_j^{N_l} &\sim \mathcal{N}\Big(0, \sigma_b^2 + \sum_{j=1}^{N_l} \phi(z_j^{l-1}(x))^2 \mathrm{Var}\big[W_{k,j}^l\big]\Big) \\
&= \mathcal{N}\Big(0, \sigma_b^2 + \frac{\sigma_w^2}{N_l} \sum_{j=1}^{N_l} \phi(z_j^{l-1}(x))^2\Big)
\end{aligned}
$$

using property of weighted sum of Gaussian:

$$
\begin{aligned}
X_i &\sim \mathcal{N}(\mu_i, \sigma_i^2), \qquad i = 1, \ldots, \\
\implies \sum_{i=1}^{n} a_i X_i &\sim \mathcal{N}\left(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \ \mathrm{Var}[X_i]\right)
\end{aligned}
$$

- Conditional Co-variance

$$
\begin{aligned}
&\mathrm{Cov}\Big[z_k^l(x^{(p)}), z_k^l(x^{(q)}) \ \Big| \ \Big\{\phi\big(z_j^{l-1}(x^{(p)})\big), \phi\big(z_j^{l-1}(x^{(q)})\big)\Big\}_{j=1}^{N_l}\Big] \\
&= \mathbb{E}\Big[z_k^l(x^{(p)}) z_k^l(x^{(q)}) \ \Big| \ \Big\{\phi\big(z_j^{l-1}(x^{(p)})\big), \phi\big(z_j^{l-1}(x^{(q)})\big)\Big\}_{j=1}^{N_l}\Big] \\
&= \sigma_b^2 + \ \mathbb{E}_{W_{k,j}^l}\Big[\sum_{j=1}^{N_l} {W_{k,j}^l}^2 \ \underbrace{\phi\big(z_j^{l-1}(x^{(p)})\big) \ \phi\big(z_j^{l-1}(x^{(q)})\big)}_{\text{constant, used as condition}}\Big] \\
&= \sigma_b^2 + \ \sum_{j=1}^{N_l} \mathrm{Var}\big[W_{k,j}^l\big] \ \phi\big(z_j^{l-1}(x^{(p)})\big) \ \phi\big(z_j^{l-1}(x^{(q)})\big) \\
&= \sigma_b^2 + \ \frac{\sigma_w^2}{N_l} \sum_{j=1}^{N_l} \ \phi\big(z_j^{l-1}(x^{(p)})\big) \ \phi\big(z_j^{l-1}(x^{(q)})\big)
\end{aligned}
$$

not using property of weighted sum of Gaussian:

- Combine all together

$$
\begin{aligned}
\mathrm{Cov}\Big[z_k^l(x^{(p)}), z_k^l(x^{(q)}) \ \Big| \ \Big\{\phi\big(z_j^{l-1}(x^{(p)})\big), \phi\big(z_j^{l-1}(x^{(q)})\big)\Big\}_{j=1}^{N_l}\Big] &= \sigma_b^2 + \sigma_w^2 \ \frac{1}{N_l} \sum_{j=1}^{N_l} \ \phi\big(z_j^{l-1}(x^{(p)})\big) \ \phi\big(z_j^{l-1}(x^{(q)})\big) \\
z_k^l(x) \big| \left\{\phi(z_j^{l-1}(x))\right\}_j^{N_l} &\sim \mathcal{N}\Big(0, \sigma_b^2 + \sigma_w^2 \ \frac{1}{N_l} \sum_{j=1}^{N_l} \phi(z_j^{l-1}(x))^2\Big) \\
\implies \begin{bmatrix} z^l(x^{(p)}) \mid \phi\big(z_j^{l-1}(x^{(p)})\big) \\ z^l(x^{(q)}) \mid \phi\big(z_j^{l-1}(x^{(q)})\big) \end{bmatrix} &\sim \mathcal{N}\Big(\mathbf{0}, G\Big(\begin{bmatrix} K^l(x^{(p)}, x^{(p)}) & K^l(x^{(p)}, x^{(q)}) \\ K^l(x^{(p)}, x^{(q)}) & K^l(x^{(q)}, x^{(q)}) \end{bmatrix}\Big)\Big)
\end{aligned}
$$

- in GP paradigm:

$$z^l(x)|K^l \sim \mathcal{GP}\big(z^l; \mathbf{0}, G(K^l)\big)$$

where

$$K^l(x^{(p)}, x^{(q)}) = \frac{1}{N_l}\sum_{j=1}^{N_l} \phi\big(z_j^{l-1}(x^{(p)})\big)\ \phi\big(z_j^{l-1}(x^{(q)})\big)$$
$$G\big(K^l(x^{(p)}, x^{(q)})\big) = \sigma_b^2 + \sigma_w^2 K^l(x^{(p)}, x^{(q)})$$

Conveniently, we use K^l as a short-notation collection of $\phi\big(z_j^{l-1}(x^{(p)})\big)\ ,\ \phi\big(z_j^{l-1}(x^{(q)})\big) \quad \forall p, q, j$

- also taking care of the layer one, which is just input x:

$$K_{p,q}^l \equiv K^l(x^{(p)}, x^{(q)}) = \begin{cases} \frac{1}{d_{\text{in}}}\sum_{j=1}^{d_{\text{in}}} x_j^{(p)} x_j^{(q)} & l = 0 \\ \frac{1}{N_l}\sum_{j=1}^{N_l} \phi\big(z_j^{l-1}(x^{(p)})\big)\phi\big(z_j^{l-1}(x^{(q)})\big) & l > 0 \end{cases}$$

- to reflect:

$$\text{Cov}(z_k^l, z_{k'}^l) = 0\ \forall\ k, k' \in \{1, \dots N_{l+1}\}$$

one may construct giant co-variance matrix with $N_{l+1} \times N_{l+1}$ diagonal blocks:

$$
\mathbf{z}^l = \left.\underbrace{\begin{bmatrix}
\color{red}{z_1^l(x^{(1)})} & \color{red}{z_1^l(x^{(2)})} & \cdots & z_1^l(x^{(|\mathcal{D}|)}) \\
\vdots & \vdots & \cdots & \vdots \\
z_j^l(x^{(1)}) & z_j^l(x^{(2)}) & \cdots & z_j^l(x^{(|\mathcal{D}|)}) \\
\vdots & \vdots & \ddots & \vdots \\
z_{N_{l+1}}^l(x^{(1)}) & z_{N_{l+1}}^l(x^{(2)}) & \cdots & z_{N_{l+1}}^l(x^{(|\mathcal{D}|)})
\end{bmatrix}}_{|\mathcal{D}|}\right\}\text{width} \implies \text{vec}(z^l) = \begin{bmatrix}
\color{red}{z_1^l(x^{(1)})} \\
z_2^l(x^{(1)}) \\
\vdots \\
z_{N_{l+1}}^l(x^{(1)}) \\
-- \\
\color{red}{z_1^l(x^{(2)})} \\
z_2^l(x^{(2)}) \\
\vdots \\
z_{N_{l+1}}^l(x^{(2)}) \\
-- \\
\vdots \\
\vdots \\
-- \\
z_1^l(x^{(|\mathcal{D}|)}) \\
z_2^l(x^{(|\mathcal{D}|)}) \\
\vdots \\
z_{N_{l+1}}^l(x^{(|\mathcal{D}|)}) \\
--
\end{bmatrix}
$$

$$
\sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix}
G(K_{1,1}^l) & \cdots & 0 & \cdots\cdots\cdots & G(K_{1,|\mathcal{D}|}^l) & \cdots & 0 \\
\vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\
0 & 0 & G(K_{1,1}^l) & \cdots\cdots\cdots & 0 & 0 & G(K_{1,|\mathcal{D}|}^l) \\
-- & -- & -- & ---- & -- & -- & -- \\
\color{red}{G(K_{2,1}^l)} & \cdots & 0 & \cdots\cdots\cdots & G(K_{2,|\mathcal{D}}^l) & \cdots & 0 \\
\vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\
0 & 0 & G(K_{2,1}^l) & \cdots\cdots\cdots & 0 & 0 & G(K_{2,|\mathcal{D}|}^l) \\
-- & -- & -- & ---- & -- & -- & -- \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
-- & -- & -- & -- & -- & -- & -- \\
G(K_{|\mathcal{D}|,1}^l) & \cdots & 0 & \cdots\cdots\cdots & G(K_{|\mathcal{D}|,|\mathcal{D}|}^l) & \cdots & 0 \\
\vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\
0 & 0 & G(K_{|\mathcal{D}|,1}^l) & \cdots\cdots\cdots & 0 & 0 & G(K_{|\mathcal{D}|,|\mathcal{D}|}^l)
\end{bmatrix}\right)
$$

$$
\begin{aligned}
\implies p(\mathbf{z}^l|K^l) &= \mathcal{N}\Big(\mathbf{0}, G\big(K^l\big) \otimes \mathbf{I}_{N_{l+1}\times N_{l+1}}\Big) \\
&= \mathcal{GP}\big(\mathbf{z}^l; 0, G(K^l)\big)
\end{aligned}
$$

5.3 $p(K^l|K^{l-1})$

Use marginal property of GP and look at: $p(K^l|K^{l-1})$:

$$
\begin{aligned}
p(K^l|K^{l-1}) &= \int_{z^{l-1}} p(K^l|z^{l-1}) p(z^{l-1}|K^{l-1}) \\
&= \int_{z^{l-1}} p(K^l|z^{l-1}) \mathcal{GP}\big(z^{l-1}; 0, G(K^{l-1})\big)
\end{aligned}
$$

- using GP property, and just look at two points $x^{(p)}$, $x^{(q)}$:

$$
\begin{aligned}
p(K^l_{p,q}|K^{l-1}_{p,q}) = \int_{z^{l-1}(x^{(p)}),\, z^{l-1}(x^{(q)})} & p\Bigg(\frac{1}{N_l}\sum_{j=1}^{N_l} \phi\big(z^l_j(x^{(p)})\big)\phi\big(z^l_j(x^{(q)})\big)\Bigg) \\
& \mathcal{N}\Bigg(\begin{bmatrix} z^{l-1}(x^{(p)}) \\ z^{l-1}(x^{(q)}) \end{bmatrix}; 0, G\Bigg(\begin{bmatrix} K^{l-1}(x^{(p)}, x^{(p)}) & K^{l-1}(x^{(p)}, x^{(q)}) \\ K^{l-1}(x^{(p)}, x^{(q)}) & K^{l-1}(x^{(q)}, x^{(q)}) \end{bmatrix}\Bigg)\Bigg)
\end{aligned}
$$

5.3.1 what happen to sum $\sum_{j=1}^{N_l} \phi\big(z^{l-1}_j(x^{(p)})\big)\phi\big(z^{l-1}_j(x^{(q)})\big)$ as $N_l \to \infty$ using CLT:

- look at $K^l_{p,q}$ and notice it's sum of iid random variable $K^{l,j}_{p,q}$:

$$
\begin{aligned}
\underbrace{K^l_{p,q}}_{\overline{X}} &= \frac{1}{N_l}\sum_{j=1}^{N_l} \underbrace{\phi\big(z^{l-1}_j(x^{(p)})\big)\phi\big(z^{l-1}_j(x^{(q)})\big)}_{X_j \equiv K^{l,j}_{p,q}} \\
\implies p(K^{l,1}_{p,q}|K^{l-1}_{p,q}) &= \int_{z^{l-1}(x^{(p)}),\, z^{l-1}(x^{(q)})} p\Big(\phi\big(z^l_j(x^{(p)})\big)\phi\big(z^l_j(x^{(q)})\big)\Big) \\
& \qquad \mathcal{N}\Bigg(\begin{bmatrix} z^{l-1}(x^{(p)}) \\ z^{l-1}(x^{(q)}) \end{bmatrix}; 0, G\Bigg(\begin{bmatrix} K^{l-1}(x^{(p)}, x^{(p)}) & K^{l-1}(x^{(p)}, x^{(q)}) \\ K^{l-1}(x^{(p)}, x^{(q)}) & K^{l-1}(x^{(q)}, x^{(q)}) \end{bmatrix}\Bigg)\Bigg) \\
&= (F \circ G)(K^{l-1}_{p,q})
\end{aligned}
$$

- using CLT, pick the most appropriate definition:

$$
(\overline{X} - \mathbb{E}[X_1]) \xrightarrow{d} \mathcal{N}\Bigg(0, \frac{\text{Var}[X_1]}{n}\Bigg)
$$

- let's see what is $\lim_{N_l \to \infty} p(K^l|K^{l-1})$:

$$
\begin{aligned}
(\overline{X} - \mathbb{E}[X_1]) &\xrightarrow{d} \mathcal{N}\Bigg(0, \frac{\text{Var}[X_1]}{n}\Bigg) \\
\implies \big(K^l_{p,q} - \mathbb{E}[K^{l,1}_{p,q}]\big) &\xrightarrow{d} \mathcal{N}\Bigg(0, \frac{\text{Var}[K^{l,1}_{p,q}]}{N_l}\Bigg) \\
\implies \big(K^l_{p,q} - (F \circ G)(K^{l-1}_{p,q})\big) &\xrightarrow{d} \mathcal{N}\Bigg(0, \frac{\text{Var}[K^{l,1}_{p,q}]}{N_l}\Bigg) \\
\implies \big(K^l_{p,q}|K^{l-1}_{p,q}\big) &\xrightarrow{d} \mathcal{N}\Bigg((F \circ G)(K^{l-1}), \frac{\text{Var}[K^{l,1}_{p,q}]}{N_l}\Bigg) \\
\implies \lim_{N_l \to \infty} p(K^l|K^{l-1}) &= \delta\big(K^l - (F \circ G)(K^{l-1})\big) \quad \text{entire matrix}
\end{aligned}
$$

- **note** using CLT, sample mean converge to δ_μ, can be exploited for other application
- note that this single step conditional is quite easy

5.4 putting in the overall objective function

let width of all layers to $\to \infty$:

$$
\begin{aligned}
p(z^L|x) &= \int p(z^L, K^0, K^1, \ldots, K^L|x)\, \mathrm{d}K^{0,\ldots,L} \\
&= \int p(z^L|K^L) \Big(\prod_{l=1}^{L} p(K^l|K^{l-1}) \Big) p(K^0|x)\, \mathrm{d}K^{0,\ldots,L} \\
\lim_{N_L \to \infty, \ldots, N_1 \to \infty} p(z^L|x) &= \int p(z^L|K^L) \Big(\prod_{l=1}^{L} \delta\big(K^l - (F \circ G)(K^{l-1})\big) \Big) p(K^0|x)\, \mathrm{d}K^{0,\ldots,L} \\
&= \int \mathcal{GP}\big(z^L; 0, G(K^L)\big) \underbrace{\Big(\prod_{l=1}^{L} \delta\big(K^l - (F \circ G)(K^{l-1})\big) \Big) \delta\Big(K^0 - \frac{1}{d_{\text{in}}} x^\top x\Big)}_{=\begin{cases} = 1 & \text{if } K^L = (F \circ G)(K^{L-1}) \\ & \qquad\quad = (F \circ G)^2(K^{L-2}) \ldots \\ & \qquad\quad = (F \circ G)^L \Big(\frac{1}{d_{\text{in}}} x^\top x\Big) \\ = 0 & \text{otherwise} \end{cases}} \mathrm{d}K^{0,\ldots,L} \\
&= \mathcal{GP}\Big(z^L; 0,\ G \circ (F \circ G)^L \big(\frac{1}{d_{\text{in}}} x^\top x\big)\Big)
\end{aligned}
$$