4. Training & Posterior Distribution

4.1 NNGPs

At infinite width, NNs define a class of GPs

prior covariance: $K_{\delta_1 \delta_2} = \lim_{n \to \infty} E\left[\phi^{(L)}_{\delta_1} \phi^{(L)}_{\delta_2} \right]$.

Bayesian inference then yields the posterior distr. as seen in L2.

↳ no fitting is required.

4.2 Training

Introduce training time t, $t=0$ initialization

Loss fn. $\alpha \in \mathcal{A}$, indices of the training set, $\mathcal{A} \subset \mathcal{D}$

MSE: $\mathcal{L}_A = \sum_\alpha \left(\phi^{(L)}_{i\alpha} - y_{i\alpha} \right)^2, \quad \alpha \in \mathcal{A}$.

Gradient Descent: $\frac{d}{dt} \theta_\mu(t) = - \partial_\mu \mathcal{L}_A$

sometimes, it can be generalized: $- \lambda_{\mu\nu} \partial_\nu \mathcal{L}_A$.

Using the chain rule:

$$\partial_\mu \mathcal{L}_A = \sum_{i,\alpha} \frac{\partial \mathcal{L}_A}{\partial \phi^{(L)}_{i\alpha}(t)} \partial_\mu \phi^{(L)}_{i\alpha}(t)$$

$$\Rightarrow \frac{d}{dt} \theta_\mu = - \sum_{\nu, i, \alpha} \lambda_{\mu\nu} \partial_\nu \phi^{(L)}_{i\alpha}(t) \, \varepsilon_{i\alpha}(t)$$

$$\varepsilon_{i\alpha}(t) = \frac{\partial \mathcal{L}_A}{\partial \phi^{(L)}_{i\alpha}(t)} = 2 \left(f^{(L)}_{i\alpha} - y_{i\alpha} \right) \quad \text{for MSE.}$$

Flow eq. for the fields $\phi^{(L)}$.

$$\frac{d}{dt}\phi^{(L)}_{i;\delta}(t) = \sum_\mu \partial_\mu \phi^{(L)}_{i;\delta} \frac{d}{dt}\theta_\mu$$

$$= -\sum_{j,\alpha}\Big\{\sum_{\mu,\nu}\lambda_{\mu\nu}\,\partial_\mu \phi^{(L)}_{i;\delta}\,\partial_\nu \phi^{(L)}_{j;\alpha}\Big\}\varepsilon_{j\alpha}$$

$$= -H_{i\delta,j\alpha}(t)\,\varepsilon_{j\alpha}(t) \qquad \text{omitted the sum}$$

$\uparrow$ $\alpha \in \mathcal{A}$!

$$\Rightarrow \quad \frac{d\phi^{(L)}(t)}{dt} = -H(t)\,\varepsilon(t) \qquad \text{omitted } \underline{\text{all}} \text{ indices.}$$

$\uparrow$ Neural Tangent Kernel.

Exercise: Show that

$$\frac{d}{dt}O(\phi^{(L)}) = -\frac{\partial O}{\partial \phi^{(L)}} H\,\varepsilon$$

(put back all indices & summations !).

4.3 1/n Expansion

$$H_{i\delta,j\alpha} = \delta_{ij}\Theta_{\delta\alpha} + O(1/n)$$

$\uparrow$ time-independent.

Correlations btw neurons only appear at $O(1/n)$!

Evolution is better understood by considering a series of fields

$$O_s(t) = O_{\delta_1 \dots \delta_s}(t)$$

such that :

$$\begin{cases} O_s(t) = \partial_\mu O_{s-1}(t)\,\partial_\mu \phi^{(L)}_{\delta_s}(t) \\ O_1(t) = \phi^{(L)}_{\delta_1}(t) \end{cases}$$

Then:

$$\frac{d}{dt} O_s(t) = \partial_\mu O_s(t) \frac{d}{dt} \theta_\mu(t)$$

$$= - \partial_\mu O_s(t) \partial_\mu \phi_\delta \varepsilon_\delta(t)$$

$$= - O_{s+1}(t) \varepsilon(t)$$

Check:

$$O_2(t) = O_{\delta_1 \delta_2}(t) = \partial_\mu O_{\delta_1}(t) \partial_\mu \phi^{(1)}_{\delta_2}(t)$$

$$= H_{\delta_1 \delta_2}(t)$$

$$\frac{d\phi^{(1)}}{dt} = - H \varepsilon \qquad \checkmark$$

And we have:

$$\frac{dH}{dt} = - O_3 \varepsilon \qquad \leftarrow \text{RHS is } O(1/n).$$

$$\frac{dO_3}{dt} = - O_4 \varepsilon \qquad \leftarrow \text{RHS is } O(1/n)$$

$$\frac{dO_4}{dt} = O\left(\frac{1}{n^2}\right) \to 0 \quad \text{if we are only interested in } O(1/n).$$

Expand all quantities in powers of $1/n$.

$$\phi = \phi^{(0)} + \phi^{(1)} + \cdots$$

$$H = H^{(0)} + H^{(1)} + \cdots$$

$$O_3 = O_3^{(1)} + \cdots$$

$$O_4 = O_4^{(1)} + \cdots$$

$O(1)$ $\quad H^{(0)}(t) = H^{(0)}$ const.

$$\Rightarrow \quad \phi^{(0)}(t) - y = e^{-H^{(0)} t} \left(\phi^{(0)}(0) - y\right).$$

$O(1/n)$: $\quad \dfrac{dO_3^{(1)}}{dt} = \dfrac{dO_3}{dt} = -O_4(0)\left(\phi^{(0)}(t) - y\right).$

$$\Rightarrow \quad O_3(t) = O_3(0) - \int_0^t dt'\, O_4(0)\left(\phi^{(0)}(t') - y\right)$$

$$= O_3(0) - O_4(0) \int_0^t dt'\, e^{-H^{(0)} t'} \left(\phi^{(0)}(0) - y\right)$$

$$= O_3(0) - O_4(0) \left[H^{(0)}\right]^{-1} \left(1 - e^{-t H^{(0)}}\right) \left(\phi^{(0)}(0) - y\right).$$

$$H^{(1)}(t) = H^{(1)}(0) - \int_0^t dt'\, O_3^{(1)}(t') \left(\phi^{(0)}(t') - y\right).$$

can also be computed explicitly.

$$\frac{d\phi^{(1)}}{dt} = -\left[H^{(0)} \phi^{(1)}(t) + H^{(1)}(t)\left(\phi^{(0)}(t) - y\right)\right].$$

$$\phi^{(1)} = e^{-H^{(0)}t}\,\hat{\phi}^{(1)}(t)$$

$$\frac{d\phi^{(1)}}{dt} = -H^{(0)} e^{-H^{(0)}t}\,\hat{\phi}^{(1)}(t) + e^{-H^{(0)}t}\,\frac{d}{dt}\hat{\phi}^{(1)}(t)$$

$$= -H^{(0)}\phi^{(1)}(t) - H^{(1)}(t)\,(\phi^{(0)}(t) - y)$$

$$\frac{d}{dt}\hat{\phi}^{(1)} = -e^{H^{(0)}t}\,H^{(1)}(t)\,e^{-H^{(0)}t}\,(\phi^{(0)}(0) - y)$$

$$\hat{\phi}^{(1)}(t) = -\int_0^t e^{+H^{(0)}t'}\,H^{(1)}(t')\,e^{-H^{(0)}t'}\,(\phi^{(0)}(0) - y)\,.$$

Finally

$$\phi(t) = y + e^{-H^{(0)}t}\left\{1 - \int_0^t dt'\, e^{H^{(0)}t'}\,H^{(1)}(t')\,e^{-H^{(0)}t'}\right\}(\phi(0) - y)\,.$$

↑ analytic description of the network during the training

Quantify over/underfitting ?

Monitor bias & variance in a document ?

Eigenvectors/eigenvalues of $H^{(0)}$?

Linear Networks. (Frozen weights f. $l < L$).

$$f^{(L)}(t) = \partial_\mu f^{(L)} \theta_\mu(t)$$

$$\dot{\theta}_\mu(t) = -\eta \sum_\alpha \partial_\mu f_\alpha^{(L)} \left(\sum_\nu \partial_\nu f_\alpha^{(L)} \theta_\nu(t) - y_\alpha \right).$$

$$\Rightarrow \theta_\mu(t) = -\partial_\mu f^{(L)} H^{-1} \left(1 - e^{-H\eta t}\right) \left(f^{(L)}(0) - y\right) + \theta_\mu(0)$$

$$\Rightarrow f_{\alpha_1}^{(L)}(t) = \left(1 - e^{-\eta H t}\right)_{\alpha_1 \alpha_2} y_{\alpha_2} + \left(e^{-\eta H t}\right)_{\alpha_1 \alpha_2} f_{\alpha_2}^{(L)}(0)$$

$$f_\beta^{(L)}(t) = f_\beta^{(L)}(0) + H_{\beta\alpha_1} H^{-1}_{\alpha_1 \alpha_2} \left(1 - e^{-\eta H t}\right)_{\alpha_2 \alpha_3} \left(y - f^{(L)}(0)\right)_{\alpha_3}$$

$$\lambda_{b^{(l)}_{i_1} b^{(l)}_{i_2}} = \delta_{i_1 i_2} \lambda_b^{(l)}, \qquad \lambda_{w^{(l)}_{i_1 j_1} w^{(l)}_{i_2 j_2}} = \delta_{i_1 i_2} \delta_{j_1 j_2} \frac{\lambda_w^{(l)}}{n_{l-1}}$$

$$\hat{H}^{(l+1)}_{i_1\alpha_1; i_2\alpha_2} = \sum_{j=1}^{n_{l+1}} \left\{ \lambda_b^{(l+1)} \frac{\partial f^{(l+1)}_{i_1\alpha_1}}{\partial b^{(l+1)}_j} \frac{\partial f^{(l+1)}_{i_2\alpha_2}}{\partial b^{(l+1)}_j} + \frac{\lambda_w^{(l+1)}}{n_l} \sum_{k=1}^{n_l} \frac{\partial f^{(l+1)}_{i_1\alpha_1}}{\partial w^{(l+1)}_{jk}} \frac{\partial f^{(l+1)}_{i_2\alpha_2}}{\partial w^{(l+1)}_{jk}} \right\} +$$

$$+ \sum_{j_1 j_2=1}^{n_l} \frac{\partial f^{(l+1)}_{i_1\alpha_1}}{\partial f^{(l)}_{j_1\alpha_1}} \frac{\partial f^{(l+1)}_{i_2\alpha_2}}{\partial f^{(l)}_{j_2\alpha_2}} \hat{H}^{(l)}_{j_1\alpha_1 j_2\alpha_2} .$$

If we only consider the parameters in the last layer, $b^{(l)}_i = 0$

$$\hat{H}^{(l)}_{i_1\alpha_1 i_2\alpha_2} = \frac{\lambda_w}{n_{l-1}} \sum_{j,k} \delta_{i_1 j} \delta_{i_2 j} \rho^{(l-1)}_{k\alpha_1} \rho^{(l-1)}_{k\alpha_2}$$

$$= \frac{\lambda_w}{n_{l-1}} \delta_{i_1 i_2} \rho^{(l-1)}_{\alpha_1} \cdot \rho^{(l-1)}_{\alpha_2}$$

if $\lambda_w = C_w$ then $H^{(l)} = G^{(l)}$

$\Rightarrow$ reproduce NNGP.