Neural Network Hessians + Free Probability

Model NN: $f(x) = W\rho(Vx)$, $\rho(t) = \max\{0,t\}$ (ReLU)

("1 1/2 layer network")

$= f(x;\theta)$ $\quad \theta = (W,V)$

Assume all dimensions $= d$, i.e.

$x \in \mathbb{R}^d$, $V, W \in \mathbb{R}^{d\times d}$, $f(x) \in \mathbb{R}^d$.

Data: $(x_1, y_1), \dots, (x_m, y_m)$

Assume: $\frac{d}{m} \longrightarrow c \in (0,\infty)$

Interpretations:

(1) "Big data" regime: $\frac{\#\text{features per data pt.}}{\#\text{ data pts.}} = O(1)$

(2) "Overparametrization": $\frac{\#\text{ model params}}{\text{total \# data dim.}} = O(1)$

$\frac{2d^2}{dm} \longrightarrow 2c.$

$\hat{y}_k(\theta)$

$$L(\theta) = \frac{1}{2}\sum_{k=1}^{m} \|f(x_k;\theta) - y_k\|^2 = \frac{1}{2}\sum_{k=1}^{m}\sum_{i=1}^{d} (\hat{y}_{ki}(\theta) - y_{ki})^2$$

minimize with gradient-type steps.

Previously: Behavior of $f(x;\theta)$ at initial val. θ.

Now: Behavior of $L(\theta)$, where does minimization stop?

Critical pts. of L: θ s.t. $\nabla L(\theta) = 0$.



Q: How does $L(\theta)$ behave near its critical points?

Specified by Hessian matrix $\nabla^2 L(\theta)$ $\overset{\theta_0}{=} H$ at $\theta = \theta_0$.

Near critical pt. θ_0, $L(\theta) \approx L(\theta_0) + \frac{1}{2}(\theta-\theta_0)^T H (\theta-\theta_0)$

Relevant property: sign of eigvals of $\nabla^2 L(\theta)$,

negative eigvals = # "descent directions".

Question: How to predict eigval distrib. of $\nabla^2 L(\theta)$ at crit. pts θ? (In enough detail to get #, maybe size, of postive (negative eigvals))

$$\nabla L(\theta)_\alpha = \partial_\alpha L(\theta)$$

$$= \sum_{j=1}^{m} \sum_{i=1}^{d} \underbrace{(\hat{y}_{ki}(\theta) - y_{ki})}_{= r_{ki}(\theta) \text{ ("residual")}} \partial_\alpha \hat{y}_{ki}(\theta)$$

α, β: abstract indices in θ, entries of W or V.

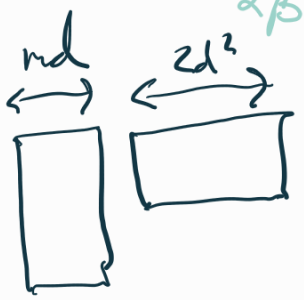
Total # $= 2d^2$.

$\alpha = V_{ab}$ or W_{cd}

$$\nabla^2 L(\theta)_{\alpha\beta} = \partial_\alpha \partial_\beta L(\theta) = \partial_\beta \partial_\alpha L(\theta)$$

$$= \underbrace{\sum_{k=1}^{m} \sum_{i=1}^{d} \partial_\alpha \hat{y}_{ki}(\theta)\, \partial_\beta \hat{y}_{ki}(\theta)}_{H^{(1)}_{\alpha\beta}} + \underbrace{\sum_{k=1}^{m} \sum_{i=1}^{d} r_{ki}(\theta)\, \partial_\alpha \partial_\beta \hat{y}_{ki}(\theta)}_{H^{(2)}_{\alpha\beta}}$$

$$H = \nabla^2 L(\theta) = H^{(1)} + H^{(2)}$$



$$H^{(1)}_{\alpha\beta} = \langle v_\alpha, v_\beta \rangle \implies H^{(1)} = W^T W \geq 0$$

(Interp: eigvals make more positive; fewer descent dir.)

Guess of eigval dist? Suppose V had iid entries.

Then, $\lambda(H^{(1)}) \overset{\sim}{\propto} \mu_{MP}\left(\frac{2d^2}{md}\right) = \mu_{MP}(c)$

↳ Marchenko-Pastur from last time.

$$H^{(2)} = \sum_{k=1}^{m} \sum_{i=1}^{d} r_{ki}(\theta)\, \nabla^2 \hat{y}_{ki}(\theta)$$

$r_{ki}(\theta)$: depends on "label" y_{ki}

$\nabla^2 \hat{y}_{ki}(\theta)$: ↳ only depends on architecture + x_k

Heuristic: behaves like randomly weighted sum of $\nabla^2 \hat{y}_{ki}(\theta) \rightsquigarrow$ symmetric eg'vl dist.

$$H = H^{(1)} + H^{(2)}$$

(???)



Later justify heuristically viewing $H^{(1)}, H^{(2)}$ indep.

Q: Given two "nice" indep rand. mxs. X, Y, if esd of $X \approx \mu$, esd $Y \approx \nu$, what can we say about esd. of $X+Y$?

Next goal: develop tools (free probability) to answer questions (under assump. on X, Y).

Intro to Free Probability Ideas

Our method for studying esd: trace moments.

Def: For $X \in \mathbb{R}^{d\times d}$, $\mathrm{tr}(X) := \frac{1}{d}\mathrm{Tr}(X)$ ("normalized trace")

Method: calculate $\lim_{d\to\infty} \mathbb{E}\,\mathrm{tr}\left((X^{(d)})^k\right)$.

Suppose can do this for $X = X^{(d)}$, $Y = Y^{(d)}$:

$$\forall k \geq 1, \begin{cases} \lim_{d\to\infty} \mathbb{E}\,\mathrm{tr}\, X^k = m_k \\ \lim_{d\to\infty} \mathbb{E}\,\mathrm{tr}\, Y^k = n_k \end{cases}$$

Q: Can we show $\lim_{d\to\infty} \mathbb{E}\,\mathrm{tr}(X+Y)^k = s_k$ (some sequence $\in \mathbb{R}$)?

Idea: combinatorial expansion, e.g.

$$(X+Y)^2 = X^2 + XY + YX + Y^2 \quad (2^k \text{ terms})$$

Unlike scalars, for X, Y mxs. need to track order of terms.

Def: Word $w \in \{X, Y\}^k$ ← length; sequence of letters

By abuse of notation, $w = w(X, Y) \in \mathbb{R}^{d \times d}$, mx. valued polynomial of X, Y.

$$t(w) := \lim_{d \to \infty} \mathbb{E}\, \mathrm{tr}\left(w(X, Y)\right)$$

"word expansion"

$$\rightsquigarrow s_k = \lim_{d \to \infty} \mathrm{tr}\,(X + Y)^k$$

$$= \lim_{d \to \infty} \sum_{w \in \{X,Y\}^k} \mathrm{tr}\left(w(X, Y)\right) = \sum_{w \in \{X,Y\}^k} t(w).$$

Evaluating limits of word traces:

assump. on X

Immediately:

$$\begin{cases} t(\underbrace{X \cdots X}_{k \text{ times}}) = \lim_{d \to \infty} \mathbb{E}\, \mathrm{tr}(X^k) = m_k \\ t(Y \cdots Y) = n_k \end{cases}$$

Assump. on X, Y:

(1) X, Y ind. random mxs.

(2) X, Y both symmetric-orthogonally invariant.

Prop: If X sym. orth. invar., $X \in \mathbb{R}^{d\times d}_{sym}$

$$\mathbb{E}X^k = (\mathbb{E}\,\mathrm{tr}\,X^k)\,I_d$$

Pf: $X \overset{d}{=} QXQ^T \quad \forall Q \in O(d)$.

$$\Rightarrow \underbrace{\mathbb{E}X^k}_{M} = \mathbb{E}(QXQ^T)^k = \mathbb{E}QX^kQ^T$$

$$= Q(\mathbb{E}X^k)Q^T \quad (M \text{ deterministic})$$

$$\Rightarrow M = QMQ^T \quad \forall Q \in O(d).$$

$$\Rightarrow M = c\cdot I_d. \qquad \text{“} t(X^aY^b)$$

Cor: $t(\underbrace{X\cdots X}_{a \text{ times}}\underbrace{Y\cdots Y}_{b \text{ times}}) = m_a n_b$, (under assump.)

cyclic property $\Rightarrow t(X^aY^bX^c) = t(X^{a+c}Y^b)$

$$= m_{a+c}\,n_b.$$

Pf: (LHS) $= \lim \mathbb{E}\,\mathrm{tr}(X^aY^b)$

$$= \lim \mathrm{tr}\left(\mathbb{E}[X^aY^b]\right) = \lim \mathrm{tr}\left(\mathbb{E}[X^a]\,\mathbb{E}[Y^b]\right)$$

$$= \lim \operatorname{tr}\Big(\underbrace{(\mathbb{E}\operatorname{tr} X^a)}_{\text{scalar}} I_d \cdot \underbrace{(\mathbb{E}\operatorname{tr} Y^b)}_{\text{scalar}} I_d \Big)^3$$

$$= \lim (\mathbb{E}\operatorname{tr} X^a)(\mathbb{E}\operatorname{tr} Y^b) \operatorname{tr}(I_d) \to 1$$

$$= m_a n_b. \quad \blacksquare$$

Simple manipulations calculate $t(w)$ for $w = X^k$, $w = X^a Y^b$, $w = Y^a X^b$, $w = X^a Y^b X^c$, etc.

Next time: issue of "tangled" words

e.g. $w = XYXY$.