Neurodynamika statystyczna głębokich sieci: geometria przestrzeni sygnałów

Mateusz Przyborowski

16 października 2020 r.

Główne wyniki z pracy:

1. Tensor metryczny, poprzez kolejne warstwy, zmienia się w sposób konforemny.
2. Krzywizna rozmaitości sygnału zbiega do stałej (pod pewnymi założeniami).
3. Wyjaśniony jest opis zmiany odległości pomiędzy dwoma sygnałami w kolejnych warstwach.

Oznaczenia

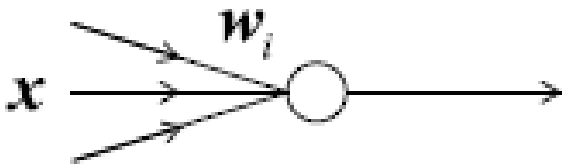
Rozważmy sieć neuronową o m neuronach i n-wymiarowym wejściu. Oznaczmy przez **x** wejście, **y** wynik sieci, $\mathbf{w_{ik}}$ waga połączenia pomiędzy **k**-tym wejściem i **i**-tym neuronem, $\mathbf{b_i}$ bias **i**-tego neuronu. Φ to funkcja błędu.

$$x = (x_1, \cdots, x_n)$$

$$y = (y_1, \cdots, y_m)$$

$$u_i = \sum_\kappa w_{i\kappa} x_\kappa + b_i$$

$$y_i = \varphi(u_i)$$



$$y = \varphi(u_i)$$

$$u_i = \sum w_{i\kappa} x_\kappa + b_i$$

$$\varphi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} \exp\left\{-\frac{v^2}{2}\right\} dv$$

Oznaczenia

Zakładamy, że wagi połączeń są niezależnymi zmiennymi losowymi o rozkładzie normalnym ze średnią 0 oraz wariancjami σ^2/n i σ_b^2 dla wag i biasu. Wówczas u_i są niezależnymi zmiennymi losowymi wspólnego rozkładu normalnego o średniej 0 i wariancji: $\tau^2 = \frac{\sigma^2}{n}\sum x_\kappa^2 + \sigma_b^2$

Dalej będziemy badali głęboką sieć neuronową, gdzie **t**-ta warstwa ma wejście x^{t-1}, które jest wynikiem działania (**t-1**)-ej warstwy: $x^t = \varphi\left(W^t x^{t-1} + b^t\right)$

Podobnie zakładamy, że wagi i bias w^t_{ij} i b^t_i są niezależnymi zmiennymi losowymi rozkładów normalnych o średniej 0 i wariancjami odpowiednio σ_t^2/n_{t-1} i σ_b^2 .

X^t to przestrzeń możliwych wyników działania sieci w **t**-tej warstwie. Przyjmując początkowy sygnał $X = X^0$, przez $\hat{X}^t$ oznaczmy jego obraz w **t**-tej warstwie.

Przetwarzanie sygnału wejściowego

Przy zadanym sygnale $\mathbf{x}^t$ rozważmy funkcję ilości aktywności, zdefiniowanej jako:

$$A^t = \frac{1}{n^t}\sum \left(x_i^t\right)^2 = \frac{1}{n^t}\sum \left\{\varphi\left(u_i^t\right)\right\}^2 = \mathrm{E}\left[\left\{\varphi\left(u_i^t\right)\right\}^2\right]$$

By badać wielkości $\mathbf{A}^t$ połóżmy również:

$$\chi_0(\sigma^2, A) = \int \varphi^2\left(\sigma_A v\right) Dv$$

$$\sigma_A^2 = \sigma^2 A + \sigma_b^2$$

$$Dv = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{v^2}{2}\right\} dv$$

$$\chi_1\left(\sigma^2, A\right) = \sigma^2 \int \left\{\varphi'\left(\sigma_A v\right)\right\}^2 Dv$$

Jako że $\mathbf{u}^t_i$ ma rozkład normalny ze średnią 0 i wariancją $\tau_t^2 = \sigma_t^2 A^{t-1} + \sigma_b^2$.

stąd mamy: $\mathrm{E}\left[\{\varphi\left(u_i^t\right)\}^2\right] = \frac{1}{\sqrt{2\pi}\tau_t}\int \{\varphi(u)\}^2 \exp\left\{-\frac{u^2}{2\tau_t^2}\right\} du = \chi_0\left(\sigma_t^2, A^{t-1}\right)$

W szczególności, przy funkcji błędu Φ zachodzi $\chi_0(\sigma^2, A) = \frac{1}{2\pi}\cos^{-1}\left(\frac{-\sigma_A^2}{1+\sigma_A^2}\right)$

Twierdzenie.

Aktywność sygnału w sieci rozwija się zgodnie ze wzorem $A^t = \chi_0\left(\sigma_t^2, A^{t-1}\right)$

Jako że $\boldsymbol{\chi_0}$ jest monotonicznie rosnącą funkcją **A**, gdy dla wszystkich **t** mamy $\boldsymbol{\sigma_t^2 = \sigma^2}$, to istnieje dokładnie jedno stabilne ekwilibrium $\hat{\mathbf{A}}$, a dla dużych **t A** zbiega do $\hat{\mathbf{A}}$.

Zdefiniowanie metryki

Niech **X** = $\mathbf{X}^0$ będzie przestrzenią euklidesową. Połóżmy $dx = \sum dx_\kappa e_\kappa$ przy wektorach bazy ortonormalnej **e**.

W kolejnych warstwach, proste z **X** stają się krzywymi w $\tilde{\mathbf{X}}^t$, stąd dla macierzy Jacobiego **B**, $B_{ij}^t = \frac{\partial \tilde{x}_i^t}{\partial \tilde{x}_j^{t-1}} = \frac{\partial \varphi\left(u_i^t\right)}{\partial \tilde{x}_j^{t-1}}$, mamy $d\tilde{x}_i^t = \sum B_{ij}^t d\tilde{x}_j^{t-1}$

Wektory bazowe przestrzeni stycznej do $\tilde{\mathbf{X}}^t$ są zdefiniowane jako:

$$e_{\kappa i}^t = \frac{\partial \tilde{x}_i^t}{\partial x_\kappa} = \sum_j B_{ij}^t e_{\kappa j}^{t-1}, \quad i = 1, \cdots, n_t$$

Stąd wprowadzamy tensor metryczny: $g_{\kappa\lambda}^t = \langle \tilde{e}_\kappa^t, \tilde{e}_\lambda^t \rangle = \sum e_{\kappa i}^t e_{\lambda i}^t$

Zauważmy, że $\langle \tilde{e}_\kappa^t, \tilde{e}_\lambda^t \rangle = \sum_{i,j,k,l} B_{ik}^t B_{jl}^t \delta_{ij} e_{\kappa l}^{t-1} e_{\lambda k}^{t-1}$

Dla dość dużych $\mathbf{n_t}$, z prawa wielkich liczb, mamy:

$$\frac{1}{n_t} \sum_i B_{ik}^t B_{il}^t = \mathrm{E}\left[\left\{\varphi'\left(u_i^t\right)\right\}^2 w_{ik}^t w_{il}^t\right] = \mathrm{E}\left[\left\{\varphi'\left(u_i^t\right)\right\}^2\right] \mathrm{E}\left[w_{ik}^t w_{il}^t\right]$$

Natomiast z poprzednich wyników mamy:

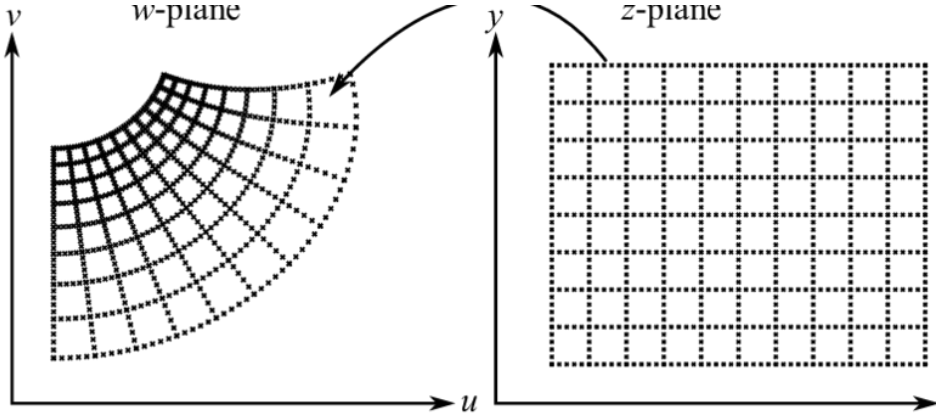
$$\chi_{1,t}\left(\sigma_t^2, A^{t-1}\right) = \sigma_t^2 \mathrm{E}\left[\left\{\varphi'\left(u_i^t\right)\right\}^2\right]$$

$$\mathrm{E}\left[w_{ik}^t w_{il}^t\right] = \frac{\sigma_t^2}{n_t} \delta_{kl}$$

Twierdzenie.

Metryka jest przekształcana zgodnie ze wzorem: $g_{\kappa\lambda}^t(x) = \chi_{1,t}(x) g_{\kappa\lambda}^{t-1}(x)$

Jest to przekształcenie konforemne.



Metryka w warstwie **t**-tej jest wyprowadzona jako $g^t_{\kappa\lambda}(x) = \delta_{\kappa\lambda} \prod_{s=1}^{t-1} \chi_{1,s}$

Dla $\sigma_t^2 = \sigma^2$ i dla dość dużych **t**, $\mathbf{A}^t$ zbiega do $\hat{\mathbf{A}}$. Stąd asymptotycznie mamy:

$$g^t_{\kappa\lambda}(x) \approx \{\bar{\chi}_1\}^{t-1} \delta_{\kappa\lambda},\ \bar{\chi}_1 = \chi_1\left(\sigma^2, \bar{A}\right)$$

Dla ϕ funkcji błędu mamy $\chi_1(\sigma^2, A) = \frac{\sigma^2}{2\pi} \frac{\sigma_b^2 + \sigma^2 A}{\sqrt{1 + 2\left(\sigma_b^2 + \sigma^2 A\right)}}$

Gdy $\bar{\chi}_1 < 1,\ g^t_{\kappa\lambda}$ zbiega do zera.

Gdy $\bar{\chi} > 1$ (dzieje się tak przy dużej wariancji), długość krzywych rośnie wraz ze wzrostem **t**, ale skoro sygnał leży w ograniczonej podprzestrzeni $\mathbf{X}^t$ (z wyjątkiem niektórych funkcji aktywacji, np. ReLU), to krzywa musi być bardzo poskręcana.

Krzywizna

Krzywizna rozmaitości $\hat{\mathbf{X}}^t$ jest mierzona przez to, jak bardzo bazowe wektory $\hat{\mathbf{e}}^t_k$ przestrzeni stycznej do $\hat{\mathbf{X}}^t$ się zmieniają gdy punkty poruszają się w kierunku $\hat{\mathbf{e}}^t_l$. Za pomocą pochodnej kierunkowej kładziemy $H^t_{\kappa\lambda} = \nabla_\lambda \tilde{e}^t_\kappa$

Zauważmy, że dla $\partial_\kappa = \frac{\partial}{\partial x_\kappa}$ mamy:

$$H^t_{\kappa\lambda i} = \partial_\lambda \partial_\kappa \varphi\left(u^t_i\right) = \partial_\lambda e^t_{\kappa i} = \partial_\kappa \left\{ \sum \varphi'\left(u_i\right) w_{ij} \partial_\lambda \tilde{x}^{t-1}_j \right\} =$$

$$= \varphi''\left(u_i\right)\left(w_i \cdot \partial_\kappa \tilde{x}^{t-1}\right)\left(w_i \cdot \partial_\lambda \tilde{x}^{t-1}\right) + \varphi'\left(u_i\right) w_i \cdot \partial_\kappa \partial_\lambda \tilde{x}^{t-1} =$$

$$= \varphi''\left(u_i\right)\left(w_i \cdot \tilde{e}^{t-1}_\kappa\right)\left(w_i \cdot \tilde{e}^{t-1}_\lambda\right) + \varphi'\left(u_i\right) w_i \cdot \boldsymbol{H}^{t-1}_{\kappa\lambda}$$

Przy ostatniej równości korzystamy z tego, że $\boldsymbol{H}^{t-1}_{\kappa\lambda} = \partial_\kappa \partial_\lambda \tilde{x}^{t-1}$

Teraz możemy położyć jako wielkość krzywizny $\left|\boldsymbol{H}^t_{\kappa\lambda}\right|^2 = \left\langle \boldsymbol{H}^t_{\kappa\lambda}, \boldsymbol{H}^t_{\kappa\lambda} \right\rangle = \sum_i \left(H^t_{\kappa\lambda i}\right)^2$

Z prawa wielkich liczb mamy:

$$\left|\boldsymbol{H}_{\kappa\lambda}^{t}\right|^{2} = n_t \mathrm{E}\left[\left\{\varphi''(u_i)\right\}^2 \left(\boldsymbol{w}\cdot\tilde{\boldsymbol{e}}_{\kappa}^{t-1}\right)^2 \left(\boldsymbol{w}\cdot\tilde{\boldsymbol{e}}_{\lambda}^{t-1}\right)^2\right] + 2n_t \mathrm{E}\left[\varphi'(u_i)\varphi''(u_i)\left(\boldsymbol{w}\cdot\tilde{\boldsymbol{e}}_{\kappa}^{t-1}\right)\left(\boldsymbol{w}\cdot\tilde{\boldsymbol{e}}_{\lambda}^{t-1}\right)\left(\boldsymbol{w}\cdot\partial_{\kappa}\tilde{\boldsymbol{e}}_{\lambda}^{t-1}\right)\right] + n_t \mathrm{E}\left[\left\{\varphi'(u_i)\right\}^2 \left(\boldsymbol{w}\cdot\boldsymbol{H}_{\kappa\lambda}^{t-1}\right)^2\right]$$

Środkowy wyraz upraszcza się z nieparzystości funkcji. Pozostałe dwa można przeformułować do postaci:

$$\left|\boldsymbol{H}_{\kappa\lambda}^{t}\right|^{2} = \chi_1\left(\sigma_t^2, A^{t-1}\right)\left|\boldsymbol{H}_{\kappa\lambda}^{t-1}\right|^{2} + \frac{1}{n_t}\left(1+2\delta_{\kappa\lambda}\right)\left\{\chi_2\left(\sigma_t^2, A^{t-1}\right)\right\}\left(g_{\kappa\kappa}^{t-1}\right)^2\left(g_{\lambda\lambda}^{t-1}\right)^2$$

Dalej upraszczając kładziemy:

$$\gamma_t^2 = \frac{1}{n_t}\sum\left(g_t^{-1}\right)^{\kappa\mu}\left(g_t^{-1}\right)^{\lambda\nu} H_{\kappa\lambda i}^{t} H_{\mu\nu j}^{t}\delta_{ij} = \frac{\gamma_{t-1}^2}{\chi_{1,t-1}} + 3\frac{\chi_{2,t-1}}{n_t\left(\chi_{1,t-1}\right)^2} = 3\frac{1}{n_t}\sum_{s=1}^{t-1}\frac{\chi_{2,t-s}}{\chi_{1,t-s}^2}\left(\prod_{r=1}^{s}\frac{1}{\chi_{1,t-r}}\right)$$

Dla dość dużych **t**, $\boldsymbol{\sigma_t^2 = \sigma^2}$ i $\bar{\chi} > 1$ mamy $\gamma_t^2 = \dfrac{3\chi_2}{n_t\chi_1\left(\chi_1-1\right)}$
co oznacza, że skalar krzywizny zbiega do stałej.
Dla $\bar{\chi}_1 < 1$ jest rozbieżny.

Odległość między sygnałami

Niech $\mathbf{x}^{t-1}$ i $\mathbf{y}^{t-1}$ ($\mathbf{x}^{t}$ i $\mathbf{y}^{t}$) oznaczają dwa sygnały na wejściu (wyjściu) **t**-tej warstwy.

Połóżmy $D_t = D\left(x^t, y^t\right) = \frac{1}{n_t}\sum\left(x_i^t - y_i^t\right)^2$ i $C_t = C\left(x, y\right) = \frac{1}{n_t\sqrt{A\left(x^t\right)A\left(y^t\right)}}\sum x_i^t y_i^t$
do badania odpowiednio odległości i nachodzenia się sygnałów. Zachodzi:

$$D\left(x, y\right) = A(x) + A\left(y\right) - 2\sqrt{A(x)A(y)}C\left(x, y\right)$$

A więc znając $C_t = \psi\left(C_{t-1}\right)$ możemy wyrazić $D_t = \xi\left(D_{t-1}\right)$

Załóżmy dodatkowo, że $\mathbf{A(x)} = \mathbf{A(y)} = \mathbf{A^t}$; wówczas mamy $D_t = 2A^t\left(1 - C_t\right)$

a więc możemy wyrazić odległość jako $\xi\left(D_{t-1}\right) = 2A^t\left\{1 - \psi\left(1 - \frac{D_{t-1}}{2A^t}\right)\right\}$

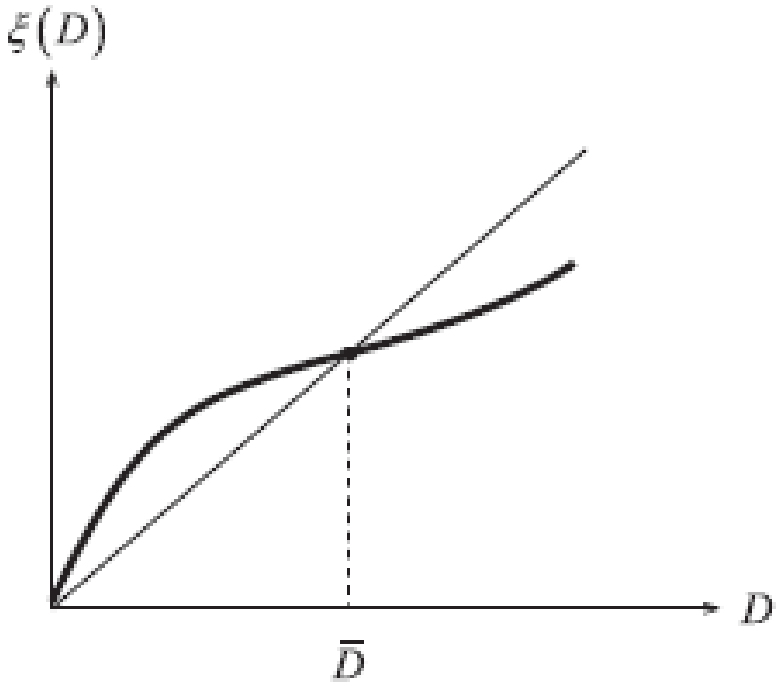
Zmienne losowe $u = \sum w_j x_j + b_j$ i $u' = \sum w_j y_j + b_j$ mają rozkład łączny normalny o zerowej średniej i macierzy kowariancji:

$$E\left[u^2\right] = E\left[u'^2\right] = \sigma_{A^t}^2,$$
$$E\left[uu'\right] = \sigma^2 A^t C\left(\boldsymbol{x}, \boldsymbol{y}\right) + \sigma_b^2 = \sigma_{A^t C}^2.$$

Wówczas sygnały nachodzą się zgodnie z: $C = \frac{1}{A^t} E\left[\varphi(u)\varphi\left(u'\right)\right]$
co daje nam ogólną postać:

$$C_t = \frac{1}{2\pi A^t} \cos^{-1}\left(-\frac{C_{t-1} A^t \sigma_t^2 + \sigma_b^2}{\sigma_{A^t}^2 \sqrt{1 + \sigma_{A^t}^2}}\right)$$

A z tego wyprowadzamy **ξ**. W szczególności **ξ** jest monotoniczną funkcją **D** i mamy **ξ**(0) = 0 oraz, przy $\sigma_t^2 = \sigma^2$, zachodzi $\xi'(0) = \bar{\chi}_1$.
Dla $\bar{\chi} > 1$ mamy jeszcze jedno, unikalne rozwiązanie $\bar{D} = \xi\left(\bar{D}\right)$, które jest jedynym stabilnym przy **t** zbiegającym do nieskończoności. W tym miejscu teoria bazuje na nieprzeliczalnej kardynalności przestrzeni **X**.



Wnioski

1. Przy $\bar{\chi}$ bliskiej **1** dynamika $\tilde{x}^t$ jest bliska chaotycznej.

2. Przy $\bar{\chi}$ równej **1** krzywizna jest rozbieżna do nieskończoności (ale pod warunkiem że **n** skończone!)

3. Otrzymane wyniki opisują zachowanie przy $t \longrightarrow \infty$ oraz $n \longrightarrow \infty$,
ale w praktyce wartości **t** oraz **n** pozostają skończone.

4. Autorzy podkreślają potrzebę badań wpływu „skończoności” na dotychczas uzyskane wyniki.

Dziękuję za uwagę

Bibliografia:

Shun-ichi Amari, Ryo Karakida, and Masafumi Oizumi.
Statistical Neurodynamics of Deep Networks: Geometryof Signal Spaces.
Technical report, 2018a.