Normalizing Flows

Zarys Problemu:

- Mamy pewien rozkład $p_x(x)$, który jest "trudny" np 
- Mamy próbki z tego rozkładu x, ale niekoniecznie potrafimy obliczać $p_x(x)$
- możemy postawić dwa problemy: → jak obliczać $p_x(x)$?
  → jak generować próbki z tego rozkładu?

Idea normalizing Flows:

- Zaczynamy od prostego rozkładu i "przetransformujmy" go w nasz żądany w sposób odwracalny

$z \underset{F^{-1}}{\overset{F}{\rightleftarrows}} x$, wtedy $p_x(x) = p(z)\left|\det\left(\frac{\partial F}{\partial x}\right)\right|$ [F - musi być wystarczająco "dobre"]

z zasady zmiennych $= p_z(z)\left|\det\left(\frac{\partial F}{\partial x}\right)\right|^{-1}$ F^{-1}, F - muszą być różniczkowalne

$\Rightarrow$ Zmiana problemu na szukanie odpowiedniej f-cji: F

Zauważmy, że F może być złożeniem $F = f_k \circ f_{k-1} \circ \dots \circ f_3 \circ f_2 \circ f_1$ i jeżeli każda z f_i:

będzie spełniać nasze założenia to F też, ponieważ mamy

- $(f_2 \circ f_1)^{-1} = f_1^{-1} \circ f_2^{-1}$ oraz $\det(J_{f_2 \circ f_1}(u)) = \det(J_{f_2}(f_1(u))) \cdot \det(J_{f_1}(u))$

Intuicja ze przekształceniem i jakobianem:

F - "rozciąganie" przestrzeni $\mathbb{R}^d$, tak żeby zamienić $p_x(x)$ w $p_z(z)$, $|\det(J_F(\ ))|$ odpowiada za zmianę objętości w "małym" otoczeniu ze względu na F. Bierzemy mechanicznie małe otoczenie du i dx, ponieważ masa prawdopodobieństwa się nie zmienia to zmiana objętości wpływa na gęstość: $V\uparrow \Rightarrow \rho\downarrow$, $V\downarrow \Rightarrow \rho\uparrow$

Czego chcemy od f_i:

- wystarczająco "elastyczne" żeby przekształcić prosty rozkład w złożony
- f_i oraz f_i^{-1} muszą istnieć i być w miarę łatwe do obliczenia
- wyznacznik jakobianu f_i^{-1} (f_i) musi być w miarę prosty do obliczenia

Żeby znaleźć funkcje spełniające ostatnią własność czysto konstrukcyjnie, z własności wyznacznika np.

- wyznacznik macierzy trójkątnej jest równy iloczynowi przekątnej

Uczenie: maksymalizujemy log-wiarygodność $\log(p_x(x)) = \log(p_u(u)) + \sum_{i=1}^{k} \log\left(\left|\det \frac{f_i^{-1}(z)}{\partial z}\right|\right)$

Lemat o wyznaczniku macierzy: Jeżeli A jest $d \times d$ i odwracalne, a V, W są macierzami $d \times m$ to $\det(A + VW^T) = \det(I + W^T A^{-1} V) \det A$

jeżeli mamy łatwo policzyć A^{-1} i $\det(A)$, oraz $m < d$ to wyznaczenie prawej strony jest obliczeniowo lepsze niż lewej $O(D^3 + D^2 M)$ [lewa strona], a dla np diagonalnej A prawa strona to $O(M^3 + DM^2)$

Planar flow

$f(z) = z + v\sigma(w^T z + b)$ [sieć neuronowa z jednym neuronem ukrytym]

$v, w \in \mathbb{R}^d$, $b \in \mathbb{R}$ są parametrami flowu (do nauczenia) a σ jest różniczkowalną funkcją aktywacji np. tangens hiperboliczny

Możemy interpretować ten flow jako "rozciągnięcie / ściśnięcie" przestrzeni w kierunku prostopadłym do hiperpłaszczyzny $w^T z + b = 0$

Jakobian dany jest wzorem $\nabla f(z) = I + \sigma'(w^T z + b) v w^T$

σ' - pochodna f-ji aktywacji

Korzystając z lematu o wyznaczniku: $\det \nabla f(z) = 1 + \sigma'(w^T z + b) w^T v$

Żeby ten flow był odwracalny musimy przyjąć dodatkowe założenie:

- σ' jest dodatnia i ograniczona
- Warunek dostateczny dla odwracalności $w^T v > -\frac{1}{\sup_t \sigma'(t)}$

Sylvester Flow (uogólnienie Planar flow na M warstwowych ukrytych)

$f(z) = z + V\sigma(W^T z + b)$, $V, W \in \mathbb{R}^{d \times m}$, $b \in \mathbb{R}^m$ to parametry do nauczenia

σ - f-ja aktywacji, nałożona po współrzędnych

Jakobian przekształcenia $J_f(z) = I + VS(z)W^T$, gdzie $S(z)$ jest macierzą diagonalną $m \times m$ o diagonali $\sigma'(W^T z + b)$, znów z lematu:

$$\det J_f(z) = \det\left(I + S(z) W^T V\right),$$ można zadać $V = QU$, $W = QL$

gdzie $Q \in \mathbb{R}^{d \times m}$ ma kolumny ortonormalne, $U \in \mathbb{R}^{m \times m}$ jest górnotrójkątna, a $L \in \mathbb{R}^{m \times m}$ dolnotrójkątna, ponieważ mamy $Q^T Q = I$ oraz $L^T U$ jest dolnotrójkątne to:

$$\det J_f(z) = \det\left(I + S(z) L^T \underbrace{Q^T Q}_{I} U\right) = \prod_{i=1}^{d} \left(1 + S_{ii}(z) L_{ii} U_{ii}\right)$$

dodatkowe założenia gwarantujące odwracalność:

- σ' jest dodatnia i ograniczona
- $L_{ii} U_{ii} > -\frac{1}{\sup_t \sigma'(t)}$, $i \in \{1, \ldots, d\}$

Flow sprzęgający

Dzielimy z na rozłączne części (z^A, z^B) i niech h będzie bijekcją $\mathbb{R}^d \to \mathbb{R}^d$

definiujemy $f(z)$ na kawałkach: $f(z)^B = h(z^B, \Theta(z^A))$

$$f(z)^A = z^A$$

$\Theta(z^B)$ - dowolna funkcja

f jest odwracalny gdyż h jest odwracalne i $f^{-1}(z)^B = h^{-1}(f(z)^B, \Theta(z^A))$

$$f^{-1}(z)^A = f(z)^A = z^A$$

Jakobian f: $\begin{array}{|c|c|}\hline I & 0 \\ \hline A & Dh \\ \hline \end{array} \Rightarrow \det(J_f) = \det(Dh)$

Najczęściej h działa po współrzędnych i wtedy Dh jest dolnotrójkątne

Zauważmy że $\Theta(x^B)$ może być dowolnie skomplikowana

Przykład: Additywna warstwa sprzęgająca

$x^A = z^A \qquad x^B = z^B + m(z^A)$ gdzie m - np. sieć neuronowa

ale zauważmy że $\det J_f = 1$ [czyli zachowuje objętość → mocne ograniczenie]

Poprawka: Przed warstwami sprzęgającymi dajemy warstwę skalującą

$X = Sz \qquad S$-diagonalna

Albo: Real NVP

$x^A = z^A \qquad x^B = s^\Theta(x^A) \odot x^B + t^\Theta(x^A)$ [łatwo odwracalne]

$s^\Theta(x^A)$: skalowanie, $t^\Theta(x^A)$: translacja

Uwagi i ogólna architektura:

- jeżeli mamy warstwy które zmieniają tylko część wejścia to pomiędzy musimy dać warstwy z permutacjami
- używamy zazwyczaj złożenie kilku warstw
- możemy badać prawdopodobieństwo próbki x: $p_x(x) = p_z(F^{-1}(x)) \det(J_{F^{-1}}(x))$
- możemy generować nowe próbki: generujemy z z $p_z(z)$ i przekształcamy $x = F(z)$