Normalizing Flows (NFs)

- direct generalization of inverse transform sampling from 1-D to high dimensions:

goal: $x \sim p^*(x)$, choose code distribution $q(z)$ $(= \mathcal{N}(0, \mathbb{I}))$

learn $g(z)$ such that $\hat{p}(x) = g_\# q(z) \approx p^*(x)$

$\Leftrightarrow z \sim q(z)$, $x = g(z) \sim \hat{p}(x)$

[recall 1-D: $q(z) = \text{uniform}(0,1)$ $f(x) = g^{-1}(x) = \text{CDF}(p^*(x))$]

- work-horse method of my team

generalize the change-of-variables formula to arbitrary dimension:

- 1-D: $p(x) = q(f(x)) \cdot |f'(x)|$
- D-dimensional: $p(X)$ with $\mathrm{dom}(X)$ domain ($p(x) > 0$ if $x \in \mathrm{dom}(X)$)

$A \subseteq \mathrm{dom}(X)$ some subset, $Pr(x \in A) = \int_A p(x)\,dx$

$z = f(x)$ and $p(z) = f_\# p(x)$, $\tilde{A} = \{ z = f(x) : x \in A \} = f(A)$ $\overset{!}{=}$

transformation must preserve probability mass: $Pr(z \in \tilde{A}) = \int_{\tilde{A}} p(z)\,dz$

$$\int_{\tilde{A} = f(A)} p(z)\,dz = \int_{f^{-1}(\tilde{A}) = A} p(z = f(x))\, |\det f'(x)|\, dx$$

$$f'(x) = \begin{bmatrix} \frac{\partial z_1}{\partial x_1} & \frac{\partial z_2}{\partial x_1} & \cdots & \frac{\partial z_D}{\partial x_1} \\ \vdots & \vdots & & \vdots \\ \frac{\partial z_1}{\partial x_D} & \frac{\partial z_2}{\partial x_D} & \cdots & \frac{\partial z_D}{\partial x_D} \end{bmatrix}$$ Jacobian matrix

to hold for all A: $\boxed{p(x) = p(z = f(x)) \cdot |\det f'(x)|}$

general law about Jacobians of invertible functions: $f(x)$, $g(z) = f^{-1}(z)$

$$f'(x) = \left(g'(z=f(x))\right)^{-1}$$, linear algebra: $\det B^{-1} = \frac{1}{\det B}$

$$|\det f'(x)| = \frac{1}{|\det g'(z=f(x))|}$$

two equivalent forms of change-of-variables formula

$$\boxed{\begin{aligned} p(X=x) &= q(Z=f(x)) \cdot |\det f'(x)| \\ p(X=g(z)) &= q(Z=z) \cdot |\det g'(z)|^{-1} \end{aligned}}$$

$\Rightarrow$ NF learning problem: (1) How to define $f(x)$ such that $g(z) = f^{-1}(z)$ exists and is easy to calculate?

(2) How can we efficiently calculate $|\det f'(x)|$ and/or $|\det g'(z)|$?

(3) How to learn $f(x)$ and/or $g(z)$?

- calculating a determinant: $D=1$: $\det f'(x) = f'(x)$ because $f'(x)$ is scalar

$$D=2 \quad \det\begin{pmatrix} a & b \\ c & d \end{pmatrix} = a\cdot d - b\cdot c$$

$$D=3 \quad \det\begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} = \begin{matrix} a\cdot(e\cdot i - f\cdot h) \\ + b\cdot(f\cdot g - d\cdot i) \\ + c\,(d\cdot h - e\cdot g) \end{matrix}$$

number of terms grows fast with $D \Rightarrow$ impractical

- solution via SVD: $f'(x) = U \cdot \Lambda \cdot V^T$ — U, V^T: orthonormal; Λ: diagonal, ≥ 0

$$|\det f'(x)| = \underbrace{|\det U|}_{=1} \cdot \underbrace{|\det \Lambda|}_{=\prod_{j=1}^{D} \lambda_{jj}} \cdot \underbrace{|\det V^T|}_{=1} = \prod_{j=1}^{D} \lambda_{jj}$$

still too expensive for training (millions of evaluations)

- solution in NF based on triangular maps: $\det\left[B = \begin{pmatrix} b_{11} & 0 & \cdots \\ b_{21} & b_{22} & 0 \cdots \\ \vdots & & \\ b_{D1} & \cdots & b_{DD} \end{pmatrix}\right] = \prod_{j=1}^{D} b_{jj}$

to make $f'(x)$ triangular, $f(x)$ must be a triangular map ("Knothe-Rosenblatt-rearrangement")

$$f(x) = \begin{bmatrix} f_1(x) \\ f_2(x) \\ \vdots \\ f_D(x) \end{bmatrix} = \begin{bmatrix} f_1(x_1) \\ f_2(x_1, x_2) \\ \\ f_D(x_1, \ldots, x_D) \end{bmatrix} \Rightarrow \frac{\partial z_j}{\partial x_{j'}} = \frac{\partial f_j(x)}{\partial x_{j'}} = 0 \quad \text{if } j' > j \Rightarrow \text{upper triangle of } f'(x) \text{ is zero}$$

(abbreviation: $f_j(x_{\leq j})$)

$$\Rightarrow \det f'(x) = \prod_{j=1}^{D} \frac{\partial f_j(x)}{\partial x_j} \leftarrow \text{diagonal elements of } f'(x)$$

However, learning a triangular map is hard & impractical $\Rightarrow$ NF use a compromise:

$f(x) = f_L \circ f_{L-1} \circ \ldots \circ f_2 \circ f_1$ is a composition of L layers, and the f_ℓ are triangular maps

$$|\det f'(x)| = |\det f_L'| \circ |\det f_{L-1}'| \cdot \ldots |\det f_1'|$$

$\leftarrow$ all easy because triangular maps

choose f^l such that
- $\det f_l'$ is cheap
- f_l^{-1} is cheap $\rightarrow g(z) = \underbrace{g_1}_{f_1^{-1}} \circ \underbrace{g_2}_{f_2^{-1}} \circ \ldots \circ \underbrace{g_L}_{f_D^{-1}}(z)$
- $f(x)$ is expressive (can learn complicated $p^*(x)$)

$\Rightarrow$ coupling layer ("affine" coupling layer)

$z^{(l)} = f_l(z^{(l-1)})$ with $z^{(0)} = x$ intermediate activations

(1) split vector $z^{(l)}$ into two halves: $z^{(l)} = \begin{bmatrix} z^{(l)}_{\le \tilde{D}} \\ z^{(l)}_{> \tilde{D}} \end{bmatrix}$ $\quad \tilde{D} = \lfloor \frac{D}{2} \rfloor$ $\quad z^{(l)}_{\le \tilde{D}} = \begin{bmatrix} z^{(l)}_1 \\ \vdots \\ z^{(l)}_{\tilde{D}} \end{bmatrix}$

(2) $z^{(l)}_{\le \tilde{D}} = z^{(l-1)}_{\le \tilde{D}}$ (skip connection for first half) $\quad z^{(l)}_{> \tilde{D}} = \begin{bmatrix} z^{l}_{\tilde{D}+1} \\ \vdots \\ z^{(l)}_{D} \end{bmatrix}$

(3) $z^{(l)}_j = s^{(l)}_j(z^{(l-1)}_{\le \tilde{D}}) \cdot z^{(l-1)}_j + t^{(l)}_j(z^{(l-1)}_{\le \tilde{D}})$ $\quad j > \tilde{D}$ (second half)

$z^{(\ell)} = f_\ell(z^{(\ell-1)})$ defined as a affine coupling is a triangular map

if $j \leq \tilde{D}$: $\dfrac{\partial z_j^{(\ell)}}{\partial z_{j'}^{(\ell-1)}} = \begin{cases} 1 & \text{if } j = j' \\ 0 & \text{otherwise} \end{cases}$ (skip connections

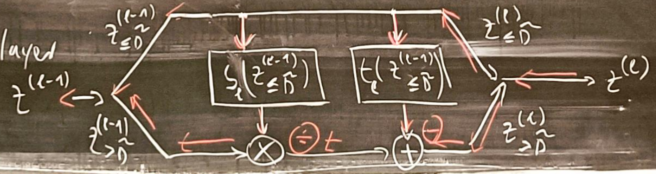
if $j > \tilde{D}$ $\dfrac{\partial z_j^{(\ell)}}{\partial z_{j'}^{(\ell-1)}} = \begin{cases} 0 & \text{if } j' > j \\ s(z^{(\ell-1)}_{\leq \tilde{D}}) & \text{if } j = j' \\ \neq 0 & \text{otherwise} \end{cases}$

$$f'(z^{(\ell-1)}) = \begin{array}{cc} \left[\begin{array}{ccc:ccc} 1 & & 0 & & & \\ & 1 & & & 0 & \\ 0 & & 1 & & & \\ \hdashline & & & s_{\tilde{D}+1}(z^{(\ell-1)}_{\leq \tilde{D}}) & & 0 \\ & \neq 0 & & & \ddots & \\ & & & 0 & & s_D(z^{(\ell-1)}_{\leq \tilde{D}}) \end{array}\right] & \tilde{D} \\ \quad\quad\quad\quad \tilde{D} & \end{array}$$

$$\left|\det f'(z^{(\ell-1)})\right| = \prod_{j=\tilde{D}+1}^{D} s_j(z^{(\ell-1)}_{\leq \tilde{D}})$$

forward: $z^{(\ell)}_{>\tilde{D}} = s_\ell(z^{(\ell-1)}_{\leq \tilde{D}}) \cdot z^{(\ell-1)}_{>\tilde{D}} + t_\ell(z^{(\ell-1)}_{\leq \tilde{D}})$

picture of layer



require $s_\ell(z^{(\ell-1)}_{\leq \tilde{D}}) \neq 0$

inverse: $z^{(\ell-1)}_{>\tilde{D}} = \dfrac{z^{(\ell)}_{>\tilde{D}} - t_\ell(z^{(\ell)}_{\leq \tilde{D}})}{s_\ell(z^{(\ell)}_{\leq \tilde{D}})}$