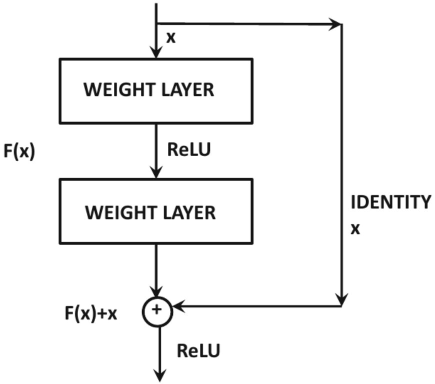
Tomáš Oberhuber

Faculty of Nuclear Sciences and Physical Engineering
Czech Technical University in Prague

Reziduální sítě



- viděli jsme, jak síť ResNet využívá tzv. *skip connections*
- zkratka obcházející jednu vrstvu má význam identity

Reziduální sítě

- aktivace jednotlivých vrstev pak lze popsat jako

$$\vec{x}^{(l+1)} = \vec{x}^{(l)} + F(\vec{x}^{(l)})$$

- to lze přepsat do tvaru

$$\vec{x}^{(l+1)} = \vec{x}^{(l)} + \Delta t \cdot F(\vec{x}^{(l)}),$$

  pro $\Delta t = 1$
- tento tvar odpovídá Eulerově metodě pro řešení ODR

$$\frac{\mathrm{d}\vec{x}}{\mathrm{d}t} = F(\vec{x}, t)$$

E. Weinan, A proposal on machine learning via dynamical systems, Communications in Mathematics and Statistics, vol. 5, no. 1, pp. 1–11, 2017

T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud, “Neural ordinary differential equations,” in Advances in Neural Information Processing Systems, pp. 6571–6583, 2018

Optimalizace s ODR

- místo učení NN se tak můžeme zabývat úlohou hledání vhodných parametrů pro ODR tak, aby např. generovala předepsaná data
- vlastně tak hledáme matematický model pro generování experimentálních dat
- jde o úlohu zvanou *physics from data* - Youtube
- optimalizace s ODR jsou také známé jako **optimální procesy**

L.S. Pontryagin, Mathematical Theory of Optimal Processes: The Mathematical Theory of Optimal Processes, Routledge, 1987.

Optimalizace s ODR

Mějme ODR tvaru

$$\begin{aligned} \dot{\vec{u}}(t, \vec{p}) &= \vec{f}(\vec{u}(t, \vec{p}), t, \vec{p}) \text{ na } (0, T), \qquad (1) \\ \vec{u}(0, \vec{p}) &= \vec{u}_0(\vec{p}), \qquad (2) \end{aligned}$$

kde

- $\vec{p}$ je vektor parametrů $\vec{p} \in P \equiv \mathbb{R}^m$
- $\vec{u}(t, \vec{p}) : \langle 0, T \rangle \times P \to \mathbb{R}^n$ je neznámá funkce
- $\vec{f}(\vec{u}(t, \vec{p}), t, \vec{p}) : \mathbb{R}^n \times \langle 0, T \rangle \times P \to \mathbb{R}^n$ je daná pravá strana

Mějme zobrazení

- $h_1(\vec{u}(t,\vec{p}),t,p) : \mathbb{R}^n \times \langle 0, T\rangle \times P \to \mathbb{R}$,
- $h_2(\vec{u}(T,\vec{p}),p) : \mathbb{R}^n \times P \to \mathbb{R}$.

Pak můžeme definovat funkcionál J jako

$$J(\vec{p}) = \int_0^T h_1(\vec{u}(t,\vec{p}),t,p)dt + h_2(\vec{u}(T,\vec{p}),\vec{p})$$

a řešit úlohu

$$\vec{p}^* = \arg\min_{p\in P} J(\vec{p}).$$

Example 1

- předpokládejme, že máme zadaná (experimentální) data $\vec{d}(t)$
- volbou

$$
\begin{aligned}
h_1(\vec{u}(t,\vec{p}),t,p) &= \left|\vec{u}(t,\vec{p}) - \vec{d}(t)\right|^2, \\
h_2(\vec{u}(T,\vec{p}),p) &= 0
\end{aligned}
$$

- dostáváme funkcionál

$$
J(\vec{p}) = \int_0^T \left|\vec{u}(t,\vec{p}) - \vec{d}(t)\right|^2 dt
$$

- a hledáme parametry $\vec{p}^*$ tak, že ODR (10) s pravou stranou $\vec{f}(\vec{u}(t,\vec{p}^*),t,\vec{p}^*)$ generuje naše data.

Example 2

- mějme zadaný finální stav $\vec{d}(T)$
- volbou

$$\begin{aligned} h_1(\vec{u}(t,\vec{p}),t,p) &= 0, \\ h_2(\vec{u}(T,\vec{p}),p) &= \left|\vec{u}(T,\vec{p}) - \vec{d}(T)\right|^2 \end{aligned}$$

- dostáváme funkcionál

$$J(\vec{p}) = \left|\vec{u}(T,\vec{p}) - \vec{d}(T)\right|^2$$

- a hledáme parametry $\vec{p}^*$ tak, že počáteční podmínka (10) vede k finálnímu stavu $\vec{d}(T)$, tj. můžeme řešit *inverzní úlohu* k úloze (10)-(10)

Optimalizace s ODR

- pro nalezení optimálních parametrů $\vec{p}^*$ opět potřebujeme spočítat

$$\nabla_{\vec{p}} J(\vec{p})$$

- to lze udělat třemi různými způsoby
  1. přímým výpočtem pomocí malých perturbací parametrů $\vec{p}$
     - pro větší počet parametrů to může být výpočetně velmi náročné
  2. algoritmicky pomocí tzv. *automatic differentiation* - Julia
     - u ODR je to v současnosti často nejjednodušší přístup
  3. odvozením tzv. **adjungovaná rovnice** (*adjoint equation*) - obdoba backpropagation

Adjungovaná rovnice

- pro odvození adjungované rovnice použijeme Lagrangeovy multiplikátory
- definujeme funkcionál $J^*(\vec{p})$ jako

$$J^*(\vec{p}) = \int_0^T h_1 + \vec{\lambda}_1^T \left(\dot{\vec{u}}(\vec{p}) - \vec{f}\right) dt + h_2 + \vec{\lambda}_2^T (\vec{u}\,|_{t=0} - \vec{u}_0)$$

  kde $\vec{\lambda}_{1,2} \in C(\langle 0, T\rangle; \mathbb{R}^n)$ jsou Lagrangeovy multiplikátory
- pokud $\vec{u}$ je řešení ODR (10)-(10) potom

$$\begin{aligned} \dot{\vec{u}} - \vec{f} &= 0 \text{ na } (0, T), \\ \vec{u}\,|_{t=0} - \vec{u}_0 &= 0, \end{aligned}$$

  a $J(\vec{p}) = J^*(\vec{p})$ platí pro všechna $\vec{\lambda}_{1,2}$

- gradient J^* má tvar

$$\nabla_{\vec{p}} J^*(\vec{p}) = \int_0^T \frac{\partial h_1}{\partial \vec{u}} \frac{\partial \vec{u}}{\partial \vec{p}} + \frac{\partial h_1}{\partial \vec{p}} + \vec{\lambda}_1^T \left(\frac{\partial \dot{\vec{u}}}{\partial \vec{p}} - \frac{\partial \vec{f}}{\partial \vec{u}} \frac{\partial \vec{u}}{\partial \vec{p}} - \frac{\partial \vec{f}}{\partial \vec{p}} \right) dt + \frac{\partial h_2}{\partial \vec{u}} \frac{\partial \vec{u}}{\partial \vec{p}} \Big|_{t=T} + \frac{\partial h_2}{\partial \vec{p}} \Big|_{t=T} + \vec{\lambda}_2^T \left(\frac{\partial \vec{u}}{\partial \vec{p}} \Big|_{t=0} + \frac{\partial \vec{u}_0}{\partial \vec{p}} \right)$$

- výpočet $\partial \vec{u}/\partial \vec{p}$ a $\partial \dot{\vec{u}}/\partial \vec{p}$ by vyžadoval opětovné řešení původní ODR a postupnými perturbacemi $\vec{p}$
- těchto členů se chceme zbavit

Adjungovaná rovnice

- začneme s výrazem $\partial\dot{\vec{u}}/\partial\vec{p}$

$$
\begin{aligned}
\int_0^T \vec{\lambda}_1^T \frac{\partial\dot{\vec{u}}}{\partial\vec{p}} dt &= \int_0^T \vec{\lambda}_1^T \frac{\partial}{\partial\vec{p}} \frac{\partial\vec{u}}{\partial t} dt \\
&= \int_0^T \vec{\lambda}_1^T \frac{\partial}{\partial t} \frac{\partial\vec{u}}{\partial\vec{p}} dt \\
&= -\int_0^T \frac{\partial\vec{\lambda}_1^T}{\partial t} \frac{\partial\vec{u}}{\partial\vec{p}} dt + \left[\vec{\lambda}_1^T \frac{\partial\vec{u}}{\partial\vec{p}}\right]_0^T \\
&= -\int_0^T \frac{\partial\vec{\lambda}_1^T}{\partial t} \frac{\partial\vec{u}}{\partial\vec{p}} dt + \vec{\lambda}_1^T \frac{\partial\vec{u}}{\partial\vec{p}} \Big|_{t=T} - \vec{\lambda}_1^T \frac{\partial\vec{u}}{\partial\vec{p}} \Big|_{t=0},
\end{aligned}
$$

- a tak dostaneme

$$
\begin{aligned}
\nabla_{\vec{p}} J^*(\vec{p}) \quad = \quad & \int_0^T \left(\frac{\partial h_1}{\partial \vec{u}} - \dot{\vec{\lambda}}_1^T - \vec{\lambda}_1^T \frac{\partial \vec{f}}{\partial \vec{u}} \right) \frac{\partial \vec{u}}{\partial \vec{p}} + \frac{\partial h_1}{\partial \vec{p}} - \vec{\lambda}_1^T \frac{\partial \vec{f}}{\partial \vec{p}} dt + \\
& \left(\vec{\lambda}_1^T + \frac{\partial h_2}{\partial \vec{u}} \right) \frac{\partial \vec{u}}{\partial \vec{p}} \Big|_{t=T} + \frac{\partial h_2}{\partial \vec{p}} \Big|_{t=T} + \left(\vec{\lambda}_2^T - \vec{\lambda}_1^T \right) \frac{\partial \vec{u}}{\partial \vec{p}} \Big|_{t=0} + \vec{\lambda}_2^T \frac{\partial \vec{u}_0}{\partial \vec{p}}
\end{aligned}
$$

Adjungovaná rovnice

- a splněním následujících vztahů

$$
\begin{aligned}
\frac{\partial h_1}{\partial \vec{u}} - \dot{\vec{\lambda}}_1^T - \vec{\lambda}_1^T \frac{\partial \vec{f}}{\partial \vec{u}} &= 0, && (3)\\
\left(\vec{\lambda}_1^T + \frac{\partial h_2}{\partial \vec{u}} \right)_{t=T} &= 0, && (4)\\
\left(\vec{\lambda}_2 - \vec{\lambda}_1 \right) \Big|_{t=0} &= 0. && (5)
\end{aligned}
$$

- můžeme nulovat výrazy před $\frac{\partial \vec{u}}{\partial \vec{p}}$

Adjungovaná rovnice

- (3)-(4) lze splnit řešením adjungované ODR

$$
\begin{aligned}
-\frac{\partial \vec{\lambda}_1}{\partial t} &= \vec{\lambda}_1 \frac{\partial \vec{f}}{\partial \vec{u}} - \frac{\partial h_1}{\partial \vec{u}} \text{ na } (0, T), && (6) \\
\vec{\lambda}_1 \,\Big|_{t=T} &= -\frac{\partial h_2}{\partial \vec{u}} \,\Big|_{t=T} && (7)
\end{aligned}
$$

- a následně nastavit

$$\vec{\lambda}_2|_{t=0} = \vec{\lambda}_1 \,\Big|_{t=0} .$$

- uvedená ODR má počáteční podmínku zadanou v bodu $t = T$

Adjungovaná rovnice

- provedeme transformaci

$$\vec{\lambda}_1^*(t) = \vec{\lambda}_1(T - t) \text{ for } t \in \langle 0, T \rangle$$

- a vidíme, že

$$\frac{\partial \vec{\lambda}_1^*}{\partial t} = -\frac{\partial \vec{\lambda}_1}{\partial t} \quad \text{a} \quad \vec{\lambda}_1^*(0) = \vec{\lambda}_1(T)$$

- adjungovanou rovnici (6)-(7) lze tedy transformovat na tvar

$$\frac{\partial \vec{\lambda}_1^*}{\partial t} = \vec{\lambda}_1^* \frac{\partial \vec{f}(\vec{u}(T-t,\vec{p}), T-t, \vec{p})}{\partial \vec{u}} - \frac{\partial h_1(T-t)}{\partial \vec{u}} \text{ na } (0, T), \tag{8}$$

$$\vec{\lambda}_1^* \Big|_{t=0} = -\frac{\partial h_2}{\partial \vec{u}} \tag{9}$$

Výpočet gradientu pak vypadá takto:

1: vyřeš primární rovnici

$$\begin{aligned}\dot{\vec{u}}(t,\vec{p}) &= \vec{f}(\vec{u}(t,\vec{p}),t,\vec{p}) \text{ na } (0,T),\\ \vec{u}(0,\vec{p}) &= \vec{u}_0(\vec{p}),\end{aligned}$$

2: za pomoci znalosti primárního řešení vyřeš adjungovanou rovnici

$$\begin{aligned}\frac{\partial\vec{\lambda}_1^*}{\partial t} &= \vec{\lambda}_1^*\frac{\partial\vec{f}(\vec{u}(T-t,\vec{p}),T-t,\vec{p})}{\partial\vec{u}} - \frac{\partial h_1(T-t)}{\partial\vec{u}} \text{ na } (0,T),\\ \vec{\lambda}_1^*\,\Big|_{t=0} &= -\frac{\partial h_2}{\partial\vec{u}}\end{aligned}$$

3: s pomocí znalosti $\vec{\lambda}_1^*$ dopočítej $\lambda_1(t)=\lambda_1^*(T-t)$ a $\vec{\lambda}_2|_{t=0}=\vec{\lambda}_1\,\Big|_{t=0}$

4: nakonec spočítej samotný gradient

$$\nabla_{\vec{p}}J^*(\vec{p}) = \int_0^T\left(\frac{\partial h_1}{\partial\vec{p}} - \vec{\lambda}_1\frac{\partial\vec{f}}{\partial\vec{p}}\right)dt + \frac{\partial h_2}{\partial\vec{p}}\,|_{t=T} + \vec{\lambda}_2\frac{\partial\vec{u}_0}{\partial\vec{p}}$$

Adjungovaná rovnice

Pro jednoduchost nyní předpokládejme, že je

$$\frac{\partial h_1}{\partial \vec{u}} = 0.$$

Pak je má adjungovaná rovnice tvar

$$\frac{\partial \vec{\lambda}_1^*}{\partial t} = \vec{\lambda}_1^* \frac{\partial \vec{f}}{\partial \vec{u}} \text{ na } (0, T),$$

což vede na exponenciální řešení.

Adjungovaná rovnice

- To je velká nevýhoda tohoto přístupu, rovnice na intervalu $(0, T)$ často diverguje/exploduje (blow-up).
- Je to proto, že pokud špatně odhadneme parametry ODR, pak na celém intervalu $(0, T)$ vzniká veliká odchylka primární rovnice od zadaných dat a adjungovaná rovnice na to pak reaguje velkou citlivostí.
- Tato metoda je tak velice citlivá na dostatečně přesný počáteční odhad parametrů ODR.

Sparse Identification of Non-linear Dynamycs

Uvažujme ODR tvaru:

$$\begin{aligned}\dot{\vec{u}}(t) &= \vec{f}(\vec{u}(t), t, \mathbb{P}) \text{ na } (0, T),\\ \vec{u}(0) &= \vec{u}_0,\end{aligned}$$

Funkci f budeme hledat ve tvaru

$$f_j(\vec{u}, t, \mathbb{P}) = \sum_{i=1}^{q_j} p_i^{(j)} \phi_i(\vec{u}, t) = (\vec{p}^{(j)})^T \vec{\phi}(\vec{u}, t).$$

SINDy

Zavedeme si matici naměřených dat v různých časech $t_1, \ldots, t_m$

$$\mathbb{X} = \begin{pmatrix} \vec{x}(t_1)^T \\ \vdots \\ \vec{x}(t_m)^T \end{pmatrix} \in \mathbb{R}^{m,n},$$

a podobně matici derivací

$$\dot{\mathbb{X}} = \begin{pmatrix} \dot{\vec{x}}(t_1)^T \\ \vdots \\ \dot{\vec{x}}(t_m)^T \end{pmatrix} \in \mathbb{R}^{m,n}.$$

Dále si zavedeme bázi funkcí, ze kterých budeme $\vec{f}$ vyjadřovat:

$$\Phi(\mathbb{X}) = \left(\vec{1}\ \mathbb{X}\ \mathbb{X}^2 \dots \mathbb{X}^d \dots \sin(\mathbb{X}) \dots\right) \in \mathbb{R}^{m,q},$$

kde např.

- $$\sin(\mathbb{X}) = \begin{pmatrix} \sin(\vec{x}(t_1)^T) \\ \vdots \\ \sin(\vec{x}(t_m)^T) \end{pmatrix} = \begin{pmatrix} \sin(x_1(t_1)) & \dots & \sin(x_n(t_1)) \\ \vdots & \ddots & \vdots \\ \sin(x_1(t_m)) & \dots & \sin(x_n(t_m)) \end{pmatrix}$$
- $\mathbb{X}^d$ je matice jejíž sloupcové indexy odpovídají všem možným polynomům složek vektoru $\vec{x}(t_i)$ pro $i = 1, \dots, m$.

Zavedeme ještě matici

$$\mathbb{F} = \begin{pmatrix} \vec{f}(\vec{u}, t_1, \vec{p})^T \\ \vdots \\ \vec{f}(\vec{u}, t_m, \vec{p})^T \end{pmatrix} \in \mathbb{R}^{m,n},$$

Pak má ale platit

$$\dot{\mathbb{X}} = \mathbb{F} = \Phi(\mathbb{X})\mathbb{P},$$

kde $\mathbb{P} = (\vec{p}_1, \ldots \vec{p}_n)$, kde $\vec{p}_i \in \mathbb{R}^q$.

Řešíme tedy úlohu

$$\mathbb{P} = \arg\min_{\mathbb{P}} \left\| \dot{\mathbb{X}} - \Phi(\mathbb{X})\mathbb{P} \right\| + \lambda \left\| \mathbb{P} \right\|_1 ,$$

což je lineární regrese.

SINDy - příklad

Jako příklad si uvedeme Lorenzův atraktor:

$$
\begin{aligned}
\dot{x} &= \sigma(y - x), \\
\dot{y} &= x(\rho - z) - y, \\
\dot{z} &= xy - \beta z.
\end{aligned}
$$

Pak dostáváme

$$
\dot{\mathbb{X}} = \begin{pmatrix} \dot{x}(t_1) & \dot{y}(t_1) & \dot{z}(t_1) \\ \dot{x}(t_2) & \dot{y}(t_2) & \dot{z}(t_2) \\ \vdots & \vdots & \\ \dot{x}(t_m) & \dot{y}(t_m) & \dot{z}(t_m) \end{pmatrix}
$$

SINDy - příklad

Jako bázi zavedeme

$$\Phi(\mathbb{X}) = \begin{pmatrix} 1 & x\,|_{t_1} & y\,|_{t_1} & z\,|_{t_1} & x^2\,|_{t_1} & y^2\,|_{t_1} & z^2\,|_{t_1} & xy\,|_{t_1} & xz\,|_{t_1} & yz\,|_{t_1} \\ 1 & x\,|_{t_2} & y\,|_{t_2} & z\,|_{t_2} & x^2\,|_{t_2} & y^2\,|_{t_2} & z^2\,|_{t_2} & xy\,|_{t_2} & xz\,|_{t_2} & yz\,|_{t_2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x\,|_{t_m} & y\,|_{t_m} & z\,|_{t_m} & x^2\,|_{t_m} & y^2\,|_{t_m} & z^2\,|_{t_m} & xy\,|_{t_m} & xz\,|_{t_m} & yz\,|_{t_m} \end{pmatrix} \in \mathbb{R}^{m,q},$$

A následně

$$\mathbb{P} = (\vec{p}_x, \vec{p}_y, \vec{p}_z) \in \mathbb{R}^{q,n}.$$

Zdroje

1. DiffEqFlux.jl – A Julia Library for Neural Differential Equations
2. Ch. Rackauckas, M. Innes, Y. Ma, J. Bettencourt, L. White, V. Dixit, *DiffEqFlux.jl - A Julia Library for Neural Differential Equations*, arXiv:1902.02376.
3. V.M Alexejev, V.M.Tichomirov, S.V.Fomin, *Matematická teorie optimálních procesů*, Academia, 1991.