Adjoint State Method

Tuesday, October 15, 2024 5:01 PM

we're going to solve inverse problems, just one example of where the adjoint state (or adjoint sensitivity method) is useful

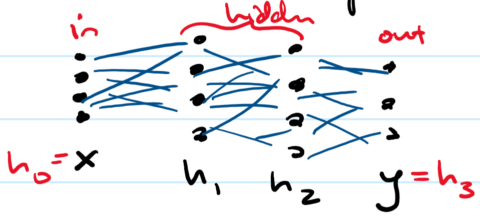
Dates back to at least Pontryagin et al. '62

widely known in some communities... and always confusing!

We'll follow "Neural Ordinary Diff. Eq" (Chen, Rubanova, Bettencourt, Duvenaud) NeurIPS '18 (Toronto)

Their setup:

Consider a 2 hidden layer neural net w/ skip/residual connections



$h_1 = \sigma(W_1 \cdot h_0) + h_0$

$h_2 = \sigma(W_2 h_1) + h_1$

$h_3 = \sigma(W_3 h_2) + h_2$

If all weight matrices are the same, $W_1 = W_2 = W_3 = \theta$,

think of this as Forward Euler to solve

$$\underbrace{\frac{d\,h(t)}{dt}}_{} = f_\theta(h(t), t) \quad \text{with } \Delta t = 1,\ T = 3$$

similarity

Forward Euler gives $h(t+\Delta t) = h(t) + \Delta t \cdot f(h(t), t)$

[observation goes back to at least Yiping Lu et al. '17]

So, let's define our neural net output to be the ODE output at some time T. Use any ODE solver!

new hyperparameter, now continuous

Q: How to train neural net weights θ?

Backpropagate? If ODE took a lot of steps, we need to store a lot of intermediate variables.

Sol'n: adjoint state method "regenerates" intermediate variables "on-the-fly" by solving a related ODE backwards in time!

Adjoint State Method (p. 2)

Tuesday, October 15, 2024 5:20 PM

Adjoint State Method

Suppose we have a loss function $\mathcal{L}(z(T))$,

weights

z solves ODE $z(0) = z_0$, $z'(t) = f(z, t; \theta)$

Abstractly, $z(T) = z_0 + \int_0^T f(z(t), t; \theta)\, dt$

What we want is

$$\frac{d\mathcal{L}}{d\theta} = \underbrace{\frac{d\mathcal{L}}{dz}}\ \underbrace{\frac{\partial z}{\partial \theta}}$$ (or a little more complicated...)

← how do we get this? i.e. $\frac{\partial z(t)}{\partial \theta}$ $\forall t$?

Running Example

want $\mathcal{L}(z(T))$

$\mathcal{L}(x) = x^2$, $z' = \theta \cdot z$, $z(0) = 3$ so $z(t) = 3 \cdot e^{\theta t}$

This is simple enough that we can check our answer

Compute $\frac{d}{d\theta}\mathcal{L}(z(T)) = \frac{d}{d\theta}(3e^{\theta T})^2 = 2 \cdot 3e^{\theta T} \cdot \frac{d}{d\theta}(3e^{\theta T})$

i.e. $g(\theta) = \mathcal{L}(z(T))$

Find $\frac{dg}{d\theta}$

$= 18 \cdot T \cdot e^{2\theta T}$

Derivation: first goal, get $\frac{d\mathcal{L}}{dz(t)} =: a(t)$ ← the "adjoint state"

Thm $a(t)$ satisfies its own ODE, $a' = -a \cdot \frac{\partial f(z, t; \theta)}{\partial z}$

We start at $t = T$ with $a(T) = \frac{\partial \mathcal{L}}{\partial z(T)}$ ("initial" condition)

proof

If ODE is "nice" (i.e. f is smooth), which is common, then

$$z(t+h) = z(t) + h \underbrace{z'(t)}_{= f(z, t; \theta)} + O(h^2)$$ Taylor Series

$$a(t) := \frac{d\mathcal{L}}{dz(t)} = \frac{d\mathcal{L}}{dz(t+h)} \cdot \frac{dz(t+h)}{dz(t)}$$ Chain Rule

$$= a(t+h) \cdot \frac{d}{dz(t)}\left(z(t) + h \cdot f(z(t), t; \theta) + O(h^2)\right)$$

$$= a(t+h) \cdot \left(I^* + h \cdot \frac{\partial f}{\partial z} + O(h^2)\right)$$

* we've been a bit loose but z could be a vector, so this is I not 1.

Think of as row vectors for now.
(since Jacobians ($= \nabla^T$) have nice direction of chain rule)

Adjoint State Method (p. 3)

Wednesday, October 16, 2024 9:46 AM

... recall...

$$a(t) = a(t+h)\cdot\left(I + h\cdot\frac{\partial f}{\partial z} + o(h^2)\right)$$

Then

$$a' := \frac{da(t)}{dt} := \lim_{h\to 0} \frac{a(t+h) - a(t)}{h}$$

$$= \lim_{h\to 0} a(t+h)\cdot \frac{\left(I - \left(I + h\frac{\partial f}{\partial z} + o(h^2)\right)\right)}{h}$$

$$= \lim_{h\to 0} -a(t+h)\cdot\frac{\partial f}{\partial z} + o(h)$$

Assuming $a(t)$ is continuous...

$$= -a(t)\cdot\frac{\partial f}{\partial z}(z(t), t; \theta) \qquad \square$$

Running Example

$\mathcal{L}(x) = x^2$, $z' = \underbrace{\theta\cdot z}_{= f(z,t,\theta)}$, $z(0) = 3$ so $z(t) = 3\cdot e^{\theta t}$

Then so $\frac{\partial f}{\partial z} = \theta$

$a(t) := \frac{d\mathcal{L}}{dz(t)}$

satisfies $a' = -a\cdot\frac{\partial f}{\partial z}$ i.e. $a' = -\theta\cdot a$

Full loss is

$\mathcal{L}(z(T))$, $\frac{d\mathcal{L}}{dz(T)} = \underbrace{2\cdot z(T)}_{= 6\cdot e^{\theta T}}$ (explicit)

(we know since we did forward solve)

So $a(T) = \; = 6\cdot e^{\theta T}$

Hence solve $a' = -\theta\cdot a$ on $t\in[0,T]$ and "I.C." $a(T) = 6e^{\theta T}$

... i.e., solve backwards! (if your ODE solver is unhappy, do a change of variables $t \mapsto \tilde{t} = T - t$ and adjust...)

So... in general, solve $a' = -a\cdot\frac{\partial f}{\partial z}$

on $[0,T]$ "starting" at $a(T) = \frac{d\mathcal{L}}{dz(T)}$ (known)

But... we want $\frac{d\mathcal{L}}{d\theta}$. What to do?

Adjoint State Method (p. 4)

Wednesday, October 16, 2024 10:04 AM

So, we defined $\frac{d\mathcal{L}}{dz(t)} = a(t)$, showed $\boxed{\begin{aligned} a' &= -a \cdot \frac{\partial f}{\partial z} \\ a(T) &= \frac{\partial \mathcal{L}}{\partial z(T)} \end{aligned}}$

i.e. $\frac{d\mathcal{L}}{dz} := a_z$, $a_z' = -a_z \frac{\partial f}{\partial z}$

where $z' = f(z, t, \theta)$

Let's augment: $\frac{d}{dt}\begin{bmatrix} z \\ \theta \end{bmatrix} = \underbrace{\begin{bmatrix} f(z,t,\theta) \\ 0 \end{bmatrix}}_{F}$) temporarily in column vectors

(i.e. $\theta' = 0$ since it doesn't depend on time)

$\frac{d\mathcal{L}}{d[z,\theta]} := A(t) = [a_z, a_\theta]$

so... $\boxed{A' = -A \cdot \frac{\partial F}{\partial [z,\theta]}}$ i.e.

$$\left[\frac{d}{dt}a_z, \frac{d}{dt}a_\theta\right] = -[a_z, a_\theta] \cdot \begin{bmatrix} \frac{\partial f}{\partial z} & \frac{\partial f}{\partial \theta} \\ \cancel{\frac{\partial 0}{\partial z}} & \cancel{\frac{\partial 0}{\partial \theta}} \end{bmatrix}$$

So solve augmented system in your ODE solver

$$= \left[\underbrace{-a_z \cdot \frac{\partial f}{\partial z}}_{\text{original adjoint ODE}}, -a_z \frac{\partial f}{\partial \theta}\right]$$

i.e. $\frac{d}{dt}a_\theta = \underbrace{-a_z \frac{\partial f}{\partial \theta}}$ doesn't depend on a_θ so can integrate

$$a_\theta(0) = \underbrace{a_\theta(T)}_{\text{set to 0 as "I.C."}} - \int_0^T a_\theta' \, dt$$

set to 0 as "I.C." since $\mathcal{L}$ doesn't depend <u>explicitly</u> on θ

$$= \int_T^0 a_\theta' \, dt = \int_T^0 -a_z \cdot \frac{\partial f}{\partial \theta} \, dt$$

calculate in adjoint ODE ("backward pass")

and $a_\theta(0) = \frac{d\mathcal{L}}{d\theta(0)}$

$= \frac{d\mathcal{L}}{d\theta}$ since θ is constant

which is what we wanted!

Adjoint State Method (p. 5)

Wednesday, October 16, 2024 10:29 AM

Running Example

$\mathcal{L}(x) = x^2, \quad z' = \underbrace{\theta \cdot z}_{= f(z,t,\theta)}, \quad z(0) = 3 \quad \text{so} \quad z(t) = 3 \cdot e^{\theta t}$

Then $\quad$ so $\dfrac{\partial f}{\partial z} = \theta$

$a(t) := \dfrac{d\mathcal{L}}{dz(t)}$

satisfies $\quad a' = -a \cdot \dfrac{\partial f}{\partial z} \quad$ i.e. $\quad a' = -\theta \cdot a$

Full loss is $\mathcal{L}(z(T)), \quad \dfrac{d\mathcal{L}}{dz(T)} = \underbrace{2 \cdot z(T)}_{= 6 \cdot e^{\theta T}}$ (explicit)

(we know since we did forward solve)

so $\quad a(T) = 6 \cdot e^{\theta T}$

Hence solve $a' = -\theta \cdot a$ on $t \in [0, T]$ and "I.C." $a(T) = 6e^{\theta T}$

... i.e., solve backwards!

solve explicitly

So...

$a = c\, e^{-\theta t}$

$a(T) = c \cdot e^{-\theta \cdot T}$

$= 6 e^{\theta T}$

so $c = 6 \cdot e^{2\theta T}$

so $a(t) = 6\, e^{2\theta T - \theta t}$

Find $\dfrac{\partial f}{\partial \theta} = z = 3e^{\theta t}$

Add in augmented dynamics...

$$a_\theta(0) = -\int_T^0 a \cdot \frac{\partial f}{\partial \theta}\, dt$$

$$= \int_0^T \underbrace{6e^{2\theta T - \theta t}}_{a} \cdot \underbrace{3e^{\theta t}}_{\frac{\partial f}{\partial \theta} (= z)}\, dt$$

$$= 18 \cdot e^{2\theta T} \cdot \int_0^T dt = \underline{18 \cdot T \cdot e^{2\theta T}}$$

matches!

Adjoint State Method (p. 6)

Wednesday, October 16, 2024 10:42 AM

In practice, 1) $z(t)$ often a vector

2) loss might not just be $\mathcal{L} = (z(T) - y_T)^2$

it might be $\mathcal{L} = \sum_{i=1}^{n} (z(t_i) - y_i)^2$

then solve (backward) adjoint equation from t_n to t_{n-1} using (explicit) $\frac{\partial \mathcal{L}}{\partial z(t_n)}$ as "initial condition",

then solve from t_{n-1} to t_n using $\frac{\partial \mathcal{L}}{\partial z(t_{n-1})}$ as "initial condition",

etc.