Differentiation through iterative solvers

Nicolas Nytko

October 31, 2022

1 Background

Consider some function $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ that describes the iterative process

$$\boldsymbol{x}^{(k+1)} = f\left(\boldsymbol{x}^{(k)}, \boldsymbol{\theta}\right), \tag{1}$$

where $\boldsymbol{x}^{(k)}$ is the state at iteration k and $\boldsymbol{\theta}$ is some set of parameters used at each step of the iteration that we would like to differentiate this process with respect to. We will assume that as $k \to \infty$, $\boldsymbol{x}^{(k)}$ converges to a *fixed point* such that

$$\boldsymbol{x}^\star = f\left(\boldsymbol{x}^\star, \boldsymbol{\theta}\right), \tag{2}$$

at least to within some small numerical tolerance. There are two ways to look at this problem depending on what information we are trying to obtain:

1. We care about the end result only. We assume that f is some reasonable method and we converge to $\boldsymbol{x}^\star$. Lets call this the *convergent* analysis.

2. We want to look at intermediate results. There is no assumption that f actually converges to $\boldsymbol{x}^\star$ — we might be optimizing the method itself here. Lets call this one the *nonconvergent* analysis.

2 Convergent Methods

We assume that we run the method f for a sufficient number of iterations such that we have a reasonable approximation to $\boldsymbol{x}^\star$, which we will denote by $\hat{\boldsymbol{x}}$. We therefore have

$$\hat{\boldsymbol{x}} = f\left(\hat{\boldsymbol{x}}, \boldsymbol{\theta}\right) + \boldsymbol{\varepsilon}\left(\boldsymbol{\theta}\right), \tag{3}$$

for some (hopefully) small $\boldsymbol{\varepsilon}$. Taking the derivative of both sides with respect to the parameter $\boldsymbol{\theta}$, we get

$$\begin{align}
\frac{\partial \hat{\boldsymbol{x}}}{\partial \boldsymbol{\theta}} &= \frac{\partial}{\partial \boldsymbol{\theta}} f\left(\hat{\boldsymbol{x}}, \boldsymbol{\theta}\right) \tag{4}\\
&= \frac{\partial f}{\partial \boldsymbol{\theta}}\left(\hat{\boldsymbol{x}}, \boldsymbol{\theta}\right) + \frac{\partial f}{\partial \boldsymbol{x}}\left(\hat{\boldsymbol{x}}, \boldsymbol{\theta}\right)\frac{\partial \hat{\boldsymbol{x}}}{\partial \boldsymbol{\theta}} + \frac{\partial \boldsymbol{\varepsilon}}{\partial \boldsymbol{\theta}}. \tag{5}
\end{align}$$

Rearranging terms gives

$$\begin{align}
\frac{\partial \hat{\boldsymbol{x}}}{\partial \boldsymbol{\theta}} &= \left(\boldsymbol{I} - \frac{\partial f}{\partial \boldsymbol{x}}\left(\hat{\boldsymbol{x}}, \boldsymbol{\theta}\right)\right)^{-1}\left(\frac{\partial f}{\partial \boldsymbol{\theta}}\left(\hat{\boldsymbol{x}}, \boldsymbol{\theta}\right) + \frac{\partial \boldsymbol{\varepsilon}}{\partial \boldsymbol{\theta}}\right) \tag{6}\\
&\approx \left(\boldsymbol{I} - \frac{\partial f}{\partial \boldsymbol{x}}\left(\hat{\boldsymbol{x}}, \boldsymbol{\theta}\right)\right)^{-1}\frac{\partial f}{\partial \boldsymbol{\theta}}\left(\hat{\boldsymbol{x}}, \boldsymbol{\theta}\right). \tag{7}
\end{align}$$

Of course, in an autogradient setting we are interested in computing the vector-Jacobian product instead of the Jacobian itself; left multiplying eq. (7) by some vector $\boldsymbol{v}$ (and dropping the $\approx$ to simplify notation) results in

$$\boldsymbol{v}^T\frac{\partial \hat{\boldsymbol{x}}}{\partial \boldsymbol{\theta}} = \boldsymbol{v}^T\left(\boldsymbol{I} - \frac{\partial f}{\partial \boldsymbol{x}}\left(\hat{\boldsymbol{x}}, \boldsymbol{\theta}\right)\right)^{-1}\frac{\partial f}{\partial \boldsymbol{\theta}}\left(\hat{\boldsymbol{x}}, \boldsymbol{\theta}\right). \tag{8}$$

Defining the intermediate vector $\boldsymbol{w}$ like

$$\boldsymbol{w}^T = \boldsymbol{v}^T \left(\boldsymbol{I} - \frac{\partial f}{\partial \boldsymbol{x}} \left(\hat{\boldsymbol{x}}, \boldsymbol{\theta} \right) \right)^{-1}, \tag{9}$$

we can rearrange terms to get

$$\boldsymbol{w}^T \left(\boldsymbol{I} - \frac{\partial f}{\partial \boldsymbol{x}} \left(\hat{\boldsymbol{x}}, \boldsymbol{\theta} \right) \right) = \boldsymbol{v}^T \tag{10}$$

$$\boldsymbol{w}^T - \boldsymbol{w}^T \frac{\partial f}{\partial \boldsymbol{x}} \left(\hat{\boldsymbol{x}}, \boldsymbol{\theta} \right) = \boldsymbol{v}^T \tag{11}$$

$$\boldsymbol{w}^T = \boldsymbol{v}^T + \boldsymbol{w}^T \frac{\partial f}{\partial \boldsymbol{x}} \left(\hat{\boldsymbol{x}}, \boldsymbol{\theta} \right), \tag{12}$$

which is itself a fixed-point iteration that we can find with intermediate vector-Jacobian products of f. Once we find $\boldsymbol{w}^T$, eq. (7) can be computed by VJP of f wrt $\boldsymbol{\theta}$ and $\boldsymbol{w}$.

3 Nonconvergent Methods

In the case that we do not run f to convergence, we will cast the iteration as an ODE like

$$\frac{d\boldsymbol{x}}{dt} = f\left(\boldsymbol{x}, \boldsymbol{\theta}\right) - \boldsymbol{x} = g\left(\boldsymbol{x}, \boldsymbol{\theta}\right). \tag{13}$$

Observe that if we integrate eq. (13) with an Euler forward timestepper with $\Delta t = 1$ and let $\boldsymbol{x}^{(k)} = \boldsymbol{x}(k)$, we recover the iteration exactly:

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + (\Delta t)\, g\left(\boldsymbol{x}^{(k)}, \boldsymbol{\theta}\right) \tag{14}$$

$$= \boldsymbol{x}^{(k)} + f\left(\boldsymbol{x}^{(k)}, \boldsymbol{\theta}\right) - \boldsymbol{x}^{(k)} \tag{15}$$

$$= f\left(\boldsymbol{x}^{(k)}, \boldsymbol{\theta}\right). \tag{16}$$

Assume we have run the above for j iterations and have obtained $\boldsymbol{x}^{(j)}$ (but not necessarily the intermediate steps). To differentiate some scalar loss (ℓ) with respect to $\boldsymbol{x}^{(0)}$ and $\boldsymbol{\theta}$, we will introduce the adjoint equation

$$\boldsymbol{a}(t) = \frac{\partial \ell}{\partial \boldsymbol{x}(t)}, \tag{17}$$

with derivative

$$\frac{d\boldsymbol{a}(t)}{dt} = -\boldsymbol{a}(t)^T \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}} \left(\boldsymbol{x}(t), \boldsymbol{\theta}\right). \tag{18}$$

Which we can integrate to get

$$\frac{d\ell}{d\boldsymbol{x}^{(0)}} = -\int_j^0 \boldsymbol{a}\left(t\right)^T \frac{\partial g\left(\boldsymbol{x}(t), \boldsymbol{\theta}\right)}{\partial \boldsymbol{x}} \tag{19}$$

$$\frac{d\ell}{d\boldsymbol{\theta}} = -\int_j^0 \boldsymbol{a}\left(t\right)^T \frac{\partial g\left(\boldsymbol{x}(t), \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}}. \tag{20}$$

Because we have discarded the intermediate results of $x^{(i)}$, we can also integrate backwards from $x^{(j)}$ at the same time using backwards Euler (see appendix A.1) to re-create the sequence of values.

3.1 Adjoint computation of gradient

We can now iteratively solve for the gradients $\frac{\partial \ell}{\partial \boldsymbol{x}^{(0)}}$ and $\frac{\partial \ell}{\boldsymbol{\theta}}$. We begin with $\boldsymbol{x}^{(j)}$ and $\boldsymbol{a}^{(j)} = \boldsymbol{a}(j) = \frac{\partial \ell}{\partial \boldsymbol{x}^{(j)}}$. We first compute $x^{(j-1)}$ by implicit Euler, which we will perform the nonlinear solve using a gradient descent.

Define the residual like

$$\boldsymbol{r} := \boldsymbol{x}^{(k)} - \boldsymbol{f}\left(\boldsymbol{y}, \boldsymbol{\theta}\right), \tag{21}$$

clearly if $\boldsymbol{y} = \boldsymbol{x}^{(k-1)}$ then $\boldsymbol{r} = \boldsymbol{0}$. To minimize this, we can optimize $\boldsymbol{r}^T\boldsymbol{r}$ by first taking the gradient wrt $\boldsymbol{y}$,

$$\begin{aligned} \nabla_{\boldsymbol{y}}\left(\boldsymbol{r}^T\boldsymbol{r}\right) &= -2\boldsymbol{x}^T J_{\boldsymbol{y}}\left(\boldsymbol{f}\left(\boldsymbol{y},\boldsymbol{\theta}\right)\right) + J_{\boldsymbol{y}}\left(\boldsymbol{f}\left(\boldsymbol{y},\boldsymbol{\theta}\right)\right)\boldsymbol{f}\left(\boldsymbol{y},\boldsymbol{\theta}\right) + \boldsymbol{f}\left(\boldsymbol{y},\boldsymbol{\theta}\right)^T J_{\boldsymbol{y}}\left(\boldsymbol{f}\left(\boldsymbol{y},\boldsymbol{\theta}\right)\right) && (22)\\ &= \left(\boldsymbol{f}\left(\boldsymbol{y},\boldsymbol{\theta}\right) - 2\boldsymbol{x}\right)^T J_{\boldsymbol{y}}\left(\boldsymbol{f}\left(\boldsymbol{y},\boldsymbol{\theta}\right)\right) + J_{\boldsymbol{y}}\left(\boldsymbol{f}\left(\boldsymbol{y},\boldsymbol{\theta}\right)\right)\boldsymbol{f}\left(\boldsymbol{y},\boldsymbol{\theta}\right), && (23)\end{aligned}$$

then descending on $\boldsymbol{y}$ until we find the previous iterate. At each gradient calculation we must compute 3 vector-Jacobian products: most autogradient software do not natively do Jacobian-vector products and instead implement it as two VJP.

$$\begin{aligned} a^{(k-1)} &= a^{(k)} + a^{(k-1)}J(x,\theta) && (24)\\ a^{(k-1)} - a^{(k-1)}J(x,\theta) - a^{(k)} &= 0 && (25)\\ a^{(k-1)}\left(I - J(x,\theta)\right) - a^{(k)} &= 0 && (26)\end{aligned}$$

$$\begin{aligned} \left(a^{(k-1)}\left(I - J(x,\theta)\right) - a^{(k)}\right)^T\left(a^{(k-1)}\left(I - J(x,\theta)\right) - a^{(k)}\right) && (27)\\ \left(\left(I - J(x,\theta)\right)^T\left(a^{(k-1)}\right)^T - \left(a^{(k)}\right)^T\right)\left(a^{(k-1)}\left(I - J(x,\theta)\right) - a^{(k)}\right) && (28)\\ \left(I - J(x,\theta)\right)^T\left(a^{(k-1)}\right)^T a^{(k-1)}\left(I - J(x,\theta)\right) - 2\left(a^{(k)}\right)^T a^{(k-1)}\left(I - J(x,\theta)\right) + \left(a^{(k)}\right)^T a^{(k)} && (29)\\ \left(a^{(k-1)}\right)^T a^{(k-1)}\left(I - J(x,\theta)^T\right)\left(I - J(x,\theta)\right) - 2\left(a^{(k)}\right)^T a^{(k-1)}\left(I - J(x,\theta)\right) + \left(a^{(k)}\right)^T a^{(k)} && (30)\end{aligned}$$

A Misc

A.1 Backward Euler, backwards is forward Euler, forwards, but backwards

Let $x(t)$ be an ODE whose derivative is defined by

$$\frac{dx}{dt} = f\left(x, t\right). \tag{31}$$

If we time-step forward from $t^{(0)}$ to $t^{(0)}$ starting from $x^{(0)} = x(0)$ using forward Euler and obtain the intermediate sequence of $x(t)$ values

$$x^{(1)}, x^{(1)}, \ldots, x^{(j)}, \tag{32}$$

we will obtain the same sequence of values if we time-step backward from $t^{(1)}$ to $t^{(0)}$ starting from $x^{(j)} = x\left(t^{(1)}\right)$ by using backward Euler.

Proof. Denote the step size taken in both forward and backward Euler by Δt. Each forward iteration is defined by

$$x_{\text{fwd}}^{(k+1)} = x_{\text{fwd}}^{(k)} + \Delta t f\left(x_{\text{fwd}}^{(k)}\right), \tag{33}$$

with each backward iteration being equivalently defined by

$$x_{\text{bwd}}^{(k)} = x_{\text{bwd}}^{(k+1)} - \Delta t f\left(x_{\text{bwd}}^{(k)}\right). \tag{34}$$

Assuming we start the backward iteration from the final result obtained by the forward iteration, we have $x_{\mathrm{bwd}}^{(j)} = x_{\mathrm{fwd}}^{(j)}$. Inductively, we get

$$x_{\mathrm{bwd}}^{(k)} = x_{\mathrm{fwd}}^{(k+1)} - \Delta t f\left(x_{\mathrm{bwd}}^{(k)}\right), \tag{35}$$

or, with some rearranging and from eq. (33),

$$x_{\mathrm{bwd}}^{(k)} + \Delta t f\left(x_{\mathrm{bwd}}^{(k)}\right) = x_{\mathrm{fwd}}^{(k+1)} = x_{\mathrm{fwd}}^{(k)} + \Delta t f\left(x_{\mathrm{fwd}}^{(k)}\right). \tag{36}$$

□

References