Adjoint-based Optimization

Yan Cheng

Department of Applied Physics and Applied Mathematics
Columbia University

July 24, 2023

Motivation

Suppose I want to solve a constrained optimization problem, where the constraint is given by a differential equation:

$$\begin{aligned} \min_{m} \quad & \Phi(m) = \frac{1}{2}\|u(m) - d\|_{\mathcal{D}}^2 \\ \text{subject to} \quad & \mathcal{L}[m](u) = 0. \end{aligned}$$

This optimization problem asks: Given data d and a belief that the physical model is described by $\mathcal{L}[m](u) = 0$, what is the optimal parameter m that fits this data when measuring data misfit by the norm $\|\cdot\|_{\mathcal{D}}$?

Here, $\mathcal{L}[m]$ is a differential operator capturing our knowledge about the underlying physics.

For instance:

- The acoustic wave equation could serve as the model with $\mathcal{L}[m] = m\partial_{tt} - \Delta$.
- In some scenarios, the elastic or viscoelastic wave equation might be more appropriate.

Suppose we want to solve it computationally via gradient descent. We would have to compute $\nabla_m \Phi$.

Challenges in Gradient Computation

We aim to compute $\frac{d\Phi}{dm}$. Using the chain rule:

$$\frac{d\Phi}{dm} = \left\langle \frac{\partial \Phi}{\partial u}, \frac{\partial u}{\partial m} \right\rangle$$

Computing $\frac{\partial u}{\partial m}$ is non-trivial. Even for linear PDEs in u, the mapping $m \to u$ is often nonlinear. For instance:

$$m\partial_{tt}(u_1 + u_2) - \Delta(u_1 + u_2) = (m\partial_{tt}u_1 - \Delta u_1) + (m\partial_{tt}u_2 - \Delta u_2)$$

yet,

$$(m_1 + m_2)\partial_{tt}u - \Delta u \neq (m_1\partial_{tt}u - \Delta u) + (m_2\partial_{tt}u - \Delta u)$$

generally.

A straightforward approach to compute $\partial u/\partial m$ is by perturbing m at each grid point to compute the updated u:

$$\lim_{h \to 0} \frac{u(m(x_i + h, y_j)) - u(m(x_i, y_j))}{h}$$

However, for large-scale issues, this method is computationally prohibitive.

Adjoint Method: The Strike of A Genius

First, differentiate the equation $\mathcal{L}u = 0$ w.r.t. m on both sides. For the wave operator:

$$LHS = \frac{d}{dm}(m\partial_{tt}u(m) - \Delta u(m)) = \partial_{tt}u + m\partial_{tt}\left(\frac{\partial u}{\partial m}\right) - \Delta\frac{\partial u}{\partial m}$$

This gives:

$$m\partial_{tt}\left(\frac{\partial u}{\partial m}\right) - \Delta\frac{\partial u}{\partial m} = -\partial_{tt}u$$

Generally,

$$\mathcal{L}\left(\frac{\partial u}{\partial m}\right) = -\frac{\partial}{\partial m}\mathcal{L}u$$

Let's define the **adjoint equation**:

$$\mathcal{L}^*\lambda = \frac{\partial \Phi}{\partial u}$$

For the L^2-norm, we have:

$$\frac{d}{dm}\|u(m) - d\|_2^2 = \left\langle u - d, \frac{\partial u}{\partial m}\right\rangle \implies \frac{\partial \Phi}{\partial u} = u - d$$

Hence,

$$\left\langle \mathcal{L}\left(\frac{\partial u}{\partial m}\right), \lambda\right\rangle = \left\langle \frac{\partial u}{\partial m}, \mathcal{L}^*\lambda\right\rangle$$

This equates to:

$$\left\langle -\frac{\partial}{\partial m}\mathcal{L}u, \lambda\right\rangle = \left\langle \frac{\partial u}{\partial m}, \frac{\partial \Phi}{\partial u}\right\rangle = \frac{d\Phi}{dm}$$

Gradient Computed via the Adjoint Method

Algorithm Gradient Computation

1: Solve for u in

$$\mathcal{L}u = 0$$

2: Solve for λ in

$$\mathcal{L}^*\lambda = 0$$

3: Compute

$$\nabla_m \Phi = -\left\langle \frac{\partial}{\partial m}\mathcal{L}u, \lambda \right\rangle$$

The primary advantage of the adjoint method is its efficiency in gradient computation. Regardless of the problem size or complexity:

- Only two PDEs need to be solved: one for u and one for λ.
- This remains constant even as the parameter space dimension increases.

In terms of the number of PDEs to solve, the complexity is $O(1)$. However, keep in mind that the computational cost of solving each PDE might increase with finer discretizations or higher dimensions.