Notes on Adjoint Methods and Automatic Differentiation

2009 Ice Sheet Modeling Summer School
Portland, Oregon

Patrick Heimbach

August 13, 2009

Contents

1 Introduction

2 A simple example

2.1 A model and an objective function

Consider the model L which maps the two-dimensional vector $\mathbf{x}$ onto the two-dimensional vector y. The model is given by

$$\mathbf{y} \; = \; L(\mathbf{x}) = \left[\begin{array}{c} y_1 \\ y_2 \end{array} \right] \; = \; \left[\begin{array}{cc} 0 & a \\ -b & 0 \end{array} \right] \cdot \left[\begin{array}{c} y_1 \\ y_2 \end{array} \right] \; = \; \left[\begin{array}{r} ax_2 \\ -bx_1 \end{array} \right] \qquad (1)$$

{eq:simplemodel}

Now, assume observations $[d_1 \quad d_2]^T$ are available for the two elements $[y_1 \quad y_2]^T$, and we can write a misfit or cost function

$$\begin{aligned} J_0 \; = \; J_0(\mathbf{y}) &= \frac{1}{\sigma_1^2} \left(y_1 - d_1 \right)^2 \; + \; \frac{1}{\sigma_2^2} \left(y_2 - d_2 \right)^2 \\ &= \frac{1}{\sigma_1^2} \left(ax_2 - d_1 \right)^2 \; + \; \frac{1}{\sigma_2^2} \left(-bx_1 - d_2 \right)^2 \end{aligned} \qquad (2)$$

{eq:simplecost}

where σ_1, σ_2 can be attributed to standard deviations, such that their squares correspond to variances $\sigma_1^2 \, \sigma_2^2$. In this form, J_0 represents a simple least-squares cost function.

We can also view J_0 as a *composite* mapping $J_0 = J_0(\mathbf{y}) = J_0(L(\mathbf{x}))$, such that

$$\begin{array}{rccccc} J_0 \; : & \mathbf{x} & \longmapsto & \mathbf{y} & \longmapsto & J_0[\mathbf{y}] \\ & \mathbf{x} & \longmapsto & L[\mathbf{x}] & \longmapsto & J_0[L[\mathbf{x}]] \end{array} \qquad (3)$$

We wish to find the gradient of J_0 with respect to the input variable $\mathbf{x}$ (note that, alternatively, or in addition, we could also be interested in the gradient of J_0 with respect to the model parameters $\mathbf{p} = [a \quad b]^T$). Of course, the example chosen is very simple, and from eqn. (2) we can readily write down the gradient:

$$\nabla_x J_0^T \; = \; \left[\begin{array}{c} \frac{\partial J_0}{\partial x_1} \\ \frac{\partial J_0}{\partial x_2} \end{array} \right] \; = \; \left[\begin{array}{r} -\frac{2b}{\sigma_2^2} \left(-bx_1 - d_2 \right) \\ \frac{2a}{\sigma_1^2} \left(ax_2 - d_1 \right) \end{array} \right] \qquad (4)$$

{eq:gradwrtx}

2.2 A conventional way for finding the gradient of J_0

Suppose, the function $J_0(L(\mathbf{x}))$ was too complicated to write down analytically, and we needed to compute the gradient numerically. Conventionally, in order to assemble the complete gradient we would perform two finite difference perturbations for each component $[x_1 \quad x_2]^T$, i.e. compute

$$\frac{\partial J_0}{\partial x_i} \; = \; \frac{J_0(\mathbf{x} + \epsilon \, \mathbf{e}_i) \; - \; J_0(\mathbf{x})}{\epsilon}$$

for small ϵ, and for each direction

$$\mathbf{e}_1 = [1 \quad 0]^T, \qquad \mathbf{e}_2 = [0 \quad 1]^T$$

This approach has serveral shortcomings:

- If the dimension of $\mathbf{x}$ was very large (e.g. 10^7 instead of just 2) and calculation of J_0 expensive, performing 10^7 perturbation calculations would be prohibitive;
- The accuracy depends on the coice of ϵ and the finite-differencing scheme used (here we just used the simplest possible)

2.3 The tangent linear and adjoint model

Consider how perturbations $\delta\mathbf{x}$ in $\mathbf{x}$ are mapped to perturbations $\delta\mathbf{y}$ in $\mathbf{y} = L\mathbf{x}$. We define the linearized model dL via the general expression $\delta\mathbf{y} = dL\,\delta\mathbf{x}$, for which we obtain:

$$\begin{aligned}
\left[\begin{array}{c} \delta x_1 \\ \delta x_2 \end{array}\right] \longmapsto \left[\begin{array}{c} \delta y_1 \\ \delta y_2 \end{array}\right] &= \left[\begin{array}{c} \frac{\partial y_1}{\partial x_1}\delta x_1 + \frac{\partial y_1}{\partial x_2}\delta x_2 \\ \frac{\partial y_2}{\partial x_1}\delta x_1 + \frac{\partial y_2}{\partial x_2}\delta x_2 \end{array}\right] \\
&= \left[\begin{array}{cc} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} \end{array}\right] \cdot \left[\begin{array}{c} \delta x_1 \\ \delta x_2 \end{array}\right] \qquad (5) \\
&= \left[\begin{array}{cc} 0 & a \\ -b & 0 \end{array}\right] \cdot \left[\begin{array}{c} \delta x_1 \\ \delta x_2 \end{array}\right] = \left[\begin{array}{c} a\,\delta x_1 \\ -b\,\delta x_2 \end{array}\right]
\end{aligned}$$

Since L is a linear mapping/model, the model Jacobian dL is identical to L (this is a choice to simplify our calculation for now).

Now, consider the total variation of J_0 with respect to $\mathbf{y}$:

$$\delta J_0 \;=\; \frac{\partial J_0}{\partial y_1}\delta y_1 \;+\; \frac{\partial J_0}{\partial y_2}\delta y_2 \;=\; \left\langle \frac{\partial J_0}{\partial \mathbf{y}}^T \;, \delta\mathbf{y} \right\rangle \qquad (6)$$

where we have used the notation for general scalar products $< \,.\,,\,.\, >$. Again, using eqn. (2), we obtain for the perturbation in J_0 due to the perturbation $\delta\mathbf{x}$:

$$\begin{aligned}
\delta J_0 &= \frac{2}{\sigma_1^2}\,(y_1 - d_1)\,\delta y_1 \;+\; \frac{2}{\sigma_2^2}\,(y_2 - d_2)\,\delta y_2 \\
&= \left[\frac{2}{\sigma_1^2}\,(y_1 - d_1) \quad \frac{2}{\sigma_2^2}\,(y_2 - d_2)\right] \cdot \left[\begin{array}{c} \delta y_1 \\ \delta y_2 \end{array}\right] \\
&= \left[\frac{2}{\sigma_1^2}\,(a x_2 - d_1) \quad \frac{2}{\sigma_2^2}\,(-b x_1 - d_2)\right] \cdot \left[\begin{array}{cc} 0 & a \\ -b & 0 \end{array}\right] \cdot \left[\begin{array}{c} \delta x_1 \\ \delta x_2 \end{array}\right] \qquad (7) \quad \texttt{\{eq:deljtlmx\}} \\
&= \left[-\frac{2b}{\sigma_2^2}\,(-b x_1 - d_2) \quad \frac{2a}{\sigma_1^2}\,(a x_2 - d_1)\right] \cdot \left[\begin{array}{c} \delta x_1 \\ \delta x_2 \end{array}\right]
\end{aligned}$$

We now see how the gradient is obtained through repeating eqn. (**??**) and using the formal definition of the adjoint as follows:

$$\begin{aligned} \delta J_0 &= \left\langle \frac{\partial J_0}{\partial \mathbf{y}}^T , \delta \mathbf{y} \right\rangle \\ &= \left\langle \frac{\partial J_0}{\partial \mathbf{y}}^T , dL \, \delta \mathbf{x} \right\rangle = \left\langle dL^T \frac{\partial J_0}{\partial \mathbf{y}}^T , \delta \mathbf{x} \right\rangle = \left\langle \frac{\partial J_0}{\partial \mathbf{x}}^T , \delta \mathbf{x} \right\rangle \end{aligned} \tag{8}$$

or, more concise, and using a unit "perturbation" $\delta J_0^T = 1$,

$$\nabla_x J_0^T = \frac{\partial \mathbf{y}}{\partial \mathbf{x}}^T \cdot \frac{\partial J_0}{\partial \mathbf{y}}^T \cdot \delta J_0^T \tag{9}$$

We obtain general expressions for the tangent linear model and its dual, the adjoint model:

$$\begin{array}{llcccc} \text{TLM} & dJ_0 : & \delta \mathbf{x} & \longrightarrow & \delta \mathbf{y} = dL \cdot \delta \mathbf{x} & \longrightarrow \quad \delta J_0 = \nabla_y J_0 \cdot \delta \mathbf{y} \\ \text{ADM} & d^* J_0 : & \delta^* \mathbf{x} = dL^T \cdot \delta^* \mathbf{y} & \longleftarrow & \delta^* \mathbf{y} = \nabla_y J_0^T & \longleftarrow \quad \delta^* J_0 = 1 \end{array} \tag{10}$$

In our new notation, we recognize eqn. (**??**) as *tangent linear model* (TLM), while the adjoint model is readily written as:

$$\begin{aligned} \delta^* \mathbf{x} \; &= \left[\begin{array}{c} \delta^* x_1 \\ \delta^* x_2 \end{array} \right] = \left[\begin{array}{c} -\frac{2b}{\sigma_2^2}(-bx_1 - d_2) \\ \frac{2a}{\sigma_1^2}(ax_2 - d_1) \end{array} \right] \\ &= \left[\begin{array}{cc} 0 & -b \\ a & 0 \end{array} \right] \cdot \left[\begin{array}{c} \frac{2}{\sigma_1^2}(ax_2 - d_1) \\ \frac{2}{\sigma_2^2}(-bx_1 - d_2) \end{array} \right] \cdot \delta^* J_0 \\ &= dL^T \cdot \delta^* \mathbf{y} \cdot \delta^* J_0 \end{aligned} \tag{11}$$

{eq:deljadmx}

with $\delta^* J_0 = 1$ and $\delta^* \mathbf{x} = \nabla_{\mathbf{x}} J_0^T$. The $*$ variables are referred to either as *dual* variables, *adjoint* variables, *sensitivities*, or *Lagrange multipliers*.

2.4 Change of control space: same model, but different adjoint

We've managed to squeeze a lot of equations out of this simple problem, and we think we're done, but not so fast. Imagine, instead of being interested in the sensitivity of J_0 with respect to $\mathbf{x}$ (which previously were considered to be uncertain "initial conditions"), we instead consider $\mathbf{x}$ to be uninteresting, and are interested in the sensitivities with respect to the *model parameters* $\mathbf{p} = [a \quad b]^T$. Everything stays the same, we still use the model L and are evaluating the cost function J_0, but now as a function of $\mathbf{p}$ for fixed $\mathbf{x}$. As might be expected, the gradient $\nabla_p J_0$ of J_0 with respect to $\mathbf{p}$ has quite a different form, compared to $\nabla_x J_0$, eqn. (4). We can derive it directly from eqn. (2):

$$\nabla_p J_0^T = \left[\begin{array}{c} \frac{\partial J_0}{\partial a} \\ \frac{\partial J_0}{\partial b} \end{array} \right] = \left[\begin{array}{c} \frac{2}{\sigma_1^2} (ax_2 - d_1) \, x_2 \\ -\frac{2}{\sigma_2^2} (-bx_1 - d_2) \, x_1 \end{array} \right] \tag{12}$$

{eq:gradwrtp}

A calculation similar to eqn. (7) yields:

$$\begin{aligned}
\delta J_0 &= \frac{\partial J_0}{\partial a}\delta a + \frac{\partial J_0}{\partial b}\delta b \\
&= \left(\frac{\partial J_0}{\partial y_1}\frac{\partial y_1}{\partial a} + \frac{\partial J_0}{\partial y_2}\frac{\partial y_2}{\partial a} \right) \delta a + \left(\frac{\partial J_0}{\partial y_1}\frac{\partial y_1}{\partial b} + \frac{\partial J_0}{\partial y_2}\frac{\partial y_2}{\partial b} \right) \delta b \\
&= \left[\begin{array}{cc} \frac{2}{\sigma_1^2}\left(y_1 - d_1\right) & -\frac{2}{\sigma_2^2}\left(y_2 - d_2\right) \end{array} \right] \cdot \left[\begin{array}{c} \delta a \\ \delta b \end{array} \right] \\
&= \left[\begin{array}{cc} \frac{2}{\sigma_1^2}\left(ax_2 - d_1\right) & -\frac{2}{\sigma_2^2}\left(-bx_1 - d_2\right) \end{array} \right] \cdot \left[\begin{array}{cc} x_2 & 0 \\ 0 & -x_1 \end{array} \right] \cdot \left[\begin{array}{c} \delta a \\ \delta b \end{array} \right]
\end{aligned} \qquad (13) \quad \texttt{\{eq:deljtlmp\}}$$

from which we can readily deduce the adjoint expression

$$\begin{aligned}
\delta^* \mathbf{p} \;=\; \left[\begin{array}{c} \delta^* a \\ \delta^* b \end{array} \right] &= \left[\begin{array}{c} \frac{2}{\sigma_1^2}(ax_2 - d_1)x_2 \\ -\frac{2}{\sigma_2^2}(-bx_1 - d_1)x_1 \end{array} \right] \\
&= \left[\begin{array}{cc} x_2 & 0 \\ 0 & -x_1 \end{array} \right] \cdot \left[\begin{array}{c} \frac{2}{\sigma_1^2}(ax_2 - d_1) \\ \frac{2}{\sigma_2^2}(-bx_1 - d_2) \end{array} \right] \cdot \delta^* J_0 \\
&= d\tilde{L}^T \cdot \delta^* \mathbf{y} \cdot \delta^* J_0
\end{aligned} \qquad (14) \quad \texttt{\{eq:deljadmp\}}$$

The mapping relationship corresponding to eqn. (11), but now with **p** as control, reads:

$$\begin{array}{llccccc}
\text{TLM} & dJ_0 \; : & \delta \mathbf{p} & \longrightarrow & \delta \mathbf{y}(\mathbf{p}) = d\tilde{L} \cdot \delta \mathbf{p} & \longrightarrow & \delta J_0 = \nabla_y J_0 \cdot \delta \mathbf{y} \\
\text{ADM} & d^* J_0 \; : & \delta^* \mathbf{p} = d\tilde{L}^T \cdot \delta^* \mathbf{y} & \longleftarrow & \delta^* \mathbf{y} = \nabla_y J_0^T & \longleftarrow & \delta^* J_0 = 1
\end{array} \qquad (15)$$

The bottom line is that the gradient, and thus "the adjoint model" looks quite different for this control problem. Several lessons have been learnt:

- There isn't such a thing as "the" adjoint model. Its form depends crucially on the control problem formulated, i.e. on the set of independent and dependent variables chosen (an issue not appreciated to this day by a large fraction of the ocean modeling community).

- One of the crucial strengths of automatic differentiation is the very fact that it can deal much more flexibly with changes to the formulation of the control problem that one wishes to solve.

- It isn't even clear, for a given control problem, which part of eqn. (11), (or, equivalently of eqn. (14)) refers to "the adjoint model". Mathematicians would refer to the entire expression $dL^T \cdot \delta^* \mathbf{y} \cdot \delta^* J_0$ as the adjoint of the mapping $J_0(L(\mathbf{x}))$, whereas physicists tend to think of L as "the model", to dL as "the model Jacobian", and thus to dL^T only as "the adjoint model“. The caveats are evident.

- Note also that in eqn. (10) and (15) the expressions for $\nabla_y J_0$ (and their transpose) remain the same, and it is really dL vs. $d\tilde{L}$ (and their transpose) which change the overall TLM and ADM.

2.5 Line-by-line: code generated via automatic differentiation

2.5.1 The tangent linear code (forward-mode AD)

2.5.2 The adjoint code (reverse-mode AD)

3 The 2nd derivative, the Hessian, and posterior error

4 The time-varying problem: the harmonic oscillator

4.1 Lagrange multipliers vs. chain rule

4.2 An example: the harmonic oscillator