Perspectives on strong constraint 4DVar

Ross Bannister

February 24, 2020

1 Introduction

Consider the following strong constraint incremental 4DVar cost function

$$\begin{aligned} J(\delta\mathbf{x}_0) &= \frac{1}{2}\delta\mathbf{x}_0^{\mathrm{T}}\mathbf{B}^{-1}\delta\mathbf{x}_0 + \frac{1}{2}\sum_{t=0}^{T}\left(\delta\mathbf{y}_t - \mathbf{H}_t\delta\mathbf{x}_t\right)^{\mathrm{T}}\mathbf{R}_t^{-1}\left(\delta\mathbf{y}_t - \mathbf{H}_t\delta\mathbf{x}_t\right), \\ &= J^{\mathrm{b}}(\delta\mathbf{x}_0) + J^{\mathrm{o}}(\delta\mathbf{x}_0). \end{aligned} \tag{1}$$

where $\delta\mathbf{x}_t = \mathbf{x}_t - \boldsymbol{\mathcal{M}}_{0\to t}\left(\mathbf{x}^{\mathrm{b}}\right)$ is the increment from the background, $\mathbf{x}_t$ is the state at time t, $\mathbf{x}^{\mathrm{b}}$ is the background state (valid at $t=0$), $\boldsymbol{\mathcal{M}}_{t\to t'}(\bullet)$ is the non-linear forecast model propagator from times t to t', $\mathbf{B}$ is the background error covariance matrix, $\mathbf{R}_t$ is the observation error covariance matrix at time t, T is the length of the 4DVar time window, $\delta\mathbf{y}_t = \mathbf{y}_t - \boldsymbol{\mathcal{H}}_t\left(\boldsymbol{\mathcal{M}}_{0\to t}\left(\mathbf{x}^{\mathrm{b}}\right)\right)$ is the observation-minus-background, $\mathbf{y}_t$ are the observations at time t, $\boldsymbol{\mathcal{H}}_t(\bullet)$ is the non-linear observation operator at time t, and $\mathbf{H}_t$ is its Jacobian. Additionally, J^{b} is the background term of the cost function, and J^{o} is the observation term. Furthermore, let $\mathbf{M}_{t\to t'}$ be the Jacobian of the forecast model from times t to t' such that

$$\delta\mathbf{x}_{t'} = \mathbf{M}_{t\to t'}\delta\mathbf{x}_t. \tag{2}$$

The problem may be posed in the following way: minimise (1) with respect to $\delta\mathbf{x}_0, \ldots, \delta\mathbf{x}_T$ subject to the constraint (2).

We will consider two methods of deriving the gradient of (1) with respect to the control variable $\delta\mathbf{x}_0$.

2 Method A of deriving the gradient: imposing the constraint explicitly

2.1 The cost function and its gradient

Method A imposes the constraint by substitution. Substituting (2) into (1) gives a cost function written explicitly in terms of $\delta\mathbf{x}_0$:

$$J(\delta\mathbf{x}_0) = \frac{1}{2}\delta\mathbf{x}_0^{\mathrm{T}}\mathbf{B}^{-1}\delta\mathbf{x}_0 + \frac{1}{2}\sum_{t=0}^{T}\left(\delta\mathbf{y}_t - \mathbf{H}_t\mathbf{M}_{0\to t}\delta\mathbf{x}_0\right)^{\mathrm{T}}\mathbf{R}_t^{-1}\left(\delta\mathbf{y}_t - \mathbf{H}_t\mathbf{M}_{0\to t}\delta\mathbf{x}_0\right). \tag{3}$$

The total derivative with respect to $\delta\mathbf{x}_0$ is

$$\frac{dJ}{d\delta\mathbf{x}_0} = \frac{dJ^{\mathrm{b}}}{d\delta\mathbf{x}_0} + \frac{dJ^{\mathrm{o}}}{d\delta\mathbf{x}_0} \tag{4}$$

$$\frac{dJ^{\mathrm{b}}}{d\delta\mathbf{x}_0} = \mathbf{B}^{-1}\delta\mathbf{x}_0 \tag{5}$$

$$\frac{dJ^{\mathrm{o}}}{d\delta\mathbf{x}_0} = -\sum_{t=0}^{T}\mathbf{M}_{0\to t}^{\mathrm{T}}\mathbf{H}_t^{\mathrm{T}}\mathbf{R}_t^{-1}\left(\delta\mathbf{y}_t - \mathbf{H}_t\mathbf{M}_{0\to t}\delta\mathbf{x}_0\right). \tag{6}$$

Note that we call this the *total* derivative because this derivative accounts for the dependency not just to $\delta\mathbf{x}_0$ but also due to $\delta\mathbf{x}_t$ ($0 < t \le T$) given that (2) holds.

2.2 Rewriting the gradient for efficiency

Equation (6) for the observation term's total gradient can be written in a way that allows it to be evaluated efficiently. First note that $\mathbf{M}^{\mathrm{T}}_{0\to t} = \mathbf{M}^{\mathrm{T}}_{0\to 1} \dots \mathbf{M}^{\mathrm{T}}_{t-1\to t}$, and use this to write the summation in (6) systematically. In the following, each row is a term in (6), and let $\mathbf{g}_t = \mathbf{H}^{\mathrm{T}}_t \mathbf{R}^{-1}_t (\delta\mathbf{y}_t - \mathbf{H}_t \mathbf{M}_{0\to t}\delta\mathbf{x}_0)$:

$$
-\frac{dJ^{\mathrm{o}}}{d\delta\mathbf{x}_0} = \begin{array}{llllll}
\mathbf{g}_0 + & & & & & \\
\mathbf{M}^{\mathrm{T}}_{0\to 1} & \mathbf{g}_1 + & & & & \\
\mathbf{M}^{\mathrm{T}}_{0\to 1} & \mathbf{M}^{\mathrm{T}}_{1\to 2} & \mathbf{g}_2 + & & & \\
\vdots & \vdots & \vdots & \mathbf{g}_{T-2} + & & \\
\mathbf{M}^{\mathrm{T}}_{0\to 1} & \mathbf{M}^{\mathrm{T}}_{1\to 2} & \cdots & \mathbf{M}^{\mathrm{T}}_{T-2\to T-1} & \mathbf{g}_{T-1} + & \\
\mathbf{M}^{\mathrm{T}}_{0\to 1} & \mathbf{M}^{\mathrm{T}}_{1\to 2} & \cdots & \mathbf{M}^{\mathrm{T}}_{T-2\to T-1} & \mathbf{M}^{\mathrm{T}}_{T-1\to T} & \mathbf{g}_T .
\end{array} \tag{7}
$$

The efficiency can be gained by observing that the common adjoint operator in each column of (7). It will be evident that the gradient is equivalent to executing the following algorithm.

1. Set $\boldsymbol{\mu}_{T+1} = 0$.
2. Loop backwards from $t = T$ to $t = 0$:
   (a) Let $\boldsymbol{\mu}_t = \mathbf{M}^{\mathrm{T}}_{t\to t+1}\boldsymbol{\mu}_{t+1} + \mathbf{g}_t$.
3. $\boldsymbol{\mu}_0$ evaluates to $-dJ^{\mathrm{o}}/d\delta\mathbf{x}_0$.

The total derivative of all terms in the cost function is then

$$
\frac{dJ}{d\delta\mathbf{x}_0} = \frac{dJ^{\mathrm{b}}}{d\delta\mathbf{x}_0} + \frac{dJ^{\mathrm{o}}}{d\delta\mathbf{x}_0} = \mathbf{B}^{-1}\delta\mathbf{x}_0 - \boldsymbol{\mu}_0 . \tag{8}
$$

3 Method B of deriving the gradient: using Lagrange multipliers

3.1 Introducing the Lagrange multipliers and finding the derivatives

Method A imposes the constraint with Lagrange multipliers. The problem of minimising (1) subject to constraint (2) is to write the following new unconstrained minimisation problem with one Lagrange multiplier, $\boldsymbol{\lambda}_t$, multiplying each constraint:

$$
L(\delta\mathbf{x}_0, \dots, \delta\mathbf{x}_T; \boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_T) = J(\delta\mathbf{x}_0, \dots, \delta\mathbf{x}_T) + \sum_{t=0}^{T-1} \boldsymbol{\lambda}^{\mathrm{T}}_{t+1} (\delta\mathbf{x}_{t+1} - \mathbf{M}_{t\to t+1}\delta\mathbf{x}_t) . \tag{9}
$$

Minimising (9) with respect to $\delta\mathbf{x}_0, \dots, \delta\mathbf{x}_T; \boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_T$ in an unconstrained way is equivalent to minimising $J(\delta\mathbf{x}_0, \dots, \delta\mathbf{x}_T)$ with respect to $\delta\mathbf{x}_0, \dots, \delta\mathbf{x}_T$ subject to the constraint that $\delta\mathbf{x}_{t+1} = \mathbf{M}_{t\to t+1}\delta\mathbf{x}_t$.

In order to minimise (9), we need the following derivatives (set to zero):

$$
\frac{\partial L}{\partial \delta\mathbf{x}_{t'}} = \frac{\partial J}{\partial \delta\mathbf{x}_{t'}} + \sum_{t=0}^{T-1} \frac{\partial}{\partial \delta\mathbf{x}_{t'}} \left(\boldsymbol{\lambda}^{\mathrm{T}}_{t+1}\delta\mathbf{x}_{t+1} \right) - \sum_{t=0}^{T-1} \frac{\partial}{\partial \delta\mathbf{x}_{t'}} \left(\boldsymbol{\lambda}^{\mathrm{T}}_{t+1}\mathbf{M}_{t\to t+1}\delta\mathbf{x}_t \right) = 0 \tag{10}
$$

$$
\frac{\partial L}{\partial \boldsymbol{\lambda}_{t'}} = \delta\mathbf{x}_{t'+1} - \mathbf{M}_{t'\to t'+1}\delta\mathbf{x}_{t'} = 0. \tag{11}
$$

Note that (i) we write (10) and (11) as *partial* derivatives (rather than the *total* derivatives in (4)) as each argument of $L(\bullet)$ is now considered an independent variable, and (ii) that the derivatives with respect to $\boldsymbol{\lambda}_{t'}$ recover the constraints that are imposed.

We now do the differentiation needed to complete the last two terms in (10). For the penultimate term:

$$
\sum_{t=0}^{T-1} \boldsymbol{\lambda}^{\mathrm{T}}_{t+1}\delta\mathbf{x}_{t+1} = \sum_{t=0}^{T-1} \sum_{i} \left[\boldsymbol{\lambda}_{t+1}\right]_i \left[\delta\mathbf{x}_{t+1}\right]_i ,
$$

so $\sum_{t=0}^{T-1} \left(\partial / \partial \left[\delta \mathbf{x}_{t'}\right]_j\right) \left(\boldsymbol{\lambda}_{t+1}^{\mathrm{T}} \delta \mathbf{x}_{t+1}\right) = \sum_{t=0}^{T-1} \sum_i \left[\boldsymbol{\lambda}_{t+1}\right]_i \delta_{t',t+1} \delta_{i,j} = \left[\boldsymbol{\lambda}_{t'}\right]_j$. So the derivative with respect to the vector $\delta \mathbf{x}_{t'}$ is $\boldsymbol{\lambda}_{t'}$. For the last term:

$$\sum_{t=0}^{T-1} \boldsymbol{\lambda}_{t+1}^{\mathrm{T}} \mathbf{M}_{t \to t+1} \delta \mathbf{x}_t = \sum_{t=0}^{T-1} \sum_i \sum_{i'} \left[\boldsymbol{\lambda}_{t+1}\right]_{i'} \left[\mathbf{M}_{t \to t+1}\right]_{i'i} \left[\delta \mathbf{x}_t\right]_i ,$$

so $\sum_{t=0}^{T-1} \left(\partial / \partial \left[\delta \mathbf{x}_{t'}\right]_j\right) \left(\boldsymbol{\lambda}_{t+1}^{\mathrm{T}} \mathbf{M}_{t \to t+1} \delta \mathbf{x}_t\right) = \sum_{t=0}^{T-1} \sum_i \sum_{i'} \left[\boldsymbol{\lambda}_{t+1}\right]_{i'} \left[\mathbf{M}_{t \to t+1}\right]_{i'i} \delta_{t',t} \delta_{i,j} = \sum_{i'} \left[\boldsymbol{\lambda}_{t'+1}\right]_{i'} \left[\mathbf{M}_{t' \to t'+1}\right]_{i'j} = \sum_{i'} \left[\mathbf{M}_{t' \to t'+1}^{\mathrm{T}}\right]_{ji'} \left[\boldsymbol{\lambda}_{t'+1}\right]_{i'}$. So the derivative with respect to the vector $\delta \mathbf{x}_{t'}$ is $\mathbf{M}_{t' \to t'+1}^{\mathrm{T}} \boldsymbol{\lambda}_{t'+1}$. Equation (10) is then

$$\frac{\partial L}{\partial \delta \mathbf{x}_{t'}} = \frac{\partial J}{\partial \delta \mathbf{x}_{t'}} + \boldsymbol{\lambda}_{t'} - \mathbf{M}_{t' \to t'+1}^{\mathrm{T}} \boldsymbol{\lambda}_{t'+1} = 0. \tag{12}$$

The Lagrange multipliers, $\boldsymbol{\lambda}_t$ (sometimes called adjoint variables), are defined only for $1 \leq t \leq T$ in (12), although we will define two extra adjoint variables, namely $\boldsymbol{\lambda}_{T+1} \equiv 0$ and $\boldsymbol{\lambda}_0$. We define $\boldsymbol{\lambda}_0$ as a variable that obeys (12), although is not associated with any constraint.

3.2 The total gradient

We are interested in the total gradient of the cost function at $t = 0$, but we only have partial derivatives at each time. The relationship between the total gradient and the partial derivatives is found from the chain rule:

$$\frac{dJ}{d\delta \mathbf{x}_0} = \sum_{t=0}^{T} \left(\frac{\partial \mathbf{x}_t}{\partial \mathbf{x}_0}\right)^{\mathrm{T}} \frac{\partial J}{\partial \delta \mathbf{x}_t} = \sum_{t=0}^{T} \mathbf{M}_{0 \to t}^{\mathrm{T}} \frac{\partial J}{\partial \delta \mathbf{x}_t}. \tag{13}$$

N.B. the transpose operator is present for the following reason. In terms of components, the chain rule is

$$\frac{dJ}{d\left[\delta \mathbf{x}_0\right]_i} = \sum_{t=0}^{T} \sum_j \frac{\partial \left[\delta \mathbf{x}_t\right]_j}{\partial \left[\delta \mathbf{x}_0\right]_i} \frac{\partial J}{\partial \left[\delta \mathbf{x}_t\right]_j} = \sum_{t=0}^{T} \sum_j \left[\mathbf{M}_{0 \to t}\right]_{ji} \frac{\partial J}{\partial \left[\delta \mathbf{x}_t\right]_j} = \sum_{t=0}^{T} \sum_j \left[\mathbf{M}_{0 \to t}^{\mathrm{T}}\right]_{ij} \frac{\partial J}{\partial \left[\delta \mathbf{x}_t\right]_j},$$

which gives component i of (13) in line with normal matrix algebra.

Substituting (12) into (13):

$$\begin{aligned}
\frac{dJ}{d\delta \mathbf{x}_0} &= \sum_{t=0}^{T} \mathbf{M}_{0 \to t}^{\mathrm{T}} \frac{\partial J}{\partial \delta \mathbf{x}_t} \\
&= \sum_{t=0}^{T} \mathbf{M}_{0 \to t}^{\mathrm{T}} \left(\mathbf{M}_{t \to t+1}^{\mathrm{T}} \boldsymbol{\lambda}_{t+1} - \boldsymbol{\lambda}_t\right) \\
&= \sum_{t=0}^{T} \mathbf{M}_{0 \to t}^{\mathrm{T}} \mathbf{M}_{t \to t+1}^{\mathrm{T}} \boldsymbol{\lambda}_{t+1} - \sum_{t=0}^{T} \mathbf{M}_{0 \to t}^{\mathrm{T}} \boldsymbol{\lambda}_t \\
&= \sum_{t=0}^{T} \mathbf{M}_{0 \to t+1}^{\mathrm{T}} \boldsymbol{\lambda}_{t+1} - \sum_{t=0}^{T} \mathbf{M}_{0 \to t}^{\mathrm{T}} \boldsymbol{\lambda}_t \\
&= \sum_{t=1}^{T+1} \mathbf{M}_{0 \to t}^{\mathrm{T}} \boldsymbol{\lambda}_t - \sum_{t=0}^{T} \mathbf{M}_{0 \to t}^{\mathrm{T}} \boldsymbol{\lambda}_t \\
&= \sum_{t=1}^{T} \mathbf{M}_{0 \to t}^{\mathrm{T}} \boldsymbol{\lambda}_t + \mathbf{M}_{0 \to T+1}^{\mathrm{T}} \boldsymbol{\lambda}_{T+1} - \boldsymbol{\lambda}_0 - \sum_{t=1}^{T} \mathbf{M}_{0 \to t}^{\mathrm{T}} \boldsymbol{\lambda}_t \\
&= -\boldsymbol{\lambda}_0,
\end{aligned} \tag{14}$$

where, to get the last line, the first and last terms cancel and we use the boundary condition that $\boldsymbol{\lambda}_{T+1} = 0$. This set of steps proves that $\boldsymbol{\lambda}_0$ is equal to minus the total gradient of the cost function.

3.3 Finding $\boldsymbol{\lambda}_0$ (and hence the total gradient)

All is left is to find $\boldsymbol{\lambda}_0$. Consider (12), and substitute $\partial J/\partial\delta\mathbf{x}_{t'}$ for the actual partial derivative from (1).

$$\mathbf{B}^{-1}\delta\mathbf{x}_0\delta_{t,0} - \mathbf{H}_t^{\mathrm{T}}\mathbf{R}_t^{-1}\left(\delta\mathbf{y}_t - \mathbf{H}_t\delta\mathbf{x}_t\right) + \boldsymbol{\lambda}_t - \mathbf{M}_{t\to t+1}^{\mathrm{T}}\boldsymbol{\lambda}_{t+1} = 0. \tag{15}$$

Generate an algorithm whereby (15) is integrated backwards from T to 0 (where it is noted that $\boldsymbol{\lambda}_{T+1} = 0$).

1. Set $\boldsymbol{\lambda}_{T+1} = 0$.
2. Loop backwards from $t = T$ to $t = 0$:
   (a) Let $\boldsymbol{\lambda}_t = \mathbf{M}_{t\to t+1}^{\mathrm{T}}\boldsymbol{\lambda}_{t+1} + \mathbf{H}_t^{\mathrm{T}}\mathbf{R}_t^{-1}\left(\delta\mathbf{y}_t - \mathbf{H}_t\delta\mathbf{x}_t\right) - \mathbf{B}^{-1}\delta\mathbf{x}_0\delta_{t,0}$.
3. $\boldsymbol{\lambda}_0$ evaluates to $-dJ/d\delta\mathbf{x}_0$.

Notice that this algorithm is, in effect, exactly the same as that given at the end of Sect. 3 for the case when the constraint is applied explicitly. Thus the two approaches A and B are equivalent.