Derivation of the Hessian for Deep Equilibrium Models

This document details the derivation of the Hessian of the Loss function with respect to parameters θ for a Deep Equilibrium Model (DEQ).

1. Fixed Point and Gradient

Fixed Point Condition

The hidden state h_* is defined implicitly by the fixed point equation:

$$h_* = f_\theta(x, h_*)$$

Implicit Differentiation (Jacobian)

Differentiating with respect to θ:

$$\frac{dh_*}{d\theta} = \frac{\partial f}{\partial \theta} + \frac{\partial f}{\partial h_*}\frac{dh_*}{d\theta}$$

Solving for the total Jacobian $Z = \frac{dh_*}{d\theta}$:

$$\left(I - \frac{\partial f}{\partial h_*}\right)\frac{dh_*}{d\theta} = \frac{\partial f}{\partial \theta}$$

$$Z = \frac{dh_*}{d\theta} = (I - J)^{-1}\frac{\partial f}{\partial \theta}$$

where $J = \frac{\partial f}{\partial h_*}$ is the Jacobian of the transformation f with respect to the state.

The Gradient

The gradient of the loss $L(h_*)$ is:

$$\nabla_\theta L = \left(\frac{dL}{d\theta}\right)^T = \left(\frac{\partial L}{\partial h_*}\frac{dh_*}{d\theta}\right)^T = Z^T \nabla_{h_*} L$$

---

2. The Hessian Setup

The Hessian H is the derivative of the gradient vector $\nabla_\theta L$ with respect to θ:

$$H = \frac{d}{d\theta}\left(\nabla_\theta L\right) = \frac{d}{d\theta}\left[Z^T \nabla_{h_*} L\right]$$

Applying the product rule, this splits into two terms:

$$H = \underbrace{Z^T \frac{d}{d\theta}(\nabla_{h_*} L)}_{\text{Term A: Loss Curvature}} + \underbrace{\left[\frac{d}{d\theta} Z^T\right] \nabla_{h_*} L}_{\text{Term B: Dynamics Curvature}}$$

---

3. Term A: Loss Curvature

We evaluate $\frac{d}{d\theta}(\nabla_{h_*} L)$ using the chain rule, since the gradient depends on θ via h_*:

$$\frac{d}{d\theta}(\nabla_{h_*} L) = \frac{\partial(\nabla_{h_*} L)}{\partial h_*} \frac{dh_*}{d\theta} = \nabla^2_{h_*} L \cdot Z$$

Substituting this back gives the standard Gauss-Newton curvature term:

$$\text{Term A} = Z^T (\nabla^2_{h_*} L) Z$$

---

4. Term B: Dynamics Curvature

This term captures how the equilibrium state itself changes curvature as parameters change.

$$\text{Term B} = \text{Contract}\left(\frac{d^2 h_*}{d\theta^2}, \nabla_{h_*} L\right)$$

4.1 Deriving the Second Derivative of the State

To find $\frac{d^2 h_*}{d\theta^2}$, we differentiate the fixed point gradient equation with respect to θ. Recall the equation for the Jacobian Z:

$$\left(I - \frac{\partial f}{\partial h_*}\right) Z = \frac{\partial f}{\partial \theta}$$

Applying the product rule to the Left Hand Side (LHS) and the total derivative to the Right Hand Side (RHS):

$$\left(I - \frac{\partial f}{\partial h_*}\right) \frac{dZ}{d\theta} + \left[\frac{d}{d\theta}\left(I - \frac{\partial f}{\partial h_*}\right)\right] Z = \frac{d}{d\theta}\left(\frac{\partial f}{\partial \theta}\right)$$

Step 1: Expand the RHS The term $\frac{d}{d\theta}\left(\frac{\partial f}{\partial \theta}\right)$ involves a total derivative. Since f depends on θ directly and via h_*:

$$\text{RHS} = \frac{\partial^2 f}{\partial \theta^2} + \text{Contract}\left(\frac{\partial^2 f}{\partial h_* \partial \theta}, Z\right)$$

Step 2: Expand the LHS Bracket The term $\frac{d}{d\theta}\left(I - \frac{\partial f}{\partial h_*}\right)$ involves differentiating the Jacobian matrix. The identity I vanishes.

$$[\dots] = -\frac{d}{d\theta}\left(\frac{\partial f}{\partial h_*}\right) = -\left(\frac{\partial^2 f}{\partial \theta \partial h_*} + \text{Contract}\left(\frac{\partial^2 f}{\partial h_*^2}, Z\right)\right)$$

Step 3: Substitute and Rearrange Substituting these expansions back into the main equation:

$$(I-J)\frac{d^2 h_*}{d\theta^2} - \left(\frac{\partial^2 f}{\partial \theta \partial h_*} + \text{Contract}\left(\frac{\partial^2 f}{\partial h_*^2}, Z\right)\right) Z = \frac{\partial^2 f}{\partial \theta^2} + \text{Contract}\left(\frac{\partial^2 f}{\partial h_* \partial \theta}, Z\right)$$

Move the negative terms to the RHS. Note that the mixed derivative terms appear on both sides (one from expanding RHS, one from expanding LHS).

$$(I-J)\frac{d^2 h_*}{d\theta^2} = \frac{\partial^2 f}{\partial \theta^2} + \underbrace{\text{Contract}\left(\frac{\partial^2 f}{\partial h_* \partial \theta}, Z\right) + \text{Contract}\left(\frac{\partial^2 f}{\partial \theta \partial h_*}, Z\right)}_{\text{Mixed Terms}} + \text{DoubleContract}\left(\frac{\partial^2 f}{\partial h_*^2}, Z, Z\right)$$

Step 4: Solve for $\frac{d^2 h_*}{d\theta^2}$ Multiply by the inverse Jacobian $(I-J)^{-1}$:

$$\frac{d^2 h_*}{d\theta^2} = (I-J)^{-1}\left[\frac{\partial^2 f}{\partial \theta^2} + 2\frac{\partial^2 f}{\partial h_* \partial \theta} Z + \frac{\partial^2 f}{\partial h_*^2} Z^2\right]$$

(Note: The notation inside the brackets represents the Total Hessian of f w.r.t θ, denoted as $\mathcal{T}_{total}$).

4.2 Applying the Adjoint Method

Now we substitute this result back into Term B:

$$\text{Term B} = \nabla_{h_*} L \cdot (I-J)^{-1} \cdot \mathcal{T}_{total}$$

We define the **Adjoint Vector** λ to avoid computing the full tensor inverse:

$$\lambda^T = (\nabla_{h_*} L)^T (I - J)^{-1}$$

Thus, Term B becomes the contraction of λ with the Total Hessian tensor:

$$\text{Term B} = \text{Contract}(\lambda, \mathcal{T}_{total})$$

This is equivalent to the Hessian of the scalar function $\lambda^T f$.

---

5. Final Formula

Combining Term A and Term B, we obtain the complete Hessian.

Block Matrix Form

Let $S(\theta, h_*) = \lambda^T f(\theta, h_*)$ (with λ fixed).

$$H = \underbrace{\left(\frac{dh_*}{d\theta}\right)^T (\nabla^2_{h_*} L) \left(\frac{dh_*}{d\theta}\right)}_{\text{Term A}} + \underbrace{\begin{pmatrix} I \\ \frac{dh_*}{d\theta} \end{pmatrix}^T \left[\nabla^2_{(\theta, h_*)} S\right] \begin{pmatrix} I \\ \frac{dh_*}{d\theta} \end{pmatrix}}_{\text{Term B}}$$

Full Expanded Formula (Symmetric)

Expanding the block matrix multiplication explicitly reveals the structure as a weighted sum over the state components $r = 1 \ldots N$. This form explicitly shows the symmetry of the mixed derivative terms:

$$H = \left(\frac{dh_*}{d\theta}\right)^T (\nabla^2_{h_*} L) \left(\frac{dh_*}{d\theta}\right) + \sum_{r=1}^{N} \lambda_r \left[\frac{\partial^2 f_r}{\partial \theta^2} + \underbrace{\left(\frac{dh_*}{d\theta}\right)^T \frac{\partial^2 f_r}{\partial h_* \partial \theta} + \left(\frac{\partial^2 f_r}{\partial h_* \partial \theta}\right)^T \frac{dh_*}{d\theta}}_{\text{Symmetrized Mixed Term}} + \left(\frac{dh_*}{d\theta}\right)^T \frac{\partial^2 f_r}{\partial h_*^2} \left(\frac{dh_*}{d\theta}\right) \right]$$

This formula separates the curvature of the cost function from the curvature of the physical constraint, coupled only by the sensitivity matrix $Z = \frac{dh_*}{d\theta}$.

Explicit Substitution

Substituting $\frac{dh_*}{d\theta} = (I - J)^{-1} \frac{\partial f}{\partial \theta}$ explicitly into the equation:

$$
\begin{aligned}
H = &\left(\frac{\partial f}{\partial \theta}\right)^T (I-J)^{-T}(\nabla^2_{h_*} L)(I-J)^{-1}\frac{\partial f}{\partial \theta} \\
&+ \sum_{r=1}^{N} \lambda_r \left[\frac{\partial^2 f_r}{\partial \theta^2} + \left(\frac{\partial f}{\partial \theta}\right)^T (I-J)^{-T}\frac{\partial^2 f_r}{\partial h_* \partial \theta} + \left(\frac{\partial^2 f_r}{\partial h_* \partial \theta}\right)^T (I-J)^{-1}\frac{\partial f}{\partial \theta}\right. \\
&\left. + \left(\frac{\partial f}{\partial \theta}\right)^T (I-J)^{-T}\frac{\partial^2 f_r}{\partial h_*^2}(I-J)^{-1}\frac{\partial f}{\partial \theta}\right]
\end{aligned}
$$

Factored Form

We can factor out the sensitivity terms to group the curvatures. Let $H^{(\lambda)}_{\theta\theta}$, $H^{(\lambda)}_{h\theta}$, and $H^{(\lambda)}_{hh}$ be the weighted sums of the Hessians of f_r (e.g., $H^{(\lambda)}_{hh} = \sum \lambda_r \frac{\partial^2 f_r}{\partial h_*^2}$).

We can combine the Loss Hessian with the state-state Dynamics Hessian:

$$
\begin{aligned}
H =& H^{(\lambda)}_{\theta\theta} f \\
&+ \left(\frac{\partial f}{\partial \theta}\right)^T (I-J)^{-T}\left(\nabla^2_{h_*} L + H^{(\lambda)}_{hh} f\right)(I-J)^{-1}\frac{\partial f}{\partial \theta} \\
&+ \left(\frac{\partial f}{\partial \theta}\right)^T (I-J)^{-T} H^{(\lambda)}_{h\theta} f + H^{(\lambda)}_{\theta h} f (I-J)^{-1}\left(\frac{\partial f}{\partial \theta}\right)
\end{aligned}
$$