Derivation of the Penalty method Hessian

Curt Da Silva, March 2015

Penalty method Gradient and Hessian

The penalty method can be written as

$$\phi_\lambda(m,u) = \frac{1}{2}\sum_{k,l} \|Pu_{k,l} - d_{k,l}\|_2^2 + \frac{\lambda^2}{2}\|H_k(m)u_{k,l} - q_{k,l}\|_2^2$$

where k indexes the frequency and l indexes the source position.

We let $\psi(m) = \min_u \phi_\lambda(m,u)$ and the optimal $u(m)$ solving this equation arises from solving

$$\nabla_u \phi_\lambda(m,u) = P^T(Pu-d) + \lambda^2 H(m)^H(H(m)u-q) = 0 \tag{1}$$

which is $u(m) = (\lambda^2 H(m)^H H(m) + P^T P)^{-1}(\lambda^2 H(m)^H q + P^T d)$. Here we suppress the dependence on k, l just for notational simplicity.

By the paper of Aravkin et al. "Estimating nuisance parameters in inverse problems", the expression for the penalty method gradient and hessian are

$$\begin{aligned} \nabla_m \psi(m) &= \nabla_m \phi_\lambda(m, u(m)) \\ \nabla_m^2 \psi(m)[\delta m] &= \nabla_m^2 \phi_\lambda(m,u(m))[\delta m] + \nabla_{m,u}\phi_\lambda(m,u(m))[Du(m)[\delta m]] \end{aligned} \tag{2}$$

Let us compute the expressions for $\nabla_m \psi(m)$ and $\nabla_m^2 \psi(m)[\delta m]$ in (2).

We use the notation * to denote an operator transpose and a ^{H} to denote standard, matrix hermitian transpose.

The gradient of $\phi_\lambda(m,u)$ with respect to m is

$$\nabla_m \phi_\lambda(m,u) = \lambda^2 (D(H(m))[\cdot])^*(H(m)u-q). \tag{3}$$

When $H(m) = A\text{diag}([Bf(m)]^2) + L$, where $f(m)$ is a pointwise applied function and A, B, L are constant matrices, then

$$D(H(m))[\delta m]y = A\text{diag}(2[Bf(m)] \odot [Bdf(m)] \odot \delta m)y. \tag{4}$$

Let $M = \text{diag}(2[Bf(m)] \odot [Bdf(m)])$, so $DH(m)[\delta m]y = AM\text{diag}(\delta m)y = AM\text{diag}(y)\delta m$

The adjoint operator, $D(H(m))[\cdot]^*$, acting on a vector Z is

$$D(H(m))[\cdot]^* Z = \text{diag}(\overline{y})\overline{M}A^H Z. \tag{5}$$

So the expression for (3) is

$$\nabla_m \psi(m) = \lambda^2 \text{diag}(\overline{u})\overline{M}A^H(H(m)u-q) \tag{6}$$

To compute $\nabla_m^2 \phi_\lambda(m,u(m))[\delta m]$, we differentiate (6) in the direction δm, with u fixed, which gives

$$\nabla_m^2 \phi_\lambda(m,u)[\delta m] = \lambda^2 \text{diag}(\overline{u})\overline{M}A^H AM\text{diag}(u)\delta m \tag{7}$$

Let us compute $\nabla_{m,u}\phi_\lambda(m,u(m))[Du(m)[\delta m]]$ by first computing $Du(m)[\delta m]$. We set $G(m) := (\lambda^2 H(m)^H H(m) + P^T P)$ and $r(m) := \lambda^2 H(m)^H q + P^T d$ for notational simplicity, so $u(m) = G(m)^{-1} r(m)$.

We differentiate $u(m)$ in the direction δm as

$$D(u(m))[\delta m] = D(G(m)^{-1})[\delta m] r(m) + G(m)^{-1} D(r(m))[\delta m] \tag{8}$$

which simplifies to

$$D(u(m))[\delta m] = -G(m)^{-1}[D(G(m))[\delta m]]G^{-1}(m)r(m) + \lambda^2 G(m)^{-1}(D(H(m)[\delta m])^H q)$$

and $D(G(m))[\delta m] = \lambda^2(D(H(m))[\delta m]^H H[m] + H[m]^H D(H(m))[\delta m])$.

Note that we can simplify (8) as

$$\begin{aligned} D(u(m))[\delta m] &= G(m)^{-1}\{\lambda^2 DH(m)[\delta m]^H(q) - \lambda^2 DH(m)[\delta m]^H H(m)u(m) - \lambda^2 H(m)^H DH(m)[\delta m]u \\ &= \lambda^2 G(m)^{-1}\{DH(m)[\delta m]^H(q - H(m)u(m)) - H(m)^H DH(m)[\delta m]u(m)\} \end{aligned}$$

so that we only have to perform another application of $G(m)$ if we've already computed $\tilde{u}(m)$.

Then we compute $\nabla_{m,u}\phi_\lambda(m,u(m))[\delta u]$, with $\delta u = Du(m)[\delta m]$, by differentiating (6) in the direction δu, which gives

$$\nabla_{m,u}\phi_\lambda(m,u(m))[\delta u] = \lambda^2 \text{diag}(\overline{\delta u})\overline{M}A^H(H(m)u(m) - q) + \lambda^2 \text{diag}(\overline{u(m)})\overline{M}A^H(H(m)\delta u).$$

In summary, the relevant expressions for the reduced gradient and hessian of $\psi(m)$ are

$$\begin{aligned} H(m) &= A\text{diag}([Bf(m)]^2) + L \\ M &= \text{diag}(2[Bf(m)] \odot [Bdf(m)]) \\ DH(m)[\delta m]y &= AM\text{diag}(y)\delta m \\ r_{PDE} &:= H(m)u(m) - q \\ G(m) &:= (\lambda^2 H(m)^H H(m) + P^T P) \\ u &= G(m)^{-1}(\lambda^2 H(m)^H q + P^T d) \\ \delta u &:= Du(m)[\delta m] = \lambda^2 G(m)^{-1}\{-DH(m)[\delta m]^H r_{PDE} - H(m)^H DH(m)[\delta m]u(m)\} \\ \nabla_m \psi(m) &= \lambda^2 \text{diag}(\overline{u})\overline{M}A^H r_{PDE} \\ \nabla_m^2 \psi(m)[\delta m] &= \lambda^2 \text{diag}(\overline{u})\overline{M}A^H DH(m)[\delta m]u \\ &\quad + \lambda^2 \text{diag}(\overline{\delta u})\overline{M}A^H r_{PDE} \\ &\quad + \lambda^2 \text{diag}(\overline{u})\overline{M}A^H(H(m)\delta u) \end{aligned}$$

An interesting remark by Zhilong is that in the expression for $\nabla_m \psi(m)$, by manipulating the equation $G(m)u(m) = r(m)$, we get that

$$\begin{aligned} \lambda^2 H(m)^H H(m)u + P^T Pu &= \lambda^2 H(m)^H q + P^T d \\ \Rightarrow r_{PDE} := H(m)u - q &= -\frac{1}{\lambda^2}(H(m)^{-H}P^T(Pu-d)) \end{aligned}$$

Plugging this expression in to $\nabla_m \psi(m)$ gives

$$\nabla_m \psi(m) = -\text{diag}(\overline{u})\overline{M}A^H(H(m)^{-H}P^T(Pu-d)),$$

which is exactly the expression for the FWI gradient, ableit with a u that solves the data-augmented equation instead of the standard helmholtz equation. Note that $\nabla_m \phi(m) = O(1)$ as $\lambda \to \infty$.

This also helps us analyze the terms in the Hessian of $\psi(m)$ as $\lambda \to \infty$, since

$$\begin{aligned} r_{PDE} &= O(\lambda^{-2}) \\ u &= O(1) \\ \lambda^2 G(m)^{-1} &= O(1) \\ Du(m)[\delta m] &= O(1) \\ \lambda^2 \text{diag}(\overline{u})\overline{M}A^H DH(m)[\delta m]u &= O(\lambda^2) \\ \lambda^2 \text{diag}(\overline{\delta u})\overline{M}A^H r_{PDE} &= O(1) \\ \lambda^2 \text{diag}(\overline{u})\overline{M}A^H(H(m)\delta u) &= O(\lambda^2) \end{aligned}$$

So the Hessian becomes ill conditioned as $\lambda \to \infty$, as you'd expect from a penalty method.

Gauss-Newton Hessian

If we want to write the reduced penalty method objective as a nonlinear least squares function, this is

$$\phi(m) = \frac{1}{2}\|F(m)\|_2^2$$

where $F(m) = \begin{bmatrix} Pu(m) - D \\ \lambda(H(m)u(m) - q) \end{bmatrix}$.

The Jacobian of $F(m)$ in the direction δm is

$$DF(m)[\delta m] = \begin{bmatrix} PDu(m)[\delta m] \\ \lambda(DH(m)[\delta m]u(m) + H(m)Du(m)[\delta m]) \end{bmatrix}$$

$Du(m)[\delta m]$ is computed above in (8). The adjoint of the operator $Du(m)[\cdot]$ acting on the vector Z is

$$Du(m)[\cdot]^* Z = \text{diag}(\overline{\overline{M}A^H(q - Hu)})V - \text{diag}(\overline{M \odot u})A^H H(m)V. \tag{9}$$

Here $V = \lambda^2 G(m)^{-1}Z$. We have also suppressed the dependence on m here, because the notation is bad enough as is.

Therefore the adjoint of $DF(m)[\cdot]$ applied to the block vector $\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix}$ is

$$Du(m)[\cdot]^*(P^T Z_1 + \lambda H(m)^H Z_2) + (DH(m)[\cdot]u(m))^*(\lambda Z_2).$$

The product of the Gauss-Newton Hessian with a vector, $H_{GN}v = DF(m)^*(DF(m)[v])$ can be derived by placing the output of $DF(m)[v]$ in to the input of $DF(m)^*(\cdot)$. These results can be extended to varying sources and frequencies by summing over the results for each source and frequency.