Regularization by Model Reparameterization

William S. Harlan

August, 1995

Geophysical inversion frequently makes use of regularization, such as the "Tikhonov regularization" used by Kenneth Bube and Bob Langan [1] for their "continuation approach." I'd like to suggest an adjustment of the objective function to allow faster convergence of regularization and the continuation approach. A damping term that discourages complexity can be replaced equivalently by a change of variables to model simplicity directly.

For an optimized inversion, an objective function typically includes a norm of the difference between a data vector $\mathbf{d}$ and a non-linear transform $\mathbf{f}(\mathbf{m})$ of a model vector $\mathbf{m}$. The global minimum of this norm is often flat, with little sensitivity to large variations in the model.

For regularization (more than simple damping), a linear operator $\underset{\sim}{\mathbf{D}}$ is chosen to remove simplicity and preserve complexity when applied to the model vector as $\underset{\sim}{\mathbf{D}} \cdot \mathbf{m}$. Most examples use a roughening operator, such as a derivative, to suppress long wavelengths and amplify short wavelengths. A regularized objective function adds a norm of this roughened model to the norm fitting the data:

$$\min_{\mathbf{m}} \quad J_1(\mathbf{m}) = \|\mathbf{d} - \mathbf{f}(\mathbf{m})\|^2 + c\|\underset{\sim}{\mathbf{D}} \cdot \mathbf{m}\|^2. \tag{1}$$

This particular l^2 objective function is easily motivated as a maximum a posteriori estimate of the model given the data. Additive noise is assumed to be Gaussian and uncorrelated with zero mean. The model is assumed to be Gaussian and zero mean, with an inverse covariance matrix equal to

$$\underset{\sim}{\mathbf{C}}_m^{-1} \equiv E(\mathbf{mm}^*)^{-1} = \underset{\sim}{\mathbf{D}}^* \cdot \underset{\sim}{\mathbf{D}}. \tag{2}$$

Asterisks indicate adjoints. The assumption that model samples are correlated is equivalent to the encouragement of simplicity. A constant c adjusts the ratio of variances assumed for noise and the model.

Bube and Langan's continuation approach begins with a large constant c, minimizes the objective function (1) for a first model, then reduces c repeatedly for a tradeoff between simplicity and accuracy in fitting the recorded data. They find the simplest model possible to explain the data adequately, without preventing the model from using complexity to fit genuinely significant features

of the data. Informative details are added to the model when justified by the data, without unnecessary distracting details that are poorly determined from the data.

Each minimization of the objective function (1) for a fixed constant c typically uses a descent method such as Gauss-Newton with conjugate gradients. The properties of the gradient are important to the rate of convergence:

$$\boldsymbol{\nabla_m} J_1(\mathbf{m}_0) \;=\; \boldsymbol{\nabla} \mathbf{f}(\mathbf{m}_0)^* \cdot [\mathbf{d} - \mathbf{f}(\mathbf{m}_0)] \tag{3}$$
$$+\;\; c\underset{\sim}{\mathbf{D}}^* \cdot \underset{\sim}{\mathbf{D}} \cdot \mathbf{m}_0. \tag{4}$$

The model is perturbed with scaled sums of successive gradients, evaluated for different reference versions $\mathbf{m}_0$ of the model. The first term (3) is able to introduce fairly arbitrary complexity into the model immediately and at any time, even if such complexity will be suppressed at the global minimum of objective function (1). The second term (4) must wait until the reference model $\mathbf{m}_0$ has been revised in later iterations to suppress this unnecessary complexity. Meanwhile, the first term (3) of later iterations can continue to introduce other unnecessary complexity into the model. The second term removes complexity in the reference model, not in the current perturbation. Convergence is slow. Slow convergence is a natural consequence of applying perturbations which do not have any of the correlations assumed for the model samples. Instead, let us introduce the appropriate correlation into all gradient perturbations.

Assume a new operator $\underset{\sim}{\mathbf{S}}$ as a partial right inverse of $\underset{\sim}{\mathbf{D}}$, so that the two operators approximate an identity: $\underset{\sim}{\mathbf{D}} \cdot \underset{\sim}{\mathbf{S}} \approx \underset{\sim}{\mathbf{I}}$. This operator should be designed to preserve simplicity and suppress complexity, although without destroying complexity entirely. If $\underset{\sim}{\mathbf{D}}$ is a roughening operator like differentiation, then $\underset{\sim}{\mathbf{S}}$ should be a smoothing operator like leaky integration.

More directly, define the smoothing operator as a factored form of the assumed covariance. (Indeed, such a factorization always exists because the covariance is positive semidefinite.)

$$\underset{\sim}{\mathbf{C}}_m \equiv E(\mathbf{mm}^*) = \underset{\sim}{\mathbf{S}}^* \cdot \underset{\sim}{\mathbf{S}}. \tag{5}$$

Minimization of the original objective function (1) is entirely equivalent to minimizing the objective function with a new variable $\mathbf{m}'$, where $\mathbf{m} = \underset{\sim}{\mathbf{S}} \cdot \mathbf{m}'$:

$$\min_{\mathbf{m}'} \quad J_1(\underset{\sim}{\mathbf{S}} \cdot \mathbf{m}') = J_2(\mathbf{m}') = \|\mathbf{d} - \mathbf{f}(\underset{\sim}{\mathbf{S}} \cdot \mathbf{m}')\|^2 + c\|\mathbf{m}'\|^2. \tag{6}$$

The second term reduces to a simple damping norm, demonstrating that the new model $\mathbf{m}'$ now has uncorrelated samples. Although we optimize this new model $\mathbf{m}'$, we keep and use the original model $\mathbf{m} = \underset{\sim}{\mathbf{S}} \cdot \mathbf{m}'$. Continuation can adjust the constant c as before, with identical results (assuming complete minimization of the objective functions [1] and [6]).

The revised gradient contains the desired correlation:

$$\boldsymbol{\nabla}_{\mathbf{m}'} J_2(\mathbf{m}'_0) = \underset{\sim}{\mathbf{S}}^* \cdot \boldsymbol{\nabla}\mathbf{f}(\underset{\sim}{\mathbf{S}} \cdot \mathbf{m}'_0)^* \cdot [\mathbf{d} - \mathbf{f}(\underset{\sim}{\mathbf{S}} \cdot \mathbf{m}'_0)] \qquad (7)$$

$$+ \quad c\mathbf{m}'_0 \qquad (8)$$

The last operation appearing in the first term of this gradient (7) is the adjoint $\underset{\sim}{\mathbf{S}}^*$ of the operator $\underset{\sim}{\mathbf{S}}$, both of which are simplification operators. (Many such operators are self-adjoint.) Unlike the first term (3) of the original gradient, the revised term (7) suppresses complexity from each new perturbation direction. The original term (3) contained arbitrary correlations. If (3) were entirely uncorrelated, then the revised term (7) would have exactly the desired correlations assumed by the covariance (5).

The two objective functions produce different results when optimization is incomplete. A descent optimization of the original objective function (1) will begin with complex perturbations of the model and slowly converge toward an increasingly simple model at the global minimum. A descent optimization of the revised objective function (6) will begin with simple perturbations of the model and slowly converge toward an increasingly complex model at the global minimum. The latter strategy is more consistent with the overall goal of the continuation approach. A more economic implementation can use fewer iterations. Insufficient iterations result in an insufficiently complex model, not in an insufficiently simplified model.

I also prefer to adjust more than a single scale factor c. Instead, assume a suite of simplification operators $\mathbf{S_i}$ which allow increasing complexity as the index i increases. (Furthermore $\forall \tilde{\mathbf{m}}'_i$ and $j > i, \exists\; \mathbf{m}'_j \ni \mathbf{S_j} \cdot \mathbf{m}'_j = \underset{\sim}{\mathbf{S_i}} \cdot \mathbf{m}'_i.$) We then can optimize a suite of possible models, $\{\mathbf{m}_i = \underset{\sim}{\mathbf{S_i}} \cdot \mathbf{m}'_i\}$ of increasing complexity as i increases. Use each optimized model $\mathbf{m}_i$ to initialize the next $\mathbf{m}_{i+1}$. As multigrid methods have shown, we can thus improve our overall convergence by optimizing the most reliable (smoothest) global features in the model before attempting finer detail.

Finally, I think it easier to choose a simplification operator $\underset{\sim}{\mathbf{S}}$ which describes the desirable features of the model, rather than an operator $\underset{\sim}{\mathbf{D}}$ which keeps only features thought to be undesirable. I see some value in constructing both, however, to check the consistency of assumptions.

References

[1] Kenneth P. Bube and Robert T. Langan. A continuation approach to regularization for traveltime tomography. In *64th Ann. Internat. Mtg., Expanded Abstracts*, pages 980–983. Soc. Expl. Geophys., 1994.