DOI:10.1190/segam2020-3425521.1

# Anderson Acceleration for Seismic Inversion

*Yunan Yang, Courant Institute of Mathematical Sciences, New York University.*


## SUMMARY

Full waveform inversion (FWI) and least-squares reverse time migration (LSRTM) are popular imaging techniques that can be solved as PDE-constrained optimization problems. Due to the large-scale nature, gradient- and Hessian-based optimization algorithms are preferred in practice to find the optimizer iteratively. However, a balance between the evaluation cost and the rate of convergence needs to be considered. We propose the use of Anderson acceleration (AA), a popular strategy to speed up the convergence of fixed-point iterations, to accelerate a gradient descent method. We show that AA can achieve fast convergence that provides competitive results with some quasi-Newton methods. Independent of the dimensionality of the unknown parameters, the computational cost of implementing the method can be reduced to an extremely low-dimensional least-squares problem, which makes AA an attractive method for seismic inversion.


## INTRODUCTION

The fast growth of computational power popularizes numerous techniques that utilize the full wavefields in seismic imaging (Tarantola and Valette, 1982). In particular, two optimization-based imaging techniques, full-waveform inversion (FWI) (Virieux and Operto, 2009) and least-squares reverse-time migration (LSRTM) (Dai and Schuster, 2013), seek a quantity of interest in an iterative fashion by successively searching in the direction of the gradient (Métivier et al., 2012). In each iteration, a line search is applied to select a step size, which can guarantee a sufficiently good decrease in the objective function (by the Armijo rule) and a sufficiently reduced slope (by the Wolfe condition). However, the computational cost of the exact (or inexact) line search can be much larger than the cost of gradient evaluation. In a practical inversion application, these issues often motivate us to avoid second-order methods or fix a step size, which can result in slow convergence.

In this paper, we aim to combine the fast convergence of second-order optimization methods with the low cost of evaluating the gradient. We do this by applying an acceleration strategy introduced by D.G. Anderson (1965) to the steepest descent algorithm. In contrast to the Picard iteration, which uses only one previous iterate, the method proceeds by linearly recombining a list of previous iterates in a way that approximately minimizes the linearized residual. In the last two decades, the so-called Anderson acceleration (AA) has been widely used in several applied fields for problems that can be regarded as involving a fixed-point iteration. applications include flow problems (Pollock et al., 2018), numerical optimization (Peng et al., 2018; Fu et al., 2019), machine learning (Geist and Scherrer, 2018) and wave propagation (Yang et al., 2020). The literature on this subject is broad, so we only mention a few papers to show the variety of results obtained by the AA.

The main contribution of the paper is to apply AA to seismic inversion to improve the convergence of the steepest descent method. We first reformulate the iterative formula as a fixed-point operator. The fixed-point solution is then the model parameter where the gradient is evaluated as zero. AA is applied to produce every new iterate as a linear combination of several previous iterates. The linear coefficients are selected optimally to achieve the best (linearized) residual reduction. Overall, AA speeds up the convergence of the steepest descent method and reduces the computational cost of computing the exact Hessian or building an approximation of the (inverse) Hessian matrix.

The paper is arranged as follows. We first introduce the theory of AA and then provides details on how to integrate it with seismic inverse problems, including FWI and LSRTM. We show the effectiveness of the new acceleration strategy by showing numerical examples on the reconstruction of wave speed in FWI and model reflectivity in LSRTM. Several comparisons of the performance in terms of computational cost and convergence rate between AA and other optimization schemes demonstrate the potential of this new strategy for optimization-based techniques for seismic imaging.

## THEORY

In this section, we first review the essential background of the gradient calculation for the large-scale inverse problem and then introduce the algorithmic details of AA.

### Gradient calculation in seismic inversion

In seismic inversion and some PDE-constrained optimization problems, one can obtain the gradient of a parameter on the entire domain by one forward, and one adjoint wave solves, based on the adjoint-state method (Plessix, 2006), independent of the dimensionality of the parameter. The PDE constraint of FWI is the wave equation (1), where $m_0$ is the velocity parameter that we aim to reconstruct, $u$ is the solution to the wave equation on the temporal domain $[0,T]$ and spatial domain $\Omega$. The source term $s$ is known.

$$\begin{cases} m_0(\mathbf{x})\dfrac{\partial^2 u(\mathbf{x},t)}{\partial t^2} - \triangle u(\mathbf{x},t) = s(\mathbf{x},t), \\ u(\mathbf{x},0) = 0, \quad \dfrac{\partial u}{\partial t}(\mathbf{x},0) = 0. \end{cases} \tag{1}$$

The gradient for FWI is then

$$\left.\frac{\partial J_{FWI}}{\partial m}\right|_{m=m_0} = -\int_0^T \frac{\partial^2 u(\mathbf{x},t)}{\partial t^2} v(\mathbf{x},t)\,dt, \tag{2}$$

where $v$ solves the adjoint wave equation in (3). Here, the source term is the Fréchet derivative of the objective function $J_{FWI}$ with respect to the synthetic data $f = Ru$, where $R$ is the projection operator onto the receiver locations.

$$\begin{cases} m_0(\mathbf{x})\frac{\partial^2 v(\mathbf{x},t)}{\partial t^2} - \triangle v(x,t) = -R^T \dfrac{\partial J}{\partial f}, \\ v(\mathbf{x},T) = 0, \quad \dfrac{\partial v}{\partial t}(\mathbf{x},T) = 0. \end{cases} \tag{3}$$



10.1190/segam2020-3425521.1

For LSRTM, the forward modeling becomes the linearized wave equation (4), which is derived by the first-order Born approximation. The solution to (4) is the scattered wavefield $u_1$. The background velocity $m_0$, and the incident wavefield $u_0$ are known. The reflectivity (perturbed velocity) $m_1$ is the value of interest for reconstruction in LSRTM.

$$\begin{cases} m_0(\mathbf{x})\dfrac{\partial^2 u_1(\mathbf{x},t)}{\partial t^2} - \triangle u_1(\mathbf{x},t) = -m_1(\mathbf{x})\dfrac{\partial^2 u_0(\mathbf{x},t)}{\partial t^2}, \\ u_1(\mathbf{x},0) = 0, \quad \dfrac{\partial u_1}{\partial t}(\mathbf{x},0) = 0. \end{cases} \tag{4}$$

The model gradient of the LSRTM objective function can be obtained in a similar manner with the FWI gradient; see Dai and Schuster (2013) for more details. In this paper, we use the $\ell^2$ norm as the objective function for both FWI and LSRTM.

**Algorithm 1** Anderson Acceleration

Given $p_0$ and $m \geq 1$.
Set $p_1 = G(p_0)$.
**for** k = 0, 1,2,... **do**
Set $m_k = \min(m,k)$.
Set $F_k = (f_{k-m_k}, \ldots, f_k)$, where $f_i = G(p_i) - p_i$ is the residual at the $i$-th iteration.
Find $\alpha^{(k)} = (\alpha_0^{(k)}, \ldots, \alpha_{m_k}^{(k)})^T$ to minimize

$$\|F_k \alpha^{(k)}\|, \text{ where } \sum_{i=0}^{m_k} \alpha_i^{(k)} = 1. \tag{5}$$

Update $p_{k+1}$ according to

$$p_{k+1} = (1-\beta_k)\sum_{i=0}^{m_k} \alpha_i^{(k)} p_{k-m_k+i} + \beta_k \sum_{i=0}^{m_k} \alpha_i^{(k)} G(p_{k-m_k+i}).$$

**end for**

**Anderson acceleration**

We start by stating AA in Algorithm 1. The memory parameter $m$ determines the number of previous iterates we use to compute the next iterate. For example, when $m = 0$, AA reduces to the Picard iteration as the $k+1$-th iterate only depends on the $k$-th iterate. That is, $p_{k+1} = G(p_k)$, where $G$ is the fixed-point operator. For a nonzero $m$, $p_{k+1}$ is a linear combination of the previous $m+1$ iterates, together with their evaluation by the fixed-point operator $G$. If $m = +\infty$ and $\beta_k$ is also chosen optimally at each iteration, then AA is essentially equivalent to the generalized minimal residual method (GMRES) (Toth and Kelley, 2015). The damping parameter $\beta_k$ controls the weights between the linear combination of the iterates $\{p_{k-m+i}\}_{i=0}^m$ and the linear combination of their evaluation by the operator $\{G(p_{k-m+i})\}_{i=0}^m$. The mixing parameter can be adaptively selected and could vary on different iterations.

At each iteration, $k$, the weighting vector $\alpha^{(k)}$ for the linear combination are determined by the constrained optimization in (5). By a change of variable, we can remove the constraints to simplify the optimization. Consider the vector

$$\gamma^{(k)} = (\gamma_0^{(k)}, \ldots, \gamma_{m_k-1}^{(k)})^T,$$

which depends on vector $\alpha^{(k)} = (\alpha_0^{(k)}, \ldots, \alpha_{m_k}^{(k)})^T$ as follows,

$$\gamma_i^{(k)} = \alpha_0^{(k)} + \cdots + \alpha_i^{(k)}, \quad 0 \leq i \leq m_k - 1. \tag{6}$$

If we also choose the $\ell^2$ norm in (5), $\gamma^{(k)}$ is then the least-squares solution to the following linear system:

$$A_k \gamma^{(k)} = f_k, \tag{7}$$

where $f_k = G(p_k) - p_k$ is the residual at iteration $k$. Here, $A_k$ is the matrix given by

$$A_k = (f_{k-m_k+1} - f_{k-m_k}, \ldots, f_k - f_{k-1}), \tag{8}$$

whose column vectors are the differences in residual between consecutive iterations. If $\beta_k = 1$ for any $k$, then we can rewrite the updating formula in terms of $\gamma^{(k)}$, i.e.,

$$p_{k+1} = G(p_k) - \sum_{i=0}^{m_k-1} \gamma_i^{(k)} \left[G(p_{k-m_k+i+1}) - G(p_{k-m_k+i})\right]. \tag{9}$$

AA is computationally efficient to implement with the $\ell^2$ norm. The matrix $A_k$ in (8) is an $n$ by $m$ matrix, where $n$ is the dimension of the parameter (i.e., the number of unknown variables), and the memory parameter $m$ is often chosen to be no more than 20. A fast rank-updated QR factorization can further reduce the cost of solving (7).

## METHOD

Within the framework of iterative methods, we consider the gradient descent algorithm as a type of fixed-point operator. The corresponding fixed-point solution is what we aim to reconstruct in seismic inversion, provided the initial model is good enough. By combining both the fixed-point operator and the objective function, AA outperforms the steepest descent method and meanwhile saves the computational cost of high-order methods that involve Hessian approximation.

**Algorithm 2** $\ell^2$-based AA for gradient descent

**Input:** Given the initial model parameter $p_0$, memory parameter $m \geq 1$, and the fixed-point operator $G$ (10).
Set $p_1 = G(p_0)$.
**for** $k = 1, 2, \ldots$ until convergence or maximum iteration **do**
**Step 1:** Set $m_k = \min(m,k)$ and $A_k$ following (8).
**Step 2:** Solve the least-squares problem:
$\gamma^{(k)} = \text{argmin}_\gamma ||A_k\gamma - f_k||_2, \quad \gamma^{(k)} = (\gamma_0^{(k)}, \ldots, \gamma_{m_k-1}^{(k)})^T$.
**Step 3:** Compute the new iterate $\widetilde{p}_{k+1}$ following (9).
**Step 4:** Backtracking line search for optimal step size $\alpha$.
**Step 5:** Set $p_{k+1} = \alpha\widetilde{p}_{k+1} + (1-\alpha)G(p_k)$.
**end for**

**The fixed-point operator**

Consider an objective function $J(p)$, where $p$ represents the unknown parameter in seismic inverse problems. The adjoint-state method offers an effective way to evaluate the model gradient by simply two wave equation solves. We are going to form a fixed-point operator based on the gradient.

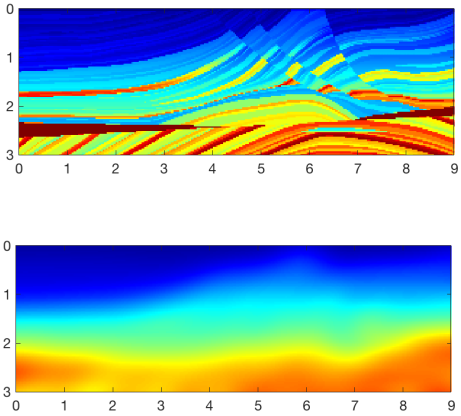
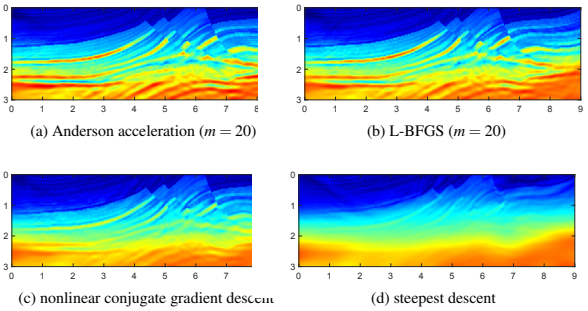


Figure 1: Marmousi model true and initial velocity.



(a) Anderson acceleration ($m = 20$) (b) L-BFGS ($m = 20$)

(c) nonlinear conjugate gradient descent (d) steepest descent

Figure 2: FWI results using AA, L-BFGS, nonlinear CG and steepest descent after $10^3$ gradient evaluations.

We are interested in minimizing the objective function $J(p)$ by the steepest descent method. The $(k+1)$-th iterate $p_{k+1}$ is obtained by the $k$-th iterate $p_k$ and the model gradient at $p_k$ with an adaptive step size $\eta_k$ that ensures a sufficient decrease in the objective function:

$$p_{k+1} = p_k - \eta_k \frac{\partial J}{\partial p}\bigg|_{p=p_k} = G(p_k). \tag{10}$$

To reduce the cost of line search, $\eta_k$ can be fixed as a constant that is small enough to guarantee the misfit decrease.

Note that the right-hand side of (10) only depends on one previous iterate $p_k$, which shares a similar property with the Picard iteration for fixed-point operators. Therefore, one can regard the iterating scheme (10) as a fixed-point iteration with $G$ being the fixed-point operator. The fixed-point solutions are then the parameters whose model gradient is zero.

Seismic inverse problems are often ill-posed, especially for FWI. Therefore, zero gradients are not enough to guarantee the global optimality of the solution. There has been extensive literature on how to mitigate the existence of local minima (Engquist and Yang, 2020; Symes, 2020). In this paper, we choose to focus on the acceleration of the convergence and not to address the cycle-skipping issues. Therefore, we assume that the initial model $p_0$ is appropriately chosen, and the nearest critical point $p^*$ where $p^* = G(p^*)$ is the optimal solution of the inverse problems. Under this (strong) assumption, the gradient-based fixed-point operator $G$ is *contractive*. We aim to find $p^*$ as fast as possible in terms of the total CPU cost (instead of the number of iterations).

**The objective function**

We have successfully translated the optimization scheme into a fixed-point problem and obtained the fixed-point operator $G$ defined in (10). Different from typical fixed-point problems where residual is the only indicator of convergence, we can also utilize our objective function in the optimization to accelerate the convergence of the model parameters further.

By combining both the residual and the objective function, we describe the new workflow of AA in Algorithm 2. We reformulate Steps 1-3 to remove the constraints in the weights optimization and set $\beta_k = 1$. As discussed earlier, we employ the $\ell^2$ norm to measure the residual in (5) for its optimization advantages over the $\ell^1$ or the $\ell^\infty$ norm. One can also use a weighted least-squares norm to enforce the bias towards certain modes of the solution. For example, AA based on the $\mathbf{H}^{-2}$ Sobolev norm is well-suited to solve fixed-point operators derived from second-order elliptic differential operators (Yang et al., 2020). Any $\mathbf{H}^s$-based residual can be optimized as a weighted least-square problem without incurring additional costs. A backtracking line search is added to the new workflow in Step 4 and Step 5 to ensure a sufficient decrease in the objective function $J$. The $(k+1)$-th iteration is set as a linear combination of the output by the gradient descent, $G(p_k)$, and the optimized new iterate by AA, $\widetilde{p}_{k+1}$; see Algorithm 2 for details.

## NUMERICAL EXAMPLE

The previous description of the algorithm and implementation illustrate the computational advantages of Anderson acceleration (AA) for large-scale seismic inverse problems. In this section, we present two particular applications of AA to demonstrate its fast convergence compared with the frequently used optimization algorithms in FWI and LSRTM.

**Full waveform inversion**

The goal of this experiment is to reconstruct the Marmousi velocity model (3 km in depth and 9 km in width) from a smoothed initial guess, as shown in Figure 1. We set 11 equally spaced sources at 150 m below the water surface, each of which is a Ricker wavelet centered at 15 Hz. The total recording time is 4 seconds. We remark that the initial model is selected to be "good" in the sense that it does not suffer from cycle-skipping issues. The convergence to the global minimum is guaranteed for all gradient-based methods.

Although it is feasible to recover the truth, the rate of convergence is our primary interest. The most time-consuming step of FWI is wave modeling, which is essential for gradient cal-

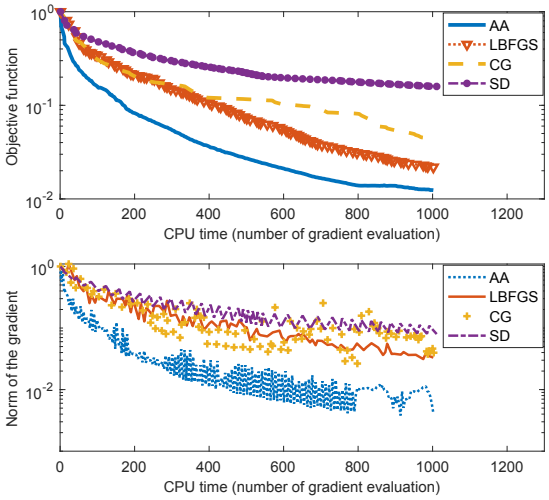



Figure 3: FWI convergence history in terms of computational time (measured by the number of FWI gradient evaluations).

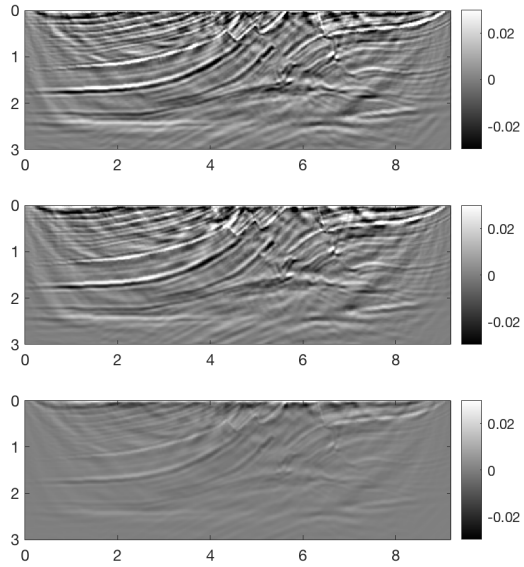



Figure 4: LSRTM results by AA, L-BFGS and gradient descent method (from top to bottom) after 10 iterations.

culation. Therefore, instead of using the *number of iteration* as a reference for comparison, we switch to the *number of gradient evaluation*, twice of which is the minimum number of wave modeling required by FWI.

After 1000 gradient evaluations, FWI results using AA ($m = 20$), L-BFGS ($m = 20$), nonlinear conjugate gradient (CG), and steepest descent are shown in Figure 2. Backtracking line search is used with the same set of parameters. Results by AA and L-BFGS have similarly good resolution, and unsurprisingly, steepest descent is the slowest in convergence. The semi-log plots for the convergence history in the relative $\ell^2$ objective function and the relative norm of the gradient in Figure 3 offer more quantitative information. In both plots, AA gives a faster convergence within fixed CPU time than L-BFGS and nonlinear CG. Known as quasi-Newton methods, L-BFGS and CG converge in a smaller number of iterations than AA. However, in each iteration, more gradient evaluations are required for line search to satisfy the Armijo-Goldstein and the Wolfe conditions. These extra gradient evaluations slow down the overall reconstruction. The comparison between the curves by AA and steepest descent in Figure 3 also illustrate the drastic improvement by the simple strategy of linearly combining a list of previous iterates for the next iteration. Considering the low cost of implementation, AA can be an attractive optimization technique for FWI.

**Least-squares reverse-time migration**

Our second example is to apply AA to LSRTM to speed up the convergence, which is to reduce the total number of required wave solves. We still use the Marmousi benchmark for illustration. The smoothed velocity model in Figure 1 is regarded as the known background velocity. The velocity perturbation $m_1$ in (4) (reflectivity) is the target of reconstruction.

Overall, 80 shots are used for this migration task. The source frequency is a Ricker wavelet centered at 25. AA ($m = 10$), L-BFGS ($m = 10$) and the steepest descent method are applied for the inversion, respectively. The entire workflow is the same as the FWI example, but with a linear wave modeling. Figure 4 shows the inversion results after 10 iterations. AA obtains a competitive migration image with the one by L-BFGS, while the image by the steepest descent method lacks the optimal resolution.

## CONCLUSION

Anderson acceleration for seismic inversion treats the method of steepest descent as a fixed-point operator. It speeds up the convergence by linearly combining a list of the previous iterates in an optimized way. AA can be considered as an alternative to L-BFGS but does not require either the storage or an approximation of the inverse Hessian at each iteration. The computational cost of implementing AA, which mainly comes from the 1D optimization for the weights, does not increase with the dimensionality of the unknown parameter. Therefore, AA is a computationally attractive method for seismic inversion, which can achieve competitive results with the widely used optimization algorithms such as L-BFGS and nonlinear CG, as demonstrated by our numerical examples.

## ACKNOWLEDGEMENTS

The author thanks the sponsors of the Texas Consortium for Computational Seismology (TCCS) for travel support. This work was also partially supported by NSF DMS-1913129.

## REFERENCES


Anderson, D. G., 1965, Iterative procedures for nonlinear integral equations: Journal of the ACM, **12**, 547–560.

Dai, W., and G. T. Schuster, 2013, Plane-wave least-squares reverse-time migration: Geophysics, **78**, no. 4, S165–S177, doi: https://doi.org/10.1145/321296.321305.

Engquist, B., and Y. Yang, 2020, Optimal transport based seismic inversion: Beyond cycle skipping: arXiv preprint arXiv:2002.00031.

Fu, A., J. Zhang, and S. Boyd, 2019, Anderson accelerated Douglas-Rachford splitting: arXiv preprint arXiv:1908.11482.

Geist, M., and B. Scherrer, 2018, Anderson acceleration for reinforcement learning: arXiv preprint arXiv:1809.09501.

Métivier, L., R. Brossier, J. Virieux, and S. Operto, 2012, The truncated Newton method for full waveform inversion: 82nd Annual International Meeting, SEG, Expanded Abstracts, 1–5, doi: https://doi.org/10.1190/segam2012-0981.1.

Peng, Y., B. Deng, J. Zhang, F. Geng, W. Qin, and L. Liu, 2018, Anderson acceleration for geometry optimization and physics simulation: ACM Transactions on Graphics, **37**, 1–14, doi: https://doi.org/10.1145/3197517.3201290.

Plessix, R. E., 2006, A review of the adjoint-state method for computing the gradient of a functional with geophysical applications: Geophysical Journal International, **167**, no. 2, 495–503, doi: https://doi.org/10.1111/j.1365-246X.2006.02978.x.

Pollock, S., L. G. Rebholz, and M. Xiao, 2019, Anderson-accelerated convergence of Picard iterations for incompressible Navier–Stokes equations: SIAM Journal on Numerical Analysis, **57**, 615–637.

Symes, W. W., 2020, Full waveform inversion by source extension: Why it works: arXiv preprint arXiv:2003.12538.

Tarantola, A., and B. Valette, 1982, Generalized nonlinear inverse problems solved using the least squares criterion: Reviews of Geophysics, **20**, 219–232, doi: https://doi.org/10.1029/RG020i002p00219.

Toth, A., and C. T. Kelley, 2015, Convergence analysis for Anderson acceleration: SIAM Journal on Numerical Analysis, **53**, 805–819, doi: https://doi.org/10.1137/130919398.

Virieux, J., and S. Operto, 2009, An overview of full-waveform inversion in exploration geophysics: Geophysics, **74**, no. 6, WCC1–WCC26, doi: https://doi.org/10.1190/1.3238367.

Yang, Y., A. Townsend, and D. Appelö, 2020, Anderson acceleration using the H-s norm: arXiv preprint arXiv:2002.03694.