

The Adaptive Variational Dual Method for Nonlinear Nonsquare Algebraic Equations

Amit Acharya* Soumya Kar† N. Sukumar‡

Abstract

We study the Adaptive Variational Dual (AVD) method as a direct solver for nonlinear nonsquare systems, $G(U) = 0$, $G : \mathbb{R}^n \rightarrow \mathbb{R}^m$, $m \neq n$. A central question is not whether least-squares optimization is effective as an optimization problem, but whether obtaining critical points of $\min_U \frac{1}{2} \|G(U)\|_2^2$ is a sufficiently reliable way to solve the original equations $G(U) = 0$. Another is to produce a first-order method with a convex variational structure for general systems of nonlinear equations, potentially useful for large systems of equations.

AVD retains the system $G(U) = 0$ as the primary object, introduces a variational dual formulation for it, and uses a sequence of convex optimization problems to solve it. In doing so, it adapts its computational mechanism to the local rectangular geometry through singular-value information. For overdetermined systems, structural dual null directions are removed from the active dual space; for underdetermined systems, primal nonuniqueness is distinguished from weak but genuinely active dual directions. The adaptive AVD scheme combines step-size controlled gradient flow and active-space Newton steps.

We present a series of numerical tests consisting of overdetermined, underdetermined, and square systems to compare the AVD scheme with nonlinear least squares. We show that the AVD method is consistently robust across overdetermined pathologies, unstable saddle selections, nonredundant neural network calibrations, and ill-conditioned systems, while least-squares often fails or terminates at non-root stationary states, operating within the standard protocol of finding critical points of the least squares functional. The numerical results indicate that the AVD method is a competitive candidate for solving nonlinear, nonsquare systems of equations.

1 Introduction

The numerical problem considered here is a nonlinear system of equations

$$G(U) = 0, \quad G : \mathbb{R}^n \rightarrow \mathbb{R}^m, \quad m \neq n. \quad (1)$$

Such systems arise in inverse problems, constrained model calibration, parameter identification, reduced-order models, and machine-learning parametrizations. When $m = n$, standard Newton and related root-finding methods have a well-developed theory [1, 2]. When $m \neq n$, the linear algebra is fundamentally rectangular [3]. A common response is to replace (1) by nonlinear least squares (see, e.g., [4]),

$$\Phi(U) = \frac{1}{2} \|G(U)\|_2^2, \quad \min_U \Phi(U), \quad (2)$$

*Department of Civil & Environmental Engineering, and Center for Nonlinear Analysis, Carnegie Mellon University, Pittsburgh, PA 15213, email: acharyaamit@cmu.edu.

†Department of Electrical & Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, email: soumyak@andrew.cmu.edu.

‡Department of Civil and Environmental Engineering, One Shields Avenue, University of California, Davis, CA 95616, email: nsukumar@ucdavis.edu.

and look for critical points of the problem as the practical alternative to solving the minimization problem. This is often computationally attractive, but it changes the problem. The stationarity condition is

$$\nabla\Phi(U) = G_U(U)^T G(U) = 0, \quad (3)$$

which does not imply $G(U) = 0$ whenever $G_U^T G = 0$ at a non-root stationary point or when the equation system is infeasible. Thus a successful optimization termination does not, by itself, guarantee a solution of (1).

The underlying dual variational construction of our approach follows [5, Sec. 2]; related developments in hidden convexity in PDE [6] and adaptive variational dual methods appear in [7, 8, 9, 10], while [11] provides a closely related algorithmic development. The present work specializes and extends these ideas to finite-dimensional, nonlinear nonsquare algebraic systems. The present manuscript asks a specific methodological question:

Is the Adaptive Variational Dual (AVD) method a better computational method for solving nonlinear nonsquare equation systems $G(U) = 0$ than replacing the equation problem by nonlinear least squares?

The answer cannot be reduced to a single ‘yes’ or ‘no’ answer. A fair comparison must distinguish feasible and infeasible systems, overdetermined and underdetermined geometry, benign and difficult nonlinear basins, and algorithmic termination from exact equation satisfaction.

The main contributions of the study are therefore practical and comparative. We formulate AVD for rectangular systems, explain the relevant singular-value geometry on linearization, document a robust adaptive scheme, and compare its behavior against least squares on a deliberately varied suite of tests.

2 The AVD formulation: from the primal equations to a dual parametrization

This section provides the theoretical basis of the AVD algorithm for nonsquare algebraic systems.

2.1 Primal equation, predual, and dual-to-primal map

The question throughout is to solve the stated primal problem

$$G(U) = 0, \quad G : \mathbb{R}^n \rightarrow \mathbb{R}^m. \quad (4)$$

Introduce a dual variable $D \in \mathbb{R}^m$, a base state $\bar{U} \in \mathbb{R}^n$, and a sufficiently smooth auxiliary potential $H(U; \bar{U})$ that is strictly convex in U and has its unique minimum at $U = \bar{U}$. The predual Lagrangian is

$$L(U, D; \bar{U}) = D \cdot G(U) - H(U; \bar{U}). \quad (5)$$

For a fixed choice of H and fixed $(D, \bar{U})$, the dual-to-primal (DtP) map $\mathcal{U}(D; \bar{U})$ is defined implicitly by stationarity of the predual with respect to the primal variable:

$$L_U(\mathcal{U}(D; \bar{U}), D; \bar{U}) = G_U(\mathcal{U}(D; \bar{U}))^T D - H_U(\mathcal{U}(D; \bar{U}); \bar{U}) = 0. \quad (6)$$

Thus, for fixed $\bar{U}$, (6) defines a local dual parametrization $D \mapsto \mathcal{U}(D; \bar{U})$ of primal states within the region of strict predual concavity; it is not itself the equation $G = 0$.

Substitution of the DtP map into the predual defines the reduced dual functional

$$\mathcal{S}(D; \bar{U}) = L(\mathcal{U}(D; \bar{U}), D; \bar{U}). \quad (7)$$

If different auxiliary potentials must be compared, we write $\mathcal{U}^{[H]}$ and $\mathcal{S}^{[H]}$; otherwise H is fixed and the label is suppressed. This avoids confusing dependence on the chosen potential with derivative subscripts. Because $L_U = 0$ on the DtP map, the chain rule gives

$$\mathcal{S}_D(D; \bar{U}) = L_D + \mathcal{U}_D^T L_U = G(\mathcal{U}(D; \bar{U})). \quad (8)$$

Consequently, stationarity of the reduced dual functional is exactly the original equation problem expressed through the dual parametrization:

$$\mathcal{S}_D(D; \bar{U}) = 0 \iff G(\mathcal{U}(D; \bar{U})) = 0. \quad (9)$$

This is the central variational identity: solving the dual stationarity equations recovers a solution of $G(U) = 0$, rather than merely a stationary point of $\frac{1}{2}\|G\|^2$. At $D = 0$, strict convexity and the choice of the minimum of H give $\mathcal{U}(0; \bar{U}) = \bar{U}$.

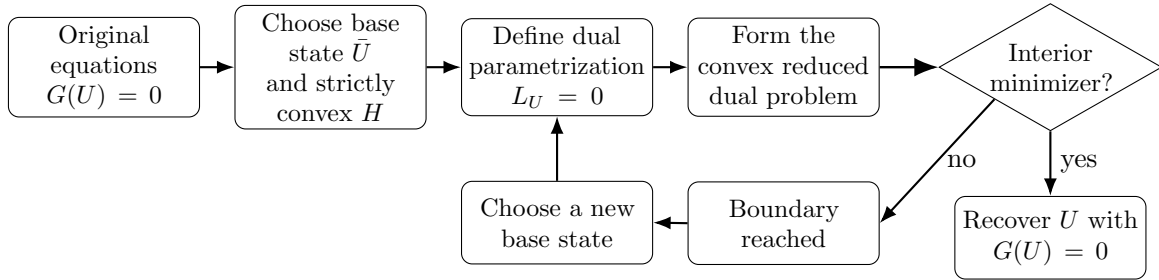


Figure 1: Conceptual structure of AVD. For a fixed base state, the DtP relation defines a dual parametrization of primal states in the region of strict predual concavity. An interior minimizer of the corresponding convex reduced dual problem recovers a solution of $G(U) = 0$. If the admissibility boundary is reached first, the base state is changed and a new dual parametrization is introduced.

In theory and practice, we go beyond stationarity. Since L is necessarily affine in D , by restricting the domain of dual variables to an open ball (and hence a convex set) around $D = 0$ in which L_{UU} is negative definite (and hence L strictly concave), the DtP solution in (6) becomes unique and the dual functional (7) convex in D . Thus, for every fixed H with its corresponding base state, (9) amounts to looking for a minimizer of a convex function. Then, one either finds a minimizer in the interior of the convex set in which case we have a solution, or the minimizer is at the boundary of the convex set. If the latter, one invokes a change of base state (along with its associated change in the auxiliary potential H) keeping $G(U)$ continuous across such switches, and repeats the process until a solution is found; Fig. 1 provides a conceptual flow-chart and [10, Sec. 3 and precursors] contains details of the scheme. Thus, our procedure amounts to working with a sequence of convex optimization problems, solved by a gradient flow (GF) and/or a Newton scheme in practice. In the case of a gradient flow, the L^2 norm of the gradient $\mathcal{S}_D = G$ cannot increase along such a flow, thus producing a sequence of real numbers bounded below by 0.

When can such a procedure stall on theoretically fundamental grounds before achieving a solution, i.e., a U such that $\|G(U)\| = 0$? This is clear from (6)—if the scheme reaches a state U for which $(G_U(U))^T = 0$ then there is no dual state D which can produce a solution other than $U = \bar{U}$, and no improvement is possible from $G(\bar{U}) \neq 0$. Essentially, if the scheme arrives at a

primal state U in the interior of a ball in primal space where $\|G\| \neq 0$ is constant, we have a stall. In a practical algorithm, this eventuality, if it occurs, is described in what follows as a circumstance when an ‘outer kick’ is invoked.

2.2 Derivative of the DtP map and the dual Jacobian

Define the primal Hessian of the predual along the DtP map by

$$K(U, D; \bar{U}) := L_{UU}(U, D; \bar{U}) = \sum_{i=1}^m D_i G_{i,UU}(U) - H_{UU}(U; \bar{U}). \quad (10)$$

Differentiating the defining equation (6) with respect to D gives

$$K \mathcal{U}_D + G_U^T = 0. \quad (11)$$

Whenever K is nonsingular, the implicit-function theorem therefore yields the derivative of the DtP map with respect to D , represented by the matrix

$$\mathcal{U}_D(D; \bar{U}) = -K^{-1} G_U^T, \quad (12)$$

where K and G_U are evaluated at $U = \mathcal{U}(D; \bar{U})$. The dual residual is $G(\mathcal{U}(D; \bar{U}))$, and its Jacobian is consequently

$$J := \frac{\partial}{\partial D} G(\mathcal{U}(D; \bar{U})) = G_U \mathcal{U}_D = -G_U K^{-1} G_U^T. \quad (13)$$

The symmetry of the exact dual stiffness (13)—the Jacobian used by dual Newton—is a direct consequence of the fact that the original system $G(U) = 0$ has been given a variational status and turned into the optimality condition of a scalar objective through an adapted dual parametrization utilizing the auxiliary potential H . This conclusion does not require G_U to be square or symmetric, nor does it require G to be the gradient of a scalar potential. Moreover, while the retained dual parametrization remains in the region where L is strictly concave in U , $K \prec 0$, so $J \succeq 0$ and its restriction to the active dual space is positive definite when G_U has full row rank there. Thus the Newton correction can use symmetric linear algebra—for example, Cholesky on a positive-definite active restriction or a symmetric-indefinite factorization when appropriate—rather than a general nonsymmetric solve. This can reduce storage and factorization cost, and makes the structural-nullspace and weak-mode diagnostics naturally spectral. In implementation, the linear solves involving K must be performed consistently enough to preserve this symmetry numerically. At $D = 0$, $K = -H_{UU}(\bar{U}; \bar{U})$ is negative definite by strict convexity of H . The current dual parametrization is continued only while this predual concavity is retained with a prescribed margin. When it loses admissibility or ceases to be effective, AVD changes the base state and reinitializes the dual parametrization; the local formula (13) is therefore one component of the global AVD scheme, not the whole algorithm.

2.3 Quadratic auxiliary potential

The most common choice is

$$H(U; \bar{U}) = \frac{1}{2}(U - \bar{U})^T M (U - \bar{U}), \quad M = M^T \succ 0, \quad (14)$$

with isotropic specialization $M = \kappa I$. In this case,

$$L_U = G_U(U)^T D - M(U - \bar{U}), \quad (15)$$

$$K = \sum_{i=1}^m D_i G_{i,UU}(U) - M, \quad (16)$$

$$\mathcal{U}_D = -K^{-1} G_U^T, \quad J = -G_U K^{-1} G_U^T. \quad (17)$$

The general formulation above is retained because anisotropic or nonlinear strictly convex choices of H can shape the dual parametrization to the problem geometry.

3 Rectangular linear algebra: structural nullspace versus weak active modes

Let

$$G_U = P \Sigma V^T \quad (18)$$

be a singular-value decomposition, with singular values

$$\sigma_1 \geq \sigma_2 \geq \dots \geq 0. \quad (19)$$

The distinction between structural null modes and weak active modes is central to the AVD method.

3.1 Overdetermined systems: $m > n$

When $m > n$, G_U is tall. If it has numerical rank $r \leq n$, then

$$\dim \ker G_U^T = m - r. \quad (20)$$

These are structural dual null directions. They do not correspond to missing physical information; they arise because the dual space is larger than the primal image.

AVD therefore works in the active dual space

$$\mathcal{A}_D = \mathcal{R}(G_U) = \ker(G_U^T)^\perp, \quad (21)$$

with an orthonormal basis supplied by the left singular vectors associated with nonzero singular values. Structural zero modes are not treated as weak physical directions and are not artificially stiffened merely to make a square linear system invertible.

3.2 Underdetermined systems: $m < n$

When $m < n$ and G_U has row rank $r \leq m$,

$$\dim \ker G_U = n - r. \quad (22)$$

There is primal nonuniqueness, but this does not by itself imply a singularity in the active dual calculation: the nonzero singular values span an r -dimensional active dual space, and AVD restricts the calculation to that space.

A more delicate case occurs when a spectral gap separates the strongly active singular values from one or more small but nonzero singular values:

$$0 < \sigma_{s+1} \ll \sigma_s, \quad s < r. \quad (23)$$

The singular directions associated with $\sigma_{s+1}, \dots, \sigma_r$ are weak active modes. They must not be confused with structural null directions. They carry genuine independent information, although that information is weak in the current dual parametrization.

This distinction is one of the principal reasons for utilizing some Singular value Decomposition (SVD) related diagnostics in AVD. The purpose is not simply to make a linear system invertible; it is to understand what the linearized equations actually constrain at any fixed DtP stage.

3.3 SVD information without a full or expensive SVD

The role of singular-value information in AVD is not a requirement to compute a detailed, dense SVD of a large Jacobian at every iteration. Indeed, for a benign problem it need not compute even one singular mode. The first regularization level can be the Levenberg–Marquardt-like shifted dual solve

$$(J + \mu I)\Delta D = -G, \quad \mu > 0, \quad (24)$$

with μ controlled by the existing acceptance and backtracking logic (see Algorithm 1). This isotropic stiffening often suffices when the dual system is merely mildly ill-conditioned. It is a computational device rather than a claim that structural null modes have become active; if it restores adequate progress, no SVD or Ritz vectors need be computed.

Only when the scalar shift is inadequate does the scheme request spectral detail. It then generally needs only a small amount of information: a numerical rank, a few extreme singular values, and, when a weak active mode is detected, the corresponding few-dimensional active subspace. In many applications this information can be obtained substantially more cheaply than a full factorization.

For a rectangular Jacobian $G_U \in \mathbb{R}^{m \times n}$, the first useful question is often whether the matrix has full row rank or full column rank and whether one or a few singular values are anomalously small. A rank-revealing QR factorization, a partial Lanczos or Golub–Kahan iteration, or a randomized range-finding procedure can supply this information without constructing all singular vectors. When only a few extreme singular values are required, iterative methods can be stopped as soon as the active/weak spectral separation is resolved. The required linear algebra can likewise be implemented through Jacobian-vector and transpose-Jacobian-vector products, so an explicit dense Jacobian need not be formed in a large-scale implementation.

For the small systems used in the present numerical study, ordinary thin SVDs are inexpensive and were used directly. This is to be viewed as a transparent implementation choice, not as a computational requirement of AVD. In larger systems, the natural implementation is a *partial* or *rank-revealing* spectral computation whose accuracy is commensurate with the stage-transition decision being made. In particular, there is no need to resolve a cluster of already well-separated large singular values to high relative accuracy merely to decide whether one additional singular direction is structurally zero or weakly active.

Thus the spectral role of AVD is modest and adaptive: first use ordinary progress diagnostics and Levenberg–Marquardt-like scalar stiffening with no modal computation; if that is insufficient, use inexpensive rank information and refine only the few weak directions in a Newton solve. SVD diagnostics is necessary only as a last resort for difficult problems. This is consistent with the broader AVD design principle that GF is the cheap default mechanism and expensive linear algebra is invoked only when the nonlinear trajectory calls for it.

4 The Adaptive Variational Dual (AVD) solution scheme

For a fixed base state, the local variational construction defines $\mathcal{U}(D; \bar{U})$ only within the corresponding admissibility region. A robust nonlinear solver must decide how to evolve within the current dual parametrization, when a local mechanism has ceased to make meaningful progress, and when to change $\bar{U}$ and reinitialize the parametrization. The globalization procedure that makes these decisions is therefore central to AVD in both theory and computation.

In the following we use the notation

$$R_2 := \|G(U)\|_2; \quad R_\infty := \|G(U)\|_\infty.$$

4.1 Gradient flow (GF)

A gradient flow

$$\frac{\partial D}{\partial s} = -\mathcal{S}_D(D; \bar{U}) = -G(\mathcal{U}(D; \bar{U})) \quad (25)$$

in a fake time-like variable s provides the inexpensive default step. Its principal advantage is that it does not require a matrix solve or an SVD. Consequently, long GF trajectories are not automatically evidence of a hard problem. A solution to $G(U) = 0$ is obtained in the GF stage if R_2 and R_∞ reach below target tolerances.

4.2 Meaningful GF progress and GF-to-Newton handoff

The implemented AVD scheme monitors residual history rather than imposing an arbitrary residual-level handoff. A history-based stall detector evaluates relative residual reduction over two windows. The logic is

$$\text{GF history stall} \implies \text{Newton}. \quad (26)$$

A further diagnostic uses

$$\eta(U) = \frac{\|G_U(U)^T G(U)\|_2}{\|G_U(U)\|_2 \|G(U)\|_2 + \varepsilon_\eta}. \quad (27)$$

This quantity is useful since away from the exact-root tolerance, $G_U^T G$ is the first-order residual-reduction signal for $\frac{1}{2}\|G\|^2$. The normalization in (27) makes η dimensionless and removes the trivial effect of the current scales of G_U and G . Thus $\eta \ll 1$ means that the residual is nearly orthogonal to the local range of G_U : first-order motion in the current dual parametrization has very little ability to reduce the residual. This is precisely the geometry of a positive-residual least-squares stationary state, where $G_U^T G = 0$ although $G \neq 0$.

The diagnostic matters operationally because residual decrease alone can be misleading. A GF trajectory can continue to make microscopic decreases for a very long time while converging toward such a non-root stationary state. When two residual-history windows show collapsing progress and η is also small, the scheme assumes that the cheap local GF mechanism is exhausted and should promote the solve to Newton. The same principle later helps decide when local recovery has ceased to be meaningful and a base-state change is required. Neither small η nor $G_U^T G \approx 0$ proves infeasibility, and η is not interpreted as a root test when $\|G\|$ is already within tolerance.

4.3 Active-space Newton

The Newton time-stepping component used in the present scheme follows the computational idea developed in [9]. Here, Newton directions are computed only on the active dual space. Weak active modes are retained, with the smallest eigenvalues monitored rather than confused with structural nullspace. The scheme therefore distinguishes:

$$\text{structural zero mode} \quad \rightarrow \text{project/remove from active dual solve}, \quad (28)$$

$$\text{weak nonzero active mode} \quad \rightarrow \text{retain and monitor}, \quad (29)$$

$$\text{ordinary active mode} \quad \rightarrow \text{solve normally}. \quad (30)$$

If the Newton iterations are deemed to be lacking in making progress towards residual reduction according to the residual-history windows showing collapsing progress and η small criteria as before, a hand-off is invoked to the ‘GF rescue and outer kick’ stages.

4.4 GF rescue and outer kicks

If the Newton solves stall, a final GF phase is invoked whose failure to reduce the residual magnitude results in an outer kick stage of the algorithm—thus, the sequence is

$$\boxed{\text{GF} \rightarrow \text{Newton} \rightarrow \text{GF rescue} \rightarrow \text{outer kick}.} \quad (31)$$

The outer kick used in the presented computations is a bounded base-state restart procedure. The scheme falls back to a prescribed finite set of reproducible directions:

$$b_{\ell j} = U + r_{\ell} q_j, \quad \|q_j\|_2 = 1. \quad (32)$$

These may be angular, coordinate, block, problem-informed, or seeded spherical directions. Each fallback candidate is evaluated using $R(b) = \|G(b)\|_2$ and the normalized primal-sensitivity diagnostic $\eta(b)$ from (27). Fallback acceptance has two ordered stages:

1. *Ordinary residual-descent fallback.* If at least one candidate satisfies $R(b_{\ell j}) < R(U)$, accept the candidate having the smallest residual. No additional η threshold is imposed.
2. *Sensitivity-restoring rescue fallback.* This stage is considered only if no candidate decreases the residual. A candidate may then be accepted if $\eta(b_{\ell j}) \geq \eta_{\text{rescue}}$; among the responsive candidates, choose the one with the smallest residual. Its residual is allowed to be temporarily larger because the purpose is to leave a positive-residual insensitive region and initialize a more useful dual parametrization.

If both fallback stages fail, the scheme advances to the next prescribed radius.

Exhaustion of the finite radius schedule is reported as a globalization failure.

When a candidate b is accepted, the state is reset according to

$$\bar{U} \leftarrow b, \quad U \leftarrow b, \quad D \leftarrow 0. \quad (33)$$

The promotion and residual histories are reset, the dual parametrization is reinitialized about the new base, and the scheme returns to GF. Thus an outer kick changes the current dual parametrization but never the equation being solved: final acceptance still requires the original unweighted residual tolerance for $G(U) = 0$.

The results presented in this work utilized uniformly spaced angular directions in the two-dimensional cases and symmetry-breaking plus seeded spherical directions in the higher-dimensional cases, followed by the two-stage fallback acceptance rule above.

We mention in passing a more principled approach to the outer kick. Consider

$$\mathcal{R}(U) = \frac{1}{2} \|G(U)\|_2^2, \quad A = G_U(U), \quad \nabla \mathcal{R}(U) = A^T G(U). \quad (34)$$

Provided $\|A^T G\|_2$ is not below a prescribed scale-aware smallness threshold, define the normalized informed direction

$$q_R = -\frac{A^T G}{\|A^T G\|_2}. \quad (35)$$

At kick level ℓ , a backtracked trial $b = U + \alpha q_R$ is begun with $\alpha = r_\ell$, where $0 < r_0 < \dots < r_{L-1}$ is a finite radius schedule. The trial is accepted only if it produces a meaningful decrease of the original unweighted L^2 residual. Normalization prevents the magnitude of $A^T G$ from setting the kick length, while the small-gradient test avoids division by a nearly vanishing vector. We note that choices of $\mathcal{R}$ as an L^p , $p > 2$, norm of G is possible, and this can lead to other possibilities for the outer-kick stage of the algorithm.

4.5 Base states and successive dual parametrizations

The base state is more than an initial guess. It is the center of $H(U; \bar{U})$ and therefore the primal state corresponding to $D = 0$ in the current dual parametrization: $\mathcal{U}(0; \bar{U}) = \bar{U}$. Replacing $\bar{U}$ changes the local dual parametrization of primal states while leaving the target equation $G(U) = 0$ unchanged. Successive base-state recentering is the mechanism by which AVD globalizes the fixed-base dual variational construction.

Definition of the parameter q : The implemented AVD scheme performs a deterministic periodic base-state promotion after a prescribed number q of healthy accepted steps: it sets $\bar{U} \leftarrow U$, resets $D \leftarrow 0$, reinitializes the dual parametrization, and resumes the scheme. The reported primary protocol first uses $q = 5$. Only after complete recovery exhaustion does it restart the original initialization with $q = 2$. This 5/2 rule is a base-state update policy; it neither changes the equations nor relaxes the final unweighted residual acceptance test.

For the presented computations, all numerical scheme constants are explicit. A trial DtP state is admissible only when

$$\frac{\lambda_{\min}(-L_{UU})}{\lambda_{\max}(H_{UU}(\bar{U}))} \geq \delta, \quad \delta = 10^{-3}. \quad (36)$$

The outer dual step is accepted only if, with $R = \|G\|_2$,

$$R_{\text{trial}} \leq (1 + \epsilon_{\text{out}})R_n, \quad \epsilon_{\text{out}} = 10^{-3}; \quad (37)$$

otherwise that outer step is halved. On the active dual space, weak nonzero modes are retained and the Newton solve uses

$$\mu_i^{\text{reg}} = \max\{\mu_i, \eta_{\text{spec}} \mu_{\text{max}}\}, \quad \eta_{\text{spec}} = 10^{-5}. \quad (38)$$

Only genuine structural directions in $\ker G_U^T$ are projected from a dual solve. The values of κ in the quadratic auxiliary potential are retained problem by problem; for example, $\kappa \in 1, 2$ was utilized in the cases presented in this work; we note that the choice is free to make in the definition of the auxiliary potential.

The numerical experiments suggest a useful interpretation of difficult nonlinear cases. When the current dual parametrization stagnates or exhausts recovery, a base-state change or successful outer

kick initializes a new dual parametrization and returns the scheme to the inexpensive GF stage. A problem-informed base state can also be supplied at initialization, but base-state evolution remains part of the solver rather than merely a preprocessing choice.

A particularly useful construction for a weak active mode is to solve a well-conditioned core and then search for a nearby base state satisfying the remaining independent information. In a simple four-parameter neural-network test, solving a two-equation core $G_1 = G_2 = 0$ and then varying one primal coordinate until an independent $G_3 = 0$ condition was found to produce base states from which the full AVD solve could recover cases that otherwise failed. This is presented as a base-state construction principle, not as a replacement for the full equation problem.

4.6 The role of the auxiliary potential

The quadratic choice

$$H = \frac{1}{2}(U - \bar{U})^T M(U - \bar{U}) \quad (39)$$

need not be isotropic. Its anisotropy can reflect strong and weak primal directions. More generally, strictly convex nonlinear H is permitted. This gives AVD a design degree of freedom absent from a fixed least-squares objective: the dual parametrization itself can be adapted to the nonlinear geometry.

5 Numerical study and comparison methodology

All primary comparisons are made on identical sets of initial conditions. AVD success means satisfaction of the original equation system to the prescribed unweighted residual tolerance. For the current studies,

$$R_2 = \|G\|_2 \leq 10^{-10}, \quad R_\infty = \|G\|_\infty \leq 10^{-10} \quad (40)$$

is the equation-solve acceptance criterion.

Algorithm 1 gives the complete AVD scheme and the numerical controls used in the reported calculations. To avoid overloading the notation, ϵ_{sp} denotes the small-progress threshold, η_{sp} denotes the normalized stationarity threshold used in the GF-to-Newton handoff, and η_{spec} denotes the distinct weak-active spectral floor. The parameter $\epsilon_\eta = 10^{-300}$ in (27) only protects the normalization against division by zero. The residual-history window length is denoted by W steps. The parameter `maxDtP`, the maximum number of DtP iterations allowed, was conservatively set. In a representative statistic comprising 35,498 DtP calls from 16 runs spanning the principal problem classes, no call reached this limit. The maximum for an accepted DtP solve was 11 iterations, and most accepted solves required at most three. This is expected, as the DtP solve finds the unique maximizer of a strictly concave problem by design. In the reported computations, the prescribed outer-kick radius schedule was $(r_0, \dots, r_5) = (0.10, 0.20, 0.35, 0.50, 0.75, 1.00)$.

For least squares, we distinguish native optimization termination from solution of $G = 0$. A least-squares run is counted as solving the original equation problem only if its returned residual satisfies the same equation tolerance. Native optimization outcomes are separately reported. All reported least-squares comparisons use `scipy.optimize.least_squares` with supplied analytic Jacobians: the trust-region reflective method (TRF) in Cases 2–6 except Case 2b, where TRF, dogbox, and Levenberg–Marquardt (LM) are each reported. Case 2c additionally uses the practical square root solver `scipy.optimize.root(method="hybr")`, the Powell hybrid method; it is distinguished from the undamped primal Newton map analyzed there.

```

1: Input: Initial state  $U_0$ ; case-specific callbacks and kick directions
2: Set  $\delta = 10^{-3}$ ,  $\epsilon_{\text{out}} = 10^{-3}$ ,  $\epsilon_{\text{sp}} = 10^{-3}$ ,  $\eta_{\text{sp}} = 2 \times 10^{-6}$ ,  $\eta_{\text{spec}} = 10^{-5}$ ,  $\eta_{\text{rescue}} = 0.25$ ,  $W = 50$ , and  $\text{maxDtP} = 80$ .
3: Set the exact-solve tolerances  $\tau_2 = \tau_{\infty} = 10^{-10}$ , the DtP stationarity tolerance to  $10^{-12}$ , and the minimum outer step to  $6 \times 10^{-5}$ 
4: for  $q \in (5, 2)$  do
    Use the  $q = 2$  pass only after complete recovery exhaustion of the  $q = 5$  pass
    Reset  $U \leftarrow U_0$ ,  $\bar{U} \leftarrow U_0$ ,  $D \leftarrow 0$ ; initialize the dual parametrization; set mode to GF
    repeat
        Form the trial dual step prescribed by the current mode
        Solve  $L_U(U, D_{\text{trial}}; \bar{U}) = 0$  by DtP Newton
        if the DtP solve does not reach tolerance within  $\text{maxDtP}$  iterations, or  $\lambda_{\min}(-L_{UU})/\lambda_{\max}(H_{UU}(\bar{U})) < \delta$  then
            Reject the trial, halve the outer step, and retry
        else if  $R_{\text{trial}} > (1 + \epsilon_{\text{out}})R_n$  then
            Reject the trial, halve the outer step, and retry
        else
            Accept  $(D, U) \leftarrow (D_{\text{trial}}, U_{\text{trial}})$ 
        if  $R_2 \leq \tau_2$  and  $R_{\infty} \leq \tau_{\infty}$  then
            return  $U$  as an exact solution of the original equations
        if  $q$  healthy accepted steps have accumulated then
            Promote the base:  $\bar{U} \leftarrow U$ ,  $D \leftarrow 0$ ; reinitialize the dual parametrization and continue in GF
        if mode is GF and either  $\eta \leq \eta_{\text{sp}}$  or both consecutive  $W$ -step windows have relative progress below  $\epsilon_{\text{sp}}$  then
            Set mode to active-space Newton
        else if mode is active-space Newton and Newton stalls then
            Set mode to GF rescue
        else if mode is GF rescue and local recovery is exhausted then
            Set mode to outer kick
        if mode is active-space Newton then
            Project only  $\ker G_U^T$  and use  $\mu_i^{\text{reg}} = \max\{\mu_i, \eta_{\text{spec}}\mu_{\text{max}}\}$ 
        if mode is outer kick then
            Accept the smallest-residual candidate among those that decrease  $R$ 
            if no candidate decreases  $R$  then
                Accept the smallest-residual candidate satisfying  $\eta \geq \eta_{\text{rescue}}$ , if one exists
            if a kick is accepted then
                Set  $\bar{U} \leftarrow U \leftarrow U_{\text{kick}}$ ,  $D \leftarrow 0$ ; reset histories; return to GF
            else if the finite prescribed kick schedule is exhausted then
                break begin  $q = 2$  restart, or report exhaustion after that pass
5: return globalization exhaustion without relaxing the exact residual test
    
```

Algorithm 1: The implemented AVD scheme. The functions G , G_U , and $D \cdot G_{UU}$ and the value $\kappa \in \{1, 2\}$ used are case-specific; the stage-transition logic is unchanged.

6 Overdetermined Problems

The AVD and LS schemes are compared as equation solvers for three overdetermined nonlinear nonsquare problems in this Section.

6.1 A feasible overdetermined LS pathology

The first benchmark is the deliberately small analytic system $G : \mathbb{R} \rightarrow \mathbb{R}^2$, so $n = 1$ and $m = 2$:

$$G(x) = \begin{bmatrix} x^2 - 1 \\ (x^2 - 1)(1 + 3x^2) \end{bmatrix}. \quad (41)$$

It is feasible, with exact roots $x = \pm 1$. Nevertheless, the least-squares objective possesses a strict nonzero-residual local minimum at $x = 0$. At such a state,

$$G_U^T G = 0, \quad G \neq 0, \quad (42)$$

so the least-squares optimization problem can terminate without solving the equation problem.

This test is a simple demonstration of why solving $G = 0$ by minimizing $\|G\|^2$ can be mathematically different from solving $G = 0$ directly. AVD retains the original equation residual as the acceptance criterion and therefore does not identify the false stationary point as a solution.

6.2 Nonlinear overdetermined basin

The second benchmark is the overdetermined map $G : \mathbb{R}^2 \rightarrow \mathbb{R}^4$

$$G(x, y) = \begin{bmatrix} (x^2 - 1)(2x + 1) \\ (x^2 - 1)(1 - 2x) \\ (y^2 - 1)(2y + 1) \\ (y^2 - 1)(1 - 2y) \end{bmatrix}. \quad (43)$$

It was evaluated from a 13×13 grid of 169 initial conditions on $[-1.5, 1.5]^2$. The AVD scheme solved all tested starts for a $169/169 = 100\%$ success rate.

The SciPy trust-region reflective least-squares implementation for the same problem reached the exact equation residual criterion from $100/169 = 59.2\%$ of the starts; the remaining $69/169 = 40.8\%$ terminated at nonzero residuals. All 169 runs had native optimizer termination. The equation-solve criterion was held fixed at $\|G\|_2, \|G\|_\infty \leq 10^{-10}$.

This test demonstrates that the mechanism in Sec 6.1 is not merely an analytic curiosity. It has a substantial basin-level effect in a nonlinear overdetermined computation.

6.3 A uniquely selected unstable critical point

This benchmark was designed specifically to test whether least squares can be trapped at unwanted critical points of a nonconvex scalar function after an additional equation makes one unstable critical point the unique solution of the rectangular system. Let

$$\phi(x, y) = \frac{1}{4}(x^2 - 1)^2 - \frac{1}{2}y^2. \quad (44)$$

Its critical points are

$$A = (1, 0), \quad B = (-1, 0), \quad C = (0, 0). \quad (45)$$

The desired point A is an unstable saddle because

$$\nabla^2 \phi(A) = \begin{bmatrix} 2 & 0 \\ 0 & -1 \end{bmatrix}. \quad (46)$$

We append the single nonlinear selection equation

$$h(x) = \frac{1}{3} + \frac{1}{2}x^2 - \frac{1}{3}x^3 - \frac{1}{2}x^4 \quad (47)$$

to the critical-point equations and solve

$$G(x, y) = \begin{bmatrix} \phi_x \\ \phi_y \\ h(x) \end{bmatrix} = \begin{bmatrix} x^3 - x \\ -y \\ h(x) \end{bmatrix} = 0, \quad G : \mathbb{R}^2 \rightarrow \mathbb{R}^3. \quad (48)$$

The augmented system has the unique exact root A : $h(1) = 0$, whereas $h(-1) = 2/3$ and $h(0) = 1/3$. At the unwanted critical points, $h'(-1) = h'(0) = 0$, and consequently

$$G(B) = \begin{bmatrix} 0 \\ 0 \\ 2/3 \end{bmatrix}, \quad G(C) = \begin{bmatrix} 0 \\ 0 \\ 1/3 \end{bmatrix}, \quad G_U(B)^T G(B) = G_U(C)^T G(C) = 0. \quad (49)$$

Thus B and C remain stationary points of the full least-squares objective $\Phi = \frac{1}{2}\|G\|_2^2$, even though neither solves the augmented equations. More strongly,

$$\nabla^2 \Phi(B) = \text{diag}(2, 1), \quad \nabla^2 \Phi(C) = \text{diag}(4/3, 1), \quad (50)$$

so both are strict positive-residual local minima.

All methods were tested from the same 9×9 grid on $[-2, 2]^2$, using $\|G\|_2 \leq 10^{-10}$ as the exact-root criterion. The standard least-squares runs used the supplied exact Jacobian of (48) and tight internal tolerances. The results are presented in Table 1.

Table 1: Selected-saddle basin. Every least-squares run reported native optimizer success, but only the A column solves $G = 0$.

Method	A exact	B false minimum	C false minimum	Total
LS-TRF	27	27	27	81
LS-dogbox	27	31	23	81
LS-LM	27	35	19	81
Full AVD scheme	81	0	0	81

The worst accepted AVD residual was 8.88×10^{-11} (and of course $R_\infty \leq R_2$ in this finite dimensional context).

7 Underdetermined Problems

We consider neural-network problems as the prototypical example. Consider the two-unit, four-parameter neural network

$$y(x; U) = a_1 \tanh(b_1 x) + a_2 \tanh(b_2 x), \quad (51)$$

with parameter vector $U = (a_1, b_1, a_2, b_2)$. The equation residuals are observation mismatches

$$G_i(U) = y(x_i; U) - y_i^{\text{target}}. \quad (52)$$

The exact synthetic data were generated from

$$U_{\text{true}} = (1.2, 0.85, -0.75, -0.42), \quad (x_1, x_2, x_3) = (0.7, 1.6, 2.4), \quad (53)$$

with $y_i^{\text{target}} = y(x_i; U_{\text{true}})$.

7.1 Regular nonredundant underdetermined problems

The regular problem uses only the first two observation equations,

$$G(U) = \begin{bmatrix} G_1(U) \\ G_2(U) \end{bmatrix}, \quad G : \mathbb{R}^4 \rightarrow \mathbb{R}^2. \quad (54)$$

Thus the final dimensions are $n = 4$ unknown network parameters and $m = 2$ exact-fit equations. The problem was tested from 169 starts. AVD solved all 169 cases, and least squares also solved all 169 at the equation residual tolerance. Thus underdetermination alone is not a difficulty for either method in this example.

Thus, on a benign neural-network-type nonsquare problem the performance of both LS and AVD are adequate.

7.2 Difficult near-redundant problems

This case has the same four unknown parameters but three equations, so $n = 4$ and $m = 3$. The third equation was deliberately made nearly redundant:

$$G(U) = \begin{bmatrix} G_1(U) \\ G_2(U) \\ G_1(U) + G_2(U) + \varepsilon G_3(U) \end{bmatrix}, \quad \varepsilon = 10^{-2}. \quad (55)$$

Hence $G : \mathbb{R}^4 \rightarrow \mathbb{R}^3$; the weak but independent information enters through the third observation G_3 , evaluated at $x_3 = 2.4$. An exact solution exists by construction. At the true parameter U_* ,

$$G_1(U_*) = G_2(U_*) = G_3(U_*) = 0, \quad (56)$$

so $G(U_*) = 0$ for every nonzero ε .

The results from the two schemes for identical 169-starts were

$$\boxed{\text{AVD: } 169/169 = 100\%}, \quad \boxed{\text{LS: } 98/169 = 58.0\%}. \quad (57)$$

Thus the AVD scheme solves every tested initialization, while least squares fails the original exact-equation criterion on 71 starts despite its native optimization termination.

7.3 Curve fitting a high-frequency function

Neural networks extend classical linear regression by mapping complex, nonlinear relationships through hidden layers of weighted transformations [12]. To ameliorate spectral bias in neural networks, Fourier feature embeddings [13] and periodic sin activation functions [14] have been proposed. To show that standard activations can capture high-frequency dynamics, we evaluate

the AVD method for 1D curve fitting. For $x \in \Omega = [0, 1]$, the target function $u(x) = \sin(2\pi f x)$ is considered ($f = 5, 10$ are chosen). The input x is mapped to $[-1, 1]$ via $T(x) = 2x - 1$, and the neural network ansatz is:

$$\mathcal{N}(x; U) = \sum_{i=1}^N c_i \tanh(w_i T(x) + b_i), \quad (58)$$

where $U = \{w_i, b_i, c_i\}_{i=1}^N \in \mathbb{R}^n$ denotes the vector of $n = 3N$ trainable parameters.

To ensure consistent coverage across target frequencies, the initial network parameters are sampled independently according to

$$w_{0j} \stackrel{\text{iid}}{\sim} \text{Unif}(1, 6f), \quad x_{0j} \stackrel{\text{iid}}{\sim} \text{Unif}(-0.99, 0.99), \quad j = 1, \dots, N.$$

Then, the initial biases are explicitly set via $b_0 = -w_0 \odot x_0$, which places the inflection points of the tanh activations within the physical domain. Collocation points $\{x_i\}_{i=1}^m$ are equispaced in Ω . We evaluate two resolution levels;: $m = 21, 41$ for $f = 5$; and $m = 41, 81$ for $f = 10$. For the predictions, Ω is discretized with 2001 equispaced points. The neural network prediction, $\mathcal{N}(x) := \mathcal{N}(x; U^*)$, where $U^* \in \mathbb{R}^n$ are the converged network parameters.

Figure 2 compares the network prediction to the target function, and Table 2 summarizes the error metrics across all tested cases. At $m = 21$ for $f = 5$ ($m = 41$ for $f = 10$), the grid provides roughly 4 points per cycle, which is below the effective sampling density required for high-accuracy representation, leading to under-resolved predictions ($\|u - \mathcal{N}\|_2 \approx [0.15, 0.18]$). On doubling the resolution to $m = 41$ ($n = 123$) for $f = 5$, and $m = 81$ ($n = 243$) for $f = 10$ (yielding ~ 8 collocation points per cycle), we observe close to an order-of-magnitude reduction in the L^2 error (falling to ~ 0.018). This shows that on reshaping the optimization landscape, the AVD method has the capability to mitigate spectral bias using tanh activation on shallow neural networks.

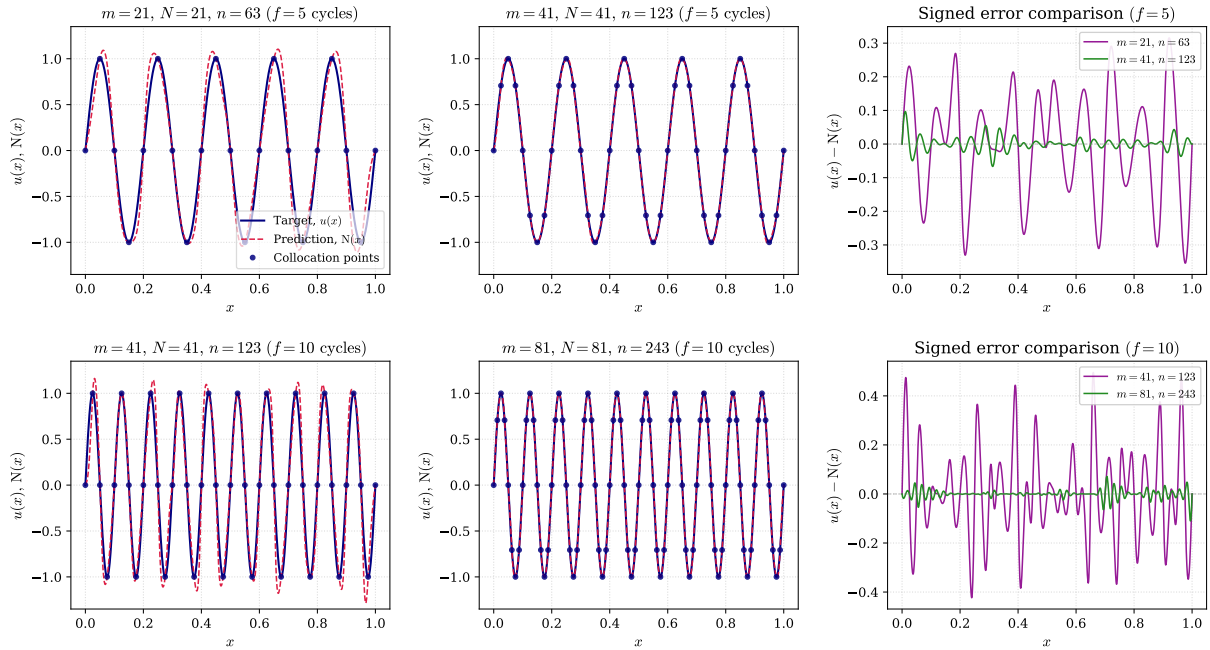


Figure 2: Panel comparison of AVD neural network predictions for $f = 5$ (top row) and $f = 10$ (bottom row) for varying grid resolutions m and number of parameters $n = 3N$.

Table 2: Residual L^2 and L_∞ error norms across frequencies and grid resolutions. Convergence is attained when both $R_2 \leq 10^{-10}$ and $R_\infty \leq 10^{-10}$ are met.

Frequency (f)	Collocation points (m)	Neurons (N)	$\ u - \mathcal{N}\ _2$	$\ u - \mathcal{N}\ _\infty$
5	21	21	1.58×10^{-1}	3.54×10^{-1}
5	41	41	2.06×10^{-2}	9.68×10^{-2}
10	41	41	1.82×10^{-1}	4.95×10^{-1}
10	81	81	1.78×10^{-2}	1.10×10^{-1}

7.4 Infeasibility and the least-squares stationarity ambiguity

The problems dealt with here have the flavor of the problem treated in Sec. 6.1, but from the opposite direction, apart from being underdetermined. That case is feasible and LS nevertheless has a non-root stationary point. The problems treated in Secs. 7.4.1–7.4.2 are infeasible, so a positive-residual least-squares stationary point is unavoidable; the key question is whether its residual is so large or so small that the computational meaning becomes obvious or ambiguous.

Consider the circle/cylinder-intersection family

$$G_1(x, y, z) = x^2 + y^2 - 1, \quad (59)$$

$$G_2(x, y, z) = (x - c)^2 + y^2 - 1, \quad (60)$$

with z free and $c > 2$. Thus both cases have $n = 3$ unknowns, $m = 2$ equations, and $G : \mathbb{R}^3 \rightarrow \mathbb{R}^2$; the free variable z makes the system underdetermined. The two unit circles are disjoint, so no exact solution exists. The least-squares minimizer occurs at $x = c/2$, $y = 0$, for which

$$G_1 = G_2 = \frac{c^2}{4} - 1, \quad (61)$$

and hence

$$\min \|G\|_2 = \sqrt{2} \left(\frac{c^2}{4} - 1 \right). \quad (62)$$

7.4.1 Grossly infeasible

For the grossly infeasible test we set $c = 3$. The full 169-start basin produced no AVD exact solutions. Least squares reported successful optimization termination from all 169 starts, but none solved the equation problem; the common residual was approximately

$$\|G\|_2 \approx 1.76777. \quad (63)$$

The lesson is straightforward: LS can solve its optimization problem perfectly well while returning no solution of $G = 0$.

7.4.2 Nearly infeasible

We chose $c = 2 + \delta$, with $\delta > 0$ determined by

$$\sqrt{2} \left(\frac{(2 + \delta)^2}{4} - 1 \right) = 10^{-7}, \quad (64)$$

so that

$$\min \|G\|_2 \approx 10^{-7}. \quad (65)$$

This is the practically interesting regime. A residual of order one obviously indicates infeasibility; a residual of order 10^{-7} can look indistinguishable from a very accurate numerical solution in a routine workflow unless one has an independent feasibility criterion or performs further analysis.

With $\kappa = 1$, the AVD scheme initially follows inexpensive GF, then uses the stationarity diagnostic to trigger Newton and subsequently applies the outer kick mechanism. It reaches residual values of $R_2 = 1.00166 \times 10^{-7}$ and $R_\infty = 7.08 \times 10^{-8}$, then quits without accepting $G \neq 0$ as a solution.

The computational lesson is not that least squares is a bad optimizer. It is that the optimization statement and the equation statement are different. The cases dealt with in Secs. 6.1-7.4.1-7.4.2 form a useful conceptual triplet:

Sec.	exact root?	non-root LS stationary state
6.1	yes	spurious local minimum exists
7.4.1	no	positive-residual minimum, obvious
7.4.2	no	positive-residual minimum, subtle

(66)

The least-squares first-order condition $G_U^T G = 0$ alone does not distinguish these cases.

8 Square Problems

The AVD scheme is evaluated as a solver for two nonlinear square problems in this Section.

8.1 Square critical-point equations with a split Newton cycle

To separate the variational mechanism from rectangular linear algebra, we also consider a square problem with a slightly larger number of degrees of freedom. Let $U = (u_1, \dots, u_{12}) \in \mathbb{R}^{12}$, impose periodic indexing, and define the coupled nonconvex energy

$$\mathcal{E}(U) = \sum_{i=1}^{12} \psi(u_i) + \varepsilon \sum_{i=1}^{12} V(u_{i+1} - u_i), \quad \varepsilon = 0.02, \quad (67)$$

where

$$\psi(z) = \frac{1}{6}z^6 - \frac{1}{4}z^5 - \frac{1}{8}z^2 + \frac{1}{4}z, \quad \psi'(z) = f(z) = z^5 - \frac{5}{4}z^4 - \frac{1}{4}z + \frac{1}{4}, \quad (68)$$

$$V(s) = \frac{1}{10}s^{10} - \frac{3}{8}s^8 + \frac{1}{2}s^6 - \frac{1}{4}s^4, \quad V'(s) = s^3(s^2 - 1)^3. \quad (69)$$

The square equation problem is the critical-point system

$$G(U) = \nabla \mathcal{E}(U) = 0, \quad G : \mathbb{R}^{12} \rightarrow \mathbb{R}^{12}. \quad (70)$$

The single-variable potential ψ has three critical points,

$$\alpha_- = -0.68304360796, \quad \alpha_0 = 0.62271110166, \quad \alpha_+ = 1.27602870005. \quad (71)$$

Consider the two complementary split initializations

$$U^{(a)} = (\underbrace{0, \dots, 0}_6, \underbrace{1, \dots, 1}_6), \quad U^{(b)} = \mathbf{1} - U^{(a)}. \quad (72)$$

They implement the intended competing-well geometry. Since $f(0) = 1/4$ and $f(1) = -1/4$, negative primal-gradient motion sends the zero block toward the negative critical point α_- and the unit block toward the positive critical point α_+ . At the same time, every nearest-neighbor difference in (72) is 0 or ± 1 , and

$$V'(0) = V'(\pm 1) = V''(0) = V''(\pm 1) = 0. \quad (73)$$

Consequently the nonlinear coupling contributes neither gradient nor Hessian at the two split states.

This makes the failure of ordinary primal Newton exact rather than a consequence of a chosen iteration limit. Since $f'(0) = f'(1) = -1/4$, the scalar Newton map $N(z) = z - f(z)/f'(z)$ satisfies

$$N(0) = 1, \quad N(1) = 0. \quad (74)$$

The full twelve-dimensional Newton iteration therefore maps $U^{(a)}$ to $U^{(b)}$ and $U^{(b)}$ back to $U^{(a)}$. It remains on this positive-residual two-cycle indefinitely, with

$$\|G(U^{(a)})\|_2 = \|G(U^{(b)})\|_2 = \frac{\sqrt{12}}{4} = 0.86602540378, \quad \|G\|_\infty = 0.25. \quad (75)$$

This exact statement concerns the undamped primal Newton map. As a practical comparison, `scipy.optimize.root(method="hybr")` was also run with the analytic Jacobian for the problem. This Powell hybrid root solver, which has globalization beyond the raw Newton map, reached the uniform root from both $U^{(a)}$ and $U^{(b)}$: the respective residual pairs were $(R_2, R_\infty) = (2.28 \times 10^{-15}, 1.17 \times 10^{-15})$ and $(4.38 \times 10^{-15}, 2.14 \times 10^{-15})$.

AVD was run with the quadratic auxiliary potential $H = \frac{1}{2} \|U - \bar{U}\|_2^2$ and the same exact-root tolerance $\|G\|_2 \leq 10^{-10}$ under the scheme parameters (36)–(38). From each split initialization, AVD reached the uniform critical point

$$U_* = (\alpha_0, \dots, \alpha_0) \quad (76)$$

in 36 accepted dual GF steps, involving seven periodic base-state promotions; neither active-space Newton nor an outer kick was invoked. The final residuals

$$\|G(U_*)\|_2 = 6.75 \times 10^{-11}, \quad \|G(U_*)\|_\infty = 2.79 \times 10^{-11}. \quad (77)$$

were achieved. The Hessian eigenvalues at the computed state are all approximately -0.70551623 , so the recovered critical point is strongly unstable for the primal energy. This square control shows that the dual parametrization can avoid an exact Newton cycle and recover an unstable critical point from initializations that are attracted toward different critical points of the primal energy.

8.2 Five-bus AC power-flow tests

We next apply the AVD scheme to the Stagg–El-Abiad five-bus AC power-flow equations [15]. Bus 1 is the slack bus with $(V_1, \theta_1) = (1.06, 0)$; bus 2 is a PV bus with $V_2 = 1.045$. The primal variable is

$$U = (\theta_2, \theta_3, \theta_4, \theta_5, V_3, V_4, V_5) \in \mathbb{R}^7. \quad (78)$$

With the supplied admittance matrix $Y = G + iB$, calculated nodal injections are

$$P_i(U) = V_i \sum_{j=1}^5 V_j (G_{ij} \cos(\theta_i - \theta_j) + B_{ij} \sin(\theta_i - \theta_j)), \quad (79)$$

$$Q_i(U) = V_i \sum_{j=1}^5 V_j (G_{ij} \sin(\theta_i - \theta_j) - B_{ij} \cos(\theta_i - \theta_j)). \quad (80)$$

The seven residual components are the scheduled-minus-calculated mismatches

$$G(U) = (P_2^{\text{sch}} - P_2, \dots, P_5^{\text{sch}} - P_5, Q_3^{\text{sch}} - Q_3, Q_4^{\text{sch}} - Q_4, Q_5^{\text{sch}} - Q_5). \quad (81)$$

The specified flat start $\theta_2 = \dots = \theta_5 = 0$, $V_3 = V_4 = V_5 = 1$ for starting base state with $\kappa = 1$ was used. The two schedules and the recovered solutions are listed in Tables 3–4.

Table 3: Five-bus scheduled injections (per unit).

Case	$(P_2^{\text{sch}}, P_3^{\text{sch}}, P_4^{\text{sch}}, P_5^{\text{sch}})$	$(Q_3^{\text{sch}}, Q_4^{\text{sch}}, Q_5^{\text{sch}})$
A	$(.20, -.45, -.40, -.60)$	$(-.15, -.05, -.10)$
B	$(.35, -.7875, -.70, -1.05)$	$(-.2625, -.0875, -.175)$

Table 4: Five-bus AVD solutions from the flat start. Angles are in radians.

Case	$(\theta_2, \theta_3, \theta_4, \theta_5)$	(V_3, V_4, V_5)	R_2	R_∞
A	$(-.084641, -.103272, -.142996, -.153904)$	$(1.011111, 1.001499, .994194)$	6.10×10^{-11}	3.60×10^{-11}
B	$(-.161465, -.191787, -.268114, -.289304)$	$(.971210, .957001, .942840)$	9.83×10^{-11}	5.83×10^{-11}

The standard base case required 226 outer operations and 45 periodic base promotions. The stressed heavy-load case required 1,099 outer operations and 219 promotions. All were inexpensive GF/DtP-stage operations. The GF progress was efficient and no Newton iterations were invoked.

To test the Newton-related code on this problem, we started directly in the unchanged AVD active-space Newton stage from the same flat base state. Table 5 shows that it reaches the same roots in five and six full AVD-Newton steps, respectively. Every step was accepted with $\alpha = 1$, without backtracking.

Table 5: Direct AVD active-space Newton diagnostic from the five-bus flat start.

Case	AVD-Newton steps	Final R_2	Final R_∞
A	5	5.11×10^{-15}	2.61×10^{-15}
B	6	5.98×10^{-15}	4.08×10^{-15}

Taken together, these results establish that the AVD scheme applies directly to a physically standard square power-flow formulation from its conventional flat initialization.

9 Conclusions

We presented the Adaptive Variational Dual (AVD) method as a direct, variational framework for solving nonlinear, nonsquare systems $G(U) = 0$, $G : \mathbb{R}^n \rightarrow \mathbb{R}^m$. By maintaining the primal residual equations as the desired objective—rather than minimizing a least-squares functional—the AVD scheme constructs a sequence of strictly convex local optimization problems that is shaped by local singular-value geometry. The computational algorithm integrates step-size controlled gradient flow, active-space Newton steps, and bounded base-state updates to ensure that the residual equations are met to user-specified tolerances.

On the basis of numerical evaluations across overdetermined, underdetermined, and square benchmark problems, we draw the following main conclusions:

1. Solving dual stationarity equations directly preserves exact equation fidelity. Unlike nonlinear least-squares methods, which frequently stall at false positive-residual minima where $G_U^\top G = 0$ ($G \neq 0$), the AVD scheme is consistently able to achieve exact root recovery.
2. By leveraging SVD-driven active-space projections, the AVD scheme is able to distinguish between structural dual null directions in overdetermined systems ($m > n$) and weak but genuinely informative active modes in near-redundant underdetermined systems ($m < n$).
3. Across initialization landscapes for overdetermined systems, selected-saddle configurations, and near-redundant neural network tests, the AVD method demonstrated superior global convergence compared to standard nonlinear least-squares algorithms.
4. The AVD scheme was also tested on two specific applications. On a 7-variable IEEE 5-bus AC power-flow system, the AVD scheme achieved exact solutions directly from flat initializations without requiring full Newton solves (AVD succeeds as well if solely Newton solves are used). When applied to 1D curve fitting with shallow (tanh activation) neural networks, the AVD method successfully achieved accurate representation of high-frequency targets, thereby revealing the potential to mitigate spectral bias.
5. A first-order AVD GF scheme has potential for use in solving large nonlinear systems of equations, particularly when they do not arise as critical point equations of a scalar objective. A large DtP solve is involved, but due to the strict concavity of the Lagrangian in the primal variables enforced by the scheme, matrix-free iterative schemes can be employed.
6. The symmetry of the system (dual) Jacobian regardless of the nature of the primal system of equations is noteworthy, and has obvious practical advantages.

These findings reveal that the AVD scheme offers a promising, mathematically consistent framework for solving general nonsquare nonlinear algebraic systems that may suffer from ill-conditioning and be prone to non-root stationary points.

Acknowledgments

It is a pleasure for AA to thank Janusz Ginster for long-term discussions on the duality method applied in this work, and for his comments on the paper.

This work owes a great deal to the LLM ChatGPT (OpenAI), and this acknowledgment is somewhat elaborate because it cannot be an author by current rules. To be sure it did not generate new ideas and needed to be kept on a tight leash by its human handler to reach the goals of this paper, without digressing or making mistakes based on its own judgment and knowledge. But within those constraints it was spectacularly efficient on routine but non-trivial technical tasks, including reasoning on matters firmly within the intersection of its own and its handler’s knowledge base. Any errors that may remain are obviously the responsibility of the humans involved in authorship, as would be the case even if it were to be an author, a role it richly deserves for this paper by the standards usually accorded to humans for authorship.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work the authors used ChatGPT (OpenAI) in order to assist with code development and preparation of the first preliminary draft of the manuscript. After using

this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the manuscript/published article.

References

- [1] J. E. Dennis, Jr. and R. B. Schnabel, “Numerical Methods for Unconstrained Optimization and Nonlinear Equations,” *SIAM Publications*, Volume 16, Classics in Applied Mathematics, 1996 [DOI]
- [2] W. C. Rheinboldt, “Methods for Solving Systems of Nonlinear Equations,” *SIAM Publications*, CBMS-NSF Regional Conference Series in Applied Mathematics, Volume 70, Second Edition, 1998 [DOI]
- [3] A. Ben-Israel and T. N. E. Greville, “Generalized Inverses: Theory and Applications,” *Springer-Verlag*, CMS Books in Mathematics, Second Edition, 2003 [DOI]
- [4] J. Nocedal and S. J. Wright, “Numerical Optimization,” Springer Series in Operations Research and Financial Engineering, *Springer*, Second Edition, New York, N. Y., 2006 [DOI]
- [5] A. Acharya, “A dual variational principle for nonlinear dislocation dynamics,” *Journal of Elasticity*, 154, 383–395 (2023). [link].
- [6] Y. Brenier. “Examples of hidden convexity in nonlinear PDEs.” HAL preprint hal-02928398, 2020. <https://hal.science/hal-02928398/document>.
- [7] A. Acharya, “A Hidden Convexity in Continuum Mechanics, with application to classical, continuous-time, rate-(in)dependent plasticity,” *Mathematics and Mechanics of Solids*, 30(3), 701–719 (2025); published online 2024. [arXiv].
- [8] S. Singh, J. Ginster, and A. Acharya. “A hidden convexity of nonlinear elasticity.” *Journal of Elasticity*, 156:975–1014, 2024.
- [9] U. Kouskiya, R. L. Pego, and A. Acharya, “Traveling wave profiles for a semi-discrete Burgers equation,” *Physica D: Nonlinear Phenomena*, 483, 134961 (2025). [link].
- [10] D. Vorotnikov and A. Acharya, “On the variational dual formulation of the Nash system and an adaptive convex gradient-flow approach to nonlinear PDEs,” 2025. To appear in *SIAM Journal of Mathematical Analysis*, [arXiv].
- [11] M. Larrain Silva and A. Acharya, “Solving the Second Law not as a constitutive restriction,” 2026. [DOI/arXiv].
- [12] K. Hornik, M. Stinchcombe, and H. White. “Multilayer feedforward networks are universal approximators.” *Neural Networks*, 2(5):359–366, 1989.
- [13] M. Tancik, P. P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. T. Barron, and R. Ng. “Fourier features let networks learn high frequency functions in low dimensional domains.” In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 7537–7547, 2020.
- [14] V. Sitzmann, J. Martel, A. Bergman, D. Lindell, and G. Wetzstein. “Implicit neural representations with periodic activation functions.” In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 7498–7509, 2020.
- [15] G. Stagg and A. El-Abiad. “Computer Methods in Power System Analysis.” McGraw-Hill, New York, 1968.