This article has been accepted for publication in IEEE Control Systems Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LCSYS.2026.3703783

Quantization Robustness of Monotone Operator Equilibrium Networks

James Li, Philip H.W. Leong, and Thomas Chaffey

***Abstract*— Monotone operator equilibrium networks (MonDEQs) are implicit models whose well-posedness and convergence rely on a single scalar margin computed from the weight matrix, which has made them attractive for learned controllers with closed-loop stability guarantees. When such models are deployed on low-precision hardware, weights are quantized, and these structural guarantees can be destroyed. We provide an explicit quantization-preservation certificate: a single computable check on the weight perturbation determines whether the quantized network retains existence, uniqueness, linear convergence, and a bounded displacement of its equilibrium. The same threshold certifies the backward solve as well as the forward solve, so it governs both training and inference. MNIST experiments confirm a phase transition at the predicted threshold: three- and four-bit post-training quantization diverge, while five-bit and above converge; quantization-aware training recovers provable convergence at four bits.**

***Index Terms*— Quantization (signal), Neural networks, Robustness, Convergence, Optimization**

I. Introduction

DEPLOYING neural networks as controllers in safety-critical applications requires rigorous behavioral guarantees. Implicit-layer architectures such as monotone operator equilibrium networks (MonDEQs) [1] and the related recurrent equilibrium networks (RENs) [2] have emerged as models providing such guarantees: RENs have been used to learn nonlinear controllers with closed-loop stability [3], [4]. However, embedded deployment requires **quantization**: representing weights and activations at low-bit precision [5]. Hardware efficiency and accuracy are conflicting goals, since rounding error grows as bit-precision is reduced. Analytic bounds relating quantization error to a network's robustness would let bit-width be selected based on deployment requirements rather than by trial and error.

This motivates the question of whether quantization error can be bounded at the model level. At present, there is no generally applicable bound on quantization error; instead, only architecture-specific analyses exist [6], [7]. Progress therefore requires restricting attention to architectures with tractable convergence guarantees — a requirement familiar in control, where quantized feedback has been modeled as a sector-bounded perturbation and stability is analyzed via small-gain conditions [8]. MonDEQs are a class of deep equilibrium models (DEQs) [9] that enforce monotonicity of the underlying operator, guaranteeing existence, uniqueness, and linear convergence of the equilibrium via operator splitting. A MonDEQ layer's well-posedness is captured by a single spectral margin: the smallest eigenvalue m of a symmetric matrix constructed from the layer's weights (defined formally in Section II). Having $m > 0$ ensures the implicit equation has a unique equilibrium that the numerical solver converges to; because quantization perturbs this matrix, the margin m provides a natural handle for analyzing quantization error. To our knowledge, MonDEQ behavior under quantization has not been analyzed.

The authors are with the School of Electrical and Computer Engineering, The University of Sydney, NSW, Australia. Emails: jali4795@uni.sydney.edu.au, {philip.leong, thomas.chaffey}@sydney.edu.au.

A. Contributions

We give the first explicit quantization-preservation certificate for a MonDEQ's structural guarantees: a single computable threshold $\|\Delta W\|_2 < m$ under which existence, uniqueness, linear convergence, and a deterministic displacement bound all hold for the quantized network. The induced perturbation of the monotonicity margin and Lipschitz constant is bounded by specialising the radius theorem for monotone mappings [10, Thm. 4] to the MonDEQ setting (Theorem 2, Section IV-A); this gives explicit conditions under which the quantized MonDEQ retains existence, uniqueness, and linear convergence (Corollary 1, instantiating the forward–backward splitting framework of [11]). The fixed-point displacement between quantized and full-precision equilibria is bounded and converted into a condition number (Theorems 3–4, Section IV-B), sharpening the closest prior MonDEQ Lipschitz analysis [12] by deriving the perturbed margin from the quantizer bit-width. The same threshold certifies the backward solve as well as the forward solve (Theorem 5, Section IV-C), so a single margin check governs both training and inference. We validate the certificate empirically on a single-layer MonDEQ trained on MNIST across bit-widths from 3 to 32 bits (Section V); the experiments test certificate predictiveness against the threshold $\|\Delta W\|_2 < m$, not deployment-scale benchmark performance. Code is available at `https://github.com/JLi-Projects/mondeq-quant`.

B. Related Work

Quantization theory. Standard quantization modeling treats the quantized weight matrix as a bounded perturbation of its full-precision counterpart [5], [13]; post-training and quantization-aware variants (Section V) trade off training

cost against achievable bit-width [5], [14]. *Inexact operator splitting.* Operator splitting methods such as forward–backward and Peaceman–Rachford admit inexact variants in which bounded per-step errors are tolerated while preserving convergence [15], [16]. In Section IV, we apply these results to quantization-induced errors in the MonDEQ solver and derive new bounds on equilibrium displacement and the associated condition number.

Numerical error analysis. Beuzeville et al. [17] prove backward stability of feedforward networks under floating-point rounding; Jonkman et al. [18] model quantized communication in distributed optimization as inexact Krasnosel'skiĭ–Mann iteration.

MonDEQ sensitivity. Pabbaraju et al. [12] derive input-output and weight-output Lipschitz bounds for MonDEQs, but their perturbation bound assumes the perturbed margin is known and does not address quantization-specific structure, convergence conditions, or condition number.

II. Preliminaries

We collect notation and standard definitions from monotone operator theory that are used throughout the paper.

We work in $\mathbb{R}^n$ with the Euclidean norm $\|\cdot\|_2$ and denote the spectral norm of a matrix by $\|\cdot\|_2$. The symmetric and skew-symmetric components of a matrix A are $\mathrm{sym}(A) := \frac{1}{2}(A + A^\top)$ and $\mathrm{skw}(A) := \frac{1}{2}(A - A^\top)$.

Monotone operators. Monotonicity generalises positive-definiteness from linear maps to nonlinear ones and is the property that guarantees well-posedness of the fixed-point equations we analyse. An operator $F : \mathbb{R}^n \to \mathbb{R}^n$ is *monotone* if $\langle F(x) - F(y), x - y\rangle \geq 0$ for all $x, y \in \mathbb{R}^n$, *maximal* if its graph is not properly contained in the graph of any other monotone operator, *m-strongly monotone* if $\langle F(x) - F(y), x - y\rangle \geq m\|x - y\|_2^2$, and *$L$-Lipschitz* if $\|F(x) - F(y)\|_2 \leq L\|x - y\|_2$. For the affine operator $F(z) = (I - W)z - (Ux + b)$, the strong monotonicity margin is $m = \lambda_{\min}(\mathrm{sym}(I - W))$ and the Lipschitz constant is $L = \|I - W\|_2$ [11], [19].

Resolvents. The resolvent $J_{\alpha G} := (I + \alpha G)^{-1}$ of a maximal monotone G plays the role of the activation function inside the splitting iteration; it is single-valued, firmly nonexpansive, and hence 1-Lipschitz. The reflected resolvent is $R_{\alpha G} := 2J_{\alpha G} - I$ [11].

The *forward–backward* iteration $z^{k+1} = J_{\alpha G}\big(z^k - \alpha F(z^k)\big)$ converges linearly for any $\alpha \in (0, 2m/L^2)$ with contraction modulus $r_{\mathrm{FB}} = \sqrt{1 - 2\alpha m + \alpha^2 L^2}$ [11], [19]. The *Peaceman–Rachford* iteration $z^{k+1} = (2J_{\alpha G} - I)\big((2J_{\alpha F} - I)(z^k)\big)$ converges linearly for any $\alpha > 0$ with contraction modulus $\rho_{\mathrm{PR}} = \sqrt{1 - \frac{4\alpha m}{(1+\alpha L)^2}}$ [11], [19].

III. Monotone Operator Equilibrium Networks

Monotone operator equilibrium networks (MonDEQs) [1] compute their output as the fixed point of a splitting map derived from a monotone inclusion. We summarize the key definitions.

Definition 1. *Fix an input $x \in \mathbb{R}^d$. Let $W \in \mathbb{R}^{n\times n}$, $U \in \mathbb{R}^{n\times d}$ and $b \in \mathbb{R}^n$ be parameters collected in a vector $\boldsymbol{\vartheta} \in \mathbb{R}^r$. Define the affine map*

$$F(z) \;:=\; (I - W)z - (Ux + b), \qquad z \in \mathbb{R}^n.$$

Let $G : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be a maximal monotone operator and let $J_{\alpha G} := (I + \alpha G)^{-1}$ denote its resolvent for any $\alpha > 0$. Considering the nonlinear fixed point iteration

$$z^{k+1} = J_{\alpha G}\big(z^k - \alpha F(z^k)\big) := \Phi(z^k; \boldsymbol{\vartheta}),$$

*suppose it has a fixed point $z^\star$. We call the mapping from the input x to fixed point $z^\star$ a **monotone operator equilibrium network** (MonDEQ).*

A MonDEQ thus replaces the layer stack of a feedforward network with a single nonlinear fixed-point equation: the iteration Φ alternates an affine pre-activation step (parameterised by W, U, b) with the resolvent of G (which plays the role of the activation), and the network's output is its equilibrium. The following equivalence [1] recasts this fixed point as the solution of a monotone inclusion, opening the door to operator-splitting theory.

Theorem 1. *Define a MonDEQ as in Definition 1. Then $z^\star \in \mathrm{Fix}(\Phi) \iff 0 \in F(z^\star) + G(z^\star)$.*

Theorem 1 reduces computation of the MonDEQ output to solving the monotone inclusion $0 \in F(z^\star) + G(z^\star)$. This reformulation is useful because the splitting algorithms of monotone operator theory apply directly and converge linearly when F is strongly monotone. The choice of G encodes the activation: when $G = \partial\rho$ for proper, closed, convex ρ, the resolvent $J_{\alpha G} = \mathrm{prox}_{\alpha\rho}$ acts as the activation function in the splitting iteration, with ρ the indicator of $\mathbb{R}^n_{\geq 0}$ recovering the rectified linear unit (ReLU) activation [1].

Any W guaranteeing F is m-strongly monotone can be written in the form below, giving a constructive recipe for training MonDEQs with a target margin built in.

Proposition 1. *$\mathrm{sym}(I - W) \succeq mI$ if and only if there exist $A, B \in \mathbb{R}^{n\times n}$ such that $W = (1 - m)I - A^\top A + B - B^\top$.*

Proof. Direct computation [1]. □

The margin m is determined by the parameterization of W. Because $m = \lambda_{\min}(\mathrm{sym}(I - W))$ is an explicit function of W, perturbing the weight matrix perturbs m in a way that can be bounded analytically. Since $m > 0$ is both necessary and sufficient for well-posedness, bounding how quantization perturbs m directly determines whether the quantized network remains well-posed.

IV. Quantization in a MonDEQ

Here, quantization replaces floating-point weights with fixed-point (low-bit) approximations, reducing memory and enabling efficient integer arithmetic at the cost of increased rounding error. We analyze the resulting error as a perturbation of the weight matrix $W \to \widehat{W} = W + \Delta W$ [13], bounding its effect on well-posedness, the equilibrium point, and the backward pass used for training.

This article has been accepted for publication in IEEE Control Systems Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LCSYS.2026.3703783

We use symmetric uniform (mid-tread) quantization: for b-bit representation with weights in $[-1, 1]$, the quantizer $Q_\Delta(w) = \Delta \cdot \mathrm{round}(w/\Delta)$ has step size $\Delta = 2^{1-b}$ and worst-case elementwise error $\Delta/2$. Uniform quantization is standard for weight compression because the evenly spaced levels map directly to fixed-point integer formats, enabling hardware-accelerated matrix arithmetic; non-uniform schemes such as logarithmic quantizers [8] sacrifice this property. Since each entry of ΔW is bounded by $\Delta/2$, we have $\|\Delta W\|_2 \le (\Delta/2)n$ for the square $n \times n$ weight matrix. This motivates modeling weight quantization as a bounded perturbation [13].

Definition 2. *Given a MonDEQ as in Definition 1, its* quantized counterpart *replaces W with $\widetilde{W} = W + \Delta W$, $\|\Delta W\|_2 \le \varepsilon_W$.*

For the symmetric uniform quantizer with step size $\Delta = 2^{1-b}$ at b bits, $\varepsilon_W = n\Delta/2$.

Weight quantization introduces a deterministic perturbation to the weight matrix. This raises the question of how large the perturbation can be before the equilibrium ceases to exist. In practice, each iterate also incurs computational errors such as finite-precision arithmetic or activation rounding, so the computed iterates obey $z^{k+1} = \widetilde{\Phi}(z^k) + \delta_k$ with bounded per-step errors δ_k. Together, the weight perturbation ΔW and the iterate errors δ_k model the two sources of error in a quantized MonDEQ.

A. Margin Perturbation and Well-Posedness

The following theorem shows that weight perturbation reduces the monotonicity margin by at most $\|\Delta W\|_2$. This theorem is a specialisation of the radius theorem for monotone mappings [10, Theorem 4].

Theorem 2. *Define a MonDEQ in accordance with Definition 1 with weights W satisfying Proposition 1. Let $\widetilde{W}$ be the quantized weights with perturbation $\|\Delta W\|_2 \le \varepsilon_W$, and let $\widetilde{F}(z) := (I - \widetilde{W})z - (Ux + b)$ denote the corresponding quantized affine operator. Then the margin $\widetilde{m} := \lambda_{\min}(\mathrm{sym}(I - \widetilde{W}))$ of $\widetilde{F}$ is bounded below by*

$$\widetilde{m} \;\ge\; m - \|\Delta W\|_2,$$

and the Lipschitz constant $\widetilde{L}$ of $\widetilde{F}$ satisfies $|L - \|\Delta W\|_2| \le \widetilde{L} \le L + \|\Delta W\|_2$. In particular, $\widetilde{F}$ is strongly monotone (with margin $\widetilde{m} > 0$) whenever $\|\Delta W\|_2 < m$.

Proof. $\mathrm{sym}(I - \widetilde{W}) = \mathrm{sym}(I - W) - \mathrm{sym}(\Delta W)$, so by Rayleigh [20],

$$\begin{aligned}\widetilde{m} &= \min_{\|x\|_2 = 1} x^\top \big[\mathrm{sym}(I - W) - \mathrm{sym}(\Delta W)\big] x \\ &\ge m - \|\mathrm{sym}(\Delta W)\|_2 \ge m - \|\Delta W\|_2 .\end{aligned}$$

For the Lipschitz constant, the triangle and reverse triangle inequalities applied to $\widetilde{L} = \left\|I - \widetilde{W}\right\|_2 = \|(I - W) - \Delta W\|_2$ give the stated bounds. □

If $\|\Delta W\|_2 < m$ then $\widetilde{m} > 0$ and the equilibrium is preserved; in the worst case the condition number $\widetilde{\kappa} = \widetilde{L}/\widetilde{m}$ degrades from both sides, slowing convergence. The margin is the binding constraint in practice: $\mathrm{sym}(I - W) = mI + A^\top A$ attains m exactly wherever $A^\top A$ has a zero eigenvalue, while $L = \|I - W\|_2$ is robust to elementwise rounding. The Peaceman–Rachford analogue substitutes $\rho_{\mathrm{PR}}(\alpha; \widetilde{m}, \widetilde{L})$.

Corollary 1. *If $\varepsilon_W < m$ and $\alpha \in (0, 2\widetilde{m}/\widetilde{L}^2)$, the quantized forward–backward map $\widetilde{\Phi}_{\mathrm{FB}}(z) := J_{\alpha G}\big(z - \alpha \widetilde{F}(z)\big)$ is a contraction with modulus $r_{\mathrm{FB}}(\alpha; \widetilde{m}, \widetilde{L})$.*

Proof. Replace (m, L) by $(\widetilde{m}, \widetilde{L})$ from Theorem 2 in the forward–backward convergence rate. □

In words, provided $\varepsilon_W < m$, weight quantization slows but does not break convergence: the solver still reaches a unique equilibrium, and the next subsection bounds how far that equilibrium moves. Larger perturbations can drive $\widetilde{m} \le 0$ and break convergence, as in the 4-bit case of Section V.

B. Equilibrium Displacement

Convergence guarantees the quantized solver reaches *some* fixed point, but a controller deployed at low precision needs to know how far that fixed point has moved from the one the controller was designed for. The next result bounds the displacement $\|\widetilde{z}^\star - z^\star\|_2$ in terms of the perturbation size and the (unperturbed) margin.

Theorem 3. *Assume $F(z) = (I - W)z - (Ux + b)$ is m-strongly monotone and $G : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is monotone. With $\widetilde{W}$ given as in Definition 2, suppose $\|\Delta W\|_2 < m$ (in particular $\widetilde{m} > 0$). Let*

$$\widetilde{F}(z) := (I - \widetilde{W})z - (Ux + b) = F(z) - \Delta W z.$$

Let $z^\star$ and $\widetilde{z}^\star$ denote the (unique) solutions of the full-precision and quantized inclusions

$$0 \in F(z^\star) + G(z^\star), \qquad 0 \in \widetilde{F}(\widetilde{z}^\star) + G(\widetilde{z}^\star).$$

Then

$$\|\widetilde{z}^\star - z^\star\|_2 \;\le\; \frac{\|\Delta W\|_2}{m}\, \|\widetilde{z}^\star\|_2. \tag{1}$$

Proof. Pick $g^\star \in G(z^\star)$, $\widetilde{g}^\star \in G(\widetilde{z}^\star)$ with $F(z^\star) + g^\star = 0$, $\widetilde{F}(\widetilde{z}^\star) + \widetilde{g}^\star = 0$. Subtracting (using $\widetilde{F} = F - \Delta W$) and taking the inner product with $\delta z := \widetilde{z}^\star - z^\star$,

$$\langle F(\widetilde{z}^\star) - F(z^\star),\, \delta z\rangle - \langle \Delta W \widetilde{z}^\star,\, \delta z\rangle + \langle \widetilde{g}^\star - g^\star,\, \delta z\rangle = 0.$$

The first term is $\ge m\,\|\delta z\|_2^2$ (m-strong monotonicity of F); the third is ≥ 0 (monotonicity of G). Hence $m\,\|\delta z\|_2^2 \le \|\Delta W\|_2\, \|\widetilde{z}^\star\|_2\, \|\delta z\|_2$ by Cauchy–Schwarz; dividing yields (1). □

The bound (1) depends on $\|\widetilde{z}^\star\|_2$ rather than $\|z^\star\|_2$ because the perturbation acts through the shifted fixed point. Exchanging F and $\widetilde{F}$ in the proof gives the symmetric bound $\|z^\star - \widetilde{z}^\star\|_2 \le (\|\Delta W\|_2 / \widetilde{m})\, \|z^\star\|_2$. An explicit *relative* bound in terms of $\|z^\star\|_2$ alone is given in Corollary 4.

For hardware deployment, U and b are also quantized; the same argument with $\Delta u := \Delta U\, x + \Delta b$ extends the bound.

Corollary 2. *Under the hypotheses of Theorem 3 with $\widetilde{U} = U + \Delta U$, $\widetilde{b} = b + \Delta b$, and $\Delta u := \Delta U\, x + \Delta b$, $\|z^\star - \widetilde{z}^\star\|_2 \le (\|\Delta W\|_2\, \|\widetilde{z}^\star\|_2 + \|\Delta u\|_2)/m$.*

Proof. The argument of Theorem 3 applies with $\widetilde{F}(z) - F(z) = -\Delta W z - \Delta u$; Cauchy–Schwarz on the additional $\langle \Delta u, \delta z \rangle$ term yields the extra $\|\Delta u\|_2$. □

In words, quantizing U and b shifts the equilibrium by at most $\|\Delta u\|_2 / m$ but does not threaten convergence: the margin and Lipschitz constant of Theorem 2 depend only on $(I - \widetilde{W})$, so Corollary 1's convergence guarantee extends verbatim. The next result accounts for the second error source, *iterate quantization*: per-step residuals from finite-precision arithmetic or activation rounding.

Corollary 3. *Let $\widetilde{\Phi}$ be a quantized map as in Corollary 1, with contraction modulus $r \in (0,1)$ and fixed point $\widetilde{z}^\star$. Then*

$$\limsup_{k\to\infty} \|z^k - \widetilde{z}^\star\|_2 \;\le\; \frac{\limsup_{k\to\infty} \|\delta_k\|_2}{1-r}.$$

If $\sum_{k=0}^{\infty} \|\delta_k\|_2 < \infty$, then $z^k \to \widetilde{z}^\star$ exactly.

Proof. Follows from standard inexact contraction results [11, Sec. 5.5]. □

In practice, bounded per-step errors do not destroy convergence: the solver reaches a neighbourhood of $\widetilde{z}^\star$ whose radius scales with the error magnitude and contraction rate. Summability $\sum \|\delta_k\|_2 < \infty$ holds, for example, under an *adaptive quantizer* (one whose step size shrinks across iterations) chosen so that $\|\delta_k\|_2$ decays geometrically [18], in which case the total error $\|z^k - z^\star\|_2 \le \|z^k - \widetilde{z}^\star\|_2 + \|\widetilde{z}^\star - z^\star\|_2$ collapses to the displacement bound alone.

The bound (1) measures displacement in absolute terms. We now derive a relative bound and extract the condition number, which separates the problem's inherent sensitivity from the perturbation size.

Corollary 4. *Under the hypotheses of Theorem 3, if $\|\Delta W\|_2 < m$ then*

$$\frac{\|z^\star - \widetilde{z}^\star\|_2}{\|z^\star\|_2} \;\le\; \frac{\|\Delta W\|_2}{m - \|\Delta W\|_2}. \tag{2}$$

Proof. From Theorem 3, $\|\widetilde{z}^\star - z^\star\|_2 \le \frac{\|\Delta W\|_2}{m} \|\widetilde{z}^\star\|_2$. Substituting $\|\widetilde{z}^\star\|_2 \le \|z^\star\|_2 + \|\widetilde{z}^\star - z^\star\|_2$ gives $\|\widetilde{z}^\star - z^\star\|_2 \le \frac{\|\Delta W\|_2}{m}(\|z^\star\|_2 + \|\widetilde{z}^\star - z^\star\|_2)$. Rearranging, $(1 - \|\Delta W\|_2/m)\,\|\widetilde{z}^\star - z^\star\|_2 \le \frac{\|\Delta W\|_2}{m}\|z^\star\|_2$, which yields (2) since $\|\Delta W\|_2 < m$. □

Corollary 4 gives a global bound: the relative displacement is at most $\|\Delta W\|_2/(m - \|\Delta W\|_2)$, which depends only on the perturbation size and margin. For example, on the trained MonDEQ of Section V at 8 bits ($\|\Delta W\|_2 = 0.035$, $m = 0.227$), the bound gives 18%; the empirical relative error is much smaller (Section V). As $\|\Delta W\|_2 \to 0$, the bound linearizes to $\|\Delta W\|_2 / m$, recovering the condition number scaling of Theorem 4.

The sensitivity of the equilibrium to small weight perturbations is captured by the condition number [13], [21].

Theorem 4. *For an unquantized MonDEQ with margin $m > 0$, the absolute condition number*

$$\kappa_{\text{abs}} := \limsup_{\|\Delta W\|_2 \to 0} \frac{\|\widetilde{z}^\star - z^\star\|_2}{\|\Delta W\|_2}$$

satisfies $\kappa_{\text{abs}} \le \|z^\star\|_2 / m$.

Proof. From Theorem 3, $\|z^\star - \widetilde{z}^\star\|_2 / \|\Delta W\|_2 \le \|\widetilde{z}^\star\|_2 / m$. By Corollary 4, $\|\widetilde{z}^\star\|_2 \le \|z^\star\|_2/(1 - \|\Delta W\|_2/m)$, so $\|z^\star - \widetilde{z}^\star\|_2 / \|\Delta W\|_2 \le \|z^\star\|_2/(m - \|\Delta W\|_2)$. Taking $\|\Delta W\|_2 \to 0$ gives $\kappa_{\text{abs}} \le \|z^\star\|_2 / m$. □

In words, the equilibrium's sensitivity to weight perturbation is governed by the ratio of its magnitude to the margin. The relative condition number $\kappa_{\text{rel}} \le \|W\|_2 / m$ gives $\|z^\star - \widetilde{z}^\star\|_2 / \|z^\star\|_2 \le \kappa_{\text{rel}}\, \eta_W$ to first order, where $\eta_W := \|\Delta W\|_2 / \|W\|_2$; on the trained MNIST model $\kappa_{\text{rel}} \approx 7.6$. A sufficient pre-deployment check is $\varepsilon_W < m$, and a single margin check $\widetilde{m} > 0$ guarantees both forward and backward convergence (Theorem 5).

Unlike feedforward networks, where rounding errors accumulate through L layers as $O(Lu)$ [17], contractivity bounds the error here regardless of iteration count: $\widetilde{z}^\star$ is exact for $I - \widetilde{W}$. The next subsection shows the backward solve inherits these guarantees verbatim.

C. Backward Pass Under Quantization

Training a MonDEQ requires gradients of the loss with respect to the parameters $\boldsymbol{\vartheta} = (W, U, b)$, computed by implicit differentiation through the equilibrium condition $0 \in F(z^\star; \boldsymbol{\vartheta}) + G(z^\star)$. The key observation is that the resulting backward problem is itself a monotone inclusion with the same linear part $(I - W)$ as the forward problem, and therefore inherits the same margin and convergence guarantees.

Differentiating $F(z^\star; \boldsymbol{\vartheta}) + g^\star = 0$ in $\boldsymbol{\vartheta}$ via the chain rule produces the backward inclusion $0 \in (I - W)p - r + G_b(p)$, where $p := \frac{\mathrm{d}z^\star}{\mathrm{d}\boldsymbol{\vartheta}}$ is the backward sensitivity, $r := \frac{\mathrm{d}W}{\mathrm{d}\boldsymbol{\vartheta}} z^\star + \frac{\mathrm{d}U}{\mathrm{d}\boldsymbol{\vartheta}} x + \frac{\mathrm{d}b}{\mathrm{d}\boldsymbol{\vartheta}}$ collects the parameter-derivatives of the affine forcing, and $G_b \in \partial_C G(z^\star) \subseteq \mathbb{R}^{n\times n}$ is an element of the Clarke generalized Jacobian — the convex hull of limits of Jacobians at points of differentiability — satisfying $\mathrm{sym}(G_b) \succeq 0$ [22]. The following theorem shows this structure is preserved under weight quantization.

Theorem 5. *Let $\widetilde{W} = W + \Delta W$ with $\|\Delta W\|_2 < m$, and let $\widetilde{z}^\star$ solve $0 \in \widetilde{F}(\widetilde{z}^\star; \boldsymbol{\vartheta}) + G(\widetilde{z}^\star)$. Define $\widetilde{p} := \frac{\mathrm{d}\widetilde{z}^\star}{\mathrm{d}\boldsymbol{\vartheta}}$, $\widetilde{r} := \frac{\mathrm{d}\widetilde{W}}{\mathrm{d}\boldsymbol{\vartheta}} \widetilde{z}^\star + \frac{\mathrm{d}U}{\mathrm{d}\boldsymbol{\vartheta}} x + \frac{\mathrm{d}b}{\mathrm{d}\boldsymbol{\vartheta}}$, and let $\widetilde{G}_b \in \partial_C G(\widetilde{z}^\star) \subseteq \mathbb{R}^{n\times n}$ with $\mathrm{sym}(\widetilde{G}_b) \succeq 0$. Then $\widetilde{p}$ solves*

$$0 \in (I - \widetilde{W})\widetilde{p} - \widetilde{r} + \widetilde{G}_b(\widetilde{p}), \tag{3}$$

and under the stepsize hypothesis of Corollary 1, the splitting method converges to $\widetilde{p}$ with the perturbed parameters $(\widetilde{m}, \widetilde{L})$ from Theorem 2. In particular, if the forward pass converges ($\widetilde{m} > 0$), then the backward pass also converges with the same contraction modulus; a single margin check suffices for both passes.

Proof. Differentiating $\widetilde{F}(\widetilde{z}^\star; \boldsymbol{\vartheta}) + \widetilde{g}^\star = 0$ with respect to $\boldsymbol{\vartheta}$ yields (3). The backward operator $(I - \widetilde{W})p - \widetilde{r}$ has the same linear part as $\widetilde{F}$, so it inherits the same $(\widetilde{m}, \widetilde{L})$ from Theorem 2. Since $\mathrm{sym}(\widetilde{G}_b) \succeq 0$ by hypothesis — equivalently, $\widetilde{G}_b$ is monotone as a linear operator — the same splitting method converges. □

This article has been accepted for publication in IEEE Control Systems Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LCSYS.2026.3703783

Theorem 5 validates quantization-aware training (QAT): whenever the forward pass converges under quantized weights, gradients can be computed at the same precision and with the same iteration budget. No additional solver resources are required for the backward pass.

The gradient error under quantization has two sources: the displaced equilibrium ($z^\star \to \widetilde{z}^\star$) and the perturbed weight matrix ($W \to \widetilde{W}$). By Theorem 5, the backward sensitivity $\widetilde{p}$ solves a monotone inclusion with the same linear operator $(I - \widetilde{W})$, so the backward equilibrium exists and can be computed by the same splitting method. Since both sources introduce perturbations of size $O(\|\Delta W\|_2)$ (the weight perturbation directly, and the equilibrium displacement via Theorem 3), the chain rule gives $\left\| \frac{\partial \ell}{\partial W} - \frac{\partial \ell}{\partial \widetilde{W}} \right\|_2 = O(\|\Delta W\|_2)$.

V. Numerical Experiments

We validate the predictions of Section IV on a single-layer MonDEQ with $n = 100$ hidden units trained on MNIST (Adam, lr $= 10^{-3}$, 15 epochs, step decay $\gamma = 0.1$ at epoch 10). Unlike [1], which fixes m, we treat m as learnable via $m = \mathrm{softplus}(m_{\mathrm{raw}})$ with $m_{\mathrm{raw}} \in \mathbb{R}$, ensuring $m > 0$. The trained model achieves 98.22% test accuracy with $m = 0.227$, $L = 1.845$, $\kappa = L/m = 8.13$. Post-training quantization (PTQ) applies symmetric uniform quantization with step $\Delta = 2^{1-b}$ and per-tensor scaling, without calibration or bias correction [5], [14]; QAT retrains from scratch with a straight-through estimator. The forward–backward solver terminates when the relative residual falls below 10^{-5} or after 2000 iterations.

Margin stability certificate. Figure 1 tests the convergence condition $\|\Delta W\|_2 < m$ from Theorem 2 across bit-widths 3–32. The non-convergence/convergence transition aligns with $\|\Delta W\|_2 / m = 1$: 3-bit (ratio 5.36) and 4-bit (2.66) diverge, 5-bit and above converge. The 5-bit case (ratio 1.25, $\widetilde{m} = 0.045 > 0$) illustrates that the condition is sufficient but not necessary: the actual margin remains positive, so the solver converges despite the sufficient condition being violated. Iteration count reflects the degraded margin (5-bit ∼1730, 8-bit ∼450). At 8 bits, weight storage drops 4× versus 32-bit floating-point with 98.24% vs. 98.22% accuracy.

QAT vs. PTQ. Theorem 5's backward-pass guarantee makes QAT well-defined (it requires differentiating through the equilibrium). Figure 2 compares PTQ and QAT at 4, 6, 8 bits. PTQ fails at 4 bits ($\widetilde{m} = -0.142$); QAT learns weights with $\widetilde{m} = 0.006 > 0$, achieving 96.78% accuracy at a smaller margin ($m = 0.184$ vs. 0.227). At 6–8 bits both methods converge, with PTQ slightly higher (98.25/98.29%) by inheriting the larger float margin.

Displacement bound validation. The preceding experiments test *convergence*; we now test the *accuracy* of the converged equilibrium. Theorem 3 bounds the displacement between exact equilibria; the forward–backward solver terminates at finite tolerance, so a Cauchy–Schwarz residual-to-state argument $\|\hat{z} - z^\star\|_2 \le \|F(\hat{z}) + \hat{g}\|_2 / m$ for any $\hat{g} \in G(\hat{z})$ (immediate from m-strong monotonicity, cf. [11, Sec. 5.5]) combined with Theorem 3 gives a corrected observable bound

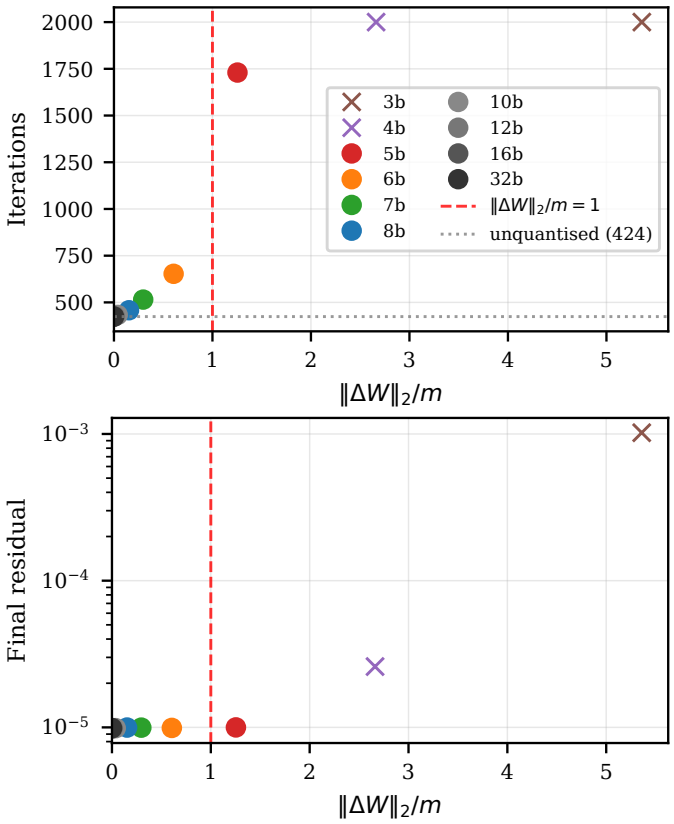

Fig. 1. Margin stability certificate. Iterations to convergence (top) and final residual (bottom) vs. normalized perturbation $\|\boldsymbol{\Delta W}\|_{\mathbf{2}} / \boldsymbol{m}$; each point is one bit-width (3–32 bits). The vertical dashed line marks the sufficient condition $\|\boldsymbol{\Delta W}\|_{\mathbf{2}} / \boldsymbol{m} = \mathbf{1}$; the horizontal dotted line in the top panel is the unquantized baseline (424 iterations). Circles: converged (relative residual $< \mathbf{10^{-5}}$); crosses: did not converge within 2000 iterations.

that absorbs the solver tolerance. Figure 3 illustrates Corollary 2's bound on 2,560 randomly sampled test inputs at 6, 8, 12, and 16 bits, with W, U, b all quantized at the same bit-width. The maximum $\|\Delta u\|_2$ ranges from 2.44 (6-bit) to 0.002 (16-bit), and the empirical displacement is 3–10× lower than the bound for every sample.

VI. Conclusions

We have analyzed the effect of weight quantization on monotone operator equilibrium networks through spectral perturbation of the monotone inclusion. The monotonicity margin m emerges as the single quantity governing robustness to quantization: convergence of the forward and backward solvers is guaranteed provided $\|\Delta W\|_2 < m$ (Theorem 2), the equilibrium displacement satisfies $\|\widetilde{z}^\star - z^\star\|_2 \le (\|\Delta W\|_2 / m)\, \|\widetilde{z}^\star\|_2$ (Theorem 3), and the relative condition number $\kappa_{\mathrm{rel}} = \|W\|_2 / m$ links bit-width to forward error (Theorem 4). Experiments confirm a phase transition at the predicted threshold and show the displacement bound holds on every tested sample across 6–16 bits, with a conservative factor of 3–10×. Quantization-aware training recovers convergence at 4 bits where post-training quantization fails, enabled by the backward-pass guarantee of Theorem 5.

The analysis is limited to uniform symmetric quantization

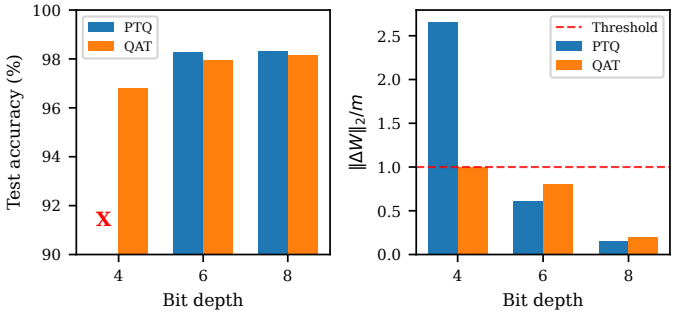

Fig. 2. QAT vs. PTQ at 4, 6, and 8 bits. Left: test accuracy (%; a red X indicates PTQ non-convergence at 4 bits). Right: $\|\boldsymbol{\Delta W}\|_2/m$; the dashed line marks $\|\boldsymbol{\Delta W}\|_2/m = 1$.

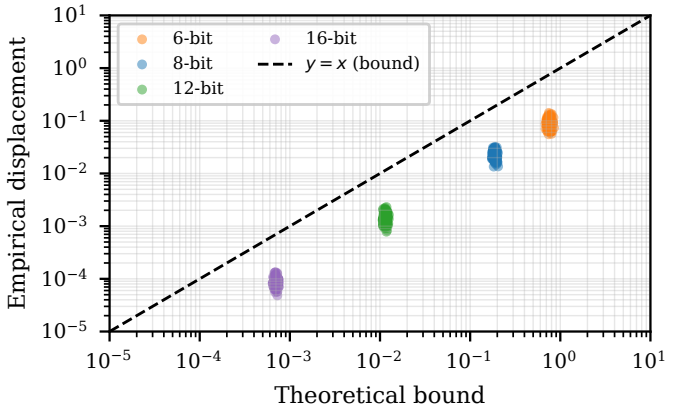

Fig. 3. Displacement bound validation at 6, 8, 12, and 16 bits with $\boldsymbol{W}$, $\boldsymbol{U}$, and $\boldsymbol{b}$ all quantized. Each point is one test sample; axes show relative quantities (log–log). The x-axis is the Corollary 2 bound $(\|\boldsymbol{\Delta W}\|_2\,\|\tilde{z}^\star\|_2 + \|\boldsymbol{\Delta u}\|_2)/(m\,\|z^\star\|_2)$ with $\boldsymbol{\Delta u} := \boldsymbol{\Delta U}\,x + \boldsymbol{\Delta b}$; the y-axis is the empirical relative displacement $\|\tilde{z}^\star - z^\star\|_2 / \|z^\star\|_2$. Points below the dashed line ($y = x$) satisfy the bound.

of a single-layer MonDEQ; natural extensions include per-channel and mixed-precision schemes, multi-layer architectures, and margin-aware regularization. An important open question is whether the structural guarantees of equilibrium-based control components survive weight quantization. Recurrent equilibrium networks (RENs) [2] — the related dynamic architecture in which equilibrium-based controllers are currently deployed [3], [4] — are the natural next target, and the bounds here are a first step. Another avenue for future work is to extend these results to realizations of MonDEQs on quantized analog hardware [23].

ACKNOWLEDGMENT

Generative AI was used to assist with the experimentation code, finding references, and checking for grammatical errors.

REFERENCES

[1] E. Winston and J. Z. Kolter, "Monotone operator equilibrium networks," in *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[2] M. Revay, R. Wang, and I. R. Manchester, "Recurrent Equilibrium Networks: Flexible Dynamic Models With Guaranteed Stability and Robustness," *IEEE Transactions on Automatic Control*, vol. 69, no. 5, pp. 2855–2870, May 2024.

[3] N. Junnarkar, H. Yin, F. Gu, M. Arcak, and P. Seiler, "Synthesis of stabilizing recurrent equilibrium network controllers," in *2022 IEEE 61st Conference on Decision and Control (CDC)*, 2022, pp. 7059–7064, arXiv:2204.00122.

[4] R. Wang, N. H. Barbara, M. Revay, and I. R. Manchester, "Learning over all stabilizing nonlinear controllers for a partially-observed linear system," *IEEE Control Systems Letters*, vol. 7, pp. 91–96, 2022, arXiv:2112.04219.

[5] M. Nagel, M. Fournarakis, R. A. Amjad, Y. Bondarenko, M. van Baalen, and T. Blankevoort, "A white paper on neural network quantization," 2021, arXiv:2106.08295.

[6] Y. Zhang, F. Song, and J. Sun, "QEBVerif: Quantization Error Bound Verification of Neural Networks," in *Computer Aided Verification*, C. Enea and A. Lal, Eds. Cham: Springer Nature Switzerland, 2023, pp. 413–437.

[7] A. Kabaha and D. D. Cohen, "Quantization with Guaranteed Floating-Point Neural Network Classifications," *Proc. ACM Program. Lang.*, vol. 9, no. OOPSLA2, pp. 340:1893–340:1920, Oct. 2025.

[8] M. Fu and L. Xie, "The sector bound approach to quantized feedback control," *IEEE Transactions on Automatic Control*, vol. 50, no. 11, pp. 1698–1711, 2005.

[9] S. Bai, J. Z. Kolter, and V. Koltun, "Deep equilibrium models," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[10] A. L. Dontchev, A. Eberhard, and R. T. Rockafellar, "Radius Theorems for Monotone Mappings," *Set-Valued and Variational Analysis*, vol. 27, no. 3, pp. 605–621, Sep. 2019.

[11] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, ser. CMS Books in Mathematics. Cham: Springer International Publishing, 2017.

[12] C. Pabbaraju, E. Winston, and J. Z. Kolter, "Estimating Lipschitz constants of monotone deep equilibrium models," in *International Conference on Learning Representations*, 2021.

[13] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, 2nd ed. SIAM, 2002.

[14] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko, "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 2704–2713.

[15] J. Eckstein and D. P. Bertsekas, "On the Douglas—Rachford splitting method and the proximal point algorithm for maximal monotone operators," *Mathematical Programming*, vol. 55, no. 1, pp. 293–318, Apr. 1992.

[16] P. L. Combettes and J.-C. Pesquet, "Proximal Splitting Methods in Signal Processing," in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, H. H. Bauschke, R. S. Burachik, P. L. Combettes, V. Elser, D. R. Luke, and H. Wolkowicz, Eds. New York, NY: Springer, 2011, pp. 185–212.

[17] T. Beuzeville, A. Buttari, S. Gratton, and T. Mary, "Deterministic and probabilistic rounding error analysis of neural networks in floating-point arithmetic," *IMA Journal of Numerical Analysis*, 2025.

[18] J. A. Jonkman, T. Sherson, and R. Heusdens, "Quantisation Effects in Distributed Optimisation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018, pp. 3649–3653.

[19] E. K. Ryu and S. Boyd, "A primer on monotone operator methods," *Applied and Computational Mathematics*, vol. 15, no. 1, pp. 3–43, 2016.

[20] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge: Cambridge University Press, 1985.

[21] T. Beuzeville, "Backward error analysis of artificial neural networks with applications to floating-point computations and adversarial attacks," Ph.D. dissertation, Université de Toulouse, 2024.

[22] P. L. Combettes and J.-C. Pesquet, "Deep neural network structures solving variational inequalities," *Set-Valued and Variational Analysis*, vol. 28, pp. 491–518, 2020.

[23] T. Chaffey, "Circuit realization and hardware linearization of monotone operator equilibrium networks," Sep. 2025, arXiv:2509.13793.