ZeroProofML: Epsilon-Free Rational Neural Layers via Transreal Arithmetic

Zsolt Döme zpml.dome@outlook.com

Independent Researcher

Editor:

Abstract

We introduce ZeroProofML, a framework for deterministic, ε-free rational neural layers based on transreal (TR) arithmetic. By totalizing division (and other singular operations) with explicit tags (REAL, $\pm\infty$, Φ), TR removes ad-hoc ε knobs and yields reproducible semantics for singularities. We formalize TR autodiff (Mask-REAL) and give stability statements (bounded updates, batch-safe steps). On 2R inverse kinematics, TR matches overall accuracy and achieves 1.5–2.5× lower error in the closest near-singularity bins (B0–B1), modest improvements in B2 (∼3–4%), and near parity elsewhere, with stable closed-loop behavior. Results extend to planar 3R and synthetic 6R, supporting robustness near rank-deficient Jacobians.

Keywords: transreal arithmetic, rational layers, singularities, robotics IK, reproducibility

Preliminaries

We use the transreal domain $\mathbb{T} = \mathbb{R} \cup \{+\infty, -\infty, \perp\}$ with tags $\{\mathrm{REAL}, \mathrm{INF}, \mathrm{NULL}\}$. Values are pairs (v, τ) with $v \in \overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$. Arithmetic on $\mathbb{T}$ follows explicit tag rules (addition/multiplication/division, integer powers, and guarded $\sqrt{\cdot}$).

Positioning & Practicality

TR totalization targets models with explicit singular structure (e.g., rational layers P/Q, guarded roots/logs, Jacobian-based control). It is *not* intended to replace standard deep models when singularities are not the failure mode. Use TR when deterministic, analyzable behavior near poles is required; otherwise classical components suffice.

Scope of Totality

Definition 1 (Admissible class $\mathcal{F}_{\mathrm{TR}}$) *The least class of total maps $f : \mathbb{T}^n \to \mathbb{T}$ that contains constants and projections and is closed under TR-totalized $+$, $-$, $\times$, $/$, integer powers, guarded $\sqrt{\cdot}$, composition, and tupling. Optionally includes a chosen finite set of transcendental primitives (e.g., $\log$) when equipped with explicit TR-totalization policies (branch/guard/tag rules).*

Proposition 2 (Totality within $\mathcal{F}_{\mathrm{TR}}$) *Every $f \in \mathcal{F}_{\mathrm{TR}}$ is total on $\mathbb{T}^n$. If inputs are REAL and the classical f_{cl} is defined, then $\mathrm{tag}(f(x)) = \mathrm{REAL}$ and $\mathrm{val}(f(x)) = f_{\mathrm{cl}}(\mathrm{val}(x))$; at poles/indeterminate forms, non-REAL tags are returned per the primitive rules.*

Remark 3 (Transcendentals) *Claims of totality are limited to $\mathcal{F}_{\mathrm{TR}}$. Primitives beyond* log *and $\sqrt{\cdot}$ are out of scope unless explicitly totalized.*

Scope & Composability

Standard components. ReLU is total and TR-consistent. For `sigmoid/tanh/softmax/layernorm` we provide: (i) TR-policy variants (explicit guards for `exp/log/div`), or (ii) rational/Padé surrogates with uniform error on compact training ranges. Mixed stacks preserve TR guarantees on the rational backbone and classical behavior elsewhere.

When TR helps. Poles/constraints/control/analytic layers $\Rightarrow$ TR; ordinary MLP/CNN without divisions $\Rightarrow$ classical.

IEEE–TR Bridge

Define $\Phi : \mathsf{IEEE} \to \mathbb{T}$ (total) and $\Psi : \mathbb{T}_{\mathrm{REAL/INF}} \to \mathsf{IEEE}$ (round-to-nearest-even; undefined on $\perp$). Mapping table: finite $\mapsto (v, \mathrm{REAL})$; $\pm 0 \mapsto (0, \mathrm{REAL})$ with recorded IEEE zero sign; $\pm\infty \mapsto (\pm\infty, \mathrm{INF})$; NaN $\mapsto (*, \mathrm{NULL})$.

Lemma 4 (Partial homomorphism) *If IEEE evaluates $x \circ y$ ($\circ \in \{+, -, \times, /\}$) without NaN, then $\Phi(x) \circ_{\mathbb{T}} \Phi(y) = \Phi(x \circ y)$. Divisions by ± 0 match signs.*

Signed zeros. We retain the IEEE zero sign in a latent flag used only when directional limits matter (e.g., $1/\pm 0$).

Export. $\Psi(v, \mathrm{REAL}) = \mathrm{round}(v)$; $\Psi(\pm\infty, \mathrm{INF}) = \pm\infty$; $\Psi(\perp)$ undefined (or map to NaN by explicit policy). The bridge is a faithful embedding on non-NaN cases and a conservative *extension* elsewhere.

Autodiff with Tags: Mask-REAL

Let nodes be $z_k = F_k(z_{i_1}, \ldots, z_{i_m})$ with $F_k \in \mathcal{F}_{\mathrm{TR}}$. Each primitive has a REAL-mask predicate $\chi_k \in \{0, 1\}$ that is 1 iff all inputs and the evaluation are REAL-tagged.

Definition 5 (Mask-REAL gradient) *Backprop uses gates: $\bar{z}_i \mathrel{+}= \chi_k \, \bar{z}_k \, \partial_{z_i} F_k|_{\mathrm{REAL}}$ along edge $z_i \to z_k$. When $\chi_k = 0$, either drop the term or use a bounded surrogate S_k (Remark* **??***).*

Lemma 6 (REAL-path equivalence) *If all nodes are REAL on an open set U and f_{cl} is C^1 on U, then Mask-REAL equals the classical gradient on U.*

Lemma 7 (Chain rule with tag gating) *For $f = g \circ h$, $\nabla f_{\mathrm{MR}}(x) = J_g(h(x)) M_g(x) J_h(x) M_h(x)$ where $M_\bullet$ are diagonal masks of local χ_k.*

Proposition 8 (Bounded update under saturation) *Assume: (i) loss Lipschitz with constant L_ℓ; (ii) REAL derivatives bounded by B_k or surrogates S_k with norm $\le G_{\max}$; (iii) step size $\eta \le \eta_{\max}$. Then $\|\Delta\theta\| \le \eta C$ with C depending on L_ℓ, depth, and $\{B_k\}, G_{\max}$. In particular, choosing $\eta_{\max} = c/(L_\ell \prod_k \max\{B_k, G_{\max}\})$ ensures $\|\Delta\theta\| \le c$.*

Remark 9 (Saturation) *Use a smooth saturator $\sigma(a) = a/\sqrt{1 + (a/G_{\max})^2}$ to keep bounded gradients when $\chi_k = 0$.*

Hybrid Switching: Mask-REAL $\leftrightarrow$ Saturated

Let Γ denote pole hypersurfaces. Diagnostics: distance $d(x) = \mathrm{dist}(x, \Gamma)$ and local sensitivity $g_k = \|\nabla_z F_k\|$ on REAL values. Choose thresholds $0 < \delta_{\text{on}} < \delta_{\text{off}}$ and $0 < g_{\text{on}} < g_{\text{off}}$.

Aggregator choice. Max/min in the triggers may be replaced by robust quantiles (e.g., 90th percentiles of d and g) or any Lipschitz aggregator without affecting the finite-switching and descent guarantees.

Definition 10 (Hysteretic hybrid) *Mode $m_t \in \{\mathrm{MR}, \mathrm{SAT}\}$. Switch to SAT if $d_t \le \delta_{\text{on}}$ or $\max_k g_k \ge g_{\text{on}}$; switch to MR if $d_t \ge \delta_{\text{off}}$ and $\max_k g_k \le g_{\text{off}}$; otherwise keep m_t.*

Lemma 11 (No chattering) *With hysteresis ($\delta_{\text{off}} > \delta_{\text{on}}$, $g_{\text{off}} > g_{\text{on}}$) and continuous trajectories between steps, the number of switches on a compact interval is finite.*

Proposition 12 (Bounded updates under hybrid) *For $\eta \le c/(L_\ell \prod_k \max\{B_k, G_{\max}\})$, we have $\|\Delta\theta\| \le c$ regardless of switching times.*

Sufficient Conditions for Finite Switching

Theorem 13 (Finite/zero-density switching) *Assume (i) hysteresis margins $\delta_{\text{off}} > \delta_{\text{on}}$, $g_{\text{off}} > g_{\text{on}}$; (ii) batch-safe steps $\eta_t \le 1/\widehat{L}_{\mathcal{B}_t}$; (iii) bounded inputs in a compact set and coverage quotas preventing persistent dwelling in $\Gamma_{\delta_{\text{on}}}$. Then with probability 1 the number of mode switches on any finite horizon is finite (or has zero density), and convergence theorems in Sec.* **??** *apply.*

Proof [Proof sketch] Hysteresis yields nonzero travel distance between triggers; batch-safe steps bound state increments; the coverage controller reduces revisit frequency to the guard band. Hybrid-systems arguments imply finite switching on compact intervals. ■

Coverage Controller

Bucket by pole proximity: $B_0 = \{d \ge \Delta_2\}$, $B_1 = \{\Delta_1 \le d < \Delta_2\}$, $B_2 = \{d < \Delta_1\}$.

Distance estimator. We estimate $d(x)$ via $|Q(x)|/\|\nabla Q(x)\|_*$ (or basis-aware surrogates); any consistent positive estimator suffices. Constrained ERM:

$$\min_\theta \mathbb{E}[\ell(f(x;\theta), y)] \quad \text{s.t.} \quad \pi_1 \ge \alpha_1,\ \pi_2 \ge \alpha_2,\ \rho_{\text{flip}} \le \rho_{\max}. \tag{1}$$

Lagrangian with hinge surrogates: $\mathcal{L} + \lambda_1[\alpha_1 - \hat{\pi}_1]_+ + \lambda_2[\alpha_2 - \hat{\pi}_2]_+ + \mu[\hat{\rho}_{\text{flip}} - \rho_{\max}]_+$. Dual ascent on (λ, μ) yields an interpretable controller increasing pressure when quotas are violated. Standard primal–dual arguments give monotone decrease (up to $\mathcal{O}(\eta)$) and bounded constraint residuals under bounded variance.

Batch-Safe Learning Rate

Let $A_i = \|\nabla_\theta f(x^{(i)};\theta)\|$ and β_ℓ be the loss smoothness. Then the batch objective is $L_{\mathcal{B}}$-smooth with $L_{\mathcal{B}} \le \frac{\beta_\ell}{m}\sum_i A_i^2 \le \frac{\beta_\ell}{m}\sum_i (A_i^{\max})^2 =: \widehat{L}_{\mathcal{B}}$. Hence GD with $\eta \le 1/\widehat{L}_{\mathcal{B}}$ satisfies the standard descent lemma. A quantile-robust alternative uses $L_{\mathcal{B}}^{(q)} = \beta_\ell\,(A^{(q)})^2$. Combine with Prop. **??** via $\eta_t = \min\{\alpha/\widehat{L}_{\mathcal{B},t},\ c/(L_\ell \prod_k \max\{B_k, G_{\max}\})\}$.

Second-Order Derivatives and Momentum Stability

Assumptions. Work on a tag-stable REAL region U (no pole crossings), or use bounded saturated surrogates S_k when $\chi_k = 0$. On U, $f_{\text{cl}} \in C^2$; primitives have bounded first/second derivatives; surrogates are bounded by $G_{\max}$ (and optionally Lipschitz).

Hessian on REAL regions. If $\chi_k \equiv 1$ on U, then $\nabla^2 f_{\text{MR}}(x) = \nabla^2 f_{\text{cl}}(x)$ for all $x \in U$.

Across guard bands. With masks $M(x)$, $\nabla^2 f_{\text{MR}}(x) = M\,\nabla^2 f_{\text{cl}}(x)\,M + (\nabla M) * (\nabla f_{\text{cl}})$. Use piecewise-constant M or bounded surrogates; operator norms are bounded by local second-derivative bounds and $G_{\max}$.

Proposition 14 (Bounded curvature with saturation) *If $\|\nabla F_k\| \le B_k$, $\|\nabla^2 F_k\| \le H_k$ on REAL inputs, and surrogates satisfy $\|S_k\| \le G_{\max}$, $\|\nabla S_k\| \le H_{\max}$, then on any batch $\|\nabla^2 \mathcal{L}\| \le C_H := C_0\big(\sum_{paths}\prod_{k\in path} c_k\big)$ with $c_k \in \{B_k^2 + H_k,\ G_{\max}^2 + H_{\max}\}$.*

Gauss–Newton & Fisher. On REAL regions MR $\equiv$ classical; in SAT regions, bounded surrogates keep curvature finite.

Momentum and Adam

Heavy-ball/Polyak. $v_{t+1} = \beta_1 v_t + \nabla\mathcal{L}_{\mathcal{B}}(\theta_t)$, $\theta_{t+1} = \theta_t - \eta v_{t+1}$. Safe region: $\eta \le 2(1-\beta_1)/\widehat{L}_{\mathcal{B}}$.

Nesterov. Same bound under smoothness; restart on tag-flip spikes.

Adam/RMSProp. With bias-corrected moments and bounded gradients, effective per-coordinate step $\eta_{t,i}^{\text{eff}} \lesssim \eta/\sqrt{\hat{L}_{\mathcal{B},i}}$. A sufficient batch-safe condition is $\eta \le (1-\beta_1)\,/(\sqrt{1-\beta_2}\,\widehat{L}_{\mathcal{B}})$.

Identifiability

Rational layer $r = P/Q$ with parameters (p, q). Invariances: scaling $(cP)/(cQ)$ and common factors. Impose leading-1 on Q and coprimeness $\gcd(P, Q) = 1$.

Proposition 15 (Identifiability a.e.) *Assume (A1) leading-1 on Q, (A2) $\gcd(P, Q) = 1$, (A3) data support S has nonempty interior in the REAL region. If $r(\cdot;\theta_1) = r(\cdot;\theta_2)$ a.e. on S (and tag patterns agree), then $\theta_1 = \theta_2$, up to a null exceptional set of parameters.*

Sketch: If $P_1/Q_1 = P_2/Q_2$ on a set with an accumulation point away from poles, then $P_1Q_2 - P_2Q_1 \equiv 0$. With gcd and leading-1, this implies equality of coefficients. Locally (tag-stable neighborhood; full-rank design), the empirical risk is strictly convex on the constraint manifold, yielding an isolated minimizer.

Identifiability under manifold support. If the data support lies on a lower-dimensional manifold, identifiability holds *modulo* factors that vanish on the manifold. Coprime regularization via the Sylvester smallest singular value or resultant barriers discourages near-common-factor regimes.

Numerical Precision and Tag Robustness

Policy note (training vs evaluation). Guard-band thresholds $\tau_Q, \tau_P = \Theta(u)$ are part of the *training-time* tag policy: they classify REAL/INF/NULL deterministically near poles and trigger hybrid switching. They do not alter TR algebra; they govern tags and mode selection. Evaluation may use identical or stricter thresholds (policy-dependent).

Floating-point perturbations can flip tags near $\Gamma = \{Q = 0\}$. Define a guard band with thresholds $\tau_Q, \tau_P = \Theta(u)$ scaled by local sensitivities (e.g., $\|\nabla Q\|$, $\|\nabla P\|$). Classifier: REAL if $|Q| \geq \tau_Q$; INF if $|Q| < \tau_Q$ and $|P| \geq \tau_P$; NULL if both below thresholds. Use hysteresis ($\tau^{\text{on}} < \tau^{\text{off}}$); retain signed zero to preserve directional limits. Batch statistics π_{band} and ρ_{flip} feed the coverage controller.

Reproducibility as Policy-Determinism

Given a declared policy (ULP bands $\tau_{Q/P}$, rounding mode, signed-zero retention, deterministic reduction trees), tag classification is deterministic across runs and devices up to the stated ULP band. Outside guard bands misclassification cannot occur by Lemmas in Sec. **??**; inside, hysteresis enforces finite flips and stable behavior.

Robustness to Floating-Point Errors

Overflow/Underflow. TR tags absorb overflow as $\pm\infty$ (INF) with sign consistency; guard bands mitigate subnormal noise.

Mixed precision. Keep denominators/tags in master precision; safe downcast only when $|Q| \geq \tau_Q^{\text{off}}$; prefer stochastic rounding for accumulators.

Stable reductions. Use compensated or pairwise reductions and a deterministic reduction tree for order invariance.

Cross-hardware. Declare a device-agnostic ULP band for tag decisions and use deterministic kernels.

Error propagation. For $r = P/Q$, $|\Delta r| \lesssim (|\Delta P| + |r|\,|\Delta Q|)/|Q|$, motivating guard bands and hybrid switching.

Layer contracts. Publish $(B_k, H_k, G_{\max}, H_{\max})$ to tie into batch-safe LR and curvature bounds.

Global Stability and Convergence

Standing assumptions. (A1) Loss $\ell(\hat{y}, y)$ is bounded below, β_ℓ-smooth and L_ℓ-Lipschitz. (A2) Primitives in $\mathcal{F}_{\mathrm{TR}}$; on REAL regions they are C^1/C^2. (A3) Hybrid policy and guard bands ensure finite switching and bounded gradients. (A4) Steps obey a diminishing or batch-safe constant rule (Sec. **??**).

Deterministic GD

For $\eta_t \le 1/\widehat{L}_{\mathcal{B}_t}$: $\mathcal{L}_{t+1} \le \mathcal{L}_t - \frac{\eta_t}{2}\|\nabla \mathcal{L}_t\|^2$, persisting across MR↔SAT switches by bounded gradients (Prop. **??**).

Theorem 16 (GD with diminishing steps) *If $\sum_t \eta_t = \infty$, $\sum_t \eta_t^2 < \infty$ and $\eta_t \le 1/\widehat{L}_{\mathcal{B}_t}$, then $\sum_t \eta_t \|\nabla \mathcal{L}_t\|^2 < \infty$ and $\liminf_t \|\nabla \mathcal{L}_t\| = 0$. If switching is finite or of zero density, every limit point is stationary for its mode.*

Theorem 17 (Linear rate under PL) *If a tag-stable neighborhood U satisfies PL and $\eta \le 1/\widehat{L}$, then with no switches in U: $\mathcal{L}(\theta_t) - \mathcal{L}^* \le (1-\mu\eta)^{t-t_0}\,(\mathcal{L}(\theta_{t_0}) - \mathcal{L}^*)$.*

SGD

With unbiased gradients, variance σ^2, and $\eta_t \le 1/\widehat{L}_{\mathcal{B}_t}$:

Theorem 18 (SGD convergence) *If $\sum_t \eta_t = \infty$, $\sum_t \eta_t^2 < \infty$, then $\liminf_t \mathbb{E}\|\nabla \mathcal{L}(\theta_t)\| = 0$. Under PL and constant $\eta \le c/\widehat{L}$: $\mathbb{E}[\mathcal{L}(\theta_t) - \mathcal{L}^*] \;\le\; (1-\mu\eta)^t(\mathcal{L}(\theta_0) - \mathcal{L}^*) + \frac{\eta\sigma^2}{2\mu}$.*

Experimental Setup

Tasks. Planar 2R IK with $|\det J| \approx |\sin\theta_2|$ (primary), planar 3R (rank drop by alignment), and synthetic 6R (serial DH).

Datasets. 2R: stratified by $|\det J|$ with edges $[0, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, \infty)$; near-pole coverage ensured in train/test. 3R: stratified by manipulability ($\sigma_1\sigma_2$). 6R: stratified by $d_1 = \sigma_{\min}(J)$.

Baselines. MLP; Rational+ε (grid); smooth surrogate $P/\sqrt{Q^2+\alpha^2}$ (grid); learnable-ε; ε-ensemble. Reference: DLS.

TR models. TR–Basic (Mask-REAL only). TR–Full: shared-Q TR–Rational heads with hybrid gradients, tag/pole heads, anti-illusion residual, coprime regularizer; coverage enforcement and TR policy hysteresis; batch-safe LR.

Metrics. Overall and per-bucket MSE (B0–B4); closed-loop tracking (task-space error, max $\|\Delta\theta\|$, failures). 3R: PLE, sign consistency across θ_2, θ_3, residual consistency. 6R: overall + selected bins.

Aggregation. 3 seeds (2R/6R), deterministic policy for TR; means±std reported across seeds. Scripts emit per-seed JSONs and LaTeX tables/figures used below.

Related Work

Rational neural networks model functions as P/Q with strong approximation guarantees (?); practical deployments often use ε-regularized denominators $Q + \varepsilon$ to avoid division-by-zero. Batch normalization and related techniques also rely on explicit ε (?). Transreal arithmetic provides totalized operations with explicit tags for infinities and indeterminate forms (??). Masking rules in autodiff have appeared in robust training and subgradient methods; our Mask-REAL rule formalizes tag-aware gradient flow, ensuring exact zeros through non-REAL nodes while preserving classical derivatives on REAL paths. Bounded (saturating) gradients near poles relate to gradient clipping and smooth surrogates, but here arise from a deterministic, tag-aware calculus under an explicit policy. We adopt standard optimizers (e.g., Adam (?)) and normalization variants (e.g., LayerNorm (?)) as needed in controlled baselines.

Limitations and Outlook

Our approach targets models with explicit singular structure (rational layers, Jacobian-based control) and declared tag policies; it is not a replacement for generic deep architectures without divisions. Extending empirical coverage to higher-DOF systems with full physics stacks (URDF/Pinocchio) and integrating TR policies with mainstream autodiff frameworks are promising directions.

Code and Data Availability

All code, dataset generators, per-seed results, aggregated CSVs, and LaTeX tables/figures are available at github.com/domezsolt/ZeroProofML. The repository records environment info and dataset hashes for reproducibility.

Conclusion

ZeroProofML replaces ε-based numerical fixes with a principled, tag-aware calculus that is total by construction. Mask-REAL autodiff, hybrid switching with bounded surrogates, coverage control, and policy determinism translate into empirical advantages: decisive near-pole accuracy (B0–B1), bounded updates and stable rollouts, and low across-seed variance under a declared policy. We expect these guarantees to benefit rational and control-oriented models where explicit singular structure is intrinsic.

Acknowledgments

We thank contributors to the open-source ZeroProofML codebase and reviewers for constructive feedback.