Supplementary Information

Local Second-Order Adjoint Dynamics for Implicit Neural Networks

Dino Vlahek[1*], Dijana Oreški[1], Matija Novak[1], Darko Andročec[1]

[1]Faculty of Organization and Informatics, University of Zagreb, Varaždin, Croatia
*Corresponding author: dvlahek@foi.hr

Supplementary Tables

Supplementary Table S1: Directed-spectrum control with the original real-interval coefficients over two local directed operators, linear and tanh problems, and five seeds. Convergence uses relative residual tolerance 10^{-6}.

ρ	AR converged	CAT converged	median AR/CAT ratio
0.80	20/20	20/20	1.42×
0.95	20/20	4/20	–
0.99	20/20	0/20	–

Supplementary Table S2: Oracle complex-spectrum diagnostics at $\rho = 0.95$. For the heterogeneous non-normal operator, the root factor is reported as the range over five independently generated local operators. Each operator is evaluated with four deterministic right-hand sides.

Operator	Parameterization	Root factor	Converged	Median actions
directed normal	real interval	1.030	0/20	–
directed normal	oracle ellipse	0.878	20/20	106
directed normal	exact spectrum	0.878	20/20	106
heterogeneous non-normal	real interval	0.973–1.025	8/20	–
heterogeneous non-normal	oracle ellipse	0.867–0.882	20/20	106.5
heterogeneous non-normal	exact spectrum	0.867–0.881	20/20	104

Supplementary Table S3: Feed-forward comparison over seven datasets and ten seeds. Differences are given relative to BP-Momentum.

Method	Δ loss	Δ accuracy	median time / BP
BP-Momentum	0	0	1.0×
AR64-Momentum	-1.6×10^{-4}	-9.5×10^{-5}	15.1×
CAT40-Momentum	$+1.07 \times 10^{-3}$	-1.38×10^{-3}	13.1×
CAT64-Momentum	$+2.4 \times 10^{-5}$	-3.2×10^{-5}	19.5×

Supplementary Methods and Analyses

Feed-Forward Experimental Details

The feed-forward study uses two moons, concentric circles, iris, wine, Wisconsin breast cancer, a synthetic anisotropic binary problem, and the scikit-learn digits dataset. A ten-class `synthetic_large` dataset with 100 input features provides the scaling condition. Data are divided into 60% training, 15% validation, and 25% test sets, with standardization fitted on the training data only.

The networks contain six affine blocks. Training uses minibatches of 64, validation-loss model selection, patience 20, minimum improvement 10^{-4}, and at most 150 epochs. Ten seeds are used for the core datasets and three for the scaling condition. The complete feed-forward study contains 861 successful runs.

The comparison reported here includes BP-Momentum, AR64, CAT40, and CAT64. AR64 and CAT64 use 64 synchronous local relaxation sweeps per backward pass, while CAT40 uses 40 sweeps. The CAT credit dynamics use $\Delta\tau_a = 0.04$ and $\gamma_a = \omega_a = 8$.

Maximum practical deviations relative to BP-Momentum are fixed at 0.005 in test accuracy and 0.001 in test cross-entropy. Uncertainty is estimated by hierarchical bootstrap resampling of datasets and seeds.

Architectures and Hyperparameters

The feed-forward architectures are listed as input width, five hidden widths, and output width:

Dataset	Architecture
moons, circles	$(2, 32, 32, 24, 16, 8, 2)$
iris	$(4, 32, 32, 24, 16, 8, 3)$
wine	$(13, 40, 40, 32, 24, 12, 3)$
breast cancer	$(30, 48, 48, 32, 24, 12, 2)$
anisotropic	$(20, 48, 48, 32, 24, 16, 2)$
digits	$(64, 64, 64, 48, 32, 16, 10)$
MNIST	$(784, 256, 256, 192, 128, 64, 10)$
CIFAR-10	$(3072, 512, 512, 384, 256, 128, 10)$
`synthetic_large`	$(100, 128, 128, 96, 64, 32, 10)$

For the fixed-operator implicit real-data experiment, the standard hidden width is 48 on the core datasets and 64 on `synthetic_large`, MNIST, and CIFAR-10. The recurrent radius is fixed at 0.95. Core training uses learning rate 0.04 and up to 35 epochs. The vision experiment uses learning rate 0.02 and up to 20 epochs. Both use momentum 0.9, batch size 64, weight decay 10^{-4}, forward tolerance 10^{-8}, backward tolerance 10^{-5}, and three seeds.

The trainable recurrent control uses hidden width 48, 20 seeds per dataset, learning rate 0.04 for the input/readout parameters, learning rate 0.005 for the recurrent edges, momentum 0.9, batch size 64, and up to 40 epochs. The recurrent row-sum cap is 0.95.

Reproducibility Details

The complete archived feed-forward suite records 861 successful runs and stores its configuration hash, package versions, and git commit. The present article reports the BP-Momentum, AR64, CAT40, and CAT64 conditions used for the feed-forward comparison.

The implicit spectral script records solver action counts, spectral intervals, paired ratios, and Chebyshev controls. The implicit dataset script stores training summaries and held-out direct-solve audits. Separate scripts reproduce the exact-error comparison with GMRES, Broyden, and Anderson, as well as the trainable and directed recurrent controls. Compact standard-run summaries are stored in `reference_results/` in the public repository.

CIFAR-10 Experiment

CIFAR-10 provides an additional higher-dimensional evaluation of the adjoint solve. The experiment uses 8000 real examples per seed, hidden width 64, three seeds, and the same fixed local recurrent operator as the main classification experiment. Mean test accuracy is 0.3728 for both AR and CAT. CAT uses 1.40× fewer local adjoint actions and is favoured in all three seed-level pairs. The audited Jacobian radius is approximately 0.518. Across the combined 48 held-out MNIST and CIFAR-10 audit samples, the Spearman association between the actual radius and the oracle CAT advantage is 0.956. CAT-oracle and Chebyshev-oracle have the same dataset-level median action count in the two image experiments.

The low absolute classification accuracy reflects the small non-convolutional architecture. The relevant result is the persistence of the exact-gradient solver behavior with a larger input map.