Eva Optimizer

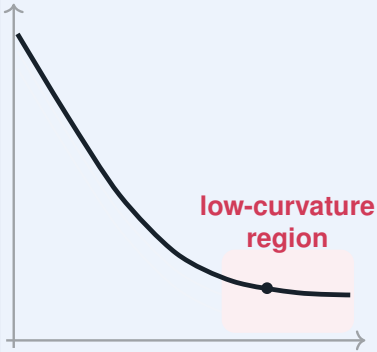
Escaping Low-Curvature Traps in Deep Learning

Preserve Adam if its step matches the local gradient scale; **correct only mismatched coordinates**.

A. Di Cecco · C. Metta · A. Papini · M. Fantozzi · S. G. Galfrè · M. Vegliò · **L. A. Bianchi** · M. Parton · F. Morandin

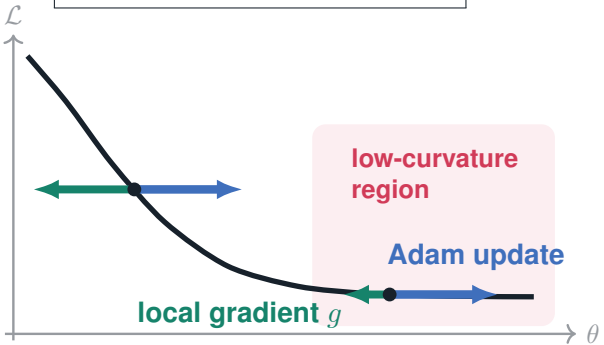
UCBM, CNR, Chalmers, Parma, Pisa, **Trento**, Chieti-Pescara; **CuriosAI**

2026.08.18, **ICPR 2026**



$$\Delta\theta^{(t)}_{\mathrm{Adam},i} = -\alpha \frac{\widehat{m}_i^{(t)}}{\sqrt{\widehat{v}_i^{(t)}} + \varepsilon}$$

$\widehat{m}_i^{(t)}$: moving average of recent gradients
$\widehat{v}_i^{(t)}$: moving average of squared gradients



In a slowly varying flat direction:

$|g_{t-1,i}|$ can become small,

$|\Delta\theta^{(t)}_{\mathrm{Adam},i}|$ need not shrink at the same rate.

One-dimensional intuition: low curvature is considered together with gradient stability.

Low-curvature / stable-gradient approximation

$$\widehat{m}_i^{(t)} \simeq g_{t-1,i}$$
$$\sqrt{v_i^{(t)}} \simeq |g_{t-1,i}|$$
$$\Rightarrow \quad |\Delta\theta_{\mathrm{Adam},i}^{(t)}| \simeq \alpha$$

$$\kappa_i^{(t)} = \frac{|\Delta\theta_{\mathrm{Adam},i}^{(t)}|}{|g_{t-1,i}|}$$

Component-wise scale mismatch indicator associated with the low-curvature regime.

$\kappa_i \leq 1$ **KEEP ADAM**

$\kappa_i > 1$ **CORRECT**

Operationally: compare the two magnitudes directly—no division is required.

$$\Delta\theta^{(t)}_{\mathrm{Eva},i} = \begin{cases} \Delta\theta^{(t)}_{\mathrm{Adam},i}, & \kappa_i^{(t)} \le 1, \\ \sigma\lambda g_{t-1,i}, & \kappa_i^{(t)} > 1 \end{cases}$$

Different coordinates, same iteration t

Adam	Adam	Eva	Adam	Eva	Adam
θ_1	θ_2	θ_3	θ_4	θ_5	θ_6

Only flagged coordinates change.

Adam is preserved everywhere else.

$\lambda = 0.1$

fixed default used in reported experiments

Eva$^+$ and Eva$^-$ differ only in the correction sign

Sign convention: $\theta_t = \theta_{t-1} + \Delta\theta_t$ First-order contribution: $g_{t,i}\Delta\theta_i = \sigma\lambda g_{t,i}g_{t-1,i}$

If $g_{t,i}g_{t-1,i} > 0$, the current and previous gradients are locally aligned.

Eva$^+$: $\sigma = +1$

$\Delta\theta_i = +\lambda g_{t-1,i}$

intended role

leave a flagged region

Eva$^-$: $\sigma = -1$

$\Delta\theta_i = -\lambda g_{t-1,i}$

intended role

reduce the flagged step

Separate variants: sign chosen before training; the gate tests magnitudes, not alignment.

The replacement is bounded by the gradient scale

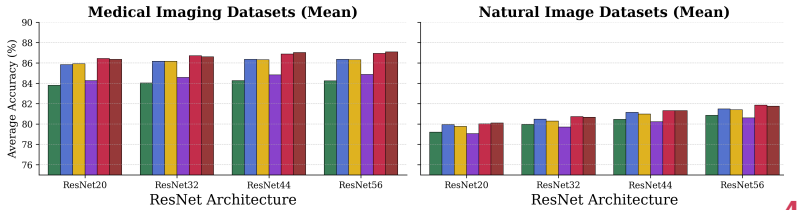
When the gate triggers and $0 < \lambda \leq 1$:

$$\boxed{|\Delta\theta_{\mathrm{Eva},i}| = \lambda|g_{t-1,i}| \leq |g_{t-1,i}| < |\Delta\theta_{\mathrm{Adam},i}|}$$

component-wise **no Hessian** **element-wise $O(d)$**

Implementation note: Eva needs access to g_{t-1}, typically via a previous-gradient buffer.

UNDER MATCHED HYPERPARAMETERS



SGD **Adam** **AdamW** **RMSProp** **Eva^{+}** **Eva^{-}**

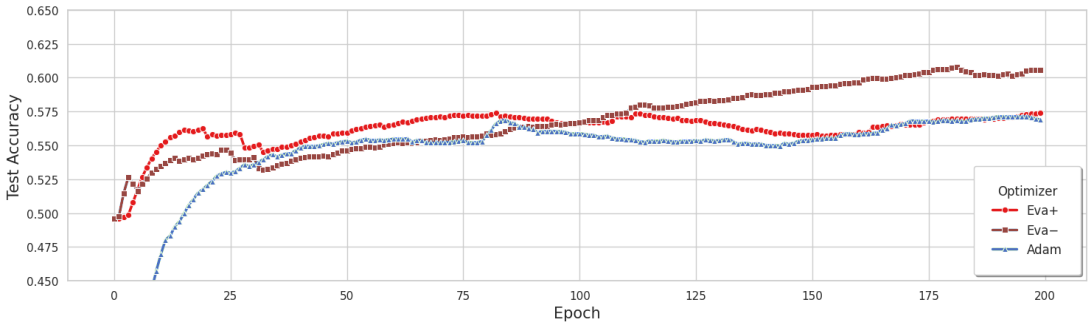
16 datasets $\times$ 4 ResNets $\cdot$ 5 independent runs

63/64
best or tied by an Eva variant

42 Eva^{+} unique best
17 Eva^{-} unique best
4 ties between Evas
1 Adam unique best

Reported sweep. Note that optimizer-specific tuning can alter rankings.

Representative run · RetinaMNIST · ResNet-56



Eva^{+}
faster early progress

Eva^{-}
accuracy in the long run

Reported gains extend beyond ResNets

Pretrained vision

4 families $\times$ 2 datasets

ConvNeXt · EfficientNet · InceptionNet
MobileNetV2 on two medical datasets

VAE reconstruction

4/4

Eva$^-$ best reported SSIM

Tabular classification

4/4

Eva$^+$ best reported accuracy

BERT / IMDB

92.3% Eva$^+$

vs AdamW 91.7% and Adam 91.5%

DETECT
compare magnitudes: adaptive update vs. previous gradient

→

ADAPT
flagged only: $+\lambda g$ or $-\lambda g$

→

PRESERVE
base adaptive update on unflagged coordinates

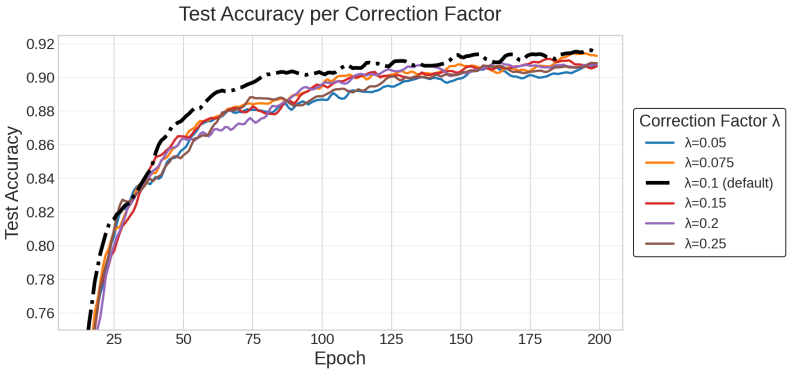
Eva changes only what the gate flags.

Flagged coordinates receive a controlled correction; unflagged coordinates use the Adam update.

CuriosAI/Eva on GitHub

We do not need to trust Adam equally in every coordinate, at every step.

Correction-factor ablation

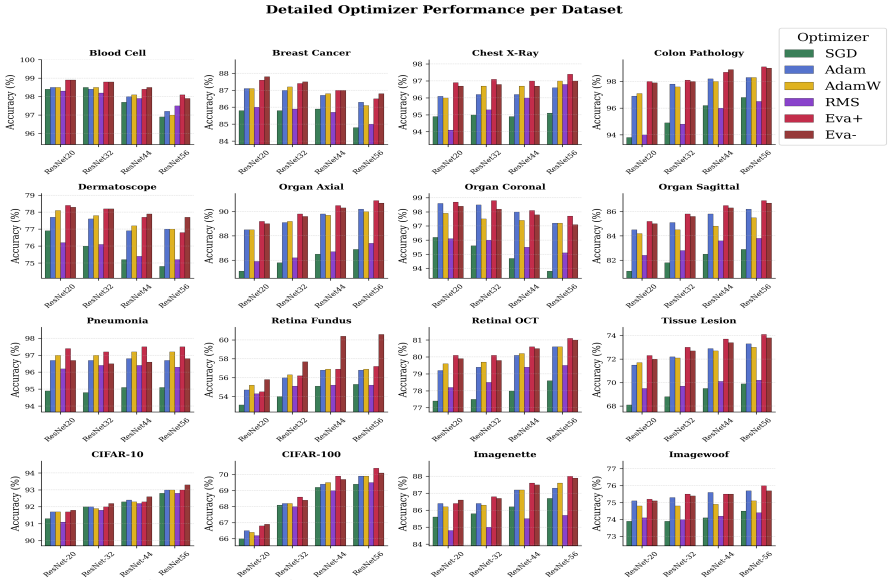


$$\lambda = 0.1$$

fixed default in the reported experiments.

This ResNet-20 / CIFAR-10 ablation supports the choice in this configuration, it is no proof of a universal optimum.

Detailed 64-setting results



Assumptions along coordinate i

- low curvature;
- gradient stability: $g_{t,i} \approx g_{t-1,i}$;
- near-zero ideal update: $\Delta\theta_i^* \approx 0$.

Local simplifying assumptions, not global training conditions.

$$|\Delta\theta_{\text{Adam},i}| \simeq \alpha$$

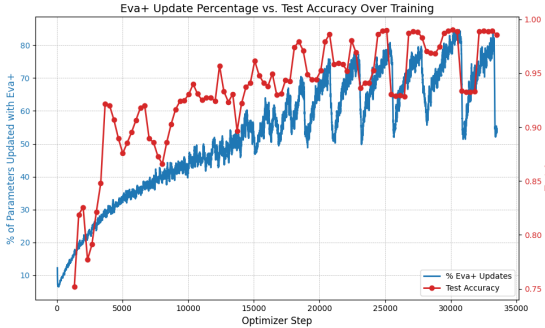
$$\text{gate active} \Rightarrow |g_{t-1,i}| < \alpha$$

$$e^2_{\text{Eva}} \approx \lambda^2 |g_{t-1,i}|^2 <$$

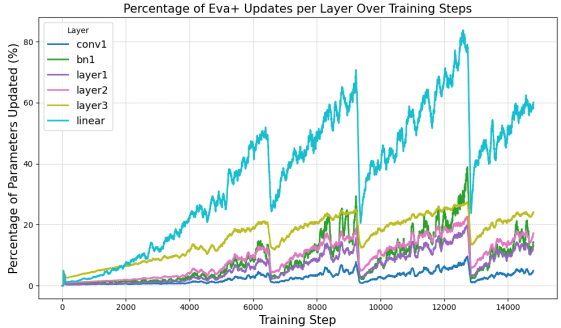
$$< \lambda^2\alpha^2 < \alpha^2 \approx e^2_{\text{Adam}}$$

For $\lambda = 0.1$, this model puts the Eva squared-error scale below 1% of α^2.

Gating dynamics by layer

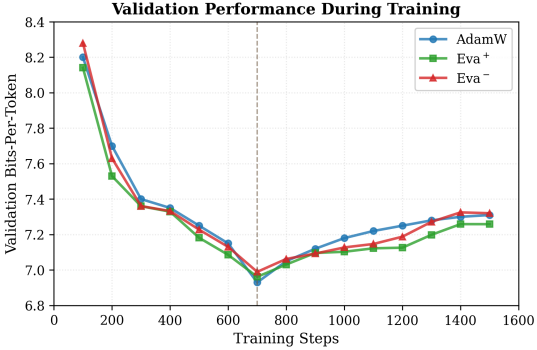


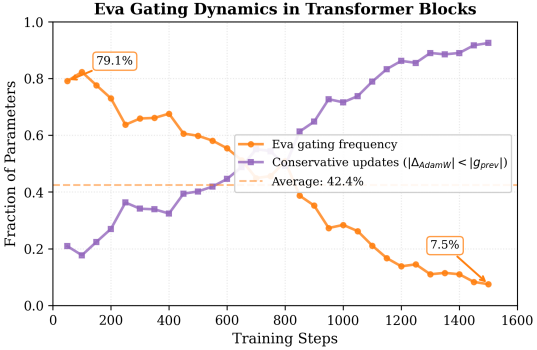
update usage vs. accuracy



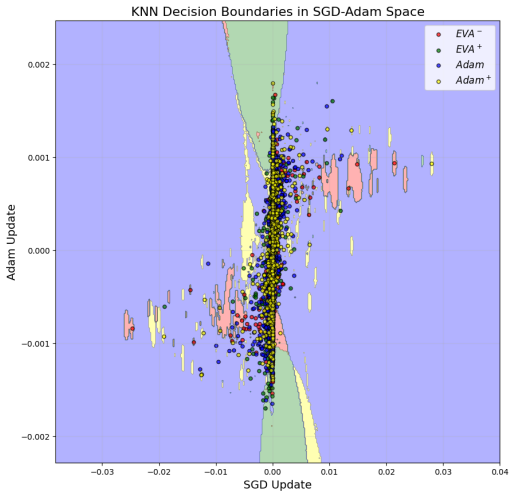
update usage across layers

In the reported PathMNIST experiment, activation evolves through training and differs substantially by layer.





Limitations: ∼10M-parameter Shakespeare model; selective gating; 50–50 blended update; heuristic blending ratio; no systematic ablation.



Protocol

50,000 recorded parameter-level triplets

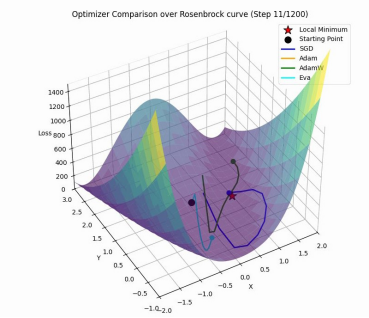
Candidate update labels: Adam, reversed Adam, SGD, reversed SGD

k-NN visualization with $k = 500$

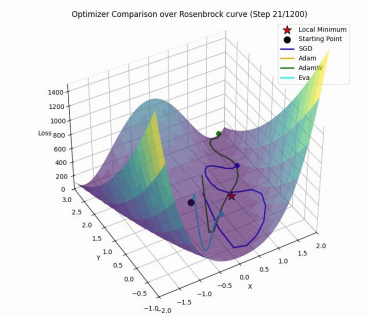
Reported pattern

Eva$^+$-like choices appear where Adam updates are large relative to gradient/SGD updates; Adam dominates where the scales are proportional.

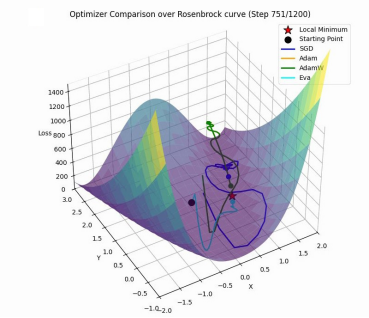
Only a supportive empirical analysis, not a proof.



step 11



step 21



step 751

Toy illustration: controlled Eva trajectories versus large Adam/AdamW excursions on the Rosenbrock surface.