CISPA
HELMHOLTZ CENTER FOR INFORMATION SECURITY

Hyperbolic Aware Minimization: Implicit Bias for Sparsity

Tom Jacobs, Advait Gadhikar , Celia Rubio-Madrigal, and Rebekka Burkholz

ICLR

Outline

1. Benefits of hyperbolic geometry
2. Problem: vanishing inverse metric of the reparameterization mw
3. Solution: Hyperbolic Aware Minimization (HAM)
4. Validation: Sharpeness Aware Minimization and Sparsity

Benefits of hyperbolic geometry

- Implicit bias: bias towards a (sparse) solution due to the geometry
- Hyperbolic: fast movement around zero
- Result: sparsity bias and sign flips

	Sparse implicit bias	Sign flips	No hard perturbations	No extra parameters
Dense training	✗	–	✓	✓
PILoT (Jacobs & Burkholz, 2025)	✓	✗	✓	✗
Sign-In (Gadhikar et al., 2025)	✓ (mild)	✓	✗	✗
HAM (ours)	✓ (mild)	✓	✓	✓

Table 1: HAM induces a mild L_1 bias and flips parameter signs more easily due to its inverse metric (see Fig. 1), which together lead to boosting sparse training without explicit overparameterization.

Riemannian gradient flow

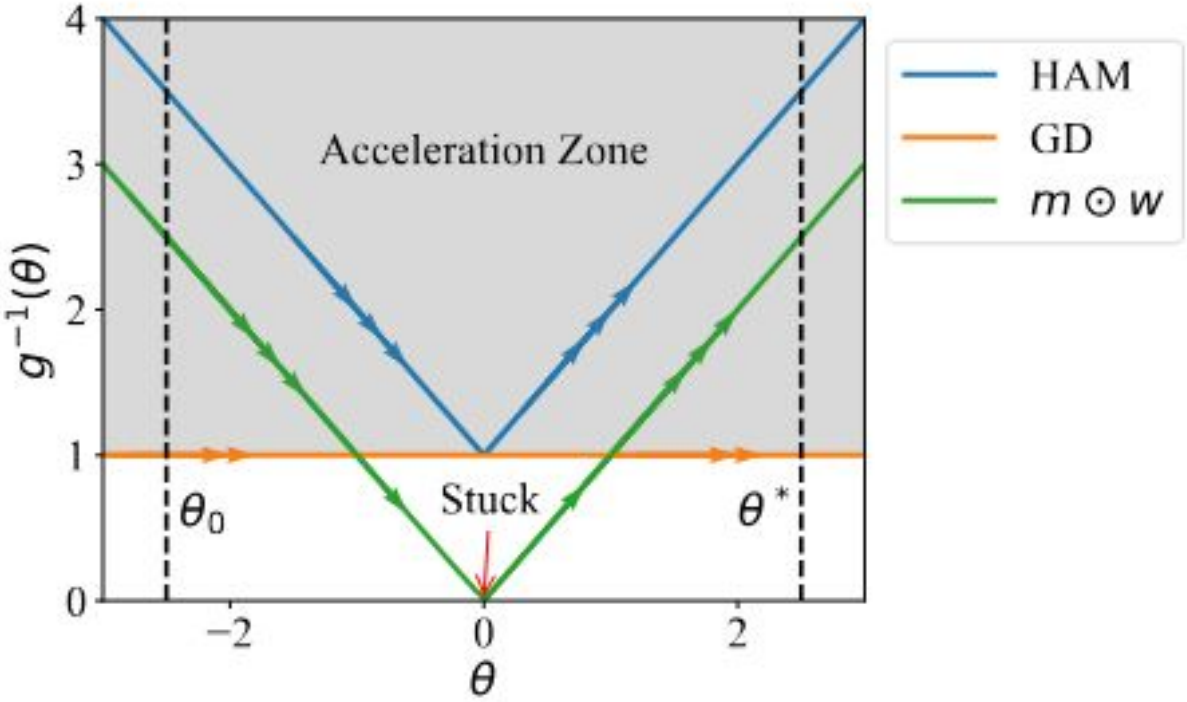
$$d\boldsymbol{\theta}_t = -g^{-1}(\boldsymbol{\theta}_t)\nabla f(\boldsymbol{\theta}_t)dt, \qquad \boldsymbol{\theta}_0 = \boldsymbol{\theta}_{init}.$$

Table 2: Inverse metrics of gradient descent, the overparameterization $\boldsymbol{m} \odot \boldsymbol{w}$, and HAM.

	GD	$\boldsymbol{m} \odot \boldsymbol{w}$	HAM
$g^{-1}(\boldsymbol{\theta})$	1	$\sqrt{\boldsymbol{\theta}^2 + \boldsymbol{\gamma}^2}$	$1 + \alpha\lvert\boldsymbol{\theta}\rvert$

Problem: vanishing inverse metric

- To little movement can hamper training by slow down and no sign flips



Solution: Hyperbolic Aware Minimization

Hyperbolic step:

$$\theta_k = \theta_{k+\frac{1}{2}} \odot \exp\left(-\eta\left(\alpha\ \mathrm{sign}(\theta_{k+\frac{1}{2}})\nabla f(\theta_k) + \beta\right)\right), \qquad \text{(HYP*)}$$

Algorithm:

Algorithm 1 HAM

Require: steps T, schedule η, initialization θ_{init}, constants $\alpha, \beta > 0$.

for $k \in 0 \dots T-1$ **do**

$\theta_{k+\frac{1}{2}} = \mathrm{OptimizerStep}(\nabla f(\theta_k), \eta)$

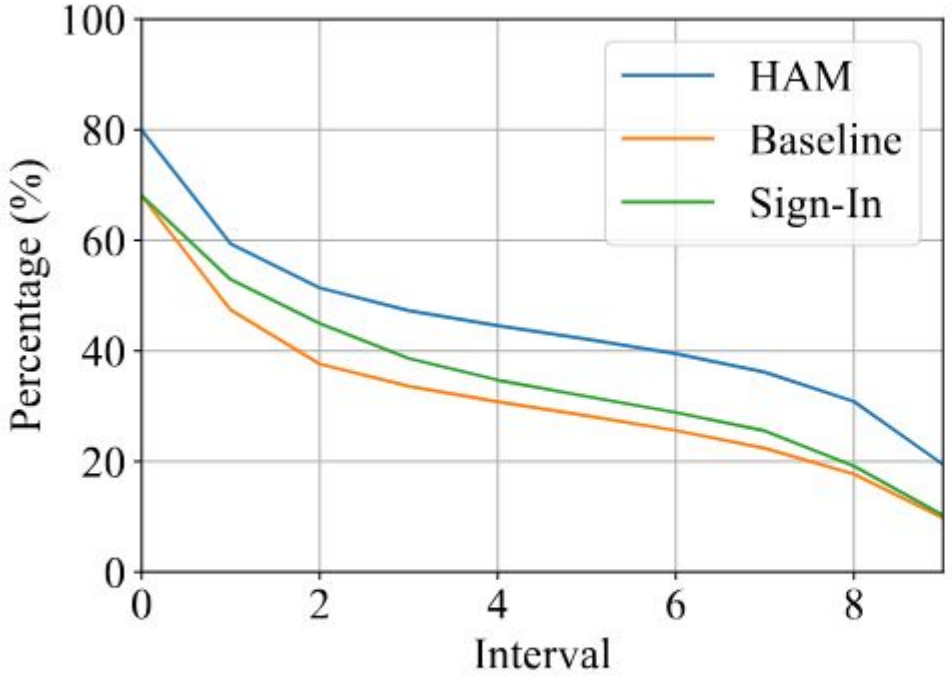
$\theta_{k+1} = \mathrm{HyperbolicStep}(\theta_{k+\frac{1}{2}}, \nabla f(\theta_k), \alpha, \beta, \eta)$ according to formula (HYP*)

end for

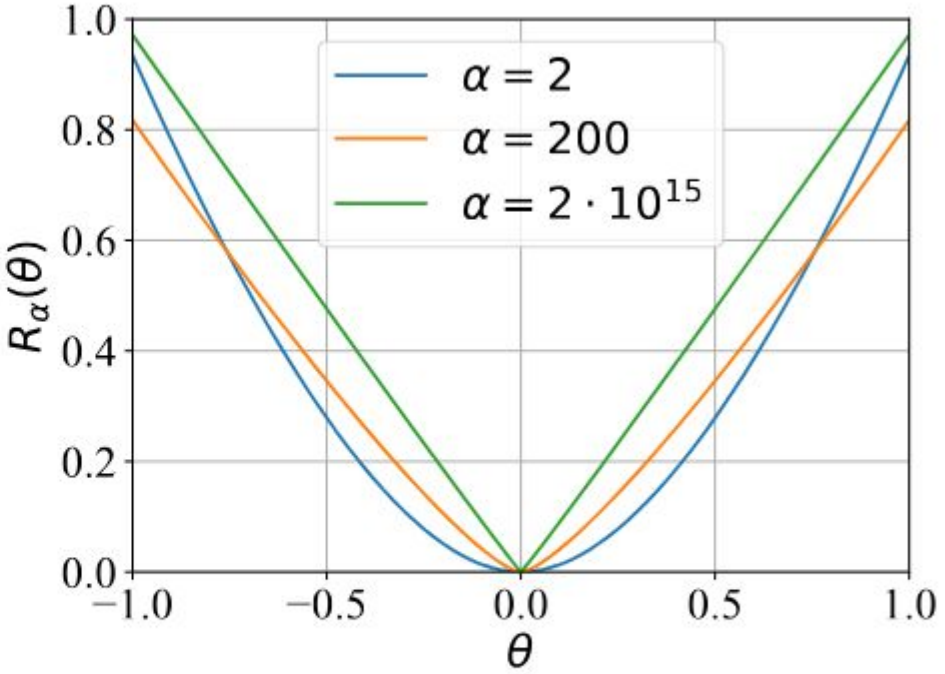
return Model weights θ_T

Theoretical mechanisms at work

Sign flip and mild sparsity promotion.



(a) Sign flips during training.



(b) HAM's Bregman function.

Compatible with SGD and SAM

- Sparsity beneficial for generalization.
- Sharpness and sparsity go hand in hand.

Sparsity	HAM	Baseline
0.9	1.384	0.104
0.8	59.008	0.424
0.7	71.940	37.776
0.5	76.824	73.404

Table 3: HAM improves dense training of a ResNet50 on ImageNet.

	100 epchs	200 epchs	+ SAM, 100 epchs	+ SAM, 200 epchs
Baseline	76.72 ± 0.19	77.27 ± 0.13	77.10 ± 0.21	77.94 ± 0.16
HAM	$\mathbf{77.51 \pm 0.11}$	$\mathbf{77.86 \pm 0.05}$	$\mathbf{77.92 \pm 0.15}$	$\mathbf{78.56 \pm 0.12}$

Improve sparse training algorithms

Table 4: Dense-to-sparse training and pruning at initialization with HAM on ImageNet with ResNet50.

Pruning type	Method	$s = 0.8$	$s = 0.9$	$s = 0.95$
PaI	Random	73.87(±0.06)	71.56(±0.03)	68.72(±0.05)
	Random + *Sign-In*	74.12(±0.09)	72.19(±0.18)	69.38(±0.1)
	Random + HAM	**74.84(±0.09)**	**72.72(±0.03)**	**70.05(±0.06)**
DtS	AC/DC	75.83(±0.02)	74.75(±0.02)	72.59(±0.11)
	AC/DC + *Sign-In*	75.9(±0.14)	74.74(±0.12)	72.88(±0.13)
	AC/DC + HAM	**77.2(±0.14)**	**76.66(±0.12)**	**75.45(±0.13)**
DST	RiGL	75.02(±0.1)	73.7(±0.2)	71.89(±0.07)
	RiGL + *Sign-In*	75.02(±0.1)	74.27(±0.08)	**73.07(±0.17)**
	RiGL + HAM	**76.22(±0.07)**	**74.83(±0.08)**	**72.93(±0.1)**
Cont. spars.	spred	72.64	71.84	69.47
	PILoT	75.62	74.73	71.3
	STR	75.49(±0.14)	72.4(±0.11)	64.94(±0.07)
	STR + HAM	**76.37(±0.18)**	**75.01(±0.02)**	**71.41(±0.1)**

Takeaway

Redesigning optimization can improve sparse training.