Adam, Gradient Descent and Zipf's law

Frederik Kunstner

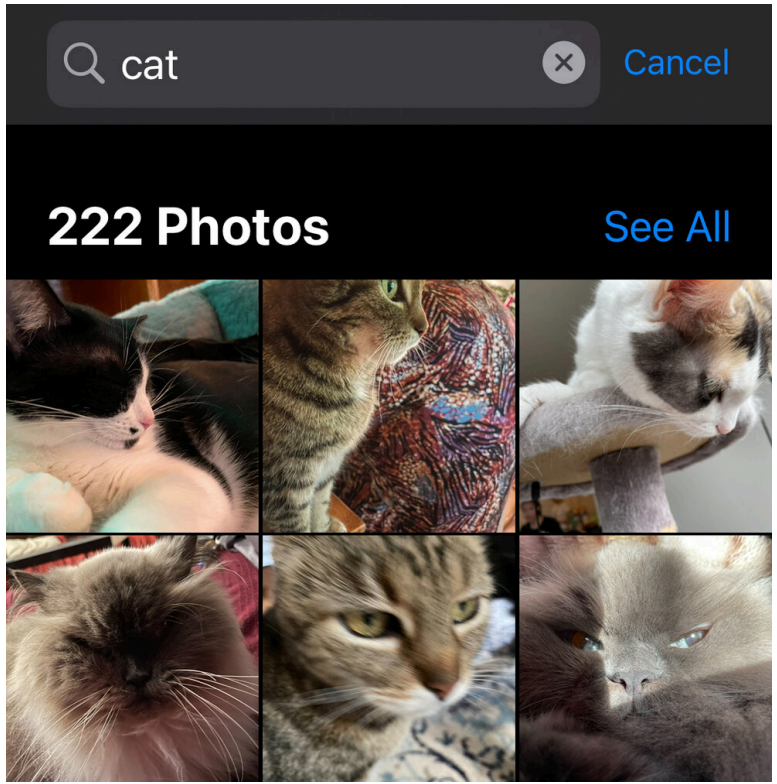
Lund April 21

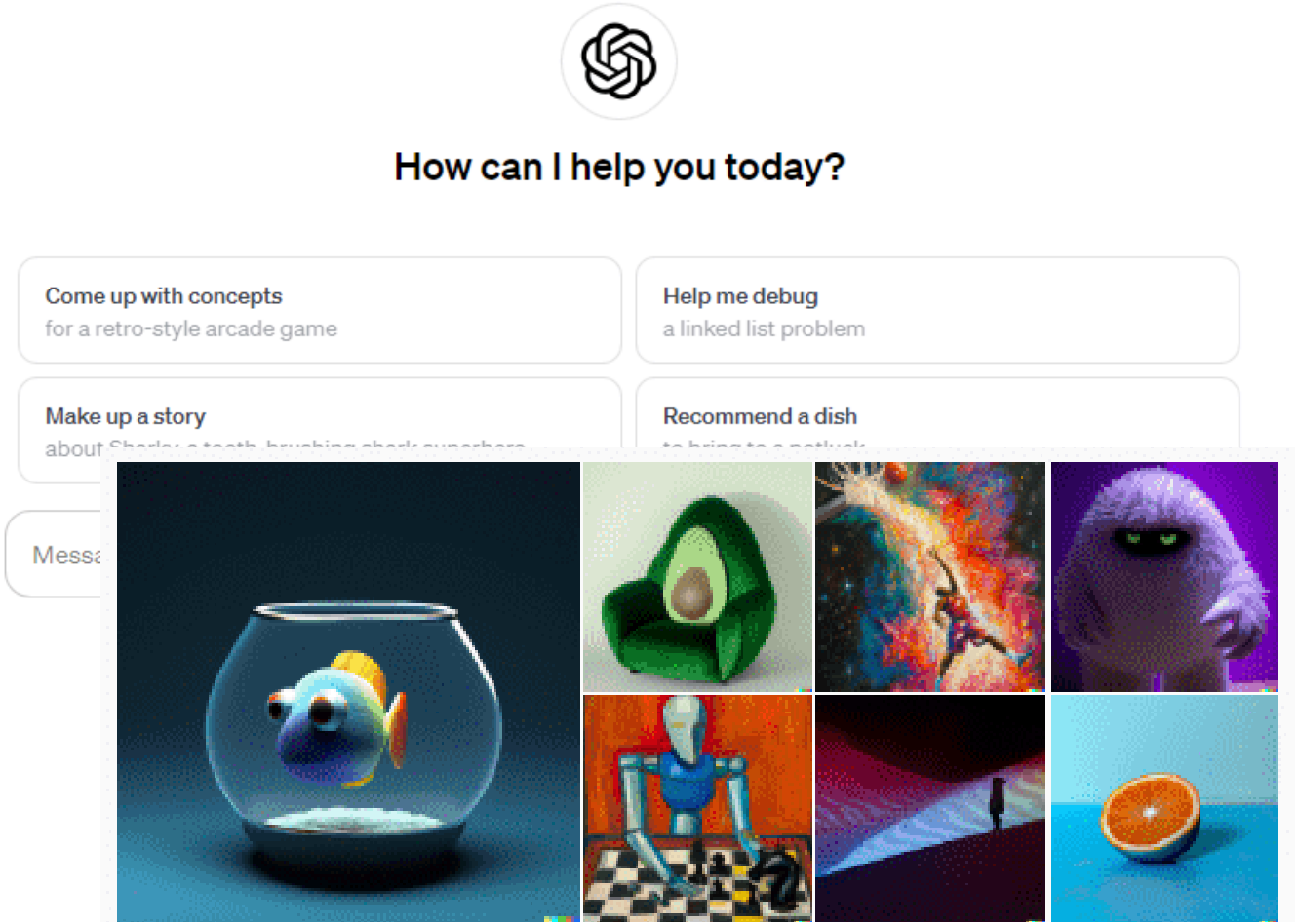
Based on work with F. Bach, L. Balles, A. Bietti, J. Chen, P. Hennig, J.W. Lavington, A. Milligan, M. Schmidt, R. Yadav

Classical ML **Deep Learning**

Linear regression SVM, Kernels & GPs Neural Networks RL Transformers

Graphical models Decision trees & forests CNNs GANs Diffusion





How to fit models?

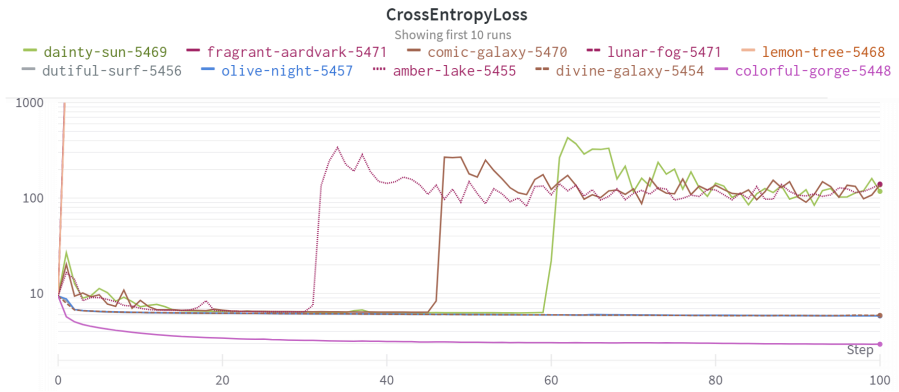
```
model.fit()
```

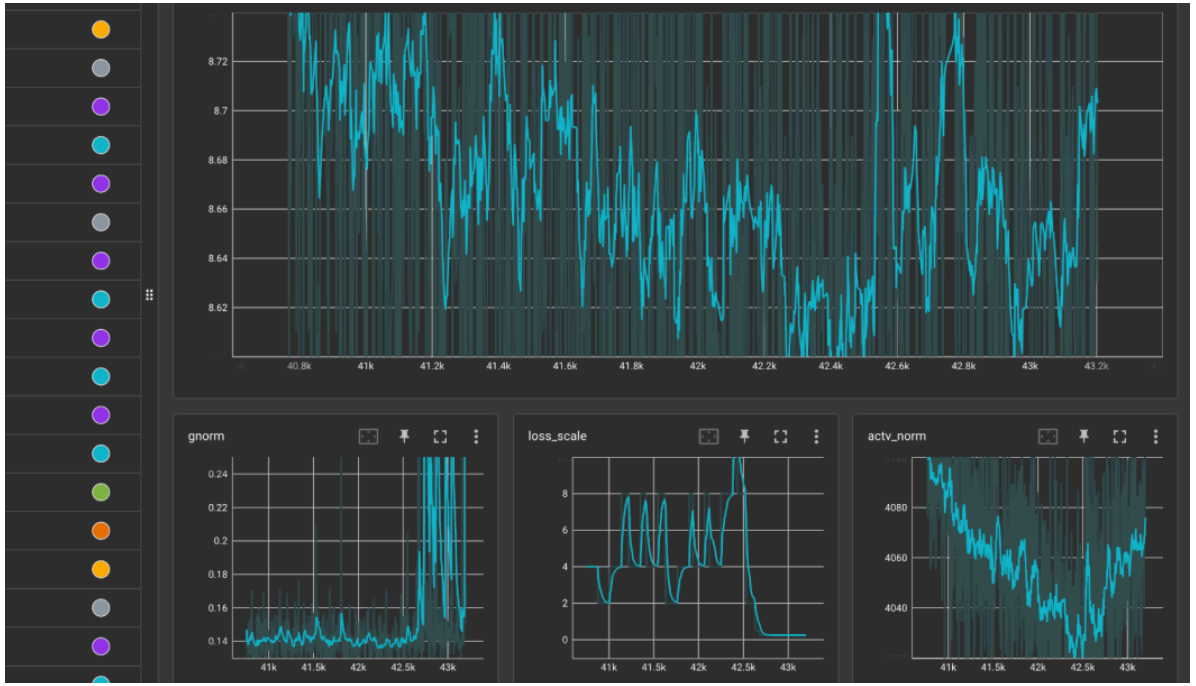
```
python train.py \
    --batch-size=128 --epochs=600 --auto-augment=ta_wide \
    --opt=sgd --lr=0.5 --momentum=0.9 \
    --lr-scheduler=cosineannealinglr --lr-warmup-epochs=5 --lr-warmup-method=linear \
    --random-erase=0.1 --dropout=0.1 \
    --weight-decay=0.00002 --norm-weight-decay=0.0 --label-smoothing=0.1 \
    --mixup-alpha=0.2 --cutmix-alpha=1.0
```

Git clone

Babysit

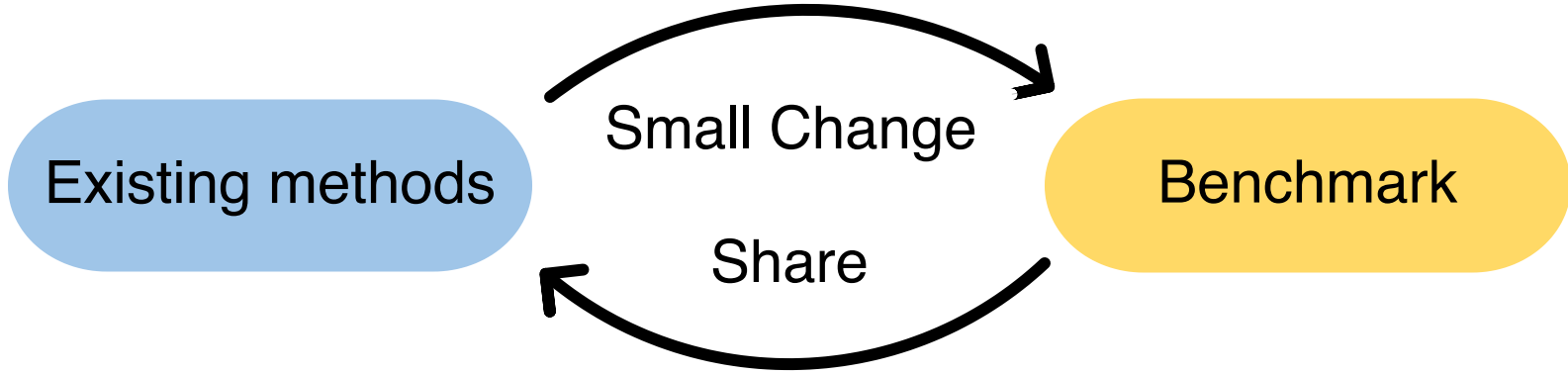
Try everything



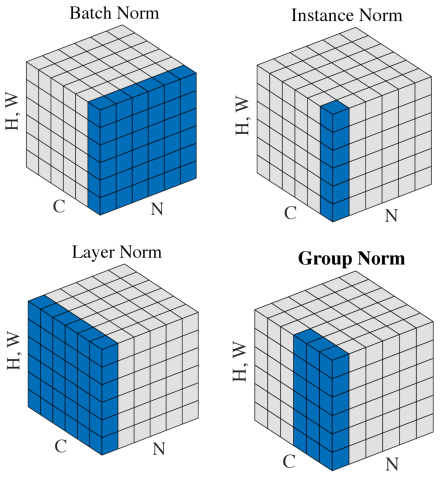


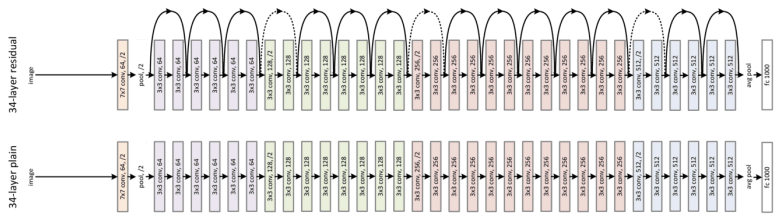
Progress in ML

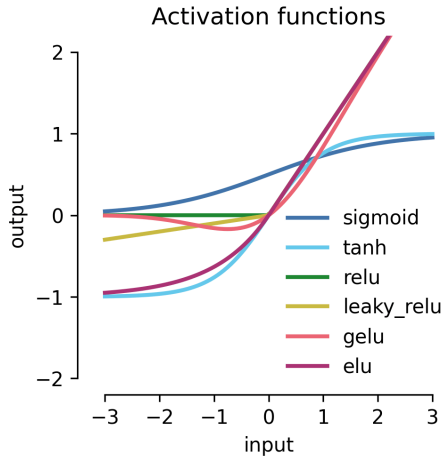
Fast iteration and good evaluation

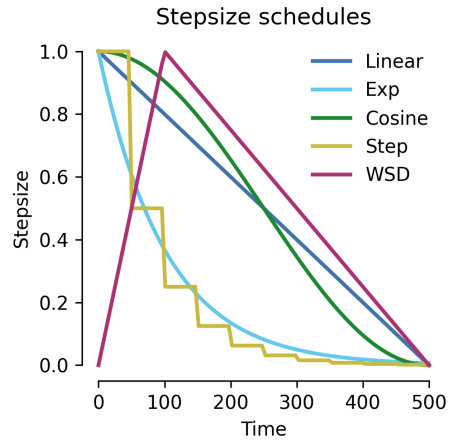


ML research community = evolutionary algorithm at a global scale









The evolutionary algorithm works

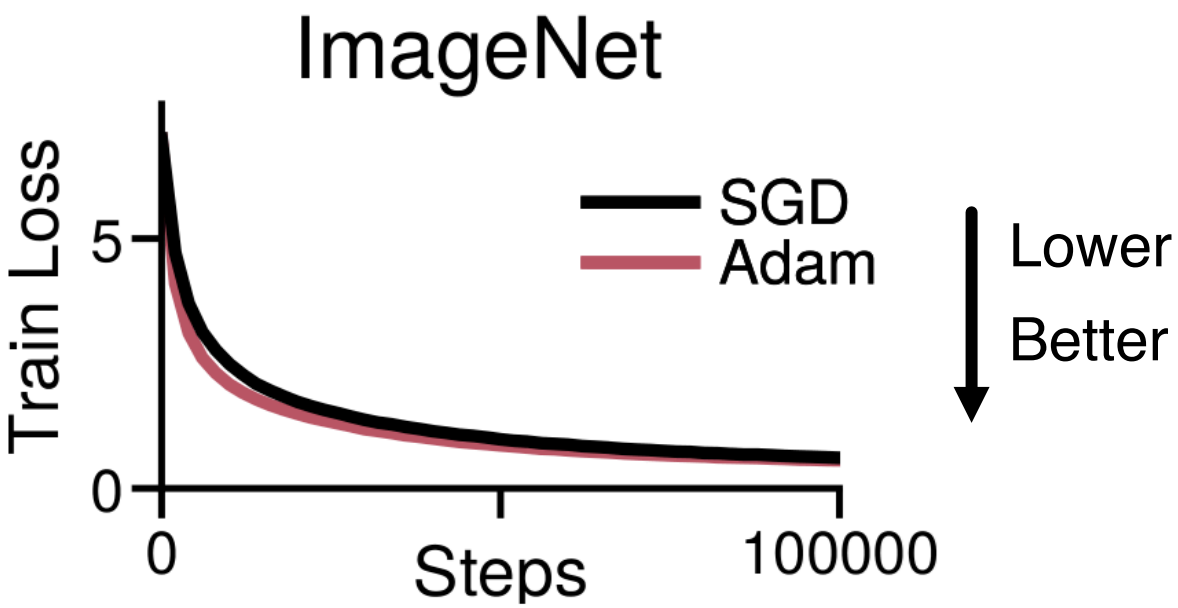
We're proposing solutions faster than we can understand them

Adan	MTAdam	Adam+	AdaAlter	AdaGC	ACGBAdam	AVGrad
AAdam	Nadam	Adam+	AdaBatch	AdaMod	FedAdam	BPGrad
BAdam	NDAdam	Adam++	AdaBayes	AdaPlus	HyperAdam	WNGrad
B	Padam	AdamAL	AdaBelief	AdaScale	L4Adam	AMSGrad
B						ad
C						d
C						ad
D						rad
E						rad
Gadam	TAdam	AdamT	AdaFix	AdaL	ProxAdam	SHAdaGrad
Gadam	VAdam	AdamW	AdaFom	AdaS	SignAdam++	AcceleGrad
Hadam	VAdam	AdamX	AdaFTRL	AdaX	SoftAdam	AngularGrad

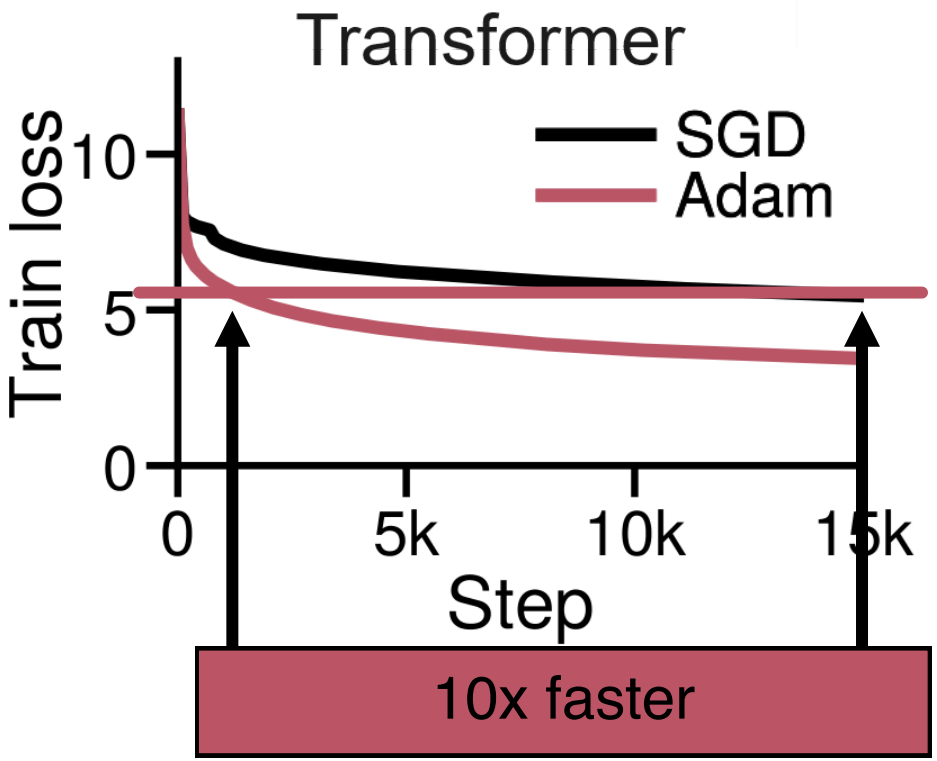
Why does Adam work?

When does Adam help?

Often small improvement



But necessary for LLMs



Gradient Descent

$$w \leftarrow w - \alpha\, g$$

Gradient Descent → Momentum

$$m \leftarrow \beta m + (1 - \beta) g$$

$$w \leftarrow w - \alpha\, m$$

Gradient Descent → Momentum → Adam

$$M \leftarrow \beta_2 M + (1-\beta_2) g^2$$

$$m \leftarrow \beta m + (1-\beta) g$$

$$w \leftarrow w - \alpha \frac{m}{\sqrt{M}}$$

Adam

$$M \leftarrow \beta_2 M + (1 - \beta_2) g^2$$

$$m \leftarrow \beta m + (1 - \beta) g$$

$$w \leftarrow w - \alpha \frac{m}{\sqrt{M}}$$

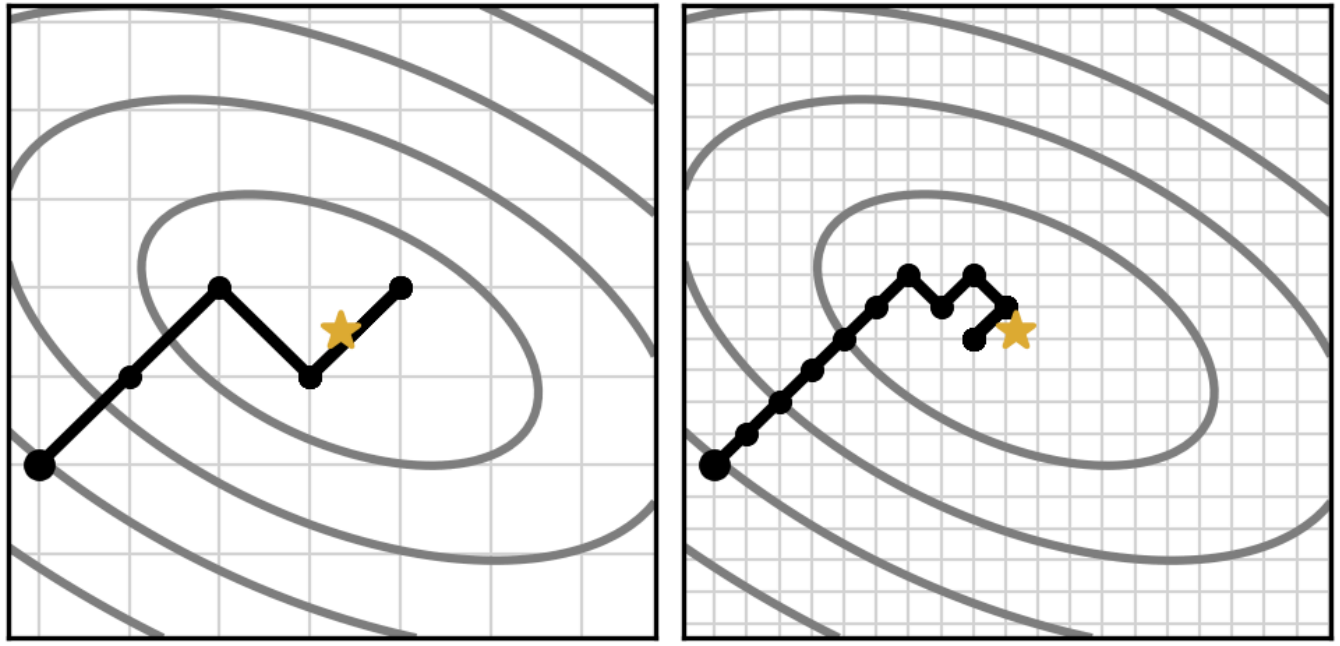
Second-order method?

$$\frac{m}{\sqrt{M}} \approx \frac{g}{\sqrt{g^2}} \approx \operatorname{sign}(g)$$

Limitations of the Empirical Fisher, NeurIPS'18

Why sign?

Throws away information
Needs small step-size

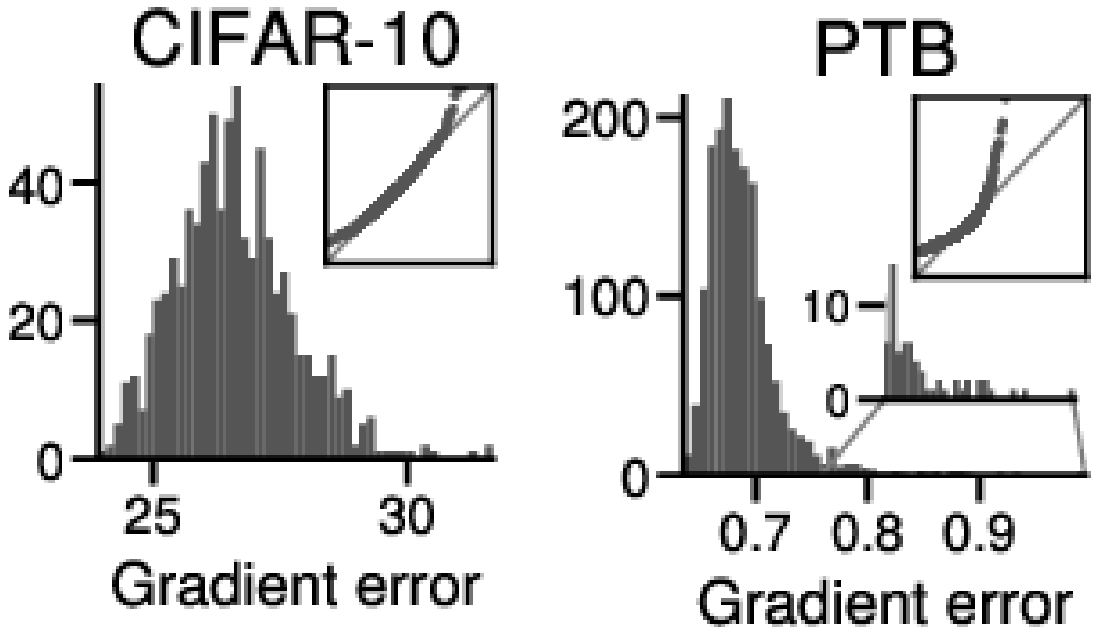


Existing hypothesis

Adam better handles noise better

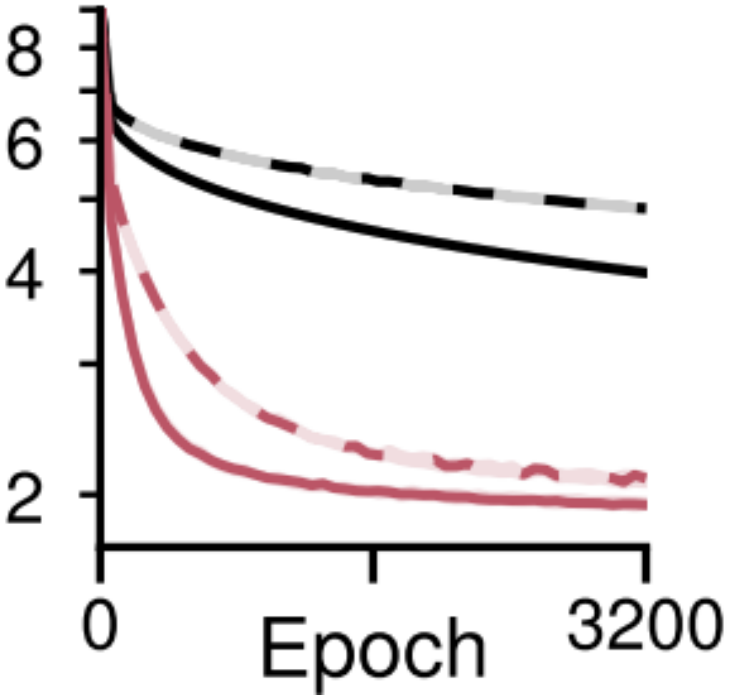
$$\|g - \nabla \mathcal{L}(w)\|$$

Normalized: $$w \leftarrow w - \alpha \frac{g}{\|g\|}$$

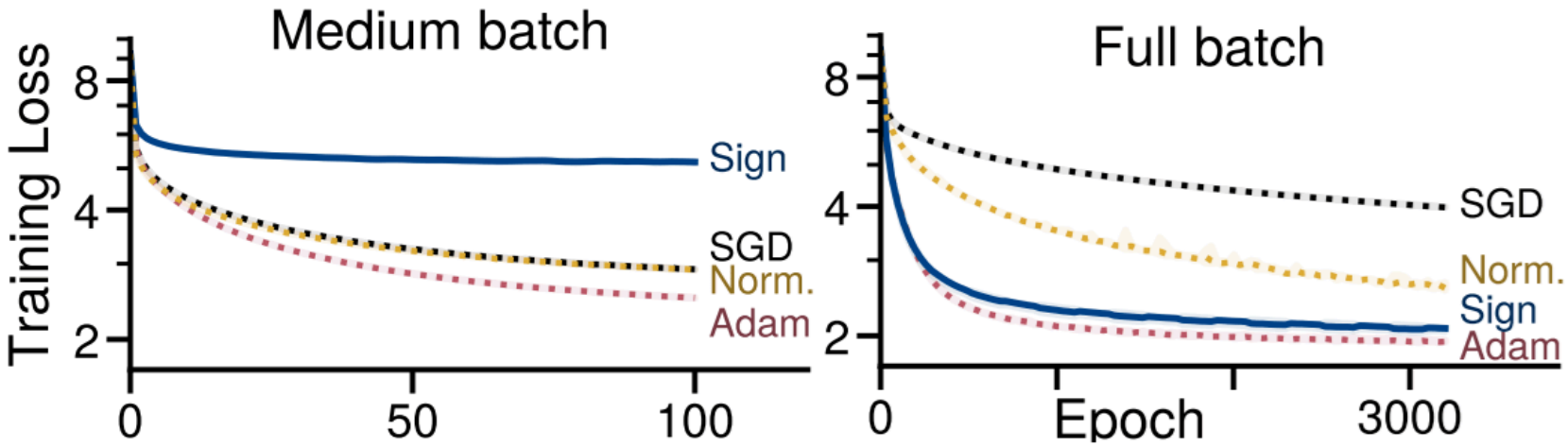


Does gradient noise really matter?

Full batch training



Large gap still appears



Noise is not the problem

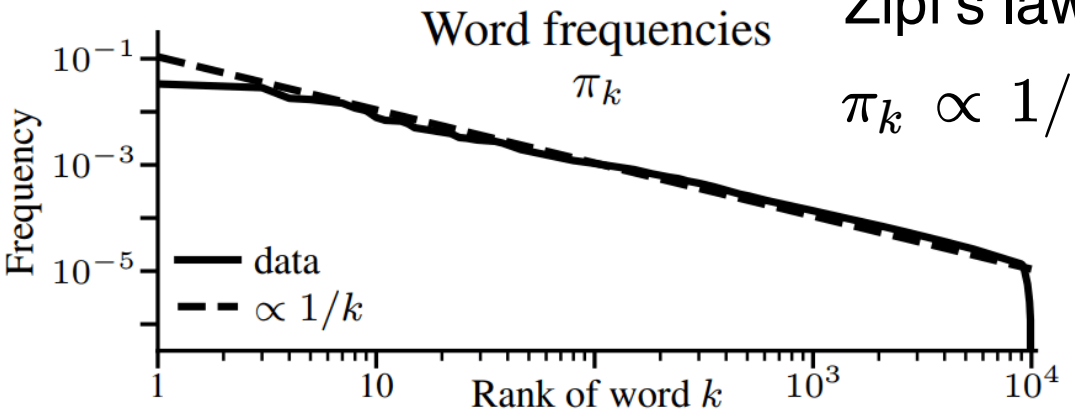
Normalization is not enough

Adam ≈ Sign descent

Why the gap?

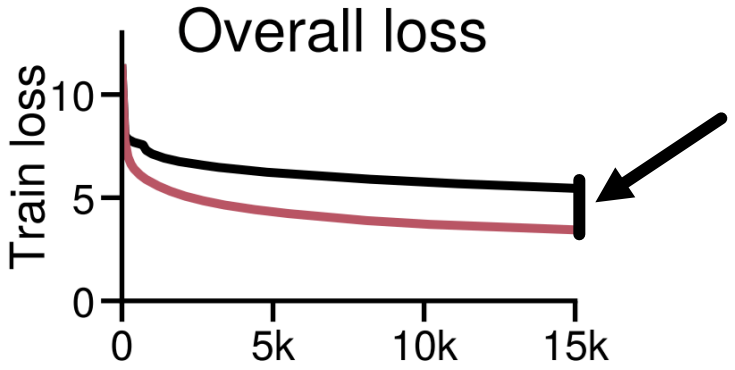
What is the problem?

Language model:
Next word classification



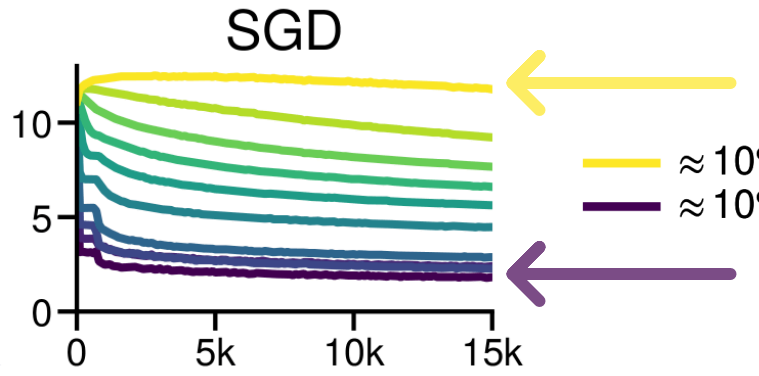
Zipf's law:

$\pi_k \propto 1/k$



Large imbalance

+ many rare words

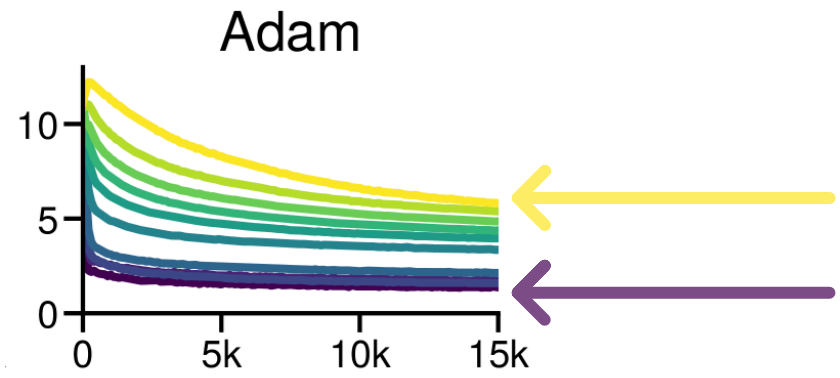


Rare words not learned

≈10% samples, least freq. classes

≈10% samples, most freq. classes

Frequent words are



Both learned

Intuition

per-class loss

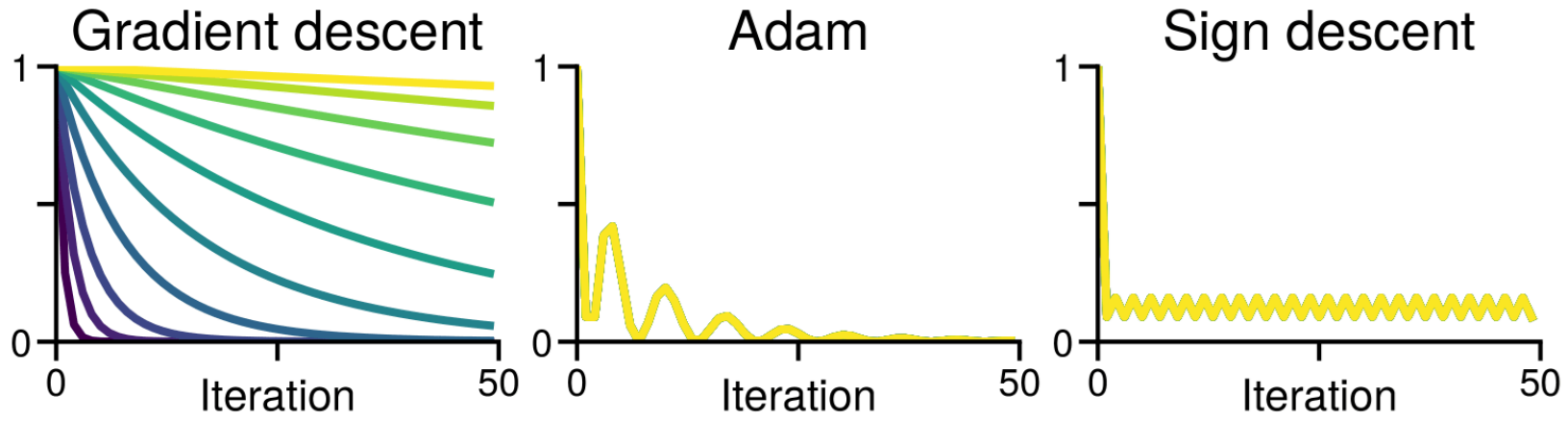
frequencies

$f_1, \ldots, f_c \quad f_i(w) = \frac{1}{2} w^2$

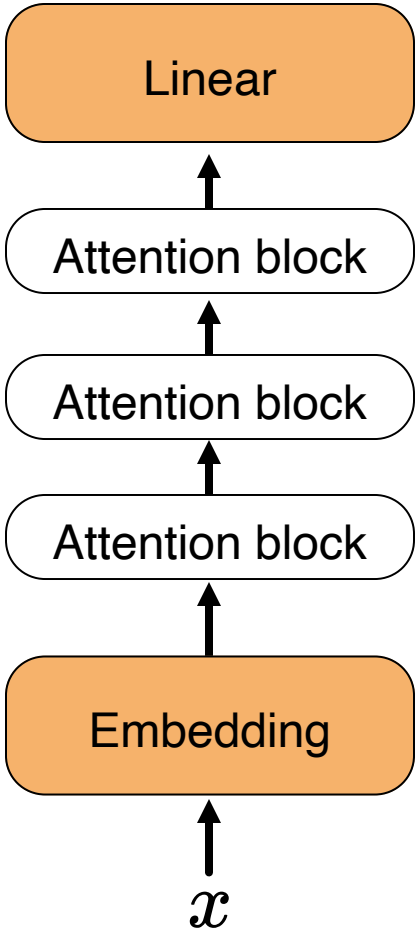
$f([w_1, \ldots w_c]) = \sum_i \pi_i f_i(w_i)$

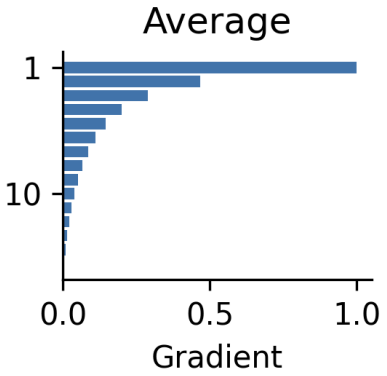
Gradient descent $w_k \leftarrow w_k - \alpha \pi_k w_k$

Sign descent $w_k \leftarrow w_k - \alpha \operatorname{sign}(\pi_k w_k)$



What about language models?

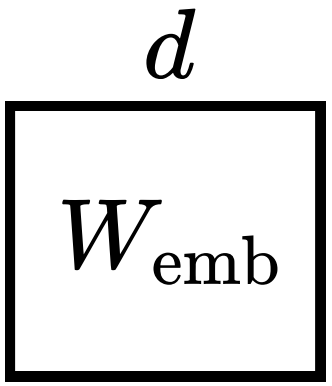


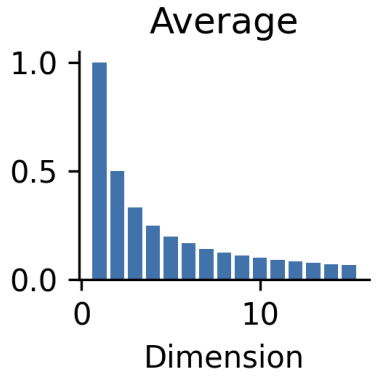


Imbalance

No signal for rare words

Weights will not change





Sign

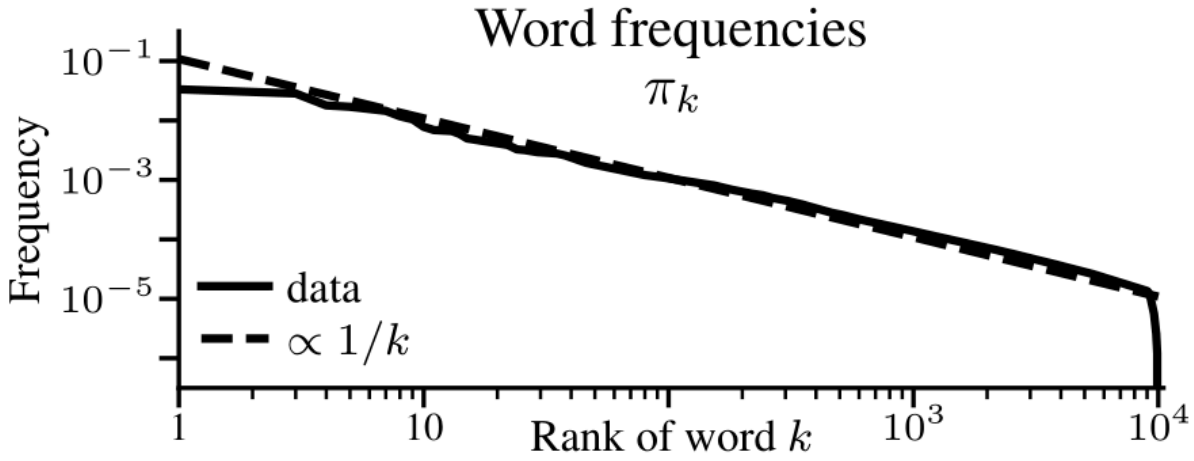
Larger updates

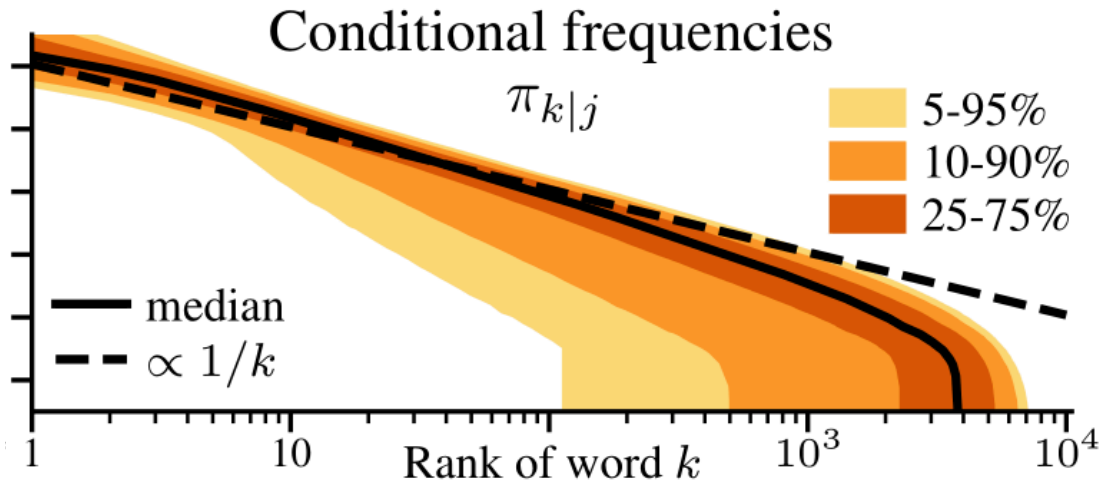
on weights for rare words

A simplified problem for theory

$$x, y \in \{0,1\}^d \qquad \mathcal{L}(W) = \mathbb{E}[\tfrac{1}{2}\|Wx - y\|^2] \qquad W \in \mathbb{R}^{d\times d}$$

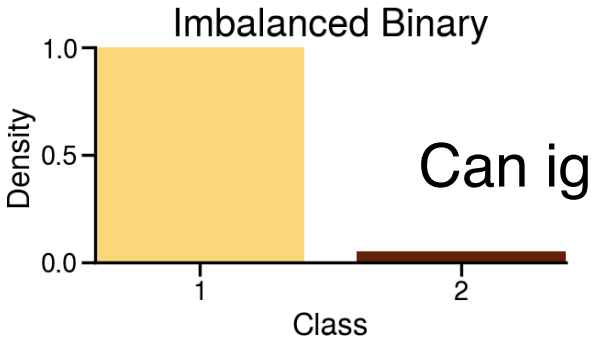
$$p(x = k) = \pi_k \propto \frac{1}{k^\alpha} \qquad p(y = k|x = i) = \pi_{k|i} \propto \frac{1}{k^\alpha}$$



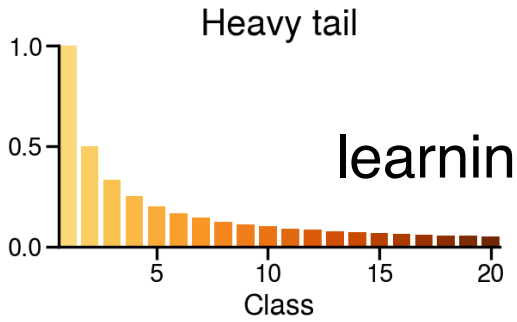


Scaling laws for bigram models, NeurIPS'25

Is text-imbalance special?



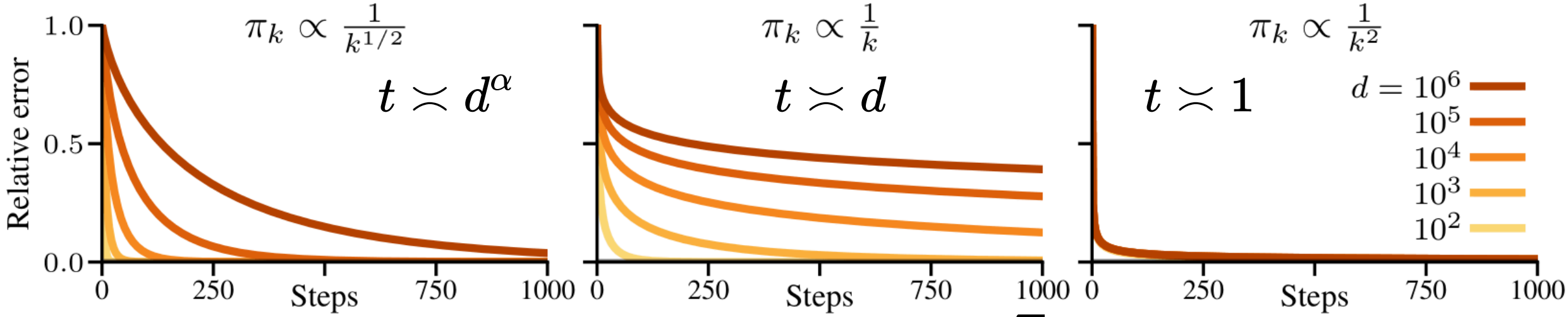
Can ignore rare class



learning the tail matters

If we increase vocab. size, should we scale runtime?

$$\frac{\mathcal{L}(W_t) - \mathcal{L}^*}{\mathcal{L}(W_0) - \mathcal{L}^*} = \varepsilon$$



Sign descent: $t \asymp \sqrt{d}$ $d = 10^4 \rightarrow 100\times$ faster

Scaling laws for bigram models, NeurIPS'25

What about inside the network?

Input/Output word imbalance is a problem in input/output layers

That imbalance is "aligned" to the axes (affects rows/columns)

Adam fixes that (to some extent)

How do we generalize that?

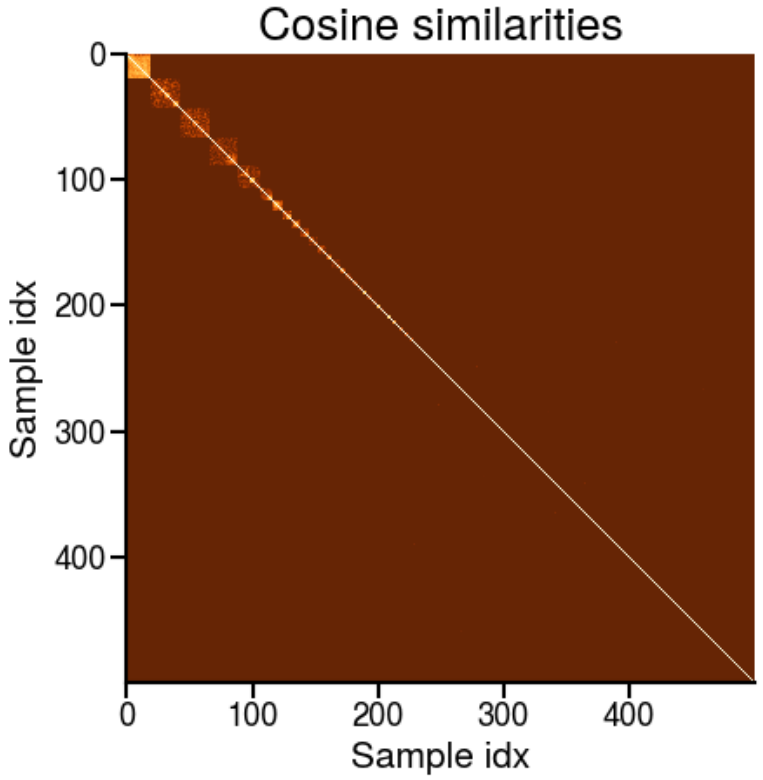
What about Muon?

Signal imbalance

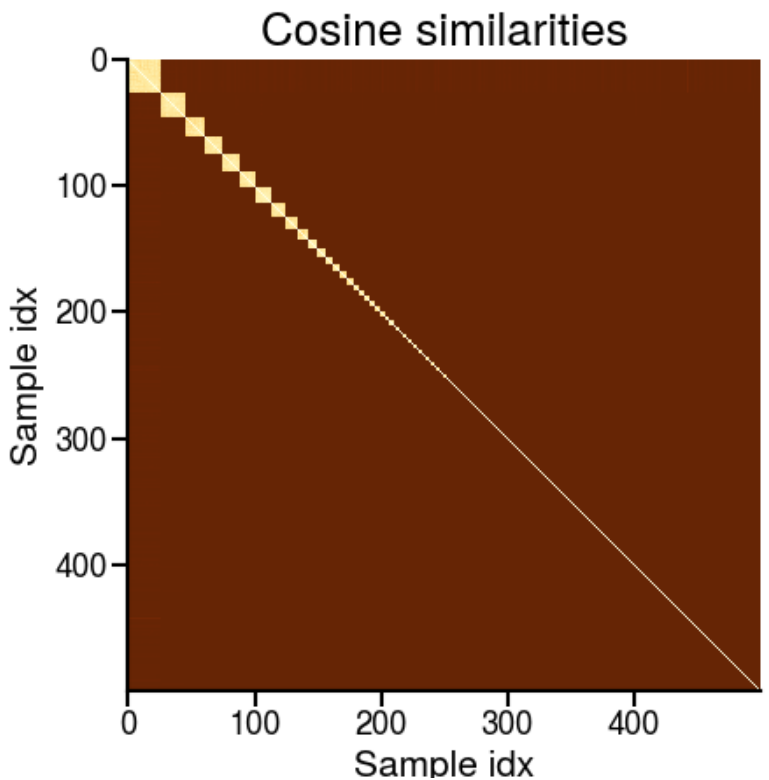
Hypothesis:
Gradients are clustered
Clusters are imbalanced

$$\frac{\langle \nabla f_i(w), \nabla f_j(w) \rangle}{\|\nabla f_i(w)\| \|\nabla f_j(w)\|} = \begin{cases} > 0.5 & \text{if } i, j \in C \\ \epsilon & \text{otherwise} \end{cases}$$

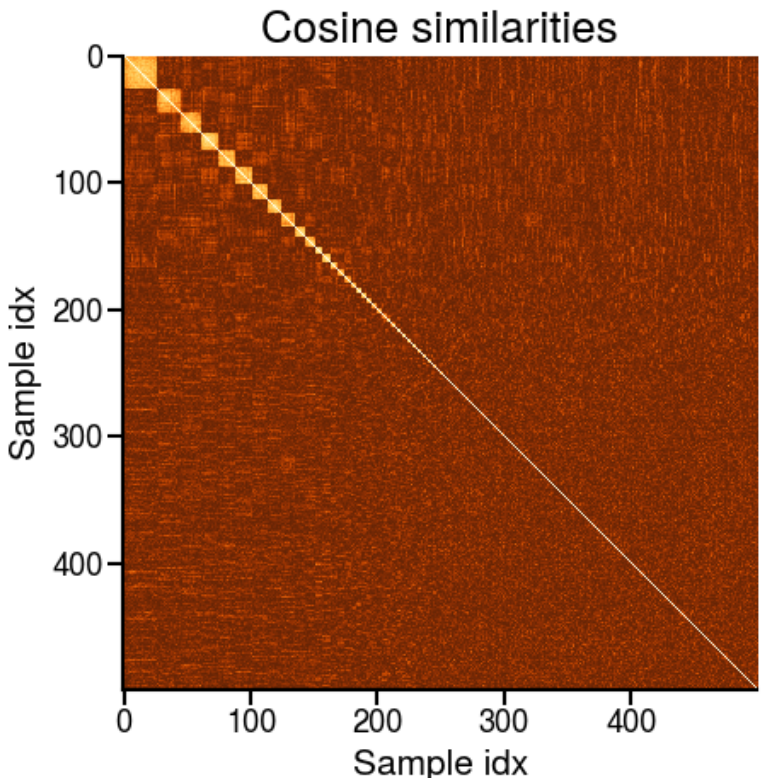
First layer



Last layer



2nd Attention V matrix



Summary

Adam not a generic improvement

Limitations of the Empirical Fisher, NeurIPS'18

Not explained by existing hypotheses

View of Adam as sign method

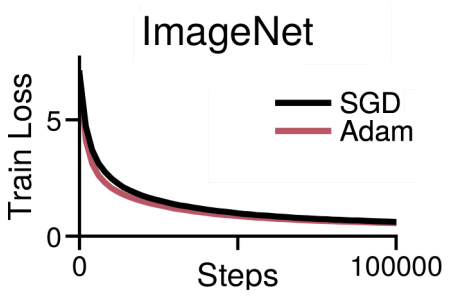
Noise is not the main factor but the sign is, ICLR'23

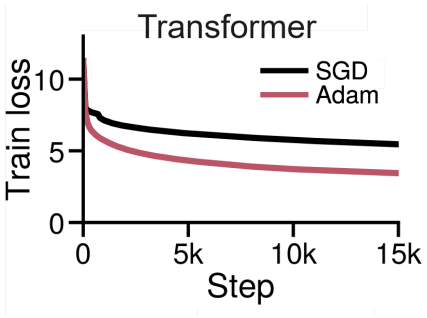
Bottleneck: Imbalance in data frequencies

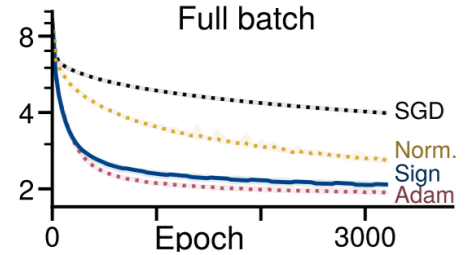
Heavy-tailed imbalance, NeurIPS'24

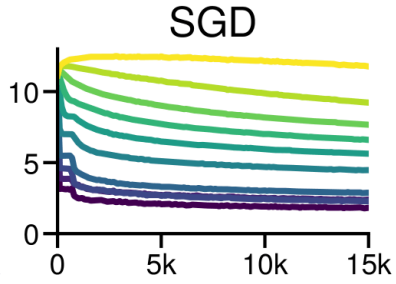
Provable benefit of sign-like methods

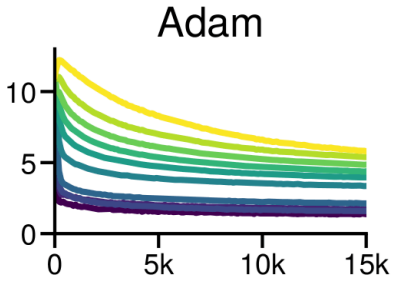
Scaling Laws for Bigram Models, NeurIPS'25

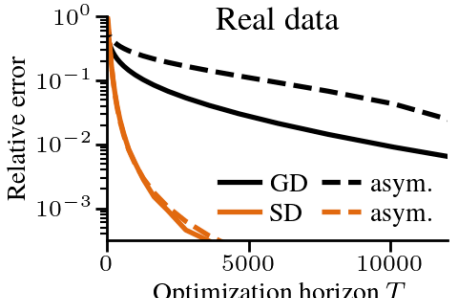


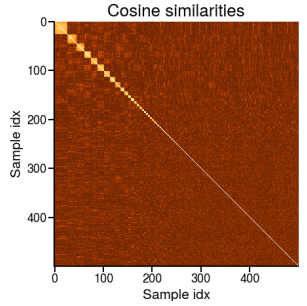












Limitations of the Empirical Fisher Approximation for Natural Gradient Descent

Frederik Kunstner **Lukas Balles** **Philipp Hennig**

NOISE IS NOT THE MAIN FACTOR BEHIND THE GAP BETWEEN SGD AND ADAM ON TRANSFORMERS, BUT SIGN DESCENT MIGHT BE

Frederik Kunstner, Jacques Chen, J. Wilder Lavington & Mark Schmidt†

Heavy-Tailed Class Imbalance and Why Adam Outperforms Gradient Descent on Language Models

Frederik Kunstner **Robin Yadav** **Alan Milligan**
Mark Schmidt **Alberto Bietti**

Scaling Laws for Gradient Descent and Sign Descent for Linear Bigram Models under Zipf's Law

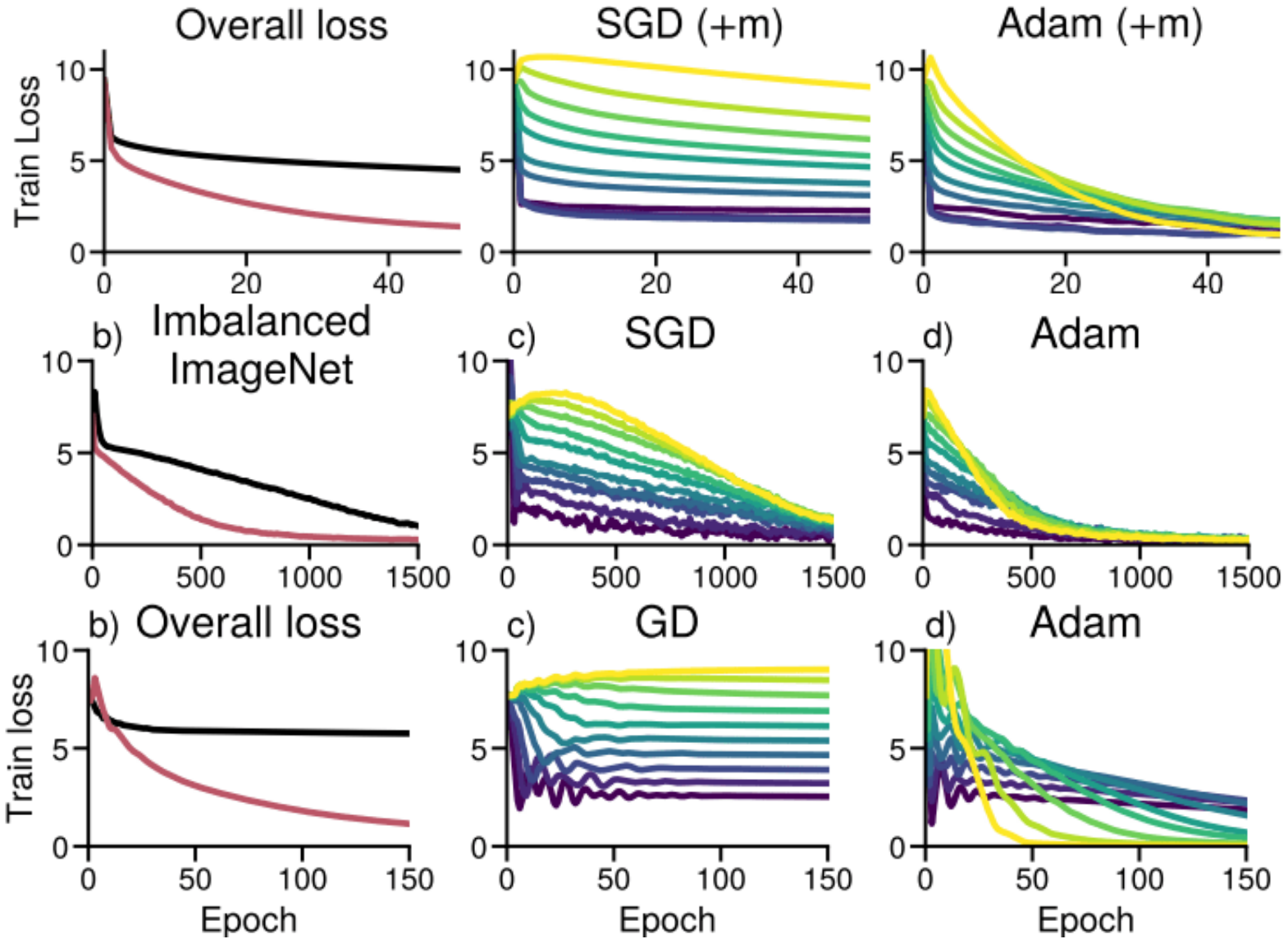
Frederik Kunstner **Francis Bach**

Is it really data imbalance?

Due to small batch size?
Small model, full batch training

Something special in language?
Imbalanced Vision task

Something else in deep models?
Linear model



Heavy-tailed imbalance, NeurIPS'24

Does sign descent work?

Uniform speed across words

Oscillates unless step-size decreases

Gradient descent: $t \asymp d$

Sign descent: $t \asymp \sqrt{d}$

$d = 10^4 \to 100\times$ faster

Only for Zipf's law ($\alpha = 1$)