# The **AlgoPerf**: Training Algorithms Benchmark

## Faster neural network training through better training algorithms

**Frank Schneider**

September 14, 2023

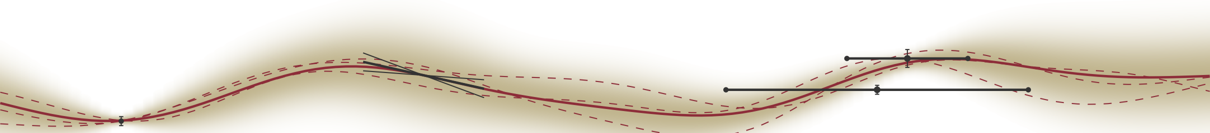
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

MAX PLANCK INSTITUTE FOR INTELLIGENT SYSTEMS

erc

Federal Ministry of Education and Research

imprs-is



# Neural Networks are extremely useful models...

...but!

In practice, neural networks are...

- ...**slow** to train (training can easily take days or weeks)
- ...**tedious** (demanding human intervention and fiddeling)
- ...**expensive** (requiring dozens of trial runs)

# Neural Networks are extremely useful models...

...but!

In practice, neural networks are...

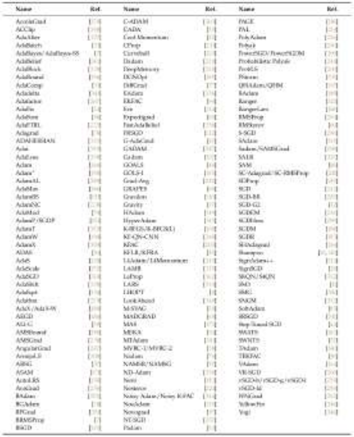
- ...**slow** to train (training can easily take days or weeks)
- ...**tedious** (demanding human intervention and fiddeling)
- ...**expensive** (requiring dozens of trial runs)

**We all want neural network training to be faster, more automatic, and more efficient!**

# The state of deep learning training methods

A confusingly crowded field of methods & hyperparameters

## A huge number of training methods...



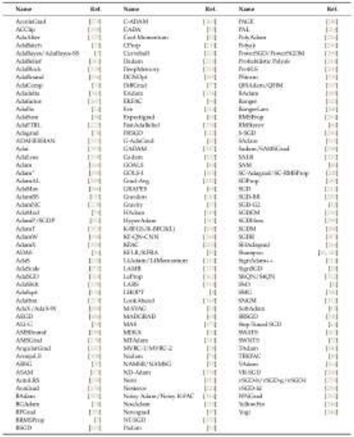
| Name | Ref. | Name | Ref. | Name | Ref. |
|---|---|---|---|---|---|
| AcceleGrad | [illegible] | C-ADAM | [illegible] | PAGE | [illegible] |
| ACClip | [illegible] | CADA | [illegible] | PAL | [illegible] |
| AdaAlter | [illegible] | Cool Momentum | [illegible] | PolyAdam | [illegible] |
| AdaBatch | [illegible] | CProp | [illegible] | Polyak | [illegible] |
| AdaBayes/AdaBayes-SS | [illegible] | Curveball | [illegible] | PowerSGD/PowerSGDM | [illegible] |
| AdaBelief | [illegible] | Dadam | [illegible] | Probabilistic Polyak | [illegible] |
| AdaBlock | [illegible] | DeepMemory | [illegible] | ProbLS | [illegible] |
| AdaBound | [illegible] | DGNOpt | [illegible] | PStorm | [illegible] |
| AdaComp | [illegible] | DiffGrad | [illegible] | QHAdam/QHM | [illegible] |
| Adadelta | [illegible] | EAdam | [illegible] | RAdam | [illegible] |
| Adafactor | [illegible] | EKFAC | [illegible] | Ranger | [illegible] |
| AdaFix | [illegible] | Eve | [illegible] | RangerLars | [illegible] |
| AdaFom | [illegible] | Expectigrad | [illegible] | RMSProp | [illegible] |
| AdaFTRL | [illegible] | FastAdaBelief | [illegible] | RMSterov | [illegible] |
| Adagrad | [illegible] | FRSGD | [illegible] | S-SGD | [illegible] |
| ADAHESSIAN | [illegible] | G-AdaGrad | [illegible] | SAdam | [illegible] |
| Adai | [illegible] | GADAM | [illegible] | Sadam/SAMSGrad | [illegible] |
| AdaLoss | [illegible] | Gadam | [illegible] | SALR | [illegible] |
| Adam | [illegible] | GOALS | [illegible] | SAM | [illegible] |
| Adam+ | [illegible] | GOLS-I | [illegible] | SC-Adagrad/SC-RMSProp | [illegible] |
| AdamAL | [illegible] | Grad-Avg | [illegible] | SDProp | [illegible] |
| AdaMax | [illegible] | GRAPES | [illegible] | SGD | [illegible] |
| AdamBS | [illegible] | Gravilon | [illegible] | SGD-BB | [illegible] |
| AdamNC | [illegible] | Gravity | [illegible] | SGD-G2 | [illegible] |
| AdaMod | [illegible] | HAdam | [illegible] | SGDEM | [illegible] |
| AdamP/SGDP | [illegible] | HyperAdam | [illegible] | SGDHess | [illegible] |
| AdamT | [illegible] | K-BFGS/K-BFGS(L) | [illegible] | SGDM | [illegible] |
| AdamW | [illegible] | KF-QN-CNN | [illegible] | SGDR | [illegible] |
| AdamX | [illegible] | KFAC | [illegible] | SHAdagrad | [illegible] |
| ADAS | [illegible] | KFLR/KFRA | [illegible] | Shampoo | [illegible] |
| AdaS | [illegible] | L4Adam/L4Momentum | [illegible] | SignAdam++ | [illegible] |
| AdaScale | [illegible] | LAMB | [illegible] | SignSGD | [illegible] |
| AdaSGD | [illegible] | LaProp | [illegible] | SKQN/S4QN | [illegible] |
| AdaShift | [illegible] | LARS | [illegible] | SM3 | [illegible] |
| AdaSqrt | [illegible] | LHOPT | [illegible] | SMG | [illegible] |
| Adathm | [illegible] | LookAhead | [illegible] | SNGM | [illegible] |
| AdaX/AdaX-W | [illegible] | M-SVAG | [illegible] | SoftAdam | [illegible] |
| AEGD | [illegible] | MADGRAD | [illegible] | SRSGD | [illegible] |
| ALI-G | [illegible] | MAS | [illegible] | Step-Tuned SGD | [illegible] |
| AMSBound | [illegible] | MEKA | [illegible] | SWATS | [illegible] |
| AMSGrad | [illegible] | MTAdam | [illegible] | SWNTS | [illegible] |
| AngularGrad | [illegible] | MVRC-1/MVRC-2 | [illegible] | TAdam | [illegible] |
| ArmijoLS | [illegible] | Nadam | [illegible] | TEKFAC | [illegible] |
| ARSG | [illegible] | NAMSB/NAMSG | [illegible] | VAdam | [illegible] |
| ASAM | [illegible] | ND-Adam | [illegible] | VR-SGD | [illegible] |
| AutoLRS | [illegible] | Nero | [illegible] | vSGD-b/vSGD-g/vSGD-l | [illegible] |
| AvaGrad | [illegible] | Nesterov | [illegible] | vSGD-fd | [illegible] |
| BAdam | [illegible] | Noisy Adam/Noisy K-FAC | [illegible] | WNGrad | [illegible] |
| BGAdam | [illegible] | NosAdam | [illegible] | YellowFin | [illegible] |
| BPGrad | [illegible] | Novograd | [illegible] | Yogi | [illegible] |
| BRMSProp | [illegible] | NT-SGD | [illegible] | | |
| BSGD | [illegible] | Padam | [illegible] | | |

*from "Descending through a Crowded Valley"*
*(Schmidt, Schneider, Hennig; 2021)*

# The state of deep learning training methods

A confusingly crowded field of methods & hyperparameters

## A huge number of training methods...



| Name | Ref. | Name | Ref. | Name | Ref. |
|---|---|---|---|---|---|
| AcceleGrad | [illegible] | C-ADAM | [illegible] | PAGE | [illegible] |
| ACClip | [illegible] | CADA | [illegible] | PAL | [illegible] |
| AdaAlter | [illegible] | Cool Momentum | [illegible] | PolyAdam | [illegible] |
| AdaBatch | [illegible] | CProp | [illegible] | Polyak | [illegible] |
| AdaBayes/AdaBayes-SS | [illegible] | Curveball | [illegible] | PowerSGD/PowerSGDM | [illegible] |
| AdaBelief | [illegible] | Dadam | [illegible] | Probabilistic Polyak | [illegible] |
| AdaBlock | [illegible] | DeepMemory | [illegible] | ProxLS | [illegible] |
| AdaBound | [illegible] | DGNOpt | [illegible] | PStorm | [illegible] |
| AdaComp | [illegible] | DiffGrad | [illegible] | QHAdam/QHM | [illegible] |
| Adadelta | [illegible] | EAdam | [illegible] | RAdam | [illegible] |
| Adafactor | [illegible] | EKFAC | [illegible] | Ranger | [illegible] |
| AdaFix | [illegible] | Eve | [illegible] | RangerLars | [illegible] |
| AdaFom | [illegible] | Expectigrad | [illegible] | RMSProp | [illegible] |
| AdaFTRL | [illegible] | FastAdaBelief | [illegible] | RMSterov | [illegible] |
| Adagrad | [illegible] | FRSGD | [illegible] | S-SGD | [illegible] |
| ADAHESSIAN | [illegible] | G-AdaGrad | [illegible] | SAdam | [illegible] |
| Adai | [illegible] | GADAM | [illegible] | Sadam/SAMSGrad | [illegible] |
| AdaLoss | [illegible] | Gadam | [illegible] | SALR | [illegible] |
| Adam | [illegible] | GOALS | [illegible] | SAM | [illegible] |
| Adam+ | [illegible] | GOLS-I | [illegible] | SC-Adagrad/SC-RMSProp | [illegible] |
| AdamAL | [illegible] | Grad-Avg | [illegible] | SDProp | [illegible] |
| AdaMax | [illegible] | GRAPES | [illegible] | SGD | [illegible] |
| AdamBS | [illegible] | Gravilon | [illegible] | SGD-BB | [illegible] |
| AdamNC | [illegible] | Gravity | [illegible] | SGD-G2 | [illegible] |
| AdaMod | [illegible] | HAdam | [illegible] | SGDEM | [illegible] |
| AdamP/SGDP | [illegible] | HyperAdam | [illegible] | SGDHess | [illegible] |
| AdamT | [illegible] | K-BFGS/K-BFGS(L) | [illegible] | SGDM | [illegible] |
| AdamW | [illegible] | KF-QN-CNN | [illegible] | SGDR | [illegible] |
| AdamX | [illegible] | KFAC | [illegible] | SHAdagrad | [illegible] |
| ADAS | [illegible] | KFLR/KFRA | [illegible] | Shampoo | [illegible] |
| AdaS | [illegible] | L4Adam/L4Momentum | [illegible] | SignAdam++ | [illegible] |
| AdaScale | [illegible] | LAMB | [illegible] | SignSGD | [illegible] |
| AdaSGD | [illegible] | LaProp | [illegible] | SKQN/S4QN | [illegible] |
| AdaShift | [illegible] | LARS | [illegible] | SM3 | [illegible] |
| AdaSqrt | [illegible] | LHOPT | [illegible] | SMG | [illegible] |
| Adathm | [illegible] | LookAhead | [illegible] | SNGM | [illegible] |
| AdaX/AdaX-W | [illegible] | M-SVAG | [illegible] | SoftAdam | [illegible] |
| AEGD | [illegible] | MADGRAD | [illegible] | SRSGD | [illegible] |
| ALI-G | [illegible] | MAS | [illegible] | Step-Tuned SGD | [illegible] |
| AMSBound | [illegible] | MEKA | [illegible] | SWATS | [illegible] |
| AMSGrad | [illegible] | MTAdam | [illegible] | SWNTS | [illegible] |
| AngularGrad | [illegible] | MVRC-1/MVRC-2 | [illegible] | TAdam | [illegible] |
| ArmijoLS | [illegible] | Nadam | [illegible] | TEKFAC | [illegible] |
| ARSG | [illegible] | NAMSB/NAMSG | [illegible] | VAdam | [illegible] |
| ASAM | [illegible] | ND-Adam | [illegible] | VR-SGD | [illegible] |
| AutoLRS | [illegible] | Nero | [illegible] | vSGD-b/vSGD-g/vSGD-l | [illegible] |
| AvaGrad | [illegible] | Nesterov | [illegible] | vSGD-fd | [illegible] |
| BAdam | [illegible] | Noisy Adam/Noisy K-FAC | [illegible] | WNGrad | [illegible] |
| BGAdam | [illegible] | NosAdam | [illegible] | YellowFin | [illegible] |
| BPGrad | [illegible] | Novograd | [illegible] | Yogi | [illegible] |
| BRMSProp | [illegible] | NT-SGD | [illegible] | | |
| BSGD | [illegible] | Padam | [illegible] | | |

*from "Descending through a Crowded Valley"*
*(Schmidt, Schneider, Hennig; 2021)*

## ...and training tricks

- ▶ OneCycle scheduler, gradient checkpointing
- ▶ Genetic Algorithm for Hyperparameters
- ▶ Avoid batches that lead to NaN/inf losses
- ▶ One cycle, low fidelity training, SGD with restarts
- ▶ Proximal optimization for regularizers
- ▶ Line searches for the maximum learning rate
- ▶ Normalized updates
- ▶ Distributed Shampoo, Normformer, GLU
- ▶ Weight averaging
- ▶ FreezeOut
- ▶ A different epsilon value!
- ▶ Check hyperparameter performance over multiple seeds
- ▶ Lowering the learning rate!
- ▶ Normalizing data works better than batch or layer norm
- ▶ Mixed precision training
- ▶ Train with a small subset
- ▶ Cyclic and one cycle LR
- ▶ Label smoothing
- ▶ ...

*from the NeurIPS "HITY Workshop"*
*(Schneider et al.; 2022)*

## We desperately need new benchmarks for neural network training algorithms.

# The state of "benchmarking" in current deep learning optimizer papers

No standardized procedure to follow

- Each paper "invents" their own evaluation protocol.
- Unreasonably hard to perform a convincing, informative, and practically relevant comparison with strong baselines.
- Lot’s of subtle pitfalls with tuning, problem-specification, etc.

# Example of pitfalls when comparing training algorithms

Which algorithms trains the fastest depends on what it means for training to be complete

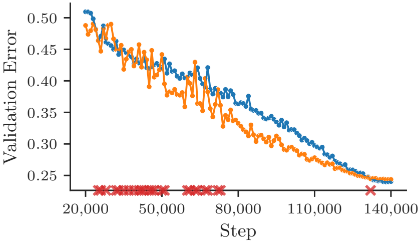


Figure 1: *Left:* Validation error for two different runs (—, —) of ADAM on RESNET-50 on IMAGENET. *Right:* The *best* validation error obtained so far. The runs intersect multiple times (✖).

# Example of pitfalls when comparing training algorithms

Which algorithms trains the fastest depends on what it means for training to be complete

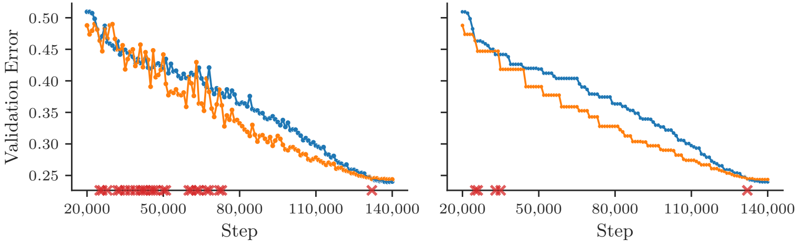


Figure 1: *Left:* Validation error for two different runs (—, —) of Adam on ResNet-50 on ImageNet. *Right:* The *best* validation error obtained so far. The runs intersect multiple times (✖).

- Directly comparing training curves is ill-posed.
- Without defining the target in advance, we can champion any method (moving the goal post after the experiment).
- Tuning goals must align.

# MLCommons introduces the **AlgoPerf**: Training Algorithm Benchmark

An unprecedented effort to find faster deep learning training algorithms

ML Commons

**Algorithms Working Group**

**AlgoPerf: Training Algorithm Benchmark**

**A standardized benchmark competition to measure neural network training speedups due to algorithmic changes.**

An open large-scale effort by 25+ researchers from Google, University of Tübingen, University of Toronto, Meta AI, etc.

Chaired by

**George Dahl** Google

**Frank Schneider** University of Tübingen

# The key features of AlgoPerf

A standardized competitive time-to-results benchmark

- **A competitive time-to-results benchmark.**
  → Everyone's submissions are everyone's **strong baselines**.

# The key features of AlgoPerf

A standardized competitive time-to-results benchmark

- **A competitive time-to-results benchmark.**
  → Everyone's submissions are everyone's **strong baselines**.
- **Fixed hardware, workloads, and process.**
  → Submissions need to provide **training algorithm improvements**.

# The key features of AlgoPerf

A standardized competitive time-to-results benchmark

- **A competitive time-to-results benchmark.**
  → Everyone's submissions are everyone's **strong baselines**.
- **Fixed hardware, workloads, and process.**
  → Submissions need to provide **training algorithm improvements**.
- **Aggregate across a variety of realistic workloads using performance profiles.**
  → No specialized solutions but **general-purpose methods**.

# The key features of AlgoPerf

A standardized competitive time-to-results benchmark

- **A competitive time-to-results benchmark.**
  → Everyone's submissions are everyone's **strong baselines**.
- **Fixed hardware, workloads, and process.**
  → Submissions need to provide **training algorithm improvements**.
- **Aggregate across a variety of realistic workloads using performance profiles.**
  → No specialized solutions but **general-purpose methods**.
- **Explicitly accounts for hyperparameter tuning by providing search spaces.**
  → **Runable training algorithms**, not algorithm templates.

# The **AlgoPerf** Training Algorithms Benchmark

We need you!

| | |
|---|---|
| **Read the Rules** | `github.com/mlcommons/algorithmic-efficiency/blob/main/RULES.md` |
| **Read the Paper** | `arxiv.org/abs/2306.07179` |
| **Submit!** | Call for Submission coming soon |

- **Benchmark experts** Join the effort and tell us how to improve!
- **ML community** Help us spread the word!
- **Algorithm researchers** Submit! AlgoPerf is the easiest way to convincingly demonstrate the capabilities of your training algorithm.