Optimization

Deep Learning KU (DAT.C302UF)

Simon Hitzginger

Nov 12, 2025

Institute of Machine Learning and Neural Computation
Graz University of Technology, Austria

Optimizers

- As soon as we have a gradient w.r.t. a mini-batch

$$\mathbf{g} = \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathcal{B}}(\boldsymbol{\theta})$$

  we can pass it to an **optimizer**
- The optimizer uses the gradient to **update our current estimate** of $\boldsymbol{\theta}$
- For example, if the optimizer is **Stochastic Gradient Descent** (SGD), it implements the update rule

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \mathbf{g}$$

  where $\eta > 0$ is the **learning rate**.

Which optimizer to choose?

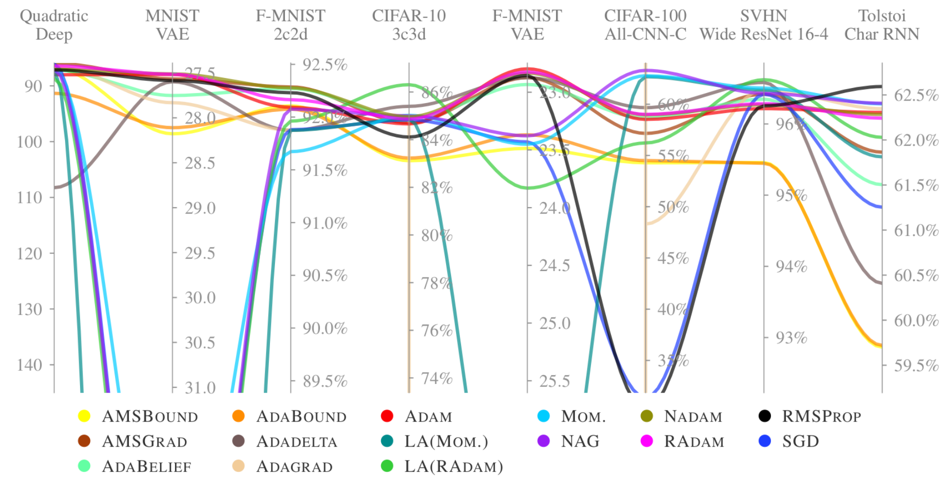
[Bosch et al., 2022]

Name	Ref.	Name	Ref.	Name	Ref.
AcceleGrad	[179]	C-ADAM	[303]	PAGE	[181]
ACClip	[348]	CADA	[55]	PAL	[214]
AdaAlter	[327]	Cool Momentum	[33]	PolyAdam	[226]
AdaBatch	[71]	CProp	[231]	Polyak	[230]
AdaBayes/AdaBayes-SS	[7]	Curveball	[120]	PowerSGD/PowerSGDM	[309]
AdaBelief	[361]	Dadam	[220]	Probabilistic Polyak	[244]
AdaBlock	[338]	DeepMemory	[320]	ProbLS	[200]
AdaBound	[196]	DGNOpt	[185]	PStorm	[331]
AdaComp	[51]	DiffGrad	[77]	QHAdam/QHM	[197]
Adadelta	[341]	EAdam	[336]	RAdam	[189]
Adafactor	[267]	EKFAC	[94]	Ranger	[321]
AdaFix	[14]	Eve	[114]	RangerLars	[106]
AdaFom	[56]	Expectigrad	[65]	RMSProp	[294]
AdaFTRL	[225]	FastAdaBelief	[356]	RMSterov	[60]
Adagrad	[78]	FRSGD	[312]	S-SGD	[280]
ADAHESSIAN	[333]	G-AdaGrad	[47]	SAdam	[315]
Adai	[329]	GADAM	[347]	Sadam/SAMSGrad	[298]
AdaLoss	[290]	Gadam	[107]	SALR	[337]
Adam	[166]	GOALS	[46]	SAM	[86]
Adam^{+}	[190]	GOLS-I	[155]	SC-Adagrad/SC-RMSProp	[211]
AdamAL	[289]	Grad-Avg	[232]	SDProp	[139]
AdaMax	[166]	GRAPES	[69]	SGD	[242]
AdamBS	[191]	Gravilon	[161]	SGD-BB	[287]
AdamNC	[238]	Gravity	[17]	SGD-G2	[12]
AdaMod	[74]	HAdam	[149]	SGDEM	[236]
AdamP/SGDP	[121]	HyperAdam	[315]	SGDHess	[299]
AdamT	[353]	K-BFGS/K-BFGS(L)	[100]	SGDM	[188]
AdamW	[194]	KF-QN-CNN	[240]	SGDR	[193]
AdamX	[300]	KFAC	[203]	SHAdagrad	[136]
ADAS	[81]	KFLR/KFRA	[35]	Shampoo	[10, 110]
AdaS	[130]	L4Adam/L4Momentum	[243]	SignAdam++	[313]
AdaScale	[152]	LAMB	[335]	SignSGD	[28]
AdaSGD	[314]	LaProp	[362]	SKQN/S4QN	[332]
AdaShift	[358]	LARS	[334]	SM3	[11]
AdaSqrt	[134]	LHOPT	[8]	SMG	[301]
Adathm	[279]	LookAhead	[349]	SNGM	[352]
AdaX/AdaX-W	[180]	M-SVAG	[20]	SoftAdam	[85]
AEGD	[186]	MADGRAD	[68]	SRSGD	[311]
ALI-G	[29]	MAS	[173]	Step-Tuned SGD	[44]
AMSBound	[196]	MEKA	[54]	SWATS	[163]
AMSGrad	[238]	MTAdam	[201]	SWNTS	[57]
AngularGrad	[247]	MVRC-1/MVRC-2	[58]	TAdam	[140]
ArmijoLS	[308]	Nadam	[76]	TEKFAC	[91]
ARSG	[57]	NAMSB/NAMSG	[57]	VAdam	[164]
ASAM	[171]	ND-Adam	[350]	VR-SGD	[266]
AutoLRS	[150]	Nero	[192]	vSGD-b/vSGD-g/vSGD-l	[259]
AvaGrad	[256]	Nesterov	[221]	vSGD-fd	[258]
BAdam	[252]	Noisy Adam/Noisy K-FAC	[344]	WNGrad	[322]
BGAdam	[18]	NosAdam	[135]	YellowFin	[346]
BPGrad	[351]	Novograd	[97]	Yogi	[340]
BRMSProp	[7]	NT-SGD	[357]		
BSGD	[133]	Padam	[53]		

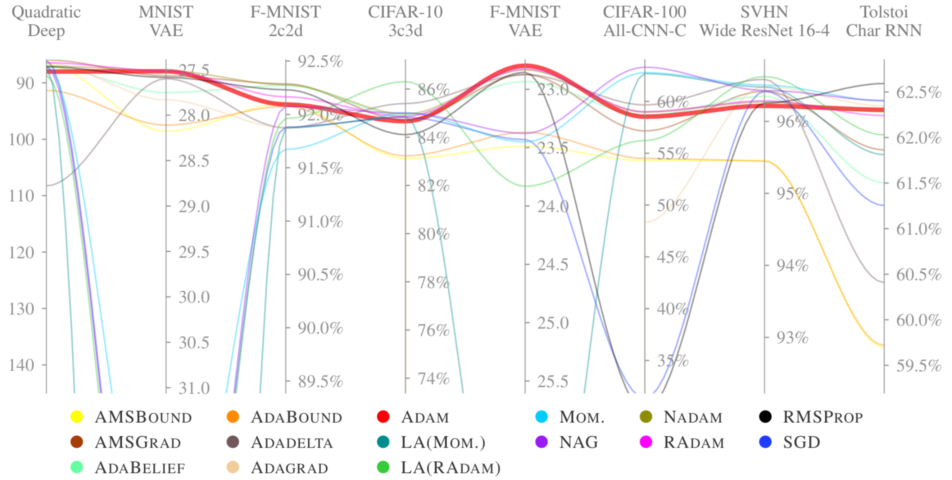
- **Many** optimizers to choose from...
- Is there a **single best** general-purpose optimizer?

Benchmarking Optimizers

[Schmidt et al., 2021]



[Schmidt et al., 2021]



- Adam is a **good default choice**
    - Main hyperparameter: **learning rate**

- ADAM is a **good default choice**
    - Main hyperparameter: **learning rate**
- However, benchmarking optimizers is **hard**...
    - Choice of **hyperparameters**
    - Performance specific to problems

DEMO: OPTIMIZERS IN `PyTorch`

Bosch, N., Grosse, J., Hennig, P., Kristiadi, A., Pförtner, M., Schmidt, J., Schneider, F., Tatzel, L., and Wenger, J. (2022).
Numerics of machine learning.
Technical report, Tübingen AI Center.

Schmidt, R. M., Schneider, F., and Hennig, P. (2021).
Descending through a crowded valley - benchmarking deep learning optimizers.
In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 9367–9376. PMLR.