Learning compositional functions via multiplicative weight updates

Jeremy Bernstein, Jiawei Zhao, Markus Meister, Ming-Yu Liu, Anima Anandkumar, Yisong Yue

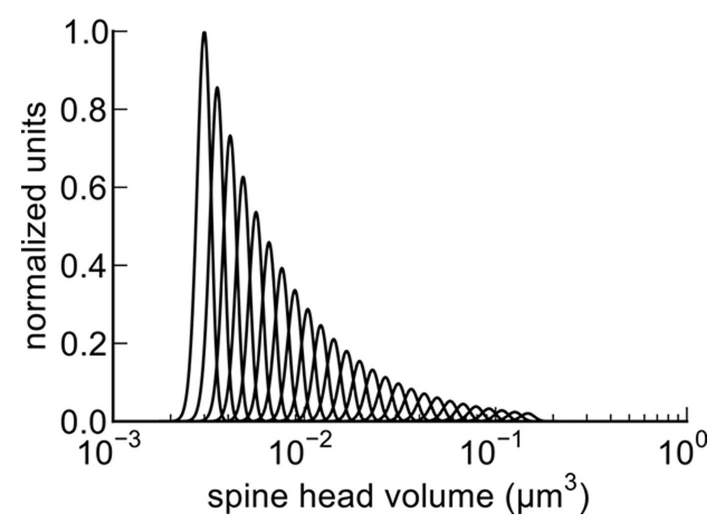
NEURAL INFORMATION PROCESSING SYSTEMS

Learning and precision in neuroscience

Biological synapses are sign-constrained, and learning occurs by adjusting synapse strengths. Neuroscientists believe that synapse strength is correlated with synapse size.

Biological precision

Bartol et al. (2015) measured the distribution of synapse sizes, arriving at the following schematic picture:

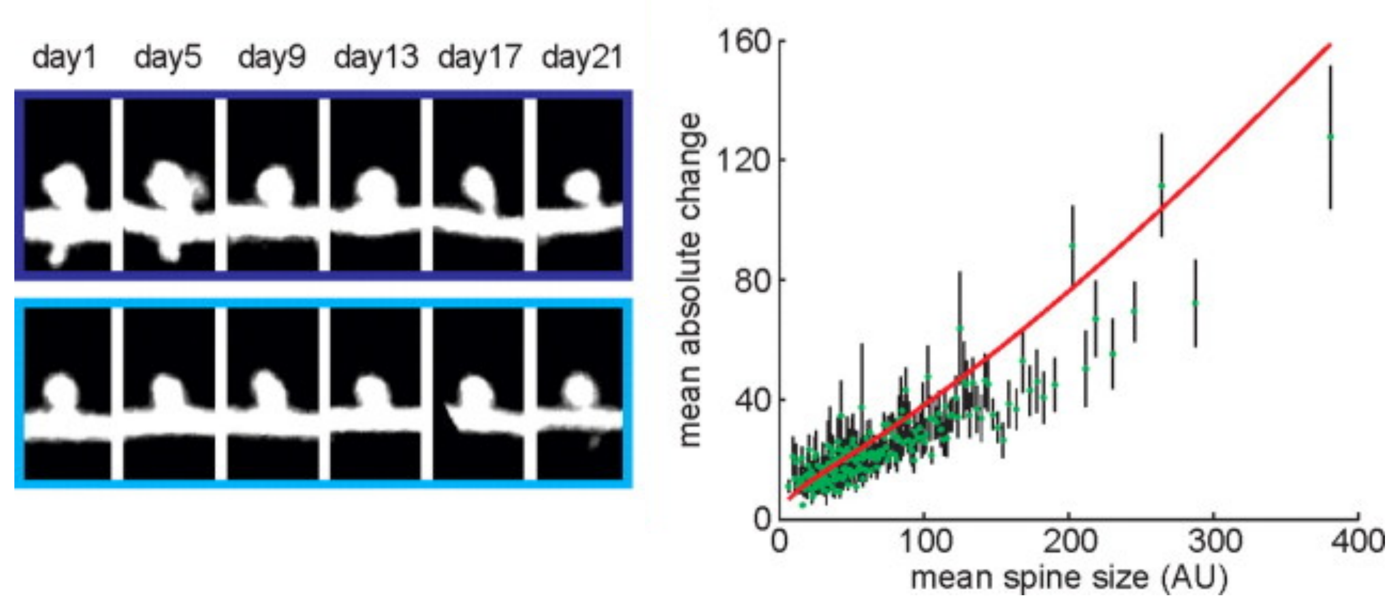


Bartol et al. (2015)

They estimated synapse precision as a function of synapse size, finding that spine volumes occupy ~ 26 distinguishable levels spread uniformly in log space, with dynamic range 60.

Biological learning rules

Loewenstein et al. (2011) found that biological synapses may adjust their strengths *multiplicatively* rather than *additively*.



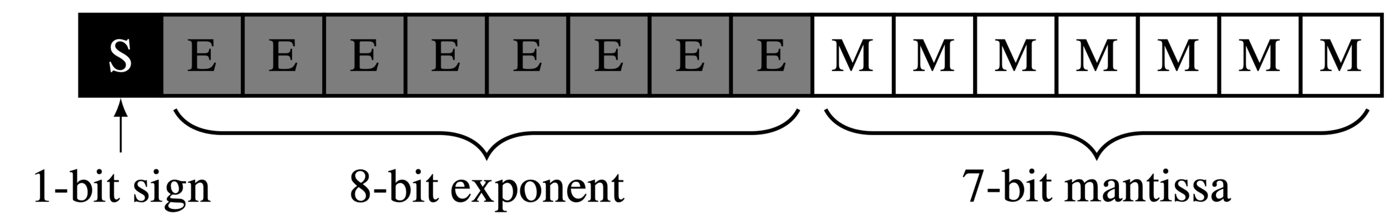
Loewenstein et al. (2011)

Learning and precision in computer science

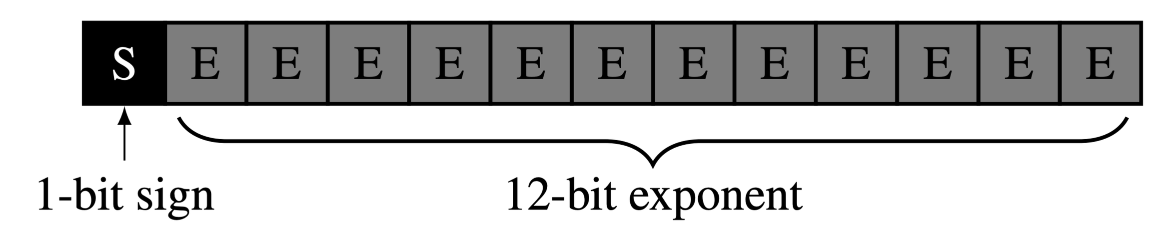
For the sake of power efficiency, we would like to reduce the bit width of deep learning hardware like GPUs and TPUs.

Computer number systems

Computers traditionally use *floating point* arithmetic. For example, the TPU employs the "bfloat16" number system:



An alternative is to use a *log number system*, where a number X is represented as (sign X, log |X|):



In terms of hardware implementation, multiplication is cheaper than addition in a log number system.

Additive learning rules

The hardware designer must choose a number system that supports stable and efficient learning. The efficiency of a number system will depend on which learning rule is used.

Most machine learning applications use additive updates:

$$w \leftarrow w - \eta \cdot g \qquad \textit{(gradient descent)},$$

for parameter w, gradient g and learning rate η.

Floating point seems like the sensible way to implement additive machine learning algorithms. For multiplicative learning rules, a log number system may be better.

Learning and precision in deep learning

Additive updates in deep learning are poorly understood, and the learning rate must be carefully tuned for each application.

Deep relative trust

Bernstein et al. (2020) studied the function and gradient of a multilayer perceptron under weight perturbation, finding the relative change in both to be roughly bounded by:

$$\prod_{l=1}^{L}\left(1+\frac{\|\Delta W_l\|_F}{\|W_l\|_F}\right)-1 \qquad \textit{(deep relative trust)}.$$

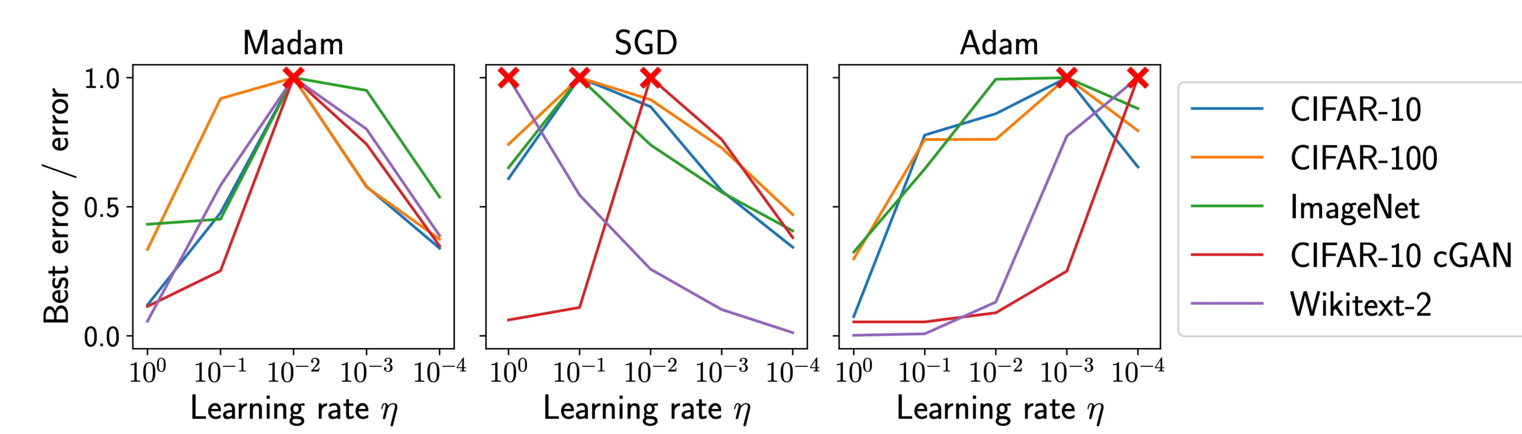
Madam learning rule

We proposed the following multiplicative learning rule:

$$w \leftarrow w \cdot \exp\left(-\eta \cdot \frac{g}{|g|_\beta} \cdot \operatorname{sign} w\right) \qquad \textit{(Madam)},$$

where $|g|_\beta$ is the RMS gradient with time constant β.

Madam respects deep relative trust, and did not require learning rate tuning across a range of experiments:



It achieved performance *close* to Adam and SGD. Madam lends itself to a log number system implementation:

Dataset	Task	FP32 Madam	12-bit	10-bit	8-bit
CIFAR-10	Resnet18	7.8 ± 0.2	7.0 ± 0.1	7.8 ± 0.3	8.6 ± 1.5
CIFAR-100	Resnet18	30.2 ± 0.1	27.6 ± 0.3	29.5 ± 0.3	33.9 ± 1.1
ImageNet	Resnet50	28.9 ± 0.1	31.1 ± 0.1	34.8 ± 0.3	50.5 ± 0.5
CIFAR-10	cGAN	19.3 ± 0.7	19.8 ± 0.8	23.4 ± 0.4	36 ± 6
Wikitext-2	Transformer	173.3 ± 0.6	182.3 ± 0.6	218.0 ± 0.6	262 ± 2

References

Bartol et al. (2015), Nanoconnectomic upper bound on the variability of synaptic plasticity.

Loewenstein et al. (2011), Multiplicative dynamics underlie the emergence of the log-normal distribution of spine sizes in the neocortex in vivo.

Bernstein et al. (2020), On the distance between two neural networks and the stability of learning.

bernstein@caltech.edu
jiawei@caltech.edu

github.com/jxbz/madam