Multiplicative noise and heavy tails in stochastic optimization

Liam Hodgkinson
Michael W. Mahoney

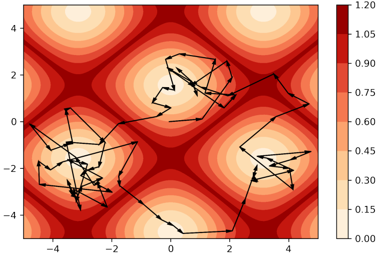
Berkeley
UNIVERSITY OF CALIFORNIA

INTERNATIONAL COMPUTER SCIENCE INSTITUTE

ICML
International Conference On Machine Learning

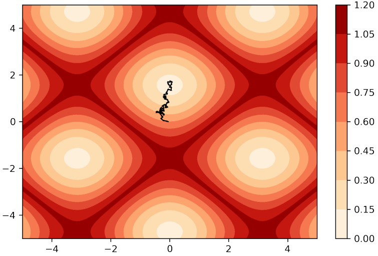
Phases of Learning

Exploration
large learning rate



(sampler)

Exploitation
small learning rate



(optimizer)

Stochastic optimization as a Markov chain

The sequence of iterated random functions

$$W_{k+1} = \Psi(W_k, X_k) \qquad X_k \overset{\text{iid}}{\sim} X.$$

Any stochastic optimization algorithm (SGD, momentum, Adam, Newton) can be written in this way.

$$W_{k+1} \approx \underbrace{\nabla\Psi(W_k, X_k)}_{\text{multiplicative}}(W_k - w^*) + \underbrace{\Psi(w^*, X_k)}_{\text{additive}}$$

Our Findings

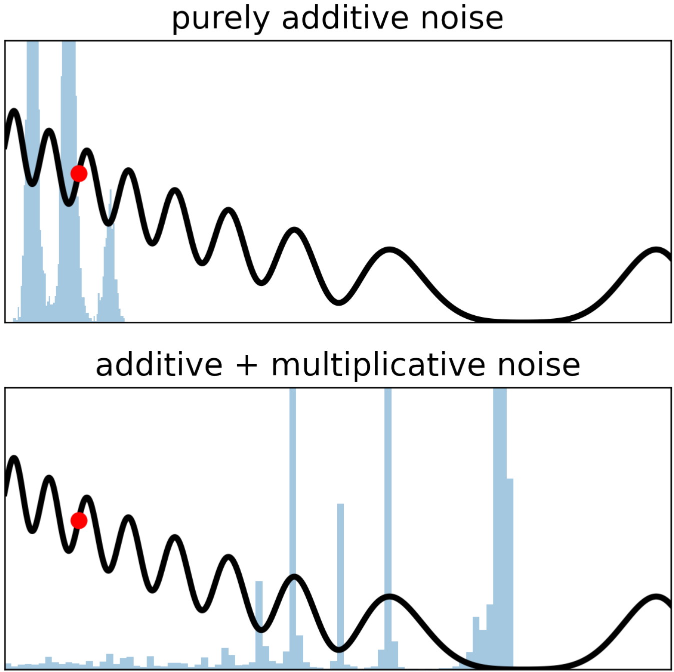
Multiplicative noise results in heavy-tailed stationary behaviour

- Decay rates in the tails that are slower than exponential are **heavy**, e.g.

$$\mathbb{P}(W > t) \approx ct^{-\alpha}$$

- Heavy-tailed fluctuations demonstrated empirically (Şimşekli et al., 2019)

Heavier tails imply wider exploration


purely additive noise
additive + multiplicative noise

Main Result

Theorem

Suppose X is non-atomic and there exist k_Ψ, K_Ψ, M_Ψ, w^* such that as $\|w\| \to \infty$,

$$k_\Psi(X) - o(1) \leq \frac{\|\Psi(w, X) - \Psi(w^*, X)\|}{\|w - w^*\|} \leq K_\Psi(X) + o(1).$$

If $\mathbb{P}(k_\Psi(X) > 1) > 0$ and $\mathbb{E} \log K_\Psi(X) < 0$, for some $\mu, \nu, C_\mu, C_\nu > 0$,

$$C_\mu (1 + t)^{-\mu} \leq \mathbb{P}(\|W_\infty\| > t) \leq C_\nu t^{-\nu}.$$

Summary

- It has been **empirically shown** that:
  - Multiplicative noise is known to be present in optimizers (Xing et al., 2018)
  - Heavy tailed fluctuations in optimizers (Şimşekli et al., 2019)
- We find that multiplicative noise induces **heavy tails** – critical to effective exploration (holds **at high generality**)
- Additive noise models are **insufficient** to analyze optimizer behaviour