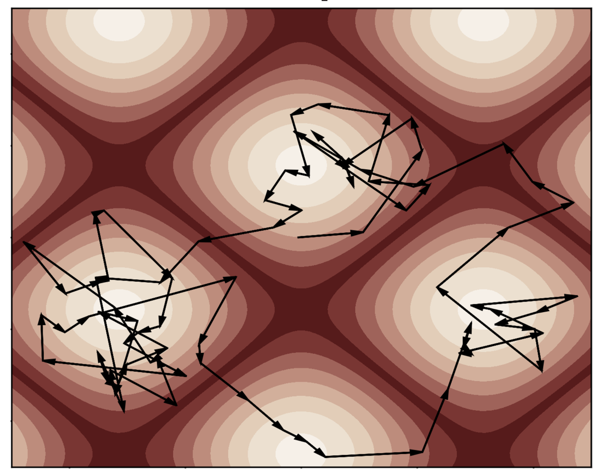
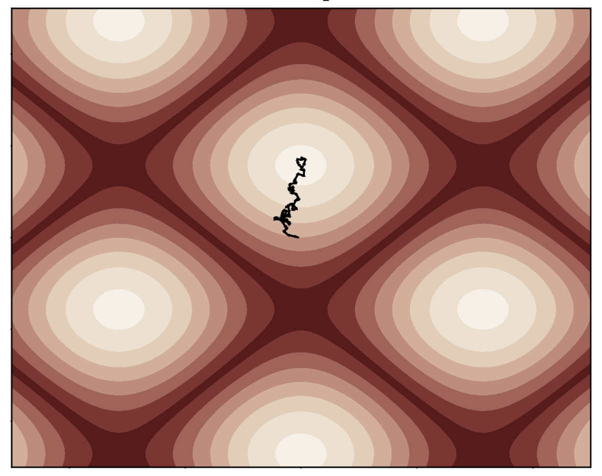
Multiplicative Noise and Heavy Tails in Stochastic Optimization

Liam Hodgkinson & Michael W. Mahoney

Phases of Learning

Exploration	***Exploitation***
large learning rate (sampler)	small learning rate ("convex" optimizer)





Every stochastic optimizer typically exhibits ***two*** phases as the learning rate is decreased. Later stages are well-studied using ***convex optimization***. ***Earlier stages*** and their effect on ***generalization*** remain elusive.

Objective

Investigate how a stochastic optimizer explores the loss landscape

1. Model stochastic optimization as a ***Markov chain***
2. Fix all hyperparameters to particular values (time-homogeneous; no annealing)
3. Examine ***stationary distribution*** (tails of the stationary distribution are an indication of capacity to ***explore***)

Empirically, fluctuations in SGD have been observed to be ***heavy-tailed*** (Şimşekli et al., 2019), i.e. $\mathbb{P}(\|\Delta W\| > w) \approx c w^{-\alpha}$ — ***why?***

Acknowledgements

We would like to acknowledge DARPA, NSF, and ONR for providing partial support of this work.

Stochastic Optimizers as Markov Chains

Problems

Minimize the expected loss over data X:

$$w^* = \arg\min_w \mathbb{E}_{X\sim\mathcal{D}}\ell(w, X),$$

for a loss ℓ depending on weights w and data X from some dataset $\mathcal{D}$.

Solve by Fixed Point Iteration

The sequence of iterated random functions

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

e.g. ***SGD*** with learning rate γ:

$$\Psi(w, x) = w - \gamma n^{-1}\textstyle\sum_{i=1}^n \nabla\ell(w, x_i).$$

Any stochastic optimizer (SGD, momentum, Adam, stochastic Newton) can be written as (1).

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

Additive vs. Multiplicative Noise

Multiplicative noise enjoys both wide ***heavy-tailed exploration*** and efficient exploitation.

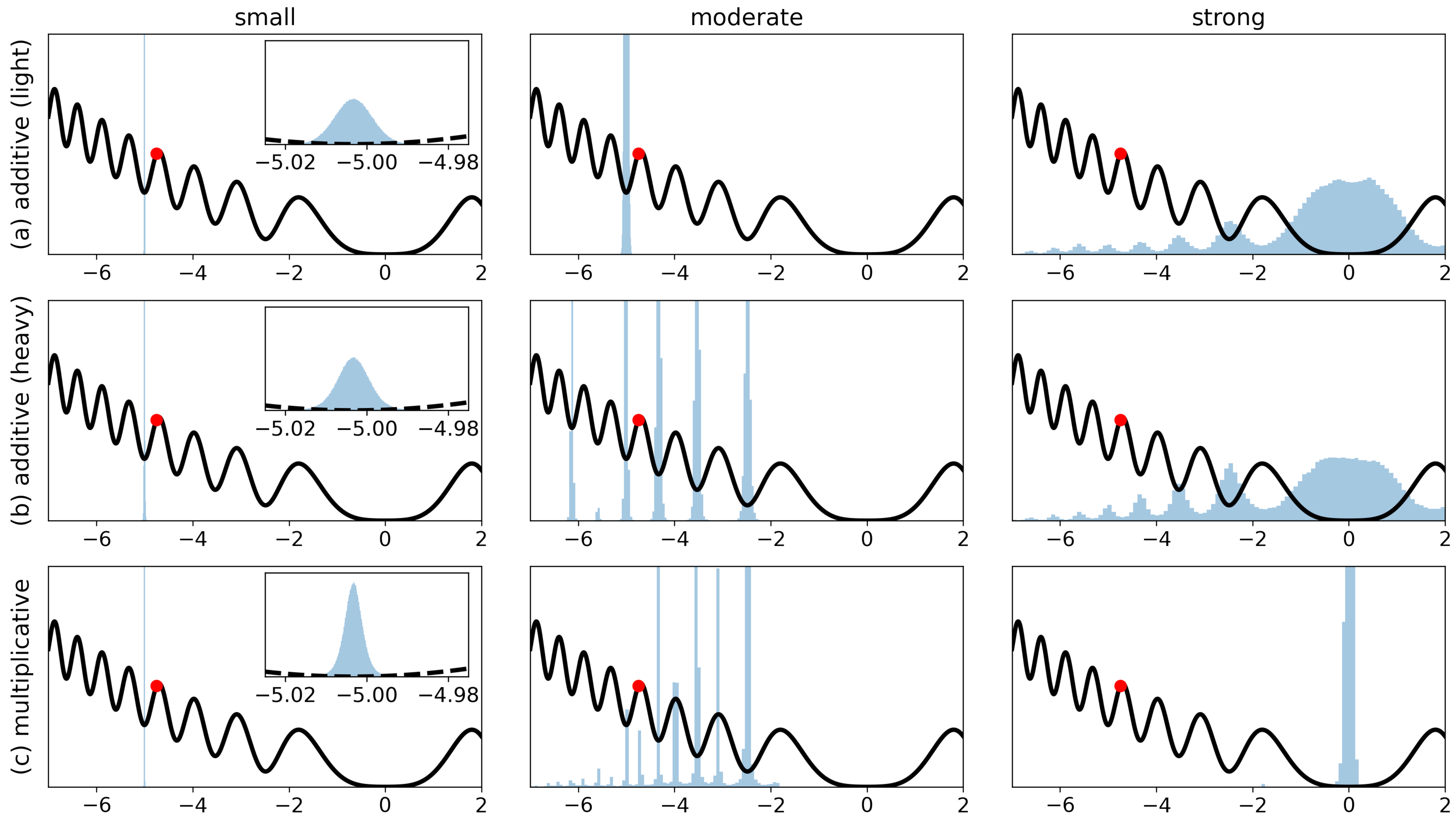


Figure 1: **Histograms** of 10^6 iterations of gradient descent with combinations of small (left), moderate (center), and strong (right) versus light additive (a), heavy additive (b), and multiplicative noise (c), applied to a **non-convex objective**. **Initial starting location** for the optimization is also shown.

Main Result

Multiplicative noise results in heavy-tailed fluctuations in stochastic optimizers

Theorem. *Suppose X is non-atomic and there exist $k_\Psi, K_\Psi, M_\Psi, w^*$ such that as $\|w\| \to \infty$,*

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

Suppose that $\mathbb{P}(\boldsymbol{k_\Psi(X)} > 1) > 0$ and $\mathbb{E}\log \boldsymbol{K_\Psi(X)} < 0$. Then **the stationary distribution is heavy-tailed**, *in particular, for some $\mu, \nu, C_\mu, C_\nu > 0$,*

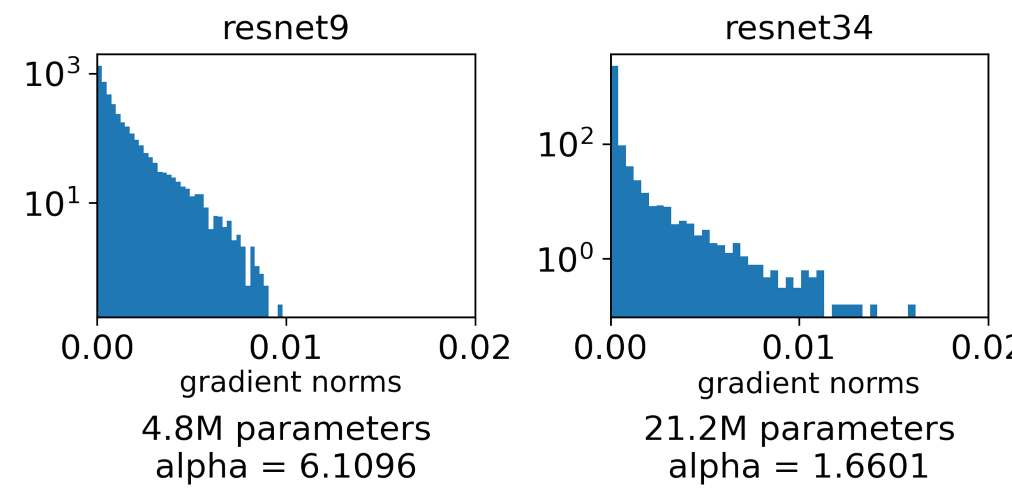
$$C_\mu(1+t)^{-\mu} \leq \mathbb{P}(\|W_\infty\| > t) \leq C_\nu t^{-\nu}.$$

e.g. holds for ***ridge regression*** when γ is large; for ***SGD***, when $\nabla^2\ell(w,X) \succ \frac{2}{\gamma}$ or $\prec \frac{2}{\gamma}$ for all w is possible.

Factors

The following results in ***heavier tails*** (and appear to correlate with improved generalization in computer vision):

- Increasing step size
- Decreasing batch size
- Increasing L^2 regularization
- Non-adaptive optimizers (SGD not Adam)
- Increasing dimension; e.g. ResNet:



Berkeley UNIVERSITY OF CALIFORNIA

ICSI

ICML International Conference On Machine Learning