Generalization Bounds using Lower Tail Exponents in Stochastic Optimization

Liam Hodgkinson, Umut Şimşekli, Rajiv Khanna & Michael W. Mahoney

Generalization Bounds

Empirical Risk Minimization

To train parameterized models, solve

$$w^* = \arg\min_{w\in\mathbb{R}^d} \mathcal{R}_n(w)$$

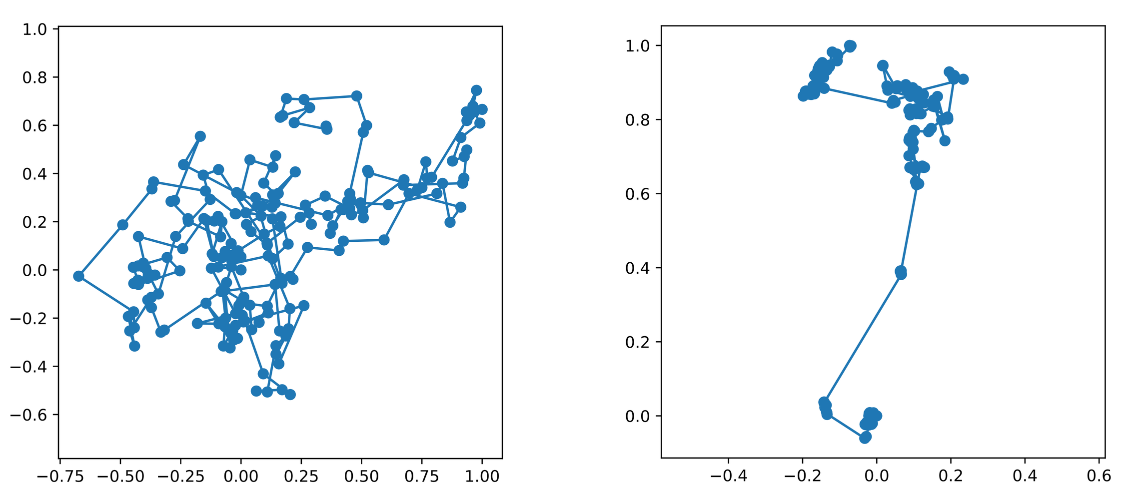
$$\mathcal{R}_n(w) := \frac{1}{n}\sum_{i=1}^{n} \ell(w, X_i),$$

for a loss ℓ depending on weights w and data $X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathcal{D}$. To quantify influence on test performance, seek bounds on the ***excess risk***

$$\mathcal{E}_n(w^*) = \mathcal{R}_n(w^*) - \overbrace{\mathbb{E}_{\mathcal{D}}\mathcal{R}_n(w^*)}^{\text{generalization}}$$

Types of Dynamics

Brownian Motion light-tailed — ***Lévy Flight*** heavy-tailed



Different stochastic optimizers (e.g. SGD, momentum, Adam) exhibit trajectories with different properties.

How do the dynamics of the optimizer influence test performance?

Heavy Tails in Machine Learning

Norms of optimizer steps in deep learning are heavy-tailed for large step sizes

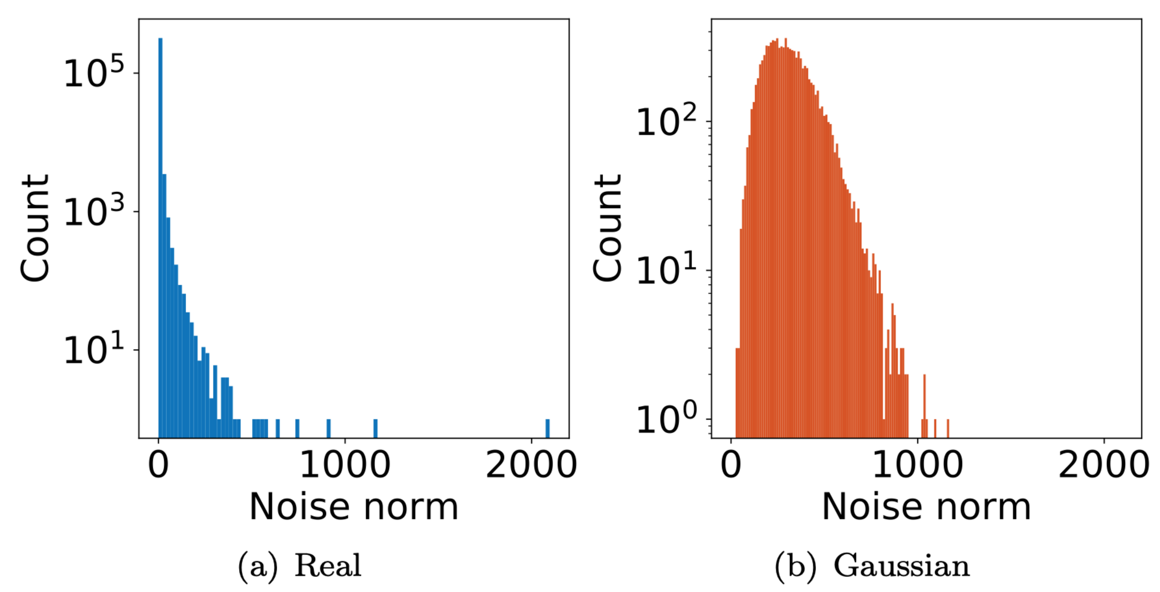


Figure 1: Histograms of (a) gradient norms from iterates of SGD on a deep learning task; (b) norms of a Gaussian random vector, as shown in [1].

As shown in [2], under a (continuous-time) Feller process model of SGD,

heavy-tailed norms $\nearrow \implies$ excess risk $\searrow$

- Optimizer trajectories exhibiting Lévy flights can be more effective
- Assumptions are complicated
- Can this be extended to ***discrete time***?

Correlations with Accuracy

Training neural networks on ***MNIST*** and ***CIFAR10*** under a variety of hyperparameters.

(FCN5) fully connected with 5 layers — ***(FCN7)*** fully connected with 7 layers — ***(CNN9)*** convolutional model with 9 layers

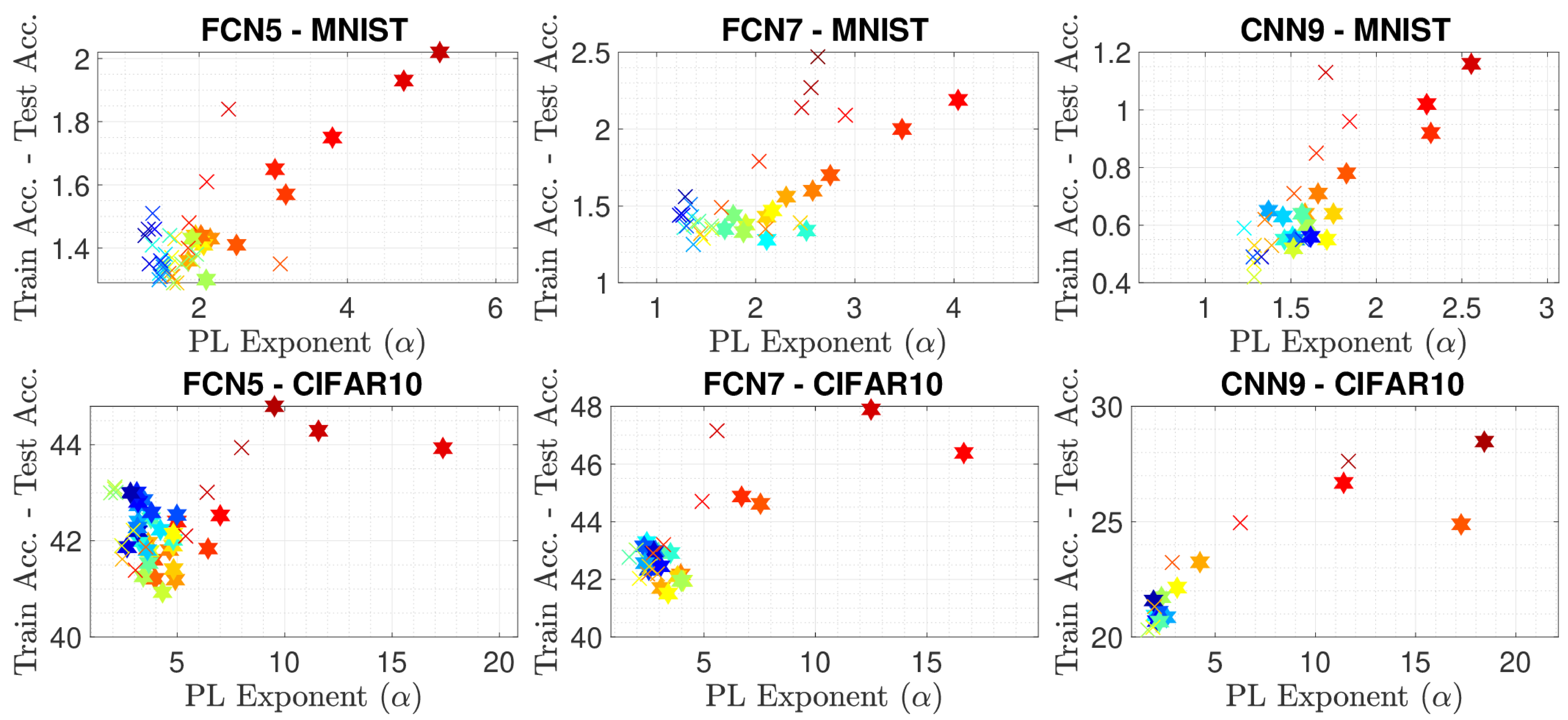


Figure 2: ***Lower tail exponents*** versus ***excess risk***. Different colors represent different step-sizes and different markers represent different batch-sizes.

Main Result

Assume that the iterates of the optimizer $W_1, W_2, \dots, W_k, \dots$, are a ***Markov chain***.

Developed a ***general proof technique*** for linking optimizer dynamics to generalization using ***generic chaining***.

Apply to Tail Exponents

- The ***upper tail exponent*** (in previous works):

$$\mathbb{P}(\|W_{k+1} - W_k\| > r) \approx \mathcal{O}(r^{-\beta}),\ r \to \infty.$$

- The ***lower tail exponent*** (we consider):

$$\mathbb{P}(\|W_{k+1} - W_k\| \le r) \approx \mathcal{O}(r^{\alpha}),\ r \to 0^+.$$

For most models of Lévy flights, $\alpha \approx \beta$.

Theorem (Informal). *Assume that the iterates W_k of an optimizer have **lower tail exponent** α in the neighbourhood of a local optimum w^*. Then an upper bound on*

$$\mathbb{E} \sup_{k=1,\dots,m} |\mathcal{E}_n(W_k)|$$

is positively correlated with α. In other words,

lower tail exponent $\searrow \implies$ excess risk $\searrow$

References

[1] Şimşekli, U., Sagun, L., & Gurbuzbalaban, M. (2019, May). A tail-index analysis of stochastic gradient noise in deep neural networks. In International Conference on Machine Learning (pp. 5827-5837). PMLR.

[2] Şimşekli, U., Sener, O., Deligiannidis, G., & Erdogdu, M. A. (2020). Hausdorff dimension, heavy tails, and generalization in neural networks. Advances in Neural Information Processing Systems, 33, 5138-5151.

Acknowledgements

We would like to acknowledge DARPA, NSF, and ONR for providing partial support of this work. U.Ş.'s research is supported by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute).

Berkeley UNIVERSITY OF CALIFORNIA — ICSI — ICML International Conference On Machine Learning