Generalization Bounds using Lower Tail Exponents in Stochastic Optimizers

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

Berkeley, University of California
International Computer Science Institute (ICSI)
Inria
ENS | PSL
Purdue University

What are generalization bounds?

Empirical Risk Minimization

To train parameterized models, solve

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

for a loss ℓ depending on weights w and data $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{D}$.

Generalization bounds

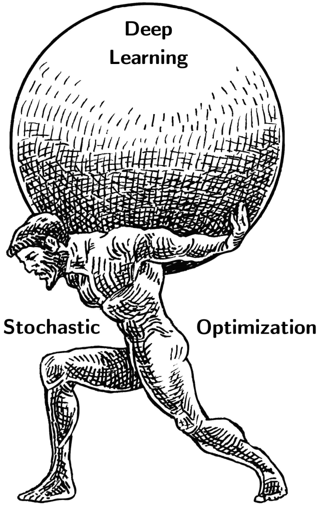
Bounds on the *generalization error*

$$\mathcal{E}_n(w^*) = \mathcal{R}_n(w^*) - \underbrace{\mathbb{E}_{\mathcal{D}}\mathcal{R}_n(w^*)}_{\text{population risk}}$$

Stochastic optimization

is the process of minimizing an objective function via the simulation of random elements.

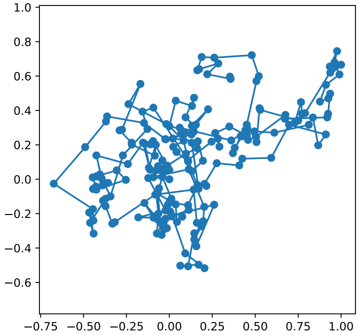
"the backbone of modern machine learning"



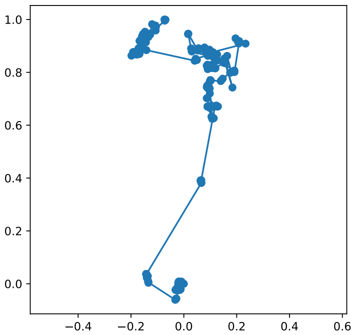
How do the dynamics of the optimizer influence generalization?

Types of Dynamics

Brownian motion
light-tailed

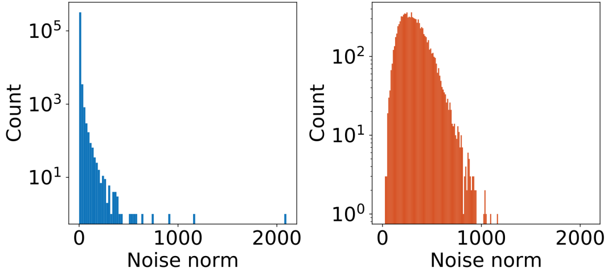


Lévy flight
heavy-tailed



Heavy Tails in Machine Learning

Norms of optimizer steps in a deep learning task



(a) Real (b) Gaussian

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.

Previous Work

Under a **(continuous-time) Feller process model** of SGD,

$$\text{heavier tails} \implies \text{smaller } \mathcal{E}_n.$$

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.

- Complicated assumptions
- What about **discrete time**, i.e. SGD itself?

Markov Assumption

Assume that the iterates of the optimizer

$$W_1, W_2, \ldots, W_k, \ldots$$

are a **Markov chain**.

Upper Tail Exponent

Previous works have considered the **upper tail exponent**:

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

as $r \to \infty$.

Lower Tail Exponent

What about the **lower tail exponent**?

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

as $r \to 0^+$.

Lower Tail Exponent

Theorem (Informal)

Assume that iterates W_k of an optimizer satisfy

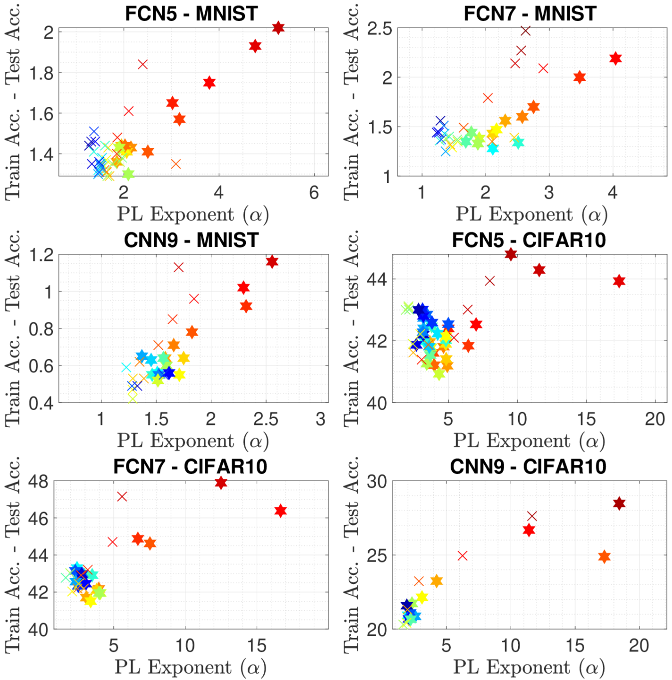
$$\mathbb{P}(\|W_{k+1} - W_k\| \leq r) \approx \mathcal{O}(r^{\alpha})$$

in the neighbourhood of a local optimum w^*. Then an upper bound on

$$\mathbb{E} \sup_{k=1,\ldots,m} |\mathcal{E}_n(W_k)| \text{ is positively correlated with } \alpha.$$

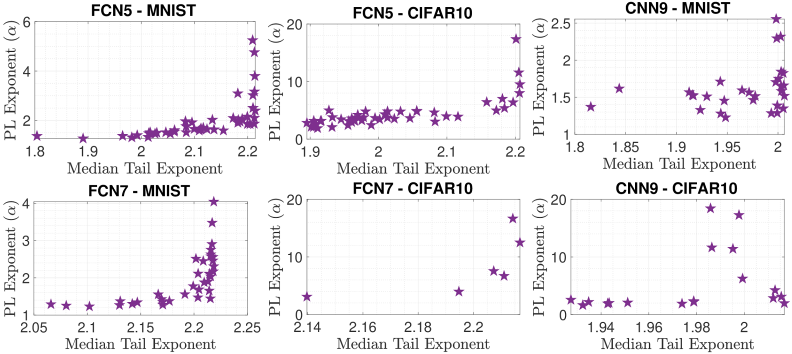
Is this true in practice?

Train NNs with varying hyperparameters & regularization


FCN5 - MNIST
FCN7 - MNIST
CNN9 - MNIST
FCN5 - CIFAR10
FCN7 - CIFAR10
CNN9 - CIFAR10
Train Acc. - Test Acc.
PL Exponent (α)

Lower Tail Exponent

Lower tail often correlates with upper tail



Contributions and Conclusions

- Developed a **general proof technique** for linking optimizer dynamics to generalization
- Extended results of Şimşekli et al., 2020.
- Lower tail exponent correlates with $\mathcal{E}_n$
  - Supported in practice
  - Lower tail correlates with upper tail