Generalization bounds for heavy-tailed SDEs

(link to preprint)

Benjamin Dupuis - Umut Şimşekli

March 27th, 2024

Why heavy-tailed algorithms?

Motivation:

1. Why heavy-tailed algorithms?
2. Why are they interesting?

A few generic notation

On a data space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ endowed with a probability distribution μ_z, we want to minimize the **population risk**

$$\min_{w \in \mathbb{R}^d} \left\{ L(w) := \mathop{\mathbb{E}}_{z \sim \mu_z} [\ell(w, z)] := \mathop{\mathbb{E}}_{(x,y) \sim \mu_z} [\mathcal{L}(h_w(x), y)] \right\},$$

Empirical risk over a dataset $S = (z_1, \dots, z_n) \sim \mu_z^{\otimes n}$

$$\widehat{L}_S(w) := \frac{1}{n} \sum_{i=1}^n \ell(w, z_i).$$

Generalization error:

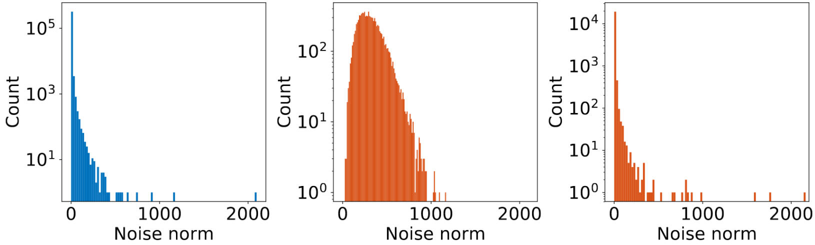
$$G_S(w) := L(w) - \widehat{L}_S(w). \tag{1}$$

1: Heavy tails as a modelisation of SGD

- SGD:

$$w_{k+1} = w_k - \frac{\eta}{b} \sum_{i \in B_k} \nabla \ell(w_k, z_i)$$

- What does the gradient noise look like [8]?

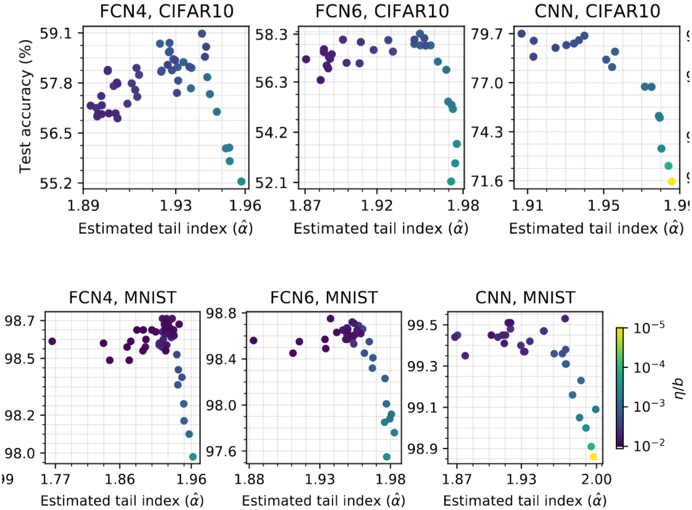


(a) Real (b) Gaussian (c) α-stable

- Other authors injected heavy-tailed noise in the algorithm to improve the generalization performance.

2: Generalization error of heavy-tailed algorithms

- **Experimental works:** complex dependence between α and the accuracy gap [1].



The simplified model we study

Simplified model

Continuous-time model:

$$dW_t = -\nabla \widehat{V}_S(W_t)dt + \sigma dL_t^{\alpha}.$$

Discrete version:

$$\widehat{W}_{k+1}^S = \widehat{W}_k^S - \eta \nabla \widehat{V}_S(\widehat{W}_k^S) + \eta^{\frac{1}{\alpha}} \sigma L_1^{\alpha}.$$

L_t^{α} is a Lévy process, for $\alpha \in (0, 2]$:

- $\alpha = 2$ corresponds to Brownian motion (Gaussian noise).
- the smaller α the higher the tail of the noise.
- We also added regularization, for technical reasons:

$$\widehat{V}_S(w) = \widehat{L}_S(w) + \frac{\lambda}{2} \|w\|^2 . \tag{2}$$

Previous works

For a fixed time horizon $T > 0$, the goal is to get a bound on:

$$G_S(W_T) := L(W_T) - \widehat{L}_S(W_T), \tag{3}$$

where $(W_t)_{t \geq 0}$ is solution of the previous equation.

Previous works

For a fixed time horizon $T > 0$, the goal is to get a bound on:

$$G_S(W_T) := L(W_T) - \widehat{L}_S(W_T), \tag{3}$$

where $(W_t)_{t \geq 0}$ is solution of the previous equation.

Previous approaches:

- **Fractal-based approaches** [9]
  - great in some settings but...
  - does not predict the observed tail-index behavior
- **Stability-based approaches** [7, 6]:
  - Only expected bound
  - Huge dependence on the dimension d
  - Can predict the non-monotonic behavior wrt α

Our work: New theoretical approach

We combine new PAC-Bayesian techniques with the study of the associated **'fractional' Fokker-Planck equation**, as done for Langevin dynamics [5, 2, 4].

$$dW_t = -\nabla \widehat{V}_S(W_t)dt + \sigma dL_t^\alpha \quad \Longrightarrow \quad \frac{\partial}{\partial_t} u_t = -\sigma_1^\alpha (-\Delta)^{\frac{\alpha}{2}} u_t + \mathrm{div}(u_t \nabla \widehat{V}_S),$$

with u_t the probability density of W_t.

Our work: New theoretical approach

We combine new PAC-Bayesian techniques with the study of the associated **'fractional' Fokker-Planck equation**, as done for Langevin dynamics [5, 2, 4].

$$dW_t = -\nabla \widehat{V}_S(W_t)dt + \sigma dL_t^\alpha \quad \Longrightarrow \quad \frac{\partial}{\partial_t} u_t = -\sigma_1^\alpha \left(-\Delta\right)^{\frac{\alpha}{2}} u_t + \mathrm{div}(u_t \nabla \widehat{V}_S),$$

with u_t the probability density of W_t.

Main result (informal, partial)

With probability at least $1 - \zeta$:

$$\mathbb{E}_{W_T \sim u_T}\left[G_S(W_T)\right] \leq 2s \sqrt{\frac{K_{\alpha,d}}{n\sigma^\alpha} \int_0^T \mathbb{E}_U \left\| \nabla \widehat{L}_S(W_t^S) \right\|^2 dt + \frac{\log(3/\zeta) + \Lambda}{n}}, \tag{4}$$

with:

$$K_{\alpha,d} = \frac{(2-\alpha)\Gamma\left(1 - \frac{\alpha}{2}\right) d\Gamma\left(\frac{d}{2}\right)}{\alpha 2^\alpha \Gamma\left(\frac{d+\alpha}{2}\right) R^{2-\alpha}},$$

Proof idea?

- Inspired from existing works in the case of Gaussian noise
- Computation of an **entropy flow**, inspired by [3]:

$$\frac{d}{dt}\mathrm{KL}(u_t, \bar{u}_\infty) = -\sigma^\alpha B^\alpha(u_t, \bar{u}_\infty) - \int u_t(x)\nabla \frac{u_t}{\bar{u}_\infty} \cdot \nabla \hat{F}_S(x)dx$$

Distribution at time t

Distribution related to the regularization

New term: ‘Bregman integral’

$$\leq \frac{1}{2C}\int \|\nabla \hat{F}_S(x)\|^2 u_t(x)dx + \frac{C}{2}\int \|\nabla \log \frac{u_t(x)}{\bar{u}_\infty}\|^2 u_t(x)dx$$

Fisher information

- Most of the complexity is contained in the term $(-\Delta)^{\frac{\alpha}{2}} u$, called the fractional Laplacian.
- The main technical idea is to bound the so-called Bregman integral term.

Proof idea? (2)

- It allows to use **PAC-Bayesian theory**
- If the loss is s-subgaussian, we prove that:

$$\mathbb{E}_{W_T \sim u_T}\left[G_S(W_T)\right] \leq 2s\sqrt{\frac{\mathrm{KL}(u_t, \bar{u}_\infty) + \log(3/\zeta)}{n}}. \tag{5}$$

Proof idea? (2)

- It allows to use **PAC-Bayesian theory**
- If the loss is s-subgaussian, we prove that:

$$\mathbb{E}_{W_T \sim u_T}\left[G_S(W_T)\right] \leq 2s\sqrt{\frac{\mathrm{KL}(u_t, \bar{u}_\infty) + \log(3/\zeta)}{n}}. \quad (5)$$

Main result (informal, partial)

With probability at least $1 - \zeta$:

$$\mathbb{E}_{W_T \sim u_T}\left[G_S(W_T)\right] \leq 2s\sqrt{\frac{K_{\alpha,d}}{n\sigma^\alpha}\int_0^T \mathbb{E}_U\left\|\nabla \widehat{L}_S(W_t^S)\right\|^2 dt + \frac{\log(3/\zeta) + \Lambda}{n}}, \quad (6)$$

with:

$$K_{\alpha,d} = \frac{(2-\alpha)\Gamma\left(1 - \frac{\alpha}{2}\right) d\Gamma\left(\frac{d}{2}\right)}{\alpha 2^\alpha \Gamma\left(\frac{d+\alpha}{2}\right) R^{2-\alpha}},$$

Quantitative analysis

- Gaussian limit: $K_{\alpha,d} \underset{\alpha \to 2^-}{\longrightarrow} \frac{1}{2}$.
- High-dimensional limit:

$$K_{\alpha,d} \underset{d \to \infty}{\sim} \frac{(2-\alpha)\Gamma\left(1-\frac{\alpha}{2}\right)}{R^{2-\alpha}\alpha 2^{\alpha/2}} d^{1-\frac{\alpha}{2}} \tag{7}$$

Quantitative analysis

- Gaussian limit: $K_{\alpha,d} \underset{\alpha\to 2^-}{\longrightarrow} \frac{1}{2}$.
- High-dimensional limit:

$$K_{\alpha,d} \underset{d\to\infty}{\sim} \frac{(2-\alpha)\Gamma\left(1-\frac{\alpha}{2}\right)}{R^{2-\alpha}\alpha 2^{\alpha/2}} d^{1-\frac{\alpha}{2}} \tag{7}$$

Phase transition

In the limit $d \to \infty$, the constant term is:

$$\frac{K_{\alpha,d}}{n\sigma_1^{\alpha}} \approx \frac{P_\alpha d_0}{n(\sigma\sqrt{d_0})^{\alpha}}, \tag{8}$$

where $d_0 := d/(R^2)$ is a "reduced dimension".

We identify two regimes whether $\sigma\sqrt{d_0} > 1$ or $\sigma\sqrt{d_0} > 1$.

Experimental results

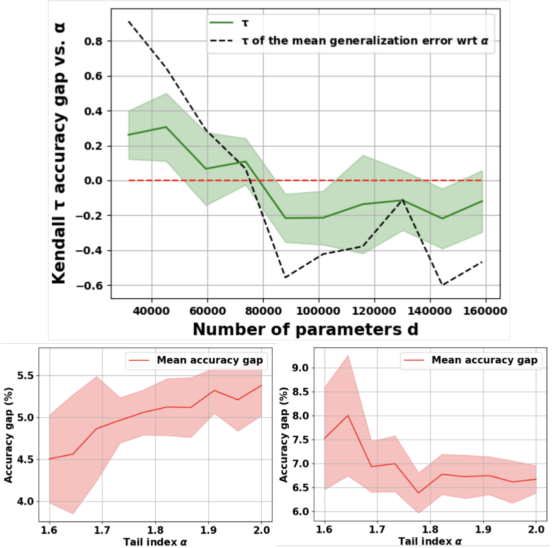


Figure: *(up)* Correlation (Kendall's τ) between α and the accuracy gap. FCN2 trained on MNIST. Green curve: average τ over 10 random seeds. Black curve is the correlation between α and the average accuracy gap over 10 seeds.

Experimental results 2

$$\widehat{G} := \sqrt{\frac{P_\alpha d^{1-\frac{\alpha}{2}}\gamma}{n\sigma_1 R^{2-\alpha}} \sum_{k=1}^{N} \left\| \nabla \widehat{F}_S(\widehat{W}_k^S) \right\|^2}. \quad (9)$$

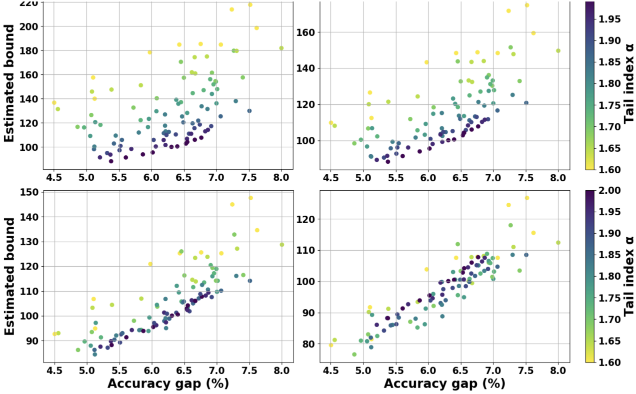


Figure: Estimated bound versus accuracy gap for a FCN2 on MNIST, for different values of R: 1 *(top left)*, 3 *(top right)*, 7 *(bottom left)*, 15 *(bottom right)*.

Experimental results 3

We perform the linear regression:

$$\log(\widehat{G}) \simeq \widehat{r} \log(d) + C, \tag{10}$$

and "estimate" α according to our model $\widehat{\alpha} := 2 - 4\widehat{r}$.

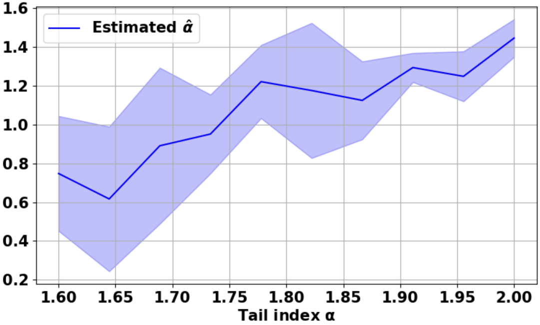
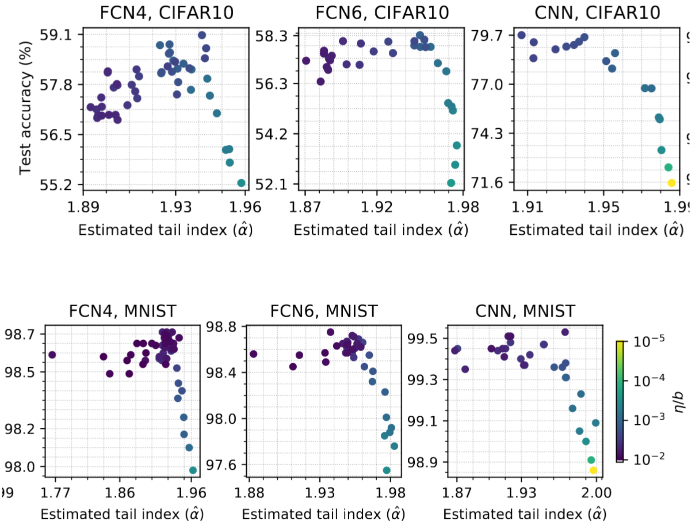


Figure: Regression of the tail-index α from the accuracy error, for a FCN2 trained on MNIST.

Conclusion



FCN4, CIFAR10
FCN6, CIFAR10
CNN, CIFAR10
Test accuracy (%)
59.1
57.8
56.5
55.2
1.89
1.93
1.96
Estimated tail index ($\hat{\alpha}$)
58.3
56.3
54.2
52.1
1.87
1.92
1.98
79.7
77.0
74.3
71.6
1.91
1.95
FCN4, MNIST
FCN6, MNIST
CNN, MNIST
98.7
98.5
98.2
98.0
1.77
1.86
1.96
98.8
98.4
98.0
97.6
1.88
1.93
1.98
99.5
99.3
99.1
98.9
1.87
1.93
2.00
10^{-5}
10^{-4}
10^{-3}
10^{-2}
η/b

[1] M. Barsbey, M. Sefidgaran, M. A. Erdogdu, G. Richard, and U. Şimşekli. Heavy Tails in SGD and Compressibility of Overparametrized Neural Networks. In *35th Conference on Neural Information Processing Systems (NeurIPS 2021)*. arXiv, June 2021.

[2] F. Futami and M. Fujisawa. Time-Independent Information-Theoretic Generalization Bounds for SGLD. In *7th Conference on Neural Information Processing Systems (NeurIPS 2023)*. arXiv, Nov. 2023.

[3] I. Gentil and C. Imbert. Logarithmic Sobolev inequalities: Regularizing effect of Lévy operators and asymptotic convergence in the Lévy-Fokker-Planck equation. *Asymptotic analysis*, Sept. 2008.

[4] J. Li, X. Luo, and M. Qiao. On Generalization Error Bounds of Noisy Gradient Methods for Non-Convex Learning. In *Published as a Conference Paper at ICLR 2020*. arXiv, Feb. 2020.

[5] W. Mou, L. Wang, X. Zhai, and K. Zheng. Generalization Bounds of SGLD for Non-convex Learning: Two Theoretical Viewpoints. In *Proceedings of the 31st Conference On Learning Theory*. arXiv, July 2017.

[6] A. Raj, M. Barsbey, M. Gürbüzbalaban, L. Zhu, and U. Şimşekli. Algorithmic Stability of Heavy-Tailed Stochastic Gradient Descent on Least Squares. In *Proceedings of The 34th International Conference on Algorithmic Learning Theory,*. arXiv, Feb. 2023.

[7] A. Raj, L. Zhu, M. Gürbüzbalaban, and U. Şimşekli. Algorithmic Stability of Heavy-Tailed SGD with General Loss Functions. In *International Conference*