Generalization Bounds for Heavy-Tailed SDEs through the Fractional Fokker-Planck Equation

Benjamin Dupuis [1 2 3] **Umut Şimşekli** [1 2 3 4]

Abstract

Understanding the generalization properties of heavy-tailed stochastic optimization algorithms has attracted increasing attention over the past years. While illuminating interesting aspects of stochastic optimizers by using heavy-tailed stochastic differential equations as proxies, prior works either provided *expected* generalization bounds, or introduced non-computable information theoretic terms. Addressing these drawbacks, in this work, we prove *high-probability* generalization bounds for heavy-tailed SDEs which do not contain any nontrivial information theoretic terms. To achieve this goal, we develop new proof techniques based on estimating the entropy flows associated with the so-called *fractional* Fokker-Planck equation (a partial differential equation that governs the evolution of the distribution of the corresponding heavy-tailed SDE). In addition to obtaining high-probability bounds, we show that our bounds have a better dependence on the dimension of parameters as compared to prior art. Our results further identify a phase transition phenomenon, which suggests that heavy tails can be either beneficial or harmful depending on the problem structure. We support our theory with experiments conducted in a variety of settings.

1. Introduction

A supervised machine learning setup consists of a data space $\mathcal{Z}$, a data distribution μ_z, and a parameter space, which will be $\mathbb{R}^d$ in our study. Given a loss function $\ell : \mathbb{R}^d \times \mathcal{Z} \longrightarrow \mathbb{R}_+$, the goal is to minimize the following population risk:

$$\min_{w\in\mathbb{R}^d} L(w), \quad L(w) := \mathbb{E}_{z\sim\mu_z}\left[\ell(w,z)\right]. \tag{1}$$

As μ_z is typically unknown in practice, the population risk L is replaced by the empirical risk, defined as follows:

$$\widehat{L}_S(w) := \frac{1}{n}\sum_{i=1}^{n} \ell(w, z_i), \tag{2}$$

where $S = (z_1, \dots, z_n) \sim \mu_z^{\otimes n}$ is a dataset and each z_i is sampled independent and identically (i.i.d.) from μ_z. Even though $\widehat{L}_S$ can be computed in practice as opposed to L, in several practical scenarios, ℓ is further replaced with a 'surrogate loss' function $f : \mathbb{R}^d \times \mathcal{Z} \longrightarrow \mathbb{R}$. For instance, in a binary classification setting, ℓ is typically chosen as the non-differentiable 0-1 loss, whereas f can be chosen as a differentiable surrogate, such as the cross-entropy loss, which would be amenable to gradient-based optimization. We accordingly define the *surrogate empirical risk*:[1]

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

Given a dataset S and a surrogate loss f, a stochastic optimization algorithm $\mathcal{A}$ aims at minimizing $\widehat{F}_S$ and can be seen as a function such that $\mathcal{A}(S, U) =: w_{S,U}$, where U is a random variable encompassing all the randomness in the algorithm. One of the major challenges of statistical learning theory is then to upper-bound the so-called generalization error, i.e., $L(w_{S,U}) - \widehat{L}_S(w_{S,U})$. Once such a bound can be obtained, it immediately provides an upper-bound on the true risk L, as $\widehat{L}_S$ can be computed numerically.

In our study, we analyze the generalization error induced by a specific class of *heavy-tailed* optimization algorithms, described by the next stochastic differential equation (SDE):

$$dW_t^S = -\nabla V_S(W_t^S)dt + \sigma_1 dL_t^\alpha + \sqrt{2}\sigma_2 dB_t, \tag{3}$$

where L_t^α is a stable Lévy process, which will be formally introduced in Section 2, $\alpha \in (1, 2)$ is the tail-index, controlling the heaviness of the tails[2] (examples of Lévy processes are provided in Fig. 1), B_t is a Brownian motion in $\mathbb{R}^d$,

[1]Inria [2]Ecole Normale Supérieure, Paris, France [3]PSL Research University, Paris, France [4]CNRS. Correspondence to: Benjamin Dupuis <benjamin.dupuis@inria.fr>, Umut Şimşekli <umut.simsekli@inria.fr>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

[1]In our theoretical setup we introduce the surrogate loss to be able to cover more general settings. However, this is not a requirement, we can set $f = \ell$.

[2]L_t^α does not admit a finite variance and as α gets smaller the process becomes heavier-tailed.

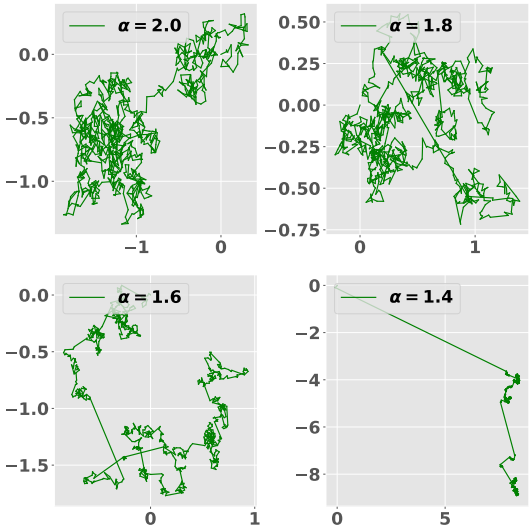


Figure 1. Simulation of L_t^α for different values of α.

$\sigma_{1,2} \geq 0$ are fixed constants, and the *potential* V_S is the ℓ_2-regularized loss that is defined as:

$$V_S(w) := \widehat{F}_S(w) + \frac{\eta}{2}\|w\|^2. \tag{4}$$

The term $\frac{\eta}{2}\|w\|^2$ corresponds to a ℓ_2 weight decay that is commonly used in the theoretical analysis of Langevin dynamics, which corresponds to (3) with $\sigma_1 = 0$ (Mou et al., 2017; Li et al., 2020; Farghly & Rebeschini, 2021).

There has been an increasing interest in understanding the theoretical properties of heavy-tailed SDEs, such as (3), due to two main reasons.

1. Recently, several studies have provided empirical and theoretical evidence that stochastic gradient descent (SGD) can exhibit heavy tails when the step-size is chosen large, or the batch size small (Simsekli et al., 2019; Gurbuzbalaban et al., 2021; Hodgkinson & Mahoney, 2020; Pavasovic et al., 2023). This has motivated several studies (see e.g., (Nguyen et al., 2019a; Şimşekli et al., 2021; Simsekli et al., 2019; Zhou et al., 2020)) to model the heavy tails, emerging in the large step-size/small batch-size regime, through heavy-tailed SDEs and analyze the resulting SDE as a proxy for SGD.
2. Injecting heavy-tailed noise to SGD in an explicit way has also been considered from several perspectives. It has been shown that heavy-tailed noise can help the algorithm avoid sharp minima (Şimşekli, 2017; Simsekli et al., 2019; Nguyen et al., 2019b;a), attain better generalization properties (Lim et al., 2022; Raj et al., 2023a) or to obtain sparse parameters in an overparametrized neural network setting (Wan et al., 2023).

Our main goal in this study is to develop *high-probability* generalization bounds for the SDE given in (3). More precisely, we will choose the learning algorithm $\mathcal{A}$ as the solution to the SDE (3), i.e., $\mathcal{A}(S,U) = W_T^S$ for some fixed time horizon $T > 0$, where in this case U will encapsulate the randomness introduced by L_t^α and B_t. We will then upper-bound the generalization gap $\mathbb{E}_U\left[G_S(W_T^S)\right]$ over $S \sim \mu_z^{\otimes n}$ under this specific choice, where

$$G_S(W_T^S) := L(W_T^S) - \widehat{L}_S(W_T^S). \tag{5}$$

Related work. In the case where the SDE (3) is only driven by a Brownian motion, i.e. $\sigma_1 = 0$, Eq. (3) reduces to the continuous Langevin dynamics, whose generalization properties have been widely studied (Mou et al., 2017; Li et al., 2020; Farghly & Rebeschini, 2021; Futami & Fujisawa, 2023), as well as its discrete-time counterpart (Raginsky et al., 2017; Pensia et al., 2018; Negrea et al., 2020; Haghifam et al., 2020; Neu et al., 2021; Farghly & Rebeschini, 2021). For instance, Mou et al. (2017) distinguished two different approaches: the first is based the concept of algorithmic stability (Bousquet, 2002; Bousquet et al., 2020), while the second is based on PAC-Bayesian theory (Shawe-Taylor & Williamson, 1997; McAllester, 1998; Catoni, 2007; Germain et al., 2009). In our work, we extend this approach to handle the presence of heavy-tailed noise.

A first step toward generalization bounds for heavy-tailed dynamics was achieved by leveraging the fractal structures generated by such SDEs (Şimşekli et al., 2021; Hodgkinson et al., 2022; Dupuis & Viallard, 2023). These studies successfully brought to light new empirical links between the sample paths of these SDEs and topological data analysis (Birdal et al., 2021; Dupuis et al., 2023; Andreeva et al., 2023). However, the uniform bounds developed in these studies contain intricate mutual information terms between the set of points of the trajectory and the dataset, which are not amenable for numerical computation to our knowledge (Dupuis et al., 2023). Closest to our work are the results recently obtained by Raj et al. (2023b;a), in the case of pure heavy-tailed noise (i.e. $\sigma_2 = 0$). Raj et al. (2023b) used an algorithmic stability argument to derive expected generalization bounds. While their approach provided more explicit bounds that do not contain mutual information terms, it still has certain drawbacks: *(i)* the proof technique cannot be directly used to derive high probability bounds and *(ii)* their bound has a strong dependence in the dimension d, rendering it vacuous in overparameterized settings.

Contributions. In our work, we aim to solve these issues by introducing new tools, taking inspiration from the PAC-Bayesian techniques already used in the case of Langevin dynamics. In particular, we will leverage recent results on fractional partial differential equations (Gentil & Imbert, 2008; Tristani, 2013), and use them to extend the analysis technique presented in Mou et al. (2017) to our heavy-tailed setting. While the presence of the heavy tails makes our task significantly more technical, our results unify both light-

tailed and heavy-tailed models around one proof technique. Our contributions are as follows:

- We derive high-probability generalization bounds, first when $\sigma_2 > 0$, then in the case $\sigma_2 = 0$, which turns out to introduce the main technical challenge to our task. Informally, our result takes the following form, with high probability over $S \sim \mu_z^{\otimes n}$,

$$\mathbb{E}_U\left[G_S(W_T^S)\right] \lesssim \sqrt{\frac{K_{\alpha,d}}{n\sigma_1^\alpha}\int_0^T \mathbb{E}_U\left\|\nabla \widehat{F}_S(W_t)\right\|^2 dt},$$

  where U denotes the randomness coming from L_t^α and B_t, and $K_{\alpha,d}$ is a constant depending on α and d. We further provide additional results where the resulting bound has a different form and is time-uniform (i.e., does not diverge with T) at the expense of introducing terms that cannot be computed in a straightforward way.
- By analyzing the constant $K_{\alpha,d}$, we study the impact of the tail-index α on our bounds. Our analysis reveals the existence of a phase transition: we identify two regimes, where in the first case heavy tails are malicious, i.e., the bound increases with the increasing heaviness of the tails, whereas, in the second regime, the heavy tails are beneficial, i.e., increasing the heaviness of the tails results in smaller bounds. Furthermore, we show that our bounds have an improved dependence on the dimension d compared to (Raj et al., 2023b).

We support our theory with various experiments conducted on several models. As our experiments require discretizing the dynamics (3), we analyze the extension of our bounds to a discrete setting, as an additional contribution, see Appendix C.6. All the proofs are presented in the Appendix. The code for our numerical experiments is available at `https://github.com/benjiDupuis/heavy_tails_generalization`.

2. Technical Background

2.1. Levy processes and Fokker-Planck equations

A Lévy process $(L_t)_{t\geq 0}$ is a stochastic process which is stochastically continuous and has stationary and independent increments, with $L_0 = 0$. We are interested in a specific class of such processes, called symmetric (strictly) α-stable processes, which we denote L_t^α. These processes are defined through the characteristic function of their increments, i.e., $\mathbb{E}\left[e^{i\xi\cdot(L_t^\alpha - L_s^\alpha)}\right] = e^{-|t-s|^\alpha\|\xi\|^\alpha}$. When the tail-index, α, is 2, then L_t^2 corresponds to $\sqrt{2}B_t$, where B_t is a standard Brownian motion in $\mathbb{R}^d$. For $\alpha < 2$, the processes have heavy-tailed distributions and exhibit jumps (see Fig. 1). We restrict our study to $\alpha > 1$, since when $\alpha \leq 1$, the expectation of L_t^α is not defined, which may introduce technical complications and does not have a clear practical interest. We provide further details on Lévy processes in Appendix A.2, see also (Schilling, 2016).

As mentioned in the introduction, the learning algorithm treated in this study consists in the SDE (3), defined in the Itô sense (Schilling, 2016, Section 12), which generalizes both Langevin dynamics (Mou et al., 2017; Li et al., 2020) and purely heavy-tailed dynamics (Raj et al., 2023b).

Inspired by (Mou et al., 2017; Li et al., 2020), our proofs will not be directly based on this SDE, but on an associated partial differential equation, called the *fractional* Fokker-Planck equation (FPE), or the forward Kolmogorov equation (Umarov et al., 2018). This equation describes the evolution of the probability density function $u_t^S(w) := u^S(t,w)$ the random variable W_t^S, that is the solution of Eq. (3). Following (Duan, 2015; Umarov et al., 2018; Schilling & Schnurr, 2010), Eq. (3) is associated with the following FPE:

$$\frac{\partial}{\partial_t}u_t^S = -\sigma_1^\alpha\left(-\Delta\right)^{\frac{\alpha}{2}}u_t^S + \sigma_2^2\Delta u_t^S + \operatorname{div}(u_t^S\nabla V_S), \quad (6)$$

where $(-\Delta)^{\frac{\alpha}{2}}$ is the (negative) fractional Laplacian operator, which is formally defined in A.2, see also (Daoud & Laamri, 2022; Schertzer et al., 2001) for introductions.

2.2. PAC-Bayesian bounds

Based on the notations of Eq. (3), the learning algorithm studied in this paper is a random map that takes the data $S \in \mathcal{Z}^n$ as input and generates W_T^S as the output. Due to the randomness introduced by B_t and L_t^α, this procedure defines a *randomized predictor*, i.e., given S, the output W_T^S follows a certain probability distribution.

Generalization properties of similar randomized predictors have been popularly studied through the PAC-Bayesian theory (see (Alquier, 2021) for a formal introduction). Informally, in PAC-Bayesian analysis, a generalization bound is typically based on some notion of distance between a *posterior* distribution over the predictors, typically denoted by ρ_S, a data-dependent probability distribution on $\mathbb{R}^d$, and a data-independent distribution over the predictors, typically denoted by π, called the *prior*, see e.g., (Catoni, 2007; McAllester, 2003; Maurer, 2004; Viallard et al., 2021).

As an additional theoretical contribution, we begin by proving a generic PAC-Bayesian bound that will be suitable for our setting. This bound has a similar form to that of Germain et al. (2009), but holds for subgaussian losses, and not only bounded losses, see Appendix B.

Theorem 2.1. *We assume that ℓ is s^2-subgaussian, in the sense of Assumption 3.1. Then, we have, with probability at least $1-\zeta$ over $S \sim \mu_z^{\otimes n}$, that*

$$\mathbb{E}_{w\sim\rho_S}\left[G_S(w)\right] \leq 2s\sqrt{\frac{\mathrm{KL}\left(\rho_S||\pi\right) + \log(3/\zeta)}{n}},$$

where $\mathrm{KL}(\rho_S||\pi)$ *is the Kullback-Leibler (KL) divergence, whose definition is recalled in Appendix A.1.*

Our main theoretical contributions will be proving upper-bounds on $\mathrm{KL}(\rho_S||\pi)$, when ρ_S is set to the distribution of W_t^S and π is chosen appropriately. Additionally, in Section 4.3, we will prove generalization bounds that are based on related but different generic PAC-Bayesian results, for which we provide a short introduction in Appendix A.1.

To end this section, we define the notion of Φ-entropy, through which we link PAC-Bayesian bounds and the study of fractional FPEs (Gentil & Imbert, 2008; Tristani, 2013).

Definition 2.2 (Φ-entropies)**.** Let μ be a non-negative measure on $\mathbb{R}^d$ and $\Phi : \mathbb{R}_+ \longrightarrow \mathbb{R}$ be a convex function. Then, for a $g : \mathbb{R}^d \longrightarrow \mathbb{R}_+$, such that $g, \Phi(g) \in L^1(\mu)$, we define:

$$\mathrm{Ent}_\mu^\Phi(g) := \int \Phi(g)\, d\mu - \Phi\left(\int g\, d\mu\right).$$

Note in particular that, if $\Phi(x) = x\log(x)$ and g is chosen to be $d\rho_S/d\pi$, the Radon-Nykodym derivative of ρ_S with respect to π, then we have $\mathrm{Ent}_\pi^\Phi(g) = \mathrm{KL}(\rho_S||\pi)$.

3. Main Assumptions

As discussed in Section 2.1, our analysis is based on Eq. (6). To avoid technical complications, we assume that it has a solution, $u_t^S = u^S(t, w)$, that is continuously differentiable in t and twice continuously differentiable in w. We provide a discussion of these properties in Appendix A.2. We also denote by ρ_t^S the corresponding probability distribution on $\mathbb{R}^d$, so that ρ_t^S is the law of W_t^S.

We first make two classical assumptions. The first is the subgaussian behavior of the objective ℓ. Besides, we make a smoothness assumption on function f ensuring the existence of strong solutions to Eq. (3) (Schilling & Schnurr, 2010). Those assumptions are made throughout the paper.

Assumption 3.1. The loss $\ell(w, z)$ is s^2-subgaussian, i.e. for all w and all $\lambda \in \mathbb{R}$, $\mathbb{E}_z\left[e^{\lambda(\ell(w,z)-\mathbb{E}_{z'}[\ell(w,z')])}\right] \leq e^{\frac{\lambda^2 s^2}{2}}$. Moreover, $\ell(w, z)$ is integrable with respect to $\rho_t^S \otimes \mu_z^{\otimes n}$.

Assumption 3.2. $f(w, z)$ is M-smooth, which means:

$$\|\nabla_w f(w, z) - \nabla_w f(w', z)\| \leq M\|w - w'\|.$$

As our proof technique is based on the use of PAC-Bayesian bounds, where we use ρ_T^S as a posterior distribution, we are required to find a pertinent choice for the prior distribution π. We define it by considering the FPE of the Lévy driven Ornstein-Uhlenbeck process associated with the regularization term, $\frac{\eta}{2}\|w\|^2$. More precisely, we consider $\bar{u}_\infty$ a solution to the following steady FPE.

$$0 = -\sigma_1^\alpha(-\Delta)^{\frac{\alpha}{2}}\bar{u}_\infty + \sigma_2^2\Delta\bar{u}_\infty + \eta\nabla\cdot(\bar{u}_\infty w). \quad (7)$$

It has been shown in (Tristani, 2013; Gentil & Imbert, 2008) that such a steady state is well-defined and regular enough. We hence denote by π the probability distribution, on $\mathbb{R}^d$, with density $\bar{u}_\infty$. We characterize further properties of the prior π in Lemma C.16.

Throughout the paper, we will use the following notation:

$$v_t^S(x) := \frac{u_t^S(x)}{\bar{u}_\infty(x)} = \frac{d\rho_t^S}{d\pi}(x). \quad (8)$$

We will often omit the dependency of v_t^S on t and/or S, hence denoting v_t, or just v.

Our theory will require a technical regularity condition on v_t^S, which we will now formalize. To achieve this goal, let $\Phi : \mathbb{R}_+ \to \mathbb{R}$ be a twice differentiable convex function. Specific choices for Φ will be made in Section 4.

Assumption 3.3. For all t and S, the functions v_t^S are positive, continuously differentiable and $\mathrm{Ent}_{\bar{u}_\infty}^\Phi\left(v_t^S\right) < \infty$. Moreover, we define:

$$a(\theta, y; s, u) := \theta\cdot\nabla v(y+s\theta)\Phi''(v(y))\theta\cdot\nabla v(y)\bar{u}_\infty(y+u\theta),$$

and we assume:

1. For each bounded interval I, there exists a non-negative integrable function χ_I s.t. $\forall t \in I$, $|\partial_t\Phi(v_t)\bar{u}_\infty| \leq \chi_I$.
2. Let us fixed t and denote $v = v_t$. For any bounded open set V of $\mathbb{R}_+^2$, the functions, defined for $(s, u) \in V$,

$$(y, \theta) \mapsto a(\theta, y; s, u),\ (y, \theta) \mapsto \frac{\partial a}{\partial s},\ (y, \theta) \mapsto \frac{\partial a}{\partial u},$$

with $(\theta, y) \in \mathbb{S}^{d-1} \times \mathbb{R}^d$, are uniformly dominated by an integrable function on $\mathbb{S}^{d-1} \times \mathbb{R}^d$.
3. Finally, the function $\mathbb{R}^d \longrightarrow \mathbb{R}$ given by:

$$\left(v|\Phi'\circ v| + |\Phi\circ v|\right)\bar{u}_\infty\left(\|\nabla V\| + \|\nabla V_S\|\right), \quad (9)$$

vanishes at infinity (along each coordinate of $x \in \mathbb{R}^d$).

We will say that the functions v_t^S are Φ-*regular*. The first condition is essentially allowing to properly differentiate $\mathrm{Ent}_{\bar{u}_\infty}^\Phi\left(v_t^S\right)$, which, following Gentil & Imbert (2008), is key to the proposed methods. The second requires local integrability of functionals naturally appearing in the computation of $\frac{d}{dt}\mathrm{Ent}_{\bar{u}_\infty}^\Phi\left(v_t^S\right)$. The third condition makes valid the integration by parts performed in the proof of Lemma 4.1, presented in Appendix C.1. Note that we do not require any uniformity in S in Assumption 3.3.

Let us informally justify the third condition when $\Phi(x)$ is either $x\log(x)$ or x^2 (which is our case). It is known that the tail behavior of $\bar{u}_\infty$ is in $1/\|x\|^{d+\alpha}$ (Tristani, 2013). Moreover, $\|\nabla V(x)\|$ and $\|\nabla V_S(x)\|$ are of order at most $1+\|x\|$ based on Assumption 3.2. The condition boils down

to $\Phi(v_t^S(x)) + v_t^S(x)\Phi'(v_t^S(x)) = \mathcal{O}\left(\|x\|^{d+\alpha-1}\right)$, which is reasonable given the definitions of u_t^S and v_t^S.

Finally, we assume that an integral appearing repeatedly in our statements and proofs is finite.

Assumption 3.4. For almost all S and all $t \geq 0$, we have:

$$\int \bar{u}_\infty \Phi''(v) v^2 \left\|\nabla \widehat{F}_S\right\|^2 dx < +\infty.$$

Let us consider $\Phi(x) = x\log(x)$. In this case, Assumption 3.4 directly holds when the surrogate f is Lipschitz in w. More generally, when $t \to \infty$, following arguments in (Tristani, 2013), it is reasonable to consider that the behavior of u_t^S near $x \to \infty$ is $\mathcal{O}\left(\|x\|^{-d-\alpha}\right)$. Therefore, the previous assumption can be informally thought as $\|\nabla \widehat{F}_S(x)\| \lesssim \|x\|^{a/2}$ with $a < \alpha$ (e.g., if ∇f is Hölder continuous), weaker conditions may also be acceptable.

4. Generalization Bounds via Multifractal Fokker-Planck Equations

In this section, we present our main theoretical contributions. The main tool is Lemma 4.1, which offers a decomposition of $\frac{d}{dt}\text{Ent}_{\bar{u}_\infty}^\Phi\left(v_t^S\right)$, that will be used throughout the proofs.

After presenting Lemma 4.1, we will start by dealing with the easier case, which is when $\sigma_2 > 0$ in (3). Then, in Section 4.2, we show how an additional assumption can be leveraged to handle the case where $\sigma_2 = 0$, which presents the most interest for us. Finally, we will extend our analysis to obtain time-uniform bounds, in Section 4.3. For notational purposes, we define, with a slight abuse of notation:

$$G_S(T) := \mathbb{E}_U\left[G_S(W_T^S)\right] = \mathbb{E}_{w\sim\rho_T^S}\left[G_S(w)\right]. \quad (10)$$

4.1. Warm-up: Noise with non-trivial Brownian part

Thanks to our PAC-Bayesian approach, our task boils down to bounding the KL divergence between the posterior ρ_T^S and the prior π. It is given by $\text{KL}\left(\rho_T^S||\pi\right) = \text{Ent}_{\bar{u}_\infty}^\Phi\left(v_T^S\right)$, where, in Sections 4.1 and 4.2, we fix the convex function Φ to be $\Phi(x) = \Phi_{\log}(x) := x\log(x)$, Assumptions 3.3 and 3.4 should be considered accordingly.

We bound the term $\text{KL}\left(\rho_T^S||\pi\right)$ by first computing the *entropy flow*, i.e. the time derivative of $\text{Ent}_{\bar{u}_\infty}^\Phi\left(v_t^S\right)$. While such an approach has already been applied in the case of pure Brownian noise ($\sigma_1 = 0$) (Mou et al., 2017), it is significantly more technical in our case, because of the presence of the fractional Laplacian in Eq. (6). The following lemma is an expression of the entropy flow for a general Φ, that we obtain by adapting the technique presented in Gentil & Imbert (2008) (in the study of the convergence to equilibrium of FPEs) to our setting.

Lemma 4.1 (Decomposition of the entropy flow)**.** *Given a convex and differentiable function $\Phi : (0,\infty) \longrightarrow \mathbb{R}$, we make Assumptions 3.3 and 3.4, relatively to this function Φ.*

$$\frac{d}{dt}\text{Ent}_{\bar{u}_\infty}^\Phi\left(v_t^S\right) = -\sigma_2^2 I_\Phi(v_t^S) - \sigma_1^\alpha B_\Phi^\alpha(v_t^S) - \int \Phi''(v_t^S) v_t^S \nabla v_t^S \cdot \nabla \widehat{F}_S \bar{u}_\infty dx,$$

where $I_\Phi(v) := \int \Phi''(v)\left\|\nabla v\right\|^2 \bar{u}_\infty dx$ is called the Φ-information, and $B_\Phi^\alpha(v)$ is called the Bregman integral, which will be formally defined in Appendix C.1.

For $\Phi(x) = x\log(x)$, the term I_Φ reduces to the celebrated *Fisher information* between ρ_T^S and π, denoted $J(\rho_T^S|\pi)$. It is commonly used in the analysis of FPEs (Chafai & Lehec, 2017, Section 1) and is defined as:

$$J(\rho_T^S|\pi) := \int \left\|\nabla \log \frac{d\rho_T^S}{d\pi}\right\|^2 d\rho_T^S.$$

Lemma 4.1 is a central tool for the derivation of our main results. As a preliminary result, we first present the simpler case where $\sigma_2 > 0$. This leads to the next corollary.

Corollary 4.2. *We make Assumption 3.3 and 3.4. With probability at least $1-\zeta$ over $\mu_z^{\otimes n}$, we have:*

$$G_S(T) \leq s\sqrt{\frac{1}{n\sigma_2^2} I(T,S) + 4\frac{\log(3/\zeta)+\Lambda}{n}},$$

where $\Lambda := \text{KL}\left(\rho_0||\pi\right)$ and I is defined by:

$$I(T,S) := \int_0^T \mathbb{E}_U\left\|\nabla \widehat{F}_S(W_t^S)\right\|^2 dt. \quad (11)$$

When f is L-Lipschitz, we have in addition $I(T,S) \leq TL^2$, recovering known bounds in the case $\sigma_1 = 0$ (Mou et al., 2017). Corollary 4.2 may seem to have no dependence on the tail-index α, however, it is implicitly playing a role through the integral term involving $\nabla \widehat{F}_S$, as W_t^S is generated by a heavy-tailed SDE. Nevertheless, this bound does not apply when $\sigma_2 = 0$, which we will now investigate.

4.2. Purely heavy-tailed case

Now we assume $\sigma_2 = 0$, which makes our task much more challenging. Indeed, in the proof of Corollary 4.2, the Φ-information, I_Φ, is used to compensate for the contribution of the third term in Lemma 4.1. As we cannot do this anymore since I_Φ does not appear with the choice of $\sigma_2 = 0$, we need to develop a finer understanding of the Bregman integral term, i.e. $B_\Phi^\alpha(v)$, which is the contribution of the stable noise L_t^α to the entropy flow.

Towards this goal, in Appendix C.3, we prove that, under Assumption 3.3, there exists a function:

$$J_{\Phi,v} : [0,+\infty) \longrightarrow [0,+\infty),$$

such that $J_{\Phi,v}$ is non-negative, continuous, satisfies $J_{\Phi,v}(0) = I_\Phi(v)$ and we have the integral representation:

$$B_\Phi^\alpha(v) = C_{\alpha,d}\frac{\sigma_{d-1}}{2d}\int_0^\infty J_{\Phi,v}(r)\frac{dr}{r^{\alpha-1}}, \qquad (12)$$

where the constant $C_{\alpha,d}$ is defined in Eq. (32), and σ_{d-1} is the area of the unit sphere, given by Eq. (30). The identification of the function J_{Φ,v_t^S} turns out to be crucial, as it illustrates that the Bregman integral term can be used for approximating a Φ-information term, and therefore re-use ideas from Corollary 4.2. Thus, $J_{\Phi,v_t^S}(r)$ can be seen as an approximation of $I_\Phi(v_t^S)$, i.e., in the case $\Phi = \Phi_{\log}$, of the Fisher information $J(\rho_T^S|\pi)$, at least for small values of r.

A takeaway of our analysis is that, for the approximation $J_{\Phi,v}(r) \approx I_\Phi(v)$ to be accurate, we need to introduce an additional condition regarding the behavior of the function $J_{\Phi,v}$ near the origin. This assumption is specified as follows:

Assumption 4.3. There exists an absolute constant $R > 0$ such that, for all $t > 0$ and $\mu_z^{\otimes n}$-almost all $S \in \mathcal{Z}^n$:

$$\forall r \in [0, R],\ J_{\Phi,v_t^S}(r) \geq \frac{1}{2}J_{\Phi,v_t^S}(0).$$

Note that, by continuity, this condition trivially holds *pointwise* for fixed $S \in \mathcal{Z}^n$ and $t > 0$; however, we essentially require it to hold uniformly in both time t and data S. On the other hand, if the dynamics (3) is initialized at its stationary distribution (like an ideal 'warm start' (Dalalyan, 2016)), then v_t^S is independent of t, in which case the statement of Assumption 4.3 can be obtained, in high probability over S, through Egoroff's theorem (Bogachev, 2007, Thm. 2.2.1).

The factor R plays an important role in our analysis, it is needed that it is positive and preferably not too small. However, we are not able to formally estimate this quantity. The exact formula for $J_{\Phi_{\log},v}$, Definition C.4, shows that, if v is a constant function, then $J_{\Phi_{\log},v}$ is a constant function, hence $R = +\infty$ (this corresponds to the trivial case where $\widehat{F}_S = 0$ and the dynamics is initialized at $\bar{u}_\infty$). Therefore, we argue that R can be large to get non-vacuous bounds when the function v is uniformly bounded away from 0 and has bounded first and second-order derivatives. A more formal version of this argument is provided in Appendix C.3.

This allows us to prove the following theorem, which is a high probability generalization bound in the case $\sigma_2 = 0$.

Theorem 4.4. *We make Assumptions 3.3, 3.4 and 4.3. Then, with probability at least $1-\zeta$ over $\mu_z^{\otimes n}$, we have*

$$G_S(T) \leq 2s\sqrt{\frac{K_{\alpha,d}}{n\sigma_1^\alpha}I(T,S) + \frac{\log(3/\zeta)+\Lambda}{n}}$$

with Λ and $I(T,S)$ as in Corollary 4.2, and:

$$K_{\alpha,d} = \frac{(2-\alpha)\Gamma\left(1-\frac{\alpha}{2}\right)d\Gamma\left(\frac{d}{2}\right)}{\alpha 2^\alpha \Gamma\left(\frac{d+\alpha}{2}\right)R^{2-\alpha}}, \qquad (13)$$

where Γ denotes the Euler's Gamma function, on which more information is provided in Appendix A.4.

Note that, in both Theorems 4.2 and 4.4, if we set the initial distribution $\rho_0 = \pi$ that has the heavy-tailed density $\bar{u}_\infty$, we get $\Lambda = 0$ and, therefore, the bound becomes tighter. This might be an argument in favor of heavy-tailed initialization, which has been considered by several studies (Favaro et al., 2020; Jung et al., 2021), and has been argued to be beneficial (Gurbuzbalaban & Hu, 2021). We will further highlight the quantitative properties of Theorem 4.4 in Section 5.

The proof of Theorem 4.4 would also apply when $\sigma_2 > 0$, however, compared to Section 4.1, it requires the additional Assumption 4.3. The bound of Corollary 4.2 was obtained by using mainly the contribution of B_t to the noise, while Theorem 4.4 corresponds to the contribution of L_t^α. It turns out that both approaches can be combined, under Assumption 4.3; it is presented in Appendix C.7.

Theorem 4.4 (valid for $1 < \alpha < 2$) should be compared with existing generalization bounds for continuous Langevin dynamics (CLD), i.e., $\alpha = 2$, where integral terms that are similar to $I(T,S)$ appear (Mou et al., 2017; Li et al., 2020; Futami & Fujisawa, 2023; Dupuis et al., 2024). Our bound features the new constant $K_{\alpha,d}/\sigma_1^\alpha$ and we show in Section 5 that we recover similar constants to the CLD case in the limit $\alpha \to 2^-$. Compared to (Mou et al., 2017) (in the case of CLD), $I(T,S)$ does not contain any exponential time decay. This point is discussed in detail in Section 4.3 and Appendix E. Despite this fact, $I(T,S)$ can still be small because the norm of the gradients may become small.

4.3. Towards time-uniform bounds

Corollary 4.2 and Theorem 4.4, while being the first high probability bounds for heavy-tailed dynamics with explicit constants, may suffer from a time-dependence issue. The reasons why this is an outcome of our proofs are discussed in Appendix E. It appears from this discussion that Theorem 4.4 can be made time-uniform, under the existence of a specific class of functional inequalities. Unfortunately, we argue that such techniques do not always apply in our case, as it is detailed in Appendix E.

Nevertheless, in this section, we take a first step towards improving the time-dependence of the bounds, derived in our setting. However, this comes at the cost of weakening the interpretability of the bound and might make it hard to compute in practice. In order to present this result, we need to make another choice for the convex function Φ: we consider $\Phi(x) = \Phi_2(x) := \frac{1}{2}x^2$, instead of $\Phi_{\log}$. This

choice is justified by the fact that it significantly changes the structure of the Bregman integral term, i.e. $B_\Phi^\alpha(v_t^S)$, in a way that is clearly presented in the proofs of Appendix C.8.

We only discuss the case $\sigma_2 = 0$, the case $\sigma_2 > 0$ can be found in Appendix C.8. Following the reasoning of Section 4.2, we make Assumption 4.3 with the convex function Φ_2 instead of $\Phi_{\log}$. We will refer to it as Assumption 4.3$-\Phi_2$. This leads to our last theoretical result.

Theorem 4.5. *Let $\sigma_2 = 0$. We make Assumptions 4.3$-\Phi_2$, 3.3 and 3.4, with the choice $\Phi = \Phi_2$. Then, with probability at least $1-\zeta$ over $S \sim \mu_z^{\otimes n}$ and $w \sim \rho_T^S$, we have*

$$G_S(w) \leq 2s\sqrt{\frac{4K_{\alpha,d}}{n\sigma_1^\alpha}\tilde{I}(T,S) + \frac{2e^{-\frac{\alpha\eta T}{2}}\Lambda + \log\frac{24}{\zeta^3}}{n}},$$

with $\Lambda = \mathrm{Ent}^{\Phi}_{\bar{u}_\infty}(\rho_0)$, and

$$\tilde{I}(T,S) := \int_0^T e^{-\frac{\alpha\eta}{2}(T-t)}\mathbb{E}_\pi\left[(v_t^S)^2\left\|\nabla\widehat{F}_S\right\|^2\right]dt.$$

While the exponential decay term, e.g. $e^{-\eta(T-t)}$ is a significant improvement over Theorem 4.4, the integral term $\tilde{I}(T,S)$ is less interpretable than the term $I(T,S)$, appearing in Corollary 4.2 and Theorem 4.4. Indeed, $I(T,S)$ is simply related to the expected gradient of the empirical risk.

5. Quantitative Analysis

We focus our qualitative and experimental analysis on the results obtained in the case of pure heavy-tailed dynamics ($\sigma_2 = 0$), namely Theorems 4.4 and 4.5, as they bring the most novelty compared to the literature. From now on, we assume that the constant R, coming from Assumption 4.3, can be taken independent of α, σ and d. This assumption has important consequences for our quantitative analysis.

Asymptotic analysis. We analyze the behavior of the constant $K_{\alpha,d}$, appearing in Theorems 4.4 and 4.5. Let $\bar{K}_{\alpha,d} = R^{2-\alpha}K_{\alpha,d}$, the following lemma provides an asymptotic formula of this constant, when the number of parameters d goes to infinity. This is pertinent as modern machine learning models typically have a lot of parameters.

Lemma 5.1. *We have that, for all $\alpha \in (1,2)$:*

$$\bar{K}_{\alpha,d} \underset{d\to\infty}{\sim} P_\alpha d^{1-\frac{\alpha}{2}}, \quad P_\alpha := \frac{(2-\alpha)\Gamma\left(1-\frac{\alpha}{2}\right)}{\alpha 2^{\alpha/2}}, \quad (14)$$

In Eq. (14), we isolated a term P_α depending only on α and a dimension dependent term, $d^{1-\frac{\alpha}{2}}$. A quick analysis (see Appendix D) shows that the pre-factor $\alpha \longmapsto P_\alpha$ is decreasing in $(1,2)$ and satisfies $\frac{1}{2} \leq P_\alpha \leq \sqrt{\frac{\pi}{2}}$.

Despite being proven for $\alpha < 2$, our bounds do not explode when $\alpha \to 2^-$, as we show in the following lemma.

Lemma 5.2. *For any $d \geq 1$, we have $K_{\alpha,d} \underset{\alpha\to 2^-}{\longrightarrow} \frac{1}{2}$.*

Phase transition. By Lemma 5.1, in the limit $d \gg 1$, Theorem 4.4 becomes:

$$G_S(T) \leq 2s\sqrt{\frac{P_\alpha d^{1-\frac{\alpha}{2}}}{n\sigma_1^\alpha R^{2-\alpha}}I(T,S) + \frac{\Lambda + \log\frac{3}{\zeta}}{n}}. \quad (15)$$

We can rewrite the constant term, multiplying $I(T,S)$, as:

$$\frac{P_\alpha d^{1-\frac{\alpha}{2}}}{n\sigma_1^\alpha R^{2-\alpha}} = \frac{P_\alpha d_0^{1-\frac{\alpha}{2}}}{n\sigma_1^\alpha} = \frac{P_\alpha d_0}{n(\sigma_1\sqrt{d_0})^\alpha},$$

where we introduced a 'reduced dimension' $d_0 := d/(R^2)$. As mentioned earlier, we have, for all $\alpha \in (1,2)$, that $\frac{1}{2} \leq P_\alpha \leq \sqrt{\frac{\pi}{2}}$. Therefore, it is clear that the main influence of the tail-index α on the generalization bounds is induced by the geometric term $(\sigma_1\sqrt{d_0})^{-\alpha}$. Based on this observation, our bounds suggest a phase transition between two regimes:

- Heavy regime: $(\sigma_1\sqrt{d_0}) < 1$, the generalization error increases with the tail, i.e. the performance should be better with heavier-tails. If we take into account the contribution of the factor P_α, this condition becomes $(\sigma_1\sqrt{d_0}) < 1/\sqrt{2\pi}$, see Appendix D.1.
- Light regime: $(\sigma_1\sqrt{d_0}) > 1$, heavy-tails are harmful for the generalization bound.

This shows that, depending on the setting and the structure of the dynamics, heavy tails may have a different impact on the generalization error.

Comparison with existing works. In (Raj et al., 2023b), the authors studied Eq. (3), with $\sigma_1 = 1$, $\sigma_2 = 0$, and f Lipschitz continuous.[3] Informally, the obtained bound is:

$$\mathbb{E}_{S,U}\left[L(W_\infty^S) - \widehat{L}_S(W_\infty^S)\right] \leq \frac{\|\ell\|_{\mathrm{Lip}}AR_{\alpha,d}}{n}, \quad (16)$$

where A is a quantity that has a complex dependence on various constants appearing in the assumptions, $\|\ell\|_{\mathrm{Lip}}$ is the Lipschitz constant of ℓ, which is assumed finite, and $R_{\alpha,d}$ is a constant, explicitly given in Appendix D.2, where we also show that it satisfies $R_{\alpha,d} = \mathcal{O}_{d\to\infty}(d^{\frac{1+\alpha}{2}})$.

We already mentioned, in Section 1, some differences between Eq. (16) and our results. We additionally emphasize that *(i)* we do not require a Lipschitz assumption, and *(ii)* The constant $R_{\alpha,d}$ has a worse dependence on the dimension d than the constant $K_{\alpha,d}$, appearing in our theorems. Eq. (16) cannot explain generalization in an overparameterized regime, i.e. when $d > n$. Moreover, in the limit $\alpha \to 2^-$, it does not yield the known dimension dependence

[3] Raj et al. (2023b) consider a Lipschitz loss ℓ and a surrogate f, that has a dissipativity property; we can frame it within our setting by assuming that f is Lipschitz in w.

for Langevin dynamics (Mou et al., 2017; Pensia et al., 2018; Farghly & Rebeschini, 2021), while Lemma 5.2 shows that $K_{\alpha,d}$ becomes independent of d when $\alpha \to 2^-$.

To have a fair comparison, we shall note that (16) does not increase with the time horizon T, whereas it is the main drawback of our bounds. Nevertheless, the results of Section 4.3 and Appendix E show that this point might have room for improvement, which we leave as future work.

6. Empirical Analysis

Setup. We numerically approximate Eq. (3), using its Euler-Maruyama discretization (Duan, 2015), $\forall k \in \{1, \dots, N\}$,

$$\widehat{W}^S_{k+1} = \widehat{W}^S_k - \gamma \nabla \widehat{F}_S(\widehat{W}^S_k) - \eta\gamma \widehat{W}^S_k + \gamma^{\frac{1}{\alpha}} \sigma_1 L^\alpha_1, \quad (17)$$

where $\gamma > 0$ and $N \in \mathbb{N}$ are fixed learning rate and number of iterations. Our main experiments were conducted with 2 layers fully-connected networks (FCN2) trained on the MNIST dataset (Lecun et al., 1998). Additional experiments, using MNIST, FasionMNIST (Xiao et al., 2017) and CIFAR10 datasets (Krizhevsky et al., 2014), as well as linear models and deeper networks, are presented in Appendix F.4. We choose the objective ℓ as the 0-1 loss and the surrogate f (that we used for training) as the cross entropy loss. These choices make our experiments as close as possible to our theoretical setting, still allowing us to have a varying number of parameters d. Each experiment is run with 10 different random seeds. All hyperparameters details can be found in Appendix F.1.

We provide, in Appendix C.6, an additional analysis justifying that our continuous-time theory is still pertinent to study the discrete one, Eq. (47), thus providing sufficient theoretical foundations for our experiments.

The estimation of the accuracy is subject to important noise, due to the jumps incurred by L^α_t. To act against this noise, we first use $\alpha \in [1.6, 2]$. This range is also coherent with estimated tail indices in practical settings by Raj et al. (2023a); Barsbey et al. (2021). Moreover, the accuracy gap is (robustly) averaged over the last iterations, see Appendix F.2.

As shown in Eq. (17), we use the full dataset S at each iteration, in accordance with the model that we study in this paper. Moreover, it has been argued in several studies (Gurbuzbalaban et al., 2021; Hodgkinson & Mahoney, 2020; Barsbey et al., 2021) that SGD may create heavy-tailed behavior, an effect whose interaction with the noise L^α_t would be unclear. Our setting allows us to isolate the effect of L^α_t on the generalization error. To make our experiments tractable, we sub-sample 10% of the MNIST and Fashion-MNIST datasets to run our main experiments. To show that our theory may stay pertinent in more practical settings, we estimated our bound when training a FCN5 on the whole MNIST dataset, with smaller batches, see Appendix F.4.5.

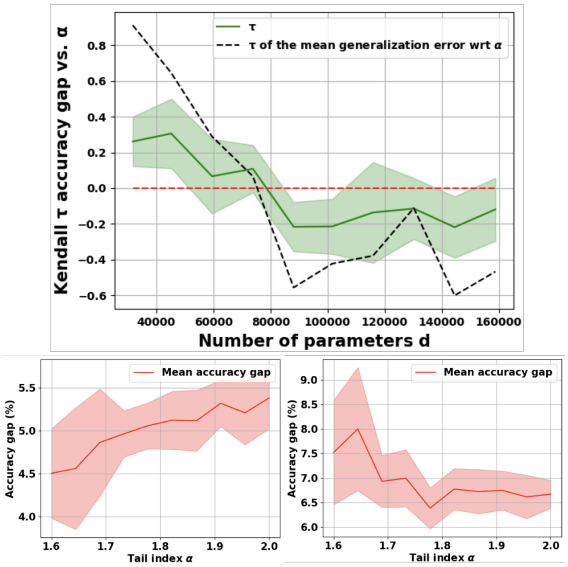


Figure 2. (up) Correlation (Kendall's τ) between α and the accuracy gap, for different values of d, with a FCN2 trained on MNIST. The green curve is the average τ over 10 random seeds. The black curve is the correlation between α and the average accuracy gap over 10 seeds. *(bottom)* Accuracy gap with respect to α for $d = 3 \cdot 10^4$ *(left)* and $d = 15 \cdot 10^4$ *(right)*.

Lévy processes simulation. As shown by Eq. (17), the numerical approximation of $\widehat{W}^S_k$ requires the simulation of the Lévy process L^α_t, which is estimated by independent realization of L^α_1. Simulating α-stable Lévy processes is standard in probabilistic simulation. In our case, we use the method described by Nolan (2013, Section 1). More precisely, we sample L^α_1 as $L^\alpha_1 = \sqrt{A}G$, where $G \sim \mathcal{N}(0, I_d)$ and A is a skewed stable distribution given by:

$$A \sim S\left(\frac{\alpha}{2}, 1, 2\cos\left(\frac{\pi\alpha}{4}\right)^{2/\alpha}, 0\right),$$

where $S(\alpha, \beta, c, \mu)$ denotes stable distributions, with β the skewness parameter and c the scale parameter, see (Duan, 2015) for more details. This model was in particular used to generate the Lévy processes in Fig. 1.

Results. We test our theory through 3 types of experiments. We present in this section their results for a FCN2 trained on MNIST. Appendix F.4 contains additional experiments.

First, on Fig. 2, we compute the correlation between α and the accuracy gap, measured in term of a Kendall's τ coefficient[4](Kendall, 1938). We use a FCN2 and let the width vary to compute τ for different values of the dimension d. The detailed procedure to obtain Fig. 2 can be found in Appendices F.1 and F.3. We observe that the phase transition between positive and negative correlation, predicted in Section 5, happens for a value of the dimension $d \simeq 8 \cdot 10^4$, which we will use to further estimate R. We observe that

[4]The sign of τ corresponds to the sign of the correlation.

the positive correlation of the heavy regime seems to be stronger than the negative correlation of the light regime. As an additional experiment, we also provide the same plot as Fig. 2 in Appendix F.4, but using the Pearson correlation coefficient instead of τ. These results, displayed in Fig. 15, yield the same empirical results than Fig. 2.

The bound of Theorem 4.4 is computable in practice, we estimate it by the formula (in that case $s = 1/2$):

$$\widehat{G} := \sqrt{\frac{P_\alpha d^{1-\frac{\alpha}{2}}\gamma}{n\sigma_1 R^{2-\alpha}} \sum_{k=1}^{N} \left\| \nabla \widehat{F}_S(\widehat{W}_k^S) \right\|^2}. \tag{18}$$

On Fig. 3, we plot Eq. (18) *w.r.t.* to the accuracy gap, for several values of R. We use $R = 1$ as a default choice, as R is unknown *a priori*, it shows, for each value of α, a good correlation with the accuracy gap. Nonetheless, based on Fig. 2, conducted with the same setting as Fig. 3, we can estimate the value of R to be in $\simeq [2.8, 7]$. If we use these values in Eq. (18), the observed correlation is much stronger. If we use a slightly larger value ($R = 15$), we see, in Fig. 3, that the correlation becomes almost perfect, which we interpret as the phase transition being correctly taken into account. This shows that the right corrective term in Eq. (49) is indeed of the form $R^{2-\alpha}$. We note that the reason why our bound over-estimates the accuracy gap, is that it increases with T. However, as the bounds presented in Section 4.3 and Appendix E are time-uniform and have similar constants and dependence on $\nabla \widehat{F}_S$ as in Theorem 4.4, we believe that this issue can be alleviated by extending Theorem 4.4 in a similar direction, which we leave as an open question.

Finally, to obtain Fig. 4, we fixed σ to 0.01 and let d vary in a fixed range. Based on Eq. (15), we expect the accuracy error, denoted G_S, to be proportional to $d^{1/2-\alpha/4}$. This suggests to perform the linear regression, $\log(G_S) \simeq \widehat{r} \log(d) + C$, to compute an estimate $\widehat{\alpha} := 2 - 4\widehat{r}$ of the tail-index α. The

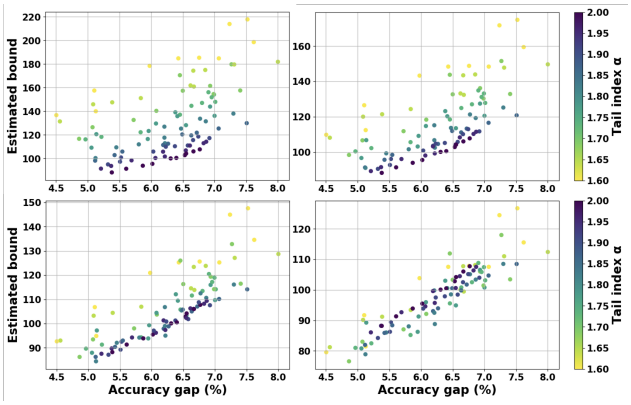


Figure 3. Bound estimated from Eq. (18) versus accuracy gap for a FCN2 on MNIST, for different values of R: 1 *(top left)*, 3 *(top right)*, 7 *(bottom left)*, 15 *(bottom right)*.

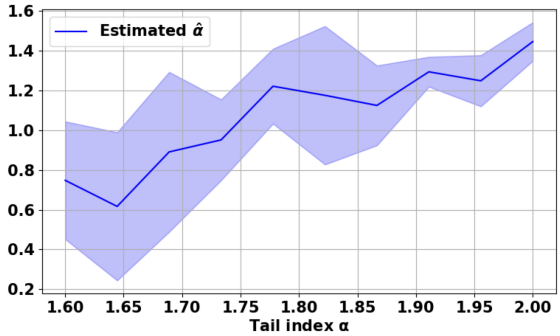


Figure 4. Regression of the tail-index α from the accuracy error, for a FCN2 trained on MNIST.

blue curve in Fig. 4 shows $\widehat{\alpha}$ in terms of α. This shows a strong correlation between the estimated and the ground-truth tail-index, in particular, we retrieve the expected monotonicity. However, $\widehat{\alpha}$ seems to underestimate the true value of α, by a term independent of α. We suspect that this may be because other terms in the bound have a dependence on d, or because our bound is not a strict equality, which we assumed to compute $\widehat{\alpha}$ from G_S.

7. Conclusion

In this paper, we proved generalization bounds for heavy-tailed SDEs. Our results are the first to be both in high-probability and computable. Moreover, they allow for a more flexible setup and have a better dimension-dependence than existing works. We analyzed the constants appearing in our theorems, which led us to predict the existence of a phase transition in terms of the effect of the tail index on the generalization. We supported our theory with various numerical experiments. Several directions remain to be studied in the future. In particular, obtaining new functional inequalities, such as presented in Appendix E, could improve the time-dependence of the bounds. Moreover, understanding the interaction, between small batches and the stable noise L_t^α, would be a natural extension of the theory.

Acknowledgments

We thank Paul Viallard, Maxime Haddouche and Isabelle Tristani for valuable discussions. U.Ş. is partially 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). B.D. and U.Ş. are partially supported by the European Research Council Starting Grant DYNASTY – 101039676.

Impact Statement

This work is largely theoretical, it does not have any direct social or ethical impact.

References

Alquier, P. User-friendly introduction to PAC-Bayes bounds, November 2021.

Andreeva, R., Limbeck, K., Rieck, B., and Sarkar, R. Metric space magnitude and generalisation in neural networks. In *Proceedings of 2nd Annual Workshop on Topology, Algebra, and Geometry in Machine Learning (TAG-ML)*, volume 221 of *Proceedings of Machine Learning Research*, pp. 242–253. PMLR, 2023.

Bakry, D., Gentil, I., and Ledoux, M. *Analysis and Geometry of Markov Diffusion Operators*. Springer, 2014.

Barsbey, M., Sefidgaran, M., Erdogdu, M. A., Richard, G., and Şimşekli, U. Heavy Tails in SGD and Compressibility of Overparametrized Neural Networks. In *35th Conference on Neural Information Processing Systems (NeurIPS 2021)*. arXiv, June 2021. doi: 10.48550/arXiv.2106.03795.

Birdal, T., Lou, A., Guibas, L., and Şimşekli, U. Intrinsic Dimension, Persistent Homology and Generalization in Neural Networks. *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, November 2021.

Bogachev, V. I. *Measure theory. Vol. I, II*. Springer-Verlag, Berlin, 2007. ISBN 978-3-540-34513-8; 3-540-34513-2. doi: 10.1007/978-3-540-34514-5.

Böttcher, B., Schilling, R., and Wang, J. *Lévy Matters III: Lévy-Type Processes: Construction, Approximation and Sample Path Properties*, volume 2099 of *Lecture Notes in Mathematics*. Springer International Publishing, Cham, 2013. ISBN 978-3-319-02683-1 978-3-319-02684-8. doi: 10.1007/978-3-319-02684-8.

Bousquet, O. Stability and generalization. *Journal of Machine Learning Research*, pp. 499–526, 2002.

Bousquet, O., Klochkov, Y., and Zhivotovskiy, N. Sharper bounds for uniformly stable algorithms. *Proceedings of Thirty Third Conference on Learning Theory*, May 2020.

Catoni, O. Pac-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning. *IMS Lecture Notes Monograph Series*, 56:1–163, 2007. ISSN 0749-2170. doi: 10.1214/074921707000000391.

Chafai, D. Entropies, convexity, and functional inequalities. *Kyoto Journal of Mathematics*, 44(2), January 2004. ISSN 2156-2261. doi: 10.1215/kjm/1250283556.

Chafai, D. and Lehec, J. Logarithmic sobolev inequalities essentials, 2017.

Chafaï, D. Binomial-poisson entropic inequalities and the m/m/infinity queue. *ESAIM: Probability and Statistics*, 10:317–339, September 2006. ISSN 1262-3318. doi: 10.1051/ps:2006013. URL http://dx.doi.org/10.1051/ps:2006013.

Dalalyan, A. S. Theoretical guarantees for approximate sampling from smooth and log-concave densities, December 2016.

Daoud, M. and Laamri, E. H. Fractional laplacians : A short survey. *Discrete and Continuous Dynamical Systems - S*, 15(1):95–116, 2022. URL https://www.aimsciences.org/article/id/5a563692-a87e-458f-8e9e-e03ce8ff49b9.

Duan, J. *An Introduction to Stochastic Dynamics*. Cambridge texts in Applied Mathematics, 2015.

Dupuis, B. and Viallard, P. From Mutual Information to Expected Dynamics: New Generalization Bounds for Heavy-Tailed SGD, December 2023.

Dupuis, B., Deligiannidis, G., and Şimşekli, U. Generalization Bounds with Data-dependent Fractal Dimensions. In *Internation Conference on Machine Learning (ICML 2023)*. arXiv, February 2023. doi: 10.48550/arXiv.2302.02766.

Dupuis, B., Viallard, P., Deligiannidis, G., and Simsekli, U. Uniform generalization bounds on data-dependent hypothesis sets via pac-bayesian theory on random sets, 2024.

Farghly, T. and Rebeschini, P. Time-independent Generalization Bounds for SGLD in Non-convex Settings. In *35th Conference on Neural Information Processing Systems (NeurIPS 2021)*. arXiv, November 2021. doi: 10.48550/arXiv.2111.12876.

Favaro, S., Fortini, S., and Peluchetti, S. Stable behaviour of infinitely wide deep neural networks. In *Proceedings of the 23rdInternational Conference on Artifi- Cial Intelligence and Statistics (AISTATS) 2020*. arXiv, February 2020.

Futami, F. and Fujisawa, M. Time-Independent Information-Theoretic Generalization Bounds for SGLD. In *7th Conference on Neural Information Processing Systems (NeurIPS 2023)*. arXiv, November 2023. doi: 10.48550/arXiv.2311.01046.

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

Germain, P., Lacasse, A., Laviolette, F., and Marchand, M. PAC-Bayesian learning of linear classifiers. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pp. 353–360, New York, NY, USA, June 2009. Association for Computing Machinery. ISBN 978-1-60558-516-1. doi: 10.1145/1553374.1553419.

Gross, L. Logarithmic Sobolev inequalities. *Amer. J. Math., 97(4):1061–1083*, 1975.

Gurbuzbalaban, M. and Hu, Y. Fractional moment-preserving initialization schemes for training deep neural networks. In *Proceedings of the 24th International Conference on Artifi- Cial Intelligence and Statistics (AISTATS) 2021*. arXiv, February 2021.

Gurbuzbalaban, M., Şimşekli, U., and Zhu, L. The Heavy-Tail Phenomenon in SGD. In *International Conference on Machine Learning (ICML 2021)*. arXiv, June 2021. doi: 10.48550/arXiv.2006.04740.

Haghifam, M., Negrea, J., Khisti, A., Roy, D. M., and Dziugaite, G. K. Sharpened Generalization Bounds based on Conditional Mutual Information and an Application to Noisy, Iterative Algorithms, October 2020.

Halperin, I. and Schwartz, L. *Introduction to the Theory of Distributions*. University of Toronto Press, 1952. ISBN 978-1-4875-9132-8.

Hodgkinson, L. and Mahoney, M. W. Multiplicative noise and heavy tails in stochastic optimization. In *Proceedings of the 38 Th International Conference on Machine Learning (ICML 2021)*. arXiv, June 2020. doi: 10.48550/arXiv.2006.06293.

Hodgkinson, L., Şimşekli, U., Khanna, R., and Mahoney, M. W. Generalization Bounds using Lower Tail Exponents in Stochastic Optimizers. *Proceedings of the 39th International Conference on Machine Learning*, July 2022.

Imbert, C. A non-local regularization of first order Hamilton-Jacobi equations. *Journal of differential equations*, 2005.

Jung, P., Lee, H., Lee, J., and Yang, H. α-Stable convergence of heavy-tailed infinitely-wide neural networks. *Advances in Applied Probability , Volume 55 , Issue 4*, June 2021.

Kendall, M. G. A new reasure of rank correlation. *Biometrika*, 1938.

Krizhevsky, A., Nair, V., and Hinton, G. E. The cifar-10 dataset, 2014.

Kühn, F. Solutions of Lévy-driven SDEs with unbounded coefficients as Feller processes. *Proceedings of the American Mathematical Society*, 146(8):3591–3604, 2018.

Lafleche, L. Fractional Fokker-Planck Equation with General Confinement Force. *SIAM Journal on Mathematical Analysis*, 52(1):164–196, January 2020. ISSN 0036-1410, 1095-7154. doi: 10.1137/18M1188331.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, November 1998. ISSN 1558-2256. doi: 10.1109/5.726791.

Li, J., Luo, X., and Qiao, M. On Generalization Error Bounds of Noisy Gradient Methods for Non-Convex Learning. In *Published as a Conference Paper at ICLR 2020*. arXiv, February 2020. doi: 10.48550/arXiv.1902.00621.

Lim, S. H., Wan, Y., and Şimşekli, U. Chaotic Regularization and Heavy-Tailed Limits for Deterministic Gradient Descent, May 2022.

Lischke, A., Pang, G., Gulian, M., Song, F., Glusa, C., Zheng, X., Mao, Z., Cai, W., Meerschaert, M. M., Ainsworth, M., and Karniadakis, G. E. What Is the Fractional Laplacian? - A comparative review with new results, November 2019.

Markowich, P. and Villani, C. On the Trend to Equilibrium for the Fokker-Planck Equation: An Interplay between Physics and Functional Analysis., 2004.

Maurer, A. A Note on the PAC Bayesian Theorem, November 2004.

McAllester, D. Some pac-bayesian theorems. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory, COLT 1998, Madison, Wisconsin, USA, July 24-26, 1998*, pp. 230–234. ACM, 1998.

McAllester, D. A. PAC-Bayesian Stochastic Model Selection. *Machine Learning*, 51(1):5–21, April 2003. ISSN 1573-0565. doi: 10.1023/A:1021840411064.

Mou, W., Wang, L., Zhai, X., and Zheng, K. Generalization Bounds of SGLD for Non-convex Learning: Two Theoretical Viewpoints. In *Proceedings of the 31st Conference On Learning Theory*. arXiv, July 2017. doi: 10.48550/arXiv.1707.05947.

Negrea, J., Haghifam, M., Dziugaite, G. K., Khisti, A., and Roy, D. M. Information-Theoretic Generalization Bounds for SGLD via Data-Dependent Estimates, January 2020.

Neu, G., Dziugaite, G. K., Haghifam, M., and Roy, D. M. Information-Theoretic Generalization Bounds for Stochastic Gradient Descent, August 2021.

Nguyen, T. H., Şimşekli, U., Gürbüzbalaban, M., and Richard, G. First Exit Time Analysis of Stochastic Gradient Descent Under Heavy-Tailed Gradient Noise. In *NIPS'19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*. arXiv, June 2019a. doi: 10.48550/arXiv.1906.09069.

Nguyen, T. H., Simsekli, U., and Richard, G. Non-asymptotic analysis of fractional langevin monte carlo for non-convex optimization. In *International Conference on Machine Learning*, pp. 4810–4819. PMLR, 2019b.

Nolan, J. P. Multivariate elliptically contoured stable distributions: theory and estimation. *Computational Statistics*, 28:2067 – 2089, 2013. URL https://api.semanticscholar.org/CorpusID:6341926.

Pavasovic, K. L., Durmus, A., and Simsekli, U. Approximate heavy tails in offline (multi-pass) stochastic gradient descent. In *Advances in Neural Information Processing Systems*, 2023.

Pensia, A., Jog, V., and Loh, P.-L. Generalization Error Bounds for Noisy, Iterative Algorithms. *2018 IEEE International Symposium on Information Theory (ISIT)*, January 2018.

Raginsky, M., Rakhlin, A., and Telgarsky, M. Non-convex learning via Stochastic Gradient Langevin Dynamics: A nonasymptotic analysis, June 2017.

Raj, A., Barsbey, M., Gürbüzbalaban, M., Zhu, L., and Şimşekli, U. Algorithmic Stability of Heavy-Tailed Stochastic Gradient Descent on Least Squares. In *Proceedings of The 34th International Conference on Algorithmic Learning Theory,*. arXiv, February 2023a. doi: 10.48550/arXiv.2206.01274.

Raj, A., Zhu, L., Gürbüzbalaban, M., and Şimşekli, U. Algorithmic Stability of Heavy-Tailed SGD with General Loss Functions. In *International Conference on Machine Learning (ICML 2023)*. arXiv, January 2023b. doi: 10.48550/arXiv.2301.11885.

Schertzer, D., Larchev, M., Duan, J., Yanovsky, V. V., and Lovejoy, S. Fractional Fokker–Planck Equation for Nonlinear Stochastic Differential Equations Driven by Non-Gaussian Levy Stable Noises. *Journal of Mathematical Physics*, 42(1):200–212, January 2001. ISSN 0022-2488, 1089-7658. doi: 10.1063/1.1318734.

Schilling, R. L. Feller Processes Generated by Pseudo-Differential Operators: On the Hausdorff Dimension of Their Sample Paths — SpringerLink. *Journal of Theoretical Probability*, 1998.

Schilling, R. L. An Introduction to Lévy and Feller Processes. Advanced Courses in Mathematics - CRM Barcelona 2014, October 2016.

Schilling, R. L. and Schnurr, A. The Symbol Associated with the Solution of a Stochastic Differential Equation. *Electr. J. Probab. 15*, December 2010.

Shawe-Taylor, J. and Williamson, R. C. A PAC analysis of a bayesian estimator. In Freund, Y. and Schapire, R. E. (eds.), *Proceedings of the Tenth Annual Conference on Computational Learning Theory, COLT 1997, Nashville, Tennessee, USA, July 6-9, 1997*, pp. 2–9. ACM, 1997.

Şimşekli, U. Fractional langevin monte carlo: Exploring lévy driven stochastic differential equations for markov chain monte carlo. In *International Conference on Machine Learning*, pp. 3200–3209. PMLR, 2017.

Simsekli, U., Sagun, L., and Gurbuzbalaban, M. A Tail-Index Analysis of Stochastic Gradient Noise in Deep Neural Networks. In *Proceedings of the 36 Th International Conference on Machine Learning (ICML 2019)*. arXiv, January 2019. doi: 10.48550/arXiv.1901.06053.

Şimşekli, U., Sener, O., Deligiannidis, G., and Erdogdu, M. A. Hausdorff Dimension, Heavy Tails, and Generalization in Neural Networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124014, December 2021. ISSN 1742-5468. doi: 10.1088/1742-5468/ac3ae7.

Teymurazyan, R. The fractional Laplacian: A primer. https://arxiv.org/abs/2310.19118v1, October 2023.

Tristani, I. Fractional Fokker-Planck equation. *Commun. Math. Sci. 13*, December 2013.

Umarov, S., Hahn, M., and Kobayashi, K. *Beyond the Triangle: Brownian Motion, Ito Calculus and Fokker-Planck Equation - Fractional Generalizations*. World scientific publishing, 2018.

van Erven, T. and Harremoës, P. Rényi Divergence and Kullback-Leibler Divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, July 2014. ISSN 0018-9448, 1557-9654. doi: 10.1109/TIT.2014.2320500.

Vershynin, R. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Number 47 in Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2020. ISBN 978-1-108-41519-4.

Viallard, P., Germain, P., Habrard, A., and Morvant, E. A General Framework for the Disintegration of PAC-Bayesian Bounds. *Machine Learning Journal*, October 2021.

Villani, C. *Optimal Transport - Old and New*. Springer, 2009.

Wan, Y., Zaidi, A., and Simsekli, U. Implicit compressibility of overparametrized neural networks trained with heavy-tailed sgd. *arXiv preprint arXiv:2306.08125*, 2023.

Wang, M. and Duan, J. Existence and regularity of a linear nonlocal Fokker–Planck equation with growing drift. *Journal of Mathematical Analysis and Applications*, 449(1):228–243, May 2017. ISSN 0022-247X. doi: 10.1016/j.jmaa.2016.12.013.

Wu, L. A new modified logarithmic Sobolev inequality for Poisson point processes and several applications. *Probability theory and related fileds - s 118, 427–438*, 2000.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-MNIST: A Novel Image Dataset for Benchmarking Machine Learning Algorithms, September 2017.

Xiao, Y. Random fractals and Markov processes. *Fractal Geometry and Applications: A jubilee of Benoît Mandelbrot - American Mathematical Society*, 72.2:261–338, 2004. doi: 10.1090/pspum/072.2/2112126.

Xie, X., Duan, J., Li, X., and Lv, G. A regularity result for the nonlocal Fokker-Planck equation with Ornstein-Uhlenbeck drift, April 2015.

Zhou, P., Feng, J., Ma, C., Xiong, C., Hoi, S. C. H., et al. Towards theoretically understanding why sgd generalizes better than adam in deep learning. *Advances in Neural Information Processing Systems*, 33:21285–21296, 2020.

Organization of the appendix: The appendix starts with a short notations section. The remainder of the document is then organized as follows:

- In Appendix A, some technical background is presented. The technical background is divided into three main topics: PAC-Bayesian bounds, Lévy processes, and Φ-entropies inequalities. We also include a small subsection on the Euler's Γ function.
- Appendix B presents the proof of our version of a PAC-Bayesian generalization bound for subgaussian losses.
- Appendix C is the core of the appendix, we prove the main result, along with all intermediary lemmas, and introduce the notations necessary to understand those proofs. Moreover, a few additional theoretical results are given, which are a refinement of the main results. In particular, the extension of the theory to a discrete setting is discussed in Appendix C.6, while Appendix C.7 presents bounds in the case $\sigma_2 > 0$, which are different than those of Section 4.1.
- In Appendix D, we provide details on how to obtain the theoretical results of Appendix D.
- In Appendix E, we discuss the time dependence of Corollary 4.2 and Theorem 4.4 and mention that this time dependence could be largely improved by assuming that a certain class of inequality holds.
- Finally, Appendix F presents some details on the experimental setting, as well as a few additional experiments.

Notations

In order to simplify the notations, we will sometimes omit the x variable when integrating with respect to the Lebesgue measure in $\mathbb{R}^d$, i.e. we will write invariable $\int f$ or $\int f dx$, instead $\int f(x)dx$. These conventions are meant to ease the notations throughout the paper.

We will use the following convention regarding the Fourier transform, for $\phi : \mathbb{R}^d \longrightarrow \mathbb{R}$, regular enough, we set:

$$\mathcal{F}\phi(\xi) := \int e^{-ix\cdot\xi}\phi(x)dx. \tag{19}$$

The partial derivative $\partial/\partial t$ will often be shortened as ∂_t. Similarly, ∂_i may denote $\partial/\partial x_i$.

Let's also precise some notations introduced in the main part of the document. As mentioned in Section 1, the data space is denoted $\mathcal{Z}$. More precisely, $\mathcal{Z}$ is a measurable space, endowed with a σ-algebra $\mathcal{F}$. The data distribution, μ_z, is a probability measure on $(\mathcal{Z}, \mathcal{F})$.

A. Technical background

A.1. Information-theoretic terms and PAC-Bayesian bounds

The concept of PAC-Bayesian analysis has been introduced in Section 2.2. In this section, we detail two particular PAC-Bayesian bounds that we use for the derivation of our main results. For a more detailed introduction to those subjects, the reader is invited to consult (Alquier, 2021).

We start by defining the information theoretic (IT) quantities appearing in the aforementioned theorems, see (van Erven & Harremoës, 2014) for more details. Let μ and ν be two probability measures, on the same space, such that μ is absolutely continuous with respect to ν. We define the Kullback-Leibler (KL) divergence as:

$$\mathrm{KL}\left(\mu||\nu\right) := \int \log\left(\frac{d\mu}{d\nu}\right) d\mu, \tag{20}$$

where $d\mu/d\nu$ denotes the Radon-Nykodym derivative between μ and ν.

Next, we define the Renyi divergences, for some $\beta > 1$, as [5]:

$$\mathrm{D}_\beta(\mu||\nu) := \frac{1}{\beta-1}\log\left(\int\left(\frac{d\mu}{d\nu}\right)^\beta d\nu\right). \tag{21}$$

By convention, we set $\mathrm{D}_1(\cdot||\cdot) := \mathrm{KL}(\cdot||\cdot)$, so that, by (van Erven & Harremoës, 2014, Theorem 3), $\mathrm{D}_\beta(\cdot||\cdot)$ is nondecreasing in β.

In the following, to mimic the notations of the rest of the paper, we consider a probability measure π on $\mathbb{R}^d$ and a family of data-dependent probability measures on $\mathbb{R}^d$, $(\rho_S)_{S\in\mathcal{Z}^n}$, where $\mathcal{Z}$ has been defined in Section 1. We mainly require this family to satisfy the following properties:

1. Absolute continuity: for (almost-)all S, we have $\rho_S \ll \pi$.
2. Markov kernel property, for all Borel set $B \subset \mathbb{R}^d$, the map $S \longmapsto \rho_S(B)$ is $\mathcal{F}^{\otimes n}$-measurable. Recall that $\mathcal{F}$ is the σ-algebra on the data space $\mathcal{Z}$.

In the following, as we do in the rest of the paper, we refer to π as a prior distribution, and to ρ_S as posterior distributions.

The next theorem is a generic PAC-bayesian bound due to Germain et al. (2009).

Theorem A.1 (General PAC-Bayesian bound)**.** *Let $\zeta \in (0,1)$ and $\varphi : \mathbb{R}^d \times \mathcal{Z}^n \to \mathbb{R}$ a measurable function, integrable with respect to the posterior distributions. With probability at least $1-\zeta$, over $S \sim \mu_z^{\otimes n}$, we have:*

$$\mathbb{E}_{w\sim\rho_S}[\varphi(w,S)] \le \log(1/\zeta) + \mathrm{KL}(\rho_S||\pi) + \log \mathbb{E}_S\mathbb{E}_{w\sim\pi}\left[e^{\varphi(w,S)}\right],$$

where the KL divergence has been defined by Eq. (20).

Theorem A.2 (Disintegrated PAC-Bayesian bound)**.** *Let $\zeta \in (0,1)$ and $\varphi : \mathbb{R}^d \times \mathcal{Z}^n \to \mathbb{R}$ a measurable function. With probability at least $1-\zeta$, over $S \sim \mu_z^{\otimes n}$ and $w \sim \rho_S$, we have:*

$$\frac{\beta}{\beta-1}\varphi(w,S) \le \frac{2\beta-1}{\beta-1}\log(2/\zeta) + \mathrm{D}_\beta(\rho_S||\pi) + \log \mathbb{E}_S\mathbb{E}_{w\sim\pi}\left[e^{\frac{\beta}{\beta-1}\varphi(w,S)}\right],$$

where the Renyi divergence has been defined by Eq. (21).

We give below one particular instance of Theorem A.1, using the notations introduced in Section 1, for μ_z, S, f, L and $\widehat{L}_S$. This theorem was first proven by (McAllester, 2003; Maurer, 2004).

Theorem A.3. *Assume that the objective f is bounded in $[0,1]$, then, with probability at least $1-\zeta$ over $S \sim \mu_z^{\otimes n}$, we have:*

$$\mathbb{E}_{\rho_S}\left[L(w) - \widehat{L}_S(w)\right] \le \sqrt{\frac{\mathrm{KL}(\rho_S||\pi) + \log\frac{2\sqrt{n}}{\zeta}}{2n}}.$$

A.2. Background on Levy processes and associated pseudo-differential operators

In this section, we introduce some basic notions related to the study of Lévy process. In particular, we insist on the case of stable Lévy processes and their associated operators, namely the Laplacian and fractional Laplacian. Therefore, we make the link between the SDE (3) and the PDE (6) as clear as possible for the reader. We will also set up several notations used throughout the sequel.

A.2.1. Levy processes

In this subsection, we recall some basic notions related to Lévy process, in order to make our main results as clear as possible. The main goal is to get an understanding of Eq. (6). For a more detailed introduction to those subjects, we refer the reader to (Schilling, 1998; Xiao, 2004; Böttcher et al., 2013).

[5] There are definitions of Renyi divergences for other values of β, but we won't need them in this paper, see (van Erven & Harremoës, 2014).

Definition A.4 (Lévy process)**.** A Lévy process $(L_t)_{t\geq 0}$, in $\mathbb{R}^d$, is a stochastic process such that:

- $L_0 = 0$,
- the increments are independents, i.e., for all $t_1 < \cdots < t_K$, the processes $L_{t_i} - L_{t_{i-1}}$ are independent,
- the increments are stationary, i.e. for $0 \leq s < t$, we have $L_t - L_s \overset{\mathbf{d}}{=} L_s$, where $\overset{\mathbf{d}}{=}$ denotes the equality in distribution,
- the process is stochastically continuous, by which we mean that for any $s \geq 0$ and $\delta > 0$, we have:

$$\lim_{t\to s} \mathbb{P}\left(\|L_t - L_s\| > \delta\right) = 0.$$

Equivalently, one can show that stochastic continuity is equivalent to the process having a modification with cadlag paths[6], therefore, the paths of Lévy processes may exhibit jumps.

Levy processes are closely related to the notion of *infinitely divisible distributions*. A probability distribution is said to be infinitely divisible if, for any $N \in \mathbb{N}^\star$, it can be seen that F is the distribution of the sum of N random variables.

Lévy processes have infinitely divisible distributions[7]. Following Schilling & Schnurr (2010, Corollary 2.5), it can be deduced that their characteristic function can be expressed as (with $\xi \in \mathbb{R}^d$):

$$\mathbb{E}\left[e^{i\xi\cdot L_t}\right] = e^{-t\psi(\xi)},$$

where the function ψ is called the *characteristic exponent*. It characterizes the Lévy process (L_t) and plays a great role in our analysis.

The following theorem is the fundamental result in the study of characteristic exponents. In particular, it introduces the notion of Lévy measure, which we use repeatedly. Several conventions or notations may exist for this formula, we follow those of (Böttcher et al., 2013, Theorem 2.2).

Theorem A.5 (Lévy-Khintchine formula)**.** *Let $(L_t)_{t\geq 0}$ be a Lévy process as above. The characteristic exponent of L has the following form:*

$$\forall \xi \in \mathbb{R}^d,\ \psi(\xi) = -il\cdot\xi + \frac{1}{2}\xi\cdot Q\xi + \int_{\mathbb{R}^d\setminus\{0\}} \left(1 - e^{iz\cdot\xi} + i\xi\cdot z\chi(\|z\|)\right) d\nu(z),$$

where $l \in \mathbb{R}^d$, $Q \in \mathbb{R}^{d\times d}$ is a symmetric positive semi-definite matrix and ν is a positive measure on $\mathbb{R}^d\setminus\{0\}$ such that

$$\int_{\mathbb{R}^d\setminus\{0\}} \min(1, \|z\|^2) d\nu(z) < +\infty.$$

Finally, χ is a truncation function*, such that $\chi(s)$ and $s\chi(s)$ are bounded and there exists a constant $\kappa > 0$ such that $0 \leq 1 - \chi(s) \leq \kappa \min(1, s)$.*

The triplet (l, Q, ν) is called the Lévy triplet *associated to L, and ν is the* Lévy measure.

Remark A.6. As it is mentioned in (Böttcher et al., 2013, Theorem 2.2), the truncation function χ is arbitrary and only influences the drift $l \in \mathbb{R}^d$. In our paper, we only consider Lévy process and infinitely divisible distributions with no drift (i.e. $l = 0$), the choice of χ therefore has no impact. A typical choice would be $\chi(s) = \min(1, s^2)$.

Remark A.7. It is clear, from the above discussion, that any infinitely divisible distribution can be associated with a Lévy triplet, as in Theorem A.5.

Example A.8 (Brownian motion). The Lévy triplet $(0, I_d, 0)$ corresponds to the standard Brownian motion in $\mathbb{R}^d$, denoted $(B_t)_{t\geq 0}$.

We end this subsection by defining stable Lévy processes, which are the main object of our study, see (Böttcher et al., 2013, Example 2.4.d).

[6] cadlag means right continuous and having a left limit everywhere.

[7] This is an equivalence: every infinitely divisible distribution is naturally associated with a Lévy process.

Definition A.9 (Stable Lévy processes)**.** Let $\alpha \in (0, 2]$, the (isotropic) α-stable Lévy process $(L_t^\alpha)_{t\geq 0}$, is defined by the following expression of its characteristic exponent: $\psi(\xi) = normof\xi^\alpha$. Its Lévy triplet is given by:

- If $\alpha = 2$, then the triplet is $(0, 2I_d, 0)$, in which case we have $L_t^2 = \sqrt{2}B_t$.
- If $\alpha \in (0, 2)$, the triplet is $(0, 0, \nu_\alpha)$, with:

$$d\nu_\alpha(z) := C_{\alpha,d}\frac{dz}{\|z\|^{d+\alpha}}, \quad C_{\alpha,d} := \alpha 2^{\alpha-1}\pi^{-d/2}\frac{\Gamma\left(\frac{\alpha+d}{2}\right)}{\Gamma\left(1-\frac{\alpha}{2}\right)}. \tag{22}$$

A.2.2. Generator of the semigroup and fractional Laplacian

In this subsection, we introduce the notion of fractional Laplacian. This is related to the study of Lévy processes through the notion of "infinitesimal generator of the semigroup", which we first define.

Given a temporally homogeneous Markov process $(X_t)_{t\geq 0}$ we define its *semigroup* $(P_t)_{t\geq 0}$ (Xiao, 2004; Schilling, 2016), as the following operators, defined for bounded measurable function f:

$$P_t f(x) = \mathbb{E}^x[f(X_t)],$$

where $\mathbb{E}^x$ denotes the initialization of the process at x (i.e. conditionally on $X_0 = x$).

If X is a Lévy process, or more generally a Feller process, see (Schilling, 2016), such a semigroup is characterized by its infinitesimal generator.

Definition A.10. Let $\mathcal{C}_\infty^0(\mathbb{R}^d)$ be the space of continuous functions vanishing to zero at infinity. As soon as it exists, we define the generator of the semigroup $(P_t)_t$ as the ensuing limit:

$$Af(x) := \lim_{t\to 0}\frac{P_t f - f}{t},$$

where the limit is understood in the uniform norm on $\mathcal{C}_\infty^0(\mathbb{R}^d)$. The domain of the generator, denoted $\mathcal{D}(A) \subset \mathcal{C}_\infty^0(\mathbb{R}^d)$, is the set of functions for which the above limit is defined[8].

It is known that the generator of $L_t^2 = \sqrt{2}B_t$ is $A\phi = \Delta\phi$, where Δ denotes the Laplacian. Following (Böttcher et al., 2013; Duan, 2015; Umarov et al., 2018), we can express the generator of (L_t^α), for $\alpha \in (0, 2)$, on the appropriated domain, which at least contains $\mathcal{C}_c^\infty(\mathbb{R}^d)$[9],

$$A\phi(x) = C_{\alpha,d}\int_{\mathbb{R}^d}(\phi(x+z) - \phi(x) - \nabla\phi(x)\cdot z\chi(\|z\|))\, d\nu_\alpha(z), \tag{23}$$

with the same notations as in Definition A.9. Following (Lischke et al., 2019; Umarov et al., 2018), Eq. (23) is one of the possible equivalent definitions of the fractional Laplacian. Note that the term fractional Laplacian is actually an abuse of notations, the correct terminology would be the negative fractional negative Laplacian, as we can see in the following definition:

$$-(-\Delta)^{\frac{\alpha}{2}}\phi(x) := C_{\alpha,d}\int_{\mathbb{R}^d}(\phi(x+z) - \phi(x) - \nabla\phi(x)\cdot z\chi(\|z\|))\, d\nu_\alpha(z), \tag{24}$$

Remark A.11 (Equivalent definitions of $(-\Delta)^{\frac{\alpha}{2}}$). There are several equivalent definitions of the fractional Laplacian, we refer the reader to (Lischke et al., 2019; Teymurazyan, 2023) for all details. We only mention two that are commonly used:

- Principal value integral this is the definition used by Raj et al. (2023b), even though their sign convention is different:

$$(-\Delta)^{\frac{\alpha}{2}}\phi(x) = C_{\alpha,d}\lim_{\epsilon\to 0}\int_{\mathbb{R}^d\setminus B_\epsilon(0)}\frac{u(x) - u(x+z)}{\|z\|^{d+\alpha}}dz.$$

- Fourier transform representation: $\mathcal{F}((-\Delta)^{\frac{\alpha}{2}}\phi)(\xi) = \|\xi\|^\alpha\mathcal{F}\phi(\xi)$.

[8] When the context allows it, the generator may be naturally extended to other spaces.

[9] $\mathcal{C}_c^\infty(\mathbb{R}^d)$ denotes the set of infinitely many times differentiable functions with compact support.

A.2.3. LÉVY-DRIVEN DIFFUSIONS

In this subsection, we consider a function $V : \mathbb{R}^d \longrightarrow \mathbb{R}$, which we call potential function. We consider the following stochastic differential equation (SDE):

$$dX_t = -\nabla V(X_t)dt + \sigma_1 dL_t^\alpha + \sigma_2\sqrt{2}dB_t, \tag{25}$$

with $\alpha \in (0, 2)$. This equation admits a strong solution (in the Itô sense), as soon as V is smooth, i.e. it satisfies Assumption 3.2, see (Schilling & Schnurr, 2010).

Let us denote by $\rho = \rho(t, x)$ the probability density of the process X. Under certain regularity conditions, this function ρ is known to satisfy the following Fokker-Planck equation, at least in a weak sense (i.e. in the sense of distributions):

$$\partial_t \rho = \nabla \cdot (\rho \nabla V) + I[\rho], \tag{26}$$

with the operator ρ being defined using the self adjoint operator of the driving process of Eq. (25) as:

$$I[\rho] = \sigma_2^2 \Delta\rho + C_{\alpha,d}\sigma_1^\alpha \int (\rho(x+z) - \rho(x) - \nabla\rho(x) \cdot z\xi(\|z\|))\, d\nu_\alpha(z), \tag{27}$$

Eq. (26) is exactly Eq. (6), with $V = F_S$.

We will use the self-adjointness of such an operator, recalled in the following lemma, of which the reader may find more precise formulations in (Gentil & Imbert, 2008; Lischke et al., 2019; Tristani, 2013).

Lemma A.12. *As soon as it is well defined, the operator I, defined by Eq. (27) is self-adjoint, i.e. for φ and ψ, regular enough, we have:*

$$\int \varphi I[\psi]\, dx = \int \psi I[\varphi]\, dx.$$

We now discuss the validity of this equation in the following two remarks. Those remarks may be skipped without hurting the general understanding of the paper. They are meant to explain what needs to be assumed to be as rigorous as possible in our treatment of the fractional Fokker-Planck equation.

Remark A.13 (Justification of the equation). Using the main result of Kühn (2018), it can be argued that, under Assumption 3.2, X is a Feller process. Its generator can be expressed as (Schilling & Schnurr, 2010; Duan, 2015; Umarov et al., 2018):

$$A\phi(x) = -\nabla V(x) \cdot \nabla\phi(x) + \sigma_2^2 \Delta\phi(x) - \sigma_1^\alpha (-\Delta)^{\frac{\alpha}{2}} \phi(x).$$

If we denote by $(P_t)_t$ the semi-group associated with X, then we have the Kolmogorov backward equation, for f in the generator $\mathcal{D}(A)$ of A:

$$\frac{d}{dt} P_t f = A P_t f = P_t A f,$$

taking the L^2-adjoint of this equation leads to $\partial_t \rho = A^\star \rho$ (at least in a weak sense). A direct computation of the adjoint $A^\star$, using properties of the fractional Laplacian, leads to Eq. (26).

Remark A.14 (Validity of Eq. (26)). Let us quickly discuss the domain of validity of the Fokker-Planck equations considered in this paper. We are in particular interested in the (local) regularity of the potential solutions, as they are necessary to give meaning to several computations made in our proofs. Let us first mention that in the simpler case of a quadratic potential, the regularity easily comes from the analytic solution of such equations (Lafleche, 2020). Moreover, equations such as Eq. (26) are known to have regularization properties and have the ability to generate smooth solutions.

More precisely, as it is mentioned in (Umarov et al., 2018), under regularity assumptions, it is known that such an equation is satisfied in a weak sense, namely in the sense of distributions (Halperin & Schwartz, 1952). Therefore, one may ask whether smooth solutions do actually exist. Smoothness in the x variable has been proven for the Ornstein-Uhlenbeck drift by (Xie et al., 2015). Space-time regularity has been proven in (Imbert, 2005) in the case of a bounded force field ∇F_S. An example of space regularity was achieved in the case of a force field with bounded derivatives of positive order, in (Wang & Duan, 2017). Finally, let us mention the work of (Lafleche, 2020), which provides further regularity conditions.

A.3. Logarithmic Sobolev inequalities and Φ-entropy inequalities

Logarithmic Sobolev inequalities (LSI) and Poincaré inequalities are central tools in probability theory. LSIs were historically introduced by Gross (1975) and have famously been applied to the analysis of Markov processes (Bakry et al., 2014), the study of evolution equations (Markowich & Villani, 2004), and have been connected to optimal transport and geometry (Villani, 2009). For a short introduction, the reader may consult the tutorials (Chafai, 2004; Chafai & Lehec, 2017). In this section, we quickly introduce a few of these results, with an emphasis on a generalization to infinitely divisible distributions, playing an important role in our study.

A.3.1. The classical inequalities

We first quickly recall the classical Poincaré inequality and LSI. They hold for the standard Gaussian measure[10], defined by $\gamma_d = \mathcal{N}(0, I_d)$. We first give Poincaré inequality:

Theorem A.15. *Let f be a function such that $\nabla f \in L^2(\gamma_d)$, we have:*

$$\int_{\mathbb{R}^d} f^2 d\gamma_d - \left(\int_{\mathbb{R}^d} f d\gamma_d\right)^2 \leq \int_{\mathbb{R}^d} \|\nabla f\|^2 d\gamma_d.$$

Example A.16. If we define the convex function $\Phi(x) = \Phi_2(x) := \frac{x^2}{2}$, as in Section 4.3, then the left hand side of the inequality of Theorem A.15 can be understood as $2\mathrm{Ent}^{\Phi}_{\gamma_d}(f)$.

The next theorem is the classical LSI for the Gaussian measure γ_d:

Theorem A.17. *Let $f \in \mathcal{C}^1(\mathbb{R}^d, R)$ be non-negative and integrable, then we have, for $\Phi(x) = \Phi_{\log}(x) = x\log(x)$:*

$$\mathrm{Ent}^{\Phi}_{\gamma_d}(f) \leq \frac{1}{2}\int_{\mathbb{R}^d} \frac{\|\nabla f\|^2}{f} d\gamma_d.$$

Example A.18. With the convex function Φ used in the above theorem, the right-hand side of the inequality is, up to the constant, the Fisher information, which we call Φ-information in Lemma 4.1.

A.3.2. Generalization to infinitely divisible distributions

Part of our analysis is based on a generalization of those inequalities to infinitely divisible distributions. Let us recall that those distributions have been defined in Appendix A.2 and can be equivalently seen as the distribution of Lévy processes. This is how we connect those inequalities to our theory.

Let us first recall the definition of Φ-entropies, i.e. Definition 2.2.

Definition 2.2 (Φ-entropies)**.** Let μ be a non-negative measure on $\mathbb{R}^d$ and $\Phi : \mathbb{R}_+ \longrightarrow \mathbb{R}$ be a convex function. Then, for a $g : \mathbb{R}^d \longrightarrow \mathbb{R}_+$, such that $g, \Phi(g) \in L^1(\mu)$, we define:

$$\mathrm{Ent}^{\Phi}_{\mu}(g) := \int \Phi(g)\, d\mu - \Phi\left(\int g\, d\mu\right).$$

Note that, by Jensen's inequality, such a term is always non-negative.

The following theorem was proved by Wu (2000) and Chafai (2004), it generalizes Poincaré and logarithmic Sobolev inequalities to infinitely divisible distributions.

Theorem A.19 (Generalized LSI)**.** *Let μ be an infinitely divisible law on $\mathbb{R}^d$, with associated triplet denoted (b, Q, ν), in the sense of Theorem A.5. We further assume that $\Phi : \mathbb{R}_+ \longrightarrow \mathbb{R}$ is a convex function that satisfies the following set of assumptions:*

$$\begin{cases} (u, v) \longmapsto \Phi(u+v) - \Phi(v) - u\Phi'(v) \text{ is convex on its domain of definition} \\ (u, x) \longmapsto \Phi''(u)x \cdot Qx \text{ is convex on its domain of definition.} \end{cases} \tag{28}$$

Then, for every smooth enough function v, we have:

$$\mathrm{Ent}^{\Phi}_{\bar{u}_\infty}(v) \leq \frac{1}{2}\int \Phi''(v)\nabla v \cdot Q\nabla v\, d\mu + \iint D_\Phi(v(x+z), v(x)) d\nu(z) d\mu(x).$$

[10] Using very simple arguments, it holds for every Gaussian, up to changes in the constants.

Note that, when $\mu = \gamma_d$, then its Lévy triplet is $(0, I_d, 0)$, so that Theorem A.19 implies Theorems A.15 and A.17. The assumptions given by (28) where in particular identified by (Chafai, 2004) and (Gentil & Imbert, 2008), they contain in particular the functions $x \mapsto x\log(x)$ (so that we recover the usual LSI in the Gaussian case), and the functions $x \mapsto x^p$, with $1 < p \leq 2$, which is the case we will consider in this study.

A.4. Some properties of the Gamma function

The Euler gamma function is classically defines by:

$$\forall x > 0,\ \Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt.$$

It has a natural extension to $\mathbb{C}\backslash(-\mathbb{N})$. In this subsection, we give a few properties of this function, which will be useful to prove the results of Section 5 in Appendix D.

One particular value:

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}.$$

Lemma A.20 (Euler's reflection formula)**.** *For all $z \in \mathbb{R}\backslash\mathbb{Z}$, we have:*

$$\Gamma(1-z)\Gamma(z) = \frac{\pi}{\sin(\pi z)}.$$

Lemma A.21 (Stirling's formulas)**.** *We have have the two following asymptotic formulas:*

$$\Gamma(x+1) \underset{x\to\infty}{\sim} \sqrt{2\pi x}\left(\frac{x}{e}\right)^x,$$

and, for all[11] *$\alpha \in \mathbb{R}$:*

$$\Gamma(x+\alpha) \underset{x\to\infty}{\sim} \Gamma(x)x^\alpha.$$

B. A PAC-Bayesian bound for sub-gaussian losses

In this section, we give proof of a PAC-Bayesian bound, which applies to sub-gaussian losses. The notations for $\mathcal{Z}$, S, μ_z, ℓ, L and $\widehat{L}_S$ are the same as in Section 1, i.e.:

$$L(w) := \mathbb{E}_{z\sim\mu_z}\left[\ell(w,z)\right], \quad ,\widehat{L}_S(w) := \frac{1}{n}\sum_{i=1}^n \ell(w,z_i),$$

with $S = (z_1, \dots, z_n) \in \mathcal{Z}^n$. We consider a prior distribution π as well as a family of posterior distributions, as it has been introduced in Appendix A.1.

By Theorem A.3, which has been proven by McAllester (2003); Maurer (2004), we know that with probability at least $1-\zeta$ under $S \sim \mu_z^{\otimes n}$, we have:

$$\mathbb{E}_{\rho_S}\left[L(w) - \widehat{L}_S(w)\right] \leq \sqrt{\frac{\mathrm{KL}\left(\rho_S||\pi\right) + \log\frac{2\sqrt{n}}{\zeta}}{2n}}.$$

Unfortunately, to the best of our knowledge, no such bounds (i.e. without the variable λ like in Theorems A.1 and A.2), exist for sub-gaussian losses, which is the assumption we make in our paper. As an additional theoretical contribution, we present the following PAC-Bayesian bound for sub-gaussian losses. We believe it may be useful for other works.

Theorem B.1. *We assume that f is s^2-subgaussian, in the sense of Assumption 3.1. Then, with probability at least $1-\zeta$ under $S \sim \mu_z^{\otimes n}$, we have:*

$$\mathbb{E}_{\rho_S}\left[L(w) - \widehat{L}_S(w)\right] \leq 2s\sqrt{\frac{\mathrm{KL}\left(\rho_S||\pi\right) + \log\frac{3}{\zeta}}{n}}.$$

[11] This formula is actually true for all $\alpha \in \mathbb{C}$, considering the extension of the Γ function into the complex plane.

Proof. The proof follows very closely that of (Vershynin, 2020, Proposition 2.5.2), which we adapt to our particular case to exhibit the exact absolute constants.

We start by fixing $0 < a < 1/(2s^2)$ and applying Theorem A.1 to the function $\phi(w, S) := an\left(L(w) - \widehat{L}_S(w)\right)^2$, which gives that, with probability at least $1-\zeta$ under $S \sim \mu_z^{\otimes n}$:

$$an\mathbb{E}_{\rho_S}\left[\left(L(w) - \widehat{L}_S(w)\right)^2\right] \leq \mathrm{KL}\left(\rho_S||\pi\right) + \log(1/\zeta) + \log \mathbb{E}_\pi \mathbb{E}_S\left[e^{an\left(L(w)-\widehat{L}_S(w)\right)^2}\right].$$

The above holds as soon as the last term is defined and finite, this will be an outcome of our computations. Let us denote $\Delta := |L(w) - \widehat{L}_S(w)|$ and estimate the expected exponential term, by Tonelli's theorem:

$$\mathbb{E}_S\left[e^{an\Delta^2}\right] = 1 + \sum_{k=1}^{\infty} \frac{a^k n^k}{k!} \mathbb{E}_S\left[\Delta^{2k}\right].$$

Let us fix some $p > 1$, we note that we have:

$$\mathbb{E}_S\left[\Delta^p\right] = \int_0^\infty \mathbb{P}\left(\Delta^p \geq \epsilon\right) d\epsilon$$

Now, by Hoeffding's inequality and several changes of variables, we have:

$$\begin{aligned}
\mathbb{E}_S\left[\Delta^p\right] &\leq \int_0^\infty \mathbb{P}\left(\Delta \geq \epsilon^{1/p}\right) d\epsilon \\
&\leq 2\int_0^\infty e^{-\frac{n\epsilon^{2/p}}{2s^2}} d\epsilon \\
&= p\int_0^\infty e^{-\frac{nv}{2s^2}} v^{\frac{p}{2}-1} dv \\
&= p\left(\frac{2s^2}{n}\right)^{p/2} \int_0^\infty e^{-t} t^{\frac{p}{2}-1} dt \\
&= p\left(\frac{2s^2}{n}\right)^{p/2} \Gamma\left(\frac{p}{2}\right) \\
&= 2\left(\frac{2s^2}{n}\right)^{p/2} \Gamma\left(1+\frac{p}{2}\right),
\end{aligned}$$

were the last two inequalities follow from the definition and properties of the Γ function, as stated in Appendix A.4. If we plug this into our previous computations, we find that:

$$\begin{aligned}
\mathbb{E}_S\left[e^{an\Delta^2}\right] &\leq 1 + 2\sum_{k=1}^{\infty} \left(\frac{2s^2}{n}\right)^k \frac{a^k n^k}{k!} k! \\
&\leq 1 + 2\sum_{k=1}^{\infty} (2s^2 a)^k \\
&= 1 + 2\frac{2s^2 a}{1-2s^2 a},
\end{aligned}$$

where the last line holds because we assumed $2as^2 < 1$. We now make the following particular choice $a := 1/(4s^2)$ and we get:

$$\mathbb{E}_S\left[e^{an\Delta^2}\right] \leq 3.$$

Now, by Jensen's inequality and Fubini's theorem, this implies that, with probability at least $1-\zeta$ over $S \sim \mu_z^{\otimes n}$:

$$\frac{n}{4s^2}\mathbb{E}_{\rho_S}\left[L(w) - \widehat{L}_S(w)\right]^2 \leq \mathrm{KL}\left(\rho_S||\pi\right) + \log(1/\zeta) + \log(3),$$

which immediately implies the desired result. □

Remark B.2. If we assume, as in Theorem A.3, that the function f is bounded in $[0,1]$, then, by Hoeffding's lemma, f is s^2-subgaussian with $s = 1/2$. Therefore, our bound implies:

$$\mathbb{E}_{\rho_S}\left[L(w) - \widehat{L}_S(w)\right] \leq \sqrt{\frac{\mathrm{KL}\left(\rho_S||\pi\right) + \log\frac{3}{\zeta}}{n}},$$

which has a slightly less good constant than Theorem A.3, but we improve the term $\log(2\sqrt{n})$ into $\log(3)$.

By combining the previous computations with the disintegrated bound of Theorem A.2, we immediately obtain:

Theorem B.3. *We assume that f is s^2-subgaussian, in the sense of Assumption 3.1. Then, with probability at least $1-\zeta$ under $S \sim \mu_z^{\otimes n}$ and $w \sim \rho_S$, we have:*

$$L(w) - \widehat{L}_S(w) \leq 2s\sqrt{\frac{\mathrm{D}_2\left(\rho_S||\pi\right) + \log\frac{24}{\zeta^3}}{n}}.$$

C. Proofs of the main theorems and additional results

In all the proofs, we use the following notation:

$$I\left[u\right] = \sigma_2^2\Delta u + C_{\alpha,d}\sigma_1^\alpha \int \left(u(x+z) - u(x) - \nabla u(x)\cdot z\xi(\|z\|)\right) d\nu_\alpha(z),$$

Note that this is the generator of the process driving Equation (3), e.g. $\sigma_2 B_t + \sigma_1 L_t^\alpha$. We will use the following notations for the drift terms:

$$V_S(w) = \widehat{F}_S(w) + \frac{\eta}{2}\left\|w\right\|^2, \qquad V(w) = \frac{\eta}{2}\left\|w\right\|^2.$$

We use the notations $u_t^S, \bar{u}_\infty, \rho_t^S, \pi$ and v_t^S in the same way as they have been introduced in Section 2. Moreover, we remind the reader that we often denote v instead of v_t^S, and u instead of u_t^S, the dependence in the time t and the data S being implicit.

C.1. The main decomposition

Before proving our main results, we define the Bregman divergence, associated with a convex function Φ. We will use it repeatedly in our proofs and statements. This notion also justifies that we call the term B_Φ^α, appearing in Lemma 4.1, the "Bregman" integral.

Definition C.1. Given a convex interval I and $\Phi : I \longrightarrow \mathbb{R}$ a convex function, we define the Bregman divergence as:

$$D_\Phi(a,b) := \Phi(a) - \Phi(b) - \Phi'(b)(a-b).$$

We first prove Lemma 4.1, which is the main decomposition that we use, in order to derive our main results. This result follows the computations of Gentil & Imbert (2008), which are adapted to the comparison of two dynamics.

Lemma 4.1 (Decomposition of the entropy flow)**.** *Given a convex and differentiable function $\Phi : (0,\infty) \longrightarrow \mathbb{R}$, we make Assumptions 3.3 and 3.4, relatively to this function Φ.*

$$\frac{d}{dt}\mathrm{Ent}_{\bar{u}_\infty}^\Phi\left(v_t^S\right) = -\sigma_2^2 I_\Phi(v_t^S) - \sigma_1^\alpha B_\Phi^\alpha(v_t^S) - \int \Phi''(v_t^S)v_t^S\nabla v_t^S\cdot\nabla\widehat{F}_S\bar{u}_\infty dx,$$

where $I_\Phi(v) := \int \Phi''(v)\left\|\nabla v\right\|^2\bar{u}_\infty dx$ is called the Φ-information, and $B_\Phi^\alpha(v)$ is called the Bregman integral, which will be formally defined in Appendix C.1.

Before proving this theorem, let us give the complete expression of the Bregman term $B_\Phi^\alpha(v)$. It is given by the following formula:

$$B_\Phi^\alpha(v) = C_{\alpha,d}\iint D_\Phi(v(x), v(x+z))\bar{u}_\infty d\nu_\alpha(z)dx.$$

Proof. In all the proof, we omit the time dependence of u and v, as already mentioned above. Let us first recall the Fokker-Planck equation satisfied by u. Rewritten with the notations of this section, this equation is:

$$\partial_t u = I\left[u\right] + \nabla \cdot (u \nabla V_S).$$

Similarly, the stationary Fokker-Planck equation satisfied by $\bar{u}_\infty$ is:

$$0 = I\left[\bar{u}_\infty\right] + \nabla \cdot (\bar{u}_\infty \nabla V).$$

We finally remind the reader that $v := u/\bar{u}_\infty$. Using the definition of v and the FPE of u and $\bar{u}_\infty$, we have:

$$\begin{aligned}\partial_t v &= \frac{1}{\bar{u}_\infty}\left(I\left[\bar{u}_\infty v\right] + \nabla \cdot (\bar{u}_\infty v \nabla V_S)\right) \\ &= \frac{1}{\bar{u}_\infty}\left(I\left[\bar{u}_\infty v\right] + \nabla \cdot (\bar{u}_\infty v \nabla V)\right) + \frac{1}{\bar{u}_\infty}\nabla \cdot (\bar{u}_\infty v \nabla (V_S - V))\end{aligned}$$

Therefore, the entropy flow is equal to:

$$\begin{aligned}\frac{d}{dt}\mathrm{Ent}^{\Phi}_{\bar{u}_\infty}(v) &= \frac{d}{dt}\int \Phi(v)\bar{u}_\infty \\ &= \int \Phi'(v)\bar{u}_\infty \partial_t v dx \\ &= \int \Phi'(v)\left(I\left[\bar{u}_\infty v\right] + \nabla \cdot (\bar{u}_\infty v \nabla V)\right) dx + \int \Phi'(v)\nabla \cdot (\bar{u}_\infty v \nabla (V_S - V)) dx \\ &=: A_1 + A_2.\end{aligned}$$

Note that the derivation under the integral is perfectly justified, from our Φ-regularity assumption.

Using the self-adjointness of I (see Appendix A.2) and the integration by parts formula, we have:

$$A_1 = \int I\left[\Phi'(v)\right] v \bar{u}_\infty dx - \int \Phi''(v) v \bar{u}_\infty \nabla v \cdot \nabla V dx.$$

Now we use the formula $r\Phi''(r) = (r\Phi'(r) - \Phi(r))'$ and get:

$$\begin{aligned}A_1 &= \int I\left[\Phi'(v)\right] v \bar{u}_\infty dx + \int (v\Phi'(v) - \Phi(v))\nabla \cdot \bar{u}_\infty \nabla V dx \\ &= \int I\left[\Phi'(v)\right] v \bar{u}_\infty dx - \int (v\Phi'(v) - \Phi(v)) I\left[\bar{u}_\infty\right] dx \\ &= \int \left(I\left[\Phi'(v)\right] v - I\left[v\Phi'(v)\right] + I\left[\Phi(v)\right]\right)\bar{u}_\infty dx.\end{aligned}$$

Therefore, we have to compute the quantity $f(v) := I\left[\Phi'(v)\right] v - I\left[v\Phi'(v)\right] + I\left[\Phi(v)\right]$. Using the definition of I, we get:

$$\begin{aligned}f(v) = \sigma_2^2 &\left(v\Delta\Phi'(v) - \Delta(v\Phi'(v)) + \Delta\Phi(v)\right) \\ &+ C_{\alpha,d}\sigma_1^\alpha \int \big\{v(x)\Phi'(v(x+z)) - v(x)\Phi'(v(x)) - v(x)\nabla\Phi'(v(x)) \cdot z\chi(\|z\|) \\ &\qquad - v(x+z)\Phi'(v(x+z)) + v(x)\Phi'(v(x)) + \nabla(v\Phi'(v))(x)) \cdot z\chi(\|z\|) \\ &\qquad + \Phi(v(x+z)) - \Phi(v(x)) - \nabla(\Phi(v))(x) \cdot z\chi(\|z\|)\big\} d\nu_\alpha(z).\end{aligned}$$

which is equal, after easy computations, to:

$$f(v) = -\sigma_2^2 \Phi''(v)\left\|\nabla v\right\|^2 - C_{\alpha,d}\sigma_1^\alpha \int D_\Phi(v(x), v(x+z)) d\nu_\alpha(z).$$

Therefore:

$$A_1 = -\sigma_2^2 \int \Phi''(v)\left\|\nabla v\right\|^2 \bar{u}_\infty dx - C_{\alpha,d}\sigma_1^\alpha \iint D_\Phi(v(x), v(x+z))\bar{u}_\infty d\nu_\alpha(z) dx.$$

for the second term, we have:

$$\begin{aligned} A_2 &= \int \Phi'(v)\nabla \cdot (\bar{u}_\infty v \nabla(V_S - V))dx \\ &= -\int \Phi''(v) v \bar{u}_\infty \nabla v \cdot \nabla(V_S - V)dx. \end{aligned}$$

The result follows. □

In the sequel, the quantity $\frac{d}{dt}\text{Ent}^{\Phi}_{\bar{u}_\infty}(v)$ will be called *entropy flow*.

Remark C.2. If, in Equations (6) and (7), we where using the multifractal process, $\sigma\sqrt{2}dB_t + \sum_{i=1}^N \sigma_i L_t^{\alpha_i}$, instead of $\sqrt{2}\sigma_2 B_t + \sigma_1 L_t^\alpha$, then the result would become:

$$\frac{d}{dt}\text{Ent}^{\Phi}_{\bar{u}_\infty}(v) = -\sigma^\alpha \int \frac{\|\nabla v\|^2}{v}\bar{u}_\infty dx - \sum_{i=1}^N C_{\alpha_i,d}\sigma_i^{\alpha_i} \iint D_\Phi(v(x), v(x+z))\bar{u}_\infty d\nu_{\alpha_i}(z)dx - \int \nabla v \cdot \nabla \widehat{F}_S \bar{u}_\infty.$$

C.2. Proof of Corollary 4.2

An easy consequence of Lemma 4.1 is the following generalization bound, namely Corollary 4.2, which was first stated in Section 4.1. It holds only when $\sigma_2 > 0$.

Corollary 4.2. *We make Assumption 3.3 and 3.4. With probability at least $1-\zeta$ over $\mu_z^{\otimes n}$, we have:*

$$G_S(T) \leq s\sqrt{\frac{1}{n\sigma_2^2}I(T,S) + 4\frac{\log(3/\zeta)+\Lambda}{n}},$$

where $\Lambda := \text{KL}(\rho_0||\pi)$ and I is defined by:

$$I(T,S) := \int_0^T \mathbb{E}_U\left\|\nabla \widehat{F}_S(W_t^S)\right\|^2 dt. \tag{11}$$

Proof. Thanks to Theorem B.1, we have, for a time t and $\zeta \in (0,1)$, with probability at least $1-\zeta$ over $S \sim \mu_z^{\otimes n}$, that:

$$\mathbb{E}_{\rho_S^t}\left[L(w) - \widehat{L}_S(w)\right] \leq 2s\sqrt{\frac{\text{KL}(\rho_S^t||\pi) + \log(3/\zeta)}{n}},$$

where the posterior ρ_S^t is the distribution with density $x \mapsto u(t,x)$, described by Equation (6), and the prior π is the distribution with density $\bar{u}_\infty$. Now we set $\Phi(u) = u\log(u)$, so that we have $\Phi''(u) = 1/u$ and, if $v = u/\bar{u}_\infty$:

$$\text{Ent}^{\Phi}_{\bar{u}_\infty}(v) = \text{KL}\left(\rho_S^t||\pi\right).$$

Therefore, by Theorem 4.1, we have:

$$\frac{d}{dt}\text{KL}\left(\rho_S^t||\pi\right) = -\sigma_2^2\int \frac{\|\nabla v\|^2}{v}\bar{u}_\infty dx - C_{\alpha,d}\sigma_1^\alpha \iint D_\Phi(v(x), v(x+z))\bar{u}_\infty d\nu_\alpha(z)dx + \int \nabla v \cdot \nabla(V - V_S)\bar{u}_\infty.$$

Using the non-negativity of the Bregman divergence of the convex function Φ, along with Cauchy-Schwarz and Young's inequalities, we have, for any constant $C > 0$:

$$\frac{d}{dt}\text{KL}\left(\rho_S^t||\pi\right) \leq -\sigma_2^2\int \frac{\|\nabla v\|^2}{v}\bar{u}_\infty dx + \frac{C}{2}\int \frac{\|\nabla v\|^2}{v}\bar{u}_\infty dx + \frac{1}{2C}\int \|\nabla V - \nabla V_S\|^2 u$$

Note that Assumptions 3.3 and 3.4 ensure that the above integrals are finite.

By choosing $C = 2\sigma_2^2$ and using the definition of u and ρ_S^t, we then have, with $\Lambda := \text{KL}\left(\rho_S^0||\pi\right) = \text{KL}(\rho_0||\pi)$:

$$\text{KL}\left(\rho_S^t||\pi\right) \leq \Lambda + \frac{1}{4\sigma^2}\int_0^t \mathbb{E}_{\rho_S^u}\left[\|\nabla V(w) - \nabla V_S(w)\|^2\right] du.$$

Finally, the results immediately follows by applying Theorem B.1, as described at the beginning of the present proof. □

C.3. Bounds on the Bregman integral - introduction of the functional $J_{\Phi,v}$

The previous computations are interesting, but we can note that the tail index α, as well as a scale σ_1 of the heavy-tailed noise, play no role in the derived bound. In particular, this approach cannot help us to derive bounds that hold in the case $\sigma_2 = 0$, i.e. a pure heavy-tailed dynamics. This section is meant to introduce the main tools for a step toward this direction. In particular, it is in this section that we justify the introduction of the functional $J_{\Phi,v}$ of Section 4.2. In this section, we fix $\Phi(x) = \Phi_{\log}(x) = x\log(x)$, while some computatios are valid in a more general setting, see C.8.

We first remark that, if we want the integral term of Corollary 4.2 to appear in the bound (in order to have a strongly interpretable bound), namely,

$$\int_0^T \mathbb{E}_U\left[\left\|\nabla \widehat{F}_S(W_t^S)\right\|^2\right] dt,$$

then we can still use Young's inequality on the last term of Lemma 4.1, and write that, for any $C > 0$:

$$\int \nabla v \cdot \nabla(V - V_S)\bar{u}_\infty \leq \frac{C}{2}\int \frac{\|\nabla v\|^2}{v}\bar{u}_\infty + \frac{1}{2C}\int \|\nabla(V - V_S)\|^2 u.$$

Therefore, in order to improve our bounds, we need to understand how the Bregman divergence integral can be used to compensate for the Fisher information given by $\int \frac{\|\nabla v\|^2}{v}\bar{u}_\infty$. The following lemma is a first step toward that direction.

Lemma C.3 (Spherical representation of the Bregman divergence integral)**.** *With the same notations and assumptions as in Theorem 4.1 and with $\Phi(x) = x\log(x)$, we have, for all $R \in (0, +\infty]$:*

$$\iint D_\Phi(v(x), v(x+z))\bar{u}_\infty d\nu_\alpha(z)dx \geq \int_0^R \int_0^r \int_s^r \int\int_{\mathbb{S}^{d-1}} \frac{\theta\cdot\nabla v(x+s\theta)\theta\cdot\nabla v(x+u\theta)}{v(x+u\theta)}\bar{u}_\infty(x)d\theta dx du ds\frac{dr}{r^{\alpha+1}}.$$

If $R = +\infty$, then this is even an equality.

The proof of the following lemma is based on a second-order approximation of the Bregman integral appearing in the Bregman integral term above. Such computations were first hinted by (Chafai, 2004; Chafaï, 2006), in particular through the notion of "Φ-calculus".

Proof. We use a spherical change of coordinates in $\mathbb{R}^d$, along with the fact that $d\nu_\alpha(z) = \|z\|^{-d-\alpha} dz$.

$$\iint D_\Phi(v(x), v(x+z))\bar{u}_\infty d\nu_\alpha(z)dx = \int\int_{\mathbb{S}^{d-1}}\int_0^\infty D_\Phi(v(x), v(x+r\theta))\bar{u}_\infty(x)\frac{dr}{r^{\alpha+1}}d\theta\, dx.$$

Let us fix some $R > 0$. By Tonelli's theorem and the positivity of the Bregman divergence, we can write that:

$$\iint D_\Phi(v(x), v(x+z))\bar{u}_\infty d\nu_\alpha(z)dx \geq \int_0^R \int_{\mathbb{S}^{d-1}}\int_{\mathbb{R}^d} D_\Phi(v(x), v(x+r\theta))\bar{u}_\infty(x)dx d\theta\frac{dr}{r^{\alpha+1}}\,.$$

Let us fix x, r and θ, by the Φ-regularity assumption, we have that v is differentiable, therefore, we have:

$$\begin{aligned}
D_\Phi(v(x), v(x+r\theta)) &= \Phi(v(x)) - \Phi(v(x+r\theta)) - \Phi'(v(x+r\theta))(v(x) - v(x+r\theta)) \\
&= -\int_0^r \Phi'(v(x+s\theta))\theta\cdot\nabla v(x+s\theta)ds + \Phi'(v(x+r\theta))\int_0^r \theta\cdot\nabla v(x+s\theta)ds \\
&= \int_0^r \theta\cdot\nabla v(x+s\theta)\left(\Phi'(v(x+r\theta)) - \Phi'(v(x+s\theta))\right)ds \\
&= \int_0^r\int_s^r \theta\cdot\nabla v(x+s\theta)\Phi''(v(x+u\theta))\theta\cdot\nabla v(x+u\theta)du\, ds
\end{aligned}$$

Now we use the fact that $\Phi(x) = x\log(x)$, hence $\Phi''(x) = 1/x$, which gives:

$$D_\Phi(v(x), v(x+r\theta)) = \int_0^r\int_s^r \frac{\theta\cdot\nabla v(x+s\theta)\theta\cdot\nabla v(x+u\theta)}{v(x+u\theta)}du\, ds,$$

Finally, thanks to the Φ-regularity assumption, the function:

$$x \longmapsto \frac{\theta \cdot \nabla v(x+s\theta)\theta \cdot \nabla v(x+u\theta)}{v(x+u\theta)} \bar{u}_\infty(x),$$

is integrable for each (s,u,θ), moreover, by the dominated convergence theorem and the Φ-regularity assumption, the function:

$$(s,u,\theta) \longmapsto \int_{\mathbb{R}^d} \left| \frac{\theta \cdot \nabla v(x+s\theta)\theta \cdot \nabla v(x+u\theta)}{v(x+u\theta)} \right| \bar{u}_\infty(x) dx,$$

is continuous. Now, as we integrate those variables over compact sets, we have:

$$\int_{\mathbb{S}^{d-1}} \int_0^r \int_s^r \int_{\mathbb{R}^d} \left| \frac{\theta \cdot \nabla v(x+s\theta)\theta \cdot \nabla v(x+u\theta)}{v(x+u\theta)} \right| \bar{u}_\infty(x) dx du ds d\theta < +\infty.$$

The result follows from the application of Fubini's theorem. □

The term appearing in the above lemma resembles a lot the Fisher information,

$$\mathbf{J}_v := \int \frac{\|\nabla v\|^2}{v} \bar{u}_\infty. \tag{29}$$

Note that, with $\Phi(x) = x\log(x)$, which is the case in all this section, we have, as introduced in Section 2:

$$I_\Phi(v) = \mathbf{J}_v.$$

This justifies the introduction of the following notion of "spherical information".

Definition C.4 (Spherical Fisher information)**.** For $r \geq 0$, we introduce the following spherical Fisher information, for $r > 0$:

$$J_v(r) := \frac{2d}{r^2 \sigma_{d-1}} \int_0^r \int_s^r \int_{\mathbb{R}^d} \int_{\mathbb{S}^{d-1}} \frac{\theta \cdot \nabla v(x+s\theta)\theta \cdot \nabla v(x+u\theta)}{v(x+u\theta)} \bar{u}_\infty(x) d\theta dx du ds,$$

where σ_{d-1} is the surface area of the $(d-1)$-dimensional hyper-sphere $\mathbb{S}^{d-1} \subset \mathbb{R}^d$, it is given by:

$$\sigma_{d-1} = \frac{2\pi^{d/2}}{\Gamma(d/2)}. \tag{30}$$

J_v is the function $J_{\Phi,v}$ introduced in the main part of the paper, for the particular case $\Phi(x) = x\log(x)$.

The following lemma justifies the normalization used in this definition.

We also define:

$$g(s,u) = \int_{\mathbb{R}^d} \int_{\mathbb{S}^{d-1}} \frac{\theta \cdot \nabla v(x+s\theta)\theta \cdot \nabla v(x+u\theta)}{v(x+u\theta)} \bar{u}_\infty(x) d\theta dx$$

We will denote by $\partial g/\partial s$ (resp. $\partial g/\partial u$) the partial derivative of g with respect to its first (resp. second) variable.

Lemma C.5. *Under Assumption 3.3, g is differentiable in the second variable and both g and $\frac{\partial g}{\partial u}$ are jointly continuous in $(s,u) \in \mathbb{R}^2_+$.*

Proof. This follows from the Φ-regularity condition. More precisely, let us denote:

$$h(s,u;x,\theta) = \frac{\theta \cdot \nabla v(x+s\theta)\theta \cdot \nabla v(x+u\theta)}{v(x+u\theta)} \bar{u}_\infty(x).$$

By the Φ-regularity condition, we know that this function is continuous in (s,u). Moreover, let $V \subset \mathbb{R}_+^2$ be a bounded open set of $\mathbb{R}_+^2$. By the Φ-regularity condition, the mappings $(s,u) \longmapsto |h(s,u;x,\theta)|$ are uniformly dominated on V by a function $\chi_V \in L^1(\mathbb{S}^{d-1} \times \mathbb{R}^d)$. Therefore, by the dominated convergence theorem, we have the joint continuity of g.

Now, again by the Φ-regularity condition, h is differentiable in the second variable and, u, and the partial derivatives $\frac{\partial g}{\partial u}$ are uniformly (*w.r.t.* $(s,u) \in V$) dominated by a function in $L^1(\mathbb{S}^{d-1} \times \mathbb{R}^d)$. Therefore, we can differentiate under the integral sign and get that g is differentiable in its second variable, u.

We get the joint continuity of $\frac{\partial g}{\partial u}$ by a very similar argument than before, it is again a consequence of our Φ-regularity condition. □

Remark C.6. The Φ-regularity assumption, Assumption 3.3, has been designed, in particular, to get enough regularity of this function g, which is a central tool in our proofs. This is the central reason why we need that much regularity of the functions v_t.

Lemma C.7. *The function $J_v(r)$ can be continuously extended to $[0,+\infty)$, by setting:*

$$J_v(0) = \mathbf{J}_v = \int \frac{\|\nabla v\|^2}{v} \bar{u}_\infty,$$

the obtained function is still denoted J_v.

Proof. We first compute:

$$g(0,0) = \int_{\mathbb{R}^d} \int_{\mathbb{S}^{d-1}} \frac{(\theta \cdot \nabla v(x))^2}{v(x)} \bar{u}_\infty(x) d\theta dx.$$

As the distribution that is considered on the sphere is the uniform distribution, the invariance by rotation, along with Tonelli's theorem, implies that:

$$g(0,0) = \int \frac{\|\nabla v\|^2}{v} \bar{u}_\infty dx \int_{\mathbb{S}^{d-1}} \theta_1^2 d\theta.$$

Now we easily compute:

$$\int_{\mathbb{S}^{d-1}} \theta_1^2 d\theta = \frac{1}{d} \int_{\mathbb{S}^{d-1}} \sum_{i=1}^d \theta_i^2 d\theta = \frac{1}{d} \int_{\mathbb{S}^{d-1}} d\theta = \frac{\sigma_{d-1}}{d},$$

so that:

$$g(0,0) = \frac{\sigma_{d-1}}{d} \int \frac{\|\nabla v\|^2}{v} \bar{u}_\infty dx.$$

Let us now define the following function:

$$H(r) := \int_0^r \int_s^r g(s,u) du ds = \int_0^r \int_0^u g(s,u) ds du$$

It is clear that $H(0) = 0$ and that:

$$H'(r) = \int_0^r g(s,r) ds.$$

Therefore we also have $H'(0) = 0$. Let us fix some $a > 0$, from the Φ-regularity condition, we justify that the function $g(s,u)$ and $\frac{\partial g}{\partial u}$ are continuous, and therefore uniformly continuous on the compact $[0,a] \times [0,a]$ (by Heine's theorem). Thus, let us fix some $\epsilon > 0$ and compute, for $0 \leq r < a$:

$$\begin{aligned}
\frac{H'(r+\epsilon) - H'(r)}{\epsilon} &= \frac{1}{\epsilon} \int_r^{r+\epsilon} g(s, r+\epsilon) ds + \frac{1}{\epsilon} \int_0^r (g(s,r+\epsilon) - g(s,r)) ds \\
&= \frac{1}{\epsilon} \int_r^{r+\epsilon} g(s,r) ds + \frac{1}{\epsilon} \int_r^{r+\epsilon} (g(s,r+\epsilon) - g(s,r)) ds + \frac{1}{\epsilon} \int_0^r (g(s,r+\epsilon) - g(s,r)) ds
\end{aligned}$$

For the first term, we clearly have, by definition of the derivative:

$$\frac{1}{\epsilon}\int_r^{r+\epsilon} g(s,r)ds \underset{\epsilon\to 0}{\longrightarrow} g(r,r).$$

From the fact that $\frac{\partial}{\partial u}$ is continuous, on $[0,a]\times[0,a]$, we deduce that it also uniformly continuous on this set. Therefore, we have:

$$\left|\frac{1}{\epsilon}\int_r^{r+\epsilon}(g(s,r+\epsilon)-g(s,r))ds\right| \leq \epsilon \left\|\frac{\partial}{\partial u}\right\|_{L^\infty([0,a]^2)} \underset{\epsilon\to 0}{\longrightarrow} 0.$$

From the dominated convergence, thanks to Lemma C.5, we can differentiate under the integral sign and get that:

$$\frac{1}{\epsilon}\int_0^{r}(g(s,r+\epsilon)-g(s,r))ds \underset{\epsilon\to 0}{\longrightarrow} \int_0^r \frac{\partial g}{\partial u}(s,r)ds.$$

Finally, we have: $H''(0)=g(0,0)$. This also implies:

$$J_v(r)=\frac{2d}{r^2\sigma_{d-1}}H(r)\underset{r\to 0}{\longrightarrow}\frac{2d}{\sigma_{d-1}}\frac{1}{2}g(0,0)=\mathbf{J}_v=\int\frac{\|\nabla v\|^2}{v}\bar{u}_\infty.$$

This is the desired result. □

We therefore have the following integral representation of the Bregman integral term:

$$\iint D_\Phi(v(x),v(x+z))\bar{u}_\infty d\nu_\alpha(z)dx=\frac{\sigma_{d-1}}{2d}\int_0^\infty J_v(r)\frac{dr}{r^{\alpha-1}}. \tag{31}$$

Note moreover that, by non-negativity of the Bregman divergence, the function J_v is non-negative.

With the notations of Lemma 4.1, we have:

$$B_\Phi^\alpha(v)=C_{\alpha,d}\frac{\sigma_{d-1}}{2d}\int_0^\infty J_v(r)\frac{dr}{r^{\alpha-1}}. \tag{32}$$

Therefore, Eq. (31) is the justification of Eq. (12) given in Section 4.2.

Remark C.8 (About the value of R in Assumption 4.3). In Section 4, we argue that the value of the parameter R introduced in Assumption 4.3 can be sufficiently large when the function v is reasonable close to a constant function (i.e., $v\equiv 1$). Let us make this argument slightly more formal. Let us fix such a function v and assume that there exists $\varepsilon_1,\varepsilon_2>0$ such that, uniformly on $\mathbb{R}^d$, we have:

$$v\geq\varepsilon_1,\quad \|\nabla v\|\leq\varepsilon_2,\quad \left\|\nabla^2 v\right\|\leq\varepsilon_2. \tag{33}$$

We also assume that v is twice continuously differentiable. Then, based on the proof of Lemma C.7, we have:

$$|J_v(r)-J_v(0)|\leq J_v(r):=\frac{2d}{r^2\sigma_{d-1}}\int_0^r\int_s^r\int_{\mathbb{R}^d}\int_{\mathbb{S}^{d-1}}\Delta(x,\theta,s,u)\bar{u}_\infty(x)d\theta dx du ds,$$

where:

$$\Delta(x,\theta,s,u):=\left|\frac{\theta\cdot\nabla v(x+s\theta)\theta\cdot\nabla v(x+u\theta)}{v(x+u\theta)}-\frac{(\theta\cdot\nabla v(x))^2}{v(x)}\right|$$

Using the conditions (33), we can easily see that:

$$\Delta(x,\theta,s,u)\leq C_1\|\theta\|^2(u+s),$$

with C_1 a constant depending on ε_1 and ε_2. Therefore, we have $|J_v(r)-J_v(0)|\leq C_2 r$, with C_2 a constant depending on ε_1, ε_2 and d. This shows that the derivative of J_v in 0 can be controlled, hence allowing R to be big enough.

C.4. Pure Levy case: $\sigma_2 = 0$ - Additional results with Bregman Fisher inequalities

Before proving our main results, i.e. the results of Section 4.2, we quickly present a more general point of view. The message of this section is the following, any inequality of the form:

$$I_\Phi(v) \lesssim B_\Phi^\alpha(v),$$

where $B_\Phi^\alpha(v)$ and $I_\Phi(v)$ have been defined in Lemma 4.1. In all this section, we assume $\sigma_2 = 0$.

We now deduce generalization bounds in the case where we do not have any Brownian part in the bounds. We denote, as before:

$$v_t^S := \frac{u_t^S}{\bar{u}_\infty}.$$

As we did repeatedly until now, we often omit the dependence of v on t and S.

In this subsection, we introduce one of the main ingredient behind our proof of generalization bounds in the pure heavy-tailed case, i.e. $\sigma_2 = 0$.

The main argument is that such bounds appear if we can use the Bregman integral to control the Fisher information coming from Young's inequality in the proof of Corollary 4.2. We formalize the connection between such a functional inequality by the following definition. We will then see how the results of the previous sections can make this inequalities happen in practice.

Definition C.9. Given a smooth convex function Φ, we introduce the notion of Bregman-Fisher inequality, denoted $\mathrm{BF}_\Phi(\gamma, \alpha)$, for $\gamma > 0$. For a (smooth enough) function v, we say that v satisfies $\mathrm{BF}_\Phi(\gamma, \alpha)$, with respect to $\bar{u}_\infty$, when:

$$\iint D_\Phi(v(x), v(x+z))\bar{u}_\infty d\nu_\alpha(z)dx \geq \frac{\gamma}{C_{\alpha,d}} \int \frac{\|\nabla v\|^2}{v}\bar{u}_\infty dx.$$

It is clear that we have the following result:

Theorem C.10. *We make Assumptions 3.1, 3.2 and 3.3, where the Φ-regularity is for the $\Phi(x) := x\log(x)$. We further assume that there exists a constant $\gamma > 0$ for each $t > 0$ and $S \in \mathcal{Z}^n$, the function v_t^S satisfies $\mathrm{BF}_\Phi(\gamma, \alpha)$ with respect to $\bar{u}_\infty$. Then, with probability at least $1 - \zeta$ over $\mu_z^{\otimes n}$, we have:*

$$\mathbb{E}_U\left[L(W_T^S) - \widehat{L}_S(W_T^S)\right] \leq 2s\sqrt{\frac{1}{4n\sigma_1^\alpha\gamma}\int_0^T \mathbb{E}_U\left[\left\|\nabla \widehat{F}_S(W_t^S)\right\|^2\right]dt + \frac{\Lambda + \log(3/\zeta)}{n}}.$$

where $\Lambda := \mathrm{KL}\left(\rho_0||\bar{u}_\infty\right)$.

Proof. From Theorem 4.1, the BF condition and Young's inequality, we immediately get, with $v = v_t^S$:

$$\begin{aligned}
\frac{d}{dt}\mathrm{Ent}_{\bar{u}_\infty}^\Phi(v) &\leq -C_{\alpha,d}\sigma_1^\alpha \iint D_\Phi(v(x), v(x+z))\bar{u}_\infty d\nu_\alpha(z)dx + \int \Phi''(v)v\nabla v \cdot \nabla(V - V_S)\bar{u}_\infty \\
&\leq -\gamma\sigma_1^\alpha \int \frac{\|\nabla v\|^2}{v}\bar{u}_\infty dx + \int \nabla v \cdot \nabla(V - V_S)\bar{u}_\infty \\
&\leq \frac{1}{4\gamma\sigma_1^\alpha}\int \|\nabla V - \nabla V_S\|^2 du_t^S.
\end{aligned}$$

Therefore, by using the same reasoning as in Corollary 4.2, we get the results. □

C.5. Pure Levy case: $\sigma_2 = 0$ - Omitted proofs of Section 4.2

Functional inequalities like $\mathrm{BF}_\Phi(\gamma, \alpha)$ are not trivial at all to get in practice. In the rest of this subsection, we will justify that, under a reasonable assumption, we can satisfy an almost identical identity. This will give us an idea of the rate that we expect for our bounds.

The idea is the following: the results of the previous section, namely Equation (31) and Lemma C.7 point us toward the following informal computation, for some $R > 0$:

$$\iint D_\Phi(v(x), v(x+z))\bar{u}_\infty d\nu_\alpha(z)dx \geq \frac{\sigma_{d-1}}{2d}\int_0^R J_v(r)\frac{dr}{r^{\alpha-1}} \simeq \frac{\sigma_{d-1}R^{2-\alpha}}{2d(2-\alpha)}J_v(0).$$

Therefore, we see that we can control the Fisher information terms, coming from Young's inequality, using the above integral. However, we need an additional assumption to control the behavior of the function J_v, uniformly with respect to the data and the time.

We formalize this idea with the following assumption:

Assumption C.11. We assume there exists an absolute constant $R > 0$ such that, for all $t > 0$ and $\mu_z^{\otimes n}$-almost all $S \in \mathcal{Z}^n$, we have:

$$\forall r \in [0, R],\ J_{v_t^S}(r) \geq \frac{1}{2}\mathbf{J}_{v_t^S}.$$

This assumption is exactly a reformulation of Assumption 4.3, with $\Phi(x) = \Phi_{\log} = x\log(x)$.

If we fix t and S, this assumption is trivial by the continuity of J_v. The above condition is essentially a kind of weak uniformity in t and S of this continuity. The uniformity in t would be justified in case of convergence of u_t^S to a limit distribution. The strongest part of the assumption is the uniformity in S. Note that it is common in the learning theory literature to assume uniformity of various constant in the data.

Theorem 4.4. *We make Assumptions 3.3, 3.4 and 4.3. Then, with probability at least $1-\zeta$ over $\mu_z^{\otimes n}$, we have*

$$G_S(T) \leq 2s\sqrt{\frac{K_{\alpha,d}}{n\sigma_1^\alpha}I(T,S) + \frac{\log(3/\zeta)+\Lambda}{n}}$$

with Λ and $I(T,S)$ as in Corollary 4.2, and:

$$K_{\alpha,d} = \frac{(2-\alpha)\Gamma\left(1-\frac{\alpha}{2}\right)d\Gamma\left(\frac{d}{2}\right)}{\alpha 2^\alpha \Gamma\left(\frac{d+\alpha}{2}\right)R^{2-\alpha}}, \tag{13}$$

where Γ denotes the Euler's Gamma function, on which more information is provided in Appendix A.4.

Proof. Let us fix $t > 0$ and $S \in \mathcal{Z}^n$ and denote v for v_t^S, as before. By Theorem 4.1, we have:

$$\frac{d}{dt}\mathrm{Ent}^\Phi_{\bar{u}_\infty}(v) \leq -C_{\alpha,d}\sigma_1^\alpha \iint D_\Phi(v(x), v(x+z))\bar{u}_\infty d\nu_\alpha(z)dx + \int \nabla v \cdot \nabla(V - V_S)\bar{u}_\infty$$

By Young's inequality, if $C > 0$ is a constant:

$$\frac{d}{dt}\mathrm{Ent}^\Phi_{\bar{u}_\infty}(v) \leq -C_{\alpha,d}\sigma_1^\alpha \iint D_\Phi(v(x), v(x+z))\bar{u}_\infty d\nu_\alpha(z)dx + \frac{C}{2}\int \frac{\|\nabla v\|^2}{v}\bar{u}_\infty dx + \frac{1}{2C}\int \|\nabla V - \nabla V_S\|^2 du_t^S.$$

By Assumption 4.3 and Lemma C.3, we have:

$$\begin{aligned}
\frac{d}{dt}\mathrm{Ent}^\Phi_{\bar{u}_\infty}(v) &\leq -C_{\alpha,d}\sigma_1^\alpha \frac{\sigma_{d-1}}{2d}\int_0^R J_v(r)\frac{dr}{r^{\alpha-1}} + \frac{C}{2}\int \frac{\|\nabla v\|^2}{v}\bar{u}_\infty dx + \frac{1}{2C}\int \|\nabla V - \nabla V_S\|^2 du_t^S. \\
&\leq -C_{\alpha,d}\sigma_1^\alpha \frac{\sigma_{d-1}}{4d}\int_0^R J_v(0)\frac{dr}{r^{\alpha-1}} + \frac{C}{2}\int \frac{\|\nabla v\|^2}{v}\bar{u}_\infty dx + \frac{1}{2C}\int \|\nabla V - \nabla V_S\|^2 du_t^S. \\
&= -\frac{C_{\alpha,d}\sigma_1^\alpha \sigma_{d-1}R^{2-\alpha}}{4d(2-\alpha)}\int \frac{\|\nabla v\|^2}{v}\bar{u}_\infty dx + \frac{C}{2}\int \frac{\|\nabla v\|^2}{v}\bar{u}_\infty dx + \frac{1}{2C}\int \|\nabla V - \nabla V_S\|^2 du_t^S.
\end{aligned}$$

So we make the choice:

$$C = \frac{C_{\alpha,d}\sigma_1^\alpha \sigma_{d-1} R^{2-\alpha}}{2d(2-\alpha)},$$

and, putting everything together, we have:

$$\frac{d}{dt}\mathrm{Ent}^{\Phi}_{\bar{u}_\infty}(v) \leq \frac{K_{\alpha,d}}{\sigma_1^\alpha}\int \|\nabla V - \nabla V_S\|^2 \, du_t^S,$$

with:

$$K_{\alpha,d} := \frac{d(2-\alpha)}{C_{\alpha,d}\sigma_{d-1}R^{2-\alpha}}$$

We conclude by the same PAC-Bayesian arguments as in the proof of Corollary 4.2, i.e., we use THeorem B.1.

Regarding the value of the constant $K_{\alpha,d}$, we remind the reader that we have:

$$C_{\alpha,d} = \alpha 2^{\alpha-1}\pi^{-d/2}\frac{\Gamma\left(\frac{\alpha+d}{2}\right)}{\Gamma\left(1-\frac{\alpha}{2}\right)}, \quad \sigma_{d-1} = \frac{2\pi^{d/2}}{\Gamma(d/2)}.$$

A simple computation gives:

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

which is the desired result. □

C.6. Extension to the discrete-time case

In this section, we quickly demonstrate that our methods, developped in the time-continuous setting, can be extended to the discrete setting. This gives more theoretical foundations to our experimental analysis. We treat this case slightly less formally than the rest of the paper, our main goal is to make a first step toward the understanding of the discrete heavy-tailed algorithms. Let us simplify the notations of Section 6 and consider the following discrete recursion:

$$w_{k+1}^S = w_k^S - \gamma g(w_k^S) - \gamma\eta w_k^S + \gamma^{1/\alpha}\sigma_1 L_1^\alpha, \tag{34}$$

where $g(w_k^S) := \nabla \widehat{F}_S(w_k^S)$ and $S \sim \mu_z^{\otimes n}$. We assume that $\gamma\eta < 1$.

Remark C.12. We could also consider that g_k is an unbiased estimate of the true gradient, i.e. using random batches independent of the stable noise. Most of our analysis would also hold in this case. However, we focus on the full-batch case, both for simplicity and to stick to the theoretical foundations of our experimental work. As mentioned in Section 6, the used of mini-batches could result in gradient with heavy-tailed noise, which could interfere in an unclear way with the stable noise L_t^α. Similarly, the same techniques could be extended to varying learning rate and noise scale, but we stick to a setting close to both the time-continuous and the experimental settings.

Extensions from the discrete In this section, we adapt the technique presented in (Mou et al., 2017, Section 5) to the heavy-tailed setting. This will highlight that our technical contribution in the continuous case are directly useful for the discrete case. The strategy is the following:

1. We will construct a Levy driven Ornstein-Uhlenbeck process interpolating between the density of two successive iterates.
2. We apply the analysis of Section 4.2 and use the associated FPE to bound the KL divergences of each iterate w_k.

Interpolating techniques have also been used by Nguyen et al. (2019a), in the study of the discretization of heavy-tailed SDEs.

Let us fix some $\sigma' > 0$ and consider the process, for a fixed $z \in \mathbb{R}^d$:

$$dX_t = -\eta X_t dt - g(z)dt + \sigma' dL_t^\alpha,$$

where, as defined above, $g = \nabla \widehat{F}_S$. Note that X depends on z. We can express the solution as:

$$X_t + \frac{g(z)}{\eta} = e^{-\eta t}\left(X_0 + \frac{g(z)}{\eta}\right) + \underbrace{\sigma' \int_0^t e^{-\eta(t-s)} dL_s^\alpha}_{:=\mathcal{O}_t}.$$

We can compute the characteristic function of the integral term, using computations similar as in (Raj et al., 2023a, Lemma 9), for all $\xi \in \mathbb{R}^d$:

$$\mathbb{E}\left[e^{i\xi\cdot\mathcal{O}_t}\right] = \exp\left\{-\int_0^t \left\|\sigma' e^{-\eta s}\xi\right\|^\alpha ds\right\} = \exp\left\{-\sigma'^\alpha \|\xi\|^\alpha \frac{1-e^{-\eta\alpha t}}{\alpha\eta}\right\}.$$

Let us now fix one iteration k and denote by u_k^S the probability density of w_k^S. We set the initial condition $X_0 \sim u_k^S$. Then, by the two previous equations, we have that, for a fixed $\tau > 0$:

$$X_\tau \sim e^{-\eta\tau} u_k^S + \frac{1-e^{-\eta\tau}}{\eta} g(u_k^S) + \sigma' \left\{\frac{1-e^{-\eta\alpha\tau}}{\alpha\eta}\right\}^{\frac{1}{\alpha}} L_1^\alpha.$$

Our goal is that X interpolates between u_k^S and u_{k+1}^S, therefore we set:

$$e^{-\eta\tau} = 1 - \gamma\eta, \quad \sigma'\left\{\frac{1-e^{-\eta\alpha\tau}}{\alpha\eta}\right\}^{\frac{1}{\alpha}} = \sigma_1 \gamma^{\frac{1}{\alpha}},$$

so that is reproduced Eq. (34). Thus X_τ and u_{k+1}^S have the same distributions.

Let us denote by $h_t^{k,S}$ the density of X_t at time t, were we made explicit its dependence on the data S and the iteration number k. Now, if we proceed as in (Mou et al., 2017, Theorem 9), and integrate the FPE of X with respect to u_k^S, we get that the density $h_t^{k,S}$ of X_t satisfies, provided we can switch the differential operators and the integration over u_k^S:

$$\partial_t h_t^{k,S} = -\sigma'^\alpha (-\Delta)^{\frac{\alpha}{2}} h_t^{k,S} + \eta\nabla\cdot(h_t^{k,S} w) + \nabla\cdot(h_t^{k,S}\mathbb{E}\left[g(w_k^S)|X_t = w\right])$$

This fractional FPE has exactly the form studied in this paper, therefore, we can express the associated entropy flow as (with a slight abuse of notation, we identify h_t with the associated probability distribution):

$$\frac{d}{dt}\mathrm{KL}\left(h_t^{k,S}||\pi\right) = -\iint B_\Phi(v_t^{k,S}(x), v_t^{k,S}(x+z))\bar{u}_\infty(x)\frac{dz}{\|z\|^{\alpha+d}}dx - \int \nabla v_t^{k,S}\cdot\mathbb{E}\left[g(w_k^S)|X_t = w\right]dw,$$

with:

$$v_t^{k,S} := \frac{h_t^{k,S}}{\bar{u}_\infty}.$$

This leads us to formulate the following assumption, which is the extension of Assumption 4.3 to the discrete case.

Assumption C.13. We assume that there exists a constant R such that, for all k, all t and all dataset S, the functions $v_t^{k,S}$, constructed by the above procedure, satisfy:

$$\forall r < R, \quad J_{\Phi_{\log}, v_t^{k,S}}(r) \geq \frac{1}{2} J_{\Phi_{\log}, v_t^{k,S}}(0).$$

We now omit, as we did in the time continuous setting, the dependence of $v_t^{k,S}$ on k and S, and just denote it v_t. We do the same for $h_t^{k,S}$, simply denoted h_t. Therefore, by the proof of Theorem 4.4, we have:

$$\frac{d}{dt}\mathrm{Ent}_{\bar{u}_\infty}^\Phi(v_t) \leq \frac{K_{\alpha,d}}{\sigma'^\alpha}\int \left\|\mathbb{E}\left[g(w_k^S)|X_t = w\right]\right\|^2 dh_t,$$

with $K_{\alpha,d}$ defined as in Theorem 4.4, using the constant R coming from Assumption C.13. Let us denote by $h_t(w,z)$ the joint density of X_t and w_k^S, by Cauchy-Schwarz's inequality, we have:

$$\begin{aligned}\frac{d}{dt}\text{Ent}_{\tilde{u}_\infty}^{\Phi}(v_t) &= \frac{K_{\alpha,d}}{\sigma'^\alpha}\int\left\|\int\frac{h_t(w,z)}{h_t(w)}g(z)dz\right\|^2 h_t(w)dw \\ &\leq \frac{K_{\alpha,d}}{\sigma'^\alpha}\int h_t(w)\left(\int\frac{h_t(w,z)}{h_t(w)^2}dz\right)\left(\int h_t(w,z)g(z)dz\right)dw \\ &= \frac{K_{\alpha,d}}{\sigma'^\alpha}\iint h_t(w,z)g(z)dzdw \\ &= \frac{K_{\alpha,d}}{\sigma'^\alpha}\mathbb{E}_U\left[\left\|g(w_k^S)\right\|^2\right].\end{aligned}$$

Let us denote by ρ_k the density of w_k. Integrating between 0 and τ, and using that h_0 is the density of w_k and that h_τ is the density of w_{k+1}, we can write that:

$$\text{KL}\left(\rho_{k+1}||\pi\right) \leq \text{KL}\left(\rho_k||\pi\right) + \tau\frac{K_{\alpha,d}}{\sigma'^\alpha}\mathbb{E}_U\left[\left\|g(w_k^S)\right\|^2\right],$$

where, as in the rest of the paper, U denotes the randomness coming from the stable noise, i.e. the noise due to L_t^α.

Therefore, a telescopic sum immediately gives that, for a fixed number N of iterations:

$$\begin{aligned}\text{KL}\left(\rho_N||\pi\right) &\leq \Lambda + \tau\frac{K_{\alpha,d}}{\sigma'^\alpha}\sum_{k=0}^{N-1}\mathbb{E}_U\left[\left\|g(w_k^S)\right\|^2\right] \\ &= \Lambda + \tau\frac{K_{\alpha,d}}{\sigma'^\alpha}\sum_{k=0}^{N-1}\mathbb{E}_U\left[\left\|g(w_k^S)\right\|^2\right] \\ &= \Lambda + \frac{K_{\alpha,d}}{\sigma_1^\alpha}\frac{1}{\gamma\eta}\log\left(\frac{1}{1-\gamma\eta}\right)\left(\frac{1-(1-\gamma\eta)^\alpha}{\alpha\eta}\right)\sum_{k=0}^{N-1}\mathbb{E}_U\left[\left\|g(w_k^S)\right\|^2\right],\end{aligned}$$

with $\Lambda = \text{KL}\left(\rho_0||\pi\right)$.

Therefore, under the subgaussian assumption, Assumption 3.1, we have proven that, with probability at least $1-\zeta$ over $S\sim\mu_z^{\otimes n}$, we have:

$$\mathbb{E}_U\left[G_S(w_N^S)\right] \leq 2s\sqrt{\frac{K_{\alpha,d}}{\sigma_1^\alpha}\frac{1}{\gamma\eta}\log\left(\frac{1}{1-\gamma\eta}\right)\left(\frac{1-(1-\gamma\eta)^\alpha}{\alpha\eta}\right)\sum_{k=0}^{N-1}\mathbb{E}_U\left[\left\|\nabla\widehat{F}_S(w_k^S)\right\|^2\right] + \frac{\Lambda+\log(3/\zeta)}{n}}. \quad (35)$$

Finally, let us notice that when $\gamma\eta$ is small, which is the case in our experiments, see Appendix F.1, we have the following asymptotic development:

$$\frac{1}{\gamma\eta}\log\left(\frac{1}{1-\gamma\eta}\right)\left(\frac{1-(1-\gamma\eta)^\alpha}{\alpha\eta}\right) \underset{\gamma\eta\to 0}{\sim} \gamma.$$

Therefore, the computations made in this section, while slightly informal, give a solid justification to Eq. (18), which we use in our experiments. The qualitative behavior with respect to α and d is unchanged.

Remark C.14. The analysis presented in this section could also be extended to SDEs with a non trivial Brownian contribution to the noise, i.e. the setting of Section 4.1 where $\sigma_2 > 0$.

C.7. Mixing of Brownian and heavy-tailed noise

Because the Fisher information term, appearing in Lemma 4.1, is non-negative, the proof of Theorem 4.4 also applies when $\sigma_2 > 0$, i.e. in the setting of Section 4.1. However, not that this requires to make Assumption 4.3, while such an assumption

is not necessary to derive Corollary 4.2. Therefore, the results of Section 4.1 still present the advantage to hold under lighter assumptions.

Nevertheless, this motivates to write down the result in the case $\sigma_2 > 0$, under Assumption 4.3. This allows to get insights on how both noise affect the generalization error. This leads to the following theorem.

Theorem C.15. *Let us consider the dynamics of Eq. (3), with both $\sigma_1 > 0$ and $\sigma_2 > 0$. We make Assumptions 3.3, 3.4 and 4.3. Then, with probability at least $1-\zeta$ over $\mu_z^{\otimes n}$, we have*

$$G_S(T) \leq 2s\sqrt{\frac{M(\sigma_1,\sigma_2,d,\alpha)}{n}\int_0^T \mathbb{E}_U\left[\left\|\widehat{F}_S(W_t^S)\right\|^2\right]dt + \frac{\log(3/\zeta)+\Lambda}{n}}$$

with $\Lambda = \mathrm{KL}\left(\rho_0||\pi\right)$, and the noise mixing constant $M(\sigma_1,\sigma_2,d)$ is given by:

$$M(\sigma_1,\sigma_2,d,\alpha) := \frac{1}{4\sigma^2 + \frac{\sigma_1^\alpha \alpha 2^\alpha \Gamma\left(\frac{d+\alpha}{2}\right)R^{2-\alpha}}{(2-\alpha)\Gamma\left(1-\frac{\alpha}{2}\right)d\Gamma\left(\frac{d}{2}\right)}}.$$

Proof. The proof follows by the exact same computations than the proofs of Corollary 4.2 and Theorem 4.4. □

Based on the above theorem, we can make an interesting observation. The constant $M(\sigma_1,\sigma_2,d,\alpha)$ clearly shows the relative contribution of both noises. Therefore, we can argue that they have the same contribution when the following condition is fulfilled:

$$4\sigma_2^2 = \frac{\sigma_1^\alpha \alpha 2^\alpha \Gamma\left(\frac{d+\alpha}{2}\right)R^{2-\alpha}}{(2-\alpha)\Gamma\left(1-\frac{\alpha}{2}\right)d\Gamma\left(\frac{d}{2}\right)}.$$

Using the asymptotic developments discussed in Section 5, and proven in Appendix D, in the limit $d \to \infty$, we can write this condition as:

$$4\sigma_2^2 \underset{d\to\infty}{\sim} \frac{1}{R^{2-\alpha}} P_\alpha \frac{\sigma_1^\alpha d^{\frac{\alpha}{2}}}{d},$$

where the pre-factor P_α has been defined and studied in Section 5. It is now meaningful to write it as:

$$(\sigma_2\sqrt{d_0})^2 = \frac{1}{4P_\alpha}(\sigma_1\sqrt{d_0})^\alpha, \tag{36}$$

with $d_0 := d/(R^2)$, the "reduced dimension introduced in Section 5. This is the equal noise contribution condition. What is particularly noticeable in Eq. (36) is that, as was the case in the study of the phase transition, in Section 5, the meaningful quantities, for both noises, is **(scale of the noise)** $\times$ $\sqrt{\textbf{dimension}}$.

C.8. Time-uniform bounds deduced from a generalized Poincaré inequality - Omitted proofs of Section 4.3

In this section, we fix the convex function Φ to be $\Phi(x) = \frac{1}{2}x^2$. In that case, the Bregman divergence becomes symmetric and satisfies:

$$D_\Phi(a,b) = D_\Phi(b,a) = \frac{1}{2}(a-b)^2.$$

Combined with the logarithmic Sobolev inequality, given by Theorem A.19, this allows us to prove a time uniform bound on the chi-squared distance between the prior and the posterior.

We denote, as above:

$$V_S(w) := \widehat{F}_S(w) + \frac{\eta}{2}\|w\|^2, \quad \text{and: } V(w) := \frac{\eta}{2}\|w\|^2,$$

with $\eta > 0$ the regularization coefficient.

Before bounding the Φ-entropy, we first need the following lemma. It is close to remarks already made in (Gentil & Imbert, 2008) and (Tristani, 2013), for instance.

Lemma C.16. *In this setting, $\bar{u}_\infty$ is the density of an infinitely divisible probability distribution, whose characteristic exponent is given by:*

$$\forall \xi \in \mathbb{R}^d,\ \psi(\xi) = \frac{\sigma_2^2}{2\eta} \|\xi\|^2 + \frac{\sigma_1^\alpha}{\alpha\eta} \|\xi\|^\alpha .$$

Proof. Let us first recall some notations. π is the (prior) density distribution whose density with respect to the Lebesgue measure is $\bar{u}_\infty$. The existence of such the steady state $\bar{u}_\infty$ was obtained in (Gentil & Imbert, 2008; Tristani, 2013). Moreover, Gentil & Imbert (2008) showed that π is infinitely divisible, therefore, it makes sense to introduce its characteristic exponent, denoted ψ, and defined by:

$$\forall \xi \in \mathbb{R}^d,\ \mathbb{E}_{X\sim\pi}\left[e^{i\xi\cdot X}\right] = e^{-\psi(\xi)}.$$

Following computations of Gentil & Imbert (2008); Tristani (2013). We quickly show how the expression of the characteristic exponent can easily be obtained in our particular case. First, because of the symmetry of Eq. (7), both $\bar{u}_\infty$ and ψ must be symmetric (i.e. even functions). Therefore, if we define the Fourier transform by:

$$\mathcal{F}(u)(\xi) = \int_{\mathbb{R}^d} e^{-iw\cdot\xi} u(w) dw,$$

then we have:

$$\mathcal{F}(\bar{u}_\infty)(\xi) = e^{-\psi(x)}.$$

We can take, at least formally (Tristani, 2013), the Fourier transform of Eq. (7) and write that, in the sense of distributions:

$$\mathcal{F}(\bar{u}_\infty)(\xi)\left(-\sigma_1^\alpha \|\xi\|^\alpha - \sigma_2^2 \|\xi\|^2\right) + \eta\mathcal{F}(\nabla\cdot(w\bar{u}_\infty(w)))(\xi) = 0.$$

In the above, we used the expression of Fourier transform for the Laplacian and the fractional Laplacian. Using properties of the Fourier transform, this leads to:

$$\mathcal{F}(\bar{u}_\infty)(\xi)\left(-\sigma_1^\alpha \|\xi\|^\alpha - \sigma_2^2 \|\xi\|^2\right) - \eta\nabla\mathcal{F}(\bar{u}_\infty)(\xi).$$

Therefore:

$$\eta\xi\cdot\nabla\psi(\xi) = \sigma_1^\alpha \|\xi\|^\alpha + \sigma_2^2 \|\xi\|^2 .$$

Given that ψ is symmetric, we can write it as $P(\|\xi\|^2)$, we can then rewrite the previous equation as (for $a > 0$):

$$P'(a) = \frac{\sigma_1^\alpha a^{\alpha/2-1} + \sigma_2^2}{2\eta}.$$

Integrating, and noting that we have $P(0) = 0$, leads to the result. □

In this setting, the Φ-entropy inequality, i.e. Theorem A.19, becomes the following statement.

Corollary C.17 (Generalized Poincaré's inequality)**.** *Let μ be an infinitely divisible law on $\mathbb{R}^d$, with associated triplet denoted (b, Q, ν), then, for every smooth enough function v, we have:*

$$\frac{1}{2}\int v^2 d\mu - \frac{1}{2}\left(\int v d\mu\right)^2 \le \int \nabla v\cdot Q\nabla v d\mu + \iint D_\Phi(v(x), v(x+z)) d\nu(z) d\mu(x).$$

where we used the symmetry of the Bregman divergence, for $\Phi(x) = \frac{1}{2}x^2$, to revert the arguments of this divergence, appearing in Theorem A.19.

If v is chosen to be the Radon-Nykodym derivative of another measure ν, with respect to μ, we recognize the chi-squared distance, defined by:

$$\chi^2\left(\nu|\,|\mu\right) := \int \left(\frac{d\nu}{d\mu}\right)^2 d\mu - 1. \tag{37}$$

We also remind the definition of Renyi entropies, for $\beta > 1$:

$$\mathrm{D}_\alpha\left(\nu||\mu\right) = \frac{1}{\beta-1}\log \int \left(\frac{d\nu}{d\mu}\right)^\beta d\mu. \tag{38}$$

Note that, by convention, we also set $\mathrm{D}_1\left(\nu||\mu\right) = \mathrm{KL}\left(\nu||\mu\right)$, see (van Erven & Harremoës, 2014) for an extensive review of the properties of those entropies.

The following inequalities are clear, the first one being proven in (van Erven & Harremoës, 2014), it is a direct consequence of Jensen's inequality:

$$\mathrm{KL}\left(\nu||\mu\right) \leq \mathrm{D}_2\left(\nu||\mu\right) = \log(\chi^2\left(\nu|\,|\mu\right)+1) \leq \chi^2\left(\nu|\,|\mu\right). \tag{39}$$

C.8.1. WARMU -UP: THE CASE $\sigma_2 > 0$

In order to mimic the resoning of Sections 4.1 and 4.2, we first handle the case where $\sigma_2 > 0$. This is an additional theoretical result.

Theorem C.18. *We make Assumptions 3.3 and 3.4, accordingly to this choice of convex function. Then, with probability at least $1-\zeta$ over $S \sim \mu_z^{\otimes n}$ and $w \sim \rho_T^S$, we have:*

$$G_S(w) \leq 2s\sqrt{\frac{2}{n\sigma_2^2}\bar{I}(T,S) + \frac{4e^{-\eta T}\Lambda + \log\frac{24}{\zeta^3}}{n}},$$

with $\Lambda = \mathrm{Ent}^\Phi_{\bar{u}_\infty}\left(\rho_0\right)$, and:

$$\bar{I}(T,S) := \int_0^T e^{-\eta(T-t)}\mathbb{E}_\pi\left[(v_t^S)^2\left\|\nabla\widehat{F}_S\right\|^2\right]dt.$$

Proof. As before, we fix t and S and ease the notations by denoting u instead of u_t^S, and similarly for v. For our particular choice of function Φ,by Lemma 4.1, the entropy flow is equal to:

$$\frac{d}{dt}\mathrm{Ent}^\Phi_{\bar{u}_\infty}\left(v\right) = -\sigma_2^2\int\|\nabla v\|^2\bar{u}_\infty - C_{\alpha,d}\sigma_1^\alpha\iint D_\Phi(v(x),v(x+z))\bar{u}_\infty d\nu_\alpha(z)dx + \int v\bar{u}_\infty\nabla v\cdot\nabla\widehat{F}_S.$$

By Young's inequality, we have, for any $C>0$ and $\gamma\in(0,1]$:

$$\frac{d}{dt}\mathrm{Ent}^\Phi_{\bar{u}_\infty}\left(v\right) \leq -\left(\sigma_2^2 - \frac{C}{2}\right)\int\|\nabla v\|^2\bar{u}_\infty - \gamma C_{\alpha,d}\sigma_1^\alpha\iint D_\Phi(v(x),v(x+z))\bar{u}_\infty d\nu_\alpha(z)dx + \frac{1}{2C}\int v^2\left\|\nabla\widehat{F}_S\right\|^2\bar{u}_\infty.$$

Thanks to Lemma C.16, we know that the Levy triplet of $\bar{u}_\infty$ is given by $(0,\sigma_2^2/(\eta),(C_{\alpha,d}\sigma_1^\alpha/(\alpha\eta))\nu_\alpha)$. Therefore, the Poincaré's inequality of Corollary C.17 is given by:

$$\mathrm{Ent}^\Phi_{\bar{u}_\infty}\left(v\right) \leq \frac{\sigma_2^2}{2\eta}\int\|\nabla v\|^2\bar{u}_\infty + \frac{C_{\alpha,d}\sigma_1^\alpha}{\alpha\eta}\iint D_\Phi(v(x),v(x+z))\bar{u}_\infty d\nu_\alpha(z)dx.$$

Therefore, we have:

$$\frac{d}{dt}\mathrm{Ent}^\Phi_{\bar{u}_\infty}\left(v\right) \leq -\gamma\alpha\eta\mathrm{Ent}^\Phi_{\bar{u}_\infty}\left(v\right) - \left(\sigma_2^2 - \frac{C}{2} - \frac{\alpha\gamma\sigma_2^2}{2}\right)\int\|\nabla v\|^2\bar{u}_\infty + \frac{1}{2C}\int v^2\left\|\nabla\widehat{F}_S\right\|^2\bar{u}_\infty. \tag{40}$$

$$\tag{41}$$

We make the choice $\gamma = 1/\alpha$ and $C = \sigma_2^2$, this gives:

$$\frac{d}{dt}\text{Ent}^{\Phi}_{\bar{u}_\infty}(v) \leq -\eta \text{Ent}^{\Phi}_{\bar{u}_\infty}(v) + \frac{1}{2\sigma_2^2}\int v^2 \left\|\nabla \widehat{F}_S\right\|^2 \bar{u}_\infty.$$

Solving this differential inequality implies that, with $\Lambda := \text{Ent}^{\Phi}_{\bar{u}_\infty}(\rho_0)$ and $T > 0$,

$$\text{Ent}^{\Phi}_{\bar{u}_\infty}\left(v_T^S\right) \leq e^{-\eta T}\Lambda + \frac{1}{2\sigma_2^2}\bar{I}(T,S),$$

with:

$$\bar{I}(T,S) := \int_0^T e^{-\eta(T-t)}\mathbb{E}_\pi\left[(v_t^S)^2\left\|\nabla \widehat{F}_S\right\|^2\right]dt.$$

From Equation (37), this implies that:

$$\frac{1}{2}\chi^2\left(\rho_T^S\middle|\,\middle|\pi\right) \leq e^{-\eta T}\Lambda + \frac{1}{2\sigma_2^2}\int_0^T e^{-\eta(T-t)}I(t,S)dt.$$

We conclude by applying Theorem B.3 and noting that, in our notations, we have:

$$\text{D}_2\left(\rho_T^S||\pi\right) \leq 2\chi^2\left(\rho_S^t\middle|\,\middle|\pi\right) = 2\text{Ent}^{\Phi}_{\bar{u}_\infty}(v).$$

□

In order to make the above theorem more interpretable, we present, as an additional theoretical result, the following corollary. It states that, if the loss is Lipschitz in w and if the regularization coefficient is big enough, then we have a simpler time-uniform bound. The main interest for the following result is that it it fully comparable with our other results. This is an additional result, it was not presented in the main part of the document.

Corollary C.19. *We make the same assumptions than in Theorem C.18 and further assume that the loss $f(w,z)$ is L_f-Lipschitz in w, uniformly with respect to z. Then, if*

$$a := \eta - \frac{L_f^2}{2\sigma_2^2} > 0, \tag{42}$$

then, with probability at least $1-\zeta$ over $S \sim \mu_z^{\otimes n}$ and $w \sim \rho_T^S$, we have, for any $\lambda > 0$:

$$G_S(w) \leq \frac{2s}{\sqrt{n}}\left\{\frac{L_f^2}{2\sigma_2^2}\frac{1-e^{-aT}}{a} + \frac{3}{2}\log\left(\frac{2}{\zeta}\right) + \frac{\Lambda}{e^{\eta T}}\right\}^{\frac{1}{2}}.$$

with $\Lambda = \text{Ent}^{\Phi}_{\bar{u}_\infty}(\rho_0)$.

Proof. We follow the exact same step than the proof of Theorem C.18, up to Equation (40). We make the same choices for γ and C and obtain:

$$\frac{d}{dt}\text{Ent}^{\Phi}_{\bar{u}_\infty}(v) \leq -\eta \text{Ent}^{\Phi}_{\bar{u}_\infty}(v) + \frac{1}{2\sigma_2^2}\int v^2 \left\|\nabla \widehat{F}_S\right\|^2 \bar{u}_\infty.$$

By the Lipschitz assumption, this is:

$$\begin{aligned}\frac{d}{dt}\text{Ent}^{\Phi}_{\bar{u}_\infty}(v) &\leq -\eta \text{Ent}^{\Phi}_{\bar{u}_\infty}(v) + \frac{L_f^2}{2\sigma_2^2}\int v^2 \bar{u}_\infty. \\ &= -\left(\eta - \frac{L_f^2}{2\sigma_2^2}\right)\text{Ent}^{\Phi}_{\bar{u}_\infty}(v) + \frac{L_f^2}{2\sigma_2^2}.\end{aligned}$$

The result is deduced by the exact same reasoning as the last steps of the proof of Theorem C.18, by applying Theorem B.3. □

C.8.2. TIME UNIFORM BOUNDS IN THE CASE $\sigma_2 = 0$ - PROOF OF THEOREM 4.5

In the previous subsection, we used heavily the presence of a non-trivial Brownian part in the operator I. However, we face the same problem that we had with Theorem C.18: while providing a generalization bound for multifractal dynamics, which is our main goal, this theorem is not very informative about the impact of the tail-index α on the generalization performance. This is why, in this section, we focus, like in Section 4.2, on the case $\sigma_2 = 0$. This follows the steps presented in Section 4.2, with a change in the choice of the convex function Φ.

To overcome this issue, we show how it is possible to derive a bound, even without the Brownian part, by using a reasoning similar to Section C.3. Under Assumption 3.3, we can apply the same reasoning with the choice of function: $\Phi(x) = x^2/2$.

With this change in the convex function Φ, the relevant quantity is no more the Fisher information, but the following L^2-norm on the gradient of v, this is the term that appears in Poincaré's inequality:

$$\mathbf{G}_v := \int \|\nabla v\|^2 \, \bar{u}_\infty dx. \tag{43}$$

This quantity, $\mathbf{G}_v$, correspond to the Φ-information term, $I_\Phi(v)$, introduced in Lemma 4.1, for the particular choice of convex function that we make in this section, namely $\Phi(x) = x^2/2$

The following lemma is proven by the exact same lines than Lemma C.3, only the function Φ, and the corresponding Φ-regularity assumption, changes.

Lemma C.20. *Assume that v is Φ-regular, for $\Phi(x) = x^2/2$. Then, we have the following representation:*

$$\iint D_\Phi(v(x), v(x+z)) d\nu_\alpha(z) \bar{u}_\infty(x) dx = \frac{\sigma_{d-1}}{2d} \int_0^\infty G_v(r) \frac{dr}{r^{\alpha-1}},$$

where the function G_v is non-negative and defined by:

$$G_v(r) := \frac{2d}{r^2 \sigma_{d-1}} \int_0^r \int_s^r \int_{\mathbb{R}^d} \int_{\mathbb{S}^{d-1}} \theta \cdot \nabla v(x + s\theta) \theta \cdot \nabla v(x + u\theta) \bar{u}_\infty(x) d\theta dx du ds.$$

This function can be continuously extended at 0 by:

$$G_v(0) = \mathbf{G}_v.$$

This function, $G_v : [0, \infty) \longrightarrow [0, \infty)$, corresponds to the function $J_{\Phi,v}$, introduced in the main part of the paper, for the particular choice of convex function $\Phi(x) = \|x\|^2/2$.

The following assumption is the equivalent of Assumption 4.3 for the function G_v, instead of J_v.

Assumption C.21. There exists an absolute constant $R > 0$ such that, for $\mu_z^{\otimes n}$-almost all S and all $t \geq 0$, we have:

$$\forall r \in [0, R], \; G_v(r) \geq \frac{1}{2} G_v(0).$$

This assumption is exactly a reformulation of Assumption 4.3, rewritten with $\Phi(x) = \frac{1}{2}x^2$. Using the denominations of Section 4.3, it is Assumption 4.3-Φ_2.

We can now prove Theorem 4.5, which is a time-uniform generalization bound obtained in the pure heavy-tailed case ($\sigma_1 = 0$).

Theorem 4.5. *Let $\sigma_2 = 0$. We make Assumptions 4.3$-\Phi_2$, 3.3 and 3.4, with the choice $\Phi = \Phi_2$. Then, with probability at least $1 - \zeta$ over $S \sim \mu_z^{\otimes n}$ and $w \sim \rho_T^S$, we have*

$$G_S(w) \leq 2s \sqrt{\frac{4K_{\alpha,d}}{n\sigma_1^\alpha} \tilde{I}(T, S) + \frac{2e^{-\frac{\alpha\eta T}{2}} \Lambda + \log \frac{24}{\zeta^3}}{n}},$$

with $\Lambda = \mathrm{Ent}^\Phi_{\bar{u}_\infty}(\rho_0)$, and

$$\tilde{I}(T, S) := \int_0^T e^{-\frac{\alpha\eta}{2}(T-t)} \mathbb{E}_\pi \left[(v_t^S)^2 \left\| \nabla \widehat{F}_S \right\|^2 \right] dt.$$

Proof. We start, again, with the expression of the entropy flow.

$$\frac{d}{dt}\text{Ent}^{\Phi}_{\bar{u}_\infty}(v) = -C_{\alpha,d}\sigma_1^\alpha \iint D_\Phi(v(x), v(x+z))\bar{u}_\infty d\nu_\alpha(z)dx + \int v\bar{u}_\infty \nabla v \cdot \nabla \widehat{F}_S.$$

We split in two the Bregman integral term and apply the generalized Poincaré inéquality, along with Young's inequality on the last term, for any $C > 0$:

$$\begin{aligned}\frac{d}{dt}\text{Ent}^{\Phi}_{\bar{u}_\infty}(v) &\leq -\frac{\alpha\eta}{2}\text{Ent}^{\Phi}_{\bar{u}_\infty}(v) - \frac{C_{\alpha,d}\sigma_1^\alpha}{2}\iint D_\Phi(v(x), v(x+z))\bar{u}_\infty d\nu_\alpha(z)dx + \int v\bar{u}_\infty \nabla v \cdot \nabla \widehat{F}_S. \\ &\leq -\frac{\alpha\eta}{2}\text{Ent}^{\Phi}_{\bar{u}_\infty}(v) - \frac{C_{\alpha,d}\sigma_1^\alpha}{2}\iint D_\Phi(v(x), v(x+z))\bar{u}_\infty d\nu_\alpha(z)dx + \frac{C}{2}\mathbf{G}_v + \frac{1}{2C}\int v^2 \left\|\nabla \widehat{F}_S\right\|^2 \bar{u}_\infty dx.\end{aligned}$$

We now use Lemma C.20 and Assumption C.21, this gives:

$$\begin{aligned}\frac{d}{dt}\text{Ent}^{\Phi}_{\bar{u}_\infty}(v) &\leq -\frac{\alpha\eta}{2}\text{Ent}^{\Phi}_{\bar{u}_\infty}(v) - \frac{C_{\alpha,d}\sigma_1^\alpha}{2}\frac{\sigma_{d-1}}{2d}\int_0^R G_v(r)\frac{dr}{r^{\alpha-1}} + \frac{C}{2}\mathbf{G}_v + \frac{1}{2C}\int v^2 \left\|\nabla \widehat{F}_S\right\|^2 \bar{u}_\infty dx \\ &\leq -\frac{\alpha\eta}{2}\text{Ent}^{\Phi}_{\bar{u}_\infty}(v) - \frac{C_{\alpha,d}\sigma_1^\alpha\sigma_1^\alpha R^{2-\alpha}}{8d(2-\alpha)}\mathbf{G}_v + \frac{C}{2}\mathbf{G}_v + \frac{1}{2C}\int v^2 \left\|\nabla \widehat{F}_S\right\|^2 \bar{u}_\infty dx.\end{aligned}$$

Therefore, we make the choice:

$$C := \frac{C_{\alpha,d}\sigma_1^\alpha\sigma_1^\alpha R^{2-\alpha}}{4d(2-\alpha)},$$

and we get:

$$\frac{d}{dt}\text{Ent}^{\Phi}_{\bar{u}_\infty}(v) \leq -\frac{\alpha\eta}{2}\text{Ent}^{\Phi}_{\bar{u}_\infty}(v) + \frac{2d(2-\alpha)}{C_{\alpha,d}\sigma_1^\alpha\sigma_1^\alpha R^{2-\alpha}}\int v^2 \left\|\nabla \widehat{F}_S\right\|^2 \bar{u}_\infty dx.$$

We introduce the same constant as the one introduced in Theorem 4.4, namely:

$$K_{\alpha,d} = \frac{d(2-\alpha)}{C_{\alpha,d}\sigma_{d-1}R^{2-\alpha}}.$$

Therefore, solving the above differential inequality gives:

$$\text{Ent}^{\Phi}_{\bar{u}_\infty}(v) \leq e^{-\frac{\alpha\eta T}{2}}\Lambda + \frac{2K_{\alpha,d}}{\sigma_1^\alpha}\bar{I}(T,S),$$

where $\bar{I}(T,S)$ and Λ are defined in the same way than in Theorem C.18.

We conclude by applying Theorem B.3 and noting that $\text{D}_2\left(\rho_S^T||\pi\right) \leq 2\text{Ent}^{\Phi}_{\bar{u}_\infty}\left(v_T^S\right)$.

□

D. Asymptotics of the constants - Omitted proofs of Section 5

This section is devoting to studying the limit behavior of the constants introduced in our generalization bounds, in terms of α and d. This corresponds to the proof of the theoretical results of Section 5.

We first study the asymptotic behavior of the constant when $d \to \infty$.

Lemma 5.1. *We have that, for all $\alpha \in (1,2)$:*

$$\bar{K}_{\alpha,d} \underset{d\to\infty}{\sim} P_\alpha d^{1-\frac{\alpha}{2}}, \quad P_\alpha := \frac{(2-\alpha)\Gamma\left(1-\frac{\alpha}{2}\right)}{\alpha 2^{\alpha/2}}, \tag{14}$$

Proof. By the definitions of Sections 4.2 and 5, we have:

$$\bar{K}_{\alpha,d} = \frac{(2-\alpha)\Gamma\left(1-\frac{\alpha}{2}\right)}{\alpha 2^{\alpha}} \frac{d\Gamma\left(\frac{d}{2}\right)}{\Gamma\left(\frac{d+\alpha}{2}\right)}.$$

By Lemma A.21, we have:

$$\begin{aligned}\frac{d\Gamma\left(\frac{d}{2}\right)}{\Gamma\left(\frac{d+\alpha}{2}\right)} &\underset{d\to\infty}{\sim} \frac{d\Gamma\left(\frac{d}{2}\right)}{\Gamma\left(\frac{d}{2}\right)\left(\frac{d}{2}\right)^{\alpha/2}} \\ &= 2^{\alpha/2} d^{1-\frac{\alpha}{2}},\end{aligned}$$

from which we immediately deduce the result. □

We now prove Lemma 5.2, which study the low-tail limit of our generalization bounds for heavy-tailed dynamics.

Lemma 5.2. *For any $d \geq 1$, we have $K_{\alpha,d} \underset{\alpha\to 2^-}{\longrightarrow} \frac{1}{2}$.*

Proof. By the definitions of Sections 4.2 and 5, we have:

$$\bar{K}_{\alpha,d} = \frac{(2-\alpha)\Gamma\left(1-\frac{\alpha}{2}\right)}{\alpha 2^{\alpha}} \frac{d\Gamma\left(\frac{d}{2}\right)}{\Gamma\left(\frac{d+\alpha}{2}\right)}.$$

We have $\alpha \in [1,2)$, therefore, we know that $\frac{\alpha}{2} \notin \mathbb{Z}$. We can then apply Euler's reflection formula, Lemma A.20, and get that:

$$\begin{aligned}\Gamma\left(1-\frac{\alpha}{2}\right)\Gamma\left(\frac{\alpha}{2}\right) &= \frac{\pi}{\sin\left(\frac{\pi\alpha}{2}\right)} \\ &= \frac{\pi}{\sin\left(\pi-\frac{\pi\alpha}{2}\right)} \\ &\underset{\alpha\to 2^-}{\sim} \frac{1}{1-\frac{\alpha}{2}}.\end{aligned}$$

On the other hand, by continuity of the Γ function and using the identity $\Gamma(x+1) = x\Gamma(x)$, we also have:

$$\Gamma\left(\frac{d+\alpha}{2}\right) \underset{\alpha\to 2^-}{\sim} \Gamma\left(\frac{d}{2}+1\right) = \frac{d}{2}\Gamma\left(\frac{d}{2}\right).$$

Putting everything together, and using that $\Gamma(1) = 1$, we get that:

$$\bar{K}_{\alpha,d} \underset{\alpha\to 2^-}{\sim} \frac{1}{2}.$$

As $K_{\alpha,d} = R^{\alpha-2}\bar{K}_{\alpha,d}$, with $R > 0$, this implies the result. □

Finally, the next lemma presents the main properties of the function $\alpha \longmapsto P_\alpha$, introduced in Section 5. As a reminder, we have:

$$P_\alpha := \frac{(2-\alpha)\Gamma\left(1-\frac{\alpha}{2}\right)}{\alpha 2^{\alpha/2}}$$

Lemma D.1. *The function $\alpha \longmapsto P_\alpha$ is, on the interval $[1,2)$, continuous, decreasing and satisfies:*

$$\forall \alpha \in [1,2), \ \frac{1}{2} \leq P_\alpha \leq \sqrt{\frac{\pi}{2}}.$$

Proof. The continuity follows from the continuity of the Γ function on $(0, \frac{1}{2}]$. Moreover, we have, using the particular values of the Γ function recalled in A.4:

$$P_1 = \frac{\Gamma\left(\frac{1}{2}\right)}{\sqrt{2}} = \frac{\sqrt{\pi}}{\sqrt{2}}.$$

Using the proof of the Lemma 5.2, we also have:

$$P_\alpha \underset{\alpha \to 2^-}{\longrightarrow} \frac{1}{2}.$$

Now we fix $\alpha \in (1, 2)$, we use Euler's reflection formula, Lemma A.20, and the formula $\Gamma(1+z) = z\Gamma(z)$ to get:

$$\begin{aligned} P_\alpha &= \frac{(2-\alpha)\Gamma\left(1-\frac{\alpha}{2}\right)}{\alpha 2^{\alpha/2}} \\ &= \frac{2-\alpha}{\alpha 2^{\alpha/2}} \frac{\pi}{\sin\left(\frac{\pi\alpha}{2}\right)\Gamma\left(\frac{\alpha}{2}\right)} \\ &= \frac{\pi\left(1-\frac{\alpha}{2}\right)}{\sin\left(\frac{\pi\alpha}{2}\right)} \frac{1}{2^{\alpha/2}\Gamma\left(1+\frac{\alpha}{2}\right)} \end{aligned}$$

From the general properties of the Γ function, we know that $\alpha \longmapsto \Gamma\left(1+\frac{\alpha}{2}\right)$ is increasing on $(1, 2)$, and therefore it is enough to show that the function

$$g(\alpha) = \frac{2-\alpha}{\sin\left(\frac{\pi\alpha}{2}\right)},$$

is decreasing, as both $\Gamma\left(1+\frac{\alpha}{2}\right)$ and g are positive on $(1, 2)$. A quick calculation reveals that:

$$\sin^2\left(\frac{\pi\alpha}{2}\right) g'(\alpha) = -\cos\left(\frac{\pi\alpha}{2}\right)\left\{\tan\left(\pi\left(\frac{\alpha}{2}-1\right)\right) - \pi\left(1-\frac{\alpha}{2}\right)\right\},$$

which implies the desired result, by noting that $\tan(x) \leq x$ on $(-\pi/2, 0)$.

□

The function $\alpha \longmapsto P_\alpha$ is represented on Fig. 5. This plot confirms the calculations of Lemma D.1.

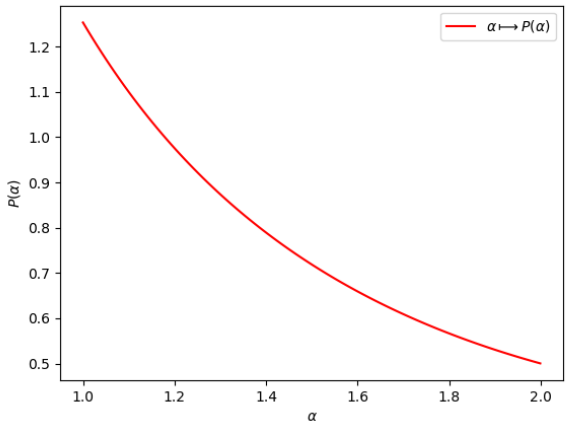


Figure 5. Graphical representation of the function $\alpha \longmapsto P_\alpha$. The leftmost value is $P_1 = \sqrt{\frac{\pi}{2}} \simeq 1.2533$ and the right limit is $P_{2^-} = \frac{1}{2}$.

D.1. Precisions on the low noise-regime

In Section 5, we concluded the existence of two regimes predicted by our bounds. We first defined the low noise regime by the following condition:

$$\frac{\sigma_1\sqrt{d}}{R} < 1.$$

However, the pre-factor $\alpha \longmapsto P_\alpha$ is decreasing. Therefore, we need to take it into account to accurately describe the regime where the generalization error is increasing with α. This happens when we have:

$$\frac{P_1 R}{\sigma_1\sqrt{d}} < \frac{P_2 R^2}{d\sigma_1^2}.$$

Using Lemma D.1, we get the condition that was mentioned briefly in Section 5:

$$\frac{\sigma_1\sqrt{d}}{R} < \frac{1}{\sqrt{2\pi}}.$$

D.2. Comparison with the constants appearing in other works

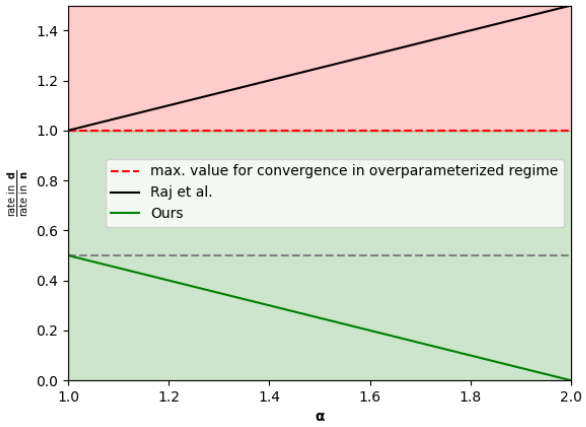


Figure 6. Comparison of the dimension dependence of our bounds and that of Raj et al. (2023b), under the assumption that ℓ is Lipschitz continuous, for both bounds.

In(Raj et al., 2023b), the authors prove expected generalization bounds for heavy-tail dynamics. The constant appearing in their bounds has a complex dependence on various constants defined in their assumptions, but we can sum up the dependence in α and d as follows (by Applying Euler's reflection formula to their Lemma 7):

$$R_{\alpha,d} = \mathcal{O}\left(1 + \frac{\sqrt{d}\Gamma\left(\frac{\alpha+d}{2}\right)}{(2-\alpha)\Gamma\left(1-\frac{\alpha}{2}\right)\Gamma\left(\frac{d}{2}\right)}\right)$$

By the proof of Lemma 5.1 and 5.2, we see that:

$$R_{\alpha,d} = \mathcal{O}\left(d^{\frac{1+\alpha}{2}}\right).$$

It is to be noted that our bound has a rate in n of $1/2$, i.e. our bound is proportional to $1/\sqrt{n}$, while the bound of Raj et al. (2023b) has a rate of 1, i.e. the bound is proportional to $1/n$. If we denote by Ξ the ratio of the rate in d to the rate in n of the bound, we find (under a Lipschitz assumption for both bounds):

$$\Xi^{(\text{Ours})} = 1 - \frac{\alpha}{2}, \quad \Xi^{(\text{Raj et al., 2023b})} = \frac{1+\alpha}{2}. \tag{44}$$

The meaning of the coefficient Ξ is that, in an overparameterized regime where $d > n$, the bound is non-vacuous only if $\Xi < 1$. We graphically represented this coefficient in Fig. 6.

E. Discussion of the time dependence of the bounds

While Theorem 4.4 is the first high probability generalization bound for heavy-tailed dynamics, one may notice the time dependence of the bound, coming from the integral term, i.e.

$$\int_0^T \mathbb{E}_U \left[\widehat{F}_S(W_t) \right] dt. \tag{45}$$

Because of this term, as explained in Section 5, the light-tail limit of this bound leads to an informal bound of the form:

$$G_S(W_t) \underset{\alpha \to 2^-}{\lesssim} \sqrt{\frac{1}{n\sigma^2} \int_0^T \mathbb{E}_U \left[\widehat{F}_S(W_t) \right] dt},$$

which is worse than the existing bounds on Langevin dynamics Mou et al. (2017); Li et al. (2020); Farghly & Rebeschini (2021), which achieve time-independent bounds. More explicitly, Mou et al. (2017) obtained a similar bound, but with an additional exponential time-decay in the integral term, i.e., informally,

$$\int_0^T e^{-a(T-t)} \mathbb{E}_U \left[\widehat{F}_S(W_t) \right] dt,$$

where $a > 0$ is a constant dependent of the problem.

To the best of our knowledge, there is no obvious theoretical argument affirming that the limit $\alpha \to 2^-$ should exactly recover the bounds for Langevin dynamics, while it is quite clear that this limit should not be infinite, as in Raj et al. (2023b). That being said, it is worth discussing why this time dependence happens in our bound, how it could be improved, and, most importantly, why a time independent bound is currently beyond the reach of the theory.

Section 4.3 presents one step in the direction of time-uniform bounds. However, the integral term appearing in Theorem 4.5 is not satisfactory, because of its lack of interpretability, compared to Eq. (45).

On the other hand, a more detailed analysis of the results of Mou et al. (2017); Li et al. (2020) shows that time-independence is a direct consequence of the use of a logarithmic Sobolev inequality (LSI, see Appendix A.3). If we translate these arguments in our setting, such an inequality would need to be satisfied by the prior distribution π. Therefore, it is natural to ask, whether we could apply Theorem A.19 in our proofs.

A similar computation has been proposed by Gentil & Imbert (2008), in their study of the rate of convergence to equilibrium of an equation of the form of Eq. (6). However, we argue that we cannot apply this reasoning in our setting. Let's explain it briefly: in the proof of Theorem 4.4, the crucial term that appears in the computation of the entropy flow is what we called the Bregman integral, it is given by:

$$B_\Phi^\alpha(v) = C_{\alpha,d} \iint D_\Phi(v(x), v(x+z)) \bar{u}_\infty d\nu_\alpha(z) dx.$$

On the other hand, Theorem A.19 uses the following term, if we apply it to π (in a similar way to how it is done in Appendix C.8.1):

$$\iint D_\Phi(v(x+z), v(x)) \bar{u}_\infty d\nu_\alpha(z) dx.$$

Unfortunately, the Bregman divergence is not commutative in general, and, to the best of our knowledge, there is no obvious way of linking the two integrals appearing above. This discussion is the motivation behind Section 4.3, where we applied Theorem A.19 and Lemma 4.1 to the case where the Bregman divergence becomes symmetric, i.e. when $\Phi(x) \propto x^2$.

Therefore, we conclude that we do not have the LSI that we would need to make our bounds time-independent. Let's end this short discussion by writing down the result that we would obtain with such an inequality.

Theorem E.1. *We consider the same setting than in Theorem 4.4, but we additionally assume that the prior π satisfies the following LSI, with reversed Bregman divergence compared to Theorem A.19, for all v smooth enough:*

$$\mathrm{Ent}^\Phi_{\bar{u}_\infty}(v) \leq \frac{C_{\alpha,d}\sigma_1^\alpha}{\alpha\eta} \iint D_\Phi(v(x), v(x+z)) \bar{u}_\infty(x) d\nu_\alpha(z) dx.$$

Then, with probability at least $1-\zeta$ *over* $\mu_z^{\otimes n}$*, we have:*

$$\mathbb{E}_U\left[G_S(W_T^S)\right] \leq \sqrt{\frac{2K_{\alpha,d}}{\sigma_1^\alpha}\int_0^T e^{-\frac{\alpha\eta}{2}(T-t)}\mathbb{E}_U\left[\left\|\nabla \widehat{F}_S(W_t^S)\right\|^2\right]dt + \frac{\Lambda+\log(3/\zeta)}{n}}$$

with $\Lambda = \mathrm{KL}\left(\rho_0||\bar{u}_\infty\right)$ *and:*

$$K_{\alpha,d} = \frac{(2-\alpha)\Gamma\left(1-\frac{\alpha}{2}\right)d\Gamma\left(\frac{d}{2}\right)}{\alpha 2^\alpha \Gamma\left(\frac{d+\alpha}{2}\right)R^{2-\alpha}}, \tag{46}$$

F. Additional experimental details

F.1. Hyperparameters details

In this section, we give a list of the exact hyperparameters used to obtain our main experiments. In general, those hyperparameters were chosen to make our experimental setting as close as possible to our theoretical setting and assumptions.

Linear model trained on MNIST (Figs. 9 and 10) : We use a linear predictor, without bias, trained with a (multivariate cross-entropy loss). All 10 classes of the MNIST dataset were used but we randomly subsample 10% of the training set, to lower the computational cost of the experiments. We simulate Eq. (17) with $T = 5.10^3$, $\gamma = 10^{-2}$, $\eta = 10^{-3}$. The last 2000 iterations were used to estimate the accuracy error, as described in Appendix F.2. During those experiments, we let α vary in $[1.6, 2]$, using a linear scale of 10 values. The parameter σ varies in a logarithmic scale of 10 values, such that $\sigma\sqrt{d} \in [0.5, 40]$, where d is the number of parameters in the model, this is inspired by the analysis of Section 5. This range was chosen to ensure both that the training is stable enough and that the heavy tail of the noise has an impact on the accuracy (i.e. σ not too small).

2-layers FCN trained on MNIST: This experiment was used to produce Figs. 2 to 4. We use a 2-layer neural network with ReLU activation, without bias, trained with a (multivariate cross-entropy loss). All 10 classes of the MNIST dataset were used but we randomly subsample 10% of the training set, to lower the computational cost of the experiments. We simulate Eq. (17) with $T = 10^4$, $\gamma = 10^{-2}$, $\eta = 10^{-3}$. The last 2000 iterations were used to estimate the accuracy error, as described in Appendix F.2. During those experiments, we let α vary in $[1.6, 2]$, using a linear scale of 10 values. The width of the network varies in a linear scale of 10 values, between 40 and 200.

FashionMNIST experiments: The same model was also used on the FashionMNIST dataset (a 10% subsample of it), to obtain Figs. 11(a), 11(b) and 12.

5-layers FCN trained on MNIST: We use a 5-layer neural network with ReLU activation, without bias, trained with a (multivariate cross-entropy loss). 100% of the MNIST dataset were used. We experimented using a batch-size in this case, i.e. Eq. (17) is replaced by:

$$\hat{W}_{k+1}^S = \hat{W}_k^S - \frac{\gamma}{b}\sum_{j=1}^b \nabla f(\hat{W}_k^S, z_{i_j}) - \eta\gamma\hat{W}_k^S + \gamma^{\frac{1}{\alpha}}\sigma_1 L_1^\alpha, \tag{47}$$

where $b \in \mathbb{N}^\star$ is the batch size and $(i_1, \dots, i_b)$ are random indices drawn at each iteration. We simulate Eq. (17) with $T = 10^5$, $\gamma = 10^{-2}$, $\eta = 10^{-3}$. The last 100 iterations were used to estimate the accuracy error, as described in Appendix F.2. We used batch sizes $b \in \{64, 128, 256, 512\}$. During those experiments, we let α vary in $[1.7, 2]$, using a linear scale of 10 values. In order for the training to be stable enough for the experiment to converge, we noted that we had to use smaller values, those values were chosen on a logarithmic scale of 6 values so that $\sigma\sqrt{d} \in [0.1, 3]$. It is probable that only the low noise regime, of Section 5, is then observable. Note that, in this case, i.e. when using mini-batches, we estimated the gradient norms $\left\|\nabla \widehat{F}_S\right\|$ using the batch gradients, appearing in Eq. (47), instead of the whole empirical risk $\widehat{F}_S$. This makes sense as one of the goals of using batches is to reduce the computational cost of the experiments.

F.2. Accuracy evolution and robust mean estimation

Figs. 7 and 8 show the evolution of both the train and test accuracy, during the first 2000 iterations of training, for models and datasets similar to the one used for our main experiments (but with less data to make it easier to visualize). We show

these curves for different values of the tail index α and of the quantity $\sigma\sqrt{d}$, which we argue in Section 5 has a strong impact on the training dynamics.

This allows us to observe two phenomenons, first, at least in these two experiments, there seems to be a transition, around the value $\sigma\sqrt{d} = 1$. Indeed, for $\sigma\sqrt{d} \lesssim 1$, we observe that the tail of the noise has little impact on the evolution of both accuracies and therefore on the generalization error. On the other hand, for higher values of $\sigma\sqrt{d}$, we see the impact of the tail index α. Indeed, the heavier the tail (i.e. the smaller α), the more jumps are observed on both accuracies, which necessarily have an impact on the observed generalization. This seems to follow our theory developed in Section 5.

More importantly, the presence of those jumps makes the estimation of the generalization error, $G_S(W_T^S)$, extremely noisy. To avoid this issue, we use the following two techniques:

1. The generalization error is averaged among the 2000 last iterations (recall that we use $N \geq 5000$ in all our experiments so that the model is close to converence when we estimate the generalization error).
2. As the random jumps introduce values of the generalization error that are completely random and often much higher than the values between the jumps, we avoid them to bias the estimation by removing the 15% upper quantile of the generalization error, in the last 2000 iterations, before computing the mean.

F.3. Experimental procedures details

In this section, we quickly give more details on how the figures of Section 6 were obtained.

F.3.1. Bound estimation - Figs. 3, 9 and 12

To get Figs. 3, 9 and 12, we estimated the computable part of Eq. (15). To do so, the gradient norm of all iterations where averaged, to get an estimate of the integral term:

$$\widehat{I} := \gamma \sum_{k=1}^{N} \left\| \nabla \widehat{F}_S(\hat{W}_k^S) \right\|^2 \simeq \int_0^T \left\| \nabla \widehat{F}_S(W_t^S) \right\|^2 dt, \tag{48}$$

where γ is the learning rate and N the number of iterations. The quantity that is plotted on the y-axis of those figures is:

$$\widehat{G} := \sqrt{\frac{P_\alpha d^{1-\frac{\alpha}{2}}}{n\sigma_1^\alpha R^{2-\alpha}} \widehat{I}}. \tag{49}$$

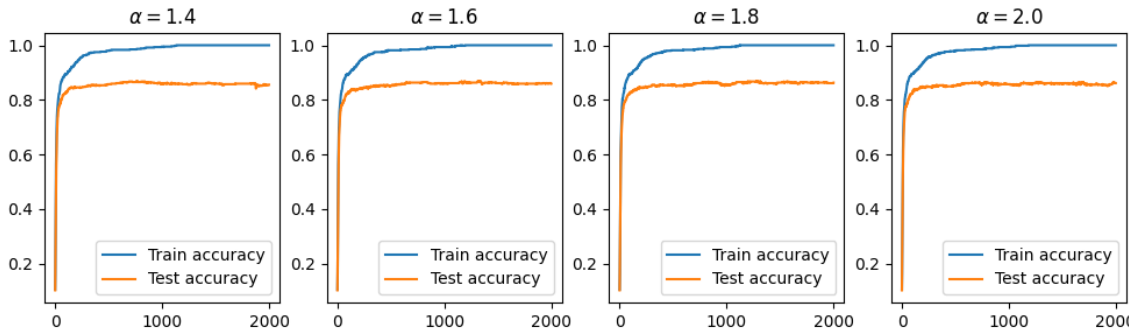
Moreover, this quantity is averaged over 10 random seeds, with the same values of the hyperparameters (σ, d, α). Unless mentioned otherwise, the constant R, which is unknown a priori, is taken to be $R = 1$ in the figures.

F.3.2. Correlation experiments - Figs. 2 and 11(b)

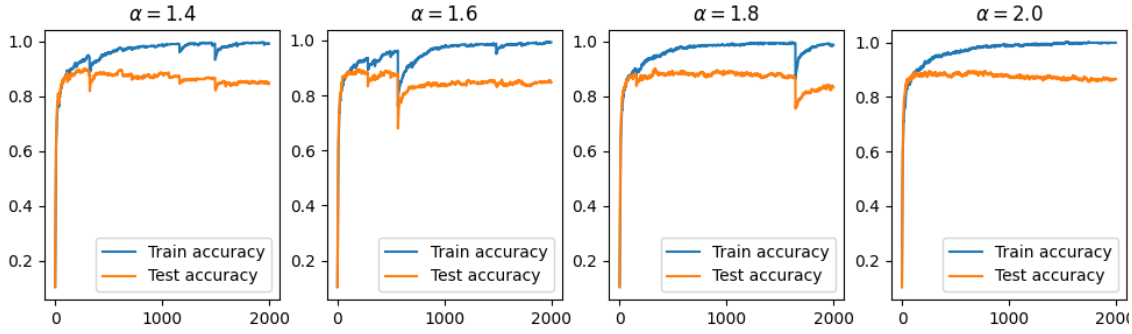
Figs. 2 and 11(b): In this experiment, we train a 2-layers neural network on MNIST. The tail-index α varies in $[1.6, 2]$ and the width of the network in $[40, 200]$. To obtain the green curve, we proceed as follows:

- For each value fixed w of the width, we compute Kendall's correlation coefficient, denoted τ, between the accuracy error and the tail-index α.
- This gives us the correlation in terms of the width, i.e. a map $\tau(w)$.
- We repeat this procedure for 10 random seeds.
- The green curve represents the mean value and standard deviation of $\tau(w)$.

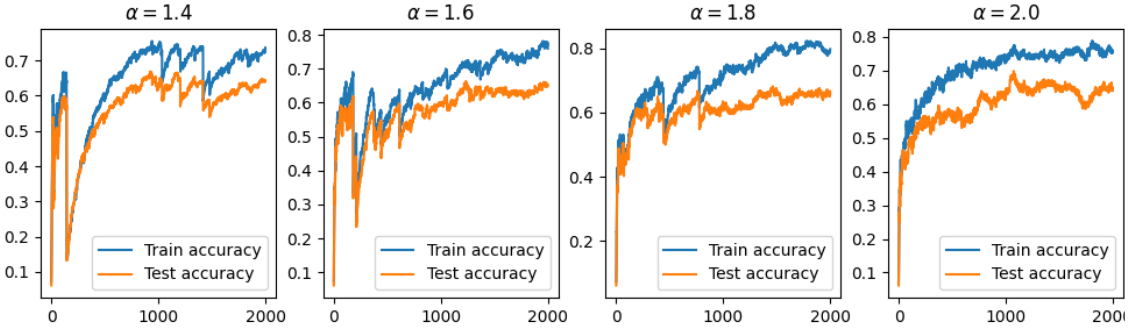
For the dot black curve, to get a less noisy coefficient, we directly used, for each width w, the average of the accuracy error across all the random seeds, and we computed the correlation between α and this mean accuracy error. This is why there are no error bars in that case.



(a) $\sigma\sqrt{d} = 0.1$



(b) $\sigma\sqrt{d} = 1$



(c) $\sigma\sqrt{d} = 10$

Figure 7. Evolution of the test and training accuracies, during the first 2000 iterations, for the simulation of (3) with a linear model trained with a cross-entropy loss, on the MNIST dataset.

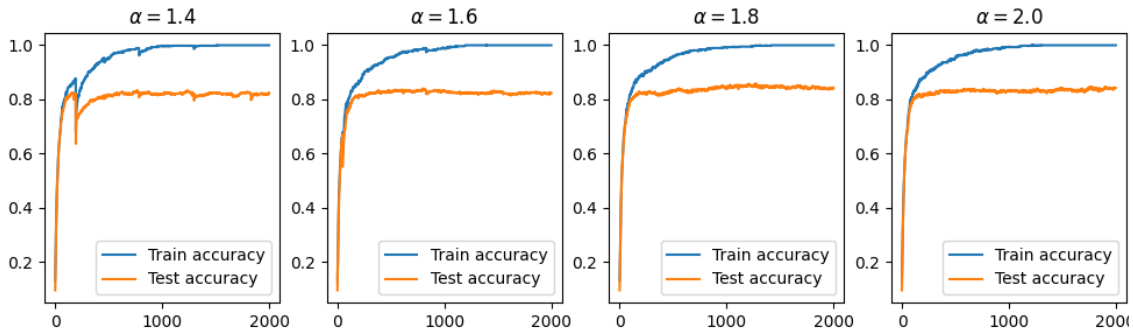
F.4. Additional experiments

In this subsection, we present a few additional experiments to further support the findings of Section 6. These additional experiments are organized as follows:

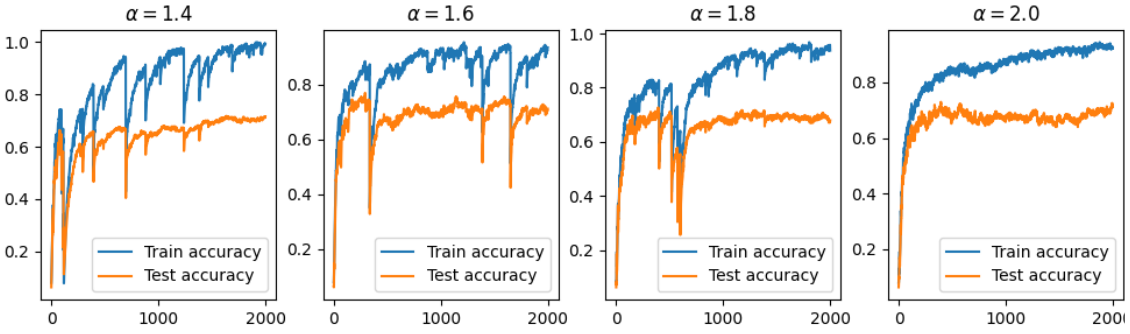
1. Additional experiments using a linear model.

2. Additional experiments conducted on the FashionMNIST dataset (the experiments presented in Section 6 were all conducted on the MNIST dataset).

3. More details figures regarding the use of mini-batches.



(a) $\sigma\sqrt{d} = 1$

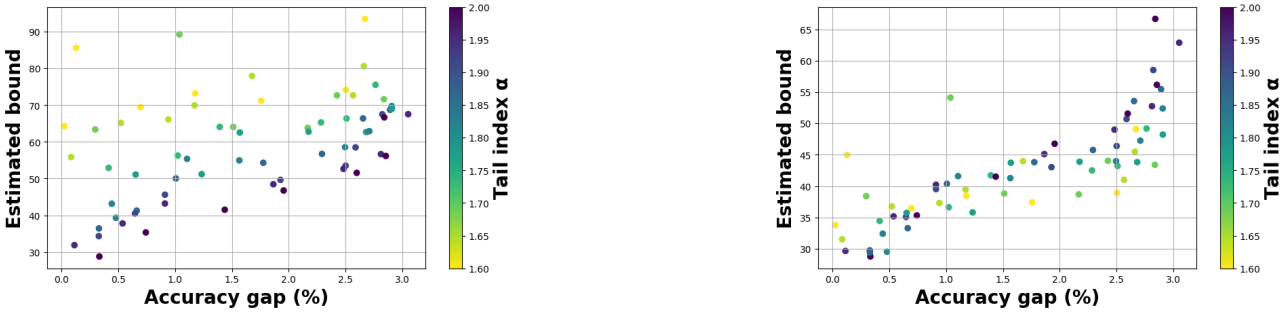


(b) $\sigma\sqrt{d} = 10$

Figure 8. Evolution of the test and training accuracies, during the first 2000 iterations, for the simulation of (3) with a 2-layer neural network, trained with a cross-entropy loss, on the MNIST dataset.

F.4.1. Additional experiment with linear models

We conducted experiments similar to that presented in Section 6 with a linear model, instead of a 2-layers FCN. The results are presented in Figs. 9 and 10. More specifically, in these experiments, we let both the noise scale σ_1 and the tail-index α vary in a fixed grid of values (see Appendix F.1). Note that, in this case, it is not possible to let the number of parameters vary without affecting the data, but we can act on σ_1 instead, and still test our theory. That being said, this model has the advantage, when using the cross-entropy as a surrogate loss, to fit nicely in our theoretical setting.



(a) $R = 1$ (b) $R = 25$

Figure 9. Linear model on MNIST

More specifically, we can see on Fig. 9 that the estimated bound, according to Eq. (49), correlates very well with the generalization error. On Fig. 10, we show how the correlation between α and the accuracy gap varies when σ_1 is varying, see Appendix F.3. The phase transition, predicted in Section 5 is visible in this figure. An additional interesting phenomenon

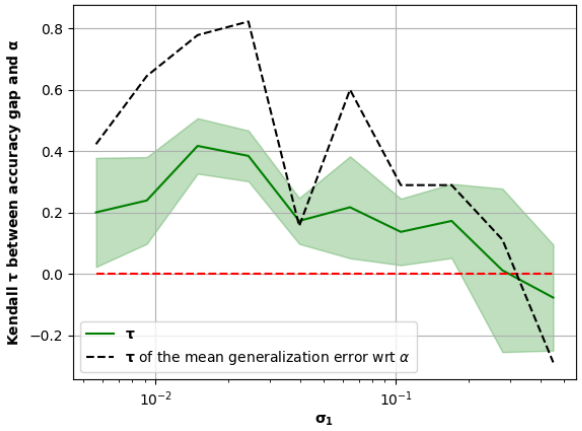


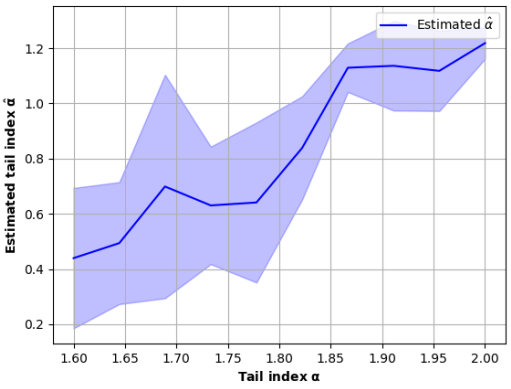
Figure 10. Correlation (Kendall's τ) between the tail index α and the accuracy gap for different values of the noise scale σ_1.

is observed on Fig. 10. Indeed, on the left of the correlation plots (both the black and green curves), we can see that the Kendall's τ coefficient, measuring the correlation between α and G_S, decreases. We interpret this as a third regime, when σ_1 is too small for our theory to hold and the algorithm starts behaving more like a deterministic gradient flow.

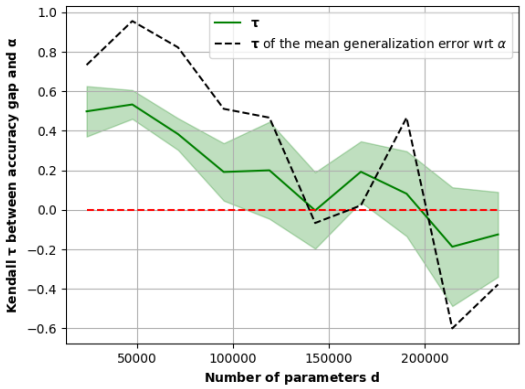
As we did on previous experiments, we can use Fig. 10 and the analysis of Section 5 to estimate the value that the parameter R takes in those experiments. Given that $d = 7840$ in this experiment and that the phase transition seems to be happening around $\sigma_1 \simeq 0.3$, we can estimate $R \simeq \sigma_1\sqrt{d} \simeq 26$. On Fig. 9, we plotted the estimated bound both for $R = 1$ (our default choice) along with the estimated bound for $R = 25$. We observe on that second figure a much improved correlation, hence supporting the experimental results of Section 6.

Remark F.1. With this linear model experiment, it is not possible to let the number of parameters vary (at least not without affecting the data), it is therefore not possible to test the dimension dependence, as in Fig. 4, in this case.

F.4.2. Additional experiments on the FashionMNIST dataset



(a) Regression of the parameter α from the accuracy error, for a 2-layers network trained on 10% of the FashionMNIST dataset.

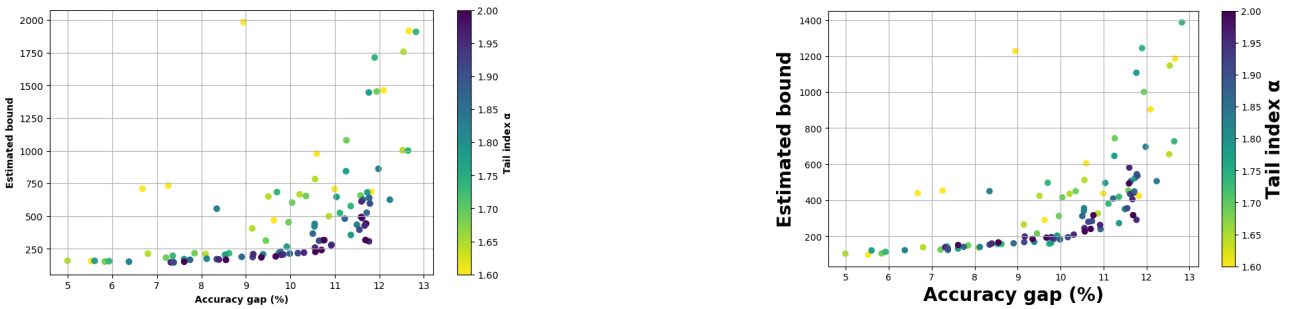


(b) Correlation between α and the accuracy error, for a 2-layers network trained on 10% of the FashionMNIST dataset, with varying width.

Figure 11. Experiments with a 2-layers FCN and the FashionMNIST dataset.

We conducted the same experiments as those presented on Figs. 2 and 4 on the Fashion MNIST dataset. The results are presented in Figs. 11(a) and 11(b). We observe the exact same behavior than the experiments conducted on the

MNIST dataset: the regression of the tail index shows a remarkable correlation with the ground truth α, with the expected monotonicity, even though the value of α is underestimated. Moreover, the phase transition, between the two regimes predicted in Section 5, is observable in Fig. 11(b). It seems to happen at a higher value of d than in the MNIST experiment.



(a) $R = 4$ (b) $R = 11$

Figure 12. FasionMNIST

Additionally, we present in Fig. 12 the value of the estimated bound, based on Eq. (49), compared to the accuracy gap, for different values of th parameter R. For the default choice $R = 1$, we already observe a very good correlation. Based on Fig. 11(b), using the same procedure than in our previous experiments, we can estimate the value of R to be about $R \simeq 11$. As we observe in all our previous experiments, in Section 6 and Appendix F.4.1, we observe that taking into account this value of R improves the observed correlation betzeen the bound and the accuracy gap, hence supporting the theory.

F.4.3. Additional experiments using the CIFAR10 dataset

All our main experiments were conducted on the MNIST and FashionMNIST datasets (10% sub-sample), and with small models (i.e., linear models and 2-layers neural networks). The fact that we focus on these small scale experiments is mainly due to three factors:

1. As mention in Section 6, the computational cost of our experiments is high. They are much more scalable for small models and datasets.
2. It is not clear whether larger DNNs satisfy our theoretical assumptions, further theoretical investigation would be needed in that direction.
3. The Lévy stable noise L_t^α can make the training dynamics highly unstable, as is shown on Fig. 8. We believe that this is mainly due to the fact that the heavy-tailed noise in Eq. (17) does not take into account the model's structure, as in (Wan et al., 2023) for instance. Further investigation would also be needed to better understand this behavior.

Despite these difficulties, we present small preliminary results toward the application of our theory to larger DNNs. More precisely we considered a small convolutional neural network (CNN) with 2 convolutional layers, 3 fully-connected (FC) layers and ReLU activation. We varied the width of these FC layers between 10 and 200 in order to vary the number of parameters d. As a dataset, we use a small sample of the CIFAR10 dataset, using only 2 classes of a 10% sub-sample of the dataset.

The results presented in Figs. 13 and 14 are coherent with those obtained in Section 6, hence hinting toward the idea that our theory could extend to more practical settings.

F.4.4. Experiments using Pearson correlation coefficient

In Fig. 2, we presented the value of the Kendall's correlation coefficient between the generalization error and the tail index α, for different values of d. We present the same experiment in Fig. 15, but using Pearson's correlation coefficient instead of Kendall's correlation coefficient. We observe that we can still observe the desired phase transition, hence showing that our experiment are robust to the choice of correlation coefficient. However, Kendall's coefficient may be more adapted to our setup, as it is not clear that a linear correlation can be observed in practice.

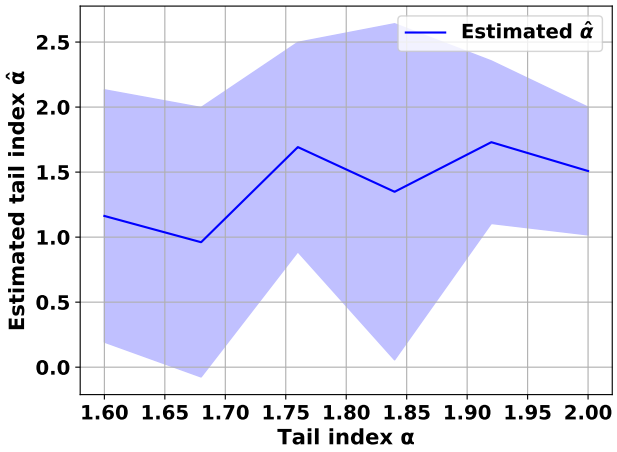
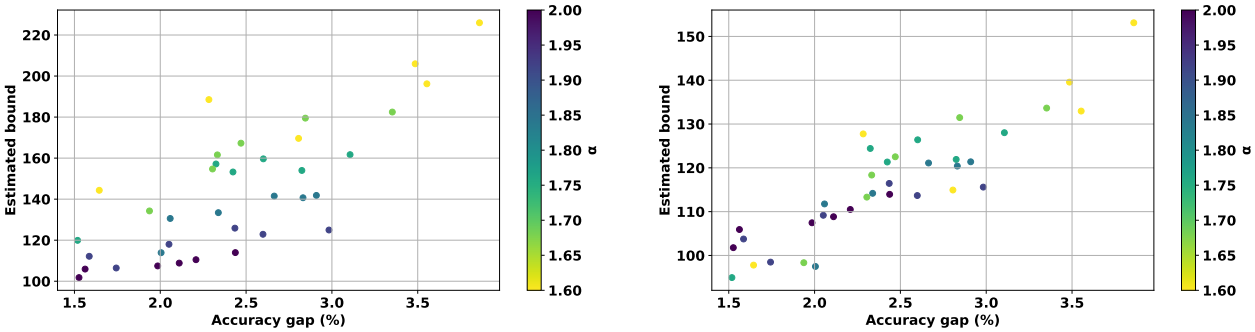


Figure 13. Regression of α from the generalization bound, for a CNN on CIFAR10.



(a) $R = 1$ (b) $R = 7$

Figure 14. Estimated bound, computed with Eq. (49), versus accuracy gap, for a CNN on CIFAR10, for $R = 1$ (*left*) and $R = 7$ (*right*).

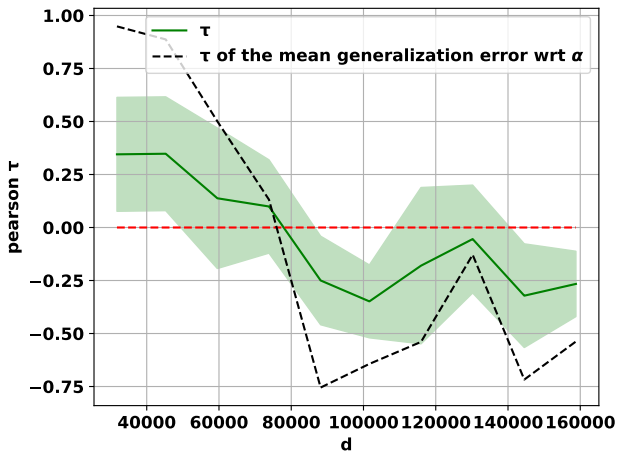
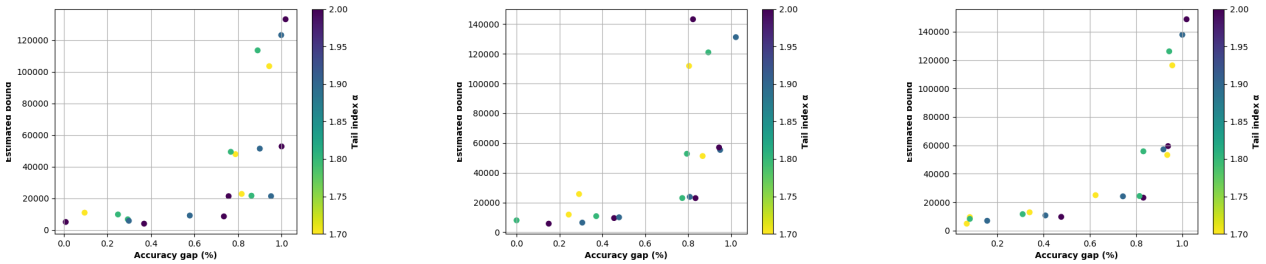


Figure 15. Same experiment as Fig. 2, but using the Pearson correlation coefficient instead of the Kendall's correlation coefficient.

F.4.5. Additional experiments with the full MNIST dataset and mini-batches

As already mentioned in Section 6, in order to argue that our bounds may also be pertinent in more practical settings than the figure presented in the rest of the paper, we computed our bound using a FCN5, on the full MNIST dataset, using mini-batches. This is to be compared eith our other experiments, were, according to the SDE we study, we take the full batch at each iteration to compute $\nabla \widehat{F}_S$. The exact procedure and hyperparameters are detailed, as for other experiments, in Appendix F.1. These results are shown on Appendix F.4.5, were we observe that, for several values of the batch size, we still observe a very good correlation between the estimated bound and the generalization error. Unfortunately, our theroy does not predict the behavior of the accuracy gap with respect to the batch-size, but, from this experiment, we understand that it is still capturing relatively well the behavior of the accuracy gap.



(a) Batch size 64 (b) Batch size 128 (c) Batch size 256

Figure 16. Estimated bound, computed with Eq. (49), versus accuracy gap, for a 5-layers FCN on the full MNIST dataset, for different value of batch size.