From Mutual Information to Expected Dynamics: New Generalization Bounds for Heavy-Tailed SGD

Benjamin Dupuis
Inria, CNRS, Ecole Normale Supérieure
PSL Research University, Paris, France
`benjamin.dupuis@inria.fr`

Paul Viallard
Inria, CNRS, Ecole Normale Supérieure
PSL Research University, Paris, France
`paul.viallard@inria.fr`

Abstract

Understanding the generalization abilities of modern machine learning algorithms has been a major research topic over the past decades. In recent years, the learning dynamics of Stochastic Gradient Descent (SGD) have been related to heavy-tailed dynamics. This has been successfully applied to generalization theory by exploiting the fractal properties of those dynamics. However, the derived bounds depend on mutual information (decoupling) terms that are beyond the reach of computability. In this work, we prove generalization bounds over the trajectory of a class of heavy-tailed dynamics, without those mutual information terms. Instead, we introduce a geometric decoupling term by comparing the learning dynamics (depending on the empirical risk) with an expected one (depending on the population risk). We further upper-bound this geometric term, by using techniques from the heavy-tailed and the fractal literature, making it fully computable. Moreover, as an attempt to tighten the bounds, we propose a PAC-Bayesian setting based on perturbed dynamics, in which the same geometric term plays a crucial role and can still be bounded using the techniques described above.

1 Introduction

In machine learning, we aim to solve a learning task represented by an unknown probability distribution μ_z over a set of data probability space $(\mathcal{Z}, \mathcal{F})$, where $\mathcal{F}$ is the associated σ-algebra. In order to solve it, we aim to learn a machine learning model parameterized by some weights $w \in \mathbb{R}^d$. For instance, the model can be obtained by solving the *Population Risk Minimization* problem defined by

$$\min_{w \in \mathbb{R}^d} \left\{ \mathcal{R}(w) := \mathbb{E}_{z \sim \mu} \ell(w, z) \right\}, \tag{PRM}$$

where $\ell : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}$ is the loss function given the weights, $z \in \mathcal{Z}$ an example and $\mathcal{R}(w)$ the population risk. However, since the probability distribution μ_z is unknown, we cannot directly solve PRM to obtain a model. Instead, we have at our disposal n data points $S = \{z_i\}_{i=1}^n$, sampled *i.i.d.* from μ_z that serve to obtain a model. Thanks to the data, a model can be obtained by (approximately) solving the *Empirical Risk Minimization* problem defined as

$$\min_{w \in \mathbb{R}^d} \left\{ \widehat{\mathcal{R}}_S(w) := \tfrac{1}{n} \sum_{i=1}^n \ell(w, z_i) \right\}, \tag{ERM}$$

where $\widehat{\mathcal{R}}_S(\cdot)$ is the *empirical risk*. Nowadays, one workhorse algorithm used to solve the ERM problem is the Stochastic Gradient Descent (SGD) algorithm. Starting from weights $W_0 \in \mathbb{R}^d$, the algorithm updates the weights iteratively by applying the following update at each time t

$$W_{t+1} = W_t - \eta \nabla \widehat{\mathcal{R}}_B(W_t),$$

where $\eta > 0$ is the learning rate and $B \subseteq S$ is a random subset of indices. Given B and η, there is theoretical and empirical evidence that the gradient noise $\nabla\widehat{\mathcal{R}}_B(W_t) - \nabla\widehat{\mathcal{R}}_S(W_t)$ can be *heavy-tailed* [30, 10, 12, 36], *i.e.*, there is a high probability that the gradient noise is large. One way of modeling SGD with such a gradient noise can be done thanks to the following stochastic differential equation (SDE) [22, 25, 31, 30]

$$dW_t = -\nabla\widehat{\mathcal{R}}_S(W_t)dt + dL_t^\alpha, \tag{1}$$

where L_t^α is a d-dimensional α-stable Lévy process (defined more formally further), where α is the "tail-index". Under this modeling assumption, *generalization bounds* (see Mohri et al. [20]), *i.e.*, upper bounds on the generalization gap $\mathcal{R}(w) - \widehat{\mathcal{R}}_S(w)$ have been derived. More precisely, most works focus on the *worst-case generalization error* over the trajectory. Şimşekli et al. [31, Thms. 1 and 2], Hodgkinson et al. [13, Thm 1. with Cor. 1], and Birdal et al. [2, Prop. 1] propose generalization bounds that have informally the following form, with probability at least $1-\zeta$

$$\sup_{w \in \mathcal{W}_S} \left(\mathcal{R}(w) - \widehat{\mathcal{R}}_S(w) \right) \lesssim \sqrt{\frac{\overline{\dim}_B + \log(1/\zeta) + \mathrm{I}_\infty(\mathcal{W}_S, S)}{n}}, \tag{2}$$

where, $\mathcal{W}_S := \{W_t, \ 0 \leq t \leq T\}$ is the trajectory of the dynamics up to a certain time horizon, $\mathrm{I}_\infty(\mathcal{W}_S, S)$ is a total mutual information term, quantifying the statistical dependence between S and $\mathcal{W}_S$, see [14, Section 2.2], and $\overline{\dim}_B$ is a notion of fractal dimension[1], which can be seen as a measure of the complexity of $\mathcal{W}_S$ (defined formally in App. A.2). The fractal geometry of those stochastic processes, like Eq. (2), has been extensively studied [37] and provided powerful tools to study the generalization properties of such heavy-tailed dynamics. Şimşekli et al. [31] noted that, under reasonable assumptions, the dimension $\overline{\dim}_B$, appearing in Eq. (2), is equal to the corresponding *Hausdorff dimension* and may be related to the tail-index α.

Even if the fractal dimension could be approximated by estimating α, or by using the method developed by Birdal et al. [2], the total mutual information term appearing in the bound is not computable, making the bound beyond the reach of any evaluation. Moreover, the total mutual information can be large or even infinite, making the bound potentially vacuous. To avoid this inconvenience, Raj et al. [25] proposed bounds for Eq. (1) without mutual information terms based on a stability approach. Their method is based on comparing two dynamics, using a surrogate loss, associated with two datasets differing by one point. However, their bounds are bounds in expectation over the last point of the dynamics and have an important dependence on the number of parameters d, while we bound the worst-case generalization gap, over the trajectory, with high probability.

In this work, we aim to close this gap by proposing in Sec. 3 a new upper bound on $\sup_{w \in \mathcal{W}_S} \mathcal{R}(w) - \widehat{\mathcal{R}}_S(w)$ using purely geometric and computable quantity, instead of the total mutual information. This is done by the comparison of the SDE in Eq. (1) with the corresponding *expected dynamics*, where we consider the population risk in the dynamic instead of the empirical one. Most importantly, we further bound the introduced geometric term using properties of the heavy-tailed dynamics. As an attempt to obtain a tighter bound, in Sec. 4, we present a PAC-Bayesian setting in which the aforementioned geometric terms naturally appear, and we prove new bounds on the generalization gap $\mathcal{R}(w) - \widehat{\mathcal{R}}_S(w)$, based on this PAC-Bayesian framework. In this case, the big constants appearing in the bounds of Sec. 3, which in particular are exponential in time, are compensated by a higher rate in n (*e.g.*, linear), therefore substantially improving the bounds. We also highlight in App. B that, under additional assumptions, time independence may be obtained in this PAC-Bayesian setting. Note that all proofs are deferred in the appendices.

2 Setting and Notations

2.1 Learning Dynamics

Let $(\Omega, \mathcal{T}, \mu_u)$ be a complete probability space, representing the external randomness added by the algorithms. To match with notations in [31, 7], the randomness will generally be denoted U. We

---

[1]We refer the reader to [8] for an introduction to fractal geometry.

consider, on Ω, the two dynamics, in $\mathbb{R}^d$, that we call respectively *empirical* and *expected* dynamics:

$$dW_t^S = -\nabla\widehat{\mathcal{R}}_S(W_t^S)dt + dL_t^\alpha, \quad \text{with} \quad W_0 = Z_0, \tag{3}$$

$$\text{and} \quad dY_t = -\nabla\mathcal{R}(Y_t)dt + dL_t^\alpha, \quad \text{with} \quad Y_0 = Z_0, \tag{4}$$

where Z_0 is a random weight independent of all other random variables. Recall that L_t^α is a strictly α-stable Lévy process; see App. A.1, or the tutorial [28], for formal definitions. Note that Eq. (3) is similar to models considered in other works [30, 22, 25] and is a particular case of [31, Eq. 4]. To ensure that the SDEs in Eqs. (3) and (4) have strong solutions and are Feller processes (see Th. 7), the following assumption imposes that the terms $-\nabla\mathcal{R}(\cdot), -\nabla\widehat{\mathcal{R}}_S(\cdot)$ are smooth enough.

Assumption 1 *The loss is differentiable, L-Lipschitz continuous and M-smooth, i.e., for all $w, w' \in \mathbb{R}^d$ and $z \in \mathcal{Z}$, we have:*

$$|\ell(w, z) - \ell(w', z)| \leq L\|w - w'\|, \quad \text{and} \quad \|\nabla\ell(w, z) - \nabla\ell(w', z)\| \leq M\|w - w'\|.$$

This assumption plays a great role in the generalization bounds presented in Secs. 3 and 4. As it is common to derive generalization bounds (see *e.g.*, [38, 31]), we make a sub-Gaussian assumption on the distribution of $\ell(w, z)$, to control its concentration around its expectation.

Assumption 2 *The function $\ell : \mathbb{R}^d \times \mathcal{Z} \longrightarrow \mathbb{R}$ is assumed to be σ-sub-Gaussian, i.e., for all $\lambda > 0$, we have $\mathbb{E}\left[e^{\lambda(\ell(w,z)-\mathbb{E}[\ell(w,z)])}\right] \leq e^{\sigma^2\lambda^2/2}$.*

In this paper, we also fix the time horizon $T > 0$ of the processes that we denote by $Y_t(U)$ and $W_t^S(U)$ by omitting the variable U when it is not ambiguous. Let us also introduce the trajectories $\mathcal{W}_{S,U}$ and $\mathcal{Y}_U$ associated with the processes $Y_t(U)$ and $W_t^S(U)$ and defined by

$$\mathcal{W}_{S,U} = \{W_t^S(U), \ 0 \leq t \leq T\}, \quad \mathcal{Y}_U = \{Y_t(U), \ 0 \leq t \leq T\},$$

which is the sets of weights $W_t^S(U)$ and $Y_t(U)$ encountered by the processes.

Remark 1 *Eq. (3) is here understood as holding at all times, almost surely with respect to $\mu_z^{\otimes n} \otimes \mu_u$. Moreover, we assume that the map $(S, U, t) \longmapsto W_t^S(U)$ is (jointly) measurable. In particular, this assumption can be true if $\mathcal{Z}$ is countable. We also implicitely use a càdlàg version of L_t^α, without loss of generality [28, Section 2], [26, Theorem 2.9].*

In addition to Asm. 1, we make the following topological assumption on the trajectory of $\mathcal{Y}$, which will help us control its fractal properties and relate them to the tail-index α, it is similar to assumptions already made in the literature, see [31, Assumption H4].

Assumption 3 (Fractal characterization of the dynamics) *We assume that, almost surely, the Hausdorff and upper box-counting dimension of $\mathcal{Y}_U$ are the same, i.e., $\overline{\dim}_B(\mathcal{Y}_U) = \dim_H(\mathcal{Y}_U)$. In all the following, this common dimension value will be denoted γ. Note that we provide technical background, regarding fractal geometry, in App. A.2.*

Roughly speaking, we implicitly assume that the process $(Y_t)_{t\geq0}$ is "regular" enough (which happens for smooth dynamics). More formally, this is particularly true if $\alpha = 2$ or if we assume the so-called Ahlfors regularity of $\mathcal{Y}_U$ ([31, Assumption H4] and [14, Corollary 1]), see App. A.2.

2.2 Heavy-tailed Properties

In our particular setting, we can conduct a precise analysis of the relation between the Hausdorff dimension of $\mathcal{Y}$ and $\mathcal{W}_{S,U}$ and the tail-index α. The following theorem describes some relationships between this dimension and the tail-index α, for the dynamics considered in this paper. It is a direct consequence of existing works on the fractal dimensions of Feller processes [27, 4, 16].

Theorem 1 (Fractal properties of heavy-tailed dynamics (informal)) *Under Asms. 1 and 2, if one of the following condition is true: (i) $\alpha \in [1, 2]$, or (ii) L_t^α is strictly stable and $\nabla\ell$ is smooth enough in w (compact support is enough for instance) so that Y and W^S satisfy the assumptions of [27, Theorem 4], then $\dim_H(\mathcal{Y}_U), \dim_H(\mathcal{W}_{S,U}) \leq \alpha$. Hence, under Asm. 3, we have $\gamma \leq \alpha$.*

In the rest of the paper, to avoid overloading the statements with technical assumptions, we will keep the notations γ to refer to the fractal dimension of the trajectories. However, γ may be replaced by α, in our upper bounds, whenever Asm. 3 and Th. 1 hold.

3

3   From Statistical to Geometric Information

In this section, we present a generalization bound, in Th. 2, on the worst-case generalization gap in terms of the fractal dimension γ of the trajectory, but without mutual information term, compared to [31, 2, 14, 7]. The following theorem introduces a Euclidean distance between the empirical and expected trajectories and uses [31, Lemma 1] to bound the generalization error over the data-dependent trajectory $\mathcal{W}_{S,U}$.

Theorem 2 *Fix some $\zeta \in (0,1)$. Under Asms. 1 to 3, there exists $N \in \mathbb{N}$, depending only on ζ such that, with probability at least $1 - 2\zeta$ over $\mu_u \otimes \mu^{\otimes n}$, for all $n \geq N$, we have:*

$$\sup_{w \in \mathcal{W}_{S,U}} \left(\mathcal{R}(w) - \widehat{\mathcal{R}}_S(w) \right) \leq 2L \sup_{0 \leq t \leq T} \|W_t^S - Y_t\| + \frac{2}{\sqrt{n}} + 2\sigma \sqrt{\frac{2\log(L\sqrt{n})\gamma + \log(1/\zeta)}{2n}}.$$

Remark 2 *Th. 2 would still be true if we replace the term $\sup_{0 \leq t \leq T} \|W_t^S - Y_t\|$ by the Hausdorff distance $d_H(\mathcal{W}_{S,U}, \mathcal{Y}_U)$ between $\mathcal{W}_{S,U}$ and $\mathcal{Y}$, see the proof for more details.*

Introducing the geometric term $\sup_{0 \leq t \leq T} \|W_t^S - Y_t\|$ allows us to transfer the estimation of the generalization error from $\mathcal{W}_{S,U}$ to $\mathcal{Y}_U$ (which does not depend on the data). Note that it is not clear, when looking at Th. 2, that the given bound is non-vacuous, as the distance between the trajectories is not divided by $\sqrt{n}$. However, it is intuitive to think that $\sup_{0 \leq t \leq T} \|W_t^S - Y_t\|$ could decrease rapidly as $n \to \infty$, because of the convergence of $\widehat{\mathcal{R}}_S(\cdot)$ to $\mathcal{R}(\cdot)$. Therefore, in the rest of this section, we will derive bounds on this term, hence making our generalization bound more pertinent. An important property, of our strategy to achieve this goal, is that we consider Asm. 1 and then leverage the fractal properties of the heavy-tailed trajectories by performing a covering argument and introducing the fractal dimension, see App. C for more details. The aforementioned covering argument is used to bound the worst-case concentration of the gradients $\nabla \ell(w, z)$, whose norm is sub-Gaussian thanks to the Lipschitz loss assumption. The drawback of this approach is to yield a big dependence of the bound on the Lipschitz constant L. However, in the best-case scenario where the distribution of the gradient norm is evenly shared among its components, we note that it would mean that each partial derivative $\frac{\partial \ell(w,z)}{\partial w_j}$ has a sub-Gaussian norm [34, Definition 2.5.6] of order $\mathcal{O}(1/\sqrt{d})$. Thus, we argue that a strong concentration of those derivatives can lead to a better bound, and, *when specified*, we shall make the following assumption.

Assumption 4 (Partial derivatives concentration) *For $1 \leq j \leq d$, the random variable $\partial_j \ell(w, z)$ is $\Sigma/\sqrt{d}$-sub-Gaussian, uniformly on j and w, where $\Sigma > 0$ is a constant.*

The following theorem provides a bound on the geometric term, $\sup_{0 \leq t \leq T} \|W_t^S - Y_t\|$, using the methods described above under different sets of assumptions.

Theorem 3 *Let us fix some $\zeta \in (0,1)$. Under Asms. 1 to 3, there exists $N \in \mathbb{N}$, depending on ζ, such that, with probability at least $1 - 2\zeta$ over $\mu_u \otimes \mu^{\otimes n}$, for all $n \geq N$, we have:*

$$\sup_{0 \leq t \leq T} \|W_t^S - Y_t\| \leq \frac{e^{MT} - 1}{M} \left(\frac{2}{\sqrt{n}} + L\sqrt{\frac{2}{n}} + 2L\sqrt{\frac{2\gamma \log(M\sqrt{n}) + \log(1/\zeta)}{2n}} \right).$$

If we further make Asm. 4, we can improve the constants to:

$$\sup_{0 \leq t \leq T} \|W_t^S - Y_t\| \leq \frac{e^{MT} - 1}{M} \left(\frac{2}{\sqrt{n}} + 2\Sigma\sqrt{\frac{2\gamma \log(M\sqrt{n}) + \log(2d/\zeta)}{2n}} \right).$$

Therefore, the tail-index α (upper-bounding γ) also has an impact on $\sup_{0 \leq t \leq T} \|W_t^S - Y_t\|$. While the bounds of Th. 3 may be made vacuous because of the dependence on T and the Lipschitz constant L, we highlight that it has the advantage of being explicit, as opposed to all existing worst-case heavy-tailed generalization bounds in high probability. In some sense, we replaced the statistical decoupling (based on mutual information terms) with a new kind of geometric term, based on the Euclidean distance between the empirical and the expected dynamics. Therefore, by combining Ths. 2 and 3, we obtain the following corollary.

Theorem 4 *Let us fix some $\zeta \in (0,1)$. Under Asms. 1 to 3, there exists $N \in \mathbb{N}$, depending only on ζ such that, with probability at least $1 - 4\zeta$ over $\mu_u \otimes \mu^{\otimes n}$, for all $n \geq N$, we have:*

$$\sup_{w \in \mathcal{W}_{S,U}} \left(\mathcal{R}(w) - \widehat{\mathcal{R}}_S(w) \right) \leq \frac{2}{\sqrt{n}} + \frac{2L}{\sqrt{n}} C_1 \frac{e^{MT} - 1}{M} + C_2 \sqrt{\frac{2\log(C_0 \sqrt{n})\gamma + \log(C_3/\zeta)}{2n}},$$

with constants

$$C_0 := \max\{M, L\}, \;\; C_1 := 2 + \sqrt{2}L, \;\; C_2 := 4L^2 \tfrac{1}{M}(e^{MT}-1) + 2\sigma, \;\; and \;\; C_3 = 1.$$

If we further make Asm. 4, we can improve the constants to (with a little loss on C_3) to

$$C_1 := 2, \quad C_2 := 4\Sigma \tfrac{L}{M}(e^{MT}-1) + 2\sigma, \quad C_3 = 2d.$$

Th. 4 corresponds to a direct way of using expected dynamics, at the cost of introducing huge constants in the derived bounds. In the next subsection, we propose a first step toward a better understanding of the geometric term $\sup_{0 \leq t \leq T} \|W_t^S - Y_t\|$, based on the PAC-Bayesian theory.

4   Tightening the Bounds Through the PAC-Bayesian Theory

While the method described in the previous section allows us to prove bounds for heavy-tailed dynamics without any mutual information term, it makes the bound rely too much on the Lipschitz continuity assumption, and has an exponential dependence on the training time, which may render the bound vacuous. In this section, we argue that the comparison between empirical and expected dynamics is able to induce better bounds. Inspired by PAC-Bayesian theory[2] [29, 19], we want to prove bounds where the supremum over the trajectory $\sup_{w \in \mathcal{W}_{S,U}} (\mathcal{R}(w) - \widehat{\mathcal{R}}_S(w))$ is replaced by a uniform sampling over the trajectory. To do so, we define a distribution for the trajectory $\mathcal{W}_{S,U}$:

$$\forall B \in \mathcal{B}(\mathbb{R}^d), \quad \tilde{\rho}_{S,U}(B) := \frac{1}{T} \int_0^T \mathbb{1}\left(W_t^S(U) \in B\right) dt, \tag{5}$$

where $\mathcal{B}(\mathbb{R}^d)$ is the Borel σ-algebra of $\mathbb{R}^d$ and $\mathbb{1}(\cdot)$ is the indicator function. This distribution can be interpreted as the distribution of the weights that are sampled uniformly over time. It is related to the notion of occupation measure in stochastic calculus (see [37, Section 6]).

Remark 3 *Within our fractal framework, the distribution $\tilde{\rho}_{S,U}$ can be replaced with the exact Hausdorff measure restricted on the trajectories, assuming those measures are finite. However, it appears in several works, like [37, Theorem 5.1], that those measures are often, up to multiplicative constants, equivalent to the occupation measures [37, Section 6] of the processes, making the sampling in time and the sampling of the points similar.*

PAC-Bayesian analysis requires the comparison between the data-dependent distribution $\tilde{\rho}_{S,U}$ with a data-independent distribution, *i.e.*, a prior distribution. In order to leverage the comparison between the empirical dynamics (3) and the expected one (4), the natural choice is to define the prior (denoted $\tilde{\pi}_U$) in the same way than $\tilde{\rho}_{S,U}$, but with support on the expected trajectory, *i.e.*, $\tilde{\pi}_U(B) = (1/T) \int_0^T \mathbb{1}\left(Y_t(U) \in B\right) dt$. While this choice is appealing, it has a major drawback. Indeed, those distributions do not satisfy the absolute continuity condition required by the PAC-Bayesian framework (*i.e.*, $\tilde{\rho}_{S,U} \not\ll \tilde{\pi}_U$). In order to overcome this issue, we chose to perturb the distribution $\tilde{\rho}_{S,U}$ by a Gaussian one. This leads to the following definitions of the distribution on both trajectories $\mathcal{W}_{S,U}$:

$$\rho_{S,U} := \mathcal{N}(0, s^2) * \tilde{\rho}_{S,U}, \quad \pi_U := \mathcal{N}(0, s^2) * \tilde{\pi}_{S,U} \tag{6}$$

where $*$ denotes the convolution. This kind of smoothed distribution has already been considered by Lugosi and Neu [17], but not for trajectories, like in Eq. (6). We will call $\rho_{S,U}$ the posterior distributions and π_U the prior distributions. Informally, sampling from $\rho_{S,U}$ means sampling points that are not too far from the true trajectory $\mathcal{W}_{S,U}$, as if we were studying a learning algorithm approximating, in some sense, those continuous trajectories. The scale parameter s then controls the quality of this approximation.

---

[2]See [1, 11] for an introduction to the PAC-Bayesian theory.

We present two results (respectively in Th. 5 and Th. 6): *(a)* a bound in expectation over $\rho_{S,U}$, like in the "classical" PAC-Bayesian theory (see *e.g.*, [9]) and *(b)* a bound in high-probability over $\rho_{S,U}$ with the "disintegrated" PAC-Bayesian theory [35]. The power of our approach is that we are able to bound the information-theoretic terms, *i.e.*, the KL (*resp.* Rényi) divergence KL $\left(\rho_{S,U}\|\pi_U\right)$ (*resp.* $\mathrm{D}_\beta\left(\rho_{S,U}\|\pi_U\right)$), appearing in those bounds by the same geometric term than in Sec. 3, see App. C.5 for details. Therefore, the analysis of Sec. 3 still applies. The next theorem presents our first bound, which is an application of Th. 10 in our setting.

Theorem 5 *Under Asms. 1 and 2, for any $\lambda>0$, with probability at least $1-\zeta$ over (S,U) we have:*

$$\lambda\mathbb{E}_{w\sim\rho_{S,U}}[\mathcal{R}(w)-\widehat{\mathcal{R}}_S(w)] \leq \frac{1}{Ts^2}\int_0^T \|W_u^S - Y_u\|^2 du + \log(1/\zeta) + \frac{C\lambda^2}{n}.$$

Note that we can further bound the integral term in Th. 5 with the same term than in Th. 2, *i.e.*,

$$\frac{1}{T}\int_0^T \|W_u^S - Y_u\|^2 du \leq \sup_{0\leq t\leq T}\|W_t^S - Y_t\|^2.$$

We now state the corresponding disintegrated bound, using the framework developed by Viallard et al. [35], in which we obtain the same geometric distances as in Ths. 2 and 5.

Theorem 6 *Under Asms. 1 and 2, for any $\lambda > 0$ and $\beta > 1$, we have, with probability at least $1 - \zeta$ over the joint distribution of (S,U) and $w \sim \rho_{S,U}$:*

$$\frac{\lambda\beta}{\beta-1}\left(\mathcal{R}(w)-\widehat{\mathcal{R}}_S(w)\right) \leq \frac{2\beta-1}{\beta-1}\log\left(\frac{2}{\zeta}\right) + \frac{\beta}{2s^2}\sup_{0\leq t\leq T}\|W_t^S - Y_t\|^2 + \left(\frac{\beta}{\beta-1}\right)^2\frac{\lambda^2\sigma^2}{2n}.$$

Th. 5 provides a bound by integrating the generalization error over $\rho_{S,U}$, which can be seen as a mean error near the empirical trajectory $\mathcal{W}_{S,U}$. Th. 6 provides more precise information, it gives a high probability bound by sampling weights near the empirical trajectory, this comes at the cost of replacing the integral distance, *i.e.*, $\frac{1}{T}\int_0^T \|W_u^S - Y_u\|^2 du$, by the supremum of the distances $\sup_{0\leq t\leq T}\|W_t^S - Y_t\|^2$, and introducing the parameter β.

Note that the same geometric distance term $\sup_{0\leq t\leq T}\|W_t^S - Y_t\|$ as in Sec. 3 appears in Ths. 5 and 6, but this time to the power 2. As mentioned in Sec. 3, the term $\sup_{0\leq t\leq T}\|W_t^S - Y_t\|^2$ can be bounded by methods relying on the fractal properties of the heavy-tailed dynamics. As it is shown in Th. 3, it is bounded by a term of magnitude $C_{L,T}^2/n$, where $C_{L,T}$ depends, among other things, on L and T. By optimizing the value parameter λ in the above bounds, we can obtain generalization bounds of order $\mathcal{O}(\sqrt{\log(1/\zeta)/n} + C_{L,T}/n)$, while the bounds in Sec. 3 are in $\mathcal{O}\left(C_{L,T}/\sqrt{n}\right)$. Therefore, the potentially big introduced constants are compensated by a higher order in n, therefore improving the bound. Our PAC-Bayesian framework has less dependence on the Lipschitz constant L, and even no dependence under Asm. 4. Moreover, in App. B, we will present, that under different assumptions, we tighten the results and obtain time-independent bounds.

5 Conclusion

In this paper, we derived generalization bounds for understanding the generalization ability of continuous dynamics driven by a stable Lévy process, and depending on the empirical risk. This highlights the generalization properties of models obtained with a stochastic gradient descent algorithm exhibiting heavy-tailed gradient noises. The main advantage of our analysis is that it does not depend on a mutual information term that is not fully understandable. Instead, we considered the expected dynamic (depending on the unknown population risk) and introduced a new geometric term in the results, that we bound using techniques from the heavy-tails and fractal literature. Several points are still to be investigated. From the theoretical perspective, it would be interesting to elaborate on the ideas presented in App. B to tighten the bounds. Moreover, we would like to extend our methods to discrete dynamics with heavy-tailed noise, and also prove non-uniform bounds (*i.e.* on the last point).

Acknowledgements

References

[1] P. Alquier. User-friendly introduction to PAC-Bayes bounds. *CoRR*, abs/2110.11216, 2021.

[2] T. Birdal, A. Lou, L. J. Guibas, and U. Şimşekli. Intrinsic dimension, persistent homology and generalization in neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6776–6789, 2021.

[3] R. Blumenthal and R. Getoor. Some theorems on stable processes. *Transactions of the American Mathematical Society*, 95(2):263–273, 1960.

[4] B. Böttcher, R. Schilling, and J. Wang. *Lévy Matters III: Lévy-Type Processes: Construction, Approximation and Sample Path Properties*, volume 2099 of *Lecture Notes in Mathematics*. Springer International Publishing, 2013.

[5] A. Camuto, G. Deligiannidis, M. Erdogdu, M. Gürbüzbalaban, U. Şimşekli, and L. Zhu. Fractal structure and generalization properties of stochastic optimization algorithms. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 18774–18788, 2021.

[6] P. Chen, C.-S. Deng, R. Schilling, and L. Xu. Approximation of the invariant measure of stable SDEs by an Euler–Maruyama scheme. *Stochastic Processes and their Applications*, 163: 136–167, 2023.

[7] B. Dupuis, G. Deligiannidis, and U. Şimşekli. Generalization bounds with data-dependent fractal dimensions. In *International Conference on Machine Learning (ICML)*, 2023.

[8] K. Falconer. *Fractal Geometry: Mathematical Foundations and Applications*. John Wiley & Sons, Inc., third edition, 2014.

[9] P. Germain, A. Lacasse, F. Laviolette, and M. Marchand. PAC-Bayesian learning of linear classifiers. In *International Conference on Machine Learning (ICML)*, volume 382 of *ACM International Conference Proceeding Series*, pages 353–360. ACM, 2009.

[10] M. Gürbüzbalaban, U. Şimşekli, and L. Zhu. The heavy-tail phenomenon in SGD. In *International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 3964–3975. PMLR, 2021.

[11] F. Hellström, G. Durisi, B. Guedj, and M. Raginsky. Generalization bounds: Perspectives from information theory and PAC-Bayes. *CoRR*, abs/2309.04381, 2023.

[12] L. Hodgkinson and M. Mahoney. Multiplicative noise and heavy tails in stochastic optimization. In *International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 4262–4274. PMLR, 2021.

[13] L. Hodgkinson, U. Şimşekli, R. Khanna, and M. Mahoney. Generalization properties of stochastic optimizers via trajectory analysis. *CoRR*, abs/2108.00781, 2021.

[14] L. Hodgkinson, U. Şimşekli, R. Khanna, and M. Mahoney. Generalization bounds using lower tail exponents in stochastic optimizers. In *International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pages 8774–8795. PMLR, 2022.

[15] N. Jain and W. Pruitt. The correct measure function for the graph of a transient stable process. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 9(2):131–138, 1968.

[16] V. Knopova, R. Schilling, and J. Wang. Lower bounds of the Hausdorff dimension for the images of Feller processes. *Statistics & Probability Letters*, 97:222–228, 2015.

[17] G. Lugosi and G. Neu. Generalization bounds via convex analysis. In *Conference on Learning Theory (COLT)*, volume 178 of *Proceedings of Machine Learning Research*, pages 3524–3546. PMLR, 2022.

[18] P. Mattila. *Geometry of Sets and Measures in Euclidean Spaces: Fractals and Rectifiability*. Cambridge University Press, first edition, 1995.

[19] D. McAllester. Some PAC-Bayesian theorems. In *Conference on Computational Learning Theory (COLT)*, pages 230–234. ACM, 1998.

[20] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine Learning*. Adaptive computation and machine learning. MIT Press, 2012.

[21] I. Molchanov. *Theory of Random Sets*, volume 87 of *Probability Theory and Stochastic Modelling*. Springer London, 2017.

[22] T. H. Nguyen, U. Şimşekli, M. Gürbüzbalaban, and G. Richard. First exit time analysis of stochastic gradient descent under heavy-tailed gradient noise. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 273–283, 2019.

[23] W. Pruitt. The Hausdorff dimension of the range of a process with stationary independent increments. *Indiana University Mathematics Journal*, 19(4):371–378, 1969.

[24] W. Pruitt and J. Taylor. Sample path properties of processes with stable components. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 12(4):267–289, 1969.

[25] A. Raj, L. Zhu, M. Gürbüzbalaban, and U. Şimşekli. Algorithmic stability of heavy-tailed SGD with general loss functions. In *International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pages 28578–28597. PMLR, 2023.

[26] D. Revuz and M. Yor. *Continuous Martingales and Brownian Motion*, volume 293 of *Grundlehren der mathematischen Wissenschaften*. Springer Berlin Heidelberg, 1999.

[27] R. Schilling. Feller processes generated by pseudo-differential operators: On the Hausdorff dimension of their sample paths. *Journal of Theoretical Probability*, 11(2):303–330, 1998.

[28] R. Schilling. An introduction to Lévy and Feller processes. Advanced courses in mathematics - CRM Barcelona 2014. *CoRR*, abs/1603.00251, 2016.

[29] J. Shawe-Taylor and R. Williamson. A PAC analysis of a Bayesian estimator. In *Conference on Computational Learning Theory (COLT)*, pages 2–9. ACM, 1997.

[30] U. Şimşekli, L. Sagun, and M. Gürbüzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pages 5827–5837. PMLR, 2019.

[31] U. Şimşekli, O. Sener, G. Deligiannidis, and M. Erdogdu. Hausdorff dimension, heavy tails, and generalization in neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12), 2021.

[32] J. Taylor. Sample path properties of a transient stable process. *Journal of Mathematics and Mechanics*, 16(11):1229–1246, 1967.

[33] T. van Erven and P. Harremoës. Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

[34] R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, first edition, 2018.

[35] P. Viallard, P. Germain, A. Habrard, and E. Morvant. A general framework for the practical disintegration of PAC-Bayesian bounds. *Machine Learning*, 2023.

[36] H. Wang, M. Gürbüzbalaban, L. Zhu, U. Şimşekli, and M. Erdogdu. Convergence rates of stochastic gradient descent under infinite noise variance. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 18866–18877, 2021.

[37] Y. Xiao. Random fractals and Markov processes. In *Proceedings of Symposia in Pure Mathematics*, volume 72.2, pages 261–338. American Mathematical Society, 2004.

[38] A. Xu and M. Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2524–2533, 2017.

The appendix is organized as follows:

- In App. A, we introduce some mathematical tools necessary for a good understanding of the paper. The main goal is to give the definition of Levy and Feller processes, as well as their relation to fractal geometry. This allows us to define precisely the fractal dimensions appearing in our proofs and prove rigorously Th. 1. We also provide a quick technical background on PAC-Bayesian theory.
- App. B presents an additional bound on the distance between empirical and expected trajectories, based on a dissipativity assumption.
- In App. C, we present all proofs of our results.

A   Technical Background

In this section, we provide a little technical background in order to present the tools and theorems that will be used in our proofs.

A.1   Lévy and Feller Processes

In this subsection, we describe some basic notions related to Lévy and Feller processes. We mention only the tools that are necessary for our results and their proofs, for more technical details and proofs, the reader may look at the tutorial of Schilling [28] or the following textbooks [4, 37].

Lévy processes: A Lévy process $(X_t)_{t\geq 0}$ is a stochastic process, in $\mathbb{R}^d$, with $X_0 = 0$ almost surely, satisfying the following properties:

- **independent increments:** for any $t_0 < t_1 < \cdots < t_k$, the random variables $X_{t_i} - X_{t_{i+1}}$, for $1 \leq i \leq k$, are independent.
- **Stationary increments:** For any $s < t$ the random variables $X_t - X_s$ and X_{t-s} have the same distribution.
- X is **stochastically continuous**, meaning:

$$\forall \epsilon > 0, \ \forall s \geq 0, \mathbb{P}\left(\|X_t - X_s\| > \epsilon\right) \xrightarrow[t\to s]{} 0.$$

One can show that such processes have modifications (*i.e.* a process $(\tilde{X}_t)_{t\geq 0}$ such that, for all t, $X_t = \tilde{X}_t$ almost surely) with almost surely *càdlàg*[3] paths, see [28, chapter 2], it is equivalent to the property of being stochastically continuous. In this work, we always assume that the considered Levy processes have almost surely càdlàg paths. Thus, Lévy processes are not continuous but admit *random jumps*. It should be noted that càdlàg paths are almost surely bounded on compact intervals, which we shall use without mentioning it.

Example 1 *Brownian motion is a Lévy process.*

Lévy processes are characterized by the following expression of their characteristic function:

$$\mathbb{E}\left[e^{i\xi\cdot X_t}\right] = e^{-t\psi(\xi)}, \tag{7}$$

where ψ is called the *characteristic exponent*, and is given by the Lévy-Khintchine formula:

$$\psi(\xi) = ib\cdot\xi + \frac{1}{2}\xi^T\Sigma\xi + \int_{\mathbb{R}^d}\left\{1 - e^{i\xi\cdot x} + i\frac{\xi\cdot x}{1+\|x\|^2}\right\}dL(x), \tag{8}$$

where $b \in \mathbb{R}^d$, Σ is a symmetric, non-negative definite, $d \times d$ matrix and L is the *Lévy measure* (on Borel sets) satisfying[4]: $\int_{x\neq 0}\frac{|x|^2}{1+|x|^2}dL(x) < +\infty$. Intuitively, the first term $ib\cdot\xi$ is a *drift* term, the second ($1/2\xi^T\Sigma\xi$) is a Gaussian term (*i.e.* the characteristic function of a multivariate Gaussian), while the last term encodes information about the jump structure and the (potential) heavy-tailed properties of the process.

---

[3]Càdlàg means that the paths are right continuous and admit left limits everywhere.
[4]Note that there exist several equivalent definitions of Levy measure.

Example 2 *If $b = 0$ and $L = 0$, the characteristic function reduces to a Gaussian, i.e., the process is a Brownian motion.*

The reason why we are interested in the characteristic exponent ψ is because it is deeply related to the Hausdorff dimension of the process, which we use in the proofs of our main results. In particular, we consider the class of *stable* Lévy processes. Those have been extensively studied for their fractal and statistical properties [3, 15, 24, 32, 37]. Stable processes are characterized by their *index* or *tail-exponent*, denoted $\alpha \in (0, 2]$, and the following expression of their Lévy measure:

$$dL(x) = \frac{dr}{r^{1+\alpha}}\nu(d\sigma), \tag{9}$$

with ν a finite measure on the sphere $\mathbb{S}^{d-1}$. One can show [37], that, in that case, the characteristic exponent may be expressed as, with a finite measure M on the sphere. It is defined by

$$\psi(\xi) = ib \cdot \xi + \int_{\mathbb{S}^{d-1}} w_\alpha(\xi, y) dM(y) \;\; (\alpha = 1), \tag{10}$$

where w_α is given by Schilling [28, Equation 3.10]

$$w_\alpha(\xi, y) = \begin{cases} |\xi \cdot y|^\alpha \left(1 - i\mathrm{sgn}(\xi \cdot y)\tan\left(\frac{\pi\alpha}{2}\right)\right), & \text{if } \alpha \neq 1, \\ |\xi \cdot y| \left(1 + \frac{2i}{\pi}\mathrm{sgn}(\xi \cdot y)\log(\xi \cdot y)\right), & \text{if } \alpha = 1. \end{cases}$$

A stable process is said to be *strictly stable* if $b = 0$ and, if $\alpha = 1$, we also require that $\int y dM(y) = 0$. We recognize in these equations the expressions of characteristic functions of stable distributions.

Remark 4 *The notion of stability has several equivalent definitions and is related to stable distribution and the self-similarity of the processes, see [28, Section 3].*

Example 3 *If the distribution of the process is rotationally invariant and has no drift ($b = 0$), then ψ reduces to a simpler form: $\psi(\xi) = C\|\xi\|^\alpha$, see [37, Section 2.1]. In the context of Eq. (3), it means that the noise is rotationally invariant.*

Feller processes: Feller processes are an extension of Lévy processes, they are Markov processes characterized by the fact that their Markov-semigroup operators, $P_t f(w) := \mathbb{E}^w[f(X_t)]$, satisfy regularity properties making them a "Feller semigroup".[5] For more technical details on Feller processes, see [28], here we just summarize some important properties useful in this work. First, note that we can assume that the considered Feller processes also have modifications with almost surely càdlàg paths, see [28, Definition 11.1]. More importantly, a Feller process $(X_t)_{t \geq 0}$ is characterized by its *symbol* $\boldsymbol{\Psi}$, which is related to the pseudo-differential operator representation of its generator. For Lévy processes, the symbol is just the characteristic exponent. The next theorem shows that we can see the dynamics given by Eq. (3) as Feller processes, see [4, Theorem 3.8].

Theorem 7 *Let L_t be a k-dimensional Lévy process, with exponent ψ, and $E : \mathbb{R}^d \longrightarrow \mathbb{R}^{d \times k}$ be a bounded and Lipschitz continuous function. Then the Itô SDE:*

$$dX_t = E(X_{t-})dL_t \quad \text{with} \quad X_0 = x, \tag{11}$$

has a unique strong solution which is a Feller process with the symbol $\boldsymbol{\Psi}(x, \xi) = \psi(E(x)^T \xi)$.

This theorem first ensures that the equations (that are strong Markov processes) considered in this paper have strong solutions. As noted by Schilling [28, Theorem 12.11], the Lipschitz continuity assumption is only needed to prove that those solutions are Feller processes. Moreover, the expression of the symbol Ψ, given in Th. 7, will be used in the proof of Th. 1, to bound the Hausdorff dimension of the processes (that is more formally defined in the next subsection).

---

[5]The notation $\mathbb{E}^w[\cdot]$ denotes that the process under the expectation was initialized with w.

A.2 Fractal Geometry

Let us very quickly describe the notions of Minkowski and Hausdorff dimension, for more details on fractal geometry, see [8, 18] for more details. One of the main goals of fractal geometry is to define various notions of dimensions, associated with any set in $\mathbb{R}^d$, extending the usual notion of dimension (*e.g.*, for vector spaces or Riemannian sub-manifolds). In this paper, those dimensions serve as a complexity measure of the sample paths (*i.e.*, trajectories) of Markov processes, which has been extensively studied in the literature [37]. In our work, we focus on two of the most popular notions of dimension: the box-counting dimension (*a.k.a.* Minkowski dimension) and the Hausdorff dimension. Both are based on the notion of δ-cover.

Definition 1 (Covering of a set) *Let $X \subseteq \mathbb{R}^d$ be a set. A δ-cover is a family (U_i) of sets of diameter at most δ, such that $X \subseteq \bigcup_i U_i$. If X is bounded, a cover of minimal cardinality is called* minimal, *and the corresponding cardinal is the* covering number, *denoted $N_\delta(X)$.*

Based on the notion of δ-cover of a set, we are now able to define the (lower and upper) Minkowski dimensions that are used to characterize the dynamics in our paper (see Asm. 3). Moreover, note that the upper Minkowski dimension is the one that appears most naturally in our proofs.

Definition 2 (Minkowski dimensions) *Given a bounded set $X \subset \mathbb{R}^d$, we define the lower and upper Minkowski dimensions as:*

$$\overline{\dim}_{\mathrm{B}}(X) := \limsup_{\delta \to 0} \frac{\log N_\delta(X)}{\log(1/\delta)}, \quad \underline{\dim}_B(X) := \liminf_{\delta \to 0} \frac{\log N_\delta(X)}{\log(1/\delta)}.$$

If their value coincide, the corresponding limit is the Minkowski dimension and denoted $\dim_B(X)$.

Informally, having a Minkowski dimension m means that, if we reduce δ by a factor of λ, then the number of δ/λ-balls necessary to cover is multiplied by λ^m.

Remark 5 *We use the fact that defining the Minkowski dimensions by using closed or open δ-balls does not change the value of the dimension.*

We end this series of definitions by defining Hausdorff measures and dimensions [8, Chapter 3].

Definition 3 (Hausdorff measures and dimensions) *Let $s \geq 0$ and $X \subset \mathbb{R}^d$, its Hausdorff measure is defined as:*

$$\mathcal{H}^s(X) = \lim_{\delta \to 0} \inf \left\{ \sum_i \mathrm{diam}(U_i)^s, \ (U_i)_i \text{ is a countable } \delta\text{-cover of } X \right\}.$$

The Hausdorff measure $s \longmapsto \mathcal{H}^s(X)$ is decreasing and only takes values ∞ and 0, except on one "jump" point, which is the Hausdorff dimension, i.e.,

$$\dim_{\mathrm{H}}(X) := \inf\{s \geq 0, \ \mathcal{H}^s(X) = 0\}.$$

In general, one has the following inequality between the two notions of dimension that we are interested in in this paper, for a bounded set $X \subset \mathbb{R}^d$:

$$\dim_{\mathrm{H}}(X) \leq \overline{\dim}_{\mathrm{B}}(X). \tag{12}$$

The following proposition gives a sufficient condition for Hausdorff and upper box-counting dimensions to be equal, see [18, Theorem 5.7]:

Proposition 1 *Let $\mathcal{W} \subset \mathbb{R}^d$ be bounded and π a probability measure on $\mathcal{W}$. If there exists, $a, b, r_0, s > 0$ such that, for all $x \in A$ and $0 < r \leq r_0$, we have $ar^s \leq \pi(B(x,r)) \leq br^s$, then:*

$$\overline{\dim}_{\mathrm{B}}(\mathcal{W}) = \dim_{\mathrm{H}}(\mathcal{W}).$$

Therefore, Prop. 1 gives a necessary condition so that Asm. 3 could be satisfied.

It is known that the Hausdorff dimension of an α-*strictly* stable Lévy process is precisely the tail-index α [32, 24, 37]. Let us now describe quickly the fractal properties of Feller processes, which we will express in terms of their symbol $\mathbf{\Psi}$. The following theorem gives a general upper bound of the Hausdorff dimension, see [4, Theorem 5.15] and [16, Theorem 1.2].

Theorem 8 *Let* (X_t) *be a Feller process with symbol* $\Psi(x, \xi)$ *and infinitesimal generator A (see [28, Chapter 5]). Assume that:*

(i) $\mathcal{C}_c^\infty \subseteq \mathcal{D}(A)$, *where* $\mathcal{D}(A)$ *is the domain of the generator A, where* $\mathcal{C}_c^\infty$ *holds for smooth functions with compact support (in* $\mathbb{R}^d$*).*

(ii) $\Psi(\cdot, 0) = 0$ *and* $|\Psi(x, \xi)| \leq c(1 + |\xi|^2)$, *for some constant c,*

then

$$\dim_{\mathrm{H}}(X([0, T])) \leq \beta_\infty := \inf \left\{ \lambda > 0, \quad \lim_{|\xi| \to \infty} \frac{\sup_{|\eta| \leq |\xi|} \sup_{x \in \mathbb{R}^d} |\Psi(x, \eta)|}{|\xi|^\lambda} = 0 \right\},$$

where β_∞ *is the* generalized Blumenthal–Getoor index at infinity.

As we will see in the proof of Th. 1, this theorem is only able to show that $\dim_{\mathrm{H}}(\mathcal{Y}) = \alpha$ for $\alpha \in [1, 2]$. To handle the case $\alpha < 1$, we can use the following theorem, which essentially, in our case, will require a lot of regularity from $w \longmapsto \nabla_w \ell(w, z)$, see [27, Theorem 4]. Note that this theorem is also the one used in [31] to relate the tail index to the Hausdorff dimension of the process, however, it does not seem to be general enough in our particular case, because of the above-mentioned regularity. The following theorem is a *slightly weaker* formulation of [27, Theorem 4].

Theorem 9 (Schilling's theorem) *Let* $(X_t)_{t \geq 0}$ *be a Feller process with symbol* $\Psi(x, \xi)$ *such that it can be decomposed as* $\Psi(x, \xi) = q_1(\xi) + q_2(x, \xi)$, *and the following conditions hold:*

(i) $\Psi(x, 0) = 0$ *and* $q_2(\cdot, \xi) \in \mathcal{C}^{d+1}(\mathbb{R}^d)$,

(ii) $1 + q_1(\xi) \leq C(1 + \|\xi\|^2)$, *where* $C > 0$ *is some constant.*

(iii) *For* $\boldsymbol{i} \in \mathbb{N}^d$, *if* $|\boldsymbol{i}| \leq d + 1$, *then* $|\partial_x^{\boldsymbol{i}} q_2(x, \xi)| \leq \Phi_{\boldsymbol{i}}(x)(1 + \|\xi\|^2)$, *with* $\Phi_0 \in L^\infty(\mathbb{R}^d)$ *and* $\Phi_{\boldsymbol{i}} \in L^1(\mathbb{R}^d)$ *for* $\boldsymbol{i} \neq 0$,

then

$$\dim_{\mathrm{H}}(X([0, T])) \leq \beta_{\text{tail}} := \inf \left\{ \lambda \geq 0, \quad \lim_{\|\xi\| \to \infty} \frac{|q_1(\xi)|}{\|\xi\|^\lambda} = 0 \right\},$$

where β_{tail} *is called the Blumenthal–Getoor index.*

A.3 PAC-Bayesian Bounds

In this section, we introduce two PAC-Bayesian bounds in the literature that are used in the proofs. We first present a PAC-Bayesian bound originally proven by Germain et al. [9, Theorem 2.1]. We will see further that the bound contains a function Φ that must be defined in order to obtain a bound with the (usual) generalization gap. In the following, we consider a family $(\rho_S)_{S \in \mathcal{Z}^n}$ (for all n), called the posteriors and satisfying the following assumption, making it a Markov kernel:

(i) For all S, ρ_S is a probability distribution on $\mathbb{R}^d$

(ii) For all $B \in \mathcal{B}(\mathbb{R}^d)$, the mapping $S \longmapsto \rho_S(B)$ is measurable, where $\mathcal{B}(\mathbb{R}^d)$ denotes the Borel σ-algebra of $\mathbb{R}^d$.

Theorem 10 (General PAC-Bayesian bound) *Given a function* $\Phi : \mathbb{R}^d \times \mathcal{Z}^n \to \mathbb{R}$, *with probability at least* $1 - \zeta$ *over* $S \sim \mu_z^{\otimes n}$ *we have*

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

where

$$\mathrm{KL}(\rho_S | |\pi) = \int \ln \left(\frac{d\rho_S}{d\pi}(w) \right) d\rho_S(w)$$

is the KL divergence between the posterior distribution ρ_S *and the prior distribution* π, *and* $\frac{d\rho_S}{d\pi}(w)$ *is the Radon-Nikodym derivative, defined when* $\rho_S \ll \pi$.

Put into words, the bound is on the expectation of Φ over ρ_S and depends mainly on the KL divergence and the term $\log \mathbb{E}_S \mathbb{E}_\pi \left[e^{\Phi(w,S)} \right]$ that is further upper-bounded when instantiating Φ. Note that the theorem of Germain et al. [9] is originally presented with a function representing a gap between the population risk and the empirical risk, however, their theorem holds with such a function $\Phi : \mathbb{R}^d \times \mathcal{Z}^n \to \mathbb{R}$.

When we consider a bound holding with high probability over $S \sim \mu_z^{\otimes n}$ and $w \sim \rho_S$, we use the bound of Viallard et al. [35] (that is called a disintegrated PAC-Bayesian bound).

Theorem 11 (Disintegrated PAC-Bayesian bound) *Given a function* $\Phi : \mathbb{R}^d \times \mathcal{Z}^n \to \mathbb{R}$, *we have, with probability at least* $1 - \zeta$ *over* $S \sim \mu_z^{\otimes n}$ *and* $w \sim \rho_S$:

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

where

$$\mathrm{D}_\beta \left(\rho_S | \, | \pi \right) = \frac{1}{\beta - 1} \log \left(\int \left(\frac{d\rho_S}{d\pi}(w) \right)^\beta d\pi(w) \right)$$

is the Rényi divergence between ρ_S *and* π.

Similarly as for Th. 10, the function Φ must be defined and the term $\log \mathbb{E}_S \mathbb{E}_{w \sim \pi} \left[e^{\frac{\beta}{\beta-1} \Phi(w,S)} \right]$ must be upper-bounded to obtain a bound with the generalization error; this is actually done in the proofs.

A.4 Egoroff's theorem

In our proofs, we will use the following theorem, to get uniform convergence of the limits defining the Minkowski dimensions. This has been used in other works [31, 5, 7].

Theorem 12 (Egoroff's theorem) *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability space and* $f, (f_n)_{n \in \mathbb{N}}$ *be measurable functions. Assume that, for almost all* $x \in \Omega$, *we have* $f_n(x) \xrightarrow[n \to \infty]{} f(x)$. *Then, for any* $\epsilon > 0$ *there exists* $\Omega_\epsilon \in \mathcal{F}$ *such that* $\mathbb{P}(\Omega_\epsilon) \ge 1 - \epsilon$ *and on which the convergence of* f_n *to* f *is uniform.*

B   Additional Results

The goal of this section is to show that, under an additional reasonable assumption, we may render our PAC-Bayesian bounds time-independent. The following assumption is similar to the so-called "dissipativity" assumption, commonly used in the stochastic calculus literature. This special form is more adapted to the simultaneous consideration of two dynamics and has recently been made by Chen et al. [6] and Raj et al. [25].

Assumption 5 (Co-dissipativity) *There exists* $m > 0$ *and* $K \ge 0$ *such that:*

$$\langle \nabla \ell(w', z) - \nabla \ell(w, z), w - w' \rangle \le K - m \| w - w' \|^2,$$

where all gradients are with respect to the w *variable.*

This assumption will be used to ensure that the difference between the empirical and expected dynamics does not become too large. Under this assumption, we can tighten the bound on the distance between the empirical and the expected dynamics.

Theorem 13 (Bound under co-dissipativity) *We make Asms. 2 and 5 and with a loss* ℓ *that is* M-*smooth in* w *(but not Lipschitz). We further assume that there exists* $w_0 \in \mathbb{R}^d$ *such that* $\nabla \ell(w_0, z) \in L^1(\mu_z)$. *Then, for* n *large enough, with probability at least* $1 - \zeta$ *over* $\mu_u \otimes \mu_z^{\otimes n}$, *we have, almost surely with respect to* (S, U):

$$\sup_{0 \le t \le T} \| W_t^S - Y_t \|^2 \le \frac{2K}{m} + \frac{4}{m^2} \left\{ \frac{2}{n} + \frac{\Sigma^2}{n} \left(2\gamma \log(M\sqrt{n}) + \log(2d/\zeta) \right) \right\}.$$

Remark 6 *If, as it is the case in the above theorem, we do not make the Lipschitz assumption on the loss ℓ, we cannot ensure anymore that W^S and Y are Feller processes, see [28, Theorem 12.11]. However, we can still argue that the process has a càdlàg modification, as a consequence of martingale regularization theorems [26, Section 2.2], and hence can show that the process has almost surely bounded paths. This makes that the upper box-counting dimensions, $\overline{\dim}_{\mathrm{B}}(W^S([0,T]))$ and $\overline{\dim}_{\mathrm{B}}(Y([0,T]))$ are well defined. However, we may not be able to formally relate it to α. In the above theorem, the notation γ therefore refers to the upper box-counting dimension.*

C Postponed Proofs

C.1 Technical Lemmas

In this subsection, we quickly prove a few technical results, especially measure theoretic ones. The goal is to ensure solid theoretical foundations for our work. Some of the results of this section are used without further notice throughout the proofs.

Lemma 1 (Almost surely càdlàg paths) *Y and W^S are almost surely càdlàg w.r.t. $\mu_z^{\otimes n} \otimes \mu_u$.*

Proof As mentioned in Rem. 1, we have almost surely, under $\mu_z^{\otimes n} \otimes \mu_u$:

$$W_t^S = W_0 - \int_0^t \nabla \widehat{\mathcal{R}}_S(W_u^S) du + L_t^\alpha.$$

By Asm. 1, the gradient $\nabla \widehat{\mathcal{R}}_S$ is bounded, so that the integral above is a continuous function of t, as L_t^α is an almost surely càdlàg version, it ensures that W^S is almost surely càdlàg with respect to $\mu_z^{\otimes n} \otimes \mu_u$. We can do the same reasoning for Y.

Remark 7 *As mentioned in [28, Section 11], we could have argued that W^S and Y are Feller processes, and therefore we may take càdlàg versions of them. However, for measure-theoretic reasons, it may be safer to directly use the Lipschitz assumption as we did.*

Remark 8 *If we do not make the Lipschitz assumption, we could still prove that the processes W^S and Y have a càdlàg modification, by using the fact that they are (local) semi-martingales, see [28, Theorem 12.11].*

We define the difference process as (the dependence on U is omitted):

$$V_t^S = W_t^S - Y_t \tag{13}$$

One important result is the following:

Lemma 2 (Joint continuity of the difference process) *The map $t \longmapsto V_t^S$ is almost surely continuous under $\mu_z^{\otimes n} \otimes \mu_u$.*

Proof We have, according to Rem. 1, almost surely under $\mu_z^{\otimes n} \otimes \mu_u$ that:

$$V_t^S = \int_0^t \left(-\nabla \widehat{\mathcal{R}}_S(W_u^S) + \nabla \mathcal{R}(Y_u) \right) du$$

Therefore $t \longmapsto V_t^S$ is almost surely continuous. It may be seen either using the boundedness of the gradient, as in the previous proof, or by the fact that càdlàg paths are bounded on compact intervals. Note that the fact that both SDEs have the same initialization is crucial here.

Lemma 3 *Under assumptions 1, the processes W^S and Y are Feller processes.*

Proof This is a direct consequence of Asm. 1 and Th. 7.

Let's continue with a remark regarding covering arguments made in this paper. We will use minimal δ-covers of the trajectory $\mathcal{Y}_U$ of Y. By using arguments similar to Dupuis et al. [7], we show that we can get measurable covering numbers that yield the correct Minkowski dimensions. This justifies the validity of subsequent covering arguments.

Lemma 4 *Let us fix some $\delta > 0$ and $T > 0$, then the mapping $U \longmapsto |N_\delta(\overline{\mathcal{Y}_U})|$ is measurable, if open balls are used for the definition of minimal coverings[6].*

Proof The proof works by constructing a kind of "Castaing representation" [21, Definition 1.3.6]. For any $q \in \mathbb{Q}_+$, we set $\xi_q(U) := Y_q(U)$. By the almost surely càdlàg property of Y, it is clear that $\overline{\{\xi_q\}} = \mathcal{Y}$, almost surely. Moreover, with $\mathbb{Q}_T := \mathbb{Q} \cap [0,T]$:

$$|N_\delta(\overline{\mathcal{Y}_U})| > N = \bigcap_{F \in \mathbb{Q}_T^N} \bigcup_{q \in \mathbb{Q}_T} \bigcap_{x \in F} \{|Y_q(U) - Y_x(U)| \geq \delta\},$$

which, together with the completeness of the space $(\Omega, \mathcal{T}, \mu_u)$, implies the desired measurability.

To end this technical section, we mention the measurability of the mapping $(S, U) \longmapsto \sup_{w \in \mathcal{W}_{S,U}} \left(\mathcal{R}(w) - \widehat{\mathcal{R}}_S(w) \right)$, which is one of the main interests of our work. From the almost surely càdlàg property of the process, we have, almost surely:

$$\sup_{w \in \mathcal{W}_{S,U}} \left(\mathcal{R}(w) - \widehat{\mathcal{R}}_S(w) \right) = \sup_{t \in \mathbb{Q} \cap [0,T]} \left(\mathcal{R}(W_t^S(U)) - \widehat{\mathcal{R}}_S(W_t^S(U)) \right),$$

which, by measurability of $(t, S, U) \longmapsto W_t^S(U))$, implies the desired measurability, if the space $\mathcal{Z}^n \times \Omega$ is complete. We can assume it is the case, in the worst case, considering its completion.

C.2 Proof of Th. 1

In the rest of the paper, we will use the notation ∂_j as a shortcut for the partial derivative $\partial / \partial w_j$, all considered partial derivatives are implicit with respect to w, even if it is not mentioned. Moreover, given a multi-index $\boldsymbol{i} \in \mathbb{N}^d$, the notation $\partial_w^{\boldsymbol{i}}$ is a shortcut for $\frac{\partial}{\partial^{i_1} w_1 \dots \partial^{i_d} w_d}$. We also denote $|\boldsymbol{i}| = i_1 + \cdots + i_d$.

Theorem 1 (Fractal properties of heavy-tailed dynamics (informal)) *Under Asms. 1 and 2, if one of the following condition is true: (i) $\alpha \in [1, 2]$, or (ii) L_t^α is strictly stable and $\nabla \ell$ is smooth enough in w (compact support is enough for instance) so that Y and W^S satisfy the assumptions of [27, Theorem 4], then $\dim_H(\mathcal{Y}_U), \dim_H(\mathcal{W}_{S,U}) \leq \alpha$. Hence, under Asm. 3, we have $\gamma \leq \alpha$.*

Before proving this theorem, let us give its more formal version.

Theorem 14 (Fractal properties of heavy-tailed dynamics) *Under Asms. 1 and 2, Eqs. (3) and (4), if one of the following condition is true:*

(i) $\alpha \in [1, 2]$, or

(ii) L_t^α is strictly stable and, for every $z \in \mathcal{Z}$ and every multi-index $\boldsymbol{i} \in \mathbb{N}^d$, with $1 \leq |\boldsymbol{i}| \leq d+1$, the function $w \longmapsto \left\| \partial_w^{\boldsymbol{i}} \nabla \ell(w, z) \right\|$ is in $L^1(\mathbb{R}^d)$,

then $\dim_H(\mathcal{Y}_U), \dim_H(\mathcal{W}_{S,U}) \leq \alpha$, hence, under Asm. 3, $\gamma \leq \alpha$.

Proof In this proof, F will denote a function that can be both $\mathcal{R}$ or $\widehat{\mathcal{R}}_S$, without changing any of the results or computations. In the same way, the Feller process solution of the associated SDE (existence ensured by Th. 7) will be denoted by (X_t), it is therefore either Y_t or W_t^S.

Let us denote by ψ_α the characteristic exponent[7] of L_t^α. The Itô SDEs Eqs. (3) and (4) may be written in the form of Eq. (11), with $L_t := (L_t^\alpha, t)$ and $E(x) = (I_d, -\nabla F(x)) \in \mathbb{R}^{d \times (d+1)}$. From Asm. 1, ∇F is Lipschitz and bounded, hence, by Th. 7, the symbol of the solution of the equation, X_t, is:

$$\boldsymbol{\Psi}(x, \xi) = \psi(\xi) - i\nabla F(x) \cdot \xi.$$

Case 1: when $\alpha \in [1, 2]$. We look at the generalized Blumenthal–Getoor index of Th. 8. Let $\lambda > \alpha$, we have, by the Lipschitz assumption (Asm. 1):

$$\sup_{|\eta| \leq |\xi|} \sup_{x \in \mathbb{R}^d} |\boldsymbol{\Psi}(x, \eta)| \leq \sup_{|\eta| \leq |\xi|} (|\psi(\eta)| + L \|\eta\|)$$

---

[6]According to Rem. 5, this does not change the value of the dimensions

[7]Like in [37], we implicitly assume that all stable distributions are non-degenerate

From the explicit expressions of the characteristic exponent of stable processes, given in [37, Section 2.1] and [24], we have

$$|\psi(\xi)| \leq C_\alpha \|\xi\| \left(1 + \log \|\xi\|\right) + \mu_0 \|\xi\|$$

(the log term handles the case of a Cauchy process), where C and μ_0 are constants. From this we deduce that $\beta_\infty \leq \alpha$, hence, by Th. 8, we have:

$$\dim_H(X([0,T])) \leq \alpha.$$

Case 2**:** In that case, even if the process is strictly stable, the drift term in the symbol $\mathbf{\Psi}$, coming from the SDE, makes that the previous reasoning proves that $\dim_H(L_t^\alpha([0,T])) \leq 1$, if $\alpha < 1$. To relate the Hausdorff dimension $\dim_H(X([0,T]))$ to α in that case, we make the assumptions of the second part of the theorem. As it is now assumed that L_t^α is now strictly stable, it is known that we have:

$$\dim_H(L_t^\alpha([0,T])) \leq \beta_{\text{tail}} = \alpha;$$

see [4, Theorem 5.12] or [23, 37]. Thanks to the smoothness assumptions on $\nabla_w \ell$, Th. 9, with $q_1 = \psi$, directly implies that:

$$\dim_H(X([0,T])) \leq \alpha.$$

C.3 Proof of Th. 2

As mentioned earlier, we will prove a slightly stronger version of Th. 2, featuring the Hausdorff distance between both trajectories. Let us recall the definition of Hausdorff distance.

Definition 4 (Hausdorff distance) *Let $A, B \subset \mathbb{R}^d$ be two non-empty bounded sets, the Hausdorff distance between A and B is defined by:*

$$d_H(A, B) = \max \left\{ \sup_{a \in A} d(a, B), \; \sup_{b \in B} d(b, A) \right\},$$

where $d(a, B)$ denotes the (Euclidean) distance between the point a and the set B i.e.

$$d(a, B) = \inf_{b \in B} \|a - b\|.$$

We are now ready to prove Th. 2.

Theorem 2 *Fix some $\zeta \in (0, 1)$. Under Asms. 1 to 3, there exists $N \in \mathbb{N}$, depending only on ζ such that, with probability at least $1 - 2\zeta$ over $\mu_u \otimes \mu^{\otimes n}$, for all $n \geq N$, we have:*

$$\sup_{w \in \mathcal{W}_{S,U}} \left(\mathcal{R}(w) - \widehat{\mathcal{R}}_S(w) \right) \leq 2L \sup_{0 \leq t \leq T} \|W_t^S - Y_t\| + \frac{2}{\sqrt{n}} + 2\sigma \sqrt{\frac{2\log(L\sqrt{n})\gamma + \log(1/\zeta)}{2n}}.$$

Proof First note that we have:

$$\sup_{w \in \mathcal{W}_{S,U}} d(w, \mathcal{Y}_U) = \sup_{0 \leq t \leq T} d(W_t^S, \mathcal{Y}_U) \leq \sup_{0 \leq t \leq T} \|W_t^S - Y_t\| = \sup_{0 \leq t \leq T} \|W_t^S - Y_t\|,$$

and idem for $d(\mathcal{W}_{S,U}, y)$, thus we have

$$d_H(\mathcal{W}_{S,U}, \mathcal{Y}_U) \leq \sup_{0 \leq t \leq T} \|W_t^S - Y_t\|.$$

Therefore, it is enough to prove the bound for the Hausdorff distance, instead of $\sup_{0 \leq t \leq T} \|W_t^S - Y_t\|$. Note that the Hausdorff distance $d_H(\mathcal{W}_{S,U}, \mathcal{Y}_U)$ is almost surely finite because of the càdlàg property of the paths (or their modification). Indeed, càdlàg paths are bounded [28, Section 9].

Let us fix some $\eta > 0$ and $S \in \mathcal{Z}^n$, by definition of the Hausdorff distance, for all $w \in \mathcal{W}_{S,U}$, there exists $y \in \mathcal{Y}_U$ such that $\|w - y\| \leq d_H(\mathcal{W}_{S,U}, \mathcal{Y}_U) + \eta$. We also have, by Lipschitz regularity:

$$\mathcal{R}(w) - \widehat{\mathcal{R}}_S(w) = \mathcal{R}(w) - \widehat{\mathcal{R}}_S(w) + \widehat{\mathcal{R}}_S(y) - \widehat{\mathcal{R}}_S(y) + \mathcal{R}(y) - \mathcal{R}(y)$$

$$\leq |\mathcal{R}(w) - \mathcal{R}(y)| + |\widehat{\mathcal{R}}_S(y) - \widehat{\mathcal{R}}_S(w)| + \mathcal{R}(y) - \widehat{\mathcal{R}}_S(y)$$

$$= 2L\|w - y\| + \mathcal{R}(y) - \widehat{\mathcal{R}}_S(y).$$

16

Therefore, we have

$$\sup_{w \in \mathcal{W}_{S,U}} \left(\mathcal{R}(w) - \widehat{\mathcal{R}}_S(w) \right) \leq 2Ld_H(\mathcal{W}_S, \mathcal{Y}) + 2L\eta + \sup_{w \in \mathcal{Y}_U} \left(\mathcal{R}(w) - \widehat{\mathcal{R}}_S(w) \right).$$

From there, we bound the term $\sup_{w \in \mathcal{Y}_U} \left(\mathcal{R}(w) - \widehat{\mathcal{R}}_S(w) \right)$ using a reasoning similar to the proof of [31, Theorem 1]. We fix some decreasing sequence (δ_n) such that $\delta_n > 0$. Let $w_1, \ldots, w_{N_{\delta_n}}$ be the centers of a minimal cover of $\mathcal{Y}_U$ with balls of radius at most δ_n (note that the covering number is random because of its dependence on U, when necessary we will denote it $N_{\delta_n}(U)$). By the same calculation as above, we have:

$$\sup_{w \in \mathcal{Y}_U} \left(\mathcal{R}(w) - \widehat{\mathcal{R}}_S(w) \right) \leq 2L\delta_n + \max_{1 \leq i \leq N_{\delta_n}} \left(\mathcal{R}(w_i) - \widehat{\mathcal{R}}_S(w_i) \right).$$

By Hoeffding's inequality, along with a union bound, we get, for any $\epsilon > 0$:

$$\mathbb{P}_S \left(\max_{1 \leq i \leq N_{\delta_n}} \left(\mathcal{R}(w_i) - \widehat{\mathcal{R}}_S(w_i) \right) \geq \epsilon \right) \leq N_{\delta_n}(U) \exp \left\{ -\frac{n\epsilon^2}{2\sigma^2} \right\}.$$

Let $\zeta \in (0, 1)$, we set:

$$\epsilon(U) := \sqrt{2}\sigma \sqrt{\frac{\log(N_{\delta_n}(U)) + \log(1/\zeta)}{n}},$$

and $\delta_n = 1/(L\sqrt{n})$. By definition of the upper Minkowski dimension, we know that:

$$\lim_{k \to \infty} \sup_{n \geq k} \frac{\log(N_{\delta_n}(U))}{\log(1/\delta_n)} = \overline{\dim}_B(\mathcal{Y}_U)$$

By Egoroff's theorem (see Th. 12), there exists $\Omega_\zeta \in \mathcal{T}$, such that $\mu_u(\Omega_\zeta) \geq 1 - \zeta$ and on which the above convergence is uniform in U. Moreover, thanks to Asm. 3, we have $\overline{\dim}_B(\mathcal{Y}_U) = \gamma > 0$. From all this we deduce that there exists N_ζ, depending only on ζ, such that, for all $n \geq N_\zeta$, we have that the limit is smaller than 2γ, and therefore:

$$\log(N_{\delta_n}(U)) \leq 2\log(L\sqrt{n})\gamma.$$

Which implies, for all $n \geq N_\zeta$:

$$\sup_{w \in \mathcal{W}_{S,U}} \left(\mathcal{R}(w) - \widehat{\mathcal{R}}_S(w) \right) \leq 2Ld_H(\mathcal{W}_S, \mathcal{Y}) + 2L\eta + \frac{2}{\sqrt{n}} + \sqrt{2}\sigma \sqrt{\frac{2\log(L\sqrt{n})\gamma + \log(1/\zeta)}{n}}.$$
$$(14)$$

As this is true for all η without changing any of the quantity appearing in Eq. (14), we can take $\eta \to 0$ and rearrange the terms to get that, for all $n \geq N_\zeta$:

$$\sup_{w \in \mathcal{W}_{S,U}} \left(\mathcal{R}(w) - \widehat{\mathcal{R}}_S(w) \right) \leq 2Ld_H(\mathcal{W}_S, \mathcal{Y}) + \frac{2}{\sqrt{n}} + 2\sigma \sqrt{\frac{2\log(L\sqrt{n})\gamma + \log(1/\zeta)}{2n}}.$$

By a union bound, the above event occurs with probability $1 - 2\zeta$ (it is the combination of the probabilities coming from Hoeffding's inequality and Egoroff's theorem).

C.4 Proof of Th. 3

Before proving Th. 3, we present a lemma to bound the following quantity, which is the worst-case deviation of the empirical gradient from its expectation over the trajectory $\mathcal{Y}_U$. It is defined by

$$G_\nabla := \sup_{w \in \mathcal{Y}_U} \|\nabla \widehat{\mathcal{R}}_S(w) - \nabla \mathcal{R}(w)\|. \tag{15}$$

Remark 9 *Note that thanks to the assumptions, we have boundedness of the gradients $\nabla_w \ell$. Therefore, by the dominated convergence theorem, we will be able to invert the expectation and the partial derivative, which we will do repeatedly in our proofs.*

Lem. 5 will be used for the proof of Th. 3.

Lemma 5 *For n big enough, with probability at least $1 - \zeta$ over $\mu_u \otimes \mu_z^{\otimes n}$, we have:*

$$\sup_{w \in \mathcal{Y}_U} \|\nabla\widehat{\mathcal{R}}_S(w) - \nabla\mathcal{R}(w)\| \leq \frac{2}{\sqrt{n}} + L\sqrt{\frac{2}{n}} + 2L\sqrt{\frac{2\gamma \log(M\sqrt{n}) + \log(1/\zeta)}{2n}}.$$

If, in addition, we make Asm. 4, then the result becomes:

$$\sup_{w \in \mathcal{Y}_U} \|\nabla\widehat{\mathcal{R}}_S(w) - \nabla\mathcal{R}(w)\| \leq \frac{2}{\sqrt{n}} + 2\Sigma\sqrt{\frac{2\gamma \log(M\sqrt{n}) + \log(2d/\zeta)}{2n}}.$$

In the proofs and the rest of the document, the notation ∂_j will be used as a shortcut for $\frac{\partial}{\partial w_j}$.

Proof We introduce the notation, for $S = (z_, \ldots, z_n)$:

$$\overline{\partial_j}(w, z_i) := \partial_j \ell(w, z_i) - \mathbb{E}_z\left[\partial_j(w, z)\right].$$

As in the proof of Th. 2, let us introduce a sequence (δ_n), positive, decreasing and converging to 0, as well as $w_1, \ldots, w_{N_{\delta_n}}$ the centers of a minimal cover of $\mathcal{Y}_U$ with balls of radius at most δ_n. As before, this covering number still depends on U. We will denote

$$D_S(w) := \|\nabla\widehat{\mathcal{R}}_S(w) - \nabla\mathcal{R}(w)\|.$$

We have, by Jensen's inequality, independence, and the Lipschitz loss assumption:

$$
\begin{aligned}
\mathbb{E}_S[D_S] &\leq \sqrt{\mathbb{E}[\|\nabla\widehat{\mathcal{R}}_S(w) - \nabla\mathcal{R}(w)\|^2]} \\
&= \frac{1}{n}\left\{\sum_{j=1}^d \mathbb{E}\left[\sum_{1 \leq i,k \leq n} \overline{\partial_j}(w, z_i)\overline{\partial_j}(w, z_k)\right]\right\}^{\frac{1}{2}} \\
&= \frac{1}{n}\left\{\sum_{j=1}^d \mathbb{E}\left[\sum_{i=1}^n \overline{\partial_j}(w, z_i)^2\right]\right\}^{\frac{1}{2}} \\
&= \frac{1}{n}\left\{\mathbb{E}\left[\sum_{i=1}^n \|\nabla\ell(w, z_i) - \nabla\mathcal{R}(w)\|^2\right]\right\}^{\frac{1}{2}} \\
&\leq L\sqrt{\frac{2}{n}}.
\end{aligned}
$$

Moreover, if S^i denotes $S \in \mathcal{Z}^n$, but with i-th element z_i replaced by some $z_i' \in \mathcal{Z}$, we have, by the inverted triangle inequality and the Lipschitz assumption on the loss:

$$|D_S(w) - D_{S^i}(w)| \leq \|\nabla\widehat{\mathcal{R}}_S(w) - \nabla\widehat{\mathcal{R}_{S^i}}(w)\| \leq \frac{2L}{n}.$$

Therefore, by McDiarmid inequality, for any w, we have:

$$\mathbb{P}_S\left(D_S(w) - \mathbb{E}\left[D_S(w)\right] \geq \epsilon\right) \leq \exp\left\{-\frac{n\epsilon^2}{2L^2}\right\}$$

The rest of this part of the proof is similar to what we did in the proof of Th. 2. We apply a covering argument similar to the proof of Th. 2, but this time using the M-smoothness of ℓ. Let $w \in \mathcal{Y}_U$, and $(w_1, \ldots, w_{N_{\delta_n}})$ be the centers of a δ_n-cover of $\mathcal{Y}_U$, we have that there exists i such that $\|w - w_i\| \leq \delta_n$, therefore, by smoothness:

$$\|\nabla\widehat{\mathcal{R}}_S(w) - \nabla\mathcal{R}(w)\| \leq 2M\delta_n + D_S(w_i) \leq 2M\delta_n + L\sqrt{\frac{2}{n}} + (D_S(w_i) - \mathbb{E}_S\left[D_S(w_i)\right]).$$

Thus:

$$\sup_{w \in \mathcal{Y}_U} \|\nabla\widehat{\mathcal{R}}_S(w) - \nabla\mathcal{R}(w)\| \leq \max_{1 \leq i \leq N_{\delta_n}} (D_S(w_i) - \mathbb{E}_S\left[D_S(w_i)\right]) + L\sqrt{\frac{2}{n}} + 2M\delta_n.$$

Using the independence between $\mathcal{Y}_U$ and a union bound, we get:

$$\mathbb{P}_S \left(\max_{1 \leq i \leq N_{\delta_n}} (D_S(w_i) - \mathbb{E}_S \left[D_S(w_i) \right]) \geq \epsilon \right) \leq N_{\delta_n} \exp \left\{ -\frac{n\epsilon^2}{2L^2} \right\}.$$

Now we set $\delta_n := 1/M\sqrt{n}$. Using the same reasoning than in Th. 2, based on Egoroff's theorem applied on μ_u, we that, for n big enough (depending on ζ), with probability at least $1 - 2\zeta$:

$$\sup_{w \in \mathcal{Y}_U} \|\nabla \widehat{\mathcal{R}}_S(w) - \nabla \mathcal{R}(w)\| \leq \frac{2}{\sqrt{n}} + L\sqrt{\frac{2}{n}} + 2L\sqrt{\frac{2\gamma \log(M\sqrt{n}) + \log(1/\zeta)}{2n}}.$$

With additional Asm. 4: We use the same covering and notations, but this time we write:

$$\sup_{w \in \mathcal{Y}_U} \|\nabla \widehat{\mathcal{R}}_S(w) - \nabla \mathcal{R}(w)\| \leq 2M\delta_n + \max_{1 \leq i \leq N_{\delta_n}} \sqrt{d \max_{1 \leq j \leq d} \left| \frac{1}{n} \sum_{i=1}^{n} \overline{\partial_j}(w, z_i) \right|^2}.$$

From our additional assumption, we know that $\|\overline{\partial_j}(w, z)\|_{\psi_2} = \Sigma/\sqrt{d}$ (where $\| \cdot \|_{\psi_2}$ is the sub-Gaussian norm), therefore by a union bound and Hoeffding's inequality:

$$\mathbb{P}_S \left(\max_{1 \leq i \leq N_{\delta_n}} \max_{1 \leq j \leq d} \left| \frac{1}{n} \sum_{i=1}^{n} \overline{\partial_j}(w, z_i) \right| \geq \epsilon \right) \leq 2dN_{\delta_n} \exp \left\{ -\frac{nd\epsilon^2}{2\Sigma^2} \right\}.$$

Therefore, we set (remembering the U dependence here):

$$\epsilon(U) := \sqrt{2}\Sigma \sqrt{\frac{\log(N_{\delta_n}(U)) + \log(2d/\zeta)}{nd}}.$$

We use again an argument based on Egoroff's theorem on μ_u to write that, for N big enough (depending on ζ), we have with probability $1 - \zeta$ that:

$$\sup_{w \in \mathcal{Y}_U} \|\nabla \widehat{\mathcal{R}}_S(w) - \nabla \mathcal{R}(w)\| \leq \frac{2}{\sqrt{n}} + 2\Sigma\sqrt{\frac{2\gamma \log(M\sqrt{n}) + \log(2d/\zeta)}{2n}}.$$

We are now ready to prove Th. 3.

Theorem 3 *Let us fix some $\zeta \in (0, 1)$. Under Asms. 1 to 3, there exists $N \in \mathbb{N}$, depending on ζ, such that, with probability at least $1 - 2\zeta$ over $\mu_u \otimes \mu^{\otimes n}$, for all $n \geq N$, we have:*

$$\sup_{0 \leq t \leq T} \|W_t^S - Y_t\| \leq \frac{e^{MT} - 1}{M} \left(\frac{2}{\sqrt{n}} + L\sqrt{\frac{2}{n}} + 2L\sqrt{\frac{2\gamma \log(M\sqrt{n}) + \log(1/\zeta)}{2n}} \right).$$

If we further make Asm. 4, we can improve the constants to:

$$\sup_{0 \leq t \leq T} \|W_t^S - Y_t\| \leq \frac{e^{MT} - 1}{M} \left(\frac{2}{\sqrt{n}} + 2\Sigma\sqrt{\frac{2\gamma \log(M\sqrt{n}) + \log(2d/\zeta)}{2n}} \right).$$

Proof We know from Th. 1, that the Feller processes W^S and Y have càdlàg paths (the variable U is omitted). Thanks to App. C.1, we know that the Feller processes W^S and Y have $\mu_z^{\otimes n} \otimes \mu_u$-almost surely càdlàg paths and that the difference $V^S := W^S - Y$ is continuous (almost surely). Therefore, this continuity will be assumed in this proof without losing generality. As a consequence, the paths are almost surely bounded, so it makes sense to introduce their upper box-counting dimension.

Let us take some $S \in \mathcal{Z}^n$, we have, almost surely, for $t \leq T$:

$$\|V_t^S\| = \left\| \int_0^t \left(-\nabla \widehat{\mathcal{R}}_S(W_\tau^S) + \nabla \mathcal{R}(Y_\tau) \right) d\tau \right\|$$

$$\leq \int_0^t \|\nabla \widehat{\mathcal{R}}_S(W_\tau^S) - \nabla \mathcal{R}(Y_\tau)\| d\tau$$

$$\leq \int_0^t \|\nabla \widehat{\mathcal{R}}_S(W_\tau^S) - \nabla \widehat{\mathcal{R}}_S(Y_\tau)\| d\tau + \int_0^t \|\nabla \widehat{\mathcal{R}}_S(Y_\tau) - \nabla \mathcal{R}(Y_\tau)\| d\tau$$

$$\leq M \int_0^t \|V_\tau^S\| d\tau + t \sup_{w \in \mathcal{Y}_U} \|\nabla \widehat{\mathcal{R}}_S(w) - \nabla \mathcal{R}(w)\|.$$

From which we deduce:

$$\|V_t^S\| + \frac{1}{M} \sup_{w \in \mathcal{Y}_U} \|\nabla \widehat{\mathcal{R}}_S(w) - \nabla \mathcal{R}(w)\| \le M \int_0^t \left(\|V_\tau^S\| + \frac{1}{M} \sup_{w \in \mathcal{Y}_U} \|\nabla \widehat{\mathcal{R}}_S(w) - \nabla \mathcal{R}(w)\| \right) d\tau$$
$$+ \frac{1}{M} \sup_{w \in \mathcal{Y}_U} \|\nabla \widehat{\mathcal{R}}_S(w) - \nabla \mathcal{R}(w)\|.$$

By Grönwall's lemma, thanks to the continuity of V_t^S:

$$\|V_t^S\| \le \frac{1}{M} \sup_{w \in \mathcal{Y}_U} \|\nabla \widehat{\mathcal{R}}_S(w) - \nabla \mathcal{R}(w)\| \left(e^{Mt} - 1 \right).$$

The results immediately follow by Lem. 5 and rearranging terms.

Remark 10 *The correct way of writing Eqs.* (3) *and* (4) *would be:*

$$dW_t^S = -\nabla \mathcal{R}(W_{t-}^S)dt + L_t^\alpha, \quad dY_t = -\nabla \mathcal{R}(Y_{t-})dt + L_t^\alpha.$$

It is clear, from the proof, that this detail does not create any issue, as the function $t \longmapsto V_t^S$ is still almost surely continuous.

C.5 Proof of our PAC-Bayesian Bounds

In all this section, the notations π_U and $\rho_{S,U}$ are the ones defined in Sec. 4, as well as the unperturbed posteriors and priors $\tilde{\rho}_{S,U}$ and $\tilde{\pi}_U$. Let us also recall what we mean by the convolution of a measure with a function. Let μ be a finite Borel measure on $\mathbb{R}^d$ and f be a bounded measurable function, we define their convolution as:

$$f * \mu(x) = \int_{\mathbb{R}^d} f(x - y)d\mu(y). \tag{16}$$

When f is a probability density, we will also denote by $f * \mu$ the associated probability measure. We now quickly recall the definition of Wasserstein distance.

Definition 5 (Wasserstein distance) *Let μ and ν be two probability measures on $\mathbb{R}^d$ and $p > 0$. A coupling between μ and ν is a probability measure on $\mathbb{R}^d \times \mathbb{R}^d$ whose marginals are μ and ν. If $\Pi(\mu, \nu)$ denote the set of all possible such couplings, then we define the p-Wasserstein distance by:*

$$W_p(\mu, \nu) := \inf_{\gamma \in \Pi(\mu, \nu)} \left\{ \iint \|x - y\|^p \, d\gamma(x, y) \right\}^{\frac{1}{p}}.$$

Remark 11 (The PAC-Bayes analysis works in this non-classical setting) *Our framework may seem slightly non-standard, as the prior π_U is random, because of its dependence on U. However, we argue that, due to the independence of π_U on S, the theorems from classical PAC-Bayesian theory, presented in App. A.3, are still valid. Moreover, one can see, from our measurability assumptions of Rem. 1 and App. C.1, that π_U and $\rho_{S,U}$ are well-defined Markov kernels.*

C.5.1 Bounds on Information-theoretic Distances

In this subsection, we prove that the quantities $\mathrm{KL}\left(\rho_{S,U} | | \pi_U\right)$ and $\mathrm{D}_\alpha\left(\rho_{S,U} | | \pi_U\right)$ can be bounded by the geometric distance. The next lemma first takes care of the KL divergence, it is a simple application of Jensen's inequality, using the fact that the KL divergence is jointly convex in its argument [33, Theorem 11], we just carefully verify that it works in a continuous setting.

Lemma 6 *With the previous notations, we have almost surely with respect to S and U:*

$$\mathrm{KL}\left(\rho_{S,U} | | \pi_U\right) \le \frac{1}{2s^2} W_2(\tilde{\rho}_{S,U}, \tilde{\pi}_U)^2 \le \frac{1}{2s^2 T} \int_0^T \left\| W_t^S - Y_t \right\|^2 dt.$$

The first inequality in this Theorem has already been used in [17, Section 4.3]. For the sake of clarity, we present a proof adapted to our framework. As mentioned before, it is a consequence of the convexity of KL divergence.

Proof We start by proving a slightly more general result: let μ and ν be any probability measures on $\mathbb{R}^d$ and f the density of $\mathcal{N}(0, s^2)$. Let us define $F(x) = x \log(x)$. We shall note that F is convex and bounded from below (this last point authorizes the subsequent use of Fubini's theorem in the proof). Let also γ be any coupling between μ and ν, we have, by Jensen's inequality:

$$
\begin{aligned}
F\left(\frac{d(f * \mu)}{d(f * \nu)}(x)\right) &= F\left(\int \frac{f(x-y)}{f * \nu(x)} d\mu(y)\right) \\
&= F\left(\iint \frac{f(x-y)}{f(x-z)} \frac{f(x-z)}{f * \nu(x)} d\gamma(y, z)\right) \\
&\leq \iint F\left(\frac{f(x-y)}{f(x-z)}\right) \frac{f(x-z)}{f * \nu(x)} d\gamma(y, z)
\end{aligned}
$$

Now we integrate over $f * \nu$ and apply Fubini's theorem (which we can do because F is bounded from below and the measures are probability measures.):

$$
\begin{aligned}
\mathrm{KL}\left(f * \mu \middle| f * \nu\right) &\leq \iiint F\left(\frac{f(x-y)}{f(x-z)}\right) f(x-z) dx d\gamma(y, z) \\
&= \iint \mathrm{KL}\left(\mathcal{N}(y, s^2) \middle| \mathcal{N}(z, s^2)\right) d\gamma(y, z) \\
&= \frac{1}{2s^2} \iint \|y - z\|^2 d\gamma(y, z),
\end{aligned}
$$

where we used the formula for KL divergence between Gaussians see [33, Equation 10]. Thus, by taking the infimum over all couplings:

$$
\mathrm{KL}\left(f * \mu \middle| f * \nu\right) \leq \frac{1}{2s^2} W_2(\mu, \nu)^2. \tag{17}
$$

Now we get back to $\rho_{S,U}$ and π_U, and as coupling, between $\tilde{\rho}_{S,U}$ and $\tilde{\pi}_U$, we chose the following

$$
\gamma = (Y_\cdot, W_\cdot^S)_\# \mathcal{U}([0, T]),
$$

so that we have

$$
W_2(\tilde{\rho}_{S,U}, \tilde{\pi}_U)^2 \leq \frac{1}{T} \int_0^T \left\|W_t^S - Y_t\right\|^2 dt.
$$

Eq. (17) then implies the result.

We now have the following corollary.

Corollary 1 *In the same setting, we have, almost surely, with respect to S and U:*

$$
\mathrm{KL}\left(\rho_{S,U} \middle| \pi_U\right) \leq \sup_{0 \leq t \leq T} \left\|W_t^S - Y_t\right\|^2.
$$

The following lemma shows a similar bound for the Rényi divergence. While Lem. 6 has already been used in the literature, this next bound is new, to our knowledge. It is slightly more complex because $\mathrm{D}_\beta\left(\cdot \middle| \cdot\right)$ is not jointly convex for $\beta > 1$, the proof will actually extend the fact that it is jointly *quasi*-convex [33, Theorem 13].

Lemma 7 *Let us fix $\beta > 1$, with the same notations as before, we have:*

$$
\mathrm{D}_\beta\left(\rho_{S,U} \middle| \pi_U\right) \leq \frac{\beta}{2s^2} \sup_{0 \leq t \leq T} \left\|W_t^S - Y_t\right\|^2.
$$

Proof We start with a similar calculation to the proof of Lem. 6. Let μ and ν be any probability measures on $\mathbb{R}^d$ and f the density of $\mathcal{N}(0, s^2)$. Let also γ be a coupling between μ and ν. By Jensen's inequality, we have:

$$
\begin{aligned}
\left\{\frac{d(f * \mu)}{d(f * \nu)}(x)\right\}^\beta &= \left\{\iint \frac{f(x-y)}{f(x-z)} \frac{f(x-z)}{(f * \nu(x))} d\gamma(y, z)\right\}^\beta \\
&\leq \iint \left(\frac{f(x-y)}{f(x-z)}\right)^\beta \frac{f(x-z)}{f * \nu(x)} d\gamma(y, z) \\
&= \frac{1}{f * \nu(x)} \iint f(x-y)^\beta f(x-z)^{1-\beta} d\gamma(y, z),
\end{aligned}
$$

21

where the inequality holds by Jensen's inequality because we have

$$\iint \frac{f(x-z)}{(f*\nu(x))} d\gamma(y,z) = \int \frac{f(x-z)}{(f*\nu(x))} d\nu(z) = 1.$$

Now we integrate with respect to $f*\nu$ and take the logarithm, by Fubini's theorem (all functions are positive), it follows that:

$$\log\left(\int (f*\mu(x))^\beta (f*\nu(x))^{1-\beta} dx\right) \le \log\left(\iiint f(x-y)^\beta f(x-z)^{1-\beta} dx d\gamma(y,z)\right)$$

$$\le \log\left(\operatorname*{ess\,sup}_{y,z\sim\gamma} \int f(x-y)^\beta f(x-z)^{1-\beta} dx\right)$$

By dividing by $\beta - 1$ and putting the $\log$ inside the essential supremum, we get:

$$\mathbf{D}_\beta\left(\mu|\,|\nu\right) \le \log\left(\operatorname*{ess\,sup}_{y,z\sim\gamma} \frac{1}{\beta-1} \int f(x-y)^\beta f(x-z)^{1-\beta} dx\right)$$

$$= \operatorname*{ess\,sup}_{y,z\sim\gamma} \left(\mathbf{D}_\beta\left(\mathcal{N}(y,s^2)|\,|\mathcal{N}(z,s^2)\right)\right)$$

$$= \operatorname*{ess\,sup}_{y,z\sim\gamma} \left(\frac{\beta}{2s^2} \|y-z\|^2\right),$$

where we used known formula, deduced from [33, Equation 10 and Theorem 28], to compute the divergence between Gaussians. Now, we get back to $\rho_{S,U}$ and π_U, and as coupling, between $\tilde\rho_{S,U}$ and $\tilde\pi_U$, we chose the same as in the proof of the previous lemma:

$$\gamma = (Y_., W_.^S)_\# \mathcal{U}([0,T]),$$

so that we have

$$\mathbf{D}_\alpha\left(\rho_{S,U}|\,|\pi_U\right) \le \sup_{0\le t\le T} \left\|W_t^S - Y_t\right\|^2.$$

C.5.2 Proof of Th. 5

Theorem 5 *Under Asms. 1 and 2, for any $\lambda > 0$, with probability at least $1-\zeta$ over (S,U) we have:*

$$\lambda\mathbb{E}_{w\sim\rho_{S,U}}[\mathcal{R}(w) - \widehat{\mathcal{R}}_S(w)] \le \frac{1}{Ts^2}\int_0^T \|W_u^S - Y_u\|^2 du + \log(1/\zeta) + \frac{C\lambda^2}{n}.$$

The random variable U is implicitly omitted when proving those bounds, by independence, it does not change the results (all the results may be true U-almost surely). Note that, to simplify notations, we denote $\mathbb{E}_{\rho_{S,U}}$ for $\mathbb{E}_{w\sim\rho_{S,U}}$ (and idem for π_U).

Proof Let $\lambda > 0$ and $\zeta \in (0,1)$. We simply write the result of Th. 10 with the previously defined priors and posteriors, and with function:

$$\Phi(w,S) := \lambda\left(\mathcal{R}(w) - \widehat{\mathcal{R}}_S(w)\right).$$

This gives with probability at least $1-\zeta$ over $S \sim \mu_z^{\otimes n}$

$$\lambda\mathbb{E}_{\rho_{S,U}}\left[\mathcal{R}(w) - \widehat{\mathcal{R}}_S(w)\right] \le \log(1/\zeta) + \mathrm{KL}\left(\rho_{S,U}|\,|\pi_U\right) + \log\mathbb{E}_S\mathbb{E}_{\pi_U}\left[e^{\lambda(\mathcal{R}(w)-\widehat{\mathcal{R}}_S(w))}\right].$$

The last term can be bounded classically, using the sub-Gaussian assumption and Fubini's theorem:

$$\mathbb{E}_S\mathbb{E}_\pi\left[e^{\lambda(\mathcal{R}(w)-\widehat{\mathcal{R}}_S(w))}\right] = \mathbb{E}_{\pi_U}\mathbb{E}_S\left[e^{\lambda(\mathcal{R}(w)-\widehat{\mathcal{R}}_S(w))}\right]$$

$$\le \exp\left\{\frac{\lambda^2\sigma^2}{2n}\right\}.$$

By Lem. 6, we have:

$$\mathrm{KL}\left(\rho_{S,U}|\,|\pi_U\right) \le \frac{1}{2s^2} W_2(\bar\pi_U, \bar\rho_S)^2 \le \frac{1}{2s^2T}\int_0^T \left\|W_t^S - Y_t\right\| dt.$$

The result follows.

C.5.3   Proof of Th. 6

Theorem 6 *Under Asms. 1 and 2, for any $\lambda > 0$ and $\beta > 1$, we have, with probability at least $1 - \zeta$ over the joint distribution of (S, U) and $w \sim \rho_{S,U}$:*

$$
\frac{\lambda \beta}{\beta - 1} \left(\mathcal{R}(w) - \widehat{\mathcal{R}}_S(w) \right) \leq \frac{2\beta - 1}{\beta - 1} \log \left(\frac{2}{\zeta} \right) + \frac{\beta}{2s^2} \sup_{0 \leq t \leq T} \| W_t^S - Y_t \|^2 + \left(\frac{\beta}{\beta - 1} \right)^2 \frac{\lambda^2 \sigma^2}{2n}.
$$

Proof It is exactly the same thing than for Th. 5, but instead of Th. 10, we use Th. 11, and instead of Lem. 6, we use Lem. 7.

C.6   Proof of Th. 13

Theorem 13 (Bound under co-dissipativity) *We make Asms. 2 and 5 and with a loss ℓ that is M-smooth in w (but not Lipschitz). We further assume that there exists $w_0 \in \mathbb{R}^d$ such that $\nabla \ell(w_0, z) \in L^1(\mu_z)$. Then, for n large enough, with probability at least $1 - \zeta$ over $\mu_u \otimes \mu_z^{\otimes n}$, we have, almost surely with respect to (S, U):*

$$
\sup_{0 \leq t \leq T} \| W_t^S - Y_t \|^2 \leq \frac{2K}{m} + \frac{4}{m^2} \left\{ \frac{2}{n} + \frac{\Sigma^2}{n} \left(2\gamma \log(M\sqrt{n}) + \log(2d/\zeta) \right) \right\}.
$$

Proof The intuition of the proof is as follows: to use Asm. 5, we would like to differentiate the function $t \longmapsto \| W_t^S - Y_t \|^2$ and then use the differential form of Grönwall Lemma, along with estimates on the quantity $\sup_{w \in \mathcal{Y}_U} \| \nabla \widehat{\mathcal{R}}_S(w) - \nabla \mathcal{R}(w) \|$ introduced in Lem. 5. However, the aforementioned function is clearly non differentiable everywhere[8], because of the jumps of the driving Lévy process. While we could chain the above argument in case the number of jumps is almost surely finite (it is the case, for example, if $\alpha = 2$), we propose here another argument, based on the regularization of the function by convolution. Let $V^S = W^S - Y$, which we continuously extend to $\mathbb{R}$ by setting the value 0 on $(-\infty, 0)$. We omit the dependence on U.

Before, describing the rest of the proof, let us take $w \in \mathbb{R}^d$, and $\mathcal{U}$ a compactly contained open set that contains w. Then we have, by the smoothness assumption:

$$
\| \nabla \ell(w, z) \| \leq \| \nabla \ell(w_0, z) \| + M \sup_{w \in \mathcal{U}} \| w - w_0 \|,
$$

which ensures, by the dominated convergence theorem, that we can write $\nabla \mathcal{R}(w) = \mathbb{E}_S [\nabla \ell(w, z)]$, without issues.

For the rest of the proof, we fix $S \in \mathcal{Z}^n$, $U \in \Omega$ as well as $T > 0$.

Before all, note that, even without the Lipschitz assumption, we can show that the paths of the processes W^S and Y are almost surely bounded, see Remark 6, as they admit cadlag modifications. Therefore, we can assume, without loss of generality, that, for all T, there exists a compact $K_T \subset \mathbb{R}^d$ such that, for all $0 \leq t \leq T$, $W_t^S, y_t \in K_T$.

Let $(\varphi)_{\epsilon > 0}$ be a regularizing sequence, *i.e.* a sequence of functions $\varphi_\epsilon : \mathbb{R} \longrightarrow \mathbb{R}$ such that:

$$
\varphi_\epsilon \in \mathcal{C}^\infty, \quad \varphi_\epsilon \geq 0, \quad \mathrm{supp}(\varphi_\epsilon) \subset (-\epsilon, \epsilon), \quad \int_{-\infty}^{\infty} \varphi_\epsilon(x) dx = 1.
$$

First note that we have:

$$
\begin{aligned}
\varphi_\epsilon * V_t^S &= \int_{-\infty}^{\infty} \varphi_\epsilon'(t - s) V_s^S ds \\
&= \int_{-\infty}^{\infty} \int_0^s \varphi_\epsilon'(t - s) \left(-\nabla \widehat{\mathcal{R}}_S(W_u^S) + \nabla \mathcal{R}(Y_u) \right) du\, ds \\
&= \int_{-\infty}^{\infty} \left\{ \int_s^{\infty} \varphi_\epsilon'(t - s) ds \right\} \left(-\nabla \widehat{\mathcal{R}}_S(W_u^S) + \nabla \mathcal{R}(Y_u) \right) du \\
&= \int_{-\infty}^{\infty} \varphi_\epsilon(t - s) \left(-\nabla \widehat{\mathcal{R}}_S(W_u^S) + \nabla \mathcal{R}(Y_u) \right) du
\end{aligned}
$$

---

[8]But we can see, from Rademacher theorem, that it is differentiable almost everywhere.

now we compute that, as, by properties of convolution, the derivative of $\varphi_\epsilon * V_t^S$ is $\varphi_\epsilon * \dot{V}_t^S$:

$$\frac{d}{dt}\|\varphi_\epsilon * V_t^S\|^2 = 2\langle \varphi_\epsilon' * V_t^S, \varphi_\epsilon * V_t^S\rangle$$

$$= 2\langle \varphi_\epsilon * \left(-\nabla\widehat{\mathcal{R}}_S(W_\cdot^S) + \nabla\mathcal{R}(Y_\cdot)\right)(t), \varphi_\epsilon * V_t^S\rangle$$

$$=: 2(A + B),$$

where, for $t \in [0, T]$ and some fixed arbitrary $\eta > 0$, for $\epsilon < \eta$:

$$B := \langle \varphi_\epsilon * \left(-\nabla\widehat{\mathcal{R}}_S(Y_\cdot) + \nabla\mathcal{R}(Y_\cdot)\right)(t), \varphi_\epsilon * V_t^S\rangle$$

$$\leq \|\varphi_\epsilon * V_t^S\| \cdot \left\|\int_{-\infty}^{\infty} \varphi_\epsilon(t-s)\left(-\nabla\widehat{\mathcal{R}}_S(W_u^S) + \nabla\mathcal{R}(Y_u)\right)du\right\|$$

$$\leq \|\varphi_\epsilon * V_t^S\|G_\nabla^\eta \int_{-\infty}^{\infty} \varphi_\epsilon(t-u)du$$

$$= \|\varphi_\epsilon * V_t^S\|G_\nabla^\eta,$$

where G_∇^η is like in Eq. (15), but on $[0, T + \eta]$.

$$G_\nabla^\eta := \sup_{0 \leq t \leq t+\eta} \|\nabla\widehat{\mathcal{R}}_S(Y_t) - \nabla\mathcal{R}(Y_t)\|.$$

Now we have:

$$A := \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \varphi_\epsilon(t-s)\varphi_\epsilon(t-u)\langle -\nabla\widehat{\mathcal{R}}_S(W_s^S) + \nabla\widehat{\mathcal{R}}_S(Y_s), W_u^S - Y_u\rangle ds du$$

$$=: C + D,$$

where, by Asm. 5:

$$C := \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \varphi_\epsilon(t-s)\varphi_\epsilon(t-u)\langle -\nabla\widehat{\mathcal{R}}_S(W_s^S) + \nabla\widehat{\mathcal{R}}_S(Y_s), W_s^S - Y_s\rangle ds du$$

$$\leq \int_{-\infty}^{\infty} \varphi_\epsilon(t-s)\left(-m\|W_s^S - Y_s\|^2 + K\right)ds$$

$$= -m\int_{-\infty}^{\infty} \varphi_\epsilon(t-s)\|W_s^S - Y_s\|^2 ds + K$$

By Cauchy-Schwarz's inequality, we have:

$$\int_{-\infty}^{\infty} \varphi_\epsilon(t-s)\|W_s^S - Y_s\|^2 ds \int_{-\infty}^{\infty} \varphi_\epsilon(t-s)ds \geq \left\{\int_{-\infty}^{\infty} \varphi_\epsilon(t-s)\|W_s^S - Y_s\|ds\right\}^2,$$

so that by the triangle inequality:

$$C \leq -m\|\varphi_\epsilon * V_t^S\|^2 + K.$$

Let us introduce, thanks to the continuity of the functions $\nabla\widehat{\mathcal{R}}_S$ and $\nabla\mathcal{R}$ and the fact that the trajectory is in the compact $K_{T+\eta}$:

$$G := \max\left\{\sup_{x \in K_{T+\eta}} \|\nabla\mathcal{R}(x)\|, \sup_{x \in K_{T+\eta}} \|\nabla\widehat{\mathcal{R}}_S(x)\|\right\}$$

On the other hand, by noting that $W_\cdot^S - Y_\cdot$ is $2L$-Lipschitz continuous on $\mathbb{R}$ and using Cauchy-Schwarz's inequality, for $\epsilon < \eta$:

$$D := \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \varphi_\epsilon(t-s)\varphi_\epsilon(t-u)\langle -\nabla\widehat{\mathcal{R}}_S(W_s^S) + \nabla\widehat{\mathcal{R}}_S(Y_s), W_u^S - Y_u - W_s^S + Y_s\rangle ds du$$

$$\leq 2G\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \varphi_\epsilon(t-s)\varphi_\epsilon(t-u)\| -\nabla\widehat{\mathcal{R}}_S(W_s^S) + \nabla\widehat{\mathcal{R}}_S(Y_s)\| \cdot |s-u| ds du$$

$$= 4G^2\int_{t-\epsilon}^{t+\epsilon}\int_{t-\epsilon}^{t+\epsilon} \varphi_\epsilon(t-s)\varphi_\epsilon(t-u)|s-u| ds du$$

$$\leq 8G^2\epsilon.$$

Therefore, we have deduced that, for $t \in [0, T]$:

$$\frac{d}{dt}\|\varphi_\epsilon * V_t^S\|^2 \leq 2\|\varphi_\epsilon * V_t^S\|G_\nabla^\eta - 2m\|\varphi_\epsilon * V_t^S\|^2 + 2K + 16G^2\epsilon.$$

By Young's inequality, we further have:

$$\frac{d}{dt}\|\varphi_\epsilon * V_t^S\|^2 \leq \frac{(G_\nabla^\eta)^2}{m} - m\|\varphi_\epsilon * V_t^S\|^2 + 2K + 16G^2\epsilon.$$

Thus, by Gronwall's Lemma:

$$\|\varphi_\epsilon * V_t^S\|^2 \leq \frac{(G_\nabla^\eta)^2 + 2Km + 16G^2 m\epsilon}{m^2}(1 - e^{-mt}) + \|\varphi_\epsilon * V_t^S\|^2\Big|_{t=0} e^{-mt}.$$

Now, as $t \longmapsto V_t^S$ is uniformly continuous (because it is Lipschitz continuous, thanks to the boundedness of $\nabla_w \ell$ on the trajectories of W^S and Y, which are included in a compact), see App. C.1, we know that $\varphi_\epsilon * V_t^S$ converges uniformly to V_t^S on $[0, T]$, hence, by taking the limit $\epsilon \to 0$, we get, for $t \in [0, T]$:

$$\|V_t^S\|^2 \leq \frac{(G_\nabla^\eta)^2 + 2Km}{m^2}(1 - e^{-mt})$$

As η is arbitrary, we can now simply apply the same reasoning as in the proof of Lem. 5. It gives us that, for n large enough, with probability at least $1 - \zeta$ over $\mu_u \otimes \mu_z^{\otimes n}$, we have:

$$\sup_{w \in \mathcal{Y}_U} \|\nabla\widehat{\mathcal{R}}_S(w) - \nabla\mathcal{R}(w)\| \leq \frac{2}{\sqrt{n}} + 2\Sigma\sqrt{\frac{2\gamma\log(M\sqrt{n}) + \log(2d/\zeta)}{2n}}.$$

So finally, for n large enough, with probability at least $1 - \zeta$ over $\mu_u \otimes \mu_z^{\otimes n}$, we have:

$$\sup_{0 \leq t \leq T} \|V_t^S\|^2 \leq \frac{2K}{m} + \frac{4}{m^2}\left\{\frac{2}{n} + \frac{\Sigma^2}{n}\left(2\gamma\log(M\sqrt{n}) + \log(2d/\zeta)\right)\right\}.$$

Remark 12 *Plugging the above bound in the PAC-Bayesian bounds presented before, with $\lambda = \sqrt{n}$, will give a higher convergence rate in n to the terms having big constants. Moreover, the constant $\frac{K}{2m}$ term is not a problem, as it will be of the order $\frac{K}{2m\sqrt{n}}$ in the PAC-Bayesian bounds.*