Journal of Machine Learning Research 25 (2024) 1-55 Submitted 4/24; Revised 9/24; Published 12/24

Uniform Generalization Bounds on Data-Dependent Hypothesis Sets via PAC-Bayesian Theory on Random Sets

Benjamin Dupuis* BENJAMIN.DUPUIS@INRIA.FR
INRIA - Département d'Informatique de l'Ecole Normale Supérieure
PSL Research University
Paris, France

Paul Viallard PAUL.VIALLARD@INRIA.FR
Univ Rennes, Inria, CNRS IRISA - UMR 6074
Rennes, France

George Deligiannidis GEORGE.DELIGIANNIDIS@STATS.OX.AC.UK
Department of Statistics
University of Oxford, Oxford, UK

Umut Şimşekli UMUT.SIMSEKLI@INRIA.FR
INRIA - Département d'Informatique de l'Ecole Normale Supérieure
PSL Research University - CNRS
Paris, France

* *Corresponding author.*

Editor: Gergely Neu

Abstract

We propose data-dependent uniform generalization bounds by approaching the problem from a PAC-Bayesian perspective. We first apply the PAC-Bayesian framework on "random sets" in a rigorous way, where the training algorithm is assumed to output a data-dependent hypothesis set after observing the training data. This approach allows us to prove data-dependent bounds, which can be applicable in numerous contexts. To highlight the power of our approach, we consider two main applications. First, we propose a PAC-Bayesian formulation of the recently developed fractal-dimension-based generalization bounds. The derived results are shown to be tighter and they unify the existing results around one simple proof technique. Second, we prove uniform bounds over the trajectories of continuous Langevin dynamics and stochastic gradient Langevin dynamics. These results provide novel information about the generalization properties of noisy algorithms.

Keywords: Uniform Generalization Bounds, PAC-Bayesian Theory, Fractal Geometry, Langevin Dynamics, SGLD.

1. Introduction

Over the past decades, providing generalization guarantees for modern machine learning algorithms has been a major research topic. In most cases, these algorithms can be framed

as the following optimization problem:

$$\min \left\{ \mathcal{R}(w) := \int_{\mathcal{Z}} \ell(w, z) d\mu_z(z), \ w \in \mathbb{R}^d \right\}, \tag{1}$$

where $(\mathcal{Z}, \mathcal{F})$ is a measurable space, μ_z a data distribution on $\mathcal{Z}$ and $\ell : \mathbb{R}^d \times \mathcal{Z} \longrightarrow \mathbb{R}_+$ is the composite loss function. The function $\mathcal{R} : \mathbb{R}^d \longrightarrow \mathbb{R}_+$ is called the population risk. The vectors $w \in \mathbb{R}^d$ are the *parameters* (or *weights* in the neural network context) of the model. For instance, in a regression setting, $\mathcal{Z}$ would be the product $\mathcal{X} \times \mathcal{Y}$ of an input space $\mathcal{X}$ and a target space $\mathcal{Y}$. In that case, ℓ would be the composition of a parametrized predictor $F_w : \mathcal{X} \longrightarrow \mathcal{Y}$ and a loss function $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \longrightarrow \mathbb{R}_+$, *i.e.*, $\ell(w, (x, y)) = \mathcal{L}(F_w(x), y)$. In practical cases, we know neither the data distribution μ_z nor the population risk $\mathcal{R}$, but we have access to independent and identically distributed (*i.i.d.*) samples $S = (z_1, \ldots, z_n) \sim \mu_z^{\otimes n}$, drawn from the data distribution, where $\mu_z^{\otimes n}$ is the product measure $\mu_z \otimes \cdots \otimes \mu_z$. Problem (1) is then replaced by the minimization of the empirical risk, defined by

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

Since Problem (1) is replaced by the function in Equation (2), it is necessary to evaluate the quantity $\mathcal{R}(w) - \widehat{\mathcal{R}}_S(w)$ in order to estimate the performance of the method; we call this quantity the generalization error. One particular class of generalization bounds (*i.e.*, upper bounds on the generalization error) that has been widely studied is the class of uniform generalization bounds (see *e.g.*, Shalev-Schwartz and Ben-David, 2014). It consists of considering a subset of the possible functions $\{\ell(w, \cdot), \ w \in \mathcal{W}\}$, from which we aim to choose the learned model. In many practical cases, we consider a set of parameters $\mathcal{W} \subseteq \mathbb{R}^d$, and we are interested in the worst generalization error over $\mathcal{W}$. For example, $\mathcal{W}$ could be the set of vectors having a certain norm. For a given fixed set $\mathcal{W} \subseteq \mathbb{R}^d$, which we will call a *hypothesis set*, the quantity of interest becomes:

$$G_S(\mathcal{W}) := \sup_{w \in \mathcal{W}} \left(\mathcal{R}(w) - \widehat{\mathcal{R}}_S(w) \right). \tag{3}$$

Several studies have derived bounds on this quantity, known as *uniform generalization error* or *worst-case generalization error*. While some of them define a notion of complexity of the hypothesis set, like Rademacher complexity (Bartlett and Mendelson, 2002), other authors introduced intrinsic dimensions of $\mathcal{W}$, like Vapnik-Chervonenkis (VC) dimension (Vapnik and Chervonenkis, 1968, 1971) or fractal dimensions (Şimşekli et al., 2020). When these quantities are significantly smaller than the ambient dimension d, they may be able to explain the good generalization performances of modern deep learning models, for which d typically takes very high values. Other notions of hypothesis set complexity have been used, even for non-uniform generalization bounds (Grünwald and Mehta, 2019).
However, as argued by Nagarajan and Kolter (2019), even the tightest uniform generalization bounds, over a fixed set $\mathcal{W}$ (that may depend on the data distribution, μ_z, but not on the data sample $S \sim \mu_z^{\otimes n}$), may be vacuous. Moreover, it is known that neural networks can fit random labels (Zhang et al., 2017, 2021), making quantities like Rademacher complexity

over vast hypothesis sets excessively large to explain the generalization performance that is observed in practice.
These examples show that one of the main drawbacks of such approaches is the lack of data dependence, *i.e.*, the fact that the hypothesis set has no dependence on S. In addition to reducing the set of hypotheses, hence tightening the bounds, considering data-dependent sets provides generalization bounds that are specific to the dataset used during training.
Such data-dependent hypothesis sets, which we will denote $\mathcal{W}_S$, also depend on the learning algorithm and naturally emerge in stochastic optimization. For instance, depending on the context, $\mathcal{W}_S$ can be the set of minimizers of the empirical risk in Equation (2), or the trajectory of an iterative algorithm minimizing Equation (2). This last setting is considered in the fractal-based generalization literature (Şimşekli et al., 2020; Dupuis et al., 2023; Hodgkinson et al., 2022). Classical machine learning models, such as Langevin diffusions (Raginsky et al., 2017; Mou et al., 2018), also generate data-dependent trajectories. Additional concrete examples will be provided in Section 3.1.

1.1 Motivation

The data-dependence of $\mathcal{W}_S$ makes the task of bounding $G_S(\mathcal{W}_S)$ much more challenging compared to the case of a fixed hypothesis set, as most classical techniques will not be valid anymore. Toward this goal, some studies imposed new assumptions on $\mathcal{W}_S$, like hypothesis set stability (Foster et al., 2019). Other works, like (Şimşekli et al., 2020; Hodgkinson et al., 2022; Camuto et al., 2021; Dupuis et al., 2023) introduced Mutual Information (MI) terms to control the statistical dependence between S and $\mathcal{W}_S$. However, all these methods require different proof techniques and are based on complicated geometric and algorithmic assumptions. We emphasize the interest in proving new generic uniform generalization bounds for data-dependent hypothesis sets in two particular cases.

1.1.1 Fractal-based generalization bounds

Several recent works (Şimşekli et al., 2020; Camuto et al., 2021; Hodgkinson et al., 2022; Birdal et al., 2021; Dupuis et al., 2023) relate the worst-case generalization error to a notion of *fractal dimension* (Falconer, 2014) of the hypothesis set. These results can be informally summarized by stating that, with probability at least $1-\zeta$ and for n big enough, one has[1]:

$$\sup_{w\in\mathcal{W}_S}\left(\mathcal{R}(w)-\widehat{\mathcal{R}}_S(w)\right)\lesssim\sqrt{\frac{(\text{Fractal Dim. of }\mathcal{W}_S)+\mathrm{I}+\log(1/\zeta)}{n}}, \tag{4}$$

which involves several notions of fractal dimensions that will be defined in the sequel. In Equation (4), I is a MI term, *i.e.*, a notion of statistical dependence between the data and the hypothesis set, which may differ among the different results.
The interest of such bounds resides in the fact that the fractal dimension of $\mathcal{W}_S$ might be much smaller than the dimension of the ambient space $\mathbb{R}^d$, making it pertinent for the study of over-parameterized models. Moreover, these fractal-based bounds have opened the door to new links between the generalization error and topological data analysis (TDA)

1. We use the symbol $\lesssim$ to specify that absolute constants or logarithmic terms are omitted from the statement.

(Boissonnat et al., 2018). Indeed, Birdal et al. (2021); Dupuis et al. (2023); Andreeva et al. (2023) have shown that the aforementioned dimensions can be estimated using tools from TDA and observed empirical correlation with the generalization error. While providing new insight into the generalization abilities of certain learning algorithms, in particular, heavy-tailed dynamics (Gürbüzbalaban et al., 2021; Şimşekli et al., 2019, 2020), these bounds suffer from several issues. They depend on MI terms that differ from paper to paper and are generally beyond the reach of computability. Moreover, these terms are often convoluted and hard to interpret, especially in (Dupuis et al., 2023) and (Şimşekli et al., 2020). In (Dupuis et al., 2023), an additional "geometric stability" assumption is needed to simplify the MI term and make it similar to other works (Hodgkinson et al., 2022); this is an intricate assumption that seems hard to verify in practice and the resulting generalization bound has a worse rate in terms of the number of data points n. Hence, our first aim is to provide a unified theoretical framework that can encompass all the existing fractal-dimension-based bounds with the correct rate of convergence without making additional non-trivial assumptions.

1.1.2 Langevin dynamics

Our second main motivation in data-dependent hypothesis sets stems from the generalization properties of Continuous Langevin Dynamics (CLD) and Stochastic Gradient Langevin Dynamics (SGLD). The CLD algorithm corresponds to a continuous-time gradient flow, perturbed with white Gaussian noise. It is described by the following Stochastic Differential Equation (SDE):

$$dW_t = -\nabla \widehat{\mathcal{R}}_S(W_t)dt + \sqrt{2\beta^{-1}}dB_t, \tag{5}$$

where $(B_t)_{t\geq 0}$ is a standard Brownian motion in $\mathbb{R}^d$. SGLD is a discrete version of the above dynamics, *i.e.*, it is a stochastic gradient descent algorithm where standard Gaussian noise is added to the unbiased estimation of the gradient at each iteration:

$$\forall k \in \mathbb{N},\ W_{k+1} = W_k - \eta_{k+1}\hat{g}_{k+1} + \sqrt{2\beta^{-1}\eta_{k+1}}\epsilon_{k+1}, \quad \epsilon_{k+1} \sim \mathcal{N}(0, I_d), \tag{6}$$

where η_k is the step size or the learning rate at iteration k, β is an inverse temperature parameter, $\hat{g}_{k+1}$ is an unbiased estimate of $\nabla \widehat{\mathcal{R}}_S(W_k)$, and ϵ_k are independent realizations of $\mathcal{N}(0, I_d)$ drawn at each iteration. As noted by Raginsky et al. (2017), in the case where $\hat{g}_k = \nabla \widehat{\mathcal{R}}_S(W_k)$, Equation (6) is exactly the Euler-Maruyama discretization of Equation (5). Several works have successfully proven generalization bounds for both CLD and SGLD by focusing on the generalization error on the last iterate, *e.g.*, (Mou et al., 2018; Farghly and Rebeschini, 2021; Neu et al., 2021), see Tables 3 and 4 for more details. However, we argue that proving data-dependent uniform generalization bounds over their trajectories might provide additional insight.

Indeed, when considering an iterative stochastic optimization algorithm, the choice of a stopping criterion is often arbitrary and without a theoretical basis. Uniform generalization bounds over the trajectory address this issue by providing performance guarantees, regardless of the stopping time. Usual generalization bounds over the last iterate do not hold for data-dependent stopping times.

Abbreviation	Meaning	Abbreviation	Meaning
B.	Loss bounded by $B > 0$	L.	Loss is L-Lipschitz $\mathcal{C}^0$
SG.	$\ell(w, z)$ subgaussian w.r.t. z	R.	ℓ_2-regularization
BS-b	Batch size is $b \in \mathbb{N}^\star$	$T \to \infty$	Long-time limit
D.	Dissipativity	$\mathbb{E}_{\mathcal{A}}$	Expectation on the noise
$\mathbb{E}_S$	Expected bound	$\mathrm{HP}^{(\zeta)}$	With high probability $(1 - \zeta)$

Table 1: Summary of the abbreviations used in the Tables 2, 3 and 4. They concern the different assumptions that can be made on the loss $\ell(w, z)$, as well as the different types of bounds that can be proven.

	Asmp. on $\ell(w, z)$	Fractal dim.	IT term	Smallest IT
Şimşekli et al. (2020)	SG., L.	Euclidean	$\simeq I_\infty(S, N_\delta)$	
Hodgkinson et al. (2022)	B., L.	Euclidean	$I_\infty(S, \mathcal{W})$	
Dupuis et al. (2023)	B.	Data-dependent	$\max_j I_\infty(S, N_{\delta,j})$	
Ours	B., L.	Euclidean	$\log \frac{d\rho_S}{d\pi}(\mathcal{W})$	✓
Ours	B.	Data-dependent	$\log \frac{d\rho_S}{d\pi}(\mathcal{W})$	✓

Table 2: Overview of the comparison between our work and existing fractal generalization bounds. The abbreviations used in this table can be found in Table 1. We refer to the respective papers for the exact definition of the introduced IT terms. We can see in this table that our method yields smaller IT terms.

Moreover, when gradient dynamics converge toward a local minimum of the empirical risk, the behavior of the algorithm around this point may be seen as a characteristic of this local minimum. By proving bounds that are uniform over the data-dependent optimization trajectory around a local minimum, the theory can capture geometric and statistical properties that are particular to this local minimum. Therefore, if the trajectory of the algorithm is considered near a local minimum, uniform bounds on the trajectory provide information that bounds over the last iterate fail to capture. These remarks, in fact, extend beyond the study of Langevin dynamics to any other gradient-based algorithm.

1.2 Contributions and Overview of Main Results

In this paper, we prove data-dependent uniform generalization bounds that address the issues mentioned above. More precisely, we propose a theoretical framework in which such uniform bounds can be derived from a single-proof technique without the need for technical assumptions specific to each problem.

Our approach will be to extend the so-called PAC-Bayesian analysis techniques (Shawe-Taylor and Williamson, 1997; McAllester, 1998; Catoni, 2007) to random hypothesis sets, which will be formally defined in Section 2.3. This framework has proven to be able to provide practical generalization bounds, even for neural networks. However, to the best of our knowledge, no equivalent for data-dependent uniform bounds has been studied from the PAC-Bayesian perspective.

Paper	Bounded quantity	Type	Asmp. on $\ell(w,z)$	Bound ($\mathcal{O}$)
Mou et al. (2018)	$\mathbb{E}_{\mathcal{A}}\mathrm{err}_S(W_T)$	$\mathbb{E}_S$	L. B.	$\frac{LB\sqrt{\beta T}}{n}$
Mou et al. (2018)	$\mathbb{E}_{\mathcal{A}}\mathrm{err}_S(W_T)$	HP	SG. R.	$\left\{\frac{\beta}{n}\int_0^T e^{-R_t^T}\mathbb{E}\left\|g_S(t)\right\|^2 dt\right\}^{\frac{1}{2}}$
Li et al. (2020)	$\mathbb{E}_{\mathcal{A}}\mathrm{err}_S(W_T)$	$\mathbb{E}_S$	L. B. $T\to\infty$	$\frac{e^{4\beta B}BL\sqrt{\beta}}{n\sqrt{\lambda}}$
Futami et al. (2023)	$\mathbb{E}_{\mathcal{A}}\mathrm{err}_S(W_T)$	$\mathbb{E}_S$	SG. D.	$\sqrt{\frac{\min(1,\eta T)}{n}}$
Ours	$\mathbb{E}_{\mathcal{A}}\sup_{0\le t\le T}\mathrm{err}_S(W_t)$	HP	L. B.	$\frac{BL\sqrt{\beta T}}{\sqrt{n}}$
Ours	$\mathbb{E}_{\mathcal{A}}\sup_{0\le t\le T}\mathrm{err}_S(W_t)$	HP	B.	$B\left\{\frac{\beta}{n}\int_0^T\mathbb{E}\left\|g_S(t)\right\|^2 dt\right\}^{\frac{1}{2}}$
Ours	$\mathbb{E}_{\mathcal{A}}\sup_{0\le t\le T}\mathrm{err}_S(W_t)$	HP	L. B.	$\frac{B}{\sqrt{n}}+B\frac{\beta TL^2}{n}$

Table 3: Comparison of our CLD bounds with some classical bounds from the literature. The notation $\mathrm{err}_S(w)$ denotes $\mathcal{R}(w)-\widehat{\mathcal{R}}_S(w)$. The abbreviations used in this table can be found in Table 1. $\mathbb{E}_{\mathcal{A}}$ denotes the expectation over the randomness of the algorithm. $g_S(t)$ is a shortcut for $\nabla\widehat{\mathcal{R}}_S(w_t)$. We refer to the corresponding papers for their exact statements and definitions.

As opposed to the classical setting where PAC-Bayesian theory relies on probability distributions defined over one predefined hypothesis set, we will consider stochastic algorithms that "generate random hypothesis sets" $\mathcal{W}_S$, which follow a data-dependent probability distribution, denoted ρ_S, over the space of possible hypothesis sets, *i.e.*, we have $\mathcal{W}_S\sim\rho_S$. We will refer to our proposed methods as being a *PAC-Bayesian framework for random sets.* In other words, we will construct a setting where a stochastic learning algorithm generates a random set of vectors rather than a single random vector.
Our contributions are detailed as follows.

1. We rigorously formalize the PAC-Bayesian theory for random sets and state uniform generalization upper bounds, which hold in this general framework. Informally, we can summarize our bounds with the following statement. With high probability, we have:
$$\sup_{w\in\mathcal{W}_S}\left(\mathcal{R}(w)-\widehat{\mathcal{R}}_S(w)\right)\lesssim\mathcal{O}\left(\sqrt{\frac{\mathrm{IT}(\rho_S,\pi)+\mathcal{C}(\mathcal{W}_S)}{n}}\right), \tag{7}$$
where π is a data-independent distribution over the space of hypothesis sets, $\mathcal{W}_S\sim\rho_S$ is a *random hypothesis set* following the distribution ρ_S, $\mathcal{C}(\mathcal{W}_S)$ represents a notion of complexity of $\mathcal{W}_S$, and $\mathrm{IT}(\rho_S,\pi)$ denotes an Information Theoretic (IT) term that measures how "far away" ρ_S and π are from each other. A typical example of an IT term is the Kullback-Leibler (KL) divergence. The quantity $\mathcal{C}(\mathcal{W}_S)$ may differ between different applications. In particular, we prove in Theorem 11 that $\mathcal{C}(\mathcal{W}_S)$ can take the form of a *data-dependent Rademacher complexity.* As an additional technical contribution, we show that our setup naturally paves the way for new data-dependent uniform lower bounds.
2. Thanks to our PAC-Bayesian framework for random sets, we derive generalization bounds using the two commonly used notions of fractal dimensions in the literature. All our

bounds share the same information-theoretic terms and the same proof technique, which is a direct consequence of the aforementioned bound with data-dependent Rademacher complexity. Therefore, our theory simplifies the previous approaches to fractal-based generalization. In accordance with Equation (7), our bounds have the following form:

$$\sup_{w\in\mathcal{W}_S}\left(\mathcal{R}(w)-\widehat{\mathcal{R}}_S(w)\right)\lesssim\mathcal{O}\left(\sqrt{\frac{(\text{Fractal Dim. of }\mathcal{W}_S)+\text{IT}(\rho_S,\pi)}{n}}\right),$$

with $S\sim\mu_z^{\otimes n}$ and $\mathcal{W}_S\sim\rho_S$. Moreover, we show that our IT terms are smaller and, therefore, yield tighter bounds, in addition to being more intuitive than the existing terms in the literature. Our main contributions to fractal generalization bounds are summarized in Table 2. As one can see in Table 2, two types of fractal dimensions are mainly used in the literature. The first one is induced by the Euclidean distance and we abbreviate it by "Euclidean" in the table. The second one is induced by a data-dependent pseudometric, and we call it "data-dependent" (Dupuis et al., 2023), see Section 5 for technical details. In particular, we can recover bounds with the data-dependent fractal dimension and the same mutual information as in (Dupuis et al., 2023, Theorem 3.8), but without the need for any stability assumption. Moreover, our proof technique makes it possible to improve the convergence rate in n, compared to (Dupuis et al., 2023, Theorem 3.8): we obtain a $n^{-1/2}$ rate, whereas the prior work had a $n^{-1/3}$ rate.

3. We use our new techniques to derive uniform generalization bounds over the trajectory of CLD. For a fixed time horizon T, our main bound states that, with high probability

$$\mathbb{E}_{\mathcal{A}}\left[\sup_{0\le t\le T}\left(\mathcal{R}(W_t)-\widehat{\mathcal{R}}_S(W_t)\right)\right]\lesssim\mathcal{O}\left(\sqrt{\frac{1}{n\sigma^2}\int_0^T\mathbb{E}_{\mathcal{A}}\big[\|\nabla\widehat{\mathcal{R}}_S(W_t)\|^2\big]dt}\right),\tag{8}$$

where $\mathbb{E}_{\mathcal{A}}$ denotes the expectation over the randomness of the algorithm, *i.e.*, the expectation over the data-dependent distribution ρ_S, describing the law of the random hypothesis set (the trajectory) generated by the Langevin equation. Our results, summarized in Table 3, yield, to our knowledge, the first uniform bounds over the trajectories of CLD. Our bounds have an order of magnitude that is coherent with the existing literature on Langevin dynamics. In particular, they relate the uniform generalization error to the average gradient norm of the empirical risk along the random trajectory. We also apply our methods to the SGLD recursion and prove high-probability uniform generalization bounds of the following form:

$$\mathbb{E}_{\mathcal{A}}\left[\max_{1\le k\le T}\left(\mathcal{R}(W_k)-\widehat{\mathcal{R}}_S(W_k)\right)\right]\lesssim\mathcal{O}\left(\sqrt{\frac{1}{n\sigma^2}\sum_{k=1}^{T}\eta_k\mathbb{E}_{\mathcal{A}}\left\|\hat{g}_k\right\|^2}\right).\tag{9}$$

These uniform generalization bounds for SGLD have a similar form as their continuous-time counterpart, presented in Equation (8). On the left-hand side of Equation (9), the expectation $\mathbb{E}_{\mathcal{A}}$ is taken outside of the maximum. Therefore, this result could not be trivially deduced from the existing results presented in Table 4.

Paper	Bounded quantity	Type	Asmp. on $\ell(w,z)$	Bound ($\mathcal{O}$)
Raginsky et al. (2017)	$\mathbb{E}_{\mathcal{A}}\mathrm{err}_S(W_T)$	$\mathbb{E}_S$	SG. D.	$\eta T + \frac{1}{n} + e^{-\eta T/c}$
Mou et al. (2018)	$\mathbb{E}_{\mathcal{A}}\mathrm{err}_S(W_T)$	$\mathbb{E}_S$	L. B.	$\frac{LB}{n}\sqrt{\beta\sum_k \eta_k}$
Mou et al. (2018)	$\mathbb{E}_{\mathcal{A}}\mathrm{err}_S(W_T)$	HP	L. SG. BS-1 R.	$\left\{\frac{\beta}{n}\sum_{k=1}^T e^{-R_k^T}\eta_k \mathbb{E}\left\|g_k\right\|^2\right\}^{\frac{1}{2}}$
Pensia et al. (2018)	$\mathrm{err}_S(W_T)$	HP$^\zeta$	L. SG. BS-1	$L\sqrt{\frac{\beta}{n\zeta}\sum_{k=1}^T \eta_k}$
Negrea et al. (2019)	$\mathbb{E}_{\mathcal{A}}\mathrm{err}_S(W_T)$	$\mathbb{E}_S$	SG. BS-b	$\left\{\frac{1}{bn}\sum_{k=1}^T \beta_k\eta_k \mathrm{Var}(g_k)\right\}^{\frac{1}{2}}$
Haghifam et al. (2020)	$\mathbb{E}_{\mathcal{A}}\mathrm{err}_S(W_T)$	$\mathbb{E}_S$	B. BS-n	$\frac{B}{n}\left\{\sum_{k=1}^T \eta_k\beta_k \left\|\zeta_k\right\|^2\right\}^{\frac{1}{2}}$
Neu et al. (2021)	$\mathbb{E}_{\mathcal{A}}\mathrm{err}_S(W_T)$	$\mathbb{E}_S$	SG. BS-1	$\left\{\frac{\beta}{n}\sum_{k=1}^T \eta_k \mathrm{Var}(g_k\|w_k)\right\}^{\frac{1}{2}}$
Farghly et al. (2021)	$\mathbb{E}_{\mathcal{A}}\mathrm{err}_S(W_T)$	$\mathbb{E}_S$	SG. D.	$\min\{1,\eta T\}\left(\sqrt{\eta}+\frac{1}{n\sqrt{\eta}}\right)$
Futami et al. (2023)	$\mathbb{E}_{\mathcal{A}}\mathrm{err}_S(W_T)$	$\mathbb{E}_S$	SG. D.	$\sqrt{\frac{\min(1,\eta T)}{n}}$
Ours	$\mathbb{E}_{\mathcal{A}}\max_{1\le k\le T}\mathrm{err}_S(W_k)$	HP	B. L. BS-b	$\frac{BL}{\sqrt{n}}\sqrt{\beta\sum_{k=1}^T \eta_k}$
Ours	$\mathbb{E}_{\mathcal{A}}\max_{1\le k\le T}\mathrm{err}_S(W_k)$	HP	B. BS-b	$B\left\{\frac{\beta}{n}\sum_{k=1}^T \eta_k \mathbb{E}\left\|g_k\right\|^2\right\}^{\frac{1}{2}}$
Ours	$\mathbb{E}_{\mathcal{A}}\max_{1\le k\le T}\mathrm{err}_S(W_k)$	HP	B. L. BS-n	$B\frac{L^2\beta}{n}\sum_{k=1}^T \eta_k$

Table 4: Comparison of our SGLD bounds with some classical bounds existing in the literature. The notation $\mathrm{err}_S(w)$ denotes $\mathcal{R}(w)-\widehat{\mathcal{R}}_S(w)$ as before. The abbreviations used in this table can be found in Table 1. $\mathbb{E}_{\mathcal{A}}$ denotes the expectation over the randomness of the algorithm. We do not formally define here the variance terms appearing in (Negrea et al., 2019) and (Neu et al., 2021), as well as the variation on the gradient ζ_t used in (Haghifam et al., 2020), see the corresponding papers.

1.3 Organization of the paper

We start, in Section 2, by introducing notation and describing two varieties of classical generalization bounds on which we build part of our theory, namely Rademacher complexity bounds (Section 2.2) and PAC-Bayesian bounds (Section 2.3). We introduce our PAC-Bayesian framework for random sets in full generality in Section 3.1 and prove our general data-dependent uniform generalization bounds in Section 4, within the introduced framework. The remainder of the paper is devoted to applications of these bounds to fractal-based generalization bounds (Section 5), CLD (Section 6), and SGLD (Section 7). All the proofs of the main results are provided in Appendix B.

2. Preliminaries

In this section, we recall some notations and a few existing uniform generalization bounds and PAC-Bayesian bounds in Sections 2.2 and 2.3, respectively. These results will be used throughout the paper.

2.1 Notations

We use the notations $\mathcal{R}$, $\widehat{\mathcal{R}}_S$, G_S, $\mathcal{W}$, ℓ, μ_z and $(\mathcal{Z}, \mathcal{F})$ as they have been defined in the introduction. With a slight abuse of notation, we will write $G_S(w)$, or $\mathrm{err}_S(w)$, for $G_S(\{w\})$. When $\mathcal{W}$ depends on the data, it is said to be data-dependent. A fixed (or data-independent) set $\mathcal{W}$ will be called a fixed hypothesis set.
Given two probability measures μ and ν, the absolute continuity of μ with respect to ν will be denoted $\mu \ll \nu$; the equivalence between μ and ν (*i.e.*, $\mu \ll \nu$ and $\nu \ll \mu$) will be denoted $\mu \sim \nu$. Given μ a probability measure and X a random variable on the same probability space, we define the image measure (or pushforward) $X_{\#}\mu$, by $X_{\#}\mu(B) := \mu(X^{-1}(B))$. To differentiate between different settings, probability measures will be denoted $(\mathcal{P}, \mathcal{Q}_S)$ when they are distributions over parameter vectors and (π, ρ_S) when they are distributions over sets. Definitions of the MI terms (in particular I_∞, which will appear in our bounds) are provided in Appendix A, along with additional technical background.
Throughout the text, we will use the notion of *Markov kernel*. A family $(\mathcal{Q}_S)_{S\in\mathcal{Z}^n}$ of probability distributions, on a measurable space $(\Omega, \mathcal{T})$ is a Markov kernel on $\Omega \times \mathcal{Z}^n$ if, for all $A \in \mathcal{T}$, the map $S \longmapsto \mathcal{Q}_S(A)$ is $\mathcal{F}^{\otimes n}$-measurable. We denote by $\mathfrak{P}(\mathbb{R}^d)$ the set of all subsets of the parameter space $\mathbb{R}^d$. Given a finite set A, its cardinality will be denoted $|A|$.

2.2 Uniform generalization bounds with data-independent hypothesis sets

In this section, we present existing results on uniform generalization bounds for fixed hypothesis sets, *i.e.*, bounding the quantity $G_S(\mathcal{W})$ defined in Equation (3). While several approaches exist in the literature (Vapnik and Chervonenkis, 1968, 1971), we focus on uniform generalization bounds based on the so-called Rademacher complexity (Koltchinskii and Panchenko, 2004; Koltchinskii, 2001; Koltchinskii and Panchenko, 2002; Bartlett and Mendelson, 2002; Bartlett et al., 2002), which we will use for the proof of some of our main results. Throughout this section, $\mathcal{W} \subset \mathbb{R}^d$ denotes a *fixed* hypothesis set.
The next theorem provides the definition of Rademacher complexity, along with a known high probability upper bound of $G_S(\mathcal{W})$, see, *e.g.*, Mohri et al. (2018, Theorem 3.3).

Theorem 1 (Uniform generalization bounds with the Rademacher complexity) *For any bounded loss function $\ell : \mathbb{R}^d \times \mathcal{Z} \to [0, B]$, where $B > 0$ is a constant, we have*

$$\mathbb{P}_{S\sim\mu_z^{\otimes n}}\left(\sup_{w\in\mathcal{W}}\left(\mathcal{R}(w) - \widehat{\mathcal{R}}_S(w)\right) \leq 2\mathbf{Rad}_S(\mathcal{W}) + 3B\sqrt{\frac{\log(1/\zeta)}{2n}}\right) \geq 1 - \zeta, \tag{10}$$

where $\mathbf{Rad}_S(\mathcal{W})$ is the empirical Rademacher complexity, defined as

$$\mathbf{Rad}_S(\mathcal{W}) := \frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{w\in\mathcal{W}}\sum_{i=1}^n \epsilon_i \ell(w, z_i)\right]. \tag{11}$$

In this equation $\epsilon := (\epsilon_1, \ldots, \epsilon_n)$ is a vector of i.i.d. Rademacher random variables, characterized by $\mathbb{P}(\epsilon_i = 1) = \mathbb{P}(\epsilon_i = -1) = 1/2$.

We also define the (expected) Rademacher complexity as $\mathbf{Rad}(\mathcal{W}) := \mathbb{E}_{S\sim\mu_z^{\otimes n}}[\mathbf{Rad}_S(\mathcal{W})]$.

The Rademacher complexity $\mathbf{Rad}(\mathcal{W})$, and its empirical counterpart $\mathbf{Rad}_S(\mathcal{W})$, can be interpreted as a complexity measure of the hypothesis set $\mathcal{W}$ (Shalev-Schwartz and Ben-David, 2014). As examples of applications of these Rademacher complexity-based bounds, by instantiating the hypothesis set $\mathcal{W}$, one can upper-bound the Rademacher complexity for linear classifiers (Bartlett and Mendelson, 2002; Kakade et al., 2008; Awasthi et al., 2020) or neural networks (Neyshabur et al., 2015; Bartlett et al., 2017).
The key ingredient in proving Theorem 1 is the so-called symmetrization lemma, *i.e.*, Lemma 25. It also plays a great role in our analysis, see Section 4.2. The fact that this lemma does not hold in the case of a data-dependent hypothesis set has already been noted by several studies (Foster et al., 2019; Dupuis et al., 2023). This is one of the main bottlenecks for having tight uniform generalization bounds. In Section 4.2, we will show how we can leverage the PAC-Bayesian theory (see Section 2.3) to remove this limitation and obtain bounds with Rademacher complexities of data-dependent hypothesis sets.

2.3 Background on PAC-Bayesian bounds

In the above section, we presented an example of uniform generalization bound over a fixed hypothesis set $\mathcal{W}$. Contrary to the uniform generalization bounds, the PAC-Bayesian framework takes a radically different viewpoint by considering *randomized* hypotheses; each hypothesis in $\mathcal{W}$ is associated with a weight characterizing its importance (see *e.g.*, Alquier, 2024). These weights are represented by probability measures. We distinguish two kinds of probability measures on $\mathcal{W}$: *(i)* $\mathcal{P}$, called the *prior distribution*[2], which does not depend on S. *(ii)* $\mathcal{Q}_S$, called the *posterior distribution*, that is learned based on the data S. In this setting, we can study the generalization gap of *randomized* hypotheses sampled from $\mathcal{Q}_S$ thanks to the *expected* generalization gap (with an expectation over $\mathcal{Q}_S$):

$$\mathbb{E}_{w\sim\mathcal{Q}_S}\left[\mathcal{R}(w) - \widehat{\mathcal{R}}_S(w)\right], \quad \text{with } S \sim \mu_z^{\otimes n}. \tag{12}$$

The PAC-Bayesian framework provides us techniques to bound the expected generalization gap: we present the following well-known bound from McAllester (2003); Maurer (2004).

Theorem 2 (McAllester (2003); Maurer (2004)) *We assume that ℓ is bounded in $[0,1]$. Let $\mathcal{P}$ be any prior distribution on $\mathcal{W}$. For any Markov kernel[3] $\mathcal{Q}_S$, if for all $S \in \mathcal{Z}^n$, we have $\mathcal{Q}_S \ll \mathcal{P}$, then with probability at least $1-\zeta$ over $S \sim \mu_z^{\otimes n}$ we have:*

$$\mathbb{E}_{w\sim\mathcal{Q}_S}\left[\mathcal{R}(w) - \widehat{\mathcal{R}}_S(w)\right] \leq \sqrt{\frac{\mathbf{KL}(\mathcal{Q}_S\|\mathcal{P}) + \log(\frac{2\sqrt{n}}{\zeta})}{2n}}$$

where $\mathbf{KL}(\mathcal{Q}_S\|\mathcal{P})$ is the Kullback Leibler (KL) divergence between $\mathcal{Q}_S$ and $\mathcal{P}$, whose definition is recalled in Appendix A.

Theorem 2 is a special case of a more general result. Considering any suitable function $\phi : \mathcal{W} \times \mathcal{Z}^n \to \mathbb{R}$, one can prove the following PAC-Bayesian bound (Germain et al., 2009).

2. We assume that the prior and posterior distributions on $\mathcal{W}$ are defined on an arbitrary σ-algebra $\Sigma_{\mathcal{W}}$.
3. This notion has been defined in Section 2.

Theorem 3 (PAC-Bayesian bound of Germain et al. (2009)) *Let $\zeta \in (0,1)$ and consider any measurable function $\phi : \mathcal{W} \times \mathcal{Z}^n \to \mathbb{R}$. Let $\mathcal{P}$ be any prior distribution on $\mathcal{W}$, such that e^{ϕ} is in $L^1(\mathcal{P} \otimes \mu_z^{\otimes n})$. We have, for any Markov kernel $\mathcal{Q}_S$ such that, for all $S \in \mathcal{Z}^n$, we have $\mathcal{Q}_S \ll \mathcal{P}$ and $\phi(\cdot, S) \in L^1(\mathcal{Q}_S)$:*

$$\mathbb{P}_S\left(\mathbb{E}_{w\sim\mathcal{Q}_S}[\phi(w,S)] \leq \mathbf{KL}(\mathcal{Q}_S\|\mathcal{P}) + \log\frac{1}{\zeta} + \log \mathbb{E}_S\mathbb{E}_{w\sim\mathcal{P}}e^{\phi(w,S)}\right) \geq 1-\zeta,$$

Put into words, Theorem 3 is an upper-bound of $\mathbb{E}_{w\sim\mathcal{Q}_S}\phi(w,S)$, holding with probability at least $1-\zeta$ w.r.t. the sample S, and depending essentially on two terms. *(i)* The KL divergence between the probability measures $\mathcal{Q}_S$ and $\mathcal{P}$, which can be interpreted as a "distance" (that is not symmetric). *(ii)* The term $\log \mathbb{E}_S\mathbb{E}_{w\sim\mathcal{P}}e^{\phi(w,S)}$ that can be further upper-bounded when the function ϕ is instantiated. For instance, when $\phi(w,S) = 2n(\mathcal{R}(w) - \widehat{\mathcal{R}}_S(w))^2$, Theorem 3 boils down to Theorem 2.
PAC-Bayesian results such as Theorem 3 have been applied to various machine learning models such as linear classifiers (Herbrich and Graepel, 2000; Langford and Shawe-Taylor, 2002; Ambroladze et al., 2006; Langford, 2005; Germain et al., 2009; Parrado-Hernández et al., 2012), majority votes (Lacasse et al., 2006; Germain et al., 2015; Zantedeschi et al., 2021), or stochastic neural networks (Langford and Caruana, 2001; Dziugaite and Roy, 2017). However, in practice, we may be interested in a single hypothesis w, and, therefore, may want to control the generalization gap

$$\left\{\mathcal{R}(w) - \widehat{\mathcal{R}}_S(w)\right\} \quad \text{where: } S \sim \mu_z^{\otimes n},\ w \sim \mathcal{Q}_S, \tag{13}$$

instead of integrating over the posterior distribution $\mathcal{Q}_S$. This is made possible, to an extent, using so-called *disintegrated* PAC-Bayesian bounds (Blanchard and Fleuret, 2007; Catoni, 2007; Rivasplata et al., 2020; Viallard et al., 2024a). We present the bound proven by Rivasplata et al. (2020, Theorem 1*(i)*).

Theorem 4 (Disintegrated PAC-Bayesian bound of Rivasplata et al. (2020)) *Let $\zeta \in (0,1)$ and consider any measurable function $\phi : \mathcal{W} \times \mathcal{Z}^n \to \mathbb{R}$. Let $\mathcal{P}$ be any prior distribution on $\mathcal{W}$, such that e^{ϕ} is in $L^1(\mathcal{P} \otimes \mu_z^{\otimes n})$. We have, for any Markov kernel $\mathcal{Q}_S$ such that, for all $S \in \mathcal{Z}^n$, we have $\mathcal{Q}_S \ll \mathcal{P}$ and $\phi(\cdot, S) \in L^1(\mathcal{Q}_S)$:*

$$\mathbb{P}_{S,w\sim\mathcal{Q}_S}\left(\phi(w,S) \leq \log\left(\frac{d\mathcal{Q}_S}{d\mathcal{P}}(w)\right) + \log\frac{1}{\zeta} + \log \mathbb{E}_S\mathbb{E}_{w\sim\mathcal{P}}e^{\phi(w,S)}\right) \geq 1-\zeta.$$

We can see two main differences with Theorem 3: Theorem 4 is an upper-bound on $\phi(w,S)$ instead of the expectation $\mathbb{E}_{w\sim\mathcal{Q}}\phi(w,S)$ and the KL divergence is replaced by the logarithm of the Radon-Nikodym derivative $\log\frac{d\rho_S}{d\pi}(w)$, evaluated at $w \sim \mathcal{Q}_S$. Since the other terms remain unchanged, the mechanism is similar when instantiating ϕ, *i.e.*, the right-most term must be upper-bounded, and different choices of ϕ may lead to different generalization bounds. The proof of the above theorem follows from the derivations of Rivasplata et al. (2020); we note that it rigorously holds under the absolute continuity assumption.

3. PAC-Bayesian Theory on Random Sets

In this section, we introduce our framework for PAC-Bayesian bounds for random sets. As mentioned in the introduction, our goal is to reformulate the PAC-Bayesian bounds for random hypothesis sets. This section may be seen as a direct consequence of classical PAC-Bayesian theory. Indeed, it will be noted that Theorems 3 and 4 are valid in a more general setting, in which $\mathcal{W}$ is replaced by an arbitrary probability space. We first present, in Section 3.1, a generic approach to generalize the PAC-Bayesian bounds of Theorems 3 and 4 for random sets. In Section 3.2, we propose a more detailed construction based on the notion of random closed sets (Molchanov, 2017, Chapter 1), therefore providing a sound theoretical foundation for the introduced methods.

3.1 Random set formalization

We consider a set $E \subseteq \mathfrak{P}(\mathbb{R}^d)$, together with a σ-algebra $\mathfrak{E}$, making $(E, \mathfrak{E})$ a measurable space. We will now rewrite known PAC-Bayesian bounds by replacing random hypothesis $w \in \mathbb{R}^d$ by random hypothesis sets[4] $\mathcal{W} \in E$. E should be interpreted as the collection of all possible hypothesis sets. According to our PAC-Bayesian approach, we consider a *learning algorithm* as a mapping generating a data-dependent probability distribution on $(E, \mathfrak{E})$ from a dataset $S \in \mathcal{Z}^n$. More formally, this leads to the following definition.

Definition 5 (Priors and posteriors) *A* prior*, π, is a data-independent probability distribution on $(E, \mathfrak{E})$. A family of posteriors $(\rho_S)_{S \in \mathcal{Z}^n}$, is defined as a Markov kernel on $E \times \mathcal{Z}^n$. We further require that the posteriors are absolutely continuous with respect to the prior, i.e. $\rho_S \ll \pi$, $\mu_z^{\otimes n}$-almost surely.*

Whenever we consider priors and posteriors in the remainder of the paper, we assume that the properties of Definition 5 are satisfied. This framework encompasses several classical settings, such as the following examples.

Example 1 (Singleton distributions) *Assume that, for $\mathcal{W} \in E$, there exists $w \in \mathbb{R}^d$, such that $\mathcal{W} = \{w\}$, π-almost surely. Then the distributions ρ_S and π naturally extend to distributions over $\mathbb{R}^d$. In that case, the presented framework reduces to the classical PAC-Bayesian setting, as given in Section 2.3, where the distributions are defined over $\mathbb{R}^d$.*

Example 2 (Stochastic Gradient Descent) *Consider the Stochastic Gradient Descent algorithm (SGD), applied over T iterations to the minimization of the empirical risk $\widehat{\mathcal{R}}_S$. For a dataset $S \in \mathcal{Z}^n$ and external randomness $\mathcal{A}$, coming from the choice of the batch indices, SGD generates the iterates $(w_1^{S,\mathcal{A}}, \dots, w_T^{S,\mathcal{A}}) \in \left(\mathbb{R}^d\right)^T$. Then ρ_S could be defined as the conditional distribution of the sets $\left\{w_1^{S,\mathcal{A}}, \dots, w_T^{S,\mathcal{A}}\right\}$, given the data S.*

Example 3 (Stochastic Differential Equations) *Consider the Stochastic Differential Equation (SDE) $dW_t^S = -\nabla \widehat{\mathcal{R}}_S(W_t^S)dt + \sigma dX_t$, where $\widehat{\mathcal{R}}_S$ is the empirical risk and $(X_t)_{t \geq 0}$ is a well-behaved stochastic process. For instance, X could be a Brownian motion in the case of Langevin dynamics (Mou et al., 2018), or a Lévy process, in the case of heavy-tailed*

4. The notation $\mathcal{W}$ will always refer to sets rather than points.

dynamics (Şimşekli et al., 2019; Gürbüzbalaban et al., 2021; Dupuis and Şimşekli, 2024). In such a setting, given a fixed time horizon $T > 0$, the posterior ρ_S describes the distributions of the sets $\{W_t^S,\ 0 \leq t \leq T\} \subset \mathbb{R}^d$. We will cover such a setting in more detail in Section 6.

We can now extend the PAC-Bayesian bounds of Theorems 3 and 4 in a straightforward manner. To do so, we only need to consider a function $\Phi : E \times \mathcal{Z}^n \longrightarrow \mathbb{R}$ measurable with respect to $\mathfrak{E} \otimes \mathcal{F}^{\otimes n}$, and apply Theorems 3 and 4. The following example illustrates a typical such function Φ; other pertinent choices will be discussed in Section 4.

Example 4 (Supremum function) *The supremum of the generalization error,*

$$\Phi_{\sup}(\mathcal{W}, S) := \sup_{w \in \mathcal{W}} \left(\mathcal{R}(w) - \widehat{\mathcal{R}}_S(w) \right), \tag{14}$$

can be used under mild assumptions over E and ℓ. For instance, it is the case for almost surely finite hypothesis sets or for random closed sets, as it will be discussed in Section 3.2.

The above example will be further refined in Section 4, which illustrates the capacity of our approach to prove worst-case generalization bounds over data-dependent hypothesis sets. The following theorem is a direct consequence of Theorems 3 and 4.

Theorem 6 (PAC-Bayesian bounds for random sets) *Let $(E, \mathfrak{E})$ be defined as before and $\Phi : E \times \mathcal{Z}^n \longrightarrow \mathbb{R}$ be a measurable function. We also consider a prior π and posteriors $(\rho_S)_{S \in \mathcal{Z}^n}$, as in Definition 5. Then we have for any $\zeta \in (0, 1)$:*

$$\mathbb{P}_S \left(\mathbb{E}_{\mathcal{W} \sim \rho_S} \Phi(\mathcal{W}, S) \leq \mathbf{KL}(\rho_S \| \pi) + \log(1/\zeta) + \log \mathbb{E}_S \mathbb{E}_{\mathcal{W} \sim \pi} \left[e^{\Phi(\mathcal{W}, S)} \right] \right) \geq 1 - \zeta, \tag{15}$$

as well as the disintegrated bound

$$\mathbb{P}_{S, \mathcal{W} \sim \rho_S} \left(\Phi(\mathcal{W}, S) \leq \log \left(\frac{d\rho_S}{d\pi}(\mathcal{W}) \right) + \log(1/\zeta) + \log \mathbb{E}_S \mathbb{E}_{\mathcal{W} \sim \pi} \left[e^{\Phi(\mathcal{W}, S)} \right] \right) \geq 1 - \zeta, \tag{16}$$

with both bounds holding as long as all expectations appearing are well-defined.

Proof It is a consequence of the results presented in Section 2.3, namely Theorems 3 and 4, adapted to the Markov kernel $(\rho_S)_{S \in \mathcal{Z}^n}$ and the probability space $(E, \mathfrak{E})$. ■

To further illustrate that our framework generalizes that of classical PAC-Bayesian theory, we provide the following example.

Example 5 *In the singleton distributions setting of Example 1, Theorem 6 is equivalent to classical PAC-Bayesian bounds found in (Alquier, 2024; Rivasplata et al., 2020; Germain et al., 2009), if we use the supremum function of Equation* (14)*. Indeed, in that case, if the loss ℓ is bounded by B, we have, by Hoeffding's lemma and Fubini's theorem:*

$$\mathbb{E}_S \mathbb{E}_{\mathcal{W} \sim \pi} \left[e^{\lambda \Phi_{\sup}(\mathcal{W}, S)} \right] = \mathbb{E}_{\{w\} \sim \pi} \mathbb{E}_S \left[e^{\lambda \left(\mathcal{R}(w) - \widehat{\mathcal{R}}_S(w) \right)} \right] \leq e^{\frac{\lambda^2 B^2}{8n}}.$$

Theorem 6 shows that, to deduce meaningful generalization bounds over data-dependent hypothesis sets, one must be able to bound both the IT terms, namely $\mathbf{KL}(\rho_S\|\pi)$ and $\frac{d\rho_S}{d\pi}(\mathcal{W})$, on one side, and the log-exp terms, $\log \mathbb{E}_S\mathbb{E}_{\mathcal{W}\sim\pi}\left[e^{\Phi(\mathcal{W},S)}\right]$, on the other side. In the following sections, we will show, through several examples, how to design well-suited functions Φ to obtain such bounds. We will also describe techniques that help analyze the IT terms in several cases. In the general case, these IT terms measure the deviation of the posterior distribution ρ_S from the prior distribution π. In particular, the Radon-Nikodym derivative term measures the ratio of posterior and prior probability on each random hypothesis set and may be seen as a "disintegrated relative entropy".
Let us make a short remark to highlight the generality of our framework.

Remark 7 *Theorem 6 is stated using a Markov kernel formulation, i.e., $(\rho_S)_{S\in\mathcal{Z}^n}$ is a Markov kernel on $E\times\mathcal{Z}^n$, according to Definition 5. However, PAC-Bayesian bounds are often stated uniformly over the choice of posterior distribution (see e.g., Alquier, 2024). As Theorem 6 is a direct extension of the existing PAC-Bayesian bounds, it is also possible to state Equation* (15) *in this fashion. More precisely, if $\mathcal{P}_0(\mathbb{R}^d)$ denotes the family of probability distributions on $\mathbb{R}^d$, Equation* (15) *becomes:*

$$\mathbb{P}_S\left(\forall\rho\in\mathcal{P}_0(\mathbb{R}^d),\ \mathbb{E}_\rho\left[\Phi(\mathcal{W},S)\right]\le\mathbf{KL}(\rho\|\pi)+\log(1/\zeta)+\log\mathbb{E}_S\mathbb{E}_\pi\left[e^{\Phi(\mathcal{W},S)}\right]\right)\ge 1-\zeta.$$

In our application, we will use the Markov Kernel formulation, but our KL-based bounds in Theorems 13, 15 and 16 can be written uniformly over the choice of posterior ρ. Note however that, to the best of our knowledge, this uniform formulation is not possible in the disintegrated case, i.e., for Equation (16).

In the same way that we extended Theorems 3 and 4 into our framework, it is clear that other parts of the PAC-Bayesian literature may be treated the same way. In particular, the disintegrated framework of Viallard et al. (2024a) could be used similarly. It would also be possible to use tighter versions of the proposed bounds. For better clarity, we will focus on the bounds presented in Theorem 6 in the sequel. Additionally, more general PAC-Bayesian bounds could be formulated within our framework. For instance, we could state PAC-Bayesian uniform generalization bounds featuring the integral probability metrics (IPM) used by Amit et al. (2022), see Appendix B.2 for more details. Note that our proofs can also be adapted for other bounds with IPM developed by Viallard et al. (2023, 2024b).

3.2 More detailed measure-theoretic construction

The methods described in Section 3.1 are valid as soon as the general PAC-Bayesian theory applies to the measurable spaces $(E,\mathfrak{E})$ and the functions Φ under consideration. We already presented, through several examples (Examples 1 to 3), settings in which this will be the case under mild assumptions on the loss $\ell(w,z)$ (*e.g.*, bounded loss or integrability assumptions). We now provide a general measure-theoretic construction that allows us to apply our framework more widely. Note that this section may be skipped without harming the general understanding of the paper.
Our goal is to define a measurable space $(E,\mathfrak{E})$ of random sets with enough structure so that the measurability conditions of Theorem 6 are satisfied. Inspired by Molchanov (2017), we

restrict the analysis of this section to the theory of *random closed sets*. Note that, as soon as the loss function $\ell(w, z)$ is continuous in w, there is no loss of generality in considering only closed sets. The reader may find in (Molchanov, 2017) an extensive overview of the theory of random sets. A similar formalization of learning algorithms through random closed sets has been considered in (Hodgkinson et al., 2022) and (Dupuis et al., 2023).
Let us denote by $\mathbf{CL}(\mathbb{R}^d)$ the set of closed sets in $\mathbb{R}^d$ and give the definition of a suitable σ-algebra on $\mathbf{CL}(\mathbb{R}^d)$, called the Effrös σ-algebra.

Definition 8 (Effrös σ-algebra) *Let $\mathcal{O}(\mathbb{R}^d)$ be the set of open sets of $\mathbb{R}^d$. The Effrös σ-algebra on $\mathbb{R}^d$, denoted $\mathfrak{E}(\mathbb{R}^d)$, is the σ-algebra on $\mathbf{CL}(\mathbb{R}^d)$, generated by:*

$$\left\{\mathcal{F}_U,\ U \in \mathcal{O}(\mathbb{R}^d)\right\}, \quad \text{with } \mathcal{F}_U := \left\{C \in \mathbf{CL}(\mathbb{R}^d),\ C \cap U \neq \emptyset\right\}.$$

Given $(\Omega, \mathcal{T})$ a measurable space, as explained in (Molchanov, 2017, Section 1.1.1), we may define a random closed set as a measurable mapping

$$\mathcal{W} : (\Omega, \mathcal{T}) \longrightarrow (\mathbf{CL}(\mathbb{R}^d), \mathfrak{E}(\mathbb{R}^d)).$$

The following lemma ensures that we now have enough structure to apply the PAC-Bayesian theory in its general form.

Lemma 9 *Let $(\Omega, \mathcal{T})$ be a measurable space and $\zeta : \mathbb{R}^d \times \Omega \longrightarrow \mathbb{R}$ a stochastic process, which is continuous in $w \in \mathbb{R}^d$. Define, for $\mathcal{W} \in \mathbf{CL}(\mathbb{R}^d)$ and $\omega \in \Omega$, the map $\Phi(\mathcal{W}, \omega) := \sup_{w \in \mathcal{W}} \zeta(w, \omega)$. Then, Φ is measurable with respect to $\mathfrak{E}(\mathbb{R}^d) \otimes \mathcal{T}$.*

In particular, this implies that the supremum of the generalization error, *i.e.*, $G_S(\mathcal{W})$ defined by Equation (3), is measurable with respect to $\mathfrak{E}(\mathbb{R}^d) \otimes \mathcal{F}^{\otimes n}$. It also implies a similar measurability result on the Rademacher complexity terms that will appear in Sections 4.2 and 4.3. The proofs of this section are discussed in Appendix B.6.

4. Uniform Generalization Bounds with Data-dependent Hypothesis Sets

In this section, we present some of our main results, which consist of applying the framework described in Section 3.1 to informed choices of the function Φ appearing in Theorem 6. Let us first define our generic assumptions. We will explicitly mention the assumptions they require for each statement in the sequel.

Assumption 1 (Bounded measurable loss) *The loss function $\ell : \mathbb{R}^d \times \mathcal{Z}$ is measurable and bounded in $[0, B]$, for some constant $B > 0$.*

Moreover, we denote, as in Section 3.1, the probability space of hypothesis sets as $(E, \mathfrak{E})$. We fix a prior π and posteriors $(\rho_S)_{S \in \mathcal{Z}^n}$, defined on E, according to Definition 5. We make the following technical assumption, which will ensure that all quantities appearing in the rest of this section are well-defined and measurable. Note that Section 3.2 justifies that this assumption holds in numerous settings.

Assumption 2 (Supremum measurability) *Both ℓ and $(E, \mathfrak{E})$ have enough regularity so that, for any coefficients $b, a_1, \dots, a_n \in \mathbb{R}$, the following is $\mathfrak{E} \otimes \mathcal{F}^{\otimes n}$-measurable:*

$$(\mathcal{W}, S) \longmapsto \sup_{w \in \mathcal{W}} \sum_{i=1}^{n} \left(a_i \ell(w, z_i) - b\mathcal{R}(w)\right).$$

We will prove three generalization bounds. First, in Section 4.1, we will build up on the supremum function given in Example 4 to derive our first bounds. While interesting, this approach appears to be inefficient in some cases, which will be made clear later. To solve this issue, we show, in Section 4.2, how a slight change in the function Φ, can lead to a generalization bound in terms of *data-dependent Rademacher complexity*. Finally, in Section 4.3, the same methods are applied to derive a data-dependent uniform lower bound. An advantage of our framework is that all these bounds come with the same interpretable IT terms and apply to a wide variety of settings.

4.1 Warm-up: a first bound with the moment generating Rademacher function

We first apply Theorem 6, to the function

$$\Phi_\lambda(\mathcal{W}, S) := \lambda \sup_{w \in \mathcal{W}} \Big(\mathcal{R}(w) - \widehat{\mathcal{R}}_S(w)\Big) = \lambda \Phi_{\sup}(S, \mathcal{W}), \quad \mathcal{W} \in E,\ S \in \mathcal{Z}^n, \tag{17}$$

for $\lambda > 0$. The introduction of the parameter λ, in the above equation, is a classical trick in PAC-Bayesian theory, as this parameter can be further optimized in particular applications to obtain generalization bounds in a more compact form, see for instance Remark 12 below. Before writing our generalization bound, we introduce the moment generating function (MGF) of the Rademacher complexity, defined as:

$$\forall \lambda > 0, \quad \Psi_{S,\mathcal{W}}(\lambda) = \mathbb{E}_\epsilon \left[\exp\left\{\frac{\lambda}{n} \sup_{w \in \mathcal{W}} \sum_{i=1}^{n} \epsilon_i \ell(w, z_i)\right\}\right], \tag{18}$$

where $\epsilon_1, \dots, \epsilon_n$ are *i.i.d.* Rademacher random variables and $\mathcal{W} \subseteq \mathbb{R}^d$ is a set.
The following theorem is a PAC-Bayesian type bound for the worst-case generalization error in terms of the MGF of the Rademacher complexity.

Theorem 10 (PAC-Bayesian bounds with Rademacher MGF) *Under Assumption 2, we have, for any $\lambda > 0$, the following bounds, as soon as the expectations are well defined:*

$$\mathbb{P}_S\left(\mathbb{E}_{\mathcal{W} \sim \rho_S}\left[\lambda \sup_{w \in \mathcal{W}} \left(\mathcal{R}(w) - \widehat{\mathcal{R}}_S(w)\right)\right] \le \mathbf{KL}(\rho_S \| \pi) + \log(1/\zeta) + M(\lambda)\right) \ge 1 - \zeta,$$

$$\mathbb{P}_{S, \mathcal{W} \sim \rho_S}\left(\lambda \sup_{w \in \mathcal{W}} \left(\mathcal{R}(w) - \widehat{\mathcal{R}}_S(w)\right) \le \log\left(\frac{d\rho_S}{d\pi}(\mathcal{W})\right) + \log(1/\zeta) + M(\lambda)\right) \ge 1 - \zeta,$$

with $M(\lambda) := \log \mathbb{E}_{\mathcal{W} \sim \pi} \mathbb{E}_S \Psi_{S,\mathcal{W}}(2\lambda)$.

The main interest in this theorem is that it does not require any boundedness assumption on the loss. The proof of Theorem 10, deferred to Appendix B.1.1, is based on an "exponential symmetrization lemma", similar to the usual symmetrization inequality for Rademacher complexity, *i.e.*, Lemma 25. We present in the following example a simple case where the term $M(\lambda)$ can be simplified and the parameter λ accordingly optimized.

Example 6 (Almost surely finite random sets) *Let us assume that Assumption 1 holds and that $\mathcal{W}$ is π-almost surely finite (note that its cardinality is still random). Then, by Fubini's theorem and Hoeffding's lemma, we have (we mimic the proof of Massart's lemma):*

$$\Psi_{S,\mathcal{W}}(2\lambda) \leq \mathbb{E}_\epsilon\left[\sum_{w\in\mathcal{W}} \exp\left\{\frac{2\lambda}{n}\sum_{i=1}^n \epsilon_i \ell(w,z_i)\right\}\right] \leq \sum_{w\in\mathcal{W}} e^{\frac{2\lambda^2B^2}{n}} = |\mathcal{W}|e^{\frac{2\lambda^2B^2}{n}}.$$

This gives that, with probability at least $1-\zeta$ over $S\sim\mu_z^{\otimes n}$ and $\mathcal{W}\sim\rho_S$:

$$\lambda \sup_{w\in\mathcal{W}}\left(\mathcal{R}(w)-\widehat{\mathcal{R}}_S(w)\right) \leq \log\left(\frac{d\rho_S}{d\pi}(\mathcal{W})\right) + \log(1/\zeta) + \log\left(\mathbb{E}_\pi\left[|\mathcal{W}|\right]\right) + \frac{2\lambda^2B^2}{n}.$$

We refer to Remark 12 for a discussion on the role and optimization of the parameter λ.

In general, the quantity $\Psi_{S,\mathcal{W}}(2\lambda)$ could be bounded using covering arguments, similarly to (Şimşekli et al., 2020) and (Dupuis et al., 2023). However, because of the expectation over the prior appearing in the definition of $M(\lambda)$, such techniques would only lead to non-data-dependent quantities. This can be seen in Example 6, where the left-hand side of the bound features the expected (over the prior π) cardinality of $\mathcal{W}$ instead of a data-dependent term. As our goal is to provide data-dependent generalization bounds, this shows that the particular choice of function Φ used in this subsection has to be improved. Nevertheless, Theorem 10 will be used to derive our uniform SGLD bounds in Section 7.
In the next subsection, we present an alternative approach towards data-dependent generalization bounds over random sets.

4.2 Generalization bounds with data-dependent Rademacher complexity

In this section, we use our framework to prove uniform generalization bounds in terms of a data-dependent Rademacher complexity, which is a term of the form $\mathbf{Rad}_S(\mathcal{W}_S)$, where the hypothesis set $\mathcal{W}_S$ depends on the data S. We remind the reader that $G_S(\mathcal{W})$ was defined in Equation (3). In this section, we apply Theorem 6 to the following function:

$$\Phi_\lambda(\mathcal{W},S) = \lambda G_S(\mathcal{W}) - 2\lambda\mathbf{Rad}_S(\mathcal{W}), \qquad \lambda>0. \tag{19}$$

This leads to the following theorem, which is a PAC-Bayesian data-dependent uniform generalization bound involving a data-dependent[5] Rademacher complexity term.

Theorem 11 (Data-dependent Rademacher complexity bound) *Suppose that Assumptions 1 and 2 hold. Then, for any $\lambda>0$ we have*

$$\mathbb{P}_S\left(\mathbb{E}_{\mathcal{W}\sim\rho_S}\left[G_S(\mathcal{W})\right] \leq \mathbb{E}_{\mathcal{W}\sim\rho_S}\left[2\mathbf{Rad}_S(\mathcal{W})\right] + \frac{\mathbf{KL}(\rho_S\|\pi)+\log(1/\zeta)}{\lambda} + \lambda\frac{9B^2}{8n}\right) \geq 1-\zeta,$$

$$\mathbb{P}_{S,\mathcal{W}\sim\rho_S}\left(G_S(\mathcal{W}) \leq 2\mathbf{Rad}_S(\mathcal{W}) + \frac{\log\frac{d\rho_S}{d\pi}(\mathcal{W})+\log(1/\zeta)}{\lambda} + \lambda\frac{9B^2}{8n}\right) \geq 1-\zeta.$$

5. We use the expression "data-dependent Rademacher complexity" to highlight the fact that the hypothesis set $\mathcal{W}$ on which it is applied is data-dependent.

The IT terms appearing in Theorem 11 are exactly the same as those in Theorem 6 and Theorem 10, which illustrates the universality of our approach. Therefore, comparing the bounds obtained within our framework can be achieved regardless of these IT terms.
The data dependence of the Rademacher complexity terms $\mathbf{Rad}_S(\mathcal{W})$, appearing in Theorem 11, might not be obvious at first sight. This data dependence comes from the fact that $\mathcal{W}$ is here drawn from the posterior ρ_S. If we look back at Example 2, such a set $\mathcal{W}$ would be a data-dependent trajectory of SGD obtained while training on the dataset S.
Foster et al. (2019) also considered data-dependent Rademacher complexity terms. However, their results rely on so-called hypothesis set stability assumptions. We point out the fact that no such assumption is needed to derive Theorem 11. Our result can also be compared with (Sachs et al., 2023), in which the authors derived generalization bounds in terms of "algorithmic-dependent Rademacher complexity", *i.e.*, a notion of complexity depending on the algorithm, but not directly on the dataset S used in training. This is to be opposed to the term $\mathbf{Rad}_S(\mathcal{W})$ in Theorem 11, which depends explicitly on the dataset $S \in \mathcal{Z}^n$.
Building on classical arguments in learning theory, this result opens the door to introducing other data-dependent terms in the bounds by further bounding the Rademacher complexity term. In particular, we may introduce VC dimension terms and/or perform covering arguments and chaining techniques. In Section 5, we will focus on fractal-dimension-based generalization bounds, where the proposed methods are particularly useful.
Let us make a remark about the role of the parameter $\lambda > 0$ appearing in the above theorem.

Remark 12 *Theorem 11, as well as our other main results (Theorems 10 and 15), along with their applications in the sequel, are valid for any $\lambda > 0$. In many contexts, this parameter λ can be optimized to simplify the expression of our generalization bounds. In particular, using a proof technique similar to (London, 2017, Lemma 9), one can see that Theorem 1 implies that with probability at least $1-\zeta$ over $S \sim \mu_z^{\otimes n}$, we have:*

$$\mathbb{E}_{\mathcal{W}\sim\rho_S}\big[G_S(\mathcal{W})\big] \leq \mathbb{E}_{\mathcal{W}\sim\rho_S}\big[2\mathbf{Rad}_S(\mathcal{W})\big] + 6B\sqrt{\frac{\mathbf{KL}(\rho_S\|\pi) + \log(2/\zeta)}{2n}}. \tag{20}$$

A similar result could be shown for the disintegrated bound. Any bound (e.g., Sections 5 to 7) with a parameter λ playing a similar role can be further optimized in this way.
However, one can notice that in the case of a data-independent hypothesis set $\mathcal{W}$, where we have $\mathbf{KL}(\rho_S\|\pi) = 0$, Equation (20) recovers worse absolute constants than the previously known generalization bounds, as given by Theorem 1. For this reason, we chose to present all our results in the more general form, with a free parameter $\lambda > 0$. Depending on the application, the reader may use any of the two formulations.

A classical question in PAC-Bayesian analysis concerns the minimization of the bound with respect to the posterior distribution (Alquier, 2024, Section 2.1.2) and leads to the consideration of the so-called Gibbs posterior. Such an analysis extends to our framework. Indeed, by Remark 7 and Theorem 11, we have that, with probability at least $1-\zeta$ over $S \sim \mu_z^{\otimes n}$, for every posterior distribution ρ, we have:

$$\mathbb{E}_{\mathcal{W}\sim\rho}\left[\sup_{w\in\mathcal{W}} \mathcal{R}(w)\right] \leq \mathbb{E}_{\mathcal{W}\sim\rho}\left[\sup_{w\in\mathcal{W}} \widehat{\mathcal{R}}_S(w) + 2\mathbf{Rad}_S(\mathcal{W})\right] + \frac{\mathbf{KL}(\rho\|\pi) + \log(1/\zeta)}{\lambda} + \lambda\frac{9B^2}{8n}.$$

By a classical application of Donsker-Varadhan's variational formula, we see that the probability distribution ρ minimizing the right-hand-side of this expression is the following "Gibbs-Rademacher posterior", defined for any $\lambda > 0$ by:

$$d\rho_S^{(\lambda)}(\mathcal{W}) \propto \exp\left\{-\lambda \sup_{w\in\mathcal{W}} \widehat{\mathcal{R}}_S(w) - 2\lambda \mathbf{Rad}_S(\mathcal{W})\right\} d\pi(\mathcal{W}), \tag{21}$$

where the symbol $\propto$ indicates that the appropriate normalization factor has been omitted. Equation (21) provides a family of posteriors that are optimal in the sense that they minimize the worst population risk $\mathcal{R}(w)$ over $\mathcal{W}$, given a prior distribution π. In the case of singleton random sets (Example 1), they generalize the Gibbs distributions classically encountered in PAC-Bayesian analysis (Alquier, 2024), which are of the form $d\mathcal{P}_S^{(\lambda)} \propto e^{-\lambda\widehat{\mathcal{R}}_S(w)} d\mathcal{Q}(w)$. The form of $\rho_S^{(\lambda)}$ suggests that the best learning algorithm (given a prior π) samples data-dependent hypothesis sets with the lowest Rademacher complexities. The proof of Theorem 11, presented in Appendix B.1, highlights that Theorem 11 is a consequence of the symmetrization lemma, *i.e.*, Lemma 25. This suggests the following general form, which could be used to derive a wide variety of generalization bounds, involving other types of functionals than the ones we use here but still using the same IT terms.

Theorem 13 (A general form of set-dependent bounds) *Let $\Phi : E \times \mathcal{Z}^n \longrightarrow \mathbb{R}$ be a measurable function. We assume that for every $\mathcal{W} \in E$, we have $\mathbb{E}_S[\Phi(\mathcal{W}, S)] \leq 0$ and $\Phi(\mathcal{W}, S) - \mathbb{E}_S[\Phi(\mathcal{W}, S)]$ is σ^2-subgaussian.*[6] *Then, there exists an absolute constant C such that, for any $\lambda > 0$:*

$$\mathbb{P}_S\left(\mathbb{E}_{\mathcal{W}\sim\rho_S}\left[\Phi(\mathcal{W}, S)\right] \leq \frac{\mathbf{KL}(\rho_S\|\pi) + \log(1/\zeta)}{\lambda} + \lambda\frac{C\sigma}{n}\right) \geq 1-\zeta,$$

$$\mathbb{P}_{S,\mathcal{W}\sim\rho_S}\left(\Phi(\mathcal{W}, S) \leq \frac{\log\frac{d\rho_S}{d\pi}(\mathcal{W}) + \log(1/\zeta)}{\lambda} + \lambda\frac{C\sigma}{n}\right) \geq 1-\zeta.$$

Remark 14 *In the proof of Theorem 11, the subgaussian condition is, in this case, a consequence of McDiarmid's inequality (McDiarmid, 1998). This opens the door for further generalizations of our framework by considering Bernstein forms of this inequality (McDiarmid, 1998, Theorem 3.8); see also the note of Ying (2004). In our case, this would allow us to remove the bounded loss assumption, Assumption 1, at the cost of introducing intricate variance terms in the bound. We leave this as a direction for future work.*

4.3 Data-dependent generalization lower bounds

Theorem 13 provides us with a general recipe to prove PAC-Bayesian bounds. To illustrate this ability, we now derive a *data-dependent generalization lower bound*, based on Rademacher complexity. To the best of our knowledge, such data-dependent uniform lower bounds have not been derived before.

In the same way that Theorem 11 was deduced from the symmetrization lemma (which fulfills the first condition of Theorem 13), the main result of this subsection is based on the so-called desymmetrization inequality (recalled in Appendix A.2).

6. X is said to be subgaussian if, for every $\lambda > 0$, we have $\mathbb{E}\left[e^{\lambda X}\right] \leq e^{\frac{\lambda^2\sigma^2}{2}}$, see (Vershynin, 2018).

This leads us to the consideration of the following set function:

$$\Phi_{\text{lower bound}}(\mathcal{W}, S) := \frac{1}{2}\mathbf{Rad}_S(\mathcal{W}) - \frac{B}{2\sqrt{n}} - \sup_{w\in\mathcal{W}} |\widehat{\mathcal{R}}_S(w) - \mathcal{R}(w)|, \tag{22}$$

from which we deduce a lower bound on $G_S^{\text{abs}}(\mathcal{W}) := \sup_{w\in\mathcal{W}} |\widehat{\mathcal{R}}_S(w) - \mathcal{R}(w)|$, which is presented in the next theorem.

Theorem 15 (Data-dependent lower bound) *Suppose that Assumptions 1 and 2 hold. There exists an absolute constant $C > 0$, such that, for any $\lambda > 0$ we have, with probability at least $1-\zeta$ over $S \sim \mu_z^{\otimes n}$:*

$$\mathbb{E}_{\mathcal{W}\sim\rho_S}\left[G_S^{\text{abs}}(\mathcal{W})\right] \geq \frac{1}{2}\mathbb{E}_{\mathcal{W}\sim\rho_S}\left[\mathbf{Rad}_S(\mathcal{W})\right] - \frac{B}{2\sqrt{n}} - \frac{\mathbf{KL}(\rho_S\|\pi) + \log(1/\zeta)}{\lambda} - \frac{CB^2\lambda}{n}.$$

Moreover, we have the disintegrated lower bound:

$$\mathbb{P}_S\mathbb{P}_{\mathcal{W}\sim\rho_S}\left(G_S^{\text{abs}}(\mathcal{W}) \geq \frac{1}{2}\mathbf{Rad}_S(\mathcal{W}) - \frac{B}{2\sqrt{n}} - \frac{\frac{d\rho_S}{d\pi}(\mathcal{W}) + \log(1/\zeta)}{\lambda} - \frac{CB^2\lambda}{n}\right) \geq 1-\zeta.$$

Proof This is a direct consequence of Theorem 13, where Φ is chosen to be $\Phi_{\text{lower bound}}(\mathcal{W}, S)$, from Equation (22). The negative expectation follows from Proposition 26 and the sub-gaussian property from McDiarmid's inequality, as in the proof of Theorem 11. ■

5. Fractal-Dimension-Based Data-Dependent Generalization Bounds

In this section, we detail the application of our framework to fractal-dimension-based generalization bounds, as mentioned in the introduction.
We start by defining covering numbers and fractal dimensions, in Section 5.1, before proving generalization bounds involving the two main kinds of fractal dimensions found in the literature, namely the "data-dependent" dimension (Dupuis et al., 2023) and the "Euclidean-based" dimension (Şimşekli et al., 2020; Hodgkinson et al., 2022). Our bounds are deduced from standard covering arguments applied to the data-dependent Rademacher complexity term appearing in Theorem 11. These data-dependent covering bounds, which extend classical covering arguments for data-independent hypothesis sets (Shalev-Schwartz and Ben-David, 2014), are presented in Appendix B.3.1. Finally, Section 5.4 is devoted to the comparison of our information-theoretic terms and the ones existing in the literature.

5.1 Covering numbers and fractal dimensions

To be able to encapsulate all the existing fractal-dimension-based generalization bounds, we define coverings in full generality in pseudometric spaces. We say that (X, ϑ) is a pseudometric space if $\vartheta : X \times X \longrightarrow \mathbb{R}_+$ is symmetric, satisfies the triangle inequality, and vanishes on the diagonal (*i.e.* $\vartheta(x, x) = 0$). Given a compact pseudometric space (X, ϑ), we define, for any $\delta > 0$, $N_\delta^\vartheta(X)$ to be the set of centers of a minimal covering of X by closed

δ-balls. The *covering number* is the cardinality of this set, denoted $|N_\delta^\vartheta(X)|$. In the case of the Euclidean distance in $\mathbb{R}^d$, the metric will be omitted in the notations.
It has been shown in (Dupuis et al., 2023) that, under mild assumptions on the measurability of the learning algorithm, we can construct measurable coverings, which will be assumed to be the case in all the following. This is formalized by the following assumption discussed in greater detail in Appendix B.6.1.

Assumption 3 (Measurable covering numbers) *The covering numbers appearing in this section are all measurable with respect to $\mathcal{F}^{\otimes n} \otimes \mathfrak{E}$.*

Our goal is to relate the generalization error to the *upper box-counting dimension* of the random hypothesis set $\mathcal{W} \sim \rho_S$. The upper box-counting dimension, or upper Minkowski dimension (Falconer, 2014), is defined for a compact pseudometric space (X, ϑ), by

$$\overline{\dim}_B^\vartheta(X) := \limsup_{\delta \to 0} \frac{\log(|N_\delta^\vartheta(X)|)}{\log(1/\delta)}. \tag{23}$$

The upper box-counting dimension is a central tool in fractal geometry. Intuitively, it may be seen as a measure of the complexity of a set, which extends the usual notion of dimension for vector spaces or Riemannian manifolds, see (Falconer, 2014; Mattila, 1999).
Other studies have used other notions of fractal dimension in learning theory, in particular the *Hausdorff dimension* (Şimşekli et al., 2020). This is made possible by leveraging technical geometrical assumptions, such as Ahlfors regularity (Mackay and Tyson, 2010), which guarantee the Minkowski and Hausdorff dimensions are equal.
Thanks to our PAC-Bayesian framework for random sets, we leverage the same proof technique to obtain generalization bounds with two variations of the upper box-counting dimension of the space $\mathcal{W}$. By varying the pseudometric ϑ in Equation (23), we define the following fractal dimensions:

- The **data-dependent fractal dimension** is based on the data-dependent pseudometric considered by Dupuis et al. (2023), and defined by:

$$\vartheta_S(w, w') := \frac{1}{n} \sum_{i=1}^n |\ell(w, z_i) - \ell(w', z_i)|, \tag{24}$$

  where $S \in \mathcal{Z}^n$ is a dataset. This dimension will be denoted $\overline{\dim}_B^{\vartheta_S}(\mathcal{W})$, accordingly with Equation (23). Note that Equation (24) defines a random pseudometric (*i.e.*, not a metric) because the map $w \longmapsto \ell(w, \cdot)$ might not be injective.

- The **Euclidean-based fractal dimension** is based on the Euclidean distance in $\mathbb{R}^d$, it will be simply denoted $\overline{\dim}_B(\mathcal{W})$. This is the intrinsic dimension studied in (Şimşekli et al., 2020; Birdal et al., 2021; Hodgkinson et al., 2022).

Let ϑ denote either ϑ_S, for some $S \in \mathcal{Z}^n$, or the Euclidean distance. It follows from classical arguments in learning theory (Shalev-Schwartz and Ben-David, 2014) that

$$\mathbf{Rad}_S(\mathcal{W}) \lesssim \inf_{\delta > 0} \left\{ \delta + \sqrt{\frac{\log(|N_\delta^\vartheta(\mathcal{W})|)}{n}} \right\}, \tag{25}$$

where absolute constants have been omitted. By combining Equation (25) with Theorem 11, we can obtain data-dependent covering bounds (see Appendix B.3.1). These results lay the foundation for our fractal-based generalization bounds, which we will now present. Note that classical covering bounds deduced from Rademacher complexity naturally have a data-dependent flavor, due to the presence of the pseudometric ϑ_S (Shalev-Schwartz and Ben-David, 2014; Dupuis et al., 2023). The novelty of our data-dependent covering bounds is to apply to data-dependent hypothesis sets, which leads to the introduction of additional information-theoretic terms.

5.2 Bounds with data-dependent fractal dimensions

In this section, we present our generalization bound based on the data-dependent fractal dimension, as defined in Section 5.1. Inspired by Dupuis et al. (2023, Theorem 3.4), our approach for obtaining these bounds is through using covering arguments and the link between covering numbers and the fractal dimension, *i.e.*, Equation (23). The drawback of this approach is that it relies on some terms whose dependence on n is unknown a priori. This issue is identified but not resolved in (Dupuis et al., 2023), leading to the introduction of constants with unknown dependence on n.
Despite these difficulties, our framework allows us to design a natural assumption to prove better generalization bounds. Indeed, the previous issue is due to the possible lack of uniformity in n of the limit in Equation (23) defining the upper box-counting dimension. In order to prove our data-dependent fractal dimension bound, we find that it is enough to assume that the convergence in probability (under $S \sim \mu_z^{\otimes n}$ and $\mathcal{W} \sim \rho_S$) of Equation (23) is uniform in n. To avoid harming the readability of this section, this assumption is presented in detail in Appendix B.3.2.
Based on Assumption 6, we can now state our generalization bounds in terms of the data-dependent fractal dimension induced by the data-dependent pseudometric of Equation (24).

Theorem 16 *Suppose that Assumptions 1, 2, 3 and 6 hold. Then there exists a constant $C > 0$ and for any $\lambda, \epsilon, \gamma > 0$, there exists $n_{\gamma,\epsilon} \in \mathbb{N}^\star$, such that, for $n \geq n_{\gamma,\epsilon}$, we have, with probability at least $1 - \zeta - \gamma$ under $S \sim \mu_z^{\otimes n}$ and $\mathcal{W} \sim \rho_S$, for any $\lambda > 0$:*

$$G_S(\mathcal{W}) \leq \frac{2}{n} + 2B\sqrt{\frac{2\left(\overline{\dim}_B^{\vartheta_S}(\mathcal{W}) + \epsilon\right)\log(n)}{n}} + \frac{\log \frac{d\rho_S}{d\pi}(\mathcal{W}) + \log(1/\zeta)}{\lambda} + C\lambda\frac{B^2}{n}.$$

The proof of this theorem is referred to Appendix B.3.2, it is significantly simpler than the proofs of Dupuis et al. (2023), hence underlining the effectiveness of our framework. It extends the results of Dupuis et al. (2023) in our PAC-Bayesian setting. Thanks to our framework and the additional Assumption 6, our bound has a more explicit dependence on n and $\overline{\dim}_B^{\vartheta_S}(\mathcal{W})$, compared to (Dupuis et al., 2023). The only non-explicit dependence is encapsulated in the information-theoretic term (the Radon-Nikodym derivative), as it is often the case for PAC-Bayesian bounds. As we will show in Section 5.4, if the prior distribution π is well chosen, the IT term $\log(d\rho_S/d\pi)$, that appears in Theorem 16, is smaller than the total mutual information term appearing in (Dupuis et al., 2023, Theorem 3.8). Moreover, Assumption 6 is much simpler than the so-called "geometric stability"

assumption used in (Dupuis et al., 2023, Definition 3.6). More importantly, Theorem 16 has a better rate of convergence in n than (Dupuis et al., 2023, Theorem 3.8), while featuring the same fractal dimension and IT term. More precisely, our bound vanishes as $n^{-1/2}$, while it is of order $n^{-2\alpha/3}$ in (Dupuis et al., 2023, Theorem 3.8), with $\alpha < 3/2$ a parameter appearing in an intricate geometric stability assumption (Dupuis et al., 2023, Definition 3.6).

5.3 Bounds with Euclidean-based fractal dimensions

In this subsection, we adapt the results of the previous subsection to the Euclidean-based fractal dimension, according to the terminology of Section 5.1. As usual in this literature (Şimşekli et al., 2020; Hodgkinson et al., 2022; Birdal et al., 2021; Camuto et al., 2021), we assume, in this subsection, that the loss $\ell(w, z)$ is L-Lipschitz continuous in w.
While this is a strong additional assumption compared to the setting of Theorem 16, it comes with the benefit of allowing for assumptions weaker than Assumption 6. Note, however, that it would still be possible to adapt Assumption 6 to the present setting.
Such weaker assumptions are made possible by the fact that the covering numbers $|N_\delta(\mathcal{W})|$, based on the Euclidean distance, have a weaker dependency on the number n of data points, compared to their "data-dependent" counterparts $|N_\delta^{\vartheta_S}(\mathcal{W})|$. More precisely, $|N_\delta(\mathcal{W})|$ depends on n only through the posterior distribution ρ_S. Thus, inspired by the law of large numbers, the issue created by the dependence in n can be addressed by assuming convergence of the posterior distribution to a data-independent distribution, in some sense. This is made precise by Assumption 7, which assumes convergence in total variation of ρ_S to a data-independent distribution when $n \to \infty$, and is formally described in Appendix B.3.3.
The next theorem is a consequence of Theorem 11 and Corollary 33.

Theorem 17 *Suppose that Assumptions 1, 2, 3 and 7 hold. We further assume that the loss $\ell(w, z)$ is L-Lipschitz continuous in w and that $\mathcal{W}$ is π-almost surely bounded. Then, for any $\epsilon, \gamma > 0$, there exists a constant $C > 0$ and $n_{\gamma,\epsilon} \in \mathbb{N}^\star$ such that, for any $\lambda > 0$ and $n \geq n_{\gamma,\epsilon}$, with probability at least $1 - \zeta - 3\gamma$ over $S \sim \mu_z^{\otimes n}$ and $\mathcal{W} \sim \rho_S$, we have:*

$$G_S(\mathcal{W}) \leq \frac{2L}{n} + 4B\sqrt{\frac{\left(\overline{\dim}_B(\mathcal{W}) + \epsilon\right)\log(n)}{2n}} + \frac{\log \frac{d\rho_S}{d\pi}(\mathcal{W}) + \log(1/\zeta)}{\lambda} + C\lambda\frac{B^2}{n}.$$

This theorem, whose proof is presented in Appendix B.3.3, is the extension of (Şimşekli et al., 2020, Theorem 2) in our setting. Compared to (Şimşekli et al., 2020, Theorem 2), our result does not require any complicated geometric information on the hypothesis set $\mathcal{W}$ and has a simpler IT term.
Moreover, it should be noted that Theorems 16 and 17 have been derived by the same proof technique, as consequences of our generalization bound with data-dependent Rademacher complexity, *i.e.*, Theorem 11. This is an important improvement over the prior art in fractal-based generalization, where all the existing results (Şimşekli et al., 2020; Dupuis et al., 2023) rely on notably different arguments.
Unlike the results of Section 5.2, the above theorem relies on a Lipschitz continuity assumption on the loss, which may be unrealistic for modern deep neural networks. Therefore, Theorem 16 may look tighter than Theorem 17. However, it was obtained under stronger assumptions, *e.g.*, uniformity in n of the limit of Equation (23). Hence, the fact that the

bound may hold under weaker assumptions is the main advantage of using the Euclidean-based fractal dimension in Theorem 17.

5.4 Comparison of the information-theoretic terms and prior optimization

This subsection is devoted to the comparison between the IT terms appearing in our work and those appearing in other studies considering fractal dimensions (Şimşekli et al., 2020; Hodgkinson et al., 2022; Dupuis et al., 2023).
In most works proving data-dependent uniform generalization bounds, especially in the fractal-based generalization literature, the data-dependent hypothesis set is represented by a random set $\mathcal{W}_S$, depending on the data S and also on some external randomness, induced by the learning algorithm, here omitted. This is the setting in which the bounds represented by Equation (4) were proven. As already discussed, these bounds typically introduce some kind of mutual information (MI) between $\mathcal{W}_S$ and S, to deal with the data-dependence of $\mathcal{W}_S$ (Şimşekli et al., 2020; Hodgkinson et al., 2022; Camuto et al., 2021; Dupuis et al., 2023). We focus in particular on the so-called total mutual information term[7] $I_\infty(\mathcal{W}_S, S)$, for which we give the following definition.

Definition 18 *Let X and Y be two random variables and denote their distributions by $\mathbb{P}_X$ and $\mathbb{P}_Y$ respectively and their joint distribution by $\mathbb{P}_{X,Y}$, then we define the total mutual information by:*

$$I_\infty(X,Y) := \log\left(\sup_{A\in\mathcal{T}} \frac{\mathbb{P}_{(X,Y)}(A)}{\mathbb{P}_X \otimes \mathbb{P}_Y(A)}\right).$$

This quantity has already been used in learning theory, see in particular (Dwork et al., 2015; Hodgkinson et al., 2022).
In this subsection, we translate the settings mentioned above in our framework by letting the posterior ρ_S be the conditional distribution of $\mathcal{W}_S$, given S.
The aforementioned works did not consider a PAC-Bayesian setting. Therefore, to provide a meaningful comparison with our framework, we may choose a prior that is, in a sense, optimal, and compute the corresponding IT terms. Classically, following Catoni (2007) and Alquier (2024) we optimize the prior with respect to the family of posterior distributions, by using the following "optimized prior", which corresponds to the marginal distribution of $\mathcal{W}_S \sim \rho_S$ under $S \sim \mu_z^{\otimes n}$.

$$\forall A \in \mathfrak{E},\ \pi(A) := \mathbb{E}_{S\sim\mu_z^{\otimes n}}[\rho_S(A)]. \tag{26}$$

Under mild assumptions, we have $\rho_S \ll \pi$, $\mu_z^{\otimes n}$-almost surely. For instance, it is the case if $\mu_z^{\otimes n}$ is a strictly positive Borel probability measure and the maps $S \mapsto \rho_S(A)$ are continuous. In the rest of this section, we will assume that this absolute continuity condition is fulfilled. Moreover, it is known (Alquier, 2024, Section 6.5.2) that $\mathbb{E}_S\left[\mathbf{KL}(\rho_S\|\pi)\right] = I_1(\mathcal{W}_S, S)$, where I_1 is the mutual information defined in Appendix A.
The following lemma, proven in Appendix B.3.4, shows that the generalization bounds implied by this optimized prior yield tighter IT terms than existing bounds.

7. Different studies use different total mutual information terms, but as noted by Dupuis et al. (2023) and Hodgkinson et al. (2022), $I_\infty(\mathcal{W}_S, S)$ is the simplest and most intuitive one that has been introduced, it is therefore pertinent for our comparison.

Lemma 19 *With the same notations, we have, for $\mu_z^{\otimes n}$-almost all S and ρ_S-almost all $\mathcal{W}$:*

$$\log \frac{d\rho_S}{d\pi}(\mathcal{W}) \leq I_\infty(\mathcal{W}_S, S).$$

This shows that our framework provides tighter generalization guarantees. In addition, it simplifies the previous fractal bounds and derives them from one proof technique.

6. Application to Langevin Dynamics

In this section, we present the application of our PAC-Bayesian framework to the derivation of uniform generalization bounds for continuous Langevin dynamics (CLD). More precisely, let us consider a measurable space $(\Omega, \mathcal{T})$ and the following stochastic differential equation (SDE), which we call the *empirical* dynamics (restatement of Equation (5)):

$$dW_t = -\nabla \widehat{\mathcal{R}}_S(W_t)dt + \sigma dB_t, \quad W_0 = w_0, \text{ with } \sigma := \sqrt{2\beta^{-1}}. \tag{27}$$

where $(B_t)_{t\geq 0}$ is a standard Brownian motion in $\mathbb{R}^d$ and $w_0 \in \mathbb{R}^d$ is the initialization[8]. The main feature of our method is to analytically express the KL divergence appearing in Theorem 6. This is done by exploiting Girsanov's theorem (Øksendal, 2003, Section 8.6), as well as semi-martingale properties of Equation (5), similar to (Dalalyan, 2017; Raginsky et al., 2017). In Section 6.1, we will specify our setting and express the IT terms in a general form. Then, in Section 6.2, we will derive the corresponding uniform bounds for CLD.

6.1 The setting

For Equation (27) to have a unique continuous and square-integrable strong solution, we make the following classical assumption.

Assumption 4 *The loss ℓ is differentiable and M-smooth in w, which means:*

$$\forall w, w' \in \mathbb{R}^d, \ \forall z \in \mathcal{Z}, \ \left\|\nabla \ell(w, z) - \nabla \ell(w', z)\right\| \leq M \left\|w - w'\right\|.$$

Let us consider a fixed time horizon $T > 0$. We introduce the *random trajectory*, *i.e.*, the set of points encountered by the process, defined by

$$\mathcal{W}(\omega) := \{W_t(\omega), \ 0 \leq t \leq T\}. \tag{28}$$

In Section 3.1, we defined both the prior and posteriors directly on the set E, containing *subsets* of $\mathbb{R}^d$. While this is effective in many applications, in the case of SDE trajectories, it is beneficial to adapt our formulation to take into account the underlying probability space Ω. More precisely, we define the prior π and the posteriors ρ_S directly on the underlying space Ω, satisfying the same Markov kernel properties, as previously defined. We additionally require that all these distributions induce a complete probability space structure on Ω and that the measures are equivalent, *i.e.*, $\rho_S \ll \pi$ and $\pi \ll \rho_S$. W is seen as a stochastic

8. Note that we could also initialize the dynamics randomly and independently from the other random variables, without changing our results.

process defined on Ω. This provides us with a rigorous measure-theoretic setup, where all relevant quantities (*e.g.*, $G_S(\mathcal{W})$, $\mathbf{Rad}_S(\mathcal{W})$) are measurable.
Thanks to these notations, we can directly restate the PAC-Bayesian bounds of Theorem 6; for instance, we can write:

$$\mathbb{P}_S\left(\mathbb{E}_{\omega\sim\rho_S}\left[\Phi(\mathcal{W}(\omega),S)\right] \leq \mathbf{KL}(\rho_S\|\pi) + \log(1/\zeta) + \log \mathbb{E}_S\mathbb{E}_{\omega\sim\pi}\left[e^{\Phi(\mathcal{W}(\omega),S)}\right]\right) \geq 1-\zeta.$$

To ease the notation, we will omit ω and denote $\mathcal{W}\sim\rho_S$ for $\mathcal{W}\sim\mathcal{W}_{\#}\rho_S$.
Note that considering, as prior and posterior, the pushforward measures, $\mathcal{W}_{\#}\pi$, and $\mathcal{W}_{\#}\rho_s$, respectively, would take us back to the exact setting of Section 3.1. The data processing inequality, *i.e.*, $\mathbf{KL}(\mathcal{W}_{\#}\rho_s\|\mathcal{W}_{\#}\pi) \leq \mathbf{KL}(\rho_S\|\pi)$, ensures that both setups are linked.
For technical reasons, we make an additional Lipschitz continuity assumption on ℓ, which is similar to prior work (Aristoff, 2012; Mou et al., 2018; Li et al., 2020; Farghly and Rebeschini, 2021). This ensures that a technical condition, Novikov's condition, holds, see Appendix B.4 for more details.

Assumption 5 *The loss ℓ is L-Lipschitz continuous in w, uniformly with respect to z.*

We will proceed in two steps: first, in Section 6.2, we will provide two expressions of the KL divergence term, depending on the choice of the prior distribution. This highlights the fact that the IT terms appearing in our main theorems can be expressed in particular cases. In the second step, we conclude the derivation of uniform generalization bounds by deriving a bound on the Rademacher complexity of Langevin dynamics, in Section 6.3. The main result of this section then follows from Theorem 11.

6.2 Expression of the KL divergence

To get an expression of the KL divergence term, appearing in our main theorems, we must make a suitable choice of prior distribution π. To leverage classical tools from stochastic calculus, namely Girsanov's theorem (see Appendix B.4), we define π as the path measure of the following *data-independent* SDE:

$$dW_t = -\nabla F(W_t)dt + \sigma dB_t, \quad W_0 = w_0.$$

We consider two types of prior, given the choice of function F:

1. The **Brownian prior** corresponds to $F = 0$.

2. The **expected dynamics** prior corresponds to $F = \mathcal{R}$ (*i.e.*, the population risk). It might be used to tighten the bounds under certain conditions.

We provide, in Theorem 34, an expression of the KL divergence $\mathbf{KL}(\rho_S\|\pi)$ that is induced by a general function F. We now detail the results that we obtained for both choices of prior distributions, which are a direct consequence of Theorem 34.

6.2.1 Brownian prior

In this subsection, we set $F = 0$, so that, under the prior distribution π, we have $W_t = w_0 + \sigma B_t$, *i.e.* W is a (scaled and translated) Brownian motion. As a consequence of Theorem 34, we have the following expression of the KL divergence:

$$\mathbf{KL}(\rho_S\|\pi) = \frac{1}{2\sigma^2}\int_0^T \mathbb{E}_{\rho_S}\left[\|\nabla\widehat{\mathcal{R}}_S(W_t)\|^2\right]dt, \tag{29}$$

from which we deduce the following corollary.

Corollary 20 *Under Assumptions 1, 4 and 5, with probability at least $1-\zeta$ over $S \sim \mu_z^{\otimes n}$, we have, for all $\lambda > 0$:*

$$\mathbb{E}_{\rho_S}\left[G_S(\mathcal{W})\right] \leq 2\mathbb{E}_{\rho_S}\left[\mathbf{Rad}_S(\mathcal{W})\right] + \frac{1}{2\lambda\sigma^2}\int_0^T \mathbb{E}_{\rho_S}\left[\|\nabla\widehat{\mathcal{R}}_S(W_t)\|^2\right]dt + \frac{\log(1/\zeta)}{\lambda} + \lambda\frac{9B^2}{8n}.$$

Note that our bound does not require any ℓ^2-regularization to hold, as in (Mou et al., 2018; Li et al., 2020; Farghly and Rebeschini, 2021). The Rademacher complexity term will be bounded in Section 6.3. We will further discuss the implications of this result after having presented a bound on the Rademacher complexity term, in Section 6.3.

6.2.2 Expected dynamics prior

We now turn to the case of the expected dynamics prior where, under π, W follows the following SDE:

$$dW_t = -\nabla\mathcal{R}(W_t)dt + \sigma dB_t, \quad W_0 = w_0. \tag{30}$$

The consideration of such expected dynamics to prove generalization bounds has already been studied in other works (Amir et al., 2022; Dupuis and Viallard, 2023), although in different settings and leveraging distinct proof techniques. According to Theorem 34, the KL divergence can now be expressed as:

$$\mathbf{KL}(\rho_S\|\pi) = \frac{1}{2\sigma^2}\int_0^T \mathbb{E}_{\rho_S}\left[\|\nabla\widehat{\mathcal{R}}_S(W_t) - \nabla\mathcal{R}(W_t)\|^2\right]dt. \tag{31}$$

Interestingly, the term appearing under the integral in Equation (31) has the form of a generalization term; it can be expected that this term decreases to 0 as $n \to \infty$, hence allowing to gain an order of convergence in the bound. The following proposition is a bound on this KL term, proven by exploiting this idea.

Proposition 21 *Suppose that Assumption 5 holds. With probability at least $1-\zeta$ over $S \sim \mu_z^{\otimes n}$:*

$$\mathbf{KL}(\rho_S\|\pi) \leq \log(1/\zeta) + \frac{L^2\beta T}{n} + \frac{2\beta^2T^2L^4}{n}.$$

By optimizing the value of λ in the corresponding bound, our result becomes:

$$\mathbb{E}_{\rho_S}\left[G_S(\mathcal{W})\right] = \mathcal{O}\left(\mathbb{E}_{\rho_S}\left[\mathbf{Rad}_S(\mathcal{W})\right] + \sqrt{\frac{\log(1/\zeta)}{n}} + \frac{\beta TL^2}{n}\right).$$

6.3 Rademacher complexity of Langevin dynamics

In this section, we prove a bound on the expected Rademacher complexity of Langevin dynamics. Combined with the results of Sections 6.2.1 and 6.2.2, this provides fully computable uniform generalization bounds for Langevin dynamics. To perform a covering-like argument, we restrict ourselves to the case of Lipschitz losses. Hence, all the quantities appearing in Theorem 11 will be then analytically bounded. Note that, without any Lipschitz assumption, it would also be possible to leverage the results of Section 5.2 and introduce a data-dependent fractal dimension in the generalization bound. For the sake of simplicity, and because it may be of independent interest, we focus here on the Lipschitz case.

Theorem 22 *Suppose that Assumptions 1, 4 and 5 hold. We consider Equation* (27)*, followed by* $(W_t)_{0\leq t\leq T}$ *under* ρ_S*, and denote* $\mathcal{W} = \{W_t,\ 0 \leq t \leq T\}$*, as before. Then there exists a universal constant* $C > 0$ *such that:*

$$\mathbb{E}_{\rho_S}\left[\mathbf{Rad}_S(\mathcal{W})\right] \leq \frac{1}{\sqrt{n}} + \max\left\{1, B\right\}\sqrt{\frac{2\log\left(2TnL^2(1+C^2d^2\sigma^2)\right)}{n}}.$$

Thus, we see that the terms dominating in our uniform generalization bounds for CLD are the IT terms rather than the Rademacher complexity term. The overall rate of the bounds can be summarized by that of these IT terms.
Therefore, by optimizing the parameter λ in Corollary 20, its results can be informally summarized as:

$$\mathbb{E}_{\rho_S}\left[G_S(\mathcal{W})\right] = \mathcal{O}\left(B\left\{\frac{1}{2n\sigma^2}\int_0^T \mathbb{E}_{\rho_S}\left[\|\nabla\widehat{\mathcal{R}}_S(W_t)\|^2\right]dt\right\}^{\frac{1}{2}}\right),$$

which is similar to the results presented in (Mou et al., 2018), except that our bound does not feature exponential time decay. However, this is expected, as Mou et al. (2018) only bound the generalization gap at time T, while we consider the worst case gap over the time interval $[0, T]$. Moreover, note that our bound does not require any convexity, dissipativity, or regularization to hold, while such assumptions are in general necessary to obtain time-uniform bounds, as highlighted by Table 3. An important aspect of our proof technique is that Equations (29) and (31) provide exact expressions for the KL divergence, depending on the choice of prior π. Therefore, the time dependence coming from the integral term can only be improved by significantly adapting our framework and relying on stronger assumptions. We leave these important questions for future work.
Moreover, using the additional Lipschitz assumption, as it is the case in (Mou et al., 2018; Li et al., 2020; Haghifam et al., 2020; Farghly and Rebeschini, 2021), we can subsequently optimize the value of the parameter λ and get that, with probability at least $1-\zeta$:

$$\mathbb{E}_{\rho_S}\left[G_S(\mathcal{W})\right] = \mathcal{O}\left(B\sqrt{\frac{TL^2\beta + \log(1/\zeta)}{n}}\right).$$

The order of magnitude of these bounds is coherent with existing literature, as one can see in Table 3, in terms of the relative influence of the quantities (β, T, L, n, B).

Finally, it is worth noticing that the application of our methods to stochastic processes differs from the martingale techniques derived by Chugg et al. (2023), in addition to being deduced from a different proof technique. If we denote by $\mathcal{A}$ the randomness of the algorithm, the bounds in (Chugg et al., 2023, Theorem 3.1) would apply on the quantity

$$\sup_{0\leq t\leq T} \mathbb{E}_{\mathcal{A}}\left[\mathcal{R}(W_t) - \widehat{\mathcal{R}}_S(W_t)\right],$$

It can be understood that the above quantity may be much smaller than the left-hand side of Corollary 20, by noticing that, in this case, the integration over the posterior ρ_S is equivalent to an expectation over the randomness of the algorithm.

7. Uniform Generalization Bounds for SGLD

In this section, we consider the case of SGLD, as described by the following recursion, which is a restatement of Equation (6):

$$\forall k \in \mathbb{N},\ W_{k+1} = W_k - \eta_{k+1}\hat{g}_{k+1} + \sigma_{k+1}\epsilon_{k+1}, \tag{32}$$

where $\sigma_{k+1} := \sqrt{2\eta_{k+1}\beta^{-1}}$, η_{k+1} is the learning rate at iteration $k+1$, $\hat{g}_{k+1}$ is an unbiased estimate of $\nabla\widehat{\mathcal{R}}_S(W_k)$ and $(\epsilon_k)_{k\geq 1}$ are *i.i.d.* $\mathcal{N}(0, I_d)$ random variables. The results of this section follow from arguments that are similar to Section 6, which can be extended to the discrete case through classical arguments (see Appendix B.5). More precisely, we derive, in Appendix B.5, an expression of the KL divergence term appearing in Theorem 10, in the case of SGLD. This leads to the next theorem, which is a uniform generalization bound for SGLD, following from our Rademacher MGF bound, *i.e.*, Theorem 10.

Theorem 23 *Suppose that Assumptions 1 and 8 hold. Then, with probability at least* $1-\zeta$ *over* $S \sim \mu_z^{\otimes n}$*, for all* $\lambda > 0$

$$\mathbb{E}_{\rho_S}\left[\max_{w\in\mathcal{W}}\left(\mathcal{R}(w) - \widehat{\mathcal{R}}_S(w)\right)\right] \leq \frac{1}{\lambda}\left(\log(T/\zeta) + \frac{\beta}{4}\sum_{k=1}^{T}\eta_k\mathbb{E}_{U,\epsilon}\left[\|\hat{g}_k\|^2\right]\right) + \lambda\frac{2B^2}{n},$$

where (U, ϵ) *denotes the randomness of the algorithm, i.e., the randomness coming from the unbiased estimation of* $\nabla\widehat{\mathcal{R}}_S(W_k)$ *and the Gaussian noise, repsectively. The dependence of* $\hat{g}$ *on* S *has been omitted to ease the notations. The expectation over* ρ_S*, on the left-hand side, may be seen as an expectation over* (U, ϵ) *as well.*

Assumption 8, which will be formally introduced in Appendix B.5, is a technical integrability assumption that is necessary for our proofs to hold. It is satisfied, for instance, if the gradients are uniformly bounded, *i.e.*, the loss is Lipschitz continuous.
To compare it with other works, let us analyze the above bound in the case where the gradients are bounded, *i.e.*, $\mathbb{E}_{U,\epsilon}\left[\|\hat{g}_k\|^2\right] \leq L^2$. When $\hat{g}$ is computed as the average gradient over a batch of data points, this corresponds to assuming that the loss ℓ is L-Lipschitz continuous. Based on this assumption, we can optimize the parameter λ in the above theorem to get a bound of the following form:

$$\mathbb{E}_{\rho_S}\left[\max_{w\in\mathcal{W}}\left(\mathcal{R}(w) - \widehat{\mathcal{R}}_S(w)\right)\right] \leq 2B\sqrt{\frac{4\log(T/\zeta) + \beta L^2\sum_{k=1}^{T}\eta_k}{2n}}. \tag{33}$$

While, to our knowledge, no uniform bound for SGLD has been proposed, this still allows for a meaningful comparison with the results of Mou et al. (2018), see Table 4. In the aforementioned work, the authors prove high probability bounds with respect to S, and in expectation over the noise (*cf.* our bound in Theorem 23), but with an additional exponential decay in the sum on the right-hand side. Our result does not feature any exponential decay, this lack of time-uniformity is expected as our bound is uniform over the whole trajectory. In the Lipschitz case, the order of magnitude of our bound is also comparable to the results of (Negrea et al., 2019; Neu et al., 2021). However, note that most works use a subgaussian assumption on the loss ℓ, while our method requires bounded loss, which is stronger. A comparison of our result with existing bounds is given in Table 3. As already mentioned in the introduction, the expectation $\mathbb{E}_{\rho_S}$ over the posterior is taken *outside* of the maximum in Theorem 23. To the best of our knowledge, there would be no trivial way to extend existing generalization bounds for SGLD to obtain Theorem 23.

Conclusion

In this paper, we introduced a PAC-Bayesian framework to prove data-dependent uniform generalization bounds. We provided a rigorous mathematical formulation of our methods and proved two upper bounds in terms of the moment-generating function of the Rademacher complexity and the data-dependent Rademacher complexity. We additionally demonstrated the ability of our methods to prove data-dependent uniform generalization lower bounds.

We successfully applied the introduced techniques in two particular contexts. First, we used our data-dependent Rademacher complexity term to derive uniform bounds in terms of the fractal dimension of the hypothesis set. Compared to prior art, our method yields tighter bounds and uses the same information-theoretic term for all kinds of fractal dimensions. Moreover, our approach greatly simplifies and unifies the proof techniques of the existing literature. Second, we established that in the context of Langevin dynamics and SGLD, the information-theoretic terms appearing in our PAC-Bayesian bounds can be further upper-bounded by closed-form quantities. This allows us to prove the first uniform generalization bounds over the trajectory of these algorithms.

Future work. Some directions remain to be studied regarding our work. First, the generality of the proposed framework opens the door to several refinements of the methods. For instance, one could apply chaining techniques (Vershynin, 2018), other PAC-Bayesian or information-theoretic bounds, such as conditional mutual information bounds (Steinke and Zakynthinou, 2020), or try to extend the "Rademacher viewpoint" of Kakade et al. (2008) and Yang et al. (2019) into our framework. As we mentioned above, our methods could be combined with concentration inequalities in the Bernstein form (McDiarmid, 1998) in order to weaken the assumptions. Beyond the use of fractal dimensions, our work may help to further bridge the gap between generalization and topological data analysis (Andreeva et al., 2024). Regarding Langevin dynamics, it would be beneficial to investigate under which assumption the time dependence of our bounds can be improved. Finally, the optimization of our PAC-Bayesian bounds with respect to the random set posterior might lead to non-vacuous bounds, extending the study of Dziugaite and Roy (2017) to random sets.

Acknowledgments

U.Ş. is partially supported by the French government under the management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute). B.D. and U.Ş. are partially supported by the European Research Council Starting Grant DYNASTY – 101039676.

The appendix is organized as follows:

- In Appendix A, we remind the reader of some notation and provide a technical background.
- Appendix B is dedicated to the omitted proofs from the main part of the paper.

Appendix A. Additional Technical Background

In this section, we remind the reader of some probabilistic technical background as well as a few technical lemmas.

A.1 Probability theory background

The goal of this subsection is to introduce notation and definitions. Let us fix some measurable space $(\Omega, \mathcal{T})$. Given two probability measures μ and ν on $(\Omega, \mathcal{T})$, the absolute continuity of μ with respect to ν will be denoted $\mu \ll \nu$. If $\mu \ll \nu$, the Kullback-Leibler (KL) divergence between μ and ν is defined by:

$$\mathbf{KL}(\mu\|\nu) := \int \log\left(\frac{d\mu}{d\nu}\right) d\mu, \tag{34}$$

where $\frac{d\mu}{d\nu}$ denotes the Radon-Nikodym derivative of μ with respect to ν. If μ is not absolutely continuous with respect to ν, we set $\mathbf{KL}(\mu\|\nu) = +\infty$, by convention.
A probability space $(\Omega, \mathcal{T}, \mathbb{P})$ is said to be *complete* if, for all $A \in \mathcal{T}$ such that $\mathbb{P}(A) = 0$, we have $\forall B \subseteq A,\ B \in \mathcal{T}$.
Given a random variable X and a probability measure $\mathbb{P}$ on $(\Omega, \mathcal{T})$, we denote by $\mathbb{P}_X$, or $X_{\#}\mathbb{P}$, the law of X, *i.e.*, the image measure of $\mathbb{P}$ under X. Given two random variables X and Y, the mutual information between X and Y is defined by:

$$\mathrm{I}_1(X, Y) := \mathbf{KL}(\mathbb{P}_{(X,Y)}\|\mathbb{P}_X \otimes \mathbb{P}_Y).$$

This is the most common notion of mutual information, which appears, for instance, in the generalization bounds of Xu and Raginsky (2017). The total mutual information has been defined in Definition 18. It satisfies $I_1 \leq I_\infty$.

A.2 A few technical lemmas

The next lemma is just a way of writing McDiarmid's inequality (McDiarmid, 1998) in an exponential form. A proof can be found in (Boucheron et al., 2013, Theorem 6.2) for instance.

Lemma 24 *Consider any function $f : \mathcal{X}^n \to \mathbb{R}$, where $\mathcal{X}$ is any measurable space. We assume that f satisfies the bounded difference inequality, i.e., for all $i \in \{1, \dots, n\}$ and all $(x_1, \dots, x_n) \in \mathcal{X}^n$, one has:*

$$\sup_{x' \in \mathcal{X}} |f(x_1, \dots, x_n) - f(x_1, \dots, x_{i-1}, x', x_{i+1}, \dots, x_n)| \leq c_i,$$

Then, given $(X_1, \dots, X_n)$ some i.i.d. random variables on $\mathcal{X}$, the random variable $Z := f(X_1, \dots, X_n)$ satisfies:

$$\mathbb{E}\left[e^{\lambda(f(Z) - \mathbb{E}f(Z))}\right] \leq e^{\frac{\lambda^2}{8}\sum_{i=1}^n c_i^2}$$

We recall below the symmetrization lemma, which is one of the key ingredients of Rademacher complexity-based bounds, presented in Section 2.2. A proof can be found, for instance, in (Shalev-Schwartz and Ben-David, 2014).

Lemma 25 (Symmetrization) *Let $\mathcal{W}$ be a data-independent (e.g. fixed) set. We have:*

$$\mathbb{E}_S\left[G_S(\mathcal{W})\right] \leq 2\mathbf{Rad}(\mathcal{W}).$$

The symmetrization technique can also be used to obtain a lower bound, often called desymmetrization inequality. A proof can be found in (Dupuis et al., 2023), the only difference is that we write here the inequality with a slightly better absolute constant than in (Dupuis et al., 2023), which can be obtained by using Khintchine's inequality instead of Massart's lemma in the last step of the proof.

Proposition 26 (Desymmetrization inequality) *Assume that $\mathcal{W}$ is a fixed set and that the loss ℓ satisfies Assumption 1. Then we have:*

$$\mathbb{E}_S\left[\sup_{w\in\mathcal{W}} |\widehat{\mathcal{R}}_S(w) - \mathcal{R}(w)|\right] \geq \frac{1}{2}\mathbf{Rad}(\mathcal{W}) - \frac{B}{2\sqrt{n}}.$$

The next theorem is Egoroff's theorem; it is a classical result from measure theory (Bogachev, 2007). It has been used in the context of fractal-based generalization bounds (Şimşekli et al., 2020; Camuto et al., 2021; Hodgkinson et al., 2022; Dupuis et al., 2023).

Theorem 27 (Egoroff's theorem) *Let $(\Omega, \mathcal{T}, \mu)$ be a finite measure space. Let $f, (f_n)_n : \Omega \longrightarrow X$ be measurable functions, with X some arbitrary separable metric space. Assume that μ-almost everywhere, we have $f_n(x) \longrightarrow f(x)$ as $n \to \infty$. Then, for all $\gamma > 0$, there exists $\Omega_\gamma \in \mathcal{T}$ such that $\mu(\Omega \backslash \Omega_\gamma) \leq \gamma$ and the convergence of (f_n) to f is uniform on Ω_γ.*

Appendix B. Omitted Proofs and Additional Results

In this section, we present the omitted proofs of all our main results.

B.1 Omitted proofs of Section 4

B.1.1 Omitted proofs of Section 4.1 - Rademacher MGF

Before proving Theorem 10, we first prove the following exponential symmetrization lemma, which is an exponential equivalent of the usual symmetrization that is used in the Rademacher complexity literature (see *e.g.*, Shalev-Schwartz and Ben-David, 2014).

Lemma 28 (Exponential symmetrization lemma) *For any set $\mathcal{W}$, for any set $\mathcal{Z}$, for any measurable function $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}$, for any $\lambda > 0$, we have,*

$$\mathbb{E}_S\left[\exp\left\{\lambda \sup_{w\in\mathcal{W}}\Big(\mathcal{R}(w) - \widehat{\mathcal{R}}_S(w)\Big)\right\}\right] \leq \mathbb{E}_S\mathbb{E}_\epsilon \exp\left\{\frac{2\lambda}{n}\sup_{w\in\mathcal{W}}\sum_{i=1}^n \epsilon_i \ell(w, z_i)\right\},$$

where the previous inequality holds as soon as the measurability of the quantities inside the expectations is ensured.

Proof Let $S' := (z'_1, \dots, z'_n) \sim \mu_z^{\otimes n}$ be independent of $S \sim \mu_z^{\otimes n}$, by Jensen's inequality:

$$\begin{aligned}
\mathbb{E}_S\left[\exp\left\{\lambda \sup_{w\in\mathcal{W}}\Big(\mathcal{R}(w) - \widehat{\mathcal{R}}_S(w)\Big)\right\}\right] &= \mathbb{E}_S\left[\exp\left\{\frac{\lambda}{n}\sup_{w\in\mathcal{W}}\sum_{i=1}^n \mathbb{E}_{S'}[\ell(w,z'_i) - \ell(w,z_i)]\right\}\right] \\
&\leq \mathbb{E}_S\left[\exp\left\{\mathbb{E}_{S'}\frac{\lambda}{n}\sup_{w\in\mathcal{W}}\sum_{i=1}^n \ell(w,z'_i) - \ell(w,z_i)\right\}\right] \\
&\leq \mathbb{E}_{S,S'}\exp\left\{\frac{\lambda}{n}\sup_{w\in\mathcal{W}}\sum_{i=1}^n (\ell(w,z'_i) - \ell(w,z_i))\right\}.
\end{aligned}$$

By the usual symmetrization trick, see (Shalev-Schwartz and Ben-David, 2014, Lemma 26.2), we can introduce $(\epsilon_1, \dots, \epsilon_n)$ some Rademacher random variables and write:

$$\begin{aligned}
\mathbb{E}_{S,S'}\exp\left\{\frac{\lambda}{n}\sup_{w\in\mathcal{W}}\sum_{i=1}^n \big(\ell(w,z'_i) - \ell(w,z_i)\big)\right\} &= \mathbb{E}_{S,S',\epsilon}\exp\left\{\frac{\lambda}{n}\sup_{w\in\mathcal{W}}\sum_{i=1}^n \epsilon_i\left[\ell(w,z'_i) - \ell(w,z_i)\right]\right\} \\
&\leq \mathbb{E}_S\mathbb{E}_{S'}\mathbb{E}_\epsilon\left[\exp\left\{\frac{\lambda}{n}\sup_{w\in\mathcal{W}}\sum_{i=1}^n \epsilon_i\ell(w,z'_i) + \frac{\lambda}{n}\sup_{w\in\mathcal{W}}\sum_{i=1}^n (-\epsilon_i)\ell(w,z_i)\right\}\right] \\
&= \mathbb{E}_S\mathbb{E}_{S'}\mathbb{E}_\epsilon\left[\exp\left\{\frac{\lambda}{n}\sup_{w\in\mathcal{W}}\sum_{i=1}^n \epsilon_i\ell(w,z'_i)\right\}\exp\left\{\frac{\lambda}{n}\sup_{w\in\mathcal{W}}\sum_{i=1}^n (-\epsilon_i)\ell(w,z_i)\right\}\right].
\end{aligned}$$

From Cauchy-Schwarz's inequality, we finally obtain

$$\begin{aligned}
&\mathbb{E}_S\mathbb{E}_{S'}\mathbb{E}_\epsilon\left[\exp\left\{\frac{\lambda}{n}\sup_{w\in\mathcal{W}}\sum_{i=1}^n \epsilon_i\ell(w,z'_i)\right\}\exp\left\{\frac{\lambda}{n}\sup_{w\in\mathcal{W}}\sum_{i=1}^n (-\epsilon_i)\ell(w,z_i)\right\}\right] \\
&\leq \left[\mathbb{E}_{S'}\mathbb{E}_\epsilon\exp\left\{\frac{2\lambda}{n}\sup_{w\in\mathcal{W}}\sum_{i=1}^n \epsilon_i\ell(w,z'_i)\right\}\right]^{\frac{1}{2}}\left[\mathbb{E}_S\mathbb{E}_\epsilon\exp\left\{\frac{2\lambda}{n}\sup_{w\in\mathcal{W}}\sum_{i=1}^n (-\epsilon_i)\ell(w,z_i)\right\}\right]^{\frac{1}{2}} \\
&= \left[\mathbb{E}_S\mathbb{E}_\epsilon\exp\left\{\frac{2\lambda}{n}\sup_{w\in\mathcal{W}}\sum_{i=1}^n \epsilon_i\ell(w,z_i)\right\}\right]^{\frac{1}{2}}\left[\mathbb{E}_S\mathbb{E}_\epsilon\exp\left\{\frac{2\lambda}{n}\sup_{w\in\mathcal{W}}\sum_{i=1}^n \epsilon_i\ell(w,z_i)\right\}\right]^{\frac{1}{2}} \\
&= \mathbb{E}_S\mathbb{E}_\epsilon\exp\left\{\frac{2\lambda}{n}\sup_{w\in\mathcal{W}}\sum_{i=1}^n \epsilon_i\ell(w,z_i)\right\}.
\end{aligned}$$

■

We can now prove Theorem 10.

Proof Let us fix some $\lambda > 0$, we apply Theorem 6 with $\Phi_\lambda(\mathcal{W}, S) = \lambda \sup_{w\in\mathcal{W}}\big(\mathcal{R}(w) - \widehat{\mathcal{R}}_S(w)\big)$. Therefore, our task boils down to a bound on the log-exp term, which we achieve

by applying Fubini's theorem and Lemma 28. This gives:

$$\begin{aligned}\mathbb{E}_S\mathbb{E}_{\mathcal{W}\sim\pi}e^{\lambda\sup_{w\in\mathcal{W}}\left(\mathcal{R}(w)-\widehat{\mathcal{R}}_S(w)\right)} &= \mathbb{E}_{\mathcal{W}\sim\pi}\mathbb{E}_S e^{\lambda\sup_{w\in\mathcal{W}}\left(\mathcal{R}(w)-\widehat{\mathcal{R}}_S(w)\right)}\\ &\leq \mathbb{E}_{\mathcal{W}\sim\pi}\mathbb{E}_S\mathbb{E}_\epsilon \exp\left\{\frac{2\lambda}{n}\sup_{w\in\mathcal{W}}\sum_{i=1}^n \epsilon_i\ell(w,z_i)\right\}\\ &= \mathbb{E}_{\mathcal{W}\sim\pi}\mathbb{E}_S\Psi_{S,\mathcal{W}}(2\lambda),\end{aligned}$$

implying the desired results. ■

B.1.2 Omitted proofs of Section 4.2

We end this section by giving the proof of Theorem 11.

Proof Let us define the function, defined for any $\lambda > 0$:

$$\Phi_\lambda(\mathcal{W},S) := \lambda\sup_{w\in\mathcal{W}}\left(\mathcal{R}(w)-\widehat{\mathcal{R}}_S(w)\right) - 2\lambda\mathbf{Rad}_S(\mathcal{W}). \tag{35}$$

Our goal is to apply the results of Theorem 6 to function Φ_λ. Our assumptions ensure that the above terms are well-defined and measurable. Therefore, for both inequalities (KL-based and disintegrated), our task boils down to bounding the following quantity:

$$\mathcal{L} = \log\mathbb{E}_S\mathbb{E}_{\mathcal{W}\sim\pi}\left[\exp\left\{\lambda\sup_{w\in\mathcal{W}}\left(\mathcal{R}(w)-\widehat{\mathcal{R}}_S(w)\right) - 2\lambda\mathbf{Rad}_S(\mathcal{W})\right\}\right].$$

The key of our reasoning is that in the above expectation, the variables $S\sim\mu_z^{\otimes n}$ and $\mathcal{W}\sim\pi$ are now independent. This justifies the following considerations.

Let us denote $(\epsilon_1,\ldots,\epsilon_n)$ some *i.i.d.* Rademacher random variables, independent of S and $\mathcal{W}$. In order to bound $\mathcal{L}$ we remark that we have (with S^i being $S=(z_1,\ldots,z_n)$ with i-th element replaced by another one, denoted z_i'):

$$\begin{aligned}|\Phi_\lambda(\mathcal{W},S)-\Phi_\lambda(\mathcal{W},S^i)| = \lambda\Big|&\sup_{w\in\mathcal{W}}\left(\mathcal{R}(w)-\widehat{\mathcal{R}}_S(w)\right)-2\mathbf{Rad}_S(\mathcal{W})\\ &- \sup_{w\in\mathcal{W}}\left(\mathcal{R}(w)-\widehat{\mathcal{R}}_{S^i}(w)\right)-2\mathbf{Rad}_{S^i}(\mathcal{W})\Big|\\ &\leq \frac{\lambda}{n}\sup_{w\in\mathcal{W}}\left|\ell(w,z_i')-\ell(w,z_i)\right| + 2\lambda\left|\mathbf{Rad}_{S^i}(\mathcal{W})-\mathbf{Rad}_S(\mathcal{W})\right|\\ &\leq \frac{B\lambda}{n}+\frac{2\lambda}{n}\mathbb{E}_\epsilon\left[\sup_{w\in\mathcal{W}}\left|\epsilon_i\big(\ell(w,z_i)-\ell(w,z_i')\big)\right|\right]\\ &\leq \frac{3\lambda B}{n},\end{aligned}$$

where we used the fact that ℓ is bounded in $[0,B]$. From Lemma 24, we deduce that:

$$\mathbb{E}_S\left[e^{\Phi_\lambda(\mathcal{W},S)-\mathbb{E}_S[\Phi_\lambda(\mathcal{W},S)]}\right] \leq \exp\left\{\frac{9\lambda^2B^2}{8n}\right\}.$$

Moreover, by the classical symmetrization inequality, Lemma 25, we know that:

$$\forall \mathcal{W} \in E,\ \mathbb{E}_S\left[\Phi_\lambda(\mathcal{W}, S)\right] \leq 0. \tag{36}$$

Therefore, by Fubini's theorem, we have:

$$\begin{aligned}
\mathcal{L} &= \log \mathbb{E}_S \mathbb{E}_{\mathcal{W}\sim\pi}\left[e^{\Phi_\lambda(\mathcal{W},S)}\right] \\
&= \log \mathbb{E}_S \mathbb{E}_{\mathcal{W}\sim\pi}\left[e^{\Phi_\lambda(\mathcal{W},S)-\mathbb{E}_S[\Phi_\lambda(\mathcal{W},S)]} e^{\mathbb{E}_S[\Phi_\lambda(\mathcal{W},S)]}\right] \\
&\leq \log \mathbb{E}_{\mathcal{W}\sim\pi}\mathbb{E}_S\left[e^{\Phi_\lambda(\mathcal{W},S)-\mathbb{E}_S[\Phi_\lambda(\mathcal{W},S)]}\right] \\
&\leq \frac{9\lambda^2 B^2}{8n}.
\end{aligned}$$

The desired bounds immediately follow. ■

The proof of Theorem 13 follows the same lines, with just a change in the function Φ, the bounded difference condition being obtained through the inverted triangular inequality.

B.2 Data-dependent uniform generalization bounds with IPMs

The goal of this subsection is to further underline the generality of our framework by briefly extending our PAC-Bayesian framework on random sets with the Integral Probability Metrics (IPM) used by Amit et al. (2022) and Viallard et al. (2023, 2024b) to derive general PAC-Bayesian bounds. As for the main results of Sections 3.1 and 4, this extension is straightforward given our measurability assumptions.
With the notations of Section 3.1, we have the following definition of IPMs on E, see (Amit et al., 2022, Definition 3).

Definition 29 *Let $\mathcal{F}$ be a family of functions $E \to \mathbb{R}$ and μ and ν be two probability distributions on $(E, \mathfrak{E})$. The IPM between μ and ν associated with $\mathcal{F}$ is defined as:*

$$\gamma_{\mathcal{F}}(\mu, \nu) := \sup_{\phi\in\mathcal{F}} \left|\mathbb{E}_\nu\left[\phi(\mathcal{W})\right] - \mathbb{E}_\mu\left[\phi(\mathcal{W})\right]\right|.$$

Using Definition 29, we can easily state the following bound based on IPMs over random sets.

Theorem 30 *Let $(E, \mathfrak{E})$ be defined as in Section 3.1, π be a fixed prior distribution on $\mathbb{R}^d$ and $\Phi : E \times \mathcal{Z}^n \longrightarrow \mathbb{R}$ be a measurable function. For any $n \in \mathbb{N}^*$ and $S \in \mathcal{Z}^n$, we consider a family $\mathcal{F}_S$ of bounded measurable functions $E \longrightarrow \mathbb{R}$. We assume that for every n and every $S \in \mathcal{Z}$, we have $\Phi(\cdot, S) \in \mathcal{F}_S$. Then, we have:*

$$\mathbb{P}_S\left(\forall \rho \in \mathcal{P}_0(\mathbb{R}^d),\ \mathbb{E}_\rho \Phi(\mathcal{W}, S) \leq \gamma_{\mathcal{F}_S}(\rho, \pi) + \log(1/\zeta) + \log \mathbb{E}_S \mathbb{E}_\pi\left[e^{\Phi(\mathcal{W},S)}\right]\right) \geq 1 - \zeta,$$

where $\mathcal{P}_0(\mathbb{R}^d)$ denotes the set of probability distributions over $\mathbb{R}^d$.

Proof The proof mimics the one of Proposition 4 in (Amit et al., 2022), with only the difference of using a general function $\Phi : E \times \mathcal{Z}^n \longrightarrow \mathbb{R}$. By definition of IPM and Jensen's inequality, we have, for any $\rho \in \mathcal{P}_0(\mathbb{R}^d)$:

$$e^{\mathbb{E}_{\mathcal{W}\sim\rho}[\Phi(\mathcal{W},S)]-\gamma_{\mathcal{F}_S}(\rho,\pi)} \leq e^{\mathbb{E}_{\mathcal{W}\sim\pi}[\Phi(\mathcal{W},S)]} \leq \mathbb{E}_{\mathcal{W}\sim\pi}\left[e^{\Phi(\mathcal{W},S)}\right].$$

Therefore, by Markov's inequality, with probability at least $1-\zeta$ over $S \sim \mu_z^{\otimes n}$ we have:

$$\sup_{\rho \in \mathcal{P}_0(\mathbb{R}^d)} e^{\mathbb{E}_{\mathcal{W}\sim\rho}[\Phi(\mathcal{W},S)] - \gamma_{\mathcal{F}_S}(\rho,\pi)} \leq \frac{1}{\zeta} \mathbb{E}_S \mathbb{E}_{\mathcal{W}\sim\pi} \left[e^{\Phi(\mathcal{W},S)} \right].$$

Applying the logarithm on both sides gives the result. ■

For instance, let us extend Theorem 16 with IPMs. The corresponding bound is given by the following corollary.

Corollary 31 *Suppose that Assumptions 1 and 2 hold. For any $n \in \mathbb{N}^*$ and $S \in \mathcal{Z}^n$, we consider a family $\mathcal{F}_S$ of bounded measurable function $E \to \mathbb{R}$. Let us consider some $\lambda > 0$ such that for every n and every $S \in \mathcal{Z}$, we have $\Phi_\lambda(\cdot, S) \in \mathcal{F}_S$, where Φ_λ is given by Equation (19). Then, with probability at least $1-\zeta$ over $S \sim \mu_z^{\otimes n}$, we have*

$$\forall \rho \in \mathcal{P}_0(\mathbb{R}^d), \quad \mathbb{E}_{\mathcal{W}\sim\rho} \Phi(\mathcal{W}, S) \leq 2\mathbb{E}_{\mathcal{W}\sim\rho} \left[\mathbf{Rad}_S(\mathcal{W}) \right] + \frac{\gamma_{\mathcal{F}_S}(\rho,\pi) + \log(1/\zeta)}{\lambda} + \lambda \frac{9B^2}{8n}.$$

Proof We use Theorem 30 and upper bound the term $\mathbb{E}_S \mathbb{E}_\pi \left[e^{\Phi_\lambda(\mathcal{W},S)} \right]$ exactly as in the proof of Theorem 16. ■

B.3 Fractal based generalization bounds - Omitted proofs of Section 5

In this section, we present the omitted proofs of Section 5. As in other fractal-based works on generalization bounds (Şimşekli et al., 2020; Dupuis et al., 2023), the bounds involving fractal dimensions are deduced from bounds involving covering numbers. Therefore, we first give generalization bounds with data-dependent covering numbers.

B.3.1 Data-dependent covering numbers

We deduce two covering bounds from the Rademacher complexity bound of Theorem 11. The first one, presented in the next corollary, uses covering numbers defined through the data-dependent pseudometric introduced in Equation (24). It is a direct consequence of Theorem 11 and Massart's Lemma, as it is classically done in the analysis of Rademacher complexity (Shalev-Schwartz and Ben-David, 2014). As these arguments are very classical, we omit the proofs to avoid harming the readability of the paper.

Corollary 32 *Under Assumptions 1, 2 and 3, there exists an absolute constant $C > 0$, such that, for any $\lambda, \delta > 0$, with probability at least $1-\zeta$ under $S \sim \mu_z^{\otimes n}$, we have, with probability at least $1-\zeta$ over $S \sim \mu_z^{\otimes n}$ and $\mathcal{W} \sim \rho_S$:*

$$G_S(\mathcal{W}) \leq 2\delta + 2B\sqrt{\frac{2\log(|N_\delta^{\vartheta_S}(\mathcal{W})|)}{n}} + \frac{\log \frac{d\rho_S}{d\pi}(\mathcal{W}) + \log(1/\zeta)}{\lambda} + C\lambda\frac{B^2}{n}.$$

In some cases, one may be interested in introducing covering numbers with respect to the Euclidean distance on $\mathbb{R}^d$. This is, for instance, the setting considered by Şimşekli et al. (2020) and Hodgkinson et al. (2022). As highlighted by these authors, this requires a Lipschitz continuity on the loss ℓ. This leads to the next corollary.

Corollary 33 *Suppose that Assumptions 1, 2 and 3 hold. We assume that $\ell(w,z)$ is L-Lipschitz in w and that $\mathcal{W}$ is π-almost surely bounded. Then there exists a constant $C>0$, such that, for any $\lambda,\delta>0$, with probability at least $1-\zeta$ under $S\sim\mu_z^{\otimes n}$ and $\mathcal{W}\sim\rho_S$:*

$$G_S(\mathcal{W}) \leq 2L\delta + 2B\sqrt{\frac{2\log(|N_\delta(\mathcal{W})|)}{n}} + \frac{\log\frac{d\rho_S}{d\pi}(\mathcal{W}) + \log(1/\zeta)}{\lambda} + C\lambda\frac{B^2}{n}$$

Thus, we are able to deduce two types of covering bounds from our data-dependent Rademacher complexity bounds. In the next two subsections, we will deduce fractal-based generalization bounds built on these results.

B.3.2 Proof of Theorem 16

Presentation of Assumption 6. As mentioned in Section 5.2, as an additional contribution, our framework is suitable for the creation of natural assumptions to handle the uniformity in n of the limit in Equation (23), defining the data-dependent fractal dimension. For $S\in\mathcal{Z}^n$ and $\mathcal{W}\sim\rho_S$, the covering number $|N^{\vartheta_S}(\mathcal{W})|$ has a dependence in n through its dependence in the dataset S. To overcome this technical difficulty, we observe that by definition of the upper box-counting dimension, we have, for all $S\in\mathcal{Z}^n$:

$$\overline{\dim}_B^{\vartheta_S}(\mathcal{W}) := \lim_{\delta\to 0}\sup_{0<r<\delta}\frac{\log(|N_r^{\vartheta_S}(\mathcal{W})|)}{\log(1/r)}.$$

We know that (almost) sure convergence implies convergence in probability. For this reason, it makes sense to assume the uniformity in n of this convergence in probability, which is formalized by the following assumption.

Assumption 6 *We assume that, for all $\epsilon>0$, one has:*

$$\sup_{n\in\mathbb{N}^\star}\int_{\mathcal{Z}^n}\rho_S\left(\sup_{0<r<\delta}\frac{\log(|N_r^{\vartheta_S}(\mathcal{W})|)}{\log(1/r)} - \overline{\dim}_B^{\vartheta_S}(\mathcal{W}) \geq \epsilon\right)d\mu_z^{\otimes n}(S) \underset{\delta\to 0}{\longrightarrow} 0.$$

We can now present the proof of Theorem 16.

Proof Let us fix $\epsilon,\gamma>0$. From Assumption 6, we know that there exists $\delta_{\gamma,\epsilon}>0$ such that, for all $n\in\mathbb{N}^\star$ and $\delta<\delta_{\gamma,\epsilon}$, with probability at least $1-\gamma$ under $\mathbb{P}_{S\sim\mu_z^{\otimes n},\mathcal{W}\sim\rho_S}$:

$$\log(|N_\delta^{\vartheta_S}(\mathcal{W})|) \leq \log(1/\delta)(\epsilon + \overline{\dim}_B^{\vartheta_S}(\mathcal{W})).$$

Now, we define the sequence $\delta_n := 1/n$, for $n\geq 1$. For $n > \lceil 1/\delta_{\gamma,\epsilon}\rceil$, we therefore have that:

$$\mathbb{P}_{S\sim\mu_z^{\otimes n},\mathcal{W}\sim\rho_S}\left(\log(|N_{1/n}^{\vartheta_S}(\mathcal{W})|) \leq \log(n)(\epsilon + \overline{\dim}_B^{\vartheta_S}(\mathcal{W}))\right) \geq 1-\gamma.$$

The result follows from a union bound and Corollary 32. ■

B.3.3 Proof of Theorem 17

Presentation of Assumption 7. Let us introduce the probability space $\mathcal{Z}^\infty$, endowed with the cylindrical σ-algebra (denoted $\mathcal{F}^{\otimes\infty}$) and the product measure $\mu_z^{\otimes\infty}$. For any $S = (z_i)_{i\geq 1} \in \mathcal{Z}^\infty$, we denote the canonical projection $\mathcal{Z}^\infty \longrightarrow \mathcal{Z}^n$ by $S_n := (z_1, \dots, z_n)$. The following assumption consists of the convergence of the posterior distributions to a limit distribution when $n \to \infty$, in the sense of the total variation distance.

Assumption 7 *There exists a probability measure $\mathbb{Q}$ on the space of hypothesis sets $(E, \mathfrak{E})$, such that[9], for $\mu_z^{\otimes\infty}$-almost all $S \in \mathcal{Z}^\infty$:*

$$\mathrm{TV}(\rho_{S_n}, \mathbb{Q}) := 2 \sup_{A\in\mathfrak{E}} \left|\rho_{S_n}(A) - \mathbb{Q}(A)\right| \underset{n\to\infty}{\longrightarrow} 0.$$

Note that we do not impose the distribution $\mathbb{Q}$ to be equal to the prior distribution π, but it may be the case in particular applications.

By Pinsker's inequality, this is weaker than assuming a convergence in the KL divergence. Based on the above assumption, we can present the proof of Theorem 17.

Proof Let us consider a decreasing sequence (δ_k) such that $\forall k,\ \delta_k > 0$ and $\delta_k \to 0$. For any bounded set $\mathcal{W} \subseteq \mathbb{R}^d$ and $\delta > 0$, we introduce the notation:

$$f_\delta(\mathcal{W}) := \sup_{0<r<\delta} \frac{\log(|N_r(\mathcal{W})|)}{\log(1/r)} - \overline{\dim}_B(\mathcal{W}). \tag{37}$$

Note that f_δ is measurable, because the supremum may be taken over rational numbers in the interval $(0, \delta)$. Let us fix $\epsilon, \gamma > 0$. For $\mu_z^{\otimes\infty}$-almost all $S \in \mathcal{Z}^\infty$, we have, because of the total variation convergence assumption:

$$\sup_{k\in\mathbb{N}} \left|\rho_{S_n}(f_{\delta_k}(\mathcal{W}) \geq \epsilon) - \mathbb{Q}(f_{\delta_k}(\mathcal{W}) \geq \epsilon)\right| \underset{n\to\infty}{\longrightarrow} 0.$$

Thanks to the Markov kernel assumption on $S \longmapsto \rho_S(\cdot)$ and by construction of the cylindrical σ-algebra $\mathcal{F}^{\otimes\infty}$, it can be seen that the mappings:

$$\mathcal{Z}^\infty \ni S \longmapsto h_n(S) := \sup_{k\in\mathbb{N}} \left|\rho_{S_n}(f_{\delta_k}(\mathcal{W}) \geq \epsilon) - \mathbb{Q}(f_{\delta_k}(\mathcal{W}) \geq \epsilon)\right|,$$

are $\mathcal{F}^{\otimes\infty}$-measurable for any $n \in \mathbb{N}^\star$, as a countable supremum of measurable functions. Therefore, we can apply Egoroff's theorem (Theorem 27) to this sequence of function[10] to find a set $\Omega_\gamma \in \mathcal{F}^{\otimes\infty}$, such that $\mu_z^{\otimes\infty}(\Omega_\gamma) \geq 1 - \gamma$, and on which the above convergence is uniform, with respect to $S \in \Omega_\gamma$. Therefore, we can find $n^1_{\gamma,\epsilon}$, such that, for every $n \geq n^1_{\gamma,\epsilon}$:

$$\forall S \in \Omega_\gamma,\ \forall k \in \mathbb{N}^\star,\ \rho_{S_n}(f_{\delta_k}(\mathcal{W}) \geq \epsilon) \leq \gamma + \mathbb{Q}(f_{\delta_k}(\mathcal{W}) \geq \epsilon).$$

By definition of the upper Minkowski dimension, we know that $f_{\delta_k}(\mathcal{W}) \to 0$, pointwise, when $k \to \infty$. Therefore, we also have the convergence in probability $\mathbb{Q}(f_{\delta_k}(\mathcal{W}) \geq \epsilon) \underset{k\to 0}{\longrightarrow} 0$. Applying this to the sequence $\delta_n = 1/n$, we find that there exists $n^2_{\gamma,\epsilon} \in \mathbb{N}^\star$, such that:

$$\forall n \geq n^2_{\gamma,\epsilon},\ \mathbb{Q}(f_{\delta_n}(\mathcal{W}) \geq \epsilon) \leq \gamma.$$

9. We define the total variation between two measures μ and ν as $2\sup_A |\mu(A) - \nu(A)|$, some authors remove the 2 from this definition. This wouldn't affect any of the results.
10. Note that Egoroff's theorem only requires almost everywhere convergence, which is the case here.

Setting $n_{\gamma,\epsilon} = \max(n^1_{\gamma,\epsilon}, n^2_{\gamma,\epsilon})$, we have that, for $n \geq n_{\gamma,\epsilon}$:

$$\begin{aligned}\int_{\mathcal{Z}^n} \rho_S(f_{\delta_n}(\mathcal{W}) \geq \epsilon) d\mu_z^{\otimes n}(S) &= \int_{\mathcal{Z}^\infty} \rho_{S_n}(f_{\delta_n}(\mathcal{W}) \geq \epsilon) d\mu_z^{\otimes \infty}(S) \\ &\leq \gamma + \int_{\Omega_\gamma} \rho_{S_n}(f_{\delta_n}(\mathcal{W}) \geq \epsilon) d\mu_z^{\otimes \infty}(S) \\ &\leq \gamma + 2\gamma \mu_z^{\otimes \infty}(\Omega_\gamma) \\ &\leq 3\gamma.\end{aligned}$$

Hence, there exists $n_{\gamma,\epsilon}$ (potentially slightly different), such that, with probability at least $1 - 3\gamma$ under the joint law of $S \sim \mu_z^{\otimes n}$ and $\mathcal{W} \sim \rho_S$, we have, for $n \geq n_{\gamma,\epsilon}$:

$$\log(|N_{\delta_n}(\mathcal{W})|) \leq \log(n)\big(\epsilon + \overline{\dim}_B(\mathcal{W})\big).$$

The result then immediately follows from Corollary 33 and a union bound. ■

B.3.4 Proof of Lemma 19

Proof Let us denote $\mathbb{P}_{S,\rho_S}$ the joint distribution of S and ρ_S. Note that there is a slight abuse of notation here, as S is a random variable and ρ_S a distribution, we use it to ease further notations. We have, for any $A \in \mathfrak{E} \otimes \mathcal{F}^{\otimes n}$:

$$\begin{aligned}\mathbb{E}_{(S,\mathcal{W})\sim\mathbb{P}_{S,\rho_S}}\left[\mathbb{1}_A(\mathcal{W}, S)\right] &= \int_{\mathcal{Z}^n} \int \mathbb{1}_A(\mathcal{W}, S) d\rho_S(\mathcal{W}) d\mu_z^{\otimes n}(S) \\ &= \int_{\mathcal{Z}^n} \int \mathbb{1}_A(\mathcal{W}, S) \frac{d\rho_S}{d\pi}(\mathcal{W}) d\pi(\mathcal{W}) d\mu_z^{\otimes n}(S).\end{aligned}$$

Therefore $\frac{d\mathbb{P}_{S,\rho_S}}{d(\mu_z^{\otimes n}\otimes\pi)}(\mathcal{W}, S) = \frac{d\rho_S}{d\pi}(\mathcal{W})$. For the purpose of this proof, let us introduce the "infinite" Rényi-divergence. For some measurable space $(\Omega, \mathcal{T})$ and two probability measures μ and ν such that $\mu \ll \nu$, we define:

$$D_\infty(\mu||\nu) := \log\left(\sup_{A\in\mathcal{T}} \frac{\mu(A)}{\nu(A)}\right).$$

From van Erven and Harremoës (2014, Theorem 6), we get:

$$\begin{aligned}I_\infty(\mathcal{W}_S, S) &= D_\infty\left(\mathbb{P}_{S,\rho_S} \| \mu_z^{\otimes n} \otimes \pi\right) \\ &= \log\left(\operatorname{esssup}\left(\frac{d\mathbb{P}_{S,\rho_S}}{d(\mu_z^{\otimes n}\otimes\pi)}(\mathcal{W}, S)\right)\right) \\ &= \log\left(\operatorname{esssup}\left(\frac{d\rho_S}{d\pi}(\mathcal{W})\right)\right),\end{aligned}$$

where the essential supremum is over $\mathbb{P}_{S,\rho_S}$. The result follows. ■

B.4 Additional details and omitted proofs of Section 6

B.4.1 Expression of the KL divergence - omitted proofs of Section 6

The next result gives the expression of the IT terms appearing in Theorem 6.

Theorem 34 *Let $F : \mathbb{R}^d \longrightarrow \mathbb{R}$ be an arbitrary function that is bounded, Lipschitz, and smooth (i.e., Assumption 4). Let $S \in \mathcal{Z}^n$, we assume that ℓ satisfies Assumptions 1, 4 and 5. We consider a probability measure π on Ω, under which we have the following SDE:*

$$dW_t = -\nabla F(W_t)dt + \sigma dB_t, \quad W_0 = w_0. \tag{38}$$

Then there exists a probability measure ρ_S on Ω such that, under ρ_S, W satisfies:

$$dW_t = -\nabla \widehat{\mathcal{R}}_S(W_t)dt + \sigma dB_t^S, \quad W_0 = w_0,$$

with $(B_t^S)_{t\geq 0}$ a ρ_S-Brownian motion. Moreover, we have $\rho_S \sim \pi$ and the following relation:

$$\mathbf{KL}(\rho_S\|\pi) = \frac{1}{2\sigma^2}\int_0^T \mathbb{E}_{\rho_S}\big[\|\nabla \widehat{\mathcal{R}}_S(W_t) - \nabla F(W_t)\|^2\big]dt,$$

The proof of this result follows from classical arguments and is an adaptation of (Aristoff, 2012). The reader may find technical background related to Girsanov's theorem and Noikov's condition in Øksendal (2003).

Proof Let us fix S and denote $U := F - \widehat{\mathcal{R}}_S$. We also denote by $\mathcal{F}_{t\geq 0}$ a right-continuous filtration of $(\Omega, \mathcal{T})$ such that the Brownian motion $(B_t)_t$ is adapted to $(\mathcal{F}_t)_t$. Without loss of generality, we assume $\mathcal{F}_T = \mathcal{T}$, where $\mathcal{T}$ is the σ-algebra on Ω.
We define the probability measure ρ_S, on the filtration $(\mathcal{F}_t)$ by:

$$\frac{d\rho_S|_{\mathcal{F}_t}}{d\pi|_{\mathcal{F}_t}} := \exp\left\{\frac{1}{\sigma}\int_0^t \nabla U(W_s)\cdot dB_s - \frac{1}{2\sigma^2}\int_0^t \|\nabla \widehat{\mathcal{R}}_S(W_s) - \nabla \mathcal{F}(W_s)\|^2 ds\right\}, \tag{39}$$

where π and ρ_S denote the restrictions to $\mathcal{F}_T$.
It is known, thanks to the Novikov condition and our assumptions on the SDEs, that this defines a continuous π-martingale. Its stochastic logarithm, $\frac{1}{\sigma}\int_0^t \nabla U(W_s)\cdot dB_s$, is also a martingale. As B_t is also a π-martingale, by Girsanov's theorem, we know that the following is a continuous local ρ_S-martingale:

$$Y_t := B_t - \frac{1}{\sigma}\int_0^t \nabla U(W_s)ds.$$

Moreover, $[Y,Y]_t = t(\delta_{ij})_{1\leq i,j\leq d}$ (the quadratic variation is the same when defining it with two equivalent probability measures), so by Lévy's theorem, it is actually a ρ_S-Brownian motion, which we will denote $B_t^S := Y_t$.
Then we have, almost surely (under either π or ρ_S):

$$\sigma B_t^S = \sigma B_t - \int_0^t \nabla F(W_s)ds + \int_0^t \nabla \widehat{\mathcal{R}}_S(W_s)ds = W_t - W_0 + \int_0^t \nabla \widehat{\mathcal{R}}_S(W_s)ds,$$

which is the desired dynamics. By a similar calculation based on Itô's lemma, we have:

$$\frac{1}{\sigma}\int_0^t \nabla U(W_s)\cdot dB_s = \frac{1}{\sigma}\int_0^t \nabla U(W_s)\cdot dB_s^S + \frac{1}{\sigma^2}\int_0^t \|\nabla U(W_s)\|^2\, ds.$$

Hence, for the KL divergence, by the martingale property and Fubini's theorem, using Equation (39), we have:

$$\begin{aligned}\mathbf{KL}(\rho_S\|\pi) &= \frac{1}{\sigma}\mathbb{E}_{\rho_S}\left[\int_0^T \nabla U(W_t)\cdot dB_t^S\right] + \frac{1}{2\sigma^2}\mathbb{E}_{\rho_S}\left[\int_0^T \|\nabla\widehat{\mathcal{R}}_S(W_t)-\nabla F(W_t)\|^2 dt\right]\\ &= \frac{1}{2\sigma^2}\int_0^T \mathbb{E}_{\rho_S}\left[\|\nabla\widehat{\mathcal{R}}_S(W_t)-\nabla F(W_t)\|^2\right]dt.\end{aligned}$$

■

Our use of Girsanov's theorem allows for more general changes of measure than what is given by Theorem 34. This is formalized by the following remark.

Remark 35 (Disintegrated bounds from Girsanov's theorem) *In addition to providing a closed-form expression of the KL divergence appearing in the KL-based bound of Theorem 6, the proof of Theorem 34 gives a formula for the Radon-Nykodym derivative $d\rho_S/d\pi$, through Equation* (39)*. This formula can therefore be used to perform more general changes of measure and in particular, it naturally provides an explicit form of the Radon-Nykodym derivative appearing in Equation* (16)*. This is a straightforward extension of the presented theory. For the sake of simplicity, we do not discuss it in more details here.*

Equation (29) and Corollary 20 are immediate consequences of the above theorem and Theorem 11, using $F = 0$. We now present the proof of Proposition 21. Let us consider the setting of Section 6.2.2, where π represents the expected dynamics prior defined in Section 6.1. Then the following results hold. To avoid harming the readability of the paper, we only sketch this proof.

Proof The idea is the following: we fix some $\alpha > 0$ and apply the PAC-Bayesian bound of Theorem 3 to the function $\phi : \Omega, \mathcal{Z}^n \longrightarrow \mathbb{R}$, given by:

$$\phi(\omega, S) := \frac{1}{2\sigma^2}\int_0^T \mathbb{E}_{\rho_S}\left[\|\nabla\widehat{\mathcal{R}}_S(W_t(\omega)-\nabla F(W_t(\omega)\|^2\right]dt.$$

This gives us that, for any $\zeta \in (0,1)$, we have, with probability at least $1-\zeta$ over $\mu_z^{\otimes n}$:

$$\alpha\mathbb{E}_{\rho_S}\left[\frac{1}{2\sigma^2}\int_0^T \|\nabla\widehat{\mathcal{R}}_S(W_t)-\nabla\mathcal{R}(W_t)\|^2 dt\right] \le \mathbf{KL}(\rho_S\|\pi) + \log(1/\zeta) + \log(E),$$

with (keep in mind that $\sigma = \sqrt{2\beta^{-1}}$):

$$E := \mathbb{E}_S\mathbb{E}_\pi\left[\exp\left\{\frac{\alpha\beta}{4}\int_0^T \|\nabla\widehat{\mathcal{R}}_S(W_t)-\nabla\mathcal{R}(W_t)\|^2 dt\right\}\right].$$

By using the previously obtained expression of the KL divergence, this can be rewritten as:

$$(\alpha - 1)\mathbf{KL}(\rho_S \| \pi) \leq \log(1/\zeta) + \log(E).$$

By Jensen's inequality, we have:

$$E \leq \frac{1}{T}\int_0^T \mathbb{E}_S \mathbb{E}_\pi \left[\exp\left\{ \frac{\alpha\beta}{4} T \|\nabla \widehat{\mathcal{R}}_S(W_t) - \nabla \mathcal{R}(W_t)\|^2 \right\} \right] dt.$$

A quick computation shows that:

$$\mathbb{E}_S[\|\nabla \widehat{\mathcal{R}}_S(W_t) - \nabla \mathcal{R}(W_t)\|^2] \leq \frac{2L^2}{n}.$$

Moreover, if S and S^i are two datasets of size n differing by only the i-th element, we have, by the inverted triangle inequality:

$$\left| \|\nabla \widehat{\mathcal{R}}_S(W_t) - \nabla \mathcal{R}(W_t)\|^2 - \|\nabla \mathcal{R}_{S^i}(W_t) - \nabla \mathcal{R}(W_t)\|^2 \right| \leq \frac{8L^2}{n}.$$

Therefore, by Lemma 24, we have:

$$\mathbb{E}_S \left[\exp\left\{ \frac{\alpha\beta T}{4} \left(\|\nabla \widehat{\mathcal{R}}_S(W_t) - \nabla \mathcal{R}(W_t)\|^2 - \mathbb{E}_S[\|\nabla \widehat{\mathcal{R}}_S(W_t) - \nabla \mathcal{R}(W_t)\|^2] \right) \right\} \right] \leq e^{\alpha^2 \frac{\beta^2 T^2 L^4}{2n}}.$$

Combining these equations and applying Fubini's theorem, we get that:

$$(\alpha - 1)\mathbf{KL}(\rho_S \| \pi) \leq \log(1/\zeta) + \alpha \frac{L^2 \beta T}{2n} + \alpha^2 \frac{\beta^2 T^2 L^4}{2n}. \tag{40}$$

The result follows by choosing $\alpha = 2$ in the above computation. ■

B.4.2 Rademacher complexity of CLD - omitted proofs of Section 6.3

To end this section, we prove our bound for the Rademacher complexity of Langevin dynamics. We use the following lemma, which is taken from (Vershynin, 2018, Exercise 2.5.10).

Lemma 36 *On a probability space $(\Omega, \mathbb{P})$, we consider almost surely non-negative random variables $(X_1, \ldots, X_N)$ (not necessarily i.i.d.) and $\Sigma > 0$ such that, for all i, we have:*

$$\forall a \geq 0,\ \mathbb{P}\left(X_i \geq a\right) \leq 2e^{-\frac{a^2}{2\Sigma}},$$

then there exists an absolute constant $A > 0$ such that $\mathbb{E}\left[\max_{1 \leq i \leq N} X_i\right] \leq A\sqrt{\Sigma \log(N)}$.

We can now prove Theorem 22.

Proof Let us fix $S \in \mathcal{Z}^n$ and some integer $K \in \mathbb{N}^\star$, let $\delta := T/K$. For $i \in \{0, \ldots, K\}$, we define $t_i := i\delta$ and suppose that K is big enough so that $\delta < 1$. From Theorem 34 and

its proof, (B_t^S) is a ρ_S Brownian motion in $\mathbb{R}^d$, we denote its coordinates (which are ρ_S-independent) by $B_t^S := (B_{1,t}^S, \dots, B_{d,t}^S)$. We also introduce some *i.i.d.* Rademacher random variables $\epsilon = (\epsilon_1, \dots, \epsilon_n)$. Let[11] $\omega \sim \rho_S$ and $\mathcal{W} = \mathcal{W}(\omega)$, we know that:

$$\mathbf{Rad}_S(\mathcal{W}) = \mathbb{E}_\epsilon \sup_{0\le t\le T} \frac{1}{n}\sum_{i=1}^n \epsilon_i \ell(W_t, z_i)$$

Let us take $t \in [0,T]$, then there exists j such that $t_j > t \ge t_{j-1}$. Therefore:

$$\frac{1}{n}\sum_{i=1}^n \epsilon_i \ell(W_t, z_i) \le L \left\|W_t - W_{t_{j-1}}\right\| + \max_{0\le j\le K-1} \frac{1}{n}\sum_{i=1}^n \epsilon_i \ell(W_{t_j}, z_i).$$

We focus on the first term of this equation, from Equation (27), we have, ρ_S-almost surely:

$$\begin{aligned}
\left\|W_t - W_{t_{j-1}}\right\| &\le \left\| \int_{t_{j-1}}^t \nabla \widehat{\mathcal{R}}_S(W_u) du \right\|_2 + \sigma \left\| B_t^S - B_{t_{j-1}}^S \right\|_2 \\
&\le L\delta + \sigma \max_{0\le j\le K-1} \sup_{t_j \le s < t_{j+1}} \left\| B_s^S - B_{t_j}^S \right\|_2 \\
&\le L\delta + \sigma \max_{0\le j\le K-1} \sup_{t_j \le s < t_{j+1}} \sum_{k=1}^d |B_{k,s}^S - B_{k,t_j}^S| \\
&\le L\delta + \sigma \sum_{k=1}^d \max_{0\le j\le K-1} \underbrace{\sup_{t_j \le s < t_{j+1}} |B_{k,s}^S - B_{k,t_j}^S|}_{:= Y_{k,j}}.
\end{aligned}$$

Each of the coordinates of B_t^S are one-dimensional standard Brownian motion. Now we fix k; from the strong Markov property, we know that the $Y_{k,j}$ are independent for $0 \le j \le K-1$. Moreover, they all have the same distribution as $Y_{k,1}$. We can also write that:

$$Y_{k,1} = \sup_{0\le t\le \delta} |B_{1,t}^S| \le \sup_{0\le t\le\delta} B_{1,t}^S + \sup_{0\le t\le\delta} \left(-B_{1,t}^S\right),$$

where the inequality follows from the fact that $\sup_{0\le t\le\delta} B_{1,t}^S$ and $\sup_{0\le t\le\delta}(-B_{1,t}^S)$ are almost-surely positive. From the reflection principle, we have that for all $0 \le j \le K-1$:

$$\mathbb{P}_{\rho_S}\left(\sup_{0\le t\le\delta} B_{1,t}^S \ge a \right) \le 2\mathbb{P}_{\rho_S}\left(B_{1,\delta}^S \ge a\right) \le 2e^{-\frac{a^2}{2\delta}},$$

where the last inequality follows from (Vershynin, 2018, Equation (2.10)). It is clear that $\sup_{0\le t\le\delta}(-B_{1,t}^S)$ satisfies the same inequality. By Lemma 36, we have that there exists an absolute constant $C > 0$, such that:

$$\mathbb{E}_{\rho_S}\left[\mathbf{Rad}_S(\mathcal{W})\right] \le L^2\delta + CLd\sigma\sqrt{\delta}\sqrt{\log(K)} + \mathbb{E}_{\rho_S}\left[\mathbf{Rad}_S(\{W_{t_0}, \dots, W_{t_{K-1}}\})\right].$$

11. There is a slight abuse of notation here, we just want to highlight that we consider the posterior distribution ρ_S on Ω.

Therefore, by Massart's lemma (Shalev-Schwartz and Ben-David, 2014, Lemma 26.8):

$$\mathbb{E}_{\rho_S}\left[\mathbf{Rad}_S(\mathcal{W})\right] \leq L^2\delta + \left(CLd\sigma\sqrt{\delta} + B\sqrt{\frac{2}{n}}\right)\sqrt{\log(T/\delta)}.$$

We choose $K := \lceil TnL^2(1 + C^2d^2\sigma^2)\rceil$, and the bound follows. ■

B.5 Omitted proofs of Section 7

B.5.1 Setting details

Let us precise our setting for SGLD. We consider a measurable space $(\Omega, \mathcal{F})$, endowed with a filtration $\mathbb{F} := (\mathcal{F}_k)_{k\geq 0}$, *i.e.* a sequence of σ-algebras, $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}_T$. By convention, we set $\mathcal{F} = \mathcal{F}_T$. We also fix a dataset $S \in \mathcal{Z}^n$ and a probability distribution $\mathbb{P}$ on $(\Omega, \mathcal{F})$. We consider $W^S \in \mathbb{R}^d$ satisfying Equation (6), *i.e.*,

$$\forall k \in \mathbb{N},\ W^S_{k+1} = W^S_k - \eta_{k+1}\hat{g}_{k+1} + \sigma_{k+1}\epsilon_{k+1}, \quad \epsilon_{k+1} \sim \mathcal{N}(0, I_d). \tag{41}$$

Note that the dependence of W^S on S comes from the fact that $\hat{g}_{k+1}$ is an unbiased estimate of $\nabla\widehat{\mathcal{R}}_S(W_k)$. We assume that:

- $(\epsilon_k)_{k\geq 1}$ are adapted to $\mathbb{F}$, *i.e.* ϵ_k is measurable with respect to the σ-algebra $\mathcal{F}_k$, and are *i.i.d.* with distribution $\mathcal{N}(0, I_d)$.
- $(\hat{g}_k)_{k\geq 1}$ are adapted to $\mathbb{F}$.
- For $k \geq 1$, ϵ_k is independent of the following σ-algebra $\tilde{\mathcal{F}}_k := \sigma\left(\sigma(\hat{g}_k) \cup \mathcal{F}_{k-1}\right)$.

Note that this also implies that $(W_k)_{k\geq 1}$ is adapted to $\mathbb{F}$. To simplify our final bounds, we assume that $\hat{g}_{k+1}$ has the form:

$$\hat{g}_{k+1} = G(W^S_k, S, U_{k+1}), \tag{42}$$

where U_{k+1} denotes a sequence of *i.i.d.* random variables which adapted to $\mathbb{F}$ and independent from ϵ_{k+1} and W_k. For example, U_{k+1} may denote a random set of indices of $\{1, \ldots, n\}$, with size $b \in \mathbb{N}^\star$, in which case $\hat{g}_{k+1} = \frac{1}{b}\sum_{i\in U_{k+1}} \nabla\ell(W_k, z_i)$. This is similar to settings adopted in other works (Mou et al., 2018; Negrea et al., 2019).

B.5.2 Proof of Theorem 23

To prove Theorem 23, we are going to write lemmas acting as a discrete version of Girsanov's theorem and Lévy's characterization theorem. While those are classical arguments, we provide proof for the sake of completeness.
Let us first make the following remark regarding this proof technique.

Remark 37 (Generality of the Girsanov approach) *To bound the KL divergence appearing in Theorem 6 for SGLD and obtain Theorem 23, it is not necessary to prove a discrete version of Girsanov theorem. However, we describe this proof to highlight the similarity with the use of Girsanov's theorem in Section 6.2 and to obtain a more general change*

of measure. Indeed, our approach does not only provide a bound on the KL divergence, it gives a closed-form expression of the Radon-Nykodym derivative between the posterior and prior distributions. This is similar to what was observed in Remark 35 for the continuous case. Therefore, our framework can be used to prove disintegrated uniform generalization bounds for SGLD. If one is only interested in the KL divergence, it can simply be obtained by decomposing the joint distribution of the path of SGLD into a product of conditional distributions, we leave the details to the reader.

Let us introduce the following random variable:

$$\forall N \geq 1,\ Z_N := \prod_{k=1}^{N} e^{E_N}, \quad \text{with: } E_N := \frac{\eta_k}{\sigma_k} \langle \hat{g}_k, \epsilon_k \rangle - \frac{\eta_k^2}{2\sigma_k^2} \|\hat{g}_k\|^2 . \tag{43}$$

For our method to work, we need the following assumption, which is, in particular, true if the stochastic gradients $\hat{g}_k$ are almost surely bounded.

Assumption 8 *The random variables W_k^S and Z_k are square integrable, i.e., in $L^2(\mathbb{P})$.*

The proof is detailed through several lemmas, which we will now present.

Lemma 38 *$(Z_k)_{k\leq 1}$ is a $\mathbb{P}$-martingale, with respect to $\mathbb{F}$, where, as a reminder, $\mathbb{F}$ denotes the filtration $\mathcal{F}_0 \subseteq \cdots \subseteq \mathcal{F}_T$.*

Proof We fix $N \geq 1$. As ϵ_N is independent of $\tilde{\mathcal{F}}_N$ (defined above), we have:

$$\mathbb{E}\left[Z_N \middle| \mathcal{F}_{N-1}\right] = Z_{N-1}\mathbb{E}\left[e^{E_N}\middle|\mathcal{F}_{N-1}\right] = Z_{N-1}\mathbb{E}\left[e^{-\frac{\eta_N^2}{2\sigma_N^2}\|\hat{g}_N\|^2}\mathbb{E}\left[e^{\frac{\eta_N}{\sigma_N}\langle \hat{g}_N, \epsilon_N\rangle}\middle|\tilde{\mathcal{F}}_N\right]\middle|\mathcal{F}_{N-1}\right].$$

From the formula for the Moment Generating Function (MGF) of multivariate Gaussian distributions, we deduce that $\mathbb{E}\left[Z_N \middle| \mathcal{F}_{N-1}\right] = Z_{N-1}$. ■

Lemma 39 *$(W_k^S Z_k)_{k\leq 1}$ is a $\mathbb{P}$-martingale, with respect to $\mathbb{F}$.*

Proof We fix $N \geq 1$. With similar arguments to the previous proof and using the definition of W^S, we have:

$$\begin{aligned}\mathbb{E}\left[W_N^S Z_N \middle| \mathcal{F}_{N-1}\right] &= W_{N-1}^S \mathbb{E}\left[Z_N \middle| \mathcal{F}_{N-1}\right] - \eta_N \mathbb{E}\left[Z_N \hat{g}_N \middle| \mathcal{F}_{N-1}\right] + \sigma_N \mathbb{E}\left[Z_N \epsilon_N \middle| \mathcal{F}_{N-1}\right] \\ &= W_{N-1}^S Z_{N-1} - \eta_N Z_{N-1}\mathbb{E}\left[e^{E_N}\hat{g}_N \middle| \mathcal{F}_{N-1}\right] + \sigma_N Z_{N-1}\mathbb{E}\left[e^{E_N}\epsilon_N \middle| \mathcal{F}_{N-1}\right].\end{aligned}$$

But we also compute separately:

$$\mathbb{E}\left[e^{E_N}\hat{g}_N \middle| \mathcal{F}_{N-1}\right] = \mathbb{E}\left[\hat{g}_N e^{-\frac{\eta_N^2}{2\sigma_N^2}\|\hat{g}_N\|^2}\mathbb{E}\left[e^{\frac{\eta_N}{\sigma_N}\langle \hat{g}_N, \epsilon_N\rangle}\middle|\tilde{\mathcal{F}}_N\right]\middle|\mathcal{F}_{N-1}\right] = \mathbb{E}\left[\hat{g}_N \middle| \mathcal{F}_{N-1}\right],$$

$$\mathbb{E}\left[e^{E_N}\epsilon_N \middle| \mathcal{F}_{N-1}\right] = \mathbb{E}\left[e^{-\frac{\eta_N^2}{2\sigma_N^2}\|\hat{g}_N\|^2}\mathbb{E}\left[e^{\frac{\eta_N}{\sigma_N}\langle \hat{g}_N, \epsilon_N\rangle}\epsilon_N\middle|\tilde{\mathcal{F}}_N\right]\middle|\mathcal{F}_{N-1}\right] = \frac{\eta_N}{\sigma_N}\mathbb{E}\left[\hat{g}_N \middle| \mathcal{F}_{N-1}\right],$$

where the last line follows from $\mathbb{E}_{X\sim\mathcal{N}(0,I_d)}\left[e^{\langle u,X\rangle}X\right] = e^{\|u\|^2/2}u$. The result follows. ■

We define a new probability measure on $(\Omega,\mathcal{F})$ by $\frac{d\mathbb{Q}}{d\mathbb{P}} = Z_T$. Thanks to Lemma 38, we have $\frac{d\mathbb{Q}|_{\mathcal{F}_N}}{d\mathbb{P}|_{\mathcal{F}_N}} = Z_N$, for $N \leq T$. By construction, it follows from Lemma 39 that $(W_k^S)_{k\geq 1}$ is a $\mathbb{Q}$-martingale with respect to $\mathbb{F}$. Now we define:

$$Y_N := \sum_{k=1}^{N}\epsilon_k - \sum_{k=1}^{N}\frac{\eta_k}{\sigma_k}\hat{g}_k, \tag{44}$$

and by convention $Y_0 = 0$. The following lemma is now clear from $Y_N - Y_{N-1} = \frac{W_N^S - W_{N-1}^S}{\sigma_N}$.

Lemma 40 *$(Y_k)_k$ is a $\mathbb{Q}$-martingale, with respect to $\mathbb{F}$.*

The most important lemma is the following, it shows that, under $\mathbb{Q}$, W^S is a data-independent normal random walk.

Lemma 41 *The variables $(Y_k - Y_{k-1})_{k\geq 1}$ are, under $\mathbb{Q}$, independent and identically distributed with distribution $\mathcal{N}(0, I_d)$.*

Proof Inspired by the proof of Lévy's theorem for the characterization of Brownian motion from its quadratic variation, we prove the lemma by computing Characteristic Functions (CF). To achieve this, let $J \subseteq \{1,\dots,T\}$ be an arbitrary set of indices. Let $M := |J|$ and $j_0 := \max(J)$. We denote $\Delta_k := Y_k - Y_{k-1}$ and compute the following CF, for $u \in (\mathbb{R}^d)^J$:

$$\mathbb{E}_{\mathbb{Q}}\left[e^{i\sum_{j\in J}\langle u_j,\Delta_j\rangle}\right] = \mathbb{E}_{\mathbb{Q}}\left[e^{i\sum_{j\in J\setminus\{j_0\}}\langle u_j,\Delta_j\rangle}\mathbb{E}_{\mathbb{Q}}\left[e^{i\langle u_{j_0},\Delta_{j_0}\rangle}\middle|\tilde{\mathcal{F}}_{j_0}\right]\right] \tag{45}$$

$$= \mathbb{E}_{\mathbb{Q}}\left[e^{i\sum_{j\in J\setminus\{j_0\}}\langle u_j,\Delta_j\rangle}e^{-i\frac{\eta_{j_0}}{\sigma_{j_0}}\langle\hat{g}_{j_0},u_{j_0}\rangle}\mathbb{E}_{\mathbb{Q}}\left[e^{i\langle u_{j_0},\epsilon_{j_0}\rangle}\middle|\tilde{\mathcal{F}}_{j_0}\right]\right]. \tag{46}$$

For any $N \leq T$, we compute, from the definition of Z_N, for any $A \in \tilde{\mathcal{F}}_N$:

$$\begin{aligned}\mathbb{E}_{\mathbb{Q}}\left[e^{i\langle u_N,\epsilon_N\rangle}\mathbb{1}_A\right] &= \mathbb{E}_{\mathbb{P}}\left[e^{i\langle u_N,\epsilon_N\rangle}e^{E_N}Z_{N-1}\mathbb{1}_A\right]\\ &= \mathbb{E}_{\mathbb{P}}\left[Z_{N-1}\mathbb{1}_A e^{-\frac{\eta_N^2}{2\sigma_N^2}\|\hat{g}_N\|^2}\mathbb{E}_{\mathbb{P}}\left[e^{\frac{\eta_N}{\sigma_N}\langle\hat{g}_N,\epsilon_N\rangle+i\langle u_N,\epsilon_N\rangle}\middle|\tilde{\mathcal{F}}_N\right]\right]\\ &= \mathbb{E}_{\mathbb{P}}\left[Z_{N-1}\mathbb{1}_A e^{i\frac{\eta_N}{\sigma_N}\langle\hat{g}_N,u_N\rangle}e^{-\frac{\|u_N\|^2}{2}}\right],\end{aligned}$$

where we used $\mathbb{E}_{X\sim\mathcal{N}(0,I_d)}\left[e^{\langle a+ib,X\rangle}\right] = e^{\frac{\|a\|^2}{2}+i\langle a,b\rangle-\frac{\|b\|^2}{2}}$, for $a,b\in\mathbb{R}^d$. Hence

$$\mathbb{E}_{\mathbb{Q}}\left[e^{i\langle u_N,\epsilon_N\rangle}\middle|\tilde{\mathcal{F}}_N\right] = e^{-\frac{\|u_N\|^2}{2}}e^{i\frac{\eta_N}{\sigma_N}\langle\hat{g}_N,u_N\rangle}.$$

Therefore, we deduce that:

$$\mathbb{E}_{\mathbb{Q}}\left[e^{i\sum_{j\in J}\langle u_j,\Delta_j\rangle}\right] = e^{-\frac{\|u_{j_0}\|^2}{2}}\mathbb{E}_{\mathbb{Q}}\left[e^{i\sum_{j\in J\setminus\{j_0\}}\langle u_j,\Delta_j\rangle}\right].$$

The result is implied by an immediate recursion (on $M = |J|$) and identifying the CF of multivariate Gaussian distributions. ■

This shows that, under $\mathbb{Q}$, W follows the dynamics, $W_{k+1} = W_k + \sigma_{k+1}\mathcal{N}(0, I_d)$, with independent realizations of $\mathcal{N}(0, I_d)$ at each iterations. Let us denote by $F_T(\mathbb{R}^d)$ the set of finite subsets of $\mathbb{R}^d$ with cardinality T. With a slight abuse of notation, we see W^S as a map $W^S : \Omega \longrightarrow F_T(\mathbb{R}^d)$. We define the posterior and prior distributions on $F_T(\mathbb{R}^d)$ by:

$$\rho_S = W^S_{\#}\mathbb{P}, \quad \pi = W^S_{\#}\mathbb{Q}. \tag{47}$$

Remark 42 *ρ_S is data-dependent by definition of the dynamics. While $\mathbb{Q}$ may depend on the data S, the pushforward $W^S_{\#}\mathbb{Q}$ is not data-dependent, as it corresponds to a data-independent dynamics, e.g., $W_{k+1} = W_k + \sigma_{k+1}\mathcal{N}(0, I_d)$.*

We can now prove the main result of Section 7, *i.e.* Theorem 23.

Proof We apply Theorem 10 and, as the random sets drawn from ρ_S and π are surely of cardinal $T < +\infty$, we apply the reasoning of Example 6 to get:

$$\mathbb{E}_{\rho_S}\left[\max_{w\in\mathcal{W}}\left(\mathcal{R}(w) - \widehat{\mathcal{R}}_S(w)\right)\right] \leq \frac{1}{\lambda}\left(\log(T/\zeta) + \mathbf{KL}(\rho_S\|\pi)\right) + \lambda\frac{2B^2}{n}.$$

By the data processing inequality, we know that we have, for a fixed $S \in \mathcal{Z}^n$, $\mathbf{KL}(\rho_S\|\pi) \leq \mathbf{KL}(\mathbb{P}\|\mathbb{Q})$, where $\mathbb{P}$ and $\mathbb{Q}$ correspond to the notations above (the dependence on S is implicit here). Using the definition of Z_T, we easily compute, from the fact that ϵ_k is independent of $\hat{g}_k$:

$$\mathbf{KL}(\mathbb{P}\|\mathbb{Q}) = -\int \log(Z_T)d\mathbb{P} = \frac{1}{2}\sum_{k=1}^{T}\frac{\eta_k^2}{\sigma_k^2}\mathbb{E}_{\mathbb{P}}\left[\|\hat{g}_k\|^2\right] = \frac{\beta}{4}\sum_{k=1}^{T}\eta_k\mathbb{E}_{\mathbb{P}}\left[\|\hat{g}_k\|^2\right].$$

■

B.6 Random closed sets formalization: omitted proofs

In this section, we prove the measurability results related to the general measure-theoretic construction that we propose in Section 3.2. This is essential to provide strong theoretical foundations for the introduced techniques.

As already mentioned, the reader may refer to (Molchanov, 2017) for a more detailed introduction to the theory of random (closed) sets. The Effrös σ-algebra was defined in Definition 8, we slightly refine its definition in the following proposition, which summarizes results from Propositions 1.1.1 and 1.1.1′ of (Molchanov, 2017).

Proposition 43 *The σ-algebra $\mathfrak{E}(\mathbb{R}^d)$ is generated by both of the following family of sets:*

$$\left\{\mathcal{F}_K,\ K \subset \mathbb{R}^d \text{ compact}\right\}, \quad \text{and: } \left\{\mathcal{F}_U,\ U \subseteq \mathbb{R}^d \text{ open}\right\},$$

where $\mathcal{F}_A := \left\{C \in \mathbf{CL}(\mathbb{R}^d), C \cap A \neq \emptyset\right\}$.

We now prove Lemma 9.

Proof It is enough to prove that, for each $\forall t \in \mathbb{R},\ \{\Phi > t\} \in \mathfrak{E}(\mathbb{R}^d) \otimes \mathcal{T}$. We remark that

$$\begin{aligned}\Phi(\mathcal{W},\omega) > t &\iff \exists w \in \mathcal{W},\ \exists \eta \in \mathbb{Q}_{>0},\ \zeta(w,\omega) \geq t+\eta \\ &\iff \exists q \in \mathbb{Q}^d,\ \exists \epsilon,\eta \in \mathbb{Q}_{>0},\ \begin{cases} \forall q' \in \mathbb{Q}^d \cap B(q,\epsilon),\ \zeta(q',\omega) \geq t+\eta \\ \mathcal{W} \cap B(q,\epsilon) \neq \emptyset, \end{cases}\end{aligned}$$

hence

$$\{\Phi(\mathcal{W},\omega) > t\} = \bigcup_{q\in\mathbb{Q}^d,\epsilon,\eta\in\mathbb{Q}_{>0}} \mathcal{F}_{B(q,\epsilon)} \cap \bigcap_{q'\in\mathbb{Q}^d\cap B(q,\epsilon)} \mathbf{CL}(\mathbb{R}^d) \times \left\{x,\ \zeta(q',\omega) \geq t+\eta\right\},$$

where, implicitely, we see $\mathcal{F}_{B(q,\epsilon)}$ as $\mathcal{F}_{B(q,\epsilon)} \times \Omega$. The above is in $\mathfrak{E}(\mathbb{R}^d) \otimes \mathcal{T}$ by countable unions, intersections, and using the previous proposition. ■

B.6.1 Covering numbers measurability

In this subsection, we justify that the random closed sets formalization also implies the measurability of the covering numbers used in Section 5 (and therefore of the fractal dimensions). Without loss of generality, we can only consider "rational" covering numbers, which we define as:

Definition 44 (Rational covering numbers) *Let $X \subset \mathbb{R}^d$ be a ρ-bounded set and $\delta > 0$. Then $N_\delta^{\mathbb{Q}} \subset \mathbb{Q}^d$ is a minimal set of points, in $\mathbb{Q}^d$, such that $X \subset \bigcup_{w\in N_\delta^{\mathbb{Q}}} \bar{B}_\delta(w)$.*

It is clear that, with the notations of Section 5.1, we have $|N_{2\delta}^{\mathbb{Q}}(X)| \leq |N_\delta(X)| \leq |N_\delta^{\mathbb{Q}}(X)|$. Hence, up to a potential small absolute constant, we do not modify the bounds presented in Section 5 by considering rational covering numbers in place of the ones used in Section 5.1. Moreover, the above inequalities imply that both notions of covering yield the same upper box-counting dimension. This leads to the following lemma.

Lemma 45 *We extend the definition of both covering numbers to be $+\infty$ on unbounded closed sets. Then the covering number $|N_\delta^{\mathbb{Q}}(X)|$ is measurable with respect to $\mathfrak{E}(\mathbb{R}^d)$.*

Proof Let $N \in \mathbb{N}^\star$ and $\delta > 0$, we just remark that:

$$\left\{\mathcal{W},\ |N_\delta^{\mathbb{Q}}(\mathcal{W})| > N\right\} = \bigcap_{m=1}^{N} \bigcap_{w_1,\ldots,w_m\in\mathbb{Q}^d} \left\{\mathcal{W},\ \mathcal{W}\backslash \bigcup_{i=1}^{m} \bar{B}_\delta(w_i) \neq \emptyset\right\},$$

which implies the desired measurability. ■

For the data-dependent covering numbers, induced by pseudometric ϑ_S, see Equation (24), and used in Section 5.1, we can perform similar reasoning and invoke Lemma 9 to conclude that the rational covering numbers, associated to pseudometric ϑ_S, are measurable with respect to $\mathfrak{E}(\mathbb{R}^d) \otimes \mathcal{F}^{\otimes n}$, as soon as the loss $\ell(w,z)$ is continuous.

References

Pierre Alquier. User-friendly Introduction to PAC-Bayes Bounds. *Foundations and Trends® in Machine Learning*, 2024.

Amiran Ambroladze, Emilio Parrado-Hernández, and John Shawe-Taylor. Tighter PAC-Bayes Bounds. In *Advances in Neural Information Processing Systems (NIPS)*, 2006.

Idan Amir, Roi Livni, and Nati Srebro. Thinking Outside the Ball: Optimal Learning with Gradient Descent for Generalized Linear Stochastic Convex Optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Ron Amit, Baruch Epstein, Shay Moran, and Ron Meir. Integral Probability Metrics PAC-Bayes Bounds. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Rayna Andreeva, Katharina Limbeck, Bastian Rieck, and Rik Sarkar. Metric Space Magnitude and Generalisation in Neural Networks. In *ICML 2023 Workshop on Topological, Algebraic and Geometric Learning*, 2023.

Rayna Andreeva, Benjamin Dupuis, Rik Sarkar, Tolga Birdal, and Umut Şimşekli. Topological Generalization Bounds for Discrete-Time Stochastic Optimization Algorithms. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

David Aristoff. Estimating Small Probabilities for Langevin Dynamics. *arXiv*, abs/1205.2400, 2012.

Pranjal Awasthi, Natalie Frank, and Mehryar Mohri. On the Rademacher Complexity of Linear Hypothesis Sets. *arXiv*, abs/2007.11045, 2020.

Peter Bartlett and Shahar Mendelson. Rademacher and Gaussian Complexities: Risk Bounds and Structural Results. *Journal of Machine Learning Research*, 2002.

Peter Bartlett, Stéphane Boucheron, and Gábor Lugosi. Model Selection and Error Estimation. *Machine Learning*, 2002.

Peter Bartlett, Dylan Foster, and Matus Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

Tolga Birdal, Aaron Lou, Leonidas Guibas, and Umut Şimşekli. Intrinsic Dimension, Persistent Homology and Generalization in Neural Networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Gilles Blanchard and François Fleuret. Occam's Hammer. In *Conference on Learning Theory (COLT)*, 2007.

Vladimir Bogachev. *Measure theory*. Springer, 2007.

Jean-Daniel Boissonnat, Frédéric Chazal, and Mariette Yvinec. *Geometric and Topological Inference*. Cambridge University Press, 2018.

Stéphane Boucheron, Gabor Lugosi, and Pascal Massart. *Concentration Inequalities - A Non-asymptotic Theory of Independence*. Oxford University Press, 2013.

Alexander Camuto, George Deligiannidis, Murat Erdogdu, Mert Gürbüzbalaban, Umut Şimşekli, and Lingjiong Zhu. Fractal Structure and Generalization Properties of Stochastic Optimization Algorithms. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Olivier Catoni. *Pac-Bayesian supervised classification: the thermodynamics of statistical learning*. Institute of Mathematical Statistics, 2007.

Ben Chugg, Hongjian Wang, and Aaditya Ramdas. A Unified Recipe for Deriving (Time-Uniform) PAC-Bayes Bounds. *Journal of Machine Learning Research*, 2023.

Arnak Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2017.

Benjamin Dupuis and Umut Şimşekli. Generalization Bounds for Heavy-Tailed SDEs through the Fractional Fokker-Planck Equation. In *International Conference on Machine Learning (ICML)*, 2024.

Benjamin Dupuis and Paul Viallard. From Mutual Information to Expected Dynamics: New Generalization Bounds for Heavy-Tailed SGD. In *NeurIPS 2023 Workshop Heavy Tails in Machine Learning*, 2023.

Benjamin Dupuis, George Deligiannidis, and Umut Şimşekli. Generalization Bounds with Data-dependent Fractal Dimensions. In *International Conference on Machine Learning (ICML)*, 2023.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. Generalization in Adaptive Data Analysis and Holdout Reuse. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

Gintare Karolina Dziugaite and Daniel Roy. Computing Nonvacuous Generalization Bounds for Deep (Stochastic) Neural Networks with Many More Parameters than Training Data. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2017.

Kenneth Falconer. *Fractal Geometry - Mathematical Foundations and Applications*. Wiley, 2014.

Tyler Farghly and Patrick Rebeschini. Time-independent Generalization Bounds for SGLD in Non-convex Settings. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Dylan Foster, Spencer Greenberg, Satyen Kale, Haipeng Luo, Mehryar Mohri, and Karthik Sridharan. Hypothesis Set Stability and Generalization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Futoshi Futami and Masahiro Fujisawa. Time-Independent Information-Theoretic Generalization Bounds for SGLD. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Pascal Germain, Alexandre Lacasse, François Laviolette, and Mario Marchand. PAC-Bayesian Learning of Linear Classifiers. In *International Conference on Machine Learning (ICML)*, 2009.

Pascal Germain, Alexandre Lacasse, François Laviolette, Mario Marchand, and Jean-Francis Roy. Risk Bounds for the Majority Vote: From a PAC-Bayesian Analysis to a Learning Algorithm. *Journal of Machine Learning Research*, 2015.

Peter D. Grünwald and Nishant A. Mehta. A Tight Excess Risk Bound via a Unified PAC-Bayesian–Rademacher–Shtarkov–MDL Complexity. In *Algorithmic Learning Theory (ALT)*, 2019.

Mert Gürbüzbalaban, Umut Şimşekli, and Lingjiong Zhu. The Heavy-Tail Phenomenon in SGD. In *International Conference on Machine Learning (ICML)*, 2021.

Mahdi Haghifam, Jeffrey Negrea, Ashish Khisti, Daniel Roy, and Gintare Karolina Dziugaite. Sharpened Generalization Bounds Based on Conditional Mutual Information and an Application to Noisy, Iterative Algorithms. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Ralf Herbrich and Thore Graepel. A PAC-Bayesian Margin Bound for Linear Classifiers: Why SVMs work. In *Advances in Neural Information Processing Systems (NIPS)*, 2000.

Liam Hodgkinson, Umut Şimşekli, Rajiv Khanna, and Michael Mahoney. Generalization Bounds Using Lower Tail Exponents in Stochastic Optimizers. In *International Conference on Machine Learning (ICML)*, 2022.

Sham Kakade, Karthik Sridharan, and Ambuj Tewari. On the Complexity of Linear Prediction: Risk Bounds, Margin Bounds, and Regularization. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.

Vladimir Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 2001.

Vladimir Koltchinskii and Dmitriy Panchenko. Rademacher processes and bounding the risk of function learning. *arXiv*, math/0405338, 2004.

Vladimir Koltchinskii and Dmitry Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics*, 2002.

Alexandre Lacasse, François Laviolette, Mario Marchand, Pascal Germain, and Nicolas Usunier. PAC-Bayes Bounds for the Risk of the Majority Vote and the Variance of the Gibbs Classifier. In *Advances in Neural Information Processing Systems (NIPS)*, 2006.

John Langford. Tutorial on Practical Prediction Theory for Classification. *Journal of Machine Learning Research*, 2005.

John Langford and Rich Caruana. (Not) Bounding the True Error. In *Advances in Neural Information Processing Systems (NIPS)*, 2001.

John Langford and John Shawe-Taylor. PAC-Bayes & Margins. In *Advances in Neural Information Processing Systems (NIPS)*, 2002.

Jian Li, Xuanyuan Luo, and Mingda Qiao. On Generalization Error Bounds of Noisy Gradient Methods for Non-Convex Learning. In *International Conference on Learning Representations (ICLR)*, 2020.

Ben London. A PAC-Bayesian Analysis of Randomized Learning with Application to Stochastic Gradient Descent. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

John Mackay and Jeremy Tyson. *Conformal Dimension: Theory and Application*. American Mathematical Society, 2010.

Pertti Mattila. *Geometry of Sets and Measures in Euclidean Spaces*. Cambridge University Press, 1999.

Andreas Maurer. A Note on the PAC Bayesian Theorem. *arXiv*, cs.LG/0411099, 2004.

David McAllester. Some PAC-Bayesian Theorems. In *Conference on Computational Learning Theory (COLT)*, 1998.

David McAllester. PAC-Bayesian Stochastic Model Selection. *Machine Learning*, 2003.

Colin McDiarmid. Concentration. In *Probabilistic methods for algorithmic discrete mathematics*. Springer, 1998.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

Ilya Molchanov. *Theory of Random Sets*. Springer, 2017.

Wenlong Mou, Liwei Wang, Xiyu Zhai, and Kai Zheng. Generalization Bounds of SGLD for Non-convex Learning: Two Theoretical Viewpoints. In *Conference On Learning Theory (COLT)*, 2018.

Vaishnavh Nagarajan and J. Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Jeffrey Negrea, Mahdi Haghifam, Gintare Karolina Dziugaite, Ashish Khisti, and Daniel Roy. Information-Theoretic Generalization Bounds for SGLD via Data-Dependent Estimates. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Gergely Neu, Gintare Karolina Dziugaite, Mahdi Haghifam, and Daniel Roy. Information-Theoretic Generalization Bounds for Stochastic Gradient Descent. In *Conference on Learning Theory (COLT)*, 2021.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-Based Capacity Control in Neural Networks. In *Conference on Learning Theory (COLT)*, 2015.

Bernt Øksendal. *Stochastic Differential Equations*. Springer, 2003.

Emilio Parrado-Hernández, Amiran Ambroladze, John Shawe-Taylor, and Shiliang Sun. PAC-Bayes Bounds with Data Dependent Priors. *Journal of Machine Learning Research*, 2012.

Ankit Pensia, Varun Jog, and Po-Ling Loh. Generalization Error Bounds for Noisy, Iterative Algorithms. *IEEE International Symposium on Information Theory (ISIT)*, 2018.

Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-Convex Learning via Stochastic Gradient Langevin Dynamics: A Nonasymptotic Analysis. In *Conference on Learning Theory (COLT)*, 2017.

Omar Rivasplata, Ilja Kuzborskij, Csaba Szepesvári, and John Shawe-Taylor. PAC-Bayes Analysis Beyond the Usual Bounds. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Sarah Sachs, Tim van Erven, Liam Hodgkinson, Rajiv Khanna, and Umut Şimşekli. Generalization Guarantees via Algorithm-dependent Rademacher Complexity. In *Conference on Learning Theory (COLT)*, 2023.

Shai Shalev-Schwartz and Shai Ben-David. *Understanding Machine Learning - From Theory to Algorithms*. Cambridge University Press, 2014.

John Shawe-Taylor and Robert Williamson. A PAC Analysis of a Bayesian Estimator. In *Conference on Computational Learning Theory (COLT)*, 1997.

Umut Şimşekli, Levent Sagun, and Mert Gürbüzbalaban. A Tail-Index Analysis of Stochastic Gradient Noise in Deep Neural Networks. In *International Conference on Machine Learning (ICML)*, 2019.

Umut Şimşekli, Ozan Sener, George Deligiannidis, and Murat Erdogdu. Hausdorff Dimension, Heavy Tails, and Generalization in Neural Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Thomas Steinke and Lydia Zakynthinou. Reasoning About Generalization via Conditional Mutual Information. In *Conference on Learning Theory (COLT)*, 2020.

Tim van Erven and Peter Harremoës. Rényi Divergence and Kullback-Leibler Divergence. *IEEE Transactions on Information Theory*, 2014.

Vladimir Vapnik and Alexey Chervonenkis. On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities. In *Doklady Akademii Nauk USSR*, 1968.

Vladimir Vapnik and Alexey Chervonenkis. On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities. *Theory of Probability and its Applications*, 1971.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

Paul Viallard, Maxime Haddouche, Umut Şimşekli, and Benjamin Guedj. Learning via Wasserstein-Based High Probability Generalisation Bounds. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Paul Viallard, Pascal Germain, Amaury Habrard, and Emilie Morvant. A general framework for the practical disintegration of PAC-Bayesian bounds. *Machine Learning*, 2024a.

Paul Viallard, Maxime Haddouche, Umut Şimşekli, and Benjamin Guedj. Tighter Generalisation Bounds via Interpolation. *arXiv*, abs/2402.05101, 2024b.

Aolin Xu and Maxim Raginsky. Information-Theoretic Analysis of Generalization Capability of Learning Algorithms. *Advances in Neural Information Processing Systems (NIPS 2017)*, 2017.

Jun Yang, Shengyang Sun, and Daniel M. Roy. Fast-rate PAC-Bayes Generalization Bounds via Shifted Rademacher Processes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Yiming Ying. McDiarmid's inequalities of Bernstein and Bennett forms, 2004.

Valentina Zantedeschi, Paul Viallard, Emilie Morvant, Rémi Emonet, Amaury Habrard, Pascal Germain, and Benjamin Guedj. Learning Stochastic Majority Votes by Minimizing a PAC-Bayes Generalization Bound. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations (ICLR)*, 2017.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 2021.