Generalization Bounds using Lower Tail Exponents in Stochastic Optimizers

Liam Hodgkinson [1]   Umut Şimşekli [2]   Rajiv Khanna [3]   Michael W. Mahoney [1]

Abstract

Despite the ubiquitous use of stochastic optimization algorithms in machine learning, the precise impact of these algorithms and their dynamics on generalization performance in realistic non-convex settings is still poorly understood. While recent work has revealed connections between generalization and heavy-tailed behavior in stochastic optimization, this work mainly relied on continuous-time approximations; and a rigorous treatment for the original discrete-time iterations is yet to be performed. To bridge this gap, we present novel bounds linking generalization to the *lower tail exponent* of the transition kernel associated with the optimizer around a local minimum, in *both* discrete- and continuous-time settings. To achieve this, we first prove a data- and algorithm-dependent generalization bound in terms of the celebrated Fernique–Talagrand functional applied to the trajectory of the optimizer. Then, we specialize this result by exploiting the Markovian structure of stochastic optimizers, and derive bounds in terms of their (data-dependent) transition kernels. We support our theory with empirical results from a variety of neural networks, showing correlations between generalization error and lower tail exponents.

1. Introduction

Fundamental to the operation of modern machine learning is *stochastic optimization*: the process of minimizing an objective function via the simulation of random elements. Its practical utility is matched by its theoretical depth; for decades, optimization theorists have sought to explain the

[1] ICSI and Department of Statistics, University of California, Berkeley, USA [2] INRIA — Département d'Informatique de l'École Normale Supérieure, PSL Research University, Paris, France [3] Department of Computer Science, Purdue University, Indiana, USA. Correspondence to: Liam Hodgkinson <liam.hodgkinson@gmail.com>.
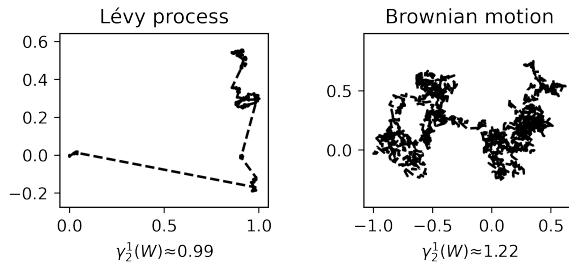
Figure 1: Discrete sample path approximations of a heavy-tailed α-stable Lévy process ($\alpha = 1.5$), and standard Brownian motion. Estimates of our normalized Fernique–Talagrand functional $\gamma_2^1(\cdot)$ is reported under each figure (see Section 2.3). Observe this functional is reduced with smaller tail index and "tighter clustering" of the trajectory.

surprising generalization ability of stochastic gradient descent (SGD) and its various extensions for non-convex problems — most recently in the context of neural networks and deep learning. Classical convex optimization-centric approaches fail to explain this phenomenon.

There has been an increasing number of attempts for developing generalization bounds for non-convex learning settings. This work has approached the problem from different perspectives, such as information theory, compression/sparsity/intrinsic dimension, or implicit (algorithmic) regularization (details to be provided in Section 1.2). Among these approaches, a promising direction has been to consider *optimization trajectories*, rather than single point estimates obtained during (or at the end of) the optimization process (e.g., Neyshabur et al. (2017); Xu & Raginsky (2017); Arora et al. (2018)). This addresses a plausible concern that single points may not necessarily be able to capture all the information regarding generalization. There has also been significant empirical evidence (using a wide variety of approaches) supporting this idea (Jastrzebski et al., 2018; Xing et al., 2018; Martin & Mahoney, 2021a; Martin et al., 2021; Jastrzebski et al., 2020; 2021). Of particular interest to us are the recent empirical developments linking *heavy-tailed fluctuations* in optimization trajectories to generalization performance (Simsekli et al., 2019b; Gurbuzbalaban et al., 2020; Hodgkinson & Mahoney, 2020).

The heavy-tailed dynamics observed in SGD exhibit qualitatively different behavior from Gaussian dynamics (see

Figure 1), and it typically coincides with improved performance (Martin & Mahoney, 2019). As a first step in providing theoretical justification for these observations, Şimşekli et al. (2020) used fractal dimension theory to prove a generalization bound involving the tail exponent of the iterates obtained from a stochastic optimizer. While they brought a new perspective, their analysis unfortunately assumes a continuous-time Feller process model as an approximation for the optimizer trajectories, and it is unclear how their techniques can be extended beyond this setting. Of course, optimization procedures used in machine learning are not continuous-time, and a rigorous treatment for the original discrete-time setting is still missing.

In this study, we address this issue and present a new mathematical framework that is sufficiently flexible to treat the discrete-time setting and to recover (and improve) the results of Şimşekli et al. (2020) in the continuous-time setting. Similar to Şimşekli et al. (2020), our underlying strategy is to explicitly include *all* the information surrounding the dynamics of the optimizer near the optimum as it applies to generalization performance. Therefore, the quantity of interest in this work is an *accumulated generalization gap* over the trajectory of the optimizer:

$$\sup_{t\in[t_1,t_2]} |\mathcal{E}_n(W_t)| = \sup_{w\in\{W_t\}_{t=t_1}^{t_2}} |\mathcal{E}_n(w)|, \qquad (1)$$

where $\{W_t\}_{t=t_1}^{t_2}$ denotes the optimizer trajectory from 'time' t_1 to t_2, $\mathcal{E}_n(w) := \mathcal{R}_n(w) - \mathcal{R}(w)$ is the generalization gap (or the generalization error), and $\mathcal{R}$ and $\mathcal{R}_n$ denote the population and empirical risk functions, respectively. As the generalization error towards the end of training is of greatest interest, t_1 will typically be chosen so that $\{W_t\}_{t=t_1}^{t_2}$ lies in the domain of attraction of a local optimum. The advantage of considering (1) is that the presence of the supremum enables tools from analytic probability theory surrounding uniform error bounds. Therefore, to bound and estimate (1), we will draw from this literature — in particular, a certain functional of Fernique (1971) and Talagrand (1996).

1.1. Contributions

Recalling that any practical stochastic optimization algorithm can be written as a Markov process (Hodgkinson & Mahoney, 2020), our main contribution is encapsulated in the following informal theorem, with the precise statement resulting from combining Theorem 1 and Corollary 2 (see Section 3).

Theorem (Informal). *Assume that an optimizer satisfies the following in the neighborhood of a local minimum w^* : there is a* lower tail exponent α *such that for any k,*

$$\mathbb{P}(\|W_{k+1} - w^*\| \leqslant r \,|\, W_k = w^*) \approx \mathcal{O}(r^\alpha), \text{ as } r \to 0^+,$$

where W_k is the kth iterate of the optimizer. Then, the expected accumulated generalization gap (1) in the neighborhood of w^* is approximately bounded by the sum of $C\sqrt{\alpha/n}$ and the mutual information between the data and the trajectory of the weights, where $C > 0$ is a constant.

To prove this result, we develop a theoretical framework for investigating the generalization properties of stochastic optimizers, in two parts. In the first part, **(I)**, the generalization gap attributable to optimizer dynamics is effectively reduced to a normalized *Fernique–Talagrand* (FT) functional. This functional is introduced in Section 2; and in Theorem 1, we obtain a sharp (up to constants) bound on the accumulated generalization gap (1) in terms of our normalized FT functional applied to the trajectory of the optimizer, and the mutual information. In the second part, **(II)**, we proceed to bound the normalized FT functional in two ways:

1. *Hausdorff dimension:* By considering the trajectory of a continuous-time stochastic optimization model and conducting a similar fractal dimension analysis to Şimşekli et al. (2020), we recover and sharpen their results in Corollary 1. In particular, using our framework, we remove many nontransparent assumptions while improving convergence rates in generalization for heavy-tailed continuous-time processes from $\mathcal{O}(n^{-1/2} \log n)$ to $\mathcal{O}(n^{-1/2})$.

2. *Transition kernel:* In Theorem 2, we bound the expected normalized FT functional of the optimizer trajectory in terms of the transition kernel of the optimizer. Similar results are obtained for continuous-time Markov models as well. Our result illustrates effective dimension reduction through properties (e.g., variance) of the kernel.

Finally, assuming that the behavior of the optimizer in a neighborhood of a local minimum can be well-approximated by a random walk, we bound the expected normalized FT functional in terms of the *lower tail exponent* of the transition kernel, thereby obtaining our main result. Smaller values of the lower tail exponent correspond to tighter "clustering" behavior in iterates of the optimizer (Figure 1). Intuitively, this translates to a form of compression in the otherwise possibly high dimensional search space, leading to better generalization performance. While our analysis relies on the randomness in stochastic optimizers, for sufficiently large step sizes, deterministic optimizers (i.e., full-batch gradient descent) are known to also exhibit stochastic behaviors (Kong & Tao, 2020), and so our results extend to these cases. Our contributions here are predominantly theoretical, motivated by recent empirical work on state-of-the-art neural network models (Martin & Mahoney, 2021a; Martin et al., 2021; Martin & Mahoney, 2021b). However, due to the relative tractability of estimating the local tail exponent, in Section 4, we demonstrate how these quantities correlates with generalization in practice.

1.2. Related Work

For motivation and comparison, we discuss some important previous efforts and adjacent concepts in the literature.

Generalization bounds. Naturally, there is a substantial literature involved in the development of "generalization bounds," which we can only briefly summarize here. For more details, see Jiang et al. (2019) and references therein. Almost universally, these bounds consider a "single-point generalization gap," that is $|\mathcal{R}_n(w) - \mathcal{R}(w)|$, for fixed weights w. Earlier bounds were typically dependent only on properties of the *model*, including Vapnik–Chervonenkis theory (Vapnik & Chervonenkis, 2015), and other norm-based bounds (Bartlett et al., 2017). Most of these can be derived or sharpened through *generic chaining* (Audibert & Bousquet, 2007), which we shall also reconsider, albeit for a different problem. Such bounds are well-known to be vacuous (Jiang et al., 2019; Bartlett & Long, 2020; Martin & Mahoney, 2021b). Non-vacuous bounds typically require some degree of *data-dependence*, with the most effective of these bounds involving measures of *sharpness* (Neyshabur et al., 2017; Jiang et al., 2019; Martin & Mahoney, 2021b). Other bounds also have some degree of *algorithm-dependence*, such as margin-based bounds (Antos et al., 2002; Sokolić et al., 2017), or bounds centered around stochastic gradient Langevin dynamics (Mou et al., 2018; Haghifam et al., 2020). Specific to explaining the generalization in neural networks, which can even fit arbitrarily labeled data points (Zhang et al., 2021), additional simplifying assumptions such as an infinite width (Arora et al., 2018), a kernelized approximation (Cao & Gu, 2020), or a simpler 2-layer setup (Farnia et al., 2018) are usually made.

Mutual information. One particular class of generalization bounds involves the *mutual information* between the data and the stochastic optimizer, quantifying the one-point generalization gap by tying it to the learning ability of the algorithm itself (Russo & Zou, 2019; Xu & Raginsky, 2017). Intuitively, the mutual information balances the trade-off between training loss and poor generalization due to overfitting. Such approaches are both data- and algorithm-dependent; and they can be made not only non-vacuous, but surprisingly tight (Asadi et al., 2018). Our Theorem 1 will also involve mutual information, extending (Xu & Raginsky, 2017) to bound the error (1). At present, applications of mutual information have tied variances in the optimizer to generalization (Pensia et al., 2018; Li et al., 2019; Negrea et al., 2019; Bu et al., 2020; Haghifam et al., 2020; Neu et al., 2021). Unfortunately, in larger models, the variance can anti-correlate with generalization (Jastrzebski et al., 2018; Martin et al., 2021; Martin & Mahoney, 2021b), which, to our knowledge, these analyses are unable to predict.

Optimization-based generalization. A body of work leans on implicit regularization effects of optimization algorithms to explain generalization, e.g., see Arora et al. (2019); Chizat & Bach (2020) and references within for certain simplified problem settings. Another line of work focuses on stability of the optimization process (Hardt et al., 2016) to bound the generalization gap. These are also one-point generalization bounds, and they do not take into account the trajectory of optimization. In lieu of the inability of convex optimization theory to explain the behavior of SGD in non-convex settings, it is common to consider the behavior of Markov process models for stochastic optimizers (Mandt et al., 2016). These models are often continuous for ease of analysis (Orvieto & Lucchi, 2018; Simsekli et al., 2019b), although discrete-time treatments have become increasingly popular (Dieuleveut et al., 2017; Hodgkinson & Mahoney, 2020; Camuto et al., 2021). Such continuous-time models are formulated as stochastic differential equations $\mathrm{d}W_t = \mu(W_t)\mathrm{d}t + \sigma(W_t)\mathrm{d}X_t$, where X_t is typically Brownian motion, or some other Lévy process, and derived through the (generalized) central limit theorem and taking learning rates to zero (Fontaine et al., 2020).

Heavy-tailed universality. Recent investigations have identified the presence of heavy tails in the dynamics of stochastic optimizers (Simsekli et al., 2019b;a; Panigrahi et al., 2019). Subsequent theoretical analyses trace the origins of these fluctuations to the presence of multiplicative noise (Hodgkinson & Mahoney, 2020; Gurbuzbalaban et al., 2020). Establishing theory connecting generalization performance to the presence of power laws has become a prominent open problem in light of the empirical and theoretical studies of Martin & Mahoney (2017; 2021a; 2019; 2020); Martin et al. (2021); these studies have explicitly tied performance to the presence of heavier tails in the spectral distributions of weights, which was then linked to generalization through compressibility, under a statistical independence assumption (Barsbey et al., 2021). Of particular note is the previous work of Şimşekli et al. (2020), correlating generalization performance with heavier-tailed dynamics; this work considered (1) in the case $(t_1, t_2) = (0, 1)$ and in the context of continuous-time stochastic optimizer models. However, their approach is bound to a continuous-time Feller process model for the optimizer; and a significant objective of this work is to extend these ideas into the natural discrete-time setting.

2. Preliminaries

2.1. Background

Let $\ell : \mathbb{R}^D \times \mathbb{R}^p \to \mathbb{R}_+$ be a non-negative loss function assessing accuracy for a model with parameters $w \in \mathbb{R}^D$ to fixed data $X \in \mathbb{R}^p$. Total model accuracy is determined

by the population risk function $\mathcal{R}(w) := \mathbb{E}_{X \sim \mathcal{D}}[\ell(w, X)]$, where $\mathcal{D}$ denotes the distribution of possible data. Therefore, the optimal choice of parameters is determined to be those solving the true risk minimization problem $\min_w \mathcal{R}(w)$. As this problem is intractable, model training is typically achieved by solving the empirical risk minimization problem: for a collection of data $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{D}$, solve

$$\min_w \mathcal{R}_n(w), \quad \text{where} \quad \mathcal{R}_n(w) := (1/n) \sum_{i=1}^{n} \ell(w, X_i)$$

is the empirical risk function. Provided this problem can be solved to near-zero empirical risk, model accuracy depends on the *generalization gap* $\mathcal{E}_n(w) = n^{-1} \sum_{i=1}^{n} R_i(w)$, where $R_i(w) = \ell(w, X_i) - \mathbb{E}_{X \sim \mathcal{D}} \ell(w, X)$.

Given a set of parameters $W \subset \mathbb{R}^D$, our objective is to bound the worst-case generalization gap over all $w \in W$, i.e., $\sup_{w \in W} |\mathcal{E}_n(w)|$. The set W is kept arbitrary for now, yet, we will be mainly interested in the case where W is part of an optimizer trajectory, e.g., $W = \{W_t\}_{t=t_1}^{t_2}$ (cf. (1)).

It is common to assume that the loss functions are Lipschitz in w, that is, for some metric $(x, y) \mapsto d(x, y)$ on $\mathbb{R}^D$, $|\ell(w, X) - \ell(w', X)| \leq L d(w, w')$ for $w, w' \in W$. Hoeffding's inequality (Boucheron et al., 2013, Theorem 2.8) then implies that the difference $R_i(w) - R_i(w')$ between two points $w, w' \in W$ is sub-Gaussian with variance parameter $(L d(w, w'))^2$. Dudley's classical method of chaining (Dudley, 1967) asserts that one can take advantage of the triangle inequality to "chain" these bounds together over particular choices of w, w' and to bound the maximal error over the set W. Talagrand (1996) later improved on this approach and developed *generic chaining*. This approach is heavily inspired by the following functional originally introduced by Fernique (1971; 1975):

$$\gamma_2(W, d) = \inf_{\mu} \sup_{w \in W} \int_0^{\text{diam}(W)} \sqrt{\log \frac{1}{\mu(B_r^d(w))}} \, dr, \quad (2)$$

where $B_r^d(w) = \{w' : d(w, w') \leq r\}$ is the ball of radius r under d around w, $\text{diam}(W) = \sup_{w, w' \in W} d(w, w')$, and the infimum is taken over all probability measures μ supported on W. We refer to the functional (2) as the *Fernique–Talagrand (FT) functional*. Generic chaining would later focus on an (equivalent, up to constants) discrete variant of this functional (Talagrand, 2001; Audibert & Bousquet, 2007; Talagrand, 2014), but for our purposes, we will find it more convenient to work with the original formulation (2).

It is also known that the approach is essentially *optimal* in the following sense: if the error $w \mapsto \mathcal{R}_n(w) - \mathcal{R}(w)$ is a Gaussian process with covariance $(w, w') \mapsto (L d(w, w'))^2$, then the FT functional both upper and lower bounds the maximal expected error in the empirical risk $\mathbb{E} \sup_{w \in W} |\mathcal{R}_n(w) - \mathcal{R}(w)|$ up to constants (Talagrand, 1996, Theorem 5.1). Therefore, in the absence of additional

information on the distribution of the empirical risk, the FT functional provides the *sharpest possible generalization bound* up to constant factors.

Broadly speaking, the FT functional simultaneously measures *variance* and *clustering*. Clustering occurs in the absence of spatial homogeneity, and it can be measured in a number of ways. For more discussion, we refer to Appendix B. Later, to draw connections to heavy-tailed theory in machine learning, e.g. (Şimşekli et al., 2020), the degree of clustering will be represented using lower tail exponents in the transition kernel.

2.2. Data-dependence with mutual information

In the classical setting, the set W is fixed and deterministic. However, in our setting of (1), W is both random and data-dependent[1]. To extend the theory to allow for data-dependent W, we shall invoke some ideas from information theory. Recall that the α-Renyi divergence is defined by

$$D_\alpha(\mu, \nu) = \frac{1}{\alpha - 1} \log \mathbb{E}\left[\frac{d\mu}{d\lambda}(Z)^\alpha \frac{d\nu}{d\lambda}(Z)^{1-\alpha} \right],$$

where $\alpha > 1$ and Z is distributed according to some (arbitrary) probability measure λ where μ and ν are absolutely continuous with respect to λ (for example, $\lambda = \frac{1}{2}(\mu + \nu)$). The α-mutual information between two random elements X, Y is defined as the α-Renyi divergence between the joint probability measure $\mathbb{P}_{X,Y}$ and the product measure $\mathbb{P}_X \otimes \mathbb{P}_Y$: $I_\alpha(X, Y) = D_\alpha(\mathbb{P}_{X,Y} \| \mathbb{P}_X \otimes \mathbb{P}_Y)$, measuring the extent of the dependence between X and Y. The standard Kullback-Leibler mutual information is obtained by taking $\alpha \to 1^+$. The α-mutual information is non-decreasing in α, that is, $I_\alpha(X, Y) \leq I_\beta(X, Y)$ for $\alpha \leq \beta$. Therefore, we may define the *total mutual information* as $I_\infty(X, Y) = \lim_{\alpha \to \infty} I_\alpha(X, Y) = \sup_\alpha I_\alpha(X, Y)$.

2.3. Normalized Fernique–Talagrand functional

It is often the case with generalization bounds that ℓ is assumed to be Lipschitz-continuous with respect to the Euclidean metric (Neyshabur et al., 2017; Mou et al., 2018). To also ensure subgaussianity of $R_i(w)$ itself, boundedness of ℓ is often assumed (Negrea et al., 2019; Şimşekli et al., 2020). Together, these two assumptions are equivalent to assuming Lipschitz continuity under the truncated metric $d_\rho(x, y) = \min\{\rho, \|x - y\|\}$, where $\rho > 0$. We refer to the corresponding FT functional $\gamma_2^\rho(W) := \gamma_2(W, d_\rho)$ as the *normalized* Fernique–Talagrand functional: for $\rho > 0$, from (2) and scaling by $1/\rho$,

$$\gamma_2^\rho(W) = \inf_{\mu} \sup_{w \in W} \frac{1}{\rho} \int_0^\rho \sqrt{\log \frac{1}{\mu(B_r(w))}} \, dr, \quad (3)$$

---

[1] We refer the reader to (Molchanov, 2005) for the definition and details of a random set.

where $B_r(w)$ denotes the Euclidean ball of radius $r > 0$ about w, and the infimum is once again over all probability measures μ on W. The functional is considered normalized as it does not grow as $\text{diam}(W) \to \infty$. Here, ρ is arbitrary, although we will find the tightest bounds to occur when $\rho = B/L$, where B and L is the upper bound, and Lipschitz constant of ℓ, respectively. It is worth noting that in the analysis to follow, the bounded+Lipschitz assumption could be relaxed (e.g. to Hölder continuity) by choosing a different metric, thus considering a different FT functional. However, we have found this choice of assumption and its corresponding FT functional (3) to yield the best compliance with our experiments.

3. Main Results

Our first main result is presented in Theorem 1 below. If W is uncountable, we interpret probabilities and expectations of suprema over uncountable sets as the corresponding supremum over all possible countable subsets.

Theorem 1. *Assume that ℓ is bounded by $B > 0$ and L-Lipschitz continuous (with respect to the Euclidean metric). There exists a universal constant $K_1 > 0$ such that for any (random) closed set $W \subset \mathbb{R}^D$, with probability at least $1 - \delta$, for any $\rho > 0$, letting $L_\rho := \max\{B, L\rho\}$,*

$$
\begin{aligned}
& \sup_{w \in W} |\mathcal{E}_n(w)| \\
& \leqslant K_1 L_\rho \left(\frac{\gamma_2^\rho(W)}{\sqrt{n}} + \sqrt{\frac{\log(1/\delta) + I_\infty(X, W)}{n}} \right).
\end{aligned} \quad (4)
$$

Furthermore, there exists $K_2 > 0$ such that

$$
\mathbb{E} \sup_{w \in W} |\mathcal{E}_n(w)| \leqslant K_2 L_\rho \left(\frac{\mathbb{E}\gamma_2^\rho(W) + \sqrt{I_1(X, W)}}{\sqrt{n}} \right).
$$

$$(5)$$

If ℓ is unbounded, then (4) holds with probability at least $1 - \delta - \mathbb{P}(\sup_{w \in W} \mathcal{R}_n(w) > B)$.

Note also that when $W = \{w^*\}$, where w^* is the location of the optimizer at some *deterministic* stopping time, we recover (up to constants) the information-theoretic bound of Xu & Raginsky (2017). Therefore, we inherit the interpretation of mutual information as a measurement of "overfitting," together with its follow-up developments (Asadi et al., 2018; Haghifam et al., 2020). The advantage of our bound is that we can better investigate the effect of dynamics on generalization through the supremum over the trajectory. We also inherit the sharpness of Theorem 1 from that of the FT functional in the event that X and W are independent.

3.1. Markov stochastic optimizers

Our remaining theoretical contributions are concerned with bounding and estimating γ_2^ρ when applied to the trajectory

of a stochastic optimizer. Such bounds on γ_2^ρ directly imply generalization bounds by applying Theorem 1.

We adopt the Markov formulation of stochastic optimizers, seen in Hodgkinson & Mahoney (2020). This formulation incorporates SGD, momentum, Adam, and stochastic Newton, among others. To summarize, approximate solutions to problems of the form $\arg\min_w \mathbb{E}\ell(w, X)$ are typically obtained by fixed point iteration: for some continuous map Ψ such that any fixed point of $\mathbb{E}\Psi(\cdot, X)$ is a minimizer of ℓ, a stochastic optimizer is constructed from the sequence of iterations $W_{k+1} = \Psi(W_k, X_{k+1})$, where X_k are independent copies of X. For example, (online) stochastic gradient descent corresponds to the choice $\Psi(w, x) = w - \gamma b^{-1} \sum_{i=1}^b \nabla \ell(w, x_i)$, where γ is a chosen learning rate and b denotes the batch size. These iterations induce a discrete-time Markov chain in W_k with transition kernel $P(w, E) = \mathbb{P}(W_{k+1} \in E \mid W_k = w)$ for a measurable set E. Under certain regimes (for example, small learning rate and large batch size), this chain is well-approximated by a continuous-time Markov process $\{W_t\}_{t \geqslant 0}$ with transition kernel $P_t(w, E) = \mathbb{P}(W_t \in E \mid W_0 = w)$ — see Fontaine et al. (2020), for example. We note that imposing a Markov assumption on the optimizer is not restrictive, as any recursive method is Markov under suitable state augmentation. Furthermore, the Markov assumption is local, so W_0 may be taken to be any point in the optimization.

Fortunately, the FT functional is sufficiently versatile that we can provide a direct generalization bound in terms of transition kernels. This is accomplished using covering arguments and the classical Dudley entropy bound (Boucheron et al., 2013, Corollary 13.2). Here is our main result for this.

Theorem 2. *For a Markov transition kernel $P(x, E)$ and any $\rho > 0$, let*

$$
\mathcal{I}_\rho[P] := \frac{1}{\rho} \int_0^\rho \sup_{x \in \mathbb{R}^D} \sqrt{\log\left(\frac{3^{D+2}}{P(x, B_r(x))} \right)} \, \mathrm{d}r.
$$

There exists a universal constant $K > 0$ such that the following bounds on the FT functional hold:

1. *Let W_k, $k \geqslant 0$ be a discrete-time homogeneous Markov chain on $\mathbb{R}^D$ with transition kernel $P^k(w, E)$. Then for the average kernel $\bar{P}_m(w, E) := m^{-1} \sum_{k=1}^m P^k(w, E)$, we have $\mathbb{E}\gamma_2^\rho(\{W_k\}_{k=0}^m) \leqslant K\mathcal{I}_\rho[\bar{P}_m]$.*

2. *Let $\{W_t\}_{t \in [0,T]}$ be a continuous-time homogeneous Markov process on $\mathbb{R}^D$ with kernel $P_t(w, E)$. Then for the average kernel $\bar{P}_T(w, E) = \frac{1}{T} \int_0^T P_t(w, E)\mathrm{d}t$, we have $\mathbb{E}\gamma_2^\rho(\{W_t\}_{t \in [0,T]}) \leqslant K\mathcal{I}_\rho[\bar{P}_T]$.*

We can use Theorems 1 and 2 to bound the generalization gap for a stochastic process as long as we can characterize the transition probability kernel of the process. Due to the

dependence on the ambient dimension, we do not expect the bounds in Theorem 2 to be sharp[2]. Fortunately, this is no concern for our purposes, as we will use Theorem 2 to imply correlations between tail properties of the transition kernel and the generalization gap.

3.2. Fractal dimensions

It has been observed by Şimşekli et al. (2020); Birdal et al. (2021) that the *fractal dimension* of the set W is often a good indicator of generalization performance. While this previous approach focused on precise covering arguments, here we show that a similar bound to that of Şimşekli et al. (2020, Theorem 2) can be readily attained using Theorem 1 under weaker assumptions. Indeed, Theorem 1 provides a remarkably straightforward illustration of the relationship between generalization performance and fractal dimension: Assuming there exists a measure μ on W such that $\mu(B_r(w)) \geqslant (cr)^\alpha$ for any $w \in W$ and $0 < r < \rho$, for some $c > 0$, $\alpha > 0$, then inserting this measure into the definition (3) reveals $\gamma_2^\rho(W) = \mathcal{O}(\sqrt{\alpha})$. If a similar upper bound also holds for μ, then most notions of fractal dimension of the set W coincide, and are *precisely equal to* α (Mattila, 1999, Chapter 5). In this way, α becomes an *effective dimension*, or *intrinsic dimension*, of W (in particular, if $W \subset \mathbb{R}^D$, then $\alpha \leqslant D$, the ambient dimension). This idea is formalized in Corollary 1. The result involves the *Hausdorff dimension* of W, i.e., $\dim_H W \in [0, D]$, which is a generalization of the usual notion of dimension to fractional orders (e.g., $\dim_H \mathbb{R}^D = D$), and the *Hausdorff measure* $\mathcal{H}^\alpha$, which is a generalization of the Lebesgue measure. We provide their exact definitions in Appendix A.

Corollary 1. *Suppose that $\dim_H W = \alpha$ and is α-Ahlfors lower regular almost surely, that is,*

$$C_\rho := \inf_{0 < r < \rho, w \in W} \frac{\mathcal{H}^\alpha(W \cap B_r(w))}{r^\alpha \mathcal{H}^\alpha(W)} > 0.$$

Then the normalized Fernique–Talagrand functional satisfies $\gamma_2^\rho(W) \leqslant (2\rho C_\rho)^{-1}\sqrt{\pi\alpha}$ almost surely.

Note that the Ahlfors regularity assumption (i.e., $C_\rho > 0$) is contained in Şimşekli et al. (2020, Assumption H4), but is a far simpler condition on its own, and does not require a Feller process construction. Furthermore, we improve on Şimşekli et al. (2020, Theorem 2), which suggests a rate $\mathcal{O}(\log n/\sqrt{n})$, whereas we obtain a rate of $\mathcal{O}(1/\sqrt{n})$.

Using a continuous-time Markov model, a precise link between Corollary 1 and stochastic optimization can be made: suppose that W_t, $t \in [0, 1]$, is a continuous-time Markov process with transition kernel $P_t(x, E)$, e.g., a

continuous-time model of a stochastic optimizer (Orvieto & Lucchi, 2018). If W_t is spatially homogeneous[3], and $\{W_t\}_{t \in [0,1]}$ is α-Ahlfors lower-regular almost surely, then by Xiao (2003, Theorem 4.2), Corollary 1 applies with $\dim_H\{W_t\}_{t \in [0,1]} = \alpha$ almost surely, where

$$\alpha = \sup\left\{\gamma \geqslant 0 : \lim_{r \to 0^+} \frac{1}{r^\gamma} \int_0^1 P_t(0, B_r(0))\mathrm{d}t < \infty\right\}. \quad (6)$$

Similar results also apply for spatially inhomogeneous Feller processes (Schilling, 1998). Therefore, Corollary 1 supports the claim that fractal dimensions (of the trajectory of a continuous Markov model of a stochastic optimizer) can be an effective measure of generalization performance. This is the strategy proposed by Şimşekli et al. (2020). However, in reality, optimization procedures are not continuous-time and optimizer trajectories are finite sets. Since all fractal dimensions are identically zero on finite sets, an alternative approach is required.

3.3. Lower tail behavior around local minima

Fortunately, we can obtain a discrete-time analogue of Corollary 1 using Theorem 2. There is one significant caveat however: while Theorem 2 successfully relates the dynamics of a Markov stochastic optimizer to generalization performance, the bound is ineffective when the increments $W_{k+1} - W_k$ are not uniformly stochastically bounded. Indeed, the bound in Theorem 2 is most tight when the optimizer exhibits random walk behavior without drift. This behavior is unlikely to occur at the global scale. However, *locally*, in the neighborhood of a local minimum, a stochastic optimizer should exhibit minimal drift. Furthermore, in practice, the behavior around a minimum is typically of greatest interest. Recall that the first objective of a stochastic optimizer is to reach and then occupy some central region Ω around a local minimum $x^* \in \Omega$ with high probability. We let ζ_m denote the probability of remaining in this region after m steps. Within this region, we assume that the optimizer behaves like a random walk $\bar{W}_{k+1} = \bar{W}_k + Z_k$, where each $Z_1, \ldots, Z_k \overset{\text{iid}}{\sim} \mu$ is independent and identically distributed. To incorporate these observations into a bound, we can appeal to approximation in total variation d_{TV}. Under these conditions, the assumption of bounded ℓ is no longer restrictive, as ℓ can be assumed to be only *locally* bounded.

With this in mind, we develop a discrete time analogue of Corollary 1, closing an open problem connecting tail exponents to generalization. Drawing inspiration from (6), we define a new α such that the kernel $P(x, B_r(x)) \approx cr^\alpha$

---

[2]We suspect this dependence can be removed using a more efficient embedding, but we leave this as an open problem. For further details, refer to the comments in the proof.

[3]There exists $K, C, r_0 > 0$ such that $P_t(x, B_{r_0}(x)) \geqslant K$ for all t, x, and $\frac{1}{C}P_t(0, B_r(0)) \leqslant P_t(x, B_r(x)) \leqslant CP_t(0, B_r(0))$ for all $t, x, r \leqslant r_0$.

as $r \to 0^+$. In particular, we can let

$$\alpha = \lim_{r \to 0^+} \frac{\log P(x, B_r(x))}{\log r}.$$

This way, α becomes an exponent on the *lower tail* of the transition kernel. Note that, *a priori*, this is *distinct from the (upper) tail exponent* of the distribution considered in Hodgkinson & Mahoney (2020); Simsekli et al. (2019b); Şimşekli et al. (2020), although the two appear to correlate in practice (see Section 4.2).

Corollary 2. *Let $\Omega \subseteq \mathbb{R}^D$ be a closed set such that $\mathbb{P}(W_k \notin \Omega$ for some $k = 0, \dots, m) \leqslant \zeta_m$ and let μ be a probability measure on $\mathbb{R}^D$. Suppose that $Z_1, \dots, Z_m \overset{iid}{\sim} \mu$ and there exists $r_0, \alpha > 0$ such that $\mathbb{P}(\|Z_1 + \dots + Z_k\| \leqslant r) \geqslant c_k r^\alpha$ for all $0 < r < r_0$, where $c_k > 0$ are constants for each $k = 1, 2, \dots$. Letting P_Ω denote the transition kernel of W_k conditioned on $W_k \in \Omega$ for $k = 1, 2, \dots, m$, there is a constant $K > 0$ such that for any $\epsilon > 0$, there exists $\rho_\epsilon > 0$ independent of α where*

$$\mathbb{E}\gamma_2^{\rho_\epsilon}(\{W_k\}_{k=0}^m) \leqslant \frac{K}{\rho_\epsilon}\sqrt{\alpha + \epsilon} +$$
$$\sqrt{\log(m+1)}\left(\zeta_m + m \sup_{x \in \Omega} d_{\text{TV}}(P_\Omega(x, x + \cdot), \mu)\right). \tag{7}$$

The second line of (7) concerns only the quality of the random walk approximation, and can mostly be ignored if we expect such an approximation to be accurate (e.g., the example below). What remains is an implied correlation between the expected normalized FT functional (itself linked to generalization through Theorem 1) and the lower tail exponent α. This can be seen in Figure 1, where a Lévy process with lower tail exponent $\alpha = 1.5$ is compared to Brownian motion with exponent $\alpha = 2$. The reduced tail exponent coincides with a corresponding reduction in γ_2^1.

Example (Perturbed Gradient Descent). Arguably the most common discrete-time model for a stochastic optimizer is the perturbed gradient descent (GD) model, which satisfies $W_{k+1} = W_k - \gamma(\nabla f(W_k) + Z_k)$, where Z_k is a Gaussian vector with zero mean and constant covariance matrix Σ representing noise in the stochastic gradient. In the neighborhood of a local optimum w^*, $\nabla f(W_k) \approx 0$, and hence the perturbed GD model resembles a Gaussian random walk $W_{k+1} = W_k + \gamma Z_k$. Here, the exponent α is *precisely* the ambient dimension, i.e., $\alpha = D$ (see Appendix D). However, we will find this exponent to be much less than the ambient dimension for an actual optimization path, suggesting an interpretation of α as a measure of *effective dimension* for the purposes of generalization.
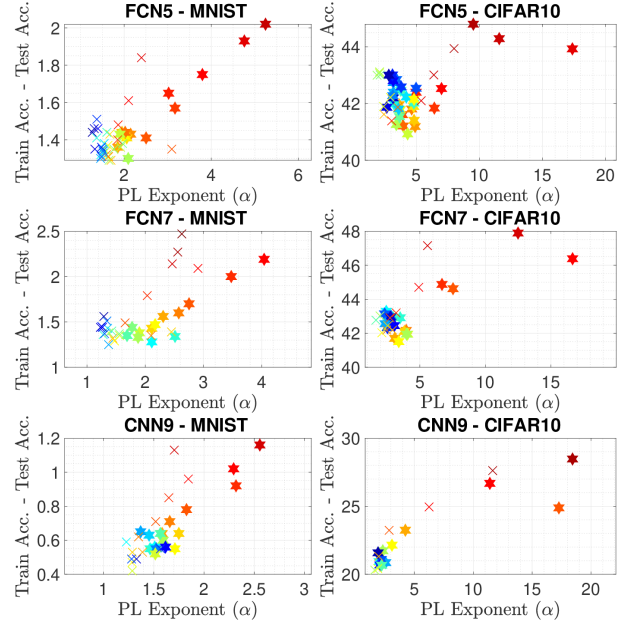

Figure 2: Lower tail exponents versus generalization gap. Different colors represent different step-sizes and different markers represent different batch-sizes.

4. Empirical results

4.1. Lower tail exponents of the transition kernel

For our experiments, we consider three architectures and two standard image classification datasets. In particular, we consider (i) a fully connected model with 5 layers (FCN5), (ii) a fully connected model with 7 layers (FCN7), and (iii) a convolutional model with 9 layers (CNN9); and two datasets (i) MNIST and (ii) CIFAR10. All models use the ReLU activation function and all are trained with constant step-size SGD, without weight-decay or momentum. Our code is implemented in `PyTorch` and executed on 5 GeForce GTX 1080 GPUs. For each architecture, we trained the networks with different step-sizes and batch-sizes, where we varied the step-size in the range $[0.002, 0.35]$ and the batch-size in the set $\{50, 100\}$. We trained all models until training accuracy reaches exactly 100%. For measuring training and test accuracies, we use standard training-test splits.

To estimate α, once training accuracy reaches 100%, we further run the algorithm for $m = 200$ additional iterations to obtain a trajectory $\{W_k\}_{k=1}^m$. Following Corollary 2, we assume *local homogeneity*, that is, the trajectory $\{W_k\}_{k=1}^m$ remains near the local minimum and each $W_{k+1} - W_k$ is approximately iid. Under this assumption, the second term in (7) can be ignored, hence, we compute the sequence $1/\|W_{k+1} - W_k\|$ for $k = 1, \dots, m-1$, and then fit a power-law (PL) distribution to this one dimensional set of observations, by using the `powerlaw` toolbox (Clauset et al., 2009). Figure 2 visualizes the results. In all configurations, we ob-
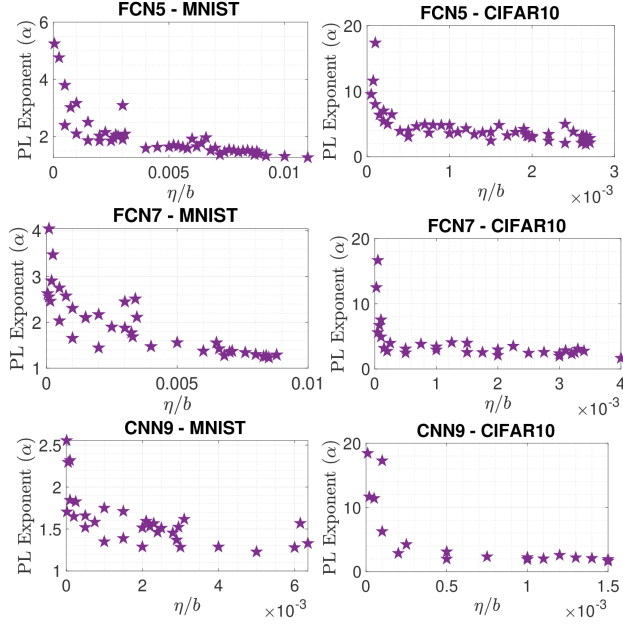
Figure 3: Lower tail exponents versus step-size/batch-size ratio (η/b). We apply the same significance test as in Figure 2: the p-value ranges from 10^{-9} to 10^{-3}.

serve that the results are in accordance with our theory: the estimated lower tail exponents α and the generalization gap exhibit significant correlations. This trend is even clearer for the CNN9 model, suggesting the geometry induced by the convolutional architecture results in a transition kernel for which the local homogeneity condition becomes more accurate, i.e., the TV term in Corollary 2 becomes smaller. Furthermore, for the significance of correlations in Figure 2, we compute the p-values under a linear model. The results show that the PL exponent and generalization error are always positively correlated and this correlation is significant: the p-value ranges from 10^{-12} to 10^{-4}.

It has been empirically demonstrated that the generalization performance can depend on the ratio of the step-size to the batch-size η/b (Jastrzebski et al., 2018). Next, we monitor the behavior of the lower tail exponent α with respect to the SGD hyperparameters: step-size and batch-size. Figure 3 shows the results. We observe that the local power-law behavior of SGD near the found local minimum also heavily depends on the ratio η/b, where we observe a clear monotonicity. This reveals an interesting behavior that the hyperparameters η and b modify the local lower tail exponent of the transition kernel, hence the effective dimension, which in turn determines the generalization error. This outcome also shows interesting similarities with the recent studies of Hodgkinson & Mahoney (2020); Gurbuzbalaban et al. (2020) that have shown that in online SGD (one pass regime with infinite data) the ratio η/b determines the "heaviness of the tails" of the stationary distribution of SGD. Figure 3 suggests that in the finite training data set-

ting (where Hodgkinson & Mahoney (2020); Gurbuzbalaban et al. (2020) are not applicable), another type of power-law behavior is still observed in the local exponent α, which also shows monotonic behavior with respect to η/b. We suspect that this monotonic behavior can be formally quantified, but we leave this as future work.

We finally note that estimating lower tail exponents accurately is a challenging task. For our estimator, it is known that larger fitted α values, e.g., greater than ≈ 5, are less reliable. However, since our empirical results involve running the same experiment multiple times for different hyperparameters that are fairly close to each other, the consistency and the clear trends in our results support our theoretical contributions, even if the estimations are to be inexact.

4.2. Correlations between types of tail exponents

Here, we shall discuss the relationship between the lower tail exponent in Corollary 2 and the tail exponent seen in Simsekli et al. (2019a); Şimşekli et al. (2020); Hodgkinson & Mahoney (2020); Gurbuzbalaban et al. (2020). While the two are not related in general, in practice, we expect them to be somewhat correlated, and we justify this through the model considered in Simsekli et al. (2019a).

Building on the perturbed GD model, Simsekli et al. (2019b) replaced the Gaussian updates by *heavy-tailed* α-stable distributed random variables, justified through the generalized central limit theorem. In this case, the model satisfies $W_{k+1} = W_k - \eta(\nabla f(W_k) + S_k(W_k))$, where $S_k(w) = (S_k^i(w))_{i=1}^n$ and each $S_k^i(w)$ is independent symmetric α-stable with scale $\sigma(w)$, that is, $S_k^i(w)$ has characteristic function $\varphi(t) = e^{-|\sigma(w)t|^\alpha}$. As $\eta \to 0^+$, this Markov chain behaves similarly to the stochastic differential equation $dW_t = -\nabla f(W_t)dt + \eta^{\frac{\alpha-1}{\alpha}}\sigma dL_t^\alpha$, where L_t^α is an α-stable Lévy process (see (Simsekli et al., 2019a) for details). If ∇f is bounded, the process satisfies the following two properties (Bogdan & Jakubowski, 2007, Lemma 3): (1) $\int_0^1 P_t(x, B_r(x))dt \sim c_1 r^\alpha$ as $r \to 0^+$; and (2) $\int_0^1 P_t(x, B_r(x))dt \sim c_2 r^{-\alpha}$ as $r \to \infty$. Therefore, the lower tail exponent in Corollary 2 and the tail exponent in (Simsekli et al., 2019a) are identical in this model.

Empirically, we find that while the two exponents are not identical, they do appear to correlate. In Figure 4, we follow the setup of Gurbuzbalaban et al. (2020) and assume that each layer of the neural network possesses a different tail-exponent, where these layer-wise tail-exponents are computed on averaged iterates under the assumption that they are distributed from an α-stable distribution. We use the same tail-exponent estimator (Mohammadi et al., 2015) as the one used in Gurbuzbalaban et al. (2020). Once the layer-wise indices are computed, we compute the median tail-exponent over the layers. As we can observe from Fig-
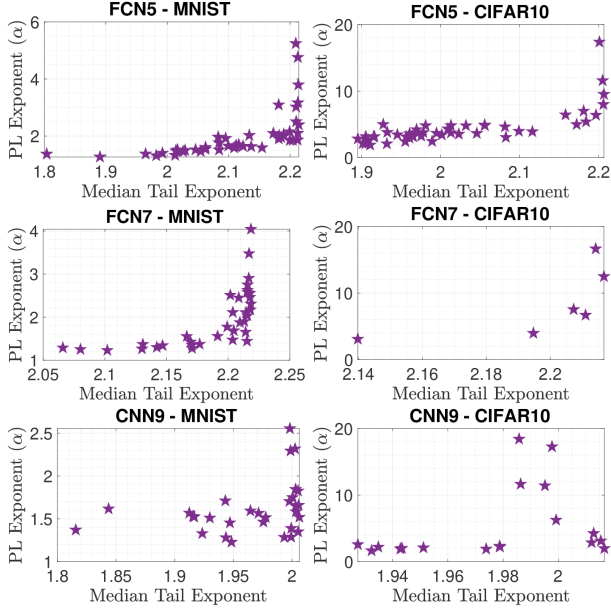
Figure 4: Lower tail exponents versus the tail-exponent estimates used in (Gurbuzbalaban et al., 2020).

ure 4, our lower tail exponent and the tail-exponent estimate show an overall strong correlation. More precisely, when we estimate the p-values under a linear model, we observe that the p-value ranges from 10^{-9} to 10^{-1}, where 10^{-1} corresponds to CNN9-CIFAR10 result, where the correlation is not as strong.

These results shed more light on the empirical observations provided in (Simsekli et al., 2019b), as the estimator in (Mohammadi et al., 2015) might be capturing the lower-tail behavior of the gradient noise, which seems to be the key factor according to our theory.

5. Conclusion

We have developed a theoretical framework for analyzing the generalization properties of stochastic optimizers by using their trajectories. We first proved generalization bounds based on the celebrated Fernique–Talagrand functional; and then, by using the Markovian structure of stochastic optimizers, we specialized these results to Markov processes. Using these results, we linked the accumulated generalization gap over an optimizer trajectory to a lower tail exponent in the transition kernel via a random walk approximation about a local minimum. This provides a discrete-time analogue of the work of Şimşekli et al. (2020) to more practical stochastic optimizer models. Finally, we supported our theory with empirical results on several simple neural network models, finding correlations between the lower tail exponent, generalization gap at the end of training, step-size/batch-size ratio, and upper tail exponents.

Our analysis raises a few unresolved questions. Firstly,

while suggested by the theory, it is untested whether the correlation between generalization gap and the lower tail exponent is universal, or holds only due to correlations between the lower tail exponent and the hyperparameters of the optimizer in our setup (which are known to affect performance). To assess this, one would need to be able to alter the lower tail exponent without changing any other commonly considered hyperparameters (e.g., step-size/batch-size). Furthermore, we have only considered fixed step sizes, and it is unclear whether our theory can be extended to hold under common step size schedules. Finally, while the subgaussian assumption on the data leads to the typical $\mathcal{O}(n^{-1/2})$ rate in Theorem 1, it is known that this rate is not reflected in real-world settings (Kaplan et al., 2020). Different distributional assumptions on the data could yield more accurate rates. We leave these three problems to be addressed in future work.

Acknowledgments

References

Antos, A., Kégl, B., Linder, T., and Lugosi, G. Data-dependent margin-based generalization bounds for classification. *Journal of Machine Learning Research*, 3(Jul): 73–98, 2002.

Arora, S., Ge, R., Neyshabur, B., and Zhang, Y. Stronger generalization bounds for deep nets via a compression approach. In *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)*, pp. 254–263. PMLR, 2018.

Arora, S., Cohen, N., Hu, W., and Luo, Y. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Asadi, A. R., Abbe, E., and Verdú, S. Chaining mutual information and tightening generalization bounds. *arXiv preprint arXiv:1806.03803*, 2018.

Audibert, J.-Y. and Bousquet, O. Combining PAC-Bayesian and generic chaining bounds. *Journal of Machine Learning Research*, 8(Apr):863–889, 2007.

Barsbey, M., Sefidgaran, M., Erdogdu, M. A., Richard, G., and Simsekli, U. Heavy tails in SGD and compressibility of overparametrized neural networks. In Beygelzimer,

A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021.

Bartlett, P., Foster, D. J., and Telgarsky, M. Spectrally-normalized margin bounds for neural networks. *arXiv preprint arXiv:1706.08498*, 2017.

Bartlett, P. L. and Long, P. M. Failures of model-dependent generalization bounds for least-norm interpolation. *arXiv preprint arXiv:2010.08479*, 2020.

Birdal, T., Lou, A., Guibas, L., and Simsekli, U. Intrinsic dimension, persistent homology and generalization in neural networks. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021.

Bogdan, K. and Jakubowski, T. Estimates of heat kernel of fractional Laplacian perturbed by gradient operators. *Communications in Mathematical Physics*, 271(1):179–198, 2007.

Borst, S., Dadush, D., Olver, N., and Sinha, M. Majorizing measures for the optimizer. *arXiv preprint arXiv:2012.13306*, 2020.

Boucheron, S., Lugosi, G., and Massart, P. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Bu, Y., Zou, S., and Veeravalli, V. V. Tightening mutual information-based bounds on generalization error. *IEEE Journal on Selected Areas in Information Theory*, 1(1): 121–130, 2020.

Camuto, A., Deligiannidis, G., Erdogdu, M. A., Gurbuzbalaban, M., Simsekli, U., and Zhu, L. Fractal structure and generalization properties of stochastic optimization algorithms. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021.

Cao, Y. and Gu, Q. Generalization error bounds of gradient descent for learning over-parameterized deep ReLU networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):3349–3356, Apr. 2020.

Chizat, L. and Bach, F. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conference on Learning Theory*, pp. 1305–1338. PMLR, 2020.

Clauset, A., Shalizi, C. R., and Newman, M. E. Power-law distributions in empirical data. *SIAM review*, 51(4): 661–703, 2009.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., and Reis, J. Modeling wine preferences by data mining from physicochemical properties. *Decision support systems*, 47(4): 547–553, 2009.

Cressie, N. *Statistics for spatial data*. John Wiley & Sons, 2015.

Dieuleveut, A., Durmus, A., and Bach, F. Bridging the gap between constant step size stochastic gradient descent and Markov chains. *arXiv preprint arXiv:1707.06386*, 2017.

Dudley, R. M. The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *Journal of Functional Analysis*, 1(3):290–330, 1967.

Farnia, F., Zhang, J., and Tse, D. A spectral approach to generalization and optimization in neural networks. In *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*, 2018.

Federer, H. *Geometric measure theory*. Springer, 2014.

Fernique, X. Régularité de processus gaussiens. *Inventiones Mathematicae*, 12(4):304–320, 1971.

Fernique, X. Regularité des trajectoires des fonctions aléatoires gaussiennes. In *Ecole d'Eté de Probabilités de Saint-Flour IV—1974*, pp. 1–96. Springer, 1975.

Fontaine, X., De Bortoli, V., and Durmus, A. Continuous and discrete-time analysis of stochastic gradient descent for convex and non-convex functions. *arXiv preprint arXiv:2004.04193*, 2020.

Gurbuzbalaban, M., Simsekli, U., and Zhu, L. The heavy-tail phenomenon in SGD. *arXiv preprint arXiv:2006.04740*, 2020.

Haghifam, M., Negrea, J., Khisti, A., Roy, D. M., and Dziugaite, G. K. Sharpened generalization bounds based on conditional mutual information and an application to noisy, iterative algorithms. *arXiv preprint arXiv:2004.12983*, 2020.

Hardt, M., Recht, B., and Singer, Y. Train faster, generalize better: Stability of stochastic gradient descent. In Balcan, M. F. and Weinberger, K. Q. (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1225–1234, New York, New York, USA, 20–22 Jun 2016. PMLR.

Hodgkinson, L. and Mahoney, M. W. Multiplicative noise and heavy tails in stochastic optimization. *arXiv preprint arXiv:2006.06293*, 2020.

Jastrzebski, S., Kenton, Z., Arpit, D., Ballas, N., Fischer, A., Bengio, Y., and Storkey, A. Three factors influencing minima in SGD. In *Artificial Neural Networks and Machine Learning*, 2018.

Jastrzebski, S., Szymczak, M., Fort, S., Arpit, D., Tabor, J., Cho, K., and Geras, K. The Break-Even Point on Optimization Trajectories of Deep Neural Networks. In *Proceedings of the 8th International Conference on Learning Representations (ICLR 2020)*, 2020.

Jastrzebski, S., Arpit, D., Astrand, O., Kerg, G. B., Wang, H., Xiong, C., Socher, R., Cho, K., and Geras, K. J. Catastrophic fisher explosion: Early phase Fisher matrix impacts generalization. In *Proceedings of the 38th International Conference on Machine Learning (ICML 2021)*, pp. 4772–4784, 2021.

Jiang, Y., Neyshabur, B., Mobahi, H., Krishnan, D., and Bengio, S. Fantastic generalization measures and where to find them. *arXiv preprint arXiv:1912.02178*, 2019.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Kong, L. and Tao, M. Stochasticity of deterministic gradient descent: Large learning rate for multiscale objective function. *Advances in Neural Information Processing Systems*, 33:2625–2638, 2020.

Li, J., Luo, X., and Qiao, M. On generalization error bounds of noisy gradient methods for non-convex learning. *arXiv preprint arXiv:1902.00621*, 2019.

Liu, L. and Xiao, Y. Hausdorff dimension theorems for self-similar Markov processes. *Probability and Mathematical Statistics*, 18, 1998.

Mandt, S., Hoffman, M., and Blei, D. A variational analysis of stochastic gradient algorithms. In *Proceedings of the 33rd International Conference on Machine Learning (ICML 2016)*, pp. 354–363. PMLR, 2016.

Martin, C. H. and Mahoney, M. W. Rethinking generalization requires revisiting old ideas: statistical mechanics approaches and complex learning behavior. *arXiv preprint arXiv:1710.09553*, 2017.

Martin, C. H. and Mahoney, M. W. Traditional and heavy-tailed self regularization in neural network models. *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*, 2019.

Martin, C. H. and Mahoney, M. W. Heavy-tailed Universality predicts trends in test accuracies for very large pre-trained deep neural networks. In *Proceedings of the 2020 SIAM International Conference on Data Mining*, pp. 505–513. SIAM, 2020.

Martin, C. H. and Mahoney, M. W. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *Journal of Machine Learning Research*, 2021a.

Martin, C. H. and Mahoney, M. W. Post-mortem on a deep learning contest: a Simpson's paradox and the complementary roles of scale metrics versus shape metrics. *arXiv preprint arXiv:2106.00734*, 2021b.

Martin, C. H., Peng, T. S., and Mahoney, M. W. Predicting trends in the quality of state-of-the-art neural networks without access to training or testing data. *Nature Communications*, 12(1):1–13, 2021.

Mattila, P. *Geometry of sets and measures in Euclidean spaces: fractals and rectifiability*. Number 44. Cambridge University Press, 1999.

Mohammadi, M., Mohammadpour, A., and Ogata, H. On estimating the tail index and the spectral measure of multivariate α-stable distributions. *Metrika*, 78(5):549–561, 2015.

Molchanov, I. *Theory of random sets*, volume 87. Springer, 2005.

Mou, W., Wang, L., Zhai, X., and Zheng, K. Generalization bounds of SGLD for non-convex learning: Two theoretical viewpoints. In *Conference on Learning Theory*, pp. 605–638. PMLR, 2018.

Negrea, J., Haghifam, M., Dziugaite, G. K., Khisti, A., and Roy, D. M. Information-theoretic generalization bounds for SGLD via data-dependent estimates. *arXiv preprint arXiv:1911.02151*, 2019.

Neu, G., Dziugaite, G. K., Haghifam, M., and Roy, D. M. Information-theoretic generalization bounds for stochastic gradient descent. In Belkin, M. and Kpotufe, S. (eds.), *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134, pp. 3526–3545. PMLR, 15–19 Aug 2021.

Neyshabur, B., Bhojanapalli, S., McAllester, D., and Srebro, N. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*, pp. 5947–5956, 2017.

Orvieto, A. and Lucchi, A. Continuous-time models for stochastic optimization algorithms. *arXiv preprint arXiv:1810.02565*, 2018.

Panigrahi, A., Somani, R., Goyal, N., and Netrapalli, P. Non-gaussianity of stochastic gradient noise. *arXiv preprint arXiv:1910.09626*, 2019.

Pensia, A., Jog, V., and Loh, P.-L. Generalization error bounds for noisy, iterative algorithms. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pp. 546–550. IEEE, 2018.

Ripley, B. D. The second-order analysis of stationary point processes. *Journal of Applied Probability*, 13(2):255–266, 1976.

Russo, D. and Zou, J. How much does your data exploration overfit? Controlling bias via information usage. *IEEE Transactions on Information Theory*, 66(1):302–323, 2019.

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

Sheu, S.-J. Some estimates of the transition density of a nondegenerate diffusion Markov process. *The Annals of Probability*, pp. 538–561, 1991.

Simsekli, U., Gürbüzbalaban, M., Nguyen, T. H., Richard, G., and Sagun, L. On the heavy-tailed theory of stochastic gradient descent for deep neural networks. *arXiv preprint arXiv:1912.00018*, 2019a.

Simsekli, U., Sagun, L., and Gurbuzbalaban, M. A tail-index analysis of stochastic gradient noise in deep neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*, pp. 5827–5837. PMLR, 2019b.

Şimşekli, U., Sener, O., Deligiannidis, G., and Erdogdu, M. A. Hausdorff dimension, stochastic differential equations, and generalization in neural networks. *arXiv preprint arXiv:2006.09313*, 2020.

Sokolić, J., Giryes, R., Sapiro, G., and Rodrigues, M. R. Generalization error of deep neural networks: Role of classification margin and data structure. In *2017 International Conference on Sampling Theory and Applications (SampTA)*, pp. 147–151. IEEE, 2017.

Talagrand, M. Majorizing measures: the generic chaining. *Annals of Probability*, 24(3):1049–1103, 1996.

Talagrand, M. Majorizing measures without measures. *Annals of Probability*, pp. 411–417, 2001.

Talagrand, M. *Upper and lower bounds for stochastic processes: modern methods and classical problems*, volume 60. Springer Science & Business Media, 2014.

Vapnik, V. N. and Chervonenkis, A. Y. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity*, pp. 11–30. Springer, 2015.

Xiao, Y. Random fractals and Markov processes. *Mathematics Preprint Archive*, 2003(6):830–907, 2003.

Xing, C., Arpit, D., Tsirigotis, C., and Bengio, Y. A walk with SGD. *arXiv preprint arXiv:1802.08770*, 2018.

Xu, A. and Raginsky, M. Information-theoretic analysis of generalization capability of learning algorithms. *arXiv preprint arXiv:1705.07809*, 2017.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

A. Hausdorff dimension

Here, we shall review facts about the Hausdorff dimension. The Hausdorff measure $\mathscr{H}^m$ on $\mathbb{R}^d$ is defined on a set $E \subseteq \mathbb{R}^d$ by

$$\mathscr{H}^m(E) = \lim_{\delta \to 0^+} \inf_{\substack{E \subseteq \bigcup_j S_j \\ \operatorname{diam}(S_j) \leqslant \delta}} \sum_j \alpha(m) 2^{-m} \operatorname{diam}(S_j)^m,$$

where the infimum is taken over countable covers S_δ of E by non-empty subsets of $\mathbb{R}^d$ with diameter not exceeding δ, and $\alpha(m) = \frac{\pi^{n/2}}{\Gamma(n/2+1)}$ is the volume of the unit m-sphere. The following facts are fundamental (Federer, 2014, §2.10.2, §2.10.35) : (1) if $\mathscr{H}^m(E) < \infty$, then $\mathscr{H}^l(E) = 0$ for any $l > m$; and (2) on $\mathbb{R}^d$, $\mathscr{H}^d \equiv \mathscr{L}^d$ the d-dimensional Lebesgue measure.

The first fact implies the existence of the *Hausdorff dimension,* which is defined for a set $E \subseteq \mathbb{R}^d$ by $\dim_{\mathrm{H}}(E) = \inf\{m : \mathscr{H}^m(E) = 0\}$. The second fact implies that any set $E \subset \mathbb{R}^d$ with $\dim_{\mathrm{H}}(E) < d$ has zero Lebesgue measure; but, importantly, the converse is not true. In differential geometry, the Hausdorff dimension is useful for identifying the dimension of submanifolds. However, it is also of significant value in the study of fractal sets, providing a measurement of "clustering" in space. One can intuit this from the definition, but it is perhaps best seen through examples. In Figure 1, a a Lévy process with sample paths possessing Hausdorff dimension 1.5 is compared to Brownian motion, whose sample paths have Hausdorff dimension 2. The Lévy process, with the smaller Hausdorff dimension, exhibits dynamics with tighter clusters separated by large jumps.

Ahlfors lower-regularity plays a key role in Corollary 2. A more commonly considered definition is Ahlfors regularity itself.

Definition 1 (Ahlfors regular). A set $W \subset \mathbb{R}^D$ is α-Ahlfors regular if there exists a measure μ on W and $c_1, c_2, r_0 > 0$ such that

$$c_1 r^\alpha \leqslant \mu(B_r(w)) \leqslant c_2 r^\alpha, \qquad \text{for all } w \in W, \, 0 < r < r_0.$$

If only the lower (upper) inequalities are satisfied, W is said to be α-Ahlfors lower-regular (upper-regular).

Ahlfors regularity is often assumed to equate several notions of fractal dimension, such as in (Şimşekli et al., 2020). For example, the lower and upper Minkowski dimensions of a set W are given by

$$\underline{\dim}_M(W) = \liminf_{r \to 0^+} \frac{\log N_r(W)}{|\log r|}, \quad \overline{\dim}_M(W) = \liminf_{r \to 0^+} \frac{\log N_r(W)}{|\log r|},$$

respectively, where $N_r(W)$ denotes the smallest number of balls of radius r needed to cover W. The following holds.

Theorem ((Mattila, 1999), Theorem 5.7). *For any α-Ahlfors regular set W, $\underline{\dim}_M(W) = \overline{\dim}_M(W) = \dim_H(W) = \alpha$.*

To link Hausdorff dimension to continuous-time optimization, we rely on Xiao (2003, Theorem 4.2), restated below.

Theorem ((Xiao, 2003), Theorem 4.2). *Let X_t be a continuous-time Markov process in $\mathbb{R}^D$ with transition kernel $P_t(x, \mathrm{d}y)$ satisfying $P_t(x, B_r(x)) \geqslant K$ for sufficiently large $r > 0$ and $c^{-1} P_t(0, B_r(0)) \leqslant P_t(x, B_r(x)) \leqslant c P_t(0, B_r(0))$ for all $r > 0$ and some $c > 0$. Then the Hausdorff dimension of $\{X_t\}_{t \in [0,1]}$ is*

$$\alpha = \sup \left\{ \gamma \geqslant 0 : \lim_{r \to 0^+} r^{-\gamma} \int_0^1 P_t(0, B_r(0)) \mathrm{d}t < +\infty \right\}$$

$$= \sup \left\{ \gamma \geqslant 0 : \int_0^1 \mathbb{E}(|X_t|^{-\gamma}) \mathrm{d}t < +\infty \right\}.$$

B. Variance and clustering in the Fernique–Talagrand functional

Now that two generalization bounds in Theorem 2 and Corollary 2 have been obtained involving the dynamics of the stochastic optimizer, we shall briefly discuss how two properties of the trajectory — variance and clustering — play a critical role in the Fernique–Talagrand functional.

First, we discuss the influence of "variance," or the average size of fluctuations of the stochastic optimizer. Drawing from Theorem 2, we consider a continuous-time stochastic optimization model in the form of a stochastic differential equation $\mathrm{d}W_t = \mu(W_t)\mathrm{d}t + \sigma(W_t)\mathrm{d}B_t$, where μ typically involves the gradient of the empirical risk $\mathcal{R}_n$. We assume
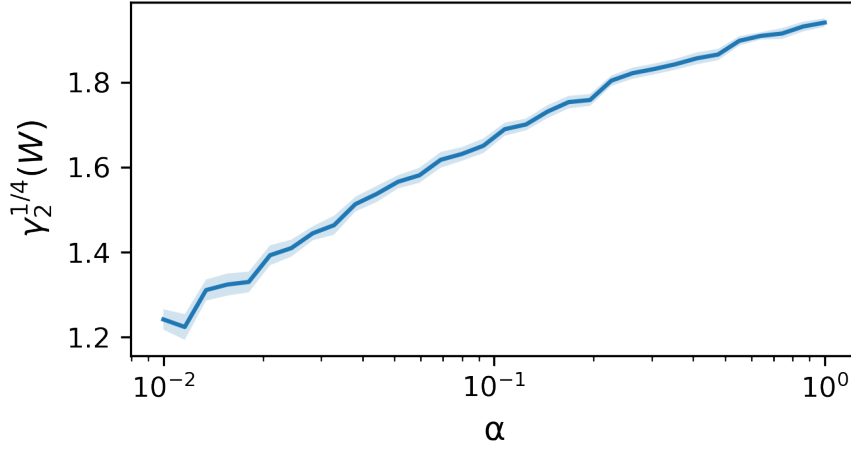
Figure 5: Lower tail exponents versus normalized Fernique–Talagrand functional $\gamma_2^{1/4}(W)$ for $\rho = 1/4$, averaged over 100 runs with 95% confidence intervals shaded.

that μ, σ are bounded with bounded second derivatives, and $\sigma(x)$ has non-zero singular values for all $x \in \mathbb{R}^D$. The Aronson estimates (Sheu, 1991) imply that for some constant $K > 0$ and "minimal variance" $\lambda > 0$, $P_t(x, E) \geqslant K t^{-D/2} \int_E \exp(-\|x - y\|^2/(2\lambda t)) \mathrm{d}y$. This, together with Theorem 2, implies the existence of constants $K_1, K_2 > 0$ independent of T, λ such that (see Appendix D)

$$\mathbb{E}\gamma_2^\rho(\{W_t\}_{t \in [0,T]}) \leqslant \frac{K_1}{\rho} \int_0^\rho \sqrt{K_2 - D \log r - \log J_{\rho,T}(\tfrac{1}{2\lambda}, D)} \mathrm{d}r, \tag{8}$$

where $J_{\rho,T}(a, D) = \frac{1}{T} \int_0^1 \int_{1/T}^\infty v^{D/2-1} s^{D/2-2} e^{-asv\rho^2} \mathrm{d}s \mathrm{d}v$ is monotone decreasing in a and monotone increasing in D. It is important to note that λ will generally increase with the magnitude of μ, σ, and hence with the size of the stochastic gradient. Therefore, the normalized FT functional should increase monotonically with the variance of the fluctuations of the optimizer. Consequently, the *sharpness* of an optimum should be reflected in γ_2^ρ: in a "flat" neighborhood, the variance of fluctuations in the optimizer is small, and hence γ_2^ρ will be smaller.

Now, we move on to discuss how the α parameter in Corollary 2 can be interpreted as a measure of "clustering" of the optimizer trajectory. To accomplish this, we invoke the K-function of Ripley (1976), a commonly used spatial statistic to determine spatial inhomogeneity. For a point process N on $\mathbb{R}^D$, the K-function is defined for each $r > 0$ as the ratio of the expected number of points within distance r of any randomly chosen point, and the average density of points. Letting $W = \{w_1, \ldots, w_n\}$ denote a realization of a point process, its K-function can be estimated consistently by Cressie (2015, eqn. 8.2.18):

$$\hat{K}(r) = \frac{\mathrm{diam}(W)}{n} \sum_{\substack{i,j=1 \\ i \neq j}}^n \mathbb{1}\{\|w_i - w_j\| \leqslant r\}, \qquad r > 0. \tag{9}$$

The K-function of a homogeneous Poisson process on $\mathbb{R}^D$ with constant intensity λ is given by $K(r) = \pi^{D/2} r^D / \Gamma(1 + \frac{D}{2})$ for $r > 0$ (Cressie, 2015, eqn. 8.3.34). Note that this function is independent of the intensity. Therefore, deviations in the K-function from an $\mathcal{O}(r^D)$ growth rate suggest spatial inhomogeneity, and therefore points which exhibit spatial clustering. Treating an optimizer trajectory $W_1, \ldots, W_n$ as the realization of a point process, we can interpret the K-function through (9) and measure spatial clustering of the trajectory. Following the assumptions in Corollary 2 which define α, we can assume that $\|W_i - W_j\| \leqslant r$ with probability approximately $c_{|i-j|} r^\alpha$ for $0 < r < r_0$, where $c_1, \ldots, c_n$ are positive constants. Under this assumption, $\hat{K}(r) \approx C \mathrm{diam}(W) r^\alpha$ for $0 < r < r_0$ and some constant $C > 0$. Therefore, we would expect $\alpha \ll D$ to be an indication that the optimization exhibits significant spatial clustering. By Corollary 2, this circumstance should coincide with a smaller Fernique–Talagrand functional, and therefore improved generalization.

C. Lower tail exponent and the Fernique–Talagrand functional

To further demonstrate the relationship between the lower tail exponent of a Markov transition kernel and the Fernique–Talagrand functional over its sample path, we consider an extension of the setup in Figure 1 and measure $\gamma_2^\rho(\{W_k\}_{k=0}^m)$ for random walks with prescribed lower tail exponents.

Consider a Markov chain W_k on $\mathbb{R}^2$ defined as follows: starting from W_0, let

$$W_{k+1} = W_k + (\cos(U_k), \sin(U_k))Z_k,$$

where for each $k = 0, 1, \ldots$, $U_k \sim \mathcal{U}(-\pi, \pi)$, and $Z_k \sim \beta'(\alpha, \beta)$ are independent, and $\beta'(\alpha, \beta)$ is the beta prime distribution with density

$$p(x) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1+x)^{-\alpha-\beta}, & \text{if } x > 0, \\ 0, & \text{otherwise.} \end{cases}$$

The lower tail exponent of this process, as defined in Section 3.3, is precisely α. To remove the effect of variance, the chain is normalized as $\bar{W}_k = W_k / S(\{W_l\}_{l=0}^k)$, where $S(W)$ is the coordinate-wise standard deviation of W. For fixed $W_0 = 0$, $\beta = 3.5$ and varying $\alpha \in [10^{-2}, 1]$, we plot the normalized Fernique–Talagrand functional $\gamma_2^{1/4}(\{\bar{W}_k\}_{k=0}^{100})$ (with $\rho = 1/4$) of the first 100 iterates of W_k, averaged over 100 runs. The result is shown in Figure 5. As expected, the FT functional grows with α. However, unlike the behaviour suggested in Corollary 2, the FT functional appears to grow like $\log \alpha$ for small α.

D. Lower tail exponent as intrinsic dimension

Here, we shall discuss the relationship between the exponent α and the dimension D. The most commonly considered discrete-time model for a stochastic optimizer is the perturbed gradient descent (GD) model, which satisfies $W_{k+1} = W_k - \gamma(\nabla f(W_k) + Z_k)$, where Z_k is a Gaussian random vector with zero mean and constant covariance matrix Σ. In the neighborhood of a local minimum w^*, $\nabla f(W_k) \approx 0$, and hence the perturbed gradient descent model resembles a Gaussian random walk $W_{k+1} = W_k + \gamma Z_k$. In this case, if D denotes the ambient dimension, the k-step transition kernel becomes that of a D-dimensional multivariate normal distribution with covariance matrix $k\Sigma$:

$$P^k(x, E) = (2\pi k)^{-\frac{D}{2}} |\det \Sigma|^{-\frac{1}{2}} \int_E \exp\left(-\tfrac{1}{2k}(y-x)^\top \Sigma^{-1}(y-x)\right) dy.$$

Since Σ is necessarily positive-definite, letting σ_1 and σ_D denote the largest and smallest singular values of Σ respectively, for any $x \in \mathbb{R}^D$ and set $E \subset \mathbb{R}^D$,

$$\frac{1}{(2\pi k\sigma_1)^{D/2}} \int_E \exp\left(-\frac{\|y-x\|^2}{2k\sigma_D}\right) dy \leqslant P^k(x, E) \leqslant \frac{1}{(2\pi k\sigma_D)^{D/2}} \int_E \exp\left(-\frac{\|y-x\|^2}{2k\sigma_1}\right) dy. \tag{10}$$

Applying Lemma 4 to (10), we see that for the perturbed GD model, α in Corollary 2 is precisely the ambient dimension D.

E. Direct estimation of the Fernique–Talagrand functional

An attractive feature of the FT functional is the availability of a low-degree polynomial time approximation algorithm for $\gamma_2(W, d)$ when W is a finite subset of $\mathbb{R}^D$. In particular, (Borst et al., 2020) shows that $\gamma_2(W, d)$ is computable to ϵ-accuracy in $\mathcal{O}((|W|^{1+\omega} + D|W|^3)\log(|W|D/\epsilon))$ time, and $\omega \leqslant 2.373$ is the matrix multiplication exponent. To our knowledge, this functional is the only object which tightly bounds (1), up to constants, and is approximable in polynomial time. This is especially attractive for our purposes, providing an effective measure of generalization performance and exploratory capacity, which does *not* require access to any test data. A general procedure to estimate $\gamma_2(W, d)$ is presented in Algorithm 1. To perform the optimization in the final step, any off-the-shelf nonlinear optimization procedure will suffice, including gradient descent. Indeed, since the objective $I_W(p)$ happens to be convex in p (see (Borst et al., 2020)), any local minimizer p^* with subgradient $\partial I_W(p^*) = 0$ will satisfy $\gamma_2(W, d) = I_W(p^*)$.
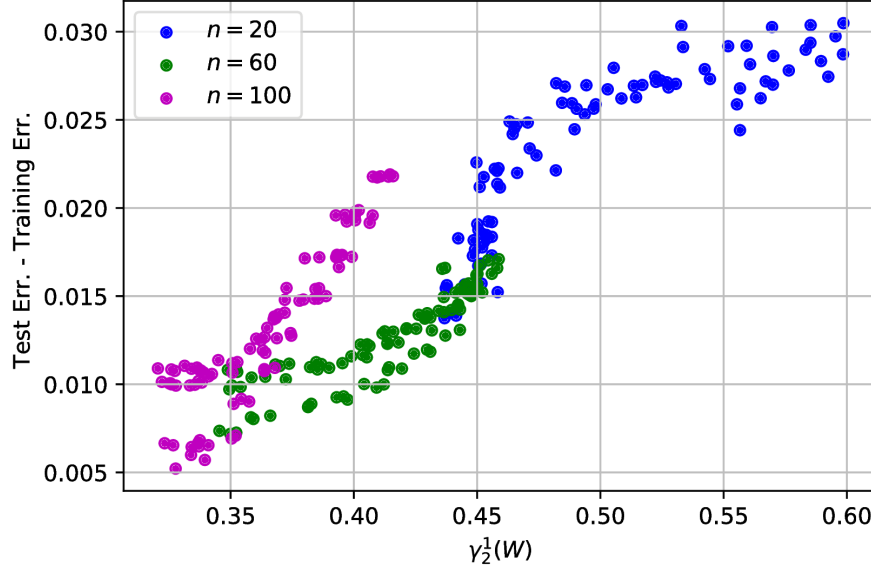
Figure 6: Normalized Fernique–Talagrand functional $\gamma_2^1(W)$ versus generalization gap. Different colors denote different batch-sizes.

Algorithm 1 Fernique–Talagrand functional

Input: A set $W = \{w_1, \ldots, w_n\}$, and metric $d(w, w')$ on W
Output: An estimate of $\gamma_2(W, d)$.
1: Compute Gram matrix $G = (d(w_i, w_j))_{i,j=1}^n$
2: **for** $i = 1, \ldots, n$
3:     sort $(G_{ij})_{j=1}^n$ in ascending order to obtain $(\tilde{G}_{ij})_{j=1}^n$ and
        sorted indices $(\iota_{ij})_{j=1}^n$
4: **end for**
5: **function** $I_W(p)$
6:     **for** $i = 1, \ldots, n$
7:     $I_W^i(p) = \sum_{j=0}^{n-1}(G_{i,j+1} - G_{ij})\sqrt{|\log \sum_{k=0}^j p_{\iota_{ik}}|}$
8:     **return** $\max\{I_W^1(p), \ldots, I_W^n(p)\}$
9: **end function**
10: **return** $\min_{z \in \mathbb{R}^n} I_W(\text{softmax}(z))$

Unfortunately, this approach becomes more challenging in high dimensions, where the optimization task becomes more difficult to solve to reasonable accuracy. Hence, we restrict our discussion in this section to smaller models. Our model of choice is a three-layer fully-connected neural network with 20 hidden units, applied to the least-squares regression task on the Wine Quality UCI dataset (Cortez et al., 2009). Models are trained from the same (random) initialization for 30 epochs (*before* reaching 100% training accuracy) using SGD with constant step size $\eta \in \{0.01, 0.005, 0.001\}$, batch size $b \in \{20, 50, 100\}$, weight decay parameter $\lambda \in [10^{-4}, 5 \times 10^{-4}, 10^{-3}]$, and added zero-mean Gaussian noise to the input data with variance σ^2 for $\sigma \in \{0, 0.05, 0.1\}$. In Figure 6, for each model, we plot test error at the end of training against the estimated normalized FT functional (using Algorithm 1) from the last 50 iterates of training. The most profound difference in trends is seen with varying batch size. Nevertheless, as expected, the FT functional shows a strong correlation to test error.

F. Proofs of our main results

F.1. Proof of Theorem 1 and Corollary 1

The total mutual information is valuable as it precisely defines the degree to which we may decouple two random elements, as shown in the following lemma.

Lemma 1. *For any Borel set B, $\mathbb{P}_{X,Y}(B) \leq \exp I_\infty(X,Y) \cdot \mathbb{P}_X \otimes \mathbb{P}_Y(B)$.*

Proof. The proof relies on the data processing inequality for α-Renyi divergence, which implies that for any $\alpha > 1$:

$$D_\alpha(\mathbf{B}(\mathbb{P}_{X,Y}(B)) \| \mathbf{B}(\mathbb{P}_X \otimes \mathbb{P}_Y(B))) \leq I_\alpha(X,Y),$$

where $\mathbf{B}(p)$ is a Bernoulli measure with success probability p. Therefore, letting $p = \mathbb{P}_{X,Y}(B)$ and $q = \mathbb{P}_X \otimes \mathbb{P}_Y(B)$, as $\alpha \to \infty$,

$$\begin{aligned}
D_\alpha(\mathbf{B}(p) \| \mathbf{B}(q)) &= \frac{1}{\alpha - 1} \log \left(q \frac{p^\alpha}{q^\alpha} + (1 - q) \frac{(1-p)^\alpha}{(1-q)^\alpha} \right) \\
&= \log \left(\left(q \frac{p^\alpha}{q^\alpha} + (1 - q) \frac{(1-p)^\alpha}{(1-q)^\alpha} \right)^{1/(\alpha-1)} \right) \to \log \max \left\{ \frac{p}{q}, \frac{1-p}{1-q} \right\}.
\end{aligned}$$

Therefore, it follows that $p/q \leq \sup_\alpha \exp I_\alpha(X,Y) = \exp I_\infty(X,Y)$. $\qquad\square$

Proof of Theorem 1. In the sequel, we shall let $K > 0$ denote a universal constant, not necessarily the same in each appearance. Let $d_\rho(w, w') = \min\{\rho, \|x - y\|\}$, and consider the alternative *generic chaining* functional $\hat{\gamma}$ given by

$$\hat{\gamma}_2(W, d_\rho) = \inf \sup_{w \in W} \sum_{k=1}^\infty 2^{n/2} d_\rho(w, T_k),$$

where the infimum is taken over all sequences of subsets $\{T_k\}_{k=1}^\infty$ such that $|T_k| \leq N_k$ (where $N_0 = 1$ and $N_k = 2^{2^k}$ otherwise). By Talagrand (2001, Theorem 1.1), there exists a universal constant $K > 0$ such that $\hat{\gamma}_2(W, d_\rho) \leq K \gamma_2^\rho(W)$. The proof proceeds in a similar fashion to Talagrand (2014, Theorem 2.2.27). For each k, let T_k^W be a set such that $|T_k^W| \leq N_k$ and

$$\sup_{w \in W} \sum_{k=1}^\infty 2^{k/2} d_\rho(w, T_k^W) \leq 2 \hat{\gamma}_2(W, d_\rho).$$

To construct an increasing sequence of subsets, let $U_k^W = \bigcup_{m \leq k} T_m^W$, so that $U_0^W = T_0^W$ and $|U_k^W| \leq 2N_k$. Now, let $Y_w = n^{1/2}[\mathcal{R}_n(w) - \mathcal{R}(w)] / \max\{B/\rho, L\}$, so that $\mathbb{E}Y_w = 0$, and by Hoeffding's inequality, for any $u > 0$ and $w, w' \in \mathbb{R}^D$,

$$\mathbb{P}(|Y_w - Y_{w'}| > u d_\rho(w, w')) \leq 2 \exp(-\tfrac{1}{2} u^2). \tag{11}$$

For $u > 0$, consider the event $\Omega(u)$ where

$$\forall k \geq 1, \forall w, w' \in U_k^W, |Y_w - Y_{w'}| \leq 2(2^{k/2} + u) d_\rho(w, w').$$

For each k, let $\tilde{U}_k^W$ be an independent copy of U_k^W. Then, for $M = \exp I_\infty(X, W)$,

$$\begin{aligned}
\mathbb{P}(\Omega \backslash \Omega(u)) &\leq \sum_{k=1}^\infty \mathbb{E}\mathbb{P}(|Y_w - Y_{w'}| \leq 2(2^{k/2} + u) d_\rho(w, w'), \forall w, w' \in U_k^W \mid U_k^W) \\
&\leq M \sum_{k=1}^\infty \mathbb{E}\mathbb{P}(|Y_w - Y_{w'}| \leq 2(2^{k/2} + u) d_\rho(w, w'), \forall w, w' \in \tilde{U}_k^W \mid \tilde{U}_k^W) \\
&\leq M \sum_{k=1}^\infty 2^{2(2^k+1)+1} \exp(-2(2^k + u^2)) \\
&\leq M \sum_{k=1}^\infty \exp(2(2^k + 1) + 1) \exp(-2(2^k + u^2)) \\
&\leq KM \exp(-2u^2).
\end{aligned}$$

For each element $w \in W$, we define a sequence of (random) integers $k(w, q)$ in the following inductive manner. First, let $k(w, 0) = 0$, and for each $q \geqslant 1$, we define

$$k(w, q) = \inf \left\{ k : k \geqslant k(w, q-1), d_\rho(w, U_k^W) \leqslant \frac{1}{2} d_\rho(w, U_{k(w,q-1)}^W) \right\}.$$

Now, consider elements $\pi_q(w) \in U_{k(w,q)}^W$ satisfying $d_\rho(w, \pi_q(w)) = d_\rho(w, U_{k(w,q)}^W)$. By induction, we find that $d_\rho(w, \pi_q(w)) \leqslant \rho 2^{-q}$. Furthermore, when $\Omega(u)$ occurs, since $\pi_q(w) \in U_{k(w,q)}^W$ and $\pi_{q-1}(w) \in U_{k(w,q-1)}^W \subset U_{k(w,q)}^W$, it follows that

$$|Y_{\pi_q(w)} - Y_{\pi_{q-1}(w)}| \leqslant 2(2^{k(w,q)/2} + u)d_\rho(\pi_q(w), \pi_{q-1}(w)).$$

Therefore, letting $w_0 \in T_0^W$, under $\Omega(u)$,

$$
\begin{aligned}
|Y_w - Y_{w_0}| &\leqslant \sum_{q=1}^{\infty} |Y_{\pi_q(w)} - Y_{\pi_{q-1}(w)}| \\
&\leqslant \sum_{q=1}^{\infty} 2(2^{k(w,q)/2} + u)d_\rho(\pi_q(w), \pi_{q-1}(w)) \\
&\leqslant \sum_{q=1}^{\infty} 2(2^{k(w,q)/2} + u)d_\rho(w, \pi_q(w)) + \sum_{q=1}^{\infty} 2(2^{k(w,q)/2} + u)d_\rho(w, \pi_{q-1}(w)).
\end{aligned}
$$

By construction,

$$\sum_{q=1}^{\infty} 2^{k(w,q)/2} d_\rho(w, \pi_q(w)) \leqslant \sum_{k=0}^{\infty} 2^{k/2} d_\rho(w, T_n^W) \leqslant 2\hat{\gamma}_2(W, d_\rho).$$

Similarly, by the definition of $k(w, q)$, it follows that $d_\rho(w, U_{k(w,q)}^W) \leqslant \frac{1}{2} d_\rho(w, U_{k(w,q-1)}^W)$ and $d_\rho(w, U_{k(w,q)-1}^W) \geqslant \frac{1}{2} d_\rho(w, U_{k(w,q-1)}^W)$. Therefore,

$$
\begin{aligned}
\sum_{q=1}^{\infty} 2^{k(w,q)/2} d_\rho(w, \pi_{q-1}(w)) &\leqslant 2 \sum_{q=1}^{\infty} 2^{k(w,q)/2} d_\rho(w, T_{k(w,q)-1}^W) \\
&\leqslant 4 \sum_{k=0}^{\infty} 2^{k/2} d_\rho(w, T_n^W) \\
&\leqslant 8\hat{\gamma}_2(W, d_\rho).
\end{aligned}
$$

Finally, we have that $\sum_{q=1}^{\infty} d_\rho(\pi_q(w), w) \leqslant \rho \sum_{q=1}^{\infty} 2^{-q} = \rho$, and $\sum_{q=1}^{\infty} d_\rho(\pi_{q-1}(w), w) \leqslant 2\rho$. Therefore, when $\Omega(u)$ occurs, for any $w \in W$,

$$|Y_w - Y_{w_0}| \leqslant K(\hat{\gamma}_2(W, d_\rho) + \rho u) \leqslant K(\gamma_2^\rho(W) + \rho u).$$

Altogether, this implies that

$$\mathbb{P}\left(\sup_{w, w' \in W} |Y_w - Y_{w'}| > K(\gamma_2^\rho(W) + \rho u) \right) \leqslant K \exp(I_\infty(X, W) - 2u^2).$$

Since $\mathbb{E}Y_w = 0$, $\sup_{w \in W} |Y_w| \leqslant \sup_{w, w' \in W} |Y_w - Y_{w'}|$, and (4) follows. We would now like to apply Xu & Raginsky (2017, Lemma 1) to show (5). To do so, it is necessary to show that $\sup_{w \in \tilde{W}} |\mathcal{R}_n(w) - \mathcal{R}(w)|$ is subgaussian, where $\tilde{W}$ is an independent copy of W. Recall that a random variable X is σ-subgaussian if $\log \mathbb{E} \exp(\lambda(X - \mathbb{E}X)) \leqslant \lambda^2 \sigma^2/2$. First, consider the case where $\tilde{W} = \mathcal{W}$ is a deterministic set of weights, and for brevity, let $\bar{\mathcal{R}}_n(w) = \mathcal{R}_n(w) - \mathcal{R}(w)$. In this case, one may apply McDiarmid's inequality to $\sup_{w \in \mathcal{W}} |\bar{\mathcal{R}}_n(w)| = f(X_1, \ldots, X_n)$, where

$$f(x_1, \ldots, x_n) = \sup_{w \in \mathcal{W}} \left| \frac{1}{n} \sum_{i=1}^{n} \ell(x_i, w) - \mathbb{E}\ell(X_i, w) \right|.$$

Since ℓ is bounded, it follows that $|\ell(x_i, w) - \ell(y_i, w)| \leqslant 2B$ for any x, y and $w \in \mathcal{W}$. Therefore,

$$|f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \leqslant \frac{2B}{n}.$$

Applying McDiarmid's inequality reveals that

$$\mathbb{P}\left(\left|\sup_{w\in\mathcal{W}}|\bar{\mathcal{R}}_n(w)| - \mathbb{E}\sup_{w\in\mathcal{W}}|\bar{\mathcal{R}}_n(w)|\right| > u\right) \leqslant 2\exp\left(-\frac{nu^2}{2B^2}\right).$$

Since the bound does not depend on $\mathcal{W}$, and $\{X_i\}_{i=1}^n$, $\tilde{W}$ are independent, we can condition on $\tilde{W}$ and apply this bound to find that

$$\mathbb{P}\left(\left|\sup_{w\in\tilde{W}}|\bar{\mathcal{R}}_n(w)| - \mathbb{E}\sup_{w\in\tilde{W}}|\bar{\mathcal{R}}_n(w)|\right| > u\right) \leqslant 2\exp\left(-\frac{nu^2}{2B^2}\right),$$

which, by Boucheron et al. (2013, Theorem 2.1), implies that $\sup_{w\in\tilde{W}}|\bar{\mathcal{R}}_n(w)|$ is $(4B/\sqrt{n})$-subgaussian. Applying Xu & Raginsky (2017, Lemma 1),

$$\mathbb{E}\sup_{w\in W}|\mathcal{R}_n(w) - \mathcal{R}(w)| \leqslant \mathbb{E}\sup_{w\in\tilde{W}}|\mathcal{R}_n(w) - \mathcal{R}(w)| + \sqrt{\frac{32B^2 I_1(X,W)}{n}}. \tag{12}$$

Using (11), an application of Talagrand (1996, Proposition 2.4) shows that

$$\mathbb{E}\sup_{w\in\tilde{W}}|Y_w| \leqslant \mathbb{E}\left[\mathbb{E}_{\tilde{W}}\sup_{w\in\tilde{W}}|Y_w|\right] \leqslant K\mathbb{E}\gamma_2^\rho(\tilde{W}) = K\mathbb{E}\gamma_2^\rho(W), \tag{13}$$

where $\mathbb{E}_{\tilde{W}}$ denotes conditional expectation, conditioned on $\tilde{W}$. The result now follows by combining (12) and (13). $\square$

Remark 1. Unfortunately, there has been little work on providing good estimates on the universal constant K. To our knowledge, only the original work of Fernique reports constants: from Fernique (1975, pg. 74), we find that

$$K \leqslant 30\sqrt{8}\left[\sqrt{2 + \frac{1}{\log 2}} + \frac{e}{2\sqrt{\log 2}}\right] \approx 296,$$

which is likely much larger than necessary.

Proof of Corollary 1. Define a probability measure μ with support on W by $\mu(E) = \mathcal{H}^\alpha(W \cap E)/\mathcal{H}^\alpha(W)$. By assumption, $\mu(B_r(w)) \geqslant (C_\rho r)^\alpha$ for $0 \leqslant r < \rho$ and any $w \in W$. Therefore,

$$\gamma_2^\rho(W) \leqslant \sup_{w\in W}\frac{1}{\rho}\int_0^\rho \sqrt{\log\frac{1}{\mu(B_r(w))}}\mathrm{d}r \leqslant (C_\rho^{-1}\sqrt{\alpha})\cdot\frac{1}{\rho}\int_0^{\rho C_\rho}\sqrt{\log\frac{1}{r}}\mathrm{d}r.$$

Since $\mu(B_r(w)) \leqslant 1$, it follows that $\rho C_\rho \leqslant 1$. The result follows upon the observation that $\int_0^1\sqrt{\log\frac{1}{r}}\mathrm{d}r = \frac{\sqrt{\pi}}{2}$. $\square$

F.2. Proof of Theorem 2

The Dudley bound is related to the Fernique–Talagrand functional through the following lemma, which combines Talagrand (2014, Corollary 2.3.2) with the discussion on Talagrand (2014, pg. 22). Let $N_r^d(W)$ denote the r-covering number of W, that is, the smallest integer N such that there exists a set of N balls of radius r under the metric d, whose union contains W.

Lemma 2 (Dudley entropy)**.** *There exists a universal constant $K > 0$ such that for any metric d and set W, $\gamma_2(W,d) \leqslant K\int_0^\infty\sqrt{\log N_r^d(W)}\mathrm{d}r$. In particular, $\gamma_2^\rho(W) \leqslant \frac{K}{\rho}\int_0^\rho\sqrt{\log N_r(W)}\mathrm{d}r$.*

If $W \subset \mathbb{R}^D$, the Dudley bound is never off by any more than a factor of $\log(d+1)$ (Talagrand, 2014, Exercise 2.3.4). Now, since $x \mapsto \sqrt{\log x}$ is concave on $[1,\infty)$, Jensen's inequality implies that $\mathbb{E}\sqrt{\log X} \leqslant \sqrt{\mathbb{E}\log X}$ for any random variable with support in $[1,\infty)$. Therefore,

$$\mathbb{E}\gamma_2^\rho(W) \leqslant \frac{K}{\rho}\int_0^\rho\sqrt{\log\mathbb{E}N_r(W)}\mathrm{d}r.$$

Covering numbers for images of Markov processes are bounded by the following fundamental lemma.

Lemma 3. *Let $\Lambda(r)$ be a fixed collection of cubes of side length r in $\mathbb{R}^d$ such that no ball of radius r can intersect more than K cubes of $\Lambda(r)$.*

1. *Suppose that X_n is a time-homogeneous Markov chain with n-step transition kernel $P^n(x, A)$. For any integer m, let $\mathcal{N}_r(m)$ denote the number of cubes in $\Lambda(r)$ hit by X_n at some time $0 \leqslant k \leqslant m$. Then*

$$\mathbb{E}\mathcal{N}_r(m) \leqslant 2K \left[\inf_{x \in \bigcup_{r>0} \Lambda(r)} \mathbb{E}_x \left(\frac{1}{m} \sum_{k=1}^{m} P^k(x, B_{r/3}(x)) \right) \right]^{-1}.$$

2. *Suppose that X_t is a time-homogeneous strong Markov process in $\mathbb{R}^d$ with transition kernel $P(t, x, A)$. For any $t \geqslant 0$, let $\mathcal{N}_r(t)$ denote the number of cubes in $\Lambda(r)$ hit by X_t at some time $s \in [0, t]$. Then*

$$\mathbb{E}\mathcal{N}_r(t) \leqslant 2K \left[\inf_{x \in \bigcup_{r>0} \Lambda(r)} \mathbb{E}_x \left(\frac{1}{t} \int_0^t P(s, x, B_{r/3}(x)) \mathrm{d}s \right) \right]^{-1}.$$

Proof. The second result is precisely Liu & Xiao (1998, Lemma 3.1), so it will suffice to show only the first. In a similar fashion, we consider a sequence of stopping times constructed in the following manner: let $\tau_0 = 0$ and for each positive integer j, we let

$$\tau_j = \min \left\{ k \geqslant \tau_{j-1} : \min_{i=0,\dots,j-1} |X_k - X_{\tau_i}| > r \right\}.$$

In other words, each τ_j is chosen to be the first time that the Markov chain is at least distance r from $X_{\tau_0}, \dots, X_{\tau_{j-1}}$. By construction, $|X_{\tau_j} - X_{\tau_k}| \geqslant r$ for $j \neq k$. By a Vitali covering argument, the balls $\{B_{r/3}(X_{\tau_j})\}_j$ are disjoint. Now, let

$$T_j = \sum_{k=\tau_j+1}^{\tau_j+m} \mathbb{1}_{X_k \in B_{r/3}(X_{\tau_j})}$$

be the sojourn time of X_k in $B_{r/3}(X_{\tau_j})$ in the interval $(\tau_j, \tau_j + m]$. Furthermore, let $\eta = \min\{k : \tau_k > m\}$ so that $\{X_k\}_{k=0}^m \subset \bigcup_{j=0}^{\eta-1} B_r(X_{\tau_j})$. Therefore, because $\{X_k\}_{k=0}^m$ is contained within the union of η balls in $\mathbb{R}^d$ and no ball in $\mathbb{R}^d$ can intersect any more than K cubes of $\Lambda(r)$, it follows that

$$\mathcal{N}_r(m) \leqslant K\eta. \tag{14}$$

Let I_j be the indicator of the event $\{\tau_j \leqslant m\}$, or equivalently, $\{\eta - 1 \geqslant j\}$. Doing so, we have that $\eta = \sum_{j=0}^{\infty} I_j$. Furthermore, since $\sum_{j=0}^{\infty} \mathbb{1}_{X_k \in B_{r/3}(X_{\tau_j})} \leqslant 1$ by the disjointness of $\{B_{r/3}(X_{\tau_j})\}_j$,

$$\sum_{j=0}^{\infty} I_j T_j = \sum_{j=0}^{\infty} \sum_{k=\tau_j+1}^{\tau_j+m} \mathbb{1}_{\tau_j \leqslant m} \mathbb{1}_{X_k \in B_{r/3}(X_{\tau_j})}$$

$$\leqslant \sum_{k=1}^{2m} \sum_{j=0}^{\infty} \mathbb{1}_{X_k \in B_{r/3}(X_{\tau_j})} = 2m.$$

By the strong Markov property, we may condition on starting the process at X_{τ_j}:

$$\mathbb{E}[I_j T_j] = \mathbb{E}\left[\mathbb{1}_{\tau_j \leqslant m} \mathbb{E}_{X_{\tau_j}} \sum_{k=1}^{m} \mathbb{1}_{X_k \in B_{r/3}(X_{\tau_j})} \right]$$

$$\geqslant \mathbb{E}I_j \cdot \inf_{x \in \mathbb{R}^d} \mathbb{E}_x \sum_{k=1}^{m} \mathbb{1}_{X_k \in B_{r/3}(x)}.$$

Therefore, by monotone convergence,

$$\mathbb{E}\eta \cdot \inf_{x \in \mathbb{R}^d} \mathbb{E}_x \sum_{k=1}^{m} \mathbb{1}_{X_k \in B_{r/3}(x)} = \sum_{j=0}^{\infty} \mathbb{E}I_j \inf_{x \in \mathbb{R}^d} \mathbb{E}_x \sum_{k=1}^{m} \mathbb{1}_{X_k \in B_{r/3}(x)}$$

$$\leqslant \sum_{j=0}^{\infty} \mathbb{E}[I_j T_j] = \mathbb{E} \sum_{j=0}^{\infty} I_j T_j \leqslant 2m,$$

and hence

$$\mathbb{E}\eta \leqslant 2\left[\inf_{x\in\mathbb{R}^d}\mathbb{E}_x\left(\frac{1}{m}\sum_{k=1}^{m}\mathbb{1}_{X_k\in B_{r/3}(x)}\right)\right]^{-1}. \tag{15}$$

The result now follows from (14) and (15). $\qquad\square$

If $\Lambda(r)$ is chosen to be the set of dyadic cubes in $\mathbb{R}^D$ of side length r, then $K = 3^D$. Combining Lemmas 2 and 3 with this choice of $\Lambda(r)$ yields Theorem 2. Note that the dimension dependence arises only due to this particular choice of $\Lambda(r)$. For our purposes, K is mostly irrelevant, but it is worth noting that this dimension dependence could feasibly be removed with a less naive choice of $\Lambda(r)$.

F.3. Proof of (8)

The proof of (8) relies on the following simple lemma.

Lemma 4. *For any $a > 0$, $x \in \mathbb{R}^D$, and $0 \leqslant r \leqslant \rho$,*

$$\frac{r^D}{2}I_\rho(a, D) \leqslant \int_{B_r(x)} e^{-a\|y-x\|^2}\mathrm{d}y \leqslant \frac{r^D}{D},$$

where $I_\rho(a, D) = \int_0^1 v^{D/2-1}e^{-av\rho^2}\mathrm{d}v$ is monotone decreasing in a, D, and ρ.

Proof. By a change of variables,

$$\int_{B_r(0)} \exp(-a\|x\|^2)\mathrm{d}x = \int_0^r u^{D-1}e^{-au^2}\mathrm{d}u = \frac{r^D}{2}\int_0^1 v^{D/2-1}e^{-ar^2v}\mathrm{d}v.$$

The bounds are obtained through $e^{-av\rho^2} \leqslant e^{-ar^2v} \leqslant 1$. $\qquad\square$

Proof of (8). In the sequel, K will be used to denote a universal constant, not necessarily the same at each appearance. First, by the Aronson estimate and Lemma 4,

$$\int_0^T P_t(x, B_r(x))\mathrm{d}t \geqslant K\int_0^T\int_{B_r(x)}\frac{1}{t^{D/2}}\exp\left(-\frac{\|x-y\|^2}{2\lambda t}\right)\mathrm{d}y\mathrm{d}t$$

$$\geqslant Kr^D\int_0^T\frac{1}{t^{D/2}}I_\rho\left(\frac{1}{2\lambda t}, D\right)\mathrm{d}t.$$

By Fubini's Theorem, and through the change of variables $t \mapsto t^{-1}$,

$$\int_0^T P_t(x, B_r(x))\mathrm{d}t \geqslant Kr^D\int_0^1\int_0^T\frac{1}{t^{D/2}}v^{D/2-1}e^{-v\rho^2/(2\lambda t)}\mathrm{d}t\mathrm{d}v$$

$$\geqslant Kr^D\int_0^1\int_{1/T}^\infty s^{D/2-2}v^{D/2-1}e^{-sv\rho^2/2\lambda t}\mathrm{d}s\mathrm{d}v$$

$$\geqslant KTr^D J_{\rho,T}\left(\frac{1}{2\lambda}, D\right),$$

where $J_{\rho,T}$ is as defined in Section 3. Equation (8) now follows from Theorem 2. $\qquad\square$

F.4. Proof of Corollary 2

The proof of Corollary 2 itself relies on the following corollary, which performs the local homogeneity approximation to $\{W_k\}_{k=0}^m$.

Corollary 3. *Let Ω be a closed set such that $\mathbb{P}(W_k \notin \Omega$ for some $k = 0, \ldots, m) \leqslant \zeta_m$. Then for any probability measure μ, letting μ^k denote k-fold convolution of μ and P_Ω the transition kernel of W_k conditioned on $W_k \in \Omega$ for $k = 1, 2, \ldots, m$,*

$$\mathbb{E}\gamma_2^\rho(\{W_k\}_{k=0}^m) \leqslant \frac{K}{\rho} \int_0^\rho \sqrt{(D+2)\log 3 - \log\left(\frac{1}{m}\sum_{k=1}^m \mu^k(B_r)\right)}\,\mathrm{d}r$$
$$+ \sqrt{\log(m+1)}\left(\zeta_m + m\sup_{x \in \Omega} d_{TV}(P_\Omega(x, x + \cdot), \mu)\right). \quad (16)$$

Proof. Let $\{\bar{W}_k\}_{k=0}^m$ denote the random walk satisfying $\bar{W}_{k+1} = \bar{W}_k + Z_k$, where each $Z_k \sim \mu$ is independent. Since $\gamma_2^\rho(W) \leqslant \sqrt{\log|W|}$, it follows that

$$|\mathbb{E}\gamma_2^\rho(\{W_k\}_{k=0}^m) - \mathbb{E}\gamma_2^\rho(\{\bar{W}_k\}_{k=0}^m)| \leqslant \sqrt{\log(m+1)} \sup_{\|f\|_\infty \leqslant 1} |\mathbb{E}f(\{W_k\}_{k=0}^m) - \mathbb{E}f(\{\bar{W}_k\}_{k=0}^m)|$$
$$\leqslant \sqrt{\log(m+1)}d_{\mathrm{TV}}(\{W_k\}_{k=0}^m, \{\bar{W}_k\}_{k=0}^m).$$

Since $\mathbb{E}\gamma_2^\rho(\{\bar{W}_k\}_{k=0}^m)$ is bounded above by the first term of (16) by Theorem 2, it suffices to show that

$$d_{\mathrm{TV}}(\{W_k\}_{k=0}^m, \{\bar{W}_k\}_{k=0}^m) \leqslant \zeta_m + m\sup_{x \in \Omega} d_{\mathrm{TV}}(P_\Omega(x, x + \cdot), \mu).$$

Let Z_k be chosen such that $\{W_k\}_{k=0}^m$ and $\{\bar{W}_k\}_{k=0}^m$ are optimally coupled under the total variation metric, that is,

$$d_{\mathrm{TV}}(\{W_k\}_{k=0}^m, \{\bar{W}_k\}_{k=0}^m) = \mathbb{P}(\{W_k\}_{k=0}^m \neq \{\bar{W}_k\}_{k=0}^m).$$

Conditioning on $W_k \in \Omega$, there is

$$d_{\mathrm{TV}}(\{W_k\}_{k=0}^m, \{\bar{W}_k\}_{k=0}^m) \leqslant \zeta_m + \mathbb{P}(\{W_k\}_{k=0}^m \neq \{\bar{W}_k\}_{k=0}^m|\{W_k\}_{k=0}^m \in \Omega^m)$$
$$\leqslant \zeta_m + \sum_{j=1}^m \mathbb{P}(W_j \neq \bar{W}_j|\{W_k\}_{k=0}^m \in \Omega^m, \{W_k\}_{k=0}^{j-1} = \{\bar{W}_k\}_{k=0}^{j-1}),$$

which implies (16). $\square$

Proof of Corollary 2. By the hypotheses, for $\bar{c}_m = m^{-1}(c_1 + \cdots + c_m)$, $m^{-1}\sum_{k=1}^m \mu^k(B_r) \geqslant \bar{c}_m r^\alpha$ for any $0 < r < r_0$. Let $\epsilon > 0$ be arbitrary. Note that if $r \leqslant \rho_\epsilon := \min\{1, r_0, (\bar{c}_m/3^{D+2})^{1/\epsilon}\}$,

$$\frac{1}{m}\sum_{k=1}^m \mu^k(B_r) \geqslant 3^{D+2}r^{\alpha+\epsilon}.$$

Therefore, considering the first term on the right hand side of (16), since $\rho_\epsilon \leqslant 1$,

$$\int_0^{\rho_\epsilon} \sqrt{(D+2)\log 3 - \log\left(\frac{1}{m}\sum_{k=1}^m \mu^k(B_r)\right)}\,\mathrm{d}r \leqslant \sqrt{\alpha+\epsilon}\int_0^1 \sqrt{\log\frac{1}{r}}\,\mathrm{d}r.$$

The result now follows from Corollary 3. $\square$