Proceedings of Machine Learning Research vol 201:1–51, 2023

Algorithmic Stability of Heavy-Tailed Stochastic Gradient Descent on Least Squares

Anant Raj ANANT.RAJ@INRIA.FR
Coordinated Science Laboratory
University of Illinois Urbana-Champaign
Inria, Ecole Normale Supérieure
PSL Research University, Paris, France

Melih Barsbey MELIH.BARSBEY@BOUN.EDU.TR
Department of Computer Engineering
Boğaziçi University, Istanbul, Turkey

Mert Gürbüzbalaban MG1366@RUTGERS.EDU
Department of Management Science and Information Systems
Rutgers University, Piscataway, NJ, United States of America

Lingjiong Zhu ZHU@MATH.FSU.EDU
Department of Mathematics
Florida State University, Tallahassee, FL, United States of America

Umut Şimşekli UMUT.SIMSEKLI@INRIA.FR
Inria, CNRS, Ecole Normale Supérieure
PSL Research University, Paris, France

Abstract

Recent studies have shown that heavy tails can emerge in stochastic optimization and that the heaviness of the tails have links to the generalization error. While these studies have shed light on interesting aspects of the generalization behavior in modern settings, they relied on strong topological and statistical regularity assumptions, which are hard to verify in practice. Furthermore, it has been empirically illustrated that the relation between heavy tails and generalization might not always be monotonic in practice, contrary to the conclusions of existing theory. In this study, we establish novel links between the tail behavior and generalization properties of stochastic gradient descent (SGD), through the lens of algorithmic stability. We consider a quadratic optimization problem and use a heavy-tailed stochastic differential equation (and its Euler discretization) as a proxy for modeling the heavy-tailed behavior emerging in SGD. We then prove uniform stability bounds, which reveal the following outcomes: (i) Without making any exotic assumptions, we show that SGD will not be stable if the stability is measured with the squared-loss $x \mapsto x^2$, whereas it in turn becomes stable if the stability is instead measured with a surrogate loss $x \mapsto |x|^p$ with some $p < 2$. (ii) Depending on the variance of the data, there exists a *'threshold of heavy-tailedness'* such that the generalization error decreases as the tails become heavier, as long as the tails are lighter than this threshold. This suggests that the relation between heavy tails and generalization is not globally monotonic. (iii) We prove matching lower-bounds on uniform stability, implying that our bounds are tight in terms of the heaviness of the tails. We support our theory with synthetic and real neural network experiments.

Keywords: Heavy tails, SGD, algorithmic stability, SDEs

1. Introduction

Over the last decade, understanding the generalization behavior in modern machine learning settings has been one of the main challenges in statistical learning theory. Here, the main goal has been deriving upper-bounds on the so-called *generalization error*, i.e., the gap between the true and the empirical risks $|F(\theta) - \hat{F}(\theta, X)|$, which are respectively defined as follows:

$$F(\theta) := \mathbb{E}_{x\sim P_X}[f(\theta, x)], \qquad \hat{F}(\theta, X) := (1/n)\sum_{i=1}^{n} f(\theta, x_i), \tag{1}$$

where $\theta \in \mathbb{R}^d$ denotes the *parameter vector*, $f : \mathbb{R}^d \times \mathcal{X} \mapsto \mathbb{R}_+$ is the *loss function*, $\mathcal{X}$ is the space of *data points*, P_X is the unknown *data distribution*, and finally $X = \{x_1, \dots, x_n\}$ denotes a (random) dataset with n points, such that each x_i is independently and identically distributed (i.i.d.) from P_X.

The past few years have witnessed the development of a variety of mathematical frameworks for analyzing the generalization error (see e.g., Liu and Theodorou (2019); He and Tao (2020) for recent surveys). In the context of empirical risk minimization (ERM), i.e., solving $\min_{\theta\in\mathbb{R}^d} \hat{F}(\theta, X)$, one promising direction has been to explicitly take into account the statistical properties of the *optimization algorithm* used during training, which is typically chosen as stochastic gradient descent (SGD) that is based on the following recursion:

$$\theta_{k+1} = \theta_k - \eta \nabla \tilde{F}_{k+1}(\theta_k, X), \tag{2}$$

where η is the step-size (or learning-rate), and $\nabla \tilde{F}_k(\theta, X) := \frac{1}{b}\sum_{i\in\Omega_k} f(\theta, x_i)$ is the stochastic gradient, with $\Omega_k \subset \{1, \dots, n\}$ being a random subset drawn with or without replacement, and $b := |\Omega_k| \ll n$ being the batch-size. In this line of research, Şimşekli et al. (2019) and Martin and Mahoney (2019) empirically demonstrated that, perhaps surprisingly, a *heavy-tailed* behavior can emerge in SGD in different ways, and the heaviness of the tails correlates with the generalization error, suggesting that heavier tails indicate better generalization.

Theoretically investigating these empirical observations, Gürbüzbalaban et al. (2021) and Hodgkinson and Mahoney (2021) explored the origins of the observed heavy-tailed behavior. They simultaneously showed that, in online SGD[1] (i.e., when the data is streaming), due the multiplicative nature of the gradient noise, i.e., $\nabla \tilde{F}_k(\theta, X) - \nabla \hat{F}(\theta, X)$, the distribution of the iterates θ_k can converge to a heavy-tailed distribution as $k \to \infty$. Furthermore, Gürbüzbalaban et al. (2021) showed that, when the loss f is a quadratic and the data distribution is Gaussian, the tails become monotonically heavier when η gets larger or b gets smaller.

Due to the fact that analyzing the heavy-tailed behavior arising from (2) can be highly non-trivial, relatively simpler heavy-tailed mathematical models have been used as a proxy for the original heavy-tailed SGD recursion in stationarity; e.g., SGD with heavy-tailed noise, i.e.,

$$\theta_{k+1} = \theta_k - \eta_{k+1}\left[\nabla \hat{F}(\theta_k, X) + \xi_{k+1}\right], \quad \text{with} \quad \mathbb{E}[\|\xi_k\|^2] = +\infty, \quad \text{for every } k = 1, 2, \dots, \tag{3}$$

where ξ_k denotes the heavy-tailed noise and η_k denotes a sequence of decreasing step-sizes. It has been revealed that another interesting situation emerges in this setting, this time in the behavior

1. The framework of Hodgkinson and Mahoney (2021) can handle stochastic optimization algorithms other than SGD as well.

of the optimization error. Notably, Zhang et al. (2020, Remark 1) pointed out that, when the loss function f is chosen as a simple quadratic, i.e., $f(\theta, x) = \|\theta\|^2$, we have that $\mathbb{E}[\|\theta_k - \theta_\star\|^2] = \mathbb{E}[\|\theta_k\|^2] = \mathbb{E}[f(\theta_k, x)] = +\infty$ for all k, where $\theta_\star = 0$ is the global minimum of f. While this result might appear daunting as it might seemingly suggest that "SGD diverges" under heavy-tailed perturbations, Wang et al. (2021) refined this result and showed that, if there exists $p \in [1, 2)$ such that $\mathbb{E}[\|\xi_k\|^p] < \infty$, then $\mathbb{E}[\|\theta_k - \theta_\star\|^p]$ converges to zero, for a class of strongly convex losses f. This result is particularly remarkable, since it shows that, even when the iterates may diverge under the 'true' loss function f (which SGD is originally trying to minimize), i.e., $\mathbb{E}[f(\theta_k, x)] = +\infty$, they might still converge to the *minimum of the original loss* $\theta_\star$ when a surrogate loss function $\tilde{f}$ is used for measuring the optimization error, which in this example is $\tilde{f}(\theta, x) = \|\theta\|^p$ with $p < 2$.

In an initial attempt for formalizing the relation between the tail behavior and generalization, Şimşekli et al. (2020) also modeled the original heavy-tailed recursion (2) by using a proxy and considered the following stochastic differential equation (SDE) as a model (which can be seen as a continuous-time version of (3)):

$$\mathrm{d}\theta_t = -\nabla \hat{F}(\theta_t, X)\mathrm{d}t + \Sigma(\theta_t)\mathrm{dL}_t^\alpha, \tag{4}$$

where $\Sigma : \mathbb{R}^d \mapsto \mathbb{R}^{d\times d}$ is a matrix-valued function and L_t^α denotes a heavy-tailed α-stable Lévy process, which is a random process parameterized by $\alpha \in (0, 2]$, such that a smaller α indicates heavier tails (we will make the definition of L_t^α precise in the next section). They showed that, under several assumptions on the SDE (4), the worst-case generalization error over the trajectory, i.e., $\sup_{t\in[0,1]} |\hat{F}(\theta_t, X) - F(\theta_t)|$ scales with the intrinsic dimension of the trajectory $(\theta_t)_{t\in[0,1]}$, which is then upper-bounded as a particular function of the tail-exponent around a local minimum, indicating that heavier tails imply lower generalization error. Their results were later extended to discrete-time recursions as well in Hodgkinson et al. (2022). More recently, Barsbey et al. (2021) linked heavy-tails to generalization through a notion of compressibility in the over-parameterized regimes. Yet, these bounds require several topological and statistical regularity assumptions that are hard to verify in realistic settings, and the experiments in Barsbey et al. (2021) illustrated that the relation between the tail-exponent and the generalization error is not always monotonic; hence, a generalization bound that requires less assumptions while being more pertinent to the practical observations is still missing.

In this study, we aim at establishing novel links between tail behavior and generalization and address the aforementioned shortcomings. We consider the problem through the lens of *algorithmic stability* (Bousquet and Elisseeff, 2002; Hardt et al., 2016), and explore the effects of heavy tails on the stability of SGD. Similar to recent work (Ali et al., 2020; Gürbüzbalaban et al., 2021; Latorre et al., 2021), in order to have a more explicit control over the problem, we limit our scope to quadratic optimization, and consider the following SDE as a proxy for heavy-tailed SGD:

$$\mathrm{d}\theta_t = -\frac{1}{n}\left(X^\top X\right)\theta_t \mathrm{d}t + \Sigma \mathrm{dL}_t^\alpha, \tag{5}$$

where $\Sigma \in \mathbb{R}^{d\times d}$ is a matrix that scales the noise and is assumed to be fixed (i.e., state-independent), and by a slight abuse of notation we represent the dataset as a matrix $X \in \mathbb{R}^{n\times d}$, such that i-th row of X is equal to x_i. This SDE naturally arises from the ERM problem with the loss function being $f(\theta, x) = (\theta^\top x)^2$.

As the learning algorithm, we first consider the case where we assume that we have a sample from the stationary distribution of (5) (i.e., the case where $t \to \infty$) and analyze the stability of this

sample. Then we extend our analysis in two directions: we analyze (i) the case where t is finite and (ii) the case where the SDE is discretized by using a constant step-size. Our contributions are as follows:

- As opposed to classical SDEs driven by a Brownian motion (rather than α-stable Lévy processes L_t^α as we consider here), the stationary distribution of (5) does not admit a simple analytical closed-form expression. As a remedy, we perform the stability analysis in the Fourier domain, and introduce new proof techniques.
- We prove upper-bounds on the stability of (5), which suggest that the algorithm will not be stable, when $\alpha < 2$ and the stability is measured with respect to the quadratic loss $(\theta^\top x)^2$. We further show that, when the stability is instead measured with respect to a surrogate loss function $|\theta^\top x|^p$ with $p < \alpha < 2$, the algorithm in turn becomes stable, where the level of stability depends on α, among several other quantities. This result reveals a similar phenomenon to that of Zhang et al. (2020) and Wang et al. (2021) as discussed above. Furthermore, our results do not require any non-trivial assumptions, compared to the existing heavy-tailed generalization bounds (Şimşekli et al., 2020; Hodgkinson et al., 2022; Barsbey et al., 2021).
- Our theory further discloses an interesting property: depending on the variance of the data distribution P_X, there exits an $\alpha_0 > 1$, such that the algorithm becomes more stable as $\alpha \in [\alpha_0, 2]$ get smaller, i.e., the tails get heavier up to a certain point determined by α_0. This result implies that the stability of the algorithm, hence the generalization error will be monotonic with respect to the tail-exponent α only when α is large enough. This outcome sheds more light on the experimental results presented in Barsbey et al. (2021), where the relation between the tail-exponent and the generalization error is only partially monotonic.
- We prove matching lower-bounds on the stability of (5), implying that our stability bounds are tight in the tail-exponent α.
- We show that the same conclusions hold for finite t, and for the Euler discretization of (5) when the step-size is small enough.

We support our theory on both synthetic data and real experiments conducted on standard benchmark datasets by using fully-connected and convolutional neural networks. All the proofs and the implementation details are provided in the Appendix.

2. Notation and Background

Notation. Consider a real-valued function $f : \mathbb{R}^d \to \mathbb{R}$ defined on $\mathbb{R}^d$. The Fourier transform of $f(\theta)$ for $\theta \in \mathbb{R}^d$ is denoted by $\mathcal{F}f(u)$ and is defined as, $\mathcal{F}f(u) := \int_{\mathbb{R}^d} f(\theta) e^{-\mathrm{i}u^\top \theta} \mathrm{d}\theta$. Similarly, the inverse Fourier transform of a function $\hat{f}(u)$ that is from $\mathbb{R}^d$ to $\mathbb{R}$ is denoted by $\mathcal{F}^{-1}\hat{f}(\theta)$ and is defined by, $\mathcal{F}^{-1}\hat{f}(\theta) := \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \hat{f}(u) e^{\mathrm{i}u^\top \theta} du$. In both of these definitions, $\mathrm{i} := \sqrt{-1}$.

α-stable distributions. The α-stable distribution appears as the limiting distribution in the generalized central limit theorems for a sum of i.i.d. random variables with infinite variance (Lévy, 1937). A scalar random variable X is called symmetric α-stable, denoted by $X \sim \mathcal{S}\alpha\mathcal{S}(\sigma)$, if its characteristic function takes the form: $\mathbb{E}\left[e^{\mathrm{i}uX}\right] = \exp\left(-\sigma^\alpha |u|^\alpha\right)$, for any $u \in \mathbb{R}$, where $\sigma > 0$ is known as the scale parameter that measures the spread of X around 0 and $\alpha \in (0, 2]$ which is known as the tail-index that determines the tail thickness of the distribution and the tail becomes heavier as α gets smaller. In general, the probability density function of a symmetric α-stable distribution, $\alpha \in (0, 2]$,

does not yield closed-form expression except for a few special cases. When $\alpha = 1$ and $\alpha = 2$, $\mathcal{S}\alpha\mathcal{S}$ reduces to the Cauchy and the Gaussian distributions, respectively. When $0 < \alpha < 2$, the moments are finite only up to the order α in the sense that $\mathbb{E}[|X|^p] < \infty$ if and only if $p < \alpha$, which implies infinite variance. Moreover, α-stable distribution can be extended to the high-dimensional case for random vectors. One natural extension is the rotationally symmetric α-stable distribution. X follows a d-dimensional rotationally symmetric α-stable distribution if it admits the characteristic function $\mathbb{E}\left[e^{\mathrm{i}\langle u, X\rangle}\right] = e^{-\sigma^\alpha \|u\|_2^\alpha}$ for any $u \in \mathbb{R}^d$. We refer to Samorodnitsky and Taqqu (1994) for the details of α-stable distributions.

Lévy processes. Lévy processes are stochastic processes with independent and stationary increments. Their successive displacements can be viewed as the continuous-time analogue of random walks. Lévy processes include the Poisson process, Brownian motion, the Cauchy process, and more generally stable processes; see e.g. Bertoin (1996); Samorodnitsky and Taqqu (1994); Applebaum (2009). Lévy processes in general admit jumps and have heavy tails which are appealing in many applications; see e.g. Cont and Tankov (2004). In this paper, we will consider the rotationally symmetric α-stable Lévy process L_t^α in $\mathbb{R}^d$ that is defined as follows.

(i) $\mathrm{L}_0^\alpha = 0$ almost surely;

(ii) For any $t_0 < t_1 < \cdots < t_N$, the increments $\mathrm{L}_{t_n}^\alpha - \mathrm{L}_{t_{n-1}}^\alpha$ are independent;

(iii) The difference $\mathrm{L}_t^\alpha - \mathrm{L}_s^\alpha$ and L_{t-s}^α have the same distribution, with the characteristic function $\exp(-(t-s)^\alpha \|u\|_2^\alpha)$ for $t > s$;

(iv) L_t^α has stochastically continuous sample paths, i.e. for any $\delta > 0$ and $s \geq 0$, $\mathbb{P}(\|\mathrm{L}_t^\alpha - \mathrm{L}_s^\alpha\| > \delta) \to 0$ as $t \to s$.

When $\alpha = 2$, $\mathrm{L}_t^\alpha = \sqrt{2}\mathrm{B}_t$, where B_t is the standard d-dimensional Brownian motion.

Ornstein-Uhlenbeck processes. Ornstein-Uhlenbeck (OU) process (Uhlenbeck and Ornstein, 1930) is a d-dimensional Markov and Gaussian process that satisfies the SDE:

$$\mathrm{d}X_t = -AX_t\mathrm{d}t + \Sigma\mathrm{d}\mathrm{B}_t, \tag{6}$$

where A and Σ are a $d \times d$ matrices and B_t is a standard d-dimensional Brownian motion. The OU process is a special case of the Langevin equation in physics (Pavliotis, 2014), and has wide applications including for example modeling the change in organismal phenotypes in evolutionary biology (Martins, 1994), and the short-rate in the interest rate modeling in finance (Vasicek, 1977). More generally, we can consider an OU process driven by a Lévy process, for example, replacing B_t in (6) by a rotationally symmetric α-stable Lévy process L_t^α so that

$$\mathrm{d}X_t = -AX_t\mathrm{d}t + \Sigma\mathrm{d}\mathrm{L}_t^\alpha. \tag{7}$$

Under some mild conditions on A and Σ, the OU process X_t in (7) admits a unique stationary distribution that can be fully characterized; see e.g. Sato and Yamazato (1984); Masuda (2004).

Algorithmic stability and generalization. In this paper, we study the generalization of the continuous-time heavy-tailed SGD by using the tools of algorithmic stability. Several notions of stability have been defined in the literature of statistical learning theory (Bousquet and Elisseeff, 2002; Elisseeff et al., 2005). We will use the notion of algorithmic stability of the randomized algorithm $\mathcal{A}$ defined in Hardt et al. (2016). We denote the set $\mathcal{X}_n$ as the set of all possible size n datapoints subsampled uniformly at random from P_X.

Definition 1 (Hardt et al. (2016), Definition 2.1) *For a loss function $f : \mathbb{R}^d \times \mathcal{X} \to \mathbb{R}$, an algorithm $\mathcal{A}$ is ε-uniformly stable if*

$$\varepsilon_{stab}(\mathcal{A}) := \sup_{X \cong \hat{X}} \sup_{z \in \mathcal{X}} \mathbb{E}\left[f(\mathcal{A}(X), z) - f(\mathcal{A}(\hat{X}), z)\right] \leq \varepsilon, \tag{8}$$

where the first supremum is taken over data $X, \hat{X} \in \mathcal{X}_n$ that differ by one element, denoted by $X \cong \hat{X}$.

Since its introduction in statistical learning theory in Bousquet and Elisseeff (2002), stability based arguments have been useful in deriving generalization bound for several learning algorithms (Belkin et al., 2004; Cortes et al., 2012; Wu and Cheng, 2021; Maurer and Jaakkola, 2005) and have also been extended to get generalization bound for randomized algorithm like SGD and SGLD (Hardt et al., 2016; Raginsky et al., 2017; Kuzborskij and Lampert, 2018; Mou et al., 2018; Chen et al., 2018; Bassily et al., 2020; Charles and Papailiopoulos, 2018; Lei and Ying, 2020; Farghly and Rebeschini, 2021). Here below, we provide a result from Hardt et al. (2016) which relates algorithmic stability with the generalization performance of a randomized algorithm.

Theorem 2 ((Hardt et al., 2016), Theorem 2.2) *Suppose that $\mathcal{A}$ is an ε-uniformly stable algorithm, then the expected generalization error is bounded by*

$$\left|\mathbb{E}_{\mathcal{A},X}\left[\hat{F}(\mathcal{A}(X), X) - F(\mathcal{A}(X))\right]\right| \leq \varepsilon. \tag{9}$$

In several of recent works (Feldman and Vondrak, 2019; Bousquet et al., 2020; Klochkov and Zhivotovskiy, 2021), high probability bounds have been obtained using algorithmic stability bounds.

3. Algorithmic Stability of Heavy-Tailed SGD on Least Squares Regression

In this section, we will investigate the effects of heavy-tails on algorithmic stability. We consider the setting of least square regression with $f(\theta, (x, y)) = (\theta^\top x - y)^2/2$. We assume that we only have the access to the data generation distribution P_X via the generated training samples and our goal is to learn a parameter vector $\theta \in \mathbb{R}^d$ which minimize the corresponding population risk. We denote the training data by the matrix $X = \left[x_1^\top, x_2^\top, \ldots, x_i^\top, \ldots, x_n^\top\right] \in \mathbb{R}^{n \times d}$ and $y = [y_1, y_2, \ldots, y_i, \ldots, y_n] \in \mathbb{R}^n$, where n is the number of data points, d is the dimension of the problem, and $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$ for all i. Training data points are i.i.d. from the distribution P_X. We consider the ERM problem as defined in (1): $\min_{\theta \in \mathbb{R}^d} \frac{1}{2n} \sum_{i=1}^n (\theta^\top x_i - y_i)^2$.

In the context of algorithmic stability, we assume that we have two training datasets (X, y) and $(\hat{X}, \hat{y})$ that differ in only one data point. Without loss of generality, we have

$$\hat{X} = \left[x_1^\top, x_2^\top, \ldots, \tilde{x}_i^\top, \ldots, x_n^\top\right] \in \mathbb{R}^{n \times d}, \qquad \hat{y} = [y_1, y_2, \ldots, \tilde{y}_i, \ldots, y_n] \in \mathbb{R}^n.$$

For our ERM problem, we consider the continuous-time heavy-tailed stochastic gradient descent, which is represented by the following two SDEs that are driven by a rotationally symmetric α-stable Lévy process L_t^α in $\mathbb{R}^d$,

$$\mathrm{d}\theta_t = -\frac{1}{n}\left(X^\top X \theta_t - X^\top y\right)\mathrm{d}t + \Sigma \mathrm{d}\mathrm{L}_t^\alpha, \tag{10}$$

$$\mathrm{d}\hat{\theta}_t = -\frac{1}{n}\left(\hat{X}^\top \hat{X}\hat{\theta}_t - \hat{X}^\top \hat{y}\right)\mathrm{d}t + \Sigma \mathrm{d}\mathrm{L}_t^\alpha, \tag{11}$$

where $\Sigma \in \mathbb{R}^{d\times d}$ is a real-valued matrix.

Under mild conditions, the SDEs (10) and (11) have unique strong solutions, which are Markov processes and they admit unique invariant distributions (Sato and Yamazato, 1984). Thanks to the linearity of the drifts of these SDEs, the stationary distribution is achieved very quickly, with an exponential rate (Xie and Zhang, 2020). Hence, to ease our analysis, we will assume that we have two samples from the stationary distributions of (10) and (11), say θ and $\hat{\theta}$. In other words, we set our learning algorithm such that it gives a random sample from the stationary distribution of the SDE determined by the dataset, i.e., $\mathcal{A}_{\text{cont}}((X,y)) = \theta$, and $\mathcal{A}_{\text{cont}}((\hat{X},\hat{y})) = \hat{\theta}$, where $\mathcal{A}_{\text{cont}}$ denotes the *continuous-time* heavy-tailed SGD algorithm. In the rest of this section, we will derive stability bounds for this learning algorithm.

3.1. Warm-up: the need for the surrogate loss

To motivate our analysis technique, let us first consider the following simple setting, where we set $d = 1$, so that we have $f(\theta,(x,y)) = (x\theta - y)^2$. In this specific case, when $\alpha > 1$, we can compute the stationary distributions of (10) and (11) in an explicit form. With a slight abuse of notation, the distribution of θ_t converges to a symmetric stable law: $(\delta/s) + \mathcal{S}\alpha\mathcal{S}((\alpha s)^{-1/\alpha})$, where $s = (1/n)\sum_{i=1}^n x_i^2$ and $\delta = (1/n)\sum_{i=1}^n x_i y_i$ with δ/s being the mean (and the mode) of the stationary distribution, which coincides with the ordinary least-squares solution. Similarly, the distribution of $\hat{\theta}_t$ converges to $(\hat{\delta}/\hat{s}) + \mathcal{S}\alpha\mathcal{S}((\alpha\hat{s})^{-1/\alpha})$, where $\hat{\delta}$ and $\hat{s}$ are defined analogously.

As a first observation, assume that we have a sample from the stationary distribution of θ_t, such that $\theta \sim (\delta/s)+\mathcal{S}\alpha\mathcal{S}((\alpha s)^{-1/\alpha})$. Considering this scheme as the algorithm, i.e., $\mathcal{A}_{\text{cont}}((X,y)) = \theta$, a simple calculation shows that

$$\mathbb{E}_{\mathcal{A}_{\text{cont}}(X,y)}\left[f\left(\mathcal{A}_{\text{cont}}((X,y)),(X,y)\right)\right] = \mathbb{E}_{\theta,(X,y)}\left[(1/n)\sum\nolimits_{i=1}^n (x_i\theta - y_i)^2\right] = +\infty,$$

since the variance of $\mathcal{S}\alpha\mathcal{S}((\alpha s)^{-1/\alpha})$ is infinite whenever $\alpha < 2$. Therefore, it is clear that we cannot expect any algorithmic stability in this scheme, as long as the stability is measured with respect to the squared loss. However, as we will show in the sequel, in turns out that if we instead measure the stability with respect to a surrogate loss function, which in this case would be $|x\theta - y|^p$ for some $p \in [1,\alpha)$, the algorithm becomes stable, even though it is based on a distribution that concentrates near the optimum for squared loss.

On the other hand, we notice that the means of the stationary distributions, i.e., δ/s and $\hat{\delta}/\hat{s}$ do not interact with the tail exponent α. Since our main goal is to investigate the interplay between the tail behavior and algorithmic stability, we will ignore this term and assume that $y_i = 0$ almost surely for all i (otherwise non-zero y_i will only introduce terms in the stability that do not depend on α). This way, we fall back to the SDE given in (5).

In the light of these two observations, for the general case where $d \geq 1$, we will use the following surrogate loss function to measure stability:

$$f(x) := f(\theta,x) := |\theta^\top x|^p, \quad \text{for some} \quad p \in [1,2], \tag{12}$$

which generalizes the original loss function. Note that, from now on we will drop the notation $\tilde{f}$ for denoting surrogate losses for simplicity and use a single notation for the loss function.

3.2. Algorithmic stability analysis in the Fourier domain

For $d \geq 2$, unfortunately we cannot identify the stationary distributions of (10) and (11) in an explicit form. However, by using the theory of the characterization of the stationary distribution for an Ornstein-Uhlenbeck process driven by a Lévy process in the literature (see Sato and Yamazato (1984); Masuda (2004) and the background review in the Appendix), in the next lemma, we show that we can characterize the stationary distribution of the Ornstein-Uhlenbeck process driven by a rotationally symmetric α-stable Lévy process in a semi-explicit way:

$$\mathrm{d}\theta_t = -A\theta_t \mathrm{d}t + \Sigma \mathrm{d}\mathrm{L}_t^{\alpha}, \tag{13}$$

where A and Σ are $d \times d$ real matrices.

Lemma 3 *Assume that A is a real symmetric matrix with all the eigenvalues being positive. Then (13) admits a unique stationary distribution π whose characteristic function is given by*

$$\int_{\mathbb{R}^d} e^{\mathrm{i}u^\top x} \pi(\mathrm{d}x) = \exp\left(- \int_0^\infty \left\| \Sigma^\top e^{-sA} u \right\|_2^\alpha \mathrm{d}s \right). \tag{14}$$

While Lemma 3 provides us information about the stationary distributions of the SDEs (10) and (11), it considers the Fourier transforms of these distributions, which makes this setting not amenable to conventional algorithmic stability analysis tools.

As a remedy, we perform the stability analysis directly in the Fourier domain and use the Fourier inversion theorem to compute stability bounds for continuous-time heavy-tailed SGD. Our main approach is based on the following observation. Let $g : \mathbb{R}^d \mapsto \mathbb{R}$ be a function, and P, Q be random variables in $\mathbb{R}^d$ with respective characteristic functions ψ_P and ψ_Q. If the Fourier inversion theorem holds on g, then g is the inverse Fourier transform of $\mathcal{F}g(\cdot)$. Hence,

$$\begin{aligned}
\mathbb{E}\left[g(P) - g(Q)\right] &= \frac{1}{(2\pi)^d} \mathbb{E}\left[\int_{\mathbb{R}^d} \left(e^{\mathrm{i}u^\top P} - e^{\mathrm{i}u^\top Q} \right) \mathcal{F}g(u)\, \mathrm{d}u \right] \\
&= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \mathbb{E}\left[e^{\mathrm{i}u^\top P} - e^{\mathrm{i}u^\top Q} \right] \mathcal{F}g(u)\, \mathrm{d}u \\
&= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \left(\psi_P(u) - \psi_Q(u) \right) \mathcal{F}g(u)\, \mathrm{d}u \\
&\leq \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} |\psi_P(u) - \psi_Q(u)| |\mathcal{F}g(u)|\, \mathrm{d}u.
\end{aligned} \tag{15}$$

Hence, (15) enables us to utilize the result given in Lemma 3 and hence gives us a way to perform stability analysis (as given in Definition 1) in the Fourier domain.

Algorithmic stability via characteristic function. By setting $y = \hat{y} = 0$ and invoking Lemma 3, the characteristic functions of stationary distributions corresponding to the SDEs (10) and (11) are respectively given as follows:

$$\psi_\theta(u) = \exp\left(- \int_0^\infty \left\| \Sigma^\top e^{-s\frac{1}{n}(X^\top X)} u \right\|_2^\alpha \mathrm{d}s \right), \tag{16}$$

$$\psi_{\hat{\theta}}(u) = \exp\left(- \int_0^\infty \left\| \Sigma^\top e^{-s\frac{1}{n}(\hat{X}^\top \hat{X})} u \right\|_2^\alpha \mathrm{d}s \right). \tag{17}$$

From the Definition 1 and from (15) and (12), we have

$$\begin{aligned}\varepsilon_{\text{stab}}(\mathcal{A}_{\text{cont}}) &= \sup_{X\cong\hat{X}}\sup_{x\in\mathcal{X}}\mathbb{E}\left[\left|\theta^\top x\right|^p - \left|\hat{\theta}^\top x\right|^p\right] \\ &= \sup_{X\cong\hat{X}}\sup_{x\in\mathcal{X}}\frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}|\psi_\theta(u)-\psi_{\hat{\theta}}(u)|\cdot\left|\mathcal{F}\left[|x^\top\cdot|^p\right](u)\right|\,\mathrm{d}u. \end{aligned} \tag{18}$$

In the remainder of this section, we will consider $\Sigma = I$ for convenience with I being the identity matrix. However, we provide bounds showing the effect of Σ in the Appendix.

One-dimensional case ($d = 1$). We first discuss the case where $d = 1$ and report it as a separate result since its proof is simpler and more instructive. Following (15), as a first step, we prove a lemma, which relates the characteristic functions of the stationary distributions by upper-bounding $|\psi_\theta(u) - \psi_{\hat{\theta}}(u)|$. For the sake of brevity, we present this result in the Appendix (Lemma 12). By using this intermediate result, we next prove upper- and lower-bounds on the stability of the continuous time heavy-tailed SGD algorithm and discuss its behavior with respect to α and p.

Theorem 4 *Consider the one-dimensional loss function $f(x) = |\theta x|^p$. For any $x \sim P_X$, if we have $|x| > R$ with probability δ_1 and for any X sampled uniformly at random from the set $\mathcal{X}_n$, if we have $\|X\|_2^2 \le \sigma^2 n$ with probability δ_2. Then,*

(i) For $\alpha \in [1, 2)$, the algorithm is not stable when $p \in [\alpha, 2]$ i.e. $\varepsilon_{stab(\mathcal{A}_{cont})}$ diverges. When $\alpha = p = 2$ then $\varepsilon_{stab}(\mathcal{A}_{cont}) \le \frac{R^4}{\pi\sigma^4 n}$ with probability at least $1 - \delta_1 - 2\delta_2$.

(ii) For $p \in [1, \alpha)$, we have the following upper bound for the algorithmic stability,

$$\varepsilon_{stab}(\mathcal{A}_{cont}) \le \frac{2R^{p+2}}{\pi\sigma^2 n}\Gamma(p+1)\cos\left(\frac{(p-1)\pi}{2}\right)\frac{1}{\alpha}\left(\frac{1}{\alpha\sigma^2}\right)^{\frac{p}{\alpha}}\Gamma\left(1-\frac{p}{\alpha}\right) =: c(\alpha),$$

which holds with probability at least $1 - \delta_1 - 2\delta_2$. Furthermore, for some $\alpha_0 > 1$, if we have

$$\sigma^2 \ge \exp\left(1 + \frac{2}{p} - \log\alpha_0 - \phi\left(1 - \frac{p}{\alpha_0}\right)\right), \tag{19}$$

where ϕ is the digamma function, then the map $\alpha \mapsto c(\alpha)$ is increasing for $\alpha \in [\alpha_0, 2)$.

(iii) The stability bound is tight in α.

Informally, this result illuminates the following facts: (i) When subject to heavy tails, i.e., $\alpha < 2$, the algorithm is stable only when a surrogate loss is used with $p < \alpha$. (ii) For $1 \le p < \alpha < 2$, the stability level $\varepsilon_{\text{stab}}$ is upper-bounded by a function of α, p, and the variance of the data distribution σ^2. Furthermore (and perhaps more surprisingly), for a given *heavy-tailedness threshold* $\alpha_0 \in (1, 2)$, if the data variance is sufficiently large as in (19), the stability bound becomes monotonically increasing for $\alpha \in [\alpha_0, 2)$, which indicates that as the algorithm becomes more stable it gets heavier-tailed. However, this relation holds as long as the heaviness of the tails does not exceed the threshold α_0. (iii) We further show that, there exists a data distribution P_X such that $\varepsilon_{\text{stab}}$ is lower-bounded by a function, which also depends on α, p, and σ^2. In the proved lower-bound, the terms depending on α have the same order as of the ones given in the upper-bound of Theorem 4. Hence our stability bound is tight in α. Combined with point (ii), this result suggests that the generalization error might not be globally monotonic with respect to the heaviness of the tails under our modeling strategy. On the other hand, for a fixed data distribution where σ^2 is given, (19) provides a 'guideline' for choosing the optimal tail index α in the sense of algorithmic stability.

Multi-dimensional case ($d \geq 2$). Now we will focus our attention to the case of d dimensions. We follow the same route as in Theorem 4, where we first relate the characteristic functions of the stationary distributions. We also present this result in the Appendix (Lemma 13). Based on Lemma 13, we next provide stability bounds for the d-dimensional case.

Theorem 5 *Consider $f(x) = |\theta^\top x|^p$ such that $\theta, x \in \mathbb{R}^d$. Assume that for almost all $x \sim P_X$, we have $\|x\|_2 \leq R$, for any X sampled uniformly at random from the set $\mathcal{X}_n$, we have $\frac{1}{n}\|X^\top X u\|_2 \geq \sigma_{\min}\|u\|_2$ for all $u \in \mathbb{R}^d$ and for any two $X \cong \hat{X}$ sampled from $\mathcal{X}_n$ generating two stochastic process given by SDEs in equations (10) and (11), we have $\|x_i x_i^\top - \tilde{x}_i \tilde{x}_i^\top\|_2 \leq 2\sigma$ holds with high probability. Then,*

(i) For $\alpha \in (1,2)$, the algorithm is not stable when $p \in [\alpha, 2]$ i.e. $\varepsilon_{stab(\mathcal{A}_{cont})}$ diverges. When $\alpha = p = 2$ then with high probability $\varepsilon_{stab}(\mathcal{A}_{cont}) \leq \frac{2R^2}{\pi}\frac{\sigma}{n\sigma_{\min}^2}$.

(ii) For $p \in [1, \alpha)$, we have the following upper bound for the algorithmic stability,

$$\varepsilon_{stab}(\mathcal{A}_{cont}) = \frac{8R^p}{\pi}\frac{\sigma}{n\alpha^2\sigma_{\min}}\Gamma(p+1)\cos\left(\frac{(p-1)\pi}{2}\right)\left(\frac{1}{\alpha\sigma_{\min}}\right)^{\frac{p}{\alpha}}\Gamma\left(1-\frac{p}{\alpha}\right) = c(\alpha),$$

which holds with high probability. Furthermore, for some $\alpha_0 > 1$, if we have

$$\sigma_{\min} \geq \exp\left(1 + \frac{4}{p} - \log\alpha_0 - \phi\left(1 - \frac{p}{\alpha_0}\right)\right),$$

where ϕ is the digamma function, then the map $\alpha \to c(\alpha)$ is increasing for $\alpha \in [\alpha_0, 2)$.

iii The stability bound is tight in α.

The conclusions of Theorem 5 are almost identical to the ones of Theorem 4, though its proof requires a more careful analysis, especially for the lower-bound in (iii). The main differences here are that, we need the smallest eigenvalue of the covariance matrix of P_X, i.e., $\sigma_{\min}$ to be large enough, and we need a different condition on the second moment σ of the distribution. Under these conditions, we obtain very similar stability and monotonicity properties.

As a final remark, we note that our results do not require any non-trivial topological or statistical assumptions in comparison with Şimşekli et al. (2020) and Barsbey et al. (2021) that suggested a globally monotonic relation for the generalization error and the tail exponent α. On the other hand, the rate $1/n$ in our bounds are in line with the existing stability literature (Hardt et al., 2016; Maurer and Jaakkola, 2005).

Finite time bound. The result presented in Theorem 5 is for the case when $t \to \infty$ i.e. θ is sampled from the stationary distribution of the stochastic process corresponding to the SDE in equation (5). However, in the Appendix B, we characterize the finite time distribution of a Lévy-driven OU process. We show that the characteristic function of the probability density corresponding to the SDE in equation (5) is given as,

$$\psi_\theta(t,u) = \exp\left(-\int_0^t \left\|\Sigma^\top e^{-s\frac{1}{n}(X^\top X)}u\right\|_2^\alpha \mathrm{d}s\right).$$

If we observe carefully, we can follow the similar procedure to get the stability bound for finite time case as we did to obtain for $t \to \infty$. In particular, in Remark 14 in the appendix, we show that whenever $t = O(\frac{1}{\alpha\sigma_{\min}})$, the same monotonicity conclusions of Theorem 5 still hold. See Remark 14 for more details.

Algorithmic stability for the Euler discretization. Previously, we have provided results for the continuous-time case which can not be implemented in practice. Now, we derive a stability bound for the Euler discretization of the SDE (5). We consider the following scheme:

$$\theta_{k+1} = \theta_k - \frac{\eta}{n}\left(X^\top X\theta_k - X^\top y\right) + \eta^{1/\alpha}\Sigma S_{k+1}, \tag{20}$$

$$\hat{\theta}_{k+1} = \hat{\theta}_k - \frac{\eta}{n}\left(\hat{X}^\top \hat{X}\hat{\theta}_k - \hat{X}^\top \hat{y}\right) + \eta^{1/\alpha}\Sigma S_{k+1}. \tag{21}$$

To provide algorithmic stability guarantees for the discretization, we first identify the characteristic function of the stationary distribution of the discretization in Appendix C (Lemma 10). We then provide a stability bound based on these characteristic functions. We only present the result for $k \to \infty$ here, however, the result for any finite k follows the same procedure as we have provided stability bound for characteristic function for any finite k in Lemma 15.

Theorem 6 *Consider $f(x) = |\theta^\top x|^p$ such that $\theta, x \in \mathbb{R}^d$. Assume that for almost all $x \sim P_X$, we have $\|x\|_2 \leq R$, for any X sampled uniformly at random from the set $\mathcal{X}_n$, it holds that $\frac{1}{n}\|X^\top X u\|_2 \geq \sigma_{\min}\|u\|_2$ for all $u \in \mathbb{R}^d$ and for any two $X \cong \hat{X}$ sampled from $\mathcal{X}_n$ generating the stochastic process given in (20) and (21), we have that $\frac{1}{n}\hat{X}^\top \hat{X}$, $\|x_i x_i^\top - \tilde{x}_i \tilde{x}_i^\top\|_2 \leq 2\sigma$ holds with high probability. Further assume that $\eta \leq \frac{1}{L}$ where L is the maximum of largest eigenvalues of $\frac{1}{n}X^\top X$. Then, for $p \in [1, \alpha)$, we have*

$$\varepsilon_{stab} \leq \frac{2R^p}{\pi}\Gamma(p+1)\cos\left(\frac{(p-1)\pi}{2}\right)\frac{\sigma\eta^{1+\frac{p}{\alpha}}(1-\eta\sigma_{\min})^{\alpha-1}}{n\alpha(1-(1-\eta\sigma_{\min})^\alpha)^{1+\frac{p}{\alpha}}}\Gamma\left(1-\frac{p}{\alpha}\right)$$

with high probability.

This theorem shows that the monotonicity behavior of algorithmic stability with respect to α can be more complicated. However, if the step-size η is chosen small enough such that we can consider the approximation $(1-\eta\sigma_{\min})^\alpha \approx 1 - \eta\alpha\sigma_{\min}$ then the result from above Theorem 6 can be written as, $\varepsilon_{\text{stab}} \leq \frac{2R^p}{\pi}\Gamma(p+1)\cos\left(\frac{(p-1)\pi}{2}\right)\frac{\sigma}{n\alpha^2\sigma_{\min}}\left(\frac{1}{\alpha\sigma_{\min}}\right)^{\frac{p}{\alpha}}\Gamma\left(1-\frac{p}{\alpha}\right)$, with high probability. This expression is almost the same as the bound given in Theorem 5. Hence, we get a similar behavior of $\varepsilon_{\text{stab}}$ with respect to α for the discretized SDEs when η is small enough. Finally, this result is independent of η, which is perhaps not surprising as similar results also exist for SGD with strong convex losses (Hardt et al., 2016, Theorem 3.9).

4. Experiments

Synthetic data. We first test the implications of the theoretical findings presented above with synthetic data experiments. We assume that $y = 0$ and P_X is a scaled uniform distribution $\mathcal{U}(-0.5a, 0.5a)$, with a determining the range of the distribution. We simulate the SDE presented in (10) by using the Euler-Maruyama discretization, which yields the following recursion:

$$\theta_{k+1} = \theta_k - \eta\frac{1}{n}\left(X^\top X\right)\theta_k - \eta^{1/\alpha}E_{k+1}, \tag{22}$$

where η is the learning-rate and each $E_k \in \mathbb{R}^d$ is a rotationally symmetric α-stable random vector.

In the experiments, we systematically vary a as well as the tail-index of the additive noise, α. For all experiments we set $\eta = 0.1$ and ran the algorithm for 3000 iterations. We set $n = 1000$

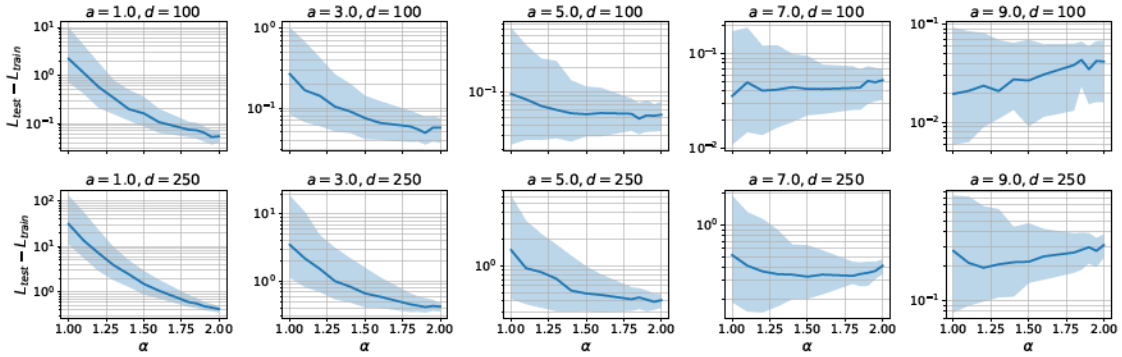

Figure 1: Results of the synthetic data experiments with varying a, α, and d. Each experiment was repeated 200 times with $n = 1000$. The lines correspond to the median, and the shaded areas are the interquartile ranges.

and varied d to be 100 or 250. The order of the loss function f was selected to be $p = 1$. For each experimental setting, we repeated the experiment 200 times, where after sampling a population $N = 100000$ observations, for each replication we sampled $n = 1000$ with replacement from within this population. The generalization error was computed to be the difference between loss computed on the replication sample of size n and the population of size N. To prevent numerical issues, the noise was scaled with a constant of 0.1 in all experiments, which corresponds to choosing $\Sigma = (1/10)I$.

The results are presented in Figure 1, and corroborate the trend predicted by Theorem 5. As a grows, the variance of the input increases, leading the map $\alpha \mapsto c(\alpha)$ to become increasing for $\alpha \in [\alpha_0, 2)$ for some α_0. Since $c(\alpha)$ is the upper bound for stability $\varepsilon_{\text{stab}}$, this leads to the observed 'V-shaped' trend in generalization error for higher a values, where the inflection point corresponds to α_0 for a given experiment setting.[2]

Experiments on image data. In our second set of experiments, we consider a real image classification task, where we use plain SGD (2) *without* adding explicit heavy-tailed noise and monitor the effect of the heavy-tails that are *inherently* introduced by SGD, as shown in Gürbüzbalaban et al. (2021); Hodgkinson and Mahoney (2021). In this context, we will view the SDE (5) as a proxy to the original SGD recursion near a local minimum, so that a quadratic approximation would be pertinent.

Here, we train two fully connected neural networks (FCN) of different depths (4 vs. 6) as well as a convolutional neural network (CNN) on the MNIST, CIFAR-10, and CIFAR-100 datasets (LeCun et al., 2010; Krizhevsky, 2009). We train these models under different, constant learning rates (η) and with batch sizes (b) of 50 or 100, producing models trained under a wide range of η/b values. The models are trained until convergence, where the convergence criteria for MNIST and CIFAR-10 is a training negative log-likelihood (NLL) of $< 5 \times 10^{-5}$ and a training accuracy of 100%, and for CIFAR-100 these are a NLL of $< 1 \times 10^{-2}$ and a training accuracy of $> 99\%$.

2. We note that the rather large error bars in Figure 1 are caused by the randomness coming from the heavy-tails (i.e., not by the randomness due to the choice of datasets). As we are essentially trying to compute the expectation of a heavy-tailed random variable by using a finite number of samples, these errors bars are not surprising as the task is notoriously difficult (Lugosi and Mendelson, 2019).

For the estimation of the trained networks' tail indices, we used the multivariate estimator proposed in (Mohammadi et al., 2015, Corollary 2.4)[3], which is previously used in various related neural network research (Şimşekli et al., 2020; Gürbüzbalaban et al., 2021; Zhou et al., 2020; Barsbey et al., 2021). Since this estimator assumes a stable distribution, after convergence we obtained 1000 iterations of SGD and computed the average to be used in this estimation, based on the generalized central limit theorem (Gürbüzbalaban et al., 2021, Corollary 11), which demonstrated that the ergodic averages of heavy-tailed SGD iterates converge to a multivariate stable distribution. Before estimating the parameters, we centered the parameters with median values. Each layer's tail-index estimation was conducted separately, which were in turn averaged to produce a single tail-index for every model, as in Barsbey et al. (2021). See the Appendix for further details.

Previous literature demonstrated that (i) training neural networks with larger η/b values lead to heavy-tailed parameters (Gürbüzbalaban et al., 2021) and (ii) networks with heavier-tailed parameters are more likely to generalize Şimşekli et al. (2020); Barsbey et al. (2021). Here, Figure 2 demonstrates that networks with highest α (light-tails) consistently perform worst in terms of generalization and the performance improves as the α decreases until some *threshold*. This outcome is in line with the predictions of our theoretical results, which suggest a 'V-shaped' behavior for the relation between generalization and α, as opposed to Şimşekli et al. (2020); Barsbey et al. (2021). As a final remark, here the values of α are larger compared with the synthetic experiments; however, we shall emphasize that such values for α still indicate strong heavy tails.

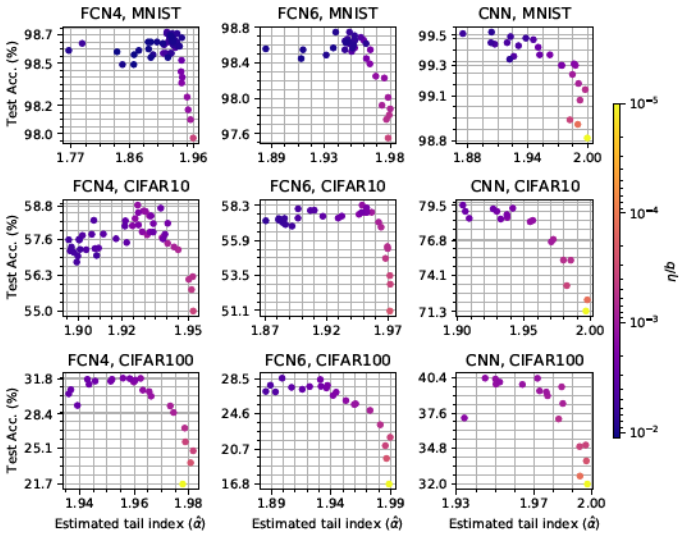

Figure 2: Test accuracy vs. mean estimated tail-index ($\hat{\alpha}$) for each model. Color: training η/b ratio.

5. Conclusion

We established novel links between the tail behavior and generalization properties of SGD building on the notion of algorithmic stability. We focused on quadratic optimization and considered a heavy-tailed SDE previously proposed as a proxy to SGD dynamics. We then proved uniform stability bounds which uncover several phenomena about the effect of the heaviness of the tails on the generalization. We also established lower bounds which show that our stability bounds are tight in terms of the heaviness of the tails. We then extended our results to the finite-time, and to the discrete-time cases and showed that similar results hold. We finally supported our theory on a

3. We note that this estimator has been shown to be consistent; yet, we do not have an non-asymptotic understanding of the esimates (Mohammadi et al., 2015).

variety of experiments. Future work includes extending our work to explore the relation between distributional robustness and heavy-tails Das et al. (2021).

Acknowledgments

A.R is supported by the a Marie Sklodowska-Curie Fellowship (project NN-OVEROPT 101030817). M.G.'s research is supported in part by the grants Office of Naval Research Award Number N00014-21-1-2244, National Science Foundation (NSF) CCF-1814888, NSF DMS-2053485, NSF DMS-1723085. U.Ş.'s research is supported by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute) and the European Research Council Starting Grant DYNASTY – 101039676. L.Z. is grateful to the support from a Simons Foundation Collaboration Grant and the grants NSF DMS-2053454, NSF DMS-2208303 from the National Science Foundation.

References

Alnur Ali, Edgar Dobriban, and Ryan Tibshirani. The implicit regularization of stochastic gradient flow for least squares. In *International Conference on Machine Learning*, pages 233–244. PMLR, 2020.

David Applebaum. *Lévy Processes and Stochastic Calculus*. Cambridge University Press, Cambridge, UK, second edition, 2009.

Melih Barsbey, Milad Sefidgaran, Murat A Erdogdu, Gaël Richard, and Umut Şimşekli. Heavy Tails in SGD and Compressibility of Overparametrized Neural Networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 29364–29378. Curran Associates, Inc., 2021.

Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, and Kunal Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. In *Advances in Neural Information Processing Systems*, volume 33, pages 4381–4391, 2020.

Mikhail Belkin, Irina Matveeva, and Partha Niyogi. Regularization and semi-supervised learning on large graphs. In *International Conference on Computational Learning Theory*, pages 624–638. Springer, 2004.

Jean Bertoin. *Lévy Processes*. Cambridge University Press, Cambridge, UK, 1996.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2(Mar):499–526, 2002.

Olivier Bousquet, Yegor Klochkov, and Nikita Zhivotovskiy. Sharper bounds for uniformly stable algorithms. In *Conference on Learning Theory*, pages 610–626. PMLR, 2020.

Zachary Charles and Dimitris Papailiopoulos. Stability and generalization of learning algorithms that converge to global optima. In *International Conference on Machine Learning*, pages 745–754. PMLR, 2018.

Yuansi Chen, Chi Jin, and Bin Yu. Stability and convergence trade-off of iterative optimization algorithms. *arXiv preprint arXiv:1804.01619*, 2018.

Rama Cont and Peter Tankov. *Financial Modelling with Jump Processes*. Chapman and Hall/CRC, 2004.

Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Algorithms for learning kernels based on centered alignment. *Journal of Machine Learning Research*, 13:795–828, 2012.

Bikramjit Das, Anulekha Dhara, and Karthik Natarajan. On the heavy-tail behavior of the distributionally robust newsvendor. *Operations Research*, 69(4):1077–1099, 2021.

Andre Elisseeff, Theodoros Evgeniou, Massimiliano Pontil, and Leslie Pack Kaelbing. Stability of randomized learning algorithms. *Journal of Machine Learning Research*, 6(3):55–79, 2005.

Tyler Farghly and Patrick Rebeschini. Time-independent generalization bounds for SGLD in non-convex settings. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Vitaly Feldman and Jan Vondrak. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In *Conference on Learning Theory*, pages 1270–1279. PMLR, 2019.

Leopold Flatto. The dixie cup problem and FKG inequality. *High Frequency*, 2(3-4):169–174, 2019.

Izrail Moiseevic Gelfand and Georgij Evgenevic Shilov. *Generalized Functions. Vol. 1, Properties and Operations*. Academic Press, 1969.

Mert Gürbüzbalaban, Umut Şimşekli, and Lingjiong Zhu. The heavy-tail phenomenon in SGD. In *International Conference on Machine Learning*, pages 3964–3975. PMLR, 2021.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 1225–1234. PMLR, 2016.

Fengxiang He and Dacheng Tao. Recent advances in deep learning theory. *arXiv preprint arXiv:2012.10931*, 2020.

Liam Hodgkinson and Michael Mahoney. Multiplicative noise and heavy tails in stochastic optimization. In *International Conference on Machine Learning*, pages 4262–4274. PMLR, 2021.

Liam Hodgkinson, Umut Simsekli, Rajiv Khanna, and Michael Mahoney. Generalization bounds using lower tail exponents in stochastic optimizers. In *International Conference on Machine Learning*, pages 8774–8795. PMLR, 2022.

Yegor Klochkov and Nikita Zhivotovskiy. Stability and deviation optimal risk bounds with convergence rate $o(1/n)$. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Ilja Kuzborskij and Christoph Lampert. Data-dependent stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 2815–2824. PMLR, 2018.

Fabian Latorre, Leello Tadesse Dadi, Paul Rolland, and Volkan Cevher. The effect of the intrinsic dimension on the generalization of quadratic classifiers. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Yann LeCun, Corinna Cortes, and CJ Burges. MNIST handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

Yunwen Lei and Yiming Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *International Conference on Machine Learning*, pages 5809–5819. PMLR, 2020.

Paul Lévy. Théorie de l'addition des variables aléatoires. *Gauthiers-Villars, Paris*, 1937.

Guan-Horng Liu and Evangelos A Theodorou. Deep learning theory review: An optimal control and dynamical systems perspective. *arXiv preprint arXiv:1908.10920*, 2019.

Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019.

Charles H Martin and Michael W Mahoney. Traditional and heavy tailed self regularization in neural network models. In *International Conference on Machine Learning*, pages 4284–4293. PMLR, 2019.

Emilia P. Martins. Estimating the rate of phenotypic evolution from comparative data. *The American Naturalist*, 144(2):193–209, 1994.

Hiroki Masuda. On multidimensional Ornstein-Uhlenbeck processes driven by a general Lévy process. *Bernoulli*, 10(1):97–120, 2004.

Andreas Maurer and Tommi Jaakkola. Algorithmic stability and meta-learning. *Journal of Machine Learning Research*, 6(6):967–994, 2005.

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

Wenlong Mou, Liwei Wang, Xiyu Zhai, and Kai Zheng. Generalization bounds of SGLD for non-convex learning: Two theoretical viewpoints. In *Conference on Learning Theory*, pages 605–638. PMLR, 2018.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

Grigorios A Pavliotis. *Stochastic Processes and Applications: Diffusion Processes, the Fokker-Planck and Langevin Equations*, volume 60. Springer, New York, 2014.

Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient Langevin dynamics: A nonasymptotic analysis. In *Conference on Learning Theory*, pages 1674–1703. PMLR, 2017.

Gennady Samorodnitsky and Murad S. Taqqu. *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance*. Chapman & Hall, New York, 1994.

Aastha M. Sathe and N. S. Upadhye. Estimation of the parameters of multivariate stable distributions. *Communications in Statistics - Simulation and Computation*, 51(10):5897–5914, 2022.

Ken-iti Sato and Makoto Yamazato. Operator-self-decomposable distributions as limit distributions of processes of Ornstein-Uhlenbeck type. *Stochastic Processes and their Applications*, 17:73–100, 1984.

Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556 [cs]*, April 2015.

Umut Şimşekli, Mert Gürbüzbalaban, Thanh Huy Nguyen, Gaël Richard, and Levent Sagun. On the heavy-tailed theory of stochastic gradient descent for deep neural networks. *arXiv preprint arXiv:1912.00018*, 2019.

Umut Şimşekli, Ozan Sener, George Deligiannidis, and Murat A Erdogdu. Hausdorff dimension, heavy tails, and generalization in neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 5138–5151. Curran Associates, Inc., 2020.

George Tzagkarakis, John P Nolan, and Panagiotis Tsakalides. Compressive sensing of temporally correlated sources using isotropic multivariate stable laws. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 1710–1714. IEEE, 2018.

George Eugene Uhlenbeck and Leonard S. Ornstein. On the theory of Brownian motion. *Physical Review*, 36(5):823–841, 1930.

Oldrich Vasicek. An equilibrium characterization of the term structure. *Journal of Financial Economics*, 5(2):177–188, 1977.

Hongjian Wang, Mert Gürbüzbalaban, Lingjiong Zhu, Umut Şimşekli, and Murat A Erdogdu. Convergence rates of stochastic gradient descent under infinite noise variance. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Xinxing Wu and Qiang Cheng. Algorithmic stability and generalization of an unsupervised feature selection algorithm. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Longjie Xie and Xicheng Zhang. Ergodicity of stochastic differential equations with jumps and singular coefficients. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 56(1): 175–229, 2020.

Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? In *Advances in Neural Information Processing Systems*, volume 33, pages 15383–15393, 2020.

Pan Zhou, Jiashi Feng, Chao Ma, Caiming Xiong, Steven Chu Hong Hoi, and Weinan E. Towards theoretically understanding why SGD generalizes better than ADAM in deep learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21285–21296. Curran Associates, Inc., 2020.

Appendix

The Appendix is organized as follows:

- In Section A, we provide the background details about characterizing the stationary distribution of a Lévy-driven OU process.
- In Section B, we characterize the finite-time distribution of a Lévy-driven OU process.
- In Section C, we characterize the distributions of a discrete-time Lévy-driven OU process.
- In Section D, we provide the proofs for the 1-dimensional case.
- In Section E, we prove the results for least square in d-dimension.
- In Section F, we provide theory for the discretized SDE.
- In Section G, we extend the d-dimensional result for general preconditioner PSD Σ.
- In Section H, we discuss useful results which we utilize in proving our results for d-dimensional case.
- In Section I, we provide further details about our experimental setup.

Appendix A. Characterizing the Stationary Distribution of a Lévy-Driven OU Process

In this section, we review the technical background of characterizing the stationary distribution of an Ornstein-Uhlenbeck process driven by a general Lévy process. Consider an Ornstein-Uhlenbeck process driven by a general Lévy process

$$\mathrm{d}\theta_t = -A\theta_t \mathrm{d}t + \mathrm{d}\mathrm{Z}_t, \tag{23}$$

where Z_t is a general Lévy process. One particular example is $\mathrm{Z}_t = \Sigma \mathrm{L}_t^\alpha$ so that

$$\mathrm{d}\theta_t = -A\theta_t \mathrm{d}t + \Sigma \mathrm{d}\mathrm{L}_t^\alpha. \tag{24}$$

Under some regularity conditions on A and the Lévy measure of Z, θ_t in (23) admits a unique invariant distribution π, and the class of all possible π's forms the class of all A-self-decomposable distributions; see Masuda (2004) and the references therein.

Let $d \in \mathbb{N}$ and Z_t is a d-dimensional Lévy process such that $Z_0 = 0$ a.s. and Z_t admits the generating triplet (b, C, ν), that is, $b \in \mathbb{R}^d$, C is a $d \times d$ symmetric non-negative definite matrix and

ν is a σ-finite measure on $\mathbb{R}^d$ satisfying $\nu(\{0\}) = 0$ and $\int_{\mathbb{R}^d} \min(1, \|z\|^2)\nu(dz) < \infty$ for which Z_t has the characteristic function

$$\varphi_t(u) := \mathbb{E}\left[e^{\mathrm{i}\langle u, \mathrm{Z}_t\rangle}\right] = \exp\left(t\left\{\mathrm{i}u^\top b - \frac{1}{2}u^\top C u + \int_{\mathbb{R}^d}\left(e^{\mathrm{i}u^\top z} - 1 - \mathrm{i}u^\top z 1_{\|z\|\le 1}\right)\nu(dz)\right\}\right) \tag{25}$$

for any $u \in \mathbb{R}^d$ and $t > 0$.

Lemma 7 (Theorems 4.1. and 4.2. in Sato and Yamazato (1984)) *Assume that A is a $d \times d$ matrix such that the real parts of all its eigenvalues are positive. Moreover, assume that*

$$\int_{\|z\|>1} \log\|z\|\nu(\mathrm{d}z) < \infty.$$

Then, θ_t in (23) admits a unique invariant distribution π whose characteristic function is given by

$$\int_{\mathbb{R}^d} e^{\mathrm{i}\langle u,x\rangle}\pi(\mathrm{d}x) = \exp\left(\int_0^\infty \log\varphi_1\left(e^{-sA^\top}u\right)\mathrm{d}s\right), \tag{26}$$

for any $u \in \mathbb{R}^d$. In particular, the generating triplet of the limiting distribution is $(b_\infty, C_\infty, \nu_\infty)$, where

$$b_\infty = A^{-1}b + \int_{\mathbb{R}^d}\int_0^\infty e^{-sA}z\left(1_{\|e^{-sA}z\|\le 1} - 1_{\|z\|\le 1}\right)\mathrm{d}s\nu(\mathrm{d}z), \tag{27}$$

$$C_\infty = \int_0^\infty e^{-sA}Ce^{-sA^\top}\mathrm{d}s, \tag{28}$$

$$\nu_\infty(E) = \int_0^\infty \nu\left(e^{sA}E\right)\mathrm{d}s, \qquad \text{for any } E \in \mathcal{B}(\mathbb{R}^d). \tag{29}$$

By using Lemma 7, we can easily obtain the following result.

Lemma *[Restatement of Lemma 3] Assume that A is a real symmetric matrix with all the eigenvalues being positive. Then (24) admits a unique stationary distribution*

$$\int_{\mathbb{R}^d} e^{\mathrm{i}\langle u,x\rangle}\pi(\mathrm{d}x) = \exp\left(-\int_0^\infty \left\|\Sigma^\top e^{-sA}u\right\|_2^\alpha \mathrm{d}s\right), \qquad \text{for any } u \in \mathbb{R}^d. \tag{30}$$

Proof In our (24), A is a real symmetric matrix with positive eigenvalues and $\mathrm{Z}_t = \Sigma\mathrm{L}_t^\alpha$, where L_t^α is a rotationally symmetric α-stable Lévy process and moreover, α-stable Lévy measure satisfies the condition $\int_{\|z\|>1}\log\|z\|\nu(dz) < \infty$, so that the condition in Lemma 7 is satisfied, and we conclude that (24) admits a unique stationary distribution say π.

Moreover, in our case, the characteristic function of $\mathrm{Z}_t = \Sigma\mathrm{L}_t^\alpha$ is given by

$$\varphi_1(u) = \mathbb{E}\left[e^{\mathrm{i}\langle u, \Sigma\mathrm{L}_1^\alpha\rangle}\right] = \mathbb{E}\left[e^{\mathrm{i}\langle \Sigma^\top u, \mathrm{L}_1^\alpha\rangle}\right]e^{-\|\Sigma^\top u\|_2^\alpha}. \tag{31}$$

Therefore, by Lemma 7, the unique invariant distribution π of (24) has the following characteristic function:

$$\int_{\mathbb{R}^d} e^{\mathrm{i}\langle u,x\rangle}\pi(\mathrm{d}x) = \exp\left(-\int_0^\infty \left\|\Sigma^\top e^{-sA}u\right\|_2^\alpha \mathrm{d}s\right). \tag{32}$$

This completes the proof. ■

Remark 8 *(i) We can also characterize the generating triplet for the limiting distribution π in the above lemma according to Lemma 7. In our case, $b = 0$ in (27), $C = 0$ in (28), and ν is the Lévy measure for $\Sigma \mathrm{L}_t^\alpha$ in (29).*

(ii) In general, it is not possible to further simplify the expression (30) except for some special cases. For example, when $A = I$. For any $s \geq 0$, $e^{-sI} = \sum_{k=0}^{\infty} \frac{(-sI)^k}{k!} = \sum_{k=0}^{\infty} \frac{(-s)^k}{k!} I = e^{-s} I$. Therefore, when $A = I$, we can compute from (30) that

$$\int_{\mathbb{R}^d} e^{\mathrm{i}\langle u,x\rangle} \pi(\mathrm{d}x) = \exp\left(- \int_0^\infty e^{-s\alpha} \left\| \Sigma^\top u \right\|_2^\alpha \mathrm{d}s \right) = e^{-\frac{1}{\alpha} \left\| \Sigma^\top u \right\|_2^\alpha}. \tag{33}$$

Hence, in this special case, the limiting distribution is $\alpha^{\frac{-1}{\alpha}} \Sigma \mathrm{L}_1^\alpha$.

Appendix B. Characterizing the Finite-Time Distribution of a Lévy-Driven OU Process

In this section, we derive the characteristic function for the finite-time distribution of a Lévy-driven OU process. We recall from equation (24)

$$\mathrm{d}\theta_t = -A\theta_t \mathrm{d}t + \Sigma \mathrm{d}\mathrm{L}_t^\alpha. \tag{34}$$

We have the following technical lemma that computes the characteristic function of θ_t at any finite time $t > 0$.

Lemma 9 *For any $t > 0$ and $u \in \mathbb{R}^d$, we have*

$$\mathbb{E}\left[e^{\mathrm{i}u^\top \theta_t} \right] = e^{\mathrm{i}u^\top e^{-At}\theta_0} e^{-\int_0^t \left\| \Sigma^\top e^{-sA} u \right\|_2^\alpha \mathrm{d}s}.$$

Proof We can solve the Lévy-driven SDE (34) and obtain

$$\theta_t = e^{-At}\theta_0 + \int_0^t e^{-A(t-s)} \Sigma \mathrm{d}\mathrm{L}_s^\alpha, \tag{35}$$

such that for any $u \in \mathbb{R}^d$, we have

$$\begin{aligned} \mathbb{E}\left[e^{\mathrm{i}u^\top \theta_t} \right] &= e^{\mathrm{i}u^\top e^{-At}\theta_0} \mathbb{E}\left[e^{\mathrm{i}\int_0^t u^\top e^{-A(t-s)} \Sigma \mathrm{d}\mathrm{L}_s^\alpha} \right] \\ &= e^{\mathrm{i}u^\top e^{-At}\theta_0} e^{-\int_0^t \left\| \Sigma^\top e^{-(t-s)A} u \right\|_2^\alpha \mathrm{d}s} = e^{\mathrm{i}u^\top e^{-At}\theta_0} e^{-\int_0^t \left\| \Sigma^\top e^{-sA} u \right\|_2^\alpha \mathrm{d}s}. \end{aligned}$$

This completes the proof. ■

We recall from (10)-(11) that

$$\mathrm{d}\theta_t = -\frac{1}{n}\left(X^\top X \theta_t - X^\top y \right) \mathrm{d}t + \Sigma \mathrm{d}\mathrm{L}_t^\alpha, \tag{36}$$

$$\mathrm{d}\hat{\theta}_t = -\frac{1}{n}\left(\hat{X}^\top \hat{X} \hat{\theta}_t - \hat{X}^\top \hat{y} \right) \mathrm{d}t + \Sigma \mathrm{d}\mathrm{L}_t^\alpha. \tag{37}$$

For the sake of simplicity, we take $y = \hat{y} = 0$ and $\theta_0 = \hat{\theta}_0 = 0$. We denote $\psi_\theta(t,u) := \mathbb{E}\left[e^{\langle \mathrm{i}u, \theta_t\rangle}\right]$ and $\psi_{\hat{\theta}}(t,u) := \mathbb{E}\left[e^{\langle \mathrm{i}u, \hat{\theta}_t\rangle}\right]$. Hence, we obtain from Lemma 9 that

$$\psi_\theta(t,u) = \exp\left(-\int_0^t \left\|\Sigma^\top e^{-s\frac{1}{n}(X^\top X)}u\right\|_2^\alpha \mathrm{d}s\right), \tag{38}$$

$$\psi_{\hat{\theta}}(t,u) = \exp\left(-\int_0^t \left\|\Sigma^\top e^{-s\frac{1}{n}(\hat{X}^\top \hat{X})}u\right\|_2^\alpha \mathrm{d}s\right). \tag{39}$$

Appendix C. Heavy-Tailed Discretized SDE on Least Squares Regression

In this section, we introduce heavy-tailed discretized SGD for the least square regression.

In the context of algorithmic stability, we assume that we have two training datasets (X, y) and $(\hat{X}, \hat{y})$ that differ in only one data point. Without loss of generality, we have

$$\hat{X} = \left[x_1^\top, x_2^\top, \ldots, \tilde{x}_i^\top, \ldots, x_n^\top\right] \in \mathbb{R}^{n\times d}$$

and

$$\hat{y} = [y_1, y_2, \ldots, \tilde{y}_i, \ldots, y_n] \in \mathbb{R}^n.$$

We consider the following discretized heavy-tailed SDE for the ERM problem as defined in (1):

$$\theta_{k+1} = \theta_k - \frac{\eta}{n}\left(X^\top X\theta_k - X^\top y\right) + \eta^{1/\alpha}\Sigma S_{k+1}, \tag{40}$$

$$\hat{\theta}_{k+1} = \hat{\theta}_k - \frac{\eta}{n}\left(\hat{X}^\top \hat{X}\hat{\theta}_k - \hat{X}^\top \hat{y}\right) + \eta^{1/\alpha}\Sigma S_{k+1}, \tag{41}$$

where $\eta > 0$ is the stepsize and $\Sigma \in \mathbb{R}^{d\times d}$ is a real-valued matrix and S_k are i.i.d. alpha-stable random vectors with the characteristic function:

$$\mathbb{E}\left[e^{\mathrm{i}\langle u, S_k\rangle}\right] = e^{-\|u\|_2^\alpha}, \qquad \text{for any } u \in \mathbb{R}^d. \tag{42}$$

We denote $\psi_\theta(k,u) := \mathbb{E}\left[e^{\langle \mathrm{i}u, \theta_k\rangle}\right]$ and $\psi_{\hat{\theta}}(k,u) := \mathbb{E}\left[e^{\langle \mathrm{i}u, \hat{\theta}_k\rangle}\right]$. For the sake of simplicity, we take $y = \hat{y} = 0$ and $\theta_0 = \hat{\theta}_0 = 0$. By Lemma 10, we obtain

$$\psi_\theta(k,u) = \exp\left(-\eta\sum_{j=0}^{k-1}\left\|\Sigma^\top\left(I - \frac{\eta}{n}\left(X^\top X\right)\right)^j u\right\|_2^\alpha\right), \tag{43}$$

$$\psi_{\hat{\theta}}(k,u) = \exp\left(-\eta\sum_{j=0}^{k-1}\left\|\Sigma^\top\left(I - \frac{\eta}{n}\left(\hat{X}^\top \hat{X}\right)\right)^j u\right\|_2^\alpha\right), \tag{44}$$

which will be the key ingredients to obtain the algorithmic stability results.

Here below, we derive the characteristic function for distributions of a discrete-time Lévy-driven OU process:

$$\theta_{k+1} = \theta_k - \eta A\theta_k + \eta^{1/\alpha}\Sigma S_{k+1}, \tag{45}$$

where A is a real symmetric matrix and S_k are i.i.d. alpha-stable random vectors with the characteristic function:

$$\mathbb{E}\left[e^{\mathrm{i}\langle u, S_k\rangle}\right] = e^{-\|u\|_2^\alpha}, \qquad \text{for any } u \in \mathbb{R}^d. \tag{46}$$

We can compute the characteristic function of the finite-time distribution of the discrete-time Lévy-driven OU process (45) as follows.

Lemma 10 *Assume that A is a real symmetric matrix. For any $k \in \mathbb{N}$ and for any $u \in \mathbb{R}^d$,*

$$\mathbb{E}\left[e^{\mathrm{i}\langle u, \theta_k\rangle}\right] = e^{\mathrm{i}\langle (I-\eta A)^k u, \theta_0\rangle} e^{-\eta \sum_{j=0}^{k-1} \left\|\Sigma^\top (I-\eta A)^j u\right\|_2^\alpha}. \tag{47}$$

Proof We can compute from (45) that for any $u \in \mathbb{R}^d$,

$$\begin{aligned}\mathbb{E}\left[e^{\mathrm{i}\langle u, \theta_{k+1}\rangle}\right] &= \mathbb{E}\left[e^{\mathrm{i}\langle u, (I-\eta A)\theta_k\rangle}\right] \mathbb{E}\left[e^{\mathrm{i}\langle u, \eta^{1/\alpha}\Sigma S_k\rangle}\right] \\ &= \mathbb{E}\left[e^{\mathrm{i}\langle (I-\eta A)u, \theta_k\rangle}\right] \mathbb{E}\left[e^{\mathrm{i}\langle \eta^{1/\alpha}\Sigma^\top u, S_k\rangle}\right] = \mathbb{E}\left[e^{\mathrm{i}\langle (I-\eta A)u, \theta_k\rangle}\right] e^{-\eta\|\Sigma^\top u\|_2^\alpha},\end{aligned} \tag{48}$$

and we can further compute that

$$\mathbb{E}\left[e^{\mathrm{i}\langle (I-\eta A)u, \theta_k\rangle}\right] = \mathbb{E}\left[e^{\mathrm{i}\langle (I-\eta A)^2 u, \theta_{k-1}\rangle}\right] e^{-\eta\|\Sigma^\top (I-\eta A)u\|_2^\alpha}. \tag{49}$$

Hence, iteratively, we obtain

$$\mathbb{E}\left[e^{\mathrm{i}\langle u, \theta_k\rangle}\right] = e^{\mathrm{i}\langle (I-\eta A)^k u, \theta_0\rangle} e^{-\eta \sum_{j=0}^{k-1} \left\|\Sigma^\top (I-\eta A)^j u\right\|_2^\alpha}. \tag{50}$$

This completes the proof. ■

We can derive from Lemma 10 the characteristic function of the stationary distribution of the discrete-time Lévy-driven OU process (45) as follows.

Corollary 11 *Assume that A is a real symmetric matrix with all the eigenvalues being positive and less than $1/\eta$. Then, for any $u \in \mathbb{R}^d$,*

$$\mathbb{E}\left[e^{\mathrm{i}\langle u, \theta_\infty\rangle}\right] = e^{-\eta \sum_{j=0}^{\infty} \|\Sigma^\top (I-\eta A)^j u\|_2^\alpha}. \tag{51}$$

Proof When A is a real symmetric matrix with all the eigenvalues being positive and less than $1/\eta$, we have $\|I - \eta A\| < 1$ and it follows that

$$\left|\left\langle (I-\eta A)^k u, \theta_0 \right\rangle\right| \le \|I-\eta A\|^k \cdot \|u\| \cdot \|\theta_0\| \to 0, \tag{52}$$

as $k \to \infty$, and moreover,

$$\left\|\Sigma^\top (I-\eta A)^j u\right\|_2^\alpha \le \left\|\Sigma^\top\right\| \cdot \|I-\eta A\|^j \cdot \|u\|, \tag{53}$$

which is summable over j and hence the result follows from Lemma 10. The proof is complete. ■

Appendix D. Proofs for the 1-Dimensional Case

In this section, we provide the proofs for the one-dimensional case. In the next lemma, we first bound the difference between the characteristic functions of the stationary distributions.

Lemma 12 *For two matrices $X \in \mathbb{R}^n$ and $\hat{X} \in \mathbb{R}^n$ as defined earlier, the absolute value of difference between the characteristic function for stationary distribution at any $u \in \mathbb{R}$ corresponding to one-dimensional rotation invariant processes in equations (16) and (17) ($d = 1$) is bounded as*

$$\left|\psi_\theta(u) - \psi_{\hat{\theta}}(u)\right| \leq \frac{x_i^2 - \tilde{x}_i^2}{\|\hat{X}\|_2^2} \left(|u|^\alpha \frac{n}{\alpha \|X\|_2^2} \exp\left(-|u|^\alpha \frac{n}{\alpha \|X\|_2^2} \right) \right).$$

Proof We can compute that

$$\begin{aligned}
&\left|\psi_\theta(u) - \psi_{\hat{\theta}}(u)\right| \\
&= \left| \exp\left(- \int_0^\infty \left| e^{-s\frac{1}{n}\|X\|_2^2} u \right|^\alpha \mathrm{d}s \right) - \exp\left(- \int_0^\infty \left| e^{-s\frac{1}{n}\|\hat{X}\|_2^2} u \right|^\alpha \mathrm{d}s \right) \right| \\
&= \left| \exp\left(- \int_0^\infty \left| e^{-s\frac{1}{n}\|X\|_2^2} u \right|^\alpha \mathrm{d}s \right) \left(1 - \exp\left(\int_0^\infty \left| e^{-s\frac{1}{n}\|X\|_2^2} u \right|^\alpha \mathrm{d}s \right.\right.\right. \\
&\qquad\qquad \left.\left.\left. - \int_0^\infty \left| e^{-s\frac{1}{n}\|\hat{X}\|_2^2} u \right|^\alpha \mathrm{d}s \right) \right) \right| \qquad (54) \\
&\leq \exp\left(- \int_0^\infty \left| e^{-s\frac{1}{n}\|X\|_2^2} u \right|^\alpha \mathrm{d}s \right) \left| \int_0^\infty \left| e^{-s\frac{1}{n}\|X\|_2^2} u \right|^\alpha \mathrm{d}s - \int_0^\infty \left| e^{-s\frac{1}{n}\|\hat{X}\|_2^2} u \right|^\alpha \mathrm{d}s \right| \\
&= \exp\left(-|u|^\alpha \int_0^\infty e^{-s\frac{\alpha}{n}\|X\|_2^2} \mathrm{d}s \right) |u|^\alpha \left| \int_0^\infty e^{-s\frac{\alpha}{n}\|X\|_2^2} \mathrm{d}s - \int_0^\infty e^{-s\frac{\alpha}{n}\|\hat{X}\|_2^2} \mathrm{d}s \right| \\
&= |u|^\alpha \exp\left(-|u|^\alpha \int_0^\infty e^{-s\frac{\alpha}{n}\|X\|_2^2} \mathrm{d}s \right) \left| \frac{n}{\alpha\|X\|_2^2} - \frac{n}{\alpha\|\hat{X}\|_2^2} \right| \\
&= |u|^\alpha \exp\left(-|u|^\alpha \frac{n}{\alpha\|X\|_2^2} \right) \left| \frac{n}{\alpha\|X\|_2^2} - \frac{n}{\alpha\|\hat{X}\|_2^2} \right| \\
&= |u|^\alpha \frac{n|x_i^2 - \tilde{x}_i^2|}{\alpha\|X\|_2^2\|\hat{X}\|_2^2} \exp\left(-|u|^\alpha \frac{n}{\alpha\|X\|_2^2} \right) \\
&= \frac{|x_i^2 - \tilde{x}_i^2|}{\|\hat{X}\|_2^2} \left(|u|^\alpha \frac{n}{\alpha\|X\|_2^2} \exp\left(-|u|^\alpha \frac{n}{\alpha\|X\|_2^2} \right) \right),
\end{aligned}$$

which completes the proof. ■

Theorem *[Restatement of Theorem 4] Consider the one-dimensional loss function $f(x) = |\theta x|^p$. For any $x \sim P_X$, if we have $|x| > R$ with probability δ_1 and for any X sampled uniformly at random from the set $\mathcal{X}_n$, if we have $\|X\|_2^2 \leq \sigma^2 n$ with probability δ_2. Then,*

(i) For $\alpha \in [1, 2)$, the algorithm is not stable when $p \in [\alpha, 2]$ i.e. $\varepsilon_{stab(\mathcal{A}_{cont})}$ diverges. When $\alpha = p = 2$ then $\varepsilon_{stab}(\mathcal{A}_{cont}) \leq \frac{R^4}{\pi\sigma^4 n}$ with probability at least $1 - \delta_1 - 2\delta_2$.

(ii) For $p \in [1, \alpha)$, we have the following upper bound for the algorithmic stability,

$$\varepsilon_{stab}(\mathcal{A}_{cont}) \leq \frac{2R^{p+2}}{\pi\sigma^2 n}\Gamma(p+1)\cos\left(\frac{(p-1)\pi}{2}\right)\frac{1}{\alpha}\left(\frac{1}{\alpha\sigma^2}\right)^{\frac{p}{\alpha}}\Gamma\left(1-\frac{p}{\alpha}\right) =: c(\alpha),$$

which holds with probability at least $1-\delta_1-2\delta_2$. Furthermore, for some $\alpha_0 > 1$, if we have

$$\sigma^2 \geq \exp\left(1+\frac{2}{p}-\log\alpha_0-\phi\left(1-\frac{p}{\alpha_0}\right)\right), \tag{55}$$

where ϕ is the digamma function, then the map $\alpha \mapsto c(\alpha)$ is increasing for $\alpha \in [\alpha_0, 2)$.

(iii) The stability bound is tight in α.

Proof Closed-form expression for the Fourier transform of function $f(\theta) = |\theta x|^p$ has been given in Gelfand and Shilov (1969). However, we provide here the result for the sake of completeness. Let us compute the Fourier transform of the function $\theta \to |\theta x|^p$ for $p \in [1, 2)$.

$$\begin{aligned}
\int_{-\infty}^{\infty}|\theta x|^p e^{-\mathrm{i}u\theta}\mathrm{d}\theta =& |x|^p\int_{-\infty}^{\infty}|\theta|^p e^{-\mathrm{i}u\theta}\,\mathrm{d}\theta\\
=& |x|^p\int_0^{\infty}(e^{\mathrm{i}u\theta}+e^{-\mathrm{i}u\theta})\theta^p\,\mathrm{d}\theta\\
=& |x|^p\left[\int_0^{\infty}e^{\mathrm{i}u\theta}\theta^p\,\mathrm{d}\theta+\int_0^{\infty}e^{-\mathrm{i}u\theta}\theta^p\,\mathrm{d}\theta\right]\\
=& 2|x|^p\Gamma(p+1)\cos\left(\frac{(p+1)\pi}{2}\right)\frac{1}{|u|^{p+1}}.
\end{aligned} \tag{56}$$

From Gelfand and Shilov (1969),

$$\int_{-\infty}^{\infty}|\theta x|^2 e^{-\mathrm{i}u\theta}\mathrm{d}\theta = \frac{-2x^2}{u^2}\delta(u),$$

where $\delta(u)$ is the Dirac-delta function.

First, we get the result for $p \in [1, 2)$. We utilize the result from Lemma 12 and equation (56) to get,

$$\begin{aligned}
\varepsilon_{\text{stab}}(\mathcal{A}_{\text{cont}}) =& \sup_{X\cong\hat{X}}\sup_{x\in\mathcal{X}}\left[\frac{1}{(2\pi)}\int_{\mathbb{R}}|\psi_\theta(u)-\psi_{\hat{\theta}}(u)|\left|\int_{-\infty}^{\infty}|\theta x|^p e^{-\mathrm{i}u\theta}\,\mathrm{d}\theta\right|\,\mathrm{d}u\right]\\
=& \sup_{X\cong\hat{X}}\sup_{x\in\mathcal{X}}\frac{|x|^p}{\pi}\Gamma(p+1)\cos\left(\frac{(p-1)\pi}{2}\right)\frac{|x_i^2-\tilde{x}_i^2|}{\|\hat{X}\|_2^2}\frac{n}{\alpha\|X\|_2^2}\\
&\qquad\cdot\int_{\mathbb{R}}\left(|u|^\alpha\exp\left(-|u|^\alpha\frac{n}{\alpha\|X\|_2^2}\right)\right)\frac{1}{|u|^{p+1}}\,\mathrm{d}u\\
\leq& \sup_{X\cong\hat{X}}\sup_{x\in\mathcal{X}}\frac{2|x|^{p+2}}{\pi\|\hat{X}\|_2^2}\Gamma(p+1)\cos\left(\frac{(p-1)\pi}{2}\right)\\
&\qquad\cdot\int_0^{\infty}u^{\alpha-p-1}\frac{n}{\alpha\|X\|_2^2}\exp\left(-|u|^\alpha\frac{n}{\alpha\|X\|_2^2}\right)\,\mathrm{d}u.
\end{aligned}$$

In the above integral, by substituting $u^{\alpha}\frac{n}{\alpha\|X\|^2}$ with t so that

$$\mathrm{d}t = u^{\alpha-1}\frac{n}{\|X\|_2^2}\mathrm{d}u, \text{ and } \frac{1}{u^p} = \left(\frac{n}{\alpha\|X\|_2^2}\right)^{\frac{p}{\alpha}} t^{-p/\alpha}, \tag{57}$$

we have,

$$\varepsilon_{\text{stab}}(\mathcal{A}_{\text{cont}}) \leq \sup_{X\cong\hat{X}}\sup_{x\in\mathcal{X}} \frac{2|x|^{p+2}}{\pi\alpha\|\hat{X}\|_2^2}\Gamma(p+1)\cos\left(\frac{(p-1)\pi}{2}\right)\left(\frac{n}{\alpha\|X\|_2^2}\right)^{\frac{p}{\alpha}}\int_0^{\infty} t^{-p/\alpha}e^{-t}\,\mathrm{d}t. \tag{58}$$

It is clear that, the above integral diverges for $p \geq \alpha$, hence the algorithm is not stable for $p \in [1, 2)$. Now, we check the case for $p = 2$. For $p = 2$, we have,

$$\varepsilon_{\text{stab}}(\mathcal{A}_{\text{cont}}) \leq \sup_{X\cong\hat{X}}\sup_{x\in\mathcal{X}} \frac{|x|^4}{\pi\|\hat{X}\|_2^2}\frac{n}{\|X\|_2^2}\int_0^{\infty} u^{\alpha-2}\exp\left(-u^{\alpha}\frac{n}{\|X\|_2^2}\right)\delta(u)\,\mathrm{d}u.$$

The above integral clearly diverges for $\alpha < 2$. However, when $\alpha = 2$, then

$$\varepsilon_{\text{stab}}(\mathcal{A}_{\text{cont}}) \leq \sup_{X\cong\hat{X}}\sup_{x\in\mathcal{X}} \frac{|x|^4}{\pi\|\hat{X}\|_2^2}\frac{n}{\|X\|_2^2}.$$

If we have $|x| > R$ for any $x \sim P_X$ with probability δ_1 and $\|X\|_2^2 \leq \sigma^2 n$ for any X sampled uniformly from the set $\mathcal{X}_n$ with probability δ_2, then for $\alpha = 2$ and $p = 2$,

$$\varepsilon_{\text{stab}}(\mathcal{A}_{\text{cont}}) \leq \frac{R^4}{\pi\sigma^4 n}, \tag{59}$$

with probability at least $1 - \delta_1 - 2\delta_2$. This proves the part (i) of our result.

Next, let us prove the part (ii). From equation (58), for $p < \alpha$, we have

$$\begin{aligned}\varepsilon_{\text{stab}}(\mathcal{A}_{\text{cont}}) &\leq \sup_{X\cong\hat{X}}\sup_{x\in\mathcal{X}} \frac{2|x|^{p+2}}{\pi\alpha\|\hat{X}\|_2^2}\Gamma(p+1)\cos\left(\frac{(p-1)\pi}{2}\right)\left(\frac{n}{\alpha\|X\|_2^2}\right)^{\frac{p}{\alpha}}\int_0^{\infty} t^{-p/\alpha}e^{-t}\,\mathrm{d}t \\ &= \sup_{X\cong\hat{X}}\sup_{x\in\mathcal{X}} \frac{2|x|^{p+2}}{\pi\alpha\|\hat{X}\|_2^2}\Gamma(p+1)\cos\left(\frac{(p-1)\pi}{2}\right)\left(\frac{n}{\alpha\|X\|_2^2}\right)^{\frac{p}{\alpha}}\Gamma\left(1-\frac{p}{\alpha}\right).\end{aligned}$$

If we have $|x| > R$ for any $x \sim P_X$ with probability δ_1 and $\|X\|_2^2 \leq \sigma^2 n$ for any X sampled uniformly at random from the set $\mathcal{X}_n$ with probability δ_2, then with probability at least $1 - \delta_1 - 2\delta_2$, the following holds:

$$\varepsilon_{\text{stab}}(\mathcal{A}_{\text{cont}}) \leq \frac{2R^{p+2}}{\pi\sigma^2 n}\Gamma(p+1)\cos\left(\frac{(p-1)\pi}{2}\right)\frac{1}{\alpha}\left(\frac{1}{\alpha\sigma^2}\right)^{\frac{p}{\alpha}}\Gamma\left(1-\frac{p}{\alpha}\right) = c(\alpha).$$

Now, consider the function,

$$\Lambda(\alpha) = \frac{1}{\alpha}\left(\frac{1}{\alpha\sigma^2}\right)^{\frac{p}{\alpha}}\Gamma\left(1-\frac{p}{\alpha}\right).$$

We can compute that

$$\partial_\alpha \log \Lambda(\alpha) = \frac{p}{\alpha^2}\left[\log\alpha + \log\sigma^2 - 1 - \frac{\alpha}{p} + \phi\left(1 - \frac{p}{\alpha}\right)\right],$$

where ϕ is the digamma function. For any arbitrary α_0, if we choose

$$\sigma^2 \geq \exp\left(1 + \frac{2}{p} - \log\alpha_0 - \phi\left(1 - \frac{p}{\alpha_0}\right)\right),$$

then $\partial_\alpha \log\Lambda(\alpha) > 0$ for $\alpha \in [\alpha_0, 2)$. Hence, for all $\alpha_1, \alpha_2 \in [\alpha_0, 2)$, $\alpha_1 < \alpha_2 \Rightarrow \Lambda(\alpha_1) \leq \Lambda(\alpha_2)$. This proves that $c(\alpha)$ is an increasing map in α.

(iii). Now we show that the bound on the stability is tight for some appropriately chosen P_X. Note that we have,

$$\begin{aligned}
\varepsilon_{\text{stab}}(\mathcal{A}_{\text{cont}}) &= \sup_{X\cong\hat{X}}\sup_{x\in\mathcal{X}}\left[\frac{1}{(2\pi)}\int_{\mathbb{R}}\left|\psi_\theta(u) - \psi_{\hat{\theta}}(u)\right|\left|\int_{-\infty}^{\infty}|\theta x|^p e^{-\mathrm{i}u\theta}\,\mathrm{d}\theta\right|\,\mathrm{d}u\right] \\
&= \sup_{X\cong\hat{X}}\sup_{x\in\mathcal{X}}\underbrace{\left[\frac{1}{\pi}\int_0^{\infty}\left|\psi_\theta(u) - \psi_{\hat{\theta}}(u)\right|\left|\int_{-\infty}^{\infty}|\theta x|^p e^{-\mathrm{i}u\theta}\,\mathrm{d}\theta\right|\,\mathrm{d}u\right]}_{:=\Phi_{x,X,\hat{X}}(\mathcal{A}_{\text{cont}})}.
\end{aligned}$$

Hence, let us consider $u \geq 0$. From equation (54), we have

$$\begin{aligned}
\left|\psi_\theta(u) - \psi_{\hat{\theta}}(u)\right| &= \left|\exp\left(-\int_0^{\infty}\left|e^{-s\frac{1}{n}\|X\|_2^2}u\right|^\alpha \mathrm{d}s\right)\right. \\
&\qquad \left.\cdot\left(1 - \exp\left(\int_0^{\infty}\left|e^{-s\frac{1}{n}\|X\|_2^2}u\right|^\alpha \mathrm{d}s - \int_0^{\infty}\left|e^{-s\frac{1}{n}\|\hat{X}\|_2^2}u\right|^\alpha \mathrm{d}s\right)\right)\right| \\
&= \exp\left(-\int_0^{\infty}\left|e^{-s\frac{1}{n}\|X\|_2^2}u\right|^\alpha \mathrm{d}s\right)\left(1 - \exp\left(-|u|^\alpha\left[\frac{n}{\alpha\|\hat{X}\|_2^2} - \frac{n}{\alpha\|X\|_2^2}\right]\right)\right) \\
&= \exp\left(-|u|^\alpha\frac{n}{\alpha\|X\|_2^2}\right)\left(1 - \exp\left(-|u|^\alpha\left[\frac{n\overbrace{(x_i^2 - \tilde{x}_i^2)}^{:=\delta}}{\alpha\|X\|_2^2\|\hat{X}\|_2^2}\right]\right)\right) \\
&= \exp\left(-|u|^\alpha\frac{n}{\alpha\|X\|_2^2}\right)\left[\sum_{k=1}^{\infty}\frac{(-1)^{k+1}}{k!}|u|^{k\alpha}\left(\frac{n\delta}{\alpha\|X\|_2^2\|\hat{X}\|_2^2}\right)^k\right].
\end{aligned}$$

Hence,

$$\begin{aligned}
\Phi_{x,X,\hat{X}}(\mathcal{A}_{\text{cont}}) &= \left[\frac{1}{\pi}\int_0^{\infty}\left|\psi_\theta(u) - \psi_{\hat{\theta}}(u)\right|\left|\int_{-\infty}^{\infty}|\theta x|^p e^{-\mathrm{i}u\theta}\,\mathrm{d}\theta\right|\,\mathrm{d}u\right] \\
&= \frac{2|x|^p}{\pi}\Gamma(p+1)\cos\left(\frac{(p+1)\pi}{2}\right)\int_0^{\infty}\exp\left(-u^\alpha\frac{n}{\alpha\|X\|_2^2}\right)
\end{aligned}$$

$$\cdot \left[\sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k!} u^{k\alpha-p-1} \left(\frac{n\delta}{\alpha\|X\|_2^2\|\hat{X}\|_2^2}\right)^k\right] \mathrm{d}u. \quad (60)$$

To simplify the above term, we do need to compute the integral:

$$\int_0^{\infty} \exp\left(-\frac{u^{\alpha}n}{\alpha\|X\|_2^2}\right) \frac{u^{k\alpha}}{u^{p+1}}\, \mathrm{d}u.$$

Let us use the substitution:

$$\frac{u^{\alpha}n}{\alpha\|X\|_2^2} = t$$

such that

$$\mathrm{d}t = \frac{u^{\alpha-1}n}{\|X\|_2^2}\mathrm{d}u \quad \text{and} \quad u = \left(\frac{\alpha\|X\|_2^2}{n}\right)^{1/\alpha} t^{1/\alpha}.$$

Hence,

$$\begin{aligned}
\int_0^{\infty} \exp\left(-\frac{u^{\alpha}n}{\alpha\|X\|_2^2}\right) \frac{u^{k\alpha}}{u^{p+1}}\, \mathrm{d}u &= \int_0^{\infty} e^{-t} u^{\alpha(k-1)-p} \frac{\|X\|_2^2}{n}\, \mathrm{d}t \\
&= \frac{\|X\|_2^2}{n}\left(\frac{\alpha\|X\|_2^2}{n}\right)^{k-1-\frac{p}{\alpha}} \int_0^{\infty} e^{-t} t^{k-1-\frac{p}{\alpha}} \mathrm{d}t \\
&= \frac{\|X\|_2^2}{n}\left(\frac{\alpha\|X\|_2^2}{n}\right)^{k-1-\frac{p}{\alpha}} \Gamma\left(k-\frac{p}{\alpha}\right).
\end{aligned}$$

This implies,

$$\begin{aligned}
\Phi_{x,X,\hat{X}}(\mathcal{A}_{\text{cont}}) &= \frac{2|x|^p}{\pi}\Gamma(p+1)\cos\left(\frac{(p+1)\pi}{2}\right) \int_0^{\infty} \exp\left(-\frac{|u|^{\alpha}n}{\alpha\|X\|_2^2}\right) \\
&\qquad \cdot \sum_{k=1}^{\infty}\left[\frac{(-1)^{k+1}}{k!} u^{k\alpha-p-1}\left(\frac{n\delta}{\alpha\|X\|_2^2\|\hat{X}\|_2^2}\right)^k\right] \mathrm{d}u \\
&= \frac{2|x|^p}{\pi}\Gamma(p+1)\cos\left(\frac{(p+1)\pi}{2}\right) \\
&\qquad \cdot \sum_{k=1}^{\infty}\left[\frac{(-1)^{k+1}}{k!}\frac{\delta^k}{\alpha\|\hat{X}\|_2^{2k}}\left(\frac{n}{\alpha\|X\|_2^2}\right)^{\frac{p}{\alpha}} \Gamma\left(k-\frac{p}{\alpha}\right)\right] \\
&= \frac{2|x|^p}{\pi}\Gamma(p+1)\cos\left(\frac{(p+1)\pi}{2}\right) \sum_{k=1}^{\infty}(-1)^{k+1}\gamma_k,
\end{aligned}$$

where $\gamma_k = \frac{1}{k!}\frac{\delta^k}{\alpha\|\hat{X}\|_2^{2k}}\left(\frac{n}{\alpha\|X\|_2^2}\right)^{\frac{p}{\alpha}} \Gamma\left(k-\frac{p}{\alpha}\right)$. Let us now compute $\frac{\gamma_{k+1}}{\gamma_k}$. We have

$$\frac{\gamma_{k+1}}{\gamma_k} = \frac{\delta}{\|\hat{X}\|_2^2}\frac{\Gamma\left(k+1-\frac{p}{\alpha}\right)}{(k+1)\Gamma\left(k-\frac{p}{\alpha}\right)} = \frac{\delta}{\|\hat{X}\|_2^2}\frac{k-\frac{p}{\alpha}}{k+1} = \frac{\delta}{\|\hat{X}\|_2^2}\left(1-\frac{1+\frac{p}{\alpha}}{k+1}\right).$$

Hence, we have the following,

$$\begin{aligned}
\Phi_{x,X,\hat{X}}(\mathcal{A}_{\text{cont}}) &= \frac{2|x|^p}{\pi}\Gamma(p+1)\cos\left(\frac{(p+1)\pi}{2}\right)\left[\gamma_1\sum_{k=1}^{\infty}\left(-\frac{\delta}{\|\hat{X}\|_2^2}\right)^{k-1}\right. \\
&\qquad \left.-\gamma_1\left(1+\frac{p}{\alpha}\right)\sum_{k=1}^{\infty}\left(-\frac{\delta}{\|\hat{X}\|_2^2}\right)^{k-1}\prod_{j=1}^{k-1}\left(\frac{1}{j+1}\right)\right] \\
&\geq \frac{2|x|^p}{\pi}\Gamma(p+1)\cos\left(\frac{(p+1)\pi}{2}\right)\gamma_1\sum_{k=1}^{\infty}\left(-\frac{\delta}{\|\hat{X}\|_2^2}\right)^{k-1} \\
&= \frac{2|x|^p}{\pi}\Gamma(p+1)\cos\left(\frac{(p+1)\pi}{2}\right)\frac{\gamma_1}{1+\frac{\delta}{\|\hat{X}\|_2^2}} \\
&\geq \frac{2|x|^p}{\pi}\Gamma(p+1)\cos\left(\frac{(p+1)\pi}{2}\right)\frac{\delta}{\alpha\|\hat{X}\|_2^2}\left(\frac{n}{\alpha\|X\|_2^2}\right)^{\frac{p}{\alpha}}\Gamma\left(1-\frac{p}{\alpha}\right). \qquad (61)
\end{aligned}$$

Now, let us assume that P_X is a distribution with discrete support in range σ^2 to R with C number of support points equally spaced. Hence, with probability $(1-1/C)$, $\delta \geq c$ for some positive constant c. Hence, with with high probability,

$$\Phi_{x,X,\hat{X}}(\mathcal{A}_{\text{cont}}) \geq \frac{2\sigma^{2p}c}{R^2\pi n}\Gamma(p+1)\cos\left(\frac{(p+1)\pi}{2}\right)\frac{1}{\alpha}\left(\frac{1}{\alpha R^2}\right)^{\frac{p}{\alpha}}\Gamma\left(1-\frac{p}{\alpha}\right).$$

This completes the proof. ■

Appendix E. Proofs for Least-Square in d-Dimension

In this section, we provide the proofs for least-square in d-dimension. We start by proving the following lemma, relating the characteristic functions of the two distributions.

Lemma 13 *For two matrices $X \in \mathbb{R}^{n\times d}$ and $\hat{X} \in \mathbb{R}^{n\times d}$ as defined earlier, the absolute value of difference between the characteristic functions of the stationary distributions at any $u \in \mathbb{R}^d$ corresponding to d-dimensional rotation invariant processes in equations (16) and (17) is bounded as*

$$\left|\psi_\theta(u) - \psi_{\hat{\theta}}(u)\right| \leq \frac{2(\sigma_1+\sigma_2)\|u\|_2^\alpha}{n\alpha^2\sigma_{\min}^2}\exp\left(-\frac{\|u\|_2^\alpha}{\alpha\sigma_{\min}}\right),$$

where $\sigma_{\min}$ is the smaller of the smallest of singular values of the matrices $\frac{1}{n}X^\top X$ and $\frac{1}{n}\hat{X}^\top\hat{X}$, and $x_i x_i^\top - \tilde{x}_i\tilde{x}_i^\top = \sigma_1 v_1 v_1^\top + \sigma_2 v_2 v_2^\top$ where v_1 and v_2 are orthogonal vectors.

Proof We can compute that

$$\begin{aligned}
&\left|\psi_\theta(u) - \psi_{\hat{\theta}}(u)\right| \\
&= \left|\exp\left(-\int_0^\infty \left\|e^{-s\frac{1}{n}X^\top X}u\right\|_2^\alpha \mathrm{d}s\right) - \exp\left(-\int_0^\infty \left\|e^{-s\frac{1}{n}\hat{X}^\top\hat{X}}u\right\|_2^\alpha \mathrm{d}s\right)\right|
\end{aligned}$$

$$\leq \underbrace{\exp\left(-\int_0^\infty \left\|e^{-s\frac{1}{n}X^\top X}u\right\|_2^\alpha \mathrm{d}s\right)}_{:=B} \underbrace{\left|\int_0^\infty \left\|e^{-s\frac{1}{n}X^\top X}u\right\|_2^\alpha \mathrm{d}s - \int_0^\infty \left\|e^{-s\frac{1}{n}\hat{X}^\top \hat{X}}u\right\|_2^\alpha \mathrm{d}s\right|}_{:=C}.$$

We first consider the term C in the above equation. From Lemma 19, we have for two positive numbers a and b, and for some $1 \leq \alpha \leq 2$, we have

$$|a^\alpha - b^\alpha| \leq |a-b|(a^{\alpha-1} + b^{\alpha-1}).$$

Now,

$$\begin{aligned}
C &= \left|\int_0^\infty \left\|e^{-s\frac{1}{n}X^\top X}u\right\|_2^\alpha \mathrm{d}s - \int_0^\infty \left\|e^{-s\frac{1}{n}\hat{X}^\top \hat{X}}u\right\|_2^\alpha \mathrm{d}s\right| \\
&= \left|\int_0^\infty \left(\left\|e^{-s\frac{1}{n}X^\top X}u\right\|_2^\alpha - \left\|e^{-s\frac{1}{n}\hat{X}^\top \hat{X}}u\right\|_2^\alpha\right) \mathrm{d}s\right| \\
&\leq \int_0^\infty \left|\left\|e^{-s\frac{1}{n}X^\top X}u\right\|_2 - \left\|e^{-s\frac{1}{n}\hat{X}^\top \hat{X}}u\right\|_2\right| \left(\left\|e^{-s\frac{1}{n}X^\top X}u\right\|_2^{\alpha-1} + \left\|e^{-s\frac{1}{n}\hat{X}^\top \hat{X}}u\right\|_2^{\alpha-1}\right) \mathrm{d}s \\
&\leq \int_0^\infty \left\|e^{-s\frac{1}{n}X^\top X}u - e^{-s\frac{1}{n}\hat{X}^\top \hat{X}}u\right\|_2 \left(\left\|e^{-s\frac{1}{n}X^\top X}u\right\|_2^{\alpha-1} + \left\|e^{-s\frac{1}{n}\hat{X}^\top \hat{X}}u\right\|_2^{\alpha-1}\right) \mathrm{d}s \\
&= \int_0^\infty \left\|e^{-s\frac{1}{n}X^\top X}\left(I - e^{s\frac{1}{n}X^\top X - s\frac{1}{n}\hat{X}^\top \hat{X}}\right)u\right\|_2 \\
&\qquad \cdot \left(\left\|e^{-s\frac{1}{n}X^\top X}u\right\|_2^{\alpha-1} + \left\|e^{-s\frac{1}{n}\hat{X}^\top \hat{X}}u\right\|_2^{\alpha-1}\right) \mathrm{d}s.
\end{aligned}$$

Now, we have from the definitions,

$$X^\top X - \hat{X}^\top \hat{X} = x_i x_i^\top - \tilde{x}_i \tilde{x}_i^\top.$$

Hence,

$$C \leq \int_0^\infty \left\|e^{-s\frac{1}{n}X^\top X}\left(I - e^{s\frac{1}{n}\left(x_i x_i^\top - \tilde{x}_i \tilde{x}_i^\top\right)}\right)u\right\|_2 \left(\left\|e^{-s\frac{1}{n}X^\top X}u\right\|_2^{\alpha-1} + \left\|e^{-s\frac{1}{n}\hat{X}^\top \hat{X}}u\right\|_2^{\alpha-1}\right) \mathrm{d}s.$$

We recall that the 2-norm $\|\cdot\|_2$ for a matrix $D \in \mathbb{R}^{d\times d}$ is defined as follows:

$$\|D\|_2 = \sup_{u\in\mathbb{R}^d : \|u\|_2 = 1} \|Du\|_2$$

for $u \in \mathbb{R}^d$. We notice that

$$\min\left(\frac{1}{n}\left\|\left(X^\top X\right)u\right\|_2, \frac{1}{n}\left\|\left(\hat{X}^\top \hat{X}\right)u\right\|_2\right) \geq \sigma_{\min}\|u\|_2.$$

Hence,

$$\begin{aligned}
C \leq \int_0^\infty &\left\|e^{-s\frac{1}{n}X^\top X}u\right\|_2 \left\|I - e^{-s\frac{1}{n}\left(\tilde{x}_i \tilde{x}_i^\top - x_i x_i^\top\right)}\right\|_2 \\
&\cdot \left(\left\|e^{-s\frac{1}{n}X^\top X}u\right\|_2^{\alpha-1} + \left\|e^{-s\frac{1}{n}\hat{X}^\top \hat{X}}u\right\|_2^{\alpha-1}\right) \mathrm{d}s
\end{aligned} \tag{62}$$

$$= \int_0^\infty \underbrace{\left\| I - e^{s\frac{1}{n}\left(x_i x_i^\top - \tilde{x}_i \tilde{x}_i^\top\right)} \right\|_2}_{:=D} \underbrace{\left(\left\| e^{-s\frac{1}{n}X^\top X} u \right\|_2^\alpha + \left\| e^{-s\frac{1}{n}X^\top X} u \right\|_2 \left\| e^{-s\frac{1}{n}\hat{X}^\top \hat{X}} u \right\|_2^{\alpha-1} \right)}_{:=E} \mathrm{d}s. \tag{63}$$

Let us consider the term D in (63) first. $\tilde{x}_i \tilde{a}_i^\top - x_i x_i^\top$ is a rank 2 matrix. Consider the two non-zero eigenvalues of this matrix are σ_1 and σ_2. Hence, $\tilde{x}_i \tilde{x}_i^\top - x_i x_i^\top = \sigma_1 v_1 v_1^\top + \sigma_2 v_2 v_2^\top$ where v_1 and v_2 are the eigenvectors. Then,

$$I - e^{-s\frac{1}{n}\left(\tilde{x}_i \tilde{a}_i^\top - x_i x_i^\top\right)} = \left(1 - e^{\frac{-s\sigma_1}{n}}\right) v_1 v_1^\top + \left(1 - e^{\frac{-s\sigma_2}{n}}\right) v_2 v_2^\top.$$

Hence,

$$\begin{aligned} D = \left\| I - e^{-s\frac{1}{n}\left(\tilde{x}_i \tilde{a}_i^\top - x_i x_i^\top\right)} \right\|_2 &\leq \left\| \left(1 - e^{\frac{-s\sigma_1}{n}}\right) v_1 v_1^\top \right\|_2 + \left\| \left(1 - e^{\frac{-s\sigma_2}{n}}\right) v_2 v_2^\top \right\|_2 \\ &\leq \frac{s\sigma_1}{n} + \frac{s\sigma_2}{n}, \end{aligned}$$

where v_1 and v_2 are orthogonal vectors with $\|v_1\|_2 = \|v_2\|_2 = 1$ and $v_1^\top v_2 = 0$. By definition, we have

$$\frac{1}{n}\left\| X^\top X u \right\|_2 \geq \sigma_{\min} \|u\|_2, \quad \text{and} \quad \frac{1}{n}\left\| \hat{X}^\top \hat{X} u \right\|_2 \geq \sigma_{\min} \|u\|_2.$$

This gives,

$$\left\| e^{-s\frac{1}{n}X^\top X} u \right\|_2 \leq e^{-s\sigma_{\min}}, \quad \text{and} \quad \left\| e^{-s\frac{1}{n}\hat{X}^\top \hat{X}} u \right\|_2 \leq e^{-s\sigma_{\min}},$$

which implies that the E term in (63) can be bounded as:

$$E \leq 2\|u\|_2^\alpha e^{-s\alpha\sigma_{\min}}. \tag{64}$$

Therefore,

$$C \leq \frac{2(\sigma_1 + \sigma_2)\|u\|_2^\alpha}{n} \int_0^\infty s e^{-s\alpha\sigma_{\min}} \mathrm{d}s = \frac{2(\sigma_1 + \sigma_2)\|u\|_2^\alpha}{n\alpha^2\sigma_{\min}^2}. \tag{65}$$

Hence, we have

$$\begin{aligned} |\psi_1(u) - \psi_2(u)| &\leq \frac{2(\sigma_1 + \sigma_2)\|u\|_2^\alpha}{n\alpha^2\sigma_{\min}^2} \exp\left(- \int_0^\infty \left\| e^{-s\frac{1}{n}X^\top X} u \right\|_2^\alpha \mathrm{d}s \right) \\ &\leq \frac{2(\sigma_1 + \sigma_2)\|u\|_2^\alpha}{n\alpha^2\sigma_{\min}^2} \exp\left(-\|u\|_2^\alpha \int_0^\infty e^{-s\alpha\sigma_{\min}} \mathrm{d}s \right), \end{aligned} \tag{66}$$

where the last inequality is due to the definition of $\sigma_{\min}$. Hence we conclude that

$$|\psi_1(u) - \psi_2(u)| \leq \frac{2(\sigma_1 + \sigma_2)\|u\|_2^\alpha}{n\alpha^2\sigma_{\min}^2} \exp\left(-\frac{\|u\|_2^\alpha}{\alpha\sigma_{\min}} \right), \tag{67}$$

which completes the proof. ■

Theorem *[Restatement of Theorem 5] Consider the d-dimensional loss function $f(x) = |\theta^\top x|^p$ such that $\theta, x \in \mathbb{R}^d$. For any $x \sim P_X$ if $\|x\|_2 \leq R$, for any X sampled uniformly at random from the set $\mathcal{X}_n$, if $\frac{1}{n}\|X^\top X u\|_2 \geq \sigma_{\min}\|u\|_2$ for $u \in \mathbb{R}^d$ and for any two $X \cong \hat{X}$ sampled from $\mathcal{X}_n$ generating two stochastic process given by SDEs in equations (10) and (11), $\|x_i x_i^\top - \tilde{x}_i \tilde{x}_i^\top\|_2 \leq 2\sigma$ holds with high probability. Then,*

(i) For $\alpha \in (1, 2)$, the algorithm is not stable when $p \in [\alpha, 2]$ i.e. $\varepsilon_{stab(\mathcal{A}_{cont})}$ diverges. When $\alpha = p = 2$ then with high probability $\varepsilon_{stab}(\mathcal{A}_{cont}) \leq \frac{2R^2}{\pi}\frac{\sigma}{n\sigma_{\min}^2}$.

(ii) For $p \in [1, \alpha)$, we have the following upper bound for the algorithmic stability,

$$\varepsilon_{stab}(\mathcal{A}_{cont}) = \frac{8R^p}{\pi}\frac{\sigma}{n\alpha^2\sigma_{\min}}\Gamma(p+1)\cos\left(\frac{(p-1)\pi}{2}\right)\left(\frac{1}{\alpha\sigma_{\min}}\right)^{\frac{p}{\alpha}}\Gamma\left(1-\frac{p}{\alpha}\right) = c(\alpha),$$

which holds with high probability. Furthermore, for some $\alpha_0 > 1$, if we have

$$\sigma_{\min} \geq \exp\left(1 + \frac{4}{p} - \log\alpha_0 - \phi\left(1 - \frac{p}{\alpha_0}\right)\right),$$

where ϕ is the digamma function, then the map $\alpha \to c(\alpha)$ is increasing for $\alpha \in [\alpha_0, 2)$.

iii The stability bound is tight in α.

Proof We have d-dimensional loss function for an $x \in \mathbb{R}^d$ sampled uniformly at random from P_X, $f(\theta) = |\theta^\top x|^p$ Let us denote the Fourier transform of f, $\mathcal{F}f(u)$ as $h(u)$. For an orthogonal matrix A such that $Ae_1 = \frac{x}{\|x\|_2}$, we have from the results in Lemma 21,

$$h(Au) = 2\|x\|_2^p(2\pi)^{d-1}\delta(u_2, \cdots, u_d)\Gamma(p+1)\cos\left(\frac{(p+1)\pi}{2}\right)\frac{1}{|u_1|^{p+1}} \quad \text{for } p \in [1, 2), \tag{68}$$

and

$$h(Au) = 2\|x\|_2^p(2\pi)^{d-1}\delta(u_1, u_2, \cdots, u_d)\frac{2}{u_1^2} \quad \text{for } p = 2, \tag{69}$$

where δ is the Dirac-delta function. Let us first consider the case when $p \in [1, 2)$. From equation (18),

$$\begin{aligned}
\varepsilon_{\text{stab}}(\mathcal{A}_{\text{cont}}) &= \sup_{X\cong\hat{X}}\sup_{x\in\mathcal{X}}\frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}\left|\psi_\theta(u) - \psi_{\hat{\theta}}(u)\right||h(u)|\,\mathrm{d}u \\
&= \sup_{X\cong\hat{X}}\sup_{x\in\mathcal{X}}\frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}\left|\psi_\theta(u) - \psi_{\hat{\theta}}(u)\right|\left|\int_{\mathbb{R}^d}|\theta^\top x|^p e^{\mathrm{i}u^\top\theta}\,\mathrm{d}\theta\right|\,\mathrm{d}u \\
&= \sup_{X\cong\hat{X}}\sup_{x\in\mathcal{X}}\frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}\frac{2(\sigma_1+\sigma_2)\|u\|_2^\alpha}{n\alpha^2\sigma_{\min}^2}\exp\left(-\frac{\|u\|_2^\alpha}{\alpha\sigma_{\min}}\right)\left|\int_{\mathbb{R}^d}|\theta^\top x|^p e^{\mathrm{i}u^\top\theta}\,\mathrm{d}\theta\right|\,\mathrm{d}u.
\end{aligned}$$

In the above equation, let us apply the change of variable $u = Av$ and use result from Lemma 21 (equations (68)) and we get the following,

$$\varepsilon_{\text{stab}}(\mathcal{A}_{\text{cont}}) = \sup_{X\cong\hat{X}}\sup_{x\in\mathcal{X}}\frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}\frac{2(\sigma_1+\sigma_2)\|Av\|_2^\alpha}{n\alpha^2\sigma_{\min}^2}\exp\left(-\frac{\|Av\|_2^\alpha}{\alpha\sigma_{\min}}\right)\left|\int_{\mathbb{R}^d}|\theta^\top x|^p e^{\mathrm{i}(Av)^\top\theta}\,\mathrm{d}\theta\right|\,\mathrm{d}v$$

$$
\begin{aligned}
&= \sup_{X\cong\hat{X}}\sup_{x\in\mathcal{X}} \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} \frac{2(\sigma_1+\sigma_2)\|v\|_2^\alpha}{n\alpha^2\sigma_{\min}^2}\exp\left(-\frac{\|v\|_2^\alpha}{\alpha\sigma_{\min}}\right)|h(Av)|\ \mathrm{d}v \\
&= \sup_{X\cong\hat{X}}\sup_{x\in\mathcal{X}} \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} \left[\left(\frac{2(\sigma_1+\sigma_2)\|v\|_2^\alpha}{n\alpha^2\sigma_{\min}^2}\exp\left(-\frac{\|v\|_2^\alpha}{\alpha\sigma_{\min}}\right)\right)\right. \\
&\qquad \left.\cdot\left(\left|2\|x\|_2^p(2\pi)^{d-1}\delta(v_2,\cdots,v_d)\Gamma(p+1)\cos\left(\frac{(p+1)\pi}{2}\right)\frac{1}{|v_1|^{p+1}}\right|\right)\right]\ \mathrm{d}v \\
&= \sup_{X\cong\hat{X}}\sup_{x\in\mathcal{X}} \frac{2\|x\|_2^p}{\pi}\Gamma(p+1)\cos\left(\frac{(p-1)\pi}{2}\right)\frac{(\sigma_1+\sigma_2)}{n\alpha^2\sigma_{\min}^2} \\
&\qquad\qquad \cdot\int_{-\infty}^{\infty}|v_1|^\alpha\exp\left(\frac{-|v_1|^\alpha}{\alpha\sigma_{\min}}\right)\frac{1}{|v_1|^{p+1}}\,\mathrm{d}v_1 \\
&= \sup_{X\cong\hat{X}}\sup_{x\in\mathcal{X}} \frac{4\|x\|_2^p}{\pi}\frac{(\sigma_1+\sigma_2)}{n\alpha^2\sigma_{\min}^2}\Gamma(p+1)\cos\left(\frac{(p-1)\pi}{2}\right) \\
&\qquad\qquad \cdot\int_0^{\infty}v_1^{\alpha-p-1}\exp\left(\frac{-v_1^\alpha}{\alpha\sigma_{\min}}\right)\,\mathrm{d}v_1.
\end{aligned}
$$

In the above integral, by substituting $\frac{v_1^\alpha}{\alpha\sigma_{\min}}$ with t so that

$$
\mathrm{d}t = v_1^{\alpha-1}\frac{1}{\sigma_{\min}}\mathrm{d}v_1,\ \text{and}\ \frac{1}{v_1^p} = \left(\frac{1}{\alpha\sigma_{\min}}\right)^{\frac{p}{\alpha}}t^{-p/\alpha}, \tag{70}
$$

we have,

$$
\begin{aligned}
\varepsilon_{\text{stab}}(\mathcal{A}_{\text{cont}}) = \sup_{X\cong\hat{X}}\sup_{x\in\mathcal{X}} &\frac{4\|x\|_2^p}{\pi}\frac{(\sigma_1+\sigma_2)}{n\alpha^2\sigma_{\min}} \\
&\cdot\Gamma(p+1)\cos\left(\frac{(p-1)\pi}{2}\right)\left(\frac{1}{\alpha\sigma_{\min}}\right)^{\frac{p}{\alpha}}\int_0^\infty t^{-p/\alpha}e^{-t}\,\mathrm{d}t.
\end{aligned} \tag{71}
$$

It is clear that, the above integral diverge for $p \geq \alpha$, hence the algorithm is not stable for $p \in [1, 2)$. Now, we check the case for $p = 2$. For $p = 2$, we have,

$$
\varepsilon_{\text{stab}}(\mathcal{A}_{\text{cont}}) = \sup_{X\cong\hat{X}}\sup_{x\in\mathcal{X}} \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}\frac{2(\sigma_1+\sigma_2)\|u\|_2^\alpha}{n\alpha^2\sigma_{\min}^2}\exp\left(-\frac{\|u\|_2^\alpha}{\alpha\sigma_{\min}}\right)\left|\int_{\mathbb{R}^d}|\theta^\top x|^2e^{\mathrm{i}u^\top\theta}\,\mathrm{d}\theta\right|\,\mathrm{d}u.
$$

In the above equation, we make change of variable $u = Av$ and use result from Lemma 21 (equations (69)), we get the following,

$$
\begin{aligned}
\varepsilon_{\text{stab}}(\mathcal{A}_{\text{cont}}) &= \sup_{X\cong\hat{X}}\sup_{x\in\mathcal{X}} \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}\frac{2(\sigma_1+\sigma_2)\|Av\|_2^\alpha}{n\alpha^2\sigma_{\min}^2} \\
&\qquad\qquad \cdot\exp\left(-\frac{\|Av\|_2^\alpha}{\alpha\sigma_{\min}}\right)\left|\int_{\mathbb{R}^d}|\theta^\top x|^2e^{\mathrm{i}(Av)^\top\theta}\,\mathrm{d}\theta\right|\,\mathrm{d}v \\
&= \sup_{X\cong\hat{X}}\sup_{x\in\mathcal{X}} \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}\frac{2(\sigma_1+\sigma_2)\|v\|_2^\alpha}{n\alpha^2\sigma_{\min}^2}\exp\left(-\frac{\|v\|_2^\alpha}{\alpha\sigma_{\min}}\right)|h(Av)|\ \mathrm{d}v \\
&= \sup_{X\cong\hat{X}}\sup_{x\in\mathcal{X}} \frac{2}{\pi}\int_{\mathbb{R}^d}\frac{(\sigma_1+\sigma_2)\|v\|_2^\alpha}{n\alpha^2\sigma_{\min}^2}\exp\left(-\frac{\|v\|_2^\alpha}{\alpha\sigma_{\min}}\right)\|x\|_2^2\delta(v_1,v_2,\cdots,v_d)\frac{2}{v_1^2}\,\mathrm{d}v.
\end{aligned}
$$

The above integral clearly diverges for $\alpha < 2$. However, when $\alpha = 2$, then

$$\varepsilon_{\text{stab}}(\mathcal{A}_{\text{cont}}) \leq \frac{\|x\|_2^2}{\pi} \frac{(\sigma_1 + \sigma_2)}{n\sigma_{\min}^2}.$$

Now, if σ is the upper bound on σ_1 and σ_2 for all $X \cong \hat{X} \in \mathcal{X}_n$ and $\|x\|_2 \leq R$ for $x \sim P_X$ with high probability, then,

$$\varepsilon_{\text{stab}}(\mathcal{A}_{\text{cont}}) \leq \frac{2R^2}{\pi} \frac{\sigma}{n\sigma_{\min}^2},$$

holds with high probability. This proves the part (i) of our claim.

Next, we will prove part (ii) when $p < \alpha$. We have from equation (71),

$$\begin{aligned}
\varepsilon_{\text{stab}}(\mathcal{A}_{\text{cont}}) &= \sup_{X \cong \hat{X}} \sup_{x \in \mathcal{X}} \frac{4\|x\|_2^p}{\pi} \frac{(\sigma_1 + \sigma_2)}{n\alpha^2\sigma_{\min}} \Gamma(p+1) \cos\left(\frac{(p-1)\pi}{2}\right) \left(\frac{1}{\alpha\sigma_{\min}}\right)^{\frac{p}{\alpha}} \int_0^\infty t^{-p/\alpha} e^{-t} \, \mathrm{d}t \\
&= \sup_{X \cong \hat{X}} \sup_{x \in \mathcal{X}} \frac{4\|x\|_2^p}{\pi} \frac{(\sigma_1 + \sigma_2)}{n\alpha^2\sigma_{\min}} \Gamma(p+1) \cos\left(\frac{(p-1)\pi}{2}\right) \left(\frac{1}{\alpha\sigma_{\min}}\right)^{\frac{p}{\alpha}} \Gamma\left(1 - \frac{p}{\alpha}\right).
\end{aligned}$$

Now, if σ is the upper bound on σ_1 and σ_2 for all $X \cong \hat{X} \in \mathcal{X}_n$ and $\|x\|_2 \leq R$ for $x \sim P_X$ with high probability then,

$$\varepsilon_{\text{stab}}(\mathcal{A}_{\text{cont}}) = \frac{8R^p}{\pi} \frac{\sigma}{n\alpha^2\sigma_{\min}} \Gamma(p+1) \cos\left(\frac{(p-1)\pi}{2}\right) \left(\frac{1}{\alpha\sigma_{\min}}\right)^{\frac{p}{\alpha}} \Gamma\left(1 - \frac{p}{\alpha}\right)$$

holds with high probability. Now, consider the function,

$$\Lambda(\alpha) = \frac{1}{\alpha^2} \left(\frac{1}{\alpha\sigma_{\min}}\right)^{\frac{p}{\alpha}} \Gamma\left(1 - \frac{p}{\alpha}\right).$$

We can compute that

$$\partial_\alpha \log \Lambda(\alpha) = \frac{p}{\alpha^2} \left[\log \alpha + \log \sigma_{\min} - 1 - \frac{2\alpha}{p} + \phi\left(1 - \frac{p}{\alpha}\right)\right],$$

where ϕ is the digamma function. For any arbitrary α_0, if we choose

$$\sigma_{\min} \geq \exp\left(1 + \frac{4}{p} - \log \alpha_0 - \phi\left(1 - \frac{p}{\alpha_0}\right)\right),$$

then $\partial_\alpha \log \Lambda(\alpha) > 0$ for $\alpha \in [\alpha_0, 2)$. Hence, for all $\alpha_1, \alpha_2 \in [\alpha_0, 2)$, $\alpha_1 < \alpha_2 \Rightarrow \Lambda(\alpha_1) \leq \Lambda(\alpha_2)$. This proves that $c(\alpha)$ is an increasing map in α.

This completes the proof till part (ii). Now, we will prove tightness result in α. Let us have the following construction. Consider a one-dimensional distribution P_X supported in a ring such that the density function $\int_A p(x)\mathrm{d}x \leq \eta$ such that $A = \{x : |x| \geq \sqrt{\sigma_{\min} d \log d} \text{ or } |x| \leq R\}$. The empirical covariance matrix $X^\top X$ is a diagonal matrix. Hence, from the results in Flatto (2019). with high probability $1 - \delta$, we have

$$\frac{1}{n} \left\| X^\top X u \right\|_2 \geq \sigma_{\min} \|u\|_2.$$

Exact expression for δ is given in Flatto (2019). Similarly, for the dataset $\hat{X}$, the similar condition holds,

$$\frac{1}{n}\left\|\hat{X}^\top \hat{X} u\right\|_2 \geq \sigma_{\min}\|u\|_2^2$$

with high probability $1-\delta$. We have,

$$\begin{aligned}
&\left|\psi_\theta(u) - \psi_{\hat{\theta}}(u)\right| \\
&= \left|\exp\left(-\int_0^\infty \left\|e^{-s\frac{1}{n}X^\top X}u\right\|_2^\alpha \mathrm{d}s\right) - \exp\left(-\int_0^\infty \left\|e^{-s\frac{1}{n}\hat{X}^\top \hat{X}}u\right\|_2^\alpha \mathrm{d}s\right)\right| \\
&= \exp\left(-\int_0^\infty \left\|e^{-s\frac{1}{n}X^\top X}u\right\|_2^\alpha \mathrm{d}s\right)\left|1 - \exp\left(-\int_0^\infty \left|\left\|e^{-s\frac{1}{n}\hat{X}^\top \hat{X}}u\right\|_2^\alpha - \left\|e^{-s\frac{1}{n}X^\top X}u\right\|_2^\alpha\right| \mathrm{d}s\right)\right|.
\end{aligned} \tag{72}$$

From equation (18),

$$\begin{aligned}
\varepsilon_{\text{stab}}(\mathcal{A}_{\text{cont}}) &= \sup_{X\cong\hat{X}}\sup_{x\in\mathcal{X}} \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} \left|\psi_\theta(u) - \psi_{\hat{\theta}}(u)\right| |h(u)| \,\mathrm{d}u \\
&= \sup_{X\cong\hat{X}}\sup_{x\in\mathcal{X}} \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} \left|\psi_\theta(u) - \psi_{\hat{\theta}}(u)\right| \left|\int_{\mathbb{R}^d} |\theta^\top x|^p e^{\mathrm{i}u^\top\theta}\,\mathrm{d}\theta\right| \,\mathrm{d}u.
\end{aligned}$$

In the above equation, let us apply the change of variable $u = Av$ where A is the orthogonal matrix defined earlier and we get the following,

$$\begin{aligned}
\varepsilon_{\text{stab}}(\mathcal{A}_{\text{cont}}) &= \sup_{X\cong\hat{X}}\sup_{x\in\mathcal{X}} \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} \left|\psi_\theta(Av) - \psi_{\hat{\theta}}(Av)\right| \left|\int_{\mathbb{R}^d} |\theta^\top x|^p e^{\mathrm{i}(Av)^\top\theta}\,\mathrm{d}\theta\right| \,\mathrm{d}v \\
&= \sup_{X\cong\hat{X}}\sup_{x\in\mathcal{X}} \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} \left|\psi_\theta(Av) - \psi_{\hat{\theta}}(Av)\right| |h(Av)| \,\mathrm{d}v
\end{aligned}$$

Since, A is orthogonal matrix, we can see that

$$\left|\psi_\theta(Av) - \psi_{\hat{\theta}}(Av)\right| = \left|\psi_\theta(v) - \psi_{\hat{\theta}}(v)\right|.$$

Hence,

$$\begin{aligned}
\varepsilon_{\text{stab}}(\mathcal{A}_{\text{cont}}) &= \sup_{X\cong\hat{X}}\sup_{x\in\mathcal{X}} \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} \left|\psi_\theta(v) - \psi_{\hat{\theta}}(v)\right| |h(Av)| \,\mathrm{d}v \\
&= \sup_{X\cong\hat{X}}\sup_{x\in\mathcal{X}} \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} \left|\psi_\theta(v) - \psi_{\hat{\theta}}(v)\right| \left(\left|2\|x\|_2^p(2\pi)^{d-1}\delta(v_2,\cdots,v_d)\right.\right. \\
&\qquad\qquad \left.\left.\cdot\,\Gamma(p+1)\cos\left(\frac{(p+1)\pi}{2}\right)\frac{1}{|v_1|^{p+1}}\right|\right)\mathrm{d}v.
\end{aligned}$$

Let us denote

$$\Phi_{x,X,\hat{X}}(\mathcal{A}_{\text{cont}}) := \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} \left|\psi_\theta(v) - \psi_{\hat{\theta}}(v)\right| \left(\left|2\|x\|_2^p(2\pi)^{d-1}\delta(v_2,\cdots,v_d)\right.\right.$$

$$\cdot \Gamma(p+1)\cos\left(\frac{(p+1)\pi}{2}\right)\frac{1}{|v_1|^{p+1}}\bigg|\bigg)\mathrm{d}v.$$

Now, we use the property of Dirac-delta function. From our construction, $X^\top X$ and $\hat{X}^\top\hat{X}$ are diagonal matrices. Let us denote $X^\top X = \mathrm{diag}(a_1, a_2, \cdots, a_d)$. Similarly, we denote $\hat{X}^\top\hat{X} = \mathrm{diag}(\hat{a}_1, \hat{a}_2, \cdots, \hat{a}_d)$. Hence, we have

$$\exp\left(-\int_0^\infty \left\|e^{-s\frac{1}{n}X^\top X}v\right\|_2^\alpha \mathrm{d}s\right) = \exp\left(-\int_0^\infty \left(\sum_{i=1}^d e^{-2(s/n)a_i}v_i^2\right)^{\frac{\alpha}{2}}\mathrm{d}s\right).$$

From the construction, the matrix $X^\top X$ and $\hat{X}^\top X$ are both diagonal and differ at two diagonal elements with probability $(1-1/d)$. They differ at one diagonal element with probability $1/d$. Let's assume that x_i has non-zero element at dimension 1 and $\tilde{x}_i$ has non-zero element either at dimension 1 or at 2 (without loss of generality). Hence, with high probability,

$$\begin{aligned}
&\left|1-\exp\left(-\int_0^\infty \left|\left\|e^{-s\frac{1}{n}\hat{X}^\top\hat{X}}u\right\|_2^\alpha - \left\|e^{-s\frac{1}{n}X^\top X}u\right\|_2^\alpha\right|\mathrm{d}s\right)\right| \\
&= 1-\exp\left(-\int_0^\infty \left|\left(\sum_{i=1}^d e^{-2(s/n)a_i}v_i^2\right)^{\alpha/2} - \left(\sum_{i=1}^d e^{-2(s/n)\hat{a}_i}v_i^2\right)^{\alpha/2}\right|\mathrm{d}s\right)
\end{aligned}$$

Combining everything together and using the property of Dirac-delta function we get,

$$\begin{aligned}
\Phi_{x,X,\hat{X}}(\mathcal{A}_{\text{cont}}) &= \frac{1}{2\pi}\int_{-\infty}^\infty \Bigg[\exp\left(-\int_0^\infty e^{-(s\alpha/n)a_1}|v_1|^\alpha \mathrm{d}s\right) \\
&\qquad \cdot \left[1-\exp\left(-\int_0^\infty \left(e^{-(s\alpha/n)a_1} - e^{-(s\alpha/n)\tilde{a}_1}\right)|v_1|^\alpha \mathrm{d}s\right)\right] \\
&\qquad \cdot \left(\left|2\|x\|_2^p\Gamma(p+1)\cos\left(\frac{(p+1)\pi}{2}\right)\frac{1}{|v_1|^{p+1}}\right|\right)\Bigg]\mathrm{d}v_1 \\
&= \frac{2\|x\|_2^p}{\pi}\Gamma(p+1)\cos\left(\frac{(p+1)\pi}{2}\right)\int_0^\infty \Bigg[\exp\left(-\int_0^\infty e^{-(s\alpha/n)a_1}v_1^\alpha \mathrm{d}s\right) \\
&\qquad \cdot \frac{1}{v_1^{p+1}}\left[1-\exp\left(-\int_0^\infty \left(e^{-(s\alpha/n)a_1} - e^{-(s\alpha/n)\tilde{a}_1}\right)v_1^\alpha \mathrm{d}s\right)\right]\Bigg]\mathrm{d}v_1 \\
&= \frac{2\|x\|_2^p}{\pi}\Gamma(p+1)\cos\left(\frac{(p+1)\pi}{2}\right)\int_0^\infty \Bigg[\exp\left(-v_1^\alpha\frac{n}{\alpha a_1}\right) \\
&\qquad \cdot \frac{1}{v_1^{p+1}}\left[1-\exp\left(-v_1^\alpha\Big[\frac{n}{\alpha a_1} - \frac{n}{\alpha\hat{a}_1}\Big]\right)\right]\Bigg]\mathrm{d}v_1 \\
&= \frac{2\|x\|_2^p}{\pi}\Gamma(p+1)\cos\left(\frac{(p+1)\pi}{2}\right)\int_0^\infty \Bigg[\frac{1}{v_1^{p+1}}\exp\left(-v_1^\alpha\frac{n}{\alpha a_1}\right) \\
&\qquad \cdot \left[1-\exp\left(\frac{n(\|x_i\|_2^2 - \|\tilde{x}_i\|_2^2)}{\alpha a_1\hat{a}_1}\right)\right]\Bigg]\mathrm{d}v_1.
\end{aligned}$$

Let us denote $\delta := \|x_i\|_2^2 - \|\tilde{x}_i\|_2^2$. Hence,

$$\begin{aligned}\Phi_{x,X,\hat{X}}(\mathcal{A}_{\text{cont}}) &= \frac{2\|x\|_2^p}{\pi}\Gamma(p+1)\cos\left(\frac{(p+1)\pi}{2}\right) \\ &\qquad \cdot \int_0^\infty \left[\frac{1}{v_1^{p+1}}\exp\left(-v_1^\alpha \frac{n}{\alpha a_1}\right)\left[1-\exp\left(\frac{n\delta}{\alpha a_1 \hat{a}_1}\right)\right]\right] \mathrm{d}v_1 \\ &= \frac{2\|x\|_2^p}{\pi\alpha}\Gamma(p+1)\cos\left(\frac{(p+1)\pi}{2}\right)\int_0^\infty \exp\left(-v_1^\alpha \frac{n}{\alpha a_1}\right) \\ &\qquad \cdot \left[\sum_{k=1}^\infty \frac{(-1)^{k+1}}{k!} v_1^{k\alpha-p-1}\left(\frac{n\delta}{\alpha a_1 \hat{a}_1}\right)^k\right] \mathrm{d}v_1.\end{aligned}$$

The above equation is just reduction to the computation of one-dimensional case which we did in equation (60). We apply similar argument that we did apply in computing the lower bound in equation (60). Hence, with high probability, we get (equation (61))

$$\Phi_{x,X,\hat{X}}(\mathcal{A}_{\text{cont}}) \geq \frac{2\|x\|_2^p}{\pi}\Gamma(p+1)\cos\left(\frac{(p+1)\pi}{2}\right)\frac{\delta}{\alpha^2 \hat{a}_1}\left(\frac{n}{\alpha a_1}\right)^{\frac{p}{\alpha}}\Gamma\left(1-\frac{p}{\alpha}\right).$$

Here, we also assume that P_X is a distribution with discrete support in range σ^2 to R with C number of support points equally spaced. Hence, with probability $(1-1/C)$, $\delta \geq c$ for some positive constant c.

By construction and the result from Flatto (2019), we know that $nCd\log d \geq a_1 \geq n\sigma_{\min}$ for some positive constant C with high probability. This also holds for $\hat{a}_1$. Hence, for some positive constant C_1 and C_2 (C_1 and C_2 has dependence on the dimension) , with high probability

$$\Phi_{x,X,\hat{X}}(\mathcal{A}_{\text{cont}}) \geq \frac{C_1}{n\alpha^2}\Gamma(p+1)\cos\left(\frac{(p+1)\pi}{2}\right)\left(\frac{1}{\alpha C_2}\right)^{\frac{p}{\alpha}}\Gamma\left(1-\frac{p}{\alpha}\right).$$

This completes the proof. ■

Remark 14 *As we have characterized the finite-time distribution of a Lévy-driven OU process in Appendix B, it is clear to see that for any finite time t, if $\psi_i^{(t)}(u)$ denotes the characteristic function at that time then following the same procedure as that in Lemma 13,*

$$\begin{aligned}&\left|\psi_1^{(t)}(u) - \psi_2^{(t)}(u)\right| \\ &\leq \frac{2(\sigma_1+\sigma_2)\|u\|_2^\alpha}{n\alpha^2\sigma_{\min}^2}\left(1-(\alpha\sigma_{\min}t+1)e^{-\alpha\sigma_{\min}t}\right)\exp\left(-\frac{\|u\|_2^\alpha}{\alpha\sigma_{\min}}\left(1-e^{-\alpha\sigma_{\min}t}\right)\right).\end{aligned}$$

And hence, the algorithmic stability can be calculated in the similar way as that given in Theorem 5 for any time instance t. From here, it is hard to analyze the monotonic behavior of algorithmic stability for all time instance t. Here, we consider two interesting cases to discuss the monotone behavior:

- *When* $t = O(\frac{1}{\alpha\sigma_{\min}})$ *or higher but finite. In this case,*

$$\left|\psi_1^{(t)}(u) - \psi_2^{(t)}(u)\right| \leq \frac{2(\sigma_1+\sigma_2)\|u\|_2^\alpha}{n\alpha^2\sigma_{\min}^2} \exp\left(-\frac{\|u\|_2^\alpha}{\alpha\sigma_{\min}}\left(1-\frac{1}{e}\right)\right).$$

The above expression differs from the result in Lemma 13 only by a constant factor in the exponential. Hence, the stability bound will have similar monotonic behavious as that for $t \to \infty$.

- *When* t *is very small i.e.* $t \ll \frac{1}{\alpha\sigma_{\min}}$ *such that* $1 - e^{-\alpha\sigma_{\min}t} \approx \alpha\sigma_{\min}t$. *Then,*

$$\left|\psi_1^{(t)}(u) - \psi_2^{(t)}(u)\right| \leq \frac{2(\sigma_1+\sigma_2)\|u\|_2^\alpha t}{n\alpha\sigma_{\min}} \exp\left(-\|u\|_2^\alpha t\right).$$

In that case, we can easily see that under similar conditions in Theorem 5,

$$\varepsilon_{stab}(\mathcal{A}_{cont}) \leq \frac{8R^p}{\pi}\frac{\sigma t^{\frac{p}{\alpha}}}{n\alpha^2\sigma_{\min}}\Gamma(p+1)\cos\left(\frac{(p-1)\pi}{2}\right)\Gamma\left(1-\frac{p}{\alpha}\right) = c(\alpha).$$

We can similarly show here that there exist some α_0 *corresponding to every* t *when* $c(\alpha)$ *is monotonic in* $[\alpha_0, 2)$.

Appendix F. Theory and Proofs for the Discretized SDE

In this section, we provide theoretical results and their proofs of the discretized SDE (40)-(41).

Lemma 15 *For two matrices* $X \in \mathbb{R}^{n\times d}$ *and* $\hat{X} \in \mathbb{R}^{n\times d}$ *as defined earlier, the absolute value of difference between the characteristic functions of the anytime distributions for* $\eta \leq \frac{1}{L}$ *where* L *is the maximum of largest eigenvalues of* $\frac{1}{n}X^\top X$ *and* $\frac{1}{n}\hat{X}^\top\hat{X}$, *at any* $u \in \mathbb{R}^d$ *corresponding to* d*-dimensional rotation invariant processes in equations (43) and (44) is bounded as*

$$\begin{aligned}&\left|\psi_\theta(k,u) - \psi_{\hat{\theta}}(k,u)\right| \\ &\leq \frac{\eta^2(\sigma_1+\sigma_2)}{n(1-\eta\sigma_{\min})}\frac{(k-1)(1-\eta\sigma_{\min})^{\alpha(k+1)} - k(1-\eta\sigma_{\min})^{\alpha k} + (1-\eta\sigma_{\min})^\alpha}{(1-(1-\eta\sigma_{\min})^\alpha)^2} \\ &\qquad\qquad \cdot \|u\|_2^\alpha \exp\left(-\frac{\eta(1-(1-\eta\sigma_{\min})^{k\alpha})}{1-(1-\eta\sigma_{\min})^\alpha}\|u\|_2^\alpha\right),\end{aligned}$$

for any $k > 0$, *where* $\sigma_{\min}$ *is the smaller of the smallest of singular values of the matrices* $\frac{1}{n}X^\top X$ *and* $\frac{1}{n}\hat{X}^\top\hat{X}$, *and* $x_i x_i^\top - \tilde{x}_i\tilde{x}_i^\top = \sigma_1 v_1 v_1^\top + \sigma_2 v_2 v_2^\top$ *where* v_1 *and* v_2 *are orthogonal vectors.*

Proof For simplicity, we consider $\Sigma = I$ here. For general PSD sigma, similar steps can be followed as in Appendix G. We can compute that

$$\begin{aligned}&\left|\psi_\theta(k,u) - \psi_{\hat{\theta}}(k,u)\right| \\ &= \left|\exp\left(-\eta\sum_{j=0}^{k-1}\left\|\left(I - \frac{\eta}{n}\left(X^\top X\right)\right)^j u\right\|_2^\alpha\right) - \exp\left(-\eta\sum_{j=0}^{k-1}\left\|\left(I - \frac{\eta}{n}\left(\hat{X}^\top\hat{X}\right)\right)^j u\right\|_2^\alpha\right)\right|\end{aligned}$$

$$\leq \underbrace{\exp\left(-\eta\sum_{j=0}^{k-1}\left\|\left(I-\frac{\eta}{n}\left(X^\top X\right)\right)^j u\right\|_2^\alpha\right)}_{:=B} \cdot \underbrace{\left|\eta\sum_{j=0}^{k-1}\left\|\left(I-\frac{\eta}{n}\left(X^\top X\right)\right)^j u\right\|_2^\alpha - \eta\sum_{j=0}^{k-1}\left\|\left(I-\frac{\eta}{n}\left(\hat{X}^\top \hat{X}\right)\right)^j u\right\|_2^\alpha\right|}_{:=C}. \quad (73)$$

We first consider bounding the term C in equation (73). From Lemma 19, we have for two positive numbers a and b, and for some $1 \leq \alpha \leq 2$, we have

$$|a^\alpha - b^\alpha| \leq |a-b|(a^{\alpha-1}+b^{\alpha-1}).$$

Utilizing the above result and triangle inequality, we have

$$\begin{aligned}
&\left|\eta\sum_{j=0}^{k-1}\left\|\left(I-\frac{\eta}{n}\left(X^\top X\right)\right)^j u\right\|_2^\alpha - \eta\sum_{j=0}^{k-1}\left\|\left(I-\frac{\eta}{n}\left(\hat{X}^\top \hat{X}\right)\right)^j u\right\|_2^\alpha\right| \\
=&\eta\left|\sum_{j=0}^{k-1}\left[\left(\left\|\left(I-\frac{\eta}{n}\left(X^\top X\right)\right)^j u - \left(I-\frac{\eta}{n}\left(\hat{X}^\top \hat{X}\right)\right)^j u\right\|_2\right)\right.\right. \\
&\left.\left. \cdot\left(\left\|\left(I-\frac{\eta}{n}\left(X^\top X\right)\right)^j u\right\|_2^{\alpha-1} + \left\|\left(I-\frac{\eta}{n}\left(\hat{X}^\top \hat{X}\right)\right)^j u\right\|_2^{\alpha-1}\right)\right]\right|.
\end{aligned}$$

By definition, we have

$$\frac{1}{n}\left\|X^\top X u\right\|_2 \geq \sigma_{\min}\|u\|_2, \quad \text{and} \quad \frac{1}{n}\left\|\hat{X}^\top \hat{X} u\right\|_2 \geq \sigma_{\min}\|u\|_2.$$

If $\eta \leq \frac{1}{L}$, where L is the maximum of largest eigenvalues of $\frac{1}{n}X^\top X$ and $\frac{1}{n}\hat{X}^\top \hat{X}$, then,

$$\left\|\left(I-\frac{\eta}{n}\left(X^\top X\right)\right)^j u\right\|_2 \leq (1-\eta\sigma_{\min})^j\|u\|_2.$$

Similarly,

$$\left\|\left(I-\frac{\eta}{n}\left(\hat{X}^\top \hat{X}\right)\right)^j u\right\|_2 \leq (1-\eta\sigma_{\min})^j\|u\|_2.$$

For any two symmetric matrices A and B with $AB = BA$, we have $A^j - B^j = (A-B)(A^{j-1} + AB^{j-2}\cdots + B^{j-1})$. It follows from the definitions that

$$X^\top X - \hat{X}^\top \hat{X} = x_i x_i^\top - \tilde{x}_i \tilde{x}_i^\top.$$

By using similar argument as before, we have

$$\left\|\left(I-\frac{\eta}{n}\left(X^\top X\right)\right)^j u - \left(I-\frac{\eta}{n}\left(\hat{X}^\top \hat{X}\right)\right)^j u\right\|_2 \leq \frac{\eta j}{n}\left\|x_i x_i^\top - \tilde{x}_i \tilde{x}_i^\top\right\|_2 (1-\eta\sigma_{\min})^{j-1}\|u\|_2.$$

Note that $\tilde{x}_i\tilde{a}_i^\top - x_i x_i^\top$ is a rank 2 matrix. Consider the two non-zero eigenvalues of this matrix are σ_1 and σ_2. Hence, $\tilde{x}_i\tilde{x}_i^\top - x_i x_i^\top = \sigma_1 v_1 v_1^\top + \sigma_2 v_2 v_2^\top$ where v_1 and v_2 are the eigenvectors. Hence, we have obtained a bound on the term C in equation (73) such that

$$C \leq \frac{\eta^2(\sigma_1+\sigma_2)}{n(1-\eta\sigma_{\min})}\|u\|_2^\alpha \sum_{j=0}^{k-1} j(1-\eta\sigma_{\min})^{j\alpha}.$$

Now, let us consider bounding the term B in equation (73). Using previous arguments,

$$\exp\left(-\eta\sum_{j=0}^{k-1}\left\|\left(I-\frac{\eta}{n}\left(X^\top X\right)\right)^j u\right\|_2^\alpha\right) \leq \exp\left(-\eta\sum_{j=0}^{k-1}(1-\eta\sigma_{\min})^{j\alpha}\|u\|_2^\alpha\right).$$

Hence, we get,

$$\begin{aligned}
&\left|\psi_\theta(k,u)-\psi_{\hat{\theta}}(k,u)\right| \\
&\leq \frac{\eta^2(\sigma_1+\sigma_2)}{n(1-\eta\sigma_{\min})}\sum_{j=0}^{k-1}[j(1-\eta\sigma_{\min})^{j\alpha}]\|u\|_2^\alpha \exp\left(-\eta\sum_{j=0}^{k-1}(1-\eta\sigma_{\min})^{j\alpha}\|u\|_2^\alpha\right) \\
&= \frac{\eta^2(\sigma_1+\sigma_2)}{n(1-\eta\sigma_{\min})}\frac{(k-1)(1-\eta\sigma_{\min})^{\alpha(k+1)} - k(1-\eta\sigma_{\min})^{\alpha k} + (1-\eta\sigma_{\min})^\alpha}{(1-(1-\eta\sigma_{\min})^\alpha)^2} \\
&\qquad\qquad \cdot \|u\|_2^\alpha \exp\left(-\frac{\eta(1-(1-\eta\sigma_{\min})^{k\alpha})}{1-(1-\eta\sigma_{\min})^\alpha}\|u\|_2^\alpha\right),
\end{aligned}$$

where we applied Lemma 20 and the proof is complete. ■

In particular, by letting $k \to \infty$ in Lemma 15, we obtain the following corollary that concerns the stability of the characteristic functions for the stationary distributions. By denoting $\psi_\theta(u) := \psi_\theta(\infty, u)$ and $\psi_{\hat{\theta}}(u) := \psi_{\hat{\theta}}(\infty, u)$, we have the following result.

Corollary 16 *Under the settings in Lemma 15, we have*

$$\left|\psi_\theta(u)-\psi_{\hat{\theta}}(u)\right| \leq \frac{\eta^2(\sigma_1+\sigma_2)(1-\eta\sigma_{\min})^{\alpha-1}}{n(1-(1-\eta\sigma_{\min})^\alpha)^2}\cdot\|u\|_2^\alpha \exp\left(-\frac{\eta}{1-(1-\eta\sigma_{\min})^\alpha}\|u\|_2^\alpha\right).$$

Proof The results directly follows from Lemma 15 by letting $k \to \infty$ and using the results for sum of geometric series. ■

Theorem *[Restatement of Theorem 6] Consider the d-dimensional loss function $f(x) = |\theta^\top x|^p$ such that $\theta, x \in \mathbb{R}^d$. For any $x \sim P_X$ if $\|x\|_2 \leq R$, for any X sampled uniformly at random from the set $\mathcal{X}_n$, if $\frac{1}{n}\|X^\top X u\|_2 \geq \sigma_{\min}\|u\|_2$ for $u \in \mathbb{R}^d$ and for any two $X \cong \hat{X}$ sampled from $\mathcal{X}_n$ generating two stochastic process given by SDEs in equations (20) and (21) for $\eta \leq \frac{1}{L}$ where L is the maximum of largest eigenvalues of $\frac{1}{n}X^\top X$ and $\frac{1}{n}\hat{X}^\top\hat{X}$, $\|x_i x_i^\top - \tilde{x}_i\tilde{x}_i^\top\|_2 \leq 2\sigma$ holds with high probability. Then, for $p \in [1, \alpha)$, we have*

$$\varepsilon_{stab} \leq \frac{2R^p}{\pi}\Gamma(p+1)\cos\left(\frac{(p-1)\pi}{2}\right)\frac{\sigma\eta^{1+\frac{p}{\alpha}}(1-\eta\sigma_{\min})^{\alpha-1}}{n\alpha(1-(1-\eta\sigma_{\min})^\alpha)^{1+\frac{p}{\alpha}}}\Gamma\left(1-\frac{p}{\alpha}\right)$$

with high probability.

Proof We have d-dimensional loss function for an $x \in \mathbb{R}^d$ sampled uniformly at random from P_X, $f(\theta) = |\theta^\top x|^p$ Let us denote the Fourier transform of f, $\mathcal{F}f(u)$ as $h(u)$. For an orthogonal matrix A such that $Ae_1 = \frac{x}{\|x\|_2}$, we have from the results in Lemma 21,

$$h(Au) = 2\|x\|_2^p(2\pi)^{d-1}\delta(u_2, \cdots, u_d)\Gamma(p+1)\cos\left(\frac{(p+1)\pi}{2}\right)\frac{1}{|u_1|^{p+1}} \quad \text{for } p \in [1,2), \tag{74}$$

and

$$h(Au) = 2\|x\|_2^p(2\pi)^{d-1}\delta(u_1, u_2, \cdots, u_d)\frac{2}{u_1^2} \quad \text{for } p = 2, \tag{75}$$

where δ is the Dirac-delta function. Let us first consider the case when $p \in [1,2)$. From equation (18),

$$\begin{aligned}
&\varepsilon_{\text{stab}}(\mathcal{A}_{\text{cont}}) \\
&= \sup_{X \cong \hat{X}} \sup_{x \in \mathcal{X}} \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \left|\psi_\theta(u) - \psi_{\hat{\theta}}(u)\right| |h(u)| \, \mathrm{d}u \\
&= \sup_{X \cong \hat{X}} \sup_{x \in \mathcal{X}} \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \left|\psi_\theta(u) - \psi_{\hat{\theta}}(u)\right| \left| \int_{\mathbb{R}^d} |\theta^\top x|^p e^{\mathrm{i}u^\top\theta} \, \mathrm{d}\theta \right| \, \mathrm{d}u \\
&= \sup_{X \cong \hat{X}} \sup_{x \in \mathcal{X}} \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \frac{\eta^2(\sigma_1+\sigma_2)(1-\eta\sigma_{\min})^{\alpha-1}}{n(1-(1-\eta\sigma_{\min})^\alpha)^2} \cdot \|u\|_2^\alpha \exp\left(-\frac{\eta}{1-(1-\eta\sigma_{\min})^\alpha}\|u\|_2^\alpha\right) \\
&\qquad \cdot \left| \int_{\mathbb{R}^d} |\theta^\top x|^p e^{\mathrm{i}u^\top\theta} \, \mathrm{d}\theta \right| \, \mathrm{d}u.
\end{aligned}$$

In the above equation, let us apply the change of variable $u = Av$ and use result from Lemma 21 (equations (74)) and we get the following,

$$\begin{aligned}
&\varepsilon_{\text{stab}}(\mathcal{A}_{\text{cont}}) \\
&= \sup_{X \cong \hat{X}} \sup_{x \in \mathcal{X}} \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \frac{\eta^2(\sigma_1+\sigma_2)(1-\eta\sigma_{\min})^{\alpha-1}}{n(1-(1-\eta\sigma_{\min})^\alpha)^2} \cdot \|Av\|_2^\alpha \exp\left(-\frac{\eta\|Av\|_2^\alpha}{1-(1-\eta\sigma_{\min})^\alpha}\right) \\
&\qquad \cdot \left| \int_{\mathbb{R}^d} |\theta^\top x|^p e^{\mathrm{i}(Av)^\top\theta} \, \mathrm{d}\theta \right| \, \mathrm{d}v \\
&= \sup_{X \cong \hat{X}} \sup_{x \in \mathcal{X}} \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \frac{\eta^2(\sigma_1+\sigma_2)(1-\eta\sigma_{\min})^{\alpha-1}}{n(1-(1-\eta\sigma_{\min})^\alpha)^2} \\
&\qquad \cdot \|v\|_2^\alpha \exp\left(-\frac{\eta\|v\|_2^\alpha}{1-(1-\eta\sigma_{\min})^\alpha}\right) |h(Av)| \, \mathrm{d}v \\
&= \sup_{X \cong \hat{X}} \sup_{x \in \mathcal{X}} \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \left[\frac{\eta^2(\sigma_1+\sigma_2)(1-\eta\sigma_{\min})^{\alpha-1}}{n(1-(1-\eta\sigma_{\min})^\alpha)^2} \cdot \|v\|_2^\alpha \exp\left(-\frac{\eta\|v\|_2^\alpha}{1-(1-\eta\sigma_{\min})^\alpha}\right)\right. \\
&\qquad \left. \cdot \left(\left|2\|x\|_2^p(2\pi)^{d-1}\delta(v_2, \cdots, v_d)\Gamma(p+1)\cos\left(\frac{(p+1)\pi}{2}\right)\frac{1}{|v_1|^{p+1}}\right|\right)\right] \, \mathrm{d}v \\
&= \sup_{X \cong \hat{X}} \sup_{x \in \mathcal{X}} \frac{\|x\|_2^p}{\pi}\Gamma(p+1)\cos\left(\frac{(p-1)\pi}{2}\right)\frac{\eta^2(\sigma_1+\sigma_2)(1-\eta\sigma_{\min})^{\alpha-1}}{n(1-(1-\eta\sigma_{\min})^\alpha)^2}
\end{aligned}$$

$$\cdot \int_{-\infty}^{\infty} |v_1|^{\alpha} \exp\left(\frac{-\eta|v_1|^{\alpha}}{1-(1-\eta\sigma_{\min})^{\alpha}}\right) \frac{1}{|v_1|^{p+1}}\, \mathrm{d}v_1$$

$$= \sup_{X\cong\hat{X}} \sup_{x\in\mathcal{X}} \frac{2\|x\|_2^p}{\pi}\Gamma(p+1)\cos\left(\frac{(p-1)\pi}{2}\right) \frac{\eta^2(\sigma_1+\sigma_2)(1-\eta\sigma_{\min})^{\alpha-1}}{n(1-(1-\eta\sigma_{\min})^{\alpha})^2}$$

$$\cdot \int_{0}^{\infty} |v_1|^{\alpha-p-1} \exp\left(\frac{-\eta|v_1|^{\alpha}}{1-(1-\eta\sigma_{\min})^{\alpha}}\right) \mathrm{d}v_1.$$

In the above integral, by substituting $\frac{\eta v_1^{\alpha}}{1-(1-\eta\sigma_{\min})^{\alpha}}$ with t so that

$$\mathrm{d}t = v^{\alpha-1}\frac{\eta\alpha}{1-(1-\eta\sigma_{\min})^{\alpha}}\mathrm{d}v_1, \text{ and } \frac{1}{v^p} = \left(\frac{\eta}{1-(1-\eta\sigma_{\min})^{\alpha}}\right)^{\frac{p}{\alpha}} t^{-p/\alpha}, \tag{76}$$

we have,

$$\varepsilon_{\text{stab}}(\mathcal{A}_{\text{cont}}) = \sup_{X\cong\hat{X}} \sup_{x\in\mathcal{X}} \frac{2\|x\|_2^p}{\pi}\Gamma(p+1)\cos\left(\frac{(p-1)\pi}{2}\right) \frac{\eta^2(\sigma_1+\sigma_2)(1-\eta\sigma_{\min})^{\alpha-1}}{n(1-(1-\eta\sigma_{\min})^{\alpha})^2}$$

$$\cdot \frac{1-(1-\eta\sigma_{\min})^{\alpha}}{\eta\alpha}\left(\frac{\eta}{1-(1-\eta\sigma_{\min})^{\alpha}}\right)^{\frac{p}{\alpha}} \int_0^{\infty} t^{-p/\alpha}e^{-t}\,\mathrm{d}t$$

$$= \sup_{X\cong\hat{X}} \sup_{x\in\mathcal{X}} \frac{2\|x\|_2^p}{\pi}\Gamma(p+1)\cos\left(\frac{(p-1)\pi}{2}\right) \frac{\eta^{1+\frac{p}{\alpha}}(\sigma_1+\sigma_2)(1-\eta\sigma_{\min})^{\alpha-1}}{n\alpha(1-(1-\eta\sigma_{\min})^{\alpha})^{1+\frac{p}{\alpha}}}\Gamma\left(1-\frac{p}{\alpha}\right). \tag{77}$$

Now, if σ is the upper bound on σ_1 and σ_2 for all $X \cong \hat{X} \in \mathcal{X}_n$ and $\|x\|_2 \le R$ for $x \sim P_X$ with high probability then,

$$\varepsilon_{\text{stab}} \le \frac{2R^p}{\pi}\Gamma(p+1)\cos\left(\frac{(p-1)\pi}{2}\right) \frac{\sigma\eta^{1+\frac{p}{\alpha}}(1-\eta\sigma_{\min})^{\alpha-1}}{n\alpha(1-(1-\eta\sigma_{\min})^{\alpha})^{1+\frac{p}{\alpha}}}\Gamma\left(1-\frac{p}{\alpha}\right). \tag{78}$$

This completes the proof. ■

Appendix G. Case for General P.S.D Σ (Preconditioning)

In this section, we would discuss the effect of general positive semidefinite matrix Σ. As in equations (79) and (80), we consider two SDEs corresponding to a rotationally symmetric α-stable Lévy process L_t^{α} in $\mathbb{R}^d$,

$$\mathrm{d}\theta_t = -\frac{1}{n}\left(X^{\top}X\right)\theta_t\mathrm{d}t + \Sigma\mathrm{d}\mathrm{L}_t^{\alpha}, \tag{79}$$

$$\mathrm{d}\hat{\theta}_t = -\frac{1}{n}\left(\hat{X}^{\top}\hat{X}\right)\hat{\theta}_t\mathrm{d}t + \Sigma\mathrm{d}\mathrm{L}_t^{\alpha}, \tag{80}$$

where $\Sigma \in \mathbb{R}^{d\times d}$ is a real valued P.S.D matrix. The corresponding characteristic functions are given by as in equations (81) and (82) (see Lemma 3),

$$\psi_{\theta}(u) = \exp\left(-\int_0^{\infty} \left\|\Sigma^{\top}e^{-s\frac{1}{n}(X^{\top}X)}u\right\|_2^{\alpha}\mathrm{d}s\right), \tag{81}$$

$$\psi_{\hat{\theta}}(u) = \exp\left(-\int_0^\infty \left\|\Sigma^\top e^{-s\frac{1}{n}(\hat{X}^\top\hat{X})}u\right\|_2^\alpha \mathrm{d}s\right). \tag{82}$$

We assume that the largest and smallest eigenvalues of the matrix Σ is $\lambda_{\max}$ and $\lambda_{\min}$.

Lemma 17 *For two matrices $X \in \mathbb{R}^{n\times d}$ and $\hat{X} \in \mathbb{R}^{n\times d}$ as defined earlier, the absolute value of difference between the characteristic functions of the stationary distributions at any $u \in \mathbb{R}^d$ corresponding to d-dimensional rotation invariant processes in equations (81) and (82) is bounded as*

$$|\psi_\theta(u) - \psi_{\hat{\theta}}(u)| \le \lambda_{\max}^\alpha \frac{2(\sigma_1+\sigma_2)\|u\|_2^\alpha}{n\alpha\sigma_{\min}} \exp\left(-\frac{\lambda_{\min}^\alpha\|u\|_2^\alpha}{\alpha^2\sigma_{\min}^2}\right),$$

where $\sigma_{\min}$ is the smaller of the smallest of singular values of the matrices $\frac{1}{n}X^\top X$ and $\frac{1}{n}\hat{X}^\top\hat{X}$, and $x_i x_i^\top - \tilde{x}_i\tilde{x}_i^\top = \sigma_1 v_1 v_1^\top + \sigma_2 v_2 v_2^\top$ where v_1 and v_2 are orthogonal vectors.

Proof We can compute that

$$\begin{aligned}
&|\psi_\theta(u) - \psi_{\hat{\theta}}(u)| \\
&= \left|\exp\left(-\int_0^\infty \left\|\Sigma^\top e^{-s\frac{1}{n}X^\top X}u\right\|_2^\alpha \mathrm{d}s\right) - \exp\left(-\int_0^\infty \left\|\Sigma^\top e^{-s\frac{1}{n}\hat{X}^\top\hat{X}}u\right\|_2^\alpha \mathrm{d}s\right)\right| \\
&\le \underbrace{\exp\left(-\int_0^\infty \left\|\Sigma^\top e^{-s\frac{1}{n}X^\top X}u\right\|_2^\alpha \mathrm{d}s\right)}_{:=B} \underbrace{\left|\int_0^\infty \left\|\Sigma^\top e^{-s\frac{1}{n}X^\top X}u\right\|_2^\alpha \mathrm{d}s - \int_0^\infty \left\|\Sigma^\top e^{-s\frac{1}{n}\hat{X}^\top\hat{X}}u\right\|_2^\alpha \mathrm{d}s\right|}_{:=C}.
\end{aligned}$$

We first consider the term C in the above equation. From Lemma 19, we have for two positive numbers a and b, and for some $1 \le \alpha \le 2$, we have

$$|a^\alpha - b^\alpha| \le |a-b|(a^{\alpha-1} + b^{\alpha-1}).$$

Now,

$$\begin{aligned}
C &= \left|\int_0^\infty \left\|\Sigma^\top e^{-s\frac{1}{n}X^\top X}u\right\|_2^\alpha \mathrm{d}s - \int_0^\infty \left\|\Sigma^\top e^{-s\frac{1}{n}\hat{X}^\top\hat{X}}u\right\|_2^\alpha \mathrm{d}s\right| \\
&= \left|\int_0^\infty \left(\left\|\Sigma^\top e^{-s\frac{1}{n}X^\top X}u\right\|_2^\alpha - \left\|\Sigma^\top e^{-s\frac{1}{n}\hat{X}^\top\hat{X}}u\right\|_2^\alpha\right)\mathrm{d}s\right| \\
&\le \int_0^\infty \left|\left\|\Sigma^\top e^{-s\frac{1}{n}X^\top X}u\right\|_2 - \left\|\Sigma^\top e^{-s\frac{1}{n}\hat{X}^\top\hat{X}}u\right\|_2\right| \\
&\qquad\qquad \cdot \left(\left\|\Sigma^\top e^{-s\frac{1}{n}X^\top X}u\right\|_2^{\alpha-1} + \left\|\Sigma^\top e^{-s\frac{1}{n}\hat{X}^\top\hat{X}}u\right\|_2^{\alpha-1}\right)\mathrm{d}s \\
&\le \int_0^\infty \left\|\Sigma^\top e^{-s\frac{1}{n}X^\top X}u - \Sigma^\top e^{-s\frac{1}{n}\hat{X}^\top\hat{X}}u\right\|_2 \left(\left\|\Sigma^\top e^{-s\frac{1}{n}X^\top X}u\right\|_2^{\alpha-1} + \left\|\Sigma^\top e^{-s\frac{1}{n}\hat{X}^\top\hat{X}}u\right\|_2^{\alpha-1}\right)\mathrm{d}s \\
&= \int_0^\infty \left\|\Sigma^\top e^{-s\frac{1}{n}X^\top X}\left(I - e^{s\frac{1}{n}X^\top X - s\frac{1}{n}\hat{X}^\top\hat{X}}\right)u\right\|_2 \\
&\qquad\qquad \cdot \left(\left\|\Sigma^\top e^{-s\frac{1}{n}X^\top X}u\right\|_2^{\alpha-1} + \left\|\Sigma^\top e^{-s\frac{1}{n}\hat{X}^\top\hat{X}}u\right\|_2^{\alpha-1}\right)\mathrm{d}s
\end{aligned}$$

$$\leq \lambda_{\max}^{\alpha} \underbrace{\int_0^{\infty} \left\| e^{-s\frac{1}{n}X^\top X} \left(I - e^{s\frac{1}{n}X^\top X - s\frac{1}{n}\hat{X}^\top \hat{X}} \right) u \right\|_2 \left(\left\| e^{-s\frac{1}{n}X^\top X} u \right\|_2^{\alpha-1} + \left\| e^{-s\frac{1}{n}\hat{X}^\top \hat{X}} u \right\|_2^{\alpha-1} \right) \mathrm{d}s}_{\text{This term has been analyzed as an upper bound on term C in Lemma 13 (Equation (63)).}}.$$

Using the result directly from equation (65), we have,

$$C \leq \lambda_{\max}^{\alpha} \frac{2(\sigma_1 + \sigma_2)\|u\|_2^{\alpha}}{n\alpha^2 \sigma_{\min}^2}. \tag{83}$$

Next, let us consider the term B. Using the similar arguments as in Lemma 13 (equation (66)), we have,

$$\begin{aligned} \exp\left(- \int_0^{\infty} \left\| \Sigma^\top e^{-s\frac{1}{n}X^\top X} u \right\|_2^{\alpha} \mathrm{d}s \right) &\leq \exp\left(-\lambda_{\min}^{\alpha} \|u\|_2^{\alpha} \int_0^{\infty} e^{-s\alpha\sigma_{\min}} \mathrm{d}s \right) \\ &= \exp\left(-\frac{\lambda_{\min}^{\alpha}\|u\|_2^{\alpha}}{\alpha\sigma_{\min}} \right). \end{aligned} \tag{84}$$

Hence, we have the final result,

$$|\psi_{\theta}(u) - \psi_{\hat{\theta}}(u)| \leq \lambda_{\max}^{\alpha} \frac{2(\sigma_1 + \sigma_2)\|u\|_2^{\alpha}}{n\alpha^2\sigma_{\min}^2} \exp\left(-\frac{\lambda_{\min}^{\alpha}\|u\|_2^{\alpha}}{\alpha\sigma_{\min}} \right), \tag{85}$$

which completes the proof. ■

Theorem 18 *Consider the d-dimensional loss function $f(x) = |\theta^\top x|^p$ such that $\theta, x \in \mathbb{R}^d$. For any $x \sim P_X$ if $\|x\|_2 \leq R$, for any X sampled uniformly at random from the set $\mathcal{X}_n$, if $\frac{1}{n}\|X^\top X u\|_2 \geq \sigma_{\min}\|u\|_2$ for $u \in \mathbb{R}^d$ and for any two $X \cong \hat{X}$ sampled from $\mathcal{X}_n$ generating two stochastic process given by SDEs in equations (79) and (80), $\|x_i x_i^\top - \tilde{x}_i \tilde{x}_i^\top\|_2 \leq 2\sigma$ holds with high probability. Then,*

(i) For $\alpha \in (1, 2)$, the algorithm is not stable when $p \in [\alpha, 2]$ i.e. $\varepsilon_{stab}(\mathcal{A}_{cont})$ diverges. When $\alpha = p = 2$ then with high probability $\varepsilon_{stab}(\mathcal{A}_{cont}) \leq \frac{2R^2}{\pi}\frac{\lambda_{\max}^2\sigma}{n\sigma_{\min}}$.

(ii) For $p \in [1, \alpha)$, we have the following upper bound for the algorithmic stability,

$$\begin{aligned} &\varepsilon_{stab}(\mathcal{A}_{cont}) \\ &\leq \frac{8R^p}{\pi}\lambda_{\min}^p \left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)^{\alpha} \frac{\sigma}{n\alpha^2\sigma_{\min}} \Gamma(p+1) \cos\left(\frac{(p-1)\pi}{2}\right) \left(\frac{1}{\alpha\sigma_{\min}}\right)^{\frac{p}{\alpha}} \Gamma\left(1 - \frac{p}{\alpha}\right) \\ &= c(\alpha), \end{aligned}$$

which holds with high probability. Furthermore, for some $\alpha_0 > 1$, if we have

$$\sigma_{\min} \geq \exp\left(1 + \frac{4}{p} - \log\alpha_0 - \phi\left(1 - \frac{p}{\alpha_0}\right) - \alpha_0^2 \log\left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)\right),$$

where ϕ is the digamma function, then the map $\alpha \to c(\alpha)$ is increasing for $\alpha \in [\alpha_0, 2)$.

Proof We have d-dimensional loss function for an $x \in \mathbb{R}^d$ sampled uniformly at random from P_X, $f(\theta) = |\theta^\top x|^p$ Let us denote the Fourier transform of f, $\mathcal{F}f(u)$ as $h(u)$. For an orthogonal matrix A such that $Ae_1 = \frac{x}{\|x\|_2}$, we have from the results in Lemma 21,

$$h(Au) = 2\|x\|_2^p (2\pi)^{d-1} \delta(u_2, \cdots, u_d) \Gamma(p+1) \cos\left(\frac{(p+1)\pi}{2}\right) \frac{1}{|u_1|^{p+1}} \quad \text{for } p \in [1,2), \tag{86}$$

and

$$h(Au) = 2\|x\|_2^p (2\pi)^{d-1} \delta(u_1, u_2, \cdots, u_d) \frac{2}{u_1^2} \quad \text{for } p = 2, \tag{87}$$

where δ is the Dirac-delta function. Let us first consider the case when $p \in [1,2)$. From equation (18) and Lemma 17,

$$\begin{aligned}
\varepsilon_{\text{stab}}(\mathcal{A}_{\text{cont}}) &= \sup_{X \cong \hat{X}} \sup_{x \in \mathcal{X}} \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} |\psi_\theta(u) - \psi_{\hat{\theta}}(u)||h(u)| \, \mathrm{d}u \\
&= \sup_{X \cong \hat{X}} \sup_{x \in \mathcal{X}} \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} |\psi_\theta(u) - \psi_{\hat{\theta}}(u)| \left| \int_{\mathbb{R}^d} |\theta^\top x|^p e^{\mathrm{i} u^\top \theta} \, \mathrm{d}\theta \right| \, \mathrm{d}u \\
&= \sup_{X \cong \hat{X}} \sup_{x \in \mathcal{X}} \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \lambda_{\max}^\alpha \frac{2(\sigma_1 + \sigma_2)\|u\|_2^\alpha}{n\alpha^2 \sigma_{\min}^2} \\
&\qquad \cdot \exp\left(-\frac{\lambda_{\min}^\alpha \|u\|_2^\alpha}{\alpha \sigma_{\min}}\right) \left| \int_{\mathbb{R}^d} |\theta^\top x|^p e^{\mathrm{i} u^\top \theta} \, \mathrm{d}\theta \right| \, \mathrm{d}u.
\end{aligned}$$

In the above equation, we make change of variable $u = Av$ and use the result from Lemma 21 (equation (86)) to get the following,

$$\begin{aligned}
\varepsilon_{\text{stab}}(\mathcal{A}_{\text{cont}}) &= \sup_{X \cong \hat{X}} \sup_{x \in \mathcal{X}} \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \lambda_{\max}^\alpha \frac{2(\sigma_1 + \sigma_2)\|Av\|_2^\alpha}{n\alpha^2 \sigma_{\min}^2} \\
&\qquad \cdot \exp\left(-\frac{\lambda_{\min}^\alpha \|Av\|_2^\alpha}{\alpha \sigma_{\min}}\right) \left| \int_{\mathbb{R}^d} |\theta^\top x|^p e^{\mathrm{i}(Av)^\top \theta} \, \mathrm{d}\theta \right| \, \mathrm{d}v \\
&= \sup_{X \cong \hat{X}} \sup_{x \in \mathcal{X}} \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \lambda_{\max}^\alpha \frac{2(\sigma_1 + \sigma_2)\|v\|_2^\alpha}{n\alpha^2 \sigma_{\min}^2} \exp\left(-\frac{\lambda_{\min}^\alpha \|v\|_2^\alpha}{\alpha \sigma_{\min}}\right) |h(Av)| \, \mathrm{d}v \\
&= \sup_{X \cong \hat{X}} \sup_{x \in \mathcal{X}} \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \left[\left(\lambda_{\max}^\alpha \frac{2(\sigma_1 + \sigma_2)\|v\|_2^\alpha}{n\alpha^2 \sigma_{\min}^2} \exp\left(-\frac{\lambda_{\min}^\alpha \|v\|_2^\alpha}{\alpha \sigma_{\min}}\right) \right) \right. \\
&\qquad \left. \cdot \left(\left| 2\|x\|_2^p (2\pi)^{d-1} \delta(v_2, \cdots, v_d) \Gamma(p+1) \cos\left(\frac{(p+1)\pi}{2}\right) \frac{1}{|v_1|^{p+1}} \right| \right) \right] \mathrm{d}v \\
&= \sup_{X \cong \hat{X}} \sup_{x \in \mathcal{X}} \frac{2\|x\|_2^p}{\pi} \Gamma(p+1) \cos\left(\frac{(p-1)\pi}{2}\right) \frac{\lambda_{\max}^\alpha (\sigma_1 + \sigma_2)}{n\alpha^2 \sigma_{\min}^2} \\
&\qquad \cdot \int_{-\infty}^{\infty} |v_1|^\alpha \exp\left(\frac{-\lambda_{\min}^\alpha |v_1|^\alpha}{\alpha \sigma_{\min}}\right) \frac{1}{|v_1|^{p+1}} \, \mathrm{d}v_1 \\
&= \sup_{X \cong \hat{X}} \sup_{x \in \mathcal{X}} \frac{4\|x\|_2^p}{\pi} \frac{\lambda_{\max}^\alpha (\sigma_1 + \sigma_2)}{n\alpha^2 \sigma_{\min}^2} \Gamma(p+1) \cos\left(\frac{(p-1)\pi}{2}\right)
\end{aligned}$$

$$\cdot \int_0^\infty v_1^{\alpha-p-1} \exp\left(\frac{-\lambda_{\min}^\alpha v_1^\alpha}{\alpha\sigma_{\min}}\right) \mathrm{d}v_1.$$

In the above integral, by substituting $\frac{\lambda_{\min}^\alpha v^\alpha}{\alpha\sigma_{\min}}$ with t so that

$$\mathrm{d}t = \lambda_{\min}^\alpha v^{\alpha-1}\frac{1}{\sigma_{\min}}\mathrm{d}v, \text{ and } \frac{1}{v^p} = \lambda_{\min}^p \left(\frac{1}{\alpha\sigma_{\min}}\right)^{\frac{p}{\alpha}} t^{-p/\alpha}, \tag{88}$$

we have,

$$\varepsilon_{\text{stab}}(\mathcal{A}_{\text{cont}}) = \sup_{X\cong\hat{X}}\sup_{x\in\mathcal{X}} \frac{4\|x\|_2^p}{\pi}\lambda_{\min}^p\left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)^\alpha \frac{(\sigma_1+\sigma_2)}{n\alpha^2\sigma_{\min}}$$
$$\cdot \Gamma(p+1)\cos\left(\frac{(p-1)\pi}{2}\right)\left(\frac{1}{\alpha\sigma_{\min}}\right)^{\frac{p}{\alpha}}\int_0^\infty t^{-p/\alpha}e^{-t}\,\mathrm{d}t. \tag{89}$$

It is clear that, the above integral diverge for $p \geq \alpha$, hence the algorithm is not stable for $p \in [1,2)$. Now, we check the case for $p = 2$. For $p = 2$, we have,

$$\varepsilon_{\text{stab}}(\mathcal{A}_{\text{cont}}) = \sup_{X\cong\hat{X}}\sup_{x\in\mathcal{X}} \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}\lambda_{\max}^\alpha \frac{2(\sigma_1+\sigma_2)\|u\|_2^\alpha}{n\alpha^2\sigma_{\min}^2}$$
$$\cdot \exp\left(-\frac{\lambda_{\min}^\alpha\|u\|_2^\alpha}{\alpha\sigma_{\min}}\right)\left|\int_{\mathbb{R}^d}|\theta^\top x|^2 e^{\mathrm{i}u^\top\theta}\,\mathrm{d}\theta\right|\,\mathrm{d}u.$$

In the above equation, we make change of variable $u = Av$ and use the result from Lemma 21 (equation (87)) to get the following,

$$\begin{aligned}
&\varepsilon_{\text{stab}}(\mathcal{A}_{\text{cont}}) \\
&= \sup_{X\cong\hat{X}}\sup_{x\in\mathcal{X}}\left\{\frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}\lambda_{\max}^\alpha\frac{2(\sigma_1+\sigma_2)\|Av\|_2^\alpha}{n\alpha^2\sigma_{\min}^2}\right. \\
&\qquad\left.\cdot\exp\left(-\frac{\lambda_{\min}^\alpha\|Av\|_2^\alpha}{\alpha\sigma_{\min}}\right)\left|\int_{\mathbb{R}^d}|\theta^\top x|^2 e^{\mathrm{i}(Av)^\top\theta}\,\mathrm{d}\theta\right|\,\mathrm{d}v\right\} \\
&= \sup_{X\cong\hat{X}}\sup_{x\in\mathcal{X}}\frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}\lambda_{\max}^\alpha\frac{2(\sigma_1+\sigma_2)\|v\|_2^\alpha}{n\alpha^2\sigma_{\min}^2}\exp\left(-\frac{\lambda_{\min}^\alpha\|v\|_2^\alpha}{\alpha\sigma_{\min}}\right)|h(Av)|\,\mathrm{d}v \\
&= \sup_{X\cong\hat{X}}\sup_{x\in\mathcal{X}}\frac{2}{\pi}\int_{\mathbb{R}^d}\lambda_{\max}^\alpha\frac{(\sigma_1+\sigma_2)\|v\|_2^\alpha}{n\alpha^2\sigma_{\min}^2}\exp\left(-\frac{\lambda_{\min}^\alpha\|v\|_2^\alpha}{\alpha\sigma_{\min}}\right)\|x\|_2^2\delta(v_1,v_2,\cdots,v_d)\frac{2}{v_1^2}\,\mathrm{d}v.
\end{aligned}$$

In the last equation, we used the result from Lemma 21. The above integral clearly diverges for $\alpha < 2$. However, when $\alpha = 2$, then

$$\varepsilon_{\text{stab}}(\mathcal{A}_{\text{cont}}) \leq \frac{\|x\|_2^2}{\pi}\frac{\lambda_{\max}^2(\sigma_1+\sigma_2)}{n\sigma_{\min}^2}.$$

Now, if σ is the upper bound on σ_1 and σ_2 for all $X \cong \hat{X} \in \mathcal{X}_n$ and $\|x\|_2 \leq R$ for $x \sim P_X$ with high probability then,

$$\varepsilon_{\text{stab}}(\mathcal{A}_{\text{cont}}) \leq \frac{2R^2}{\pi}\frac{\lambda_{\max}^2\sigma}{n\sigma_{\min}^2},$$

holds with high probability. This proves part (i) of our claim.

Next, we will prove part (ii) when $p < \alpha$. We have from equation (89),

$$\begin{aligned}
&\varepsilon_{\text{stab}}(\mathcal{A}_{\text{cont}}) \\
&= \sup_{X \cong \hat{X}} \sup_{x \in \mathcal{X}} \left\{ \frac{4\|x\|_2^p}{\pi} \lambda_{\min}^p \left(\frac{\lambda_{\max}}{\lambda_{\min}} \right)^{\alpha} \frac{(\sigma_1 + \sigma_2)}{n\alpha^2 \sigma_{\min}} \Gamma(p+1) \right. \\
&\qquad \left. \cdot \cos\left(\frac{(p-1)\pi}{2} \right) \left(\frac{1}{\alpha \sigma_{\min}} \right)^{\frac{p}{\alpha}} \int_0^{\infty} t^{-p/\alpha} e^{-t} \, \mathrm{d}t \right\} \\
&= \sup_{X \cong \hat{X}} \sup_{x \in \mathcal{X}} \frac{4\|x\|_2^p}{\pi} \lambda_{\min}^p \left(\frac{\lambda_{\max}}{\lambda_{\min}} \right)^{\alpha} \frac{(\sigma_1 + \sigma_2)}{n\alpha^2 \sigma_{\min}} \Gamma(p+1) \cos\left(\frac{(p-1)\pi}{2} \right) \left(\frac{1}{\alpha \sigma_{\min}} \right)^{\frac{p}{\alpha}} \Gamma\left(1 - \frac{p}{\alpha}\right).
\end{aligned}$$

Now, if σ is the upper bound on σ_1 and σ_2 for all $X \cong \hat{X} \in \mathcal{X}_n$ and $\|x\|_2 \leq R$ for $x \sim P_X$ with high probability then,

$$\varepsilon_{\text{stab}}(\mathcal{A}_{\text{cont}}) = \frac{8R^p}{\pi} \lambda_{\min}^p \left(\frac{\lambda_{\max}}{\lambda_{\min}} \right)^{\alpha} \frac{\sigma}{n\alpha^2 \sigma_{\min}} \Gamma(p+1) \cos\left(\frac{(p-1)\pi}{2} \right) \left(\frac{1}{\alpha \sigma_{\min}} \right)^{\frac{p}{\alpha}} \Gamma\left(1 - \frac{p}{\alpha}\right)$$

holds with high probability. Now, consider the function,

$$\Lambda(\alpha) = \frac{1}{\alpha^2} \left(\frac{\lambda_{\max}}{\lambda_{\min}} \right)^{\alpha} \left(\frac{1}{\alpha \sigma_{\min}} \right)^{\frac{p}{\alpha}} \Gamma\left(1 - \frac{p}{\alpha}\right).$$

We can compute that

$$\partial_\alpha \log \Lambda(\alpha) = \log\left(\frac{\lambda_{\max}}{\lambda_{\min}} \right) + \frac{p}{\alpha^2} \left[\log \alpha + \log \sigma_{\min} - 1 - \frac{2\alpha}{p} + \phi\left(1 - \frac{p}{\alpha}\right) \right],$$

where ϕ is the digamma function. For any arbitrary α_0, if we choose

$$\sigma_{\min} \geq \exp\left(1 + \frac{2}{p} - \log \alpha_0 - \phi\left(1 - \frac{p}{\alpha_0}\right) - \alpha_0^2 \log\left(\frac{\lambda_{\max}}{\lambda_{\min}} \right) \right),$$

then $\partial_\alpha \log \Lambda(\alpha) > 0$ for $\alpha \in [\alpha_0, 2)$. Hence, for all $\alpha_1, \alpha_2 \in [\alpha_0, 2)$, $\alpha_1 < \alpha_2$ it follows that $\Lambda(\alpha_1) \leq \Lambda(\alpha_2)$. This proves that $c(\alpha)$ is an increasing map in α. This completes the proof. ■

Appendix H. Useful Results

Here below, we provide a few technical results which are used in the proofs of the main results.

Lemma 19 *For any two positive numbers a and b, and for some $0 < \alpha \leq 2$, we have*

$$|a^\alpha - b^\alpha| \leq |a - b|(a^{\alpha-1} + b^{\alpha-1}). \tag{90}$$

Proof When $a = b > 0$, the result is obviously true. Without loss of generality, let us assume that $a > b > 0$ and by considering the RHS of (90), we get

$$|a - b|(a^{\alpha-1} + b^{\alpha-1}) = (a - b)(a^{\alpha-1} + b^{\alpha-1})$$

$$= a^\alpha + ab^{\alpha-1} - a^{\alpha-1}b - b^\alpha$$
$$= |a^\alpha - b^\alpha| + ab^{\alpha-1} - a^{\alpha-1}b.$$

Since, we have assumed that $a > b > 0$ and $\alpha > 0$, hence $ab^{\alpha-1} - a^{\alpha-1}b > 0$ always which essentially means,

$$|a^\alpha - b^\alpha| \leq |a-b|(a^{\alpha-1} + b^{\alpha-1}).$$

Same argument can be given while assuming $b > a > 0$. This completes the proof. ■

Lemma 20 *For any $a > 0$, and $k \in \mathbb{N}$,*

$$\sum_{j=0}^{k-1} ja^j = \frac{(k-1)a^{k+1} - ka^k + a}{(a-1)^2}.$$

In particular, for any $0 < a < 1$,

$$\sum_{j=0}^{\infty} ja^j = \frac{a}{(a-1)^2}.$$

Proof We can compute that

$$\sum_{j=0}^{k-1} ja^j = a\sum_{j=1}^{k-1} ja^{j-1} = a\frac{d}{da}\sum_{j=1}^{k-1} a^j = a\frac{d}{da}\left(\frac{a^k - a}{a-1}\right) = \frac{(k-1)a^{k+1} - ka^k + a}{(a-1)^2}.$$

The proof is complete. ■

Lemma 21 (Fourier transform of $|\theta^\top x|^p$) *Consider the function $f(\theta) = |\theta^\top x|^p$ for $p \in [1,2]$ and $h(u)$ denotes the Fourier transform of $f(\theta)$ where $u = [u_1, \cdots, u_d]$ is a vector in d-dimension. Given an unitary matrix $A \in \mathbb{R}^{d\times d}$ such that $A^\top A = AA^\top = I$ where I is an identity matrix in $\mathbb{R}^{d\times d}$ and $Ae_1 = \frac{x}{\|x\|_2}$ where e_i is vector in $\mathbb{R}^d$ with all entries set to 0 except ith entry which is set to 1, we have*

$$h(Au) = 2\|x\|_2^p(2\pi)^{d-1}\delta(u_2, \cdots, u_d)\Gamma(p+1)\cos\left(\frac{(p+1)\pi}{2}\right)\frac{1}{|u_1|^{p+1}} \quad \text{for } p \in [1,2),$$

and

$$h(Au) = 2\|x\|_2^p(2\pi)^{d-1}\delta(u_1, u_2, \cdots, u_d)\frac{2}{u_1^2} \quad \text{for } p = 2,$$

where δ is the Dirac-delta function.

Proof We recall that the Fourier transform is given by

$$\mathcal{F}f(u) = \int_{\mathbb{R}^d} f(\theta)e^{-\mathrm{i}u^\top\theta}\mathrm{d}\theta.$$

Let

$$h(u) := \mathcal{F}[|\langle x, \cdot\rangle|^p] = \|x\|_2^p \mathcal{F}\left[\left|\left\langle \frac{x}{\|x\|_2}, \cdot \right\rangle\right|^p\right].$$

We consider now an unitary matrix $A \in \mathbb{R}^{d\times d}$ such that $A^\top A = AA^\top = I$ where I is an identity matrix in $\mathbb{R}^{d\times d}$ and $Ae_1 = \frac{x}{\|x\|_2}$ where e_i is vector in $\mathbb{R}^d$ with all entries set to 0 except ith entry which is set to 1. Now let us compute $h(Au)$.

$$\begin{aligned} h(Au) &= \|x\|_2^p \int_{\mathbb{R}^d} \left|\left\langle \frac{x}{\|x\|_2}, \theta \right\rangle\right|^p e^{-\mathrm{i}(Au)^\top\theta}\, \mathrm{d}\theta \\ &= \|x\|_2^p \int_{\mathbb{R}^d} |\langle Ae_1, \theta\rangle|^p\, e^{-\mathrm{i}(Au)^\top\theta}\, \mathrm{d}\theta. \end{aligned}$$

In the above integral we substitute, $\beta = A^\top\theta$. Hence, when $p \in [1, 2)$, we have

$$\begin{aligned} h(Au) &= \|x\|_2^p \int_{\mathbb{R}^d} |\langle e_1, \beta\rangle|^p e^{-\mathrm{i}u^\top\beta}\, d\beta \\ &= \|x\|_2^p (2\pi)^{d-1}\delta(u_2, \cdots, u_d) \int_{-\infty}^{\infty} |\beta_1|^p\, e^{-\mathrm{i}u_1\beta_1}\, d\beta_1 \\ &= \|x\|_2^p (2\pi)^{d-1}\delta(u_2, \cdots, u_d) \int_{0}^{\infty} \left(e^{-\mathrm{i}u_1\beta_1} + e^{\mathrm{i}u_1\beta_1}\right) \beta_1^p\, d\beta_1 \\ &= 2\|x\|_2^p (2\pi)^{d-1}\delta(u_2, \cdots, u_d)\Gamma(p+1)\cos\left(\frac{(p+1)\pi}{2}\right)\frac{1}{|u_1|^{p+1}}. \end{aligned}$$

When $p = 2$, we have

$$\begin{aligned} h(Au) &= \|x\|_2^p \int_{\mathbb{R}^d} |\langle e_1, \beta\rangle|^2 e^{-\mathrm{i}u^\top\beta}\, d\beta \\ &= \|x\|_2^p (2\pi)^{d-1}\delta(u_2, \cdots, u_d) \int_{-\infty}^{\infty} |\beta_1|^2\, e^{-\mathrm{i}u_1\beta_1}\, d\beta_1 \\ &= 2\|x\|_2^p (2\pi)^{d-1}\delta(u_1, u_2, \cdots, u_d)\frac{2}{u_1^2}. \end{aligned} \tag{91}$$

This completes the proof. ■

Appendix I. Further Details on Experiment Settings and Resources

This section contains further details regarding the experiments presented in the main paper. As the synthetic data experiment setting was fully described in the text, most of the information below will pertain to the real data experiments with the exception of additional synthetic data results that include mean estimates. See the accompanying code regarding the implementation of the experiments described.

I.1. Additional synthetic data results

In addition to median and interquartile range based results presented in the paper, we add the following results in Figure 3 with a robust mean estimate of the results, demonstrating a similar pattern to that observed in the main paper.

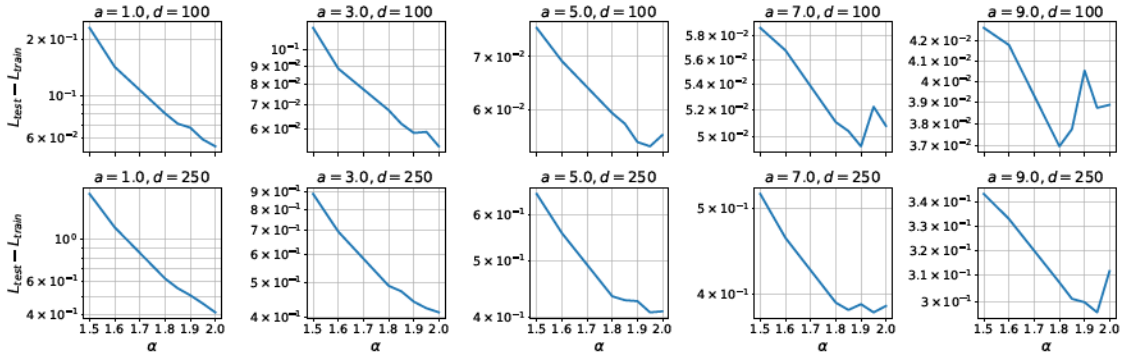

Figure 3: Results of the synthetic data experiments with varying a, α, and d. Each experiment was repeated 500 times with $n = 1000$. The lines correspond to the robust mean estimates with the samples with losses above the 90% quantile removed from among the respective experiments.

I.2. Datasets

The real data experiments involved a supervised learning setting, where images are classified into a number of predefined class labels. Each model architecture with given hyperparameters were trained on MNIST (LeCun et al., 2010), CIFAR10, and CIFAR100 (Krizhevsky, 2009) data sets[4]. The MNIST data set includes 28×28 black and white handwritten digits, with digits ranging from 0 to 9. The data set in its original form includes 60000 training and 10000 test samples. CIFAR10 and CIFAR100 are also image classification dataset comprising 32×32 color images of objects or animals, making up 10 and 100 classes respectively. There are 50000 training and 10000 test images in either of these data sets, and the instances are divided among classes equally. We used the standard train-test splits in all data sets.

I.3. Models

We used three different architectures in our experiments: a fully connected network with 4 hidden layers (FCN4), another fully connected network with 6 hidden layers (FCN6), and a convolutional neural network (CNN). In both FCN architectures, all hidden layer widths were 2048. All architectures featured ReLU activation functions. Batch normalization, dropout, residual layers, or any explicit regularization term in the loss function were not used in any part of the experiments. The architecture we chose for our CNN model closely follows that of VGG11 model (Simonyan and Zisserman, 2015), with the significant difference that only a single linear layer with a softmax output follows the convolutional layers presented below:

$$64, M, 128, M, 256, 256, M, 512, 512, M, 512, 512, M.$$

Here, integers describe the number of filters for 2-dimensional convolutional layers - for which the kernel sizes are 3×3. M stands for 2×2 max-pooling operations with a stride value of 2. This

4. MNIST and CIFAR10/100 data sets have been shared under Creative Commons Attribution-Share Alike 3.0 license and MIT License respectively.

architecture was slightly modified for the MNIST experiments by removing the first max-pooling layer due to the smaller dimensions of the MNIST images. The Table 1 describes the number of different parameters used for each model-dataset combination.

	FCN4	FCN6	CNN
MNIST	14,209,024	22,597,632	9,221,696
CIFAR10	18,894,848	27,283,456	9,222,848
CIFAR100	18,899,456	27,288,064	9,227,456

Table 1: Number of parameters for model-dataset combinations.

I.4. Training and hyperparameters

As described in the main text, the models were trained with SGD until convergence on the training set. The convergence criteria for MNIST and CIFAR-10 is a training negative log-likelihood (NLL) of $< 5 \times 10^{-5}$ and a training accuracy of 100%, and for CIFAR-100 these are a NLL of $< 1 \times 10^{-2}$ and a training accuracy of $> 99\%$. We use two different batch sizes ($b = 50, 100$) and a diversity of learning rates (η) to generate a large range of η/b values. Table 2 presents the η/b values created for each experiment setting. The varying nature of these ranges are due to the fact that different η/b values might lead to heavy-tailed behavior or divergence under different points in this hyperparameter space. Source code includes the enumerations of specific combinations of these hyperparameters for all settings.

	FCN4	FCN6	CNN
MNIST	5×10^{-5} to 1.14×10^{-2}	5×10^{-5} to 8.8×10^{-3}	1×10^{-5} to 6.35×10^{-3}
CIFAR10	5×10^{-5} to 2.7×10^{-3}	2.5×10^{-5} to 4×10^{-3}	1×10^{-5} to 1.5×10^{-3}
CIFAR100	1×10^{-5} to 1.6×10^{-3}	1×10^{-5} to 2.25×10^{-3}	1×10^{-5} to 7×10^{-4}

Table 2: The ranges of η/b for all experiments.

I.5. Tail-index estimation

The multivariate estimator proposed by Mohammadi et al. (2015) was used for tail-index estimation:

Theorem 22 ((Mohammadi et al., 2015, Corollary 2.4)) *Let $\{X_i\}_{i=1}^{K}$ be a collection of i.i.d. random vectors where each X_i is multivariate strictly stable with tail-index α, and $K = K_1 \times K_2$. Define $Y_i := \sum_{j=1}^{K_1} X_{j+(i-1)K_1}$ for $i \in \{1, \dots, K_2\}$. Then, the estimator*

$$\widehat{\frac{1}{\alpha}} \triangleq \frac{1}{\log K_1} \Big(\frac{1}{K_2} \sum_{i=1}^{K_2} \log \|Y_i\| - \frac{1}{K} \sum_{i=1}^{K} \log \|X_i\| \Big) \tag{92}$$

converges to $1/\alpha$ almost surely, as $K_2 \to \infty$.

Previous deep learning research such as Tzagkarakis et al. (2018); Şimşekli et al. (2019); Barsbey et al. (2021) have also used this estimator. As described in the main text, tail-index estimation

is conducted on the ergodic averaged version of the parameters, an operation which does not change the tail-index of the parameters, to conform to this estimator's assumptions. We use the columns of parameters in FCN's and specific filter parameters in CNN as the random vectors instances for the multivariate distribution. Before conducting the tail-index estimation we center the parameters using the index-wise median values. We observe that (i) centering with mean values, and/or (ii) using the alternative univariate tail-index estimator (Mohammadi et al., 2015, Corollary 2.2) from the same paper produces qualitatively identical results. We also observe that using alternative tail index estimators with symmetric α-stable assumption produces no qualitatively significant differences in the estimated values (Sathe and Upadhye, 2022).

I.6. Hardware and software resources

The computational resources for the experiments were provided by a research institute. The bulk of the resources were expended on the real data experiments, where a roughly equal division of labor between Nvidia Titan X, 1080 Ti, and 1080 model GPU's. Our results rely on 273 models, training of which brings about a GPU-heavy computational workload. The training of a single model took approximately 4.5 hours, with an approximate estimated total GPU time for the ultimate results 1270 hours. This total also includes the training time for the 40 models which diverged during training, with the training stopping around 1 hour mark on average. The computational time expended for tail-index estimation in real data experiments and the totality of synthetic experiments amounted to approximately 20 hours of computation with similar hardware as described above.

The experiments were implemented in the Python programming language. For the real data experiments, the deep learning framework PyTorch (Paszke et al., 2019) was extensively used, including the implementation methodology in some of its tutorials[5]. PyTorch is shared under the Modified BSD License.

5. HTTPS://GITHUB.COM/PYTORCH/VISION/BLOB/MASTER/TORCHVISION/MODELS/VGG.PY