# Algorithmic Stability of Heavy-Tailed SGD with General Loss Functions

**Anant Raj** [*1 2] **Lingjiong Zhu** [*3] **Mert Gürbüzbalaban** [4 5] **Umut Şimşekli** [2]

## Abstract

Heavy-tail phenomena in stochastic gradient descent (SGD) have been reported in several empirical studies. Experimental evidence in previous works suggests a strong interplay between the heaviness of the tails and generalization behavior of SGD. To address this empirical phenomena theoretically, several works have made strong topological and statistical assumptions to link the generalization error to heavy tails. Very recently, new generalization bounds have been proven, indicating a non-monotonic relationship between the generalization error and heavy tails, which is more pertinent to the reported empirical observations. While these bounds do not require additional topological assumptions given that SGD can be modeled using a heavy-tailed stochastic differential equation (SDE), they can only apply to simple quadratic problems. In this paper, we build on this line of research and develop generalization bounds for a more general class of objective functions, which includes non-convex functions as well. Our approach is based on developing Wasserstein stability bounds for heavy-tailed SDEs and their discretizations, which we then convert to generalization bounds. Our results do not require any nontrivial assumptions; yet, they shed more light to the empirical observations, thanks to the generality of the loss functions.

## 1. Introduction

Many supervised learning problems can be expressed as an instance of the *risk minimization problem*

$$\min_{\theta\in\mathbb{R}^d} \left\{F(\theta) := \mathbb{E}_{x\sim\mathcal{D}}[f(\theta,x)]\right\}, \tag{1}$$

where $x \in \mathcal{X}$ is a random data point, distributed according to an unknown probability distribution $\mathcal{D}$ and taking values in the data space $\mathcal{X}$, $\theta$ denotes the parameter vector of the model to be learned and $f(\theta, x)$ is the instantaneous loss of misprediction with parameters $\theta$ corresponding to the data point $x$. With different choices of the function $f$, we can recover many problems in supervised learning from deep learning to logistic regression or support vector machines (Shalev-Shwartz & Ben-David, 2014).

As $\mathcal{D}$ is unknown in many scenarios, directly attacking (1) is often not possible. Assuming we have access to a training dataset $X_n = \{x_1, \dots, x_n\} \subset \mathcal{X}^n$ with $n$ independent and identically distributed (i.i.d.) observations, in practice, we can consider the *empirical risk minimization* (ERM) problem instead, given as follows:

$$\min_{\theta\in\mathbb{R}^d} \left\{\hat{F}(\theta, X_n) := \frac{1}{n}\sum_{i=1}^{n} f(\theta, x_i)\right\}.$$

One of the most popular algorithms for attacking the ERM problem is stochastic gradient descent (SGD) that is based on the following recursion:

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

where $\eta$ is the step-size (or learning-rate) and

$$\nabla\tilde{F}_k(\theta, X) := \frac{1}{b}\sum_{i\in\Omega_k} \nabla f(\theta, x_i)$$

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.

Understanding the generalization properties of SGD has been a major challenge in modern machine learning. In this context, the goal is to bound the so-called generalization error: $|\hat{F}(\theta, X_n) - F(\theta)|$, either in expectation or in high probability.

[*]Equal contribution [1]Coordinated Science Laboraotry, University of Illinois Urbana-Champaign, IL, USA [2]Inria, Ecole Normale Supérieure, PSL Research University, Paris, France [3]Department of Mathematics, Florida State University, FL, USA [4]Department of Management Science and Information Systems, Rutgers University, NJ, USA [5]Princeton University, NJ, USA. Correspondence to: Umut Şimşekli <umut.simsekli@inria.fr>.

*Proceedings of the $40^{th}$ International Conference on Machine Learning*, Honolulu, Hawaii, USA. PMLR 202, 2023. 

While a plethora of approaches have been proposed to address this task (Cao & Gu, 2019; Lei & Ying, 2020; Neu et al., 2021; Park et al., 2022), a promising approach among those has been based on the theoretical and empirical observations which showed that SGD can exhibit a *heavy-tailed* behavior, depending on the choice of hyperparameters ($\eta$ and $b$), the data distribution $\mathcal{D}$, and the geometry of the loss function $f$ (Gürbüzbalaban et al., 2021; Hodgkinson & Mahoney, 2021). This has motivated the use of 'heavy-tailed proxies' for SGD, which –to some extent– facilitated the analysis of SGD in terms of its generalization error. Examples of such proxies include gradient descent with additive heavy-tailed noise:

$$\theta_{k+1} = \theta_k - \eta \nabla \hat{F}(\theta_k, X_n) + \xi_{k+1}, \tag{3}$$

where $(\xi_k)_{k\geq 1}$ is a sequence of heavy-tailed random vectors, potentially with unbounded higher-order moment, i.e., $\mathbb{E}\|\xi_k\|^p = +\infty$ for some $p > 1$ (see e.g., (Nguyen et al., 2019; Zhang et al., 2020; Wang et al., 2021)).

Another popular proxy for heavy-tailed SGD is based on a *continuous-time* version of (3), which is expressed by the following stochastic differential equation (SDE):

$$d\theta_t = -\nabla \hat{F}(\theta_t, X_n)dt + \sigma d\mathrm{L}_t^\alpha, \tag{4}$$

where $\sigma > 0$ is a scale parameter, $\mathrm{L}_t^\alpha$ is a $d$-dimensional $\alpha$-stable Lévy process, which has heavy-tailed increments and will be formally defined in the next section[1], and $\alpha \in (0, 2]$ denotes the 'tail-exponent' such that as $\alpha$ gets smaller the process $\mathrm{L}_t^\alpha$ becomes heavier-tailed.

Within this mathematical framework, Şimşekli et al. (2020) proved an upper-bound (which was then improved in (Hodgkinson et al., 2021)) for the worst-case generalization error over the trajectories of (4). The bound informally reads as follows: with probability at least $1 - \delta$, it holds that

$$\sup_{\theta\in\Theta} \left| \hat{F}(\theta, X_n) - F(\theta) \right| \lesssim \sqrt{\frac{\alpha + I(\Theta, X_n) + \log(1/\delta)}{n}},$$

where $\Theta$ denotes the trajectory of (4), i.e.,

$$\Theta := \left\{ \theta \in \mathbb{R}^d : \exists t \in [0, 1], \theta = \theta_t \right\},$$

with $\theta_t$ being the solution of (4), and $I(\Theta, X_n)$ denotes a form of 'mutual information' between the trajectory $\Theta$ and the data sample $X_n$ (cf. (Xu & Raginsky, 2017)). This result suggests that the generalization error is essentially determined by two terms: (i) the tail exponent $\alpha$, as the tails get heavier the generalization error will be lower, (ii) the statistical dependency between the trajectory and the data sample, the lower the dependency the better the generalization performance.

While these results illuminated an interesting connection between heavy-tails and generalization, they unfortunately rely on nontrivial topological assumptions on $\Theta$ and the mutual information term cannot be controlled in an interpretable way in general. On the other hand, Barsbey et al. (2021) empirically illustrated that the relation between the tail exponent and the generalization error might not be monotonic in practical applications; an observation which cannot be directly supported by the bound in (Şimşekli et al., 2020) and (Hodgkinson et al., 2021).

Aiming to alleviate these issues, very recently, Raj et al. (2023) considered the same problem from the lens of algorithmic stability (Bousquet & Elisseeff, 2002; Hardt et al., 2016). They considered the SDE (4) and further simplified it by choosing the loss function as a simple quadratic, i.e., $f(\theta, x) = (\theta^\top x)^2$. They showed that any parameter vector $\theta$ provided by (4) (or its Euler-Maruyama discretization with small enough small step-size) cannot be algorithmically stable. However, when the algorithmic stability is measured by a *surrogate loss function* instead (reminiscent of (Wang et al., 2021)), the parameter vector $\theta$ becomes algorithmically stable, which immediately implies generalization. Their bound further illustrated that the relation between $\alpha$ and the generalization error might not be monotonic, which is in line with the observations provided in (Barsbey et al., 2021).

While the bounds in (Raj et al., 2023) do not require additional topological assumptions and do not contain a mutual information term as opposed to (Şimşekli et al., 2020; Hodgkinson et al., 2021), their analysis technique heavily relies on the fact that $f$ is a quadratic, hence cannot be directly extended beyond quadratic loss functions.

In this paper, we aim at filling this gap and prove algorithmic stability bounds the SDE (4) (and its Euler-Maruyama discretization) with general loss functions, which can be even non-convex. Our contributions are as follows:

- We first focus on the continuous-time setting and prove Wasserstein stability bounds for two SDEs of the form of (4) with different drift functions. Our results cover both the finite-time case, i.e., $t < \infty$ and the stationary case, i.e., $t \to \infty$. We build upon recently introduced stochastic analysis tools for uniform-in-time Wasserstein error bounds for Euler-Maruyama discretization (Chen et al., 2022) to obtain a novel Wasserstein stability bound for two $\alpha$-stable Lévy-driven SDEs. Our analysis relies on an additional pseudo-Lipschitz like condition for the underlying process and the dataset (Assumption 3.1) and careful adaption of the tools in (Chen et al., 2022) to our context (Lemma C.5 and Theorem B.1 in the Supplementary Doc-

[1] This type of SDEs have also received some attention in terms of limits of deterministic gradient descent with dynamical regularization (Lim et al., 2022).

ument) as well as additional analysis (Lemma 3.5) that allows us to characterize the dependence of our bounds on the tail-index $\alpha$. Our derived bounds would be interesting on their own to a much broader scope.

- By following (Raginsky et al., 2016), we translate the derived Wasserstein stability bounds to algorithmic stability bounds. Similar to (Raj et al., 2023), our approach necessitates surrogate loss functions to measure algorithmic stability. Our results reveal that the relation between heaviness of the tail $\alpha$ and the generalization error might not be monotonic, indicating that the conclusions of (Raj et al., 2023) extends to the general case.

- By combining our results with (Chen et al., 2022), we extend our bounds to the Euler-Maruyama discretization of (4) (that is of the form of (3)) and show that for small enough step-sizes the discrete-time process achieves almost identical stability bounds.

Contrary to (Şimşekli et al., 2020; Hodgkinson et al., 2021; Lim et al., 2022), our bounds do not rely on any topological regularity assumptions and they further do not contain a mutual information term. Moreover, our results shed more light to the non-monotonic relation between heavy tails and the generalization error, as empirically observed in (Barsbey et al., 2021; Raj et al., 2023), since they are applicable to non-convex losses, as opposed to (Raj et al., 2023). We also note that our generalization bounds and Wasserstein bounds are independent of time. Such a result was previously shown in (Farghly & Rebeschini, 2021) in the context of Brownian-motion driven SDEs and their discretizations, our work uses different techniques considering Lévy-driven SDEs and studies the link between the generalization and the coefficient of heavy tail.

## 2. Notations and Technical Background

**Gradients and Hessians.** For any twice continuously differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, we denote by $\nabla f$ and $\nabla^2 f$ the gradient and the Hessian of $f$. First-order and second-order directional derivatives of $f$ are defined as

$$\nabla_v f(x) := \lim_{\epsilon \to 0} \frac{f(x+\epsilon v) - f(x)}{\epsilon},$$
$$\nabla_{v_2} \nabla_{v_1} f(x) := \lim_{\epsilon \to 0} \frac{\nabla_{v_1} f(x+\epsilon v_2) - \nabla_{v_1} f(x)}{\epsilon}, \quad (5)$$

for any directions $v, v_1, v_2 \in \mathbb{R}^d$. If $f$ is three times continuously differentiable, then third-order derivatives along the directions $v_1, v_2$ are given by

$$\nabla_{v_2} \nabla_{v_1} \nabla f(x) := \lim_{\epsilon \to 0} \frac{\nabla_{v_1} \nabla f(x+\epsilon v_2) - \nabla_{v_1} \nabla f(x)}{\epsilon}. \quad (6)$$

**Wasserstein distance.** For $p \geq 1$, the $p$-Wasserstein distance between two probability measures $\mu$ and $\nu$ on $\mathbb{R}^d$ is defined as (Villani, 2009):

$$\mathcal{W}_p(\mu, \nu) = \{\inf \mathbb{E}\|X - Y\|^p\}^{1/p}, \quad (7)$$

where the infimum is taken over all coupling of $X \sim \mu$ and $Y \sim \nu$. In particular, the 1-Wasserstein distance has the following dual representation (Villani, 2009):

$$\mathcal{W}_1(\mu, \nu) = \sup_{h \in \mathrm{Lip}(1)} \left| \int_{\mathbb{R}^d} h(x)\mu(dx) - \int_{\mathbb{R}^d} h(x)\nu(dx) \right|, \quad (8)$$

where $\mathrm{Lip}(1)$ denotes the set of functions $h : \mathbb{R}^d \to \mathbb{R}$ that are 1-Lipschitz.

**Algorithmic stability.** Algorithmic stability is an important notion in learning theory, which has pave the way for several important theoretical results (Bousquet & Elisseeff, 2002; Hardt et al., 2016). Let us first state the notion of algorithmic stability as defined in (Hardt et al., 2016).

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

$$\sup_{X \cong \hat{X}} \sup_{z \in \mathcal{X}} \mathbb{E}\left[\ell(\mathcal{A}(X), z) - \ell(\mathcal{A}(\hat{X}), z)\right] \leq \varepsilon, \quad (9)$$

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

Here, we intentionally use a different notation for the loss $\ell$ (as opposed to $f$), as our theory will require the algorithmic stability to be measured by using a surrogate loss function, which might be different than the original loss $f$.

We now provide a result from (Hardt et al., 2016) which relates algorithmic stability to the generalization performance of a randomized algorithm. Before stating the result, similar to $\hat{F}$ and $F$, we respectively define the empirical and population risks with respect to the loss $\ell$ as follows:

$$\hat{R}(w, X_n) := \frac{1}{n} \sum_{i=1}^{n} \ell(w, x_i), \quad R(w) := \mathbb{E}_{x \sim \mathcal{D}}[\ell(w, x)].$$

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

$$\left| \mathbb{E}_{\mathcal{A}, X_n} \left[ \hat{R}(\mathcal{A}(X_n), X_n) - R(\mathcal{A}(X_n)) \right] \right| \leq \varepsilon. \quad (10)$$

**Alpha-stable distributions.** A scalar random variable $X$ is said to follow a symmetric $\alpha$-stable distribution, denoted by $X \sim \mathcal{S}\alpha\mathcal{S}(\sigma)$, if its characteristic function takes the form: $\mathbb{E}\left[e^{iuX}\right] = \exp\left(-\sigma^\alpha |u|^\alpha\right)$, for any $u \in \mathbb{R}$, where $\sigma > 0$ is

known as the scale parameter that measures the spread of $X$ around 0 and for $\alpha \in (0, 2]$ which is known as the tail-index that determines the tail thickness of the distribution. The tail becomes heavier as $\alpha$ gets smaller. The $\alpha$-stable distribution $\mathcal{S}\alpha\mathcal{S}$ 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). The probability density function of a symmetric $\alpha$-stable distribution, $\alpha \in (0, 2]$, does not yield closed-form expression in general except for a few special cases; for example $\mathcal{S}\alpha\mathcal{S}$ reduces to the Cauchy and the Gaussian distributions, respectively, when $\alpha = 1$ and $\alpha = 2$. When $0 < \alpha < 2$, the moments are finite only up to the order $\alpha$ in the sense that $\mathbb{E}[|X|^p] < \infty$ if and only if $p < \alpha$, which implies infinite variance.

Finally, $\alpha$-stable distribution can be extended to the high-dimensional case for random vectors. One of the most commonly used extension is the rotationally symmetric $\alpha$-stable distribution. $X$ follows a $d$-dimensional rotationally symmetric $\alpha$-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 $u \in \mathbb{R}^d$. For further details of $\alpha$-stable distributions, we refer to (Samorodnitsky & Taqqu, 1994).

**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, the Brownian motion, the Cauchy process, and more generally stable processes; see e.g. (Bertoin, 1996; Samorodnitsky & Taqqu, 1994; Applebaum, 2009). Lévy processes in general admit jumps and have heavy tails which are appealing in many applications; see e.g. (Cont & Tankov, 2004).

In this paper, we will consider the rotationally symmetric $\alpha$-stable Lévy process, denoted by $\mathrm{L}_t^\alpha$ in $\mathbb{R}^d$ and is defined as follows.

- $\mathrm{L}_0^\alpha = 0$ almost surely;
- For any $t_0 < t_1 < \cdots < t_N$, the increments $\mathrm{L}_{t_n}^\alpha - \mathrm{L}_{t_{n-1}}^\alpha$ are independent;
- The difference $\mathrm{L}_t^\alpha - \mathrm{L}_s^\alpha$ and $\mathrm{L}_{t-s}^\alpha$ have the same distribution, with the characteristic function $\exp(-(t-s)^\alpha \|u\|_2^\alpha)$ for $t > s$;
- $\mathrm{L}_t^\alpha$ 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 $\mathrm{B}_t$ is the standard $d$-dimensional Brownian motion.

## 3. Main Results

In this section, we present our main theoretical results. To ease the notation, we will consider the following SDE in lieu of (4):

$$d\theta_t = -\nabla \hat{F}(\theta_t, X_n)dt + d\mathrm{L}_t^\alpha, \tag{11}$$

in the rest of the paper[2].

Our road map is as follows. We will first consider the continuous-time case (11), i.e., we will set the learning algorithm as $\mathcal{A}(X_n) = \theta_t$ for some $t \in [0, +\infty]$, where $\theta_\infty$ denotes a sample from the stationary distribution of the SDE (11). As our aim is to prove algorithmic stability bounds for this choice of algorithm, we then consider another dataset $\hat{X}_n \cong X_n$, which differ from $X_n$ by one element, accordingly define the following SDE:

$$d\hat{\theta}_t = -\nabla \hat{F}(\hat{\theta}_t, \hat{X}_n)dt + d\mathrm{L}_t^\alpha, \tag{12}$$

such that $\mathcal{A}(\hat{X}_n) = \hat{\theta}_t$. Then, we will argue that, for any time $t$, the laws of $\theta_t$ and $\hat{\theta}_t$ will be close to each other in the 1-Wasserstein metric.

By considering a surrogate loss function $\ell$, which we will assume to be $\mathcal{L}$-Lipschitz, our bound on the Wasserstein distance between $\mathrm{Law}(\theta_t)$ and $\mathrm{Law}(\hat{\theta}_t)$ (Theorem 3.3) will immediately provide us a generalization bound thanks to the dual representation of the 1-Wasserstein distance (cf. (Raginsky et al., 2016, Lemma 3)):

$$\begin{aligned}\left|\mathbb{E}_{\theta_t, X_n}\left[\hat{R}(\theta_t, X_n) - R(\theta_t)\right]\right| \\ \leq \mathcal{L} \sup_{X_n \cong \hat{X}_n} \mathcal{W}_1\left(\mathrm{Law}(\theta_t), \mathrm{Law}(\hat{\theta}_t)\right),\end{aligned} \tag{13}$$

where $\mathrm{Law}(\theta_t)$ and $\mathrm{Law}(\hat{\theta}_t)$ respectively depend on $X_n$ and $\hat{X}_n$ due to the form of the SDEs. The reason why we require a surrogate loss function is the fact that we need the Lipschitz continuity of the loss to be able to derive the bound in (13). However, as we will detail in the next subsection, our assumptions on the true loss $f$ will be incompatible with the Lipschitz continuity of $f$.

After proving a generalization bound of the form (13), we will further investigate the behavior of the bound with respect to the heaviness of the tail which is characterized by the tail-index $\alpha$. Finally, we will consider the discrete-time case, where we will show that almost identical results hold for the Euler-Maruyama discretizations of (11) and (12), as long as a sufficiently small step-size is chosen.

### 3.1. Assumptions

In this section we state our main assumptions and we will assume that they hold throughout the paper. For any $w \in \mathbb{R}^d$, we use $\theta_t^w$ to denote the process $\theta_t$ that starts at $\theta_0 = w$.

[2]Note that the stationary distribution of (4) is the same as the stationary distribution of $d\theta_t = -\sigma^{-\alpha}\nabla \hat{F}(\theta_t, X_n)dt + d\mathrm{L}_t^\alpha$, and so that we can easily adapt our main result (Theorem 3.3) to the general case $\sigma > 0$.

**Assumption 3.1.** There exist universal constants $K_1$, $K_2$ such that for every $x, \hat{x} \in \mathcal{X}$ and $\theta, \hat{\theta} \in \mathbb{R}^d$:

$$\begin{aligned}&\|\nabla f(\theta, x) - \nabla f(\hat{\theta}, \hat{x})\| \\ &\leq K_1\|\theta - \hat{\theta}\| + K_2\|x - \hat{x}\|(\|\theta\| + \|\hat{\theta}\| + 1). \end{aligned} \tag{14}$$

This assumption is a pseudo-Lipschitz like condition (similar to the one in (Erdogdu et al., 2018)) on the loss $f$. Under this assumption, for two datasets $X_n$ and $\hat{X}_n$, we immediately have the following property:

$$\begin{aligned}\left\|\nabla \hat{F}(\theta, X_n) - \nabla \hat{F}\left(\hat{\theta}, \hat{X}_n\right)\right\| &\leq K_1\|\theta - \hat{\theta}\| \\ &+\rho(X_n, \hat{X}_n)K_2\left(\|\theta\| + \|\hat{\theta}\| + 1\right), \end{aligned} \tag{15}$$

for any $\theta, \hat{\theta} \in \mathbb{R}^d$, where

$$\rho(X_n, \hat{X}_n) := \frac{1}{n}\sum_{i=1}^{n}\|x_i - \hat{x}_i\|. \tag{16}$$

We will show that the term $\rho(X_n, \hat{X}_n)$ will have an important role in terms of Wasserstein stability.

By following (Chen et al., 2022), we also make the following assumption.

**Assumption 3.2.** For every $x \in \mathcal{X}$, $f(\cdot, x)$ is three-times continuously differentiable, and there exist universal positive constants $B$, $m$, $K$, $L$, and $M$ such that for any $\theta_1, \theta_2 \in \mathbb{R}^d$ and $x \in \mathcal{X}$:

$$\begin{aligned}&\|\nabla f(0, x)\| \leq B, \\ &\langle \nabla f(\theta_1, x) - \nabla f(\theta_2, x), \theta_1 - \theta_2\rangle \\ &\qquad\qquad \geq m\|\theta_1 - \theta_2\|^2 - K,\end{aligned}$$

and for any $\theta, v, v_1, v_2 \in \mathbb{R}^d$ and $x \in \mathcal{X}$:

$$\begin{aligned}&\|\nabla_v \nabla f(\theta, x)\| \leq L\|v\|, \\ &\|\nabla_{v_1}\nabla_{v_2}\nabla f(\theta, x)\| \leq M\|v_1\|\|v_2\|.\end{aligned}$$

The first part of this assumption is common in stochastic analysis and often referred to as dissipativity (Raginsky et al., 2017; Gao et al., 2022). The second part of the assumption amounts to requiring the drift $\nabla f(x)$ to have bounded third-order directional derivatives (see also (Chen et al., 2022)). This would be satisfied for instance if $f$ has bounded third-order derivatives on the set $\mathcal{X}$.

### 3.2. Continuous-Time Dynamics

Now, we are ready to state our first theorem that characterizes the 1-Wasserstein distance between $\theta_t$ and $\hat{\theta}_t$ at any finite time $t$, which is uniform in $t$. As a result, we also obtain an upper-bound on the 1-Wasserstein distance between the unique invariant distribution $\mu$ of $(\theta_t)_{t\geq 0}$ and the unique invariant distribution $\hat{\mu}$ of $(\hat{\theta}_t)_{t\geq 0}$.[3]

The full statement of the theorem is rather lengthy and is given in the Section B.1 in the Supplementary Document. For clarity, in the next theorem, we provide our upper-bound on the distance between the invariant distributions, i.e., $t \to \infty$. The finite $t$ case is handled in the Supplementary Document.

**Theorem 3.3.** *Suppose that Assumptions 3.1 and 3.2 hold. Denote by $\mu, \hat{\mu}$ the unique invariant distributions of $(\theta_t)_{t\geq 0}$ and $(\hat{\theta}_t)_{t\geq 0}$, respectively. Then, the following inequality holds:*

$$\mathcal{W}_1(\mu, \hat{\mu}) \leq \left(C_1\lambda^{-1}e^{\lambda} + 1\right)e^{L}\rho(X_n, \hat{X}_n)K_2\left(2C_0 + 1\right), \tag{17}$$

*where $K_2$ and $L$ are defined in Assumption 3.1 and Assumption 3.2 and $C_0, C_1$ and $\lambda$ are some positive real constants.*

This theorem shows that, as long as the datasets $X_n$ and $\hat{X}_n$ are close to each other, i.e., $\rho(X_n, \hat{X}_n)$ is small, the distance between the solutions of the SDEs (11) and (12) will be small as well for any time $t$. This result can be seen as a heavy-tailed version of the results presented in (Raginsky et al., 2017; Farghly & Rebeschini, 2021).

**Generalization Bound.** By combining Theorem 3.3 and (13), we can now easily obtain generalization bound under a Lipschitz surrogate loss function.

**Corollary 3.4.** *Suppose that Assumptions 3.1 and 3.2 hold. Assume that $\ell$ is $\mathcal{L}$-Lipschitz in $\theta$ and $\sup_{x,y\in\mathcal{X}}\|x - y\| \leq D$ for some $D < \infty$. Then the following inequality holds:*

$$\begin{aligned}&\left|\mathbb{E}_{\theta_\infty, X_n}\left[\hat{R}(\theta_\infty, X_n) - R(\theta_\infty)\right]\right| \\ &\qquad\leq \frac{\mathcal{L}D\left(C_1\lambda^{-1}e^{\lambda} + 1\right)e^{L}K_2\left(2C_0 + 1\right)}{n}.\end{aligned}$$

The proof of this corollary is straightforward, hence omitted. Similar to Theorem 3.3, we presented Corollary 3.4 for the stationary case, where $t \to \infty$; yet, we shall underline that our theory holds for any finite time $t$.

Lower bounds on algorithmic stability have been discussed in (Raj et al., 2023) for Ornstein-Uhlenbeck process with $\alpha$-stable Lévy noise. While comparing with the bound obtained in this work, we can see that the obtained bound has optimal dependence on the number of samples $n$.

Next, we will investigate how the constants in Theorem 3.3 behave with respect to varying $\alpha$.

**Constants in Theorem 3.3.** In Theorem 3.3, we provided an upper bound on $\mathcal{W}_1(\mu, \hat{\mu})$ which depends on various

[3] Here, we know that under our assumptions by the results of (Chen et al., 2022), invariant distributions $\mu$ and $\hat{\mu}$ exist.

quantities, and our next goal is to figure out how the parameters $C_1, \lambda$, $L, K_2$ and $C_0$ depend on the tail-index $\alpha$.

First, we notice that the parameters $L$ and $K_2$ only depend on the loss function. Second, the parameters $C_1, \lambda$ come from the 1-Wasserstein contraction in Lemma C.2 in the Supplementary Document which is a restatement of Proposition 2.2 in (Chen et al., 2022); that is, for any $w, y \in \mathbb{R}^d$:

$$\mathcal{W}_1\left(\text{Law}\left(\theta_t^w\right), \text{Law}\left(\theta_t^y\right)\right) \leq C_1 e^{-\lambda t}\|w-y\|, \tag{18}$$

$$\mathcal{W}_1\left(\text{Law}\left(\hat{\theta}_t^w\right), \text{Law}\left(\hat{\theta}_t^y\right)\right) \leq C_1 e^{-\lambda t}\|w-y\|, \tag{19}$$

where $\theta_t^w$ to denote the process $\theta_t$ that starts at $\theta_0 = w$. Furthermore, Proposition 2.2 in (Chen et al., 2022) follows from Theorem 1.2. in (Wang, 2016). A careful look at Theorem 1.2. in (Wang, 2016) reveals that $C_1$, $\lambda$ are independent of the tail-index $\alpha$.

Finally, $C_0$ depends on $\alpha$ and it comes from Lemma C.1 in the Supplementary Document which is a restatement of Proposition 2.1. in (Chen et al., 2022); that is, which says that for any $w \in \mathbb{R}^d$:

$$\mathbb{E}\|\theta_t^w\| \leq C_0(1+\|w\|), \qquad \text{for any } t > 0, \tag{20}$$

$$\mathbb{E}\|\hat{\theta}_t^w\| \leq C_0(1+\|w\|), \qquad \text{for any } t > 0. \tag{21}$$

Notice that Proposition 2.1. in (Chen et al., 2022) does not provide an explicit formula for $C_0$, and in the next result, we provide a more refined estimate to spell out the dependence of $C_0$ on the tail-index $\alpha$.

**Lemma 3.5.** *Suppose that Assumptions 3.1 and 3.2 hold. For any $w \in \mathbb{R}^d$, we have*

$$\mathbb{E}\|\theta_t^w\| \leq C_0(1+\|w\|), \qquad \textit{for any } t > 0, \tag{22}$$

$$\mathbb{E}\|\hat{\theta}_t^w\| \leq C_0(1+\|w\|), \qquad \textit{for any } t > 0, \tag{23}$$

*where we can take*

$$C_0 := 3 + \frac{2\left(K+B\right)}{m} + \frac{2^{\alpha+1}\Gamma\left(\frac{d+\alpha}{2}\right)\pi^{-d/2}\sigma_{d-1}}{|\Gamma(-\alpha/2)|m}\left(\frac{\sqrt{d}}{2-\alpha} + \frac{1}{\alpha-1}\right), \tag{24}$$

*where $\Gamma(\cdot)$ is the gamma function and $\sigma_{d-1} = 2\pi^{\frac{d}{2}}/\Gamma(d/2)$ is the surface area of the unit sphere in $\mathbb{R}^d$, and $K, B, m$ are defined in Assumption 3.2.*

Hence, it follows from Theorem 3.3 and Lemma 3.5 that the dependence on the tail-index $\alpha$ is only via the function:

$$g(\alpha; d) := \frac{2^\alpha \Gamma\left(\frac{d+\alpha}{2}\right)\sqrt{d}}{|\Gamma(-\alpha/2)|(2-\alpha)} + \frac{2^\alpha \Gamma\left(\frac{d+\alpha}{2}\right)}{|\Gamma(-\alpha/2)|(\alpha-1)}.$$

The next result formalizes how the function $g(\alpha; d)$ depends on the tail-index $\alpha$.

**Proposition 3.6.** *Let $\alpha_0 := 2(c_0 - 1) \in (0, 2)$, where $c_0$ is the unique critical value in $(1, 2)$ such that the gamma function $\Gamma(x)$ is increasing for any $x > c_0$ and decreasing for any $1 < x < c_0$. Then, the following holds.*

*(i) For any $d \geq d_0$, where $d_0 := \max\left(2, \frac{1}{(\log 2)^2(\alpha_0-1)^4}\right)$, the map $\alpha \mapsto g(\alpha; d)$ is increasing in $\alpha \in [\alpha_0, 2]$.*

*(ii) For any fixed $d \in \mathbb{N}$, the map $\alpha \mapsto g(\alpha; d)$ is decreasing in $\alpha \in [1, \alpha_0']$, where $\alpha_0' \leq \alpha_0$ is defined as $\alpha_0' := \min\left(\alpha_0, 1 + \frac{-1+\sqrt{1+4y_0^{-1}\sqrt{d}}}{2\sqrt{d}}\right)$, with $y_0 := \log(2) + \frac{1}{2}\psi(d + \frac{\alpha}{2}) + \frac{3-\alpha_0}{2-\alpha_0}$, where $\psi(\cdot)$ is the digamma function.*

*Proof.* Let us first prove part (i). First, we can re-write $g(\alpha; d)$ as

$$g(\alpha; d) = \frac{2^\alpha \Gamma\left(\frac{d+\alpha}{2}\right)}{|\Gamma(-\alpha/2)|(2-\alpha)}\left(\sqrt{d} + \frac{2-\alpha}{\alpha-1}\right). \tag{25}$$

By the properties of the gamma function, we have

$$\begin{aligned}\Gamma\left(2-\frac{\alpha}{2}\right) &= \left(1-\frac{\alpha}{2}\right)\Gamma\left(1-\frac{\alpha}{2}\right) \\ &= \left(1-\frac{\alpha}{2}\right)\frac{-\alpha}{2}\Gamma(-\alpha/2).\end{aligned}$$

Therefore, we have

$$|\Gamma(-\alpha/2)|(2-\alpha) = \frac{4}{\alpha}\Gamma\left(2-\frac{\alpha}{2}\right). \tag{26}$$

Moreover, by the properties of the gamma function,

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

Hence, we conclude that

$$\begin{aligned}g(\alpha; d) = {}& 2^{\alpha-2}\alpha\Gamma\left(\frac{d+\alpha}{2}\right)\Gamma\left(\frac{\alpha}{2}\right) \\ &\cdot \frac{\sin\left(\pi\left(1-\frac{\alpha}{2}\right)\right)}{\pi\left(1-\frac{\alpha}{2}\right)}\left(\sqrt{d} + \frac{2-\alpha}{\alpha-1}\right),\end{aligned}$$

where we used $\sin(-x) = -\sin(x)$ for any $x \in \mathbb{R}$. Let us define $h(x) := \frac{\sin(x)}{x}$ for any $0 \leq x \leq \pi/2$. We can compute that $h'(x) = \frac{x\cos(x)-\sin(x)}{x^2}$. Let $p(x) := x\cos(x) - \sin(x)$. Then $p(0) = 0$ and $p'(x) = -x\sin(x) < 0$ for any $0 < x < \pi/2$ which implies that $p(x) < 0$ and thus $h'(x) < 0$ for any $0 < x < \pi/2$. Hence $h(x)$ is decreasing in $x$ for any $0 \leq x \leq \pi/2$. As a result, the map

$$\alpha \mapsto \frac{\sin\left(\pi\left(1-\frac{\alpha}{2}\right)\right)}{\pi\left(1-\frac{\alpha}{2}\right)} \quad \text{is increasing in } \alpha \text{ for any } 1 < \alpha < 2. \tag{27}$$

It is well known that gamma function $x \mapsto \Gamma(x)$ is log-convex for $x > 0$ and thus convex for any $x > 0$. Since $\Gamma(1) = \Gamma(2) = 1$, there exists a unique critical value $c_0 \in (1, 2)$ such that the gamma function $x \mapsto \Gamma(x)$ is increasing for any $x \geq c_0$ and decreasing for any $1 \leq x \leq c_0$.

Next, for any given $\alpha_0 \in (1, 2)$ such that $1 + \frac{\alpha_0}{2} \geq c_0$, we have for any $2 \geq \alpha_2 > \alpha_1 \geq \alpha_0$ and $d \geq 2$,

$$\begin{aligned}
&\frac{g(\alpha_2; d)}{g(\alpha_1; d)} \\
&= \frac{2^{\alpha_2 - 2}\alpha_2 \Gamma\left(\frac{d+\alpha_2}{2}\right)\Gamma\left(\frac{\alpha_2}{2}\right)\frac{\sin\left(\pi\left(1-\frac{\alpha_2}{2}\right)\right)}{\pi\left(1-\frac{\alpha_2}{2}\right)}\left(\sqrt{d} + \frac{2-\alpha_2}{\alpha_2 - 1}\right)}{2^{\alpha_1 - 2}\alpha_1 \Gamma\left(\frac{d+\alpha_1}{2}\right)\Gamma\left(\frac{\alpha_1}{2}\right)\frac{\sin\left(\pi\left(1-\frac{\alpha_1}{2}\right)\right)}{\pi\left(1-\frac{\alpha_1}{2}\right)}\left(\sqrt{d} + \frac{2-\alpha_1}{\alpha_1 - 1}\right)} \\
&= \frac{2^{\alpha_2}\Gamma\left(\frac{d+\alpha_2}{2}\right)\Gamma\left(1+\frac{\alpha_2}{2}\right)\frac{\sin\left(\pi\left(1-\frac{\alpha_2}{2}\right)\right)}{\pi\left(1-\frac{\alpha_2}{2}\right)}\left(\sqrt{d} + \frac{2-\alpha_2}{\alpha_2 - 1}\right)}{2^{\alpha_1}\Gamma\left(\frac{d+\alpha_1}{2}\right)\Gamma\left(1+\frac{\alpha_1}{2}\right)\frac{\sin\left(\pi\left(1-\frac{\alpha_1}{2}\right)\right)}{\pi\left(1-\frac{\alpha_1}{2}\right)}\left(\sqrt{d} + \frac{2-\alpha_1}{\alpha_1 - 1}\right)} \\
&\geq 2^{\alpha_2 - \alpha_1}\frac{\sqrt{d} + \frac{2-\alpha_2}{\alpha_2 - 1}}{\sqrt{d} + \frac{2-\alpha_1}{\alpha_1 - 1}}, \qquad (28)
\end{aligned}$$

where we used (27) and the fact that the gamma function $x \mapsto \Gamma(x)$ is increasing in $x \geq 1 + \frac{\alpha_0}{2} \geq c_0$.

Next, let us define the function:

$$q(x) := 2^{x - \alpha_1}\frac{\sqrt{d} + \frac{2-x}{x-1}}{\sqrt{d} + \frac{2-\alpha_1}{\alpha_1 - 1}}, \qquad (29)$$

where $2 \geq x \geq \alpha_1 \geq \alpha_0$. It is clear that $q(\alpha_1) = 1$ and moreover, we can compute that

$$\begin{aligned}
q'(x) &= \log(2)2^{x-\alpha_1}\frac{\sqrt{d} + \frac{2-x}{x-1}}{\sqrt{d} + \frac{2-\alpha_1}{\alpha_1 - 1}} - \frac{2^{x-\alpha_1}}{(x-1)^2}\frac{1}{\sqrt{d} + \frac{2-\alpha_1}{\alpha_1 - 1}} \\
&= \frac{2^{x-\alpha_1}}{\sqrt{d} + \frac{2-\alpha_1}{\alpha_1 - 1}}\left(\log(2)\left(\sqrt{d} + \frac{2-x}{x-1}\right) - \frac{1}{(x-1)^2}\right) \\
&\geq \frac{2^{x-\alpha_1}}{\sqrt{d} + \frac{2-\alpha_1}{\alpha_1 - 1}}\left(\log(2)\sqrt{d} - \frac{1}{(\alpha_0 - 1)^2}\right) \geq 0, \qquad (30)
\end{aligned}$$

provided that

$$d \geq \frac{1}{(\log 2)^2(\alpha_0 - 1)^4}. \qquad (31)$$

This implies that $q(x)$ is increasing for $2 \geq x \geq \alpha_1 \geq \alpha_0$ provided that $d \geq 2$, $1 + \frac{\alpha_0}{2} \geq c_0$ and (31) holds. Hence, we conclude that $g(\alpha_2; d) \geq g(\alpha_1; d)$ for any $d \geq d_0 = \max\left(2, \frac{1}{(\log 2)^2(\alpha_0 - 1)^4}\right)$, and $2 \geq \alpha_2 \geq \alpha_1 \geq \alpha_0$.

Due to the space constraint, the proof of part (ii) will be provided in the Supplementary Document. The proof is complete. □

This result reveals an interesting fact. Depending on the dimension of the parameter vector $d$, the Wasserstein stability bound (Theorem 3.3) and the generalization bound (Corollary 3.4) exhibit different behaviors with respect to varying $\alpha$. We observe that for sufficiently large $d$, there exists a critical value $\alpha_0$ such that the bound is monotonically increasing for $\alpha \geq \alpha_0$. This suggests that for $d$ large enough, increasing the heaviness of the tails (i.e., smaller $\alpha$) can be beneficial unless $\alpha$ is smaller than $\alpha_0$.

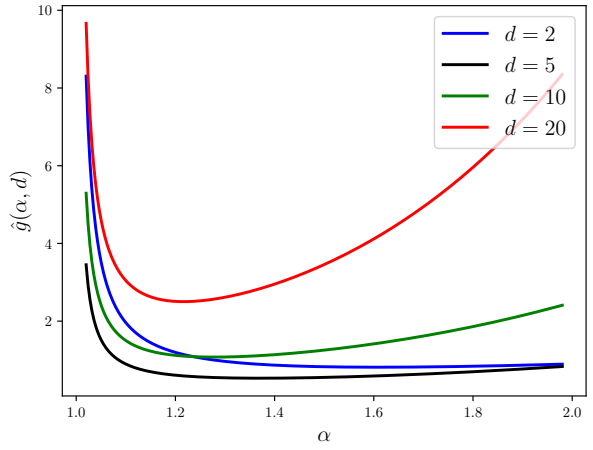


*Figure 1.* Behavior of $g(\alpha; d)$ with respect to $\alpha$. We scale $g(\alpha; d)$ appropriately to fit all the plots in the same frame which we denote as $\hat{g}(\alpha; d)$.

For visualization, we also provide a pictorial illustration of the function $g(\alpha; d)$ in Figure 1. The figure shows the behavior of $g(\alpha; d)$ with respect to $\alpha$ for various dimensions $d$. The observed non-globally monotonic behavior for large $d$ indicates that the conclusions of (Raj et al., 2023) extend beyond quadratic loss functions.

On the other hand, Barsbey et al. (2021) and Raj et al. (2023) reported several experimental results conducted on neural networks, which illustrated the existence of a non-globally monotonic relation between the generalization error and the heaviness of the tails in practical settings (see Figure 7 in (Barsbey et al., 2021) and Figure 2 in (Raj et al., 2023)). Our result brings a stronger theoretical justification to these empirical observations thanks to the generality of our theoretical framework.

### 3.3. Infeasibility of $p$-Wasserstein Distance for $p \geq \alpha$

Now, that we have provided result for the 1-Wasserstein distance between the distribution of $\theta$ and $\hat{\theta}$. A natural question to ask is whether similar results could be obtained more generally in the $p$-Wasserstein distance for some arbitrary $p$.

Not surprisingly, the following result says that in general we do not expect to control the $p$-Wasserstein distance when $p$

is larger than the tail-index $\alpha$.

**Proposition 3.7.** *Let $d = 1$, $\alpha > 1$, $\mathcal{X} \subset \mathbb{R}$, and $f(\theta, x) = (\theta x)^2$. Denote $\mu$ and $\nu$ as the invariant measures of (11) and (12), respectively. Then for any $p > \alpha$, $\mathcal{W}_p(\mu, \nu) = +\infty$.*

*Proof.* Due to our choice of the loss function $f$, the SDEs (11) and (12) reduce to Ornstein-Uhlenbeck processes driven by a symmetric $\alpha$-stable Lévy process. Hence, we can characterize the invariant distributions of the SDEs as follows (see e.g. (Raj et al., 2023)):

$$\theta_\infty =^{\mathrm{d}} \mu + \sigma\xi, \qquad \text{and} \qquad \hat{\theta}_\infty =^{\mathrm{d}} \hat{\mu} + \hat{\sigma}\hat{\xi}, \tag{32}$$

for some $\mu, \hat{\mu} \in \mathbb{R}$ and $\sigma, \hat{\sigma} \in \mathbb{R}_+$. Here, $\xi$ and $\hat{\xi}$ are $\mathcal{S}\alpha\mathcal{S}(1)$ distributed (see Section 2 for definition) and $=^{\mathrm{d}}$ denotes equality in distribution. Now recall that $\mu = \mathrm{Law}(\theta_\infty)$ and $\nu = \mathrm{Law}(\hat{\theta}_\infty)$, and the $p$-Wasserstein metric for one-dimensional distributions is given by,

$$\mathcal{W}_p^p(\mu, \nu) = \inf_{\gamma \in \Gamma(\mu,\nu)} \mathbb{E}_{(x,y)\sim\gamma(x,y)} |x - y|^p,$$

where $\Gamma(\mu, \nu)$ is the set of all couplings of $\mu$ and $\nu$. In our case, $x \in \mathbb{R}$ and $y \in \mathbb{R}$. For any coupling $\gamma^\star \in \Gamma(\mu, \nu)$, we have

$$\begin{aligned}
&\int_{\mathbb{R}\times\mathbb{R}} |x - y|^p \, d\gamma^\star(x, y) \\
&= \int_{\mathbb{R}\times\mathbb{R}} \left[|x - y|^2\right]^{p/2} d\gamma^\star(x, y) \\
&= \int_{\mathbb{R}\times\mathbb{R}} (x^2 + y^2 - 2xy)^{p/2} \, d\gamma^\star(x, y) \\
&\geq \int_{\mathbb{R}_+\times\mathbb{R}_-} (x^2 + y^2 - 2xy)^{p/2} \, d\gamma^\star(x, y) \\
&\geq \int_{\mathbb{R}_+\times\mathbb{R}_-} (|x|^p + |y|^p + |2xy|^{p/2}) \, d\gamma^\star(x, y) \\
&\geq \int_{\mathbb{R}_+\times\mathbb{R}_-} |x|^p \, d\gamma^\star(x, y) + \int_{\mathbb{R}_+\times\mathbb{R}_-} |y|^p \, d\gamma^\star(x, y) \\
&= C_1 \int_{\mathbb{R}_+} |x|^p \, d\mu(x) + C_2 \int_{\mathbb{R}_-} |y|^p \, d\nu(y) = +\infty,
\end{aligned}$$

where $C_1$ and $C_2$ are some finite, positive constants. The last equation comes from the properties of the $\alpha$-stable distribution. Since it holds for any $\gamma^\star \in \Gamma(\mu, \nu)$, we conclude that $\mathcal{W}_p^p(\mu, \nu) = \infty$. This completes the proof. □

### 3.4. Discrete-Time Dynamics

Finally, we will illustrate that our theory also extends to the discretizations of the SDEs (11) and (12). Consider the following Euler-Maruyama discretization:

$$\theta_{k+1} = \theta_k - \eta\nabla\hat{F}(\theta_k, X_n) + \eta^{1/\alpha} S_{k+1}, \tag{33}$$

where $S_k$ are i.i.d. rotationally invariant alpha-stable random vectors with the characteristic function:

$$\mathbb{E}\left[e^{\mathrm{i}\langle u, S_k\rangle}\right] = e^{-\|u\|_2^\alpha}, \qquad \text{for any } u \in \mathbb{R}^d. \tag{34}$$

Similarly, with input data $\hat{X}_n$, we have

$$\hat{\theta}_{k+1} = \hat{\theta}_k - \eta\nabla\hat{F}(\hat{\theta}_k, \hat{X}_n) + \eta^{1/\alpha} S_{k+1}. \tag{35}$$

Let $\mu$ and $\hat{\mu}$ denote the stationary distributions of continuous-time $\theta_t$ and $\hat{\theta}_t$ as $t \to \infty$. Moreover, let $\nu$ and $\hat{\nu}$ denote the stationary distributions of discrete-time $\theta_k$ and $\hat{\theta}_k$ as $k \to \infty$. It is proved in (Chen et al., 2022) that the 1-Wasserstein distance of the discretization error is of order $\eta^{2/\alpha-1}$. More precisely, they showed the following result.

**Lemma 3.8** (Theorem 1.2. in Chen et al. (2022))**.** *Suppose that Assumptions 3.1 and 3.2 hold. Let $m$, $L$ be as in Assumption 3.2. Then, there exists some constant $Q$ (that may depend on $B, m, K, L, M$ from Assumption 3.2) such that for every $\eta < \min\{1, m/L^2, 1/m\}$, one has*

$$\mathcal{W}_1(\mu, \nu) \leq Q\eta^{2/\alpha-1}, \tag{36}$$

$$\mathcal{W}_1(\hat{\mu}, \hat{\nu}) \leq Q\eta^{2/\alpha-1}. \tag{37}$$

By applying the above 1-Wasserstein bound for the discretization error in Lemma 3.8 and Theorem 3.3, we obtain the following corollary, which provides the 1-Wasserstein distance of the stationary distributions of the discrete-time $(\theta_k)_{k\geq 0}$ and $(\hat{\theta}_k)_{k\geq 0}$ processes.

**Corollary 3.9.** *Under the assumptions in Theorem 3.3 and Lemma 3.8, we have*

$$\begin{aligned}
\mathcal{W}_1(\nu, \hat{\nu}) &\leq 2Q\eta^{2/\alpha-1} \\
&+ \left(C_1\lambda^{-1}e^\lambda + 1\right) e^L \rho(X_n, \hat{X}_n) K_2 \left(2C_0 + 1\right).
\end{aligned} \tag{38}$$

By using the same approach as we used in Corollary 3.4, we can easily obtain a generalization bound for the discrete-time as well. Note that the upper bound in Corollary 3.9 depends on the tail-index $\alpha$ only via $\eta^{2/\alpha-1}$, which is increasing in $\alpha$ (since $\eta < 1$ as assumed in Lemma 3.8), and the constant $C_0$ which depends on $\alpha$ via $g(\alpha; d)$ function. Therefore, by Proposition 3.6, the upper bound in Corollary 3.9 is increasing in $\alpha \in [\alpha_0, 2]$, for any $d \geq d_0$, where $d_0$ and $\alpha_0$ are given in Proposition 3.6. Moreover, the proof of Proposition 3.6 reveals that $\frac{\partial}{\partial\alpha} g(\alpha; d)$ tends to $-\infty$ as $\alpha$ tends to 1 and thus there exists some $\alpha_0'' < \alpha_0'$, where $\alpha_0'$ is defined in Proposition 3.6, such that for any fixed $d \in \mathbb{N}$, the upper bound in Corollary 3.9 is decreasing in $\alpha \in [1, \alpha_0'']$. Hence, the conclusions that we obtained for the continuous-time processes remain valid for the discretizations as well.

## 4. Conclusion

In this work, we studied the relation between the generalization behavior and the heavy tails arising in the SGD

dynamics. Previous work on the topic obtained monotonic relationship under strong topological and statistical regularity assumptions, with the exception of the approach in (Raj et al., 2023) which was limited to only quadratic losses. Following the literature, we considered heavy-tailed SDEs and their discretization for modeling the heavy-tailed behavior of SGD, and showed that the relation is non-monotonic for a general class of losses satisfying a dissipativity condition which generalizes the results of (Raj et al., 2023) beyond quadratic losses. Our proof technique is based on a novel 1-Wasserstein stability bound for the symmetric $\alpha$-stable Lévy-driven SDEs, that model the SGD dynamics. Furthermore, our results, when combined with the results of (Raginsky et al., 2016), yield directly a generalization bound for the class of Lipschitz functions.

**Future Directions:** As a future research direction, we would like to obtain similar stability bounds without making the dissipativity assumption on the objective function as being done for Langevin Monte Carlo in (Kinoshita & Suzuki, 2022). We would also like to consider specific class of functions (e.g. one-layer neural network) and study the effect of tail-index with other parameters and its effect on the generalization.

## Acknowledgments

Anant Raj is supported by the a Marie Sklodowska-Curie Fellowship (project NN-OVEROPT 101030817). Lingjiong Zhu is partially supported by the grants NSF DMS-2053454, NSF DMS-2208303, and a Simons Foundation Collaboration Grant. Mert Gürbüzbalaban'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. Umut Şimşekli'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.

## References

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

Barsbey, M., Sefidgaran, M., Erdogdu, M. A., Richard, G., and Şimşekli, U. Heavy Tails in SGD and Compressibility of Overparametrized Neural Networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29364–29378. Curran Associates, Inc., 2021.

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

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

Cao, Y. and Gu, Q. Generalization bounds of stochastic gradient descent for wide and deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Chen, P., Deng, C., Schilling, R., and Xu, L. Approximation of the invariant measure of stable SDEs by an Euler–Maruyama scheme. *arXiv preprint arXiv:2205.01342*, 2022.

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

Erdogdu, M. A., Mackey, L., and Shamir, O. Global nonconvex optimization with discretized diffusions. In *Advances in Neural Information Processing Systems*, 2018.

Farghly, T. and Rebeschini, P. Time-independent generalization bounds for SGLD in non-convex settings. In *Advances in Neural Information Processing Systems*, volume 34, pp. 19836–19846, 2021.

Gao, X., Gürbüzbalaban, M., and Zhu, L. Global convergence of stochastic gradient Hamiltonian Monte Carlo for nonconvex stochastic optimization: Nonasymptotic performance bounds and momentum-based acceleration. *Operations Research*, 70(5):2931–2947, 2022.

Gürbüzbalaban, M., Şimşekli, U., and Zhu, L. The heavy-tail phenomenon in SGD. In *International Conference on Machine Learning*, pp. 3964–3975. PMLR, 2021.

Hardt, M., Recht, B., and Singer, Y. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pp. 1225–1234. PMLR, 2016.

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

Hodgkinson, L., Şimşekli, U., Khanna, R., and Mahoney, M. W. Generalization properties of stochastic optimizers via trajectory analysis. *arXiv preprint arXiv:2108.00781*, 2021.

Kinoshita, Y. and Suzuki, T. Improved convergence rate of stochastic gradient Langevin dynamics with variance reduction and its application to optimization. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

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

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

Lim, S. H., Wan, Y., and Simsekli, U. Chaotic regularization and heavy-tailed limits for deterministic gradient descent. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Neu, G., Dziugaite, G. K., Haghifam, M., and Roy, D. M. Information-theoretic generalization bounds for stochastic gradient descent. In *Conference on Learning Theory*, pp. 3526–3545. PMLR, 2021.

Nguyen, T. H., Simsekli, U., Gurbuzbalaban, M., and Richard, G. First exit time analysis of stochastic gradient descent under heavy-tailed gradient noise. In *Advances in Neural Information Processing Systems*, pp. 273–283, 2019.

Park, S., Simsekli, U., and Erdogdu, M. A. Generalization bounds for stochastic gradient descent via localized $\varepsilon$-covers. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Raginsky, M., Rakhlin, A., Tsao, M., Wu, Y., and Xu, A. Information-theoretic analysis of stability and bias of learning algorithms. In *2016 IEEE Information Theory Workshop (ITW)*, pp. 26–30. IEEE, 2016.

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

Raj, A., Barsbey, M., Gürbüzbalaban, M., Zhu, L., and Şimşekli, U. Algorithmic stability of heavy-tailed stochastic gradient descent on least squares. In *International Conference on Algorithmic Learning Theory*. PMLR, 2023.

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

Shalev-Shwartz, S. and Ben-David, S. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

Şimşekli, U., Sener, O., Deligiannidis, G., and Erdogdu, M. A. Hausdorff dimension, heavy tails, and generalization in neural networks. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 5138–5151. Curran Associates, Inc., 2020.

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

Wang, H., Gürbüzbalaban, M., Zhu, L., Şimşekli, U., and Erdogdu, M. A. Convergence rates of stochastic gradient descent under infinite noise variance. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Wang, J. $L^p$-Wasserstein distance for stochastic differential equations driven by Lévy processes. *Bernoulli*, 22:1598–1616, 2016.

Xu, A. and Raginsky, M. Information-theoretic analysis of generalization capability of learning algorithms. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Zhang, J., Karimireddy, S. P., Veit, A., Kim, S., Reddi, S., Kumar, S., and Sra, S. Why are adaptive methods good for attention models? In *Advances in Neural Information Processing Systems*, volume 33, pp. 15383–15393, 2020.

# Algorithmic stability of heavy-tailed SGD with general loss functions

**Supplementary Document**

## A. Background on Markov Semigroups

In this section, we introduce the concept of Markov semigroups, that will be used in the proofs of main results in Section B.

For a continuous-time Markov process $(X_t^w)_{t\geq 0}$ that starts at $X_0 = w$, its Markov semigroup $P_t$ is defined as for any bounded measurable function $f : \mathbb{R}^d \to \mathbb{R}$,

$$P_t f(w) = \mathbb{E} f(X_t^w), \qquad t \geq 0. \tag{39}$$

Similarly, for a discrete-time Markov process $(Y_k^w)_{k=0}^{\infty}$ that starts at $Y_0 = w$, its Markov semigroup $Q_k$ is defined as for any bounded measurable function $f : \mathbb{R}^d \to \mathbb{R}$,

$$Q_k f(w) = \mathbb{E} f(Y_k^w), \qquad k = 0, 1, 2, \ldots. \tag{40}$$

## B. Proofs of Main Results

In this section, we provide the proofs of main results in our paper.

### B.1. Proof of Theorem 3.3

We first provide the theorem statement with all the details.

**Theorem B.1** (Restatement of Theorem 3.3)**.** *Assume $\theta_0 = \hat{\theta}_0 = w$. Denote by $\mu, \hat{\mu}$ the unique invariant distributions of $(\theta_t)_{t\geq 0}$ and $(\hat{\theta}_t)_{t\geq 0}$ respectively. The following two statements hold:*

*(i) For every $N \geq 1$ and $\eta \in (0, 1)$, we have the following statements:*

*(I) If $N = 1$, then*

$$\begin{aligned}
&\mathcal{W}_1\left(Law(\theta_{\eta N}), Law(\hat{\theta}_{\eta N})\right) \\
&\leq \left(K_1 + \rho(X_n, \hat{X}_n)K_2\right)(2C) \cdot \left[(1 + \|w\|)\eta^{1+\frac{1}{\alpha}} + \rho(X_n, \hat{X}_n)K_2(2\|w\| + 1)\eta\right].
\end{aligned} \tag{41}$$

*(II) If $2 \leq N \leq \eta^{-1} + 1$, then*

$$\begin{aligned}
&\mathcal{W}_1\left(Law(\theta_{\eta N}), Law(\hat{\theta}_{\eta N})\right) \\
&\leq \left(K_1 + \rho(X_n, \hat{X}_n)K_2\right)(2C)(1 + C_0(1 + \|w\|))\eta^{1+\frac{1}{\alpha}} + \rho(X_n, \hat{X}_n)K_2(2C_0(1 + \|w\|) + 1)\eta \\
&\quad + e^L\left(K_1 + \rho(X_n, \hat{X}_n)K_2\right)(2C)\left(1 + C_0(1 + \|w\|)\right)\eta^{\frac{1}{\alpha}} \\
&\quad + e^L\rho(X_n, \hat{X}_n)K_2(2C_0(1 + \|w\|) + 1).
\end{aligned} \tag{42}$$

*(III) If $N > \eta^{-1} + 1$, then*

$$\begin{aligned}
&\mathcal{W}_1\left(Law(\theta_{\eta N}), Law(\hat{\theta}_{\eta N})\right) \\
&\leq \left(K_1 + \rho(X_n, \hat{X}_n)K_2\right)(2C)\left(1 + C_0(1 + \|w\|)\right)\eta^{1+\frac{1}{\alpha}} + \rho(X_n, \hat{X}_n)K_2(2C_0(1 + \|w\|) + 1)\eta \\
&\quad + \left(C_1\lambda^{-1}e^{\lambda} + 1\right)e^L\left(K_1 + \rho(X_n, \hat{X}_n)K_2\right)(2C)\left(1 + C_0(1 + \|w\|)\right)\eta^{\frac{1}{\alpha}} \\
&\quad + \left(C_1\lambda^{-1}e^{\lambda} + 1\right)e^L\rho(X_n, \hat{X}_n)K_2(2C_0(1 + \|w\|) + 1).
\end{aligned} \tag{43}$$

*(ii) We have*

$$\mathcal{W}_1(\mu, \hat{\mu}) \leq \left(C_1\lambda^{-1}e^{\lambda} + 1\right)e^L\rho(X_n, \hat{X}_n)K_2\left(2C_0 + 1\right). \tag{44}$$

*Here, $K_1, K_2$ and $L$ are defined in Assumption 3.1 and Assumption 3.2 and $C_0, C_1$ and $\lambda$ are some positive real constants.*

*Proof of Theorem 3.3.* (i) We first prove part (i). For any $h \in \text{Lip}(1)$, by the semigroup property, we have

$$\begin{aligned} P_{N\eta}h(w) - \hat{P}_{N\eta}h(w) &= \sum_{i=1}^{N} \left( \hat{P}_{(i-1)\eta} P_{(N-i+1)\eta} h(w) - \hat{P}_{i\eta} P_{(N-i)\eta} h(w) \right) \\ &= \sum_{i=1}^{N} \hat{P}_{(i-1)\eta} \left( P_\eta - \hat{P}_\eta \right) P_{(N-i)\eta} h(w). \end{aligned}$$

Therefore, we can compute that

$$\begin{aligned} &\mathcal{W}_1 \left( \text{Law}\left(\theta_{\eta N}\right), \text{Law}\left(\hat{\theta}_{\eta N}\right) \right) \\ &= \sup_{h \in \text{Lip}(1)} \left| P_{N\eta}h(w) - \hat{P}_{N\eta}h(w) \right| \\ &\le \sup_{h \in \text{Lip}(1)} \left| \hat{P}_{(N-1)\eta} \left( P_\eta - \hat{P}_\eta \right) h(w) \right| + \sum_{i=1}^{N-1} \sup_{h \in \text{Lip}(1)} \left| \hat{P}_{(i-1)\eta} \left( P_\eta - \hat{P}_\eta \right) P_{(N-i)\eta} h(w) \right|. \end{aligned} \tag{45}$$

Let us first bound the first term in (45). For any $h \in \text{Lip}(1)$ and $\eta < 1$, by applying Lemma C.5, we get

$$\left| \left( P_\eta - \hat{P}_\eta \right) h(w) \right| \le \left( K_1 + \rho(X_n, \hat{X}_n) K_2 \right) (2C) \left(1 + \|w\|\right) \eta^{1+\frac{1}{\alpha}} + \rho(X_n, \hat{X}_n) K_2 (2\|w\| + 1)\eta.$$

Hence, we have

$$\begin{aligned} &\sup_{h \in \text{Lip}(1)} \left| \hat{P}_{(N-1)\eta} \left( P_\eta - \hat{P}_\eta \right) h(w) \right| \\ &\le \left( K_1 + \rho(X_n, \hat{X}_n) K_2 \right) (2C) \left(1 + \mathbb{E}\|\hat{\theta}^w_{(N-1)\eta}\|\right) \eta^{1+\frac{1}{\alpha}} + \rho(X_n, \hat{X}_n) K_2 \left( 2\mathbb{E}\|\hat{\theta}^w_{(N-1)\eta}\| + 1 \right) \eta \\ &\le \left( K_1 + \rho(X_n, \hat{X}_n) K_2 \right) (2C) \left(1 + C_0(1 + \|w\|)\right) \eta^{1+\frac{1}{\alpha}} + \rho(X_n, \hat{X}_n) K_2 (2C_0(1 + \|w\|) + 1)\eta, \end{aligned} \tag{46}$$

where we applied Lemma C.1 to obtain the last inequality above.

Next, let us bound the second term in (45) and hence bound the 1-Wasserstein distance $\mathcal{W}_1 \left( \text{Law}(\theta_{\eta N}), \text{Law}(\hat{\theta}_{\eta N}) \right)$.

We consider three cases: (I) $N = 1$; (II) $2 \le N \le \eta^{-1} + 1$ and (III) $N > \eta^{-1} + 1$.

Case (I): $N = 1$. One can apply (46) and obtain

$$\begin{aligned} &\mathcal{W}_1 \left( \text{Law}(\theta_\eta), \text{Law}(\hat{\theta}_\eta) \right) \\ &\le \left( K_1 + \rho(X_n, \hat{X}_n) K_2 \right) (2C) \left(1 + \mathbb{E}\|\hat{\theta}^w_0\|\right) \eta^{1+\frac{1}{\alpha}} + \rho(X_n, \hat{X}_n) K_2 \left( 2\mathbb{E}\|\hat{\theta}^w_0\| + 1 \right) \eta \\ &= \left( K_1 + \rho(X_n, \hat{X}_n) K_2 \right) (2C) \left(1 + \|w\|\right) \eta^{1+\frac{1}{\alpha}} + \rho(X_n, \hat{X}_n) K_2 (2\|w\| + 1)\eta. \end{aligned} \tag{47}$$

This completes the proof of part (I).

Case (II): $2 \le N \le \eta^{-1} + 1$. By Lemma C.5, for any $i \ge 1$, we have

$$\begin{aligned} &\left| \left( P_\eta - \hat{P}_\eta \right) P_{(N-i)\eta} h(w) \right| \\ &\le \|\nabla P_{(N-i)\eta} h\|_\infty \left[ \left( K_1 + \rho(X_n, \hat{X}_n) K_2 \right) (2C) \left(1 + \|w\|\right) \eta^{1+\frac{1}{\alpha}} + \rho(X_n, \hat{X}_n) K_2 (2\|w\| + 1)\eta \right] \\ &\le e^L \left[ \left( K_1 + \rho(X_n, \hat{X}_n) K_2 \right) (2C) \left(1 + \|w\|\right) \eta^{1+\frac{1}{\alpha}} + \rho(X_n, \hat{X}_n) K_2 (2\|w\| + 1)\eta \right], \end{aligned} \tag{48}$$

where we used Lemma C.3 and the fact that for any $i \ge 1$ and $2 \le N \le \eta^{-1} + 1$ we have $(N - i)\eta \le 1$ in the inequality (48).

By applying Lemma C.1, we obtain

$$\begin{aligned}
&\sup_{h\in\text{Lip}(1)}\left|\tilde{P}_{(i-1)\eta}\left(P_\eta-\hat{P}_\eta\right)P_{(N-i)\eta}h(w)\right| \\
&\le e^L\left[\left(K_1+\rho(X_n,\hat{X}_n)K_2\right)(2C)\left(1+\mathbb{E}\|\hat{\theta}_{(i-1)\eta}\|\right)\eta^{1+\frac{1}{\alpha}}+\rho(X_n,\hat{X}_n)K_2\left(2\mathbb{E}\|\hat{\theta}_{(i-1)\eta}\|+1\right)\eta\right] \\
&\le e^L\left(K_1+\rho(X_n,\hat{X}_n)K_2\right)(2C)\left(1+C_0(1+\|w\|)\right)\eta^{1+\frac{1}{\alpha}} \\
&\qquad\qquad +e^L\rho(X_n,\hat{X}_n)K_2(2C_0(1+\|w\|)+1)\eta. \qquad (49)
\end{aligned}$$

Hence, we conclude that

$$\begin{aligned}
&\mathcal{W}_1\left(\text{Law}(\theta_{\eta N}),\text{Law}(\hat{\theta}_{\eta N})\right) \\
&\le \sup_{h\in\text{Lip}(1)}\left|\hat{P}_{(N-1)\eta}\left(P_\eta-\hat{P}_\eta\right)h(w)\right|+\sum_{i=1}^{N-1}\sup_{h\in\text{Lip}(1)}\left|\hat{P}_{(i-1)\eta}\left(P_\eta-\hat{P}_\eta\right)P_{(N-i)\eta}h(w)\right| \\
&\le \left(K_1+\rho(X_n,\hat{X}_n)K_2\right)(2C)\left(1+C_0(1+\|w\|)\right)\eta^{1+\frac{1}{\alpha}}+\rho(X_n,\hat{X}_n)K_2\left(2C_0(1+\|w\|)+1\right)\eta \\
&\quad +(N-1)e^L\left(K_1+\rho(X_n,\hat{X}_n)K_2\right)(2C)\left(1+C_0(1+\|w\|)\right)\eta^{1+\frac{1}{\alpha}} \\
&\qquad\qquad +(N-1)e^L\rho(X_n,\hat{X}_n)K_2\left(2C_0(1+\|w\|)+1\right)\eta \\
&\le \left(K_1+\rho(X_n,\hat{X}_n)K_2\right)(2C)\left(1+C_0(1+\|w\|)\right)\eta^{1+\frac{1}{\alpha}}+\rho(X_n,\hat{X}_n)K_2\left(2C_0(1+\|w\|)+1\right)\eta \\
&\quad +e^L\left(K_1+\rho(X_n,\hat{X}_n)K_2\right)(2C)\left(1+C_0(1+\|w\|)\right)\eta^{\frac{1}{\alpha}} \\
&\qquad\qquad +e^L\rho(X_n,\hat{X}_n)K_2\left(2C_0(1+\|w\|)+1\right). \qquad (50)
\end{aligned}$$

This completes the proof of (I).

Case (III): $N>\eta^{-1}+1$. We can compute that

$$\begin{aligned}
&\sup_{h\in\text{Lip}(1)}\left|\hat{P}_{(i-1)\eta}\left(P_\eta-\hat{P}_\eta\right)P_{(N-i)\eta}h(w)\right| \\
&=\sup_{h\in\text{Lip}(1)}\left|\hat{P}_{(i-1)\eta}\left(P_\eta-\hat{P}_\eta\right)P_1P_{(N-i)\eta-1}h(w)\right| \\
&\le \sup_{g\in\text{Lip}(1)}\left|\hat{P}_{(i-1)\eta}\left(P_\eta-\hat{P}_\eta\right)P_1g(w)\right|\sup_{h\in\text{Lip}(1)}\|\nabla P_{(N-i)\eta-1}h\|_\infty.
\end{aligned}$$

By Lemma C.2, for any $h\in\text{Lip}(1)$,

$$|P_th(w)-P_th(y)|\le C_1e^{-\lambda t}\|w-y\|, \qquad (51)$$

for any $t\ge 0$ and $w,y\in\mathbb{R}^d$. This implies that for any $h\in\text{Lip}(1)$,

$$\|\nabla P_th\|_\infty\le C_1e^{-\lambda t}. \qquad (52)$$

Hence, we conclude that

$$\sup_{h\in\text{Lip}(1)}\left|\hat{P}_{(i-1)\eta}\left(P_\eta-\hat{P}_\eta\right)P_{(N-i)\eta}h(w)\right|\le C_1e^{-\lambda((N-i)\eta-1)}\sup_{g\in\text{Lip}(1)}\left|\hat{P}_{(i-1)\eta}\left(P_\eta-\hat{P}_\eta\right)P_1g(w)\right|,$$

where $i\le\lfloor N-\eta^{-1}\rfloor$.

Moreover, by Lemma C.5, we have

$$\begin{aligned}
&\left|P_\eta P_1g(w)-\hat{P}_\eta P_1g(w)\right| \\
&\le\|\nabla P_1g\|_\infty\left[\left(K_1+\rho(X_n,\hat{X}_n)K_2\right)(2C)\left(1+\|w\|\right)\eta^{1+\frac{1}{\alpha}}+\rho(X_n,\hat{X}_n)K_2(2\|w\|+1)\eta\right] \\
&\le e^L\left[\left(K_1+\rho(X_n,\hat{X}_n)K_2\right)(2C)\left(1+\|w\|\right)\eta^{1+\frac{1}{\alpha}}+\rho(X_n,\hat{X}_n)K_2(2\|w\|+1)\eta\right], \qquad (53)
\end{aligned}$$

which by Lemma C.1 implies that

$$\begin{aligned}
&\sup_{g\in\text{Lip}(1)}\left|\hat{P}_{(i-1)\eta}(P_\eta-\hat{P}_\eta)P_1 g(w)\right| \\
&\le e^L\left[\left(K_1+\rho(X_n,\hat{X}_n)K_2\right)(2C)\left(1+\mathbb{E}\|\theta^w_{(i-1)\eta}\|\right)\eta^{1+\frac{1}{\alpha}}+\rho(X_n,\hat{X}_n)K_2\left(2\mathbb{E}\|\theta^w_{(i-1)\eta}\|+1\right)\eta\right] \\
&\le e^L\left[\left(K_1+\rho(X_n,\hat{X}_n)K_2\right)(2C)\left(1+C_0(1+\|w\|)\right)\eta^{1+\frac{1}{\alpha}}+\rho(X_n,\hat{X}_n)K_2\left(2C_0(1+\|w\|)+1\right)\eta\right].
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
&\sum_{i=1}^{\lfloor N-\eta^{-1}\rfloor}\sup_{h\in\text{Lip}(1)}\left|\hat{P}_{(i-1)\eta}\left(P_\eta-\hat{P}_\eta\right)P_{(N-i)\eta}h(w)\right| \\
&\le\sum_{i=1}^{\lfloor N-\eta^{-1}\rfloor}C_1e^{-\lambda((N-i)\eta-1)}e^L\left(K_1+\rho(X_n,\hat{X}_n)K_2\right)(2C)\left(1+C_0(1+\|w\|)\right)\eta^{1+\frac{1}{\alpha}} \\
&\quad+\sum_{i=1}^{\lfloor N-\eta^{-1}\rfloor}C_1e^{-\lambda((N-i)\eta-1)}e^L\rho(X_n,\hat{X}_n)K_2\left(2C_0(1+\|w\|)+1\right)\eta \\
&\le C_1\lambda^{-1}e^\lambda e^L\left(K_1+\rho(X_n,\hat{X}_n)K_2\right)(2C)\left(1+C_0(1+\|w\|)\right)\eta^{\frac{1}{\alpha}} \\
&\quad+C_1\lambda^{-1}e^\lambda e^L\rho(X_n,\hat{X}_n)K_2\left(2C_0(1+\|w\|)+1\right),
\end{aligned}$$

where we used the fact that

$$\sum_{i=1}^{\lfloor N-\eta^{-1}\rfloor}e^{-\lambda((N-i)\eta-1)}\le e^\lambda\int_{\lfloor\eta^{-1}\rfloor-1}^{N-1}e^{-\lambda\eta r}dr\le e^\lambda\eta^{-1}\int_0^\infty e^{-\lambda r}dr=\lambda e^\lambda\eta^{-1}.$$

Next, when $i\ge\lfloor N-\eta^{-1}\rfloor+1$, by applying (49), we have

$$\begin{aligned}
&\sum_{i=\lfloor N-\eta^{-1}\rfloor+1}^{N-1}\sup_{h\in\text{Lip}(1)}\left|\hat{P}_{(i-1)\eta}\left(P_\eta-\hat{P}_\eta\right)P_{(N-i)\eta}h(w)\right| \\
&\le\sum_{i=\lfloor N-\eta^{-1}\rfloor+1}^{N-1}e^L\left(K_1+\rho(X_n,\hat{X}_n)K_2\right)(2C)\left(1+C_0(1+\|w\|)\right)\eta^{1+\frac{1}{\alpha}} \\
&\qquad+\sum_{i=\lfloor N-\eta^{-1}\rfloor+1}^{N-1}e^L\rho(X_n,\hat{X}_n)K_2\left(2C_0(1+\|w\|)+1\right)\eta \\
&\le e^L\left(K_1+\rho(X_n,\hat{X}_n)K_2\right)(2C)\left(1+C_0(1+\|w\|)\right)\eta^{\frac{1}{\alpha}} \\
&\qquad+\sum_{i=\lfloor N-\eta^{-1}\rfloor+1}^{N-1}e^L\rho(X_n,\hat{X}_n)K_2\left(2C_0(1+\|w\|)+1\right).
\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}
&\sum_{i=1}^{N-1}\sup_{h\in\text{Lip}(1)}\left|\hat{P}_{(i-1)\eta}\left(P_\eta-\hat{P}_\eta\right)P_{(N-i)\eta}h(w)\right| \\
&\le\left(C_1\lambda^{-1}e^\lambda+1\right)e^L\left(K_1+\rho(X_n,\hat{X}_n)K_2\right)(2C)\left(1+C_0(1+\|w\|)\right)\eta^{\frac{1}{\alpha}} \\
&\quad+\left(C_1\lambda^{-1}e^\lambda+1\right)e^L\rho(X_n,\hat{X}_n)K_2\left(2C_0(1+\|w\|)+1\right).
\end{aligned}$$

Hence, we conclude that

$$\begin{aligned}
&\mathcal{W}_1\left(\text{Law}\left(\theta_{\eta N}\right), \text{Law}\left(\hat{\theta}_{\eta N}\right)\right) \\
&\leq \sup_{h\in \text{Lip}(1)} \left|\hat{P}_{(N-1)\eta}\left(P_\eta - \hat{P}_\eta\right) h(w)\right| + \sum_{i=1}^{N-1} \sup_{h\in \text{Lip}(1)} \left|\hat{P}_{(i-1)\eta}\left(P_\eta - \hat{P}_\eta\right) P_{(N-i)\eta} h(w)\right| \\
&\leq \left(K_1 + \rho(X_n, \hat{X}_n)K_2\right)(2C)\left(1 + C_0(1+\|w\|)\right)\eta^{1+\frac{1}{\alpha}} + \rho(X_n, \hat{X}_n)K_2\left(2C_0(1+\|w\|)+1\right)\eta \\
&\quad + \left(C_1\lambda^{-1}e^{\lambda}+1\right)e^L\left(K_1 + \rho(X_n, \hat{X}_n)K_2\right)(2C)\left(1 + C_0(1+\|w\|)\right)\eta^{\frac{1}{\alpha}} \\
&\quad\quad + \left(C_1\lambda^{-1}e^{\lambda}+1\right)e^L\rho(X_n, \hat{X}_n)K_2\left(2C_0(1+\|w\|)+1\right). \qquad (54)
\end{aligned}$$

This completes the proof of part (III).

(ii) Now, we are ready prove part (ii). By triangle inequality for 1-Wasserstein distance,

$$\mathcal{W}_1\left(\mu, \hat{\mu}\right) \leq \mathcal{W}_1\left(\text{Law}(\theta_{\eta N}), \mu\right) + \mathcal{W}_1\left(\text{Law}(\theta_{\eta N}), \text{Law}(\hat{\theta}_{\eta N})\right) + \mathcal{W}_1\left(\text{Law}(\hat{\theta}_{\eta N}), \hat{\mu}\right).$$

It follows from Lemma C.1 that by letting $N \to \infty$, we have

$$\begin{aligned}
&\mathcal{W}_1\left(\mu, \hat{\mu}\right) \\
&\leq \limsup_{N\to\infty} \mathcal{W}_1\left(\text{Law}(\theta_{\eta N}), \text{Law}(\hat{\theta}_{\eta N})\right) \\
&\leq \left(K_1 + \rho(X_n, \hat{X}_n)K_2\right)(2C)\left(1 + C_0(1+\|w\|)\right)\eta^{1+\frac{1}{\alpha}} + \rho(X_n, \hat{X}_n)K_2\left(2C_0(1+\|w\|)+1\right)\eta \\
&\quad + \left(C_1\lambda^{-1}e^{\lambda}+1\right)e^L\left(K_1 + \rho(X_n, \hat{X}_n)K_2\right)(2C)\left(1 + C_0(1+\|w\|)\right)\eta^{\frac{1}{\alpha}} \\
&\quad\quad + \left(C_1\lambda^{-1}e^{\lambda}+1\right)e^L\rho(X_n, \hat{X}_n)K_2\left(2C_0(1+\|w\|)+1\right), \qquad (55)
\end{aligned}$$

where we used part (III) from part (i). Since $\mathcal{W}_1\left(\mu, \hat{\mu}\right)$ is independent of $\eta$ and the initial state $x \in \mathbb{R}^d$, we can set $\eta = 0$ and $x = 0$ in (55) and conclude that

$$\mathcal{W}_1\left(\mu, \hat{\mu}\right) \leq \left(C_1\lambda^{-1}e^{\lambda}+1\right)e^L\rho(X_n, \hat{X}_n)K_2\left(2C_0+1\right).$$

The proof is complete. □

### B.2. Proof of Lemma 3.5

*Proof of Lemma 3.5.* First of all, the infinitesimal generator of $\theta_t$ process is given by

$$\mathcal{L}^\alpha f(\theta) = \left\langle -\nabla\hat{F}(\theta, X_n), \nabla f(\theta)\right\rangle + (-\Delta)^{\alpha/2} f(\theta), \qquad (56)$$

where $(-\Delta)^{\alpha/2}$ is the fractional Laplacian operator defined as a principal value integral:

$$(-\Delta)^{\alpha/2} f(\theta) = d_\alpha \cdot \text{p.v.} \int_{\mathbb{R}^d} (f(\theta+y) - f(\theta)) \frac{dy}{\|y\|^{\alpha+d}}, \qquad (57)$$

where (see e.g. (Wang, 2016))

$$d_\alpha := \frac{2^\alpha \Gamma\left(\frac{d+\alpha}{2}\right)\pi^{-d/2}}{|\Gamma(-\alpha/2)|}. \qquad (58)$$

We derive from Assumption 3.2 that for any dataset $X_n \in \mathcal{X}^n$, we have the following property:

$$\begin{aligned}
&\left\|\nabla\hat{F}(0, X_n)\right\| \leq B, \\
&\left\langle \nabla\hat{F}(\theta_1, X_n) - \nabla\hat{F}(\theta_2, X_n), \theta_1 - \theta_2\right\rangle \geq m\|\theta_1 - \theta_2\|^2 - K,
\end{aligned}$$

and

$$\left\|\nabla_v \nabla \hat{F}(\theta, X_n)\right\| \leq L\|v\|, \qquad \left\|\nabla_{v_1} \nabla_{v_2} \nabla \hat{F}(\theta, X_n)\right\| \leq M\|v_1\|\|v_2\|,$$

for any $\theta, \theta_1, \theta_2, v, v_1, v_2 \in \mathbb{R}^d$ so that we can apply Proposition 2.1. in (Chen et al., 2022).

Next, let $V(w) := (1 + \|w\|^2)^{1/2}$. It is shown in the proof of Proposition 2.1. in (Chen et al., 2022) that $V \in \mathcal{D}(\mathcal{L}^\alpha)$, i.e. the domain of the infinitesimal generator $\mathcal{L}^\alpha$ and moreover

$$\mathcal{L}^\alpha V(w) \leq -\lambda_1 V(w) + q_1, \tag{59}$$

where

$$\lambda_1 := \frac{1}{2}m, \qquad q_1 := m + K + B + C_{d,\alpha}, \tag{60}$$

where

$$C_{d,\alpha} := \frac{d_\alpha \sqrt{d}\sigma_{d-1}}{2-\alpha} + \frac{d_\alpha \sigma_{d-1}}{\alpha - 1} = \frac{2^\alpha \Gamma\left(\frac{d+\alpha}{2}\right)\pi^{-d/2}\sqrt{d}\sigma_{d-1}}{|\Gamma(-\alpha/2)|(2-\alpha)} + \frac{2^\alpha \Gamma\left(\frac{d+\alpha}{2}\right)\pi^{-d/2}\sigma_{d-1}}{|\Gamma(-\alpha/2)|(\alpha-1)}, \tag{61}$$

where $\sigma_{d-1} := 2\pi^{\frac{d}{2}}/\Gamma(d/2)$ is the surface area of the unit sphere in $\mathbb{R}^d$, a positive constant that depends only on $d$.

Next, let us define the extended infinitesimal generator $\mathcal{L}_t^\alpha$:

$$\mathcal{L}_t^\alpha f(t, \theta) := \partial_t f(t, \theta) + \mathcal{L}^\alpha f(t, \theta). \tag{62}$$

Then, it follows from (59) that

$$\mathcal{L}_t^\alpha e^{\lambda_1 t} V(\theta) = \lambda_1 e^{\lambda_1 t} V(\theta) + e^{\lambda_1 t}\mathcal{L}^\alpha V(\theta) \leq \lambda_1 e^{\lambda_1 t} V(\theta) + e^{\lambda_1 t}\left(-\lambda_1 V(w) + q_1\right) = q_1 e^{\lambda_1 t}. \tag{63}$$

By Dynkin's formula,

$$\begin{aligned}\mathbb{E}\left[e^{\lambda_1 t} V\left(\theta_t^w\right)\right] &= V(w) + \mathbb{E}\left[\int_0^t \mathcal{L}_s^\alpha e^{\lambda_1 s} V\left(\theta_s^w\right) ds\right] \\ &\leq V(w) + \int_0^t q_1 e^{\lambda_1 s} ds = V(w) + q_1 \frac{e^{\lambda_1 t} - 1}{\lambda_1},\end{aligned}$$

which implies that

$$\mathbb{E}\|\theta_t^w\| \leq \mathbb{E}\left[V\left(\theta_t^w\right)\right] \leq e^{-\lambda_1 t} V(w) + q_1 \frac{1 - e^{-\lambda_1 t}}{\lambda_1} \leq 1 + \|w\| + \frac{q_1}{\lambda_1}, \tag{64}$$

where we used the definition $V(w) = (1 + \|w\|^2)^{1/2}$ and the inequality $\|w\| \leq (1 + \|w\|^2)^{1/2} \leq 1 + \|w\|$. Hence, we have

$$\mathbb{E}\|\theta_t^w\| \leq C_0(1 + \|w\|), \tag{65}$$

where we take

$$\begin{aligned}C_0 &:= 1 + \frac{q_1}{\lambda_1} = 1 + \frac{2\left(m + K + B + C_{d,\alpha}\right)}{m} \\ &= 3 + \frac{2\left(K + B\right)}{m} + \frac{2}{m}\left(\frac{2^\alpha \Gamma\left(\frac{d+\alpha}{2}\right)\pi^{-d/2}\sqrt{d}\sigma_{d-1}}{|\Gamma(-\alpha/2)|(2-\alpha)} + \frac{2^\alpha \Gamma\left(\frac{d+\alpha}{2}\right)\pi^{-d/2}\sigma_{d-1}}{|\Gamma(-\alpha/2)|(\alpha-1)}\right).\end{aligned}$$

Since $C_0$ in the above equation is uniform in the dataset, similarly, we also have

$$\mathbb{E}\|\hat{\theta}_t^w\| \leq C_0(1 + \|w\|), \tag{66}$$

which completes the proof. □

### B.3. Proof of Proposition 3.6 (ii)

*Proof of Proposition 3.6 (ii).* Now, let us prove part (ii) of Proposition 3.6. We recall from (25) that

$$g(\alpha;d)=\frac{2^{\alpha}\Gamma\left(\frac{d+\alpha}{2}\right)}{|\Gamma(-\alpha/2)|(2-\alpha)}\left(\sqrt{d}+\frac{2-\alpha}{\alpha-1}\right). \tag{67}$$

We can compute that

$$\frac{\partial}{\partial\alpha}g(\alpha;d)=\frac{2^{\alpha}\Gamma\left(\frac{d+\alpha}{2}\right)}{|\Gamma(-\alpha/2)|(2-\alpha)}\frac{-1}{(\alpha-1)^2}+\frac{\partial}{\partial\alpha}\left\{\frac{2^{\alpha}\Gamma\left(\frac{d+\alpha}{2}\right)}{\Gamma(-\alpha/2)(\alpha-2)}\right\}\left(\sqrt{d}+\frac{2-\alpha}{\alpha-1}\right),$$

where we can further compute that

$$\begin{aligned}\frac{\partial}{\partial\alpha}\left\{\frac{2^{\alpha}\Gamma\left(\frac{d+\alpha}{2}\right)}{\Gamma(-\alpha/2)(\alpha-2)}\right\}=&\frac{\log(2)2^{\alpha}\Gamma\left(\frac{d+\alpha}{2}\right)+2^{\alpha-1}\Gamma\left(\frac{d+\alpha}{2}\right)\psi\left(\frac{d+\alpha}{2}\right)}{\Gamma(-\alpha/2)(\alpha-2)}\\&-\frac{2^{\alpha}\Gamma\left(\frac{d+\alpha}{2}\right)\left(-\frac{1}{2}(\alpha-2)\psi(-\frac{\alpha}{2})+1\right)}{\Gamma(-\alpha/2)(\alpha-2)^2},\end{aligned}$$

where $\psi(\cdot)$ denotes the digamma function. This implies that

$$\frac{\partial}{\partial\alpha}g(\alpha;d)=\frac{2^{\alpha}\Gamma\left(\frac{d+\alpha}{2}\right)}{|\Gamma(-\alpha/2)|(2-\alpha)}p(\alpha;d), \tag{68}$$

where

$$\begin{aligned}p(\alpha;d):=&\frac{-1}{(\alpha-1)^2}+\left(\log(2)+\frac{1}{2}\psi\left(\frac{d+\alpha}{2}\right)\right)\left(\sqrt{d}+\frac{2-\alpha}{\alpha-1}\right)\\&+\left(\frac{1}{2}\psi\left(-\frac{\alpha}{2}\right)+\frac{1}{2-\alpha}\right)\left(\sqrt{d}+\frac{2-\alpha}{\alpha-1}\right).\end{aligned}$$

By the property of the digamma function, we have $\psi(-\frac{\alpha}{2})=\psi(1-\frac{\alpha}{2})+\frac{2}{\alpha}$ and $\psi(x)$ is increasing in $x>0$ and $\psi(-1/2)<0$. Therefore, for any $1<\alpha\le\alpha_0$, we have

$$\begin{aligned}p(\alpha;d)=&\frac{-1}{(\alpha-1)^2}+\left(\log(2)+\frac{1}{2}\psi\left(\frac{d+\alpha}{2}\right)\right)\left(\sqrt{d}+\frac{2-\alpha}{\alpha-1}\right)\\&+\left(\frac{1}{2}\psi\left(1-\frac{\alpha}{2}\right)+\frac{1}{\alpha}+\frac{1}{2-\alpha}\right)\left(\sqrt{d}+\frac{2-\alpha}{\alpha-1}\right)\\\le&\frac{-1}{(\alpha-1)^2}+\left(\log(2)+\frac{1}{2}\psi\left(\frac{d+\alpha_0}{2}\right)\right)\left(\sqrt{d}+\frac{1}{\alpha-1}\right)\\&+\left(1+\frac{1}{2-\alpha_0}\right)\left(\sqrt{d}+\frac{1}{\alpha-1}\right).\end{aligned}$$

It follows that $p(\alpha;d)\le 0$ holds if

$$y_0\sqrt{d}(\alpha-1)^2+y_0(\alpha-1)-1\le 0, \tag{69}$$

where

$$y_0:=\log(2)+\frac{1}{2}\psi\left(d+\frac{\alpha}{2}\right)+\frac{3-\alpha_0}{2-\alpha_0},$$

and it is easy to compute that (69) holds provided that

$$\alpha\le 1+\frac{-1+\sqrt{1+4y_0^{-1}\sqrt{d}}}{2\sqrt{d}}. \tag{70}$$

Hence, we conclude that $p(\alpha;d)$ is non-positive and thus $\frac{\partial}{\partial\alpha}g(\alpha;d)$ is non-positive (by (68)) and therefore $g(\alpha;d)$ is decreasing for any $\alpha\in[1,\alpha_0']$, where $\alpha_0':=\min\left(\alpha_0,1+\frac{-1+\sqrt{1+4y_0^{-1}\sqrt{d}}}{2\sqrt{d}}\right)$. The proof is complete. □

### B.4. Proof of Corollary 3.9

**Corollary B.2** (Restatement of Corollary 3.9)**.** *Under the assumptions in Theorem B.1 and Lemma C.6, we have:*

*(i) For any* $2 \leq N \leq \eta^{-1}+1$,

$$\begin{aligned}
&\mathcal{W}_1\left(Law(\theta_{\eta N}), Law(\hat{\theta}_{\eta N})\right) \\
&\leq \left(K_1+\rho(X_n,\hat{X}_n)K_2\right)(2C)(1+C_0(1+\|w\|))\eta^{1+\frac{1}{\alpha}}+\rho(X_n,\hat{X}_n)K_2(2C_0(1+\|w\|)+1)\eta \\
&\quad + e^L\left(K_1+\rho(X_n,\hat{X}_n)K_2\right)(2C)\left(1+C_0(1+\|w\|)\right)\eta^{\frac{1}{\alpha}} \\
&\quad\quad + e^L\rho(X_n,\hat{X}_n)K_2(2C_0(1+\|w\|)+1)+2Q(1+\|w\|)\eta^{2/\alpha-1},
\end{aligned} \tag{71}$$

*and for any* $N > \eta^{-1}+1$,

$$\begin{aligned}
&\mathcal{W}_1\left(Law(\theta_{\eta N}), Law(\hat{\theta}_{\eta N})\right) \\
&\leq \left(K_1+\rho(X_n,\hat{X}_n)K_2\right)(2C)\left(1+C_0(1+\|w\|)\right)\eta^{1+\frac{1}{\alpha}}+\rho(X_n,\hat{X}_n)K_2(2C_0(1+\|w\|)+1)\eta \\
&\quad\quad + \left(C_1\lambda^{-1}e^{\lambda}+1\right)e^L\left(K_1+\rho(X_n,\hat{X}_n)K_2\right)(2C)\left(1+C_0(1+\|w\|)\right)\eta^{\frac{1}{\alpha}} \\
&\quad\quad + \left(C_1\lambda^{-1}e^{\lambda}+1\right)e^L\rho(X_n,\hat{X}_n)K_2(2C_0(1+\|w\|)+1)+2Q(1+\|w\|)\eta^{2/\alpha-1}.
\end{aligned} \tag{72}$$

*(ii) We have*

$$\mathcal{W}_1(\mu,\hat{\mu}) \leq \left(C_1\lambda^{-1}e^{\lambda}+1\right)e^L\rho(X_n,\hat{X}_n)K_2\left(2C_0+1\right)+2Q\eta^{2/\alpha-1}. \tag{73}$$

*Proof.* Let us prove part (ii) and the proof for part (i) is similar. It follows directly from Lemma 3.8 and Theorem 3.3 and the triangle inequality for 1-Wasserstein distance:

$$\mathcal{W}_1(\nu,\hat{\nu}) \leq \mathcal{W}_1(\nu,\mu)+\mathcal{W}_1(\hat{\nu},\hat{\mu})+\mathcal{W}_1(\mu,\hat{\mu}). \tag{74}$$

The proof is complete. □

## C. Technical Lemmas

In this section, we provide some technical results that are used in the proofs of main results in Section B. First, we have the following technical result from (Chen et al., 2022).

**Lemma C.1** (Proposition 2.1. in Chen et al. (2022))**.** *Under Assumption 3.2,* $(\theta_t^w)_{t\geq 0}$ *and* $(\hat{\theta}_t^w)_{t\geq 0}$ *admit unique invariant probability measures* $\mu$ *and* $\hat{\mu}$ *respectively such that*

$$\sup_{|f|\leq V}\left|\mathbb{E}[f(\theta_t^w)]-\mu(f)\right| \leq c_1V(w)e^{-c_2t}, \qquad \text{for any } t>0, \tag{75}$$

$$\sup_{|f|\leq V}\left|\mathbb{E}[f(\hat{\theta}_t^w)]-\hat{\mu}(f)\right| \leq c_1V(w)e^{-c_2t}, \qquad \text{for any } t>0, \tag{76}$$

*for some constants* $c_1, c_2 > 0$ *where* $V(w) := (1+\|w\|^2)^{1/2}$ *is a Lyapunov function. In particular, there exists a constant* $C_0 > 0$ *such that*

$$\mathbb{E}\|\theta_t^w\| \leq C_0(1+\|w\|), \qquad \text{for any } t>0, \tag{77}$$

$$\mathbb{E}\|\hat{\theta}_t^w\| \leq C_0(1+\|w\|), \qquad \text{for any } t>0. \tag{78}$$

Moreover, we recall the following technical lemma.

**Lemma C.2** (Proposition 2.2 in Chen et al. (2022))**.** *There exist constants* $C_1, \lambda > 0$ *such that for any* $t > 0$ *and* $w, y \in \mathbb{R}^d$*, we have*

$$\mathcal{W}_1\left(Law\left(\theta_t^w\right), Law\left(\theta_t^y\right)\right) \leq C_1e^{-\lambda t}\|w-y\|, \tag{79}$$

$$\mathcal{W}_1\left(Law\left(\hat{\theta}_t^w\right), Law\left(\hat{\theta}_t^y\right)\right) \leq C_1e^{-\lambda t}\|w-y\|. \tag{80}$$

Let $P_t$ and $\hat{P}_t$ denote the Markov semigroups of $\theta_t$ and $\hat{\theta}_t$ processes respectively, that is, for any bounded function $f : \mathbb{R}^d \to \mathbb{R}$,

$$P_t f(x) = \mathbb{E} f(\theta_t^w), \qquad \hat{P}_t f(x) = \mathbb{E} f(\hat{\theta}_t^w). \tag{81}$$

We have the following technical lemma from (Chen et al., 2022).

**Lemma C.3** (Lemma 3.1 in Chen et al. (2022)). *For any $h \in Lip(1)$ and $v, w \in \mathbb{R}^d$ and $t \in (0, 1]$, we have*

$$\|\nabla_v P_t h(w)\| \le e^L \|v\|, \qquad \left\|\nabla_v \hat{P}_t h(w)\right\| \le e^L \|v\|, \tag{82}$$

*where $L$ is defined in Assumption 3.2.*

We recall the following technical lemma from (Chen et al., 2022).

**Lemma C.4** (Lemma 3.2 in Chen et al. (2022)). *There exist constants $C > 0$ such that for all $w \in \mathbb{R}^d$, $t \ge 0$, we have*

$$\mathbb{E}\|\theta_t^w - w\| \le C(1 + \|w\|) \left(t \vee t^{1/\alpha}\right), \tag{83}$$

$$\mathbb{E}\|\hat{\theta}_t^w - w\| \le C(1 + \|w\|) \left(t \vee t^{1/\alpha}\right). \tag{84}$$

Next, we state and prove the following key technical lemma.

**Lemma C.5.** *There exist constants $C > 0$ such that for all $w \in \mathbb{R}^d$, $\eta \in (0, 1)$, $f : \mathbb{R}^d \to \mathbb{R}$ with $\|\nabla f\|_\infty < \infty$, we have*

$$\begin{aligned}
&\left|P_\eta f(w) - \hat{P}_\eta f(w)\right| \\
&\le \|\nabla f\|_\infty \left[\left(K_1 + \rho(X_n, \hat{X}_n) K_2\right) 2C(1 + \|w\|)\eta^{1+\frac{1}{\alpha}} + \rho(X_n, \hat{X}_n) K_2 (2\|w\| + 1)\eta\right].
\end{aligned} \tag{85}$$

*Proof of Lemma C.5.* We can compute that

$$\begin{aligned}
\left|P_\eta f(w) - \hat{P}_\eta f(w)\right| &= \left|\mathbb{E}\left[f\left(\theta_\eta^w\right) - f\left(\hat{\theta}_\eta^w\right)\right]\right| \\
&= \left|\mathbb{E}\left[f\left(w + \int_0^\eta \nabla \hat{F}\left(\theta_r^w, X_n\right) dr + L_\eta^\alpha\right) - f\left(w + \int_0^\eta \nabla \hat{F}\left(\hat{\theta}_r^w, \hat{X}_n\right) dr + L_\eta^\alpha\right)\right]\right| \\
&\le \|\nabla f\|_\infty \mathbb{E}\left\|\int_0^\eta \nabla \hat{F}\left(\theta_r^w, X_n\right) dr - \int_0^\eta \nabla \hat{F}\left(\hat{\theta}_r^w, \hat{X}_n\right) dr\right\| \\
&\le \|\nabla f\|_\infty \mathbb{E} \int_0^\eta \left\|\nabla \hat{F}(\theta_r^w, X_n) - \nabla \hat{F}\left(\hat{\theta}_r^w, \hat{X}_n\right)\right\| dr \\
&\le \|\nabla f\|_\infty \mathbb{E} \int_0^\eta \left(K_1 \|\theta_r^w - \hat{\theta}_r^w\| + \rho(X_n, \hat{X}_n) K_2 \left(\|\theta_r^w\| + \|\hat{\theta}_r^w\| + 1\right)\right) dr \\
&= \|\nabla f\|_\infty \left[K_1 \int_0^\eta \mathbb{E}\|\theta_r^w - \hat{\theta}_r^w\| dr + \rho(X_n, \hat{X}_n) K_2 \int_0^\eta \mathbb{E}\left(\|\theta_r^w\| + \|\hat{\theta}_r^w\| + 1\right) dr\right].
\end{aligned}$$

By Lemma C.4, we have

$$\begin{aligned}
\int_0^\eta \mathbb{E}\|\theta_r^w - \hat{\theta}_r^w\| dr &\le \int_0^\eta \mathbb{E}\|\theta_r^w - w\| dr + \int_0^\eta \mathbb{E}\|\hat{\theta}_r^w - w\| dr \\
&\le C(1 + \|w\|) \int_0^\eta r^{1/\alpha} dr + C(1 + \|w\|) \int_0^\eta r^{1/\alpha} dr \\
&\le 2C(1 + \|w\|)\eta^{1+\frac{1}{\alpha}}.
\end{aligned}$$

By applying Lemma C.4 again, we have

$$\begin{aligned}
\int_0^\eta \mathbb{E}\left(\|\theta_r^w\| + \|\hat{\theta}_r^w\| + 1\right) dr &\le \int_0^\eta \mathbb{E}\left(\|\theta_r^w - w\| + \|\hat{\theta}_r^w - w\| + 2\|w\| + 1\right) dr \\
&\le \int_0^\eta \left(C(1 + \|w\|) r^{1/\alpha} + C(1 + \|w\|) r^{1/\alpha} + 2\|w\| + 1\right) dr \\
&\le 2C(1 + \|w\|)\eta^{1+\frac{1}{\alpha}} + (2\|w\| + 1)\eta.
\end{aligned}$$

Hence, we conclude that

$$\begin{aligned}\left|P_\eta f(w) - \hat{P}_\eta f(w)\right| &= \left|\mathbb{E}\left[f(\theta_\eta^w) - f(\hat{\theta}_\eta^w)\right]\right| \\ &\leq \|\nabla f\|_\infty \left[\left(K_1 + \rho(X_n, \hat{X}_n)K_2\right)(2C)\left(1 + \|w\|\right)\eta^{1+\frac{1}{\alpha}} + \rho(X_n, \hat{X}_n)K_2(2\|w\| + 1)\eta\right].\end{aligned}$$

This completes the proof. □

**Lemma C.6** (Restatement of Lemma 3.8 (Theorem 1.2. in Chen et al. (2022))). *Let $\mu_t$ and $\hat{\mu}_t$ denote the distributions of continuous-time $\theta_t$ and $\hat{\theta}_t$ and $\mu$ and $\hat{\mu}$ denote the distributions of continuous-time $\theta_\infty$ and $\hat{\theta}_\infty$. Moreover, let $\nu_k$ and $\hat{\nu}_k$ denote the distributions of discrete-time $\theta_k$ and $\hat{\theta}_k$ and $\nu$ and $\hat{\nu}$ denote the distributions of discrete-time $\theta_\infty$ and $\hat{\theta}_\infty$. Assume the dynamics start at $w$ at time $0$. Let $m$, $L$ be as in Assumption 3.2.*

*Then, there exists some constant $Q$ (that may depend on $B, m, K, L, M$ from Assumption 3.2) such that the followings hold.*

*(i) For every $N \geq 2$ and $\eta < \min\{1, m/(8L^2), 1/m\}$, one has*

$$\mathcal{W}_1(\mu_{N\eta}, \nu_N) \leq Q(1 + \|w\|)\eta^{2/\alpha - 1}, \tag{86}$$

$$\mathcal{W}_1(\hat{\mu}_{N\eta}, \hat{\nu}_N) \leq Q(1 + \|w\|)\eta^{2/\alpha - 1}. \tag{87}$$

*(ii) For every $\eta < \min\{1, m/L^2, 1/m\}$, one has*

$$\mathcal{W}_1(\mu, \nu) \leq Q\eta^{2/\alpha - 1}, \tag{88}$$

$$\mathcal{W}_1(\hat{\mu}, \hat{\nu}) \leq Q\eta^{2/\alpha - 1}. \tag{89}$$