Neural network compression with heavy-tailed SGD

Yijun Wan[1], **Abdellatif Zaidi**[1,2], **Umut Şimşekli**[3]

[1]Paris Research Center, Huawei Technologies France
{wan.yijun, abdellatif.zaidi}@huawei.com
[2]Université Gustave Eiffel, France
{abdellatif.zaidi@univ-eiffel.fr
[3]Inria, CNRS, Ecole Normale Supérieure, PSL Research University, Paris, France
umut.simsekli@inria.fr

Abstract

Neural network compression has been an increasingly important subject, due to its practical implications in terms of reducing the computational requirements and its theoretical implications, as there is an explicit connection between compressibility and the generalization error. Recent studies have shown that the choice of the hyperparameters of stochastic gradient descent (SGD) can have an effect on the compressibility of the learned parameter vector. Even though these results have shed some light on the role of the training dynamics over compressibility, they relied on unverifiable assumptions and the resulting theory does not provide a practical guideline due to its implicitness. In this study, we propose a simple modification for SGD, such that the outputs of the algorithm will be provably compressible without making any nontrivial assumptions. We consider a one-hidden-layer neural network trained with SGD and we inject additive heavy-tailed noise to the iterates at each iteration. We then show that, for *any* compression rate, there exists a level of overparametrization (i.e., the number of hidden units), such that the output of the algorithm will be compressible with high probability. We illustrate our approach on experiments, where the results suggest that the proposed approach achieves compressibility with a slight compromise from the training and test error.

1 Introduction

Obtaining compressible neural networks has become an increasingly important task in the last decade, and it has essential implications from both practical and theoretical perspectives. From a practical point of view, as the modern network architectures might contain an excessive number of parameters, compression has a crucial role in terms of deployment of such networks in resource-limited environments [O'N20, BOFG20]. On the other hand, from a theoretical perspective, several studies have shown that compressible neural networks should achieve a better generalization performance due to their lower-dimensional structure [AGNZ18, SAM$^+$20, SAN20, HJTW21, BSE$^+$21, SGRS22].

Despite their evident benefits, it is still not yet clear how to obtain compressible networks with provable guarantees. In an empirical study [FC18], introduced the 'lottery ticket hypothesis', which indicated that a randomly initialized neural network will have a sub-network that can achieve a performance that is comparable to the original network; hence, the original network can be compressed to the smaller sub-network. This empirical study has formed a fertile ground for subsequent theoretical research, which showed that such a sub-network can indeed exist (see e.g., [MYSSS20, BLMG21, dCNV22]); yet, it is not clear how to develop an algorithm that can find it in a feasible amount of time.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Another line of research has developed methods to enforce compressibility of neural networks by using sparsity enforcing regularizers, see e.g., [PRSE18, ACA19, CJS+20, Led23, KW23]. While they have led to interesting algorithms, the resulting algorithms typically require higher computational needs due to the increased complexity of the problem. On the other hand, due to the nonconvexity of the overall objective, it is also not trivial to provide theoretical guarantees for the compressibility of the resulting network weights.

Recently it has been shown that the training dynamics can have an influence on the compressibility of the algorithm output. In particular, motivated by the empirical and theoretical evidence that heavy-tails might arise in stochastic optimization (see e.g., [MM19, SSG19, ŞGN+19, ŞZTG20, ZFM+20, ZKV+20, CWZ+21]), [BSE+21, Shi21] showed that the network weights learned by stochastic gradient descent (SGD) will be compressible if we assume that they are heavy-tailed and there exists a certain form of statistical independence within the network weights. These studies illustrated that, even *without* any modification to the optimization algorithm, the learned network weights can be compressible depending on the algorithm hyperparameters (such as the step-size or the batch-size). Even though the tail and independence conditions were recently relaxed in [LAJ+22], the resulting theory relies on unverifiable assumptions, and hence does not provide a practical guideline.

In this paper, we focus on single-hidden-layer neural networks with a fixed second layer (i.e., the setting used in previous work [DBDFS20]) trained with vanilla SGD, and show that, when the iterates of SGD are simply perturbed by heavy-tailed noise with infinite variance (similar to the settings considered in [Şim17, NSR19, ŞZTG20, HMW21, ZZ23]), the assumption made in [BSE+21] in effect holds. More precisely, denoting the number of hidden units by n and the step-size of SGD by η, we consider the *mean-field limit*, where n goes to infinity and η goes to zero. We show that in this limiting case, the columns of the weight matrix will be independent and identically distributed (i.i.d.) with a common *heavy-tailed* distribution. Then, we focus on the finite n and η regime and we prove that for *any* compression ratio (to be precised in the next section), there exists a number N, such that if $n \geq N$ and η is sufficiently small, the network weight matrix will be compressible with high probability. Figure 1 illustrates the overall approach and precises our notion of compressibility.

After deriving the Euler-Maruyama-type guarantee for approximation SGD by its mean-field limit, we prove a high-probability compression bound by invoking [GCD12, AUM11], which essentially shows that an i.i.d. sequence of heavy-tailed random variables will have a small proportion of elements that will dominate the whole sequence in terms of absolute values (to be stated formally in the next section). Here, we shall note that similar mean-field regimes have already been considered in machine learning (see e.g., [MMN18, CB18, RVE18, JŠS19, MMM19, DBDFS20, SS22]). However, these studies all focused on particle SDE systems that either converge to deterministic systems or that are driven by Brownian motion. While they have introduced interesting analysis tools, we cannot directly benefit from their analysis in this paper, since the heavy-tails are crucial for obtaining compressibility, and the Brownian-driven SDEs cannot produce heavy-tailed solutions in general.

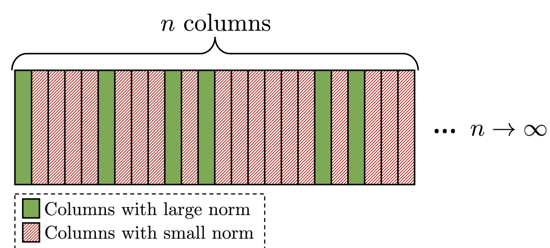

Figure 1: The illustration of the overall approach. We consider a one-hidden-layer neural network with n hidden units, which results in a weight matrix of n columns (first layer). We show that, when SGD is perturbed with heavy-tailed noise, as $n \to \infty$, each column will follow a multivariate heavy-tailed distribution in an i.i.d. fashion. This implies that a small number of columns will have significantly larger norms compared to the others; hence, the norm of the overall weight matrix will be determined by such columns [GCD12]. As a result, the majority of the columns can be removed (i.e., set to zero), which we refer to as compressibility.

To validate our theory, we conduct experiments on single-hidden-layer neural networks on different datasets. Our results show that, even with a minor modification to SGD (i.e., injecting heavy-tailed noise), the proposed approach can achieve compressibility with a negligible computational overhead and with a slight compromise from the training and test error. For instance, on a classification task with the MNIST dataset, when we set $n = 10$K, with vanilla SGD, we obtain a test accuracy of

94.69%, whereas with the proposed approach, we can remove 44% of the columns of the weight matrix, while maintaining a test accuracy of 94.04%. We provide all the proofs in the appendix.

2 Preliminaries

Notation. For a vector $u \in \mathbb{R}^d$, denote by $\|u\|$ its Euclidean norm, and by $\|u\|_p$ its ℓ_p norm. For a function $f \in C(\mathbb{R}^{d_1}, \mathbb{R}^{d_2})$, denote by $\|f\|_\infty := \sup_{x \in \mathbb{R}^{d_1}} \|f(x)\|$ its L^∞ norm. For a family of n (or infinity) vectors, the indexing $\cdot^{i,n}$ denotes the i-th vector in the family. In addition, for random variables, $\stackrel{(\mathrm{d})}{=}$ means equality in distribution, and the space of probability measures on $\mathbb{R}^d$ is denoted by $\mathcal{P}(\mathbb{R}^d)$. For a matrix $A \in \mathbb{R}^{d_1 \times d_2}$, its Frobenius norm is denoted by $\|A\|_F = \sqrt{\sum_{i=1}^{d_1} \sum_{j=1}^{d_2} |a_{i,j}|^2}$. Without specifically mentioning, $\mathbb{E}$ denotes the expectation over all the randomness taken into consideration.

2.1 Alpha-stable processes

A random variable $X \in \mathbb{R}^d$ is called α-*stable* with the stability parameter $\alpha \in (0, 2]$, if $X_1, X_2, \ldots$ are independent copies of X, then $n^{-1/\alpha} \sum_{j=1}^n X_j \stackrel{(\mathrm{d})}{=} X$ for all $n \geq 1$ [Sam17]. Stable distributions appear as the limiting distribution in the generalized central limit theorem (CLT) [gne54]. In the one-dimensional case ($d = 1$), we call the variable X a symmetric α-stable random variable if its characteristic function is of the following form: $\mathbb{E}[\exp(i\omega X)] = \exp(-|\omega|^\alpha)$ for $\omega \in \mathbb{R}$.

For symmetric α-stable distributions, the case $\alpha = 2$ corresponds to the Gaussian distribution, while $\alpha = 1$ corresponds to the Cauchy distribution. An important property of α-stable distributions is that in the case $\alpha \in (1, 2)$, the p-th moment of an α-stable random variable is finite if and only if $p < \alpha$; hence, the distribution is heavy-tailed. In particular, $\mathbb{E}[|X|] < \infty$ and $\mathbb{E}[|X|^2] = \infty$, which can be used to model phenomena with heavy-tailed observations.

There exist different types of α-stable random vectors in $\mathbb{R}^d$. In this study we will be interested in the following three variants, whose characteristic functions (for $u \in \mathbb{R}^d$) are given as follows:

- **Type-I.** Let $Z \in \mathbb{R}$ be a symmetric α-stable random variable. We then construct the random vector X such that all the coordinates of X is equated to Z. In other words $X = \mathbf{1}_d Z$, where $\mathbf{1}_d \in \mathbb{R}^d$ is a vector of ones. With this choice, X admits the following characteristic function: $\mathbb{E}\left[\exp(i\langle u, X\rangle\right] = \exp(-|\langle u, \mathbf{1}_d\rangle|^\alpha)$;
- **Type-II.** X has i.i.d. coordinates, such that each component of X is a symmetric α-stable random variable in $\mathbb{R}$. This choice yields the following characteristic function: $\mathbb{E}\left[\exp(i\langle u, X\rangle\right] = \exp(-\sum_{i=1}^d |u_i|^\alpha)$;
- **Type-III.** X is rotationally invariant α-stable random vector with the characteristic function $\mathbb{E}\left[\exp(i\langle u, X\rangle\right] = \exp(-\|u\|^\alpha)$.

Note that the Type-II and Type-III noises reduce to a Gaussian distribution when $\alpha = 2$ (i.e., the characteristic function becomes $\exp(-\|u\|^2)$).

Similar to the fact that stable distributions extend the Gaussian distribution, we can define a more general random process, called the α-*stable Lévy process*, that extends the Brownian motion. Formally, α-stable processes are stochastic processes $(\mathrm{L}_t^\alpha)_{t \geq 0}$ with independent and stationary α-stable increments, and have the following definition:

- $\mathrm{L}_0^\alpha = 0$ almost surely,
- For any $0 \leq t_0 < t_1 < \cdots < t_N$, the increments $\mathrm{L}_{t_n}^\alpha - \mathrm{L}_{t_{n-1}}^\alpha$ are independent,
- For any $0 \leq s < t$, the difference $\mathrm{L}_t^\alpha - \mathrm{L}_s^\alpha$ and $(t-s)^{1/\alpha} \mathrm{L}_1^\alpha$ have the same distribution,
- L_t^α is stochastically continuous, i.e. for any $\delta > 0$ and $s \geq 0$, $\mathbb{P}(\|\mathrm{L}_t^\alpha - \mathrm{L}_s^\alpha\| > \delta) \to 0$ as $t \to s$.

To fully characterize an α-stable process, we further need to specify the distribution of L_1^α. Along with the above properties, the choice for L_1^α will fully determine the process. For this purpose, we will again consider the previous three types of α-stable vectors: We will call the process L_t^α a Type-I process if L_1^α is a Type-I α-stable random vector. We define the Type-II and Type-III processes analogously. Note that, when $\alpha = 2$, Type-II and Type-III processes reduce to the Brownian motion. For notational clarity, occasionally, we will drop the index α and denote the process by L_t.

3 Problem Setting and the Main Result

We consider a single hidden-layer overparametrized network of n units and use the setup provided in [DBDFS20]. Our goal is to minimize the expected loss in a supervised learning regime, where for each data $z = (x, y)$ distributed according to $\pi(\mathrm{d}x, \mathrm{d}y)$;[1] the feature x is included in $\mathcal{X} \subset \mathbb{R}^d$ and the label y is in $\mathcal{Y}$. We denote by $\theta^{i,n} \in \mathbb{R}^p$ the parameter for the i-th unit, and the parametrized model is denoted by $h_x : \mathbb{R}^p \to \mathbb{R}^l$. The mean-field network is the average over models for n units:

$$f_{\Theta^n}(x) = \frac{1}{n} \sum_{i=1}^{n} h_x(\theta^{i,n}),$$

where $\Theta^n = (\theta^{i,n})_{i=1}^n \in \mathbb{R}^{p \times n}$ denotes the collection of parameters in the network and $x \in \mathcal{X}$ is the feature variable for the data point. In particular, the mean-field network corresponds to a two-layer neural network with the weights of the second layer are fixed to be $1/n$ and Θ^n is the parameters of the first layer. While this model is less realistic than the models used in practice, nevertheless, we believe that it is desirable from theoretical point of view, and this defect can be circumvented upon replacing $h_x(\theta^{i,n})$ by $h_x(c^{i,n}, \theta^{i,n}) = c^{i,n} h_x(\theta^{i,n})$, where $c^{i,n}$ and $\theta^{i,n}$ are weights corresponding to different layers. However, in order to obtain similar results in this setup as in our paper, stronger assumptions are inevitable and the proof should be more involved, which are left for future work.

Given a loss function $\ell : \mathbb{R}^l \times \mathcal{Y} \to \mathbb{R}^+$, the goal (for each n) is to minimize the expected loss

$$R(\Theta^n) = \mathbb{E}_{(x,y)\sim\pi} \left[\ell \left(f_{\Theta^n}(x), y \right) \right]. \tag{1}$$

Let us set the notation for the proposed algorithm. Let $\hat{\theta}_0^{i,n}$, $i = 1, \ldots, n$, be the initial values of the iterates, which are n random variables in $\mathbb{R}^d$ distributed independently according to a given initial probability distribution μ_0. Then, we consider the gradient descent updates with stepsize ηn, which is perturbed by i.i.d. α-stable noises $\sigma \cdot \eta^{1/\alpha} X_k^{i,n}$ for each unit $i = 1, \ldots, n$ and some $\sigma > 0$:

$$\begin{cases} \hat{\theta}_{k+1}^{i,n} = \hat{\theta}_k^{i,n} - \eta n \left[\partial_{\theta^{i,n}} R(\Theta_k^n) \right] + \sigma \cdot \eta^{1/\alpha} X_k^{i,n} \\ \hat{\theta}_0^{i,n} \sim \mu_0 \in \mathcal{P}(\mathbb{R}^d), \end{cases} \tag{2}$$

where the scaling factor $\eta^{1/\alpha}$ in front of the stable noise enables the discrete dynamics of the system homogenize to SDEs as $\eta \to 0$. At this stage, we do not have to determine which type of stable noise (e.g., Type-I, II, or III) that we shall consider as they will all satisfy the requirements of our theory. However, our empirical findings will illustrate that the choice will affect the overall performance.

We now state the assumptions that will imply our theoretical results. The following assumptions are similar to [DBDFS20, Assumption A1].

Assumption 1.

- Regularity of the model: for each $x \in \mathcal{X}$, the function $h_x : \mathbb{R}^p \to \mathbb{R}^l$ is two-times differentiable, and there exists a function $\Psi : \mathcal{X} \to \mathbb{R}_+$ such that for any $x \in \mathcal{X}$,

$$\|h_x(\cdot)\|_\infty + \|\nabla h_x(\cdot)\|_\infty + \|\nabla^2 h_x(\cdot)\|\infty \leq \Psi(x).$$

- Regularity of the loss function: there exists a function $\Phi : \mathcal{Y} \to \mathbb{R}_+$ such that

$$\|\partial_1 \ell(\cdot, y)\|_\infty + \|\partial_1^2 \ell(\cdot, y)\|_\infty \leq \Phi(y)$$

- Moment bounds on $\Phi(\cdot)$ and $\Psi(\cdot)$: there exists a positive constant B such that

$$\mathbb{E}_{(x,y)\sim\pi}[\Psi^2(x)(1 + \Phi^2(y))] \leq B^2.$$

Let us remark that these are rather standard smoothness assumptions that have been made in the mean field literature [MMN18, MMM19] and are satisfied by several smooth activation functions, including the sigmoid and hyper-tangent functions.

[1] Note that for finite datasets, π can be chosen as a measure supported on finitely many points.

We now proceed to our main result. Let $\hat{\Theta}_k^n \in \mathbb{R}^{p\times n}$ be the concatenation of all parameters $\hat{\theta}_k^{i,n}$, $i = 1, \dots, n$ obtained by the recursion (2) after k iterations. We will now compress $\hat{\Theta}_k^n$ by pruning its columns with small norms. More precisely, fix a compression ratio $\kappa \in (0, 1)$, compute the norms of the columns of $\hat{\Theta}_k^n$, i.e., $\|\hat{\theta}_k^{i,n}\|$. Then, keep the $\lfloor \kappa n \rfloor$ columns, which have the largest norms, and set all the other columns to zero, in all their entirety. Finally, denote by $\hat{\Theta}_k^{(\kappa n)} \in \mathbb{R}^{p\times n}$, the pruned version of $\hat{\Theta}_k^n$.

Theorem 3.1. *Suppose that Assumption 1 holds. For any fixed $t > 0$, $\kappa \in (0, 1)$ and $\epsilon > 0$ sufficiently small, with probability $1 - \epsilon$, there exists $N \in \mathbb{N}_+$ such that for all $n \geq N$ and η such that $\eta \leq n^{-\alpha/2-1}$, the following upper bound on the relative compression error for the parameters holds:*

$$\frac{\left\| \hat{\Theta}^{(\kappa n)}_{\lfloor t/\eta \rfloor} - \hat{\Theta}^{n}_{\lfloor t/\eta \rfloor} \right\|_F}{\left\| \hat{\Theta}^{n}_{\lfloor t/\eta \rfloor} \right\|_F} \leq \epsilon.$$

This bound shows that, thanks to the heavy-tailed noise injections, the weight matrices will be compressible at *any* compression rate, as long as the network is sufficiently overparametrized and the step-size is sufficiently small. We shall note that this bound also enables us to directly obtain a generalization bound by invoking [BSE$^+$21, Theorem 4].

4 Empirical Results

In this section, we validate our theory with empirical results. Our goal is to investigate the effects of the heavy-tailed noise injection in SGD in terms of compressibility and the train/test performance. We consider a single-hidden-layer neural network with ReLU activations and the cross entropy loss, applied on classifications tasks. We chose the Electrocardiogram (ECG) dataset [YE] and the MNIST datasets. By slightly streching the scope of our theoretical framework, we also train the weights of the second layer instead of fixing them to $1/n$. All the experimentation details are given in Appendix D and we present additional experimental results in Appendix E.

α	Train Acc.	Test Acc.	Pruning Ratio	Train Acc. a.p.	Test Acc. a.p.
no noise	0.974	0.957	11.45	0.97	0.954
1.75	0.97 ± 0.007	0.955 ± 0.003	48.07 ± 7.036	0.944 ± 0.03	0.937 ± 0.022
1.8	0.97 ± 0.007	0.955 ± 0.003	44.68 ± 5.4	0.95 ± 0.025	0.963 ± 0.016
1.9	0.966 ± 0.008	0.959 ± 0.01	39.37 ± 2.57	0.962 ± 0.012	0.953 ± 0.005

Table 1: ECG5000, Type-I noise, $n = 2$K.

In our first experiment, we consider the ECG500 dataset and choose the Type-I noise. Our goal is to investigate the effects α and n over the performance. Tables 1-2 illustrate the results. Here, for different cases, we monitor the training and test accuracies (over 1.00), the pruning ratio: the percentage of the weight matrix that can be pruned while keeping the 90% of the norm of the original matrix[2], and training/test accuracies after pruning (a.p.) the network with the specified pruning ratio.

The results show that, even for a moderate number of neurons $n = 2$K, the heavy-tailed noise results in a significant improvement in the compression capability of the neural network. For $\alpha = 1.9$, we can see that the pruning ratio increases to 39%, whereas vanilla SGD can only be compressible with a rate 11%. Besides, the compromise in the test accuracy is almost negligible, the proposed approach achieves 95.3%, whereas vanilla SGD achieves 95.7% accuracy. We also observe that decreasing α (i.e., increasing the heaviness of the tails) results in a better compression rate; yet, there is a tradeoff between this rate and the test performance. In Table 2, we repeat the same experiment for $n = 10$K. We observe that the previous conclusions become even clearer in this case, as our theory applies to large n. For the case where $\alpha = 1.75$, we obtain a pruning ratio of 52% with test accuracy 95.4%, whereas for vanilla SGD the ratio is only 11% and the original test accuracy is 96.3%.

In our second experiment, we investigate the impact of the noise type. We set $n = 10$K and use the same setting as in Table 2. Tables 3-4 illustrate the results. We observe that the choice of the

[2]The pruning ratio has the same role of κ, whereas we fix the compression error to 0.1 and find the largest κ that satisfies this error threshold.

α	Train Acc.	Test Acc.	Pruning Ratio	Train Acc. a.p.	Test Acc. a.p.
no noise	0.978	0.963	11.46	0.978	0.964
1.75	0.978 ± 0.001	0.964 ± 0.001	52.59 ± 6.55	0.95 ± 0.03	0.954 ± 0.022
1.8	0.978 ± 0.001	0.964 ± 0.001	52.59 ± 6.55	0.95 ± 0.03	0.954 ± 0.022
1.9	0.978 ± 0.001	0.964 ± 0.001	40.85 ± 2.89	0.96 ± 0.021	0.958 ± 0.013

Table 2: ECG5000, Type-I noise, $n = 10$K.

noise type can make a significant difference in terms of both compressibility and accuracy. While the Type-III noise seems to obtain a similar accuracy when compared to Type-I, it achieves a worse compression rate. On the other hand, the behavior of Type-II noise is perhaps more interesting: for $\alpha = 1.9$ it both increases compressibility and also achieves a better accuracy when compared to unpruned, vanilla SGD. However, we see that its behavior is much more volatile, the performance quickly degrades as we decrease α. From these comparisons, Type-I noise seems to achieve a better tradeoff.

α	Train Acc.	Test Acc.	Pruning Ratio	Train Acc. a.p.	Test Acc. a.p.
1.75	0.986 ± 0.003	0.982 ± 0.005	52.13 ± 27.78	0.865 ± 0.261	0.866 ± 0.251
1.8	0.985 ± 0.003	0.980 ± 0.005	39.9 ± 21.55	0.971 ± 0.025	0.972 ± 0.023
1.9	0.982 ± 0.003	0.976 ± 0.006	20.95 ± 6.137	0.982 ± 0.004	0.977 ± 0.006

Table 3: ECG5000, Type-II noise, $n = 10$K.

α	Train Acc.	Test Acc.	Pruning Ratio	Train Acc. a.p.	Test Acc. a.p.
1.75	0.97 ± 0.007	0.957 ± 0.005	33.48 ± 7.33	0.969 ± 0.008	0.957 ± 0.011
1.8	0.97 ± 0.007	0.956 ± 0.007	26.81 ± 4.72	0.963 ± 0.008	0.952 ± 0.008
1.9	0.97 ± 0.005	0.955 ± 0.005	17.59 ± 1.56	0.968 ± 0.004	0.954 ± 0.96

Table 4: ECG5000, Type-III noise, $n = 10$K.

In our next experiment, we consider the MNIST dataset, set $n = 10$K and use Type-I noise. Table 5 illustrates the results. Similar to the previous results, we observe that the injected noise has a visible benefit on compressibility. When $\alpha = 1.9$, our approach doubles the compressibility of the vanilla SGD (from 10% to 21%), whereas the training and test accuracies almost remain unchanged. On the other hand, when we decrease α, we observe that the pruning ratio goes up to 44%, while only compromising 1% of test accuracy. To further illustrate this result, we pruned vanilla SGD by using this pruning ratio (44%). In this case, the test accuracy of SGD drops down to 92%, where as our simple noising scheme achieves 94% of test accuracy with the same pruning ratio.

α	Train Acc.	Test Acc.	Pruning Ratio	Train Acc. a.p.	Test Acc. a.p.
no noise	0.95	0.9487	10.59	0.9479	0.9476
1.75	0.95 ± 0.0001	0.9454 ± 0.0005	44.42 ± 7.16	0.944 ± 0.0025	0.9409 ± 0.0019
1.8	0.95 ± 0.0001	0.9457 ± 0.0007	34.49 ± 5.07	0.9453 ± 0.0015	0.9397 ± 0.0036
1.9	0.95 ± 0.0001	0.9463 ± 0.0004	21.31 ± 1081	0.9478 ± 0.0008	0.9444 ± 0.0009

Table 5: MNIST, Type-I noise, $n = 10$K until reaching 95% training accuracy.

5 Conclusion

We provided a methodological and theoretical framework for provably obtaining compressibility in mean-field neural networks. Our approach requires minimal modification for vanilla SGD and has the same computational complexity. We showed that the resulting algorithm is guaranteed to provide compressible parameters. We illustrated our approach on several experiments, where we showed that, in most cases, the proposed approach achieves a high compressibility ratio, while slightly compromising from the accuracy.

The limitations of our approach are as follows: (i) we consider mean-field networks, it would be of interest to generalize our results to more sophisticated architectures, (ii) we focused on the compressibility; yet, the noise injection also has an effect on the train/test accuracy. Finally, due to the theoretical nature of our paper, it does not have a direct negative social impact.

Acknowledgements

The authors thank Alain Durmus for fruitful discussions. U.S. is partially supported by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA0001 (PRAIRIE 3IA Institute) and by the European Research Council Starting Grant DYNASTY – 101039676.

References

[ACA19] Caglar Aytekin, Francesco Cricri, and Emre Aksu. Compressibility loss for neural network weights. *arXiv preprint arXiv:1905.01044*, 2019.

[AGNZ18] Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 254–263. PMLR, 10–15 Jul 2018.

[AUM11] Arash Amini, Michael Unser, and Farokh Marvasti. Compressibility of deterministic and random infinite sequences. *IEEE Transactions on Signal Processing*, 59(11):5193–5201, 2011.

[BLMG21] Rebekka Burkholz, Nilanjana Laha, Rajarshi Mukherjee, and Alkis Gotovos. On the existence of universal lottery tickets. *arXiv preprint arXiv:2111.11146*, 2021.

[BOFG20] Davis Blalock, Jose Javier Gonzalez Ortiz, Jonathan Frankle, and John Guttag. What is the State of Neural Network Pruning? *arXiv:2003.03033 [cs, stat]*, March 2020.

[BSE$^+$21] Melih Barsbey, Milad Sefidgaran, Murat A Erdogdu, Gael Richard, and Umut Simsekli. Heavy tails in SGD and compressibility of overparametrized neural networks. *Advances in Neural Information Processing Systems*, 34:29364–29378, 2021.

[Cav23] Thomas Cavallazzi. Well-posedness and propagation of chaos for Lévy-driven Mckean-Vlasov SDEs under Lipschitz assumptions, 2023.

[CB18] Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in neural information processing systems*, 31, 2018.

[CJS$^+$20] Tianyi Chen, Bo Ji, Yixin Shi, Tianyu Ding, Biyi Fang, Sheng Yi, and Xiao Tu. Neural network compression via sparse optimization. *arXiv preprint arXiv:2011.04868*, 2020.

[CWZ$^+$21] Alexander Camuto, Xiaoyu Wang, Lingjiong Zhu, Chris Holmes, Mert Gürbüzbalaban, and Umut Şimşekli. Asymmetric heavy tails and implicit bias in gaussian noise injections. In *ICML*, 2021.

[DBDFS20] Valentin De Bortoli, Alain Durmus, Xavier Fontaine, and Umut Simsekli. Quantitative propagation of chaos for SGD in wide neural networks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, 2020.

[dCNV22] Arthur da Cunha, Emanuele Natale, and Laurent Viennot. Proving the strong lottery ticket hypothesis for convolutional neural networks. In *ICLR 2022-10th International Conference on Learning Representations*, 2022.

[FC18] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*, 2018.

[GCD12] Rémi Gribonval, Volkan Cevher, and Mike E. Davies. Compressible distributions for high-dimensional statistics. *IEEE Transactions on Information Theory*, 58(8):5016–5034, 2012.

[gne54] *Limit Distributions for Sums of Independent Random Variables*. Cambridge, Addison-Wesley, 1954.

[HJTW21] Daniel Hsu, Ziwei Ji, Matus Telgarsky, and Lan Wang. Generalization bounds via distillation. In *International Conference on Learning Representations*, 2021.

[HMW21] Lu-Jing Huang, Mateusz B Majka, and Jian Wang. Approximation of heavy-tailed distributions via stable-driven SDEs. 2021.

[JŠS19] Jean-François Jabir, David Šiška, and Łukasz Szpruch. Mean-field neural ODEs via relaxed optimal control. *arXiv preprint arXiv:1912.05475*, 2019.

[KW23] William Kengne and Modou Wade. Sparse-penalized deep neural networks estimator under weak dependence. *arXiv preprint arXiv:2303.01406*, 2023.

[LAJ+22] Hoil Lee, Fadhel Ayed, Paul Jung, Juho Lee, Hongseok Yang, and François Caron. Deep neural networks with dependent weights: Gaussian process mixture limit, heavy tails, sparsity and compressibility. *arXiv preprint arXiv:2205.08187*, 2022.

[Led23] Johannes Lederer. Statistical guarantees for sparse deep learning. *AStA Advances in Statistical Analysis*, pages 1–28, 2023.

[MM19] Charles H. Martin and Michael W. Mahoney. Traditional and Heavy-Tailed Self Regularization in Neural Network Models. *arXiv:1901.08276 [cs, stat]*, January 2019.

[MMM19] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Conference on Learning Theory*, pages 2388–2464. PMLR, 2019.

[MMN18] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

[MMR+17] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017*, volume 54, pages 1273–1282, 2017.

[MYSSS20] Eran Malach, Gilad Yehudai, Shai Shalev-Schwartz, and Ohad Shamir. Proving the lottery ticket hypothesis: Pruning is all you need. In *International Conference on Machine Learning*, pages 6682–6691. PMLR, 2020.

[NSR19] Than Huy Nguyen, Umut Simsekli, and Gaël Richard. Non-asymptotic analysis of fractional Langevin Monte Carlo for non-convex optimization. In *International Conference on Machine Learning*, pages 4810–4819. PMLR, 2019.

[O'N20] James O'Neill. An overview of neural network compression. *arXiv:2006.03669 [cs, stat]*, August 2020.

[PRSE18] Vardan Papyan, Yaniv Romano, Jeremias Sulam, and Michael Elad. Theoretical foundations of deep learning via sparse representations: A multilayer sparse model and its connection to convolutional neural networks. *IEEE Signal Processing Magazine*, 35(4):72–89, 2018.

[RM17] Daniel Ramage and Brendan McMahan. Federated learning: Collaborative machine learning without centralized training data, 2017.

[RVE18] Grant M Rotskoff and Eric Vanden-Eijnden. Trainability and accuracy of neural networks: An interacting particle system approach. *arXiv preprint arXiv:1805.00915*, 2018.

[Sam17] G. Samoradnitsky. *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance*. CRC Press, 2017.

[SAM+20] Taiji Suzuki, Hiroshi Abe, Tomoya Murata, Shingo Horiuchi, Kotaro Ito, Tokuma Wachi, So Hirai, Masatoshi Yukishima, and Tomoaki Nishimura. Spectral pruning: Compressing deep neural networks via spectral analysis and its generalization error. In *International Joint Conference on Artificial Intelligence*, pages 2839–2846, 2020.

[SAN20] Taiji Suzuki, Hiroshi Abe, and Tomoaki Nishimura. Compression based bound for non-compressed network: unified generalization error analysis of large compressible deep neural network. In *International Conference on Learning Representations*, 2020.

[ŞGN+19] 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:1912.00018 [cs, math, stat]*, November 2019.

[SGRS22] Milad Sefidgaran, Amin Gohari, Gael Richard, and Umut Simsekli. Rate-distortion theoretic generalization bounds for stochastic learning algorithms. In *Conference on Learning Theory*, pages 4416–4463. PMLR, 2022.

[Shi21] John Y Shin. Compressing heavy-tailed weight matrices for non-vacuous generalization bounds. *arXiv preprint arXiv:2105.11025*, 2021.

[Şim17] Umut Şimşekli. Fractional Langevin Monte Carlo: Exploring Lévy driven stochastic differential equations for Markov chain Monte Carlo. In *International Conference on Machine Learning*, pages 3200–3209. PMLR, 2017.

[SS22] Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of deep neural networks. *Mathematics of Operations Research*, 47(1):120–152, 2022.

[SSG19] Umut Simsekli, Levent Sagun, and Mert Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning*, pages 5827–5837. PMLR, 2019.

[ŞZTG20] Umut Şimşekli, Lingjiong Zhu, Yee Whye Teh, and Mert Gürbüzbalaban. Fractional underdamped langevin dynamics: Retargeting sgd with momentum under heavy-tailed gradient noise. In *ICML*, 2020.

[YE] Chen Yanping and Keogh Eamonn. ECG5000. `http://www.timeseriesclassification.com/description.php?Dataset=ECG5000`. data retrieved from World Development Indicators.

[ZFM+20] 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.

[ZKV+20] 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 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.

[ZZ23] Xiaolong Zhang and Xicheng Zhang. Ergodicity of supercritical SDEs driven by α-stable processes and heavy-tailed sampling. *Bernoulli*, 29(3):1933–1958, 2023.

Implicit Compressibility of Overparametrized Neural Networks via Heavy-Tailed Noisy Gradient Descent

APPENDIX

The Appendix is organized as follows.

- In Section A, we point out the intermediate results established for proving 3.1.
- In Section B, we provide technical lemmas for proving Theorem 3.1, Theorem A.1 and Theorem A.2.
- In Section C, we give the proofs to the theorems in the main paper.
- In Section D and E, we present experimental details and results of additional experiments.
- In Section F, implications of our compressibility studies on federated learning are discussed.

A Intermediate Results

In this section, we gather the main technical contributions with the purpose of demonstrating Theorem 3.1. We begin by rewriting (2) in the following form:

$$\begin{cases} \hat{\theta}^{i,n}_{k+1} - \hat{\theta}^{i,n}_{k} = \eta b(\hat{\theta}^{i,n}_{k}, \hat{\mu}^n_k) + \sigma \cdot \eta^{1/\alpha} X^{i,n}_k \\ \hat{\theta}^{i,n}_0 \sim \mu_0 \in \mathcal{P}(\mathbb{R}^d), \end{cases} \tag{3}$$

where $\hat{\mu}^n_k = \frac{1}{n}\delta_{\hat{\theta}^{i,n}_k}$ is the empirical distribution of parameters at iteration k and δ is the Dirac measure, and the drift is given by $b(\theta^{i,n}_k, \mu^n_k) = -\mathbb{E}[\partial_1 \ell(\mu^n_k(h_x(\cdot), y)\nabla h_x(\theta^{i,n}_k)]$, where ∂_1 denotes the partial derivative with respect to the first parameter and

$$\mu^n_k(h_x(\cdot)) := \int h_x(\theta) \mathrm{d}\mu^n_k(\theta) = \sum_{i=1}^n h_x(\theta^{i,n}_k) = f_{\Theta^n_k}(x).$$

It is easy to check that $b(\theta^{i,n}_k, \mu^n_k) = -n\partial_{\theta^{i,n}} R(\Theta^n_k)$. By looking at the dynamics from this perspective, we can treat the evolution of the parameters as a system of evolving probability distributions μ^n_k: the empirical distribution of the parameters during the training process will converge to a limit as η goes to 0 and n goes to infinity.

We start by linking the recursion (2) to its limiting case where $\eta \to 0$. The limiting dynamics can be described by the following system of SDEs:

$$\begin{cases} \mathrm{d}\theta^{i,n}_t = b(\theta^{i,n}_t, \mu^n_t)\mathrm{d}t + \sigma \mathrm{d}\mathrm{L}^{i,n}_t \\ \theta^{i,n}_0 \sim \mu_0 \in \mathcal{P}(\mathbb{R}^d), \end{cases} \tag{4}$$

where $\mu^n_t = \frac{1}{n}\delta_{\theta^{i,n}_t}$ and $(\mathrm{L}^{i,n}_t)_{t\geq 0}$ are independent α-stable processes such that $\mathrm{L}^{i,n}_1 \overset{(\mathrm{d})}{=} X^{i,n}_1$. We can now see the original recursion (2) as an Euler discretization of (4) and then we have the following strong uniform error estimate for the discretization.

Theorem A.1. *Let $(\theta^{i,n}_t)_{t\geq 0}$ be the solutions to SDE (4) and $(\hat{\theta}^{i,n}_k)_{k\in\mathbb{N}_+}$ be given by SGD (2) with the same initial condition $\theta^{i,n}_0$ and α-stable Lévy noise $\mathrm{L}^{i,n}_\cdot$, i=1,…,n. Under Assumption 1, for any $T > 0$, if $\eta k \leq T$, there exists a constant C depending on B, T, α such that the approximation error*

$$\mathbb{E}\left[\sup_{i\leq n} \|\theta^{i,n}_{\eta k} - \hat{\theta}^{i,n}_k\|\right] \leq C(\eta n)^{1/\alpha}.$$

In comparison to the standard error estimates in the Euler-Maruyama scheme concerning only the stepsize η, the additional n-dependence is due to the fact that here we consider the supremum of the approximation error over all $i \leq n$, which involves the expectation of the supremum of the modulus of n independent α-stable random variables.

Next, we start from the system (4) and consider the case where $n \to \infty$. In this limit, we obtain the following McKean-Vlasov-type stochastic differential equation:

$$\begin{cases} \mathrm{d}\theta_t^\infty = b(\theta_t^\infty, [\theta_t^\infty])\mathrm{d}t + \mathrm{dL}_t \\ [\theta_0^\infty] = \mu \in \mathcal{P}(\mathbb{R}^d), \end{cases} \tag{5}$$

where $(\mathrm{L}_t)_{t\geq 0}$ is an α-stable process and $[\theta_t^\infty]$ denotes the distribution of θ_t^∞. The existence and uniqueness of a strong solution to (5) are given in [Cav23]. Moreover, for any positive T, $\mathbb{E}\left[\sup_{t\leq T}\|\theta_t^\infty\|^\alpha\right] < +\infty$. This SDE with measure-dependent coefficients turns out to be a useful mechanism for analyzing the behavior of neural networks and provides insights into the effects of noise on the learning dynamics.

In this step, we will link the system (4) to its limit (5), which is a strong uniform propagation of chaos result for the weights. The next result shows that, when n is sufficiently large, the trajectories of weights asymptotically behave as i.i.d. solutions to (5).

Theorem A.2. *Following the existence and uniqueness of strong solutions to* (4) *and* (5)*, let* $(\theta_t^{i,\infty})_{t\geq 0}$ *be solutions to the McKean-Vlasov equation* (5) *and* $(\theta_t^{i,n})_{t\geq 0}$ *be solutions to* (4) *associated with same realization of* α*-stable processes* $(\mathrm{L}_t^i)_{t\geq 0}$ *for each* i*. Suppose that* $(\mathrm{L}_t^i)_{t\geq 0}$ *are independent. Then there exists* C *depending on* T, B *such that*

$$\mathbb{E}\left[\sup_{t\leq T}\sup_{i\leq n}|\theta_t^{i,n} - \theta_t^{i,\infty}|\right] \leq \frac{C}{\sqrt{n}}$$

Our result differs from the existing literature by taking the supremum over the indices i before taking the expectation, which is obviously stronger than taking the supremum over i outside the expectation. It is also worth mentioning that the $O(n^{-1/2})$ decreasing rate here is better, if $\alpha < 2$, than the state of the art [Cav23] with classical Lipschitz assumptions on the coefficients of SDEs. The reason is that here, thanks to Assumption 1, we can take into account the specific structure of the one-hidden layer neural networks.

Finally, we are interested in the distributional properties of the McKean-Vlasov equation (5). The following result establishes that the marginal distributions of (5) will have diverging second-order moments, hence, they will be heavy-tailed.

Theorem A.3. *Let* $(\mathrm{L}_t)_{t\geq 0}$ *be an* α*-stable process. For any time* t*, let* θ_t *be the solution to* (5) *with initialization* θ_0 *which is independent of* $(\mathrm{L}_t)_{t\geq 0}$ *such that* $\mathbb{E}\left[\|\theta_0\|\right] < \infty$*, then the following holds*

$$\mathbb{E}\left[\|\theta_t^\infty\|^2\right] = +\infty.$$

We remark that the result is weak in the sense that details on the tails of θ_t with respect to α and t are implicit. However, it renders sufficient for our compressibility result in Theorem 3.1.

Now, having proved all the necessary ingredients, Theorem 3.1 is obtained by accumulating the error bounds proven in Theorems A.1 and A.2, and applying [GCD12, Proposition 1] along with Theorem A.3.

B Technical Lemmas

Lemma B.1. *Under Assumption 1,*

$$\|b(\theta_1,\mu_1) - b(\theta_2,\mu_2)\| \leq B\cdot\left(\|\theta_1-\theta_2\| + \mathbb{E}_{x\sim\pi}\left[|\mu_1(h_x(\cdot)) - \mu_2(h_x(\cdot))|^2\right]^{\frac{1}{2}}\right).$$

Moreover, $\|b(\cdot,\cdot)\|_\infty \leq B$*, and if* $\mu_1 = \frac{1}{n}\sum_{i=1}^n \delta_{\theta_1^i}$*,* $\mu_2 = \frac{1}{n}\sum_{i=1}^n \delta_{\theta_2^i}$*,*

$$\|b(\theta_1,\mu_1) - b(\theta_2,\mu_2)\| \leq B\|\theta_1-\theta_2\| + \frac{B}{n}\sum_{i=1}^n \|\theta_1^i - \theta_2^i\|.$$

Proof. Recall that

$$b(\theta,\mu) = -\mathbb{E}\left[\partial_1 l(\mu(h_x(\cdot)), y)\nabla h_x(\theta)\right].$$

Then it follows from triangular inequality that

$$\|b(\theta_1,\mu_1)-b(\theta_2,\mu_2)\| \le \|b(\theta_1,\mu_1)-b(\theta_2,\mu_1)\| + \|b(\theta_2,\mu_1)-b(\theta_2,\mu_2)\| \tag{6}$$

The first term is upper bounded by

$$\begin{aligned}\|b(\theta_1,\mu_1)-b(\theta_2,\mu_1)\| &\le \mathbb{E}\left[\|\partial_1 l(\cdot,y)\|_\infty \cdot \|\nabla^2 h_x\|_\infty\right] \cdot \|\theta_2-\theta_1\| \\ &\le \mathbb{E}\left[\Phi(y)\Psi(x)\right] \cdot \|\theta_1-\theta_2\| \\ &\le \left(\mathbb{E}\left[\Phi^2(y)\Psi^2(x)\right]\right)^{1/2} \cdot \|\theta_1-\theta_2\| \\ &\le B \cdot \|\theta_1-\theta_2\|\end{aligned} \tag{7}$$

The second term is upper bounded by

$$\begin{aligned}\|b(\theta_2,\mu_1)-b(\theta_2,\mu_2)\| &\le \mathbb{E}\left[\|\partial_1^2 l(\cdot,y)\|_\infty \cdot \|\nabla h_x(\cdot)\|_\infty \cdot |\mu_1(h_x(\cdot))-\mu_2(h_x(\cdot))|\right] \\ &\le \left(\mathbb{E}\left[\Phi^2(y)\Psi^2(x)\right]\right)^{1/2} \mathbb{E}\left[|\mu_1(h_x(\cdot))-\mu_2(h_x(\cdot))|^2\right]^{1/2} \\ &\le B \cdot \mathbb{E}\left[|\mu_1(h_x(\cdot))-\mu_2(h_x(\cdot))|^2\right]^{1/2}\end{aligned} \tag{8}$$

We conclude the first inequality by combining (6), (7) and (8).

For the boundedness of b in the norm infinity, it is not hard to observe that

$$b(\theta,\mu) = -\mathbb{E}[\partial_1 l(\mu(h_x(\cdot)),y)\nabla h_x(\theta)] \le \mathbb{E}\left[\Phi(y)\Psi(x)\right] \le B.$$

For the last one, it follows from the first bound and Cauchy-Schwarz inequality that

$$\begin{aligned}\|b(\theta_1,\mu_1)-b(\theta_2,\mu_2)\| &\le B\|\theta_1-\theta_2\| + \frac{1}{n}\mathbb{E}_{x\sim\pi}\left[\left(\sum_{i=1}^n h_x(\theta_1^i)-h_x(\theta_2^i)\right)^2\right]^{1/2} \\ &\le B\|\theta_1-\theta_2\| + \frac{1}{n}\mathbb{E}_{x\sim\pi}\left[\|\nabla h_x\|_\infty\left(\sum_{i=1}^n \|\theta_1^i-\theta_2^i\|\right)^2\right]^{1/2} \\ &\le B\|\theta_1-\theta_2\| + \frac{1}{n}\mathbb{E}_{x\sim\pi}[\Psi^2(x)]^{1/2} \cdot \sum_{i=1}^n \|\theta_1^i-\theta_2^i\| \\ &\le B\|\theta_1-\theta_2\| + \frac{B}{n}\sum_{i=1}^n \|\theta_1^i-\theta_2^i\|.\end{aligned}$$

Then the proof is completed. □

B.1 Propagation of Chaos

Lemma B.2. *Let $(\mathrm{L}_t)_{t\ge 0}$ be an α-stable Lévy process and let $(\mathcal{F}_t)_{t\ge 0}$ be the filtration generated by $(\mathrm{L}_t)_{t\ge 0}$. Then under Assumption 1, given the initial condition $X_0=\xi$, there exists a unique adapted process $(X_t)_{t\in[0,T]}$ for all integrable datum $\xi \in L^1(\mathbb{R}^p)$ such that*

$$X_t = \xi + \int_0^t b(X_t,[X_t])\mathrm{d}t + \mathrm{L}_t.$$

Moreover the first moment of the supremum of the process is bounded

$$\mathbb{E}\left[\sup_{t\le T}\|X_t\|\right] < +\infty.$$

Proof. It follows from Theorem 1 in [Cav23] by Lemma B.1 where β is taken to be 1. □

B.2 Compression

Lemma B.3. *Consider a non-integrable probability distribution μ taking values in $\mathbb{R}_+$ such that $\mathbb{E}_{X\sim\mu}[X] = +\infty$. Let $X_1, \dots, X_n$ be n i.i.d. copies distributed according to μ. Then for any C positive,*

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} X_i \leq C\right] \xrightarrow{n\to\infty} 0.$$

Proof. Using the assumption that μ is non-integrable, let K be a cutoff level for μ such that

$$\mathbb{E}_{X\sim\mu}[\max(X, K)] = C + 1.$$

Therefore by the law of large numbers, when goes to infinity,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \max(X_i, K) = C + 1 \qquad \text{almost surely.}$$

To conclude, we remark that

$$\frac{1}{n}\liminf_{n\to\infty}\sum_{i=1}^{n} X_i \geq \frac{1}{n}\lim_{n\to\infty}\sum_{i=1}^{n}\max(X_i, K),$$

which is lower bounded by $C + 1$ almost surely. Thus the probability that $\frac{1}{n}\sum_{i=1}^{n} X_i \leq C$ goes to 0 when n goes to infinity. $\square$

C Proofs

C.1 Proof of Theorem A.3

Proof. Recall that $\theta_t = \theta_0 + \int_0^t b(\theta_s, [\theta_s])\mathrm{d}s + \mathrm{L}_t$, then

$$\begin{aligned}
\mathbb{E}\left[\|\theta_t\|^2\right] &= \mathbb{E}\left[\left\langle \theta_0 + \int_0^t b(\theta_s, [\theta_s])\mathrm{d}s + \mathrm{L}_t, \theta_0 + \int_0^t b(\theta_s, [\theta_s])\mathrm{d}s + \mathrm{L}_t\right\rangle\right] \\
&= \mathbb{E}\left[\left\|\theta_0 + \int_0^t b(\theta_s, [\theta_s])\mathrm{d}s\right\|^2\right] + 2\mathbb{E}\left[\left\langle \theta_0 + \int_0^t b(\theta_s, [\theta_s])\mathrm{d}s, \mathrm{L}_t\right\rangle\right] + \mathbb{E}\left[\|\mathrm{L}_t\|^2\right] \\
&\geq \mathbb{E}\left[\|\mathrm{L}_t\|^2\right] - 2\mathbb{E}\left[\|\theta_0\| \cdot \|\mathrm{L}_t\|\right] - 2\mathbb{E}\left[t\|b(\cdot)\|_\infty \cdot \|\mathrm{L}_t\|\right] \\
&\geq \mathbb{E}\left[\|\mathrm{L}_t\|^2\right] - 2\mathbb{E}\left[\|\theta_0\|\right]\mathbb{E}\left[\|\mathrm{L}_t\|\right] - 2Bt \cdot \mathbb{E}\left[\cdot\|\mathrm{L}_t\|\right],
\end{aligned}$$

where the last relation follows from the independence between the initialization θ_0 and the difusion noise $(\mathrm{L}_t)_{t\geq 0}$ and Lemma B.1. The proof is completed by noting that

$$\mathbb{E}\left[\|\mathrm{L}_t\|^2\right] = \infty \quad \text{and } \mathbb{E}\left[\|\theta_0\|\right], \mathbb{E}\left[\|\mathrm{L}_t\|\right] < \infty.$$

$\square$

C.2 Proof of Theorem A.2

Proof. By identification of the diffusion process $(\mathrm{L}_t^{i,n})_{t\geq 0}$ in (4) and (5), the difference of their solutions $\theta_t^{i,n}$ and $\theta_t^{i,\infty}$ for all $t \in [0, T]$ satisfies

$$\theta_t^{i,n} - \theta_t^{i,\infty} = \int_0^t [b(\theta_s^{i,n}, \mu_s^n) - b(\theta_s^{i,\infty}, [\theta_s^{i,\infty}])]\mathrm{d}s,$$

where $\mu_t = \frac{1}{n}\sum_{i=1}^{n} \delta_{\theta_t^{i,n}}$ and $[\theta_t^{i,\infty}]$ denotes the distribution of $\theta_t^{i,\infty}$. Using Lemma B.1,

$$\begin{aligned}
&\|\theta_t^{i,n} - \theta_t^{i,\infty}\| \\
\leq& B\int_0^t \|\theta_s^{i,n} - \theta_s^{i,\infty}\|\mathrm{d}s + B\int_0^t \mathbb{E}_{x\sim\pi}\left[|\mu_s^n(h_x(\cdot)) - [\theta_s^{i,\infty}](h_x(\cdot))|^2\right]^{1/2}\mathrm{d}s \\
\leq& B\int_0^t \sup_{i\leq n}\|\theta_s^{i,n} - \theta_s^{i,\infty}\|\mathrm{d}s + B\int_0^t \mathbb{E}_{x\sim\pi}\left[|\mu_s^n(h_x(\cdot)) - \bar{\mu}_s^n(h_x(\cdot))|^2\right]^{1/2}\mathrm{d}s \\
&+ B\int_0^t \mathbb{E}_{x\sim\pi}\left[|\bar{\mu}_s^n(h_x(\cdot)) - [\theta_s^{i,\infty}](h_x(\cdot))|^2\right]^{1/2}\mathrm{d}s
\end{aligned} \tag{9}$$

where $\bar{\mu}_s^n := \frac{1}{n}\sum_{i=1}^{n}\delta_{\theta_s^{i,\infty}}$, the empirical measure of $\theta_s^{i,\infty}$ for $i = 1, \ldots, n$. the last inequality follows from Cauchy-Schwarz inequality. Moreover we have

$$\begin{aligned}
\mathbb{E}_{x\sim\pi}[|\mu_s^n(h_x(\cdot) - \bar{\mu}_s^n(h_x(\cdot))|^2]^{1/2} &\leq \mathbb{E}_{x\sim\pi}\left[\left|\frac{\|\nabla h_x\|_\infty}{n}\sum_{i=1}^{n}\|\theta_s^{i,n} - \theta_s^{i,\infty}\|\right|^2\right]^{1/2} \\
&\leq \mathbb{E}_{x\sim\pi}[\Psi^2(x)]^{1/2}\cdot\frac{1}{n}\sum_{i=1}^{n}\|\theta_s^{i,n} - \theta_s^{i,\infty}\| \\
&\leq B\sup_{i\leq n}\|\theta_s^{i,n} - \theta_s^{i,\infty}\|.
\end{aligned}$$

Plug the estimate above into (9):

$$\begin{aligned}
\|\theta_t^{i,n} - \theta_t^{i,\infty}\| \leq& B(1+B)\int_0^t \sup_{i\leq n}\|\theta_s^{i,n} - \theta_s^{i,\infty}\|\mathrm{d}s \\
&+ B\int_0^t \mathbb{E}_{x\sim\pi}\left[|\bar{\mu}_s^n(h_x(\cdot)) - [\theta_s^{i,\infty}](h_x(\cdot))|^2\right]^{1/2}\mathrm{d}s
\end{aligned} \tag{10}$$

By taking the supremum over $i = 1, \ldots, n$ and t, and using the fact that

$$\sup\int_\cdot(\cdot) \leq \int_\cdot \sup(\cdot),$$

we arrive at

$$\begin{aligned}
\sup_{t\leq T}\sup_{i\leq n}\|\theta_t^{i,n} - \theta_t^{i,\infty}\| \leq& B(1+B)\int_0^T \sup_{t\leq s}\sup_{i\leq n}\|\theta_t^{i,n} - \theta_t^{i,\infty}\|\mathrm{d}s \\
&+ B\int_0^t \mathbb{E}_{x\sim\pi}\left[|\bar{\mu}_s^n(h_x(\cdot)) - [\theta_s^{i,\infty}](h_x(\cdot))|^2\right]^{1/2}\mathrm{d}s
\end{aligned} \tag{11}$$

Let us now estimate $\mathbb{E}\left[|\bar{\mu}_s^n(h_x(\cdot)) - [\theta_s^{i,\infty}](h_x(\cdot))|^2|x\right]^{1/n}$, the expectation under the stable diffusion, rather than the expectation over the data distribution, where the $1/\sqrt{n}$ convergence rate comes from. In deed for fixed x, $h_x(\theta_s^{i,\infty})$, $i = 1, \ldots, n$ are bounded i.i.d. random variables with mean $[\theta_s^{i,\infty}](h_x(\cdot))$. Therefore

$$\begin{aligned}
\mathbb{E}\left[|\bar{\mu}_s^n(h_x(\cdot)) - [\theta_s^{i,\infty}](h_x(\cdot))|^2|x\right]^{1/2} &= \mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n}h_x(\theta_s^{i,\infty}) - [\theta_s^{i,\infty}](h_x(\cdot))\right|^2 \middle| x\right]^{1/2} \\
&\leq \frac{1}{\sqrt{n}}\|h_x(\cdot)\|_\infty \leq \frac{\Psi(x)}{\sqrt{n}}.
\end{aligned} \tag{12}$$

Finally, combining (11), (12), the integrability condition Lemma B.2 and using Fubini's theorem, we arrive at

$$\mathbb{E}\left[\sup_{r\leq t}\sup_{i\leq n}\|\theta_r^{i,n} - \theta_r^{i,\infty}\|\right] \leq B(1+B)\int_0^t \mathbb{E}\left[\sup_{r\leq s}\sup_{i\leq n}\|\theta_r^{i,n} - \theta_r^{i,\infty}\|\right]\mathrm{d}s + \frac{Bt\mathbb{E}_{x\sim\pi}[\Psi(x)]}{\sqrt{n}}.$$

We conclude by Gronwall's inequality that

$$\mathbb{E}\left[\sup_{t\le T}\sup_{i\le n}\|\theta_t^{i,n}-\theta_t^{i,\infty}\|\right]\le(1+B)\left(\frac{BT}{\sqrt{n}}+\frac{B^2T^2\exp(BT(1+\mathbb{E}_{x\sim\pi}[\Psi(x)]))}{2\sqrt{n}}\right).$$

Then the proof of Theorem A.2 is completed. □

C.3 Proof of Theorem A.1

Proof. Similarly as in the proof above, we have

$$\begin{aligned}\sup_{i\le n}\|\theta_\eta^{i,n}-\hat\theta_1^{i,n}\| &\le \sup_{i\le n}\int_0^\eta \|b(\theta_t^{i,n},\mu_t^{i,n})-b(\hat\theta_0^{i,n},\mu_0^{i,n})\|\mathrm{d}t\\ &\le B\int_0^\eta \sup_{i\le n}\|\theta_t^{i,n}-\theta_0^{i,n}\|+\frac{1}{n}\sum_{j=1}^n\|\theta_t^{j,n}-\theta_0^{j,n}\|\mathrm{d}t\\ &\le B\int_0^\eta 2\|b\|_\infty\cdot t+\sup_{i\le n}\|\mathrm{L}_t^{i,n}\|+\frac{1}{n}\sum_{j=1}^n\|\mathrm{L}_t^{j,n}\|\mathrm{d}t\end{aligned}$$

Recall that $\|b\|_\infty\le B$, therefore by taking the expectation and the scaling of the stable process $\mathrm{L}_t^{i,n}$,

$$\begin{aligned}\mathbb{E}\left[\sup_{i\le n}\|\theta_\eta^{i,n}-\hat\theta_1^{i,n}\|\right] &\le B\int_0^\eta 2Bt+2t^{1/\alpha}\cdot\mathbb{E}\left[\sup_{i\le n}\|\mathrm{L}_1^{i,n}\|+\frac{1}{n}\sum_{j=1}^n\|\mathrm{L}_1^{j,n}\|\right]\mathrm{d}t\\ &\le B^2\eta^2+\frac{B\alpha\cdot\mathbb{E}\left[\sup_{i\le n}\|\mathrm{L}_1^{i,n}\|+\|\mathrm{L}_1^\alpha\|\right]}{\alpha+1}\eta^{1+1/\alpha}.\end{aligned}\tag{13}$$

Denote by $C':=\mathbb{E}\left[\sup_{i\le n}\|\mathrm{L}_1^{i,n}\|+\|\mathrm{L}_1^\alpha\|\right]$, and $\psi_t(\xi)$ the solution of (4) at time t with initial condition $\xi\in\mathbb{R}^{p\times n}$, which is the concatenation of n vectors $\psi_t^{i,n}(\xi)\in\mathbb{R}^p$, $i=1,\dots,n$. At time T which is a multiple of η,

$$\theta_T^{i,n}-\hat\theta_{T/\eta}^{i,n}=\sum_{k=0}^{T/\eta-1}\psi_{T-\eta k}^{i,n}(\hat\Theta_k^n)-\psi_{T-\eta(k+1)}^{i,n}(\hat\Theta_{k+1}^n),\tag{14}$$

where $\hat\Theta_k^n$ is the concatenation of $\hat\theta_k^{i,n}$. Similarly, for each of the term inside the summation above,

$$\begin{aligned}&\psi_{T-\eta k}^{i,n}(\hat\Theta_k^n)-\psi_{T-\eta(k+1)}^{i,n}(\hat\Theta_{k+1}^n)\\ =&\left[\int_{\eta k}^{\eta(k+1)} b^{i,n}(\psi_{t-\eta k}(\hat\Theta_k^n))\mathrm{d}t+\mathrm{dL}_t^{i,n}-(\hat\theta_{k+1}^{i,n}-\hat\theta_k^{i,n})\right]\\ &-\int_{\eta(k+1)}^T\Big(b^{i,n}(\psi_{t-\eta k}(\hat\Theta_k^n))-b^{i,n}(\psi_{t-\eta(k+1)}(\hat\Theta_{k+1}^n))\Big)\mathrm{d}t,\end{aligned}\tag{15}$$

where if no confusion arises, we write $b^{i,n}(\Theta^n)=-n\partial_{\theta^{i,n}}R(\Theta^n)$. Note that the first term in the big bracket is the difference of one-step increment started from $\hat\Theta_k^n$. Then, it follows from (13) that

$$\mathbb{E}\left[\sup_{i\le n}\left\|\int_{\eta k}^{\eta(k+1)} b^{i,n}(\psi_t(\hat\Theta_k^n))\mathrm{d}t+\mathrm{dL}_t^{i,n}-(\hat\theta_{k+1}^{i,n}-\hat\theta_k^{i,n})\right\|\right]\le B^2\eta^2+\frac{B\alpha\cdot C'}{\alpha+1}\eta^{1+1/\alpha}.\tag{16}$$

The second integral term similarly,

$$\begin{aligned}&\mathbb{E}\left[\sup_{i\le n}\|b^{i,n}(\psi_{t-\eta k}(\hat\Theta_k^n))-b^{i,n}(\psi_{t-\eta(k+1)}(\hat\Theta_{k+1}^n))\|\right]\\ \le& B\cdot\mathbb{E}\left[\sup_{i\le n}\|\psi_{t-\eta k}^{i,n}(\hat\Theta_k^n)-\psi_{t-\eta(k+1)}^{i,n}(\hat\Theta_{k+1}^n)\|\right]+\frac{B}{n}\sum_{j=1}^n\mathbb{E}\left[\|\psi_{t-\eta k}^{j,n}(\hat\Theta_k^n)-\psi_{t-\eta(k+1)}^{j,n}(\hat\Theta_{k+1}^n)\|\right]\end{aligned}\tag{17}$$

If we combine (15), (16), (17):

$$\begin{aligned}
&\mathbb{E}\left[\sup_{i\le n}\|\psi^{j,n}_{T-\eta k}(\hat{\Theta}^n_k)-\psi^{j,n}_{T-\eta(k+1)}(\hat{\Theta}^n_{k+1})\|\right]\\
\le& B^2\eta^2+\frac{2B\alpha\cdot C'}{\alpha+1}\eta^{1+1/\alpha}+2B\cdot\int_{\eta(k+1)}^{T}\mathbb{E}\left[\sup_{i\le n}\|\psi^{i,n}_{t-\eta k}(\hat{\Theta}^n_k)-\psi^{i,n}_{t-\eta(k+1)}(\hat{\Theta}^n_{k+1})\|\right]\mathrm{d}t.
\end{aligned}$$

Next it follows from Gronwall's inequality that

$$\mathbb{E}\left[\sup_{i\le n}\|\psi^{j,n}_{T-\eta k}(\hat{\Theta}^n_k)-\psi^{j,n}_{T-\eta(k+1)}(\hat{\Theta}^n_{k+1})\|\right]\le\exp(2BT)\left(B^2\eta^2+\frac{2B\alpha\cdot C'}{\alpha+1}\eta^{1+1/\alpha}\right).$$

Finally, combined with (14), we obtain

$$\mathbb{E}\left[\sup_{i\le n}\|\theta^{i,n}_T-\hat{\theta}^{i,n}_{T/\eta}\|\right]\le T\exp(2BT)\left(B^2\eta+\frac{2B\alpha\cdot C'}{\alpha+1}\eta^{1/\alpha}\right).$$

Then the result follows by Lemma C.1 that

$$C'=\mathbb{E}\left[\sup_{i\le n}\|\mathrm{L}^{i,n}_1\|+\|\mathrm{L}^{\alpha}_1\|\right]\le(8C_\alpha+1)(n^{1/\alpha}+1).$$

The proof of Theorem A.1 is therefore completed. □

Lemma C.1. *Take n i.i.d. α-stable random variables X^i (rmk:distributed as L^α_1. there exists $C_\alpha>0$ such that for t sufficiently large and any $i=1,\dots,n$, $\mathbb{P}[\|X^i\|\ge t]\ge C_\alpha t^{-\alpha}$.) such that $\mathbb{E}\left[\exp(itX^i)\right]=\exp\left(-|t|^\alpha\right)$ then*

$$\mathbb{E}\left[\sup_{i\le n}\|X^i\|\right]\le(8C_\alpha+1)n^{1/\alpha}$$

Proof. It is not hard to see from the condition $\mathbb{P}[\|X^i\|\ge t]\ge C_\alpha t^{-\alpha}$ that

$$\mathbb{P}\left[\sup_{i\le n}\|X^i\|\ge t\right]=1-\prod_{i=1}^{n}\mathbb{P}[\|X^i\|<t]\le 1-\left(1-C_\alpha t^{-\alpha}\right)^n$$

$$\begin{aligned}
\mathbb{E}\left[\sup_{i\le n}\|X^i\|\right]&=\int_0^\infty\mathbb{P}\left[\sup_{i\le n}\|X^i\|\ge t\right]\mathrm{d}t\\
&=\sum_{k=0}^{-\infty}\int_{(n/2^{k+1})^{1/\alpha}}^{(n/2^k)^{1/\alpha}}\mathbb{P}\left[\sup_{i\le n}\|X^i\|\ge t\right]\mathrm{d}t+\int_0^{n^{1/\alpha}}\mathbb{P}\left[\sup_{i\le n}\|X^i\|\ge t\right]\mathrm{d}t\\
&\le 2n^{1/\alpha}\sum_{k=0}^{-\infty}\mathbb{P}\left[\sup_{i\le n}\|X^i\|\ge(n/2^{k+1})^{1/\alpha}\right]+n^{1/\alpha}\\
&\le 2C_\alpha n^{1/\alpha}\sum_{k=0}^{-\infty}2^{k+1}+n^{1/\alpha}\\
&\le(8C_\alpha+1)n^{1/\alpha}.
\end{aligned}$$

The proof of Lemma C.1 is completed. □

C.4 Proof of Theorem 3.1

Definition C.1 (k-term approximation error [GCD12]). The best k-term approximation error $\overline{\sigma}_k(\mathbf{x})$ of a vector $\mathbf{x}$ is defined by

$$\sigma_k(\mathbf{x})=\inf_{\|\mathbf{y}\|_0\le k}\|\mathbf{x}-\mathbf{y}\|,$$

where $\|\mathbf{y}\|_0$ is the l^0-norm of $\mathbf{y}$, which counts the non-zero coefficients of $\mathbf{y}$. Without mentioned explicitly, $\|\mathbf{x}\|$ denotes the square norm of $\mathbf{x}$.

Proof. Denote by $\hat{\mathbf{w}}_t^n = (\|\hat{\theta}^{1,n}_{\lfloor t/\eta \rfloor}\|, \dots, \|\hat{\theta}^{n,n}_{\lfloor t/\eta \rfloor}\|)$ and $\mathbf{w}_t^* = (\|\theta_t^{1,\infty}\|, \dots, \|\theta_t^{n,\infty}\|)$, where the components $\theta_t^{i,\infty}$ are independent solutions to (5) in Theorem A.2. Note that the definition of Frobenius matrix norm $\|\cdot\|_F$ gives that

$$\|\hat{\Theta}^{\{\kappa n\}}_{\lfloor t/\eta \rfloor} - \hat{\Theta}^n_{\lfloor t/\eta \rfloor}\|_F = \|\sigma_{\langle \kappa n \rangle}(\hat{\mathbf{w}}_t^n)\|, \quad \|\hat{\Theta}^n_{\lfloor t/\eta \rfloor}\|_F = \|\mathbf{w}_t^\star\|, \tag{18}$$

Therefore it suffices to prove Theorem 3.1 for $\hat{\mathbf{w}}_t^n$. It follows from Theorem A.2 and Theorem A.1 that there exists a constant C independent of n such that

$$\mathbb{E}\left[\sup_{i \le n} \|\hat{\theta}^{i,n}_{\lfloor t/\eta \rfloor} - \theta_t^{i,\infty}\|\right] \le \frac{C}{\sqrt{n}}$$

Then by Markov's inequality,

$$\mathbb{P}\left[\sup_{i \le n} \|\hat{\theta}^{i,n}_{\lfloor t/\eta \rfloor} - \theta_t^{i,\infty}\| > \frac{C}{\epsilon\sqrt{n}}\right] \le \epsilon/3. \tag{19}$$

Denote by E the event $E := \left\{\sup_{i \le n} \|\hat{\theta}^{i,n}_{\lfloor t/\eta \rfloor} - \theta_t^{i,\infty}\| \le \frac{C}{\epsilon\sqrt{n}}\right\}$. If $\sup_{i \le n} \|\hat{\theta}^{i,n}_{\lfloor t/\eta \rfloor} - \theta_t^{i,\infty}\| \le \frac{C}{\epsilon\sqrt{n}}$ and $\|\sigma_{\lfloor \kappa n \rfloor}(\hat{\mathbf{w}}_t^n)\| \ge \epsilon\|\hat{\mathbf{w}}_t^n\|$, then

$$\begin{aligned}
\|\sigma_{\lfloor \kappa n \rfloor}(\mathbf{w}_t^\star)\| &\ge \|\sigma_{\lfloor \kappa n \rfloor}(\hat{\mathbf{w}}_t^n)\| - \kappa n \frac{C}{\epsilon\sqrt{n}} \\
&\ge \epsilon\|\hat{\mathbf{w}}_t^n\| - C\sqrt{n}\kappa/\epsilon \\
&\ge \epsilon(\|\mathbf{w}_t^\star\| - C\sqrt{n}/\epsilon) - C\sqrt{n}\kappa/\epsilon \\
&= \epsilon\|\mathbf{w}_t^\star\| - C\sqrt{n}(1 + \kappa/\epsilon)
\end{aligned}$$

Therefore plugging in (19),

$$\begin{aligned}
&\mathbb{P}\big[\|\sigma_{\lfloor \kappa n \rfloor}(\hat{\mathbf{w}}_t^n)\| \ge \epsilon\|\sigma_{\lfloor \kappa n \rfloor}(\hat{\mathbf{w}}_t^n)\|\big] \\
\le&\mathbb{P}\big[\|\sigma_{\lfloor \kappa n \rfloor}(\hat{\mathbf{w}}_t^n)\| \ge \epsilon\|\sigma_{\lfloor \kappa n \rfloor}(\hat{\mathbf{w}}_t^n)\|, E^c\big] + \mathbb{P}\big[\|\sigma_{\lfloor \kappa n \rfloor}(\hat{\mathbf{w}}_t^n)\| \ge \epsilon\|\sigma_{\lfloor \kappa n \rfloor}(\hat{\mathbf{w}}_t^n)\|, E\big] \\
\le&\mathbb{P}\left[\sup_{i \le n} \|\hat{\theta}^{i,n}_{\lfloor t/\eta \rfloor} - \theta_t^{i,\infty}\| > \frac{C}{\epsilon\sqrt{n}}\right] + \mathbb{P}\left[\|\sigma_{\lfloor \kappa n \rfloor}(\mathbf{w}_t^\star)\| \ge \epsilon\|\mathbf{w}_t^\star\| - C\sqrt{n}(1 + \kappa/\epsilon)\right] \\
\le&\epsilon/3 + \mathbb{P}\left[\|\sigma_{\lfloor \kappa n \rfloor}(\mathbf{w}_t^\star)\| \ge \epsilon\|\mathbf{w}_t^*\| - C\sqrt{n}(1 + \kappa/\epsilon)\right]
\end{aligned} \tag{20}$$

Moreover, there exists $N' > 0$ such that for all $n \ge N'$,

$$\begin{aligned}
&\mathbb{P}\left[\|\sigma_{\lfloor \kappa n \rfloor}(\mathbf{w}_t^\star)\| \ge \epsilon\|\mathbf{w}_t^*\| - C\sqrt{n}(1 + \kappa/\epsilon)\right] \\
\le&\mathbb{P}\left[\|\mathbf{w}_t^\star\| \le 2C\sqrt{n}(1 + \kappa/\epsilon)\right] + \mathbb{P}\left[\|\sigma_{\lfloor \kappa n \rfloor}(\mathbf{w}_t^\star)\| \ge \frac{\epsilon}{2}\|\mathbf{w}_t^\star\|\right] \\
=&\mathbb{P}\left[\frac{1}{n}\|\mathbf{w}_t^\star\|^2 \le 4C^2(1 + \kappa/\epsilon)^2\right] + \mathbb{P}\left[\|\sigma_{\lfloor \kappa n \rfloor}(\mathbf{w}_t^\star)\| \ge \frac{\epsilon}{2}\|\mathbf{w}_t^\star\|\right] \\
\le&\epsilon/3 + \mathbb{P}\left[\|\sigma_{\lfloor \kappa n \rfloor}(\mathbf{w}_t^\star)\| \ge \frac{\epsilon}{2}\|\mathbf{w}_t^\star\|\right],
\end{aligned} \tag{21}$$

where the last inequality follows from Lemma B.3. By the independence of the n coordinates of the vetor $\mathbf{w}_t^\star$, Theorem A.3 and [GCD12, Proposition 1, Part 2], there exists $N'' > 0$, for all $n \ge N''$,

$$\mathbb{P}\left[\|\sigma_{\lfloor \kappa n \rfloor}(\mathbf{w}_t^\star)\| \ge \frac{\epsilon}{2}\|\mathbf{w}_t^\star\|\right] \le \epsilon/3. \tag{22}$$

We conclude the proof by combining (18), (20), (21) and (22). □

D Experimental Details

For SGD, we fix the batch-size to be one tenth of the number of training data points, the step-size is chosen to be small enough to approximate the continuous dynamics given by the McKean-Vlasov equation in order to stay close to the theory, but also not too small so that SGD converges in a reasonable amount of time. As for the noise level σ, we have tried a range of values for each dataset and n, and we chose the largest σ such that the perturbed SGD converges. Intuitively, we can expect that smaller α with heavier tails will lead to lower relative compression error. However, it does not guarantee better test performance: one has to fine tune the parameters appropriately to achieve a favorable trade-off between compression error and the test performance. We repeat all the experiment 5 times and report and average and the standard deviation. For the noiseless case (vanilla SGD), the results of the different runs were almost identical, hence we did not report the standard deviations.

The code, implemented in PyTorch, takes about 90 hours to run on the MNIST dataset with five different seeds for n = 2K, 5K, 10K on a NVIDIA Tesla P100 GPU. With the same system configuration, it takes 5 minutes to run on ECG5000 with Type-I noise; 3 hours with Type II noise and 30 minutes with Type-III noise.

The ECG5000 dataset consists of 5000 20-hour long electrocardiograms interpolated by sequences of length 140 to discriminate between normal and abnormal heart beats of a patient that has severe congestive heart failure. After random shuffling, we use 500 sequences for the training phase and 4500 sequences for the test phase. The hyperparameters used in the ECG5000 classification experiments are summarized in Table 6 and Table 7 that follow.

n	stepsize	noise level σ	batch size	number of epochs
2K	3e-4	0.5	500	600
5K	1e-4	0.5	500	1500
10K	5e-5	0.5	500	3000

Table 6: ECG5000 classification with Type-I and Type-III noises for n = 2K, 5K and 10K.

n	stepsize	noise level σ	batch size	number of epochs
2K	3e-4	0.15	500	600
5K	1e-4	0.15	500	1500
10K	5e-5	0.15	500	3000

Table 7: ECG5000 classification with Type-II noise for n = 2K, 5K and 10K.

The MNIST database of handwritten digits consists of a training set of 60,000 examples and a test set of 10,000 examples of dimension 784. The hyperparameters used in the MNIST classification experiments until 95% training accuracy are specified in Table 8.

n	stepsize	noise level σ	batch size
2K	1e-2	5e-4	5000
5K	5e-3	4e-4	5000
10K	5e-3	6e-4	5000

Table 8: MNIST classification with Type-I noise for n = 2K, 5K and 10K.

E Additional Experiments

In this section, we provide additional experiments for the classification task with the one hidden-layer neural network trained using the ECG5000 and MNIST datasets. We conducted prunability tests further for various values of the number of neurons n, the index α and the noise type.

E.1 Further results for the ECG5000 classification

For the classification of the ECG5000 dataset using the one hidden-layer neural network, we conducted prunability tests for the following values of parameters: number of neurons, n = 2K, 5K and 10K, index α = 1.75, 1.8 and 1.9 and noises Type-I, II and III. The results, which complement those in Tables 1, 2, 3, 4, are reported in Tables 9, 10, 11, 12, 13 that follow.

α	Train Acc.	Test Acc.	Pruning Ratio	Train Acc. a.p.	Test Acc. a.p.
no noise	0.974	0.957	11.46	0.974	0.958
1.75	0.974 ± 0.001	0.959 ± 0.002	46.6 ± 5.22	0.959 ± 0.016	0.951 ± 0.014
1.8	0.974 ± 0.001	0.959 ± 0.002	42.72 ± 3.78	0.96 ± 0.017	0.952 ± 0.013
1.9	0.974 ± 0.001	0.959 ± 0.002	36.84 ± 1.51	0.969 ± 0.005	0.956 ± 0.003

Table 9: ECG5000 with Type-I noise and $n = 5$K.

α	train acc	test acc	pruning ratio	train acc a.p.	test acc a.p.
1.75	0.979 ± 0.003	0.970 ± 0.003	39.51 ± 12.18	0.959 ± 0.031	0.957 ± 0.024
1.8	0.975 ± 0.003	0.967 ± 0.003	30.33 ± 7.93	0.9692 ± 0.01	0.9618 ± 0.008
1.9	0.974 ± 0.004	0.962 ± 0.003	18.37 ± 2.54	0.974 ± 0.004	0.963 ± 0.007

Table 10: ECG5000 with Type-II noise and $n = 2$K.

α	train acc	test acc	pruning ratio	train acc a.p.	test acc a.p.
1.75	0.984 ± 0.004	0.978 ± 0.005	46.23 ± 30.14	0.945 ± 0.073	0.943 ± 0.072
1.8	0.982 ± 0.002	0.976 ± 0.005	38.1 ± 27.23	0.948 ± 0.072	0.946 ± 0.069
1.9	0.98 ± 0.002	0.971 ± 0.005	20.79 ± 8.57	0.976 ± 0.003	0.972 ± 0.006

Table 11: ECG5000 with Type-II noise and $n = 5$K.

α	train acc	test acc	pruning ratio	train acc a.p.	test acc a.p.
1.75	0.98 ± 0.004	0.973 ± 0.005	35.39 ± 7.42	0.932 ± 0.033	0.934 ± 0.022
1.8	0.981 ± 0.004	0.971 ± 0.005	27.98 ± 4.82	0.956 ± 0.024	0.960 ± 0.02
1.9	0.978 ± 0.004	0.968 ± 0.005	17.83 ± 1.59	0.976 ± 0.003	0.97 ± 0.004

Table 12: ECG5000 with Type-III noise and $n = 2$K.

α	train acc	test acc	pruning ratio	train acc a.p.	test acc a.p.
1.75	0.971 ± 0.004	0.96 ± 0.005	30.67 ± 2.78	0.965 ± 0.012	0.954 ± 0.008
1.8	0.971 ± 0.004	0.959 ± 0.005	24.92 ± 1.75	0.970 ± 0.004	0.959 ± 0.003
1.9	0.972 ± 0.002	0.957 ± 0.005	16.9 ± 0.001	0.973 ± 0.001	0.958 ± 0.002

Table 13: ECG5000 with Type-III noise and $n = 5$K.

E.2 Further results for the MNIST classification

For the classification of the MNIST dataset using the one hidden-layer neural network, we conducted prunability tests for the following values of parameters: number of neurons, $n = 2$K, 5K and 10K, index $\alpha = 1.75$, 1.8 and 1.9 and noises Type-I, II and III. The results for smaller $n = 2$K and 5K are reported in Tables 14 and 15 that follow; and they complement those for $n = 10$K in Table 5 in the main body of the text.

α	no noise	1.9	1.8	1.75
pruning ratio	10.85	15.95 ± 0.788	23.17 ± 4.26	30.58 ± 7.91
train accuracy a.p.	0.9486	0.9482 ± 0.0008	0.9468 ± 0.0004	0.9449 ± 0.0018
test accuracy a.p.	0.9474	0.9474 ± 0.0008	0.946 ± 0.0012	0.9439 ± 0.0022

Table 14: MNIST with Type-I noise and $n = 2$K until reaching 95% training accuracy.

α	no noise	1.9	1.8	1.75
pruning ratio	10.64	14.91 ± 1.56	24.16 ± 8.9	42.77 ± 21.88
train accuracy a.p.	0.9485	0.9486 ± 0.0003	0.9467 ± 0.0012	0.939 ± 0.008
test accuracy a.p.	0.9473	0.9462 ± 0.0004	0.9448 ± 0.002	0.938 ± 0.007

Table 15: MNIST with Type-I noise and $n = 5$K until reaching 95% training accuracy.

In Tables 16, 17 and 18 we report train and test accuracies that are obtained in the following way: (i) the one-hidden-layer neural network is first trained on the MNIST dataset using vanilla SGD, i.e., SGD with *no* noise injection; (ii) then we perform pruning with the same pruning ratios as those given in Tables 14 and 15; and (iii) finally we evaluate the accuracy after pruning on both train and test sets of the used MNIST dataset. It can be observed that the larger the value of n (the size of the neural network), the less compressible is the neural network trained with vanilla SGD, especially for $n = 10$K. In the latter case the test accuracy of the neural network trained using vanilla SGD drops down to 92%, while the noising scheme achieves 94% of test accuracy with the same pruning ratio, as can be seen from Table 5.

Pruning ratio	15.95	23.17	30.58
train acc a.p.	0.9481	0.9468	0.9455
test acc a.p.	0.9484	0.9455	0.9447

Table 16: MNIST accuracies for $n = 2K$, after pruning.

Pruning ratio	14.91	24.16	42.77
train acc a.p.	0.9464	0.9455	0.9332
test acc a.p.	0.9466	0.9447	0.9323

Table 17: MNIST accuracies for $n = 5K$, after pruning.

Pruning ratio	21.31	34.49	44.42
train acc a.p.	0.9452	0.9380	0.9221
test acc a.p.	0.9436	0.9385	0.9223

Table 18: MNIST accuracies for $n = 10K$, after pruning.

E.3 Effect of heavy-tailed noise injection during SGD on the performance after pruning

Table 19 reports accuracy results for the two-layer neural network trained on the CIFAR10 dataset with heavy-tailed SGD ($\alpha = 1.8$), for various levels of the variance of the added Type-I noise. In this case, the effect of pruning seems to require a larger value of n to start to be visible – the results reported in the table, which suggest that the noise injection may have a non-negligible effect on the train/test accuracy after pruning especially for large values of the noise variance, are obtained with relatively small $n = 5$K for CIFAR10 dataset with samples of dimension 3072.

σ	10^{-5}	2×10^{-5}	3×10^{-5}
pruning ratio	16.32	23.86	33.94
train accuracy a.p.	0.9236	0.8713	0.7265
test accuracy a.p.	0.5606	0.5302	0.4779

Table 19: CIFAR10 with $n = 5$K and Type-I 1.8-stable noise for various noise levels.

F Implications on Federated Learning

The federated learning (FL) setting [MMR$^+$17, RM17] is one in which there are a number of devices or clients, say n; all equipped with the same neural network model and each holding an independent own dataset. Every client learns an individual (or local) model from its own dataset, e.g., via

Stochastic Gradient Descent (SGD). The individual models are aggregated by a *parameter server* (PS) into a global model and then sent back to the devices, possibly over multiple rounds of communication between them. The rationale is that the individually learned models are refined progressively by taking into account the data held by other devices; and, at the end the training process, all relevant features of all devices' datasets are captured by the final aggregated model.

The results of this paper are useful towards a better understanding of the *compressibility* of the models learned by the various clients in this FL setting. Specifically, viewing each neuron of the hidden layer of the setup of this paper as if it were a distinct client, the results that we establish *suggest* that if the local models are learned via heavy-tailed SGD this would enable a better compressibility of them. This is particularly useful for resource-constrained applications of FL, such as in telecommunication networks where bandwidth is scarce and latency is important.