The Heavy-Tail Phenomenon in SGD

SUPPLEMENTARY DOCUMENT

Mert Gürbüzbalaban [1] **Umut Şimşekli** [2] **Lingjiong Zhu** [3]

The supplementary document is organized as follows.

1. The supplementary document begins with a discussion of different choices of stochastic differential equation (SDE) representations for SGD (Section A).
2. We then discuss the tail-index estimation in Section B.
3. In Section C, we provide the proofs of the main results in the main paper; and provide the supporting lemmas in Section D.

A. A Note on Stochastic Differential Equation Representations for SGD

In recent years, a popular approach for analyzing the behavior of SGD has been viewing it as a discretization of a continuous-time stochastic process that can be represented via a stochastic differential equation (SDE) (Mandt et al., 2016; Jastrzębski et al., 2017; Li et al., 2017; Hu et al., 2019; Zhu et al., 2019; Chaudhari & Soatto, 2018; Şimşekli et al., 2019b). While these SDEs have been useful for understanding different properties of SGD, their differences and functionalities have not been clearly understood. In this section, in light of our theoretical results, we will discuss in which situation their choice would be more appropriate. We will restrict ourselves to the case where $f(x)$ is a quadratic function; however, the discussion can be extended to more general f.

The SDE approximations are often motivated by first rewriting the SGD recursion as follows:

$$x_{k+1} = x_k - \eta \nabla \tilde{f}_{k+1}(x_k) = x_k - \eta \nabla f(x_k) + \eta U_{k+1}(x_k), \tag{A.1}$$

where $U_k(x) := \nabla \tilde{f}_k(x) - \nabla f(x)$ is called the 'stochastic gradient noise'. Then, based on certain statistical assumptions on U_k, we can view (A.1) as a discretization of an SDE. For instance, if we assume that the gradient noise follows a Gaussian distribution, whose covariance does not depend on the iterate x_k, i.e., $\eta U_k \approx \sqrt{\eta} Z_k$ where $Z_k \sim \mathcal{N}(0, \sigma_z \eta I)$ for some constant $\sigma_z > 0$, we can see (A.1) as the Euler-Maruyama discretization of the following SDE with stepsize η (Mandt et al., 2016):

$$\mathrm{d}x_t = -\nabla f(x_t)\mathrm{d}t + \sqrt{\eta \sigma_z}\mathrm{dB}_t, \tag{A.2}$$

where B_t denotes the d-dimensional standard Brownian motion. This process is called the Ornstein-Uhlenbeck (OU) process (see e.g. Øksendal (2013)), whose invariant measure is a Gaussian distribution. We argue that this process can be a good proxy to (3.5) only when $\alpha \geq 2$, since otherwise the SGD iterates will exhibit heavy-tails, whose behavior cannot be captured by a Gaussian distribution. As we illustrated in Section 4, to obtain large α, the stepsize η needs to be small and/or the batch-size b needs to be large. However, it is clear that this approximation will fall short when the system exhibits heavy tails, i.e., $\alpha < 2$. Therefore, for the large η/b regime, which appears to be more interesting since it often yields improved test performance (Jastrzębski et al., 2017), this approximation would be inaccurate for understanding the behavior of SGD. This problem mainly stems from the fact that the additive isotropic noise assumption results in a deterministic M_k matrix for all k. Since there is no *multiplicative noise* term, this representation cannot capture a potential heavy-tailed behavior.

A natural extension of the state-independent Gaussian noise assumption is to incorporate the covariance structure of U_k. In our linear regression problem, we can easily see that the covariance matrix of the gradient noise has the following form:

$$\Sigma_U(x) = \mathrm{Cov}(U_k|x) = \frac{\sigma^2}{b}\mathrm{diag}(x \circ x), \tag{A.3}$$

where $\circ$ denotes element-wise multiplication and σ^2 is the variance of the data points. Therefore, we can extend the previous assumption by assuming $Z_k|x \sim \mathcal{N}(0, \eta\Sigma_U(x))$. It has been observed that this approximation yields a more accurate representation (Cheng et al., 2020; Ali et al., 2020; Jastrzębski et al., 2017). Using this assumption in (A.1), the SGD recursion coincides with the Euler-Maruyama discretization of the following SDE:

$$\begin{aligned} dx_t &= -\nabla f(x_t)dt + \sqrt{\eta\Sigma_U(x_t)}d\mathrm{B}_t \\ &\overset{\mathrm{d}}{=} -\left(A^\top Ax_t - A^\top y\right)dt + \sqrt{\frac{\sigma^2\eta}{b}}\mathrm{diag}(x_t)d\mathrm{B}_t, \end{aligned} \tag{A.4}$$

where $\overset{\mathrm{d}}{=}$ denotes equality in distribution. The stochasticity in such SDEs is called often called *multiplicative*. Let us illustrate this property by discretizing this process and by using the definition of the gradient and the covariance matrix, we observe that (noting that $N_k \sim \mathcal{N}(0, I)$)

$$\begin{aligned} x_{k+1} &= x_k - \eta\left(A^\top Ax_k - A^\top y\right) + \sqrt{\frac{\sigma^2\eta^2}{b}}\mathrm{diag}(x_k)N_{k+1} \\ &= \left(I - \eta A^\top A + \sqrt{\sigma^2\eta^2/b}\,\mathrm{diag}(N_{k+1})\right)x_k - \eta A^\top y, \end{aligned} \tag{A.5}$$

where we can clearly see the multiplicative effect of the noise, as indicated by its name. On the other hand, we can observe that, thanks to the multiplicative structure, this process would be able to capture the potential heavy-tailed structure of SGD. However, there are two caveats. The first one is that, in the case of linear regression, the process is called a geometric (or modified) Ornstein-Uhlenbeck process which is an extension of geometric Brownian motion. One can show that the distribution of the process at any time t will have lognormal tails. Hence it will be accurate only when the tail-index α is close to the one of the lognormal distribution. The second caveat is that, for a more general cost function f, the covariance matrix is more complicated and hence the invariant measure of the process cannot be found analytically, hence analyzing these processes for a general f can be as challenging as directly analyzing the behavior of SGD.

The third way of modeling the gradient noise is based on assuming that it is heavy-tailed. In particular, we can assume that $\eta U_k \approx \eta^{1/\alpha}L_k$ where $[L_k]_i \sim \mathcal{S}\alpha\mathcal{S}(\sigma_L\eta^{(\alpha-1)/\alpha})$ for all $i = 1, \dots, d$. Under this assumption the SGD recursion coincides with the Euler discretization of the following Lévy-driven SDE (Şimşekli et al., 2019b):

$$dx_t = -\nabla f(x_t)dt + \sigma_L\eta^{(\alpha-1)/\alpha}d\mathrm{L}_t^\alpha, \tag{A.6}$$

where L_t^α denotes the α-stable Lévy process with independent components (see Section A.1 for technical background on Lévy processes and in particular α-stable Lévy processes). In the case of linear regression, this processes is called a fractional OU process (Fink & Klüppelberg, 2011), whose invariant measure is also an α-stable distribution with the same tail-index α. Hence, even though it is based on an isotropic, state-independent noise assumption, in the case of large η/b regime, this approach can mimic the heavy-tailed behavior of the system with the exact tail-index α. On the other hand, Buraczewski et al. (2016) (Theorem 1.7 and 1.16) showed that if U_k is assumed to heavy tailed with index α (not necessarily $\mathcal{S}\alpha\mathcal{S}$) then the process x_k will inherit the same tails and the ergodic averages will still converge to an $\mathcal{S}\alpha\mathcal{S}$ random variable in distribution, hence generalizing the conclusions of the $\mathcal{S}\alpha\mathcal{S}$ assumption to the case where U_k follows an arbitrary heavy-tailed distribution.

A.1. Technical background: Lévy processes

Lévy motions (processes) are stochastic processes with independent and stationary increments, which include Brownian motions as a special case, and in general may have heavy-tailed distributions (see e.g. Bertoin (1996) for a survey). Symmetric α-stable Lévy motion is a Lévy motion whose time increments are symmetric α-stable distributed. We define L_t^α, a d-dimensional symmetric α-stable Lévy motion as follows. Each component of L_t^α is an independent scalar α-stable Lévy process defined as follows:

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

(ii) For any $t_0 < t_1 < \cdots < t_N$, the increments $\mathrm{L}_{t_n}^\alpha - \mathrm{L}_{t_{n-1}}^\alpha$ are independent, $n = 1, 2, \dots, N$;

(iii) The difference $\mathrm{L}_t^\alpha - \mathrm{L}_s^\alpha$ and L_{t-s}^α have the same distribution: $\mathcal{S}\alpha\mathcal{S}((t-s)^{1/\alpha})$ for $s < t$;

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

When $\alpha = 2$, we obtain a scaled Brownian motion as a special case, i.e. $\mathrm{L}_t^\alpha = \sqrt{2}\mathrm{B}_t$, so that the difference $\mathrm{L}_t^\alpha - \mathrm{L}_s^\alpha$ follows a Gaussian distribution $\mathcal{N}(0, 2(t-s))$.

B. Tail-Index Estimation

In this study, we follow Tzagkarakis et al. (2018); Şimşekli et al. (2019b), and make use of the recent estimator proposed by Mohammadi et al. (2015).

Theorem 12 (Mohammadi et al. (2015) Corollary 2.4)**.** *Let $\{X_i\}_{i=1}^K$ be a collection of strictly stable random variables in $\mathbb{R}^d$ with tail-index $\alpha \in (0, 2]$ and $K = K_1 \times K_2$. Define $Y_i = \sum_{j=1}^{K_1} X_{j+(i-1)K_1}$ for $i \in [\![1, K_2]\!]$. Then, the estimator*

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

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

As this estimator requires a hyperparameter K_1, at each tail-index estimation, we used several values for K_1 and we used the median of the estimators obtained with different values of K_1. We provide the codes in `github.com/umutsimsekli/sgd_ht`, where the implementation details can be found. For the neural network experiments, we used the same setup as provided in the repository of Şimşekli et al. (2019b).

C. Proofs of Main Results

C.1. Proof of Theorem 2

Proof of Theorem 2. The proof follows from Theorem 4.4.15 in Buraczewski et al. (2016) which goes back to Theorem 1.1 in Alsmeyer & Mentemeier (2012) and Theorem 6 in Kesten (1973). See also Goldie (1991); Buraczewski et al. (2015). We recall that we have the stochastic recursion:

$$x_k = M_k x_{k-1} + q_k, \tag{C.1}$$

where the sequence (M_k, q_k) are i.i.d. distributed as (M, q) and for each k, (M_k, q_k) is independent of x_{k-1}. To apply Theorem 4.4.15 in Buraczewski et al. (2016), it suffices to have the following conditions being satisfied:

1. M is invertible with probability 1.
2. The matrix M has a continuous Lebesgue density that is positive in a neighborhood of the identity matrix.
3. $\rho < 0$ and $h(\alpha) = 1$.
4. $\mathbb{P}(Mx + q = x) < 1$ for every x.
5. $\mathbb{E}\left[\|M\|^\alpha(\log^+\|M\| + \log^+\|M^{-1}\|)\right] < \infty$.
6. $0 < \mathbb{E}\|q\|^\alpha < \infty$.

All the conditions are satisfied under our assumptions. In particular, Condition 1 and Condition 5 are proved in Lemma 21, and Condition 2 and Condition 4 follow from the fact that M and q have continuous distributions. Condition 3 is part of the assumption of Theorem 2. Finally, Condition 6 is satisfied by the definition of q and by the Assumptions **(A1)–(A2)**. □

C.2. Proof of Theorem 3

Proof of Theorem 3. To prove (i), according to the proof of Theorem 2, it suffices to show that if $\rho < 0$, then there exists a unique positive α such that $h(\alpha) = 1$. Note that if $\rho < 0$, then by Lemma 17, we have $h(0) = 1$, $h'(0) = \rho < 0$ and $h(s)$ is convex in s, and moreover by Lemma 18, we have $\liminf_{s\to\infty} h(s) > 1$. Therefore, there exists some $\alpha \in (0, \infty)$ such that $h(\alpha) = 1$. Finally, (ii) follows from Lemma 16. □

C.3. Proof of Theorem 4

Proof of Theorem 4. We will split the proof of Theorem 4 into two parts:

(I) We will show that the tail-index α is strictly decreasing in stepsize η and variance σ^2 provided that $\alpha \geq 1$.

(II) We will show that the tail-index α is strictly increasing in batch-size b provided that $\alpha \geq 1$.

(III) We will show that the tail-index α is strictly decreasing in dimension d.

First, let us prove (I). Let $a := \eta\sigma^2 > 0$ be given. Consider the tail-index α as a function of a, i.e.

$$\alpha(a) := \min\{s : h(a,s) = 1\}\,,$$

where $h(a,s) = h(s)$ with emphasis on dependence on a.

By assumption, $\alpha(a) \geq 1$. The function $h(a,s)$ is convex function of a (see Lemma 22 for $s \geq 1$ and a strictly convex function of s for $s \geq 0$). Furthermore, it satisfies $h(a,0) = 1$ for every $a \geq 0$ and $h(0,s) = 1$ for every $s \geq 0$. We consider the curve

$$\mathcal{C} := \{(a,s) \in (0,\infty) \times [1,\infty] : h(a,s) = 1\}\,.$$

This is the set of the choice of a, which leads to a tail-index s where $s \geq 1$. Since h is smooth in both a and s, we can represent s as a smooth function of a, i.e. on the curve

$$h(a, s(a)) = 0\,,$$

where $s(a)$ is a smooth function of a. We will show that $s'(a) < 0$; i.e. if we increase a; the tail-index $s(a)$ will drop. Pick any $(a_*, s_*) \in \mathcal{C}$, it will satisfy $h(a_*, s_*) = 1$. We have the following facts:

- (i) The function $h(a,s) = 1$ for either $a = 0$ or $s = 0$. This is illustrated in Figure 4 with a blue marker.
- (ii) $h(a_*, s) < 1$ for $s < s_*$. This follows from the convexity of $h(a_*, s)$ function and the fact that $h(a_*, 0) = 1$, $h(a_*, s_*) = 1$. From here, we see that the function $h(a_*, s)$ is increasing at $s = s_*$ and we have its derivative

$$\frac{\partial h}{\partial s}(a_*, s_*) > 0.$$

- (iii) The function $h(a, s_*)$ is convex as a function of a by Lemma 22, it satisfies $h(0, s_*) = h(a_*, s_*) = 1$. Therefore, by convexity $h(a, s_*) < 1$ for a $\in (0, s_*)$; otherwise the function $h(a, s_*)$ would be a constant function. We have therefore necessarily.

$$\frac{\partial h}{\partial a}(a_*, s_*) > 0.$$

  By convexity of the function $h(a, s_*)$, we have also $h(a, s_*) \geq h(a_*, s_*) + \frac{\partial h}{\partial a}(a_*, s_*)(a - a_*) > h(a_*, s_*) = 1$. Therefore, $h(a, s_*) > 1$ for $a > a_*$. Then, it also follows that $h(a, s) > 1$ for $a > a_*$ and $s > s_*$ (otherwise if $h(a, s) \leq 1$, we get a contradiction because $h(0, s) = 1$, $h(a_*, s) > 1$ and $h(a, s) \leq 1$ is impossible due to convexity). This is illustrated in Figure 4 where we mark this region as a rectangular box where $h > 1$.

- (iv) By similar arguments we can show that the function $h(a, s) < 1$ if $(s, a) \in (0, a_*) \times [1, s_*)$. Indeed, if $h(a, s) \geq 1$ for some $(s, a) \in [1, s_*) \times (0, a_*)$, this contradicts the fact that $h(0, s) = 1$ and $h(a_*, s) < 1$ proven in part (ii). This is illustrated in Figure 4 where inside the rectangular box on the left-hand side, we have $h < 1$.

Geometrically, we see from Figure 4 that the curve $s(a)$ as a function of a, is sandwiched between two rectangular boxes and has necessarily $s'(a) < 0$. This can also be directly obtained rigorously from the implicit function theorem; if we differentiate the implicit equation $h(a, s(a)) = 0$ with respect to a, we obtain

$$\frac{\partial h}{\partial a}(a_*, s_*) + \frac{\partial h}{\partial s}(a_*, s_*)s'(a_*) = 0\,.$$

From parts $(ii) - (iii)$, we have $\frac{\partial h}{\partial a}(a_*, s_*)$ and $\frac{\partial h}{\partial s}(a_*, s_*) > 0$. Therefore, we have

$$s'(a_*) = -\frac{\frac{\partial h}{\partial a}(a_*, s_*)}{\frac{\partial h}{\partial s}(a_*, s_*)} < 0\,, \tag{C.2}$$

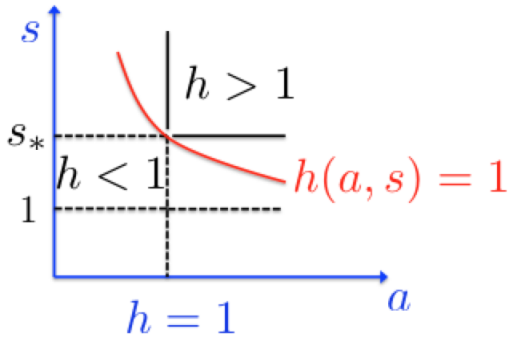

Figure 4. The curve $h(a,s)=1$ in the (a,s) plane

which completes the proof for $s_* \geq 1$.

Next, let us prove (II). With slight abuse of notation, we define the function $h(b,s)=h(s)$ to emphasize the dependence on b. We have

$$h(b,s) = \mathbb{E}\left\|\left(I - \frac{\eta}{b}\sum_{i=1}^{b} a_i a_i^T\right) e_1\right\|^s. \tag{C.3}$$

where we used Lemma 16. When $s \geq 1$, the function $x \mapsto \|x\|^s$ is convex, and by Jensen's inequality, we get for any $b \geq 2$ and $b \in \mathbb{N}$,

$$\begin{aligned} h(b,s) &= \mathbb{E}\left\|\frac{1}{b}\sum_{i=1}^{b}\left(I - \frac{\eta}{b-1}\sum_{j\neq i} a_j a_j^T\right) e_1\right\|^s \\ &\leq \mathbb{E}\left[\frac{1}{b}\sum_{i=1}^{b}\left\|\left(I - \frac{\eta}{b-1}\sum_{j\neq i} a_j a_j^T\right) e_1\right\|^s\right] \\ &= \frac{1}{b}\sum_{i=1}^{b}\mathbb{E}\left[\left\|\left(I - \frac{\eta}{b-1}\sum_{j\neq i} a_j a_j^T\right) e_1\right\|^s\right] = h(b-1,s), \end{aligned}$$

where we used the fact that a_i are i.i.d. Indeed, from the condition for equality to hold in Jensen's inequality, and the fact that a_i are i.i.d. random, the inequality above is a strict inequality. Hence when $d \in \mathbb{N}$ for any $s \geq 1$, $h(b,s)$ is strictly decreasing in b. By following the same argument as in the proof of (I), we conclude that the tail-index α is strictly increasing in batch-size b.

Finally, let us prove (III). Let us show the tail-index α is strictly decreasing in dimension d. Since a_i are i.i.d. and $a_i \sim \mathcal{N}(0, \sigma^2 I_d)$, by Lemma 19,

$$h(s) = \mathbb{E}\left[\left(1 - \frac{2a}{b}X + \frac{a^2}{b^2}X^2 + \frac{a^2}{b^2}XY\right)^{s/2}\right], \tag{C.4}$$

where X, Y are independent chi-square random variables with degree of freedom b and $d-1$ respectively. Notice that $h(s)$ is strictly increasing in d since the only dependence of $h(s)$ on d is via Y, which is a chi-square distribution with degree of freedom $(d-1)$. By writing $Y = Z_1^2 + \cdots + Z_{d-1}^2$, where $Z_i \sim N(0,1)$ i.i.d., it follows that $h(s)$ is strictly increasing in d. Hence, by similar argument as in (I), we conclude that α is strictly decreasing in dimension d. □

Remark 13. *When $d=1$ and a_i are i.i.d. $N(0,\sigma^2)$, we can provide an alternative proof that the tail-index α is strictly increasing in batch-size b. It suffices to show that for any $s \geq 1$, $h(s)$ is strictly decreasing in the batch-size b. By Lemma 19 when $d=1$,*

$$h(b,s) = \mathbb{E}\left[\left(1 - \frac{2\eta\sigma^2}{b}X + \frac{\eta^2\sigma^4}{b^2}X^2 + \frac{\eta^2\sigma^4}{b^2}XY\right)^{s/2}\right], \tag{C.5}$$

where $h(b,s)$ is as in (C.3) and X, Y are independent chi-square random variables with degree of freedom b and $d-1$ respectively. When $d=1$, we have $Y \equiv 0$, and

$$h(b,s) = \mathbb{E}\left[\left(1 - \frac{2\eta\sigma^2}{b}X + \frac{\eta^2\sigma^4}{b^2}X^2\right)^{s/2}\right] = \mathbb{E}\left[\left|1 - \frac{\eta\sigma^2}{b}X\right|^s\right]. \tag{C.6}$$

Since X is a chi-square random variable with degree of freedom b, we have

$$h(b,s) = \mathbb{E}\left[\left|1 - \frac{\eta\sigma^2}{b}\sum_{i=1}^{b} Z_i\right|^s\right], \tag{C.7}$$

where Z_i are i.i.d. $N(0,1)$ random variables. When $s \geq 1$, the function $x \mapsto |x|^s$ is convex, and by Jensen's inequality, we get for any $b \geq 2$ and $b \in \mathbb{N}$

$$\begin{aligned} h(b,s) &= \mathbb{E}\left[\left|\frac{1}{b}\sum_{i=1}^{b}\left(1 - \frac{\eta\sigma^2}{b-1}\sum_{j\neq i} Z_j\right)\right|^s\right] \\ &\leq \mathbb{E}\left[\frac{1}{b}\sum_{i=1}^{b}\left|1 - \frac{\eta\sigma^2}{b-1}\sum_{j\neq i} Z_j\right|^s\right] = \frac{1}{b}\sum_{i=1}^{b}\mathbb{E}\left[\left|1 - \frac{\eta\sigma^2}{b-1}\sum_{j\neq i} Z_j\right|^s\right] = h(b-1,s), \end{aligned}$$

where we used the fact that Z_i are i.i.d. Indeed, from the condition for equality to hold in Jensen's inequality, and the fact that Z_i are i.i.d. $N(0,1)$ distributed, the inequality above is a strict inequality. Hence when $d=1$ for any $s \geq 1$, $h(b,s)$ is strictly decreasing in b.

C.4. Proof of Proposition 5

Proof of Proposition 5. We first prove (i). When $\eta = \eta_{crit} = \frac{2b}{\sigma^2(d+b+1)}$, that is $\eta\sigma^2(d+b+1) = 2b$, we can compute that

$$\rho \leq \frac{1}{2}\log\mathbb{E}\left[1 - \frac{2\eta\sigma^2}{b}\sum_{i=1}^{b} z_{i1}^2 + \frac{\eta^2\sigma^4}{b^2}\sum_{i=1}^{b}\sum_{j=1}^{b}(z_{i1}z_{j1} + \cdots + z_{id}z_{jd})z_{i1}z_{j1}\right] = 0, \tag{C.8}$$

where z_{ij} are i.i.d. $N(0,1)$ random variables. Note that since $1 - \frac{2\eta\sigma^2}{b}\sum_{i=1}^{b} z_{i1}^2 + \frac{\eta^2\sigma^4}{b^2}\sum_{i=1}^{b}\sum_{j=1}^{b}(z_{i1}z_{j1} + \cdots + z_{id}z_{jd})z_{i1}z_{j1}$ is random, the inequality above is a strict inequality from Jensen's inequality. Thus, when $\eta = \eta_{crit}$, i.e. $\eta\sigma^2(d+b+1) = 2b$, $\rho < 0$. By continuity, there exists some $\delta > 0$ such that for any $2b < \eta\sigma^2(d+b+1) < 2b+\delta$, i.e. $\eta_{crit} < \eta < \eta_{max}$, where $\eta_{max} := \eta_{crit} + \frac{\delta}{\sigma^2(d+b+1)}$, we have $\rho < 0$. Moreover, when $\eta\sigma^2(d+b+1) > 2b$, i.e. $\eta > \eta_{crit}$, we have

$$h(2) = \mathbb{E}\left[\left(1 - \frac{2\eta\sigma^2}{b}\sum_{i=1}^{b} z_{i1}^2 + \frac{\eta^2\sigma^4}{b^2}\sum_{i=1}^{b}\sum_{j=1}^{b}(z_{i1}z_{j1} + \cdots + z_{id}z_{jd})z_{i1}z_{j1}\right)\right] = 1 - 2\eta\sigma^2 + \frac{\eta^2\sigma^4}{b}(d+b+1) \geq 1,$$

which implies that there exists some $0 < \alpha < 2$ such that $h(\alpha) = 1$.

Finally, let us prove (ii) and (iii). When $\eta\sigma^2(d+b+1) \leq 2b$, i.e. $\eta \leq \eta_{crit}$, we have $h(2) \leq 1$, which implies that $\alpha > 2$. In particular, when $\eta\sigma^2(d+b+1) = 2b$, i.e. $\eta = \eta_{crit}$, the tail-index $\alpha = 2$. □

C.5. Proof of Theorem 6 and Corollary 7

Proof of Theorem 6. We recall that

$$x_k = M_k x_{k-1} + q_k, \tag{C.9}$$

which implies that

$$\|x_k\| \leq \|M_k x_{k-1}\| + \|q_k\|. \tag{C.10}$$

(i) If the tail-index $\alpha \leq 1$, then for any $0 < p < \alpha$, we have $h(p) = \mathbb{E}\|M_k e_1\|^p < 1$ and moreover by Lemma 23,

$$\|x_k\|^p \leq \|M_k x_{k-1}\|^p + \|q_k\|^p. \tag{C.11}$$

Due to spherical symmetry of the isotropic Gaussian distribution, the distribution of $\frac{\|M_k x\|}{\|x\|}$ does not depend on the choice of $x \in \mathbb{R}^d \backslash \{0\}$. Therefore, $\frac{\|M_k x_{k-1}\|}{\|x_{k-1}\|}$ and $\|x_{k-1}\|$ are independent, and $\frac{\|M_k x_{k-1}\|}{\|x_{k-1}\|}$ has the same distribution as $\|M_k e_1\|$, where e_1 is the first basis vector. It follows that

$$\mathbb{E}\|x_k\|^p \leq \mathbb{E}\|M_k e_1\|^p \mathbb{E}\|x_{k-1}\|^p + \mathbb{E}\|q_k\|^p, \tag{C.12}$$

so that

$$\mathbb{E}\|x_k\|^p \leq h(p)\mathbb{E}\|x_{k-1}\|^p + \mathbb{E}\|q_1\|^p, \tag{C.13}$$

where $h(p) \in (0, 1)$. By iterating over k, we get

$$\mathbb{E}\|x_k\|^p \leq (h(p))^k \mathbb{E}\|x_0\|^p + \frac{1 - (h(p))^k}{1 - h(p)} \mathbb{E}\|q_1\|^p. \tag{C.14}$$

(ii) If the tail-index $\alpha > 1$, then for any $1 < p < \alpha$, by Lemma 23, for any $\epsilon > 0$, we have

$$\|x_k\|^p \leq (1+\epsilon)\|M_k x_{k-1}\|^p + \frac{(1+\epsilon)^{\frac{p}{p-1}} - (1+\epsilon)}{\left((1+\epsilon)^{\frac{1}{p-1}} - 1\right)^p} \|q_k\|^p, \tag{C.15}$$

which (similar as in (i)) implies that

$$\mathbb{E}\|x_k\|^p \leq (1+\epsilon)\mathbb{E}\|M_k e_1\|^p \mathbb{E}\|x_{k-1}\|^p + \frac{(1+\epsilon)^{\frac{p}{p-1}} - (1+\epsilon)}{\left((1+\epsilon)^{\frac{1}{p-1}} - 1\right)^p} \mathbb{E}\|q_k\|^p, \tag{C.16}$$

so that

$$\mathbb{E}\|x_k\|^p \leq (1+\epsilon)h(p)\mathbb{E}\|x_{k-1}\|^p + \frac{(1+\epsilon)^{\frac{p}{p-1}} - (1+\epsilon)}{\left((1+\epsilon)^{\frac{1}{p-1}} - 1\right)^p} \mathbb{E}\|q_1\|^p. \tag{C.17}$$

We choose $\epsilon > 0$ so that $(1+\epsilon)h(p) < 1$. By iterating over k, we get

$$\mathbb{E}\|x_k\|^p \leq ((1+\epsilon)h(p))^k \mathbb{E}\|x_0\|^p + \frac{1 - ((1+\epsilon)h(p))^k}{1 - (1+\epsilon)h(p)} \frac{(1+\epsilon)^{\frac{p}{p-1}} - (1+\epsilon)}{\left((1+\epsilon)^{\frac{1}{p-1}} - 1\right)^p} \mathbb{E}\|q_1\|^p. \tag{C.18}$$

The proof is complete. □

Remark 14. *In general, there is no closed-form expression for $\mathbb{E}\|q_1\|^p$ in Theorem 6. We provide an upper bound as follows. When $p > 1$, by Jensen's inequality, we can compute that*

$$\mathbb{E}\|q_1\|^p = \eta^p \mathbb{E}\left\|\frac{1}{b}\sum_{i=1}^{b} a_i y_i\right\|^p \leq \frac{\eta^p}{b}\sum_{i=1}^{b}\mathbb{E}\left\|a_i y_i\right\|^p = \eta^p \mathbb{E}\left[|y_1|^p \left\|a_1\right\|^p\right], \tag{C.19}$$

and when $p \leq 1$, by Lemma 23, we can compute that

$$\mathbb{E}\|q_1\|^p = \frac{\eta^p}{b^p}\mathbb{E}\left\|\sum_{i=1}^{b} a_i y_i\right\|^p \leq \frac{\eta^p}{b^p}\mathbb{E}\left[\left(\sum_{i=1}^{b}\|a_i y_i\|\right)^p\right] \leq \frac{\eta^p}{b^p}\sum_{i=1}^{b}\mathbb{E}\left\|a_i y_i\right\|^p = \eta^p \mathbb{E}\left[|y_1|^p \left\|a_1\right\|^p\right]. \tag{C.20}$$

Proof of Corollary 7. It follows from Theorem 6 by letting $k \to \infty$ and applying Fatou's lemma. □

C.6. Proof of Theorem 8, Corollary 9, Proposition 10 and Corollary 11

Proof of Theorem 8. For any $\nu_0, \tilde{\nu}_0 \in \mathcal{P}_p(\mathbb{R}^d)$, there exists a couple $x_0 \sim \nu_0$ and $\tilde{x}_0 \sim \tilde{\nu}_0$ independent of $(M_k, q_k)_{k\in\mathbb{N}}$ and $\mathcal{W}_p^p(\nu_0, \tilde{\nu}_0) = \mathbb{E}\|x_0 - \tilde{x}_0\|^p$. We define x_k and $\tilde{x}_k$ starting from x_0 and $\tilde{x}_0$ respectively, via the iterates

$$x_k = M_k x_{k-1} + q_k, \tag{C.21}$$
$$\tilde{x}_k = M_k \tilde{x}_{k-1} + q_k, \tag{C.22}$$

and let ν_k and $\tilde{\nu}_k$ denote the probability laws of x_k and $\tilde{x}_k$ respectively. For any $p < \alpha$, since $\mathbb{E}\|M_k\|^\alpha = 1$ and $\mathbb{E}\|q_k\|^\alpha < \infty$, we have $\nu_k, \tilde{\nu}_k \in \mathcal{P}_p(\mathbb{R}^d)$ for any k. Moreover, we have

$$x_k - \tilde{x}_k = M_k(x_{k-1} - \tilde{x}_{k-1}), \tag{C.23}$$

Due to spherical symmetry of the isotropic Gaussian distribution, the distribution of $\frac{\|M_k x\|}{\|x\|}$ does not depend on the choice of $x \in \mathbb{R}^d \backslash \{0\}$. Therefore, $\frac{\|M_k(x_{k-1}-\tilde{x}_{k-1})\|}{\|x_{k-1}-\tilde{x}_{k-1}\|}$ and $\|x_{k-1} - \tilde{x}_{k-1}\|$ are independent, and $\frac{\|M_k(x_{k-1}-\tilde{x}_{k-1})\|}{\|x_{k-1}-\tilde{x}_{k-1}\|}$ has the same distribution as $\|M_k e_1\|$, where e_1 is the first basis vector. It follows from (C.23) that

$$\mathbb{E}\|x_k - \tilde{x}_k\|^p \leq \mathbb{E}\left[\|M_k(x_{k-1} - \tilde{x}_{k-1})\|^p\right] = \mathbb{E}\left[\|M_k e_1\|^p\right] \mathbb{E}\left[\|x_{k-1} - \tilde{x}_{k-1}\|^p\right] = h(p)\mathbb{E}\left[\|x_{k-1} - \tilde{x}_{k-1}\|^p\right],$$

which by iterating implies that

$$\mathcal{W}_p^p(\nu_k, \tilde{\nu}_k) \leq \mathbb{E}\|x_k - \tilde{x}_k\|^p \leq (h(p))^k \mathbb{E}\|x_0 - \tilde{x}_0\|^p = (h(p))^k \mathcal{W}_p^p(\nu_0, \tilde{\nu}_0). \tag{C.24}$$

By letting $\tilde{\nu}_0 = \nu_\infty$, the probability law of the stationary distribution x_∞, we conclude that

$$\mathcal{W}_p(\nu_k, \nu_\infty) \leq \left((h(p))^{1/q}\right)^k \mathcal{W}_p(\nu_0, \nu_\infty). \tag{C.25}$$

Finally, notice that $1 \leq p < \alpha$, and therefore $h(p) < 1$. The proof is complete. □

Proof of Corollary 9. When $\eta\sigma^2 < \frac{2b}{d+b+1}$, by Proposition 5, the tail-index $\alpha > 2$, by taking $p = 2$, and using $h(2) = 1 - 2\eta\sigma^2 + \frac{\eta^2\sigma^4}{b}(d + b + 1) < 1$ (see Proposition 5), it follows from Theorem 8 that

$$\mathcal{W}_2(\nu_k, \nu_\infty) \leq \left(1 - 2\eta\sigma^2\left(1 - \frac{\eta\sigma^2}{2b}(d + b + 1)\right)\right)^{k/2} \mathcal{W}_2(\nu_0, \nu_\infty). \tag{C.26}$$

□

Remark 15. *Consider the case a_i are i.i.d. $\mathcal{N}(0, \sigma^2 I_d)$. In Theorem 6, Corollary 7 and Theorem 8, the key quantity is $h(p) \in (0, 1)$, where $p < \alpha$. We recall that*

$$h(p) = \mathbb{E}\left[\left(1 - \frac{2a}{b}X + \frac{a^2}{b^2}X^2 + \frac{a^2}{b^2}XY\right)^{p/2}\right], \tag{C.27}$$

where $a = \eta\sigma^2$, X, Y are independent chi-square random variables with degree of freedom b and $d-1$ respectively. The first-order approximation of $h(p)$ is given by

$$h(p) \sim 1 + \frac{p}{2}\mathbb{E}\left[-\frac{2a}{b}X + \frac{a^2}{b^2}X^2 + \frac{a^2}{b^2}XY\right] = 1 + \frac{p}{2}\left[-2a + \frac{a^2}{b}(b+2) + \frac{a^2}{b}(d-1)\right] < 1, \tag{C.28}$$

provided that $a = \eta\sigma^2 < \frac{2b}{d+b+1}$ which occurs if and only if $\alpha > 2$. In other words, when $\eta\sigma^2 < \frac{2b}{d+b+1}$, $\alpha > 2$ and

$$h(p) \sim 1 - p\eta\sigma^2\left(1 - \frac{\eta\sigma^2(b+d+1)}{2b}\right) < 1. \tag{C.29}$$

On the other hand, when $\eta\sigma^2 \geq \frac{2b}{d+b+1}$, $p < \alpha \leq 2$, and the second-order approximation of $h(p)$ is given by

$$h(p) \sim 1 + \frac{p}{2}\mathbb{E}\left[-\frac{2a}{b}X + \frac{a^2}{b^2}X^2 + \frac{a^2}{b^2}XY\right] + \frac{\frac{p}{2}(\frac{p}{2}-1)}{2}\mathbb{E}\left[\left(-\frac{2a}{b}X + \frac{a^2}{b^2}X^2 + \frac{a^2}{b^2}XY\right)^2\right]$$
$$= 1 + qa\left(\frac{a(b+d+1)}{2b} - 1\right) - \frac{2-p}{8}\mathbb{E}\left[\left(-\frac{2a}{b}X + \frac{a^2}{b^2}X^2 + \frac{a^2}{b^2}XY\right)^2\right],$$

and for small $a = \eta\sigma^2$ and large d,

$$\mathbb{E}\left[\left(-\frac{2a}{b}X + \frac{a^2}{b^2}X^2 + \frac{a^2}{b^2}XY\right)^2\right] \sim \frac{4a^2}{b}(b+2) + \frac{a^4}{b^3}(b+2)d^2 - \frac{4a^3}{b^2}(b+2)d, \tag{C.30}$$

and therefore with $a = \eta\sigma^2$,

$$h(p) \sim 1 - pa\left(\frac{-a(b+d+1)}{2b} + 1 + \frac{(2-p)a(b+2)}{2qb}\left(1 + \frac{a^2}{4b^2}d^2 - \frac{a}{b}d\right)\right) < 1, \tag{C.31}$$

provided that $1 \leq \frac{a(b+d+1)}{2b} < 1 + \frac{(2-p)a(b+2)}{2qb}\left(1 + \frac{a^2}{4b^2}d^2 - \frac{a}{b}d\right)$.

Proof of Proposition 10. First, we notice that it follows from Theorem 2 that $\mathbb{E}\|x_\infty\|^\alpha = \infty$. To see this, notice that $\lim_{t\to\infty} t^\alpha \mathbb{P}(e_1^T x_\infty > t) = e_\alpha(e_1)$, where e_1 is the first basis vector in $\mathbb{R}^d$, and $\mathbb{P}(\|x_\infty\| \geq t) \geq \mathbb{P}(e_1^T x_\infty \geq t)$, and thus

$$\mathbb{E}\|x_\infty\|^\alpha = \int_0^\infty t\mathbb{P}(\|x_\infty\|^\alpha \geq t)dt = \int_0^\infty t\mathbb{P}(\|x_\infty\| \geq t^{1/\alpha})dt = \infty. \tag{C.32}$$

By following the proof of Theorem 6 by letting $q = \alpha$ in the proof, one can show the following.

(i) If the tail-index $\alpha \leq 1$, then we have

$$\mathbb{E}\|x_\infty\|^\alpha \leq \mathbb{E}\|x_0\|^\alpha + k\mathbb{E}\|q_1\|^\alpha, \tag{C.33}$$

which grows linearly in k.

(ii) If the tail-index $\alpha > 1$, then for any $\epsilon > 0$, we have

$$\mathbb{E}\|x_k\|^\alpha \leq (1+\epsilon)^k\mathbb{E}\|x_0\|^\alpha + \frac{(1+\epsilon)^k - 1}{\epsilon}\frac{(1+\epsilon)^{\frac{\alpha}{\alpha-1}} - (1+\epsilon)}{\left((1+\epsilon)^{\frac{1}{\alpha-1}} - 1\right)^\alpha}\mathbb{E}\|q_1\|^\alpha = O(k), \tag{C.34}$$

which grows exponentially in k for any fixed $\epsilon > 0$. By letting $\epsilon \to 0$, we have

$$\mathbb{E}\|x_k\|^\alpha = (1+\epsilon)^k\mathbb{E}\|x_0\|^\alpha + (1+O(\epsilon))\left((1+\epsilon)^k - 1\right)\frac{(\alpha-1)^{\alpha-1}}{\epsilon^\alpha}\mathbb{E}\|q_1\|^\alpha.$$

Therefore, it holds for any sufficiently small $\epsilon > 0$ that,

$$\mathbb{E}\|x_k\|^\alpha \leq \frac{(1+\epsilon)^k}{\epsilon^\alpha}\left(\mathbb{E}\|x_0\|^\alpha + (\alpha-1)^{\alpha-1}\mathbb{E}\|q_1\|^\alpha\right).$$

We can optimize $\frac{(1+\epsilon)^k}{\epsilon^\alpha}$ over the choice of $\epsilon > 0$, and by choosing $\epsilon = \frac{\alpha}{k-\alpha}$, which goes to zero as k goes to ∞, we have $\frac{(1+\epsilon)^k}{\epsilon^\alpha} = (1 + \frac{\alpha}{k-\alpha})^k(\frac{k-\alpha}{\alpha})^\alpha = O(k^\alpha)$, and hence

$$\mathbb{E}\|x_k\|^\alpha = O(k^\alpha), \tag{C.35}$$

which grows polynomially in k. The proof is complete. □

Proof of Corollary 11. The result is obtained by a direct application of Theorem 1.15 in Mirek (2011) to the recursions (3.5) where it can be checked in a straightforward manner that the conditions for this theorem hold. □

D. Supporting Lemmas

In this section, we present a few supporting lemmas that are used in the proofs of the main results of the paper as well as the additional results in the Supplementary Document.

First, we recall that the iterates are given by $x_k = M_k x_{k-1} + q_k$, where (M_k, q_k) are i.i.d. and M_k is distributed as $I - \frac{\eta}{b}H$, where $H = \sum_{i=1}^b a_i a_i^T$ and q_k is distributed as $\frac{\eta}{b}\sum_{i=1}^b a_i y_i$, where $a_i \sim \mathcal{N}(0, \sigma^2 I_d)$ and y_i are i.i.d. satisfying the Assumptions **(A1)**–**(A3)**.

We can compute ρ and $h(s)$ as follows where ρ and $h(s)$ are defined by (3.7) and (3.6).

Lemma 16. *Under Assumptions **(A1)–(A3)**, ρ can be characterized as:*

$$\rho = \mathbb{E}\left[\log\left\|\left(I - \frac{\eta}{b}H\right)e_1\right\|\right], \tag{D.1}$$

and $h(s)$ can be characterized as:

$$h(s) = \mathbb{E}\left[\left\|\left(I - \frac{\eta}{b}H\right)e_1\right\|^s\right], \tag{D.2}$$

provided that $\rho < 0$. Furthermore, we have

$$\hat{\rho} = \mathbb{E}\log\left\|\left(I - \frac{\eta}{b}H\right)e_1\right\|, \quad \hat{h}(s) = \mathbb{E}\left[\left\|\left(I - \frac{\eta}{b}H\right)e_1\right\|^s\right] \tag{D.3}$$

where $\hat{\rho}$ and $\hat{h}(s)$ are defined in (3.10).

Proof. It is known that the Lyapunov exponent defined in (3.7) admits the alternative representation

$$\rho := \lim_{k\to\infty}\frac{1}{k}\log\|\tilde{x}_k\|, \tag{D.4}$$

where $\tilde{x}_k := \Pi_k \tilde{x}_0$ with $\Pi_k := M_k M_{k-1}\ldots M_1$ and $\tilde{x}_0 := x_0$ (see Equation (2) in Newman (1986)). We will compute the limit on the right-hand side of (D.4). First, we observe that due to spherical symmetry of the isotropic Gaussian distribution, the distribution of $\frac{\|M_k x\|}{\|x\|}$ does not depend on the choice of $x \in \mathbb{R}^d\backslash\{0\}$ and is i.i.d. over k with the same distribution as $\|Me_1\|$ where we chose $x = e_1$. This observation would directly imply the equality (D.3). In addition,

$$\frac{1}{k}\log\|\tilde{x}_k\| - \frac{1}{k}\log\|\tilde{x}_0\| = \frac{1}{k}\sum_{i=1}^k \log\frac{\|\tilde{x}_i\|}{\|\tilde{x}_{i-1}\|} = \frac{1}{k}\sum_{i=1}^k \log\frac{\|M_i\tilde{x}_{i-1}\|}{\|\tilde{x}_{i-1}\|}$$

is an average of i.i.d. random variables and by the law of large numbers we obtain

$$\rho = \lim_{k\to\infty}\frac{1}{k}\log\|\tilde{x}_k\| = \mathbb{E}\left[\log\left\|\left(I - \frac{\eta}{b}H\right)e_1\right\|\right].$$

From (D.4), we conclude that this proves (D.1).

It remains to prove (D.2). We consider the function

$$\tilde{h}(s) = \lim_{k\to\infty}\left(\mathbb{E}\frac{\|\tilde{x}_k\|^s}{\|\tilde{x}_0\|^s}\right)^{1/k},$$

where the initial point $\tilde{x}_0 = x_0$ is deterministic. In the rest of the proof, we will show that for $\rho < 0$, $h(s) = \tilde{h}(s)$ where $h(s)$ is given by (3.6) and $\tilde{h}(s)$ is equal to the right-hand side of (D.2); our proof is inspired by the approach of Newman (1986). We will first compute $\tilde{h}(s)$ and show that it is equal to the right-hand side of (D.2). Note that we can write

$$\frac{\|\tilde{x}_k\|^s}{\|\tilde{x}_0\|^s} = \prod_{i=1}^k \frac{\|M_i\tilde{x}_{i-1}\|^s}{\|\tilde{x}_{i-1}\|^s}.$$

This is a product of i.i.d. random variables with the same distribution as that of $\|Me_1\|^s$ due to the spherical symmetry of the input a_i. Therefore, we can write

$$\tilde{h}(s) = \lim_{k\to\infty}\left(\mathbb{E}\frac{\|\tilde{x}_k\|^s}{\|\tilde{x}_0\|^s}\right)^{1/k} = \lim_{k\to\infty}\left(\mathbb{E}\prod_{i=1}^{k}\|M_ie_1\|^s\right)^{1/k} = \mathbb{E}\left[\|Me_1\|^s\right] = \mathbb{E}\left[\left\|\left(I-\frac{\eta}{b}H\right)e_1\right\|^s\right], \tag{D.5}$$

where we used the fact that M_ie_1 are i.i.d. over i. It remains to show that $h(s)=\tilde{h}(s)$ for $\rho<0$. Note that $\frac{\|\tilde{x}_k\|^s}{\|\tilde{x}_0\|^s}\leq\|\Pi_k\|^s$, and therefore from the definition of $h(s)$ and $\tilde{h}(s)$, we have immediately

$$h(s)\geq\tilde{h}(s) \tag{D.6}$$

for any $s>0$. We will show that $h(s)\leq\tilde{h}(s)$ when $\rho<0$. We assume $\rho<0$. Then, Theorem 2 is applicable and there exists a stationary distribution x_∞ with a tail-index α such that $h(\alpha)=1$. We will show that $\tilde{h}(\alpha)=1$. First, the tail density admits the characterization (3.8), and therefore $x_\infty\in L_s$ for $s<\alpha$, i.e. the s-th moment of x_∞ is finite. Similarly due to (3.8), $x_\infty\notin L_s$ for $s>\alpha$. Since $h(\alpha)=1$, it follows from (D.6) that we have $\tilde{h}(\alpha)\leq 1$. However if $\tilde{h}(\alpha)<1$, then by the continuity of the $\tilde{h}$ function there exists ε such that $h(s)<1$ for every $s\in(\alpha-\varepsilon,\alpha+\varepsilon)\subset(0,1)$. From the definition of $\tilde{h}(s)$ then this would imply that $\mathbb{E}(\|x_k\|^s)\to 0$ for every $s\in(\alpha-\varepsilon,\alpha+\varepsilon)$. On the other hand, by following a similar argument to the proof technique of Corollary 7, it can be shown that the s-th moment of x_∞ has to be bounded,[8] which would be a contradiction with the fact that $x_\infty\notin L_s$ for $s>\alpha$. Therefore, $\tilde{h}(\alpha)\geq 1$. Since $h(\alpha)=1$, (D.6) leads to

$$h(\alpha)=\tilde{h}(\alpha)=1. \tag{D.7}$$

We observe that the function h is homogeneous in the sense that if the iterations matrices M_i are replaced by cM_i where $c>0$ is a real scalar, $h(s)$ will be replaced by $h_c(s):=c^sh(s)$. In other words, the function

$$h_c(s):=\lim_{k\to\infty}\left(\mathbb{E}\left\|(cM_k)(cM_{k-1})\dots(cM_1)\right\|^s\right)^{1/k} \tag{D.8}$$

clearly satisfies $h_c(s)=c^sh(s)$ by definition. A similar homogeneity property holds for $\tilde{h}(s)$: If the iterations matrices M_i are replaced by cM_i, then $\tilde{h}(s)$ will be replaced by $\tilde{h}_c(s):=c^s\tilde{h}(s)$. We will show that this homogeneity property combined with the fact that $h(\alpha)=\tilde{h}(\alpha)=1$ will force $h(s)=\tilde{h}(s)$ for any $s>0$. For this purpose, given $s>0$, we choose $c=1/\sqrt[s]{h(s)}$. Then, by considering input matrix cM_i instead of M_i and by following a similar argument which led to the identity (D.7), we can show that $h_c(s)=c^sh(s)=1$. Therefore, $\tilde{h}_c(s)=\tilde{h}_c(s)=1$. This implies directly $\tilde{h}(s)=h(s)$.

□

Next, we show the following property for the function h.

Lemma 17. *We have $h(0)=1$, $h'(0)=\rho$ and $h(s)$ is strictly convex in s.*

Proof. By the expression of $h(s)$ from Lemma 16, it is easy to check that $h(0)=1$. Moreover, we can compute that

$$h'(s)=\mathbb{E}\left[\log\left(\left\|\left(I-\frac{\eta}{b}H\right)e_1\right\|\right)\left\|\left(I-\frac{\eta}{b}H\right)e_1\right\|^s\right], \tag{D.9}$$

and thus $h'(0)=\rho$. Moreover, we can compute that

$$h''(s)=\mathbb{E}\left[\left(\log\left(\left\|\left(I-\frac{\eta}{b}H\right)e_1\right\|\right)\right)^2\left\|\left(I-\frac{\eta}{b}H\right)e_1\right\|^s\right]>0, \tag{D.10}$$

which implies that $h(s)$ is strictly convex in s. □

In the next result, we show that $\liminf_{s\to\infty}h(s)>1$. This property, together with Lemma 17 implies that if $\rho<0$, then there exists some $\alpha\in(0,\infty)$ such that $h(\alpha)=1$. Indeed, in the proof of Lemma 18, we will show that $\liminf_{s\to\infty}h(s)=\infty$.

Lemma 18. *We have* $\liminf_{s\to\infty}h(s)>1$.

[8] Note that the proof of Corollary 7 establishes first that x_∞ has a bounded s-th moment provided that $\tilde{h}(s)=\mathbb{E}\left[\|Me_1\|^s\right]<1$ and then cites Lemma 16 regarding the equivalence $h(s)=\tilde{h}(s)$.

Proof. We recall from Lemma 16 that

$$h(s) = \mathbb{E}\left\|\left(I - \frac{\eta}{b}H\right)e_1\right\|^s, \tag{D.11}$$

where e_1 is the first basis vector in $\mathbb{R}^d$ and $H = \sum_{i=1}^b a_i a_i^T$, and $a_i = (a_{i1}, \ldots, a_{id})$ are i.i.d. distributed as $\mathcal{N}(0, \sigma^2 I_d)$. We can compute that

$$\begin{aligned}
\mathbb{E}\left\|\left(I - \frac{\eta}{b}H\right)e_1\right\|^s &= \mathbb{E}\left(\left\|\left(I - \frac{\eta}{b}H\right)e_1\right\|^2\right)^{s/2} \\
&= \mathbb{E}\left[\left(e_1^T\left(I - \frac{\eta}{b}\sum_{i=1}^b a_i a_i^T\right)\left(I - \frac{\eta}{b}\sum_{i=1}^b a_i a_i^T\right)e_1\right)^{s/2}\right] \\
&= \mathbb{E}\left[\left(1 - \frac{2\eta}{b}e_1^T\sum_{i=1}^b a_i a_i^T e_1 + \frac{\eta^2}{b^2}e_1^T\sum_{i=1}^b a_i a_i^T\sum_{i=1}^b a_i a_i^T e_1\right)^{s/2}\right] \\
&= \mathbb{E}\left[\left(1 - \frac{2\eta}{b}\sum_{i=1}^b a_{i1}^2 + \frac{\eta^2}{b^2}\sum_{i=1}^b\sum_{j=1}^b(a_{i1}a_{j1} + \cdots + a_{id}a_{jd})a_{i1}a_{j1}\right)^{s/2}\right] \\
&= \mathbb{E}\left[\left(\left(1 - \frac{\eta}{b}\sum_{i=1}^b a_{i1}^2\right)^2 + \frac{\eta^2}{b^2}\sum_{i=1}^b\sum_{j=1}^b(a_{i2}a_{j2} + \cdots + a_{id}a_{jd})a_{i1}a_{j1}\right)^{s/2}\right] \\
&\geq \mathbb{E}\left[2^{s/2}1_{\frac{\eta^2}{b^2}\sum_{i=1}^b\sum_{j=1}^b(a_{i2}a_{j2}+\cdots+a_{id}a_{jd})a_{i1}a_{j1}\geq 2}\right] \\
&= 2^{s/2}\mathbb{P}\left(\frac{\eta^2}{b^2}\sum_{i=1}^b\sum_{j=1}^b(a_{i2}a_{j2} + \cdots + a_{id}a_{jd})a_{i1}a_{j1} \geq 2\right) \to \infty,
\end{aligned}$$

as $s \to \infty$. □

Next, we provide alternative formulas for $h(s)$ and ρ for the Gaussian data which is used for some technical proofs.

Lemma 19. *For any $s > 0$,*

$$h(s) = \mathbb{E}\left[\left(\left(1 - \frac{\eta\sigma^2}{b}X\right)^2 + \frac{\eta^2\sigma^4}{b^2}XY\right)^{s/2}\right],$$

and

$$\rho = \frac{1}{2}\mathbb{E}\left[\log\left(\left(1 - \frac{\eta\sigma^2}{b}X\right)^2 + \frac{\eta^2\sigma^4}{b^2}XY\right)\right],$$

where X, Y are independent and X is chi-square random variable with degree of freedom b and Y is a chi-square random variable with degree of freedom $(d-1)$.

Proof. We can compute that

$$
\begin{aligned}
h(s) &= \mathbb{E}\left[\left(1-\frac{2\eta\sigma^2}{b}\sum_{i=1}^{b} z_{i1}^2+\frac{\eta^2\sigma^4}{b^2}\sum_{i=1}^{b}\sum_{j=1}^{b}(z_{i1}z_{j1}+\cdots+z_{id}z_{jd})z_{i1}z_{j1}\right)^{s/2}\right] \\
&= \mathbb{E}\left[\left(1-\frac{2\eta\sigma^2}{b}\sum_{i=1}^{b} z_{i1}^2+\frac{\eta^2\sigma^4}{b^2}\sum_{i=1}^{b}\sum_{j=1}^{b}\left(z_{i1}^2z_{j1}^2+z_{i1}z_{j1}\sum_{k=2}^{d}z_{ik}z_{jk}\right)\right)^{s/2}\right] \\
&= \mathbb{E}\left[\left(1-\frac{2\eta\sigma^2}{b}\sum_{i=1}^{b} z_{i1}^2+\frac{\eta^2\sigma^4}{b^2}\left(\sum_{i=1}^{b} z_{i1}^2\right)^2+\frac{\eta^2\sigma^4}{b^2}\sum_{k=2}^{d}\left(\sum_{i=1}^{b} z_{i1}z_{ik}\right)^2\right)^{s/2}\right] \\
&= \mathbb{E}\left[\left(\left(1-\frac{\eta\sigma^2}{b}\sum_{i=1}^{b} z_{i1}^2\right)^2+\frac{\eta^2\sigma^4}{b^2}\sum_{k=2}^{d}\left(\sum_{i=1}^{b} z_{i1}z_{ik}\right)^2\right)^{s/2}\right],
\end{aligned}
$$

where z_{ij} are i.i.d. $N(0,1)$ random variables. Note that conditional on z_{i1}, $1\le i\le b$,

$$
\sum_{i=1}^{b} z_{i1}z_{ik}\sim\mathcal{N}\left(0,\sum_{i=1}^{b} z_{i1}^2\right), \tag{D.12}
$$

are i.i.d. for $k=2,\ldots,d$. Therefore, we have

$$
\begin{aligned}
h(s) &= \mathbb{E}\left[\left(\left(1-\frac{\eta\sigma^2}{b}\sum_{i=1}^{b} z_{i1}^2\right)^2+\frac{\eta^2\sigma^4}{b^2}\sum_{k=2}^{d}\left(\sum_{i=1}^{b} z_{i1}z_{ik}\right)^2\right)^{s/2}\right] \\
&= \mathbb{E}\left[\left(\left(1-\frac{\eta\sigma^2}{b}\sum_{i=1}^{b} z_{i1}^2\right)^2+\frac{\eta^2\sigma^4}{b^2}\sum_{i=1}^{b} z_{i1}^2\sum_{k=2}^{d} x_k^2\right)^{s/2}\right],
\end{aligned}
$$

where x_k are i.i.d. $N(0,1)$ independent of z_{i1}, $i=1,\ldots,b$. Hence, we have

$$
\begin{aligned}
h(s) &= \mathbb{E}\left[\left(\left(1-\frac{\eta\sigma^2}{b}\sum_{i=1}^{b} z_{i1}^2\right)^2+\frac{\eta^2\sigma^4}{b^2}\sum_{i=1}^{b} z_{i1}^2\sum_{k=2}^{d} x_k^2\right)^{s/2}\right] \\
&= \mathbb{E}\left[\left(\left(1-\frac{\eta\sigma^2}{b}X\right)^2+\frac{\eta^2\sigma^4}{b^2}XY\right)^{s/2}\right],
\end{aligned}
$$

where X, Y are independent and X is chi-square random variable with degree of freedom b and Y is a chi-square random variable with degree of freedom $(d-1)$.

Similarly, we can compute that

$$
\begin{aligned}
\rho &= \frac{1}{2}\mathbb{E}\left[\log\left[\left(1-\frac{\eta\sigma^2}{b}\sum_{i=1}^{b} z_{i1}^2\right)^2+\frac{\eta^2\sigma^4}{b^2}\sum_{i=1}^{b}\sum_{j=1}^{b} z_{i1}z_{j1}\sum_{k=2}^{d} z_{ik}z_{jk}\right]\right] \\
&= \frac{1}{2}\mathbb{E}\left[\log\left[\left(1-\frac{\eta\sigma^2}{b}\sum_{i=1}^{b} z_{i1}^2\right)^2+\frac{\eta^2\sigma^4}{b^2}\sum_{k=2}^{d}\left(\sum_{i=1}^{b} z_{i1}z_{ik}\right)^2\right]\right] \\
&= \frac{1}{2}\mathbb{E}\left[\log\left(\left(1-\frac{\eta\sigma^2}{b}X\right)^2+\frac{\eta^2\sigma^4}{b^2}XY\right)\right],
\end{aligned}
$$

where X, Y are independent and X is chi-square random variable with degree of freedom b and Y is a chi-square random variable with degree of freedom $(d-1)$. The proof is complete. □

In the next result, we show that the inverse of M exists with probability 1, and provide an upper bound result, which will be used to prove Lemma 21.

Lemma 20. *Let a_i satisfy Assumption **(A1)**. Then, M^{-1} exists with probability 1. Moreover, we have*

$$\mathbb{E}\left[\left(\log^+ \|M^{-1}\|\right)^2\right] \leq 8.$$

Proof. Note that M is a continuous random matrix, by the assumption on the distribution of a_i. Therefore,

$$\mathbb{P}\left(M^{-1} \text{ does not exist}\right) = \mathbb{P}(\det M = 0) = 0. \tag{D.13}$$

Note that the singular values of M^{-1} are of the form $|1-\frac{\eta}{b}\sigma_H|^{-1}$ where σ_H is a singular value of H and we have

$$\left(\log^+ \left\|M^{-1}\right\|\right)^2 = \begin{cases} 0 & \text{if} \quad \frac{\eta}{b}H \succ 2I\,, \\ \left(\left\|(I-\frac{\eta}{b}H)^{-1}\right\|\right)^2 & \text{if} \quad 0 \preceq \frac{\eta}{b}H \preceq 2I\,. \end{cases} \tag{D.14}$$

We consider two cases $0 \preceq \frac{\eta}{b}H \preceq I$ and $I \preceq \frac{\eta}{b}H \preceq 2I$. We compute the conditional expectations for each case:

$$\mathbb{E}\left[\left(\log^+ \left\|M^{-1}\right\|\right)^2 \mid 0 \preceq \frac{\eta}{b}H \preceq I\right] = \mathbb{E}\left[\left(\log\left\|\left(I-\frac{\eta}{b}H\right)^{-1}\right\|\right)^2 \mid 0 \preceq \frac{\eta}{b}H \prec I\right] \tag{D.15}$$

$$\leq \mathbb{E}\left[\left(2\frac{\eta}{b}\|H\|\right)^2 \mid 0 \preceq \frac{\eta}{b}H \preceq I\right] \tag{D.16}$$

$$\leq 4\,, \tag{D.17}$$

where in the first inequality we used the fact that

$$\log(I-X)^{-1} \preceq 2X \tag{D.18}$$

for a symmetric positive semi-definite matrix X satisfying $0 \preceq X \prec I$ (the proof of this fact is analogous to the proof of the scalar inequality $\log(\frac{1}{1-x}) \leq 2x$ for $0 \leq x < 1$). By a similar computation,

$$\begin{aligned}
\mathbb{E}\left[(\log^+ \|M^{-1}\|)^2 \mid I \preceq \frac{\eta}{b}H \preceq 2I\right] &= \mathbb{E}\left[\log\left\|\left(I-\frac{\eta}{b}H\right)^{-1}\right\| \mid I \preceq \frac{\eta}{b}H \prec 2I\right] \\
&= \mathbb{E}\left[\log^2\left\|\left(\frac{\eta}{b}H\right)^{-1}\left[I-\left(\frac{\eta}{b}H\right)^{-1}\right]^{-1}\right\| \mid I \preceq \frac{\eta}{b}H \prec 2I\right] \\
&\leq \mathbb{E}\left[\log^2\left(\left\|\left(\frac{\eta}{b}H\right)^{-1}\right\| \cdot \left\|\left[I-\left(\frac{\eta}{b}H\right)^{-1}\right]^{-1}\right\|\right) \mid I \preceq \frac{\eta}{b}H \prec 2I\right] \\
&\leq \mathbb{E}\left[\log^2\left(\left\|\left[I-\left(\frac{\eta}{b}H\right)^{-1}\right]^{-1}\right\|\right) \mid I \preceq \frac{\eta}{b}H \prec 2I\right] \\
&= \mathbb{E}\left[\log^2\left(\left\|\left[I-\left(\frac{\eta}{b}H\right)^{-1}\right]^{-1}\right\|\right) \mid \frac{1}{2}I \preceq \left(\frac{\eta}{b}H\right)^{-1} \prec I\right],
\end{aligned}$$

where in the last inequality we used the fact that $(\frac{\eta}{b}H)^{-1} \preceq I$ for $I \preceq \frac{\eta}{b}H \prec 2I$. If we apply the inequality (D.18) to the last inequality for the choice of $X = (\frac{\eta}{b}H)^{-1}$, we obtain

$$\mathbb{E}\left[\log^2\left\|\left[I-\left(\frac{\eta}{b}H\right)^{-1}\right]^{-1}\right\| \mid \frac{1}{2}I \preceq \left(\frac{\eta}{b}H\right)^{-1} \prec I\right] \leq \mathbb{E}\left[\left\|2\left(\frac{\eta}{b}H\right)^{-1}\right\|^2 \mid \frac{1}{2}I \preceq \left(\frac{\eta}{b}H\right)^{-1} \prec I\right] \leq 4\,. \tag{D.19}$$

Combining (D.17) and (D.19), it follows from (D.14) that $\mathbb{E}\log^+ \|M^{-1}\| \leq 8$. □

In the next result, we show that a certain expected value that involves the moments and logarithm of $\|M\|$, and logarithm of $\|M^{-1}\|$ is finite, which is used in the proof of Theorem 2.

Lemma 21. *Let a_i satisfy Assumption (A1). Then,*

$$\mathbb{E}\left[\|M\|^{\alpha}\left(\log^{+}\|M\|+\log^{+}\|M^{-1}\|\right)\right]<\infty\,.$$

Proof. Note that $M=I-\frac{\eta}{b}H$, where $H=\sum_i^b a_i a_i^T$ in distribution. Therefore for any $s>0$,

$$\mathbb{E}[\|M\|^s]=\mathbb{E}\left[\left\|I-\frac{\eta}{b}\sum_{i=1}^{b}a_i a_i^T\right\|^s\right]\leq\mathbb{E}\left[\left(1+\frac{\eta}{b}\sum_{i=1}^{b}\|a_i\|^2\right)^s\right]<\infty\,, \tag{D.20}$$

since all the moments of a_i are finite by the Assumption **(A1)**. This implies that

$$\mathbb{E}\left[\|M\|^{\alpha}\left(\log^{+}\|M\|\right)\right]<\infty\,.$$

By Cauchy-Schwarz inequality,

$$\mathbb{E}\left[\|M\|^{\alpha}\left(\log^{+}\|M^{-1}\|\right)\right]\leq\left(\mathbb{E}\left[\|M\|^{2\alpha}\right]\mathbb{E}\left[\left(\log^{+}\|M^{-1}\|\right)^2\right]\right)^{1/2}<\infty,$$

where we used Lemma 20. □

In the next result, we show a convexity result, which is used in the proof of Theorem 4 to show that the tail-index α is strictly decreasing in stepsize η and variance σ^2.

Lemma 22. *For any given positive semi-definite symmetric matrix H fixed, the function $F_H:[0,\infty)\to\mathbb{R}$ defined as*

$$F_H(a):=\|(I-aH)\,e_1\|^s$$

is convex for $s\geq 1$. It follows that for given b and d with $\tilde{H}:=\frac{1}{b}\sum_{i=1}^{b}a_i a_i^T$, the function

$$h(a,s):=\mathbb{E}\left[F_{\tilde{H}}(a)\right]=\mathbb{E}\left\|\left(I-a\tilde{H}\right)e_1\right\|^s \tag{D.21}$$

is a convex function of a for a fixed $s\geq 1$.

Proof. We consider the case $s\geq 1$ and consider the function

$$G_H(a):=\|(I-aH)\,e_1\|\,,$$

and show that it is convex for $H\succeq 0$ and it is strongly convex for $H\succ 0$ over the interval $[0,\infty)$. Let $a_1,a_2\in[0,\infty)$ be different points, i.e. $a_1\neq a_2$. It follows from the subadditivity of the norm that

$$G_H\left(\frac{a_1+a_2}{2}\right)=\left\|\left(I-\frac{a_1+a_2}{2}H\right)e_1\right\|\leq\left\|\left(\frac{I}{2}-\frac{a_1}{2}H\right)e_1\right\|+\left\|\left(\frac{I}{2}-\frac{a_2}{2}H\right)e_1\right\|=\frac{1}{2}G_H(a_1)+\frac{1}{2}G_H(a_2)\,,$$

which implies that $G_H(a)$ is a convex function. On the other hand, the function $g(x)=x^s$ is convex for $s\geq 1$ on the positive real axis, therefore the composition $g(G_H(a))$ is also convex for any H fixed. Since the expectation of random convex functions is also convex, we conclude that $h(s)$ is also convex. □

The next result is used in the proof of Theorem 6 to bound the moments of the iterates.

Lemma 23. *(i) Given $0<p\leq 1$, for any $x,y\geq 0$,*

$$(x+y)^p\leq x^p+y^p. \tag{D.22}$$

(ii) Given $p>1$, for any $x,y\geq 0$, and any $\epsilon>0$,

$$(x+y)^p\leq(1+\epsilon)x^p+\frac{(1+\epsilon)^{\frac{p}{p-1}}-(1+\epsilon)}{\left((1+\epsilon)^{\frac{1}{p-1}}-1\right)^p}y^p. \tag{D.23}$$

Proof. (i) If $y = 0$, then $(x + y)^p \leq x^p + y^p$ trivially holds. If $y > 0$, it is equivalent to show that

$$\left(\frac{x}{y} + 1\right)^p \leq \left(\frac{x}{y}\right)^p + 1, \tag{D.24}$$

which is equivalent to show that

$$(x + 1)^p \leq x^p + 1, \qquad \text{for any } x \geq 0. \tag{D.25}$$

Let $F(x) := (x + 1)^p - x^p - 1$ and $F(0) = 0$ and $F'(x) = p(x + 1)^{p-1} - px^{p-1} \leq 0$ since $p \leq 1$, which shows that $F(x) \leq 0$ for every $x \geq 0$.

(ii) If $y = 0$, then the inequality trivially holds. If $y > 0$, by doing the transform $x \mapsto x/y$ and $y \mapsto 1$, it is equivalent to show that for any $x \geq 0$,

$$(1 + x)^p \leq (1 + \epsilon)x^p + \frac{(1 + \epsilon)^{\frac{p}{p-1}} - (1 + \epsilon)}{\left((1 + \epsilon)^{\frac{1}{p-1}} - 1\right)^p}. \tag{D.26}$$

To show this, we define

$$F(x) := (1 + x)^p - (1 + \epsilon)x^p, \qquad x \geq 0. \tag{D.27}$$

Then $F'(x) = p(1 + x)^{p-1} - p(1 + \epsilon)x^{p-1}$ so that $F'(x) \geq 0$ if $x \leq ((1 + \epsilon)^{\frac{1}{p-1}} - 1)^{-1}$, and $F'(x) \leq 0$ if $x \geq ((1 + \epsilon)^{\frac{1}{p-1}} - 1)^{-1}$. Thus,

$$\max_{x \geq 0} F(x) = F\left(\frac{1}{(1 + \epsilon)^{\frac{1}{p-1}} - 1}\right) = \frac{(1 + \epsilon)^{\frac{p}{p-1}} - (1 + \epsilon)}{\left((1 + \epsilon)^{\frac{1}{p-1}} - 1\right)^p}. \tag{D.28}$$

The proof is complete. □