A Technical Background

Wasserstein metric. For any $p \geq 1$, define $\mathcal{P}_p(\mathbb{R}^d)$ as the space consisting of all the Borel probability measures ν on $\mathbb{R}^d$ with the finite p-th moment (based on the Euclidean norm). For any two Borel probability measures $\nu_1, \nu_2 \in \mathcal{P}_p(\mathbb{R}^d)$, we define the standard p-Wasserstein metric as (Villani, 2009):

$$\mathcal{W}_p(\nu_1, \nu_2) := \left(\inf \mathbb{E}\left[\|Z_1 - Z_2\|^p\right]\right)^{1/p},$$

where the infimum is taken over all joint distributions of the random variables Z_1, Z_2 with marginal distributions ν_1, ν_2.

B Technical Results

B.1 Stochastic Gradient Descent with Constant Stepsizes

In this section, let us recall some technical results from Gürbüzbalaban et al. (2021) for the SGD with constant stepsizes. When the stepsizes $\eta_k \equiv \eta$ are constant, the SGD iterates are given by

$$x_{k+1} = x_k - \eta \tilde{\nabla} f_{k+1}(x_k), \tag{16}$$

where $\eta > 0$ is the stepsize and $\tilde{\nabla} f_k(x) := \frac{1}{b}\sum_{i\in\Omega_k} \nabla f_i(x)$. We first observe that SGD (16) is an iterated random recursion of the form

$$x_k = \Psi(x_{k-1}, \Omega_k), \tag{17}$$

where the map $\Psi : \mathbb{R}^d \times \mathcal{S} \to \mathbb{R}^d$, $\mathcal{S}$ denotes the set of all subsets of $\{1, 2, \ldots, n\}$ and Ω_k is random and i.i.d. If we write $\Psi_\Omega(x) = \Psi(x, \Omega)$ for simplicity where Ω has the same distribution as Ω_k, and assume that the random map Ψ_Ω is Lipschitz on average, i.e. $\mathbb{E}[L_\Omega] < \infty$ with $L_\Omega := \sup_{x,y\in\mathbb{R}^d} \frac{\|\Psi_\Omega(x) - \Psi_{\Omega,\eta}(y)\|}{\|x-y\|}$, and is mean-contractive, i.e. if $\mathbb{E}\log(L_\Omega) < 0$ then it can be shown under further technical assumptions that the distribution of the iterates converges to a unique stationary distribution x_∞ geometrically fast (Diaconis & Freedman, 1999). We recall the following result from Gürbüzbalaban et al. (2021) that characterize the tail-index for x_∞.

Theorem 8 (Theorem 1 in Gürbüzbalaban et al. (2021), see also Mirek (2011))**.** *Assume stationary solution to $x_k = \Psi_{\Omega_k}(x_{k-1})$ exists and:*

(i) There exists a random matrix $M(\Omega)$ and a random variable $B(\Omega) > 0$ such that for a.e. Ω, $|\Psi_\Omega(x) - M(\Omega)x| \leq B(\Omega)$ for every x;

(ii) The conditional law of $\log|M(\Omega)|$ given $M(\Omega) \neq 0$ is non-arithmetic; i.e. its support is not equal to $a\mathbb{Z}$ for any scalar a where $\mathbb{Z}$ is the set of integers.

(iii) There exists $\alpha_c > 0$ such that $\mathbb{E}|M(\Omega)|^{\alpha_c} = 1$, $\mathbb{E}|B(\Omega)|^{\alpha_c} < \infty$ and

$$\mathbb{E}[|M(\Omega)|^{\alpha_c} \log^+ |M(\Omega)|] < \infty,$$

where $\log^+(x) := \max(\log(x), 0)$.

Then, it holds that $\lim_{t\to\infty} t^{\alpha_c}\mathbb{P}(|x_\infty| > t) = c_{0,c}$ for some constant $c_{0,c} > 0$.

When the objective is quadratic, it is possible to characterize the tail-index α_c in a more explicit way and also go beyond the one-dimensional case. For the quadratic objective, we can rewrite SGD iterations (16) as

$$x_{k+1} = (I - (\eta/b)H_{k+1})\, x_k + q_{k+1}, \tag{18}$$

where $H_k := \sum_{i\in\Omega_k} a_i a_i^T$ and $q_k := \frac{\eta}{b}\sum_{i\in\Omega_k} y_i$. Let us introduce

$$h_c(s) := \lim_{k\to\infty} \left(\mathbb{E}\|M_k M_{k-1} \ldots M_1\|^s\right)^{1/k}, \tag{19}$$

where $M_k := I - \frac{\eta}{b} H_k$, which arises in stochastic matrix recursions (see e.g. Buraczewski et al. (2014)) where $\|\cdot\|$ denotes the matrix 2-norm (i.e. largest singular value of a matrix). Since $\mathbb{E}\|M_k\|^s < \infty$ for all k and $s > 0$, we have $h_c(s) < \infty$. Let us also define

$$\rho_c := \lim_{k\to\infty} (2k)^{-1} \log\left(\text{largest eigenvalue of } \Pi_k^T \Pi_k\right) , \tag{20}$$

where $\Pi_k := M_k M_{k-1} \dots M_1$. In (20), the quantity ρ_c is called the top Lyapunov exponent of the stochastic recursion (5). Furthermore, if ρ_c exists and is negative, it can be shown that a stationary distribution of the recursion (5) exists. Indeed, we have the following result from Gürbüzbalaban et al. (2021) that characterizes the tail-index for the stationary distribution.

Theorem 9 (Theorem 2 in Gürbüzbalaban et al. (2021)). *Consider the SGD iterations (5). If $\rho_c < 0$ and there exists a unique positive α_c such that $h_c(\alpha_c) = 1$, where h_c and ρ_c are defined in (19) and (20), then (5) admits a unique stationary solution x_∞ and the SGD iterations converge to x_∞ in distribution, where the distribution of x_∞ satisfies*

$$\lim_{t\to\infty} t^{\alpha_c} \mathbb{P}\left(u^T x_\infty > t\right) = e_{\alpha_c}(u) , \quad u \in \mathbb{S}^{d-1} , \tag{21}$$

for some positive and continuous function e_α on $\mathbb{S}^{d-1}$.

In general, the tail-index α_c does not have a simple formula since $h_c(s)$ function lacks a simple expression. A lower bound $\hat{\alpha}_c \le \alpha_c$ holds where $\hat{\alpha}_c$ is the unique positive solution to $\hat{h}_c(\hat{\alpha}_c) = 1$, where $\hat{h}_c(s) := \mathbb{E}\left[\left\|I - \frac{\eta}{b} H_1\right\|^s\right]$, provided that $\hat{\rho}_c := \mathbb{E}\log\left\|I - \frac{\eta}{b} H_1\right\| < 0$.

B.2 Stochastic Gradient Descent with i.i.d. Stepsizes

In this section, we consider the stochastic gradient descent method with i.i.d. stepsizes. We first observe that SGD (3) is an iterated random recursion of the form

$$x_k = \Psi(x_{k-1}, \Omega_k, \eta_k), \tag{22}$$

where the map $\Psi : \mathbb{R}^d \times \mathcal{S} \times \mathbb{R}_+ \to \mathbb{R}^d$, $\mathcal{S}$ denotes the set of all subsets of $\{1, 2, \dots, n\}$ and Ω_k is random and i.i.d. When the stepsize η_k are i.i.d., if we write $\Psi_{\Omega,\eta}(x) = \Psi(x, \Omega, \eta)$ for simplicity where (Ω, η) has the same distribution as (Ω_k, η_k), and assume that the random map $\Psi_{\Omega,\eta}$ is Lipschitz on average, i.e. $\mathbb{E}[L_{\Omega,\eta}] < \infty$ with $L_{\Omega,\eta} := \sup_{x,y\in\mathbb{R}^d} \frac{\|\Psi_{\Omega,\eta}(x) - \Psi_{\Omega,\eta}(y)\|}{\|x-y\|}$, and is mean-contractive, i.e. if $\mathbb{E}\log(L_{\Omega,\eta}) < 0$ then it can be shown under further technical assumptions that the distribution of the iterates converges to a unique stationary distribution x_∞ geometrically fast (Diaconis & Freedman, 1999). We have the following result that characterizes the tail-index under such assumptions for dimension $d = 1$, which can be derived from Mirek (2011) by adapting it to our setting (see also Buraczewski et al. (2016).

Theorem 10 (Adaptation of Mirek (2011)). *Assume stationary solution to*

$$x_k = \Psi_{\Omega_k,\eta_k}(x_{k-1})$$

exists and: (i) There exists a random matrix $M(\Omega, \eta)$ and a random variable $B(\Omega, \eta) > 0$ such that for a.e. Ω, η, $|\Psi_{\Omega,\eta}(x) - M(\Omega, \eta)x| \le B(\Omega, \eta)$ for every x; (ii) The conditional law of $\log|M(\Omega, \eta)|$ given $M(\Omega, \eta) \neq 0$ is non-arithmetic ;i.e. its support is not equal to $a\mathbb{Z}$ for any scalar a where $\mathbb{Z}$ is the set of integers. (iii) There exists $\alpha > 0$ such that $\mathbb{E}|M(\Omega, \eta)|^\alpha = 1$, $\mathbb{E}|B(\Omega, \eta)|^\alpha < \infty$ and $\mathbb{E}[|M(\Omega, \eta)|^\alpha \log^+ |M(\Omega, \eta)|] < \infty$, where $\log^+(x) := \max(\log(x), 0)$. Then, it holds that $\lim_{t\to\infty} t^\alpha \mathbb{P}(|x_\infty| > t) = c_0$ for some constant $c_0 > 0$.

Theorem 10 shows that heavy tails arises for general losses that has an almost linear growth outside compact sets, however it does not characterize how the tail-index α depends on the stepsize, furthermore it is highly non-trivial how to verify its assumptions in general. Also, it works only in the one dimensional setting; Mirek (2011) studies more general d but requires the matrices $M(\Omega, \eta)$ form an orthogonal group which is not satisfied by SGD in general. This motivates us to study more structured losses in high dimensional settings where more insights can be obtained. We next study quadratic f which corresponds to linear regression to obtain finer characterizations. In this case, we have the iterates:

$$x_{k+1} = \left(I - \frac{\eta_{k+1}}{b} H_{k+1}\right) x_k + q_{k+1} , \tag{23}$$

where $H_k := \sum_{i\in\Omega_k} a_i a_i^T$ are i.i.d. Hessian matrices and $q_k := \frac{\eta_k}{b}\sum_{i\in\Omega_k} y_i$, and η_k are i.i.d. with a distribution supported on an interval $[\eta_l, \eta_u]$, where $\eta_u > \eta_l > 0$. Under some mild conditions, by following the same arguments as in Gürbüzbalaban et al. (2021), x_k converges to x_∞ in distribution, where x_∞ exhibits the heavy-tail behavior with the tail-index α which is the unique positive value such that $h(\alpha) = 1$, where

$$h(s) := \lim_{k\to\infty} \left(\mathbb{E}\|M_k M_{k-1}\dots M_1\|^s\right)^{1/k}, \tag{24}$$

provided that

$$\rho := \lim_{k\to\infty} (2k)^{-1} \log\left(\text{largest eigenvalue of } \Pi_k^T \Pi_k\right) < 0, \tag{25}$$

where $\Pi_k := M_k M_{k-1}\dots M_1$.

Similar to the SGD with constant stepsize case (Theorem 9), we have the following result that states that the iterations converge to a stationary distribution with heavy tails.

Theorem 11. *Consider the SGD iterations with i.i.d. stepsizes (23). If $\rho < 0$ and there exists a unique positive α such that $h(\alpha) = 1$, where h and ρ are defined in (24)-(25), then (23) admits a unique stationary solution x_∞ and the SGD iterations with cyclic stepsizes converge to x_∞ in distribution, where the distribution of x_∞ satisfies*

$$\lim_{t\to\infty} t^\alpha \mathbb{P}\left(u^T x_\infty > t\right) = e_\alpha(u), \quad u \in \mathbb{S}^{d-1}, \tag{26}$$

for some positive and continuous function e_α on $\mathbb{S}^{d-1}$.

Theorem 11 says the tail-index α is the unique positive value such that $h(\alpha) = 1$ provided that $\rho < 0$. However, the expressions of $h(s)$ and ρ are not very explicit. Under Assumption **(A3)**, we can simplify the expressions for $h(s)$ and ρ (see Lem. 7 and Lem. 8 in the Appendix). Moreover, under Assumption **(A3)**, we have the following result which characterizes the dependence of the tail-index α on the batch-size and the dimension.

Theorem 12. *Assume **(A3)** holds and $\rho < 0$. Then we have: (i) the tail-index α is strictly increasing in batch-size b provided $\alpha \geq 1$. (ii) The tail-index α is strictly decreasing in dimension d.*

In Theorem 12, we showed that that smaller batch-sizes lead to (smaller tail-index) heavier tail provided that $\alpha \geq 1$ and higher dimension leads to (smaller tail-index) heavier tail. On the other hand, it is also natural to conjecture that the tail-index gets smaller if the distribution of η is more spread out. The formalize our intuition, we assume that the stepsize is uniformly distributed with mean $\bar\eta$ and range R, i.e. the stepsize is uniformly distributed on the interval $(\bar\eta - R, \bar\eta + R)$. Next, we show that the tail-index decreases as the range R increases provided the tail-index α is greater than 1.

Theorem 13. *Assume **(A3)** holds and $\rho < 0$. Assume η is uniformly distributed on $(\bar\eta - R, \bar\eta + R)$. Then, the tail-index α is decreasing in the range R provided that $\alpha \geq 1$.*

Under Assumption **(A3)**, our next result characterizes the tail-index α depending on the choice of the batch-size b, the variance σ^2, which determines the curvature around the minimum and the stepsize; in particular we provide critical threshold such that the stationary distribution will become heavy tailed with an infinite variance.

Proposition 6. *Assume **(A3)** holds. Define*

$$c := 1 - 2\mathbb{E}[\eta]\sigma^2 + \frac{\mathbb{E}[\eta^2]\sigma^4}{b}(d + b + 1). \tag{27}$$

The following holds: (i) There exists $\delta > 0$ such that for any $1 < c < 1 + \delta$, tail-index $0 < \alpha < 2$. (ii) If $c = 1$, tail-index $\alpha = 2$. (iii) If $c < 1$, then tail-index $\alpha > 2$.

Theorem 11 is of asymptotic nature which characterizes the stationary distribution x_∞ of SGD iterations with a tail-index α. Next, we provide non-asymptotic moment bounds for x_k at each k-th iterate for p such that $h(p) < 1$.

Lemma 1. *Assume **(A3)** holds.*

(i) For any $p \leq 1$ and $h(p) < 1$,

$$\mathbb{E}\|x_k\|^p \leq (h(p))^k \mathbb{E}\|x_0\|^p + \frac{1-(h(p))^k}{1-h(p)} \mathbb{E}\|q_1\|^p. \tag{28}$$

(ii) For any $p > 1$, $\epsilon > 0$ and $(1+\epsilon)h(p) < 1$,

$$\mathbb{E}\|x_k\|^p \leq ((1+\epsilon)h(p))^k \mathbb{E}\|x_0\|^p + \frac{1-((1+\epsilon)h(p))^k}{1-(1+\epsilon)h(p)} \frac{(1+\epsilon)^{\frac{p}{p-1}} - (1+\epsilon)}{\left((1+\epsilon)^{\frac{1}{p-1}} - 1\right)^p} \mathbb{E}\|q_1\|^p. \tag{29}$$

Next, we will study the speed of convergence of the k-th iterate x_k to its stationary distribution x_∞ in the Wasserstein metric $\mathcal{W}_p$ for any p such that $h(p) < 1$.

Theorem 14. *Assume **(A3)** holds. Let ν_k, ν_∞ denote the probability laws of x_k and x_∞ respectively. Then*

$$\mathcal{W}_p(\nu_k, \nu_\infty) \leq (h(p))^{k/p} \mathcal{W}_p(\nu_0, \nu_\infty), \tag{30}$$

for any $p \geq 1$ and $h(p) < 1$, where the convergence rate $(h(p))^{1/p} \in (0,1)$.

When the tail-index $\alpha > 2$, by Lemma 1, the second moments of the iterates x_k are finite, in which case central limit theorem (CLT) says that if the cumulative sum of the iterates $S_K := \sum_{k=1}^K x_k$ is scaled properly, the resulting distribution is Gaussian. In the case where $\alpha < 2$, the variance of the iterates is not finite; however in this case, we derive the following generalized CLT (GCLT) which says if the iterates are properly scaled, the limit will be an α-stable distribution. This is stated in a more precise manner as follows.

Corollary 1. *Assume **(A3)** holds and the conditions of Theorem 11 are satisfied. Then, we have the following:*

(i) If $\alpha \in (0,1) \cup (1,2)$, then there is a sequence $d_K = d_K(\alpha)$ and a function $C_\alpha : \mathbb{S}^{d-1} \mapsto \mathbb{C}$ such that as $K \to \infty$ the random variables $K^{-\frac{1}{\alpha}}(S_K - d_K)$ converge in law to the α-stable random variable with characteristic function $\Upsilon_\alpha(tv) = \exp(t^\alpha C_\alpha(v))$, for $t > 0$ and $v \in \mathbb{S}^{d-1}$.

(ii) If $\alpha = 1$, then there are functions $\xi, \tau : (0,\infty) \mapsto \mathbb{R}$ and $C_1 : \mathbb{S}^{d-1} \mapsto \mathbb{C}$ such that as $K \to \infty$ the random variables $K^{-1}S_K - K\xi\left(K^{-1}\right)$ converge in law to the random variable with characteristic function $\Upsilon_1(tv) = \exp\left(tC_1(v) + it\langle v, \tau(t)\rangle\right)$, for $t > 0$ and $v \in \mathbb{S}^{d-1}$.

(iii) If $\alpha = 2$, then there is a sequence $d_K = d_K(2)$ and a function $C_2 : \mathbb{S}^{d-1} \mapsto \mathbb{R}$ such that as $K \to \infty$ the random variables $(K \log K)^{-\frac{1}{2}}(S_K - d_K)$ converge in law to the random variable with characteristic function $\Upsilon_2(tv) = \exp\left(t^2 C_2(v)\right)$, for $t > 0$ and $v \in \mathbb{S}^{d-1}$.

(iv) If $\alpha \in (0,1)$, then $d_K = 0$, and if $\alpha \in (1,2]$, then $d_K = K\bar{x}$, where $\bar{x} = \int_{\mathbb{R}^d} x\nu_\infty(dx)$.

In addition to its evident theoretical interest, Corollary 1 has also an important practical implication: estimating the tail-index of a *generic* heavy-tailed distribution is a challenging problem (see e.g. Clauset et al. (2009); Goldstein et al. (2004); Bauke (2007)); however, for the specific case of α-stable distributions, accurate and computationally efficient estimators, which *do not* require the knowledge of the functions C_α, τ, ξ, have been proposed (Mohammadi et al., 2015). Thanks to Corollary 1, we will be able to use such estimators in our numerical experiments in Section 5.

B.3 Technical Results for SGD with Cyclic Stepsizes

In this section, we provide some additional technical results for SGD with cyclic stepsizes.

If we assume that the random map $\Psi^{(m)}$ is Lipschitz on average, i.e. $\mathbb{E}\left[L^{(m)}\right] < \infty$ with $L^{(m)} := \sup_{x,y \in \mathbb{R}^d} \frac{\|\Psi^{(m)}(x) - \Psi^{(m)}(y)\|}{\|x-y\|}$, and is mean-contractive, i.e. if $\mathbb{E}\log\left(L^{(m)}\right) < 0$ then it can be shown under further technical assumptions that the iterates converges to a unique stationary distribution x_∞ geometrically fast (Diaconis & Freedman, 1999).

First, we have the following analogue of Theorem 1 which is a special case of Theorem 1.

Theorem 15 (Adaptation of Mirek (2011)). *Assume stationary solution to (12) exists and: (i) There exists a random variable $M^{(m)}$ and a random variable $B^{(m)} > 0$ such that a.s. $|\Psi^{(m)}(x) - M^{(m)}x| \leq B^{(m)}$ for every x; (ii) The conditional law of $\log|M^{(m)}|$ given $M^{(m)} \neq 0$ is non-arithmetic; i.e. its support is not equal to $a\mathbb{Z}$ for any scalar a where $\mathbb{Z}$ is the set of integers. (iii) There exists $\alpha^{(m)} > 0$ such that $\mathbb{E}[|M^{(m)}|^{\alpha^{(m)}}] = 1$, $\mathbb{E}[|B^{(m)}|^{\alpha^{(m)}}] < \infty$ and $\mathbb{E}[|M^{(m)}|^{\alpha^{(m)}} \log^+ |M^{(m)}|] < \infty$, where $\log^+(x) := \max(\log(x), 0)$. Then there exists some constant $c_0^{(m)} > 0$ such that $\lim_{t\to\infty} t^{\alpha^{(m)}} \mathbb{P}(|x_\infty| > t) = c_0^{(m)}$.*

Next, we consider the setting of the linear regression. We can iterate the SGD from (5) to obtain $x_{(k+1)m} = M_{k+1}^{(m)} x_{km} + q_{k+1}^{(m)}$, where $M_{k+1}^{(m)}$ is defined in (14) and $q_{k+1}^{(m)} := \sum_{i=km+1}^{(k+1)m} \left(I - \frac{\eta_{(k+1)m}}{b} H_{(k+1)m}\right)\left(I - \frac{\eta_{(k+1)m-1}}{b} H_{(k+1)m-1}\right)\cdots\left(I - \frac{\eta_{i+1}}{b} H_{i+1}\right) q_i$. We showed in Theorem 5 that x_∞ has heavy tails with a tail-index $\alpha^{(m)}$ and further properties of the tail-index $\alpha^{(m)}$ were obtained under Assumption **(A3)** in Theorem 7.

Under Assumption **(A3)**, our next result characterizes the tail-index $\alpha^{(m)}$ depending on the choice of the batch-size b, the variance σ^2, which determines the curvature around the minimum and the stepsize; in particular we provide critical threshold such that the stationary distribution will become heavy tailed with an infinite variance.

Proposition 7. *Assume **(A3)** holds. Define*

$$c^{(m)} := \prod_{i=1}^{m} \left(1 - 2\eta_i\sigma^2 + \frac{\eta_i^2\sigma^4}{b}(d+b+1)\right). \tag{31}$$

The following holds: (i) There exists $\delta > 0$ such that for any $1 < c^{(m)} < 1+\delta$, tail-index $0 < \alpha^{(m)} < 2$. (ii) If $c^{(m)} = 1$, tail-index $\alpha^{(m)} = 2$. (iii) If $c^{(m)} < 1$, then tail-index $\alpha^{(m)} > 2$.

In Section 3, Theorem 5 is of asymptotic nature which characterizes the stationary distribution x_∞ of SGD iterations with a tail-index $\alpha^{(m)}$. Next, we provide non-asymptotic moment bounds for x_{mk} at each mk-th iterate for p such that $h^{(m)}(p) < 1$.

Lemma 2. *Assume **(A3)** holds.*

(i) For any $p \leq 1$ and $h^{(m)}(p) < 1$,

$$\mathbb{E}\|x_{mk}\|^p \leq \left(h^{(m)}(p)\right)^k \mathbb{E}\|x_0\|^p + \frac{1 - (h^{(m)}(p))^k}{1 - h^{(m)}(p)} \mathbb{E}\left\|q_1^{(m)}\right\|^p. \tag{32}$$

(ii) For any $p > 1$, $\epsilon > 0$ and $(1+\epsilon)h^{(m)}(p) < 1$,

$$\mathbb{E}\|x_{mk}\|^p \leq \left((1+\epsilon)h^{(m)}(p)\right)^k \mathbb{E}\|x_0\|^p + \frac{1 - ((1+\epsilon)h^{(m)}(p))^k}{1 - (1+\epsilon)h^{(m)}(p)} \frac{(1+\epsilon)^{\frac{p}{p-1}} - (1+\epsilon)}{\left((1+\epsilon)^{\frac{1}{p-1}} - 1\right)^p} \mathbb{E}\left\|q_1^{(m)}\right\|^p. \tag{33}$$

Next, we will study the speed of convergence of the mk-th iterate x_{mk} to its stationary distribution x_∞ in the Wasserstein metric $\mathcal{W}_p$ for any p such that $h^{(m)}(p) < 1$.

Theorem 16. *Assume **(A3)** holds. Let ν_{mk}, ν_∞ denote the probability laws of x_{mk} and x_∞ respectively. Then*

$$\mathcal{W}_p(\nu_{mk}, \nu_\infty) \leq \left(h^{(m)}(p)\right)^{k/p} \mathcal{W}_p(\nu_0, \nu_\infty), \tag{34}$$

for any $p \geq 1$ and $h^{(m)}(p) < 1$, where the convergence rate $\left(h^{(m)}(p)\right)^{1/p} \in (0,1)$.

Similar as in Corollary 1, we have the following generalized CLT (GCLT) result for $S_K^{(m)} := \sum_{k=1}^K x_{mk}$ when it is scaled properly so that the limit will be an alpha-stable distribution.

Corollary 2. *Assume **(A3)** holds and the conditions of Theorem 5 are satisfied. Then, we have the following:*

(i) If $\alpha^{(m)} \in (0,1) \cup (1,2)$*, then there is a sequence* $d_K = d_K(\alpha^{(m)})$ *and a function* $C_{\alpha^{(m)}} : \mathbb{S}^{d-1} \mapsto \mathbb{C}$ *such that as* $K \to \infty$ *the random variables* $K^{-\frac{1}{\alpha^{(m)}}} \left(S_K^{(m)} - d_K \right)$ *converge in law to the* $\alpha^{(m)}$*-stable random variable with characteristic function* $\Upsilon_{\alpha^{(m)}}(tv) = \exp(t^{\alpha^{(m)}} C_{\alpha^{(m)}}(v))$*, for* $t > 0$ *and* $v \in \mathbb{S}^{d-1}$*.*

(ii) If $\alpha^{(m)} = 1$*, then there are functions* $\xi, \tau : (0,\infty) \mapsto \mathbb{R}$ *and* $C_1 : \mathbb{S}^{d-1} \mapsto \mathbb{C}$ *such that as* $K \to \infty$ *the random variables* $K^{-1} S_K^{(m)} - K\xi\left(K^{-1}\right)$ *converge in law to the random variable with characteristic function* $\Upsilon_1(tv) = \exp\left(tC_1(v) + it\langle v, \tau(t)\rangle\right)$*, for* $t > 0$ *and* $v \in \mathbb{S}^{d-1}$*.*

(iii) If $\alpha^{(m)} = 2$*, then there is a sequence* $d_K = d_K(2)$ *and a function* $C_2 : \mathbb{S}^{d-1} \mapsto \mathbb{R}$ *such that as* $K \to \infty$ *the random variables* $(K \log K)^{-\frac{1}{2}} \left(S_K^{(m)} - d_K \right)$ *converge in law to the random variable with characteristic function* $\Upsilon_2(tv) = \exp\left(t^2 C_2(v)\right)$*, for* $t > 0$ *and* $v \in \mathbb{S}^{d-1}$*.*

(iv) If $\alpha^{(m)} \in (0,1)$*, then* $d_K = 0$*, and if* $\alpha^{(m)} \in (1,2]$*, then* $d_K = K\bar{x}$*, where* $\bar{x} = \int_{\mathbb{R}^d} x \nu_\infty(dx)$*.*

For the specific case of α-stable distributions, accurate and computationally efficient estimators, which *do not* require the knowledge of the functions C_α, τ, ξ, have been proposed (Mohammadi et al., 2015). Thanks to Corollary 2, we will be able to use such estimators in our numerical experiments in Section 5.

B.4 Technical Results for SGD with Markovian Stepsizes

In this section, we provide some additional technical results for SGD with Markovian stepsizes. In Section 3, we restricted our discussions to the finite state space. In the following section, we provide some technical results for the general state space.

B.4.1 Markovian Stepsizes with General State Space

When the objective is quadratic, we recall that the iterates of the SGD are given by:

$$x_{k+1} = M_{k+1} x_k + q_{k+1}. \tag{35}$$

In this case, $M_k = I - \frac{\eta_k}{b} H_k$, where η_k is a stationary Markov chain with a common distribution η supported on an interval $[\eta_l, \eta_u]$, where $\eta_u > \eta_l > 0$, and H_k are i.i.d. Hessian matrices.

To the best of our knowledge, there is no general stochastic linear recursion theory for Markovian coefficients, except for some special cases, e.g. with heavy-tail coefficient (Hay et al., 2011). Nevertheless, using a direct approach, we can obtain a lower bound for the tail-index for the limit of the SGD with Markovian stepsizes as follows. Since η_k is stationary and H_k are i.i.d., M_k is stationary, we have:

$$h^{(r)}(s) \leq \hat{h}^{(g)}(s) := \mathbb{E}\left[\|M_1\|^s\right] = \mathbb{E}\left[\left\|I - \frac{\eta_1}{b} H_1\right\|^s\right], \qquad \text{for any } s \geq 0, \tag{36}$$

where $\hat{h}^{(g)}(s)$ is an upper bound on $h^{(r)}(s)$ (defined in (10)) and we also define

$$\hat{\rho}^{(g)} := \mathbb{E}\left[\log \|M_1\|\right] = \mathbb{E}\left[\log \left\|I - \frac{\eta_1}{b} H_1\right\|\right]. \tag{37}$$

While having a grasp of the exact value of the tail-index for the stationary distribution of x_∞ is difficult when the stepsizes are Markovian, in the next result, based on a technical lemma (Lem. 3 in the Appendix) for the moment bounds for x_k, we can characterize a lower bound $\hat{\alpha}^{(g)}$ for the tail-index to control how heavy tailed SGD iterates can be, in the sense that we have $\mathbb{P}(\|x_\infty\| > t) < C_p/t^p$ for some constant C_p as long as $p < \hat{\alpha}^{(g)}$.

Proposition 8. *Let* $\hat{\alpha}^{(g)}$ *be the unique positive value such that* $\hat{h}^{(g)}(\hat{\alpha}^{(g)}) = 1$*, provided that* $\hat{\rho}^{(g)} < 0$*, where* $\hat{h}^{(g)}$ *and* $\hat{\rho}^{(g)}$ *are defined in* (36)-(37)*. Then, for any* $p \leq 1$ *and* $\hat{h}^{(g)}(p) < 1$*,*

$$\mathbb{P}(\|x_\infty\| \geq t) \leq \frac{1}{1 - \hat{h}^{(g)}(p)} \frac{\mathbb{E}\|q_1\|^p}{t^p}, \qquad \textit{for any } t > 0, \tag{38}$$

and for any $p > 1$, $\epsilon > 0$ *and* $(1+\epsilon)\hat{h}^{(g)}(p) < 1$,

$$\mathbb{P}(\|x_\infty\| \geq t) \leq \frac{1}{1-(1+\epsilon)\hat{h}^{(g)}(p)} \frac{(1+\epsilon)^{\frac{p}{p-1}} - (1+\epsilon)}{\left((1+\epsilon)^{\frac{1}{p-1}} - 1\right)^p} \frac{\mathbb{E}\|q_1\|^p}{t^p}, \qquad \text{for any } t > 0, \tag{39}$$

Next, in the following result, we discuss how the tail-index (lower bound) estimate $\hat{\alpha}^{(g)}$ depends on the batch-size and how it compares with the tail-index (lower bound) estimate $\hat{\alpha}_c$ with constant stepsize $\mathbb{E}[\eta]$.

Theorem 17. *(i) The lower bound for the tail-index* $\hat{\alpha}^{(g)}$ *is strictly increasing in batch-size* b *provided that* $\hat{\alpha}^{(g)} \geq 1$. *(ii) The lower bound for the tail-index* $\hat{\alpha}^{(g)}$ *is strictly less than the lower bound for the tail-index* $\hat{\alpha}_c$ *with constant stepsize* $\mathbb{E}[\eta]$ *provided that* $\hat{\alpha}^{(g)} \geq 1$.

Under Assumption **(A3)**, our next result characterizes the tail-index $\alpha^{(r)}$ depending on the choice of the batch-size b, the variance σ^2, which determines the curvature around the minimum and the stepsize; in particular we provide critical threshold such that the stationary distribution will become heavy tailed with an infinite variance.

Proposition 9. *Assume* ***(A3)*** *holds. Define*

$$c^{(r)} := \mathbb{E}\left[\prod_{i=1}^{r_1}\left(1 - 2\eta_i\sigma^2 + \frac{\eta_i^2\sigma^4}{b}(d+b+1)\right)\right]. \tag{40}$$

The following holds: (i) There exists $\delta > 0$ *such that for any* $1 < c^{(r)} < 1+\delta$, *tail-index* $0 < \alpha^{(r)} < 2$. *(ii) If* $c^{(r)} = 1$, *tail-index* $\alpha^{(r)} = 2$. *(iii) If* $c^{(r)} < 1$, *then tail-index* $\alpha^{(r)} > 2$.

In Section 3, Theorem 2 is of asymptotic nature which characterizes the stationary distribution x_∞ of SGD iterations with a tail-index $\alpha^{(r)}$. Next, we provide non-asymptotic moment bounds for the finite iterates when $\hat{h}^{(g)}(p) < 1$, where we recall that the definition of $\hat{h}^{(g)}(s)$ from (36).

Lemma 3. *(i) For any* $p \leq 1$ *and* $\hat{h}^{(g)}(p) < 1$,

$$\mathbb{E}\|x_k\|^p \leq \left(\hat{h}^{(g)}(p)\right)^k \mathbb{E}\|x_0\|^p + \frac{1-(\hat{h}^{(g)}(p))^k}{1-\hat{h}^{(g)}(p)}\mathbb{E}\|q_1\|^p. \tag{41}$$

(ii) For any $p > 1$, $\epsilon > 0$ *and* $(1+\epsilon)\hat{h}^{(g)}(p) < 1$,

$$\mathbb{E}\|x_k\|^p \leq \left((1+\epsilon)\hat{h}^{(g)}(p)\right)^k \mathbb{E}\|x_0\|^p + \frac{1-((1+\epsilon)\hat{h}^{(g)}(p))^k}{1-(1+\epsilon)\hat{h}^{(g)}(p)} \frac{(1+\epsilon)^{\frac{p}{p-1}} - (1+\epsilon)}{\left((1+\epsilon)^{\frac{1}{p-1}} - 1\right)^p}\mathbb{E}\|q_1\|^p. \tag{42}$$

Next, we provide the convergence rate to the stationary distribution in p-Wasserstein distance provided that $\hat{h}^{(g)}(p) < 1$.

Theorem 18. *Let* ν_k, ν_∞ *denote the probability laws of* x_k *and* x_∞ *respectively. Then*

$$\mathcal{W}_p(\nu_k, \nu_\infty) \leq \left(\hat{h}^{(g)}(p)\right)^{k/p} \mathcal{W}_p(\nu_0, \nu_\infty), \tag{43}$$

for any $p \geq 1$ *and* $\hat{h}^{(g)}(p) < 1$, *where the convergence rate* $(\hat{h}^{(g)}(p))^{1/p} \in (0,1)$.

B.4.2 Markovian Stepsizes with Finite State Space

In this section, we provide additional technical results for SGD with Markovian stepsizes with finite state space. It is natural to conjecture that the tail-index gets smaller if the distribution of η is more spread out. The formalize our intuition, we assume that the stepsize is uniformly distributed with mean $\bar{\eta}$. Without loss

of generality, we assume that K is an odd number, and the stepsizes are equally spaced with distance $\delta > 0$ in the sense that the state space of the stepsizes is given by

$$\{\bar{\eta}, \bar{\eta} \pm \delta, \bar{\eta} \pm 2\delta, \ldots, \bar{\eta} \pm (K-1)\delta/2\}\,. \tag{44}$$

Then, the range of the stepsizes is $(K-1)\delta$, which increases as either δ or K increases. The stationary distribution of the simple random walk is uniform on the set (44). The following result shows that if the range of stepsizes increases, the tails gets heavier in the sense that tails admit a smaller lower bound $\hat{\alpha}^{(g)}$, which is the unique positive value such that $\hat{h}^{(g)}(\hat{\alpha}^{(g)}) = 1$, where $\hat{h}^{(g)}(s) := \mathbb{E}\left[\|M_1\|^s\right] = \mathbb{E}\left[\left\|I - \frac{\eta_1}{b} H_1\right\|^s\right]$ (see Prop. 8 in the Appendix for detailed discussions).

Theorem 19. *Assume the stationary distribution of the Markovian stepsizes is uniform on the set (44). Then, the lower bound for the tail-index $\hat{\alpha}^{(g)}$ is decreasing in the range, i.e. decreasing in δ and K, provided that $\hat{\alpha}^{(g)} \geq 1$.*

Next, we assume that the range $\frac{K-1}{2}\delta = R$ is fixed, so that given K, we have $\delta = \frac{2R}{K-1}$. For simplicity, we assume that $K = 2^n + 1$ for some $n \in \mathbb{N}$ such that the state space of the stepsizes is:

$$\left\{\bar{\eta}, \bar{\eta} \pm \left(R2^{-(n-1)}\right), \bar{\eta} \pm 2\left(R2^{-(n-1)}\right), \ldots, \bar{\eta} \pm 2^{n-1}\left(R2^{-(n-1)}\right)\right\}. \tag{45}$$

Note that the larger the value of $K = 2^n + 1$, the finer the grid for stepsizes is. We are interested in studying how the lower bound for the tail-index $\hat{\alpha}^{(g)}$ depends on $K = 2^n + 1$. We have the following result that shows that the lower bound for the tail-index $\hat{\alpha}^{(g)}$ is increasing in the $K = 2^n + 1$.

Theorem 20. *Assume the stationary distribution of the Markovian stepsizes is uniform on the set (45). Then, $\hat{\alpha}^{(g)}$ is increasing in the $K = 2^n + 1$ provided that $\hat{\alpha}^{(g)} \geq 1$.*

This result shows that the finer the grid for stepsizes is, the larger the lower bound for the tail-index so that the tail gets lighter, that is, the lower bound on the tail gets lighter. In Theorem 20, if we write $\hat{\alpha}_n^{(g)} := \hat{\alpha}^{(g)}$ to emphasize the dependence on n, then we showed that $\hat{\alpha}_n^{(g)}$ is increasing in $n \in \mathbb{N}$. However, we also showed in Theorem 17 that for any $n \in \mathbb{N}$, $\hat{\alpha}_n^{(g)}$ is less than the lower bound $\hat{\alpha}_c$ for the tail-index for the SGD with the constant stepsize $\bar{\eta}$.

The following result shows that Markovian stepsizes in fact can lead to heavier tails (in the sense of lower bound for the tail-index $\hat{\alpha}^{(g)}$ values) compared to cyclic stepsizes.

Proposition 10. *Assume the stationary distribution of the Markovian stepsizes is uniform on the set (6). Then, the lower bound for the tail-index $\hat{\alpha}^{(g)}$ is strictly less than the lower bound for the tail-index $\hat{\alpha}^{(m)}$ for the SGD with cyclic stepsizes.*

Theorem 2 in the main text is of asymptotic nature which characterizes the stationary distribution x_∞ of SGD iterations with a tail-index $\alpha^{(r)}$. Next, we provide non-asymptotic moment bounds for x_{r_k} at each r_k-th iterate, and also for the limit x_∞.

Lemma 4. *Assume **(A3)** holds.*

(i) For any $p \leq 1$ and $h^{(r)}(p) < 1$,

$$\mathbb{E}\|x_{r_k}\|^p \leq \left(h^{(r)}(p)\right)^k \mathbb{E}\|x_0\|^p + \frac{1 - (h^{(r)}(p))^k}{1 - h^{(r)}(p)} \mathbb{E}\left\|q_1^{(r)}\right\|^p. \tag{46}$$

(ii) For any $p > 1$, $\epsilon > 0$ and $(1+\epsilon)h^{(r)}(p) < 1$,

$$\mathbb{E}\|x_{r_k}\|^p \leq \left((1+\epsilon)h^{(r)}(p)\right)^k \mathbb{E}\|x_0\|^p + \frac{1 - ((1+\epsilon)h^{(r)}(p))^k}{1 - (1+\epsilon)h^{(r)}(p)} \frac{(1+\epsilon)^{\frac{p}{p-1}} - (1+\epsilon)}{\left((1+\epsilon)^{\frac{1}{p-1}} - 1\right)^p} \mathbb{E}\left\|q_1^{(r)}\right\|^p. \tag{47}$$

Next, we will study the speed of convergence of the SGD to its stationary distribution x_∞ in the Wasserstein metric $\mathcal{W}_p$ for any p such that $h^{(r)}(p) < 1$.

Theorem 21. *Assume **(A3)** holds. Let ν_{r_k}, ν_∞ denote the probability laws of x_{r_k} and x_∞ respectively. Then*

$$\mathcal{W}_p(\nu_{r_k}, \nu_\infty) \leq \left(h^{(r)}(p)\right)^{k/p} \mathcal{W}_p(\nu_0, \nu_\infty), \tag{48}$$

for any $p \geq 1$ and $h^{(r)}(p) < 1$, where the convergence rate $\left(h^{(r)}(p)\right)^{1/p} \in (0,1)$.

Similar as in Corollary 1, we have the following generalized CLT (GCLT) result for $S_K^{(r)} := \sum_{k=1}^K x_{r_k}$ when it is scaled properly so that the limit will be an alpha-stable distribution.

Corollary 3. *Assume **(A3)** holds and the conditions of Theorem 5 are satisfied. Then, we have the following:*

(i) If $\alpha^{(r)} \in (0,1) \cup (1,2)$, then there is a sequence $d_K = d_K(\alpha^{(r)})$ and a function $C_{\alpha^{(r)}} : \mathbb{S}^{d-1} \mapsto \mathbb{C}$ such that as $K \to \infty$ the random variables $K^{-\frac{1}{\alpha^{(r)}}} \left(S_K^{(r)} - d_K\right)$ converge in law to the $\alpha^{(r)}$-stable random variable with characteristic function $\Upsilon_{\alpha^{(r)}}(tv) = \exp(t^{\alpha^{(r)}} C_{\alpha^{(r)}}(v))$, for $t > 0$ and $v \in \mathbb{S}^{d-1}$.

(ii) If $\alpha^{(r)} = 1$, then there are functions $\xi, \tau : (0,\infty) \mapsto \mathbb{R}$ and $C_1 : \mathbb{S}^{d-1} \mapsto \mathbb{C}$ such that as $K \to \infty$ the random variables $K^{-1} S_K^{(r)} - K\xi\left(K^{-1}\right)$ converge in law to the random variable with characteristic function $\Upsilon_1(tv) = \exp\left(tC_1(v) + it\langle v, \tau(t)\rangle\right)$, for $t > 0$ and $v \in \mathbb{S}^{d-1}$.

(iii) If $\alpha^{(r)} = 2$, then there is a sequence $d_K = d_K(2)$ and a function $C_2 : \mathbb{S}^{d-1} \mapsto \mathbb{R}$ such that as $K \to \infty$ the random variables $(K \log K)^{-\frac{1}{2}} \left(S_K^{(r)} - d_K\right)$ converge in law to the random variable with characteristic function $\Upsilon_2(tv) = \exp\left(t^2 C_2(v)\right)$, for $t > 0$ and $v \in \mathbb{S}^{d-1}$.

(iv) If $\alpha^{(r)} \in (0,1)$, then $d_K = 0$, and if $\alpha^{(r)} \in (1,2]$, then $d_K = K\bar{x}$, where $\bar{x} = \int_{\mathbb{R}^d} x\nu_\infty(dx)$.

For the specific case of α-stable distributions, accurate and computationally efficient estimators, which *do not* require the knowledge of the functions C_α, τ, ξ, have been proposed (Mohammadi et al., 2015). Thanks to Corollary 3, we will be able to use such estimators in our numerical experiments in Section 5.

We end the discussions of this section by providing some additional technical results concerning the stationary distribution of the Markovian stepsizes, and provide a more explicit formula for the function $h^{(r)}(s)$ that plays a central role of defining the tail-index $\alpha^{(r)}$. We recall from (6) that the state space is given by

$$\{\eta_1, \eta_2, \ldots, \eta_m, \eta_{m+1}\} = \{c_1, c_2, \ldots, c_{K-1}, c_K, c_{K-1}, \ldots, c_2, c_1\},$$

where $m = 2K - 2$. The stepsize goes from η_1 to η_2 with probability 1 and it goes from η_K to η_{K-1} with probability 1. In between, for any $i = 2, 3, \ldots, K-1, K+1, \ldots, m$, the stepsize goes from η_i to η_{i+1} with probability p and from η_i to η_{i-1} with probability $1-p$ with the understanding that $\eta_{m+1} := \eta_1$. Therefore, $p = 1$ reduces to the case of cyclic stepsizes. The Markov chain exhibits a unique stationary distribution $\pi_i := \mathbb{P}(\eta_0 = \eta_i)$ that is characterized in the following lemma.

Lemma 5. *The Markov chain exhibits a unique stationary distribution $\pi_i := \mathbb{P}(\eta_0 = \eta_i)$, where*

$$\pi_1 = (1-p)\frac{p-1}{2p-1}\left(\frac{1-p}{p}\right)^{K-2}\pi_m + \frac{p^2}{2p-1}\pi_m, \tag{49}$$

and for any $2 \leq i \leq K-1$,

$$\pi_i = \frac{p-1}{2p-1}\left(\frac{1-p}{p}\right)^{K-i}\pi_m + \frac{p}{2p-1}\pi_m, \tag{50}$$

and

$$\pi_K = \frac{p(p-1)}{2p-1}\left(\frac{1-p}{p}\right)^{m-K}\pi_m + \frac{p^2}{2p-1}\pi_m, \tag{51}$$

and for any $K+1 \leq i \leq m$,

$$\pi_i = \frac{p-1}{2p-1}\left(\frac{1-p}{p}\right)^{m-i}\pi_m + \frac{p}{2p-1}\pi_m, \tag{52}$$

where

$$\pi_m = \left(\frac{4p^3 + 2(m-3)p^2 - (m-3)p - 1}{(2p-1)^2} + \frac{2p^3}{(2p-1)^2} \left(\frac{1-p}{p} \right)^{K+1} + \frac{2p(p-1)^2}{(2p-1)^2} \left(\frac{1-p}{p} \right)^{m-K} \right)^{-1}. \quad (53)$$

Next, let us provide an analytic expression for $h^{(r)}(s)$. Under Assumption **(A3)**, we define:

$$h^{(r)}(s; \eta_i, \eta_j) := \mathbb{E}_{\eta_0 = \eta_i} \left[\prod_{i=1}^{r_1(\tau_j)} \mathbb{E}_H \left[\left\| \left(I - \frac{\eta_i}{b} H \right) e_1 \right\|^s \right] \right], \quad (54)$$

where $r_1(\tau_j) := \inf\{k \geq 1 : \eta_k = \eta_j\}$, and we have the following result.

When the initialization η_0 follows the stationary distribution, i.e., $\mathbb{P}(\eta_0 = \eta_i) = \pi_i$, we conclude that

$$h^{(r)}(s) = \sum_{i=1}^{m} \mathbb{P}(\eta_0 = \eta_i) h^{(r)}(s; \eta_i, \eta_i) = \sum_{i=1}^{m} \pi_i h^{(r)}(s; \eta_i, \eta_i), \quad (55)$$

where π_i are given in Lemma 5 and $h^{(r)}(s; \eta_i, \eta_i)$ is defined in (54). In the next proposition, we compute out $h^{(r)}(s; \eta_i, \eta_i)$ explicitly and hence we obtain an explicit formula for $h^{(r)}(s)$ using (55) and Lemma 5.

Proposition 11. *Under Assumption **(A3)**, for any $1 \leq i, j \leq m$,*

$$h^{(r)}(s; \eta_i, \eta_j) = \left((I - Q^j)^{-1} p^j \right)_i, \quad (56)$$

where $p^j := [p_{1j}, p_{2j}, \ldots, p_{mj}]^T$, where for any $i = 2, \ldots, K-1, K+1, \ldots, m$

$$p_{ij} := p\mathbb{E}_H \left[\left\| \left(I - \frac{\eta_{i+1}}{b} H \right) e_1 \right\|^s \right] 1_{j=i+1} + (1-p)\mathbb{E}_H \left[\left\| \left(I - \frac{\eta_{i-1}}{b} H \right) e_1 \right\|^s \right] 1_{j=i-1}, \quad (57)$$

and

$$p_{1j} := \mathbb{E}_H \left[\left\| \left(I - \frac{\eta_2}{b} H \right) e_1 \right\|^s \right] 1_{j=2}, \qquad p_{Kj} := \mathbb{E}_H \left[\left\| \left(I - \frac{\eta_{K+1}}{b} H \right) e_1 \right\|^s \right] 1_{j=K+1}, \quad (58)$$

and $Q^j := (Q^j_{i\ell})_{1 \leq i, \ell \leq m}$ such that for any $i = 2, \ldots, K-1, K+1, \ldots, m$

$$Q^j_{i\ell} := p\mathbb{E}_H \left[\left\| \left(I - \frac{\eta_{i+1}}{b} H \right) e_1 \right\|^s \right] 1_{j \neq i+1} 1_{\ell = i+1} + (1-p)\mathbb{E}_H \left[\left\| \left(I - \frac{\eta_{i-1}}{b} H \right) e_1 \right\|^s \right] 1_{j \neq i-1} 1_{\ell = i-1}, \quad (59)$$

and

$$Q^j_{1\ell} := 1_{j \neq 2} 1_{\ell = 2}, \qquad Q^j_{K\ell} := 1_{j \neq K+1} 1_{\ell = K+1}. \quad (60)$$

B.4.3 Markovian Stepsizes with Two-State Space

In this section, we study the SGD with Markovian stepsizes with two-state space. With the general finite state space, we have seen previously that the tail-index $\alpha^{(r)}$ is the unique positive value such that $h^{(r)}\left(\alpha^{(r)}\right) = 1$. However, the expression for $h^{(r)}(s)$ is quite complicated. We are able to characterize $h^{(r)}(s)$ in a more explicit way for the two-state space case. First, we recall from Lemma 11 that $h^{(r)}(s) = \tilde{h}^{(r)}(s)$ and $\rho^{(r)} = \tilde{\rho}^{(r)}$, with $\tilde{h}^{(r)}(s)$ and $\tilde{\rho}^{(r)}$ given in Lemma 11. We have the following result, which plays a central role in order to obtain Proposition 4.

Lemma 6. *Consider the two-state Markov chain, i.e. $\mathbb{P}(\eta_1 = \eta_u | \eta_0 = \eta_l) = p$ and $\mathbb{P}(\eta_1 = \eta_l | \eta_0 = \eta_u) = p$ and assume that $(1-p)\mathbb{E}_H \left[\left\| \left(I - \frac{\eta_l}{b} H \right) e_1 \right\|^s \right] < 1$ and $(1-p)\mathbb{E}_H \left[\left\| \left(I - \frac{\eta_u}{b} H \right) e_1 \right\|^s \right] < 1$. Then, we have*

$$\tilde{h}^{(r)}(s) = \frac{\mathbb{E}_H \left[\left\| \left(I - \frac{\eta_l}{b} H \right) e_1 \right\|^s \right] \left(1 - p + (2p-1)\mathbb{E}_H \left[\left\| \left(I - \frac{\eta_u}{b} H \right) e_1 \right\|^s \right] \right)}{2\left(1 - (1-p)\mathbb{E}_H \left[\left\| \left(I - \frac{\eta_u}{b} H \right) e_1 \right\|^s \right] \right)} + \frac{\mathbb{E}_H \left[\left\| \left(I - \frac{\eta_u}{b} H \right) e_1 \right\|^s \right] \left(1 - p + (2p-1)\mathbb{E}_H \left[\left\| \left(I - \frac{\eta_l}{b} H \right) e_1 \right\|^s \right] \right)}{2\left(1 - (1-p)\mathbb{E}_H \left[\left\| \left(I - \frac{\eta_l}{b} H \right) e_1 \right\|^s \right] \right)}, \quad (61)$$

and

$$\tilde{\rho}^{(r)} = \mathbb{E}_H\left[\log\left\|\left(I - \frac{\eta_l}{b}H\right)e_1\right\|\right] + \mathbb{E}_H\left[\log\left\|\left(I - \frac{\eta_u}{b}H\right)e_1\right\|\right]. \tag{62}$$

In particular, when $p = 1$, *we get* $\tilde{h}^{(r)}(s) = \mathbb{E}_H\left[\left\|\left(I - \frac{\eta_l}{b}H\right)e_1\right\|^s\right]\mathbb{E}_H\left[\left\|\left(I - \frac{\eta_u}{b}H\right)e_1\right\|^s\right] = h^{(m)}(s)$.

In Proposition 9, we can write $c^{(r)}$ as $c^{(r)} = \tilde{h}^{(r)}(2)$. Therefore, we immediately obtain the following result by applying Lemma 6.

Corollary 4. *Consider stepsizes following the two-state Markov chain, i.e.* $\mathbb{P}(\eta_1 = \eta_u|\eta_0 = \eta_l) = p$ *and* $\mathbb{P}(\eta_1 = \eta_l|\eta_0 = \eta_u) = p$. *In Proposition 9, we have*

$$\begin{aligned} c^{(r)} = &\frac{\left(1 - 2\eta_l\sigma^2 + \frac{\eta_l^2\sigma^4}{b}(d+b+1)\right)\left(1 - p + (2p-1)\left(1 - 2\eta_u\sigma^2 + \frac{\eta_u^2\sigma^4}{b}(d+b+1)\right)\right)}{2\left(1 - (1-p)\left(1 - 2\eta_u\sigma^2 + \frac{\eta_u^2\sigma^4}{b}(d+b+1)\right)\right)} \\ &+ \frac{\left(1 - 2\eta_u\sigma^2 + \frac{\eta_u^2\sigma^4}{b}(d+b+1)\right)\left(1 - p + (2p-1)\left(1 - 2\eta_l\sigma^2 + \frac{\eta_l^2\sigma^4}{b}(d+b+1)\right)\right)}{2\left(1 - (1-p)\left(1 - 2\eta_l\sigma^2 + \frac{\eta_l^2\sigma^4}{b}(d+b+1)\right)\right)}. \end{aligned} \tag{63}$$

We recall from Proposition 9 that (i) There exists $\delta > 0$ such that for any $1 < c^{(r)} < 1 + \delta$, tail-index $0 < \alpha^{(r)} < 2$. (ii) If $c^{(r)} = 1$, tail-index $\alpha^{(r)} = 2$. (iii) If $c^{(r)} < 1$, then tail-index $\alpha^{(r)} > 2$.

C Technical Lemmas

Lemma 7. *Assume* ***(A3)*** *holds. Then, we have*

$$\rho = \tilde{\rho}, \qquad h(s) = \tilde{h}(s), \quad \textit{for every } s \geq 0, \tag{64}$$

where

$$\tilde{\rho} := \mathbb{E}\left[\log\left\|\left(I - \frac{\eta_1}{b}H_1\right)e_1\right\|\right], \tag{65}$$

and

$$\tilde{h}(s) := \mathbb{E}\left[\|M_1 e_1\|^s\right] = \mathbb{E}\left[\left\|\left(I - \frac{\eta_1}{b}H_1\right)e_1\right\|^s\right]. \tag{66}$$

Lemma 8. *Assume* ***(A3)*** *holds. For any* $s \geq 0$, $h(s) = \tilde{h}(s)$ *and* $\rho = \tilde{\rho}$, *where*

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

and

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

where η, 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* $(d-1)$.

Lemma 9. *Assume* ***(A3)*** *holds. For any* $s \geq 0$,

$$\left(h^{(m)}(s)\right)^{1/m} = \tilde{h}^{(m)}(s), \tag{67}$$

where

$$\tilde{h}^{(m)}(s) := \left(\prod_{i=1}^{m}\mathbb{E}\left[\left\|\left(I - \frac{\eta_i}{b}H\right)e_1\right\|^s\right]\right)^{1/m}. \tag{68}$$

Moreover

$$\rho^{(m)} = \tilde{\rho}^{(m)} := \sum_{i=1}^{m}\mathbb{E}\left[\log\left\|\left(I - \frac{\eta_i}{b}H\right)e_1\right\|\right]. \tag{69}$$

Lemma 10. *Assume **(A3)** holds. For any $s \geq 0$, we have $h^{(m)}(s) = \tilde{h}^{(m)}(s)$ and $\rho^{(m)} = \tilde{\rho}^{(m)}$, where*

$$\tilde{h}^{(m)}(s) = \left(\prod_{i=1}^{m} \mathbb{E}\left[\left(\left(1 - \frac{\eta_i \sigma^2}{b} X \right)^2 + \frac{\eta_i^2 \sigma^4}{b^2} XY \right)^{s/2} \right] \right)^{1/m},$$
$$\tilde{\rho}^{(m)} = \frac{1}{2} \sum_{i=1}^{m} \mathbb{E}\left[\log \left(\left(1 - \frac{\eta_i \sigma^2}{b} X \right)^2 + \frac{\eta_i^2 \sigma^4}{b^2} XY \right) \right],$$

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 $(d-1)$.

Lemma 11. *For any $s \geq 0$,*

$$h^{(r)}(s) = \tilde{h}^{(r)}(s) := \mathbb{E}\left[\prod_{i=1}^{r_1} \mathbb{E}_H \left[\left\| \left(I - \frac{\eta_i}{b} H \right) e_1 \right\|^s \right] \right], \tag{70}$$

and moreover,

$$\rho^{(r)} = \tilde{\rho}^{(r)} := \mathbb{E}\left[\sum_{i=1}^{r_1} \mathbb{E}_H \left[\log \left\| \left(I - \frac{\eta_i}{b} H \right) e_1 \right\| \right] \right], \tag{71}$$

where r_1 is defined in (8).

Lemma 12. *Assume **(A3)** holds. For any $s \geq 0$, we have $h^{(r)}(s) = \tilde{h}^{(r)}(s)$ and $\rho^{(r)} = \tilde{\rho}^{(r)}$, where*

$$\tilde{h}^{(r)}(s) = \mathbb{E}\left[\prod_{i=1}^{r_1} \mathbb{E}_{X,Y} \left[\left(\left(1 - \frac{\eta_i \sigma^2}{b} X \right)^2 + \frac{\eta_i^2 \sigma^4}{b^2} XY \right)^{s/2} \right] \right],$$
$$\tilde{\rho}^{(r)} = \frac{1}{2} \mathbb{E}\left[\sum_{i=1}^{r_1} \mathbb{E}_{X,Y} \left[\log \left(\left(1 - \frac{\eta_i \sigma^2}{b} X \right)^2 + \frac{\eta_i^2 \sigma^4}{b^2} XY \right) \right] \right],$$

where $\mathbb{E}_{X,Y}$ denotes the expectation w.r.t. X, Y, 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 $(d-1)$ and X, Y are independent of $(\eta_k)_{k \in \mathbb{N}}$.

D Technical Proofs

D.1 Proof of results in Section 3

Proof of Theorem 3

It follows from the proof of Theorem 4 in Gürbüzbalaban et al. (2021) that for any $s \geq 1$, conditional on η_i, $\mathbb{E}_H \left[\left\| I - \frac{\eta_i}{b} H \right\|^s \right]$ is strictly decreasing in b. Therefore, $\hat{h}^{(r)}(s)$ is strictly decreasing in b. It thus follows from the arguments in the proof of Theorem 4 in Gürbüzbalaban et al. (2021) that $\hat{\alpha}^{(r)}$ is strictly increasing in batch-size b provided that $\hat{\alpha}^{(r)} \geq 1$. The proof is complete. □

Proof of Theorem 4

Given $\rho^{(r)} < 0$, the tail-index $\alpha^{(r)}$ is the unique positive value such that $\tilde{h}^{(r)}(s)\left(\alpha^{(r)}\right) = 1$. It follows from Theorem 4 in Gürbüzbalaban et al. (2021) that conditional on η_i, $\mathbb{E}_H \left[\left\| \left(I - \frac{\eta_i}{b} H \right) e_1 \right\|^s \right]$ is strictly decreasing in batch-size b for any $s \geq 1$, and it is strictly increasing in dimension d. Therefore, $\tilde{h}^{(r)}(s)$ is strictly decreasing in batch-size b for any $s \geq 1$, and it is strictly increasing in dimension d, and the conclusion follows. □

Proof of Theorem 6

It follows from the proof of Theorem 4 in Gürbüzbalaban et al. (2021) that for any $s \geq 1$, the function

$$\mathbb{E}\left[\left\|I - \frac{\eta_i}{b} H\right\|^s\right]$$

is strictly decreasing in b. Therefore, $\hat{h}^{(m)}(s)$ is strictly decreasing in b. It thus follows from the arguments in the proof of Theorem 4 in Gürbüzbalaban et al. (2021) that $\hat{\alpha}^{(m)}$ is strictly increasing in batch-size b provided that $\hat{\alpha}^{(m)} \geq 1$. The proof is complete. □

Proof of Theorem 7

Given that $\rho^{(m)} < 0$, the tail-index $\alpha^{(m)}$ is the unique positive value such that $\tilde{h}^{(m)}\left(\alpha^{(m)}\right) = 1$. It follows from Theorem 4 in Gürbüzbalaban et al. (2021) that $\mathbb{E}\left[\left\|\left(I - \frac{\eta_i}{b} H\right) e_1\right\|^s\right]$ is strictly decreasing in batch-size b for any $s \geq 1$, and it is strictly increasing in dimension d. Therefore, $\tilde{h}^{(m)}(s)$ is strictly decreasing in batch-size b for any $s \geq 1$, and it is strictly increasing in dimension d, and the conclusion follows. □

D.2 Proofs of Results in Section 4

Proof of Proposition 1

By Lemma 13, 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 $s \geq 1$. By tower property and Jensen's inequality,

$$\begin{aligned} h(s) &= \mathbb{E}\left[\mathbb{E}\left[\left\|\left(I - \frac{\eta}{b} H\right) e_1\right\|^s \middle| H\right]\right] \\ &\geq \mathbb{E}\left[\left\|\mathbb{E}\left[\left(I - \frac{\eta}{b} H\right) e_1 \middle| H\right]\right\|^s\right] = \mathbb{E}\left[\left\|\left(I - \frac{\mathbb{E}[\eta]}{b} H\right) e_1\right\|^s\right], \end{aligned}$$

which is the h function with constant stepsize $\mathbb{E}[\eta]$. Since η is random, the above inequality is strict, hence we conclude that the tail-index α is strictly less than the tail-index α_c with constant stepsize $\mathbb{E}[\eta]$ provided that $\alpha \geq 1$. The proof is complete. □

Proof of Proposition 2

We recall that the tail-index $\alpha^{(m)}$ for the SGD with cyclic stepsizes is the unique positive value such that $h^{(m)}\left(\alpha^{(m)}\right) = 1$. By the inequality of arithmetic and geometric means, we obtain

$$h^{(m)}(s) \leq \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}\left[\left\|\left(I - \frac{\eta_i}{b} H\right) e_1\right\|^s\right] = h(s). \tag{72}$$

Since η_i is not constant, the above inequality is strict. Therefore, we conclude that the tail-index α of SGD with i.i.d. stepsizes is strictly less than the tail-index $\alpha^{(m)}$ for the SGD with cyclic stepsizes. The proof is complete. □

Proof of Proposition 3

Under the assumption **(A3)**, we have $h_c(\alpha_c) = 1$ and $h^{(m)}\left(\alpha^{(m)}\right) = 1$, where by Lemma 10

$$h_c(s) := h\left(s; \frac{1}{m} \sum_{i=1}^{m} \eta_i\right), \qquad \text{and} \qquad h^{(m)}(s) = \left(\prod_{i=1}^{m} h(s; \eta_i)\right)^{1/m},$$

where

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

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

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

and

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

and therefore

$$\begin{aligned}&h(s;0)\frac{\partial^2}{\partial \eta^2} h(s;0) - \left(\frac{\partial}{\partial \eta} h(s;0)\right)^2 \\ &= \mathbb{E}\left[\frac{s}{2}\left(\frac{2\sigma^4}{b^2}X^2 + \frac{2\sigma^4}{b^2}XY\right)\right] + \mathbb{E}\left[\frac{s}{2}\left(\frac{s}{2}-1\right)\frac{4\sigma^4}{b^2}X^2\right] - \frac{s^2}{4}\frac{4\sigma^4}{b^2}(\mathbb{E}[X])^2 \\ &= \frac{s\sigma^4}{b}(d+b+1) + s(s-2)\frac{\sigma^4}{b}(b+2) - s^2\sigma^4 = \frac{s\sigma^4}{b}\left(d-b+2s-3\right) > 0,\end{aligned}$$

for any $s > 0$ provided that $d \geq b+3$. This implies that under the assumption $d \geq b+3$ and the stepsize $\eta > 0$ is sufficiently small, $h(s;\eta)$ is log-convex in η and hence by Jensen's inequality, $h^{(m)}(s) \geq h_c(s)$, which implies that $\alpha^{(m)} \leq \alpha_c$. This completes the proof. □

Proof of Proposition 4

Let us denote

$$x := \mathbb{E}_H\left[\left\|\left(I - \frac{\eta_l}{b}H\right)e_1\right\|^s\right], \qquad y := \mathbb{E}_H\left[\left\|\left(I - \frac{\eta_u}{b}H\right)e_1\right\|^s\right]. \tag{73}$$

We also define:

$$F(p) := \frac{x(1-p+(2p-1)y)}{2(1-(1-p)y)} + \frac{y(1-p+(2p-1)x)}{2(1-(1-p)x)}. \tag{74}$$

Then, it follows from Lemma 6 that $\tilde{h}^{(r)}(s) = F(p)$ provided that $(1-p)x < 1$ and $(1-p)y < 1$. For any $p \in \mathcal{P}$ where $\mathcal{P}$ is defined in (15) and $s \in \mathcal{S}$ where $\mathcal{S}$ is a sufficiently small interval that contains $\alpha^{(r)}$, we have $(1-p)x < 1$ and $(1-p)y < 1$. We can compute that

$$\begin{aligned}\frac{\partial F}{\partial p} &= \frac{x(-1+2y)(1-(1-p)y) - x(1-p+(2p-1)y)y}{2(1-(1-p)y)^2} \\ &\quad + \frac{y(-1+2x)(1-(1-p)x) - y(1-p+(2p-1)x)x}{2(1-(1-p)x)^2} \\ &= \frac{-x(1-y)^2}{2(1-(1-p)y)^2} + \frac{-y(1-x)^2}{2(1-(1-p)x)^2} < 0,\end{aligned}$$

so that $\tilde{h}^{(r)}(s)$ is decreasing in $p \in \mathcal{P}$ for any $s \in \mathcal{S}$ and hence the tail-index $\alpha^{(r)}$ is increasing in $p \in \mathcal{P}$. Finally, $p = 1 \in \mathcal{P}$ and $\alpha^{(r)}$ reduces to $\alpha^{(m)}$ when $p = 1$ which implies that $\alpha^{(r)} \leq \alpha^{(m)}$. The proof is complete. □

Proof of Proposition 5

First of all, we recall that α is the tail-index for SGD with i.i.d. stepsizes which is the unique position value such that $h(\alpha) = 1$ and $\alpha^{(m)}$ is the tail-index for SGD with cyclic stepsizes which is the unique position value such that $h^{(m)}\left(\alpha^{(m)}\right) = 1$. It is easy to see that

$$\begin{aligned} h^{(m)}(s) &= \mathbb{E}_H\left[\left\|\left(I - \frac{\eta_l}{b}H\right)e_1\right\|^s\right]\mathbb{E}_H\left[\left\|\left(I - \frac{\eta_u}{b}H\right)e_1\right\|^s\right] \\ &\leq \left(\frac{\mathbb{E}_H\left[\left\|\left(I - \frac{\eta_l}{b}H\right)e_1\right\|^s\right] + \mathbb{E}_H\left[\left\|\left(I - \frac{\eta_u}{b}H\right)e_1\right\|^s\right]}{2}\right)^2 = (h(s))^2, \end{aligned}$$

which implies that $\alpha \leq \alpha^{(m)}$.

Note that α and $\alpha^{(m)}$ are independent of p and by Proposition 4, $\alpha^{(r)}$ is increasing in p, and in particular, $\alpha^{(r)} = \alpha^{(m)}$ when $p = 1$. Moreover, as

$$p \to \max\left(1 - \frac{1}{\mathbb{E}_H\left[\left\|\left(I - \frac{\eta_l}{b}H\right)e_1\right\|^s\right]}, 1 - \frac{1}{\mathbb{E}_H\left[\left\|\left(I - \frac{\eta_u}{b}H\right)e_1\right\|^s\right]}\right), \tag{75}$$

by Lemma 6, we have $\tilde{h}^{(r)}(s) \to \infty$, and hence we conclude that there exists some critical $p_c \in (0,1)$ such that for any $p_c < p < 1$, we have $\alpha < \alpha^{(r)} < \alpha^{(m)}$ and for any $p < p_c$, we have $\alpha^{(r)} < \alpha < \alpha^{(m)}$.

Indeed one can determine the critical p_c explicitly. Note that p_c is the critical value such that $\alpha = \alpha^{(r)}$, which is equivalent to the critical value p_c such that $\tilde{h}^{(r)}(\alpha) = 1$. Hence, p_c is determined by the equation:

$$\begin{aligned} &\frac{\mathbb{E}_H\left[\left\|\left(I - \frac{\eta_l}{b}H\right)e_1\right\|^\alpha\right](1 - p_c + (2p_c - 1)\mathbb{E}_H\left[\left\|\left(I - \frac{\eta_u}{b}H\right)e_1\right\|^\alpha\right])}{2(1 - (1-p_c)\mathbb{E}_H\left[\left\|\left(I - \frac{\eta_u}{b}H\right)e_1\right\|^\alpha\right])} \\ &+ \frac{\mathbb{E}_H\left[\left\|\left(I - \frac{\eta_u}{b}H\right)e_1\right\|^\alpha\right](1 - p_c + (2p_c - 1)\mathbb{E}_H\left[\left\|\left(I - \frac{\eta_l}{b}H\right)e_1\right\|^\alpha\right])}{2(1 - (1-p_c)\mathbb{E}_H\left[\left\|\left(I - \frac{\eta_l}{b}H\right)e_1\right\|^\alpha\right])} = 1. \end{aligned} \tag{76}$$

After some algebraic computations, one can rewrite the above equation for p_c as a quadratic equation in p_c:

$$\begin{aligned} &\left(2(yx^2 + y^2x) - (x+y)^2\right)p_c^2 - \left(3(yx^2 + y^2x) + 3(x+y) - 4xy - 2(x+y)^2\right)p_c \\ &\quad + 3(x+y) - 2xy + yx^2 + y^2x - (x+y)^2 - 2 = 0, \end{aligned} \tag{77}$$

where

$$x := \mathbb{E}_H\left[\left\|\left(I - \frac{\eta_l}{b}H\right)e_1\right\|^\alpha\right], \qquad y := \mathbb{E}_H\left[\left\|\left(I - \frac{\eta_l}{b}H\right)e_1\right\|^\alpha\right]. \tag{78}$$

By the definition of α, we have

$$h(\alpha) = \frac{1}{2}x + \frac{1}{2}y = 1, \tag{79}$$

which implies that $x + y = 2$ so that the quadratic equation (77) can be simplified as:

$$(4xy - 4)p_c^2 - (2xy - 2)p_c = 0, \tag{80}$$

which yields that $p_c = \frac{1}{2}$. The proof is complete. □

D.3 Proofs of Results in Section B.2

Proof of Theorem 11

The proof is similar to the proof of Theorem 2 in Gürbüzbalaban et al. (2021) and is omitted here. □

Proof of Theorem 12

By following the proof of Theorem 4 in Gürbüzbalaban et al. (2021), it suffices to show that for any $s \geq 1$, $h(s)$ is decreasing in batch-size $b \in \mathbb{N}$ and for any $s \geq 0$, $h(s)$ is increasing in dimension $d \in \mathbb{N}$. Under the assumption of the Gaussian input data, by tower property,

$$h(s) = \mathbb{E}[h(s|\eta)], \qquad h(s|\eta) := \mathbb{E}\left[\left\|\left(I - \frac{\eta}{b}H\right)e_1\right\|^s \middle| \eta\right]. \tag{81}$$

In the proof of Theorem 4 in Gürbüzbalaban et al. (2021), it showed that for any given η, for any $s \geq 1$, $h(s|\eta)$ is decreasing in batch-size $b \in \mathbb{N}$ and for any $s \geq 0$, $h(s|\eta)$ is increasing in dimension $d \in \mathbb{N}$. Since $h(s) = \mathbb{E}[h(s|\eta)]$, we conclude that $h(s)$ is decreasing in batch-size $b \in \mathbb{N}$ and for any $s \geq 0$, $h(s)$ is increasing in dimension $d \in \mathbb{N}$. Hence, by following the same arguments as in the proof of Theorem 4 in Gürbüzbalaban et al. (2021), we conclude that the tail-index α is strictly increasing in batch-size b provided that $\alpha \geq 1$ and the tail-index α is strictly decreasing in dimension d. The proof is complete. □

Proof of Theorem 13

When η is uniformly distributed on $(\bar{\eta} - R, \bar{\eta} + R)$,

$$h(s) = \frac{1}{2R}\int_{\bar{\eta}-R}^{\bar{\eta}+R} \mathbb{E}\left[\left\|\left(I - \frac{x}{b}H\right)e_1\right\|^s\right] dx. \tag{82}$$

It suffices to show that $h(s)$ is increasing in R for any $s \geq 1$. We can compute that

$$\begin{aligned}\frac{\partial}{\partial R}h(s) = &\frac{-1}{2R^2}\int_{\bar{\eta}-R}^{\bar{\eta}+R} \mathbb{E}\left[\left\|\left(I - \frac{x}{b}H\right)e_1\right\|^s\right] dx \\ &+ \frac{1}{2R}\left(\mathbb{E}\left[\left\|\left(I - \frac{\bar{\eta}+R}{b}H\right)e_1\right\|^s\right] + \mathbb{E}\left[\left\|\left(I - \frac{\bar{\eta}-R}{b}H\right)e_1\right\|^s\right]\right).\end{aligned} \tag{83}$$

Then, it suffices to show that

$$R(f(\bar{\eta}+R) + f(\bar{\eta}-R)) \geq \int_{\bar{\eta}-R}^{\bar{\eta}+R} f(x)dx, \tag{84}$$

where

$$f(x) := \mathbb{E}\left[\left\|\left(I - \frac{x}{b}H\right)e_1\right\|^s\right] \tag{85}$$

is convex in x for any $s \geq 1$ according to Lemma 13. Note that (84) is equivalent to

$$F\left(\bar{\eta}+R; \bar{\eta}-R\right) \geq 0, \tag{86}$$

where

$$F(x;a) := \frac{x-a}{2}(f(x)+f(a)) - \int_a^x f(y)dy. \tag{87}$$

Then we have $F(a;a) = 0$ and

$$\frac{\partial}{\partial x}F(x;a) = \frac{f(a) - f(x) + (x-a)f'(x)}{2} \geq 0, \tag{88}$$

which holds since $f(x)$ is convex in x. This implies that $F(x;a) \geq 0$ for any $x \geq a > 0$. and thus $F(\bar{\eta}+R; \bar{\eta}-R) \geq 0$, which implies (84). This completes the proof. □

Proof of Proposition 6

We first prove (i). Let us first recall from Lemma 8 that

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

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 $(d-1)$, and X, Y are independent of η. When $c = 1 - 2\mathbb{E}[\eta]\sigma^2 + \frac{\mathbb{E}[\eta^2]\sigma^4}{b}(d+b+1) = 1$, we can compute that

$$\begin{aligned}\tilde{\rho} &\leq \frac{1}{2}\log\mathbb{E}\left[1-\frac{2\eta\sigma^2}{b}X + \frac{\eta^2\sigma^4}{b^2}(X^2+XY)\right] \\ &= \frac{1}{2}\log\left(1-2\mathbb{E}[\eta]\sigma^2 + \frac{\mathbb{E}[\eta^2]\sigma^4}{b}(d+b+1)\right) = 0.\end{aligned} \tag{89}$$

Note that since $1 - \frac{2\eta\sigma^2}{b}X + \frac{\eta^2\sigma^4}{b^2}(X^2+XY)$ is random, the inequality in (89) is a strict inequality from Jensen's inequality. Thus, when $c = 1$, we have $\tilde{\rho} < 0$. By continuity, there exists some $\delta > 0$ such that for any $1 < c < 1+\delta$ we have $\tilde{\rho} < 0$. Moreover, when $c > 1$, we have

$$\begin{aligned}\tilde{h}(2) &= \mathbb{E}\left[1-\frac{2\eta\sigma^2}{b}X + \frac{\eta^2\sigma^4}{b^2}(X^2+XY)\right] \\ &= 1 - 2\mathbb{E}[\eta]\sigma^2 + \frac{\mathbb{E}[\eta^2]\sigma^4}{b}(d+b+1) = c > 1,\end{aligned}$$

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

Finally, let us prove (ii) and (iii). When $c \leq 1$, we have $\tilde{h}(2) \leq 1$, which implies that $\alpha \geq 2$. In particular, when $c = 1$, the tail-index $\alpha = 2$. The proof is complete. □

Proof of Lemma 1

We recall that

$$x_k = M_k x_{k-1} + q_k, \tag{90}$$

which implies that

$$\|x_k\| \leq \|M_k x_{k-1}\| + \|q_k\|. \tag{91}$$

(i) For any $p \leq 1$ and $h(p) < 1$, by Lemma 14,

$$\|x_k\|^p \leq \|M_k x_{k-1}\|^p + \|q_k\|^p. \tag{92}$$

Since M_k is independent of x_{k-1} and conditional on x_{k-1} the distribution of $\|M_k x_{k-1}\|$ is the same as $\|M_k e_1\| \cdot \|x_{k-1}\|$, we have

$$\mathbb{E}\|x_k\|^p \leq \mathbb{E}\|M_k e_1\|^p \mathbb{E}\|x_{k-1}\|^p + \mathbb{E}\|q_k\|^p, \tag{93}$$

where e_1 is the first basis vector in $\mathbb{R}^d$, so that

$$\mathbb{E}\|x_k\|^p \leq h(p)\mathbb{E}\|x_{k-1}\|^p + \mathbb{E}\|q_1\|^p. \tag{94}$$

By iterating over k, we get

$$\mathbb{E}\|x_k\|^p \leq (h(p))^k \mathbb{E}\|x_0\|^p + \frac{1-(h(p))^k}{1-h(p)}\mathbb{E}\|q_1\|^p. \tag{95}$$

(ii) For any $p > 1$ and $h(p) < 1$, by Lemma 14, for any $\epsilon > 0$, we have

$$\|x_k\|^p \leq (1+\epsilon)\|M_k x_{k-1}\|^p + \frac{(1+\epsilon)^{\frac{p}{p-1}} - (1+\epsilon)}{\left((1+\epsilon)^{\frac{1}{p-1}} - 1\right)^p}\|q_k\|^p, \tag{96}$$

which (similar as in (i)) implies that

$$\mathbb{E}\|x_k\|^p \leq (1+\epsilon)\mathbb{E}\|M_k e_1\|^p \mathbb{E}\|x_{k-1}\|^p + \frac{(1+\epsilon)^{\frac{p}{p-1}} - (1+\epsilon)}{\left((1+\epsilon)^{\frac{1}{p-1}} - 1\right)^p}\mathbb{E}\|q_k\|^p, \tag{97}$$

so that

$$\mathbb{E}\|x_k\|^p \leq (1+\epsilon)h(p)\mathbb{E}\|x_{k-1}\|^p + \frac{(1+\epsilon)^{\frac{p}{p-1}} - (1+\epsilon)}{\left((1+\epsilon)^{\frac{1}{p-1}} - 1\right)^p}\mathbb{E}\|q_1\|^p. \tag{98}$$

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

$$\mathbb{E}\|x_k\|^p \leq ((1+\epsilon)h(p))^k \mathbb{E}\|x_0\|^p + \frac{1 - ((1+\epsilon)h(p))^k}{1-(1+\epsilon)h(p)} \frac{(1+\epsilon)^{\frac{p}{p-1}} - (1+\epsilon)}{\left((1+\epsilon)^{\frac{1}{p-1}} - 1\right)^p}\mathbb{E}\|q_1\|^p. \tag{99}$$

The proof is complete. □

Proof of Theorem 14

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

$$x_k = M_k x_{k-1} + q_k, \tag{100}$$
$$\tilde{x}_k = M_k \tilde{x}_{k-1} + q_k, \tag{101}$$

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

$$x_k - \tilde{x}_k = M_k(x_{k-1} - \tilde{x}_{k-1}), \tag{102}$$

which yields that

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

where e_1 is the first basis vector in $\mathbb{R}^d$, which by iterating implies that

$$\mathcal{W}_p^p(\nu_k, \tilde{\nu}_k) \leq \mathbb{E}\|x_k - \tilde{x}_k\|^p \leq (h(p))^k \mathbb{E}\|x_0 - \tilde{x}_0\|^p = (h(p))^k \mathcal{W}_p^p(\nu_0, \tilde{\nu}_0). \tag{103}$$

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

$$\mathcal{W}_p(\nu_k, \nu_\infty) \leq \left((h(p))^{1/q}\right)^k \mathcal{W}_p(\nu_0, \nu_\infty). \tag{104}$$

The proof is complete. □

Proof of Corollary 1

The result is obtained by a direct application of Theorem 1.15 in Mirek (2011) to the recursions (23), where it can be checked in a straightforward manner that the conditions for this theorem hold. □

D.4 Proofs of Results in Section B.3

Proof of Proposition 7

We first prove (i). Let us first recall from Lemma 10 that

$$\tilde{h}^{(m)}(s) = \left(\prod_{i=1}^{m} \mathbb{E}\left[\left(\left(1 - \frac{\eta_i \sigma^2}{b} X \right)^2 + \frac{\eta_i^2 \sigma^4}{b^2} XY \right)^{s/2} \right] \right)^{1/m},$$
$$\tilde{\rho}^{(m)} = \frac{1}{2} \sum_{i=1}^{m} \mathbb{E}\left[\log \left(\left(1 - \frac{\eta_i \sigma^2}{b} X \right)^2 + \frac{\eta_i^2 \sigma^4}{b^2} XY \right) \right],$$

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 $(d-1)$. When

$$c^{(m)} = \prod_{i=1}^{m} \left(1 - 2\eta_i \sigma^2 + \frac{\eta_i^2 \sigma^4}{b}(d+b+1) \right) = 1,$$

we can compute that

$$\begin{aligned} \tilde{\rho}^{(m)} &\leq \frac{1}{2} \sum_{i=1}^{m} \log \mathbb{E}\left[1 - \frac{2\eta_i \sigma^2}{b} X + \frac{\eta_i^2 \sigma^4}{b^2}(X^2 + XY) \right] \\ &= \frac{1}{2} \sum_{i=1}^{m} \log \left(1 - 2\eta_i \sigma^2 + \frac{\eta_i^2 \sigma^4}{b}(d+b+1) \right) = 0. \end{aligned} \tag{105}$$

Note that since $1 - \frac{2\eta_i \sigma^2}{b} X + \frac{\eta_i^2 \sigma^4}{b^2}(X^2 + XY)$ is random, the inequality in (105) is a strict inequality from Jensen's inequality. Thus, when $c^{(m)} = 1$, we have $\tilde{\rho}^{(m)} < 0$. By continuity, there exists some $\delta > 0$ such that for any $1 < c^{(m)} < 1 + \delta$ we have $\tilde{\rho}^{(m)} < 0$. Moreover, when $c^{(m)} > 1$, we have

$$\begin{aligned} \left(h^{(m)}(2) \right)^m &= \prod_{i=1}^{m} \mathbb{E}\left[1 - \frac{2\eta_i \sigma^2}{b} X + \frac{\eta_i^2 \sigma^4}{b^2}(X^2 + XY) \right] \\ &= \prod_{i=1}^{m} \left(1 - 2\eta_i \sigma^2 + \frac{\eta_i^2 \sigma^4}{b}(d+b+1) \right) = c^{(m)} > 1, \end{aligned}$$

which implies that there exists some $0 < \alpha^{(m)} < 2$ such that $h^{(m)}\left(\alpha^{(m)}\right) = 1$.

Finally, let us prove (ii) and (iii). When $c^{(m)} \leq 1$, we have $\tilde{h}^{(m)}(2) \leq 1$, which implies that $\alpha^{(m)} \geq 2$. In particular, when $c^{(m)} = 1$, the tail-index $\alpha^{(m)} = 2$. The proof is complete. □

Proof of Lemma 2

The proof is similar to the proof of Lemma 1 and is hence omitted here. □

Proof of Theorem 16

The proof is similar to the proof of Theorem 14 and is hence omitted here. □

Proof of Corollary 2

The proof is similar to the proof of Corollary 1 and is hence omitted here. □

D.5 Proofs of Results in Section B.4

Proof of Proposition 8

For any $p < \hat{\alpha}^{(g)}$, we have $\hat{h}^{(g)}(p) < 1$. By Lemma 3 and Fatou's lemma, we have that for any $p \leq 1$ and $\hat{h}^{(g)}(p) < 1$,

$$\mathbb{E}\|x_\infty\|^p \leq \frac{1}{1-\hat{h}^{(g)}(p)}\mathbb{E}\|q_1\|^p, \tag{106}$$

and for any $p > 1$, $\epsilon > 0$ and $(1+\epsilon)\hat{h}^{(g)}(p) < 1$,

$$\mathbb{E}\|x_\infty\|^p \leq \frac{1}{1-(1+\epsilon)\hat{h}^{(g)}(p)} \frac{(1+\epsilon)^{\frac{p}{p-1}} - (1+\epsilon)}{\left((1+\epsilon)^{\frac{1}{p-1}} - 1\right)^p}\mathbb{E}\|q_1\|^p. \tag{107}$$

Finally, by applying Chebyshev's inequality inequality, we complete the proof. □

Proof of Theorem 17

By following the proof of Theorem 4 in Gürbüzbalaban et al. (2021), it suffices to show that for any $s \geq 1$, $\hat{h}^{(g)}(s)$ is decreasing in batch-size $b \in \mathbb{N}$. By tower property,

$$\hat{h}^{(g)}(s) = \mathbb{E}\left[\hat{h}^{(g)}(s|\eta)\right], \qquad \hat{h}^{(g)}(s|\eta) := \mathbb{E}\left[\left\|I - \frac{\eta}{b}H\right\|^s \middle| \eta\right]. \tag{108}$$

With slight abuse of notation, we define the function $\hat{h}^{(g)}(b, s|\eta) = \hat{h}^{(g)}(s|\eta)$ to emphasize the dependence on b. We have

$$\hat{h}^{(g)}(b, s|\eta) = \mathbb{E}\left[\left\|I - \frac{\eta}{b}\sum_{i=1}^{b} a_i a_i^T\right\|^s \middle| \eta\right]. \tag{109}$$

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 $b \in \mathbb{N}$,

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

where we used the fact that a_i are i.i.d. independent of the distribution of η. 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 $s \geq 1$, $\hat{h}^{(g)}(b, s|\eta)$ is strictly decreasing in b. Since $\hat{h}^{(g)}(s) = \mathbb{E}[\hat{h}^{(g)}(s|\eta)]$, we conclude that $\hat{h}^{(g)}(s)$ is decreasing in batch-size $b \in \mathbb{N}$. Hence, by following the same arguments as in the proof of Theorem 4 in Gürbüzbalaban et al. (2021), we conclude that the lower bound for the tail-index $\hat{\alpha}^{(g)}$ is strictly increasing in batch-size b provided that $\hat{\alpha}^{(g)} \geq 1$.

Moreover, by adapting the proof of Lemma 13 (Lemma 22 in Gürbüzbalaban et al. (2021)), one can show that for any given positive semi-definite symmetric matrix H fixed, the function $F_H : [0, \infty) \to \mathbb{R}$ defined as $F_H(a) := \|(I - aH)\|^s$ is convex for $s \geq 1$. The rest of the proof follows from the similar arguments as in the proof of Theorem 12. The proof is complete. □

Proof of Proposition 9

We first prove (i). Let us first recall from Lemma 12 that

$$\tilde{h}^{(r)}(s) = \mathbb{E}\left[\prod_{i=1}^{r_1} \mathbb{E}_{X,Y}\left[\left(\left(1-\frac{\eta_i\sigma^2}{b}X\right)^2 + \frac{\eta_i^2\sigma^4}{b^2}XY\right)^{s/2}\right]\right],$$
$$\tilde{\rho}^{(r)} = \frac{1}{2}\mathbb{E}\left[\sum_{i=1}^{r_1} \mathbb{E}_{X,Y}\left[\log\left(\left(1-\frac{\eta_i\sigma^2}{b}X\right)^2 + \frac{\eta_i^2\sigma^4}{b^2}XY\right)\right]\right],$$

where r_1 is defined in (8), and 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 $(d-1)$. When $c^{(r)} = \mathbb{E}\left[\prod_{i=1}^{r_1}\left(1-2\eta_i\sigma^2+\frac{\eta_i^2\sigma^4}{b}(d+b+1)\right)\right] = 1$, we can compute that

$$\begin{aligned}\tilde{\rho}^{(r)} &\le \frac{1}{2}\mathbb{E}\left[\sum_{i=1}^{r_1}\log \mathbb{E}_{X,Y}\left[\left(\left(1-\frac{\eta_i\sigma^2}{b}X\right)^2 + \frac{\eta_i^2\sigma^4}{b^2}XY\right)\right]\right] \\ &= \frac{1}{2}\mathbb{E}\left[\sum_{i=1}^{r_1}\log\left(1-2\eta_i\sigma^2+\frac{\eta_i^2\sigma^4}{b}(d+b+1)\right)\right] = 0.\end{aligned} \tag{110}$$

Note that since $1-\frac{2\eta_i\sigma^2}{b}X + \frac{\eta_i^2\sigma^4}{b^2}(X^2+XY)$ is random, the inequality in (110) is a strict inequality from Jensen's inequality. Thus, when $c^{(r)} = 1$, we have $\tilde{\rho}^{(r)} < 0$. By continuity, there exists some $\delta > 0$ such that for any $1 < c^{(r)} < 1+\delta$ we have $\tilde{\rho}^{(r)} < 0$. Moreover, when $c^{(r)} > 1$, we have

$$\begin{aligned}h^{(r)}(2) &= \mathbb{E}\left[\prod_{i=1}^{r_1}\mathbb{E}\left[1-\frac{2\eta_i\sigma^2}{b}X+\frac{\eta_i^2\sigma^4}{b^2}(X^2+XY)\right]\right] \\ &= \mathbb{E}\left[\prod_{i=1}^{r_1}\left(1-2\eta_i\sigma^2+\frac{\eta_i^2\sigma^4}{b}(d+b+1)\right)\right] = c^{(r)} > 1,\end{aligned}$$

which implies that there exists some $0 < \alpha^{(r)} < 2$ such that $h^{(r)}\left(\alpha^{(r)}\right) = 1$.

Finally, let us prove (ii) and (iii). When $c^{(r)} \le 1$, we have $\tilde{h}^{(r)}(2) \le 1$, which implies that $\alpha^{(r)} \ge 2$. In particular, when $c^{(r)} = 1$, the tail-index $\alpha^{(r)} = 2$. The proof is complete. □

Proof of Lemma 3

We recall that

$$x_k = M_k x_{k-1} + q_k, \tag{111}$$

which implies that

$$\|x_k\| \le \|M_k x_{k-1}\| + \|q_k\|. \tag{112}$$

(i) For any $p \le 1$ and $\hat{h}^{(g)}(p) < 1$, by Lemma 14,

$$\|x_k\|^p \le \|M_k x_{k-1}\|^p + \|q_k\|^p. \tag{113}$$

Since M_k is independent of x_{k-1}, we have

$$\mathbb{E}\|x_k\|^p \le \mathbb{E}\|M_k\|^p \mathbb{E}\|x_{k-1}\|^p + \mathbb{E}\|q_k\|^p, \tag{114}$$

so that

$$\mathbb{E}\|x_k\|^p \le \hat{h}^{(g)}(p)\mathbb{E}\|x_{k-1}\|^p + \mathbb{E}\|q_1\|^p. \tag{115}$$

By iterating over k, we get

$$\mathbb{E}\|x_k\|^p \leq (\hat{h}^{(g)}(p))^k \mathbb{E}\|x_0\|^p + \frac{1-(\hat{h}^{(g)}(p))^k}{1-\hat{h}^{(g)}(p)} \mathbb{E}\|q_1\|^p. \tag{116}$$

(ii) For any $p > 1$ and $\hat{h}^{(g)}(p) < 1$, by Lemma 14, for any $\epsilon > 0$, we have

$$\|x_k\|^p \leq (1+\epsilon)\|M_k x_{k-1}\|^p + \frac{(1+\epsilon)^{\frac{p}{p-1}} - (1+\epsilon)}{\left((1+\epsilon)^{\frac{1}{p-1}} - 1\right)^p} \|q_k\|^p, \tag{117}$$

which (similar as in (i)) implies that

$$\mathbb{E}\|x_k\|^p \leq (1+\epsilon)\mathbb{E}\|M_k\|^p \mathbb{E}\|x_{k-1}\|^p + \frac{(1+\epsilon)^{\frac{p}{p-1}} - (1+\epsilon)}{\left((1+\epsilon)^{\frac{1}{p-1}} - 1\right)^p} \mathbb{E}\|q_k\|^p, \tag{118}$$

so that

$$\mathbb{E}\|x_k\|^p \leq (1+\epsilon)\hat{h}^{(g)}(p)\mathbb{E}\|x_{k-1}\|^p + \frac{(1+\epsilon)^{\frac{p}{p-1}} - (1+\epsilon)}{\left((1+\epsilon)^{\frac{1}{p-1}} - 1\right)^p} \mathbb{E}\|q_1\|^p. \tag{119}$$

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

$$\mathbb{E}\|x_k\|^p \leq ((1+\epsilon)\hat{h}^{(g)}(p))^k \mathbb{E}\|x_0\|^p + \frac{1-((1+\epsilon)\hat{h}^{(g)}(p))^k}{1-(1+\epsilon)\hat{h}^{(g)}(p)} \frac{(1+\epsilon)^{\frac{p}{p-1}} - (1+\epsilon)}{\left((1+\epsilon)^{\frac{1}{p-1}} - 1\right)^p} \mathbb{E}\|q_1\|^p. \tag{120}$$

The proof is complete. □

Proof of Theorem 18

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

$$x_k = M_k x_{k-1} + q_k, \tag{121}$$
$$\tilde{x}_k = M_k \tilde{x}_{k-1} + q_k, \tag{122}$$

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

$$x_k - \tilde{x}_k = M_k(x_{k-1} - \tilde{x}_{k-1}), \tag{123}$$

which yields that

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

which by iterating implies that

$$\mathcal{W}_p^p(\nu_k, \tilde{\nu}_k) \leq \mathbb{E}\|x_k - \tilde{x}_k\|^p \leq (\hat{h}^{(g)}(p))^k \mathbb{E}\|x_0 - \tilde{x}_0\|^p = (\hat{h}^{(g)}(p))^k \mathcal{W}_p^p(\nu_0, \tilde{\nu}_0). \tag{124}$$

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

$$\mathcal{W}_p(\nu_k, \nu_\infty) \leq \left(\left(\hat{h}^{(g)}(p)\right)^{1/q}\right)^k \mathcal{W}_p(\nu_0, \nu_\infty). \tag{125}$$

The proof is complete. □

Proof of Theorem 19

When the stationary distribution of the Markovian stepsizes is uniform on the set (44), we have

$$\hat{h}^{(g)}(s) = \frac{1}{K}\mathbb{E}\left[\left\|I - \frac{\bar{\eta}}{b}H\right\|^s\right] + \frac{1}{K}\sum_{j=1}^{\frac{K-1}{2}}\left(\mathbb{E}\left[\left\|I - \frac{\bar{\eta}-j\delta}{b}H\right\|^s\right] + \mathbb{E}\left[\left\|I - \frac{\bar{\eta}+j\delta}{b}H\right\|^s\right]\right). \tag{126}$$

It suffices to show that for any $s \geq 1$, $\hat{h}^{(g)}(s)$ is increasing in δ. It suffices to show that for any $s \geq 1$ and $j = 1, \ldots, \frac{K-1}{2}$,

$$\hat{h}_j^{(g)}(s) := \mathbb{E}\left[\left\|I - \frac{\bar{\eta}-j\delta}{b}H\right\|^s\right] + \mathbb{E}\left[\left\|I - \frac{\bar{\eta}+j\delta}{b}H\right\|^s\right] \tag{127}$$

is increasing in δ. By adapting the proof of Lemma 13 (Lemma 22 in Gürbüzbalaban et al. (2021)), one can show that the function

$$f(x) := \mathbb{E}\left[\left\|I - \frac{x}{b}H\right\|^s\right] \tag{128}$$

is convex in x for any $s \geq 1$. It remains to show that $f(\bar{\eta} - j\delta) + f(\bar{\eta} + j\delta)$ is increasing in δ. We claim that

$$F(x; a) := f(x - a) + f(x + a) \tag{129}$$

is increasing in x for any $x \geq a > 0$. To see this, we can compute that $F'(a; a) = 0$ and $F''(x; a) = f''(x - a) + f''(x + a) \geq 0$ since $f(x)$ is convex in x, which implies that $F'(x; a) \geq 0$ for any $x \geq a$ and thus $F(x; a)$ is increasing in x for any $x \geq a > 0$. Hence, the lower bound for the tail-index $\hat{\alpha}^{(g)}$ is decreasing δ provided that $\hat{\alpha}^{(g)} \geq 1$.

Next, let us show that $\hat{\alpha}^{(g)}$ is increasing in K (where we recall that K is odd without loss of generality) for any $\hat{\alpha}^{(g)} \geq 1$. Let $\hat{h}^{(g)}(s; K) = \hat{h}^{(g)}(s)$ that emphasizes the dependence on K. Let us show that $\hat{h}^{(g)}(s; K + 2) \geq \hat{h}^{(g)}(s; K)$ for any odd K and $s \geq 1$. We can compute that

$$\begin{aligned}
\hat{h}^{(g)}(s; K+2) - \hat{h}^{(g)}(s; K) = &\left(\frac{1}{K+2} - \frac{1}{K}\right)\mathbb{E}\left[\left\|I - \frac{\bar{\eta}}{b}H\right\|^s\right] \\
&+ \left(\frac{1}{K+2} - \frac{1}{K}\right)\sum_{j=1}^{\frac{K-1}{2}}\left(\mathbb{E}\left[\left\|I - \frac{\bar{\eta}-j\delta}{b}H\right\|^s\right] + \mathbb{E}\left[\left\|I - \frac{\bar{\eta}+j\delta}{b}H\right\|^s\right]\right) \\
&+ \frac{1}{K+2}\left(\mathbb{E}\left[\left\|I - \frac{\bar{\eta}-\frac{K+1}{2}\delta}{b}H\right\|^s\right] + \mathbb{E}\left[\left\|I - \frac{\bar{\eta}+\frac{K+1}{2}\delta}{b}H\right\|^s\right]\right).
\end{aligned}$$

Therefore, it suffices to show that

$$\begin{aligned}
&\mathbb{E}\left[\left\|I - \frac{\bar{\eta}-\frac{K+1}{2}\delta}{b}H\right\|^s\right] + \mathbb{E}\left[\left\|I - \frac{\bar{\eta}+\frac{K+1}{2}\delta}{b}H\right\|^s\right] \\
&\geq \frac{2}{K}\mathbb{E}\left[\left\|I - \frac{\bar{\eta}}{b}H\right\|^s\right] + \frac{2}{K}\sum_{j=1}^{\frac{K-1}{2}}\left(\mathbb{E}\left[\left\|I - \frac{\bar{\eta}-j\delta}{b}H\right\|^s\right] + \mathbb{E}\left[\left\|I - \frac{\bar{\eta}+j\delta}{b}H\right\|^s\right]\right).
\end{aligned} \tag{130}$$

Since the function $f(x)$ defined in (128) is convex for any $s \geq 1$, for any $j = 0, 1, 2, \ldots, \frac{K-1}{2}$,

$$\begin{aligned}
&\mathbb{E}\left[\left\|I - \frac{\bar{\eta}-j\delta}{b}H\right\|^s\right] + \mathbb{E}\left[\left\|I - \frac{\bar{\eta}+j\delta}{b}H\right\|^s\right] \\
&\leq \mathbb{E}\left[\left\|I - \frac{\bar{\eta}-\frac{K+1}{2}\delta}{b}H\right\|^s\right] + \mathbb{E}\left[\left\|I - \frac{\bar{\eta}+\frac{K+1}{2}\delta}{b}H\right\|^s\right],
\end{aligned} \tag{131}$$

which implies that

$$\begin{aligned}
&\frac{2}{K}\mathbb{E}\left[\left\|I-\frac{\bar{\eta}}{b}H\right\|^s\right]+\frac{2}{K}\sum_{j=1}^{\frac{K-1}{2}}\left(\mathbb{E}\left[\left\|I-\frac{\bar{\eta}-j\delta}{b}H\right\|^s\right]+\mathbb{E}\left[\left\|I-\frac{\bar{\eta}+j\delta}{b}H\right\|^s\right]\right)\\
&\leq\left(\frac{1}{K}+\frac{2}{K}\frac{K-1}{2}\right)\left(\mathbb{E}\left[\left\|I-\frac{\bar{\eta}-\frac{K+1}{2}\delta}{b}H\right\|^s\right]+\mathbb{E}\left[\left\|I-\frac{\bar{\eta}+\frac{K+1}{2}\delta}{b}H\right\|^s\right]\right)\\
&=\mathbb{E}\left[\left\|I-\frac{\bar{\eta}-\frac{K+1}{2}\delta}{b}H\right\|^s\right]+\mathbb{E}\left[\left\|I-\frac{\bar{\eta}+\frac{K+1}{2}\delta}{b}H\right\|^s\right],
\end{aligned}$$

which proves (130). Hence, the lower bound for the tail-index $\hat{\alpha}^{(g)}$ is decreasing K provided that $\hat{\alpha}^{(g)}\geq 1$. The proof is complete. □

Proof of Theorem 20

When the stationary distribution of the Markovian stepsizes is uniform on the set (45), we have

$$\begin{aligned}
\hat{h}^{(g)}(s)=&\frac{1}{2^n+1}\mathbb{E}\left[\left\|I-\frac{\bar{\eta}}{b}H\right\|^s\right]\\
&+\frac{1}{2^n+1}\sum_{j=1}^{2^{n-1}}\left(\mathbb{E}\left[\left\|I-\frac{\bar{\eta}-j\frac{R}{2^{n-1}}}{b}H\right\|^s\right]+\mathbb{E}\left[\left\|I-\frac{\bar{\eta}+j\frac{R}{2^{n-1}}}{b}H\right\|^s\right]\right).
\end{aligned}$$

Let us use the notation $\hat{h}^{(g)}(s;n):=\hat{h}^{(g)}(s)$ to emphasize the dependence on n. We can compute that

$$\begin{aligned}
&\hat{h}^{(g)}(s;n)-\hat{h}^{(g)}(s;n+1)\\
&=\left(\frac{1}{2^n+1}-\frac{1}{2^{n+1}+1}\right)\mathbb{E}\left[\left\|I-\frac{\bar{\eta}}{b}H\right\|^s\right]\\
&\quad+\frac{1}{2^n+1}\sum_{j=1}^{2^{n-1}}\left(\mathbb{E}\left[\left\|I-\frac{\bar{\eta}-j\frac{R}{2^{n-1}}}{b}H\right\|^s\right]+\mathbb{E}\left[\left\|I-\frac{\bar{\eta}+j\frac{R}{2^{n-1}}}{b}H\right\|^s\right]\right)\\
&\quad-\frac{1}{2^{n+1}+1}\sum_{j=1}^{2^n}\left(\mathbb{E}\left[\left\|I-\frac{\bar{\eta}-j\frac{R}{2^n}}{b}H\right\|^s\right]+\mathbb{E}\left[\left\|I-\frac{\bar{\eta}+j\frac{R}{2^n}}{b}H\right\|^s\right]\right)\\
&=\left(\frac{1}{2^n+1}-\frac{1}{2^{n+1}+1}\right)\mathbb{E}\left[\left\|I-\frac{\bar{\eta}}{b}H\right\|^s\right]\\
&\quad+\left(\frac{1}{2^n+1}-\frac{1}{2^{n+1}+1}\right)\sum_{j=1}^{2^{n-1}}\left(\mathbb{E}\left[\left\|I-\frac{\bar{\eta}-j\frac{R}{2^{n-1}}}{b}H\right\|^s\right]+\mathbb{E}\left[\left\|I-\frac{\bar{\eta}+j\frac{R}{2^{n-1}}}{b}H\right\|^s\right]\right)\\
&\quad-\frac{1}{2^{n+1}+1}\sum_{j=1}^{2^{n-1}}\left(\mathbb{E}\left[\left\|I-\frac{\bar{\eta}-(2j-1)\frac{R}{2^n}}{b}H\right\|^s\right]+\mathbb{E}\left[\left\|I-\frac{\bar{\eta}+(2j-1)\frac{R}{2^n}}{b}H\right\|^s\right]\right).
\end{aligned}$$

By adapting the proof of Lemma 13 (Lemma 22 in Gürbüzbalaban et al. (2021)), one can show that the function

$$f(x):=\mathbb{E}\left[\left\|I-\frac{x}{b}H\right\|^s\right] \tag{132}$$

is convex in x for any $s\geq 1$. Therefore, by Jensen's inequality,

$$\mathbb{E}\left[\left\|I-\frac{\bar{\eta}-(2j-1)\frac{R}{2^n}}{b}H\right\|^s\right]\leq\frac{1}{2}\mathbb{E}\left[\left\|I-\frac{\bar{\eta}-(j-1)\frac{R}{2^{n-1}}}{b}H\right\|^s\right]+\frac{1}{2}\mathbb{E}\left[\left\|I-\frac{\bar{\eta}-j\frac{R}{2^{n-1}}}{b}H\right\|^s\right],$$

and similarly

$$\mathbb{E}\left[\left\|I-\frac{\bar{\eta}+(2j-1)\frac{R}{2^n}}{b}H\right\|^s\right]\leq\frac{1}{2}\mathbb{E}\left[\left\|I-\frac{\bar{\eta}+(j-1)\frac{R}{2^{n-1}}}{b}H\right\|^s\right]+\frac{1}{2}\mathbb{E}\left[\left\|I-\frac{\bar{\eta}+j\frac{R}{2^{n-1}}}{b}H\right\|^s\right],$$

which implies that

$$\begin{aligned}
&\hat{h}^{(g)}(s;n)-\hat{h}^{(g)}(s;n+1)\\
&\geq\left(\frac{1}{2^n+1}-\frac{2}{2^{n+1}+1}\right)\mathbb{E}\left[\left\|I-\frac{\bar{\eta}}{b}H\right\|^s\right]\\
&\quad+\left(\frac{1}{2^n+1}-\frac{2}{2^{n+1}+1}\right)\sum_{j=1}^{2^{n-1}-1}\left(\mathbb{E}\left[\left\|I-\frac{\bar{\eta}-j\frac{R}{2^{n-1}}}{b}H\right\|^s\right]+\mathbb{E}\left[\left\|I-\frac{\bar{\eta}+j\frac{R}{2^{n-1}}}{b}H\right\|^s\right]\right)\\
&\quad+\left(\frac{1}{2^n+1}-\frac{\frac{3}{2}}{2^{n+1}+1}\right)\left(\mathbb{E}\left[\left\|I-\frac{\bar{\eta}-2^{n-1}\frac{R}{2^{n-1}}}{b}H\right\|^s\right]+\mathbb{E}\left[\left\|I-\frac{\bar{\eta}+2^{n-1}\frac{R}{2^{n-1}}}{b}H\right\|^s\right]\right)\\
&=-\frac{1}{(2^n+1)(2^{n+1}+1)}\mathbb{E}\left[\left\|I-\frac{\bar{\eta}}{b}H\right\|^s\right]\\
&\quad-\frac{1}{(2^n+1)(2^{n+1}+1)}\sum_{j=1}^{2^{n-1}-1}\left(\mathbb{E}\left[\left\|I-\frac{\bar{\eta}-j\frac{R}{2^{n-1}}}{b}H\right\|^s\right]+\mathbb{E}\left[\left\|I-\frac{\bar{\eta}+j\frac{R}{2^{n-1}}}{b}H\right\|^s\right]\right)\\
&\quad+\left(\frac{1}{2^n+1}-\frac{\frac{3}{2}}{2^{n+1}+1}\right)\left(\mathbb{E}\left[\left\|I-\frac{\bar{\eta}-2^{n-1}\frac{R}{2^{n-1}}}{b}H\right\|^s\right]+\mathbb{E}\left[\left\|I-\frac{\bar{\eta}+2^{n-1}\frac{R}{2^{n-1}}}{b}H\right\|^s\right]\right).
\end{aligned}$$

Since we proved in the proof of Theorem 19 that $f(x-a)+f(x+a)$ is increasing in x for any $x\geq a>0$, we have

$$\begin{aligned}
&\hat{h}^{(g)}(s;n)-\hat{h}^{(g)}(s;n+1)\\
&\geq-\frac{1}{(2^n+1)(2^{n+1}+1)}\frac{1}{2}\left(\mathbb{E}\left[\left\|I-\frac{\bar{\eta}-2^{n-1}\frac{R}{2^{n-1}}}{b}H\right\|^s\right]+\mathbb{E}\left[\left\|I-\frac{\bar{\eta}+2^{n-1}\frac{R}{2^{n-1}}}{b}H\right\|^s\right]\right)\\
&\quad-\frac{1}{(2^n+1)(2^{n+1}+1)}\cdot\sum_{j=1}^{2^{n-1}-1}\left(\mathbb{E}\left[\left\|I-\frac{\bar{\eta}-2^{n-1}\frac{R}{2^{n-1}}}{b}H\right\|^s\right]+\mathbb{E}\left[\left\|I-\frac{\bar{\eta}+2^{n-1}\frac{R}{2^{n-1}}}{b}H\right\|^s\right]\right)\\
&\quad+\left(\frac{1}{2^n+1}-\frac{\frac{3}{2}}{2^{n+1}+1}\right)\cdot\left(\mathbb{E}\left[\left\|I-\frac{\bar{\eta}-2^{n-1}\frac{R}{2^{n-1}}}{b}H\right\|^s\right]+\mathbb{E}\left[\left\|I-\frac{\bar{\eta}+2^{n-1}\frac{R}{2^{n-1}}}{b}H\right\|^s\right]\right)=0.
\end{aligned}$$

Hence $\hat{h}^{(g)}(s;n)$ is decreasing in n provided that $s\geq 1$ and therefore the lower bound for the tail-index $\hat{\alpha}^{(g)}$ is increasing in n provided that $\hat{\alpha}^{(g)}\geq 1$. This completes the proof. □

Proof of Proposition 10

Under the assumption that the stationary distribution of the Markovian stepsizes is uniform on the set (6), we have

$$\mathbb{P}(\eta=\eta_i)=\frac{1}{m},\qquad i=1,2,\ldots,m,\tag{133}$$

so that

$$\hat{h}^{(g)}(s)=\mathbb{E}\left[\left\|I-\frac{\eta}{b}H\right\|^s\right]=\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left[\left\|I-\frac{\eta_i}{b}H\right\|^s\right].\tag{134}$$

On the other hand, we recall that the lower bound for the tail-index $\hat{\alpha}^{(m)}$ for the SGD with cyclic stepsizes is the unique positive value such that $\hat{h}^{(m)}\left(\hat{\alpha}^{(m)}\right)=1$. By the inequality of arithmetic and geometric means,

we obtain

$$\hat{h}^{(m)}(s) \leq \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left[\left\|I-\frac{\eta_i}{b}H\right\|^s\right] = \hat{h}^{(g)}(s). \tag{135}$$

Since η_i is not constant, the above inequality is strict. Therefore, we conclude that the lower bound for the tail-index $\hat{\alpha}^{(g)}$ is strictly less than the lower bound for the tail-index $\hat{\alpha}^{(m)}$ for SGD with cyclic stepsizes. The proof is complete. □

Proof of Lemma 4

The proof is similar to the proof of Lemma 1 and is hence omitted here. □

Proof of Theorem 21

The proof is similar to the proof of Theorem 14 and is hence omitted here. □

Proof of Corollary 3

The proof is similar to the proof of Corollary 1 and is hence omitted here. □

Proof of Lemma 5

First of all, the Markov chain exhibits a unique stationary distribution $\pi_i := \mathbb{P}(\eta_0 = \eta_i)$ that satisfy the equations:

$$\begin{aligned}
&\pi_1 = (1-p)\pi_2 + p\pi_m, \qquad \pi_2 = \pi_1 + (1-p)\pi_3,\\
&\pi_3 = p\pi_2 + (1-p)\pi_4,\\
&\cdots\cdots\\
&\pi_{K-2} = p\pi_{K-3} + (1-p)\pi_{K-1}, \qquad \pi_{K-1} = p\pi_{K-2},\\
&\pi_K = p\pi_{K-1} + (1-p)\pi_{K+1}, \qquad \pi_{K+1} = \pi_K + (1-p)\pi_{K+2},\\
&\pi_{K+2} = p\pi_{K+1} + (1-p)\pi_{K+3},\\
&\cdots\cdots\\
&\pi_{m-1} = p\pi_{m-2} + (1-p)\pi_m, \qquad \pi_m = p\pi_{m-1}.
\end{aligned}$$

Let us solve for $(\pi_i)_{i=1}^m$. First, $\pi_{m-1} = \frac{\pi_m}{p}$ and for any $K+1 \leq i \leq m-2$, we have

$$\pi_{i+1} = p\pi_i + (1-p)\pi_{i+2}, \tag{136}$$

and we can solve the characteristic equation:

$$(1-p)x^2 - x + p = 0, \tag{137}$$

to obtain $x = \frac{p}{1-p}$ or $x = 1$, which implies that for any $K+1 \leq i \leq m-2$,

$$\pi_i = d_1\left(\frac{p}{1-p}\right)^i + d_2, \tag{138}$$

where d_1 and d_2 can be determined via the equations:

$$d_1\left(\frac{p}{1-p}\right)^m + d_2 = \pi_m, \tag{139}$$

$$d_1\left(\frac{p}{1-p}\right)^{m-1} + d_2 = \frac{\pi_m}{p}, \tag{140}$$

so that

$$d_1 = \frac{p-1}{2p-1}\left(\frac{1-p}{p}\right)^m \pi_m, \qquad d_2 = \frac{p}{2p-1}\pi_m. \tag{141}$$

Hence, for any $K+1 \le i \le m-2$, we have

$$\pi_i = \frac{p-1}{2p-1}\left(\frac{1-p}{p}\right)^{m-i} \pi_m + \frac{p}{2p-1}\pi_m. \tag{142}$$

Therefore,

$$\begin{aligned}
\pi_K &= \pi_{K+1} - (1-p)\pi_{K+2} \\
&= \frac{p-1}{2p-1}\left(\frac{1-p}{p}\right)^{m-K-1} \pi_m + \frac{p}{2p-1}\pi_m - (1-p)\left(\frac{p-1}{2p-1}\left(\frac{1-p}{p}\right)^{m-K-2} \pi_m + \frac{p}{2p-1}\pi_m\right) \\
&= \frac{p(p-1)}{2p-1}\left(\frac{1-p}{p}\right)^{m-K} \pi_m + \frac{p^2}{2p-1}\pi_m,
\end{aligned}$$

and

$$\begin{aligned}
\pi_{K-1} &= \frac{\pi_K}{p} - \frac{1-p}{p}\pi_{K+1} \\
&= \frac{p-1}{2p-1}\left(\frac{1-p}{p}\right)^{m-K} \pi_m + \frac{p}{2p-1}\pi_m - \frac{p-1}{2p-1}\left(\frac{1-p}{p}\right)^{m-K} \pi_m - \frac{1-p}{2p-1}\pi_m \\
&= \pi_m.
\end{aligned}$$

Similar as before, we obtain that $\pi_{K-2} = \frac{\pi_m}{p}$ and for any $2 \le i \le K-3$,

$$\pi_i = \frac{p-1}{2p-1}\left(\frac{1-p}{p}\right)^{K-i} \pi_m + \frac{p}{2p-1}\pi_m. \tag{143}$$

Moreover, we can compute that

$$\begin{aligned}
\pi_1 &= \pi_2 - (1-p)\pi_3 \\
&= \frac{p-1}{2p-1}\left(\frac{1-p}{p}\right)^{K-2} \pi_m + \frac{p}{2p-1}\pi_m - (1-p)\left(\frac{p-1}{2p-1}\left(\frac{1-p}{p}\right)^{K-3} \pi_m + \frac{p}{2p-1}\pi_m\right) \\
&= (1-p)\frac{p-1}{2p-1}\left(\frac{1-p}{p}\right)^{K-2} \pi_m + \frac{p^2}{2p-1}\pi_m.
\end{aligned}$$

Finally, the constraint $\sum_{i=1}^m \pi_i = 1$ yields that

$$\begin{aligned}
&(1-p)\frac{p-1}{2p-1}\left(\frac{1-p}{p}\right)^{K-2} \pi_m + \frac{p^2}{2p-1}\pi_m + \sum_{i=2}^{K-1}\left(\frac{p-1}{2p-1}\left(\frac{1-p}{p}\right)^{K-i} \pi_m + \frac{p}{2p-1}\pi_m\right) \\
&\quad + \frac{p(p-1)}{2p-1}\left(\frac{1-p}{p}\right)^{m-K} \pi_m + \frac{p^2}{2p-1}\pi_m + \sum_{i=K+1}^{m}\left(\frac{p-1}{2p-1}\left(\frac{1-p}{p}\right)^{m-i} \pi_m + \frac{p}{2p-1}\pi_m\right) = 1,
\end{aligned}$$

which implies that

$$\begin{aligned}
&-\frac{(1-p)^2}{2p-1}\left(\frac{1-p}{p}\right)^{K-2} + \frac{2p^2}{2p-1} + \frac{(m-2)p}{2p-1} + \frac{(1-p)^2}{(2p-1)^2}\left(\left(\frac{1-p}{p}\right)^{K-2} - 1\right) \\
&\quad + \frac{p(p-1)}{2p-1}\left(\frac{1-p}{p}\right)^{m-K} + \frac{p(p-1)}{(2p-1)^2}\left(1 - \left(\frac{1-p}{p}\right)^{m-K}\right) = \frac{1}{\pi_m},
\end{aligned}$$

so that

$$\frac{2p^2+(m-2)p}{2p-1}+\frac{2(1-p)^3}{(2p-1)^2}\left(\frac{1-p}{p}\right)^{K-2} + \frac{2p(p-1)^2}{(2p-1)^2}\left(\frac{1-p}{p}\right)^{m-K}+\frac{p-1}{(2p-1)^2}=\frac{1}{\pi_m},$$

which implies that

$$\pi_m=\left(\frac{4p^3+2(m-3)p^2-(m-3)p-1}{(2p-1)^2}+\frac{2p^3}{(2p-1)^2}\left(\frac{1-p}{p}\right)^{K+1}+\frac{2p(p-1)^2}{(2p-1)^2}\left(\frac{1-p}{p}\right)^{m-K}\right)^{-1}.$$

This completes the proof. □

Proof of Proposition 11

We can compute that

$$h^{(r)}(s;\eta_1,\eta_j)=\mathbb{E}_H\left[\left\|\left(I-\frac{\eta_2}{b}H\right)e_1\right\|^s\right]\left(1_{j=2}+1_{j\neq 2}h^{(r)}(s;\eta_2,\eta_j)\right),$$
$$h^{(r)}(s;\eta_K,\eta_j)=\mathbb{E}_H\left[\left\|\left(I-\frac{\eta_{K+1}}{b}H\right)e_1\right\|^s\right]\left(1_{j=K+1}+1_{j\neq K+1}h^{(r)}(s;\eta_{K+1},\eta_j)\right),$$

and for any $i=2,\ldots,K-1,K+1,\ldots,m$,

$$h^{(r)}(s;\eta_i,\eta_j)=p\mathbb{E}_H\left[\left\|\left(I-\frac{\eta_{i+1}}{b}H\right)e_1\right\|^s\right]\left(1_{j=i+1}+1_{j\neq i+1}h^{(r)}(s;\eta_{i+1},\eta_j)\right) + (1-p)\mathbb{E}_H\left[\left\|\left(I-\frac{\eta_{i-1}}{b}H\right)e_1\right\|^s\right]\left(1_{j=i-1}+1_{j\neq i-1}h^{(r)}(s;\eta_{i-1},\eta_j)\right).$$

To simplify the notation, we define:

$$h_{ij}:=h^{(r)}(s;\eta_i,\eta_j),\qquad a_i:=\mathbb{E}_H\left[\left\|\left(I-\frac{\eta_i}{b}H\right)e_1\right\|^s\right]. \tag{144}$$

Then, we have

$$h_{1j}=a_2\left(1_{j=2}+1_{j\neq 2}h_{2j}\right),$$
$$h_{Kj}=a_{K+1}\left(1_{j=K+1}+1_{j\neq K+1}h_{(K+1)j}\right),$$

and for any $i=2,\ldots,K-1,K+1,\ldots,m$,

$$h_{ij}=pa_{i+1}\left(1_{j=i+1}+1_{j\neq i+1}h_{(i+1)j}\right)+(1-p)a_{i-1}\left(1_{j=i-1}+1_{j\neq i-1}h_{(i-1)j}\right).$$

Let us define the vectors $h^j=[h_{1j},h_{2j},\ldots,h_{mj}]^T$, $p^j=[p_{1j},p_{2j},\ldots,p_{mj}]^T$, where for any $i=2,\ldots,K-1,K+1,\ldots,m$

$$p_{ij}=pa_{i+1}1_{j=i+1}+(1-p)a_{i-1}1_{j=i-1}, \tag{145}$$

and

$$p_{1j}=a_2 1_{j=2},\qquad p_{Kj}=a_{K+1}1_{j=K+1},$$

and the matrices $Q^j=(Q^j_{i\ell})_{1\leq i,\ell\leq m}$ such that for any $i=2,\ldots,K-1,K+1,\ldots,m$

$$Q^j_{i\ell}=pa_{i+1}1_{j\neq i+1}1_{\ell=i+1}+(1-p)a_{i-1}1_{j\neq i-1}1_{\ell=i-1}, \tag{146}$$

and

$$Q^j_{1\ell}=1_{j\neq 2}1_{\ell=2},\qquad Q^j_{K\ell}=1_{j\neq K+1}1_{\ell=K+1}.$$

Thus, we have

$$h^j=p^j+Q^jh^j, \tag{147}$$

such that

$$h^j=(I-Q^j)^{-1}p^j. \tag{148}$$

This completes the proof. □

Proof of Lemma 6

It is easy to compute that:

$$\mathbb{P}(r_1 = 1) = 1 - p, \qquad \mathbb{P}(r_1 = k) = p^2(1-p)^{k-2}, \qquad k = 2, 3, \ldots, \tag{149}$$

where r_1 is defined in (8). Conditional on $\eta_0 = \eta_l$, we have

$$\begin{aligned}
&\mathbb{E}_{\eta_0=\eta_l}\left[\prod_{i=1}^{r_1}\mathbb{E}_H\left[\left\|\left(I-\frac{\eta_i}{b}H\right)e_1\right\|^s\right]\right] \\
&= (1-p)\mathbb{E}_H\left[\left\|\left(I-\frac{\eta_l}{b}H\right)e_1\right\|^s\right] + \sum_{k=2}^{\infty}p^2(1-p)^{k-2}\mathbb{E}_H\left[\left\|\left(I-\frac{\eta_l}{b}H\right)e_1\right\|^s\right]\left(\mathbb{E}_H\left[\left\|\left(I-\frac{\eta_u}{b}H\right)e_1\right\|^s\right]\right)^{k-1} \\
&= \frac{\mathbb{E}_H\left[\left\|\left(I-\frac{\eta_l}{b}H\right)e_1\right\|^s\right]\left(1-p+(2p-1)\mathbb{E}_H\left[\left\|\left(I-\frac{\eta_u}{b}H\right)e_1\right\|^s\right]\right)}{1-(1-p)\mathbb{E}_H\left[\left\|\left(I-\frac{\eta_u}{b}H\right)e_1\right\|^s\right]},
\end{aligned} \tag{150}$$

where we used the assumption that $(1-p)\mathbb{E}_H\left[\left\|\left(I-\frac{\eta_u}{b}H\right)e_1\right\|^s\right] < 1$ and moreover

$$\begin{aligned}
&\mathbb{E}_{\eta_0=\eta_l}\left[\sum_{i=1}^{r_1}\mathbb{E}_H\left[\log\left\|\left(I-\frac{\eta_i}{b}H\right)e_1\right\|\right]\right] \\
&= (1-p)\mathbb{E}_H\left[\log\left\|\left(I-\frac{\eta_l}{b}H\right)e_1\right\|\right] \\
&\quad + \sum_{k=2}^{\infty}p^2(1-p)^{k-2}\left(\mathbb{E}_H\left[\log\left\|\left(I-\frac{\eta_l}{b}H\right)e_1\right\|\right] + (k-1)\mathbb{E}_H\left[\log\left\|\left(I-\frac{\eta_u}{b}H\right)e_1\right\|\right]\right) \\
&= \mathbb{E}_H\left[\log\left\|\left(I-\frac{\eta_l}{b}H\right)e_1\right\|\right] + \mathbb{E}_H\left[\log\left\|\left(I-\frac{\eta_u}{b}H\right)e_1\right\|\right],
\end{aligned}$$

where we applied Lemma 15 to obtain the last equality above.

Similarly, we can compute that

$$\begin{aligned}
&\mathbb{E}_{\eta_0=\eta_u}\left[\prod_{i=1}^{r_1}\mathbb{E}_H\left[\left\|\left(I-\frac{\eta_i}{b}H\right)e_1\right\|^s\right]\right] \\
&= \frac{\mathbb{E}_H\left[\left\|\left(I-\frac{\eta_u}{b}H\right)e_1\right\|^s\right]\left(1-p+(2p-1)\mathbb{E}_H\left[\left\|\left(I-\frac{\eta_l}{b}H\right)e_1\right\|^s\right]\right)}{1-(1-p)\mathbb{E}_H\left[\left\|\left(I-\frac{\eta_l}{b}H\right)e_1\right\|^s\right]},
\end{aligned} \tag{151}$$

where we used the assumption that $(1-p)\mathbb{E}_H\left[\left\|\left(I-\frac{\eta_l}{b}H\right)e_1\right\|^s\right] < 1$ and moreover

$$\mathbb{E}_{\eta_0=\eta_u}\left[\sum_{i=1}^{r_1}\mathbb{E}_H\left[\log\left\|\left(I-\frac{\eta_i}{b}H\right)e_1\right\|\right]\right] = \mathbb{E}_H\left[\log\left\|\left(I-\frac{\eta_l}{b}H\right)e_1\right\|\right] + \mathbb{E}_H\left[\log\left\|\left(I-\frac{\eta_u}{b}H\right)e_1\right\|\right].$$

Since the Markov chain exhibits a unique stationary distribution $\mathbb{P}(\eta_0 = \eta_\ell) = \mathbb{P}(\eta_0 = \eta_u) = \frac{1}{2}$, we conclude that

$$\begin{aligned}
\mathbb{E}\left[\prod_{i=1}^{r_1}\mathbb{E}_H\left[\left\|\left(I-\frac{\eta_i}{b}H\right)e_1\right\|^s\right]\right] &= \frac{\mathbb{E}_H\left[\left\|\left(I-\frac{\eta_l}{b}H\right)e_1\right\|^s\right]\left(1-p+(2p-1)\mathbb{E}_H\left[\left\|\left(I-\frac{\eta_u}{b}H\right)e_1\right\|^s\right]\right)}{2\left(1-(1-p)\mathbb{E}_H\left[\left\|\left(I-\frac{\eta_u}{b}H\right)e_1\right\|^s\right]\right)} \\
&\quad + \frac{\mathbb{E}_H\left[\left\|\left(I-\frac{\eta_u}{b}H\right)e_1\right\|^s\right]\left(1-p+(2p-1)\mathbb{E}_H\left[\left\|\left(I-\frac{\eta_l}{b}H\right)e_1\right\|^s\right]\right)}{2\left(1-(1-p)\mathbb{E}_H\left[\left\|\left(I-\frac{\eta_l}{b}H\right)e_1\right\|^s\right]\right)},
\end{aligned} \tag{152}$$

and

$$\mathbb{E}\left[\sum_{i=1}^{r_1}\mathbb{E}_H\left[\log\left\|\left(I-\frac{\eta_i}{b}H\right)e_1\right\|\right]\right] = \mathbb{E}_H\left[\log\left\|\left(I-\frac{\eta_l}{b}H\right)e_1\right\|\right] + \mathbb{E}_H\left[\log\left\|\left(I-\frac{\eta_u}{b}H\right)e_1\right\|\right].$$

The proof is complete. □

Proof of Corollary 4

Since $c^{(r)} = \tilde{h}^{(r)}(2)$, it immediately follows from Lemma 6 that

$$c^{(r)} = \frac{\mathbb{E}_H\left[\left\|\left(I - \frac{\eta_l}{b}H\right)e_1\right\|^2\right](1-p+(2p-1)\mathbb{E}_H\left[\left\|\left(I-\frac{\eta_u}{b}H\right)e_1\right\|^2\right])}{2(1-(1-p)\mathbb{E}_H\left[\left\|\left(I-\frac{\eta_u}{b}H\right)e_1\right\|^2\right])} + \frac{\mathbb{E}_H\left[\left\|\left(I - \frac{\eta_u}{b}H\right)e_1\right\|^2\right](1-p+(2p-1)\mathbb{E}_H\left[\left\|\left(I-\frac{\eta_l}{b}H\right)e_1\right\|^2\right])}{2(1-(1-p)\mathbb{E}_H\left[\left\|\left(I-\frac{\eta_l}{b}H\right)e_1\right\|^2\right])}. \tag{153}$$

Moreover, we can compute that

$$\begin{aligned}\mathbb{E}_H\left[\left\|\left(I-\frac{\eta_l}{b}H\right)e_1\right\|^2\right] &= \mathbb{E}\left[1 - \frac{2\eta_l\sigma^2}{b}X + \frac{\eta_l^2\sigma^4}{b^2}(X^2+XY)\right] \\ &= 1 - 2\eta_l\sigma^2 + \frac{\eta_l^2\sigma^4}{b}(d+b+1),\end{aligned} \tag{154}$$

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 have

$$\mathbb{E}_H\left[\left\|\left(I-\frac{\eta_u}{b}H\right)e_1\right\|^2\right] = 1 - 2\eta_u\sigma^2 + \frac{\eta_u^2\sigma^4}{b}(d+b+1). \tag{155}$$

Finally, by plugging (154) and (155) into (153), we complete the proof. □

D.6 Proofs of Results in Section C

Proof of Lemma 7

If we have i.i.d. Guassian data, i.e. $a_i \sim \mathcal{N}(0, \sigma^2 I_d)$ are Gaussian distributed for every i, then conditional on the stepsize η_k, 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 $\|M_1 e_1\|$ where we chose $x = e_1$, where e_1 is the first basis vector in $\mathbb{R}^d$.

To see this, for any $x \in \mathbb{R}^d$ with $\|x\| = 1$, we can write $x = Re_1$ for some orthonormal matrix R, where e_1 is the first basis vector in $\mathbb{R}^d$. Define $b_i := R^T a_i$, here $a_i \sim \mathcal{N}(0, \sigma^2 I_d)$, and since R is orthonormal, b_i are also i.i.d. $\mathcal{N}(0, \sigma^2 I_d)$ distributed. Then, we can compute that

$$\begin{aligned}\|M_k x\| = \left\|\left(I - \frac{\eta_k}{b}\sum_{i\in\Omega_k} a_i a_i^T\right)x\right\| &= \left\|\left(RR^T - \frac{\eta_k}{b}\sum_{i\in\Omega_k} R b_i b_i^T R^T\right)Re_1\right\| \\ &= \left\|R\left(I - \frac{\eta_k}{b}\sum_{i\in\Omega_k} b_i b_i^T\right)R^T Re_1\right\| \\ &= \left\|\left(I - \frac{\eta_k}{b}\sum_{i\in\Omega_k} b_i b_i^T\right)e_1\right\|,\end{aligned}$$

which has the same distribution as $\|M_k e_1\|$. By following the similar arguments as the proof of Theorem 3 in Gürbüzbalaban et al. (2021), the conclusion follows. □

Proof of Lemma 8

Conditional on the stepsize η, it follows from Lemma 19 in Gürbüzbalaban et al. (2021) that for any $s \geq 0$,

$$\mathbb{E}\left[\left\|\left(I-\frac{\eta}{b}H\right)e_1\right\|^s \middle| \eta\right] = \mathbb{E}\left[\left(\left(1-\frac{\eta\sigma^2}{b}X\right)^2 + \frac{\eta^2\sigma^4}{b^2}XY\right)^{s/2} \middle| \eta\right],$$

and

$$\mathbb{E}\left[\log\left\|\left(I-\frac{\eta}{b}H\right)e_1\right\|\Big|\eta\right]=\frac{1}{2}\mathbb{E}\left[\log\left(\left(1-\frac{\eta\sigma^2}{b}X\right)^2+\frac{\eta^2\sigma^4}{b^2}XY\right)\Big|\eta\right],$$

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 $(d-1)$. Hence, the conclusion follows. □

Proof of Lemma 9

We follow the similar arguments as the proof of Theorem 3 in Gürbüzbalaban et al. (2021) and the key observation is that the distribution of $\left\|M_1^{(m)}\right\|/\|x\| = \|M_m M_{m-1}\cdots M_1 x\|/\|x\|$ is the same for every $x \in \mathbb{R}^d\backslash\{0\}$. For any $x\in\mathbb{R}^d$ with $\|x\|=1$, we can write $x=Re_1$ for some orthonormal matrix R, where e_1 is the first basis vector in $\mathbb{R}^d$. Define $b_i := R^T a_i$, here $a_i \sim \mathcal{N}(0,\sigma^2 I_d)$, and since R is orthonormal, b_i are also i.i.d. $\mathcal{N}(0,\sigma^2 I_d)$ distributed. Then, we can compute that

$$\begin{aligned}
\left\|M_1^{(m)}\right\| &= \|M_m M_{m-1}\cdots M_1 x\| \\
&= \left\|\left(I-\frac{\eta_m}{b}\sum_{i\in\Omega_m}a_i a_i^T\right)\left(I-\frac{\eta_{m-1}}{b}\sum_{i\in\Omega_{m-1}}a_i a_i^T\right)\cdots\left(I-\frac{\eta_1}{b}\sum_{i\in\Omega_1}a_i a_i^T\right)x\right\| \\
&= \left\|R\left(I-\frac{\eta_m}{b}\sum_{i\in\Omega_m}b_i b_i^T\right)R^TR\left(I-\frac{\eta_{m-1}}{b}\sum_{i\in\Omega_{m-1}}b_i b_i^T\right)R^T\cdots R\left(I-\frac{\eta_1}{b}\sum_{i\in\Omega_1}b_i b_i^T\right)R^TRe_1\right\| \\
&= \left\|R\left(I-\frac{\eta_m}{b}\sum_{i\in\Omega_m}b_i b_i^T\right)\left(I-\frac{\eta_{m-1}}{b}\sum_{i\in\Omega_{m-1}}b_i b_i^T\right)\cdots\left(I-\frac{\eta_1}{b}\sum_{i\in\Omega_1}b_i b_i^T\right)e_1\right\| \\
&= \left\|\left(I-\frac{\eta_m}{b}\sum_{i\in\Omega_m}b_i b_i^T\right)\left(I-\frac{\eta_{m-1}}{b}\sum_{i\in\Omega_{m-1}}b_i b_i^T\right)\cdots\left(I-\frac{\eta_1}{b}\sum_{i\in\Omega_1}b_i b_i^T\right)e_1\right\|,
\end{aligned}$$

which has the same distribution as $\|M_m M_{m-1}\cdots M_1 x\|/\|e_1\|$. By following the similar arguments as the proof of Theorem 3 in Gürbüzbalaban et al. (2021), we obtain:

$$h^{(m)}(s)=\mathbb{E}\left[\left\|\left(I-\frac{\eta_m}{b}H_m\right)\left(I-\frac{\eta_{m-1}}{b}H_{m-1}\right)\cdots\left(I-\frac{\eta_1}{b}H_1\right)e_1\right\|^s\right]. \tag{156}$$

By tower property and the fact that the distribution of $\|M_m M_{m-1}\cdots M_1 x\|/\|x\|$ is the same for every $x\in\mathbb{R}^d\backslash\{0\}$ and (η_i, H_i) are i.i.d., we have

$$\begin{aligned}
h^{(m)}(s) &= \mathbb{E}\left[\mathbb{E}\left[\left\|\left(I-\frac{\eta_m}{b}H_m\right)\left(I-\frac{\eta_{m-1}}{b}H_{m-1}\right)\cdots\left(I-\frac{\eta_1}{b}H_1\right)e_1\right\|^s\Big|\eta_1,H_1\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left\|\left(I-\frac{\eta_m}{b}H_m\right)\cdots\left(I-\frac{\eta_2}{b}H_2\right)e_1\right\|^s\Big|\eta_1,H_1\right]\left\|\left(I-\frac{\eta_1}{b}H_1\right)e_1\right\|^s\right] \\
&= \mathbb{E}\left[\left\|\left(I-\frac{\eta_m}{b}H_m\right)\cdots\left(I-\frac{\eta_2}{b}H_2\right)e_1\right\|^s\right]\mathbb{E}\left[\left\|\left(I-\frac{\eta_1}{b}H_1\right)e_1\right\|^s\right],
\end{aligned}$$

and therefore inductively we get

$$h^{(m)}(s)=\mathbb{E}\left[\left\|\left(I-\frac{\eta_m}{b}H_m\right)e_1\right\|^s\right]\mathbb{E}\left[\left\|\left(I-\frac{\eta_{m-1}}{b}H_{m-1}\right)e_1\right\|^s\right]\cdots\mathbb{E}\left[\left\|\left(I-\frac{\eta_1}{b}H_1\right)e_1\right\|^s\right].$$

Hence, we conclude that

$$\left(h^{(m)}(s)\right)^{1/m}=\tilde{h}^{(m)}(s), \tag{157}$$

where

$$\tilde{h}^{(m)}(s):=\left(\prod_{i=1}^m\mathbb{E}\left[\left\|\left(I-\frac{\eta_i}{b}H\right)e_1\right\|^s\right]\right)^{1/m}. \tag{158}$$

Similarly, we can derive that

$$\rho^{(m)} = \tilde{\rho}^{(m)}, \tag{159}$$

where

$$\tilde{\rho}^{(m)} := \sum_{i=1}^{m} \mathbb{E}\left[\log\left\|\left(I - \frac{\eta_i}{b}H\right)e_1\right\|\right]. \tag{160}$$

The proof is complete. □

Proof of Lemma 10

It follows from Lemma 19 in Gürbüzbalaban et al. (2021) that

$$\mathbb{E}\left[\left\|\left(I - \frac{\eta_i}{b}H\right)e_1\right\|^s\right] = \mathbb{E}\left[\left(\left(1 - \frac{\eta_i\sigma^2}{b}X\right)^2 + \frac{\eta_i^2\sigma^4}{b^2}XY\right)^{s/2}\right], \tag{161}$$

$$\mathbb{E}\left[\log\left\|\left(I - \frac{\eta_i}{b}H\right)e_1\right\|\right] = \frac{1}{2}\mathbb{E}\left[\log\left(\left(1 - \frac{\eta_i\sigma^2}{b}X\right)^2 + \frac{\eta_i^2\sigma^4}{b^2}XY\right)\right], \tag{162}$$

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 conclusion follows. □

Proof of Lemma 11

We follow the similar arguments as the proof of Theorem 3 in Gürbüzbalaban et al. (2021) and the key observation is that conditional on $(\eta_i)_{i=1}^{r_1}$ the distribution of $\left\|M_1^{(r)}x\right\|/\|x\|$ is the same for every $x \in \mathbb{R}^d\backslash\{0\}$, where r_1 is defined in (8). By tower property, we have

$$\begin{aligned} h^{(r)}(s) &= \mathbb{E}\left[\mathbb{E}\left[\left\|\left(I - \frac{\eta_{r_1}}{b}H_{r_1}\right)\left(I - \frac{\eta_{r_1-1}}{b}H_{r_1-1}\right)\cdots\left(I - \frac{\eta_1}{b}H_1\right)e_1\right\|^s \Big| (\eta_i)_{i=1}^{r_1}\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}_{H_{r_1}}\left[\left\|\left(I - \frac{\eta_{r_1}}{b}H_{r_1}\right)e_1\right\|^s\right]\mathbb{E}_{H_{r_1-1}}\left[\left\|\left(I - \frac{\eta_{r_1-1}}{b}H_{r_1-1}\right)e_1\right\|^s\right]\mathbb{E}_{H_1}\left[\left\|\left(I - \frac{\eta_1}{b}H_1\right)e_1\right\|^s\right]\right], \end{aligned}$$

and therefore inductively we conclude that

$$h^{(r)}(s) = \tilde{h}^{(r)}(s) := \mathbb{E}\left[\prod_{i=1}^{r_1}\mathbb{E}_H\left[\left\|\left(I - \frac{\eta_i}{b}H\right)e_1\right\|^s\right]\right]. \tag{163}$$

Similarly, we can derive that $\rho = \rho^{(r)}$, where

$$\rho^{(r)} := \mathbb{E}\left[\sum_{i=1}^{r_1}\mathbb{E}_H\left[\log\left\|\left(I - \frac{\eta_i}{b}H\right)e_1\right\|\right]\right]. \tag{164}$$

The proof is complete. □

Proof of Lemma 12

It follows from Lemma 19 in Gürbüzbalaban et al. (2021) that conditional on η_i,

$$\mathbb{E}_H\left[\left\|\left(I - \frac{\eta_i}{b}H\right)e_1\right\|^s\right] = \mathbb{E}_{X,Y}\left[\left(\left(1 - \frac{\eta_i\sigma^2}{b}X\right)^2 + \frac{\eta_i^2\sigma^4}{b^2}XY\right)^{s/2}\right], \tag{165}$$

$$\mathbb{E}_H\left[\log\left\|\left(I - \frac{\eta_i}{b}H\right)e_1\right\|\right] = \frac{1}{2}\mathbb{E}_{X,Y}\left[\log\left(\left(1 - \frac{\eta_i\sigma^2}{b}X\right)^2 + \frac{\eta_i^2\sigma^4}{b^2}XY\right)\right], \tag{166}$$

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 conclusion follows. □

E Supporting Lemmas

In this section, we provide a few supporting technical lemmas that are used in the proofs of the main results in the paper.

Lemma 13 (Lemma 22 in Gürbüzbalaban et al. (2021)). *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 in $a \geq 0$ for any $s \geq 1$.

Lemma 14 (Lemma 23 in Gürbüzbalaban et al. (2021)). *(i) Given $0 < p \leq 1$, for any $x, y \geq 0$,*

$$(x + y)^p \leq x^p + y^p. \tag{167}$$

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

$$(x + y)^p \leq (1 + \epsilon)x^p + \frac{(1 + \epsilon)^{\frac{p}{p-1}} - (1 + \epsilon)}{\left((1 + \epsilon)^{\frac{1}{p-1}} - 1\right)^p} y^p. \tag{168}$$

Lemma 15. *For any $a > 0$, and $k \in \mathbb{N}$,*

$$\sum_{i=1}^{k} i a^i = \frac{k a^{k+2} - (k + 1)a^{k+1} + a}{(a - 1)^2}.$$

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

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

Proof of Lemma 15

We can compute that

$$\sum_{i=1}^{k} i a^i = a \sum_{i=1}^{k} i a^{i-1} = a \frac{d}{da} \sum_{i=1}^{k} a^i = a \frac{d}{da} \frac{a^{k+1} - a}{a - 1} = \frac{k a^{k+2} - (k + 1)a^{k+1} + a}{(a - 1)^2}.$$

The proof is complete. □

F Additional Results

In this section, our purpose is to extend our analysis beyond linear regression, where we will assume that component functions $f_i(x) = f(x, z_i)$ arising in the empirical risk minimization problem (2) are twice continuously differentiable, and that $F(x)$ is bounded below so that a minimizer x_* of $F(x)$ exists. In this case, by Taylor's formula, we can write

$$\nabla f_i(x) = \big(\bar{H}_i(x_k)\big)(x_k - x_*) + \nabla f_i(x_*) \quad \text{where} \quad \bar{H}_i(x) := \int_{t=0}^{1} \nabla^2 f_i\left(x^* + t(x - x^*)\right) dt$$

is an averaged Hessian of the function f_i. We then introduce the following *stochastic estimate of the averaged Hessian of F*, defined analogously to the stochastic gradient, according to the formula

$$H_{k+1}(x_k) := \sum_{i \in \Omega_k} \bar{H}_i(x_k).$$

With this notation, SGD updates are equivalent to

$$x_{k+1} - x_* = \left(M_{k+1}(x_k)\right)(x_k - x_*) + \tilde{q}_{k+1}\,, \quad M_{k+1}(x_k) := I - \frac{\eta_{k+1}}{b} H_{k+1}(x_k), \tag{169}$$

with $\tilde{q}_k := \frac{-\eta_k}{b}\sum_{i\in\Omega_k}\nabla f_i(x_*)$, $\Omega_k := \{b(k-1)+1, b(k-1)+2, \dots, bk\}$ and $|\Omega_k| = b$. Here, the distribution of the stochastic Hessian estimate $H_{k+1}(x_k)$ depends on the iterate x_k; therefore the update (169) can be thought as a generalization of the update rule (5) that arises for linear regression (where the Hessian's distribution did not depend on x_k).

We first consider the case that the stepsizes are cyclic with a cycle length m, lying on a grid $(c_1, c_2, \dots, c_K)$. We consider the products

$$\overline{\sigma}^{(m)} := \prod_{j=1}^{m} \sup_{z\in\mathbb{R}^d} \|M_j(z)\|, \quad \underline{\sigma}^{(m)} := \prod_{j=1}^{m} \left(\liminf_{\|z\|\to\infty} \sigma_{\min}\big(M_j(z)\big) \right), \tag{170}$$

which are random quantities (as $M_{k+1}(z)$ is random when z is fixed, due to the randomness in the data) that roughly speaking measure the maximal and minimal growth of $M_{k+1}(x_k)$ in a cycle of length m where $\sigma_{\min}(\cdot)$ denotes the smallest singular value. The following result shows that the distributions can be heavy-tailed at stationarity with cyclic stepsizes provided that the minimal growth is large enough, i.e. if $\mathbb{P}(\underline{\sigma}^{(m)} > 1) > 0$.

Proposition 12. *Let batch-size b be given and fixed. Consider the SGD recursion with cyclic stepsize of period m when f_i are twice continuously differentiable and lower bounded for every $i = 1, 2, \dots, m$. Assume $\mathbb{E}(\log \overline{\sigma}^{(m)}) < 0$, $\mathbb{E}(\overline{\sigma}^{(m)}) < \infty$ and $\mathbb{P}(\underline{\sigma}^{(m)} > 1) > 0$ where $\underline{\sigma}^{(m)}$ and $\overline{\sigma}^{(m)}$ are defined according to (170). Then, there exists positive constants $\underline{\alpha}, \overline{\alpha}$ such that the tail-index α lies in the interval $[\underline{\alpha}, \overline{\alpha}]$, i.e. for every $\delta > 0$, $\limsup_{t\to\infty} t^{\underline{\alpha}+\delta}\mathbb{P}\left(\|x^{(\infty)}\| > t\right) > 0$, and[3] $\limsup_{t\to\infty} t^{\overline{\alpha}-\delta}\mathbb{P}\left(\|x_\infty\| > t\right) < \infty$ where x_∞ is the stationary distribution of the SGD recursion with cyclic stepsize of period m. Furthermore, we have $\mathbb{E}\big[(\overline{\sigma}^{(m)})^{\overline{\alpha}}\big] = 1$ and $\mathbb{E}\big[(\underline{\sigma}^{(m)})^{\underline{\alpha}}\big] = 1$.*

Proof. If we introduce $z_k = x_k - x_*$, then from (169),

$$z_{k+1} = \Phi_{k+1}(z_k) \quad \text{where} \quad \Phi_{k+1}(z_k) := \left(M_{k+1}(z_k + x_*)\right) z_k + \tilde{q}_{k+1}.$$

In particular, the map Φ_{k+1} admits a linear growth and Lipschitz behavior satisfying

$$\underline{s}_{k+1}\|z\| \le \|\Phi_{k+1}(z) - \Phi_{k+1}(0)\| = \|\left(M_{k+1}(z + x_*)\right) z\| \le \overline{s}_{k+1}\|z\|, \tag{171}$$

where the first inequality holds for $\|z\|$ large enough, whereas the second inequality holds for every z and

$$\underline{s}_{k+1} := \liminf_{\|z\|\to\infty} \sigma_{\min}\big(M_{k+1}(z)\big) \quad \text{and} \quad \overline{s}_{k+1} = \sup_{z\in\mathbb{R}^d} \|M_{k+1}(z)\|.$$

Then, we follow a similar approach to the proof of Theorem 5 and introduce

$$z_{(k+1)m} = \mathcal{F}_{k+1}(z_{km}) \quad \text{where} \quad \mathcal{F}_{k+1}(z_{km}) = \Phi_{(k+1)m} \circ \Phi_{(k+1)m-1} \circ \cdots \circ \Phi_{km+1}(z_{km})$$

is the composition of consecutive m iterations. Then, the composition $\mathcal{F}_{k+1}$ will also be Lipschitz satisfying

$$\underline{\sigma}^{(m)}\|z\| \le \|\mathcal{F}_{k+1}(z) - \mathcal{F}_{k+1}(0)\| \le \overline{\sigma}^{(m)}\|z\|,$$

for $\|z\|$ large enough, and the second inequality will be satisfied for every z. Or equivalently, there exists a non-negative random variable y_{k+1} (that depends on the sampled data points at steps km to $(k+1)m$) such that for every z we have

$$\underline{\sigma}^{(m)}\|z\| - y_{k+1} \le \|\mathcal{F}_{k+1}(z) - \mathcal{F}_{k+1}(0)\| \le \overline{\sigma}^{(m)}\|z\|.$$

Using this inequality, the result follows from (Hodgkinson & Mahoney, 2021, Thm. 1). □

[3] We use the convention that $\infty > 0$.

Remark 2. *Consider the smoothed Lasso loss with* $f_i(x) = \frac{1}{2}(a_i^T x - y_i)^2 + \lambda pen(x)$ *where the function* $x \mapsto pen(x)$ *is a smoothed version of the* ℓ_1 *loss and* $\lambda > 0$ *is the penalty parameter. We take* $pen(x) = \sqrt{\|x\|^2 + 1}$ *here, but many other versions are proposed in the literature (see e.g. Haselimashhadi (2019)). Then, by straightforward calculations it follows that the Hessian matrix* $\nabla^2 pen(x)$ *is uniformly bounded and satisfies* $-\frac{c_1}{R} I \preceq \nabla^2 pen(x) \preceq \frac{c_1}{R} I$ *for a positive constant* c_1 *whenever* $\|x\| \geq R$*. Under similar assumptions to* ***(A1)*** *and* ***(A2)*** *on the data, it can be checked that when the stepsizes* $(\eta_1, \eta_2, \ldots, \eta_m)$ *are small enough, the assumptions behind Propositions 12 and 13 will hold.*

Next, we assume as in (6) that the stepsizes follow a Markov chain with the finite state space

$$\{\eta_1, \eta_2, \ldots, \eta_m, \eta_{m+1}\} = \{c_1, c_2, \ldots, c_{K-1}, c_K, c_{K-1}, \ldots, c_2, c_1\}, \tag{172}$$

and let r_1 be the regeneration time such that $r_1 = \inf\{j > 0 : \eta_j = \eta_0\}$. Similar to (170), we define the products:

$$\overline{\sigma}^{(r)} := \prod_{j=1}^{r_1} \sup_{z \in \mathbb{R}^d} \|M_j(z)\|, \quad \underline{\sigma}^{(r)} := \prod_{j=1}^{r_1} \left(\liminf_{\|z\| \to \infty} \sigma_{\min}\big(M_j(z)\big) \right). \tag{173}$$

By using the similar argument as in the proof of Proposition 12, we have the following analogue of Proposition 12 for the Markovian stepsizes.

Proposition 13. *Let batch-size* b *be given and fixed. Consider the SGD recursion with Markovian stepsizes with finite state space (6) when* f_i *are twice continuously differentiable and lower bounded for every* $i = 1, 2, \ldots, m$*. Assume* $\mathbb{E}(\log \overline{\sigma}^{(r)}) < 0$*,* $\mathbb{E}(\overline{\sigma}^{(r)}) < \infty$ *and* $\mathbb{P}(\underline{\sigma}^{(r)} > 1) > 0$ *where* $\underline{\sigma}^{(r)}$ *and* $\overline{\sigma}^{(r)}$ *are defined according to (173). Then, there exists positive constants* $\underline{\alpha}, \overline{\alpha}$ *such that the tail-index* α *lies in the interval* $[\underline{\alpha}, \overline{\alpha}]$*, i.e. for every* $\delta > 0$*,* $\limsup_{t \to \infty} t^{\underline{\alpha} + \delta} \mathbb{P}\left(\|x^{(\infty)}\| > t\right) > 0$*, and* $\limsup_{t \to \infty} t^{\overline{\alpha} - \delta} \mathbb{P}\left(\|x_\infty\| > t\right) < \infty$ *where* x_∞ *is the stationary distribution of the SGD recursion with Markovian stepsizes. Furthermore, we have* $\mathbb{E}\big[(\overline{\sigma}^{(r)})^{\overline{\alpha}}\big] = 1$ *and* $\mathbb{E}\big[(\underline{\sigma}^{(r)})^{\underline{\alpha}}\big] = 1$*.*