APPENDIX
SGD: General Analysis and Improved Rates

A. Elementary Results

In this section we collect some elementary results; some of them we use repeatedly.

Proposition A.1. Let $\phi : \mathbb{R}^d \to \mathbb{R}$ be L_ϕ–smooth, and assume it has a minimizer x^* on $\mathbb{R}^d$. Then

$$\|\nabla \phi(x) - \nabla \phi(x^*)\|^2 \leq 2L_\phi(\phi(x) - \phi(x^*)).$$

Proof. Lipschitz continuity of the gradient implies that

$$\phi(x+h) \leq \phi(x) + \langle \nabla \phi(x), h \rangle + \frac{L_\phi}{2}\|h\|^2.$$

Now plugging $h = -\frac{1}{L_\phi}\nabla \phi(x)$ into the above inequality, we get $\frac{1}{2L_\phi}\|\nabla \phi(x)\|^2 \leq \phi(x) - \phi(x+h) \leq \phi(x) - \phi(x^*)$. It remains to note that $\nabla \phi(x^*) = 0$. □

In this section we summarize some elementary results which we use often in our proofs. We do not claim novelty; we but we include them for completeness and clarity.

Lemma A.2 (Double counting)**.** Let $a_{i,C} \in \mathbb{R}$ for $i = 1, \dots, n$ and $C \in \mathcal{C}$, where $\mathcal{C}$ is some collection of subsets of $[n]$. Then

$$\sum_{C \in \mathcal{C}} \sum_{i \in C} a_{i,C} \quad = \quad \sum_{i=1}^{n} \sum_{C \in \mathcal{C} \,:\, i \in C} a_{i,C}. \tag{44}$$

Lemma A.3 (Complexity bounds)**.** Let $E > 0$, $0 < \rho \leq 1$ and $0 \leq c < 1$. If $k \in \mathbb{N}$ satisfies

$$k \geq \frac{1}{1-\rho} \log\left(\frac{E}{(1-c)}\right), \tag{45}$$

then

$$\rho^k \leq (1-c)E. \tag{46}$$

Proof. Taking logarithms and rearranging (46) gives

$$\log\left(\frac{E}{1-c}\right) \leq k \log\left(\frac{1}{\rho}\right). \tag{47}$$

Now using that $\log\left(\frac{1}{\rho}\right) \geq 1-\rho$, for $0 < \rho \leq 1$ gives (45). □

A.1. The iteration complexity (12) of Theorem 3.1

To analyse the iteration complexity, let $\epsilon > 0$ and choosing the stepsize so that $\frac{2\gamma\sigma^2}{\mu} \leq \frac{1}{2}\epsilon$, gives (11). Next we choose k so that

$$(1-\gamma\mu)^k \|r^0\|^2 \leq \frac{1}{2}\epsilon.$$

Taking logarithms and re-arranging the above gives

$$\log\left(\frac{2\|r^0\|^2}{\epsilon}\right) \leq k \log\left(\frac{1}{1-\gamma\mu}\right). \tag{48}$$

Now using that $\log\left(\frac{1}{\rho}\right) \geq 1-\rho$, for $0 < \rho \leq 1$ gives

$$\begin{aligned} k &\geq \frac{1}{\gamma\mu}\log\left(\frac{2\|r^0\|^2}{\epsilon}\right) \\ &\overset{(11)}{=} \frac{1}{\mu}\max\left\{2\mathcal{L},\ \frac{4\sigma^2}{\epsilon\mu}\right\}\log\left(\frac{2\|r^0\|^2}{\epsilon}\right). \end{aligned} \tag{49}$$

Which concludes the proof.

B. Proof of Lemma 2.4

For brevity, let us write $\mathbb{E}[\cdot]$ instead of $\mathbb{E}_{\mathcal{D}}[\cdot]$. Then

$$\begin{aligned} \mathbb{E}\|\nabla f_v(x)\|^2 &= \mathbb{E}\|\nabla f_v(x) - \nabla f_v(x^*) + \nabla f_v(x^*)\|^2 \\ &\leq 2\mathbb{E}\|\nabla f_v(x) - \nabla f_v(x^*)\|^2 + 2\mathbb{E}\|\nabla f_v(x^*)\|^2 \\ &\leq 4\mathcal{L}[f(x) - f(x^*)] + 2\mathbb{E}\|\nabla f_v(x^*)\|^2. \end{aligned}$$

The first inequality follows from the estimate $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$, and the second inequality follows from (7).

C. Proof of Theorem 3.1

Proof. Let $r^k = x^k - x^*$. From (6), we have

$$\begin{aligned} \|r^{k+1}\|^2 &\overset{(6)}{=} \|x^k - x^* - \gamma\nabla f_{v^k}(x^k)\|^2 \\ &= \|r^k\|^2 - 2\gamma\langle r^k, \nabla f_{v^k}(x^k)\rangle + \gamma^2\|\nabla f_{v^k}(x^k)\|^2. \end{aligned}$$

Taking expectation conditioned on x^k we obtain:

$$\begin{aligned} \mathbb{E}_{\mathcal{D}}\|r^{k+1}\|^2 &\overset{(5)}{=} \|r^k\|^2 - 2\gamma\langle r^k, \nabla f(x^k)\rangle \\ &\quad + \gamma^2\mathbb{E}_{\mathcal{D}}\|\nabla f_{v^k}(x^k)\|^2 \\ &\overset{(2)}{\leq} (1-\gamma\mu)\|r^k\|^2 - 2\gamma[f(x^k) - f(x^*)] \\ &\quad + \gamma^2\mathbb{E}_{\mathcal{D}}\|\nabla f_{v^k}(x^k)\|^2. \end{aligned}$$

Taking expectations again and using Lemma 2.4:

$$\begin{aligned} \mathbb{E}\|r^{k+1}\|^2 &\overset{(9)}{\leq} (1-\gamma\mu)\mathbb{E}\|r^k\|^2 + 2\gamma^2\sigma^2 \\ &\quad + 2\gamma(2\gamma\mathcal{L} - 1)\mathbb{E}[f(x^k) - f(x^*)] \\ &\leq (1-\gamma\mu)\mathbb{E}\|r^k\|^2 + 2\gamma^2\sigma^2, \end{aligned}$$

where we used in the last inequality that $2\gamma\mathcal{L} \leq 1$ since $\gamma \leq \frac{1}{2\mathcal{L}}$. Recursively applying the above and summing up the resulting geometric series gives

$$\begin{aligned} \mathbb{E}\|r^k\|^2 &\leq (1-\gamma\mu)^k\|r^0\|^2 + 2\sum_{j=0}^{k-1}(1-\gamma\mu)^j\gamma^2\sigma^2 \\ &\leq (1-\gamma\mu)^k\|r^0\|^2 + \frac{2\gamma\sigma^2}{\mu}. \end{aligned} \tag{50}$$

To obtain an iteration complexity result from the above, we use standard techniques as shown in Section A.1. □

D. Proof of Theorem 3.2

Proof. Let $\gamma_k := \frac{2k+1}{(k+1)^2\mu}$ and let k^* be an integer that satisfies $\gamma_{k^*} \leq \frac{1}{2\mathcal{L}}$. In particular this holds for

$$k^* \geq \lceil 4\mathcal{K} - 1\rceil.$$

Note that γ_k is decreasing in k and consequently $\gamma_k \leq \frac{1}{2\mathcal{L}}$ for all $k \geq k^*$. This in turn guarantees that (50) holds for all $k \geq k^*$ with γ_k in place of γ, that is

$$\mathbb{E}\|r^{k+1}\|^2 \leq \frac{k^2}{(k+1)^2}\mathbb{E}\|r^k\|^2 + \frac{2\sigma^2}{\mu^2}\frac{(2k+1)^2}{(k+1)^4}. \tag{51}$$

Multiplying both sides by $(k+1)^2$ we obtain

$$\begin{aligned}(k+1)^2\mathbb{E}\|r^{k+1}\|^2 &\leq k^2\mathbb{E}\|r^k\|^2 + \frac{2\sigma^2}{\mu^2}\left(\frac{2k+1}{k+1}\right)^2 \\ &\leq k^2\mathbb{E}\|r^k\|^2 + \frac{8\sigma^2}{\mu^2},\end{aligned}$$

where the second inequality holds because $\frac{2k+1}{k+1} < 2$. Rearranging and summing from $t = k^* \dots k$ we obtain:

$$\sum_{t=k^*}^{k}\left[(t+1)^2\mathbb{E}\|r^{t+1}\|^2 - t^2\mathbb{E}\|r^t\|^2\right] \leq \sum_{t=k^*}^{k}\frac{8\sigma^2}{\mu^2}. \tag{52}$$

Using telescopic cancellation gives

$$(k+1)^2\mathbb{E}\|r^{k+1}\|^2 \leq (k^*)^2\mathbb{E}\|r^{k^*}\|^2 + \frac{8\sigma^2(k-k^*)}{\mu^2}.$$

Dividing the above by $(k+1)^2$ gives

$$\mathbb{E}\|r^{k+1}\|^2 \leq \frac{(k^*)^2}{(k+1)^2}\mathbb{E}\|r^{k^*}\|^2 + \frac{8\sigma^2(k-k^*)}{\mu^2(k+1)^2}. \tag{53}$$

For $k \leq k^*$ we have that (50) holds, which combined with (53), gives

$$\begin{aligned}\mathbb{E}\|r^{k+1}\|^2 &\leq \frac{(k^*)^2}{(k+1)^2}\left(1-\frac{\mu}{2\mathcal{L}}\right)^{k^*}\|r^0\|^2 \\ &+ \frac{\sigma^2}{\mu^2(k+1)^2}\left(8(k-k^*) + \frac{(k^*)^2}{\mathcal{K}}\right).\end{aligned} \tag{54}$$

Choosing k^* that minimizes the second line of the above gives $k^* = 4\lceil\mathcal{K}\rceil$, which when inserted into (54) becomes

$$\begin{aligned}\mathbb{E}\|r^{k+1}\|^2 &\leq \frac{16\lceil\mathcal{K}\rceil^2}{(k+1)^2}\left(1-\frac{1}{2\mathcal{K}}\right)^{4\lceil\mathcal{K}\rceil}\|r^0\|^2 \\ &\quad+\frac{\sigma^2}{\mu^2}\frac{8(k-2\lceil\mathcal{K}\rceil)}{(k+1)^2} \\ &\leq \frac{16\lceil\mathcal{K}\rceil^2}{e^2(k+1)^2}\|r^0\|^2 + \frac{\sigma^2}{\mu^2}\frac{8}{k+1},\end{aligned} \tag{55}$$

where we have used that $\left(1-\frac{1}{2x}\right)^{4x} \leq e^{-2}$ for all $x \geq 1$.

$\square$

E. Proof of Theorem 3.6

Proof. Since $v_i = v_i(S) = \mathbf{1}_{(i\in S)}\frac{1}{p_i}$. and since f_i is $\mathbf{M}_i$-smooth, the function

$$f_v(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)v_i = \frac{1}{n}\sum_{i\in S}\frac{f_i(x)}{p_i}, \tag{56}$$

is L_S–smooth where

$$L_S := \frac{1}{n}\lambda_{\max}\left(\sum_{i\in S}\frac{\mathbf{M}_i}{p_i}\right).$$

We also define the following smoothness related quantities

$$\mathcal{L}_i := \sum_{C\,:\,i\in C}\frac{p_C}{p_i}L_C, \quad \mathcal{L}_{\max} := \max_i \mathcal{L}_i, \text{ and;} \quad L_{\max} = \max_{i\in[n]}\lambda_{\max}(\mathbf{M}_i). \tag{57}$$

Since the f_i's are convex and the sampling vector $v\in\mathbb{R}^d_+$ has positive elements, each realization of f_v is convex and smooth, thus it follows from equation (2.1.7) in Theorem 2.1.5 in (Nesterov, 2013) that

$$\|\nabla f_v(x)-\nabla f_v(y)\|^2 \le 2L_S\left(f_v(x)-f_v(y)-\langle\nabla f_v(y), x-y\rangle\right). \tag{58}$$

Taking expectation in (58) gives

$$\begin{aligned}
\mathbb{E}[\|\nabla f_v(x)-\nabla f_v(y)\|^2] &\le 2\sum_C p_C L_C\left(f_{v(C)}(x)-f_{v(C)}(y)-\langle\nabla f_{v(C)}(y),x-y\rangle\right)\\
&\overset{(56)}{=} 2\sum_C p_C L_C\sum_{i\in C}\frac{1}{np_i}\left(f_i(x)-f_i(y)-\langle\nabla f_i(y),x-y\rangle\right)\\
&\overset{\text{Lemma } A.2}{=} \frac{2}{n}\sum_{i=1}^n\sum_{C:i\in C}p_C\frac{1}{p_i}L_C\left(f_i(x)-f_i(y)-\langle\nabla f_i(y),x-y\rangle\right)\\
&\overset{(18)}{\le} \frac{2}{n}\sum_{i=1}^n\mathcal{L}_{\max}\left(f_i(x)-f_i(y)-\langle\nabla f_i(y),x-y\rangle\right)\\
&= 2\mathcal{L}_{\max}\left(f(x)-f(y)-\langle\nabla f(y),x-y\rangle\right).
\end{aligned}$$

Furthermore, for each i,

$$\begin{aligned}
\mathcal{L}_i = \sum_{C:i\in C}\frac{p_C}{p_i}L_C &= \frac{1}{n}\sum_{C:i\in C}\frac{p_C}{p_i}\lambda_{\max}\left(\sum_{j\in C}\frac{\mathbf{M}_j}{p_j}\right)\\
&\le \frac{1}{n}\sum_{C:i\in C}\frac{p_C}{p_i}\sum_{j\in C}\frac{\lambda_{\max}(\mathbf{M}_j)}{p_j}\\
&\overset{\text{Lemma } A.2}{=} \frac{1}{n}\sum_{j=1}^n\sum_{C:i\in C\ \&\ j\in C}\frac{p_C}{p_ip_j}\lambda_{\max}(\mathbf{M}_j)\\
&= \frac{1}{n}\sum_{j=1}^n\frac{\mathbf{P}_{ij}}{p_ip_j}\lambda_{\max}(\mathbf{M}_j).
\end{aligned} \tag{59}$$

Hence,

$$\mathcal{L}_{\max} \le \frac{1}{n}\max_{i\in[n]}\left\{\sum_{j\in[n]}\mathbf{P}_{ij}\frac{\lambda_{\max}(\mathbf{M}_j)}{p_ip_j}\right\}. \tag{60}$$

Let $y=x^*$ and notice that $\nabla f(x^*)=0$, which gives (18). We prove (19) in the following slightly more comprehensive Lemma F.1. □

F. Bounds on the Expected Smoothness Constant $\mathcal{L}$

Below we establish some lower and upper bounds on the expected smoothness constant $\mathcal{L}=\mathcal{L}_{\max}$. These bounds were referred to in the main paper in Section 2.3. We also make use of notation introduced in Section 3.3.

Lemma F.1. Assume that there exists $\tau \in [n]$ such that $|S| = \tau$ with probability 1. Let

$$\mathcal{L}_i := \mathbb{E}\left[L_S \mid i \in S\right] = \sum_{C\,:\,i\in C} \frac{p_C}{p_i} L_C,$$

and

$$\bar{\mathcal{L}}_S := \frac{1}{|S|}\sum_{i\in S}\mathcal{L}_i.$$

Then $\mathbb{E}\left[\bar{\mathcal{L}}_S\right] = \mathbb{E}\left[L_S\right]$. Moreover,

$$L \le \mathbb{E}\left[\bar{\mathcal{L}}_S\right] \le \mathcal{L}_{\max} \le L_{\max}. \tag{61}$$

Proof. Define $\mathbf{M}_S := \frac{1}{n}\sum_{i\in S}\frac{\mathbf{M}_i}{p_i}$ and note that f is $\frac{1}{n}\sum_{i\in[n]}\mathbf{M}_i$–smooth. Furthermore

$$\mathbb{E}\left[\mathbf{M}_S\right] = \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^n \frac{\mathbf{M}_i}{p_i}\mathbf{1}_{(i\in S)}\right] = \frac{1}{n}\sum_{i=1}^n\frac{\mathbf{M}_i}{p_i}\mathbb{E}\left[\mathbf{1}_{(i\in S)}\right] = \frac{1}{n}\sum_{i\in[n]}\mathbf{M}_i.$$

We will now establish the inequalities in (61) starting from left to the right.

(Part I $L \le \mathbb{E}\left[L_S\right]$). Recalling that $L_S = \lambda_{\max}(\mathbf{M}_S)$ and by Jensen's inequality,

$$L = \lambda_{\max}\left(\mathbb{E}\left[\mathbf{M}_S\right]\right) \le \mathbb{E}\left[\lambda_{\max}(\mathbf{M}_S)\right] = \mathbb{E}\left[L_S\right].$$

Furthermore

$$\begin{aligned}
\mathbb{E}\left[\bar{\mathcal{L}}_S\right] &= \mathbb{E}\left[\frac{1}{\tau}\sum_{i\in S}\mathcal{L}_i\right] = \frac{1}{\tau}\sum_i p_i\mathcal{L}_i \\
&\overset{(57)}{=} \frac{1}{\tau}\sum_i\sum_{C\,:\,i\in C} p_C L_i \overset{\text{Lemma } A.2}{=} \frac{1}{\tau}\sum_C\sum_{i\in C} p_C L_C \\
&= \frac{1}{\tau}\sum_C |C| p_C L_C = \sum_C p_C L_C = \mathbb{E}\left[L_S\right]
\end{aligned}$$

(Part II $\mathbb{E}\left[\bar{L}_S\right] \le \mathcal{L}_{\max}$). We have that

$$\bar{L}_S = \frac{1}{|S|}\sum_{i\in S}\mathcal{L}_i \le \frac{1}{|S|}\sum_{i\in S}\max_{i\in[n]}\mathcal{L}_i = \mathcal{L}_{\max}.$$

(Part III $\mathcal{L}_{\max} \le L_{\max}$). Finally, since

$$L_C \le \frac{1}{\tau}\sum_{j\in C} L_j \le L_{\max}, \tag{62}$$

we have that

$$\mathcal{L}_i \overset{(57)+(62)}{\le} \sum_{C\,:\,i\in C}\frac{p_C}{p_i}\frac{1}{\tau}\sum_{j\in C} L_j \overset{(62)}{\le} \sum_{C\,:\,i\in C}\frac{p_C}{p_i}L_{\max} = L_{\max}.$$

Consequently taking the maximum over $i \in [n]$ in the above gives $\mathcal{L}_{\max} \le L_{\max}$. □

G. Proof of Proposition 3.7

Proof. First note that by combining (18) and (59) we have that

$$\begin{aligned}
\mathcal{L}_{\max} &\overset{(18)}{=} \max_{i\in[n]}\left\{\sum_{C:i\in C}\frac{p_C}{p_i}L_C\right\} \\
&\overset{(59)}{=} \max_{i\in[n]}\left\{\frac{1}{n}\sum_{C:i\in C}\frac{p_C}{p_i}\lambda_{\max}\left(\sum_{j\in C}\frac{\mathbf{M}_j}{p_j}\right)\right\}.
\end{aligned} \tag{63}$$

(i) By straight forward calculation from (63) and using that each set C is a singleton.

(ii) For every partition sampling we have that $p_i = p_C$ if $i \in C$, hence

$$\begin{aligned}
\mathcal{L}_{\max} & \overset{(63)}{=} \max_{i\in[n]} \left\{ \frac{1}{n} \sum_{C:i\in C} \frac{p_i}{p_i} \lambda_{\max}\left(\sum_{j\in C} \frac{\mathbf{M}_j}{p_C} \right) \right\} \\
& \overset{(59)}{=} \frac{1}{n} \max_{i\in[n]} \left\{ \sum_{C:i\in C} \frac{1}{p_C} \lambda_{\max}(\sum_{j\in C} \mathbf{M}_j) \right\} \\
& = \frac{1}{n} \max_{C\in\mathcal{G}} \left\{ \frac{1}{p_C} \lambda_{\max}(\sum_{j\in C} \mathbf{M}_j) \right\}.
\end{aligned}$$

□

H. Proof of Proposition 3.8

Proof. First, since f_i is L_i-smooth with $L_i = \lambda_{\max}(\mathbf{M}_i)$ and convex, it follows from equation (2.1.7) in Theorem 2.1.5 in (Nesterov, 2013) that

$$\|\nabla f_i(x) - \nabla f_i(y)\|^2 \leq 2L_i(f_i(x) - f_i(y) - \langle \nabla f_i(y), x - y\rangle). \tag{64}$$

Since f is L-smooth, we have

$$\|\nabla f(x) - \nabla f(y)\|^2 \leq 2L(f(x) - f(y) - \langle \nabla f(y), x - y\rangle). \tag{65}$$

Noticing that

$$\begin{aligned}
\|\nabla f_v(x) - \nabla f_v(y)\|^2 & = \frac{1}{n^2} \left\| \sum_{i\in S} \frac{1}{p_i} (\nabla f_i(x) - \nabla f_i(y)) \right\|^2 \\
& = \sum_{i,j\in S} \left\langle \frac{1}{np_i}(\nabla f_i(x) - \nabla f_i(y)), \frac{1}{np_j}(\nabla f_j(x) - \nabla f_j(y)) \right\rangle,
\end{aligned}$$

we have

$$\begin{aligned}
\mathbb{E}[\|\nabla f_v(x) - \nabla f_v(y)\|^2] & = \sum_C p_C \sum_{i,j\in C} \left\langle \frac{1}{np_i}(\nabla f_i(x) - \nabla f_i(y)), \frac{1}{np_j}(\nabla f_j(x) - \nabla f_j(y)) \right\rangle \\
& = \sum_{i,j=1}^n \sum_{C:i,j\in C} p_C \left\langle \frac{1}{np_i}(\nabla f_i(x) - \nabla f_i(y)), \frac{1}{np_j}(\nabla f_j(x) - \nabla f_j(y)) \right\rangle \\
& = \sum_{i,j=1}^n \frac{\mathbf{P}_{ij}}{p_i p_j} \left\langle \frac{1}{n}(\nabla f_i(x) - \nabla f_i(y)), \frac{1}{n}(\nabla f_j(x) - \nabla f_j(y)) \right\rangle.
\end{aligned}$$

Now consider the case where $\mathbf{P}_{ij}/(p_i p_j) = c_2$ for $i \neq j$. Recalling that $\mathbf{P}_{ii} = p_i$ we have from the above that

$$
\begin{aligned}
\mathbb{E}[\|\nabla f_v(x)-\nabla f_v(y)\|^2] &= \sum_{i\neq j} c_2 \left\langle \frac{1}{n}(\nabla f_i(x)-\nabla f_i(y)), \frac{1}{n}(\nabla f_j(x)-\nabla f_j(y)) \right\rangle + \sum_{i=1}^n \frac{1}{n^2}\frac{1}{p_i}\|\nabla f_i(x)-\nabla f_i(y))\|_2^2 \\
&= \sum_{i,j=1}^n c_2 \left\langle \frac{1}{n}(\nabla f_i(x)-\nabla f_i(y)), \frac{1}{n}(\nabla f_j(x)-\nabla f_j(y)) \right\rangle \\
&\quad + \sum_{i=1}^n \frac{1}{n^2}\frac{1}{p_i}(1-p_i c_2)\|\nabla f_i(x)-\nabla f_i(y))\|_2^2 \\
&\overset{(64)}{\leq} c_2 \|\nabla f(x)-\nabla f(y)\|_2^2 \\
&\quad +2\sum_{i=1}^n \frac{1}{n^2}\frac{L_i}{p_i}(1-p_i c_2)(f_i(x)-f_i(y)-\langle \nabla f_i(y), x-y\rangle) \\
&\overset{(65)}{\leq} 2\left(c_2 L + \max_{i=1,\ldots,n}\frac{L_i}{np_i}(1-p_i c_2)\right)(f(x)-f(y)-\langle \nabla f(y), x-y\rangle).
\end{aligned}
$$

Substituting $y = x^*$ and comparing the above to the definition of expected smoothness (7) we have that

$$
\mathcal{L} \quad \leq \quad c_2 L + \max_{i=1,\ldots,n}\frac{L_i}{np_i}(1-p_i c_2). \tag{66}
$$

(i) For independent sampling, we have that $\mathbf{P}_{ij} = p_i p_j$ for $i \neq j$, consequently $c_2 = 1$. Thus (66) gives (22).

(ii) For τ-nice sampling, we have that $\mathbf{P}_{ij} = \frac{\tau(\tau-1)}{n(n-1)}$ for $j \neq i$ and $\mathbf{P}_{ii} = p_i = \frac{\tau}{n}$, hence $c_2 = \frac{n(\tau-1)}{\tau(n-1)}$ and (66) gives (23). □

I. Proof of Theorem 3.9

Proof.

$$
\begin{aligned}
\sigma^2 = \mathbb{E}[\|\nabla f_v(x^*)\|^2] &= \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \nabla f_i(x^*)v_i\right\|^2\right] = \frac{1}{n^2}\mathbb{E}\left[\left\|\sum_{i=1}^n \nabla f_i(x^*)v_i\right\|^2\right] = \frac{1}{n^2}\mathbb{E}\left[\left\|\sum_{i\in S}\frac{1}{p_i}h_i\right\|^2\right] \\
&= \frac{1}{n^2}\mathbb{E}\left[\left\|\sum_{i=1}^n 1_{i\in S}\frac{1}{p_i}h_i\right\|^2\right] = \frac{1}{n^2}\mathbb{E}\left[\sum_{i=1}^n\sum_{j=1}^n 1_{i\in S}1_{j\in S}\langle \frac{1}{p_i}h_i, \frac{1}{p_j}h_j\rangle\right] \\
&= \frac{1}{n^2}\sum_{i,j}\frac{\mathbf{P}_{ij}}{p_i p_j}\langle h_i, h_j\rangle.
\end{aligned}
$$

□

J. Proof of Proposition 3.10

Proof. (i) By straight calculation from (24).

(ii) For independent sampling S, $\mathbf{P}_{ij} = p_i p_j$ for $i \neq j$, hence,

$$
\begin{aligned}
\sigma^2 &= \frac{1}{n^2}\sum_{i,j\in[n]}\frac{\mathbf{P}_{ij}}{p_i p_j}\langle h_i, h_j\rangle = \frac{1}{n^2}\sum_{i,j\in[n]}\langle h_i, h_j\rangle + \frac{1}{n^2}\sum_{i\in[n]}\left(\frac{1}{p_i}-1\right)\|h_i\|^2 \\
&= \frac{1}{n^2}\|\nabla f(x^*)\|^2 + \frac{1}{n^2}\sum_{i\in[n]}\left(\frac{1}{p_i}-1\right)\|h_i\|^2 = \frac{1}{n^2}\sum_{i\in[n]}\left(\frac{1}{p_i}-1\right)\|h_i\|^2.
\end{aligned}
$$

(iii) For τ-nice sampling S, if $\tau = 1$, it is obvious. If $\tau \geq 1$, then $\mathbf{P}_{ij} = \frac{C_{n-2}^{\tau-2}}{C_n^\tau}$ for $i \neq j$, and $p_i = \frac{\tau}{n}$ for all i. Hence,

$$
\begin{aligned}
\sigma^2 &= \frac{1}{n^2}\sum_{i,j\in[n]} \frac{\mathbf{P}_{ij}}{p_i p_j}\langle h_i, h_j\rangle \\
&= \frac{1}{n^2}\sum_{i\neq j} \frac{\tau(\tau-1)}{n(n-1)}\cdot\frac{n^2}{\tau^2}\langle h_i, h_j\rangle + \frac{1}{n^2}\sum_{i\in[n]}\frac{n}{\tau}\|h_i\|^2 \\
&= \frac{1}{n\tau}\left(\sum_{i\neq j}\frac{\tau-1}{n-1}\langle h_i, h_j\rangle + \sum_{i\in[n]}\|h_i\|^2\right) \\
&= \frac{1}{n\tau}\left(\sum_{i,j\in[n]}\frac{\tau-1}{n-1}\langle h_i, h_j\rangle + \sum_{i\in[n]}\frac{n-\tau}{n-1}\|h_i\|^2\right) \\
&= \frac{1}{n\tau}\cdot\frac{n-\tau}{n-1}\sum_{i\in[n]}\|h_i\|^2.
\end{aligned}
$$

(iv) For partition sampling, $\mathbf{P}_{ij} = p_C$ if $i, j \in C$, and $\mathbf{P}_{ij} = 0$ otherwise. Hence,

$$
\sigma^2 = \frac{1}{n^2}\sum_{i,j\in[n]}\frac{\mathbf{P}_{ij}}{p_i p_j}\langle h_i, h_j\rangle = \frac{1}{n^2}\sum_{C\in\mathcal{G}}\sum_{i,j\in C}\frac{1}{p_C}\langle h_i, h_j\rangle = \frac{1}{n^2}\sum_{C\in\mathcal{G}}\frac{1}{p_C}\|\sum_{i\in C} h_i\|^2.
$$

□

K. Importance sampling

K.1. Single element sampling

From (20) it is easy to see that the probabilities that minimize $\mathcal{L}_{\max}$ are $p_i^{\mathcal{L}} = L_i/\sum_{j\in[n]} L_j$, for all i, and consequently $\mathcal{L}_{\max} = \overline{L}$. On the other hand the probabilities that minimize (25) are given by $p_i^{\sigma^2} = \|h_i\|/\sum_{j\in[n]}\|h_j\|$, for all i, with $\sigma^2 = (\sum_{i\in[n]}\|h_i\|/n)^2 := \sigma^2_{opt}$.

Importance sampling. From $p_i^{\mathcal{L}}$ and $p_i^{\sigma^2}$, we construct interpolated probabilities p_i as follows:

$$
p_i = p_i(\alpha) = \alpha p_i^{\mathcal{L}} + (1-\alpha)p_i^{\sigma^2}, \tag{67}
$$

where $\alpha \in (0, 1)$. Then $0 < p_i < 1$ and from (20) we have

$$
\mathcal{L}_{\max} \leq \frac{1}{\alpha}\cdot\frac{1}{n}\max_{i\in[n]}\frac{L_i}{p_i^{\mathcal{L}}(\tau)} = \frac{1}{\alpha}\overline{L}.
$$

Similarly, from (25) we have that $\sigma^2 \leq \frac{1}{1-\alpha}\sigma^2_{opt}$. Now by letting $p_i = p_i(\alpha)$, from (29) in Theorem 3.1, we get an upper bound of the right hand side of (12):

$$
\max\left\{\frac{2\overline{L}}{\alpha\mu}, \frac{4\sigma^2_{opt}}{(1-\alpha)\epsilon\mu^2}\right\}. \tag{68}
$$

By minimizing this bound in α we can get

$$
\alpha = \frac{\overline{L}}{2\sigma^2_{opt}/\epsilon\mu + \overline{L}}, \tag{69}
$$

and then the upper bound (68) becomes

$$
\frac{4\sigma^2_{opt}}{\epsilon\mu^2} + \frac{2\overline{L}}{\mu} \leq 2\max\left\{\frac{2\overline{L}}{\mu}, \frac{4\sigma^2_{opt}}{\epsilon\mu^2}\right\}, \tag{70}
$$

where the right hand side comes by setting $\alpha = 1/2$. Notice that the minimum of the iteration complexity in (12) is not less than $\max\left\{\frac{2\overline{L}}{\mu}, \frac{4\sigma_{opt}^2}{\epsilon\mu^2}\right\}$. Hence, the iteration complexity of this importance sampling(left hand side of (70)) is at most two times larger than the minimum of the iteration complexity in (12) over p_i.

K.2. Independent sampling

For the independent sampling S, in this section we will use the following upper bound on $\mathcal{L}$ given by

$$\mathcal{L}_{\max} \leq L + \max_{i\in[n]} \tfrac{1-p_i}{p_i}\tfrac{L_i}{n}, \tag{71}$$

from (22). Denote $\overline{L} := \frac{1}{n}\sum_{i=1}^n L_i$.

Calculating $p_i^{\mathcal{L}}(\tau)$. Minimizing the upper bound of $\mathcal{L}_{\max}$ in (71) boils down to minimizing $\max_{i\in[n]}(\frac{1}{p_i}-1)L_i$, which is not easy generally. Instead, as a proxy we obtain the probabilities p_i by solving

$$\begin{array}{ll} \min & \max_{i\in[n]} \frac{L_i}{p_i} \\ \text{s.t.} & \sum_{i\in[n]} p_i = \tau,\ 0 < p_i \leq 1, \forall i. \end{array} \tag{72}$$

Let $q_i = \frac{L_i}{\sum_{j\in[n]} L_j}\cdot\tau$ for all i, and $T = \{i | q_i > 1\}$. If $T = \emptyset$, it is easy to see $p_i = p_i^{\mathcal{L}}(\tau) = q_i$ solves (72). Otherwise, in order to solve (72), we can choose $p_i = p_i^{\mathcal{L}}(\tau) = 1$ for $i \in T$, and $q_i \leq p_i = p_i^{\mathcal{L}}(\tau) \leq 1$ for $i \notin T$ such that $\sum_{i\in[n]} p_i^{\mathcal{L}}(\tau) = \tau$. By letting $p_i = p_i^{\mathcal{L}}(\tau)$, and noticing that $(\frac{1}{p_i}-1)L_i = 0$ for $p_i = 1$, we have that

$$\mathcal{L}_{\max} \leq L + \frac{1}{n}\cdot\frac{1}{\tau}\sum_{j\in[n]} L_j = L + \frac{1}{\tau}\overline{L}. \tag{73}$$

Calculating $p_i^{\sigma^2}(\tau)$. For σ^2, from (26), we need to solve

$$\begin{array}{ll} \min & \sum_{i\in[n]} \frac{\|h_i\|^2}{p_i} \\ \text{s.t.} & \sum_{i\in[n]} p_i = \tau,\ 0 < p_i \leq 1, \forall i. \end{array} \tag{74}$$

Let $q_i = \frac{\|h_i\|}{\sum_{j\in[n]}\|h_j\|}\cdot\tau$ for all i, and let $T = \{i | q_i > 1\}$. If $T = \emptyset$, it is easy to see that $p_i = p_i^{\sigma^2}(\tau) = q_i$ solve (74). Otherwise, it is a little complicated to find the optimal solution. For simplicity, if $T \neq \emptyset$, we choose $p_i = p_i^{\sigma^2}(\tau) = 1$ for $i \in T$, and $q_i \leq p_i = p_i^{\sigma^2}(\tau) \leq 1$ for $i \notin T$ such that $\sum_{i\in[n]} p_i^{\sigma^2}(\tau) = \tau$. By letting $p_i = p_i^{\sigma^2}(\tau)$, from (26), we have

$$\begin{aligned} \sigma^2 &\leq \frac{1}{n^2}\sum_{i\notin T}\left(\frac{\|h_i\|\sum_{j\in[n]}\|h_j\|}{\tau} - \|h_i\|^2\right) \\ &\leq \frac{1}{\tau}\left(\frac{\sum_{i\in[n]}\|h_i\|}{n}\right)^2 := \sigma_{opt}^2(\tau). \end{aligned}$$

Importance sampling. Since by (73) we have that $\mathcal{L}_{\max} \leq L + \frac{1}{\tau}\overline{L}$ and $\sigma^2 = \sigma_{opt}^2(\tau)$ are obtained by using the upper bounds in (71) and (26), and the upper bounds are nonincreasing as p_i increases, we get the following property.

Proposition K.1. If $p_i \geq p_i^{\mathcal{L}}(\tau)$ for all i, then $\mathcal{L}_{\max} \leq L + \frac{1}{\tau}\overline{L}$, and if $p_i \geq p_i^{\sigma^2}(\tau)$, then $\sigma^2 \leq \sigma_{opt}^2(\tau)$.

From Proposition K.1, we can get the following result.

Proposition K.2. For $0 < \alpha < 1$, let $p_i(\alpha)$ satisfy

$$\left\{\begin{array}{l} 1 \geq p_i(\alpha) \geq \min\{1, p_i^{\mathcal{L}}(\alpha\tau) + p_i^{\sigma^2}((1-\alpha)\tau)\}, \quad \forall i, \\ \sum_{i\in[n]} p_i(\alpha) = \tau. \end{array}\right. \tag{75}$$

If $p_i = p_i(\alpha)$ where $p_i(\alpha)$ satisfies (75), then we have

$$\mathcal{L}_{\max} \leq L + \frac{1}{\alpha\tau}\overline{L},$$

and

$$\sigma^2 \leq \sigma^2_{opt}((1-\alpha)\tau) = \frac{1}{(1-\alpha)\tau}(\frac{\sum_{i\in[n]}\|h_i\|}{n})^2.$$

Proof. First , we claim that $p_i(\alpha)$ can be constructed to satisfy (75). Since $0 < p_i^{\mathcal{L}}(\alpha) \leq 1$ and $0 < p_i^{\sigma^2}((1-\alpha)\tau) \leq 1$, we know

$$0 < \min\{1, p_i^{\mathcal{L}}(\alpha\tau) + p_i^{\sigma^2}((1-\alpha)\tau)\} \leq 1,$$

for all i. Hence, we can first construct $\tilde{q}_i$ such that

$$1 \geq \tilde{q}_i \geq \min\{1, p_i^{\mathcal{L}}(\alpha\tau) + p_i^{\sigma^2}((1-\alpha)\tau)\},$$

for all i. Furthermore, since $\sum_{i\in[n]} p_i^{\mathcal{L}}(\alpha\tau) = \alpha\tau$ and $\sum_{i\in[n]} p_i^{\sigma^2}((1-\alpha)\tau) = (1-\alpha)\tau$, we know $\sum_{i\in[n]}\tilde{q}_i \leq \tau$. At last, we increase some $\tilde{q}_i$ which is less than one to make the sum equal to τ, and hence, by letting $p_i(\alpha) = \tilde{q}_i$, $p_i(\alpha)$ satisfies (75).

From (75), we have $p_i = p_i(\alpha) \geq p_i^{\mathcal{L}}(\alpha\tau)$. Then by Proposition K.1, we have

$$\mathcal{L}_{\max} \leq L + \frac{1}{\alpha\tau}\overline{L}.$$

We also have $p_i(\alpha) \geq p_i^{\sigma^2}((1-\alpha)\tau)$, hence, by Proposition K.1, we get

$$\sigma^2 \leq \sigma^2_{opt}((1-\alpha)\tau) = \frac{1}{(1-\alpha)\tau}\left(\frac{\sum_{i\in[n]}\|h_i\|}{n}\right)^2.$$

□

From (12) in Theorem 3.1, by letting $p_i = p_i(\alpha)$ in Proposition K.2, we get an upper bound of the right hand side of (12):

$$\max\left\{\frac{2(L+\frac{1}{\alpha\tau}\overline{L})}{\mu}, \frac{4\sigma^2_{opt}((1-\alpha)\tau)}{\epsilon\mu^2}\right\}.$$

By minimizing this upper bound, we get

$$\alpha = \frac{\tau - a - \overline{L}/L + \sqrt{4\tau\overline{L}/L + (\tau - a - \overline{L}/L)^2}}{2\tau}, \tag{76}$$

and the upper bound becomes

$$\frac{2(L+\frac{1}{\alpha\tau}\overline{L})}{\mu}$$

where $a = 2(\frac{\sum_{i\in[n]}\|h_i\|}{n})^2/(\epsilon\mu L)$. So suboptimal probabilities

$$p_i = \min\{1, p_i^{\mathcal{L}}(\alpha\tau) + p_i^{\sigma^2}((1-\alpha)\tau)\}, \tag{77}$$

where α is given in Equation (76).

Partially biased sampling. In practice, we do not know $\|h_i\|$ generally. But we can use $p_i^{\mathcal{L}}(\tau)$ and the uniform probability $\frac{\tau}{n}$ to construct a new probability just as that in Proposition K.2. More specific, we have the following result.

Table 1. Comparison of the upper bounds of $\mathcal{L}_{\max}$ and σ^2 for τ-nice sampling, τ-partially biased independent sampling, and τ-uniform independent sampling.

	$\mathcal{L}_{\max}$	σ^2
τ-NICE SAMPLING	$\frac{n}{\tau}\cdot\frac{\tau-1}{n-1}L+\frac{1}{\tau}(1-\frac{\tau-1}{n-1})L_{\max}$	$\frac{1}{\tau}\cdot\frac{n-\tau}{n-1}\bar{h}$
τ-UNIFORM IS	$L+(\frac{1}{\tau}-\frac{1}{n})L_{\max}$	$(\frac{1}{\tau}-\frac{1}{n})\bar{h}$
τ-PBA-IS	$L+\frac{2}{\tau}\overline{L}$	$(\frac{2}{\tau}-\frac{1}{n})\bar{h}$

Proposition K.3. Let p_i satisfy

$$\begin{cases} 1 \geq p_i \geq \min\{1, p_i^{\mathcal{L}}(\frac{\tau}{2}) + \frac{1}{2}\cdot\frac{\tau}{n}\}, \quad \forall i, \\ \sum_{i\in[n]} p_i = \tau. \end{cases} \tag{78}$$

Then we have

$$\mathcal{L}_{\max} \leq \left(L + \frac{2}{\tau}\overline{L}\right),$$

and

$$\sigma^2 \leq \left(\frac{2}{\tau} - \frac{1}{n}\right)\cdot\frac{1}{n}\sum_{i\in[n]}\|h_i\|^2.$$

Proof. The proof for $\mathcal{L}_{\max}$ is the same as Proposition K.2. For σ^2, from (26), since $p_i \geq \tau/2n$, we have

$$\sigma^2 = \frac{1}{n^2}\sum_{i\in[n]}\left(\frac{1}{p_i}-1\right)\|h_i\|^2 \leq \frac{1}{n^2}\sum_{i\in[n]}\left(\frac{2n}{\tau}-1\right)\|h_i\|^2 = \left(\frac{2}{\tau}-\frac{1}{n}\right)\cdot\frac{1}{n}\sum_{i\in[n]}\|h_i\|^2.$$

□

This sampling is very nice in the sense that it can maintain $\mathcal{L}_{\max}$ has nearly linear speed up when $\tau \leq 2\overline{L}/L$, and meanwhile, can acheive nearly linear speedup in σ^2 by increasing τ. We can compare the upper bounds of $\mathcal{L}_{\max}$ and σ^2 for this sampling, τ-nice sampling, and τ-uniform independent sampling when $1 < \tau = \mathcal{O}(1)$ in the following table.

From Table 1, compared to τ-nice sampling and τ-uniform independent sampling, the iteration complexity of this τ-partially biased independent sampling is at most two times larger, but could be about $\frac{2\tau}{n}$ smaller in some extremely case where $L_{\max} \approx n\bar{L}$ and $2\mathcal{L}/\mu$ dominates in (12).

L. Additional Experiments

L.1. From fixed to decreasing stepsizes: analysis of the switching time

Here we evaluate the choice of the switching moment from a constant to a decreasing step size according to (13) from Theorem 3.2. We are using synthetic data that was generated in the same way as it had been in the Section 6 for the ridge regression problem ($n = 1000, d = 100$). In particular we evaluate 4 different cases: (i) the theoretical moment of regime switch at moment k as predicted from the Theorem, (ii) early switch at $0.3\times k$, (iii) late switch at $0.7\times k$ and (iv) the optimal k for switch, where the optimal k is obtained using one-dimensional numerical minimization of (54) as a function of k^*.

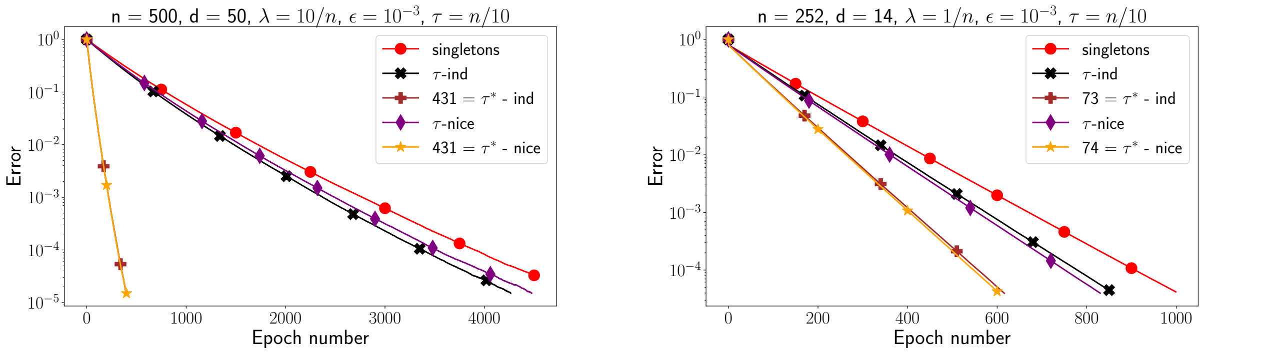

Figure 5. Performance of SGD with several minibatch strategies for ridge regression. On the left: the real data-set bodyfat from LIBSVM. On the right: synthetic data.

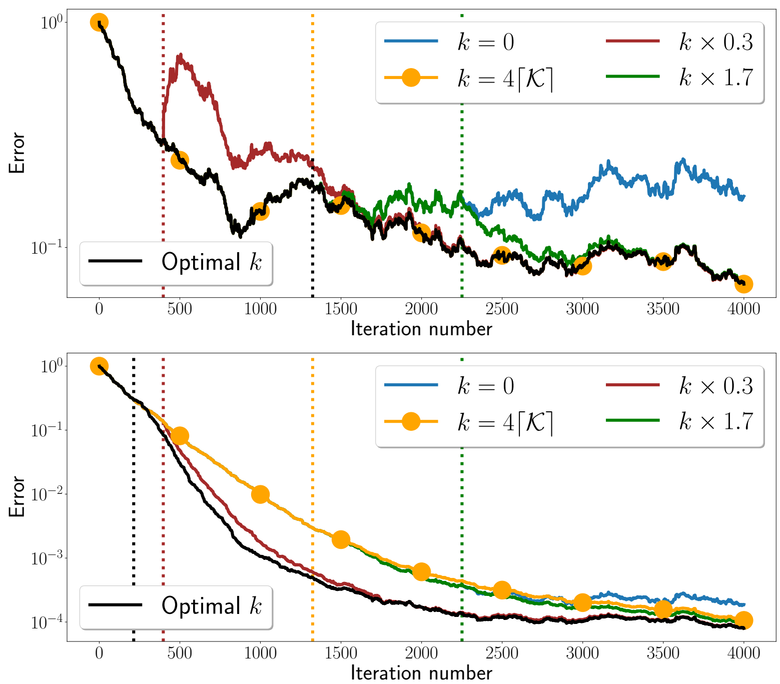

Figure 4. The first plot refers to situation when x^0 is close to x^* (for our data $\left\|r^0\right\|^2 = \left\|x^0 - x^*\right\|^2 \approx 1.0$). The second one covers the opposite case ($\left\|r^0\right\|^2 \approx 864.6$). Dotted verticals denote the moments of regime switch for the curves of the corresponding colour. The blue curve refers to constant step size $\frac{1}{2\mathcal{L}}$. Notice that in the upper plot optimal and theoretical k are very close

According to Figure 4, when x^0 is close to x^*, the moment of regime switch does not play a significant role in minimizing the number of iteration except for a very early switch, which actually also leads to almost the same situation in the long run. The case when x^0 is far from x^* shows that preliminary one-dimensional optimization makes sense and allows to reduce the error at least during the early iterations.

L.2. More on minibatches

Figure 5 reports on the same experiment as that described in Section 6.2 (Figure 2) in the main body of the paper, but on ridge regression instead of logistic regression, and using different data sets. Our findings are similar, and corroborate the conclusions made in Section 6.2.

L.3. Stepsize as a function of the minibatch size

In our last experiment we calculate the stepsize γ as a function of the minibatch size τ for τ-nice sampling using equation (36). Figure 6 depicts three plots, for three synthetic data sets of sizes $(n, d) \in \{(50, 5), (100, 10), (500, 50)\}$. We consider regularized ridge regression problems with $\lambda = 1/n$. Note that the stepsize is an increasing function of τ.

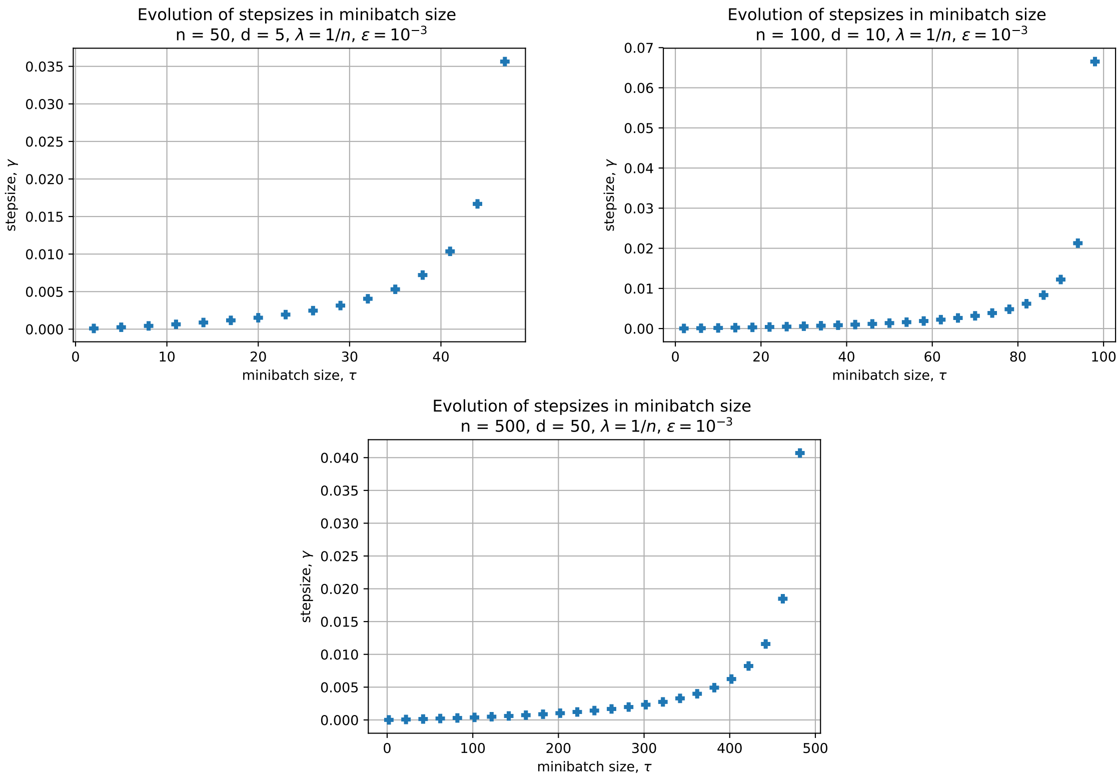

Figure 6. Evolution of stepsize with minibatch size τ for τ nice sampling.