Symbols	Meanings
i, j, l	counter variables
k	the number of sampled instances in SAMPLEVR
s	the counter of epochs
t	the counter of iterations in an epoch
i_t	the index of an instance which is sampled from training data randomly
α	the level of significance
δ	the rate of convergence
ϵ, ζ, ρ	positive real numbers
ν	the variance between the full gradient and its estimation
p	the number of dimensions
d_t	the Euclidean distance between the current parameter ω_t and its most recent snapshot
a_{ij}	the j_{th} entry of of d_i
b_{ij}	the j_{th} entry of of $\nabla f_i(\omega)$
$\check{\mu}$	the lower bound of the strongly-convex coefficient of loss function in S2GD
M	the maximal epoch size in S2GD
Δ_s	the decaying positive real number in EMGD
$\mathbb{B}_{\Delta_s}$	ω_{i_t} in EMGD is updated with $\parallel \omega_{i_{t+1}} - \omega_{i_t} \parallel \leq \Delta_s$
$\mathcal{H}_k$	the largest k elements of all dimensions of ω_{i_t} is kept and the other elements are set to be 0

Figure 1: Symbols used in the paper and their notations.

Symbol notations

The symbols used in the paper and their notations are presented in Figure 1. The symbols, whose meanings are straightforward to understand the paper, are not presented. Those symbols includes F, f_i with $i \in \{1, 2, ..., n\}$, ω and its variants, n, m and so on.

Proofs

In order to make the proofs of the theorems in this paper easy to read, the optimisation objective and assumptions are re-presented here.

The optimisation objective is:

$$\min F(\omega), \quad F(\omega) = \frac{1}{n}\sum_{i=1}^{n} f_i(\omega) + R(\omega) \tag{1}$$

. The assumptions are shown as follows:

Assumption 1. *Each a function f_i with $i \in \{1, 2, ..., n\}$ in Equ. 1 is L-Liptchiz continuous, that is, for any two parameters ω_i and ω_j:*

$$f_i(\omega_i) \leq f_i(\omega_j) + \nabla f_i(\omega_j)^{\mathrm{T}}(\omega_i - \omega_j) + \frac{L}{2} \parallel \omega_i - \omega_j \parallel^2 \tag{2}$$

.

Assumption 2. *The function F in Equ. 1 is μ-strongly convex, that is, for any two parameters ω_i and ω_j:*

$$F(\omega_i) \geq F(\omega_j) + \nabla F(\omega_j)^{\mathrm{T}}(\omega_i - \omega_j) + \frac{\mu}{2} \parallel \omega_i - \omega_j \parallel^2 \tag{3}$$

.

After that, all the proofs of the theorems presented in the main document are illustrated as follows:

Theorem 1. *After t iterations in an epoch, the distance d_t holds that $d_t = \eta^2 \sum_{j=1}^{p} \left(\sum_{i=1}^{t} a_{ij} \right)^2$. Furthermore, d_t has an upper bound such that $d_t \leq \eta^2 t^2 p \left(\frac{1}{tp} \sum_{i=1}^{t} \sum_{j=1}^{p} a_{ij}^2 \right)$, and a lower bound such that $d_t \geq \eta^2 t^2 p \left(\frac{1}{tp} \sum_{i=1}^{t} \sum_{j=1}^{p} a_{ij} \right)^2$.*

Proof.

$$\begin{aligned} d_t &= \parallel \omega_t - \tilde{\omega} \parallel^2 = \parallel \omega_{t-1} - \eta\gamma_{t-1} - \tilde{\omega} \parallel^2 \\ &= \parallel \omega_0 - \tilde{\omega} - \sum_{i=1}^{t} \eta\gamma_i \parallel^2 = \parallel - \sum_{i=1}^{t} \eta\gamma_i \parallel^2 \\ &= \eta^2 \sum_{j=1}^{p} \left(\sum_{i=1}^{t} a_{ij} \right)^2 \end{aligned} \tag{4}$$

. Taking the expectation of i_{i_t}, $\mathbb{E}(\gamma_t) = \mathbb{E}(\nabla f_{i_t}(\omega_t) - \nabla f_{i_t}(\tilde{\omega}) + \nabla F(\tilde{\omega})) = \nabla F(\omega_t)$ holds, and we thus obtain the upper bound of the distance:

$$\begin{aligned} d_t &= \eta^2 t^2 \sum_{j=1}^{p} \left(\frac{1}{t} \sum_{i=1}^{t} a_{ij} \right)^2 \leq \eta^2 t^2 \sum_{j=1}^{p} \left(\frac{1}{t} \sum_{i=1}^{t} a_{ij}^2 \right) \\ &= \eta^2 t \left(\sum_{i=1}^{t} \sum_{j=1}^{p} a_{ij}^2 \right) = \eta^2 t^2 p \left(\frac{1}{tp} \sum_{i=1}^{t} \sum_{j=1}^{p} a_{ij}^2 \right) \end{aligned} \tag{5}$$

, and the lower bound of the distance:

$$\begin{aligned} d_t &= \eta^2 p \left(\frac{1}{p} \sum_{j=1}^{p} \left(\sum_{i=1}^{t} a_{ij} \right)^2 \right) \geq \eta^2 p \left(\frac{1}{p} \sum_{j=1}^{p} \sum_{i=1}^{t} a_{ij} \right)^2 \\ &= \frac{\eta^2}{p} \left(\sum_{i=1}^{t} \sum_{j=1}^{p} a_{ij} \right)^2 = \eta^2 t^2 p \left(\frac{1}{tp} \sum_{i=1}^{t} \sum_{j=1}^{p} a_{ij} \right)^2 \end{aligned} \tag{6}$$

. □

Lemma 1. *Given $\nu = \frac{1}{k} \sum_{t=1}^{k} \nabla f_{i_t}(\omega) - \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(\omega)$, the inequality $\mathbb{E} \parallel \nu \parallel^2 \leq 2L(F(\omega) - F(\omega_*))$ holds for any an arbitrary parameter ω.*

Proof. Given any two arbitrary parameters ω_i and ω_j, $f_{i_t}(\omega_i) \leq f_{i_t}(\omega_j) + \nabla f_{i_t}(\omega_j)^{\mathrm{T}}(\omega_i - \omega_j) + \frac{L}{2} \parallel \omega_i - \omega_j \parallel^2$ holds according to Assumption 1. Let $\omega_i = \omega - \frac{1}{L}\nabla f(\omega)$, and $\omega_j = \omega$, and then we obtain

$$f_{i_t}(\omega_i) \leq f_{i_t}(\omega) - \frac{1}{2L} \parallel \nabla f_{i_t}(\omega) \parallel^2 \tag{7}$$

. i_t is taken on the expectation, and we thus obtain

$$\begin{aligned} &\mathbb{E}\left(\parallel \nabla f_{i_t}(\omega) \parallel^2 \right) \leq 2L\mathbb{E}\left(f_{i_t}(\omega) - f_{i_t}(\omega_i) \right) \\ &\leq 2L(F(\omega) - F(\omega_i)) \leq 2L(F(\omega) - F(\omega_*)) \end{aligned} \tag{8}$$

. Here, ω_* is the optimum of the loss function, i.e. F in Equation 1.

$$\begin{aligned}&\mathbb{E}\parallel\nu\parallel^2=\mathbb{E}\left(\parallel\frac{1}{k}\sum_{t=1}^{k}\nabla f_{i_t}(\omega)-\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\omega)\parallel^2\right)\\&\leq\mathbb{E}\left(\parallel\frac{1}{k}\sum_{t=1}^{k}\nabla f_{i_t}(\omega)\parallel^2\right)\leq\frac{1}{k}\sum_{t=1}^{k}\mathbb{E}\left(\parallel\nabla f_{i_t}(\omega)\parallel^2\right)\\&\leq 2L(F(\omega)-F(\omega_*))\end{aligned}\tag{9}$$

. The first inequality holds because of $\mathbb{E}(\frac{1}{k}\sum_{t=1}^{k}\nabla f_{i_t}(\tilde{\omega}_s))=\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\tilde{\omega}_s)$ and $\mathbb{E}(\xi-\mathbb{E}\xi)^2\leq\mathbb{E}\xi^2$. The second inequality holds because that the square of arithmetic mean is not greater than the arithmetic mean of the squares for a set of numbers. □

Theorem 2. *ω_* denotes the optimum of the parameter. m can be large enough so that $\delta=\frac{8L\eta^2m}{\eta(1-2\eta L)m-\frac{1}{\gamma}}<1$ holds.* SAMPLEVR *thus makes the optimisation objective converge linearly as follows: $F(\tilde{\omega}_{s+1})-F(\omega_*)\leq\delta[F(\tilde{\omega}_s)-F(\omega_*)]$.*

Proof. Construct an auxiliary function $h_i(\omega)=f_i(\omega)-f_i(\omega_*)-\nabla f_i(\omega_*)^{\mathrm{T}}(\omega-\omega_*)$, and $h_i(\omega_*)=\min_{\omega}h_i(\omega)$ holds because of $\nabla h_i(\omega_*)=0$. Thus, $h_i(\omega_*)\leq\min_{\eta}[h_i(\omega-\eta\nabla h_i(\omega))]$ holds. According to Assumption 1, we obtain $h_i(\omega_*)\leq\min_{\eta}[h_i(\omega)-\eta\parallel\nabla h_i(\omega)\parallel^2+\frac{1}{2}L\eta^2\parallel\nabla h_i(\omega)\parallel^2]=h_i(\omega)-\frac{1}{2L}\parallel\nabla h_i(\omega)\parallel^2$. That is, $\parallel\nabla f_i(\omega)-\nabla f_i(\omega_*)\parallel^2\leq 2L[f_i(\omega)-f_i(\omega_*)-\nabla f_i(\omega_*)^{\mathrm{T}}(\omega-\omega_*)]$. By summing this inequality over $i=\{1,2,...,n\}$, and using the fact that $\nabla F(\omega_*)=0$, we obtain $\frac{1}{n}\sum_{i=1}^{n}\parallel\nabla f_i(\omega)-\nabla f_i(\omega_*)\parallel^2\leq 2L[F(\omega)-F(\omega_*)]$.

Let $\dot{\gamma}_t=\nabla f_{i_t}(\omega_t)-\nabla f_{i_t}(\tilde{\omega}_s)+\nabla F(\tilde{\omega}_s)+\nu$ denote the update gradient of the parameter for the t_{th} iteration of the s_{th} epoch.

$$\begin{aligned}&\mathbb{E}\parallel\dot{\gamma}_t\parallel^2=\mathbb{E}\parallel\nabla f_{i_t}(\omega_t)-\nabla f_{i_t}(\tilde{\omega}_s)+\nabla F(\tilde{\omega}_s)+\nu\parallel^2\\&\leq 2\mathbb{E}\parallel\nabla f_{i_t}(\omega_t)-\nabla f_{i_t}(\omega_*)\parallel^2+\\&2\mathbb{E}\parallel\nabla f_{i_t}(\tilde{\omega}_s)-\nabla f_{i_t}(\omega_*)-\nabla F(\tilde{\omega}_s)-\nu\parallel^2\\&\leq 2\mathbb{E}\parallel\nabla f_{i_t}(\omega_t)-\nabla f_{i_t}(\omega_*)\parallel^2+\\&4\mathbb{E}\parallel\nabla f_{i_t}(\tilde{\omega}_s)-\nabla f_{i_t}(\omega_*)-\nabla F(\tilde{\omega}_s)\parallel^2+4\mathbb{E}\parallel\nu\parallel^2\\&\leq 2\mathbb{E}\parallel\nabla f_{i_t}(\omega_t)-\nabla f_{i_t}(\omega_*)\parallel^2+4\mathbb{E}\parallel\nabla f_{i_t}(\tilde{\omega}_s)-\nabla f_{i_t}(\omega_*)\\&-\mathbb{E}\left(\nabla f_{i_t}(\tilde{\omega}_s)-\nabla f_{i_t}(\omega_*)\right)\parallel^2+8L(F(\tilde{\omega}_s)-F(\omega_*))\\&\leq 2\mathbb{E}\parallel\nabla f_{i_t}(\omega_t)-\nabla f_{i_t}(\omega_*)\parallel^2+\\&4\mathbb{E}\parallel\nabla f_{i_t}(\tilde{\omega}_s)-\nabla f_{i_t}(\omega_*)\parallel^2+8L(F(\tilde{\omega}_s)-F(\omega_*))\\&\leq 4L[F(\omega_t)-F(\omega_*)]+16L[F(\tilde{\omega}_s)-F(\omega_*)]\end{aligned}\tag{10}$$

, according to Lemma 1 in the third inequality, and $\mathbb{E}[\xi-\mathbb{E}\xi]^2\leq\mathbb{E}\xi^2$ in the fourth inequality.

$$\begin{aligned}&\mathbb{E}\parallel\omega_{t+1}-\omega_*\parallel^2\\&=\parallel\omega_t-\omega_*\parallel^2-2\eta(\omega_t-\omega_*)^{\mathrm{T}}\mathbb{E}\dot{\gamma}_t+\eta^2\mathbb{E}\parallel\dot{\gamma}_t\parallel^2\\&\leq\parallel\omega_t-\omega_*\parallel^2-2\eta(\omega_t-\omega_*)^{\mathrm{T}}\nabla F(\omega_t)+\\&\eta^2\left(4L[F(\omega_t)-F(\omega_*)]+16L[F(\tilde{\omega}_s)-F(\omega_*)]\right)\\&\leq\parallel\omega_t-\omega_*\parallel^2-2\eta(F(\omega_t)-F(\omega_*))+\\&\eta^2\left(4L[F(\omega_t)-F(\omega_*)]+16L[F(\tilde{\omega}_s)-F(\omega_*)]\right)\\&=\parallel\omega_t-\omega_*\parallel^2-2\eta(1-2\eta L)[F(\omega_t)-F(\omega_*)]+\\&16L\eta^2[F(\tilde{\omega}_s)-F(\omega_*)]\end{aligned}\tag{11}$$

. Summing the above inequality over $t=0,1,...,m-1$, we obtain

$$\begin{aligned}&\mathbb{E}\parallel\omega_m-\omega_*\parallel^2\\&\leq\parallel\omega_0-\omega_*\parallel^2-2\eta(1-2\eta L)\sum_{i=0}^{m-1}[F(\omega_i)-F(\omega_*)]\\&+16L\eta^2m[F(\tilde{\omega}_s)-F(\omega_*)]\end{aligned}\tag{12}$$

. When $\tilde{\omega}_{s+1}$ is randomly identified from the sequence $\{\omega_0,...,\omega_{m_s-1}\}$. Taking expectation on t, we obtain $\tilde{\omega}_{s+1}=\mathbb{E}(\omega_t)$ with $t=\{0,1,...,m-1\}$. Therefore,

$$\begin{aligned}&\mathbb{E}\parallel\omega_m-\omega_*\parallel^2+2\eta(1-2\eta L)m[F(\tilde{\omega}_{s+1})-F(\omega_*)]\\&\leq\mathbb{E}\parallel\omega_0-\omega_*\parallel^2+16L\eta^2m[F(\tilde{\omega}_s)-F(\omega_*)]\\&=\parallel\tilde{\omega}_{s+1}-\omega_*\parallel^2+16L\eta^2m[F(\tilde{\omega}_s)-F(\omega_*)]\\&\leq\frac{2}{\mu}\parallel F(\tilde{\omega}_{s+1})-F(\omega_*)\parallel^2+16L\eta^2m[F(\tilde{\omega}_s)-F(\omega_*)]\end{aligned}\tag{13}$$

. The second equality holds because of $\omega_{s+1}=\mathbb{E}(\omega_t)$ with $t=\{0,1,...,m-1\}$. The third inequality holds due to the Assumption 3. Thus,

$$\begin{aligned}&\left(\eta(1-2\eta L)m-\frac{1}{\mu}\right)[F(\tilde{\omega}_{s+1})-F(\omega_*)]\\&\leq 8L\eta^2m[F(\tilde{\omega}_s)-F(\omega_*)]\end{aligned}\tag{14}$$

. That is, $F(\tilde{\omega}_{s+1})-F(\omega_*)\leq\frac{8L\eta^2m}{\eta(1-2\eta L)m-\frac{1}{\mu}}[F(\tilde{\omega}_s)-F(\omega_*)]$. Thus, the Theorem 2 have been proved. □

Theorem 3. SAMPLEVR *requires at least $\frac{\ln\zeta}{\ln\delta}m+\left(-\frac{\log\frac{\alpha}{2}}{2\epsilon}(\frac{\ln\zeta}{\ln\delta}+1)(\frac{\ln\zeta}{\ln\delta})\right)$ atomic gradient calculations with $\delta=\frac{8L\eta^2m}{\eta(1-2\eta L)m-\frac{1}{\mu}}$ to achieve $F(\tilde{\omega}_s)-F(\omega_*)\leq\zeta[F(\omega_0)-F(\omega_*)]$.*

Proof. The required atomic gradient calculations for the s_{th} epoch is denoted by G_s. We obtain $G_s=k+m=-\frac{s\log\frac{\alpha}{2}}{\epsilon}+m$. If $F(\tilde{\omega}_s)-F(\omega_*)\leq\zeta[F(\omega_0)-F(\omega_*)]$ holds, then we obtain $\delta^s=\zeta$ according to Theorem 2, that is, $s=\frac{\ln\zeta}{\ln\delta}$. Therefore, the total gradient complexity is $\frac{\ln\zeta}{\ln\delta}m+\left(-\frac{\log\frac{\alpha}{2}}{2\epsilon}(\frac{\ln\zeta}{\ln\delta}+1)(\frac{\ln\zeta}{\ln\delta})\right)$ atomic gradient calculations with $\delta=\frac{8L\eta^2m}{\eta(1-2\eta L)m-\frac{1}{\mu}}$.

□