Symbols	Meanings
i, j, l	the 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 $<x_{i_t}, y_{i_t}>$ which is sampled randomly
α	the level of significance
ρ	$[-\rho, \rho]$ is the $(1-\alpha)$-confidence interval
δ	the rate of convergence
p	the number of dimensions
$\omega, \tilde{\omega}, \omega_*$	ω_* is the optimum. ω is a parameter, and $\tilde{\omega}$ is its snapshot
d, d_t	the variance, $d_t = \parallel \omega_t - \tilde{\omega}_t \parallel^2$
a_{ij}	the j_{th} entry of of d_i
b_{ij}	the j_{th} entry of of $\nabla f_i(\omega)$
$\gamma_t, \dot{\gamma_t}$	the update gradient, and $\dot{\gamma_t}$ is its estimation.
m_s, m	the epoch size, and m is a constant.
η	the constant learning rate
ϵ, ζ	the positive real numbers
$\parallel \cdot \parallel$	the 2-norm of a vector
$g, \dot{g}$	the full gradient g and its estimation $\dot{g}$
ν	$\nu = \dot{g} - g$

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.

Proofs

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

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

. The assumptions are shown as follows:

Assumption 1. *Each differentiable function f_{i_t} with $i_t \in \{1, 2, ..., n\}$ in Equation 1 is L-Liptchiz continuous, that is, $\parallel f_{i_t}(\omega_i) - f_{i_t}(\omega_j) \parallel \leq L \parallel \omega_i - \omega_j \parallel$ holds for any two parameters ω_i and ω_j. Equivalently, we obtain*

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

.

Assumption 2. *The function F in Equation11 is μ-strongly convex. That is, for any two parameters ω_i and ω_j, we obtain*

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

.

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.

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

. Taken the expectation of 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:

$$\begin{aligned} d_t &= \eta^2 t^2 \sum_{j=1}^{p} \left(\frac{1}{t}\sum_{i=1}^{t} a_{ij}\right)^2 \leq \eta^2 t^2 \sum_{j=1}^{p} \left(\frac{1}{t}\sum_{i=1}^{t} a_{ij}^2\right) \\ &= \eta^2 t \left(\sum_{i=1}^{t}\sum_{j=1}^{p} a_{ij}^2\right) = \eta^2 t^2 p \left(\frac{1}{tp}\sum_{i=1}^{t}\sum_{j=1}^{p} a_{ij}^2\right) \end{aligned} \tag{3}$$

, and the lower bound of the distance:

$$\begin{aligned} d_t &= \eta^2 p \left(\frac{1}{p}\sum_{j=1}^{p}\left(\sum_{i=1}^{t} a_{ij}\right)^2\right) \geq \eta^2 p \left(\frac{1}{p}\sum_{j=1}^{p}\sum_{i=1}^{t} a_{ij}\right)^2 \\ &= \frac{\eta^2}{p}\left(\sum_{i=1}^{t}\sum_{j=1}^{p} a_{ij}\right)^2 = \eta^2 t^2 p \left(\frac{1}{tp}\sum_{i=1}^{t}\sum_{j=1}^{p} a_{ij}\right)^2 \end{aligned} \tag{4}$$

. □

Lemma 1. *Given $\nu = \frac{1}{k}\sum_{t=1}^{k} \nabla f_{i_t}(\omega) - \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\omega)$, we obtain $\parallel \nu \parallel^2 \leq \frac{2(n-k)L^2}{nk}$.*

Proof. Since f_{i_t} is L-Liptchiz continuous according to Assumption 1, we obtain $\parallel f_{i_t}(\omega_i) - f_{i_t}(\omega_j) \parallel \leq L \parallel \omega_i - \omega_j \parallel$. Thus, $\parallel \nabla f_{i_t}(\omega) \parallel \leq L$ holds for an arbitrary parameter ω. Without loss of generality, suppose that indices of the sampled k instances are $i_t = t$ with $t \in \{1, 2, ..., k\}$.

$$\begin{aligned} &\parallel \nu \parallel^2 = \parallel \frac{1}{k}\sum_{t=1}^{k} \nabla f_t(\omega) - \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\omega) \parallel^2 \\ &= \frac{1}{(nk)^2} \parallel (n-k)\sum_{t=1}^{k} \nabla f_t(\omega) - k\sum_{i=k+1}^{n} \nabla f_i(\omega) \parallel^2 \\ &\leq \frac{1}{(nk)^2}(2(n-k)^2 k^2 \parallel \frac{1}{k}\sum_{t=1}^{k} \nabla f_t(\omega) \parallel^2 \\ &+ 2k^2(n-k)^2 \parallel \frac{1}{n-k}\sum_{i=k+1}^{n} \nabla f_i(\omega) \parallel^2) \\ &\leq \frac{1}{(nk)^2}(2(n-k)^2 k \sum_{t=1}^{k} \parallel \nabla f_t(\omega) \parallel^2 \\ &+ 2k^2(n-k)\sum_{i=k+1}^{n} \parallel \nabla f_i(\omega) \parallel^2) \\ &\leq \frac{4(n-k)^2 L^2}{n^2} \end{aligned} \tag{5}$$

. □

Theorem 2. *Given* $\delta = \frac{1+4L\mu m\eta^2}{\mu m\eta(1-2\eta L)} < 1$ *holds with* $\frac{1}{12L}\left(1-\sqrt{\frac{\mu m-24L}{\mu m}}\right) < \eta < \frac{1}{12L}\left(1+\sqrt{\frac{\mu m-24L}{\mu m}}\right)$, SAMPLEVR *makes the training loss converge as* $\mathbb{E}[F(\tilde{\omega}_{s+1})-F(\omega_*)] \leq \delta\mathbb{E}[F(\tilde{\omega}_s)-F(\omega_*)]+\frac{8(\epsilon n+s\log\frac{\alpha}{2})^2L^2\eta}{\epsilon^2n^2(1-2\eta L)}$.

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. 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 $\frac{1}{n}\sum_{i=1}^{n} \parallel \nabla f_i(\omega)-\nabla f_i(\omega_*) \parallel^2 \leq 2L[F(\omega)-F(\omega_*)]$. i_t is a random variable which is sampled from $\{1,2,...,n\}$ randomly. Taking the expectation of i_t, we obtain

$$\begin{aligned}&\mathbb{E}_{i_t}(\parallel \nabla f_{i_t}(\omega)-\nabla f_{i_t}(\omega_*) \parallel^2)\\ &= \frac{1}{n}\sum_{i=1}^{n} \parallel \nabla f_i(\omega)-\nabla f_i(\omega_*) \parallel^2\leq 2L[F(\omega)-F(\omega_*)]\end{aligned} \tag{6}$$

. Therefore,

$$\begin{aligned}&\mathbb{E}_{i_t} \parallel \dot{\gamma}_t \parallel^2= \mathbb{E}_{i_t} \parallel \nabla f_{i_t}(\omega_t)-\nabla f_{i_t}(\tilde{\omega}_s)+\nabla F(\tilde{\omega}_s)+\nu \parallel^2\\ &\leq 2\mathbb{E}_{i_t} \parallel \nabla f_{i_t}(\omega_t)-\nabla f_{i_t}(\omega_*) \parallel^2 +\\ &2\mathbb{E}_{i_t} \parallel \nabla f_{i_t}(\tilde{\omega}_s)-\nabla f_{i_t}(\omega_*)-\nabla F(\tilde{\omega}_s)-\nu \parallel^2\\ &\leq 2\mathbb{E}_{i_t} \parallel \nabla f_{i_t}(\omega_t)-\nabla f_{i_t}(\omega_*) \parallel^2 +\\ &4\mathbb{E}_{i_t} \parallel \nabla f_{i_t}(\tilde{\omega}_s)-\nabla f_{i_t}(\omega_*)-\nabla F(\tilde{\omega}_s) \parallel^2 +4\mathbb{E}_{i_t} \parallel \nu \parallel^2\\ &\leq 2\mathbb{E}_{i_t} \parallel \nabla f_{i_t}(\omega_t)-\nabla f_{i_t}(\omega_*) \parallel^2 +4\mathbb{E}_{i_t} \parallel \nabla f_{i_t}(\tilde{\omega}_s)-\nabla f_{i_t}(\omega_*)\\ &-\mathbb{E}_{i_t}(\nabla f_{i_t}(\tilde{\omega}_s)-\nabla f_{i_t}(\omega_*)) \parallel^2 +\frac{16(n-k)^2L^2}{n^2}\\ &\leq 2\mathbb{E}_{i_t} \parallel \nabla f_{i_t}(\omega_t)-\nabla f_{i_t}(\omega_*) \parallel^2 +\\ &4\mathbb{E}_{i_t} \parallel \nabla f_{i_t}(\tilde{\omega}_s)-\nabla f_{i_t}(\omega_*) \parallel^2 +\frac{16(n-k)^2L^2}{n^2}\\ &\leq 4L[F(\omega_t)-F(\omega_*)]+8L[F(\tilde{\omega}_s)-F(\omega_*)]+\frac{16(n-k)^2L^2}{n^2}\end{aligned} \tag{7}$$

. The third inequality uses Lemma 1, and the fourth inequality uses $\mathbb{E}[\xi-\mathbb{E}\xi]^2 \leq \mathbb{E}\xi^2$, and the fifth inequality uses (6). Therefore, we obtain

$$\begin{aligned}&\mathbb{E}_{i_t} \parallel \omega_{t+1}-\omega_* \parallel^2= \mathbb{E}_{i_t} \parallel \omega_t-\eta\dot{\gamma}_t-\omega_* \parallel^2\\ &=\parallel \omega_t-\omega_* \parallel^2 -2\eta(\omega_t-\omega_*)^{\mathrm{T}}\mathbb{E}_{i_t}\dot{\gamma}_t+\eta^2\mathbb{E}_{i_t} \parallel \dot{\gamma}_t \parallel^2\\ &\leq\parallel \omega_t-\omega_* \parallel^2 -2\eta(\omega_t-\omega_*)^{\mathrm{T}}\nabla F(\omega_t)+\\ &\eta^2\left(4L[F(\omega_t)-F(\omega_*)]+8L[F(\tilde{\omega}_s)-F(\omega_*)]+\frac{16(n-k)^2L^2}{n^2}\right)\\ &\leq\parallel \omega_t-\omega_* \parallel^2 -2\eta(F(\omega_t)-F(\omega_*))+\\ &\eta^2\left(4L[F(\omega_t)-F(\omega_*)]+8L[F(\tilde{\omega}_s)-F(\omega_*)]+\frac{16(n-k)^2L^2}{n^2}\right)\\ &=\parallel \omega_t-\omega_* \parallel^2 -2\eta(1-2\eta L)[F(\omega_t)-F(\omega_*)]+\\ &8L\eta^2[F(\tilde{\omega}_s)-F(\omega_*)]+\frac{16(n-k)^2L^2\eta^2}{n^2}\end{aligned} \tag{8}$$

. The first inequality uses $\mathbb{E}_{i_t}(\dot{\gamma}_t)=\nabla F(\omega_t)$ and (7), and the second inequality holds because that $F(\omega)$ is convex. We thus obtain

$$\begin{aligned}&\parallel \omega_m-\omega_* \parallel^2\\ &\leq\parallel \omega_0-\omega_* \parallel^2 -2\eta(1-2\eta L)\sum_{t=0}^{m-1}[F(\omega_t)-F(\omega_*)]+\\ &8Lm\eta^2[F(\tilde{\omega}_s)-F(\omega_*)]+\frac{16m(n-k)^2L^2\eta^2}{n^2}\\ &=\parallel \tilde{\omega}_s-\omega_* \parallel^2 -2\eta(1-2\eta L)m\left(\frac{1}{m}\sum_{t=0}^{m-1}[F(\omega_t)-F(\omega_*)]\right)+\\ &8Lm\eta^2[F(\tilde{\omega}_s)-F(\omega_*)]+\frac{16m(n-k)^2L^2\eta^2}{n^2}\\ &=\parallel \tilde{\omega}_s-\omega_* \parallel^2 -2\eta(1-2\eta L)m\mathbb{E}_t[F(\tilde{\omega}_{s+1})-F(\omega_*)]+\\ &8Lm\eta^2[F(\tilde{\omega}_s)-F(\omega_*)]+\frac{16m(n-k)^2L^2\eta^2}{n^2}\end{aligned} \tag{9}$$

. The second equality holds because of $\omega_0 = \tilde{\omega}_s$. The third equality holds when we take expectation of t. The reason is that $\tilde{\omega}_{s+1}$ is identified by picking ω_t with $t \in \{0,1,...,m-1\}$ randomly, and $\tilde{\omega}_s$ is a constant in an epoch. Thus,

$$\begin{aligned}&2\eta(1-2\eta L)m\mathbb{E}[F(\tilde{\omega}_{s+1})-F(\omega_*)]\\ &\leq\parallel \tilde{\omega}_s-\omega_* \parallel^2 +8Lm\eta^2\mathbb{E}[F(\tilde{\omega}_s)-F(\omega_*)]+\frac{16m(n-k)^2L^2\eta^2}{n^2}\\ &\leq \frac{2}{\mu}\mathbb{E}[F(\tilde{\omega}_s)-F(\omega_*)]+8Lm\eta^2\mathbb{E}[F(\tilde{\omega}_s)-F(\omega_*)]\\ &+\frac{16m(n-k)^2L^2\eta^2}{n^2}\end{aligned} \tag{10}$$

. The second inequality holds due to the Assumption 2. Therefore, we obtain $\delta = \frac{1+4L\mu m\eta^2}{\mu m\eta(1-2\eta L)} < 1$ with $\frac{1}{12L}\left(1-\sqrt{\frac{\mu m-24L}{\mu m}}\right) < \eta < \frac{1}{12L}\left(1+\sqrt{\frac{\mu m-24L}{\mu m}}\right)$, and thus the training loss converges such that $\mathbb{E}[F(\tilde{\omega}_{s+1})-F(\omega_*)] \leq \delta\mathbb{E}[F(\tilde{\omega}_s)-F(\omega_*)]+\frac{8(n-k)^2L^2\eta}{n^2(1-2\eta L)}$. Considering $k=\frac{-s\log\frac{\alpha}{2}}{\epsilon}$, we obtain $\mathbb{E}[F(\tilde{\omega}_{s+1})-F(\omega_*)] \leq \delta\mathbb{E}[F(\tilde{\omega}_s)-F(\omega_*)]+\frac{8(\epsilon n+s\log\frac{\alpha}{2})^2L^2\eta}{\epsilon^2n^2(1-2\eta L)}$. Thus, the Theorem 2 have been proved. □

Theorem 3. *Let* α *be small enough, so that* $\frac{8(\epsilon n+\log\frac{\alpha}{2})^2L^2\eta}{\epsilon^2n^2(1-2\eta L)} \leq F(\tilde{\omega}_0)-F(\omega_*)$ *holds,* SAMPLEVR *requires* $O(\ln^2\frac{1}{\epsilon})$ *atomic gradient calculations to achieve* $\mathbb{E}[F(\tilde{\omega}_s)-F(\omega_*)] \leq \zeta$.

Proof. $\mathbb{E}[F(\tilde{\omega}_{s+1})-F(\omega_*)] \leq \delta^s(2\mathbb{E}[F(\tilde{\omega}_0)-F(\omega_*)])$, according to Theorem 2. If $\mathbb{E}[F(\tilde{\omega}_{s+1})-F(\omega_*)] \leq \zeta$ holds, we obtain $s \geq \frac{1}{\ln\delta}\ln\frac{\zeta}{2(F(\tilde{\omega}_0)-F(\omega_*))}$ which can be denoted by $s = O(\ln\frac{1}{\zeta})$. Here, ω_* is the optimum of the loss function. The required atomic gradient calculations for the s_{th} epoch is denoted by G_s. We obtain $G_s = k+m = -\frac{s\log\frac{\alpha}{2}}{\epsilon}+m$, which can be denoted by $O(\ln\frac{1}{\zeta})$ because of $s = O(\ln\frac{1}{\zeta})$. Thus, the total gradient complexity is denoted by $O(\ln^2\frac{1}{\zeta})$. □