Theorem 1. *If γ_t is p-dimensional, and can be denoted by $\gamma_t = (a_t^1, a_t^2 ..., a_t^p)$. $\bar{a_j} = \sum_{i=1}^{p} a_j^i$. After t iterations in the epoch, the distance d_t holds that $d_t \geq p\eta^2 \sum_{j=1}^{t} a_j^2$ when the learning rate, i.e. η_j with $j = 1, 2, ...m^s$ is a constant.*

Proof.

$$
\begin{aligned}
d_t &= \parallel \omega_{i_t}^s - \tilde{\omega}^s \parallel^2 \\
&= \parallel \omega_{i_{t-1}^s} - \eta_t \gamma_t^s - \tilde{\omega}^s \parallel^2 \\
&= \parallel d_{t-1} - \eta_t \gamma_t^s \parallel^2 \\
&= \parallel d_0 - \sum_{j=1}^{t} \eta_j \gamma_j^s \parallel^2 \\
&= \parallel - \sum_{j=1}^{t} \eta_j \gamma_j^s \parallel^2
\end{aligned}
\tag{1}
$$

. Since $\mathbb{E}\gamma_t = \nabla F(\omega_{i_t}^s)$, $\mathbb{E}d_t = -\sum_{j=1}^{t} \eta_j \nabla F(\omega_{i_t}^s)$ holds. Therefore, we obtain

$$
\begin{aligned}
d_t &= p \frac{\parallel \sum_{j=1}^{t} \eta_j \gamma_j^s \parallel^2}{p} = p \frac{\sum_{j=1}^{t} \sum_{i=1}^{p} (\eta_j a_j^i)^2}{p} \\
&= p \sum_{j=1}^{t} \eta_j^2 \frac{\sum_{i=1}^{p} (a_j^i)^2}{p} \geq p \sum_{j=1}^{t} \eta_j^2 \left(\frac{\sum_{i=1}^{p} a_j^i}{p} \right)^2 \\
&\overset{\eta_j = \eta}{=} p\eta^2 \sum_{j=1}^{t} (\bar{a_j})^2
\end{aligned}
\tag{2}
$$

. The above inequality uses $\frac{x_1 + x_2 +, ..., + x_n}{n} \leq \sqrt{\frac{x_1^2 + x_2^2 +, ..., + x_n^2}{n}}$. □

Theorem 2. *Assume the objective needs s epochs to achieve to the ϵ, and then the gradient complexity is $(1 - \alpha) \left(m - \frac{1}{6} \frac{\log \frac{\alpha}{2}}{\rho_0^2} s(s+1)(2s+1) \right) + \alpha n$ atomic gradient computation.*

Proof. The average atomic gradient for an epoch is denoted by G_{avg}, and we obtain

$$
\begin{aligned}
G_s &= (1 - \alpha)k + \alpha n + m \\
&= -(1 - \alpha) \frac{\log \frac{\alpha}{2}}{\rho^2} + \alpha n + m \\
&= -(1 - \alpha) \frac{(\log \frac{\alpha}{2}) s^2}{\rho_0^2} + \alpha n + m
\end{aligned}
\tag{3}
$$

. So the total gradient computation of all the s epochs is $s(m + \alpha n) + (1 - \alpha) \left(-\frac{\log \frac{\alpha}{2} s(s+1)(2s+1)}{6\rho_0^2} \right)$.

If the $F(\tilde{\omega}^s) - F(\omega_*) \leq \epsilon[\tilde{\omega}^0) - F(\omega_*)]$, then we obtain $\delta^s = \epsilon$, that is, $s = \ln \frac{\epsilon}{\delta}$. Therefore, the total gradient complexity is $(\ln \frac{\epsilon}{\delta})(m + \alpha n) + (1 - \alpha) \left(-\frac{\log \frac{\alpha}{2}}{\rho_0^2} (\ln \frac{\epsilon}{\delta})(\ln \frac{\epsilon}{\delta} + 1)(2 \ln \frac{\epsilon}{\delta} + 1) \right)$ atomic gradient computation.

□

Theorem 3. *Assume the objective needs s epochs to achieve to the ϵ, and then the gradient complexity is $(1 - \alpha) \left(m - \frac{1}{6} \frac{\log \frac{\alpha}{2}}{\rho_0^2} s(s+1)(2s+1) \right) + \alpha n$ atomic gradient computation.*

Proof. The average atomic gradient for an epoch is denoted by G_{avg}, and we obtain

$$
\begin{aligned}
G_s &= (1 - \alpha)k + \alpha n + m \\
&= -(1 - \alpha) \frac{\log \frac{\alpha}{2}}{\rho^2} + \alpha n + m \\
&= -(1 - \alpha) \frac{(\log \frac{\alpha}{2}) s^2}{\rho_0^2} + \alpha n + m
\end{aligned}
\tag{4}
$$

. So the total gradient computation of all the s epochs is $s(m + \alpha n) + (1 - \alpha) \left(-\frac{\log \frac{\alpha}{2} s(s+1)(2s+1)}{6\rho_0^2} \right)$.

If the $F(\tilde{\omega}^s) - F(\omega_*) \leq \epsilon[\tilde{\omega}^0) - F(\omega_*)]$, then we obtain $\delta^s = \epsilon$, that is, $s = \ln \frac{\epsilon}{\delta}$. Therefore, the total gradient complexity is $(\ln \frac{\epsilon}{\delta})(m + \alpha n) + (1 - \alpha) \left(-\frac{\log \frac{\alpha}{2}}{\rho_0^2} (\ln \frac{\epsilon}{\delta})(\ln \frac{\epsilon}{\delta} + 1)(2 \ln \frac{\epsilon}{\delta} + 1) \right)$ atomic gradient computation. □

Lemma 1. *ω_* denotes the optimum of the parameter. m^s can be large enough, so that $\delta = \frac{4L\eta^2 m^s}{\eta(1-2\eta L)m^s - \frac{1}{\gamma}} < 1$, EstimateVR converges at the rate as follows:*
$F(\tilde{\omega}^{s+1}) - F(\omega_*) \leq \delta[F(\tilde{\omega}^s) - F(\omega_*)] + \frac{\delta}{2L} d^s$.

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. That is,

$$
\begin{aligned}
& h_i(\omega_*) \leq \min_{\eta}[h_i(\omega) - \eta \parallel \nabla h_i(\omega) \parallel^2 + \tfrac{1}{2} L\eta^2 \parallel \nabla h_i(\omega) \parallel)] \\
& = h_i(\omega) - \frac{(}{1})(2L) \parallel \nabla g_i(\omega) \parallel^2
\end{aligned}
\tag{5}
$$

. That is, $\parallel \nabla f_i(\omega) - \nabla f_i(\omega_*) \parallel \leq 2L[f_i(\omega) - f_i(\omega_*) - \nabla f_i(\omega_*)^T(\omega - \omega_*)]$. By summing the above inequality over $i = 1, 2, ..., n$, and using the fact that $\nabla F(\omega_*) = 0$, we obtain

$$
\frac{1}{n} \sum_{i=1}^{n} \parallel \nabla f_i(\omega) - \nabla f_i(\omega_*) \parallel^2 \leq 2L[F(\omega) - F(\omega_*)]
\tag{6}
$$

. Let $v_{t+1} = \nabla f_{i_{t+1}}(\omega_t) - \nabla f_{i_{t_1}}(\tilde{\omega}^s) + g^s + n^s$ during the $t + 1_{th}$ round in the epoch of the s_{th} iteration. Conditioned on ω_t, i_{t+1} is taken on the expectation, and we get

$$
\begin{aligned}
& \mathbb{E} \parallel v_{t+1} \parallel^2 = \mathbb{E} \parallel \nabla f_{i_{t+1}}(\omega_t) - \nabla f_{i_{t+1}}(\tilde{\omega}^s) + g^s + n^s \parallel \\
& \leq 2\mathbb{E} \parallel \nabla f_{i_{t+1}}(\omega_t) - \nabla f_{i_{t+1}}(\omega_*) \parallel^2 + 2\mathbb{E} \parallel \nabla f_{i_{t+1}}(\tilde{\omega}^s) - \nabla f_{i_{t+1}}(\omega_*) \\
& \leq 2\mathbb{E} \parallel \nabla f_{i_{t+1}}(\omega_t) - \nabla f_{i_{t+1}}(\omega_*) \parallel^2 + 4\mathbb{E} \parallel \nabla f_{i_{t+1}}(\tilde{\omega}^s) - \nabla f_{i_{t+1}}(\omega_*) \\
& = 2\mathbb{E} \parallel \nabla f_{i_{t+1}}(\omega_t) - \nabla f_{i_{t+1}}(\omega_*) \parallel^2 + 4\mathbb{E} \parallel \nabla f_{i_{t+1}}(\tilde{\omega}^s) - \nabla f_{i_{t+1}}(\omega_*) \\
& \leq 2\mathbb{E} \parallel \nabla f_{i_{t+1}}(\omega_t) - \nabla f_{i_{t+1}}(\omega_*) \parallel^2 + 4\mathbb{E} \parallel \nabla f_{i_{t+1}}(\tilde{\omega}^s) - \nabla f_{i_{t+1}}(\omega_*) \\
& \leq 2\mathbb{E} \parallel \nabla f_{i_{t+1}}(\omega_t) - \nabla f_{i_{t+1}}(\omega_*) \parallel^2 + 4\mathbb{E} \parallel \nabla f_{i_{t+1}}(\tilde{\omega}^s) - \nabla f_{i_{t+1}}(\omega_*) \\
& \leq 4L[F(\omega_t) - F(\omega_*)] + 8L[F(\tilde{\omega}^s) - F(\omega_*)] + 4d^s
\end{aligned}
\tag{7}
$$

, because of holding that $\| n^s \|^2 = d^s$ in the third inequality, and $\mathbb{E}[\xi - \mathbb{E}\xi]^2 \leq \mathbb{E}\xi^2$ in the fourth inequality.

$$\begin{aligned}
&\mathbb{E} \| \omega_{t+1} - \omega_* \|^2 = \| \omega_t - \omega_* \|^2 - 2\eta(\omega_t - \omega_*)^{\mathrm{T}} \mathbb{E} v_t + \eta^2 \| v_t \|^2 \\
&\leq \| \omega_t - \omega_* \|^2 - 2\eta(\omega_t - \omega_*)^{\mathrm{T}} \nabla F(\omega_t) + \eta^2 \left(4L[F(\omega_t) - F(\omega_*)] + 8L[F(\tilde{\omega}^s) - F(\omega_*)] + 4d^s\right) \\
&\leq \| \omega_t - \omega_* \|^2 - 2\eta(F(\omega_t) - F(\omega_*)) + \eta^2 \left(4L[F(\omega_t) - F(\omega_*)] + 8L[F(\tilde{\omega}^s) - F(\omega_*)] + 4d^s\right) \\
&= \| \omega_t - \omega_* \|^2 - 2\eta(1 - 2\eta L)[F(\omega_t) - F(\omega_*)] + 8L\eta^2[F(\tilde{\omega}^s) - F(\omega_*)] + 4\eta^2 d^s
\end{aligned} \tag{8}$$

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

$$\begin{aligned}
&\mathbb{E} \| \omega_{m^s} - \omega_* \|^2 \\
&\leq \| \omega_0 - \omega_* \|^2 - 2\eta(1 - 2\eta L) \sum_{i=0}^{m^s-1} [F(\omega_i) - F(\omega_*)] + 8L\eta^2 m^s [F(\tilde{\omega}^s) - F(\omega_*)] + 4\eta^2 m^s d^s
\end{aligned} \tag{9}$$

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

$$\begin{aligned}
&\mathbb{E} \| \omega_{m^s} - \omega_* \|^2 + 2\eta(1 - 2\eta L) m^s [F(\tilde{\omega}^{s+1}) - F(\omega_*)] \\
&\leq \| \omega_0 - \omega_* \|^2 + 8L\eta^2 m^s [F(\tilde{\omega}^s) - F(\omega_*)] + 4\eta^2 m^s d^s \\
&= \| \tilde{\omega}^{s+1} - \omega_* \|^2 + 8L\eta^2 m^s [F(\tilde{\omega}^s) - F(\omega_*)] + 4\eta^2 m^s d^s \\
&\leq \frac{2}{\gamma} \| F(\tilde{\omega}^{s+1} - F(\omega_*)) \|^2 + 8L\eta^2 m^s [F(\tilde{\omega}^s) - F(\omega_*)] + 4\eta^2 m^s d^s
\end{aligned} \tag{10}$$

. Thus,

$$\begin{aligned}
&\left(\eta(1 - 2\eta L) m^s - \tfrac{1}{\gamma}\right) [F(\tilde{\omega}^{s+1}) - F(\omega_*)] \\
&\leq 4L\eta^2 m^s [F(\tilde{\omega}^s) - F(\omega_*)] + 2\eta^2 m^s d^s
\end{aligned} \tag{11}$$

. That is, $F(\tilde{\omega}^{s+1}) - F(\omega_*) \leq \frac{4L\eta^2 m^s}{\eta(1-2\eta L)m^s - \frac{1}{\gamma}} [F(\tilde{\omega}^s) - F(\omega_*)] + \frac{2\eta^2 m^s d^s}{\eta(1-2\eta L)m^s - \frac{1}{\gamma}}$. Thus, the Lemma 1 have been proved. $\square$