Clarification of our assumptions

First, we would like to thank all the reviewers for their efforts in the review process, and for acknowledging the novelty/presentation of our work.

Both reviewers bGwb and 93YJ share the concern that the assumptions are made on stochastic functions and not the expectation ones, which could be potentially a limitation and make it hard to tell the source of improvement.

We appreciate the insightful feedback from the reviewers. We concur that assumptions on the expected function, $\mathbb{E}[f(x,\xi)]$, are indeed more standard and weaker compared to those on the stochastic functions $f(x,\xi)$. However, transitioning to assumptions on $\mathbb{E}[f(x,\xi)]$ can be achieved in a standard manner. **While this change might lead to small adjustments to some constants in the complexity bounds, we assure that the dependency on both K and τ remain unaffected.** To convince the reviewers, we will highlight the key difference in the updated analysis for DSGD and DSPL.

We believe this suffices to convince the reviewers that these adjustments are primarily related to the constants in a few inequalities, leaving the core conclusions of our work intact. Furthermore, we promise that the revisions will detail the adjustments made in our proof to incorporate these changes.

Our revised assumptions will be:

Relaxed B1 (Bounded subgradient and weak convexity)

> $f(x)$ is λ-weakly convex and $\mathbb{E}_\xi[\|f'(x,\xi)\|^2] \le L_f^2$. ($f'(x,\xi) = g \in \partial f(x,\xi)$) for all $x \in \operatorname{dom} \omega$

Relaxed C1 (Bounded gradient)

> $c(x,\xi)$ is $C(\xi)$-smooth and $\mathbb{E}_\xi[\|\nabla c(x,\xi)\|^2] \le L_c^2, \mathbb{E}_\xi[\|C(\xi)\|^2] \le C^2$ for all $x \in \operatorname{dom} \omega$.

These assumptions are standard in stochastic optimization literature and is also used in [1, 2]. And we start to show how to modify our proof to accommodate this assumption.

Step 0. Properties of $f_y(x,\xi)$ for DSPL

Given **C1**, we know that

$$\begin{aligned} & |\mathbb{E}_\xi[f(x,\xi)] - f_y(x,\xi)| \\ = & |\mathbb{E}_\xi[h(c(x,\xi)) - h(c(y,\xi) + \langle \nabla c(y,\xi), x-y\rangle)]| \\ \le & L_h \mathbb{E}_\xi[|c(x,\xi) - c(y,\xi) - \langle \nabla c(y,\xi), x-y\rangle|] \\ \le & \frac{L_h \mathbb{E}_\xi[C(\xi)]}{2}\|x-y\|^2 \le \frac{L_h C}{2}\|x-y\|^2 \end{aligned} \tag{1}$$

by Jensen's inequality and at the same time, $f_y(x,\xi)$ is $L_f(\xi) = L_h \cdot \|\nabla c(x,\xi)\|$ Lipschitz-continuous.

Step 1. Bounding $\|x^{k+1} - x^k\|$ on expectation.

The first consequence of assuming **Relaxed B1** is that our original bound

$$\|x^{k+1} - x^k\| \le \frac{2(L_f + L_\omega)}{\gamma} \tag{2}$$

no longer holds uniformly, and now we instead show it on expectation. Recall that

$$x^{k+1} = \arg\min_x \left\{ \langle g^{k-\tau_k}, x - x^k \rangle + \omega(x) + \frac{\gamma}{2}\|x - x^k\|^2 \right\}$$

in our proximal update. And the by the three-point lemma, we have

$$\langle g^{k-\tau_k}, x^{k+1} - x^k \rangle + \omega(x^{k+1}) + \frac{\gamma}{2}\|x^{k+1} - x^k\|^2 \leq \omega(x^k) - \frac{\gamma}{2}\|x^{k+1} - x^k\|^2 \tag{3}$$

Re-arranging the terms and applying Cauchy-Schwartz inequality,

$$\gamma\|x^{k+1} - x^k\|^2 \leq (L_\omega + \|g^{k-\tau_k}\|)\|x^{k+1} - x^k\| \tag{4}$$

Dividing both sides by $\|x^{k+1} - x^k\|$,

$$\|x^{k+1} - x^k\| \leq \frac{1}{\gamma}(L_\omega + g^{k-\tau_k}) \tag{5}$$

Conditioned on the history information, taking expectations on both sides with respect to $\xi^{k-\tau_k}$, we have

$$\mathbb{E}_k[\|x^{k+1} - x^k\|] \leq \frac{1}{\gamma}\mathbb{E}_{\xi^{k-\tau_k}}[L_\omega + \|g^{k-\tau_k}\|] \leq \frac{L_f + L_\omega}{\gamma}, \tag{6}$$

where the last inequality is from Jensen's inequality $\mathbb{E}_\xi[\|g\|]^2 \leq \mathbb{E}_\xi[\|g\|^2] \leq L_f^2$. Now we get a counterpart of (1). A more tricky part is about $\mathbb{E}_k[\|x^{k+1} - x^k\|^2]$, and we note that

$$\begin{aligned}
\gamma\mathbb{E}_k[\|x^{k+1} - x^k\|^2] &\leq \mathbb{E}_k[(L_\omega + \|g^{k-\tau_k}\|)\|x^{k+1} - x^k\|] \\
&= \mathbb{E}_k[\|g^{k-\tau_k}\| \cdot \|x^{k+1} - x^k\|] + \mathbb{E}_k[L_\omega\|x^{k+1} - x^k\|] && (7) \\
&\leq \mathbb{E}_k[(2\gamma)^{-1}\|g^{k-\tau_k}\|^2] + \frac{\gamma}{2}\mathbb{E}_k[\|x^{k+1} - x^k\|^2] + \frac{2L_\omega(L_\omega + L_f)}{\gamma} && (8) \\
&\leq \frac{L_f^2 + 4L_\omega(L_f + L_\omega)}{2\gamma} + \frac{\gamma}{2}\mathbb{E}_k[\|x^{k+1} - x^k\|^2]
\end{aligned}$$

And the last inequality uses the inequality $2xy \leq ax^2 + \frac{y^2}{a}$ to decouple the correlated $\|g^{k-\tau_k}\|$ and x^{k+1}; and $\mathbb{E}_k[L_\omega\|x^{k+1} - x^k\|]$ is bounded using the previously established relation. Finally, we re-arrange the terms to cancel $\frac{\gamma}{2}\mathbb{E}_k[\|x^{k+1} - x^k\|^2]$ on both sides, giving

$$\mathbb{E}_k[\|x^{k+1} - x^k\|^2] \leq \frac{L_f^2 + 4L_\omega(L_f + L_\omega)}{\gamma^2} \tag{9}$$

Compared to our original result, the bound on $\mathbb{E}_k[\|x^{k+1} - x^k\|^2]$ becomes slightly worse in the enumerator. Besides, the above analysis applies to DSPL replacing $\langle g^{k-\tau_k}, x^{k+1} - x^k \rangle$ by $f_{x^{k-\tau_k}}(x, \xi^{k-\tau_k})$.

Step 2. Relaxing uniform assumption on the stochastic function

For DSGD, our assumption on uniform weak convexity appears on Line 561 in the appendix, where we have

$$\mathbb{E}_{\xi^{k-\tau_k}}[\langle g^{k-\tau_k}, \hat{x}^k - x^{k-\tau_k} \rangle] + f(x^{k-\tau_k}) - f(\hat{x}^k) \leq \frac{\lambda}{2}\|\hat{x}^k - x^{k-\tau_k}\|^2 \tag{10}$$

using λ-weak convexity. However, since $\mathbb{E}_{\xi^{k-\tau_k}}[g^{k-\tau_k}] \in \partial f(x^{k-\tau_k})$, we actually only need λ-weak convexity of f instead. For DSPL, the same result applies to the last inequality of line 649 after using (1). We can similarly modify Lemma 7 exploiting relation (6) and (9).

$$
\begin{aligned}
&|\mathbb{E}_{\xi'}[\mathbb{E}_{\xi}[f_z(\mathcal{A}(z,x,\xi'),\xi) - f_z(\mathcal{A}(z,x,\xi'),\xi')]]| \\
&= \left|\int_{\xi'\sim\Xi}\int_{\xi\sim\Xi} f_z(\mathcal{A}(z,x,\xi'),\xi) - f_z(\mathcal{A}(z,x,\xi'),\xi')d\mu_{\xi}d\mu_{\xi'}\right| \\
&= \left|\int_{\xi'\sim\Xi}\int_{\xi\sim\Xi} f_z(\mathcal{A}(z,x,\xi'),\xi)d\mu_{\xi}d\mu_{\xi'} - \int_{\xi'\sim\Xi} f_z(\mathcal{A}(z,x,\xi'),\xi')d\mu_{\xi'}\right| \\
&= \left|\int_{\xi'\sim\Xi}\int_{\xi\sim\Xi} f_z(\mathcal{A}(z,x,\xi'),\xi)d\mu_{\xi}d\mu_{\xi'} - \int_{\xi\sim\Xi} f_z(\mathcal{A}(z,x,\xi),\xi)d\mu_{\xi}\right| \\
&= \left|\int_{\xi'\sim\Xi}\int_{\xi\sim\Xi} f_z(\mathcal{A}(z,x,\xi'),\xi) - f_z(\mathcal{A}(z,x,\xi),\xi)d\mu_{\xi}d\mu_{\xi'}\right| \\
&\le \int_{\xi'\sim\Xi}\int_{\xi\sim\Xi} L_f(\xi)\|\mathcal{A}(z,x,\xi') - \mathcal{A}(z,x,\xi)\|d\mu_{\xi}d\mu_{\xi'} \\
&\le \int_{\xi'\sim\Xi}\int_{\xi\sim\Xi} \tfrac{1}{2\gamma}L_f(\xi)^2 + \tfrac{\gamma}{2}\|\mathcal{A}(z,x,\xi') - \mathcal{A}(z,x,\xi)\|^2 d\mu_{\xi}d\mu_{\xi'} \\
&\le \mathcal{O}(\tfrac{1}{\gamma}) \qquad (11)
\end{aligned}
$$

Now we are finally ready to go to the delay-relevant part.

Step 3. Putting things together

After the two steps above, we can smoothly go through, where our last piece of proof (Line 572, Line 662) needs a slight modification of constant in bounding $\sum_{k=1}^{K}\mathbb{E}[\|x^k - x^{k-\tau_k}\|^2]$.

$$
\begin{aligned}
\sum_{k=1}^{K}\mathbb{E}[\|x^k - x^{k-\tau_k}\|^2] &= \sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\sum_{l=1}^{\tau_k} x^{k+1-l} - x^{k-l}\Big\|^2\Big] \\
&\le \sum_{k=1}^{K}\tau_k\sum_{l=1}^{\tau_k}\mathbb{E}[\|x^{k+1-l} - x^{k-l}\|^2] \\
&\le \frac{L_f^2 + 4L_\omega(L_\omega + L_f)}{\gamma^2}\sum_{k=1}^{K}\tau_k^2,
\end{aligned}
$$

where the first inequality uses the inequality $\|\sum_{i=1}^{k}a_i\|^2 \le k\sum_{i=1}^{k}\|a_i\|^2$ and the second inequality applies our modified bound on $\mathbb{E}_k[\|x^{k+1} - x^k\|^2]$. The bound on $\sum_{k=1}^{K}\mathbb{E}[\|x^k - x^{k-\tau_k}\|]$ remains unchanged.

At this stage, **our analysis is fully compatible with the standard assumptions**, at the cost of a slightly worse constant in the bound. Our convergence rate and main results are unaffected.

Theorem (informal) Under the relaxed assumptions, the convergence rates of DSGD, DSPL and their robust version remain unchanged

We again thank all the reviewers for the efforts and for making valuable suggestions.

Reference

[1] Xu, Yangyang, et al. "Distributed stochastic inertial-accelerated methods with delayed derivatives for nonconvex problems." *SIAM Journal on Imaging Sciences* 15.2 (2022): 550-590.

[2] Davis, Damek, and Dmitriy Drusvyatskiy. "Stochastic model-based minimization of weakly convex functions." *SIAM Journal on Optimization* 29.1 (2019): 207-239.