We thank the reviewers for their time and helpful suggestions, which we will use to improve the paper's presentation. **R1, R2 (relevance of interpolation):** Please refer to lines 39-42 for examples where interpolation is satisfied. Recent work [5,7,42,46,77] views interpolation as key to understanding the effectiveness of SGD for deep learning. Moreover, we utilized this assumption to make algorithmic contributions that result in better empirical performance.

R2 (comparison with [5], [77]): They both use a constant step-size of $\frac{1}{L}$, which is either unknown or gives an overly-conservative, small step-size. Our initial experiments confirmed that it lead to worse empirical performance and we will mention this.

R2, R3 (wall-clock): For the line-search, we did ensure that the number of additional function evaluations is not large (Section 7). In the Fig. 1 below, we show the wall-clock time per iteration averaged across training for the three datasets.

R3 (error bars): The figures in Section 7.3 do have error bars, but they unfortunately look like spikes in the submitted version. We include one figure below with clearer error bars and will similarly update the remaining figures.

R3 (Fixing typo for Theorem 1): We correct the proof and statement of Theorem 1 below. Starting from the line justified by Equation 2 in Appendix B (recall that $\mu_{ik} = 0$ if the f_{ik} is not strongly-convex),

$$\mathbb{E}\left[\|w_{k+1} - w^*\|^2\right] \leq \left(1 - \mathbb{E}_{ik}\left[\mu_{ik}\ \min\left\{\frac{1}{L_{ik}}, \eta_{\text{max}}\right\}\right]\right)\|w_k - w^*\|^2$$

We consider the following two cases: $\eta_{\text{max}} < 1/L_{\text{max}}$ and $\eta_{\text{max}} \geq 1/L_{\text{max}}$. When $\eta_{\text{max}} < 1/L_{\text{max}}$,

$$\mathbb{E}\left[\|w_{k+1} - w^*\|^2\right] \leq (1 - \mathbb{E}_{ik}\left[\mu_{ik}\ \eta_{\text{max}}\right])\|w_k - w^*\|^2 = (1 - \bar{\mu}\ \eta_{\text{max}})\|w_k - w^*\|^2$$

When $\eta_{\text{max}} \geq 1/L_{\text{max}}$, we use $\min\left\{\frac{1}{L_{ik}}, \eta_{\text{max}}\right\} \geq \min\left\{\frac{1}{L_{\text{max}}}, \eta_{\text{max}}\right\}$ to obtain

$$\mathbb{E}\left[\|w_{k+1} - w^*\|^2\right] \leq \left(1 - \mathbb{E}_{ik}\left[\mu_{ik}\ \frac{1}{L_{\text{max}}}\right]\right)\|w_k - w^*\|^2 = \left(1 - \frac{\bar{\mu}}{L_{\text{max}}}\right)\|w_k - w^*\|^2.$$

Combining the two cases gives us the theorem statement with L_{max} instead of L. We will make this change in Theorem 1 statement. Note that Theorem 4's proof will be changed similarly.

R3 (Requested) rigorous proof for Theorem 3: We can prove an $O(1/T)$ rate by bounding $\eta_{\text{max}} \leq \frac{3}{2\rho L}$ as follows:

$$\frac{f(w_{k+1}) - f(w_k)}{\eta_k} \leq \frac{L\eta_k}{2}\|\nabla f_{ik}(w_k)\|^2 - \langle \nabla f(w_k), \nabla f_{ik}(w_k)\rangle \qquad \text{(Using smoothness and dividing by } \eta_k\text{)}$$

$$\implies \mathbb{E}\left[\frac{f(w_{k+1}) - f(w_k)}{\eta_k}\right] \leq \left(\frac{L\eta_{\text{max}}\rho}{2} - 1\right)\|\nabla f(w_k)\|^2 \qquad \text{(Since } \eta_k \leq \eta_{\text{max}} \text{ and using the SGC)}$$

$$\|\nabla f(w_k)\|^2 \leq \frac{1}{1 - \frac{L\eta_{\text{max}}\rho}{2}}\mathbb{E}\left[\frac{f(w_k) - f(w_{k+1})}{\eta_k}\right] \qquad \text{(Rearranging and upper-bounding } \eta_{\text{max}} \leq \tfrac{2}{L\rho}.\text{)}$$

$$\implies \|\nabla f(w_k)\|^2 \leq \left(\frac{1}{1 - \frac{L\eta_{\text{max}}\rho}{2}}\right)\left(\frac{1}{\eta_{\text{max}}} + \frac{L_{max}}{2(1-c)}\right)\mathbb{E}\left[f(w_k) - f(w_{k+1})\right].$$

(Bounding η_k using the line-search similar to Appendix C)

Telescoping and setting $c = 1/2$ and $\eta_{\text{max}} \leq \frac{3}{2\rho L}$ completes the proof. It is non-trivial to avoid the dependence of ρ,L in η_{max} and we leave it as future work. Regardless of this result, we believe that this paper's contributions are impactful.

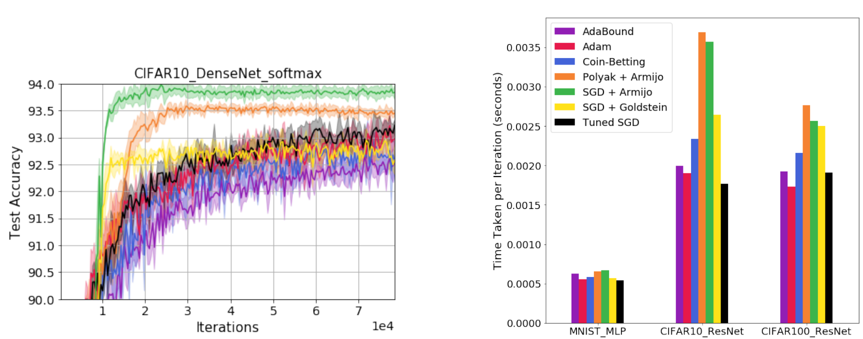

Figure 1: **Left**: CIFAR-10 with new error-bar style. **Right**: Average iteration times on CIFAR-10.