To **Reviewers**, we will make all suggested minor corrections in the final version and address main concerns below.

R1 (and **R3, R5:**) Thanks for your constructive comments! The idea of integrating several key optimization techniques, dual averaging + variance reduction + acceleration, is novel and nontrivial for three reasons. **First**, for the general convex setting, the proposed SVR-ADA gives the rate $O\big(n\log\log n+\frac{\sqrt{nL}}{\sqrt{\epsilon}}\big)$, which improves the SOTA rate $O\big(n\log\frac{1}{\epsilon}+\frac{\sqrt{nL}}{\sqrt{\epsilon}}\big)$ of Katyusha, not substantially improved for over 4 years; for the well-conditioned strongly convex setting (the case of the number of samples n far greater the condition number, *i.e.*, $n \gg \kappa$), SVR-ADA has the rate $O\big(n\log\log n+\frac{n}{\log(n/\kappa)}\log\frac{1}{n\epsilon}\big)$ which improves the SOTA rate $O(n\log\frac{1}{\epsilon})$ of SAG, unchanged for 8 years; for the ill-conditioned strongly convex setting ($n \leq O(\kappa)$), SVR-ADA matches the lower bound in [25]. **Second**, besides the improved or optimal convergence results, SVR-ADA shares the simplicity of MiG [29] and the unification of Varag (see comments of **R2**) *simultaneously*. **Third**, this work is the first to show that in the finite sum setting, combining accelerated dual averaging (ADA) with variance reduction (VR) gives better convergence rates than that of combining accelerated mirror descent (AMD) with VR as adopted in Katyusha, MiG and Varag. This provides new perspectives to acceleration.

In terms of experiments, SVR-ADA is compared with SOTA finite sum solvers. The results send one consistent message: *SVR-ADA performs well in all the three settings, namely general convex, ill-conditioned or well-conditioned strongly convex*. In comparison, the existing accelerated algorithms Katyusha and MiG do not perform so well for the well-conditioned strongly convex setting; the non-accelerated algorithm SVRG does not perform well for the general convex and ill-conditioned strongly convex settings. In terms of the choice of regularization, we use different two-norm regularizations to represent the phenomenon of the three settings compactly by Figure 1 and Figure 2. If we use one-norm, then it can only represent the general convex setting. For completeness, we will show the numerical results of one-norm in the final version. The different regularization parameters of the two figures are used to better represent the obvious change in experimental phenomena. In the final version, we will represent the results of $\{0, 10^{-8}, 10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}\}$ for both experiments at least in the supplementary material.

Thanks for your suggestions about writing! In the final version, we will rewrite the abstract to make it more clear. Per your suggestions, we will use "general convex" to replace "nonstrongly convex" and make the statements in Line 30 and 32 more precise. We will add a paragraph to make a simple introduction to stochastic algorithms that do not use VR.

R2: We are very glad that you obviously recognize significance of our contributions! We were also aware of Varag after our submission! In the final version, we will make systematical comparisons with Varag from three perspectives: 1) **Efficiency.** SVR-ADA improves the SOTA rates for general convex and well-conditioned strongly convex settings, while Varag does not improve any SOTA rates. 2) **Simplicity.** SVR-ADA only uses two-point coupling in the inner iteration, while Varag uses three-point coupling. SVR-ADA uses a uniform average in the outer iteration, while Varag uses a weighted one. 3) **Unification.** SVR-ADA unifies the general convex and strongly convex settings with much simplified parameter settings, which can be simply specified in algorithm description. However, to adapt to both settings, the parameter settings of Varag are very complicated and not explicitly and clearly stated in the algorithmic description.

In terms of techniques, **first**, we use the combination of VR and ADA, while they use the combination of VR and AMD; **second**, we use completely different and concise convergence analysis by estimation sequence; **third**, we only use negative momentum, while they use both Nesterov's momentum and negative momentum.

In terms of applicability in the constrained setting, we can always reformulate a convex constrained convex problem as an unconstrained convex problem by the indicator function of the constrained set. Thus if the indicator function admits an efficient proximal operator, then we can apply SVR-ADA in the constrained setting. Meanwhile, in the final version, we will give a complete proof for the bound of A_k. We will cite the latest reference Joulani et al. (ICML 2020)!

R3: Thanks for your constructive comments! We fully understand your concern with the applicability for nonconvex problems such as deep neural networks. However, it is an open problem for all control variate finite sum solvers, not only for ours. Meanwhile, the *primary area* of our contribution is in convex optimization, which is valuable for many problems in machine learning, signal processing, operational research. Thus we cannot undervalue the widely studied control variate approaches as they significantly reduce the overall complexity for finite sum convex optimization. It is fair to say broad applicability is the main reason why the control variate finite sum solvers are so widely studied!

In terms of **theoretical contributions**, as highlighted in our response to **R1**, we have improved rates in two regimes that remain unimproved for many years, in a very actively studied area. In terms of **empirical performance**, as in our response to **R1**, SVR-ADA is the first algorithm that performs well in all the three settings. We will conduct more experimental evaluation as per the request by you and **R1** and report in the final version.

R5: Thanks for your constructive comments! For the general convex setting, the best-known result $O(n\log\frac{1}{\epsilon}+\frac{\sqrt{nL}}{\sqrt{\epsilon}})$ is optimal up to a $\log$ factor. In convex optimization, a main endeavor is to shave off $\log$ factors to match the corresponding lower bounds. We have made a substantial step forward by reducing $\log n$ to $\log\log n$. Meanwhile, as in our response to **R1**, we also improve the rate for the well-conditioned strongly convex setting. As **R2** commented, the simplicity and unification merits of SVR-ADA are remarkable! We believe our work is of significant value to this area.