We appreciate the positive feedbacks from all the reviewers and provide a detailed response as follows.

R1: "Room for improvement in presenting numerical experiments ... it is better to include test error results"

Our initial intention was to conduct experiments reflecting the convergence rate in the theoretical analysis. We appreciate your suggestions, and will include the test error in the supplementary material.

R1/R2: "It would be interesting to include results using a variance reduction technique"

We thank both reviewers for mentioning variance reduction techniques, this is indeed very related. To apply variance reduction methods, we need to further assume that the loss function on each node can be decomposed into a finite-sum structure, say each loss is a sum of m functions. In this case, combining IDEAL+SVRG will provide a complexity of order $\tilde{O}\left((m+\kappa_f)\sqrt{\kappa_W}\log(1/\epsilon)\right)$, where acceleration is achieved with respect to the graph condition number κ_W but not with respect to objective's condition number κ_f (see also 'Two-fold acceleration' on page 6). It is conceivable that there might exist a better design for finite sum structure, e.g., by communicating the full gradient evaluated in SVRG, etc. This is definitely an interesting direction which we leave to future work.

R2: "I wonder whether the communication cost is lower than that of SSDA in practice. I suggest the authors to plot the communication cost and computation cost separately. "

When the computation/communication τ is large, the dominant term is the communication cost. Hence the third column ($\tau = 10$) in Figure 1 roughly reflects the communication cost. Thanks for the suggestion, we will include a separate plot in the supplementary material.

R2: "I wonder whether SSDA+AGD+warm start can achieve the same computation cost by using the proof technique proposed in this paper. "

We appreciate the reviewer raising this question, this is indeed one of the core messages we would like to convey: IDEAL improves upon SSDA in a non-trivial way. In short, we have applied the same warm-start strategy as IDEAL+AGD in the complexity analysis of SSDA+AGD, the higher computation cost of SSDA is due to an intrinsic weakness of the method. The high level intuition is that the regularization parameter ρ improves the condition number of the Moreau-envelope, which reduces the number of subproblems. More explicitly, the number of subproblems to be solved by IDEAL/SSDA is given by the formula

$$K = O\left(\sqrt{\frac{L_\rho}{\mu_\rho}}\log\left(\frac{C_\rho \Delta_{dual}}{\epsilon}\right)\right), \text{ where } L_\rho = \frac{\lambda_{\max}(W)}{\mu + \rho\lambda_{\max}(W)}, \mu_\rho = \frac{\lambda^+_{\min}(W)}{L + \rho\lambda^+_{\min}(W)}. \qquad \text{(Eq. 5 on page 5)}$$

Ignoring the log factor, this quantity is proportional to the regularized condition number $\sqrt{\frac{L_\rho}{\mu_\rho}}$.

- For SSDA (which is equivalent to $\rho = 0$), we have $\sqrt{L_\rho/\mu_\rho} = \sqrt{\kappa_f \kappa_W}$;
- for IDEAL, by choosing $\rho = \frac{L}{\lambda_{\max}(W)}$, we have $\sqrt{L_\rho/\mu_\rho} \le \sqrt{2\kappa_W}$.

Hence, IDEAL saves a factor of order $\sqrt{\kappa_f}$ compared to SSDA in the number of subproblems. Moreover, with the proposed choice of ρ, the cost of inexactly solving the subproblems is essentially the same for IDEAL and SSDA. Therefore we obtain the improvement in computation cost.

R2: "How to determine the regularization parameter, why choose $\rho = L/\lambda_{max}$"

The global complexity of IDEAL is given by the product of K (number of subproblems) and T (number of iterations for each subproblem), both of them are functions of ρ. The way we determine the regularization parameter is to minimize the global complexity with respect to the parameter ρ, then simplify it in an asymptotic manner.

R2: "References"

Thanks for pointing out these references, we will include them in the revision and some other recent works as well.

R3: "The contribution of this article is strong, but my main issue is I don't believe this is the best format for presenting these results... it would be very difficult for me to derive and verify the formulas in this table 2"

We are grateful to the reviewer's suggestion and will provide a more detailed explanation of this derivation in the revision. Please kindly check our response to R2's question starting in line 18 for a high level justification.

R3:" The main text of the article is not sufficient to understand the setup for these experiments"

All the experiments we conducted are for binary logistic regression. For the MNIST experiment, we directly use the normalized image as input feature. For the CIFAR experiment, the dataset is much more complicated, so a linear model is not rich enough. Hence, we first apply an unsupervised learning model (convolutional kernel network) to extract the feature and then apply the logistic regression on top of it. This can be approximately viewed as training the last layer of a conventional neural network by freezing the (well-trained) first layers. We will improve our presentation on experimental settings with more details in the revision.

R4: "Experiments are only conducted on one task. More tasks could be more compelling..."

Due to limited space, we have only presented the MNIST experiment in the main paper, however the CIFAR experiment and other ablation studies can be found in the supplementary material.