We thank all the reviewers for their comments and feedback which do help us improve the quality of our paper. We explain how we address your concerns and revise our paper based on your comments. Based on **R1** and **R3** concerns about the code, we are happy to share it. If they wish to see the code right away, we can share the code through AC/PC.

Reviewer 1: *"There are relatively few stable facts. This paper does not necessarily reduce the entropy."* The reviewer raised a very important point. We agree that there are tremendous amount of papers on adaptive gradient based optimization with few stable facts. Most variants are proposed by tweaking parameters. We expect our work to bring new insights to this field, especially in understanding the generalization through the lens of differential privacy.

"plots in Figure 2" We thank the reviewer for pointing this out. We will improve the quality of the plots in the revision.

"broader impact" This paragraph will be improved to reflect the idea of avoiding over-fitting (see plot (b) Figure 2).

Reviewer 2: *"I would have liked to see more thorough and rigorous experiments." "both ResNet18 and VGG19 should be reaching slightly higher test accuracies with SGD/Adam"* We mainly follow the method in [Wilson et al., 2017] to tune the step size, since they highlight that the initial step size and the scheme of decaying step sizes have a considerable impact. We agree that the mini-batch size would also play a important role in the performance of the training algorithms. Still, we think that our experiments have provided an extensive experimental evaluation of variants of training algorithms for tasks such as image classification and language modeling task. We believe our experiments offered a fair comparison among the baselines since the same effort is done to tune the hyper-parameters (step size).

"does RMSProp offer any particular advantage..." We agree that DPG-LAG/DPG-SPARSE can be used with any first order optimization algorithm. The RMSProp can be viewed as SGD when $\beta_2 = 1$. In the Appendix, we plan to provide a generic stable adaptive algorithm that encapsulates many popular adaptive and *non-adaptive* methods.

"How do the high probability bounds change when using mini-batches of size m?" The high probability bounds on the gradient mainly follow the generalization guarantee of differential privacy, which shows that an (ϵ, δ)-algorithm can guarantee a certain generalization error if (ϵ, δ) and sample size n used to evaluate the gradient satisfy some conditions (Lemma 1). In the case of mini-batch, the sample size becomes m and the value of (ϵ, δ) is modified. Thus, the sample complexity for the high probability bound changes. We have provided details in the proof of Theorem 5.

"Is data augmentation used in the experiments?" We used data augmentation for MNIST and CIFAR-10. For MNNIST, we normalized the value of each feature to [0,1]. For CIFAR-10, we normalized and rotated the images, using standard functions such as transforms.RandomCrop, transforms.RandomHorizontalFlip, and transforms.Normalize.

Reviewer 3: *"It is unclear how guaranteeing stationary points that have small gradient norms translates to good generalization"* Our main theoretical results provide the convergence to the *population stationary point*. Note that Theorem 2, 4 and 5 show the convergence of the norm of the *population gradient* instead of the empirical gradient. Specifically, while SAGD only has access to n samples, it converges to the *population stationary point*. Also, based on PL condition, one will be able to establish the generalization error of the value function.

"the Hoeffding's bound holds true as long as the samples are drawn independently". Yes, Hoeffding's bound holds as long as the samples are drawn independently. However, in the setting of optimization with *sample reuse* (setting in this paper), the reused samples are not independent anymore for any iteration $t > 0$. This is because the posterior distributions of samples change after training on the finite set of n observations.

"The bounds in Theorem 1 have a dependence on d". The reviewer raised a very interesting question. Yes, the dependence on d is a known result for differential privacy (DP) and is hard to avoid (see ref. [1]). Some works on DP try to improve this dependence on d by leveraging special structures of the gradients. This will be considered in the future.

"do not depend on the initialization $\mathbf{w}_0$ *but on* $\mathbf{w}_1$.*"* We thank the reviewer for this typo: should be $\mathbf{w}_0$ instead of $\mathbf{w}_1$.

"For Penn-Tree bank,[...] algorithms are not stable w.r.t. train perplexity." With respect to train perplexity, all methods stabilize around a target value (which is of course different given the highly nonconvex loss). We note that the test perplexity increases after several epochs for most baselines while our method keeps a low and steady one.

Reviewer 4: *"experiment design mainly follows [Wilson et al., 2017]"* The design is different from [Wilson et al., 2017] (except for the stepsize tuning, see **Reviewer 2**). Indeed, we study the *generalization* performance of each algorithm with an *increasing* training sample size n (see Fig. 1, x-axis is n). This is consistent with our theoretical results which show the convergence of SAGD in terms of n and compare the performance of those algorithms when n is small. However, [Wilson et al., 2017] mainly plotted the training/test accuracy against the the number of epochs elapsed. We agree that it would be interesting to add experiments to compare RMSProp with differential privacy.

"SGD with gradient corrupted by Gaussian noise performs well or not" Excellent question! Actually, one can also use Gaussian noise to design a differentially private algorithm (namely Gaussian Mechanism [7]). There are papers showing the connection between SGLD (Stochastic Gradient Langevin Dynamics) and differential privacy. Yet, the existing generalization bound of SGLD is established by the techniques of algorithmic stability [23, 26], which scales with $(\sqrt{T})$. We believe it is of great interest to provide a theoretical analysis of SGLD via the generalization of differential privacy. It is also interesting to show how Gaussian noise works in our setting. We will add a discussion in the paper. We consider the theoretical details and experimental results as a future work.

"whether the proposed method works well for small datasets in terms of generalization" Figure 1 shows that SAGD has a slightly better test accuracy than other algorithms when the training sample size n is small (x-axis).