We thank the reviewers for their positive and constructive feedback. We address several points in the review below.

Connecting Section 5 to DP-SGD. The bias reduction technique in Section 5 is designed for DP-SGD with clipping. When it is applied to DP-SGD, the update rule is shown below.

$$x_{t+1} = x_t - \alpha\left(\left(\frac{1}{|S_t|}\sum_{i\in S_t}\text{clip}(\nabla f(x_t) + \xi_{t,i} + k\zeta_{t,i}, c)\right) + Z_t\right),$$

where $\zeta_{t,i} \sim \mathcal{N}(0, I)$ is the noise added to reduce clipping bias and $Z_t \sim \mathcal{N}(0, \sigma^2 I)$ is the noise added for privacy (see (8) in the paper for other notations). The privacy guarantee of this algorithm is the same as original DP-SGD since the noise added before clipping does not scale with the corresponding L_2 sensitivity and does not provide formal privacy guarantee. We also compared original DP-SGD and the above algorithm with $\sigma = 1$ in the 3 experiment settings in Section 5, the results are similar to thoise in the figure in Section 5 (which used $\sigma = 0$ for all algorithms). We will include the experiments for $\sigma = 1$ in the paper.

Typos: Thank you for pointing them out, we will correct the typos.

Reviewer 1. *Q1: Discussion on the convergence rate of DP-SGD.* This is indeed a very important point to discuss in the paper. DP-SGD achieves convergence rate of $O(\sqrt{d}/(n\epsilon))$ in the existing literature. As shown in Theorem 5, with gradient clipping, the rate becomes $O(\sqrt{d}/(n\epsilon)) +$ clipping bias. When gradient distribution is symmetric, the convergence rate can recover to $O(\sqrt{d}/(n\epsilon))$. We will add a more detailed discussion around Theorem 5.

Q2: Better connect the clipping bias reduction method with DP-SGD. Please see "Connecting Section 5 to DP-SGD".

Q3: Additional feedback. Thank you for the suggestions, we will revise accordingly.

Reviewer 2. *Q1: Experiments to support the Theorems in Section 2.* We will add such experiments in the paper. Theorem 1 is a standard convergence bound for optimization; we will verify it on MNIST. Theorem 2 is our new contribution, we conducted some new experiments for it. We considered a 1-dimensional example and choose a symmetric noise $\xi_t \sim \mathcal{N}(0, 1)$ and set $c = 1$. Then we compare $\mathbb{E}_{\xi_t\sim\tilde{p}}[\langle\nabla f(x_t), g_t\rangle]$ (estimated by averaging 10^5 samples) with the lower bound in Theorem 2 for different $\|\nabla f(x_t)\|$. The results below verify our lower bound.

$\|\nabla f(x_t)\|$	0.05	0.1	1	2	10	100
$\mathbb{E}_{\xi_t\sim\tilde{p}}[\langle\nabla f(x_t), g_t\rangle]$	1.7e-4	6.6e-3	0.612	1.83	10	100
lower bound	4e-5	2e-3	0.148	0.3	1.48	14.8

Reviewer 3. *Q1: Interpretation and verification about $\tilde{p}$ in practice.* We will provide a few examples to illustrate good choices of $\tilde{p}$ and add more discussion on what is the desired property for $\tilde{p}$. Essentially, we want to choose $\tilde{p}$ such that the Wasserstein distance term in Theorem 5 is small and the probability term in Theorem 5 is large. For verification of symmetricity, we found the random projection method used in Figure 1 and 2 to be quite effective in practice. It is interesting to investigate better ways to measure symmetricity for very high-dimensional distributions.

Q2: The symmetric distribution may not hold for neural nets such as LSTM. This is an interesting point. In general, it will be very interesting to investigate the geometric structure among the gradients for different types of neural network, even absent privacy or clipping concerns. We will put this in our future work discussion.

Reviewer 4. *Q1: Unbounded bias terms make the bounds less useful.* The bias terms in the bounds are bounded by constants in worst case, and could be 0 for symmetric distributions. For example, in Theorem 5, the clipping bias is given by the Wasserstein distance term. For iteration t, $W_{\nabla f(x_t),c}(\tilde{p}_t, p_t) \le 2\|\nabla f(x_t)\|c$ always holds. More importantly, $W_{\nabla f(x_t),c}(\tilde{p}_t, p_t) = 0$ if $\tilde{p}_t = p_t$ which can happen when p_t (gradient noise distribution) is symmetric.

Q2: Application of the technique in Section 5 to DP-clipping. Please see Section "Connecting section 5 to DP-SGD".

Q3: How different the conclusions of this paper are from previous work. As noticed by the reviewer, the use of symmetric noise distributions in analysis for clipping is new. The key conclusion of our paper is the 0 (small) clipping bias for symmetric (approximate symmetric) gradient noise distributions. We also proposed a trick to symmetrify gradient noise distributions to reduce clipping bias. The most related previous works are Pichapati et al. [2019] and Zhang et al. [2019]. Both works give bounds on clipping bias essentially similar to $c||\nabla f(x_t)||\mathbb{P}$[gradient is clipped] which could be large when the clipping probability is large. Compare with these works, we give sharper bounds for (approximate) symmetric gradient noise distributions. Another closely related work is [1], which was uploaded to arxiv after the submission deadline. The work proves that clipping can lead to constant regret for DP-SGD, like in Example 1 in our paper. We will add a detailed discussion on comparison with these works.

Q4: Smoothness assumption limits the applicability, symmetricity assumption may not hold for RNNs, show least symmetric projections. Smooth assumption is commonly used for DP-SGD in literature and it holds for smooth approximations for relu such as softplus and swish. Extending the analysis of DP-SGD to non-smooth problems is important and interesting even absent clipping. The gradient distribution for RNNs is also a very interesting point to explore. We will mention these in our future work discussion. We will add figures for projections with least symmetry.

[1] Shuang Song, Om Thakkar, Abhradeep Thakurta. Characterizing private clipped gradient descent on convex generalized linear problems. *arXiv preprint arXiv:2006.06783, 2020.*