We thank the reviewers for their constructive feedback! We have detailed our response below and will improve the manuscript accordingly. We are pleased that reviewers appreciate our approach ‘interesting’ (**R#5**) , ‘not only improves the accuracy but also simplifies the usage ’(**R#5**), ‘improves baseline significantly in all metrics’ (**R#2**) and ‘gives many applications’ (**R#2**), and that the experiments are ‘convincing’ (**R#3**), ‘extensive’ (**R#2**,**R#4**), ‘well organized’ (**R#4**) and ‘answers every question brought in the methods section’ (**R#4**).

**[R#2] ["bounding" the norm empirically rather than theoretical? Line 294 uses the word "estimate" which has a technical/statistical meaning]** It is theoretical: norm of $\boldsymbol{g}_G^{\text{upstream}}$ is 1 almost everywhere (Line 151-155). We will further clarify. The proof can be found in [17] (Proposition 1 and Corollary 1, where $\boldsymbol{g}_G^{\text{upstream}}$ corresponds to $\nabla f^*$).

**[Would C=1 change for multiple layers? new datasets, architectures, etc.?]** No (Line 136-139). Note that we only need to bound $\|\boldsymbol{g}_G^{\text{upstream}}\|_2$. $C = 1$ follows from the theoretical property of Wasserstein objective and is independent on all other factors including architectures and datasets.

**[Other works using WGAN used other C values ... need to tune C?]** Our approach does not require tuning $C$. In previous DP-SGD GAN framework, the discriminator parameter gradients $\nabla_{\boldsymbol{\theta}_D}\mathcal{L}_D(\boldsymbol{\theta}_D)$ are sanitized, instead of $\boldsymbol{g}_G^{\text{upstream}}$. In comparison, $\nabla_{\boldsymbol{\theta}_D}\mathcal{L}_D(\boldsymbol{\theta}_D)$ do not have any bounded norm guarantee (empirically the gradient norms exhibit a heavy-tailed distribution with high variance as shown in Figure 2) and previous works need to tune $C$ carefully.

**[Not sure of the connection from Wasserstein section to the theorem]** The DP privacy analysis relies on a sensitivity value that depends on $C$. We show that by selectively applying sanitization to a necessary and sufficient subset of gradients for preserving privacy, we are able to exploit the theoretical property of WGAN and bound the sensitivity value without extensive tuning of $C$, while introducing almost no clipping bias.

**[Is Eq (11) an identical optimization-step rule for the generator as standard DP-GAN?]** No. As mentioned above, the standard DP-SGD GAN sanitize different gradients as done in our framework, and their DP sanitization process can *not* be modelled by Eq (11) ( $\boldsymbol{g}_G^{\text{upstream}}$ does depend on $\nabla_{\boldsymbol{\theta}_D}\mathcal{L}_D(\boldsymbol{\theta}_D)$ but the dependence is represented by the discriminator network, which can not be explicitly formularized).

**[The empirical improvement comes primarily in better and more complex discriminator?]** We attribute the improvements to a combination of complex discriminator and our approach which allows optimizing discriminator’s parameters *without* clipping its gradients and consequently leads to a better trained discriminator. In contrast, standard DPGANs introduce much larger clipping bias than our method, which also explain our superior performance.

**[Whether both gradient-sanitization and Wasserstein GAN or just gradient-sanitization is the contribution]** In comparison to prior works on DPGAN, our proposed gradient sanitization scheme enable us to exploit the theoretical property of WGAN. To the best of our knowledge, we are the first that consider the intrinsic property of WGAN and propose a framework that can use this property for obtaining a theoretically grounded choice of $C$.

**[Line 34: note that DP can work for a large network for large data]** We thank for the reference and will rephrase.

**[Figure 4(c): why epsilon<2 is not in the Figure?]** We will make it consistent. Note that numerical issues arise from the required noise scale for small epsilon values.

**[Release of source code]** As stated in Line 312-313 and Appendix Line 1-3, we will submit our source code together with our final version and also publish our code and setups at the time of publication.

**[R#3] [Experiments: what if using other GAN architectures]** We conduct additional experiments and show that our method with a basic DCGAN structure still yields consistent improvement over prior works, as shown in Table 1 (“Previous Best” denotes the best score achieved by previous methods; The new experiments are repeated twice and the average is reported). Moreover, we would like to emphasize that the usage of large network is prohibited by previous methods due to heavy hyperparameter search and a large clipping bias, which are well addressed by our framework.

| | | IS↑ | FID ↓ | MLP Acc↑ | CNN Acc↑ | Avg Acc↑ | Calibrated Acc↑ |
|---|---|---|---|---|---|---|---|
| MNIST | Previous Best | 4.76 | 161.11 | 0.63 | 0.68 | 0.57 | 66% |
| | Ours (DCGAN) | **8.74** | **75.83** | **0.79** | **0.79** | **0.59** | **68%** |
| Fashion-MNIST | Previous Best | 3.68 | 205.78 | 0.56 | 0.62 | **0.51** | **65%** |
| | Ours (DCGAN) | **5.59** | **134.74** | **0.67** | **0.66** | **0.51** | **65%** |

Table 1: Quantitative Results on MNIST and Fashion-MNIST ($\varepsilon = 10, \delta = 10^{-5}$).

**[Meaning of figure titles ]** The title named ‘noise scale’ means that we control the $\varepsilon$ by only changing the noise scale while keeping all other factors fixed to be their default values. As stated in Line 243-246, we indeed consider the different factors that affect $\varepsilon$ and investigate all of them .

**[R#4] [Bigger picture of differences between proposed approach and existing works]** We will enrich our discussion and include a high-level comparison between our proposed method and the existing works.

**[Background section contains only definitions ... add more connections and intuitions]** We will improve the background section to better equip readers with intuitions required to understand our approach.

**[R#5] [Using more complicated datasets, such as on nature images]** It would be ideal to evaluate on complex high-dimensional data. However, due to the data and model complexity, it is currently challenging to obtain reasonable performances for low privacy costs ($\varepsilon \leq 10$). Consequently, we use the same datasets and evaluation methodology as prior works, in which we show substantial improvements and achieve state-of-the-art performances.