We would like to thank the reviewers for their effort in reviewing and providing feedback during these hectic times.

Reviewer 1

Q: Missing Carlini et al. USENIX? **A**: We omitted this accidentally, but will definitely reference this in our revised version. Carlini et al. demonstrate privacy risks on models trained with standard SGD. Their attacks do not hold even with very weak differential privacy guarantees. Our work, by contrast, designs attacks that are effective against DP-SGD, and also shows how to use these attacks to better understand privacy guarantees offered by DP-SGD.

Q: Understanding of variance overfits to logistic regression **A**: Our attack works on multiple models, not only logistic regression. In fact, we also evaluate our attack using two-layer neural networks, and the performance is similar. See Figure 2 (d), (e), (f), and Tables 1 and 3.

Q: What does it mean for y_p to be the smallest probability class on x_p? **A**: The class which the model predicts with the smallest probability. If the model's class predictions are [0.4, 0.1, 0.5] (for classes 0, 1, and 2, respectively) then y_p would be class 1. We can update the paper to clarify this.

Q: Why does $\varepsilon = \infty$ use clipping? **A**: For consistency, we use clipping for all our experiments. Even with clipping, setting the noise $\sigma = 0$ would still not give ε-DP for any finite ε. Additionally, Secret Sharer (Carlini et al., USENIX) demonstrates the privacy risks of training without clipping/noise, and shows that gradient clipping already prevents their identified privacy leakage. Our results establishing a privacy risk when gradient clipping is used are new.

Q: Fit to security venue? **A**: It's true that many papers studying privacy attacks appear in security venues, but we believe our work is also solidly in scope for NeurIPS because it addresses questions that practitioners of machine learning have to answer: which learning algorithm should they choose and how to set its privacy level in order to achieve best trade-offs between privacy and accuracy.

Q: Why use multiple poisoning points and not the least k singular vectors? **A**: Using different singular vectors as poisoning would impact the model in uncorrelated ways. Using k copies of the same data point causes the gradients of each attack point to be correlated, making the poisoned and unpoisoned models more distinguishable.

Q: Why is membership inference not monotone? **A**: Despite averaging results over 10 trials for membership inference, we still observe some variance. Note that the inferred lower bounds are very small for membership inference and we are plotting the results in log scale, which might amplify the variance.

Reviewer 2

Q: Focus on SGD **A**: SGD (and other stochastic first-order methods) are among the most widely used methods for private ML, and one of the few with high quality public implementations, so we felt this was the most important single algorithm to study. However, in the supplementary material we also show that a suitable variant of our attack is highly effective against the output perturbation algorithm for private linear regression (Chaudhuri et al. JMLR '11). More generally, our underlying approach is applicable for any class of models for which poisoning attacks are known, which currently includes linear models (logistic regression, SVM) and non-linear models (decision trees, neural networks).

Q: Scale to bigger datasets/more complicated architectures? **A**: Our poisoning attack works for any fixed-dimensionality input. Our approach does still require training multiple models, which can be expensive as architectures get larger. Tuning the confidence parameter and number of poisoning points make it possible to observe privacy leakage with fewer training runs. The architectures we consider are state-of-the-art for DP training.

Q: Work on transformers/RNNs? **A**: RNN-specific techniques are required for training DP-RNNs, and so our attacks need to be modified accordingly. To our knowledge, DP transformers have not been considered in the literature.

Reviewer 3

Q: Multiclass problems? **A**: Our methods handle multi-class problems. In fact, we tested our techniques on all 100 classes of Purchase-100, and achieve a large lower bound.

Q: Worst case dataset? **A**: For output perturbation, our theoretical analysis suggests that datasets with small variance directions will allow for better lower bounds than more spherically distributed datasets. Finding the worst-case dataset is an interesting problem for future research in this space.

Q: Could impact of clipping norm be attack specific? **A**: We know that with clipping norm 0, the algorithm satisfies 0-DP, and with clipping norm ∞, the algorithm is wildly non-private (Carlini et al., USENIX). So it is reasonable to expect that the clipping norm affects the true privacy level in less extreme parameter regimes, although we do not have a rigorous justification for it. We note that our work is the first to identify the clipping norm as a relevant parameter, and we expect this issue will attract future study regardless of what the answer ultimately proves to be.