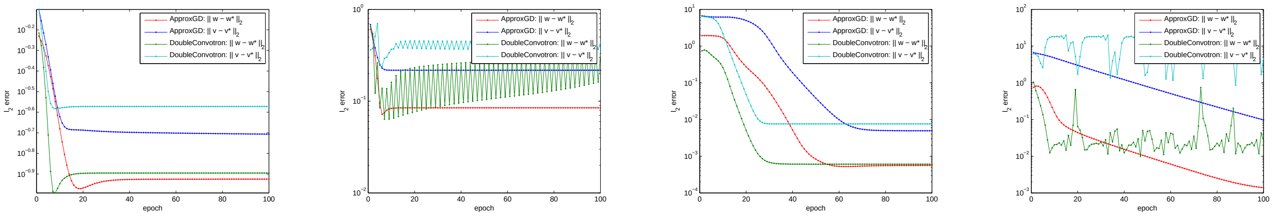
To all reviewers: *on the limitation of problem setting: Gaussian inputs and non-overlapping filters*
Our current result is for Gaussian inputs, and single hidden layer non-overlapping CNNs. However, this is also the case for most existing results, and even in this very restrictive setting, existing work falls short of tight sample complexity and generality in terms of activation functions. Therefore the problem we studied in this paper is still largely unsolved before our work. As is shown in Table 1 in the paper, our result has its unique strength and outperforms the state-of-the-art results in many aspects. We believe our work is a cornerstone of this line of research, and serves as a foundation towards more practical results.

To Reviewer1:

- *"it would be much better to write down the proof of the claim in line 209... are required. "*
  Thanks for your suggestion. We will add the derivation in line 209. It is mainly direct calculation of expectations, and therefore your suggestion can be easily addressed.
- *"the proof indicates that η_w and η_v are both less than 1...", "how to get line 518 and line 528 from assumptions", "where and how the assumptions in Theorem 4.3 are used and to explain what the conditions mean."*
  Thank you for pointing out these issues. We will revise the statement in Theorem 4.3 to make the conditions $\eta_w, \eta_v < 1$ clear. We apologize that we missed the assumptions in line 518 and line 528 when merging and simplifying the conditions on n. We will revise the conditions, clarify all the derivations, and discuss their high level implications in the revision. Here we wish to emphasize that such revision will not affect the overall validity of our analysis. Except adding several lines of detailed derivations, most parts of our proof will remain unchanged.
- *"Can this algorithm and the analysis be extended to deeper CNN or more layers?"*
  Our current analysis is specific to two-layer CNNs, and we have some preliminary idea to extend it to three-layer CNNs. However, it is not trivial to extend it to deeper CNNs with arbitrary number of layers.

To Reviewer2:

- *"plots similar to Figure 1 for non-Gaussian data, or even non-Isotropic Gaussian data."*
  Thanks for your suggestion. Here we present experiments with two types of input distributions: uniform distribution over unit sphere and a transelliptical distribution (the distribution of a Gaussian random vector after an entry-wise monotonic transform $y = x^3$). The experiments are conducted in the setting $k = 15$, $r = 5$ for ReLU and hyperbolic tangent activation functions, where $\mathbf{w}^*$ and $\mathbf{v}^*$ are generated in the same way as Line 253 in our paper. Specifically, Figures 1(a), 1(b) show the results for uniform distribution over unit sphere, while Figures 1(c), 1(d) are for the transelliptical distribution. Moreover, Figures 1(a), 1(c) are for ReLU networks, and the results for hyperbolic tangent networks are given in Figures 1(b), 1(d). From these figures, we can see that although it is not directly covered in our theoretical results, the approximate gradient descent algorithm proposed in our paper is still capable of handling non-Gaussian distributions. In specific, from Figures 1(a), 1(c), we can see that our proposed algorithm is competitive with Double Convotron for symmetric distributions and ReLU activation, which is the specific setting Double Convotron is designed for. Moreover, Figures 1(b), 1(d) clearly show that for hyperbolic tangent activation function, Double Convotron fails to converge, while our approximated gradient descent still converges linearly.



(a) ReLU+Unif. Sphere (b) tanh+Unif. Sphere (c) ReLU+Transelliptical (d) tanh+Transelliptical

Figure 1: Experimental results for non-Gaussian input distributions.

To Reviewer3:

- *"... works in the ideal setting (as the authors demonstrate during the experimental results, where the unit sphere and standard gaussian vector is used)... "*
  We would like to clarify that $\|\mathbf{w}^*\|_2 = 1$ is a reasonable and necessary assumption to ensure generality of our theory. Note that our theoretical result covers positive homogeneous activation functions like ReLU, leaky ReLU or linear activation. Without the assumption that $\|\mathbf{w}^*\|_2 = 1$, the true parameters for such networks are *not identifiable*–for any $c > 0$, the parameters $c \cdot \mathbf{w}^*$ and $c^{-1} \cdot \mathbf{v}^*$ give exactly the same network, and therefore recovering $\mathbf{w}^*$ and $\mathbf{v}^*$ is impossible. Therefore this assumption essentially specifies a particular set of parameters among a class of equivalent parameters, and is still a reasonable and necessary assumption even for real setting. We would also like to stress that, although in our experimental results we generate $\mathbf{v}^*$ as a standard Gaussian vector, this is not essential at all, and other distributions of $\mathbf{v}^*$, or even manually chosen values of $\mathbf{v}^*$, can always be recovered up to statistical accuracy by our algorithm. We will add these experimental results in the revision.