We would like to thank all the reviewers for their detailed and enlightening reviews, especially during these challenging times.

Response to Reviewer #1

Limited experimental verification. Thank you for pointing this out. When we were preparing this manuscript, we were having a hard time deciding how many empirical evaluations to include. After some thoughts, we finally decided to only include the current Fig. 1 for the following reasons: (1) since this paper is primarily on the theoretical side and in view of the 8-page limit, we prefer to use the limited spaces to explain the intuitions behind our technical results; (2) the empirical success of data augmentation is already established in many other works, so we feel that presenting a comprehensive empirical study may be of secondary importance in this paper.

We will do our best to do the Fashion-MNIST experiment with left/right flip. We need to figure out how to estimate the variance reduction term (see next answer), but we will think about this and do our best to address it, at least in the special case of the left/right flip.

Estimating the variance reduction term. Good point! This is indeed a super interesting direction and we have been thinking about it from the very beginning of this project. The major difficulty along this direction is that, to estimate the asymptotic variance reduction (say, based on Eq. 10) requires knowledge about the ground truth parameter θ_0, which may be hard to obtain. Some ideas we have in mind include: (1) replacing θ_0 with the trained weights, but this requires training and violates the original goal of "judge the quality of an augmentation without training"; (2) replacing θ_0 with a random initialization, which may be accurate in the neural tangent kernel regime when the network is very wide. Another related idea is to start with a "candidate augmentation" g, and estimate $D((gX, Y), (X, Y))$ from the data for some distance measure D between probability distributions. Then the estimated $\hat{D}((gX, Y), (X, Y))$ can be taken as an estimate of "how invariant our data are w.r.t. g". More concretely, for example, we can sample $\{(x_i, y_i)\}_{i=1}^m$ from our training data, apply g to each of them, and calculate the Wasserstein distance between the two empirical distributions (which can be solved by a linear program). Somewhat related to the above idea, we may adopt a hypothesis testing framework and try to test $H_0 : D((gX, Y), (X, Y)) \le \varepsilon$ v.s. $H_1 : D((gX, Y), (X, Y)) > \varepsilon$. We have not experimented with these ideas, but we believe these are interesting future directions, and we plan to explore them further in the future.

Miscellaneous questions. Sorry for the confusion on θ_0. And the statement from line 148-149 indeed lacks a minus sign. We will correct these two issues in the final version of this paper.

Response to Reviewer #2

Beyond group transformations. Thank you for bringing this up. Indeed, the group structure is lacking in many "real-world" transformations used by practitioners. The reason that we work with a compact group G is because we can endow it with a Haar probability measure, so we do not have measure-theoretic complications when computing averages over the group. However, we would like to point out that all of our current results would hold if G is only a semi-group (i.e., a set of transformations, not necessarily invertible), provided we can endow it with a uniform probability measure (which holds, for example, when G is discrete). We will include a discussion on this point in the final version of this paper.

Defective orbits. It would be very interesting to characterize, in a quantitative fashion, the performance loss induced by defective orbits compared to the full orbits. We will explore along this direction in future works.

Response to Reviewer #3

Abrupt halt at the end of Sec. 4.1. Thank you for the concrete suggestion (to trim some of the section on the circular shift model in favor of having more space for a final discussion section) and sorry for the abrupt halt, which is a compromise we made in view of the 8-page limit. We will include a discussion section in the final version of this paper, also incorporating some of the feedback from Reviewer #2.

Clarification on Fig. 1b. Sorry for the ambiguity of the purpose of Fig. 1b. We will at least explicitly define relative efficiency and provide some intuition for it in the caption of Fig. 1b in the final version of this paper to better explain the purpose of this plot. The suggestion of training a two-layer net with Gaussian inputs seems very interesting, and we will do our best to include it in the final version of this manuscript (perhaps replacing the current Fig. 1b).

Other quantitative works on data augmentation. Sorry, this is again a compromise of the 8-page limit. We will comment more on other quantitative works in the final version of this paper.