We thank all reviewers for their comments. Below, we respond to the questions raised by each reviewer.

Reviewer #1

Performance measure: for DGAN trained on CIFAR 10, we reported the widely used Inception score (IS) and Fréchet Inception distance (FID) (line 276 - 279), and our results are close to those of unconditional DCGANs. For EDGAN on the toy example, we reported the likelihood of generated samples under the true model and vice-versa, which is a more straightforward performance measure for generative models. Our EDGAN outperformed AdaGan on the two likelihood-based measures.

For EDGAN on real-world examples, though, we cannot directly compute the likelihood, and it is unclear how to evaluate the learned mixture weights. We will also report IS and FID scores in the final version. In fact, depending on the choice of embedding networks, IS and FID scores can be very favorable. However, we do not believe that either IS or FID fully captures the key objective of mixing GANs. We like the suggestion of the reviewer to report performance results for a subsequent task based on generated data and will do so. We expect our DGAN algorithms to achieve a high performance according to this metric since the discrepancy measure used for their training is precisely meant to capture that. Note, however, if we only reported such results, others would argue that we had to report IS scores or discrepancies, since that is what we optimize.

Selection of the hypothesis set: with a fixed embedding, the hypothesis set $\mathcal{H}$ used in experiments is always the set of linear mappings with bounded ℓ_2 norm, thus there is no further "selection of hypothesis set". Even though there are infinitely many linear mappings in $\mathcal{H}$, discrepancy takes a supremum over them (see Eq. 1) and thus is well defined. Proposition 3 actually gives a closed-form expression for the discrepancy when $\mathcal{H}$ is the set of linear mappings. We adopted linear functions and the squared loss in our experiments since the discrepancy admits a closed-form solution in that case. However, the learner could choose other hypothesis sets $\mathcal{H}$ and loss functions ℓ relevant to the learning tasks and our theoretical results would still apply.

Reviewer #2

DGAN experiments: we will include more experimental results for DGAN on more datasets and report their IS and FID scores.

Other ensemble methods: the ICLR'19 work is interesting and we could potentially use discrepancy instead of the Wasserstein distance there to come up with new ensemble algorithms.

Reviewer #3

Compare with McGan: indeed, in the specific case where $\mathcal{H}$ is the family of bounded norm linear functions and the loss is the squared loss, DGAN coincides with one of the objectives sought by McGan, that of matching the empirical covariance matrices of the two distributions, however the norm used in McGan is different (nuclear norm). Nevertheless, our theoretical analysis can serve as a justification of McGan in that case. We thank the reviewer for the connection and will reference the paper. In general though, for other hypothesis sets and loss functions, the objectives and techniques are distinct.

Fair comparison: in the DGAN experiments, we used the DCGAN architecture and trained our own embedding layer (line 258 - 261). We used pre-trained embedding only in EDGAN experiments, where the goal is to mix pre-trained GANs, and thus it is less *unfair* to assume a pre-trained embedding.

Ensemble weights: we reported the learned ensemble weights in Table 4 in Appendix C. In most cases, the weights are not very sparse and at least two GANs are assigned non-zero weights.