Reviewer 1

- **Sparsity level vs latent code dimension of same signal:** We can expect the latent code dimensionality of a GAN to be smaller than the sparsity level of the corresponding image with respect to a wavelet basis. For example, consider a set of images that correspond to a single train going down a single track. This set of images form a one dimensional submanifold of the manifold of natural images. If properly parameterized by a generative model, then it would have a latent dimensionality of approximately 1, whereas the number of wavelet coefficients needed to describe any of those images is much greater. The work of Bora et al. shows that compresed sensing can be done with 5-10x fewer measurements than sparsity models. This provides evidence for the more economical representation of generative models than of sparsity models. Additionally, it is more natural to view the natural signal manifold as a low-dimensional manifold, as opposed to being the combinatorially-many union of low dimensional spaces. Performance gains are provided by the fact that the natural signal manifold can be directly exploited, whereas the union of subspaces can only be indirectly exploited via convex relaxations. We will add exposition to this effect to the paper.

Reviewer 2

- **Sparsity level vs latent code dimension of same signal:** Please see response to Reviewer 1 regarding sparsity level and latent code dimension of same the signal.
- **A2 and A3 conditions in the theorem statements:** Thank you for your note. By $n_i = \Omega(n_{i-1})$, up to a log factor, we mean that there exists an absolute constant $c > 0$ such that $n_i \geq cn_{i-1}\log(n_{i-1})$. We will make this clear in the paper.
- **Truncation of the rows of the last layers:** Thank you for the insightful observation. We agree that on an event of high probability the truncation of the rows have no effect. This will make the theorem shorter and easier to understand. We will change the assumptions of the theorem and its proof to reflect this observation. It will result in a 1-2 line addition to the proof.
- **Convergence of the algorithm:** The focus of the paper was to show that the empirical risk objective function, under certain conditions, has a descent direction for every point outside a neighborhood of four hyperbolic curves, one of which contains the set of global minimizers. Based on this landscape of the objective function, we propose a gradient descent scheme, which will converge to a point in one of the four neighborhoods. In principle, a convergence result to the global minimizer by gradient descent is possible, though it would be considerably challenging because it would require showing a convexity-like property around the hyperbola. We believe that such an exposition would significantly increase the technicality of the paper, and we leave this for possible future work. We will clarify this in the paper.
- **Hyperbolic curve with lowest objective value:** The correct hyperbolic curve has the lowest objective because the objective is non-negative and the objective value at the correct hyperbolic curve is zero. This is the reason by which the negation-checking tweak of the gradient descent algorithm works.

Reviewer 4

- **Random gaussian matrices:** In the paper we provide two deterministic conditions that are sufficient to characterize the landscape of the objective function, and show that Gaussian matrices satisfy these conditions. In essence, we only require approximate Gaussian matrices. We also note that the state-of-the-art literature for provable convergence of the training of neural networks (for regression and classification) admit proofs only in the case that the final trained weights are close to their random initialized. Thus, our neural network assumptions are consistent with the best known cases for which networks can be provably trained. We will add a few sentences to this effect in the paper.
- **Generative model assumptions:** The benefit of using a generative prior is that it is a reusable prior and can be used in multiple inverse problems. For example, a generative prior can be trained once and used in in-painting, image-to-image translation, compressed sensing, blind deconvolution, etc.
- **Convergence of the algorithm:** Please see response to reviewer 2 regarding convergence of a gradient descent scheme to a global minimizer.
- **The set $\mathcal{K}$ as the set of local maximizers:** Thank you for the observation. We agree that the directional derivative at a point along any direction being non-negative does not rule out the possibility of that point being a local minimizer. In the proof of Theorem 2, we show that for every point in the set $\mathcal{K}$, the directional derivative along a set of directions is zero and for every other direction, the directional derivative is strictly negative. Thus, the points in the set $\mathcal{K}$ are local maximizers. We will modify the Theorem statement and the proof to reflect this clearly in the final manuscript.