We thank the reviewers for their valuable feedback.

R1-*More analysis of global landscape on LSGAN, WGAN, etc:* 1) Note Remark 1 (L155): the results can be extended to a class of separable GANs and relativistic GANs (R-GANs). More specifically, Thm. 3 (L1501) and Thm. 4 (L1513) show that separable GANs in Eq. (30) have bad basins while R-GANs in Eq. (31) have no bad basins. These results only require minor conditions on the loss (Assumption J.1-J.5), covering logistic loss, hinge loss, squared loss, etc.; 2) To cover LS-GAN (min-max version), two minor changes suffice: change the two h functions to h_1 and h_2 in Eq. (30); change Assumption J.1-J.3 accordingly. We'll modify to include LS-GAN and R-LS-GAN. 3) W-GAN is difficult to analyze. See App. J.2 for a discussion. The difficulty also indicates that our contribution goes beyond a global landscape analysis in that we identify the losses (R-GANs) that are amenable to rigorous analysis.

R1&R4-*Experiments to validate proposed GAN losses.* As stated in L230, "the effectiveness of relativistic GANs has been justified (to some extent)" and "our goal is to use experiments to support the landscape theory." For this, we focus on: a) **advantage in narrow nets**; b) **robustness to initialization**. Our paper validates a) and b) in four ways: **1)** On L265-269 we show that for MNIST and a certain initial point, RS-GAN outperforms JS-GAN by 30 FID scores (around 30 vs. 60). **2)** On L256, we show RS-GAN outperforms JS-GAN by 9 FID scores (45 vs. 53) when using a ResNet (bottleneck) on STL. **3)** In Tab. 11 (in appendix) we show that R-hinge-GAN outperforms hinge-GAN with 1/4 width (24 vs. 33 FID on CIFAR10). Both SN-GAN and BigGAN papers use hinge-GAN, so we check hinge loss. **4)** In experiments (Tab. 2, Tab. 11 in paper), separable versions (JS-GAN, hinge-GAN) do not beat their relativistic counterparts (RS-GAN, R-hinge-GAN) in any case. These points show: R-GANs are more robust to initialization and architecture. In new experiments we show: **5)** R-LS-GAN outperforms LS-GAN by 6 FID (42 vs. 48) with 1/4 width (Tab. 1 below); **6)** RS-GAN outperforms WGAN-GP (Tab. 2 below); **7)** experiments on LSUN (higher resolution than CIFAR10 - Fig. 1). These new experiments further justify the advantage of R-GANs. We will explain in the main text.

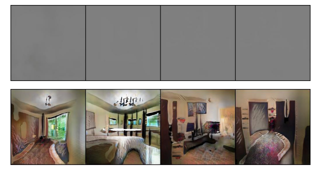

Figure 1: LSUN (256×256) generation with CNN structure for JS-GAN (above) and RS-GAN (bottom).

R2-*existence of reachable path doesn't mean SGD can follow it; there are still questions.* There're two branches of results for neural nets: one branch [60,43] only discusses paths and basins assuming width n. A drawback, as you point out: convergence of GD is not proved. Nevertheless, this is enough to distinguish RS and JS-GAN. Another branch [2,25,37] proves convergence of SGD assuming width $\geq n^6$. Drawback: assumption of width n^6 is impractical. An ideal result that SGD converges for width n is a huge open question for neural nets (attempts exist, but all have strong limitations). We do not intend to solve this open question here. We combine Thm. 1,2 with the first branch since it is cleaner and already non-trivial. It is possible to combine with the second branch (on convergence), but it will make this paper much longer. Future advances for neural nets can be potentially combined with our function space result.

R2-*parameterized result is vague. More detail of connections.* ***and***
R3-*Intuition on proof of Prop. 1,2; how proof differs from supervised learning.* The proof strategy of Prop. 1,2 is an adaptation of those of [60,43,59] to GANs. References [60,43,59] consider a convex loss (e.g. quadratic) in function space and "transfer" decreasing paths in function space to decreasing paths in parameter space. To achieve this "transfer," some assumptions on the architecture (e.g. width large enough) are needed [60,43,59]. We apply this approach to the GAN loss. In our proof, we state the general requirement of "transfer" in Assumption I.1-I.3, and then prove when these assumptions hold in Appendix I.2 and I.3 (using architecture assumptions of [60,43,59]). We'll discuss in the main text.

R3-*proof sketch of Thm. 1, Thm. 2.* For Thm. 1, careful computation suffices. For Thm. 2, we build a graph with nodes being input data, decompose the graph into cycles and trees, compute the loss by grouping the terms according to cycles and trees, and add each term. We'll sketch the proof in the main text.

R4-*study WGAN.* Thanks for suggesting. i) See the first response, point 3) in L8 of the rebuttal. ii) We add simulation showing that RS-GAN outperforms WGAN-GP for standard datasets (Tab. 2 below).

R4-*study DRAGAN architecture.* Thanks for pointing out the reference, which we read with great interest. It suggested that mode collapse may be due to bad equilibria. However, there is no formal statement or proof. We will cite it and discuss the connection with our work. DRAGAN adds a penalty of the gradient which may help eliminate some basins, but it likely creates other basins. A formal analysis requires much effort, and is an interesting future direction.

R4-*Sec. 3 (two-cluster) a bit redundant with analysis of 2-point distribution.* Thanks. The analysis of a 2-point distribution is about the landscape, not the training process. In contrast, Sec. 3 shows that the basin really appears in training, and the theoretical values 0.48 and 0.35 really play a role in understanding the training process. Following your comment, we will reduce the length of Sec. 3.

	Regular	channel/2	channel/4
LS-GAN	**32.93**	37.83	48.63
R-LS-GAN	34.78	**34.34**	**42.86**

Table 1: FID results on CIFAR-10 for LS-GAN and R-LS-GAN with CNN structure given in Tab. 5 of the appendix.

		Regular	channel/2	channel/4
CNN	WGAN-GP	39.66	42.39	50.56
	RS-GAN	**27.16**	**32.74**	**49.74**
ResNet	WGAN-GP	21.33	23.80	40.45
	RS-GAN	**19.31**	**21.78**	**31.26**

Table 2: FID results on CIFAR-10 for WGAN-GP and RS-GAN with CNN and ResNet.