We thank the four reviewers for their careful reading, detailed feedback, and helpful comments. Below, we begin with some clarifications on our contributions and minor extensions, and end with specific responses to each reviewer.

Why SDP? The adversarial examples problem is relatively unique in ML in that *global optimality* matters a lot. (In training a model, this would just overfit the model.) A central weakness of attack algorithms based on local optimization (e.g. PGD of Madry et al.) is that they cannot *prove* that an adversarial example they've found is globally optimal—even if it really is globally optimal. On the other hand, provable guarantees of global optimality are certainly obtainable if we accept exp-time (e.g. Katz et al.) or exp-time in the worst-case (e.g. Tjeng et al.). Within this context, *SDP is interesting because it is the best tool for proving global optimality in poly-time*. Goemans and Williamson revolutionized combinatorial optimization when they used SDP to prove near-global optimality bounds right at the cliff edge of poly-time (assuming $P \neq NP$). Candes, Tao and their coauthors revolutionized compressed sensing when they used SDP to solve problems previous thought NP-hard to guaranteed global optimality in poly-time.

Our contributions. Amongst the adversarial examples literature, we are the first to guarantee a globally optimal certificate in polynomial time. Within this context, others have used SDP, and have obtained good empirical results, but we are the first to give an end-to-end theoretical proof (of any kind) for SDP. *Our paper is a first step towards global optimality in poly-time; our contribution is the proof technique to get there*. The standard technique of analyzing the SDP dual (e.g. Candes and Tao) immediately runs into painful, possibly insurmountable issues. Instead, we developed a nonconvex technique (Appendix A) that reduces the SDP into a sequence of possibly nonconvex projections. *Viewing our primary contribution as the proof technique, we have taken meticulous care in communicating the technique in a clear, clean, pedagogical way, by proving clear-cut results on simple examples.* We did this because we wanted to make it as easy as possible for future researchers to build off of our work. This is especially important because the naive SDP relaxation doesn't work well on its own, but it has the potential to be made to work well by future researchers. (Much like how Lasserre built on top of Sherali-Adams and Lovasz-Schrijver)

Problem (A) vs Problem (B). Our proof technique works equally well for both (A) and its *convex restriction* (B). But (A) is almost always loose, so we cannot prove anything on (A) other than to state reasons for why the SDP is loose. In comparison, problem (B) is tight for a sufficiently small ρ (it is essentially a trace regularization to induce a low-rank solution) so we're able to prove predictive bounds and verify them numerically. This is a fairly common route in SDP.

Multiple layers. Our proof technique easily extends to the $\ell > 1$ case, as we discussed in Appendix E. The resulting problem (E.1) can be "unrolled" by recursively applying the one-layer argument. But the issue here is that after the first layer, we begin projecting onto hyperbolas. Conditions for hyperbola-on-hyperbola to be collinear can be derived but are difficult to interpret and verify (they are themselves LMIs). Nevertheless, we believe this is a direction future researchers can build off, because LMIs can always be simplified by assuming structure.

Reviewer 1. We would like to emphasize the fact that our result is the *first proof of tightness* within this context. Global optimality is exceedingly important within the context of adversarial examples for obvious reasons, and we give the *first provable method* that attains global optimality in poly-time. Our contribution is in the proof technique used to achieve this; we had to diverge substantially from the compressed sensing SDP literature to get here. Choice of ℓ_2 oracle over ℓ_∞ oracle. Both oracles are common in the literature, but generate essentially the same adversarial examples (see e.g. Carlini and Wagner). Lipschitz constants techniques are ℓ_2 methods that become considerably more conservative on ℓ_∞. SDP easily accommodate ℓ_∞; it is the only the theoretical analysis that becomes complicated. Large ρ regime. In the regime of radius $\rho \to \infty$, we view (A) as (B) with $\hat{\mathbf{z}} = -\rho\mathbf{w}/\|\mathbf{w}\|$, but this means the center of the ball $\hat{\mathbf{z}} \to \infty$ as well. Our tightness guarantees for (B) require $\hat{\mathbf{z}}$ to remain bounded. Quantative measure of tightness. On the one-neuron example, the relaxation is tight if $|\langle \mathbf{e}, \mathbf{x}\rangle| = \|\mathbf{x}\|$. But if $|\langle \mathbf{e}, \mathbf{x}\rangle| < \|\mathbf{e}\|$, then the incidence angle $\theta = \arccos(|\langle \mathbf{e}, \mathbf{x}\rangle|/\|\mathbf{e}\|)$ gives essentially the condition number of the high-rank solution. This is an insight that only becomes clear through our proof technique; we promise to add this point to the paper.

Reviewer 2. We regret the cluttering noted by the reviewer. We endeavor to reduce clutter in a future revision.

Reviewer 3. We thank the reviewer for a number of key insights, and for catching several bugs in the Appendix. We hope our common response written above adequately addresses the reviewer's concerns regarding our contributions. Proof of Lemma A.3: agreed. Line 706 should read $\hat{\mathbf{z}} = \mathbf{u} - \rho\mathbf{w}/\|\mathbf{w}\|$. The next line should read $\mathrm{tr}(\mathbf{X}_\ell) - 2\langle \hat{\mathbf{z}}, \mathbf{x}_\ell\rangle + \|\hat{\mathbf{z}}\|^2 - \rho^2 = \mathrm{tr}(\mathbf{X}_\ell) - 2\langle \mathbf{u}, \mathbf{x}_\ell\rangle + \|\mathbf{u}\|^2 + 2\rho[\langle \mathbf{w}, \mathbf{z}\rangle + b]$(the term $-\rho^2$ was lost). For nonnegativity, note that $\mathrm{tr}(\mathbf{X}) - 2\langle \mathbf{u}, \mathbf{x}\rangle + \|\mathbf{u}\|^2 = \mathrm{tr}(\mathbf{X} - \mathbf{u}\mathbf{x}^T - \mathbf{x}\mathbf{u}^T + \mathbf{u}\mathbf{u}^T) = \mathrm{tr}(\mathbf{X} - \mathbf{x}\mathbf{x}^T + (\mathbf{x}-\mathbf{u})(\mathbf{x}-\mathbf{u})^T) = \mathrm{tr}(\mathbf{X} - \mathbf{x}\mathbf{x}^T) + \|\mathbf{x}-\mathbf{u}\|^2$ but we have $\mathbf{X} - \mathbf{x}\mathbf{x}^T \succeq 0$ and therefore $\mathrm{tr}(\mathbf{X} - \mathbf{x}\mathbf{x}^T) \geq 0$. The claim that (A-lb) is a relaxation of (B-lb) follows from the corrected version of the equation above, which shows that a feasible point $\mathbf{X}_\ell, \mathbf{x}_\ell$ satisfying $\mathrm{tr}(\mathbf{X}_\ell) - 2\langle \hat{\mathbf{z}}, \mathbf{x}_\ell\rangle + \|\hat{\mathbf{z}}\|^2 \leq \rho^2$ would then immediately satisfy $2\rho[\langle \mathbf{w}, \mathbf{z}\rangle + b] \leq 0$. These points will be clarified and fixed.

Reviewer 4. We agree with the reviewer and promise to completely rewrite the introduction to better reflect our contributions. Comparison to Raghunathan et al. This previous paper was almost entirely empirical. Global optimality was not at all their focus; most of their paper was spent comparing SDP vs LP.