We thank the reviewers for their feedback and helpful comments, however there are important misunderstandings which we now clarify and politely ask for another evaluation using this rebuttal as context. We appreciate that R1, R2 and R5 acknowledged the interest, novelty and challenges of the No-Box threat model as well as the soundness of the Adversarial Example Games formulation. We believe these two contributions —i.e., proposing a novel and more realistic threat model and at the same time providing a framework to study and solve it—are extremely relevant to a venue that cares for new research perspectives. We are also encouraged that R1, R2, R4 and R5 felt that our submission was well written and the main thrust of the paper was clear. Finally, we are pleased by the interest of R1 and R4 in the proofs provided in the appendix. We start by clarifying their concern about Prop. 1.

Concern regarding Proposition 1 (R1, R4):. The concern regarding the absence of convex-concavity of the objective function. As mentioned L550, we need the loss function ℓ to be convex (this is the case for any standard loss function such as the cross entropy-loss or the mean-square loss). In (12), since $(x', y) = (g(x, y, z), y) \sim p_g$ we have,

$$\varphi(f, g) = \varphi(f, p_g) = \mathbb{E}_{(x',y)\sim p_g}[\ell(f(x'), y)]\,. \tag{1}$$

Then, because the loss function ℓ is convex for any g, we have $f \mapsto \varphi(f, g)$ that is convex. Regarding the concavity of the payoff with respect to the second variable, by linearity of the expectation with respect to p_g the function $p_g \mapsto \varphi(f, p_g)$ is linear and thus concave. Overall the payoff $(f, p_g) \mapsto \varphi(f, p_g)$ is convex concave and the set to which belongs p_g is convex and compact (c.f. L566 and L569). Thus, we can apply Fan's Theorem (Theorem 1 L573).

The NoBox attack demands the training set (R2, R5):. We appreciate the concern echoed by R2, R5 regarding the ability of the adversary to sample a training set from the same distribution as the target model as a strong requirement. We disagree. We argue that every single blackbox transfer attack paper, including all baselines we use, requires an available training set in order to first train a source model which can then be used to learn transferable attacks. Indeed, without access to such a training set it would not be possible to train a source model thus preventing possible attacks. Also, we note that Defn 3. in the seminal work by Tramèr et al., [2018] (ICLR 2018) standardizes this requirement.

Experimental setup and results (R1, R2, R5):. We value the feedback shared by R1 and R2 regarding the utility of AEG over other blackbox transfer attack strategies. Our work is the first to propose a principled way to craft transferable adversarial examples to function classes while all other prior transfer strategies, while powerful, are heuristically motivated. In terms of raw performance, AEG is SOTA or within 1% of the best transfer method for known architecture attacks, which is the precisely the setting our theory applies to. When transferring to different architectures and possibly differing function classes AEG is still SOTA when using a RN-18 or DN-121 as the source model. For robust models, we expect a drop in performance for ensemble adversarial training as the theory suggests our representative classifier cannot cope with a set union of multiple function classes. While for PGD-Adversarial attacks we are again SOTA for all transfer attacks highlighting how our generator is optimized against a worst case adversary. Thus we would like to politely push back against the assertion by R1 that there is no benefit to AEG in any setting. Regarding R5's concern on attacking non-differentiable robust models, the elegance of our framework gives an affirmative answer as long as this robust model is within the same function class. In practice this amounts to training with a non-differentiable f_c. We would also like to gently clarify to R1 that the ϵ-budget in our CIFAR10 experiments is identical to the Madry challenge as the range for pixel values $0 - 255$ is inclusive of zero.

How can we converge in practice (R1, R2):. The convergence of the optimization process in a game with general non-convex losses is an *open question in the field* (see Lin, Jin & Jordan [2019] or Kodali et al. [2017] for discussions). Similarly as the original GAN paper [Goodfellow et al, 2014] and any practical GAN paper, proving the convergence of the gradient based optimization algorithm is beyond the scope of this paper. However, note that we use an extrapolation based method (Extra-Adam) as extrapolation is a principled method for minimax optimization proposed by Gidel et al. [2019] that, unlike Adam, has convergence guarantees in the convex-concave setting.

How the architecture solves the problem (R1):. The architecture Fig. 4 corresponds exactly to the theory. The generator takes as input x, y and a latent variable z (that depends on x in order to exploit the structure of the input and outputs $x <= x + g(x, y, z)$ such that $\|g(x, y, z)\|_\infty \leq \epsilon$. The critic f takes as input x' and y and occurs a loss $\ell(f(x'), y)$. That loss is exaclty the one described in (AEG). We will further clarify these points in our submission.

The AEG method is expensive (R2):. We agree that AEG is expensive to train. But, in the context of adversarial attack what matters more is the computational cost of crafting an adversarial example at test time (the attacker may pay a cost for taking too much time to craft an adversarial example). In that context, our method is by far among the fastest since it only requires the inference of the generator network while other standard methods require to solve an optimization problem by querying many gradients (or function values) of a predictor function.

Q3 of Section 5 claims to attack robust classifiers (R4). We acknowledge R4 comment regarding our method against robust classifiers. While it is true that these approaches are not provably robust, to the best of our knowledge no such method exists to certify deep architectures such as Wide Resnet on MNIST and CIFAR10. Furthermore, both the defense methods studied in our work are the defacto gold standards for transferable adversarial examples thus we feel our investigation against robust models in the NoBox setting is appropriate.