We thank the reviewers for finding our problem relevant (R4), the game-theoretic formulation of adversarial attacks and defenses novel (R1, R3, R4) and the paper clearly written (R1, R2, R3, R4). The main critique is that the assumptions (local-linearity, restrictions on defender strategies) are strong. However, the reviewers did not point to prior work that is able to handle our setting. We believe our solution under these assumptions is a necessary stepping stone which lays the foundations for future theoretical advancements into additive attacks and defenses. We now provide detailed responses:

[R1, R2, R3, R4] Validity of locally-linear assumption. While the assumption of local-linearity might seem strong at first, there is ample empirical evidence of its validity for neural networks. Fig. A shows the decision boundaries of a CNN trained on CIFAR-10 in a ϵ neighbourhood of many randomly-selected images, where white denotes the predicted class and other shades denote other classes. It can be seen that, locally, the boundary is approximately linear. This *linearity hypothesis* was proposed in [C1] and further explored in [C2] (showing empirical evidence) and [24] (linking to the existence of universal adversarial perturbations). Recent studies [C3], [C4] improve Deep Neural Networks' robustness by promoting local-linearity. Hence, we stress that our work *does* partially apply to modern real-world classifiers, and is certainly not limited to linear classifiers. [C1]: Goodfellow et. al. Explaining and harnessing adversarial examples. [C2]: Warde-Farley et. al. Adversarial Perturbations of Deep Neural Networks. [C3]: Lee et. al. Towards Robust, Locally Linear Deep Networks. [C4]: Qin et. al. Adversarial Robustness through Local Linearization.

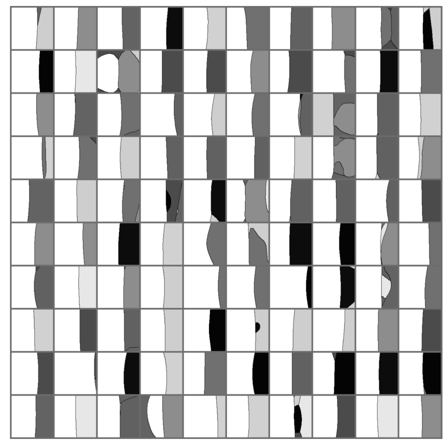

Figure A: Church-Window plots for a CNN f reproduced from Fig. 11.2 of [C2]. Each plot shows $f(\mathbf{x} + a\mathbf{u} + b\mathbf{v})$ for $a, b \in [-\epsilon, \epsilon]$, where $\mathbf{u}$ is the FGM direction, $\mathbf{v}$ is a random direction orthogonal to $\mathbf{u}$ and $\mathbf{x}$ is a random data-point from CIFAR-10. White denotes the class $f(\mathbf{x})$, and other shades denote other classes.

[R1, R2] Do images typically lie near non-linear parts of the decision boundary? Yes, empirical evidence on real world networks as in Fig. A suggests that we almost never find a decision-boundary corner in a 2ϵ neighbourhood around the image. However, these networks partition the input space into thousands of polyhedra, and the problem of efficiently verifying whether a given input image lies in a linear-region far from any vertices or edges of these polyhedra is an open research question [C3].

[R1] Strong assumption on defender's knowledge. The design of a game theoretic framework for analyzing attacks and defenses in a level playing field requires imposing constraints that prevent the attacker or the defender from always winning. We thus intentionally do not work in a situation where the classifier is publicly released for attack. Instead, we want to model the reality that the defender will train the classifier knowing the possible attacks, and in turn the attacker will create new attacks knowing that the publisher was aware of possible attacks, and so on. This precisely leads to the game-theoretic notion of *perfect knowledge* that we use in Section 2.

[R3, R4] Strong assumptions on defender's strategies. As pointed out by R4, obtaining the optimal defense in our current robust set is already a hard problem. Extending the set to include perturbations dependent on the data-point x will lead to more complex robust sets, and this is a direction for future work. A first step could be to let $\mathcal{A}_d$ contain all linear transformations projected to the set of allowed perturbations i.e. $\mathcal{A}_d = \{f_M | f_M : \mathcal{X} \to \mathcal{V} \text{ s.t. } f_M(x) = \Pi_{\mathcal{V}}(Mx)\}$.

[R3, R4] Scalability of the optimization procedure. As R4 correctly notes, the optimization problem (7) is hard, and we present an approximate solution which currently works for small datasets as shown in Table A (it takes < 10 seconds for MNIST, FMNIST). In ongoing work, we are scaling it to larger datasets by exploiting the fact that (7) is a convex-maximization problem and applying techniques from the classical literature on efficient approximations to (7).

[R2] Extensions to multiple classes, non-ReLU activations, incorporating test accuracy. We thank R2 for the suggestions. They are excellent directions for future work. We will add [Sengupta et. al.] to prior work.

Table A: Mean (Variance) over $\binom{10}{2}$ pairs.

Attack	Defense	MNIST (%)	FMNIST (%)
-	-	99.9 (0.0)	99.9 (0.1)
FGM	-	53.3 (10.0)	47.4 (5.1)
FGM	SMOOTH	71.2 (14.2)	67.4 (9.0)
PGD	-	71.9 (12.0)	74.7 (7.3)
PGD	SMOOTH	94.0 (4.0)	90.3 (8.5)

[R2, R3] Experiments on other datasets. In our experiments we used $\epsilon = 4$. Table A shows the result of repeating our experiment in Table 1 of the paper (the column Approximate Accuracy is shown) over all the 55 pairs of classes in MNIST and FMNIST. It can be seen that the trends observed in the paper hold even when the experiment is repeated over multiple pairs.

[R2, R3] Minor writing issues We will fix the typos in Eq. (33), CW method reference, and clarify that we are not using an isotropic Gaussian for randomized smoothing.

[R2] Is PGD better than FGM against our defense? No. Under the locally linear assumption, FGM performs better than PGD (Table 1: $48.3 < 85.6$ and $94.5 < 99.1$) as expected by Lemma 2. The same trend is seen in Table A.

[R4] Is there a PAC style argument? Yes, Sec. 5 establishes a PAC-style bound: The estimated solution is v_n^*, and the optimal one is v^*. Eq. (16)-(18) upper bound the difference between the objectives by a quantity α, i.e. $\phi(v^*) - \phi(v_n^*) \leq \alpha$. Eq. (19) establishes that the expectation of α is upper-bounded by a small quantity β, i.e. $\mathbb{E}[\alpha] \leq \beta$. Using α in Eq. (10) now yields $\Pr[|\phi(v^*) - \phi(v_n^*) - \beta| > \epsilon] \leq \exp(-2n\epsilon^2)$, which is a PAC bound.