We sincerely thank the reviewers for their time and valuable feedback on our work. We are pleased to see that the reviewers find our work interesting, thorough and well-written. We thank **R1** and **R3** for their motivating comments on the proposed single-step defense. We will emphasize this more in the final version. We sincerely apologize for the error in sign of the max-margin term in the loss (Eq.1 in main paper). We understand that this has led to significant confusion. The corrected loss which is maximized for attack generation is : $L = -f_\theta^y(\widetilde{x}) + \max_{j \neq y} f_\theta^j(\widetilde{x}) + \lambda \cdot ||\boldsymbol{f_\theta}(\widetilde{x}) - \boldsymbol{f_\theta}(x)||_2^2$

Discussion on the proposed regularizer: We justify the significance of proposed regularizer for GAT defense in Sec.1 of the Suppl. This can be extended to attacks as well. The local Lipschitz constant ($\mathcal{L}$) of adversarially trained models is low compared to standard models. Based on Eq.5 in the Suppl., $\mathcal{L}$ acts as an upper bound to the ℓ_2 term upto a constant factor. Hence, a low value of $\mathcal{L}$ leads to a low value of the ℓ_2 term. Therefore, while finding an adversary, maximization of the ℓ_2 term additionally leads the optimization to move towards the direction of worst case local smoothness. The use of ℓ_2 term is also motivated by the use of a better optimization objective initially as discussed in L168-L180 of main paper. We will explain these in more detail in the final version. The plot of CE loss vs. iterations (will be included in final version) for the proposed attack shows a larger increase in CE loss in presence of the ℓ_2 term. We will also draw parallels with the theory of graduated optimization (On Graduated Optimization for Stochastic Non-Convex Problems, Hazan et al.), which shows that such methods can lead to improved optimization for the family of σ-nice functions.

[R1] Too many variations of proposed method: We thank **R1** for the feedback. We will certainly work on improving the clarity of experimental setup. Although we proposed multiple variants, we would like to clarify that the main attack, GAMA-PGD uses the same loss function (max-margin, ℓ_2 term) and optimizer (PGD) across all experiments. Also, the main defense, GAT uses the same optimizer (single-step PGD) and loss (CE, ℓ_2 reg) across all experiments.
[R1] Loss change in alternate iterations seems hacky: Results in Table-2 of the Suppl. show that impact of alternating losses is marginal. The AA accuracy is 46.37% without alternation and 46.72% with alternation. (L169-172 of Suppl.)
[R1] Stability of GAT across reruns: We get similar results with low variance (SD = 0.224). The PGD100 CIFAR10 acc across reruns are 52.14, 51.7, 52.02, 52.35, 51.96, 51.74. Unlike FBF, even in the last epoch, we obtain robust models.
[R1] Use of APGD framework: We thank **R1** for the valuable suggestion. We will certainly investigate this in future.

[R2] Objective function in Eq.1: We request **R2** to kindly reconsider the contributions of our paper after the correction of loss function in L3-L5 above. The ℓ_2 regularizer is maximized for attack generation and minimized in the defense.
[R2] Comparison to CW attack, significance of ℓ_2 term in attack: CW attack uses max-margin loss in logit space, while we use this in softmax space. We introduce the ℓ_2 regularizer which is decayed to 0 over a few iterations. The advantage of the proposed approach is not only the addition of ℓ_2 loss term, but also in decaying it to 0 over a few iterations. Therefore, from Table-2 in main paper, the difference (100-step, 1 run) w.r.t. CW attack is 0.9% and advantage from the ℓ_2 regularizer and its schedule is 0.65%, both of which are significant relative to the trends on attack leaderboards. We get a significant boost over CW attack across all defenses in Table-1. We will include these results in the final version.
[R2] Significance of ℓ_2 term in defense: Table-2 in the Suppl. shows that without the ℓ_2 term in adversary generation, the AA accuracy is 43.37%, while it increases to 46.37% with the ℓ_2 term included. Similarly, by replacing the ℓ_2 term in defense with CE on adv samples, AA accuracy drops to 30.2%, which is 16.52% lower than the proposed method.

[R3] SPSA, ℓ_2-attacks: We thank **R3** for the valuable suggestions. We report results against the gradient-free attack, Square in the paper. We will certainly include results on SPSA and the suggested ℓ_2 attacks in the final version.
[R4, R3] Choice of λ and sensitivity for the attack: Kindly refer to Section-3.2 of the Suppl. and Fig.1(a) of the Suppl.
[R4, R3] Results on CIFAR-10 defense by Madry et al.: We consider the ResNet-50 (not WRN-34) architecture for reporting results on the defense by Madry et al. We use the pretrained model available in their *robustness* GitHub repo. However, the numbers reported in FAB, MT and AA papers are on the WRN-34 model by Madry et al. We apologize for missing the architecture details of defenses in Table-1. We will certainly include it in the final version. We use ResNet-18 architecture for the PGD-AT model in Table-3 since the same architecture is used across all defenses.
[R4] MT baseline results: For the plot in Fig.2(a), we cycled through the other 9 classes of CIFAR-10 in a random order and the 10^{th} restart was an untargeted max-margin attack. With 2^{nd} highest logit as the first target, the single restart acc is 54.33%. There is no change in the 10-restart accuracy as expected. We use Adam (without sign of gradient) and other hyperparameters as used by the MT authors. For the 5-restart results (4 random targets + 1 untargeted) reported in Table-1, we see marginal improvement with use of highest logits. For the Trades defense, MT attack acc improves from 53.57% to 53.32% with the use of highest logits. GAMA-PGD achieves 53.17% and an MT version of GAMA attack achieves 53.09% for 5 restarts. We thank **R4** for this feedback. We will update the table and plot in the final version.
[R4, R1] Loss landscape: Fig.3(c) in Suppl. shows that the loss landscape of the single-step defense GAT is smooth.
[R4] Clean acc of GAT: We use 40k-10k train-val split for GAT (single-step) training, whereas for other defenses, full 50k train set was used. With 49k-1k split on CIFAR10 WRN-34, we get clean acc $= 85.17\%$ and AA acc $= 50.27\%$.
[R4] We thank **R4** for the suggestions. We will explore the use of Adam for GAMA attack, include the baselines CURE (41.4% PGD-20 acc on WRN28-10, CIFAR10) and LLR (44.5% MT acc on WRN28-8, CIFAR10) and organize the tables better. The proposed method GAT is significantly better than these baselines under limited budget constraints.

We look forward to more insightful discussions on our work at NeurIPS 2020.