We appreciate the insightful and constructive comments by all reviewers. All comments will be carefully addressed in the final version. Below, we provide detailed responses to major concerns.

To Reviewer 1:

> Comparison to [40].

Our paper and [40] differ in the following three aspects. First, [40] considers a Lagrangian relaxation of the local worst-case risk $R_{\epsilon,p}(P,h) = \sup_{Q:W_p(P,Q))\leq\epsilon} R_Q(h)$ for a fixed penalty parameter λ, which can be reformulated as $\sup_Q\{R_Q(h) - \lambda W_p(P,Q)\} = \mathbb{E}_P[\varphi_{\lambda,f}(Z)]$. Then an adversarial training procedure is developed to minimize $\mathbb{E}_P[\varphi_{\lambda,f}(Z)]$ in order to achieve distributional robustness. However we focus on the original $R_{\epsilon,p}(P,h)$ and expect to find the optimal λ which minimizes $\{\lambda\epsilon_{\mathcal{B}} + \mathbb{E}_{P_n}[\varphi_{\lambda,f}(Z)]\}$. Specifically, we show that the optimal λ falls in $[\zeta^-_{f,P_n}, \zeta^+_{f,P_n}]$ in Lemma 4. In this way, we are able to obtain a much tighter bound. Moreover, Lemma 4 suggests that the λ in [40] can be selected from the interval $[\zeta^-_{f,P_n}, \zeta^+_{f,P_n}]$. Second, the penalty parameter λ must be set to a large number at the training procedure in [40], and so their method can only deal with gentle adversarial attacks. In contrast, our bound can deal with arbitrarily large and general adversarial attacks. Finally, the proof techniques in two papers are different. [40] first relaxes the original local worst-case risk and then provides a generalization bound for the relaxed problem. In contrast, we prove a uniform bound for the local worst-case risk.

> Discuss how both bounds of theorem 1 compare to each other, ... and how does it relate to the adversarial risk.

The first two terms in both bounds can be deemed as a relaxation of the empirical adversarial risk. They correspond to the empirical risk and the effect of adversary on empirical risk, respectively. Although the first bound is tighter, it is hard to optimize because of the inner minimization problem. Therefore, we only discuss the second bound in the paper.

> Update the conclusion to more clearly discuss ..., and how would such an approach differ from what [40] does.

Our bound has two data dependent terms: $1/n\sum_{i=1}^n f(z_i)$ and $\lambda^+_{f,P_n}\epsilon_{\mathcal{B}}$, corresponding to the empirical risk and the effect of adversary on empirical risk, respectively. However, in practice, we cannot minimize the sum of the two terms because λ^+_{f,P_n} is computationally intractable. Instead, we consider a heuristic method.

We consider a data-dependent upper bound for λ^+_{f,P_n} which is usually easy to obtain. Instead of using the exact λ^+_{f,P_n} in the objective function, we consider a regularization parameter $\eta \in [0,1]$ which can be selected via a grid search. For a fixed η, we multiply it by the upper bound for λ^+_{f,P_n} and use this product as a surrogate of the true λ^+_{f,P_n} in the objective function. Afterward, we minimize this surrogate objective function and obtain the optimal solution for this specific η. Each such η corresponds to a solution. Finally we choose the best one from these candidates.

The proposed approach is largely different from the training procedure in [40]. In [40], a fixed parameter λ is chosen in advance. Then the training procedure aims to find the optimal f which minimize $\mathbb{E}_{P_n}[\varphi_{\lambda,f}(Z)]$. Their paper focuses on developing an algorithm for optimizing $\varphi_{\lambda,f}(Z)$. However, our method puts more efforts on finding a good λ which depends on the data and f, i.e., λ^+_{f,P_n}. Once finding λ^+_{f,P_n}, the optimization would become relatively easier, because $\psi_{f,P_n}(\lambda)$ in our objective function is 0 when λ is set to λ^+_{f,P_n}.

To Reviewer 2:

> This paper ...better characterize the excess risk bound, ... further provide detail comparisons with related results.

The excess risk bound for adversarial learning can be derived using Theorem 1 and Hoeffding's inequality. Here we provide the general form of excess risk bound. When applying it to SVMs and deep neural networks, the desired bounds can be derived. Denote $\bar{f} = \arg\min_{f\in\mathcal{F}} R_{P_n}(f,\mathcal{B})$ and $f^* = \arg\inf_{f\in\mathcal{F}} R_P(f,\mathcal{B})$. The general excess risk bound can be expressed as $R_P(\bar{f},\mathcal{B}) - R_P(f^*,\mathcal{B}) \leq \lambda^+_{\bar{f},P_n}\epsilon_{\mathcal{B}} + 24\mathfrak{C}(\mathcal{F})/\sqrt{n} + 12\sqrt{\pi}\Lambda_{\epsilon_{\mathcal{B}}} diam(Z)/\sqrt{n} + 2M\sqrt{log(2/\delta)/2n}$.

We compare our bounds with related results. For SVMs, our bound is the same as related bounds (e.g., Corollary 4.1, [34]) except for a dimension dependent factor $\sqrt{d}$, because we use covering number analysis instead of Rademacher complexity in deriving the bounds. This has been explained in the conclusion. For neural networks, both our bounds and the bounds in [6, 35] have an explicit dependency on the network size, i.e., W. [35] have an additional factor of the number of layers of networks in their bound. Our work and [35] use spectral norm and Frobenius norm of the weight matrices, whereas the bound in [6] is given in terms of spectral norm and $(2,1)$ matrix norm. Although the results are similar in these papers, the proof techniques are different.

To Reviewer 3:

> Provide hints on how the bounds might be useful for Deep Neural Networks design.

Our adversarial risk bounds for deep neural networks might be helpful in the design of neural networks for resisting adversarial attacks. First, our results show that the adversary would introduce an additional contribution to the empirical risk. And from the expression for λ^+_{f,P_n} in Corollary 2, we can see that this effect could be weakened by the margin factor γ. This makes sense since margin can be regarded as a mechanism to defend against adversarial attacks. Therefore, choosing a relatively large margin value γ for which the empirical risk is small can improve the adversarial robustness. Second, the value λ^+_{f,P_n} is closely related to the Lipschitz constant of the function f. Our bound indicates that training the networks with the Lipschitz regularization term (e.g., Virmaux and Scaman, 2018) might be helpful for resisting adversarial attacks.