We would like to thank all the four reviewers for their comments. In this document, we try to briefly respond to the concerns and questions raised by Reviewers 2 and 4.

Reviewer 2: –(**major comments**) We have tried to avoid making any claims about "explaining the help from unlabeled data". Theorem 3 only provides a generalization bound. We revise the paper again to remove any remaining claims about this issue. However, using our framework, one can (at least theoretically) characterize cases where unlabeled data can provably help. First, it should be noted that by using only the labeled data for learning (as suggested by the reviewer), the residual generalization error in the classical learning framework would be $O\left(n^{-1/2}\eta^{-1/2}\right)$. But residual error terms of Theorem 3 are both $O\left(n^{-1/2}\right)$ (note that $\sqrt{\eta}+\sqrt{1-\eta}\leq\sqrt{2}$). Therefore, we can guarantee a much smaller residual error when supervision ratio is very small, i.e. $\eta\ll 1$. Second, for a highly compatible pair of model set $\boldsymbol{\Phi}$ and data distribution P_0, the condition $\mathrm{MSR}_{\boldsymbol{\Phi},P_0}\left(\lambda,\zeta\right)<\eta$ can be satisfied even for very small (and generally negative) values of λ. For a sufficiently small λ, our $\hat{R}_{\mathrm{SSAR}}$ becomes smaller than the average risk computed over only the labeled data. Let us discuss this matter, mathematically: For simplicity, assume the asymptotic case of $n\to+\infty$ (similar arguments hold for $n<+\infty$). Then, with a little abuse of notation and for any $\phi\in\boldsymbol{\Phi}$, we have:

$$\lim_{n\to+\infty}\hat{R}_{\mathrm{SSAR}}\left(\phi;\boldsymbol{D}\right)\overset{a.s.}{=}\mathbb{E}_{\boldsymbol{X}\sim P_{0_{\boldsymbol{X}}}}\left\{\eta\mathbb{E}_{y\sim P_{0_{|\boldsymbol{X}}}}\left\{\phi\left(\boldsymbol{X},y\right)\right\}+\left(1-\eta\right)\overset{(\lambda)}{\underset{y\in\mathcal{Y}}{\mathrm{softmin}}}\left\{\phi\left(\boldsymbol{X},y\right)\right\}\right\}\overset{*}{\leq}\mathbb{E}_{\boldsymbol{X},y\sim P_0}\left\{\phi\left(\boldsymbol{X},y\right)\right\},$$

where $*$ holds for sufficiently small values of λ, since $\mathbb{E}_{y\sim P_{0_{|\boldsymbol{X}}}}$ is an expectation operator but $\mathrm{softmin}^{(\lambda)}$ can go as far as being the $\min$ operator. Therefore, one can establish a set of theoretical conditions under which unlabeled data is guaranteed to be helpful, since all the three terms in the r.h.s. of the bound in Theorem 3 become smaller than their traditional counterparts. The above-mentioned conditions are very general, but at the same time very implicit. In any case, we will add a lemma to our appendix to highlight this issue for interested readers.

–(**minor comments**) 1. Yes, our SSM measure can also be used when $\epsilon=0$ (i.e. no distributional robustness). To the best of our knowledge, there are no similar theoretical treatments of this problem in the existing works. 2. Please refer to our response to Reviewer 4.

Reviewer 4:–(**major comments**) Considering reviewer's comments, first let us emphasize on some of our contributions that might have been missed during the review: we have tested our method on three different datasets and outperformed state-of-the-art in at least one of them. Also, we theoretically showed that a model set with a bounded VC-dimension is also *adversarially-learnable* (Lemma E.3), even in a *semi-supervised* scenario, where a corresponding generalization bound is given in Theorem 3. We agree with both Reviewers 2 and 4 that MSR in (C.11) is very implicit and hard to evaluate. However, please note that our framework is completely general, and thus providing a way to evaluate MSR in a general scenario might lead to solving several open problems in statistics (similar to providing a general way to evaluate VC-dimension or Rademacher complexity for any model set). For example, consider the loss function set $\boldsymbol{\Phi}=\{-\log P_\theta\left(\cdot,\cdot\right)\mid\theta\in\Theta\}$, where P_θ can be any parametric distribution family over $\mathcal{X}\times\mathcal{Y}$. Also, assume dataset is sampled from P_{θ_0}, where $\theta_0\in\Theta$. Then, it can be easily seen that the proposed risk in Theorem 1 when $\lambda=-1$, is in fact the ML estimator (which is also the optimal estimator). Characterizing MSR in this case can shed light on the sample complexity of ML in a general semi-supervised setting which is still an open problem. However, let us give a quick example of how fast MSR can be computed in some very specific and simple cases: Assume the *cluster assumption*, where data distribution P_0 is a mixture of two distributions whose supports do not overlap over $\mathcal{X}$, and correspond to only $y=-1$ and $+1$ over $\mathcal{Y}$, respectively. Consider the loss function set $\boldsymbol{\Phi}$ which is associated with a family of arbitrary binary classifiers, where for each $\phi\in\boldsymbol{\Phi}$ we have $\phi\left(\boldsymbol{X},y\right)=\infty\cdot\phi_{acc}\left(\boldsymbol{X},y\right)+\phi_{mar}\left(\boldsymbol{X}\right)$. Here, $\phi_{acc}\in\{0,1\}$ checks whether the label y matches with the positioning of $\boldsymbol{X}$ w.r.t. the classifier of ϕ, and $\phi_{mar}\left(\boldsymbol{X}\right)\in\mathbb{R}$ penalizes the margin, i.e. distance of $\boldsymbol{X}$ from the classifier's border. Now, let $\psi\subseteq\boldsymbol{\Phi}$ correspond to a subset of classifiers that classify all the data correctly ($\mathbb{E}_{P_0}\phi_{acc}=0$), but have different expected margins. Also, assume ϕ^* (the minimum loss associated with the optimal classifier) is also inside ψ. Then, some simple calculations reveal that for every $\phi\in\psi$ and any λ we have $\rho_\lambda\left(\phi\right)=0$ (C.6) and thus $\Lambda\left(\psi\right)=-\infty$ (C.10). Also, we have $\Gamma\left(\psi;\lambda\right)\geq 0$ (C.9), again for any λ, while $\mathrm{GAP}\left(\psi\right)$ (C.9) is strictly positive for any non-trivial $\boldsymbol{\Phi}$ (recall that $\phi^*\in\psi$). Considering the fact that we can have $\zeta=O\left(n^{-1/2}\right)$ according to Theorem 3, then for a sufficiently large n, $\mathrm{MSR}_{\boldsymbol{\Phi},P_0}\left(\lambda,O\left(n^{-1/2}\right)\right)$ becomes zero for any $\lambda\in\mathbb{R}\cup\pm\infty$. This result is in full agreement with the previous bounds that are specifically derived for generic learnability of statistical models when non-overlapping cluster assumption holds (For absolute learnability, at least one data point with a label is needed to decide which cluster is which).

–(**minor comments**) 1. We have rephrased the sentences to avoid any possible confusions. 2. Yes, $n=n_{\mathrm{l}}+n_{\mathrm{ul}}$. Reviewer is correct and notations w.r.t. $\boldsymbol{D}$ will be corrected. 5. The model used in our experiments is a deep neural network whose structure is completely explained in the supplementary document. Unfortunately, we cannot give more info in the main text due to the page limit. 3.4.6. We will correct all the grammatical mistakes, and also update the references.