We thank the reviewers for their feedback. We address their comments in the order of R2&R5, R2, R3, and R5.

REVIEWER #2 AND REVIEWER #5

#2 3.2, #5 3. Insight in classification in relation to [37]: Our main contribution of the paper is a framework for thinking about meta-augmentation. Although the importance of label shuffling is known in few-shot classification, we show that these results are consistent with being a special case of our meta-augmentation framework. Also, on top of [37], we go on to show differences between intra-shuffling and inter-shuffling, and that memorization overfitting and learner overfitting are both possible, but not guaranteed to occur for non-mutually-exclusive tasks.

REVIEWER #2

3.1.(i). Better linking CE-increasing augmentations to few-shot classification: In our few shot classification benchmarks, ϵ is a random variable for a uniformly sampled permutation from S_N, and $y' = g(\epsilon, y)$ is the application of the permutation. This augmentation increases the conditional entropy $H(Y'_q|X_q)$ as the amount of information required to describe Y' given X has now increased due to the unknown ϵ. This can be viewed as an encryption key because given just y' we cannot infer y, but y', ϵ recovers y exactly. We will make this link clear and add it in the camera ready version of the paper.

3.1.(ii). Mechanics of CE-increasing augmentation in MAML and CNP: We will add an explanation of the actual mechanics in MAML and CNP to the camera ready version of the paper. In general, CE-increasing augmentations force high loss if the meta-learner does not learn to adapt using the support set.

REVIEWER #3

3.A.1. Experimental setup of the Sinusoidal regression task: There was a mistake in our experiment description. In our experiments, x is always sampled from the disjoint intervals $[-5, -4.5], [-4, -3.5], \ldots, [4, 4.5]$, not uniformly from $[-5, 5]$ as mentioned in the paper. So x will never be sampled from $(-4.5, -4)$. The gaps between intervals means there is a continuous function over $[-5, 5]$ that exactly matches the piecewise function over the sub-intervals where the piecewise function is defined. The value of the continuous function outside those intervals can be arbitrary. We will correct this in the camera ready version.

3.A.2. Quantifying the effect of increase in conditional entropy on generalization Figure 8a in Appendix B displays test loss as a function of the number of discrete noise values added to y. Since we always use augmentations that satisfy the conditions in line 127, this quantifies the $H(Y|X)$ increase as $\log_2 n$ bits, where n is the number of discrete noise values used. All the three methods (CNP, MAML and MR-MAML) follow a U-shaped curve with best performance at an intermediate amount of added noise, not too low and not too high. As noted, CE-increasing augmentation is not a sufficient condition for generalization.

3.A.3. Inner-loop optimization and memorization overfitting: Yes, we intended memorization overfitting to mean the base learner relying too little on the support set. We will update the wording to clarify this.

3.B.[1,3]. wording comments: We agree with the reviewer. We will make β's meaning more explicit, and also update lines 68-69 as recommended.

3.B.2. $H(\epsilon)$ **proof**: We will add a proof to the Appendix. We noticed our original statement was not as precise as it should have been. The updated statement follows: Let ϵ be a noise variable independent from X, Y, let $g : \epsilon, Y \to Y$ be the augmentation function. Define $g_\epsilon(y)$ and $g_y(\epsilon)$ as $g(\epsilon, y)$ with ϵ or y fixed. If g_ϵ and g_y are one-to-one for all ϵ, y, then $H(Y'|X) = \min(H(Y|X) + H(\epsilon), H(uniform))$. In other words, the CE increases by $H(\epsilon)$, but $H(Y'|X)$ is upper-bounded by the max entropy distribution, the uniform distribution over the codomain. Proof: Ignoring the above edge case, $H(Y'|X) = H(Y, \epsilon|X)$, and independence gives $H(Y, \epsilon|X) = H(Y|X) + H(\epsilon|X) = H(Y|X) + H(\epsilon)$.

3.B.4 Shuffle CE-increase proof: We will add this to the appendix. A brief proof sketch follows: let ϵ be a random variable for a uniformly sampled permutation from S_N. Given any initial label distribution, augmenting with $Y' = g(\epsilon, Y)$ gives a uniform $Y'|X$, and since $H(Y'|X)$ is the highest possible conditional entropy, CE must increase unless $Y|X$ was already uniform.

3.C.[1-4], 6.F.1.: We agree with all the reviewer's suggestions under "minor concerns", and will also make sure to mention the augmentations used in Omniglot in the camera-ready version.

REVIEWER #5

3. Multiple instantiations of proposed framework We proposed and evaluated two different augmentation methods for pose regression: discrete noise and uniform noise (see Appendix B). We also found 8 different augmentations (shifting, scaling, sign flipping, etc.) to work well on the sinusoid regression task, although we did not include these results for compactness sake, and did not try all 8 on the pose regression task. We believe that the primary contribution of our work lies not in the specific augmentations we used for regression tasks, but in the general information-theoretic framework for meta-augmentation. We demonstrate this framework to be consistent across multiple datasets, models, classification and regression problems, and augmentation strategies.