@R1 **Q: Main contributions.** *Setting:* The idea to minimize the max loss is a novel contribution to meta-learning, as are our notions of task and task instance, needed for the min-max formulation to make sense. *Algorithm:* Our algorithm is an application of the Robust Stochastic Mirror-Prox algorithm [17], but the gradient computations are nontrivial because the task instances must be sampled in such a way that allows for unbiased gradients. *Convergence Analysis:* In the nonconvex setting, we show that our proposed method reaches an (ϵ, δ)-stationary point (formally defined in (9)) at a rate of $\mathcal{O}(\epsilon^{-5}, \delta^{-5})$ stochastic gradient evaluations, a novel result for the considered stochastic min-max setting. This result is achieved by utilizing novel techniques for evaluating unbiased stochastic gradients of the meta-objective (3) and characterizing their variance, e.g., to prove Theorem 3 we choose a batch size dependent on the number of iterations to show diminishing variance. *Generalization bounds:* The generalization bounds use standard results, but are the first to show convex hull generalization in meta-learning. *Experiments:* Our experiments are the first to evaluate worst-case task performance and show improvements in this metric over a canonical meta-learning algorithm. They also show that TR-MAML can generalize better to new tasks than MAML when the meta-training task distribution is skewed.

Q: Arguments for minimizing the max-loss not fully convincing. In many cases, a model that performs well across all tasks is desired, even those considered outlying or “too hard”, as well as rare tasks. MAML tends to perform poorly on these tasks since it focuses on average instead of worst-case performance, whereas TR-MAML prioritizes performance on them. TR-MAML may also generalize better since it focuses on a wider range of meta-training tasks.

Q: Generalization bounds. The convex hull generalization result highlights the advantage of minimizing the max loss in the sense that a wider range of tasks are prioritized during meta-training, and does not hold for MAML. Assuming nonnegative and M-smooth f_i and $\alpha < 1/M$, we have $0 \leq f_i(w - \alpha \nabla f_i(w)) \leq f_i(w)$, so ignoring gradient noise, the Rademacher complexity of the function after one step of gradient descent is at most the standard Rademacher complexity, which is well-known for many classes of functions. We will include this as a corollary in the revised paper.

@R2 **Q: Need to make the task.** Indeed, we must construct the tasks such that the min-max problem over them is tractable. This construction is natural in many settings, for example in the Omniglot experiment, where we aim to meta-learn how to classify characters from the same alphabet. For continuous distributions of task instances, it is reasonable to expect that the discretization into tasks would also be naturally suggested by the setting.

Q: Case showing worse performance than MAML, analyze more deeply. Assuming reference to the $(5, 1)$ Omniglot case, our claim is that minimizing the max loss leads to better average generalization in *some* cases, especially those with very different average meta-train and meta-test task instances. We suspect this is not true for the $(5, 1)$ case.

Q: Better performance on OOD cases. We provably show generalization within the convex hull of meta-training tasks (Theorem 4), but do not believe it is possible to provably show generalization beyond this.

@R4 **Q: Mean performance on Omniglot exceeds MAML**. *Regarding meta-train performance* (also concerns @R2 **compare to method with task average loss**): ‘Weighted Mean’ is the uniform average over task instances (i.e. is the surrogate for the expected loss over tasks given in Equation 1), and weighs the average accuracy on each task (alphabet) by the number of instances it contains. MAML aims to minimize this metric, and always outperforms TR-MAML on it. ‘Mean’ is the uniform average accuracy across tasks. These two statistics show that TR-MAML treats the tasks more uniformly than MAML, performing worse on the most frequent tasks (yielding a smaller ‘Weighted Mean’) but better on the rare tasks (larger ‘Mean’). *Regarding meta-test performance*: these results support our claim that TR-MAML can generalize better than MAML because it prioritizes performance on all the meta-training tasks. We include an experiment below on Mini-ImageNet in which the tasks contain a broader range of images (a random selection of image classes, instead of only characters from the same alphabet, as well as many more images per class), so we expect that MAML overfitting to the most popular tasks is less likely, but similar relative performance is observed.

@R1 R2 R5 **Q: More experiments.** Thanks for the feedback. In the revision, we’ll include experiments on Mini-ImageNet. We split the image classes into two subsets: 64 classes used for meta-training, and the remaining 36 for meta-testing. We create tasks as follows: we randomly group the 64 meta-training classes into 8 meta-train tasks, with the numbers of classes/task being $\{6, 7, 7, 8, 8, 9, 9, 10\}$. Likewise, the 36 meta-test classes are randomly split into 4 tasks, each with 9 classes/task. Each task instance is constructed by sampling 1 image each from 5 distinct classes within a task: thus, this is 5-way 1-shot problem. We meta-train for 60k iterations with a batch size of 2 task instances, and 5 steps of gradient descent for local adaptation. Our results show the Weighted Mean accuracy (aka average case over task instances) and the worst-case performance (aka worst accuracy over the tasks). The first two columns are generated by testing on *new task instances* from the meta-training classes; the second two columns are generated by testing on task instances from the previously unseen meta-test classes. We give 95% confidence intervals over 3 trials.

Table 1: Mini-ImageNet 5-way, 1-shot accuracies

		Eight Meta-Training Tasks		Four Meta-Testing Tasks	
(N, K)	Algorithm	Weighted Mean	Worst	Weighted Mean	Worst
(5,1)	MAML	$\mathbf{70.1 \pm 2.2}$	48.0 ± 4.5	$46.6 \pm .4$	$44.7 \pm .7$
	TR-MAML	63.2 ± 1.3	$\mathbf{60.7 \pm 1.6}$	$\mathbf{48.5 \pm .6}$	$\mathbf{45.9 \pm .8}$