We thank all the reviewers for their insightful reviews. We are going to fix all the reported typos in the final version.

Reviewer 1

"why exactly AMAX *-PEM and* QUAD *-PEM did not converge"* In the Search and Rescue experiments, both AMAX -PEM and QUAD -PEM actually converged to a policy while training. The failures reported in the experiment table refer to the configurations in which the agents do not complete the problem within **????????** steps and they only occur when these models (AMAX -PEM and QUAD -PEM) are evaluated on a different scenario than the one seen in training.

When this kind of failure happens at evaluation time, the algorithm starts oscillating between alternative two or more possible assignments, thus preventing some tasks to be solved (i.e., some victims are never rescued). We hypothesize that this kind of behavior is due to the fact that the receptive fields of the convolutional layers contains a different number of agents and tasks than during training on average, which can cause over or under saturation of the filters, since in this case we would be significantly out of the training support.

Reviewer 2

"directly learn assignment policies for test scenarios" In the attempt of identifying the best performance achievable in different SC scenarios, we ran a test experiment in one of the hard scenarios (w65v67), with a training budget twice as large as the one allocated to the smaller scenario we originally trained on (w15v17). None of the models we introduce in the paper managed to achieve a win-rate above 5%, which is significantly worse than the performance obtained by generalizing from simple to hard. It is indeed possible that the performance we achieved through generalization could be matched or even improved, but it would require much larger computation budgets. Also note that the quadratic problem gets bigger and bigger, and while this does not prevent to run in real-time at evaluation time it does slow down the training significantly.

Reviewer 3

Does the algorithm can achieves the performance of the current state of the art algorithms after continuing training? We do not claim to obtain the best performance achievable in the bigger scenarios in SC, as a topline is not available. As mentioned to Rev.2, achieving a satisfactory performance in hard scenarios seems extremely hard, and to the best of our knowledge our generalization approach is the most promising direction to solve complex MAC problems, which are currently out of reach of state-of-the-art algorithms.

Does the algorithm has the ability to continue training on new tasks? This is an interesting question and a direction worth investigating as future work. In this paper, we focused on enabling zero-shot transfer, which we believe is of practical interest, for example when you can't afford to train in the target environment.

...the model with more inductive bias can not be used. Does this affect the final performance of the algorithm after it has been transfered to a new task? Although we did not run enough tests to provide a conclusive answer, we believe that the DM should still be superior to PEM even when fine-tuning after generalization is performed. While PEM may allow expressing richer scoring functions, our results show that coordination procedures such as LP or QP are sophisticated enough to allow representing rich coordination patterns.

Correlated exploration... effectiveness of this method is not verified in the experiment. The idea of using auto-correlated noise for exploration in continuous action spaces is not novel. Although its application to MAC may be novel, it has been used in DDPG before [Lillicrap et al.], and thus we did not put too much focus on it. As a concrete example on the necessity of this, consider the search and rescue task: the reward is completely non-informative (-1 at each step) and does not guide the agent towards solving all the tasks. This creates an exploration problem since uncorrelated exploration is not likely to solve all the tasks even once because the agents will keep oscillating between tasks. As a result, in our experiments, we have never been able to learn any meaningful policy without it.

References

Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning.