Single-head OOD Experiments

The single-head gaussian experiments tell us

- Naively combining two datasets can increase the risk on target task
- However, weighted combination of the datasets guarantees that the risk decreases.
- We also get diminishing returns as the tasks become more dissimilar.

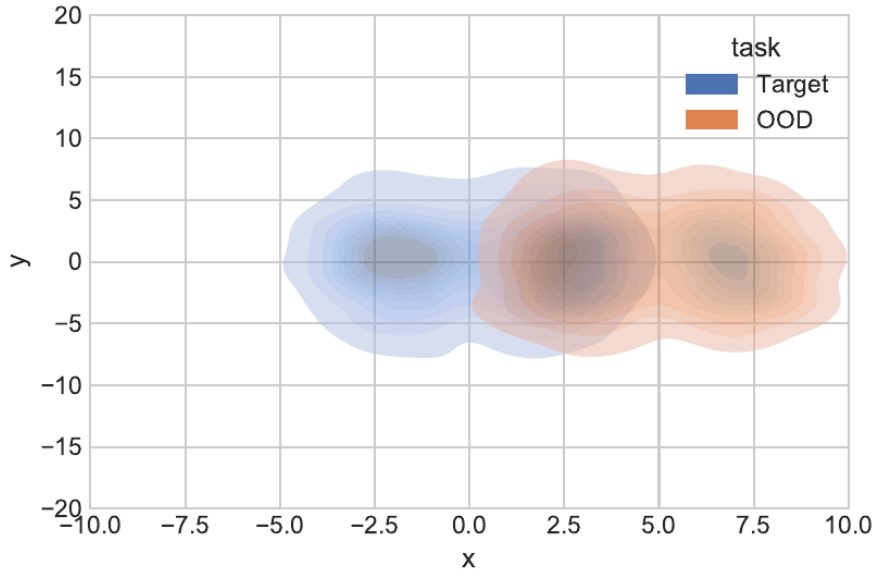
Accuracy on task [2, 3]

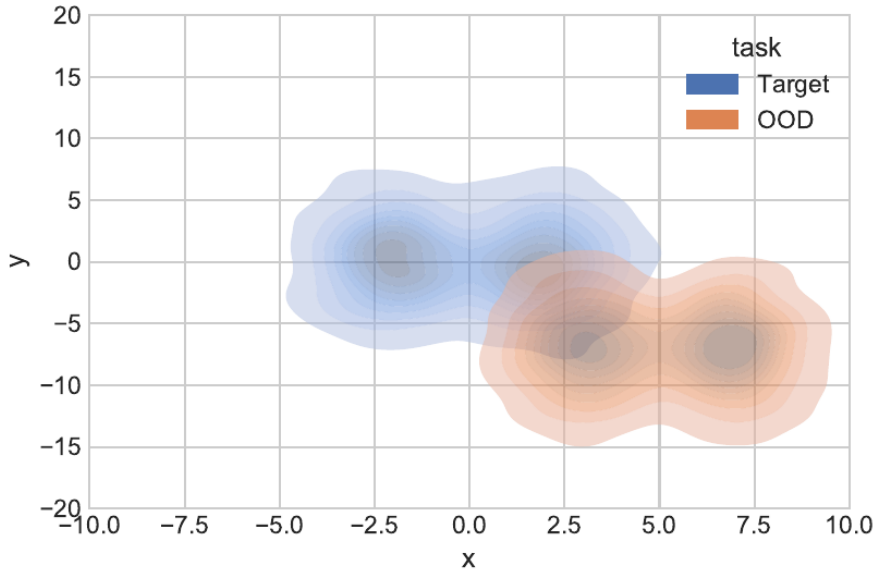
Samples	Acc
50	34.5
125	29.25
250	25.94
500	23.29
1000	20.49
1500	19.79

Lower accuracies are probably due to data augmentation but that shouldn't affect the trends.

Multi-head

Where does multi-head fit into all this? The key idea is all tasks share a low-dimensional representation.





To understand this idea, we consider tasks constructed using two-dimensional Gaussians

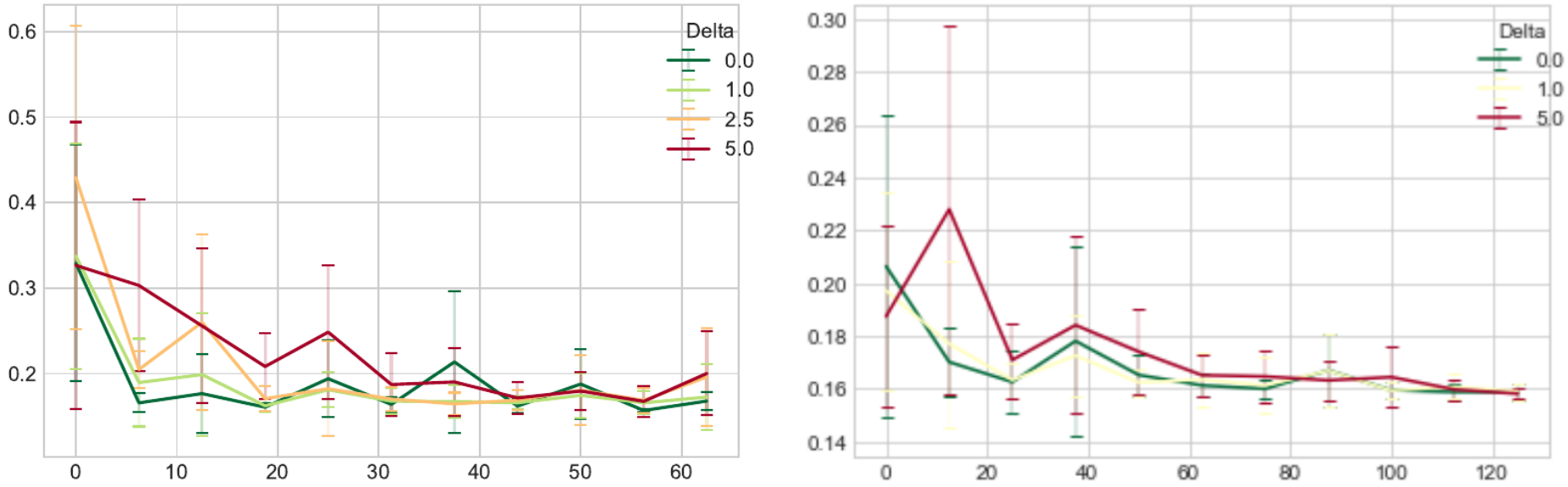
Multi-head: Gaussian Experiments

We want to find a 1-dimensional embedding that is useful to both tasks. We consider the following neural net architecture

$$\text{input} \to \text{FC}(2, 100) \to \text{FC}(100, 1) \to \text{FC}(1, 1)_i$$

Multi-head is useful as long as the OOD and target task are simple translations of each other. In this case, both tasks share an optimal 1-dimensional embedding.

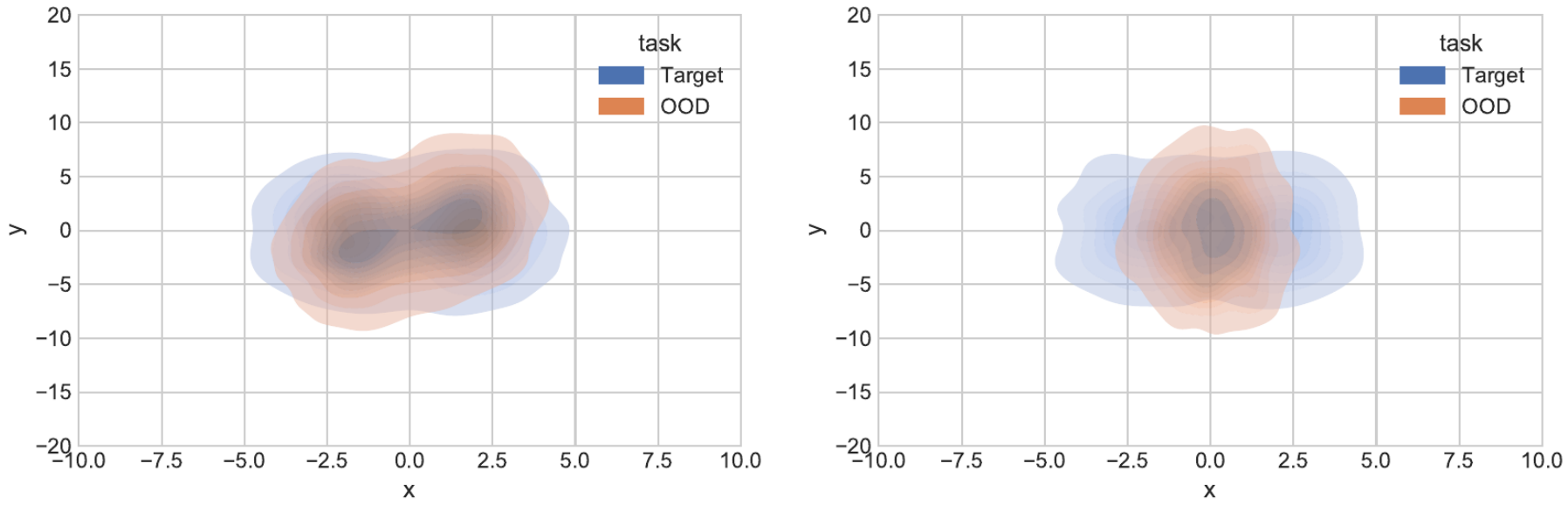
Multi-head: Translated OOD task



Left(n=2) and right(n=8). In both cases, the accuracy dips with more OOD samples. This happens regardless of the value of Δ.

Multi-head: Rotated OOD task

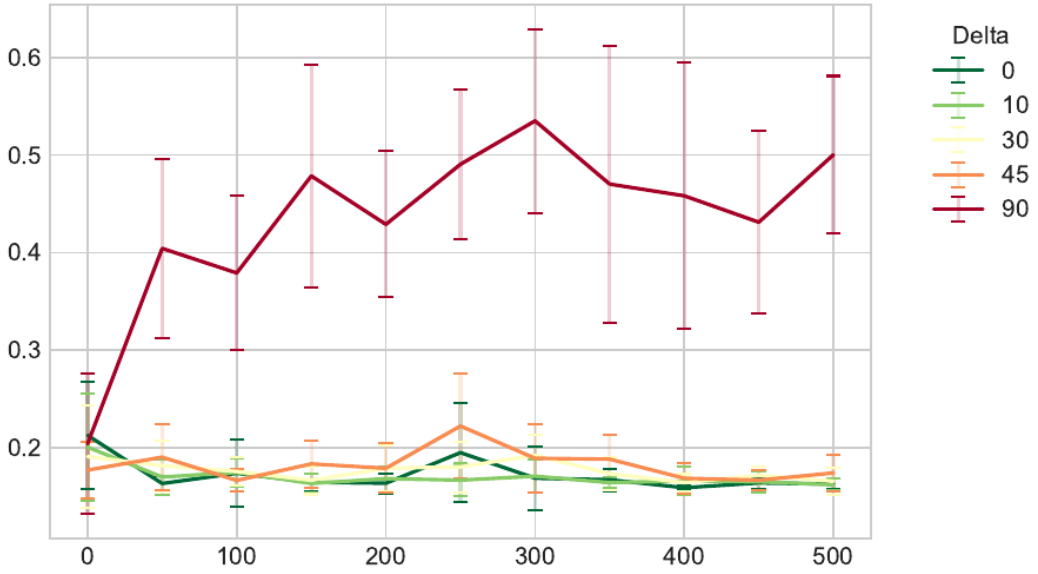
Next, we consider rotated OOD-tasks



Caption: (Left) 45 degree rotation and (right) 90 degree rotation

Here there exists no optimal 1D embedding.

Multi-head: Rotated OOD tasks



In general, Baxter's model only works if there exists a $\mathbb{R}^k$ dimensional embedding that as useful to separate the classes from any single task. Such an embedding need to exist for dissimilar tasks.

More questions on Singlhead/Multihead

- How does more tasks affect both models?
- Weighted sampling for multihead?

Next?

- Single-head and multi-head are both useful
  - We can understand augmentations as a version of single-head
- Do tasks like CIFAR100 have this low-dimensional embedding. How small is this dimension (can we estimate it?).
- How do we think of a task with more/less classes. Is it more informative if we have more classes?