We have access to data from two tasks: a target task and an OOD task. We study how task-relatedness and the choice of algorithm influence the generalization error on the target task.

# 1 Single-head results

The first model we consider is the single-head learner. Given data from two tasks, we consider the union of all data points, and train a single hypothesis $h$ on both tasks.

**Data from an OOD task can both improve or worsen the generalization error on the target task**: To illustrate this point, we consider a binary classification problem with samples of either class originating from Gaussians with different means. Formally

$$p(x, y) = \begin{cases} \mathcal{N}(+\mu + \Delta, 1) & \text{if } y = 1 \\ \mathcal{N}(-\mu + \Delta, 1) & \text{if } y = -1 \end{cases}$$

We set $\Delta = 0$ for the target dataset and $\Delta > 0$ for the OOD task.

We consider the Fisher-discriminant model with equal prior probabilities ($p_1 = p_0$). We train a single hypothesis using data from both the OOD and target tasks. However, we evaluate only on the target task.

The OOD task helps if $\Delta$ is small and hurts if $\Delta$ is large (see fig. 1). Furthermore, if $\Delta$ is large, more OOD data increases the generalization error on the target task.

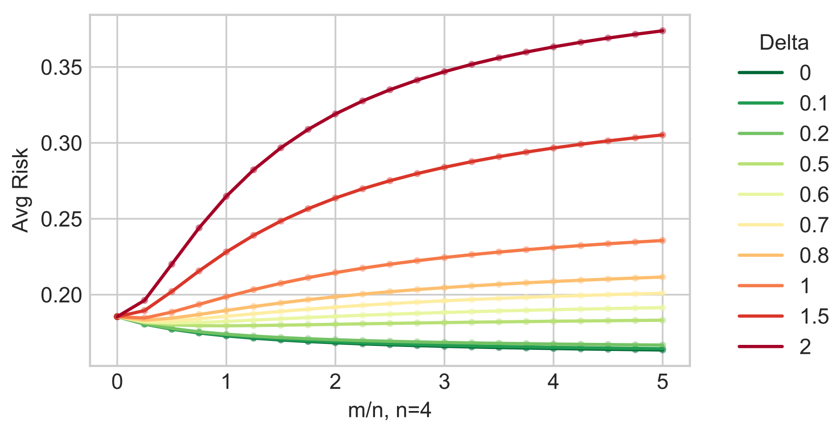


Figure 1: We consider $n = 4$ from the target task and vary the number of samples from the OOD task (x-axis). The y-axis plots the average population risk (computed using an analytic expression). The OOD task hurts performance for $\Delta$ as small as 0.5.

From fig. 1, an OOD task is helpful if it is close to the target task. The relatedness of two tasks depend on the distance between the probability distributions $p_{ood}(x, y)$ and $p_{target}(x, y)$.

**If we weight the data points from both tasks, then there exists a weighting scheme that guarantees that the generalization of the target task always improves with OOD data.** Why does this result hold? If we had the flexibility to weight the OOD and target samples, then we can choose to discard OOD samples if they hurt the generalization error of the target task.

In fig. 2, we consider the same model as earlier (Fisher discriminant) but the algorithm sees a weighted version of the samples. We weight the target samples by $\alpha$ and the OOD samples by $1-\alpha$ where $\alpha \in [0, 1]$.

The initial set of experiments considered the empirical distribution

$$\hat{S} = \sum_{x \in D_{ood} \cup D_{target}} \frac{1}{m_{ood} + m_{target}} \delta_x$$

We instead consider the empirical distribution to be

$$\hat{S}_\alpha = \sum_{x \in D_{target}} \frac{\alpha}{(1-\alpha) m_{ood} + \alpha m_{target}} \delta_x + \sum_{x \in D_{ood}} \frac{1-\alpha}{(1-\alpha) m_{ood} + \alpha m_{target}} \delta_x$$

The flexibility of $\alpha$ guarantees better or identical empirical risk, with more OOD data as seen in fig. 2. For dissimilar tasks, we select an $\alpha$ close to 1 while for similar tasks, we select an $\alpha$ close to $0.5$.

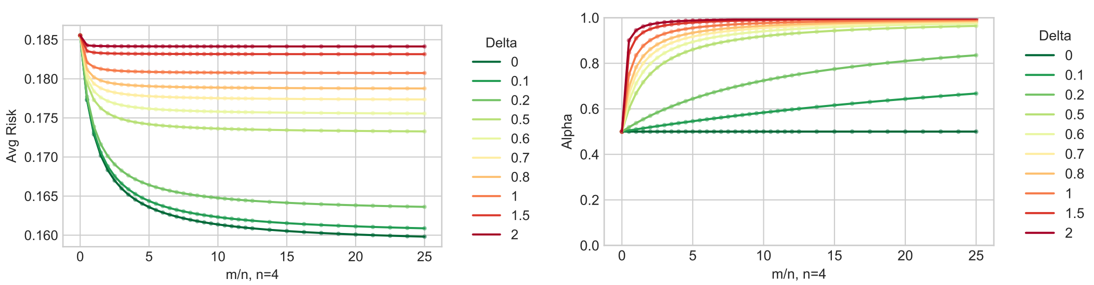


Figure 2: **(Left)**: The generalization error on the target task. More OOD data never hurts the generalization error of the target task **(Right)**: The values of the weight $\alpha$, used to visualize the curve on the left. The target samples are weighted by $\alpha$ and the OOD samples are weighted by $1-\alpha$.

**Lemma 1.1.** *Consider any algorithm $A : \mathcal{S} \mapsto h$, that produces a hypothesis from a collection of samples. Let the samples be a mixture of data points from both an OOD and target task, weighted by $1-\alpha$ and $\alpha$ respectively. Then there exists an $\alpha$ such that the generalization error with OOD data is never worse than the generalization error without any OOD data. This holds true regardless of the amount of OOD data.*

*Proof.* The proof follows from the inequality

$$\begin{aligned}\mathcal{E}\left[A(\hat{S}_{target})\right] &= \mathcal{E}\left[A\left(1\times\hat{S}_{target}+(1-1)S_{ood}\right)\right]\\ &\geq \min_{\alpha}\ \mathcal{E}\left[A(\alpha\hat{S}_{target}+(1-\alpha)S_{ood})\right]\end{aligned}$$

□

In order to prove a more general result regarding the monotonic decrease in error with more OOD samples, we need to make regularity assumptions about algorithm $A$.

## 2 A Theory of Learning from Domains

This section discusses results from Ben-david's work. The first salient result is as follows:

**Theorem 2.1.** *Let $\mathcal{H}$ be a hypothesis space of VC dimension $d$. If $\mathcal{U}_s$ and $\mathcal{U}_t$ are samples of size $m'$ from $D_s$ and $D_t$ respectively, then with probability atleast $1-\delta$ (over choice of the samples) for every $h \in H$.*

$$e_T(h) \leq e_S(h) + \frac{1}{2}\hat{d}_{H\Delta H}(\mathcal{U}_s,\mathcal{U}_t) + 4\sqrt{\frac{2d\log(2m')+\log(2/\delta)}{m'}} + e_S(h^*) + e_T(h^*)$$

The theorem states that we can never asymptotically guarantee zero error on the target task even if we have 0 error on the source task. This margin depends on the similarity of tasks, more specifically the $H\Delta H$ divergence.

However, if we have few target samples, then a related source task can help with the generalization error of a target task. This is the next salient result from Ben-david's paper. We consider an algorithm uses the loss function $\alpha\hat{\epsilon}_T+(1-\alpha)\hat{\epsilon}_S$ and trains on a combination of the source and target data. The upper-bound for the generalization error of the target under this algorithm is given by the following theorem.

**Theorem 2.2.** *Let $\mathcal{H}$ be a hypothesis space of VC dimension $d$. Let $\mathcal{U}_s$ and $\mathcal{U}_t$ be unlabeled samples of size m' each drawn from $D_s$ and $D_t$. Let $S$ be a labeled sample of size $m$ generated by drawing $\beta m$ points from $D_T$ and $(1-\beta)m$ points from $D_s$. If $\hat{h} \in \mathcal{H}$ is the empirical minimizer of $\hat{\epsilon}_\alpha(\hat{h}) = \alpha\hat{\epsilon}_T + (1-\alpha)\hat{\epsilon}_S$ on $S$, then with a probability of $1-\delta$*

$$\begin{aligned}e_T(\hat{h}) &\leq e_T(h_T^*)\\ &\quad + 4\sqrt{\frac{\alpha^2}{\beta}+\frac{(1-\alpha)^2}{1-\beta}}\sqrt{\frac{2d\log(2(m+1))+2\log(8/\delta)}{m}}\\ &\quad + 2(1-\alpha)\left(\frac{1}{2}\hat{d}_{H\Delta H}(U_s,U_t)+4\sqrt{\frac{2d\log(2m')+\log(8/\delta)}{m'}}+e_S(h^*)+e_T(h^*)\right)\end{aligned}$$

The second term is a variance or sample complexity term. The third term is a measure of distance between the two tasks which we refer to as $A$. If we define $D = \frac{\sqrt{d}}{A}$ and approximate the second square-root in the second term by $\sqrt{\frac{d}{m}}$, we get

$$e_T(\hat{h}) \leq 4\sqrt{\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{1-\beta}}\sqrt{\frac{2d}{m}} + 2(1-\alpha)A.$$

Ben-david et al. derive the expression

$$\alpha^*(m_T, m_S; D) = \begin{cases} 1 & m_T \geq D^2 \\ \frac{m_T}{m_T+m_S}\left(1 + \frac{m_s}{\sqrt{D^2(m_s+m_t)-m_s m_t}}\right) & m_T \leq D^2 \end{cases}$$

Hence if you have a hypothesis space with small-VC dimension or if the distance between source and target is large, then we train only on the target. Otherwise, it is best to use some samples from the source dataset too. In practice, we can approximate each term in the above equation using a finite number of samples.

# 3 Multi-head Model

We consider the 1D gaussian dataset to show the utility of multi-head. We setup the multi-head model in the following fashion: 1) Train $w$ using both the OOD and the target task 2) Train a task-specific threshold $c$.

## 3.1 Revisiting Fisher's discriminant

Any presentation of Fisher's discriminant – even in a classic textbook like Bishop – assumes equal prior probabilities for both classes. If we know that the classes are imbalanced, we can incorporate this into our model using the parameter $p$ (like in Ashwin's notes). This has the effect of shifting the threshold away from the majority class mean.

If $p = 1/2$, then the 1D classifier reduces to

$$h = \mathbf{1}\left[x > \frac{\hat{m}_{0,t} + \hat{m}_{1,t}}{2}\right]$$

i.e., the projection $w$ does not feature in the final hypothesis.

If we include the prior probability like in Ashwin's model, then the hypothesis for the 1D case is

$$h = \mathbf{1}\left[x > \frac{\hat{m}_{0,t} + \hat{m}_{1,t}}{2} + \frac{\sigma^2}{\hat{m}_1 - \hat{m}_0}\log\left(\frac{1-p_t}{p_t}\right)\right]$$

The first term is identical to the case where $p = 1/2$. The second term is a shift in the threshold due to class imbalance.

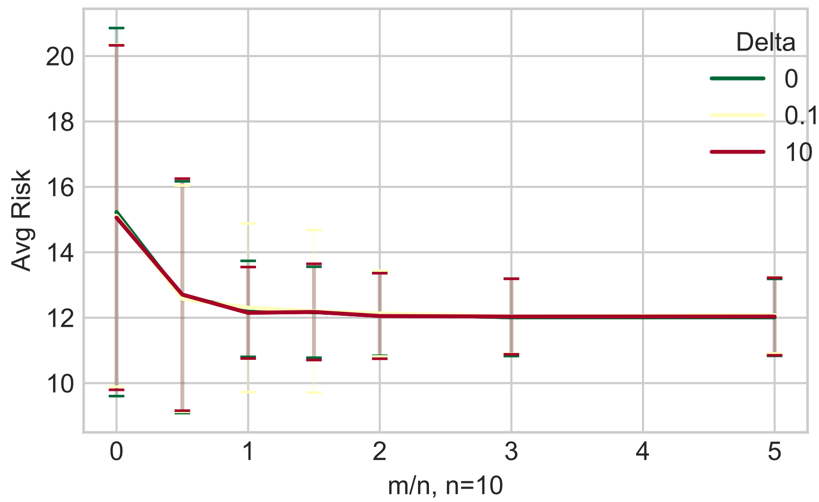


Figure 3: Plot of the number of OOD samples (x-axis) against the population risk (y-axis). The curves are nearly identical for all values of $\Delta$.

The OOD data improves the estimate of $\hat{m}_1 - \hat{m}_0$. Both $\hat{m}_{1,t} - \hat{m}_{0,t}$ and $\hat{m}_{1,ood} - \hat{m}_{0,ood}$ are identical in our toy datasets and hence we can estimate them together. As a result, the second term converges to the optimal value at a faster rate.

In fig. 3, we evaluate the multi-head learner in this setup. We consider 100,000 runs where we sample $n = 10$ target samples and $m$ OOD samples. We set $p = 0.8$, which over-samples the class $y = 1$. We use the multi-head setup and compute a shared $w$. Then, we compute a separate threshold $c_t$ using samples from only the target task.

As long as the OOD task is a translation of the target task, it always helps to have more OOD data in the multi-head setup. We hypothesize that if the distance between the means are different for the OOD and target tasks, then the OOD task is detrimental to the generalization error of the target task.

To test this hypothesis, consider the target task to be $\mathcal{N}(\pm 1, 1)$ and the OOD task to be $\mathcal{N}(\pm \mu, 1)$. The OOD task is no longer a translation of the target task; Instead, it is a scaled version of the target task along the x-axis. Note that a large or small value of $\mu$ hurts the generalization on the target task. **Hence, there exists tasks where the OOD task is not beneficial in the multi-head setup**.

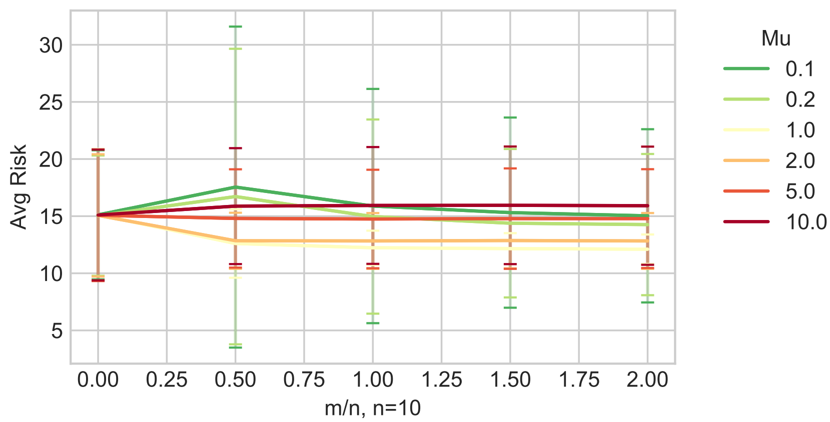


Figure 4: We consider the target task to have class means to be $\pm 1$, while the OOD task has class means to be $\pm\mu$. The variance of all gaussians is set to 1. The OOD task hurts the generalization error if $\pm\mu_{ood}$ is different from $\pm\mu_{target}$.