The Value of Out-of-Distribution Data

Ashwin De Silva[1,†], **Rahul Ramesh**[2,†],
Carey E. Priebe[1], **Pratik Chaudhari**[2,*], **Joshua T. Vogelstein**[1,*]
[1]Johns Hopkins University, [2]University of Pennsylvania
[†]Equal Contribution, [*]Equal Contribution

JOHNS HOPKINS UNIVERSITY

Penn UNIVERSITY of PENNSYLVANIA

Does more data always help?

Generalization error for a desired target task always improves with more in-distribution data.

But real data is often **heterogeneous**. Even a curated dataset can contain out-of-distribution (OOD) samples.

Does more data always help?

Generalization error for a desired target task always improves with more in-distribution data.

But real data is often **heterogeneous**. Even a curated dataset can contain out-of-distribution (OOD) samples.

For a model trained on such data, we expect the generalization error on the target task to be *monotonic* in the number of OOD samples.

Modeling heterogeneity in a dataset

Suppose dataset D consists of n target samples and m OOD samples without the knowledge of sample identities.

We seek a hypothesis h that minimizes the generalization error on the target task $e_t(h)$.

Modeling heterogeneity in a dataset

Suppose dataset D consists of n target samples and m OOD samples without the knowledge of sample identities.

We seek a hypothesis h that minimizes the generalization error on the target task $e_t(h)$.

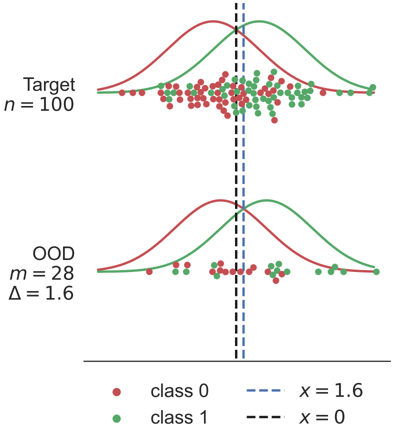
The hypothesis is selected by minimizing the empirical loss,

$$\hat{e}(h) = \sum_{i=1}^{n+m} \ell(h(x_i), y_i)$$

An example using Fisher’s Linear Discriminant

The target and OOD tasks are both gaussian mixture models.

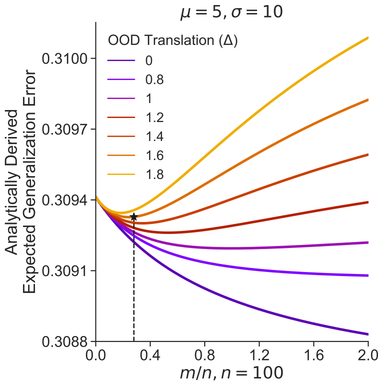
We consider a family of OOD task distributions which are translations of the target distribution.



An example using Fisher's Linear Discriminant

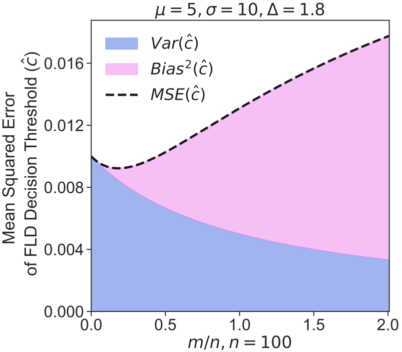
OOD data from the same distribution can both improve or deteriorate the target generalization depending on the number of OOD samples.

Generalization error on the target task can be **non-monotonic** in the number of OOD samples.



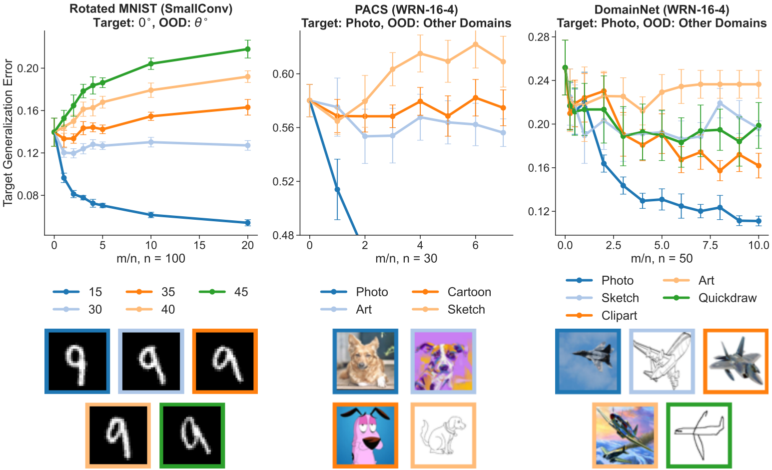
Why does non-monotonicity occur?

More OOD samples decrease the variance but increase the bias. The trade-off depends on the distance between the two distributions.



Non-monotonic trends also occur in popular benchmark datasets

Non-monotonic trends occur due to geometric, semantic nuisances and distribution shifts.



Exploiting the non-monotonic trends in generalization error

Assuming that the target and OOD samples are separable, we consider the objective

$$\hat{e}_\alpha(h) = \alpha \hat{e}_t(h) + (1-\alpha)\hat{e}_o(h).$$

We can compute the optimal α using an upper bound of the generalization error[1]

$$\alpha^* = \min\left(1, \frac{n}{n+m} \times \left(1 + \sqrt{\frac{m^2}{4\rho^2(n+m) - nm}}\right)\right).$$

ρ is the ratio between task distance and model capacity.

[1] Ben-David et al., "A theory of learning from different domains".

Exploiting the non-monotonic trends in generalization error

Assuming that the target and OOD samples are separable, we consider the objective

$$\hat{e}_\alpha(h) = \alpha \hat{e}_t(h) + (1-\alpha)\hat{e}_o(h).$$

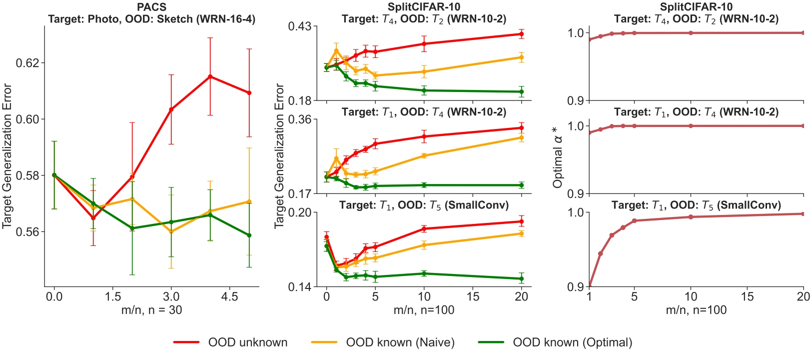
We can compute the optimal α using an upper bound of the generalization error[1]

$$\alpha^* = \min\left(1, \frac{n}{n+m} \times \left(1 + \sqrt{\frac{m^2}{4\rho^2(n+m) - nm}}\right)\right).$$

ρ is the ratio between task distance and model capacity.
In practice we consider, α to be a hyper-parameter.

[1] Ben-David et al., "A theory of learning from different domains".

Exploiting the non-monotonic trends in generalization error



Concluding Thoughts

Generalization error can be a non-monotonic function of the number of OOD samples.

A weighted objective between the OOD and target samples can mitigate this non-monotonicity.

For more details and experiments, check out our paper on arXiv
arxiv.org/abs/2208.10967