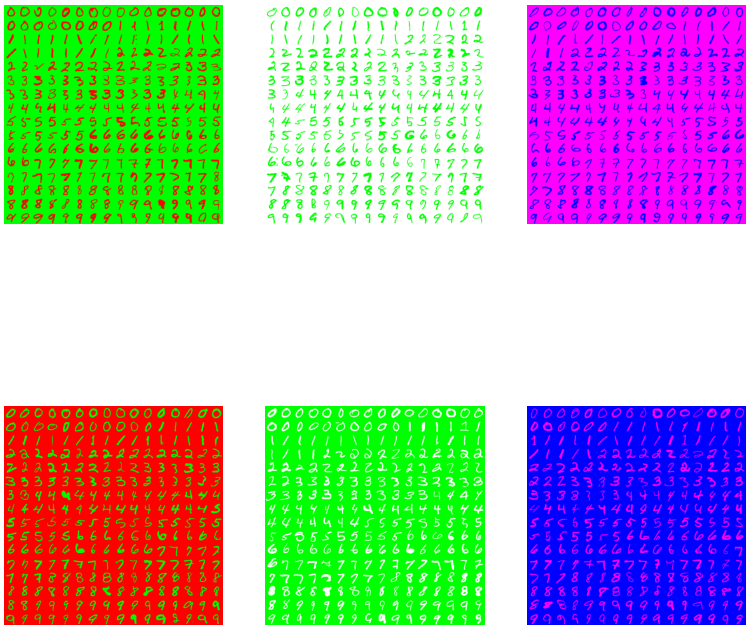
—SUPPLEMENTARY MATERIAL—
GENERALIZED INVARIANT RISK MINIMIZATION: RELATING ADAPTATION AND INVARIANT REPRESENTATION LEARNING

Extended Colored MNIST Task

Example training and testing environments for the extended colored MNIST task. Background and foreground can optionally be correlated with the label. We utilize this task as an intermediate dataset between the original colored MNIST dataset, and more complex data distributions like SVHN/Synth Digits, PACS or VLCS.



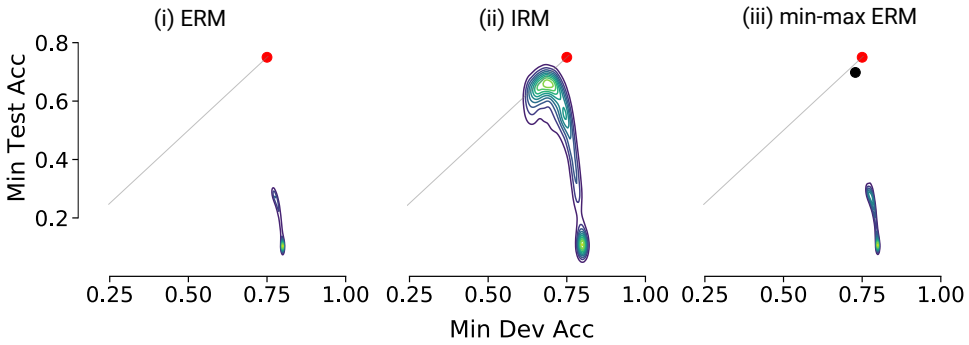
IRM Reproduction Study I

Reproduction study of the original IRM experiments with train environments $\mathcal{E}_t = \{0.1, 0.2\}$, and test on $e = 0.9$, to visualize our motivation for proposing hyperparameter selection schemes $S_{3a,b}$ in the main paper. Selecting hyperparameters based on the worst case training performance (upper table) yields only slight improvements over ERM. Selection based on the test performance is necessary (lower table) to observe the originally reported gain.

Plotting all samples using a contour plot (red dot indicates maximum possible performance at 75%; training is done with 25% label noise) demonstrate a slightly negative slope between the train and test set accuracies, making it impossible to select good hyperparameters based on the training set accuracies. This motivates alternative selection schemes based on the regularizer values of IRM as outlined in our proposal.

$N = 89889$	Epoch	worst train		best train		test	
Model (on train)	≥ 500	mean	std	mean	std	mean	std
erm	500	0.8052	0.0043	0.8988	0.0033	0.1034	0.0054
min-max erm	500	0.8055	0.0044	0.9016	0.0029	0.0978	0.0061
irm	800	0.7926	0.0014	0.8643	0.0028	**0.2520**	0.0077

$N = 89889$	Epoch	worst train		best train		test	
Model (on test)	≥ 500	mean	std	mean	std	mean	std
erm	900	0.7251	0.0606	0.7661	0.0951	0.4236	0.1477
min-max erm	500	0.7263	0.0754	0.7703	0.1122	0.4232	0.2144
irm	500	0.6933	0.0013	0.6993	0.0033	0.6840	0.0059



IRM Reproduction Study II

We now modify the training environments to $\mathcal{E}_t = \{0.0, 0.05, 0.075, 0.3, 0.35, 0.4\}$, and test on $e \in \{0.15, 0.5, 0.7, 0.8, 0.9\}$. The training now includes environments with lower correlation between color and label than between digit shape and label (75 %, cf. red dot); we kept all search ranges except for the maximum number of epochs (increased to 1000) according to the original IRM experiment.

Selecting hyperparameters based on the worst case training performance (upper table) now yields a comparable performance to selection based on the test set, and the contour plots of all considered samples reveal a slightly positive correlation.

Note that it is crucial to consider a better baseline than ERM in this case: Using the min-max formulation of ERM, i.e., minimizing the worst case expected error across training environments, results in effectively training the model on the environment $e = 0.4$ which the weakest correlation between color and label, improving the overall performance. In our protocol, we reflect this by considering the optimal weighting of environment risks for the ERM optimizer.

$N = 38033$	Epoch	worst train		best train		test	
Model (on train)	≥ 100	mean	std	mean	std	mean	std
erm	200	0.8349	0.0115	0.9134	0.0043	0.7307	0.0255
min-max erm	700	0.8629	0.0046	0.8987	0.0066	0.8123	0.0036
irm	800	0.8683	0.0034	0.8859	0.0058	0.8443	0.0052

$N = 38033$	Epoch	worst train		best train		test	
Model (on test)	≥ 100	mean	std	mean	std	mean	std
erm	200	0.8349	0.0115	0.9134	0.0043	0.7307	0.0255
min-max erm	200	0.8587	0.0105	0.8921	0.0078	0.8262	0.0266
irm	200	0.8605	0.0015	0.8660	0.0015	0.8540	0.0040