A. Experimental Results with 10 Repetitions

Table 1: Mean accuracy (%) of competing methods on four test environments in simulation study with 10 repetitions.

Env partition	(p_s^-, p_s^+)	(0.999, 0.7)				(0.999, 0.8)				(0.999, 0.9)			
	$p_v(t)$	0.9		0.8		0.9		0.8		0.9		0.8	
	TEST ACC	Mean	Worst	Mean	Worst	Mean	Worst	Mean	Worst	Mean	Worst	Mean	Worst
False	ERM	76.22	58.81	59.80	25.95	69.34	43.06	55.96	15.60	60.62	23.30	53.10	8.04
	EIIL	39.43	18.22	64.95	48.45	50.26	47.02	68.86	54.91	61.33	52.70	69.82	58.58
	HRM	76.52	59.78	59.98	26.97	69.87	44.49	56.40	16.85	60.57	23.46	53.16	8.37
	TIVA	82.54	76.74	75.82	70.97	81.53	73.05	69.78	56.23	71.42	49.95	59.47	30.77
	ZIN	87.70	85.86	**78.33**	76.60	86.78	84.86	77.42	75.12	83.42	78.62	74.03	67.45
	MinMax-TV-ℓ_1	**88.67**	**87.83**	78.14	**76.68**	**88.55**	**87.62**	**78.74**	**77.56**	**87.01**	**85.74**	**77.31**	**74.54**
True	groupDRO	72.42	54.90	63.74	43.37	71.09	51.60	62.78	40.21	69.67	47.72	61.81	36.44
	IRM	87.84	86.20	78.33	76.58	86.84	84.42	77.48	74.80	84.16	77.89	74.53	68.72
	IRM-TV-ℓ_1	**88.03**	**86.40**	**78.49**	**76.88**	**87.10**	**84.90**	**77.95**	**75.65**	**84.84**	**80.06**	**75.55**	**70.77**

Table 2: Standard deviation (%) of competing methods on four test environments in simulation study with 10 repetitions.

Env partition	(p_s^-, p_s^+)	(0.999, 0.7)				(0.999, 0.8)				(0.999, 0.9)			
	$p_v(t)$	0.9		0.8		0.9		0.8		0.9		0.8	
	TEST ACC	Mean	Worst	Mean	Worst	Mean	Worst	Mean	Worst	Mean	Worst	Mean	Worst
False	ERM	1.17	2.06	1.04	2.06	1.23	2.47	0.76	1.42	1.10	2.01	0.62	0.95
	EIIL	1.52	3.18	1.46	1.72	1.70	3.09	1.43	2.26	2.46	1.99	1.58	2.04
	HRM	1.35	2.71	0.94	2.43	0.75	1.83	0.71	2.33	0.84	1.29	0.45	0.93
	TIVA	6.12	11.09	3.55	7.18	4.83	9.19	6.46	13.96	5.18	10.34	6.32	13.66
	ZIN	1.05	2.19	1	1.43	1.67	2.73	1.43	2.13	3.52	6.72	2.09	3.86
	MinMax-TV-ℓ_1	0.57	0.60	0.84	1.03	0.45	0.50	0.67	0.74	1.28	1.66	0.65	1.13
True	groupDRO	8.45	18.08	6.99	16.84	8.42	19.03	6.71	17.27	8.27	18.51	6.52	16.45
	IRM	0.82	2.01	0.91	1.49	1.16	2.34	1.82	3.01	1.98	4.11	3.14	4.52
	IRM-TV-ℓ_1	0.86	2.08	0.74	1.33	1.35	2.67	1.24	2.22	2.19	4.77	2.92	4.31

Table 3: Average mean squared error of competing methods in house price prediction with 10 repetitions.

		Average			STD		
Env Partition	Methods	Train	Test	Worst	Train	Test	Worst
False	ERM	0.1057	0.4409	0.6206	0.0017	0.0435	0.0641
	EIIL	0.1103	0.3939	0.5581	0.0020	0.0305	0.0460
	HRM	0.5578	0.5949	0.7250	0.0593	0.0025	0.0052
	TIVA	0.2575	0.4418	0.6145	0.0002	0.0019	0.0062
	ZIN	0.2241	0.4293	0.6198	0.1137	0.1994	0.2869
	MinMax-TV-ℓ_1	0.2168	**0.3395**	**0.4983**	0.0652	0.0638	0.0958
True	groupDRO	0.1271	0.7358	1.0611	0.0029	0.0877	0.1287
	IRM	0.5663	0.8168	1.1168	0.1389	0.3115	0.4511
	IRM-TV-ℓ_1	0.3261	**0.4420**	**0.6096**	0.1279	0.2503	0.3342

Table 4: Mean accuracy (%) of competing methods on CelebA with 10 repetitions.

		Mean			STD		
Env Partition	Methods	Train	Test	Worst	Train	Test	Worst
False	ERM	0.6376	0.6399	0.6205	0.1445	0.1416	0.1416
	EIIL	0.5912	0.5815	0.5422	0.0874	0.0848	0.1023
	LfF	0.5750	0.5773	0.5618	0.0012	0.0024	0.0057
	TIVA	0.6436	0.6423	0.6163	0.0168	0.0199	0.0147
	ZIN	0.7832	0.7673	0.7619	0.0116	0.0087	0.0085
	Minmax-TV-ℓ_1	0.8512	**0.8368**	**0.8145**	0.0092	0.0033	0.0043
True	groupDRO	0.8150	0.8119	0.7927	0.0031	0.0048	0.0074
	IRM	0.8559	0.8254	0.8075	0.0149	0.0135	0.0099
	IRM-TV-ℓ_1	0.8479	**0.8347**	**0.8121**	0.0059	0.0048	0.0067

Table 5: Mean accuracy (%) of competing methods on Landcover with 10 repetitions.

	Mean				STD			
Methods	Train	IID Test	OOD Test	Worst	Train	IID Test	OOD Test	Worst
ERM	0.6661	0.6644	0.6154	0.6080	0.0182	0.0156	0.0092	0.0077
EIIL	0.6411	0.6381	0.6043	0.5953	0.0166	0.0172	0.0088	0.0121
LfF	0.5812	0.5789	0.5576	0.5507	0.0273	0.0245	0.0196	0.0193
TIVA	0.6749	0.6479	0.5202	0.5146	0.0028	0.0062	0.0098	0.0109
ZIN	0.7002	0.6942	0.6222	0.6187	0.0109	0.0114	0.0109	0.0121
Minmax-TV-ℓ_1	0.7359	**0.7195**	**0.6377**	**0.6325**	0.0069	0.0063	0.0117	0.0137