

[Aug 11th, 2024] Update Hiroki / OOD Experiments

☰ [June 24th, 2024] / Hiroki + Lyle

Baseline

<https://openreview.net/pdf?id=ipe0IMglFF>

H.4 Best OOD Performance Comparison against with Existing Benckmark

In order to confirm the soundness of our experiments, we compared our results with existing benchmarks to see how well they actually performed, in addition to the hyperparameter search ranges in the previous section. In particular, we compared our experimental results with those of DomainBed ([Gulrajani & Lopez-Paz, 2021](#)), an existing oracle benchmark that uses Adam.

Table 15: OOD accuracy (%) comparison of our experimental results with the benchmark results reported in DomainBed ([Gulrajani & Lopez-Paz, 2021](#))

Dataset	OOD Domain	Existing Benchmark Results(Adam)	Our Results(Adam)
ColoredMNIST	0.9	30.0 $\pm$ 0.3	73.92
RotatedMNIST	0	96.0 $\pm$ 0.2	96.40
VLCS	C	97.7 $\pm$ 0.3	99.36
PACS	A	87.8 $\pm$ 0.4	89.30
OfficeHome	A	61.2 $\pm$ 1.4	63.12
TerraIncognita	L100	59.9 $\pm$ 1.0	61.35
DomainNet	clipart	58.4 $\pm$ 0.3	58.48

Table 1: Comparison of the best OOD accuracy (%) of ERM between five optimizers. We use **Oracle** (see definition at 3.1 in [Gulrajani & Lopez-Paz \(2021\)](#)) as model selection method in this table. The model selection results in the training-domain validation set are shown in Table 13. Except for a small set of problems, non-adaptive optimizer outperforms adaptive optimizer in all but 10 out of 12 tasks. As a soundness check, we confirm that our Adam results outperform all existing benchmark results using Adam. Details are given in Appendix G.4.

Model	OOD Dataset	Non-Adaptive Optimizer			Adaptive Optimizer	
		SGD	Momentum	Netsterov	RMSProp	Adam
4-Layer CNN	ColoredMNIST	34.01%	34.23%	40.56%	89.30%	73.92%
	RotatedMNIST	90.00%	95.41%	94.06%	96.27%	96.40%
ResNet50	VLCS	99.43%	99.43%	99.29%	99.29%	99.29%
	PACS	88.67%	89.55%	89.25%	88.81%	89.30%
	OfficeHome	64.64%	65.01%	63.82%	62.91%	63.82%
	TerraIncognita	63.21%	62.41%	62.85%	62.31%	61.35%
	DomainNet	58.38%	61.91%	62.24%	55.74%	58.48%
	BackgroundChallenge	-	80.09%	-	-	77.90%
DistilBERT	Amazon-WILDS	52.00%	54.66%	54.66%	53.33%	52.00%
	CivilComment-WILDS	51.66%	57.69%	60.07%	45.39%	46.82%
ResNet-20	ColoredMNIST	-	33.50%	-	-	31.47%
ViT	PACS	-	91.80%	-	-	91.26%

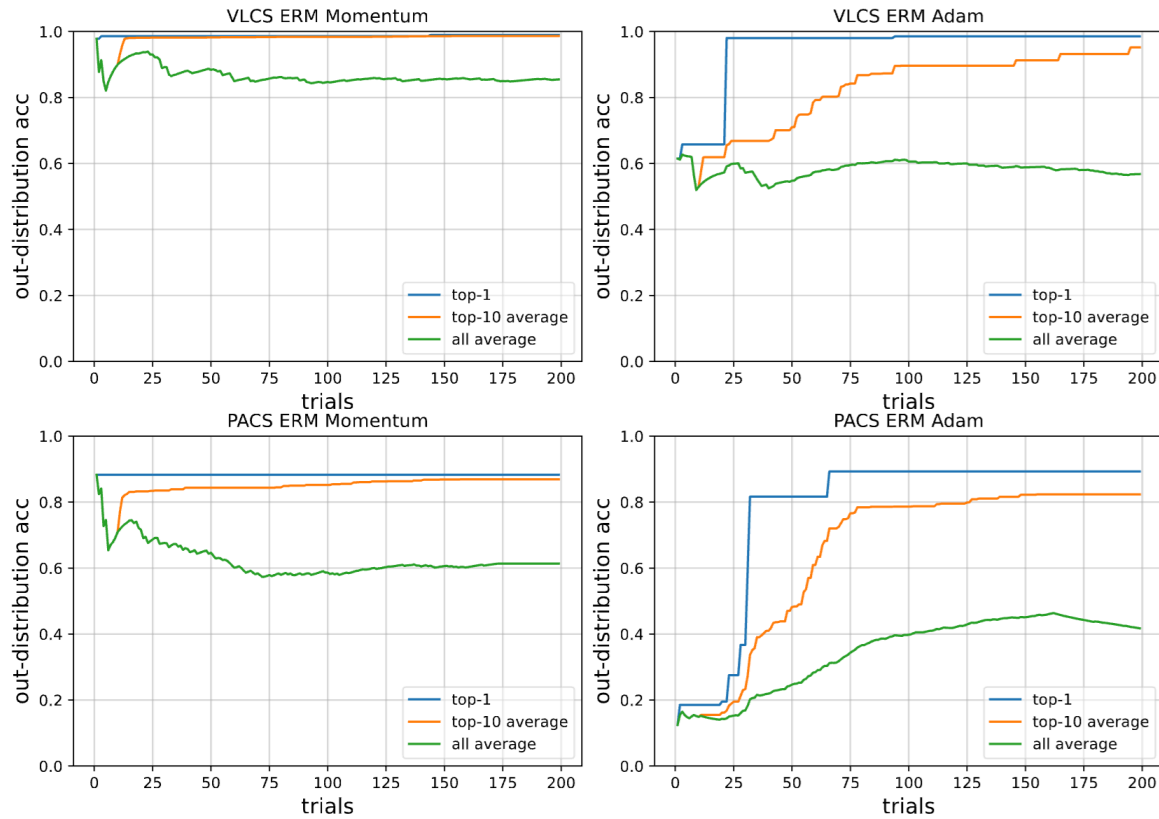


Figure 26. ERM ColoredMNIST, RotatedMNIST, VLCS and PACS / OOD accuracy when horizontal axis is the trial budget

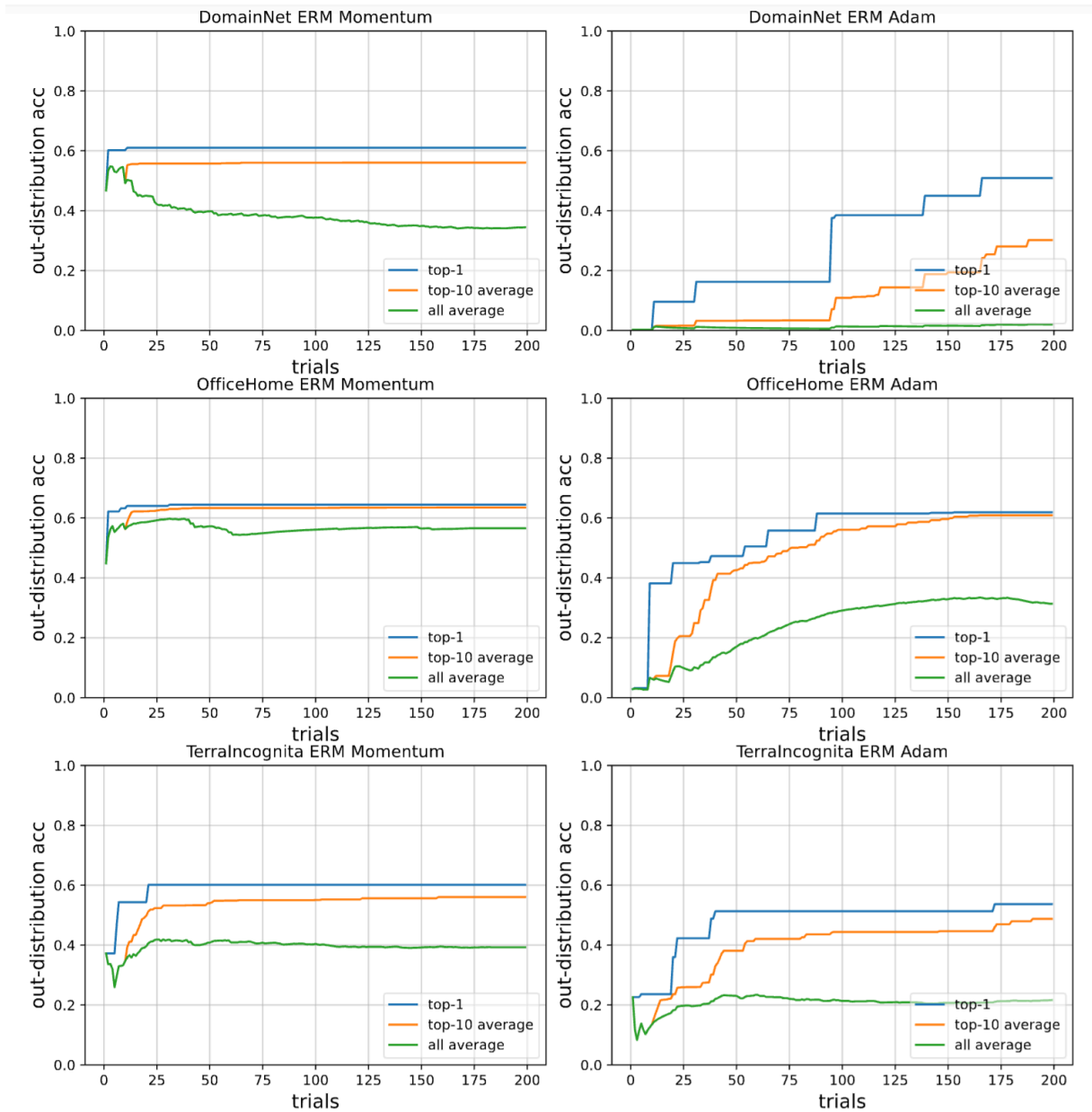


Figure 25. ERM DomainNet, OfficeHome and TerraIncognita / OOD accuracy when horizontal axis is the trial budget

Experimental Control

Results

OK (WASASAM is the best choice among SAM, MomentumSGD)

- PACS https://wandb.ai/rlopt/v0_pacs_opt_autotune/table?nw=nwuserhiroki11x
- VLCS https://wandb.ai/rlopt/v0_vlcs_opt_autotune/table?nw=nwuserhiroki11x
- TerraIncognita
https://wandb.ai/rlopt/v0_terraincognita_opt_autotune?nw=nwuserhiroki11x

Still need to explore

- OfficeHome
https://wandb.ai/rlopt/v0_officehome_opt_autotune/table?nw=nwuserhiroki11x
- DomainBed https://wandb.ai/rlopt/v0_domainnet_opt_autotune?nw=nwuserhiroki11x

Memo in Japanese

[OK!] PACS -> Run on Beluga

Rho=0.02 / SASAM is the best! (Averaged Weight) (89.221) (Existing Benchmark には勝ってる)
SAM の方が単体(89.039)では勝ってる / SASAM は (85.313)
https://wandb.ai/rlopt/v0_pacs_opt_autotune/table?nw=nwuserhiroki11x

=====

====

はいパラ

- SAM 0.05, 0.01, 0.005, 0.001 0.0005, 0.0001 / rho = 0.05
- Momentum 0.05, 0.01, 0.005, 0.001 0.0005, 0.0001
- SASAM LR=0.000001 rho =0.02

結果の見える方 () は +SWA / Max EigenH

- SAM 89.039 (88.306) / 286.73 (254.826)
- Momentum 83.847 (87.482) / 1227.9 (1145.832)
- SASAM 85.313 (**89.221**) / 204.89 (202.667)

全体的に良い！

=====

[OK!] VLCS -> Run on Beluga

Best? (99.558) (Existing Benchmark には勝ってる)

SAM の方が単体では勝ってる???

https://wandb.ai/rlopt/v0_vlcs_opt_autotune/table?nw=nwuserhiroki11x

-> 更新後

WASASAMが勝った (99.558) / 単体では SAM (99.47) が勝つ

=====

はいパラ

- SAM 0.05, 0.01, 0.005, 0.001 0.0005, 0.0001 / rho = 0.05
- Momentum 0.05, 0.01, 0.005, 0.001 0.0005, 0.0001
- SASAM LR=0.0001 rho =0.02

結果の見える方 () は +SWA / Max EigenH

- SAM 99.47 (99.47) / 112.345 (877.438)
- Momentum 98.719 (98.807) / 496.504 (2071.329)
- SASAM 96.511 (**99.558**) / 229.639 (598.141)

さちってるから、ちょっと挙動違う

平均化したモデルが一番曲率低いのは SASAM であることには変わらない！

=====

[NG] TerraIncognita -> Run on Narval

SASAM is the best (57.996) / lower than previous benchmark (59.9)

SAM (45.168) 単体 / Momentum (53.909) 単体 / SASAM 単体 (55.201)

https://wandb.ai/rlopt/v0_terraincognita_opt_autotune?nw=nwuserhiroki11x

問題点

- Previous benchmark より低いのがダメ (62% 出るはずなので、この結果はダメかも)

現在の進捗

- T4 で v02,v03 として実験を投げた
- もういいかなという気持ち

=====

=====

はいパラ

- SAM 0.05, 0.01, 0.005, 0.001 0.0005, 0.0001 / rho = 0.05
- Momentum 0.05, 0.01, 0.005, 0.001 0.0005, 0.0001
- SASAM LR=0.0001 rho =0.05

結果の見える方 () は +SWA / Max EigenH

- SAM 45.168 (40.87) / 418.644 (335.687)
- Momentum 53.909 (53.395) / 5771.888 (5090.749)
- SASAM 44.997 (**57.996**) / 261.032 (254.093)

全体的に良い！

=====

=====

[NG] OfficeHome -> Run on Narval

SASAM is worst (65.765) / higher than previous benchmark (61.2)

SAM (66.127) 平均化すると (65.276)

https://wandb.ai/rlopt/v0_officehome_opt_autotune/table?nw=nwuserhiroki11x

問題点

- SAM 単体より悪い
 - 悪くない気がする????
 - チェリーピックっぽい、、

現在の進捗

- T4 で v02,v03 として実験を投げた

解決策??

=====
=====
LR の探索範囲を絞る

- SAM 0.05, 0.01, 0.005, 0.001 / rho = 0.05
- Momentum 0.05, 0.01, 0.005, 0.001
- SASAM LR=0.00001 で fix / amp = rho = 0.005 で fix or rho = 0.05, 0.01, 0.005 で search でも OK

そうすると結果の見え方は () は +SWA / Max EigenH

- SAM 64.297 (64.528) / 372.578 (391.194)
- Momentum 61.643 (63.704) / 1319.399 (1226.569)
- SASAM 61.283 (65.276) / 709.146 (632.651)

Rho の違いで、SASAM が Curvature だと負ける

SASAM平均化したモデルで曲率が下がるというトレンドには変わらない！

=====
=====
[NG] DomainBed -> Run on Narval

SASAM is worst (56.141) / lower than previous benchmark (58.4)

SAM (58.245)

https://wandb.ai/rlopt/v0_domainnet_opt_autotune?nw=nwuserhiroki11x

問題点

- SAM より悪い

現在の進捗

- T4 で v02,v03 として実験を投げた
- **T4** でもう少し探索をする必要

(8/4 Update)

-> **R10 Cluster** で **v5** として探索

-> **Beluga** でも **v6** を実行

(8/11 Update)

-> チューニングのレンジの話ではなく、なぜこのデータセットがうまくいかないか考える必要

=====
=====

はいパラ

- SAM 0.05, 0.01, 0.005, 0.001 0.0005, 0.0001 / rho = 0.05
- Momentum 0.05, 0.01, 0.005, 0.001 0.0005, 0.0001
- SASAM LR=0.0005 rho = 0.005

結果の見える方 () は +SWA / Max EigenH

- SAM 58.245 (56.235) / 68.613 (55.69)
- Momentum 57.64 (58.58) / 59.043 (68.165)
- SASAM 52.673 (56.244) / 59.297 (65.889) まだ探索途中

=====