AdaBest: Minimizing Client Drift in Federated Learning via ADAptive Bias ESTimation

ECCV EUROPEAN CONFERENCE ON COMPUTER VISION TEL AVIV 2022 — October 23-27, 2022, Tel Aviv

Farshid Varno[1,2], Marzie Saghayi[1], Laya Rafiee Sevyeri[2,3], Sharut Gupta[2,4], Stan Matwin[1,5], Mohammad Havaei[2]

[1]Dalhousie University, [2]Imagia, [3]Concordia University, [4]IIT Delhi, [5]Polish Academy of Sciences

Background

- Federated Learning (FL): train models locally, aggregate globally, no data shared.
- Data heterogeneity among clients ⇨ *client drift*
- Reduced Variance SGD solutions

$\theta_i^{t,k} \leftarrow \theta_i^{t,k-1} - \eta \nabla L_i$ ➡ $\theta_i^{t,k} \leftarrow \theta_i^{t,k-1} - \eta(\nabla L_i + \boldsymbol{h} - \boldsymbol{h}_i)$

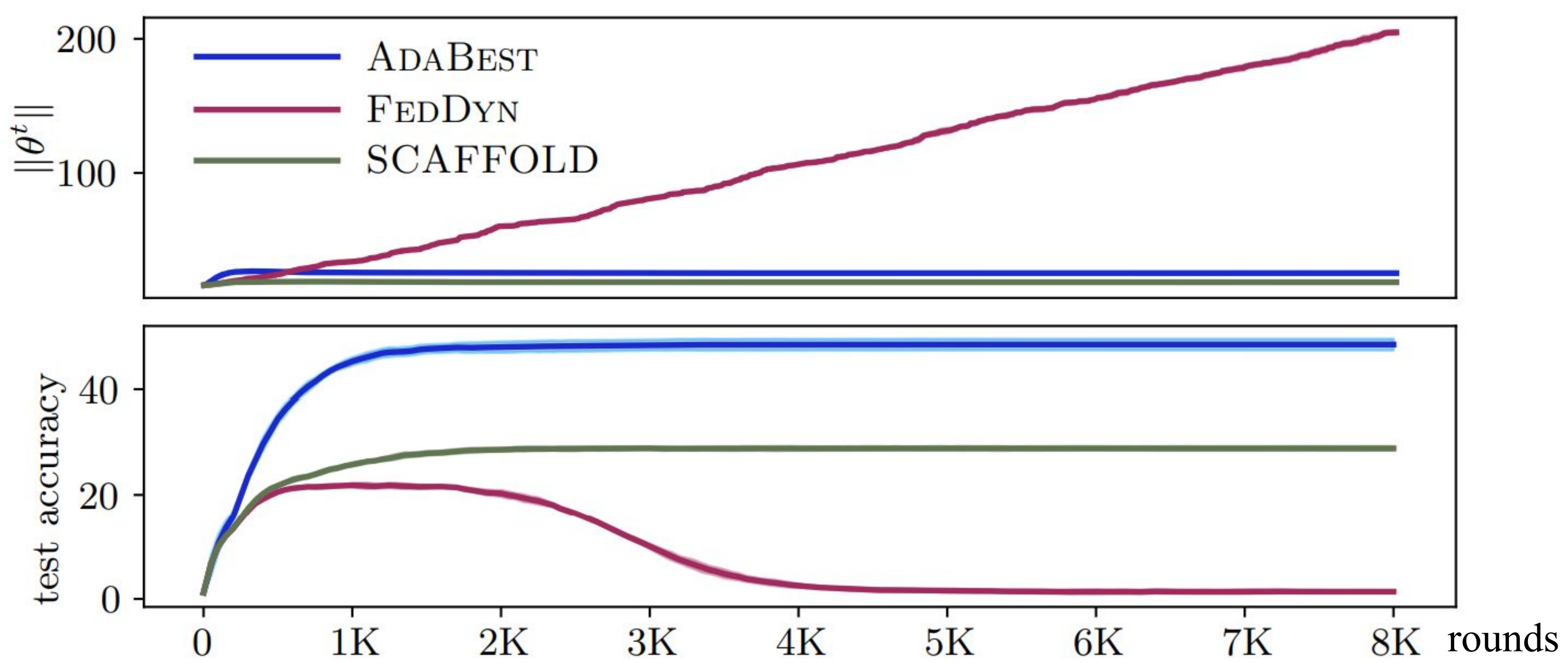
$\boldsymbol{h}_i$: estimate of grads of local samples
$\boldsymbol{h}$: estimate of grads of all samples

Prior Works

- SCAFFOLD: extra communication cost
- FedDyn:
  - ✓ apply $\boldsymbol{h}$ on the server: no extra communication!
  - ✖ not proper adaption of gradient estimates
  - ✖ parameter explosion!

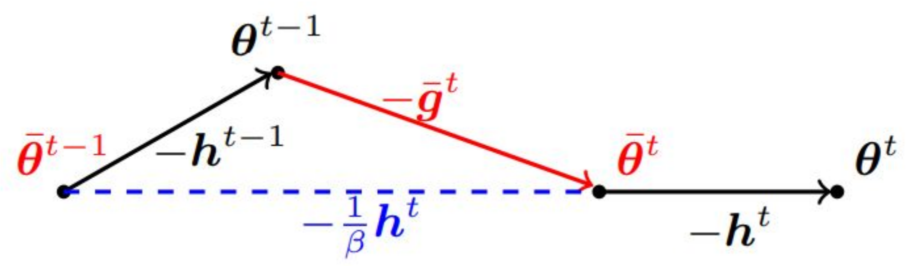
AdaBest

- ✓ Adaptive discounting gradient estimates: stability
- ✓ Improved scalability tolerance
- ✖ Extra discounting factor to tune (β)



Norm of cloud model parameters (up) and test accuracy (down) during training.

Algorithm



Geometric interpretation of AdaBest's bias correction applied to the server update.

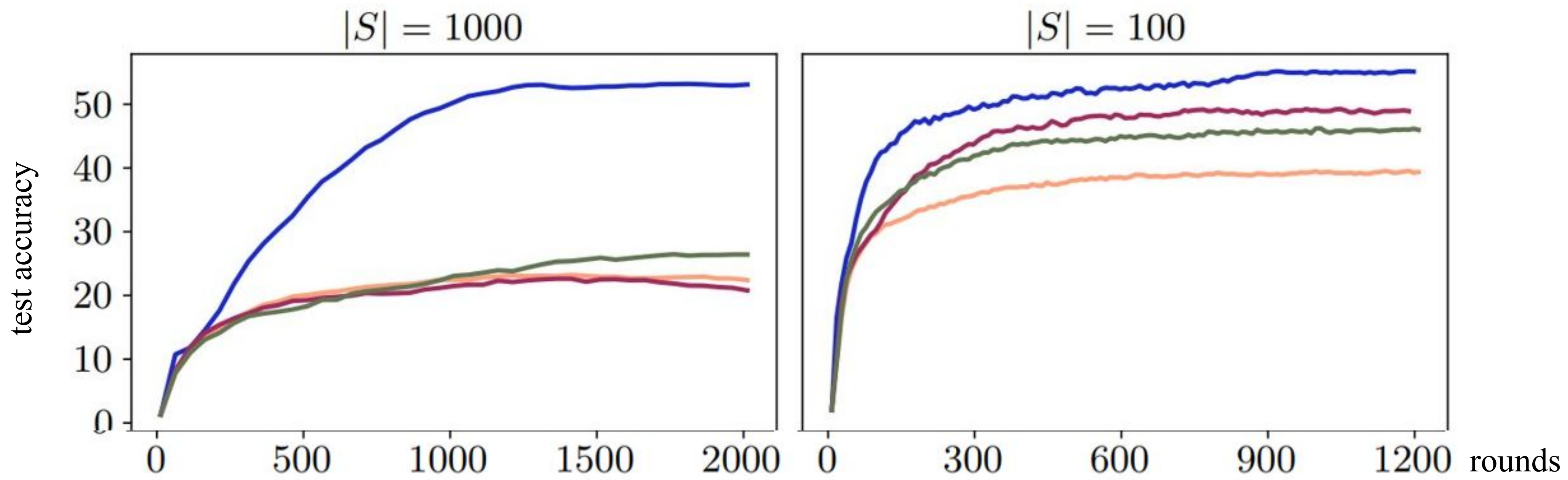
FedDyn | AdaBest

for $t = 1$ **to** T **do**
  Sample clients $\mathcal{P}^t \subseteq S^t$.
  Transmit θ^{t-1} to each client in $\mathcal{P}^t$
  for each client $i \in \mathcal{P}^t$ **in parallel do**
    Optimize θ^{t-1} locally on client i ➡
  /* aggregate received models */
  $\bar{\theta}^t \leftarrow \frac{1}{|\mathcal{P}^t|}\sum_{i\in\mathcal{P}^t}\theta_i^t$
  $\boldsymbol{h}^t \leftarrow \boldsymbol{h}^{t-1} + \frac{|\mathcal{P}^t|}{|S^t|}(\theta^{t-1} - \bar{\theta}^t)$ (FedDyn)
  $\boldsymbol{h}^t \leftarrow \beta(\bar{\theta}^{t-1} - \bar{\theta}^t)$ (AdaBest)
  /* update cloud model */
  $\theta^t \leftarrow \bar{\theta}^t - \boldsymbol{h}^t$

/* receive cloud model */
$\theta_i^{t,0} \leftarrow \theta^{t-1}$
/* locally optimize for K local steps */
for $k = 1$ **to** K **do**
  Compute mini-batch gradients $L_i(\theta_i^{t,k-1})$
  $\boldsymbol{g}_i^{t,k-1} \leftarrow \nabla L_i(\theta_i^{t,k-1}) - \boldsymbol{h}_i^{t_i'} - \mu(\theta^{t-1} - \theta_i^{t,k-1})$ (FedDyn)
  $\boldsymbol{g}_i^{t,k-1} \leftarrow \nabla L_i(\theta_i^{t,k-1}) - \boldsymbol{h}_i^{t_i'}$ (AdaBest)
  $\theta_i^{t,k} \leftarrow \theta_i^{t,k-1} - \eta \boldsymbol{g}_i^{t,k-1}$
end for
/* update local gradient estimates */
$\boldsymbol{g}_i^t \leftarrow \theta^{t-1} - \theta_i^{t,K}$
$\boldsymbol{h}_i^t \leftarrow \boldsymbol{h}_i^{t_i'} + \mu \boldsymbol{g}_i^t$ (FedDyn) | $\boldsymbol{h}_i^t \leftarrow \frac{1}{t-t_i'}\boldsymbol{h}_i^{t_i'} + \mu \boldsymbol{g}_i^t$ (AdaBest)
$t_i' \leftarrow t$
Transmit client model $\theta_i^t := \theta_i^{t,K}$.

Tuning β, Expensive?

- Experiment: scale # clients but keeping all setting (including learning rate) the same



Test accuracy for FL-CIFAR100 (alpha=0.03) with 1000 (left) and 100 (right) clients.

Test accuracy scores for FL-CIFAR100 (alpha=0.03) with 100 clients.

Top-1 Test Accuracy

Dataset	Setting	FedAvg	FedDyn	SCAFFOLD	AdaBest
EMNIST-L	α=0.03	93.58±0.25	93.57±0.20	94.29±0.11	**94.62±0.17**
	α=0.3	94.04±0.04	93.54±0.22	94.54±0.11	**94.64±0.11**
	IID	94.32±0.10	93.60±0.35	94.62±0.16	**94.70±0.24**
CIFAR10	α=0.03	74.04±0.88	76.85±0.91	77.19±1.10	**79.64±0.58**
	α=0.3	79.74±0.07	81.91±0.19	82.26±0.38	**84.15±0.36**
	IID	81.35±0.23	83.56±0.31	83.50±0.15	**85.78±0.14**
CIFAR100	α=0.03	39.18±0.56	44.24±0.66	45.80±0.36	**48.56±0.45**
	α=0.3	38.78±0.35	48.92±0.37	46.34±0.43	**54.51±0.35**
	IID	37.45±0.57	49.60±0.24	44.30±0.22	**55.58±0.14**

Results

- ✓ AdaBest shows superior performance & convergence speed in all benchmarks with partial client participation
- ✓ The greatest performance boost in CIFAR100 settings.
- ✓ Full client participation $\beta \to 1$, AdaBest $\approx$ FedDyn

Conclusion & Future Work

- Adapting gradient estimates: significant improvement in performance & convergence speed of RV-SGD based LocalSGD.
- AdaBest best works for large scale, partial participation FL setting.
- Future works: Analytical bounds for convergence rate of AdaBest & β auto-tune