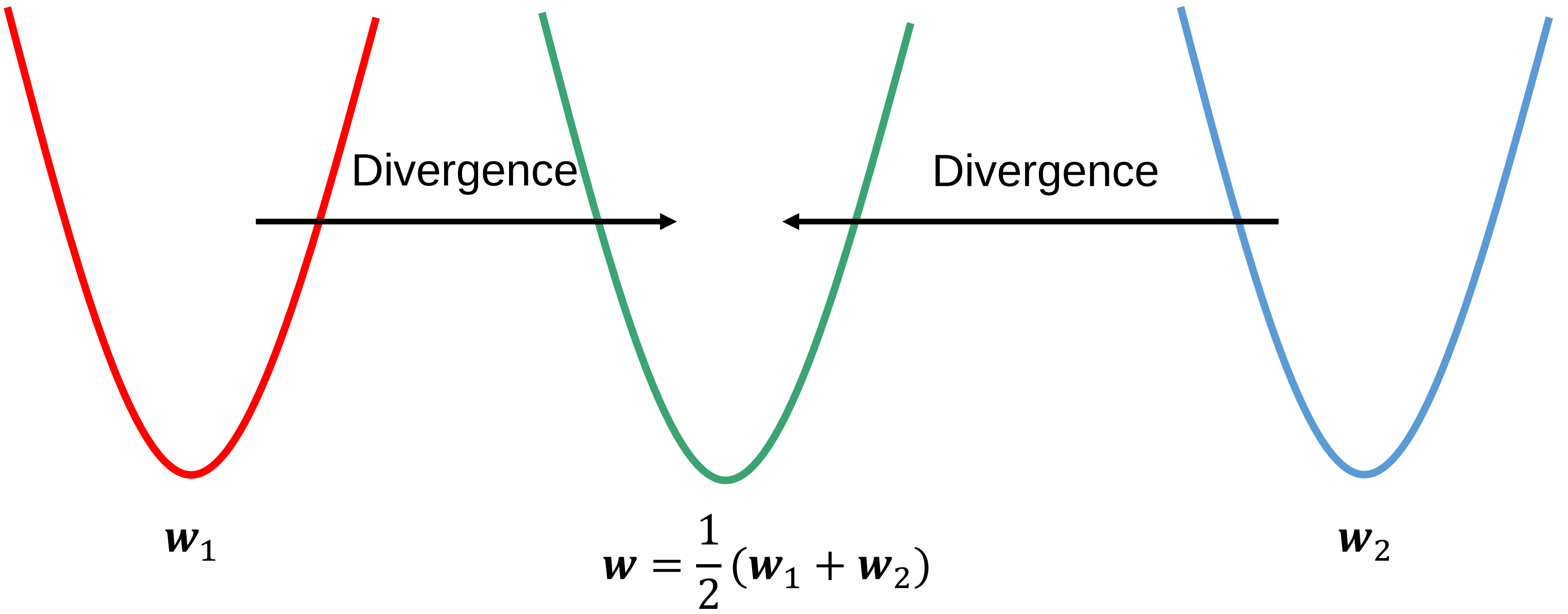
Generalized Federated Learning via Sharpness Aware Minimization

Zhe Qu [ł], Xingyu Li [‡], Rui Duan [ł], Yao Liu [ł], Bo Tang [‡], Zhuo Lu [ł]

[ł] University of South Florida

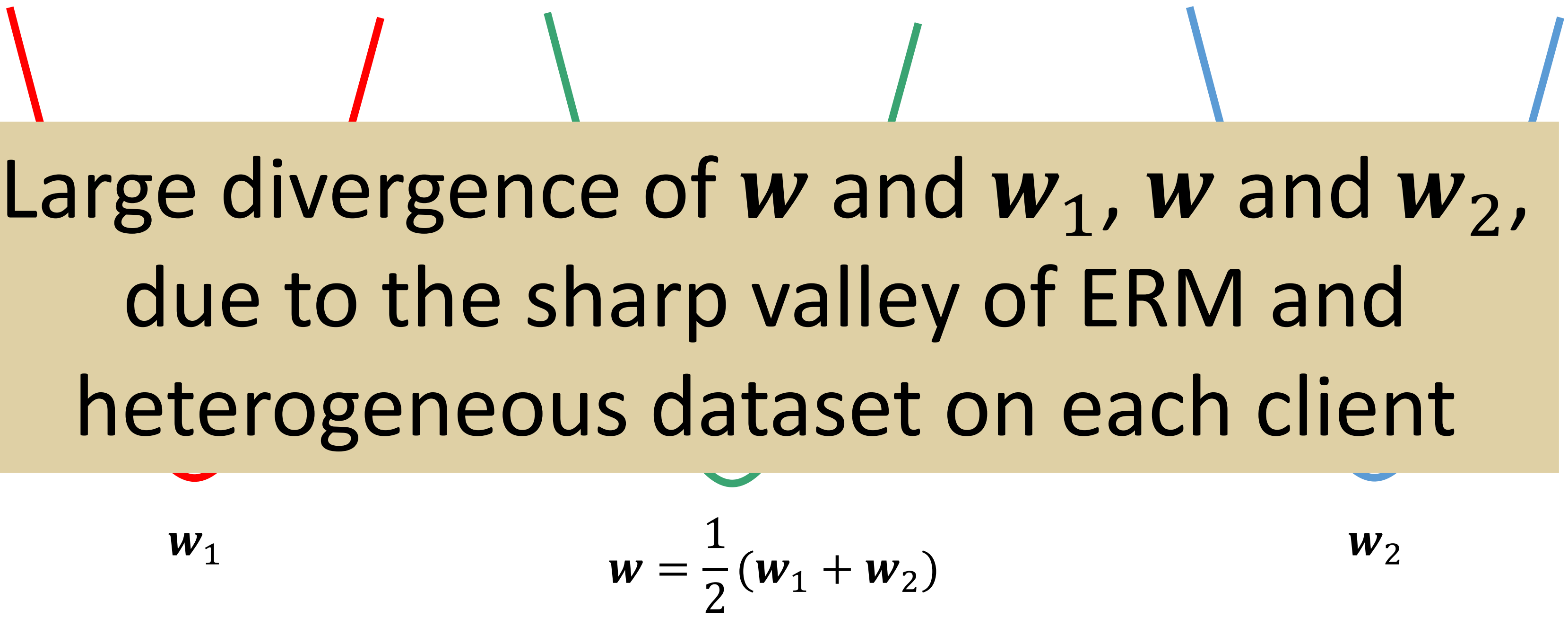
[‡] Mississippi State University

Motivation



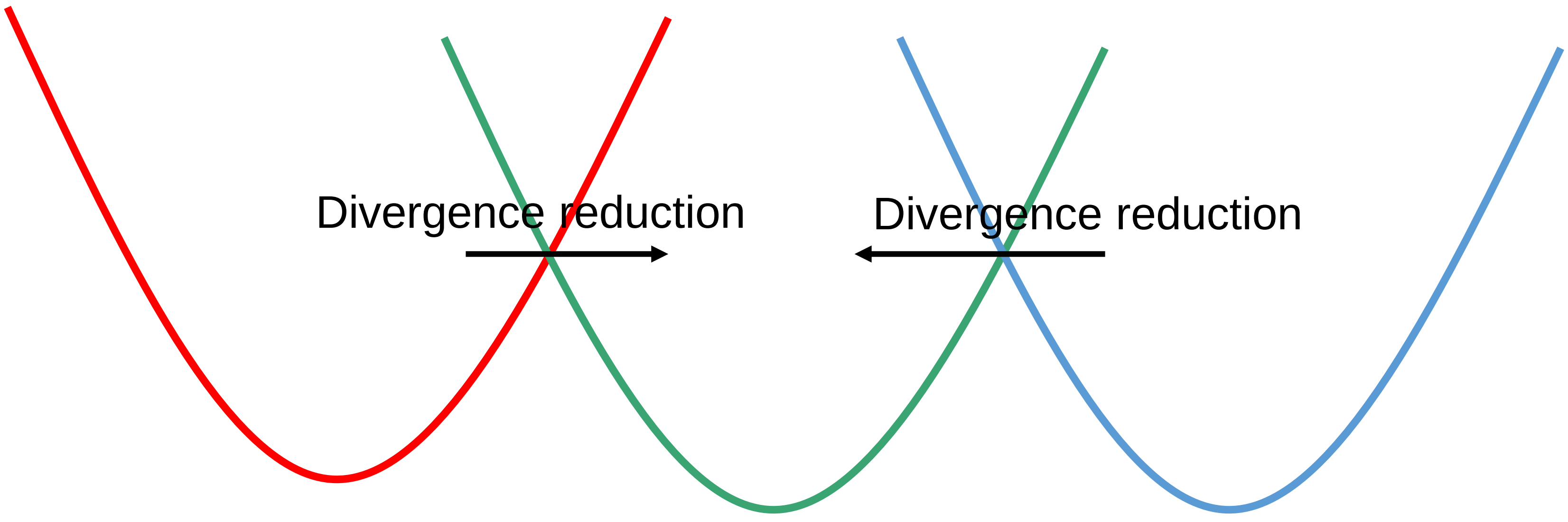
Divergence
Divergence
$\boldsymbol{w}_1$
$\boldsymbol{w} = \frac{1}{2}(\boldsymbol{w}_1 + \boldsymbol{w}_2)$
$\boldsymbol{w}_2$

Motivation



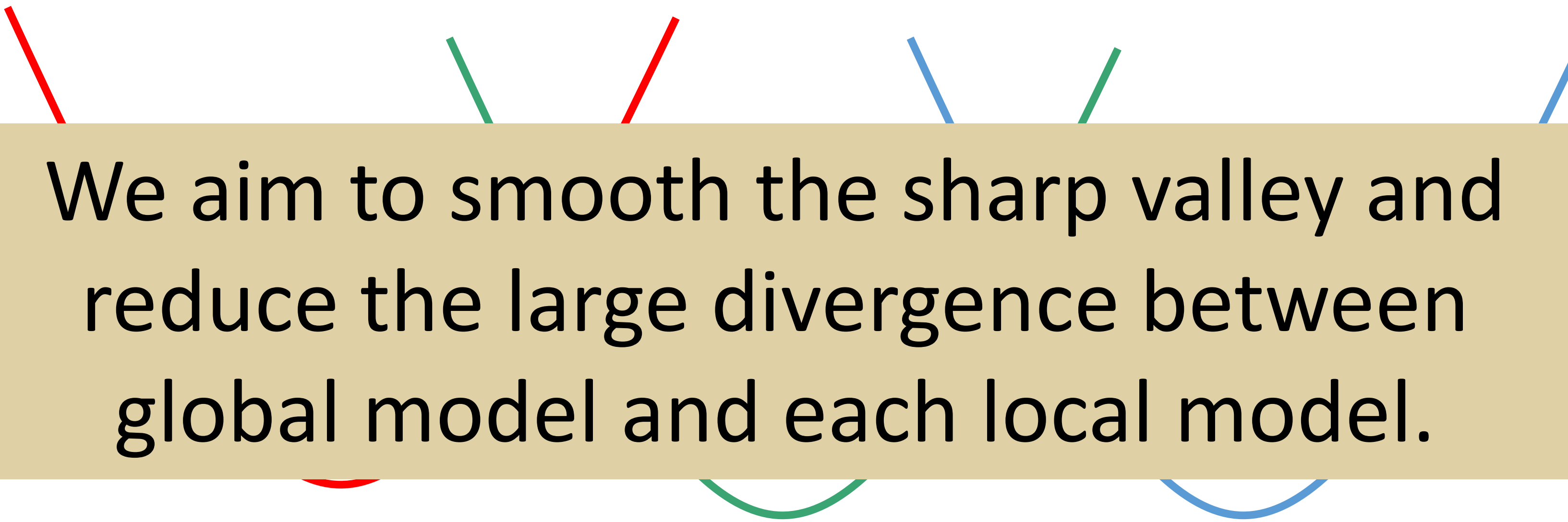
Large divergence of $\boldsymbol{w}$ and $\boldsymbol{w}_1$, $\boldsymbol{w}$ and $\boldsymbol{w}_2$, due to the sharp valley of ERM and heterogeneous dataset on each client

Motivation



$\boldsymbol{w}_1$ $\quad$ $\boldsymbol{w} = \frac{1}{2}(\boldsymbol{w}_1 + \boldsymbol{w}_2)$ $\quad$ $\boldsymbol{w}_2$

Motivation



We aim to smooth the sharp valley and reduce the large divergence between global model and each local model.

$\boldsymbol{w}_1$ $\quad$ $\boldsymbol{w} = \frac{1}{2}(\boldsymbol{w}_1 + \boldsymbol{w}_2)$ $\quad$ $\boldsymbol{w}_2$

Algorithm

Objective function:

$$\min_{\boldsymbol{w}} \left\{ F(\boldsymbol{w}) := \frac{1}{N} \sum_{i \in [N]} F_i(\boldsymbol{w}) \right\}$$

Change to

$$\min_{\boldsymbol{w}} \max_{||\delta_i|| \leq \rho} \left\{ F(\widetilde{\boldsymbol{w}}) := \frac{1}{N} \sum_{i \in [N]} F_i(\widetilde{\boldsymbol{w}}) \right\}$$

Algorithm: FedSAM

Sharpness Aware Minimization (SAM) to be local optimizer

$$\tilde{\boldsymbol{w}}_{i,k}^{r} = \boldsymbol{w}_{i,k}^{r} + \rho \frac{\boldsymbol{g}_{i,k}^{r}}{||\boldsymbol{g}_{i,k}^{r}||}$$

(find the optimal value $\tilde{\boldsymbol{w}}_{i,k}^{r} = \boldsymbol{w}_{i,k}^{r} + \delta_i^r$)

$$\boldsymbol{w}_{i,k+1}^{r} = \boldsymbol{w}_{i,k}^{r} - \eta_l \tilde{\boldsymbol{g}}_{i,k}^{r}$$

(local training)

Algorithm: FedSAM

Sharpness Aware Minimization (SAM) to be local optimizer

The loss function *f* is smoother, when *L* is smaller. For ERM based FL with the original loss surface, *L* is very high. SAM based FL can reduce the *L* significantly.

(local training)

Algorithm: FedSAM

Algorithm 1 FedAvg and FedSAM

Initialization: w_0, ρ_0 $\Delta^0 = 0$, learning rates η_l, η_g and the number of epochs K.

for $r = 0, \dots, R-1$ **do**

  Sample subset $\mathcal{S}^r \subseteq [N]$ of clients.

  $w_{i,0}^t = w^r$.

  for each client $i \in \mathcal{S}^r$ in parallel **do**

    for $k = 0, \dots, K-1$ **do**

      Compute a local training estimate $g_{i,k}^r = \nabla F_i(w_{i,k}^r, \xi_{i,k}^r)$ of $\nabla F_i(w_{i,k}^r)$.

      $w_{i,k}^r = w_{i,k}^r - \eta_l g_{i,k}^r$.

      Compute local model $w_{i,k}^r$ from (4).

    end for

    $\Delta_i^r = w_{i,K}^r - w^r$.

  end for

  $\Delta^{r+1} = \frac{1}{S} \sum_{i \in \mathcal{S}^r} \Delta_i^r$.

  $w^{r+1} = w^r + \eta_g \Delta^r$.

end for

Theoretical Results: FedSAM

Full participation

$$O\left(\frac{LF}{\sqrt{RKN}}+\frac{\sigma_g^2}{R}+\frac{L^2\sigma_l^2}{R^{3/2}\sqrt{KN}}+\frac{L^2}{R^2}\right)$$

Partial participation

$$O\left(\frac{LF}{\sqrt{RKS}}+\frac{\sqrt{K}\sigma_g^2}{\sqrt{RS}}+\frac{L^2\sigma_l^2}{R^{3/2}K}+\frac{L^2}{R^2}\right)$$

(The convergence results match the best rates in existing studies)

Theoretical Results: FedSAM

Generalization bound

$$\mathcal{L}^{SAM}\big(F(\boldsymbol{w})\big) \leq \tilde{\mathcal{L}}_{\gamma}^{SAM}\big(F(\boldsymbol{w}+\delta)\big) + O\left(\frac{32Ad^2h\log(dh)Q\big(F(\boldsymbol{w})\big)+d\log\frac{Nmd\log(M)}{\xi}}{\gamma^2 m}\right)$$

This result indicates the dependence of the perturbation δ and the different neural network parameters in which we can enforce the loss surface around a point in order to guarantee the smoothness.

Algorithm: MoFedSAM

- The local optimizer SAM cannot directly affect the global model Δ^r .
- Reusing the information Δ^r can guide the local training on the participated clients in next communication round.

$$\widetilde{\boldsymbol{w}}_{i,k}^{r} = \boldsymbol{w}_{i,k}^{r} + \rho \frac{\boldsymbol{g}_{i,k}^{r}}{||\boldsymbol{g}_{i,k}^{r}||}$$

$$\boldsymbol{v}_{i,k}^{r} = \beta \boldsymbol{g}_{i,k}^{r} + (1-\beta)\Delta^{r}$$

$$\boldsymbol{w}_{i,k+1}^{r} = \boldsymbol{w}_{i,k}^{r} + \eta \boldsymbol{v}_{i,k}^{r}$$

Algorithm: MoFedSAM

Algorithm 2 `MoFedSAM` algorithm.

Initialization: w^0, $\Delta^0 = 0$, ρ^0, momentum parameter β the number of local updates K.
for $r = 0, \ldots, R-1$ **do**
  Sample subset $\mathcal{S}^r \subseteq [N]$ of clients.
  $w^t_{i,0} = w^r$.
  for each client $i \in \mathcal{S}^r$ in parallel **do**
    for $k = 0, \ldots, K-1$ **do**
      Compute a local training estimate $g^r_{i,k} = \nabla F_i(w^r_{i,k}, \xi^r_{i,k})$ of $\nabla F_i(w^r_{i,k})$.
      Compute local model $w^r_{i,k}$ from (6).
    end for
    $\Delta^r_i = w^r_{i,K} - w^r$.
  end for
  $\Delta^{r+1} = -\frac{1}{\eta_l KS} \sum_{i \in \mathcal{S}^r} \Delta^r_i$.
  $w^{r+1} = w^r - \eta_g \Delta^{r+1}$.
end for

Theoretical Results: MoFedSAM

Full participation

$$O\left(\frac{\beta LF}{\sqrt{RKN}}+\frac{\beta\sigma_g^2}{RL^2}+\frac{L^2\sigma_l^2}{R^2\beta}+\frac{\beta L^2}{R^2}\right)$$

Partial participation

$$O\left(\frac{\beta LF}{\sqrt{RKS}}+\frac{\beta\sqrt{K}\sigma_g^2}{\sqrt{RS}}+\frac{L^2\sigma_l^2}{R^{3/2}K}+\frac{\sqrt{K}L^2}{R^{3/2}\sqrt{S}}\right)$$

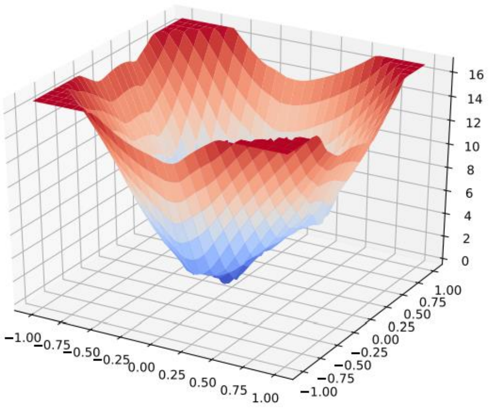
(The convergence results achieve a linear speedup compared to the existing studies.)

Experimental Results

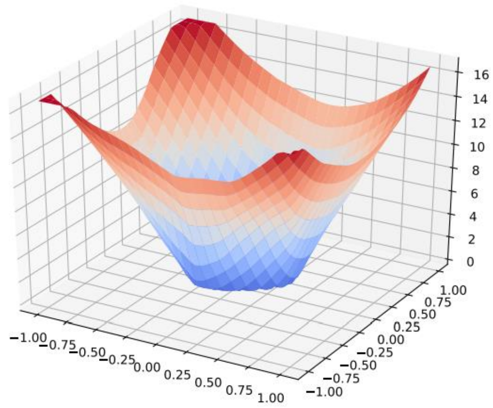
Table 1. Average (standard deviation) training accuracy and testing accuracy. Communication round to achieve the targeted testing accuracy: EMNIST 80%, CIFAR-10 80% and CIFAR-100 50%.

Algorithm	EMNIST			CIFAR-10			CIFAR-100		
	Train	Validation	Round	Train	Validation	Round	Train	Validation	Round
`FedAvg`	95.07 (0.94)	84.38 (4.03)	43	93.15 (1.44)	81.87 (5.09)	307	79.57 (1.84)	53.57 (5.40)	302
`SCAFFOLD`	93.85 (1.31)	84.09 (4.56)	69	91.76 (1.89)	80.61 (5.64)	546	78.49 (2.02)	51.49 (5.87)	551
`FedRobust`	93.17 (0.62)	83.70 (3.37)	91	90.82 (1.27)	79.63 (4.21)	847	76.80 (1.70)	49.06 (4.75)	893
`FedCM`	96.16 (1.14)	84.85 (4.11)	28	95.61 (1.50)	83.30 (4.77)	136	82.13 (1.96)	55.50 (5.04)	182
`MimeLite`	96.22 (1.16)	84.88 (4.22)	25	95.73 (1.56)	83.18 (4.65)	152	82.46 (2.00)	55.73 (5.11)	189
`FedSAM`	95.73 (0.49)	84.75 (3.04)	38	94.20 (1.08)	83.06 (3.87)	269	81.04 (1.59)	54.69 (4.36)	245
`MoFedSAM`	96.42 (0.42)	85.07 (2.95)	24	95.67 (1.16)	83.92 (3.65)	124	82.62 (1.53)	56.60 (4.42)	124

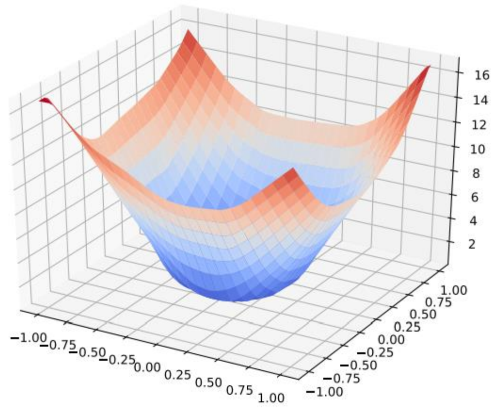
Experimental Results



(a) FedAvg.



(b) FedSAM.



(c) MoFedSAM.

Thank you !