# FEDBE: Making Bayesian Model Ensemble Applicable to Federated Learning

Hong-You Chen, Wei-Lun Chao
The Ohio State University

## *Highlights*

- A novel federated **aggregation** framework based on **Bayesian ensemble**
- Construct global posterior with client models and perform Bayesian model ensemble by sampling
- Leverage **unlabeled data** on the server for **knowledge distillation** for aggregation, bypassing weight average
- Apply stochastic weight average (SWA) [Izmailov et al., 2018] for robust distillation

## *Introduction*

- *Background: FEDAVG* [McMahan et al., 2017]:
  **Weight average:**

$$\bar{\boldsymbol{w}} \leftarrow \sum_i \frac{|\mathcal{D}_i|}{|\mathcal{D}|} \boldsymbol{w}_i.$$

➢ Non-i.i.d local data → *clients drift away from the ideal global model*
➢ Element-wise average → *negative effects on over-parameterized and permutation-invariant models like neural networks*

- *Our observation: model ensemble is significantly better than FEDAVG*
  **Bayesian model ensemble:**

$$p(y|\boldsymbol{x};\mathcal{D}) \approx \frac{1}{M}\sum_{m=1}^{M} p(y|\boldsymbol{x};\boldsymbol{w}^{(m)}), \text{ where } \boldsymbol{w}^{(m)} \sim p(\boldsymbol{w}|\mathcal{D}).$$

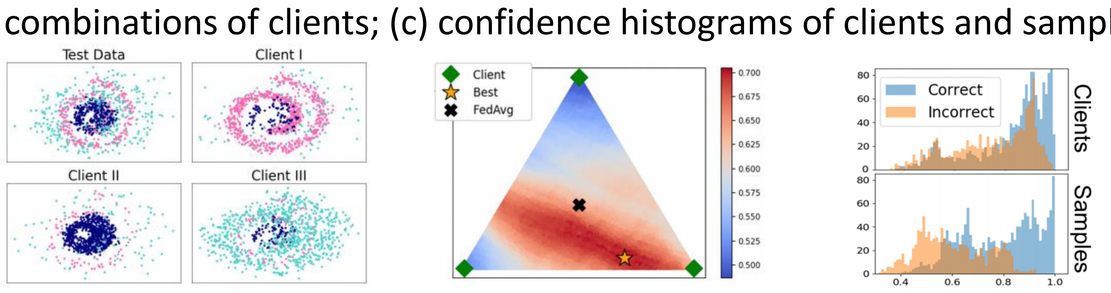
➢ How to construct global model distribution? → *stochastic weight average-Gaussian (SWAG) [Maddox et al., 2019]*
➢ How to translate the ensemble to a global model for multi-round federated learning? → *knowledge distillation on server unlabeled data*
➢ Distillation on noisy pseudo labels? → *stochastic weight average (SWA)*

**Incorporate Bayesian ensemble into federated learning by distilling on more robust aggregation of the outputs of a wide spectrum of global models**

## *Federated Bayesian Ensemble*

- *A swiss roll three-client federated learning toy example*

(a) non-i.i.d clients, balanced test set; (b) accuracy heatmap of convex combinations of clients; (c) confidence histograms of clients and samples.




(a) (b) (c)

- *Construct global model distribution* $p(\boldsymbol{w}|\mathcal{D})$

**Gaussian** $\boldsymbol{w}^{(m)} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma_{\text{diag}})$**:**

$$\boldsymbol{\mu} = \sum_i \frac{|\mathcal{D}_i|}{|\mathcal{D}|}\boldsymbol{w}_i, \qquad \Sigma_{\text{diag}} = \text{diag}\left(\sum_i \frac{|\mathcal{D}_i|}{|\mathcal{D}|}(\boldsymbol{w}_i - \boldsymbol{\mu})^2\right)$$

**Dirichlet:**

$$\boldsymbol{w}^{(m)} = \sum_i \frac{\gamma_i^{(m)}|\mathcal{D}_i|}{\sum_{i'}\gamma_{i'}^{(m)}|\mathcal{D}_{i'}|}\boldsymbol{w}_i, \qquad \boldsymbol{\gamma}^{(m)} \sim p(\boldsymbol{\gamma}) = p(\gamma_1, \cdots, \gamma_{|\mathcal{S}|}) = \frac{1}{\text{B}(\boldsymbol{\alpha})}\prod_i \gamma_i^{\alpha_i - 1}$$

- *Construct pseudo labels on server unlabeled dataset*

$$\mathcal{T} = \{(\boldsymbol{x}_j, \hat{\boldsymbol{p}}_j)\}_{j=1}^{J} \qquad \hat{\boldsymbol{p}}_j = \frac{1}{M}\sum_{m=1}^{M} p(y|\boldsymbol{x}_j;\boldsymbol{w}^{(m)})$$

- *Robust knowledge distillation with SWA*
  Momentum SGD with cyclical learning rate schedule, averaging checkpoints of ends of cycles.

## *Algorithm*

**Algorithm 1:** FEDBE (**Fed**erated **B**ayesian **E**nsemble)

**Server input** : initial global model $\boldsymbol{w}$, SWA scheduler $\eta_{\text{SWA}}$, unlabeled data $\mathcal{U} = \{\boldsymbol{x}_j\}_{j=1}^{J}$;
**Client $i$'s input** : local step size $\eta_l$, local labeled data $\mathcal{D}_i$;
**for** $r \leftarrow 1$ **to** $R$ **do**
  **Sample** clients $\mathcal{S} \subseteq \{1, \cdots, N\}$;
  **Communicate** $\boldsymbol{w}$ to all clients $i \in \mathcal{S}$;
  **for** *each client* $i \in \mathcal{S}$ *in parallel* **do**
    **Initialize** local model $\boldsymbol{w}_i \leftarrow \boldsymbol{w}$;
    $\boldsymbol{w}_i \leftarrow$ **Client training**$(\boldsymbol{w}_i, \mathcal{D}_i, \eta_l)$;
    **Communicate** $\boldsymbol{w}_i$ to the server;
  **end**
  **Construct** $\bar{\boldsymbol{w}} = \sum_{i\in\mathcal{S}} \frac{|\mathcal{D}_i|}{\sum_{i'\in\mathcal{S}}|\mathcal{D}_i'|}\boldsymbol{w}_i$;
  **Construct** global model distribution $p(\boldsymbol{w}|\mathcal{D})$ from $\{\boldsymbol{w}_i; i \in \mathcal{S}\}$;
  **Sample** $M$ global models $\{\boldsymbol{w}^{(m)} \sim p(\boldsymbol{w}|\mathcal{D})\}_{m=1}^{M}$;
  **Construct** $\{\boldsymbol{w}^{(m')}\}_{m'=1}^{M'} = \{\bar{\boldsymbol{w}}\} \cup \{\boldsymbol{w}_i; i \in \mathcal{S}\} \cup \{\boldsymbol{w}^{(m)}\}_{m=1}^{M}$;
  **Construct** $\mathcal{T} = \{\boldsymbol{x}_j, \hat{\boldsymbol{p}}_j\}_{j=1}^{J}$, where $\hat{\boldsymbol{p}}_j = \frac{1}{M'}\sum_{m'} p(y|\boldsymbol{x}_j;\boldsymbol{w}^{(m')})$;
  **Knowledge distillation:** $\boldsymbol{w} \leftarrow \text{SWA}(\bar{\boldsymbol{w}}, \mathcal{T}, \eta_{\text{SWA}})$;
**end**
**Server output** : $\boldsymbol{w}$.

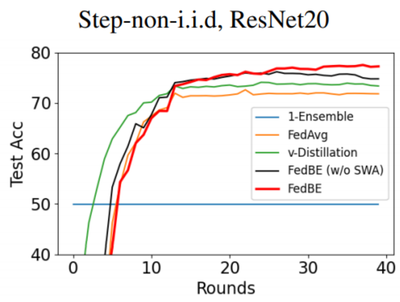
## *Experiments and Analysis*

- *Non-i.i.d CIFAR-10* (10 clients, 40 rounds, 20 local epochs, 10K unlabeled CIFAR-10 as the server data, Gaussian global distribution)

| Non-i.i.d. Type | Method | ConvNet | ResNet20 | ResNet32 | ResNet44 | ResNet56 |
|---|---|---|---|---|---|---|
| Step | 1-Ensemble | 60.5±0.28 | 49.9±0.46 | 35.5±0.50 | 32.8±0.38 | 23.3±0.52 |
| | FEDAVG | 72.0±0.25 | 70.2±0.17 | 66.5±0.36 | 60.5±0.26 | 51.4±0.15 |
| | v-Distillation | 69.2±0.18 | 72.6±0.62 | 68.4±0.33 | 60.4±0.53 | 56.4±1.10 |
| | FEDBE (w/o SWA) | 72.1±1.21 | 74.9±1.41 | 71.1±0.75 | 61.0±0.75 | 56.6±0.85 |
| | **FEDBE** | **74.5**±0.51 | **77.5**±0.42 | **72.7**±0.27 | **65.5**±0.32 | **60.7**±0.45 |
| Dirichlet | 1-Ensemble | 63.3±0.56 | 45.2±1.06 | 39.5±0.78 | 31.5±0.77 | 27.2±0.65 |
| | FEDAVG | 72.3±0.12 | 74.4±0.36 | 73.4±0.23 | 67.1±0.54 | 62.2±0.45 |
| | v-Distillation | 67.7±0.98 | 73.1±0.78 | 70.8±0.64 | 66.9±0.85 | 62.8±0.66 |
| | FEDBE (w/o SWA) | 70.1±0.42 | 75.9±0.56 | 73.9±0.55 | 68.2±0.72 | 63.2±0.71 |
| | **FEDBE** | **73.9**±0.45 | **78.2**±0.36 | **77.7**±0.45 | **71.5**±0.38 | **67.0**±0.30 |
| Centralized⋆ | SGD | 84.5 | 91.7 | 92.6 | 93.1 | 93.4 |

- *Compatibility of FEDBE with FEDAVGM and FEDPROX*

| Non-i.i.d. Type | Method | ConvNet | ResNet20 | ResNet32 | ResNet44 | ResNet56 |
|---|---|---|---|---|---|---|
| Step | FEDPROX | 72.5±0.71 | 71.1±0.52 | 67.7±0.26 | 60.4±0.71 | 54.9±0.66 |
| | **FEDBE** +FEDPROX | **74.9**±0.38 | **77.7**±0.45 | **72.9**±0.44 | **64.5**±0.37 | **60.1**±0.62 |
| | FEDAVGM | 72.3±0.55 | 73.2±0.57 | 70.0±0.62 | 59.9±0.65 | 52.7±0.49 |
| | **FEDBE** +FEDAVGM | **74.5**±0.47 | **78.0**±0.46 | **73.6**±0.50 | **65.5**±0.40 | **59.7**±0.51 |
| Dirichlet | FEDPROX | 72.6±0.38 | 76.1±0.49 | 73.4±0.51 | 68.1±0.79 | 60.9±0.46 |
| | **FEDBE** +FEDPROX | **74.6**±0.35 | **78.7**±0.49 | **77.3**±0.60 | **71.7**±0.43 | **66.5**±0.41 |
| | FEDAVGM | 73.0±0.43 | 76.5±0.44 | 75.5±0.79 | 67.7±0.46 | 58.9±0.72 |
| | **FEDBE** +FEDAVGM | **74.4**±0.49 | **78.5**±0.66 | **78.5**±0.26 | **72.0**±0.51 | **67.0**±0.55 |

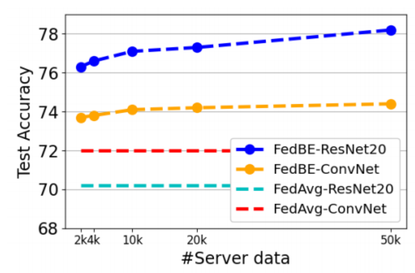



- FEDBE with SWA consistently outperforms baselines and improves on other methods
- We observe deeper models suffer from performance drop in federated learning
- FEDBE performs well with small and out-of-domain unlabeled server data

- *Effects of unlabeled data for distillation*

| Non-i.i.d. Type | $\mathcal{U}$ | $\lvert\mathcal{U}\rvert$ | ConvNet | ResNet20 |
|---|---|---|---|---|
| Step | CIFAR-10 | 10K | 74.5±0.51 | 77.5±0.42 |
| | CIFAR-100 | 50K | 74.4±0.45 | 78.2±0.58 |
| | Tiny-ImageNet | 100K | 74.5±0.64 | 77.1±0.51 |
| Dirichlet | CIFAR-10 | 10K | 73.9±0.45 | 78.2±0.36 |
| | CIFAR-100 | 50K | 73.5±0.41 | 78.6±0.63 |
| | Tiny-ImageNet | 100K | 74.0±0.35 | 78.2±0.72 |

Different sizes of server data CIFAR-100




- *FL with systems heterogeneity*
  Local training epochs E ~(0, 20]

| Method | ConvNet | ResNet20 | ResNet32 |
|---|---|---|---|
| FEDAVG | 70.6±0.46 | 69.9±0.59 | 64.0±0.50 |
| FEDPROX | 71.2±0.55 | 69.4±0.48 | 65.9±0.63 |
| **FEDBE** | 73.3±0.56 | 77.1±0.61 | 70.2±0.39 |
| **+FEDPROX** | **73.7**±0.24 | **77.5**±0.51 | **71.6**±0.37 |

- *Conclusion*: FEDBE is robust to non-i.i.d. clients and network architectures, compatible to regularized client training and server momentum methods