Sparsity in Neural Networks

Berivan Isik
Electrical Engineering, Stanford University

December 30th 2022

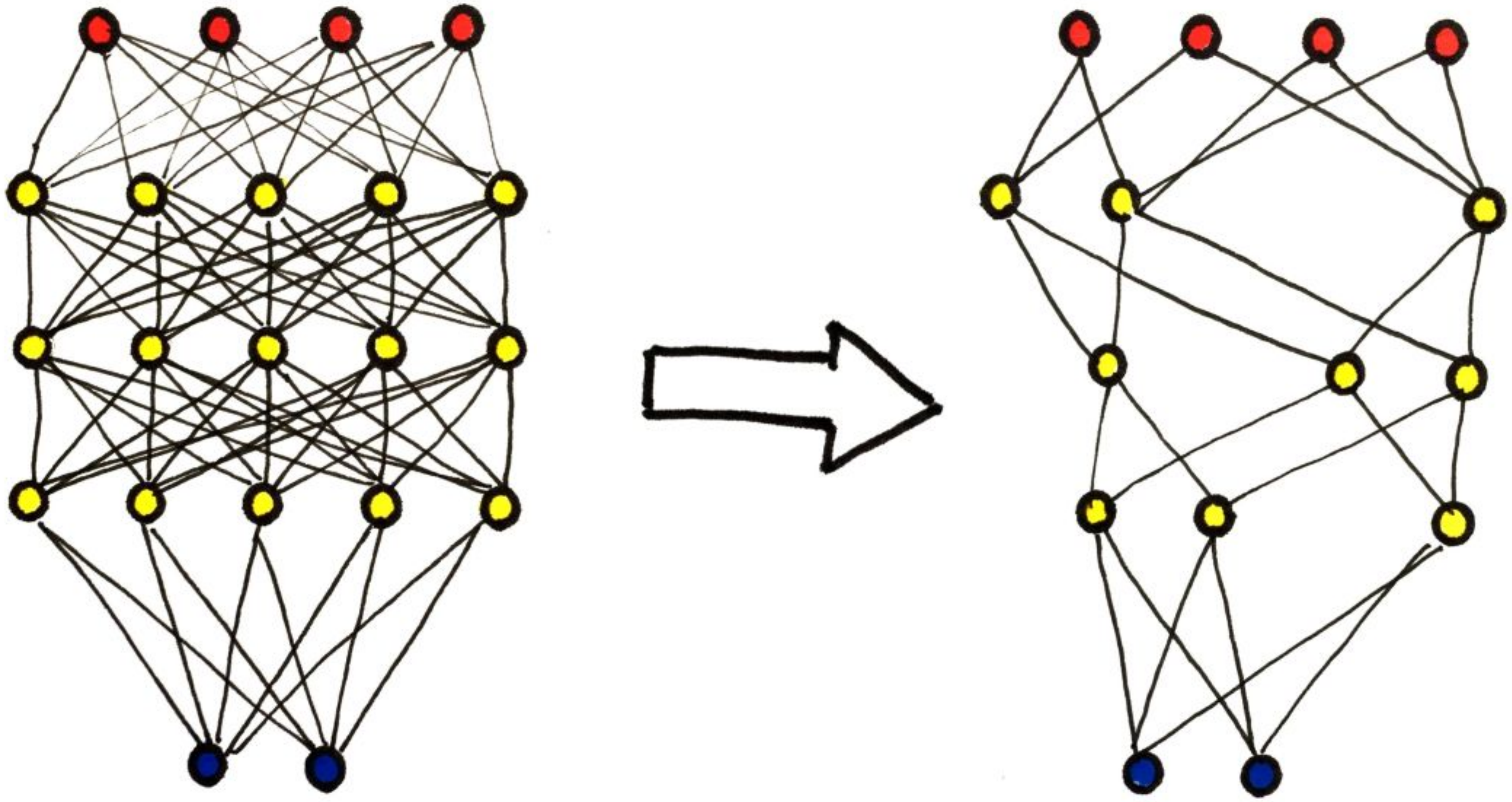
Outline

1. **What is sparsification for model compression and federated learning?**

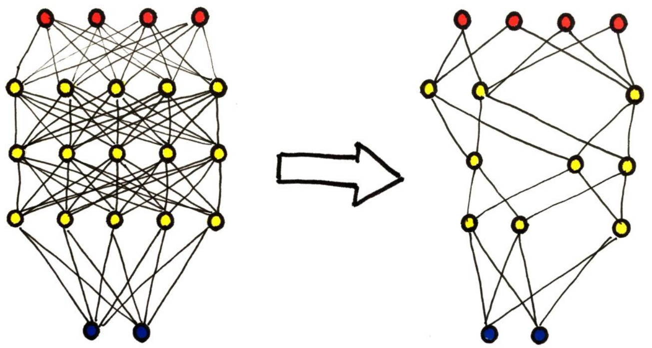
2. **Model compression**
    - An information-theoretic justification for model pruning

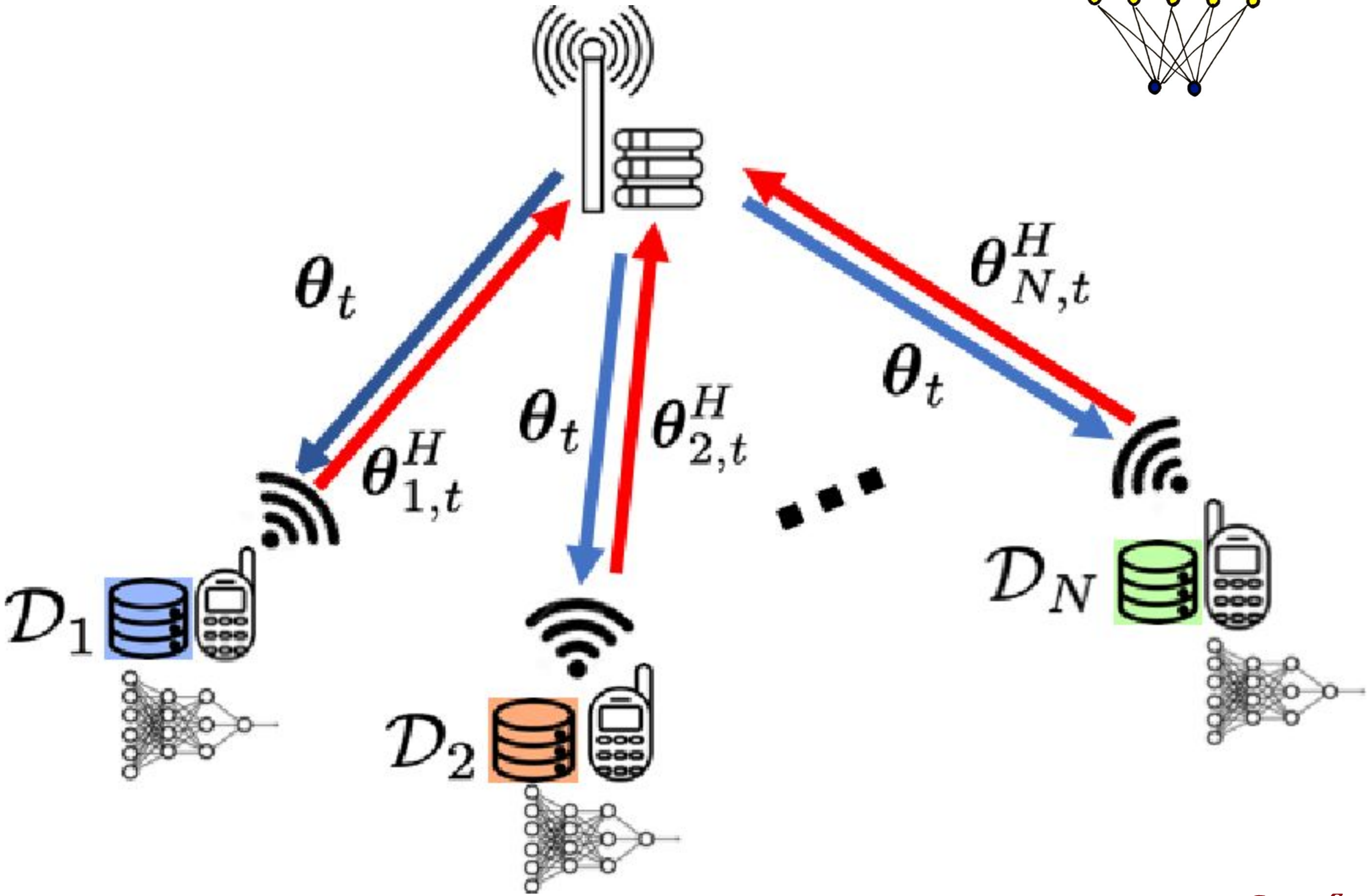
3. **Compression for federated learning**
    - Sparse random networks for communication-efficient federated learning

Sparsification



Federated Learning





An Information-Theoretic Justification for Model Pruning

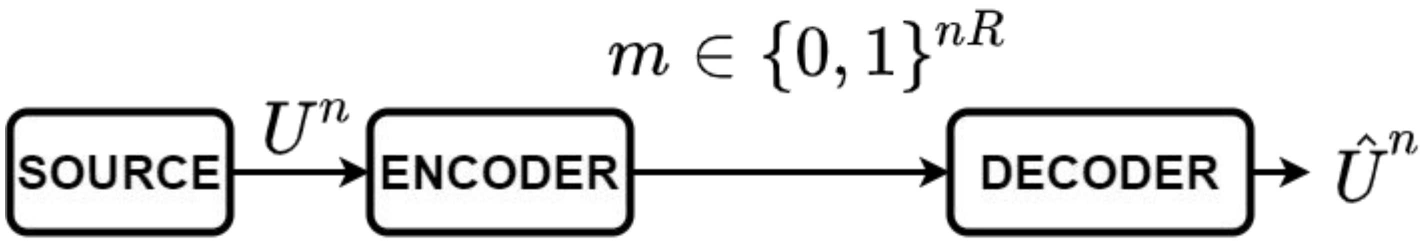
AISTATS'22

Joint work with

Tsachy Weissman (Stanford University)

Albert No (Hongik University)

Lossy Source Coding



$$R(D) = \min_{p(\hat{u}|u): E[d(u,\hat{u})] \le D} I\left(U; \hat{U}\right)$$

Distortion Metric

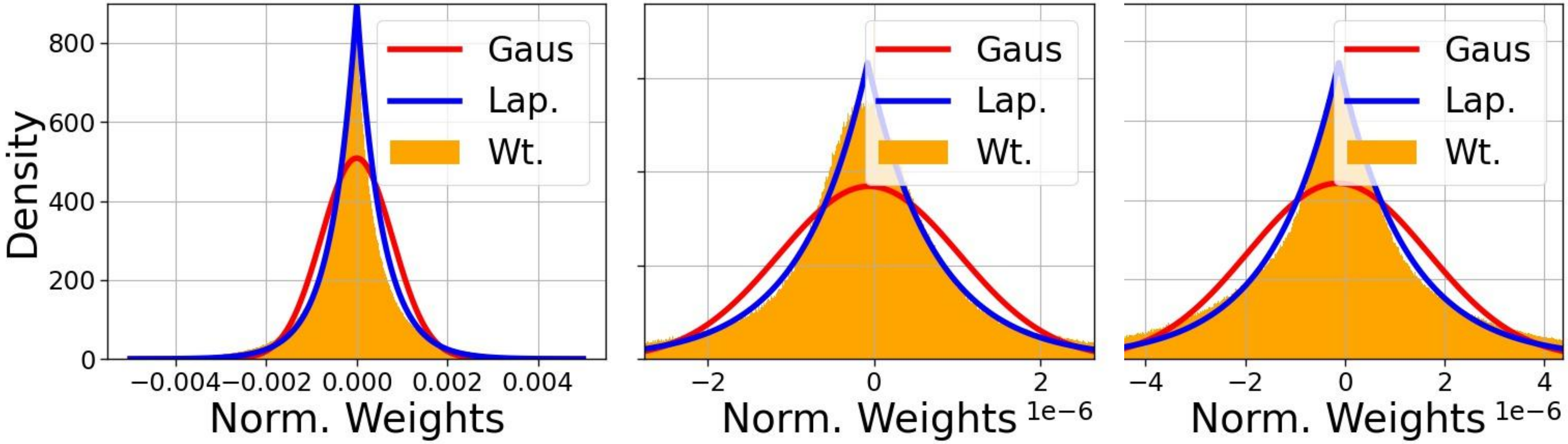
Theorem: Suppose $f(\cdot;\mathbf{w})$ is a fully-connected neural network model with d layers and 1-Lipschitz activations, e.g., ReLU. Let $\widehat{\mathbf{w}}$ be the reconstructed weight (after compression) where all layers are subject to compression. Then, we have the following bound on the output perturbation:

$$\sup_{|x|_1 \le 1} ||f(\mathbf{x},\mathbf{w}) - f(\mathbf{x},\widehat{\mathbf{w}})||_1 \le \left(\sum_{l=1}^{d} \frac{\left\|\mathbf{w}^{(\mathbf{l})} - \widehat{\mathbf{w}}^{(\mathbf{l})}\right\|_1}{\left\|\mathbf{w}^{(\mathbf{l})}\right\|_1}\right)\left(\prod_{k=1}^{d} \left\|\mathbf{w}^{(\mathbf{k})}\right\|_1\right)$$

Distortion function:

$$d(\boldsymbol{u},\widehat{\boldsymbol{u}}) = \frac{1}{n}\sum_{i=1}^{n} |u_i - \widehat{u}_i|, \quad \text{where } \boldsymbol{u}^{(l)} = \frac{\boldsymbol{w}^{(l)}}{|\boldsymbol{w}^{(l)}|_1}.$$

Density: Laplacian



Rate-Distortion Function

For an i.i.d. Laplacian source sequence distributed according to $f(u;\lambda) = \frac{\lambda}{2} e^{-\lambda|u|}$, maximum compression for l_1 distortion D is:

$$\mathrm{R(D)} = \begin{cases} -\log(\lambda D), & 0 \le \mathrm{D} \le \frac{1}{\lambda} \\ 0, & D > \frac{1}{\lambda} \end{cases}$$

$$
\begin{aligned}
I(U;V) &= H(U) - H(U|V)) \\
&= \log(\frac{2e}{\lambda}) - H(U|V) \\
&= \log(\frac{2e}{\lambda}) - H(U-V|V) \\
&\boxed{\geq \log(\frac{2e}{\lambda}) - H(U-V)} \\
&\boxed{\geq \log(\frac{2e}{\lambda}) - \log(2e\mathbb{E}[|U-V|])} \\
&= \log(\frac{2e}{\lambda}) - \log(2eD) \\
&= -\log(\lambda D)
\end{aligned}
$$

Maximum entropy theorem tells us that Laplace distribution with parameter α has the maximum differential entropy h(f) over all probability densities f satisfying $E[x] = 0$ and E[|x|] = $\frac{1}{\alpha}$ where $x \sim f$.

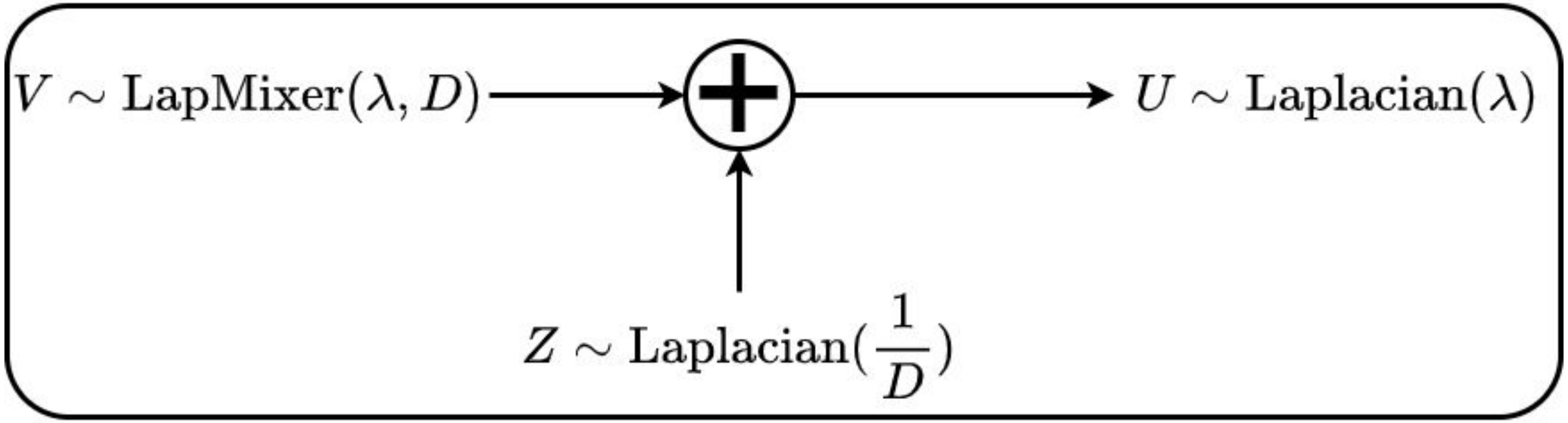
$U - V$ and V must be independent.

$U - V$ must be Laplace distributed with parameter $\frac{1}{D}$.

Conditions for Optimal Compression

1) $U - V$ and V must be independent.

2) $U - V$ must be Laplace distributed with parameter $\frac{1}{D}$.



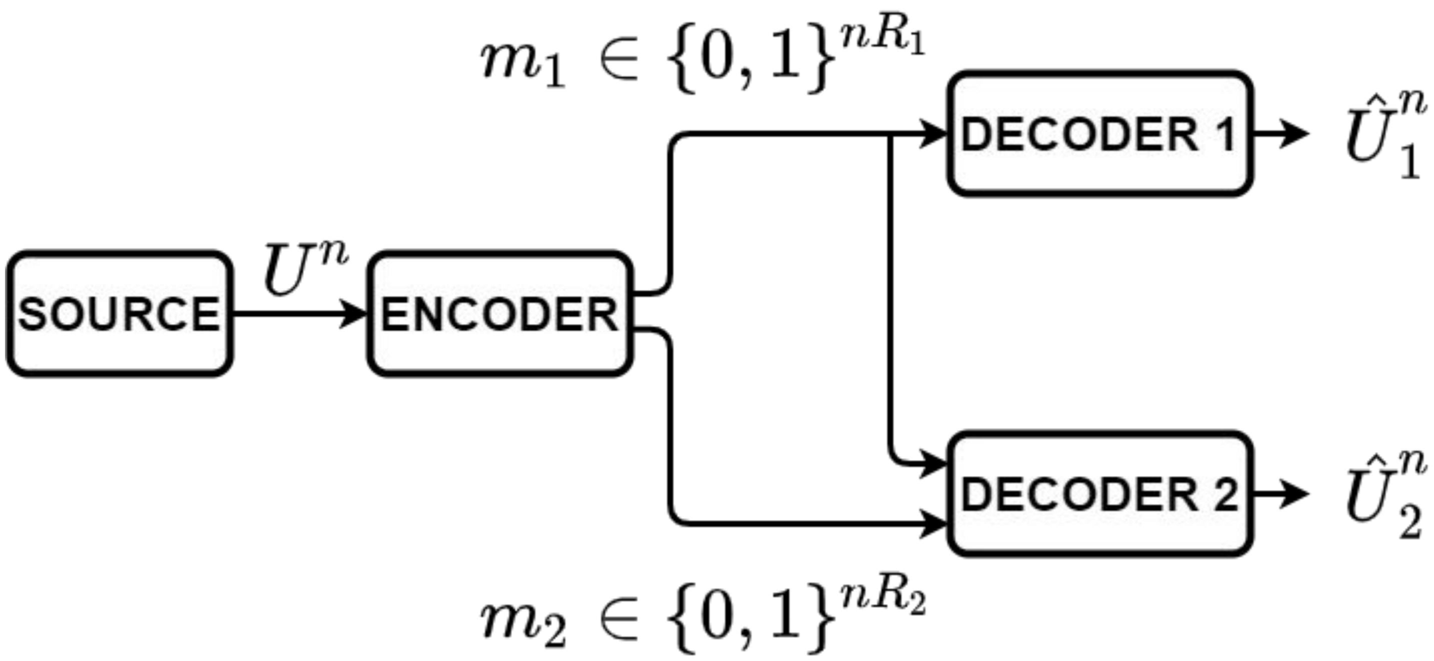
To achieve the maximum compression, we need a compression scheme with the following conditions:

1.Conditional probability distribution: $f_{U|V}(u|v) = \frac{1}{2D} e^{-|u-v|/D}$

2.Marginal probability distribution: $f_V(v) = \lambda^2 D^2 \cdot \delta(v) + (1 - \lambda^2 D^2) \cdot \frac{\lambda}{2} e^{-\lambda|v|}$

A new pruning algorithm:
Successive Refinement for Pruning (SuRP)

Successive Refinement



Recall the optimal marginal distribution:

$$f_{\mathbf{V}}(v) = \lambda^2 D^2 \cdot \delta(v) + (1 - \lambda^2 D^2) \cdot \frac{\lambda}{2} e^{-\lambda|v|}$$

First Attempt

- Consider successive refinement with L decoders.
- Let $\lambda = \lambda_1 < \cdots < \lambda_L$ where $D_t = 1/\lambda_{t+1}$ is the target distortion at the t-th decoder.
- Set $\boldsymbol{U}^{(1)} = u^{(n)}$.
- At the t-th iteration,
  - The encoder finds $V^{(t)}$ that minimizes $d(\boldsymbol{U}^{(t)}, \boldsymbol{V}^{(t)})$ from a codebook $\boldsymbol{C}^{(t)}$.
  - The encoder computes the residual $\boldsymbol{U}^{(t+1)} = \boldsymbol{U}^{(t)} - \boldsymbol{V}^{(t)}$.
  - The decoder reconstructs $\widehat{\boldsymbol{U}}^{(t)} = \sum_{\tau=1}^{t} \boldsymbol{V}^{(\tau)}$.

$$f_{\mathbf{V}^{(t)}}(v) = \frac{\lambda_t^2}{\lambda_{t+1}^2} \cdot \delta(v) + \left(1 - \frac{\lambda_t^2}{\lambda_{t+1}^2}\right) \cdot \frac{\lambda_t}{2} e^{-\lambda_t|v|}$$

- Complexity is $L \cdot 2^{nR/L}$, lower than the naïve random coding strategy with complexity 2^{nR}.

Successive Refinement for Pruning (SuRP)

Iteration 0:

$$\mathbf{U}^{(0)} \qquad \hat{\mathbf{U}}^{(0)} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Iteration 1:

$$\mathbf{U}^{(0)} \rightarrow \boxed{\mathsf{E}} \xrightarrow{\mathbf{m}_1} \boxed{\mathsf{D}} \longrightarrow \hat{\mathbf{U}}^{(1)} = \mathbf{V}^{(1)} = \begin{bmatrix} 0 \\ 1/\lambda_0 \\ -1/\lambda_0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Iteration 2:

$$\mathbf{U}^{(1)} = \mathbf{U}^{(0)} - \mathbf{V}^{(1)} \rightarrow \boxed{\mathsf{E}} \xrightarrow{\mathbf{m}_2} \boxed{\mathsf{D}} \rightarrow \hat{\mathbf{U}}^{(2)} = \hat{\mathbf{U}}^{(1)} + \mathbf{V}^{(2)} = \begin{bmatrix} 0 \\ 1/\lambda_0 + 1/\lambda_1 \\ -1/\lambda_0 \\ -1/\lambda_1 \\ 0 \\ 0 \end{bmatrix}$$

Iteration 3:

$$\mathbf{U}^{(2)} = \mathbf{U}^{(1)} - \mathbf{V}^{(2)} \rightarrow \boxed{\mathsf{E}} \xrightarrow{\mathbf{m}_3} \boxed{\mathsf{D}} \rightarrow \hat{\mathbf{U}}^{(3)} = \hat{\mathbf{U}}^{(2)} + \mathbf{V}^{(3)} = \begin{bmatrix} 0 \\ 1/\lambda_0 + 1/\lambda_1 \\ -1/\lambda_0 \\ -1/\lambda_1 - 1/\lambda_2 \\ 0 \\ 1/\lambda_2 \end{bmatrix}$$

Successive Refinement for Pruning (SuRP)

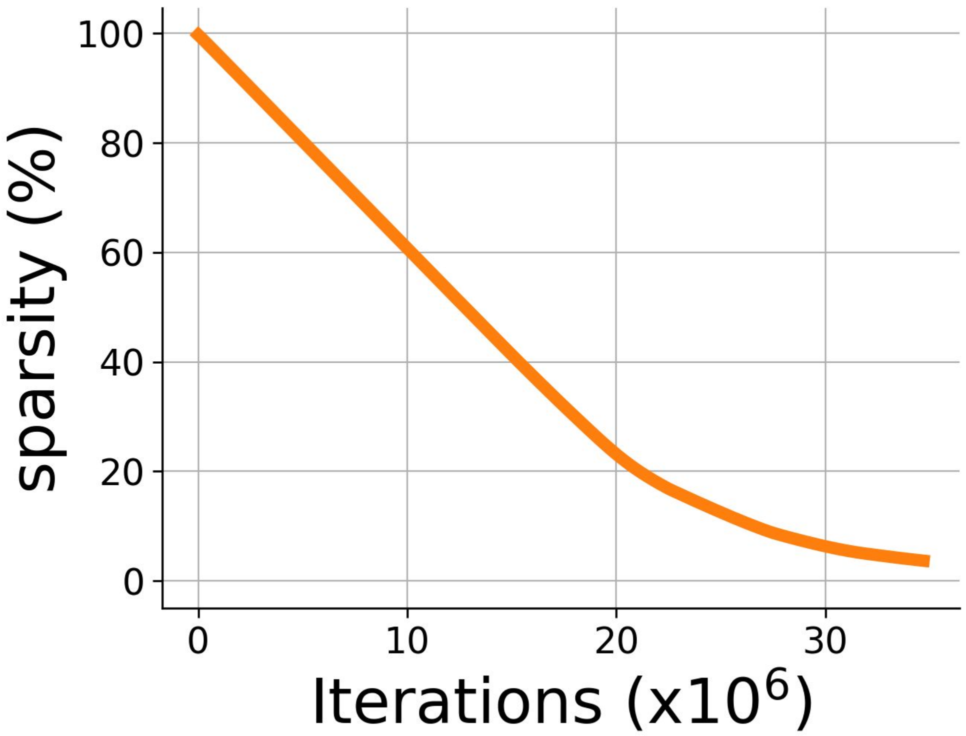
- Consider the successive refinement problem with L decoders.
- Each decoder corresponds to one iteration of SuRP (different from a pruning iteration).
- Set $\mathbf{U}^{(1)} = u^n$. For iteration $1 \le t \le L-1$:

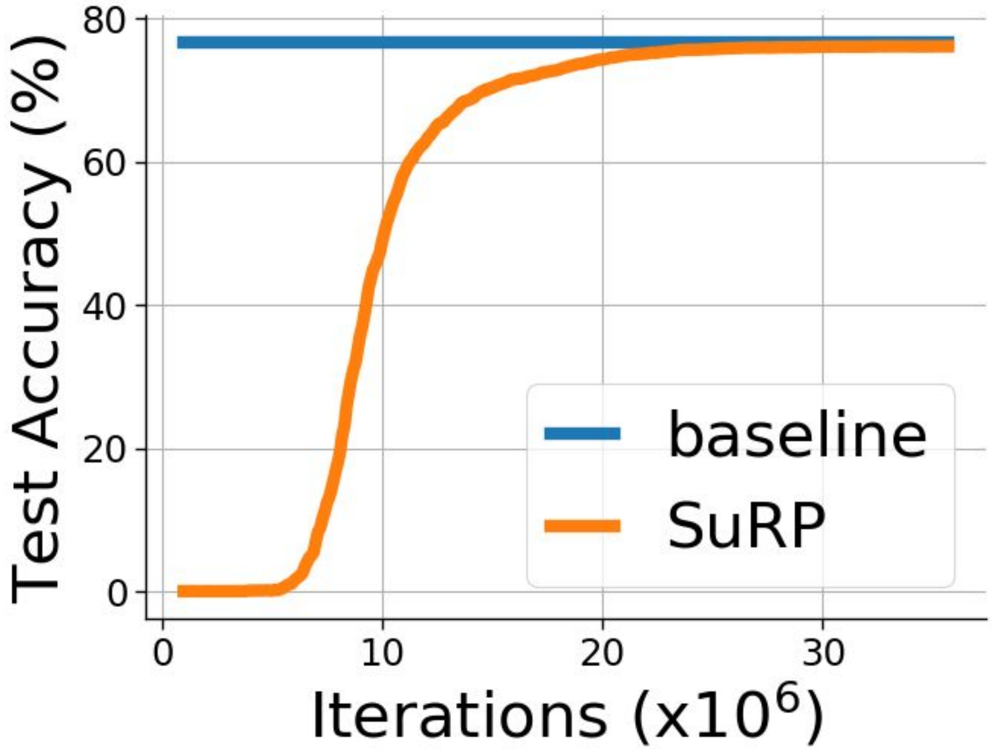
1. Find index i and j such that $\mathbf{U}_i^{(\mathbf{t})} \ge \frac{1}{\lambda_t}\log\frac{n}{2\beta}$ and $\mathbf{U}_j^{(\mathbf{t})} \le -\frac{1}{\lambda_t}\log\frac{n}{2\beta}$. If there are more than one such indices, pick an index i (or j) randomly. Encode (i, j) as m_t.
2. Let $\mathbf{V}^{(\mathbf{t})}$ be an n-dimensional all-zero vector except $\mathbf{V}_{\mathbf{i}}^{(\mathbf{t})} = \frac{1}{\lambda_t}\log\frac{n}{2\beta}$ and $\mathbf{V}_j^{(\mathbf{t})} = -\frac{1}{\lambda_t}\log\frac{n}{2\beta}$.
3. Let $\mathbf{U}^{(\mathbf{t+1})} = \mathbf{U}^{(\mathbf{t})} - \mathbf{V}^{(\mathbf{t})}$.
4. Set $\lambda_{t+1}^2 = \frac{n}{n-2\log\frac{n}{2\beta}} \cdot \lambda_t^2$.

Recall the optimal marginal distribution:

$$f_{\mathbf{V}}(v) = \lambda^2 D^2 \cdot \delta(v) + (1-\lambda^2 D^2)\cdot\frac{\lambda}{2}e^{-\lambda|v|}$$

Results


100
80
60
40
20
0
sparsity (%)
0
10
20
30
Iterations (x10^6)


80
60
40
20
0
Test Accuracy (%)
0
10
20
30
Iterations (x10^6)
baseline
SuRP

Results

CIFAR-10

	Pruning Ratio:	95.60%	98.20%	98.85%	99.26%	99.53%	99.81%
VGG-16	Global [3]	90.80	85.55	81.56	54.58	41.91	21.87
	Uniform [4]	90.78	84.17	55.68	38.51	26.41	11.58
	Adaptive [1]	91.20	89.44	87.85	86.53	84.84	74.54
	LAMP [2]	92.06	91.66	91.07	90.49	89.64	87.07
	SuRP (ours)	**92.13**	**91.72**	**91.21**	**90.73**	**90.65**	**87.28**
	Pruning Ratio:	86.58%	94.50%	96.48%	97.75%	98.56%	99.41%
ResNet-20	Global [3]	86.97	85.02	83.15	80.52	76.28	47.47
	Uniform [4]	86.70	84.53	82.05	77.19	64.24	20.45
	Adaptive [1]	87.00	85.00	83.23	80.40	76.40	52.06
	LAMP [2]	87.12	85.64	84.18	81.56	78.63	67.01
	SuRP (ours)	**90.44**	**88.87**	**87.05**	**83.98**	**79.00**	**70.64**

ResNet-50 on ImageNet

Pruning Ratio:	80%	90%
Adaptive [27]	75.60	73.90
SNIP [59]	72.00	67.20
DSR [74]	73.30	71.60
SNFS [18]	74.90	72.90
RiGL [22]	74.60	72.00
SuRP (ours)	**75.54**	**73.93**

Sparse Random Networks for Communication-Efficient Federated Learning

Joint work with

Francesco Pase (University of Padova)

Deniz Gunduz (Imperial College London)

Tsachy Weissman (Stanford University)

Michele Zorzi (University of Padova)

Contributions

1. Existence of subnetworks inside larger networks with **random weights** that perform well on clients' **non-iid** dataset.

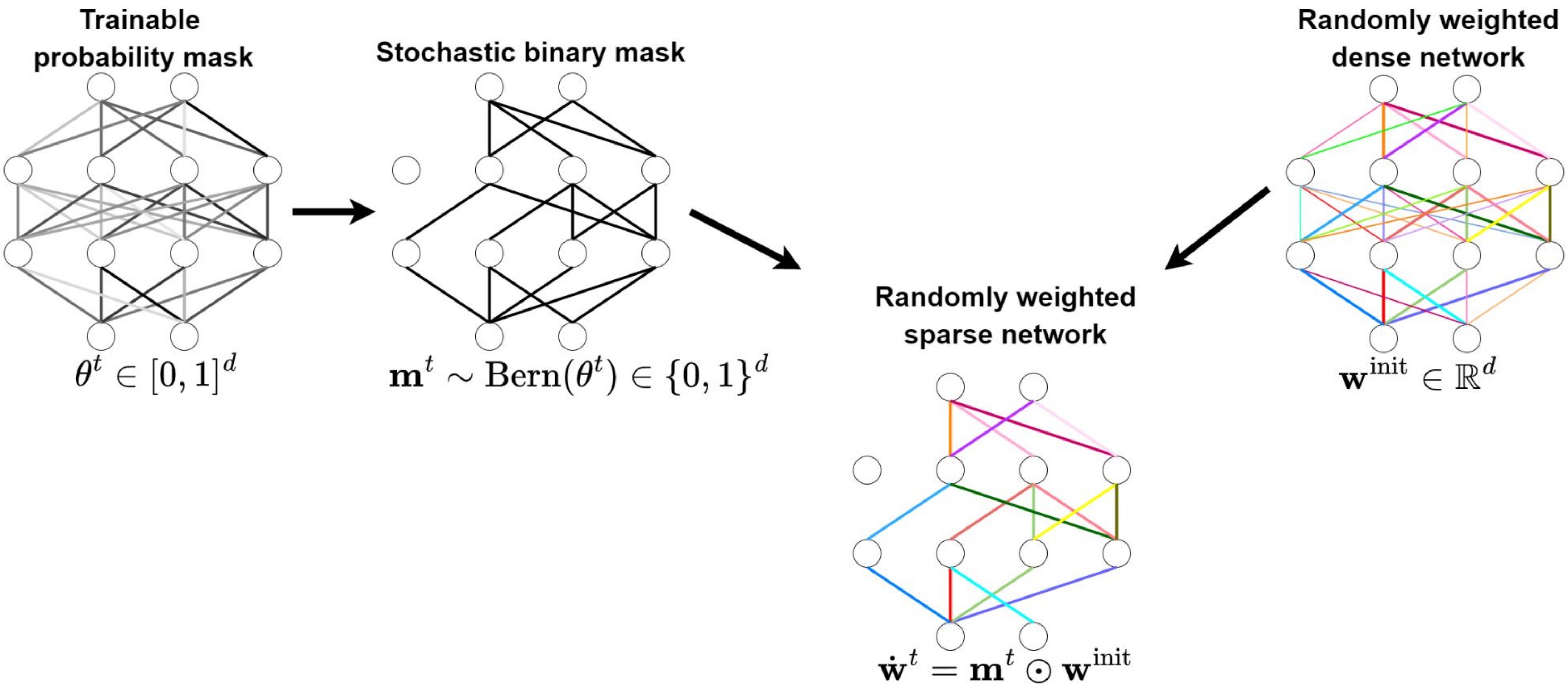
2. Finding these subnetworks in a **communication-efficient** way. (less than 1 bpp)

3. **Fast convergence**.

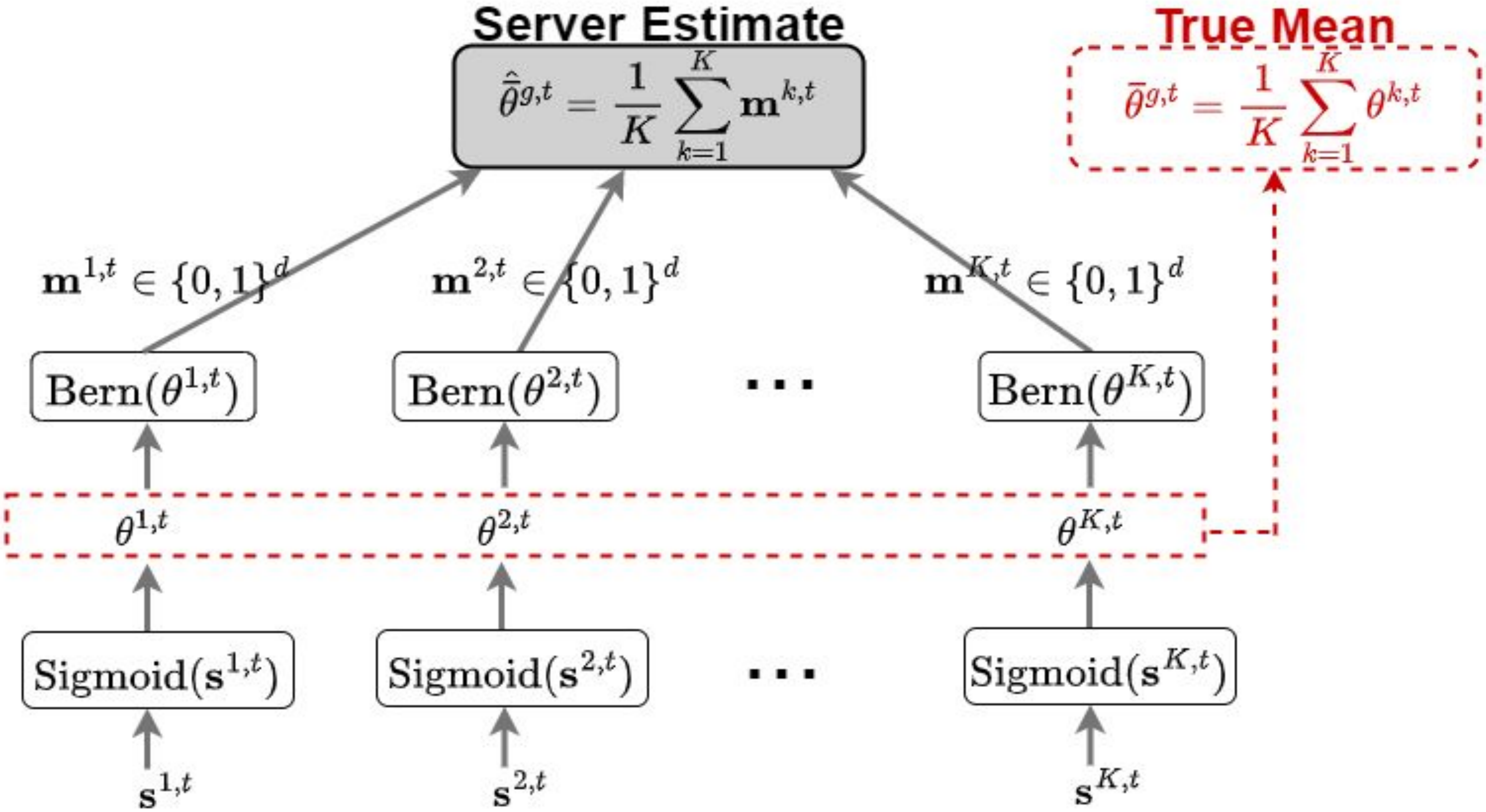
4. **Efficient representation** of the final model. (less than 1 bpp)

5. **Privacy** amplification in the presence of LDP mechanisms.

FedPM


Trainable probability mask
$\theta^t \in [0,1]^d$
Stochastic binary mask
$\mathbf{m}^t \sim \mathrm{Bern}(\theta^t) \in \{0,1\}^d$
Randomly weighted dense network
$\mathbf{w}^{\mathrm{init}} \in \mathbb{R}^d$
Randomly weighted sparse network
$\dot{\mathbf{w}}^t = \mathbf{m}^t \odot \mathbf{w}^{\mathrm{init}}$

Communication Strategy



Communication Strategy

True Mean: $\bar{\theta}^{g,t} = \frac{1}{K}\sum^{K}\theta^{k}$

Estimated Mean: $\hat{\bar{\theta}}^{g,t} = \frac{1}{K}\sum_{k=1}^{K} m^{k,t}$

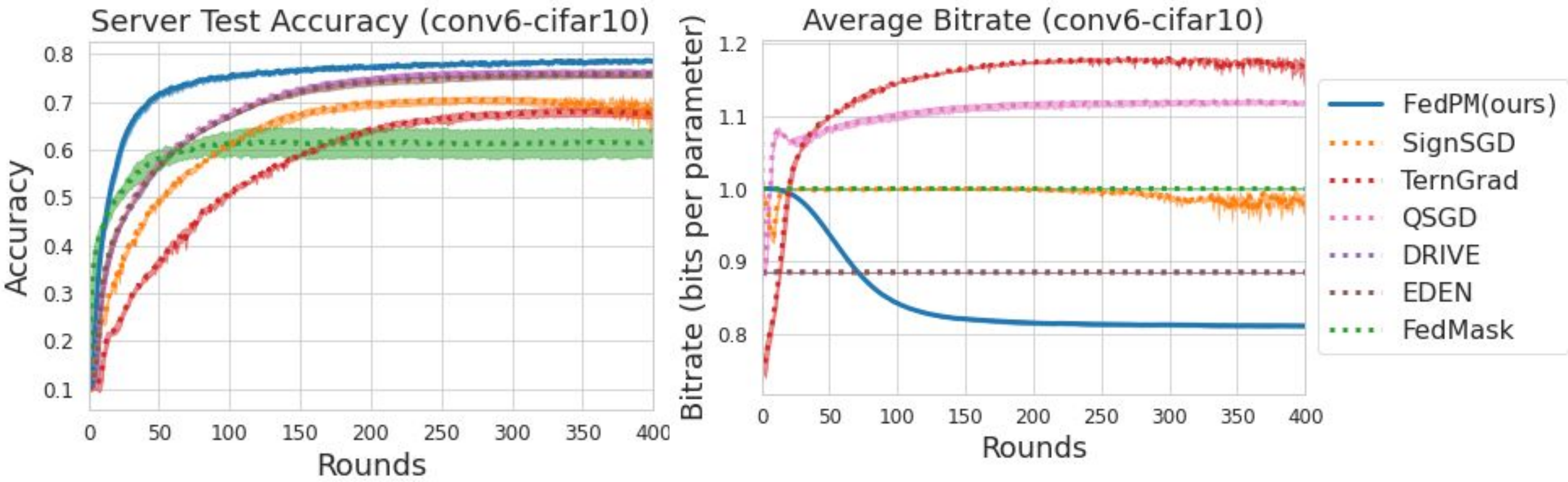
- Unbiased Estimate:

$$\mathbb{E}_{M^{k,t}\sim \mathrm{Bern}(\theta^{k,t})\ \forall k\in\mathcal{K}_t}[\hat{\bar{\theta}}^{g,t}] = \bar{\theta}^{g,t}$$

- Error:

$$\mathbb{E}_{M^{k,t}\sim \mathrm{Bern}(\theta^{k,t})\ \forall k\in\mathcal{K}_t}\left[||\hat{\bar{\theta}}^{g,t} - \bar{\theta}^{g,t}||_2^2\right] \leq \frac{d}{4K}$$

Results (CIFAR-10)



Stanford University

Bayesian Aggregation

We can model the probability mask with a Beta distribution.

$$\alpha^{g,t} = \alpha^{g,t-1} + M^{\text{agg},t}$$

$$\beta^{g,t} = \beta^{g,t-1} + K \cdot \mathbf{1} - M^{\text{agg},t}$$

where

$$M^{\text{agg},t} = \sum_{k \in \mathcal{K}_t} M^{k,t}$$

$$\boxed{\theta^{g,t} = \frac{\alpha^{g,t} - 1}{\alpha^{g,t} + \beta^{g,t} - 2}}$$

Results

	Algorithm	$\rho = 1$	$\rho = 0.5$	$\rho = 0.1$
$c_{\max} = 4$	DRIVE (Vargaftik et al., 2021)	0.739 ± 0.005	0.632 ± 0.010	0.405 ± 0.018
	EDEN (Vargaftik et al., 2022)	0.717 ± 0.006	0.665 ± 0.012	0.360 ± 0.016
	QSGD (Alistarh et al., 2017)	0.709 ± 0.006	0.644 ± 0.014	0.399 ± 0.020
	FedMask (Li et al., 2021)	0.531 ± 0.044	0.435 ± 0.057	0.362 ± 0.024
	FedPM (Ours)	$\mathbf{0.748 \pm 0.003}$	$\mathbf{0.720 \pm 0.007}$	$\mathbf{0.496 \pm 0.007}$
$c_{\max} = 2$	DRIVE (Vargaftik et al., 2021)	0.434 ± 0.025	0.376 ± 0.014	0.221 ± 0.003
	EDEN (Vargaftik et al., 2022)	0.535 ± 0.050	0.461 ± 0.016	0.219 ± 0.005
	QSGD (Alistarh et al., 2017)	0.476 ± 0.033	0.464 ± 0.002	0.243 ± 0.014
	FedMask (Li et al., 2021)	0.420 ± 0.028	0.387 ± 0.062	0.197 ± 0.030
	FedPM (Ours)	$\mathbf{0.643 \pm 0.016}$	$\mathbf{0.556 \pm 0.031}$	$\mathbf{0.277 \pm 0.003}$

Table 1: Average final accuracy $\pm\sigma$ in non-IID data split with $c_{\max} = 4$ and $c_{\max} = 2$, and partial participation with ratios $\rho = \{0.1, 0.5, 1\}$, for FedPM, FedMask, and the strongest baselines in the IID experiments: EDEN, DRIVE, and QSGD. The training duration was set to $t_{\max} = 200$ rounds.

Thank You!

An Information-Theoretic Justification for Model Pruning (AISTATS'22)

https://arxiv.org/pdf/2102.08329.pdf

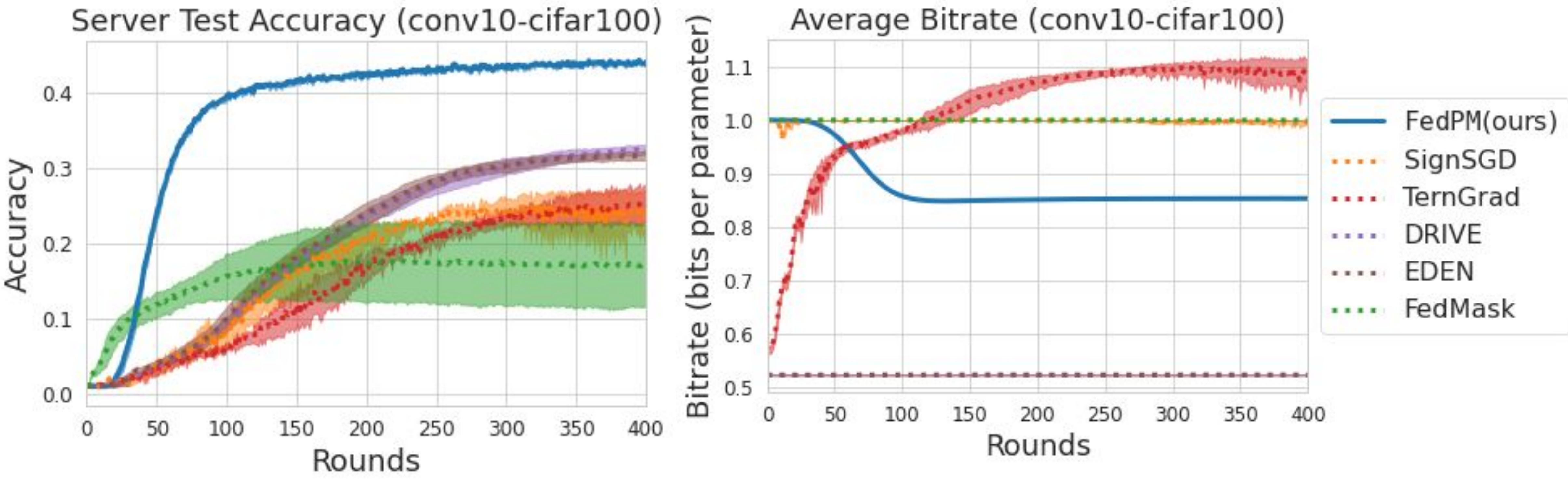
Arxiv **Code**

Sparse Random Networks for Communication-Efficient Federated Learning

https://arxiv.org/pdf/2209.15328.pdf

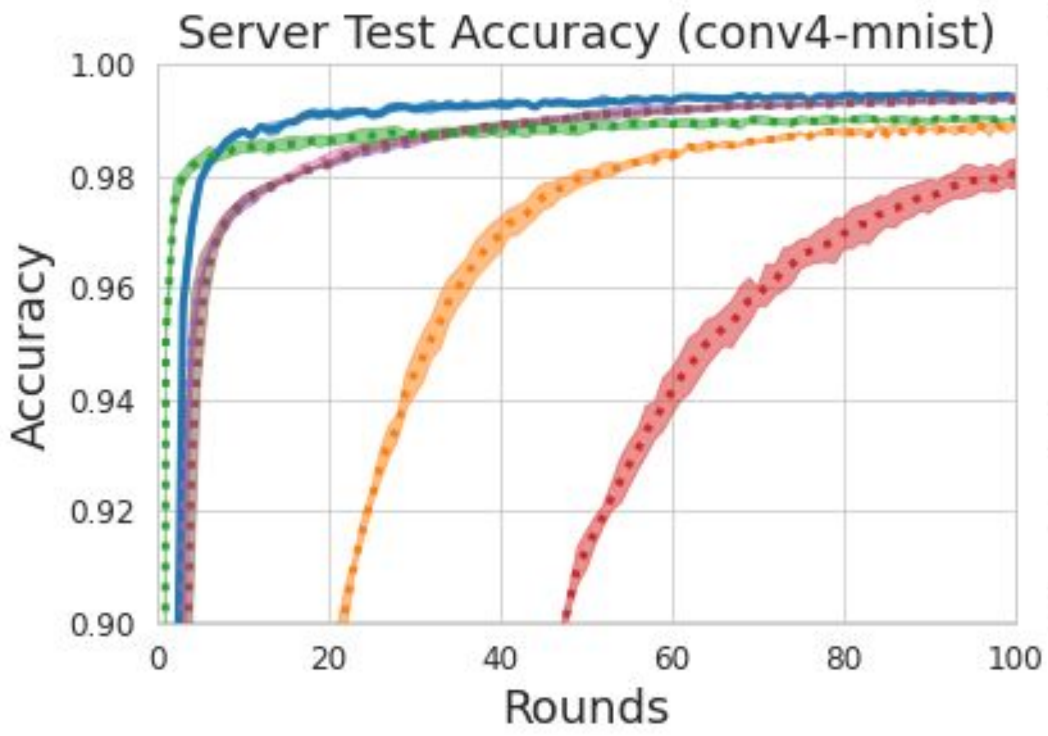
Arxiv

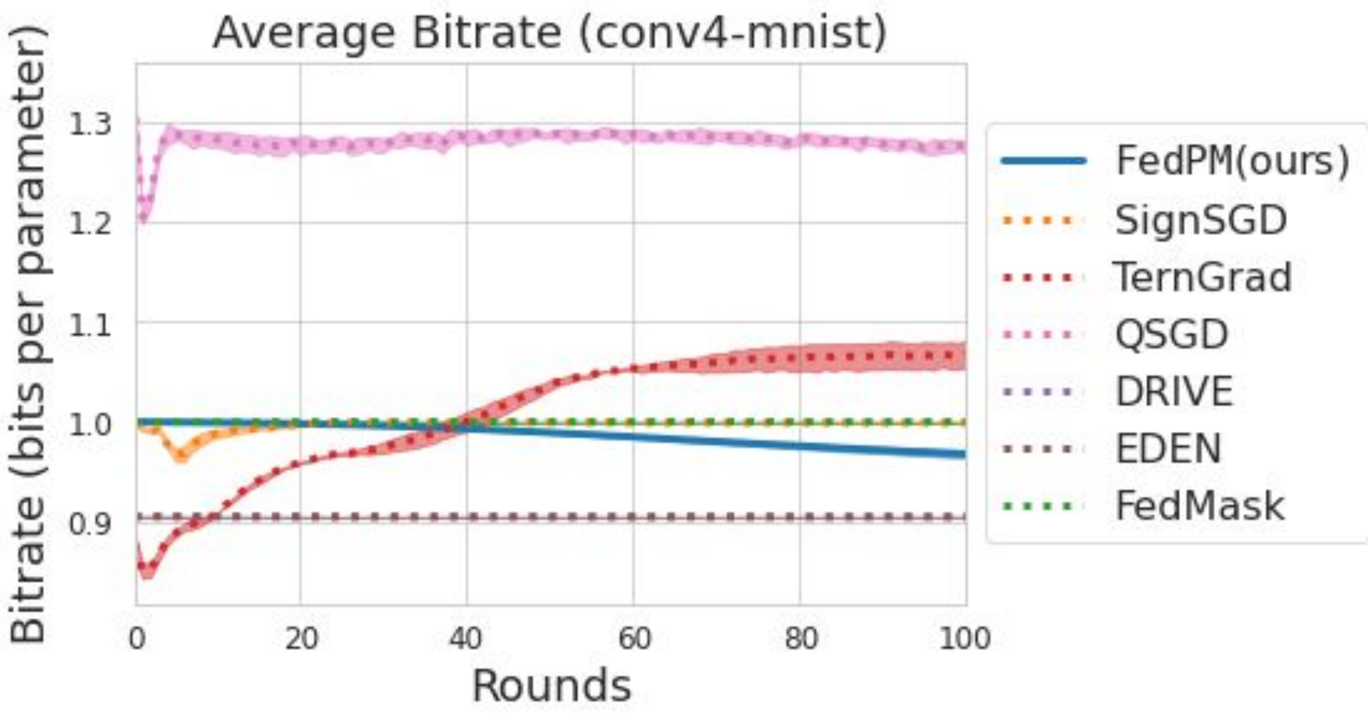
Results (CIFAR-100)



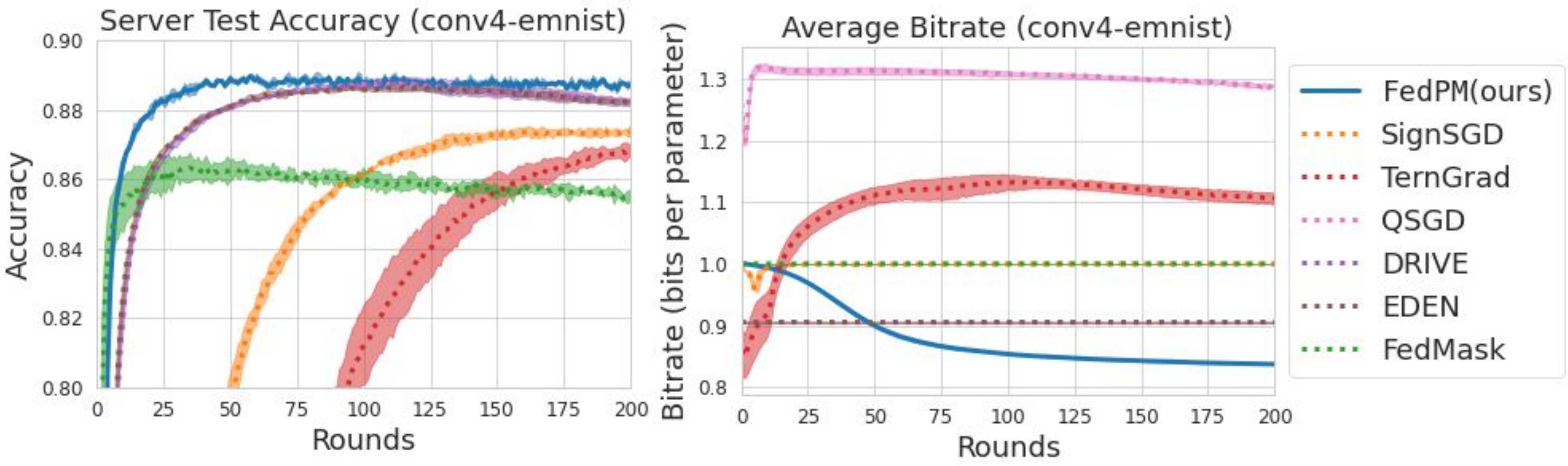
Stanford University

Results (MNIST)





Results (EMNIST)



Results

	Algorithm	$\rho = 1$	$\rho = 0.5$	$\rho = 0.1$
	DRIVE (Vargaftik et al., 2021)	$0.885 \pm 9 \cdot 10^{-5}$	$\mathbf{0.885 \pm 1 \cdot 10^{-4}}$	$\mathbf{0.885 \pm 1 \cdot 10^{-4}}$
	EDEN (Vargaftik et al., 2022)	$0.885 \pm 1 \cdot 10^{-4}$	$\mathbf{0.885 \pm 1 \cdot 10^{-4}}$	$\mathbf{0.885 \pm 1 \cdot 10^{-4}}$
$c_{\max} = 4$	QSGD (Alistarh et al., 2017)	0.982 ± 0.027	0.923 ± 0.029	0.91 ± 0.05
	FedMask (Li et al., 2021)	$1 \pm 3 \cdot 10^{-6}$	$1 \pm 8 \cdot 10^{-8}$	$1 \pm 6 \cdot 10^{-7}$
	FedPM (Ours)	$\mathbf{0.863 \pm 0.077}$	0.912 ± 0.056	0.996 ± 0.003
	DRIVE (Vargaftik et al., 2021)	$0.885 \pm 7 \cdot 10^{-5}$	$\mathbf{0.885 \pm 2 \cdot 10^{-4}}$	$\mathbf{0.885 \pm 2 \cdot 10^{-4}}$
	EDEN (Vargaftik et al., 2022)	$0.885 \pm 1 \cdot 10^{-4}$	$\mathbf{0.885 \pm 7 \cdot 10^{-5}}$	$\mathbf{0.885 \pm 7 \cdot 10^{-5}}$
$c_{\max} = 2$	QSGD (Alistarh et al., 2017)	1.230 ± 0.043	1.234 ± 0.038	1.082 ± 0.01
	FedMask (Li et al., 2021)	$1 \pm 2 \cdot 10^{-6}$	$1 \pm 2 \cdot 10^{-6}$	$1 \pm 2 \cdot 10^{-7}$
	FedPM (Ours)	$\mathbf{0.868 \pm 0.076}$	0.904 ± 0.063	0.997 ± 0.01

Table 3: Average bitrate $\pm\sigma$ over the whole training process in non-IID data split with $c_{\max} = 4$ and $c_{\max} = 2$, and partial participation with ratios $\rho = \{0.1, 0.5, 1\}$, for FedPM, FedMask, and the strongest baselines in the IID experiments: EDEN, DRIVE, and QSGD. The training duration was set to $t_{\max} = 200$ rounds.