# Bayesian SignSGD Optimizer for Federated Learning

**Paulo Abelha Ferreira** **Pablo Nascimento da Silva** **Vinicius Gottin**

**Roberto Stelling** **Tiago Calmon**

Dell Technologies, Research Office, Rio de Janeiro, RJ, Brazil
{paulo.ferreira, pablo.dasilva, vinicius.gottin, roberto.stelling, tiago.calmon}@dell.com

## A Majority voting SIGNSGD equivalence to BB-SIGNSGD %1

In this section we show how majority voting SIGNSGD (i.e. combined with FEDAVG) is a special case of the BB-SignSGD framework (please see section 4[1]). The three choices that gets us there are: (1) a uniform prior $\alpha_j = \beta_j = 1$ at every cycle; (2) $\theta_j$ as the mode of the beta posterior; and (3) $g_j = \mathbb{E}[\mathsf{g_j}; \theta_j]$ as the expected value of the updated Bernoulli distribution.

The posterior parameters for the beta distribution can be recovered easily as $\alpha_j = \alpha_j + S_j^1$ and $\beta_j = \beta_j + S_j^0$, where $S_j^0$ and $S_j^1$ are the respective total sums of $0s$ and $1s$ observed across the $M$ workers for gradient component $j$. Please note that $S_j^0 + S_j^1 = M$. Also please note that the mode of a posterior beta distribution can be obtained from the beta prior parameters as

$$\text{mode}(\text{beta}(\theta_j; \alpha_j, \beta_j)) = \frac{\alpha_j + S_j^1 - 1}{\alpha_j + \beta_j + M - 2}. \tag{1}$$

Assuming a uniform prior (i.e. $\alpha_j = \beta_j = 1$), equation 9 reduces to $\text{mode}(\text{beta}(\theta_j; \alpha_j, \beta_j)) = \frac{S_j^1}{M}$ and hence by choosing $\theta_j$ to be the mode of the posterior, and take $g_j$ as the expected value of the Bernoulli distribution, we have that $\mathbb{E}[\mathsf{g_j}; \theta_j] = \theta_j = \frac{S_j^1}{M}$.

Let $\bar{g}_j^* = \frac{1}{M} \sum_{i=1}^M d(g_{i,j}^c)$ be the global gradient component $j$ for majority voting SIGNSGD, prior to central compression.

Let $\bar{g}_j^+ = d\left(\frac{S_j^1}{M}\right)$ be the global gradient component $j$ obtained from the three choices above in the BB-SIGNSGD framework, prior to central compression.

We need only show that these two gradients $\bar{g}_j^*$ and $\bar{g}_j^+$ are equal, since central compression and further steps are the same in both frameworks.

We start by noting that $S_j^1 = \sum_{i=1}^M g_{i,j}^c$, hence $\bar{g}_j^+ = d\left(\frac{1}{M} \sum_{i=1}^M g_{i,j}^c\right)$. Through simple algebraic manipulation: $d\left(\frac{1}{M} \sum_{i=1}^M g_{i,j}^c\right) = \frac{2}{M}\left(\sum_{i=1}^M g_{i,j}^c\right) - 1 = \frac{1}{M}\left(\sum_{i=1}^M 2g_{i,j}^c - \sum_{i=1}^M 1\right) = \frac{1}{M} \sum_{i=1}^M \left(2g_{i,j}^c - 1\right) = \frac{1}{M} \sum_{i=1}^M d(g_{i,j}^c)$. Hence $\bar{g}_j^+ = \bar{g}_j^* = \frac{1}{M} \sum_{i=1}^M d(g_{i,j}^c)$. Please note that this proof works for any component $j \in \{1, 2, ...\text{D}\}$, since they are all assumed to be independent. □

[1] Main content might be separated from this file, but should be available along with it.

1st NeurIPS Workshop on New Frontiers in Federated Learning (NFFL 2021), Virtual Meeting.

## B Additional Experimental Results

In this section we show tables and figures related to all the performed experiments for this work. There are two types of tables: convergence and accuracy tables. Tables 3 and 4 are convergence tables. In these, we show, for different datasets, we show the relative convergence of all methods to BB-SIGNSGD %1. To calculate this, we first find the maximum accuracy achieved by BB-SIGNSGD %1 across all 100 cycles. Then, for any particular method's accuracy evolution, we report the cycle at which it achieves at least 99% of BB-SIGNSGD %1's maximum accuracy. Please note that, for the case of BB-SIGNSGD %1 itself, we report the cycle at which it achieves at least 99% of its own maximum accuracy. We report the mean and standard deviation for three experiments running for three different starting seeds.

Table 3: Convergence relative to BB-SIGNSGD %1 (i.e. SIGNSGD), with central learning rate $\eta_c = 1\text{e}{-3}$. We report the cycle at which the given method achieves at least 99% of BB-SIGNSGD %1's maximum accuracy. Values are mean and standard deviation for three different starting seeds.

| | | BB-SIGNSGD | | | | |
|---|---|---|---|---|---|---|
| Dataset | No compr. | %1 | %2 | %5 | %10 | %Inf |
| FashionMNIST | $13 \pm 0$ | $\mathbf{22 \pm 2}$ | $23 \pm 2$ | $\mathbf{22 \pm 2}$ | $24 \pm 0$ | $48 \pm 1$ |
| CIFAR10 | $14 \pm 1$ | $26 \pm 2$ | $24 \pm 3$ | $\mathbf{22 \pm 2}$ | $\mathbf{22 \pm 2}$ | $34 \pm 2$ |
| EMNIST-Digits | $4 \pm 0$ | $\mathbf{10 \pm 0}$ | $\mathbf{10 \pm 0}$ | $\mathbf{10 \pm 0}$ | $12 \pm 0$ | $18 \pm 1$ |
| EMNIST-Letters | $13 \pm 1$ | $26 \pm 0$ | $26 \pm 1$ | $\mathbf{24 \pm 0}$ | $26 \pm 1$ | $70 \pm 4$ |
| EMNIST-Balanced | $13 \pm 1$ | $28 \pm 1$ | $\mathbf{27 \pm 2}$ | $\mathbf{27 \pm 2}$ | $29 \pm 0$ | – |

Table 4: Convergence relative to BB-SIGNSGD %1 (i.e. SIGNSGD), with central learning rate $\eta_c = 5\text{e}{-3}$. We report the cycle at which the given method achieves at least 99% of BB-SIGNSGD %1's maximum accuracy. Values are mean and standard deviation for three different starting seeds.

| | | BB-SIGNSGD | | | | |
|---|---|---|---|---|---|---|
| Dataset | No compr. | %1 | %2 | %5 | %10 | %Inf |
| FashionMNIST | $7 \pm 0$ | $57 \pm 2$ | $19 \pm 1$ | $\mathbf{10 \pm 1}$ | $13 \pm 0$ | $\mathbf{10 \pm 2}$ |
| CIFAR10 | $6 \pm 0$ | $48 \pm 6$ | $28 \pm 5$ | $8 \pm 0$ | $\mathbf{7 \pm 0}$ | $\mathbf{7 \pm 0}$ |
| EMNIST-Digits | $3 \pm 0$ | $25 \pm 3$ | $\mathbf{2 \pm 0}$ | $5 \pm 1$ | $4 \pm 1$ | $4 \pm 1$ |
| EMNIST-Letters | $9 \pm 0$ | $58 \pm 8$ | $44 \pm 5$ | $\mathbf{11 \pm 5}$ | $14 \pm 1$ | $14 \pm 0$ |
| EMNIST-Balanced | $6 \pm 0$ | $40 \pm 4$ | $12 \pm 2$ | $\mathbf{7 \pm 1}$ | $13 \pm 0$ | $10 \pm 1$ |

Tables 5, 6 and 7 are accuracy tables. In these, we show, for different datasets, the maximum accuracy for each method. We report the mean and standard deviation for three experiments running with different starting seeds. All results in convergence and accuracy tables refer to the same set of experiments.

Table 5: Max Accuracy (%), with central learning rate $\eta_c = 1\text{e}{-3}$. Values are mean and standard deviation for three different starting seeds.

| | | BB-SIGNSGD | | | | |
|---|---|---|---|---|---|---|
| Dataset | No compr. | %1 | %2 | %5 | %10 | %Inf |
| FashionMNIST | $87.8 \pm 0.2$ | $87.6 \pm 0.2$ | $87.6 \pm 0.1$ | $87.6 \pm 0.2$ | $87.7 \pm 0.1$ | $87.5 \pm 0.0$ |
| CIFAR10 | $58.4 \pm 0.3$ | $58.3 \pm 0.3$ | $58.5 \pm 0.1$ | $58.8 \pm 0.3$ | $59.0 \pm 0.1$ | $58.7 \pm 0.3$ |
| EMNIST-Digits | $98.7 \pm 0.1$ | $98.4 \pm 0.0$ | $98.5 \pm 0.0$ | $98.6 \pm 0.0$ | $98.6 \pm 0.0$ | $98.4 \pm 0.0$ |
| EMNIST-Letters | $89.2 \pm 0.2$ | $89.2 \pm 0.1$ | $89.2 \pm 0.1$ | $89.4 \pm 0.1$ | $89.3 \pm 0.1$ | $88.8 \pm 0.1$ |
| EMNIST-Balanced | $82.4 \pm 0.1$ | $82.6 \pm 0.2$ | $82.6 \pm 0.2$ | $82.5 \pm 0.1$ | $82.6 \pm 0.2$ | $81.3 \pm 0.1$ |

Table 6: Max Accuracy (%), with central learning rate $\eta_c = 2\text{e}{-3}$. Values are mean and standard deviation for three different starting seeds.

| | | BB-SIGNSGD | | | | |
|---|---|---|---|---|---|---|
| Dataset | No compr. | %1 | %2 | %5 | %10 | %Inf |
| FashionMNIST | $87.8 \pm 0.2$ | $87.3 \pm 0.2$ | $87.6 \pm 0.1$ | $87.4 \pm 0.1$ | $87.7 \pm 0.2$ | $87.4 \pm 0.1$ |
| CIFAR10 | $58.4 \pm 0.3$ | $57.1 \pm 0.2$ | $57.9 \pm 0.2$ | $58.2 \pm 0.1$ | $58.7 \pm 0.3$ | $58.5 \pm 0.5$ |
| EMNIST-Digits | $98.7 \pm 0.1$ | $98.3 \pm 0.0$ | $98.4 \pm 0.0$ | $98.5 \pm 0.0$ | $98.5 \pm 0.0$ | $98.3 \pm 0.0$ |
| EMNIST-Letters | $89.2 \pm 0.2$ | $88.8 \pm 0.2$ | $89.1 \pm 0.1$ | $88.9 \pm 0.3$ | $89.2 \pm 0.3$ | $88.7 \pm 0.0$ |
| EMNIST-Balanced | $82.4 \pm 0.1$ | $81.9 \pm 0.1$ | $82.2 \pm 0.2$ | $81.9 \pm 0.1$ | $82.2 \pm 0.1$ | $81.2 \pm 0.2$ |

Table 7: Max Accuracy (%), with central learning rate $\eta_c = 5\text{e}{-3}$. Values are mean and standard deviation for three different starting seeds.

| | | BB-SIGNSGD | | | | |
|---|---|---|---|---|---|---|
| Dataset | No compr. | %1 | %2 | %5 | %10 | %Inf |
| FashionMNIST | $87.8 \pm 0.2$ | $86.1 \pm 0.2$ | $86.5 \pm 0.0$ | $86.9 \pm 0.2$ | $87.0 \pm 0.2$ | $87.0 \pm 0.1$ |
| CIFAR10 | $58.4 \pm 0.3$ | $53.5 \pm 0.2$ | $53.9 \pm 0.8$ | $55.5 \pm 0.1$ | $56.1 \pm 0.0$ | $57.1 \pm 0.6$ |
| EMNIST-Digits | $98.7 \pm 0.1$ | $97.9 \pm 0.0$ | $98.2 \pm 0.0$ | $98.2 \pm 0.0$ | $98.2 \pm 0.1$ | $98.3 \pm 0.1$ |
| EMNIST-Letters | $89.2 \pm 0.2$ | $88.1 \pm 0.1$ | $88.0 \pm 0.1$ | $87.8 \pm 0.2$ | $88.1 \pm 0.1$ | $88.3 \pm 0.0$ |
| EMNIST-Balanced | $82.4 \pm 0.1$ | $79.3 \pm 0.1$ | $79.3 \pm 0.1$ | $80.4 \pm 0.2$ | $80.4 \pm 0.0$ | $80.5 \pm 0.2$ |

Figures 4-21 in this section are shown in triplets, one for each dataset. Each row shows the accuracy evolution for the same dataset, for all methods. And each column has a figure for a different central learning rate $\eta_c$.

Figures 22 and 23 show, for different datasets, the relative convergence ratio of all methods as relative to SIGNSGD. We have the no compression method, where full 32-bit precision is used during communication, and then variations of BB-SIGNSGD, where BB-SIGNSGD %1 is equivalent to SIGNSGD. We can see in Figure 22 that a small central learning rate of $\eta_c = 1\text{e}{-3}$ leads to the following two Bayesian variations being equivalent: BB-SIGNSGD %1 and %2. In Figure 23, using the best learning rate for BB-SignSGD %1 of $\eta_c = 2\text{e}{-3}$,we still have the same two Bayesian variations being comparable for most datasets. Please note that, as shown in Figure 1[2], we can increase the learning rate to $\eta_c = 5\text{e}{-3}$ and have Bayesian variation BB-SIGNSGD %5 outperform BB-SIGNSGD %1 by a very large margin (see also Table 4), being competitive even with the no compression method.

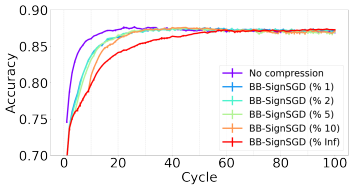


Figure 4: FashionMNIST, $\eta_c = 1\text{e}{-3}$.

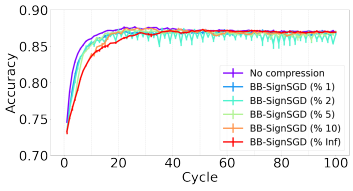


Figure 5: FashionMNIST, $\eta_c = 2\text{e}{-3}$.

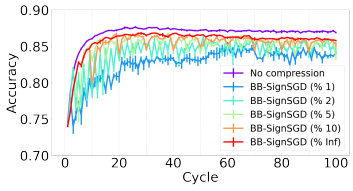


Figure 6: FashionMNIST, $\eta_c = 5\text{e}{-3}$.

[2]Main content might be separated from this file, but should be available along with it.

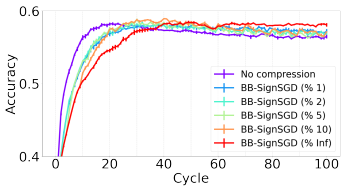


Figure 7: CIFAR10, $\eta_c = 1\text{e}{-3}$.

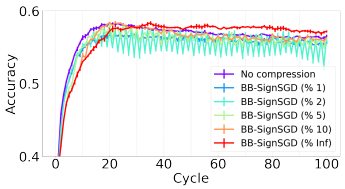


Figure 8: CIFAR10, $\eta_c = 2\text{e}{-3}$.

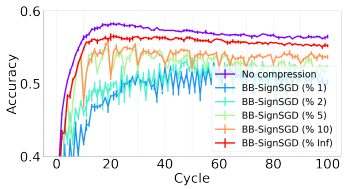


Figure 9: CIFAR10, $\eta_c = 5\text{e}{-3}$.

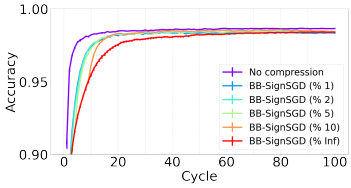


Figure 10: EMNIST-Digits, $\eta_c = 1\text{e}{-3}$.

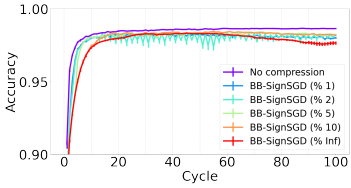


Figure 11: EMNIST-Digits, $\eta_c = 2\text{e}{-3}$.

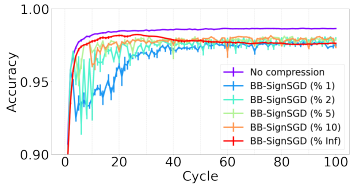


Figure 12: EMNIST-Digits, $\eta_c = 5\text{e}{-3}$.

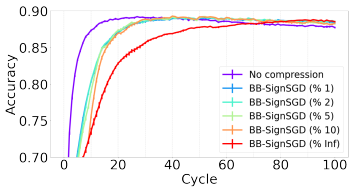


Figure 13: EMNIST-Letters, $\eta_c = 1\text{e}{-3}$.

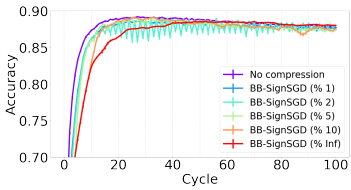


Figure 14: EMNIST-Letters, $\eta_c = 2\text{e}{-3}$.

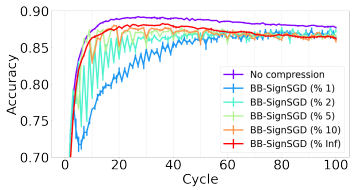


Figure 15: EMNIST-Letters, $\eta_c = 5\text{e}{-3}$.

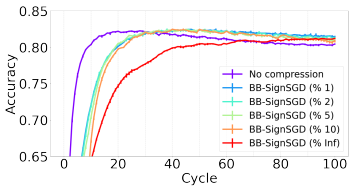


Figure 16: EMNIST-Balanced, $\eta_c = 1\text{e}{-3}$.

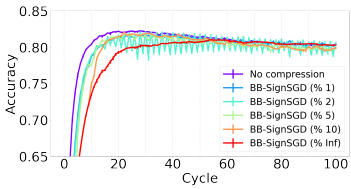


Figure 17: EMNIST-Balanced, $\eta_c = 2\text{e}{-3}$.

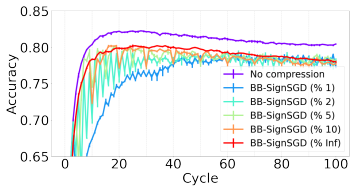


Figure 18: EMNIST-Balanced, $\eta_c = 5\text{e}{-3}$.

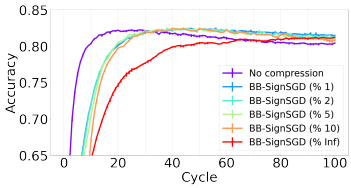


Figure 19: EMNIST-Balanced, $\eta_c = 1\text{e}{-3}$.

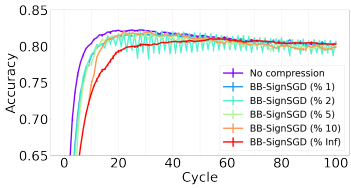


Figure 20: EMNIST-Balanced, $\eta_c = 2\text{e}{-3}$.

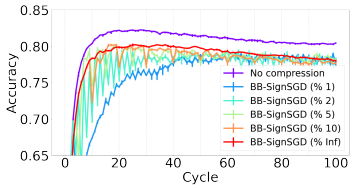


Figure 21: EMNIST-Balanced, $\eta_c = 5\text{e}{-3}$.

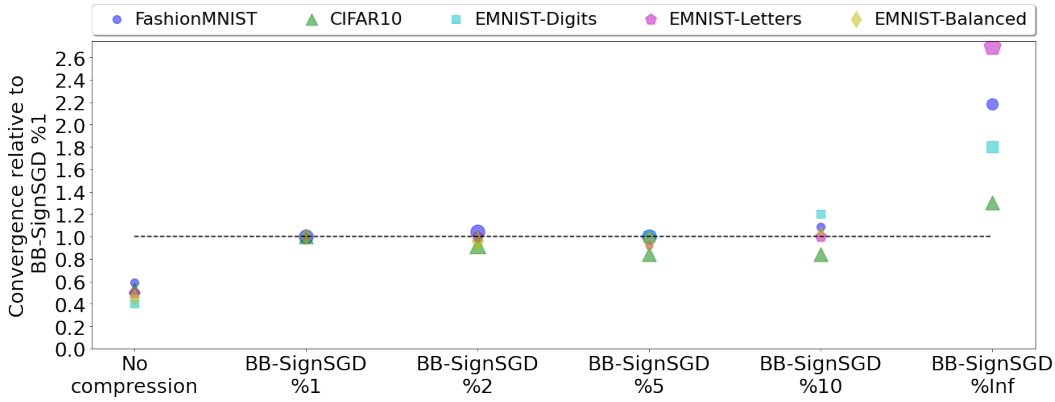


Figure 22: Converge (cycles) ratio relative to SIGNSGD (i.e. BB-SIGNSGD %1), measured as how fast the method achieved an accuracy at least 99% that of the maximum accuracy achieved by BB-SIGNSGD %1 (0.25 implies a method 4x faster). Size of marker indicates variance in experiments. Central learning rate $\eta_c = 1\text{e}{-3}$.

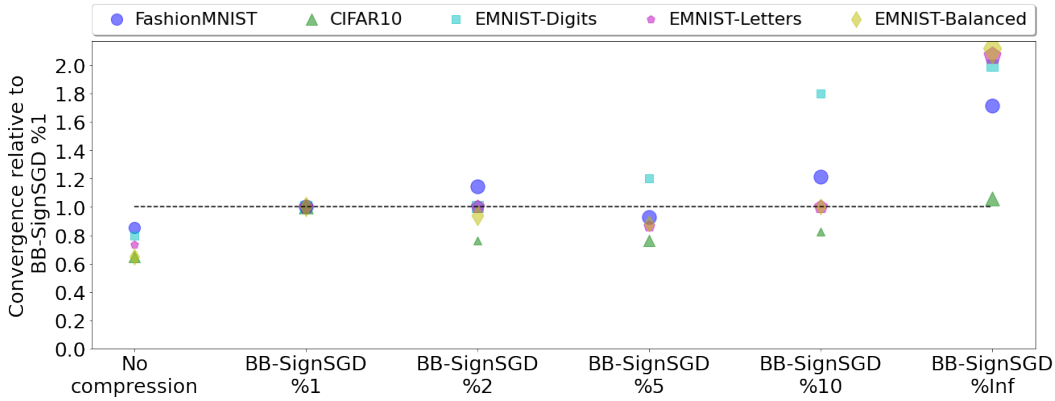


Figure 23: Converge (cycles) ratio relative to SIGNSGD (i.e. BB-SIGNSGD %1), measured as how fast the method achieved an accuracy at least 99% that of the maximum accuracy achieved by BB-SIGNSGD %1 (0.25 implies a method 4x faster). Size of marker indicates variance in experiments. Central learning rate $\eta_c = 2\text{e}{-3}$.