
Distributed and Private Stochastic EM Methods via Quantized and Compressed MCMC

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 To be completed

2 1 SAEM algorithm for Federated Learning

- 3 • SA of the expectations
- 4 • Monte Carlo approximations
- 5 • Exponential Family
- 6 • Statistics stored on devices and parameters computed on device
- 7 • FedAvg on central server — > boils down to the Averaging of parameters that we may
- 8 do in standard central settings (Polyak)
- 9 • Applications to PK-PD modeling

10 2 Notations and Algorithm

11 We minimize the negated log incomplete data likelihood

$$\min_{\theta \in \Theta} \bar{L}(\theta) := L(\theta) + r(\theta) \quad \text{with} \quad L(\theta) = \frac{1}{n} \sum_{i=1}^n L_i(\theta) := \frac{1}{n} \sum_{i=1}^n \{ -\log g(y_i; \theta) \} , \quad (1)$$

12 Consider a curved exponential family

$$f(z_i, y_i; \theta) = h(z_i, y_i) \exp(\langle S(z_i, y_i) | \phi(\theta) \rangle - \psi(\theta)) , \quad (2)$$

13 Then EM reads

$$\bar{s}_i(\theta) := \int_{\mathcal{Z}} S(z_i, y_i) p(z_i | y_i; \theta) \mu(dz_i) , \quad (3)$$

14 and the M -step is given by

$$\bar{\theta}(\bar{s}(\theta)) := \arg \min_{\vartheta \in \Theta} \{ R(\vartheta) + \psi(\vartheta) - \langle \bar{s}(\theta) | \phi(\vartheta) \rangle \} . \quad (4)$$

15 In the case where the expectations are intractable, then (3) becomes:

$$\tilde{S}^{(k+1)} := \frac{1}{n} \sum_{i=1}^n \tilde{S}_i^{(k+1)} = \frac{1}{n} \sum_{i=1}^n \frac{1}{M_k} \sum_{m=1}^{M_k} S(z_{i,m}^{(k)}, y_i) , \quad (5)$$

16 For computational purposes and privacy enhanced matter, I have chosen to study and develop the
 17 second algorithms that I proposed in my last week’s report. In that algorithm, one does not compute
 18 a periodic averaging of the local models (this would requires performing as many M-steps as there
 19 are workers). Rather, workers compute local statistics and send them to the central server for a
 20 periodic averaging of those vectors and the latter computes one M-step to update the global model.

Algorithm 1 FL-SAEM with Periodic Statistics Averaging

1: **Input:** **TO COMPLETE**
 2: Init: $\theta_0 \in \Theta \subseteq \mathbb{R}^d$, as the global model and $\bar{\theta}_0 = \frac{1}{n} \sum_{i=1}^n \theta_0$.
 3: **for** $r = 1$ to R **do**
 4: **for parallel for device** $i \in D^r$ **do**
 5: Set $\hat{\theta}_i^{(0,r)} = \hat{\theta}^{(r)}$.
 6: Draw M samples $z_{i,m}^{(r)}$ under model $\hat{\theta}_i^{(r)}$
 7: Compute the surrogate sufficient statistics $\tilde{S}_i^{(r+1)}$
 8: Workers send local statistics $\tilde{S}_i^{(k+1)}$ to server.
 9: **end for**
 10: Server computes **global model using the aggregated statistics:**

$$\hat{\theta}^{(r+1)} = \bar{\theta}(\tilde{S}^{(r+1)})$$

where $\tilde{S}^{(r+1)} = (\tilde{S}_i^{(r+1)}, i \in D_r)$ and send global model back to the devices.

11: **end for**

21 2.1 Challenges with Algorithm 1

22 While Algorithm 1 is a distributed variant of the SAEM, it is neither (a) private nor (b)
 23 communication-efficient. Indeed, we remark that broadcasting the vector of statistics are a po-
 24 tential breach to the data observations as their expression is related y and the latent data z . With a
 25 simple knowledge of the model used, the data could be retrieved if one extracts those statistics. Also
 26 regarding (b), the broadcast of n vector of statistics $S(y_i, z_i)$ can be cumbersome when the size of
 27 the latent space and the parameter space of the model are huge.

28 2.2 Algorithmic solutions

29 **Line 6 – Quantization:** The first step is to quantize the gradient in the Stochastic Langevin Dynam-
 30 ics step used in our sampling scheme Line 6 of Algorithm 1. Inspired by [Alistarh et al., 2017], we
 31 use an extension of the QSGD algorithm for our latent samples. Define the quantization operator as
 32 follows:

$$C_j^{(\ell)}(g, \xi_j) = \|v\| \cdot \text{sign}(g_j) \cdot (\lfloor \ell |g_j| / \|v\| \rfloor + \mathbf{1}_{\{\xi_j \leq \ell |g_j| / \|v\| - \lfloor \ell |g_j| / \|v\| \rfloor\}}) / \ell \quad (6)$$

33 where ℓ is the level of quantization and $j \in [d]$ denotes the dimension of the gradient.

34 Hence, for the sampling step, Line 6, we use the modified SGLD below, to be compliant with the
 35 privacy of our method.

Algorithm 2 Langevin Dynamics with Quantization for worker i

- 1: **Input:** Current local model $\hat{\theta}_i^{(r)}$ for worker $i \in \llbracket 1, n \rrbracket$.
- 2: Draw M samples $\{z_i^{(r,m)}\}_{m=1}^M$ from the posterior distribution $p(z_i|y_i; \hat{\theta}_i^{(k)})$ via Langevin diffusion with a quantized gradient:
- 3: **for** $k = 1$ to K **do**
- 4: Compute the quantized gradient of $\nabla \log p(z_i|y_i; \hat{\theta}_i^{(k)})$:

$$g_i(k, m) = C_j^{(\ell)} \left(\nabla_j f_{\theta_t}(z_i^{(k-1,m)}), \xi_j^{(k)} \right) \quad (7)$$

where $\xi_j^{(k)}$ is a realization of a uniform random variable.

- 5: Sample the latent data using the following chain:

$$z_i^{(k,m)} = z_i^{(k-1,m)} + \frac{\gamma_k}{2} g_i(k, m) + \sqrt{\gamma_k} B_k, \quad (8)$$

where B_t denotes the Brownian motion and $m \in [M]$ denotes the MC sample.

- 6: **end for**
 - 7: Assign $\{z_i^{(r,m)}\}_{m=1}^M \leftarrow \{z_i^{(K,m)}\}_{m=1}^M$.
 - 8: **Output:** latent data $z_{i,m}^{(k)}$ under model $\hat{\theta}_i^{(t,k)}$
-

36 **Line 7 – Compression MCMC output:** We use the notorious **Top- k** operator that we define as
 37 $\mathcal{C}(x)_i = x_i$, if $i \in \mathcal{S}$; $\mathcal{C}(x)_i = 0$ otherwise and where \mathcal{S} is defined as the size- k set of $i \in [p]$.
 38 Recall that after Line 6 we compute the local statistics $\tilde{S}_i^{(k+1)}$ using the output latent variables from
 39 Algorithm 2. We now use those statistics and compress them using Algorithm 3 as follows:

Algorithm 3 Sparsified Statistics with **Top- k**

- 1: **Input:** Current local statistics $\tilde{S}_i^{(k+1)}$ for worker $i \in \llbracket 1, n \rrbracket$. Sparsification level k .
- 2: Apply **Top- k** :

$$\ddot{S}_i^{(k+1)} = \mathcal{C} \left(\tilde{S}_i^{(k+1)} \right) \quad (9)$$

- 3: **Output:** Compressed local statistics for worker i denoted $\ddot{S}_i^{(k+1)}$.
-

40 **References**

- 41 D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic. Qsgd: Communication-efficient sgd
42 via gradient quantization and encoding. In *Advances in Neural Information Processing Systems*,
43 pages 1709–1720, 2017.