Expectation Maximization algorithm for Federated Learning

Gersende Fort
(IMT & CNRS)

Joint work with Aymeric Dieuleveut (CMAP, Ecole Polytechnique), Eric Moulines (CMAP, Ecole Polytechnique) and Geneviève Robin (LAMME, CNRS)

Publication: "Federated Expectation Maximization with heterogeneity mitigation and variance reduction" NeurIPS 2021

I. The Expectation-Maximization algorithm in the finite sum setting

Dempster, Laird, Rubin (1977)

EM, designed for ...

- Designed for solving

$$\mathrm{argmin}_{\theta \in \Theta} F(\theta) \qquad F(\theta) := -\frac{1}{N} \sum_{i=1}^{N} \log \int_{\mathsf{Z}} p_i(z_i; \theta) \mu(\mathsf{d} z_i)$$

  - $\Theta \subseteq \mathbb{R}^d$
  - F has **no closed form**
  - positive integrals ($p_i > 0$)

- Iterative algorithm: $\theta_{t+1} = \mathrm{EM\text{-}MAP}(\theta_t)$

- Limiting values: the fixed points of the $\mathrm{EM\text{-}MAP}$ operator.

An example ? (1/2) Statistical inference in **latent variable models**

- Independent observations $\mathbf{Y} := (Y_1, \ldots, Y_N)$
- Parametric statistical model, indexed by $\theta \in \Theta$
- Negative normalized Log-Likelihood (loss) function $F_{\mathbf{Y}}$

- Latent variable models

$$F_{\mathbf{Y}}(\theta) = -\frac{1}{N} \sum_{i=1}^{N} \log \int_{\mathsf{Z}} p(Y_i, z_i; \theta) \mu(\mathsf{d} z_i)$$

Of the form

$$\theta \mapsto -\frac{1}{N} \sum_{i=1}^{N} \log \int_{\mathsf{Z}} p_i(z_i; \theta) \mu(\mathsf{d} z_i)$$

An example ? (2/2) Mixture models

- The statistical task
  - i.i.d. observations with distribution $y \mapsto \sum_{g=1}^{G} \pi_g \, f_g(y; \vartheta)$
  - Learn the parameters $\theta := (\pi_{1:G}, \vartheta)$.
- Latent model: for each Y_i, a hidden label variable $Z_i \in \{1, \cdots, G\}$

$$\text{-LogLike}_{Y_i}(\theta) := \sum_{z=1}^{G} \underbrace{f_z(Y_i; \vartheta)}_{\text{Dist. } Y_i | Z_i = z} \overbrace{\pi_z}^{\text{Dist. } Z_i = z}$$

- The optimization problem:

$$\operatorname{argmin}_{\theta \in \Theta \subseteq \mathbb{R}^d} \; -\frac{1}{N} \sum_{i=1}^{N} \log \int \pi_z \; f_z(Y_i; \vartheta) \; \mathrm{d}\mu(z)$$

where μ is the counting measure on $\{1, \cdots, G\}$.

The EM algorithm

$\mathrm{argmin}_\theta F(\theta) \quad F(\theta) := -N^{-1} \sum_{i=1}^N \log \int_{\mathsf{Z}} p_i(z_i;\theta)\mu(\mathsf{d}z_i)$

At ieration $\#(t+1)$, given θ_t
define a *current* knowledge of $\mathbf{Z}$ through the *(a posteriori) distribution*

$$\pi_{\theta_t}(\mathsf{d}\mathbf{z}) := \prod_{i=1}^N \frac{p_i(z_i;\theta_t)}{\int p_i(u_i;\theta_t)\mu(\mathsf{d}u_i)} \mu(\mathsf{d}z_i)$$

- **E-step.** Compute the function

$$\theta \mapsto Q(\theta, \theta_t) := -\frac{1}{N} \sum_{i=1}^N \int \log p_i(z_i;\theta) \; \pi_{\theta_t}(\mathsf{d}\mathbf{z})$$

- **M-step.** Minimize this function: $\theta_{t+1} = \mathrm{argmin}_\theta \; Q(\theta, \theta_t)$

Rmk: EM is a Majorize-Minimization algorithm Lange (2016), and E-step $\equiv$ compute the majorizing function.

Implementation of EM

$\text{argmin}_\theta F(\theta) \quad F(\theta) := -N^{-1} \sum_{i=1}^N \log \int_Z p_i(z_i;\theta)\mu(\mathrm{d}z_i)$

- **Assumed (often, in the literature)** "exponential family" *(for the complete data model)*

$$p_i(z_i;\theta) = h_i(z_i) \ \exp\left(\langle S_i(z_i), \phi(\theta)\rangle - \psi(\theta)\right)$$

$$Q(\theta, \theta_t) = \psi(\theta) - \left\langle \frac{1}{N} \sum_{i=1}^N \bar{\mathsf{s}}_i(\theta_t), \phi(\theta) \right\rangle$$

where

$$\bar{\mathsf{s}}_i(\theta_t) := \int S_i(z_i) \ \pi_{\theta_t}(\mathbf{dz})$$

- **Assumed (here, in our works)** the argmin exists and is unique

$$\mathsf{T}(s) := \text{argmin}_\theta \ \psi(\theta) - \langle s, \phi(\theta)\rangle$$

Limiting points of EM: fixed points of the EM map

- in the θ-space E-M-E-M-...

$$\theta_{t+1} = \mathsf{T}\left(\frac{1}{N}\sum_{i=1}^{N}\bar{\mathsf{s}}_i(\theta_t)\right)$$

In the θ-space, the fixed points solve: $\mathsf{T}\left(\frac{1}{N}\sum_{i=1}^{N}\bar{\mathsf{s}}_i(\theta)\right) - \theta = 0$

- in the s-space M-E-M-E-...

$$s_t := \frac{1}{N}\sum_{i=1}^{N}\bar{\mathsf{s}}_i(\theta_t) \qquad s_{t+1} = \frac{1}{N}\sum_{i=1}^{n}\bar{\mathsf{s}}_i\left(\mathsf{T}(s_t)\right)$$

In the s-space, the fixed points solve: $\frac{1}{N}\sum_{i=1}^{N}\bar{\mathsf{s}}_i \circ \mathsf{T}(s) - s = 0$

II. From EM to "EM in Federated Learning"

EM as a root-finding algorithm

- The root-finding problem, *finite sum* setting

$$s \in \mathbb{R}^q \qquad \text{such that} \qquad \frac{1}{N}\sum_{i=1}^{N} \bar{\mathsf{s}}_i \circ \mathsf{T}(s) - s = 0$$

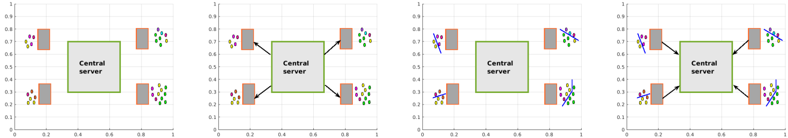
in the setting

- **sum over a large number** of functions $\longleftrightarrow$ sum over the examples
- each function $\bar{\mathsf{s}}_i \circ \mathsf{T}$ is (assumed) explicit

- Incremental EMs based on Stochastic Approximation algorithms Benveniste et al. (1990)

$$\widehat{S}_{t+1} = \widehat{S}_t + \gamma_{t+1}\,\mathsf{S}_{t+1} \qquad \mathsf{S}_{t+1} := \mathrm{approx}\left(\frac{1}{N}\sum_{i=1}^{N} \bar{\mathsf{s}}_i \circ \mathsf{T}(\widehat{S}_t) - \widehat{S}_t\right)$$

The Federated Learning setting (FL)



- The central server coordinates the participation of the local devices/clients/workers
- Local training data sets, **never** uploaded to the server
- FL reduces privacy and security risks
- Global model maintained by the central server: sent to the devices
- Each worker computes an update of the global model
- Only this update is communicated to the central server; aggregation by the central server

- Local data sets, heterogeneous, unbalanced
- Partial participation of the clients (charged devices, plugged-in, free wi-fi connection, · · ·)
- Massively distributed: large nbr devices w.r.t. the size of the local data sets

Communication cost >> Computational cost

Batch EM is not adapted for Federated Learning

$$s \in \mathbb{R}^q \qquad \text{such that} \qquad \frac{1}{N}\sum_{i=1}^{N} \bar{\mathsf{s}}_i \circ \mathsf{T}(s) - s = 0$$

- The optimization step T: run by the central server.
- The expectation step **can not be** run by the central server:

$$\frac{1}{N}\sum_{i=1}^{N} \bar{\mathsf{s}}_i \circ \mathsf{T}(s) = \frac{1}{n}\sum_{c=1}^{n} \frac{1}{m_c}\sum_{i=1}^{m_c} \bar{\mathsf{s}}_{ci} \circ \mathsf{T}(s)$$

Naive idea

Naive algorithm

- Design parameters: $k_{\max}$ $\gamma > 0$.
- Initialization: $\widehat{S}_0$
- For $k = 0, \ldots, k_{\max} - 1$:
  - *(active workers)* For each worker $\#c$ do
    - · Sample $\mathsf{S}_{k+1,c}$ an approximation of $m_c^{-1} \sum_{i=1}^{m_c} \bar{\mathsf{s}}_{ci} \circ \mathsf{T}(\widehat{S}_k)$
    - · Send $\Delta_{k+1,c} := \mathsf{S}_{k+1,c} - \widehat{S}_k$ to the central server
  - *(central server)*
    - · Update: $\widehat{S}_{k+1} = \widehat{S}_k + \gamma \frac{1}{n} \sum_{c=1}^{n} \Delta_{k+1,c}$
    - · Send $\widehat{S}_{k+1}$ and $\mathsf{T}(\widehat{S}_{k+1})$ to the n workers.
- **Return**: $\widehat{S}_k$, $0 \leq k \leq k_{\max}$

Communication cost
Partial participation of the local agents.

III. FedEM - Federated EM
and
VR-FedEM - Variance Reduced FedEM

FedEM

roots of $\mathsf{h}(s) := n^{-1} \sum_{c=1}^{n} \left\{ m_c^{-1} \sum_{i=1}^{m_c} \bar{\mathsf{s}}_{ci} \circ \mathsf{T}(s) - s \right\}$.

FedEM with partial participation $p \in (0,1)$

- Design parameters: $k_{\max}$, $\alpha > 0$, $\gamma > 0$.
- Initialization: $V_{0,c}, \widehat{S}_0$; $V_0 := n^{-1} \sum_{c=1}^{n} V_{0,c}$
- For $k = 0, \ldots, k_{\max} - 1$:
  - "Sample" workers $\mathcal{A}_{k+1}$ "with participation probability p"
  - *(active local workers)* For $c \in \mathcal{A}_{k+1}$ do
    - · Sample $\mathsf{S}_{k+1,c}$ an approximation of $m_c^{-1} \sum_{i=1}^{m_c} \bar{\mathsf{s}}_{ci} \circ \mathsf{T}(\widehat{S}_k)$
    - · Set $\Delta_{k+1,c} = \mathsf{S}_{k+1,c} - \widehat{S}_k - V_{k,c}$
    - · Set $V_{k+1,c} = V_{k,c} + \alpha \, \mathrm{Quant}(\Delta_{k+1,c})$
    - · Send $\mathrm{Quant}(\Delta_{k+1,c})$ to the central server
  - *(inactive local workers)* For $c \notin \mathcal{A}_{k+1}$, set $V_{k+1,c} = V_{k,c}$
  - *(central server)*
    - · Set $\widehat{S}_{k+1} = \widehat{S}_k + \frac{\gamma}{np} \sum_{c \in \mathcal{A}_{k+1}} \mathrm{Quant}(\Delta_{k+1,c}) + \gamma V_k$
    - · Set $V_{k+1} = V_k + \alpha n^{-1} \sum_{c=1}^{n} \mathrm{Quant}(\Delta_{k+1,c})$.
    - · Send $\widehat{S}_{k+1}$ and $\mathsf{T}(\widehat{S}_{k+1})$ to the n workers.
- **Return**: $\widehat{S}_k$, $0 \le k \le k_{\max}$

- Possible partial participation of the workers
- Federated E-step
- Random quantization w. variance reduction

(Mishchenko et al, 2019)

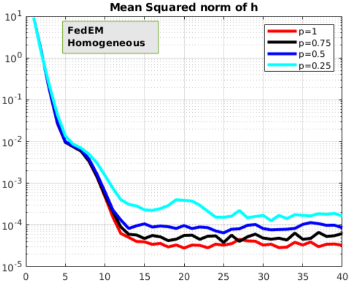
- M-step **only** at the central server

Robustness

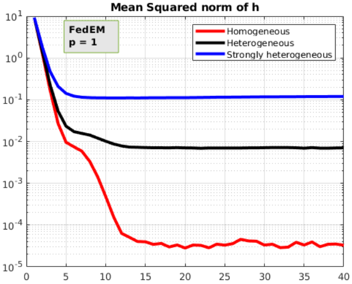
FedEM is designed to find the roots of h

Toy example: inference of a $\mathbb{R}^2$-valued Gaussian mixture model with 2 components

- **Robustness to partial participation**
  $k \mapsto \mathbb{E}\left[\|h(\widehat{S}_k)\|^2\right]$ vs the nbr of epochs. Estimated by Monte Carlo



- **Robustness to heterogeneity**
  $k \mapsto \mathbb{E}\left[\|h(\widehat{S}_k)\|^2\right]$ vs the nbr of epochs. Estimated by Monte Carlo



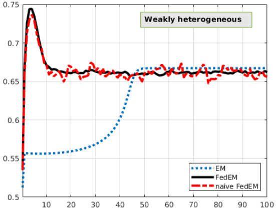
Robustness

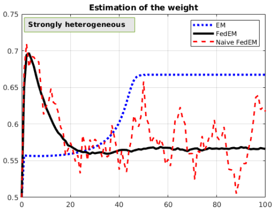
FedEM is designed to find the roots of h

Toy example: inference of a $\mathbb{R}^2$-valued Gaussian mixture model with 2 components

- FedEM vs naive-FedEM ? Estimation of the weight vs the nbr epoch; Case "homogeneous" and case "strongly heterogeneous"

In `naive-FedEM`: remove the variables $V_{\cdot c}$'s – i.e. the control variates introduced to control the variance of the quantization step.





How to reduce the variance of the local approximations ?

- in FedEM

$$\mathsf{S}_{k+1,c} = \text{approx}\left(m^{-1}\sum_{i=1}^{m} \bar{\mathsf{s}}_{ci} \circ \mathsf{T}(\widehat{S}_k)\right).$$

Consider the case "mini batch"

$$\mathsf{S}_{k+1,c} = \frac{1}{\mathsf{b}} \sum_{i \in \mathcal{B}_{k+1,c}} \bar{\mathsf{s}}_{ci} \circ \mathsf{T}(\widehat{S}_k)$$

- the SPIDER variance reduction technique

$$\begin{aligned}\mathsf{S}_{k+1,c} &= \frac{1}{\mathsf{b}} \sum_{i \in \mathcal{B}_{k+1,c}} \bar{\mathsf{s}}_{ci} \circ \mathsf{T}(\widehat{S}_k) + \mathsf{S}_{k,c} - \mathsf{b}^{-1} \sum_{i \in \mathcal{B}_{k+1,c}} \bar{\mathsf{s}}_{ci} \circ \mathsf{T}(\widehat{S}_{k-1}) \\ &\approx \frac{1}{\mathsf{b}} \sum_{i \in \mathcal{B}_{k+1,c}} \bar{\mathsf{s}}_{ci} \circ \mathsf{T}(\widehat{S}_k) + \widehat{0}\end{aligned}$$

add a **control variate**

VR-FedEM

in the case $\bar{\mathsf{s}}_c(\tau) = m^{-1} \sum_{i=1}^m \bar{\mathsf{s}}_{ci}(\tau)$

Iteration index (cycles of length k_{in})

$$k+1 \leftarrow (t-1)k_{\text{in}} + \tau \qquad t \geq 1, \tau \in \{1, \cdots, k_{\text{in}}\}.$$

Variance Reduction on $\mathsf{S}_{k+1,c}$ (case $p=1$)

- Initialization: $\mathsf{S}_{1,0,c} := m^{-1} \sum_{i=1}^m \bar{\mathsf{s}}_{ci} \circ \mathsf{T}(\widehat{S}_{\text{init}})$ and $\widehat{S}_{1,0} = \widehat{S}_{1,-1} := \widehat{S}_{\text{init}}$
- At time $\#(t-1)k_{\text{in}} + \tau$, at each local server $\#c$
  - Sample a mini-batch $\mathcal{B}_{t,\tau,c}$ of size b in $\{1, \cdots, m\}$
  - Approximate $\bar{\mathsf{s}}_c \circ \mathsf{T}(\widehat{S}_{t,\tau-1})$ with

$$\mathsf{S}_{t,\tau,c} := \mathsf{b}^{-1} \sum_{i \in \mathcal{B}_{t,\tau,c}} \bar{\mathsf{s}}_{ci} \circ \mathsf{T}(\widehat{S}_{t,\tau-1}) + \mathsf{S}_{t,\tau-1,c} - \mathsf{b}^{-1} \sum_{i \in \mathcal{B}_{t,\tau,c}} \bar{\mathsf{s}}_{ci} \circ \mathsf{T}(\widehat{S}_{t,\tau-2})$$

- At time $\#tk_{\text{in}}$, refresh the *control variate*
  - (central server) $\widehat{S}_{t,0} = \widehat{S}_{t,-1} := \widehat{S}_{t-1,k_{\text{in}}}$
  - (local workers) $\mathsf{S}_{t,0,c} := m^{-1} \sum_{i=1}^m \bar{\mathsf{s}}_{ci} \circ \mathsf{T}(\widehat{S}_{t,0})$

- A **control variate** scheme reduces the variability of the approximations of $\bar{\mathsf{s}}_c \circ \mathsf{T}(\widehat{S}.)$
- The control variate is biased: it is **refreshed every k_{in} iterations**.

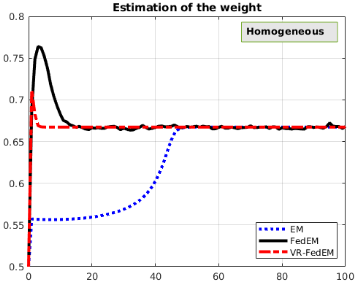
Same variance reduction as in `SPIDER-EM`, Fort et al. (2020) – SPIDER = Stochastic Path-Integrated Differential EstimatoR.

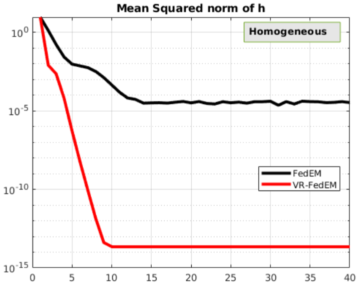
VR-FedEM

FedEM is designed to find the roots of h

Toy example: inference of a $\mathbb{R}^2$-valued Gaussian mixture model with 2 components

- Estimation of the weight vs the nbr epoch
- $k \mapsto \mathbb{E}\left[\|h(\widehat{S}_k)\|^2\right]$ vs the nbr of epochs. Estimated by Monte Carlo





IV. Explicit control of convergence
Complexity analysis

Assumptions

$$\mathrm{argmin}_{\theta \in \Theta \subseteq \mathbb{R}^d} \; F(\theta) \Longrightarrow \mathrm{argmin}_{s \in \mathbb{R}^q} \; F \circ \mathsf{T}(s) \Longrightarrow s : \mathsf{h}(s) = 0$$

$$F(\theta) := \frac{1}{n} \sum_{c=1}^{n} \frac{1}{m_c} \sum_{i=1}^{m_c} \mathcal{L}_{ci}(\theta), \qquad \mathcal{L}_{ci}(\theta) = -\log \int \exp(-\psi(\theta) + \langle S_{ci}(z), \phi(\theta) \rangle) \, \mathsf{d}\mu(z)$$

- On the model
- For the existence of a Lyapunov function
- On the local workers / local data sets
- On the quantization step
- On the participation of the workers

Assumptions

$$\mathrm{argmin}_{\theta \in \Theta \subseteq \mathbb{R}^d} \; F(\theta) \Longrightarrow \mathrm{argmin}_{s \in \mathbb{R}^q} \; F \circ \mathsf{T}(s) \Longrightarrow s : \mathsf{h}(s) = 0$$

$$F(\theta) := \frac{1}{n} \sum_{c=1}^{n} \frac{1}{m_c} \sum_{i=1}^{m_c} \mathcal{L}_{ci}(\theta), \qquad \mathcal{L}_{ci}(\theta) = -\log \int \exp(-\psi(\theta) + \langle S_{ci}(z), \phi(\theta) \rangle) \, \mathrm{d}\mu(z)$$

- On the model
  - A1 $\Theta \subset \mathbb{R}^d$ is open convex. Finite loss $\mathcal{L}_{ci}$.
  - A2 The conditional expectations $\bar{\mathsf{s}}_{ci}(\theta)$ are well defined $\forall c, i$ and $\theta \in \Theta$.
  - A3 The map T: $s \mapsto \mathrm{argmin}_{\theta \in \Theta} \psi(\theta) - \langle s, \phi(\theta) \rangle$ exists and is unique.
- For the existence of a Lyapunov function
- On the local workers / local data sets
- On the quantization step
- On the participation of the workers

Assumptions

$$\mathrm{argmin}_{\theta \in \Theta \subseteq \mathbb{R}^d} \; F(\theta) \Longrightarrow \mathrm{argmin}_{s \in \mathbb{R}^q} \; F \circ \mathsf{T}(s) \Longrightarrow s : \mathsf{h}(s) = 0$$

$$F(\theta) := \frac{1}{n} \sum_{c=1}^{n} \frac{1}{m_c} \sum_{i=1}^{m_c} \mathcal{L}_{ci}(\theta), \qquad \mathcal{L}_{ci}(\theta) = -\log \int \exp(-\psi(\theta) + \langle S_{ci}(z), \phi(\theta) \rangle) \, \mathsf{d}\mu(z)$$

- On the model
- For the existence of a Lyapunov function
  - A4 $W := F \circ \mathsf{T}$ is C^1, with globally Lipschitz gradient (constant $L_{\dot{W}}$). Furthermore, $\nabla W(s) = -B(s)\mathsf{h}(s)$ for a positive definite matrix $B(s)$ with spectrum in $[v_{\min}, v_{\max}]$ for any s, and $v_{\min} > 0$.
- On the local workers / local data sets
- On the quantization step
- On the participation of the workers

Assumptions

$\operatorname{argmin}_{\theta \in \Theta \subseteq \mathbb{R}^d} F(\theta) \Longrightarrow \operatorname{argmin}_{s \in \mathbb{R}^q} F \circ \mathsf{T}(s) \Longrightarrow s : \mathsf{h}(s) = 0$

$$F(\theta) := \frac{1}{n} \sum_{c=1}^{n} \frac{1}{m_c} \sum_{i=1}^{m_c} \mathcal{L}_{ci}(\theta), \qquad \mathcal{L}_{ci}(\theta) = -\log \int \exp(-\psi(\theta) + \langle S_{ci}(z), \phi(\theta) \rangle) \, \mathsf{d}\mu(z)$$

- On the model
- For the existence of a Lyapunov function
- On the local workers / local data sets
  - A5 There exists L_c such that for any s, s', $\|\bar{\mathsf{s}}_{c\cdot} \circ \mathsf{T}(s) - s - \bar{\mathsf{s}}_{c\cdot} \circ \mathsf{T}(s') - s'\| \leq L_c \|s - s'\|$.
  - A7 For any k, the local approximations $\mathsf{S}_{k,c}$ are independent, unbiased $\mathbb{E}\left[\mathsf{S}_{k+1,c} | \mathcal{F}_k\right] = \bar{\mathsf{s}}_c \circ \mathsf{T}(\widehat{S}_k)$ and heteregeneous variance: $\mathbb{E}\left[\|\mathsf{S}_{k+1,c} - \bar{\mathsf{s}}_c \circ \mathsf{T}(\widehat{S}_k)\|^2 | \mathcal{F}_k\right] \leq \sigma_c^2$.
- On the quantization step
- On the participation of the workers

Assumptions

$$\mathrm{argmin}_{\theta \in \Theta \subseteq \mathbb{R}^d} \ F(\theta) \Longrightarrow \mathrm{argmin}_{s \in \mathbb{R}^q} \ F \circ \mathsf{T}(s) \Longrightarrow s : \mathsf{h}(s) = 0$$

$$F(\theta) := \frac{1}{n} \sum_{c=1}^{n} \frac{1}{m_c} \sum_{i=1}^{m_c} \mathcal{L}_{ci}(\theta), \qquad \mathcal{L}_{ci}(\theta) = -\log \int \exp(-\psi(\theta) + \langle S_{ci}(z), \phi(\theta) \rangle) \, \mathsf{d}\mu(z)$$

- On the model
- For the existence of a Lyapunov function
- On the local workers / local data sets
- On the quantization step

  A6 Unbiased quantization operator $\mathbb{E}[\mathrm{Quant}(x)] = x$.
  There exists $\omega > 0$ s.t. $\mathbb{E}[\|\mathrm{Quant}(x)\|^2] \leq (1+\omega)\|x\|^2$.

  e.g. random dithering; see also Aslistarh et al. (2018); Horvath et al. (2019); Mishchenko et al. (2019)
- On the participation of the workers

Assumptions

$$\mathrm{argmin}_{\theta \in \Theta \subseteq \mathbb{R}^d} \; F(\theta) \Longrightarrow \mathrm{argmin}_{s \in \mathbb{R}^q} \; F \circ \mathsf{T}(s) \Longrightarrow s : \mathsf{h}(s) = 0$$

$$F(\theta) := \frac{1}{n} \sum_{c=1}^{n} \frac{1}{m_c} \sum_{i=1}^{m_c} \mathcal{L}_{ci}(\theta), \qquad \mathcal{L}_{ci}(\theta) = -\log \int \exp(-\psi(\theta) + \langle S_{ci}(z), \phi(\theta) \rangle) \, \mathsf{d}\mu(z)$$

- On the model
- For the existence of a Lyapunov function
- On the local workers / local data sets
- On the quantization step
- On the participation of the workers

  A8 I.i.d. Bernoulli r.v. with participation probability p.

Explicit control for FedEM

Set

$$L^2 := n^{-1} \sum_{i=1}^{n} L_i^2, \qquad \sigma^2 := n^{-1} \sum_{i=1}^{n} \sigma_i^2;$$

Theorem Dieuleveut, F., Moulines, Robin (2021)

Let $\{\widehat{S}_k, k \geq 1\}$ be given by `FedEM`, run with $V_{c0} := \bar{\mathsf{s}}_{c\cdot} \circ \mathsf{T}(\widehat{S}_0) - \widehat{S}_0$, $\alpha := (1+\omega)^{-1}$ and $\gamma_k = \gamma \in (0, \gamma_{\max}]$, where

$$\gamma_{\max} := \frac{v_{\min}}{2L_{\dot{W}}} \wedge \frac{p\sqrt{n}}{2\sqrt{2}L(1+\omega)\sqrt{\omega + (1-p)(1+\omega)/p}}.$$

Denote by K the uniform random variable on $\{0, \cdots, k_{\max}-1\}$. Then,

$$v_{\min}\left(1 - \gamma\frac{L_{\dot{W}}}{v_{\min}}\right)\mathbb{E}\left[\|\mathsf{h}(\widehat{S}_K)\|^2\right] \leq \frac{\left(W(\widehat{S}_0) - \min W\right)}{\gamma k_{\max}} + \gamma L_{\dot{W}} \frac{1 + 5\left(\omega + (1-p)(1+\omega)/p\right)}{n}\sigma^2.$$

Complexity analysis (when $p = 1$)

Given an accuracy level ϵ, how to choose the design parameters in order to minimize the number of optimization ?

- Results valid when heterogeneous data sets
- The number of optimization is $k_{\max}$ chosen in order to reach the accuracy level ϵ:

$$\mathcal{K}_{\text{opt}}(\epsilon) = O\left(\frac{1}{\epsilon^2}\frac{(1+\omega)\sigma^2}{n}\right) \vee O\left(\frac{1}{\epsilon\,\gamma_{\max}}\right)$$

1st term is leader iff $\epsilon << \gamma_{\max}(1+\omega)\sigma^2/n$ (high noise regime)

- **Compression effect:** γ is impacted by compression iff $n << \omega^3$.
On $\mathcal{K}_{\text{opt}}$:

	Complexity regime:	$\frac{(1+\omega)\sigma^2}{n\epsilon^2}$	$\frac{1}{\gamma_{\max}\epsilon}$
$\gamma_{\max}$ **regime:**	**E.g. case when**	High noise σ^2, small ϵ	Low σ^2 larger ϵ
$\frac{v_{\min}}{2L_{\dot{W}}}$	large ratio n/ω^3	$\times\omega$	$\times 1$
$\frac{\sqrt{n}}{2\sqrt{2}L(1+\omega)\sqrt{\omega}}$	low ratio n/ω^3	$\times\omega$	$\times\omega^{3/2}/\sqrt{n}$

Explicit control for VR-FedEM

Set $(m_c = m)$

$$L^2 := n^{-1} m^{-1} \sum_{c=1}^{n} \sum_{i=1}^{m} L_{ci}^2$$

Theorem Dieuleveut, F., Moulines, Robin (2021)

Let $\{\widehat{S}_{t,k}, t \geq 1, 1 \leq k \leq k_{\mathrm{in}}\}$ be given by `VR-FedEM` run with $\alpha := 1/(1+\omega)$, $V_{1,0,c} := \bar{\mathsf{s}}_c \circ \mathsf{T}(\widehat{S}_{1,0}) - \widehat{S}_{1,0}$, $\mathsf{b} := \lceil \frac{k_{\mathrm{in}}}{(1+\omega)^2} \rceil$ and

$$\gamma_{t,k} = \gamma := \frac{v_{\min}}{L_{\dot{W}}} \left(1 + 4\sqrt{2} \frac{v_{\max}}{L_{\dot{W}}} \frac{L}{\sqrt{n}} (1+\omega) \left(\omega + \frac{1+10\omega}{8} \right)^{1/2} \right)^{-1}.$$

Let (τ, K) be the uniform random variable on $\{1, \cdots, k_{\mathrm{out}}\} \times \{1, \cdots, k_{\mathrm{in}}\}$, independent of $\{\widehat{S}_{t,k}, t \geq 1, k \in \{1, \cdots, k_{\mathrm{in}}\}\}$. Then, it holds

$$\mathbb{E}\left[\|H_{\tau,K}\|^2 \right] \leq \frac{2\left(\mathbb{E}\left[W(\widehat{S}_{1,0}) \right] - \min W \right)}{v_{\min} \gamma k_{\mathrm{in}} k_{\mathrm{out}}}$$

$$\mathbb{E}\left[\|\mathsf{h}(\widehat{S}_{\tau,K-1})\|^2 \right] \leq 2\left(1 + \gamma^2 \frac{L^2 (1+\omega)^2}{n} \right) \mathbb{E}\left[\|H_{\tau,K}\|^2 \right].$$

Complexity analysis

- First result on Federated EM including variance reduction techniques, being robust to distribution heterogeneity.
- The recommended batch size b decreases as $1/(1+\omega)^2$.
- The number of optimization is $k_{\text{out}}\, k_{\text{in}}$ chosen in order to reach the accuracy level ϵ:

$$\mathcal{K}_{\text{opt}}(\epsilon) = \left(\frac{1}{\epsilon\,\gamma}\right)$$

- **Compression effect on $\mathcal{K}_{\text{opt}}$**

	Complexity:	$1/(\gamma\epsilon)$
γ **regime:**	**e.g. case when**	
$v_{\min}/L_{\dot{W}}$	large ratio n/ω^3	$\times 1$
$v_{\min}\sqrt{n}/(v_{\max}L\omega^{3/2})$	low ratio n/ω^3	$\times \omega^{3/2}/\sqrt{n}$

Contributions

The Expectation Maximization (EM) algorithm *with complete data model in the curved exponential family* is a root-finding algorithm Delyon et al. (1999).

- Emphasis on EM in Federated Learning.
- A new algorithm: `FedEM` supporting communication compression, partial participation and data heterogeneity.
- A variance reduced version `VR-FedEM`, progressively alleviating the variance brought by the random oracles on which updates of the local workers are based.
- Convergence guarantees of `FedEM` and `VR-FedEM`.
- Pioneering work in the litterature "EM in Federated Learning". contemporaneous works with different goals: Marfoq et al. (2021), Louizos et al. (2021)
  As a root finding algorithm, `VR-FedEM` state of the art (compared to `VR-DIANA` Horvath et al. (2019)).

V. Bibliography

Bibliography

D. Alistarh, T. Hoefler, M. Johansson, N. Konstantinov, S. Khirirat, and C. Renggli
The convergence of sparsified gradient methods.
In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 31, pages 5973–5983. Curran Associates, Inc., 2018.

A. Benveniste, M. Métivier, and P. Priouret.
Adaptive Algorithms and Stochastic Approximations.
Springer Verlag, 1990.

A. Dempster, N. Laird, and D. Rubin.
Maximum Likelihood from Incomplete Data via the EM Algorithm.
J. Roy. Stat. Soc. B Met., 39(1):1–38, 1977.

G. Fort, E. Moulines, and H.-T. Wai.
A Stochastic Path Integral Differential EstimatoR Expectation Maximization Algorithm.
In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, Advances in Neural Information Processing Systems, volume 33, pages 16972–16982. Curran Associates, Inc., 2020.

S. Horváth, D. Kovalev, K. Mishchenko, S. Stich, and P. Richtárik.
Stochastic distributed learning with gradient quantization and variance reduction.
arXiv preprint arXiv:1904.05115, 2019.

K. Lange.
MM Optimization Algorithms.
SIAM-Society for Industrial and Applied Mathematics, 2016.

K. Mishchenko, E. Gorbunov, M. Takáč, and P. Richtárik.
Distributed learning with compressed gradient differences.
arXiv preprint arXiv:1901.09269, 2019.