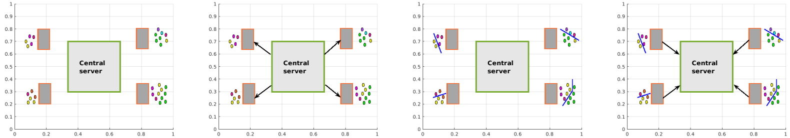
# Federated Expectation Maximization with heterogeneity mitigation and variance reduction

Gersende Fort
(IMT & CNRS, France)

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

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

# 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, $\cdots$)
- Massively distributed: large nbr devices w.r.t. the size of the local data sets

Communication cost >> Computational cost

# In this talk

- Design a novel algorithm for the optimization problem:

$$\text{find } s_\star \in \mathbb{R}^q \text{ s.t.} \qquad \mathsf{h}(s_\star) = 0$$

resulting from: finding the fixed points $G(s) = s$ of an iterative algorithm $S_{n+1} = G(S_n)$

- in the **Federated Learning** setting,

$$\mathsf{h}(s) = \frac{1}{n}\sum_{c=1}^{n} \mathsf{h}_c(s) \qquad\qquad \text{Ex. } \mathsf{h}_c(s) = G_c \circ \mathsf{T}(s) - s$$

- part of $\mathsf{h}_c$ is known by the local worker $\#c$ and depends on **local** data
- and the other part is known by the central server.

- tool: Stochastic optimization combining
  - the Stochastic Approximation method Robbins and Monro (1951); Benveniste et al. (1990)

$$\widehat{S}_{n+1} = \widehat{S}_n + \gamma_{n+1} H_{n+1} \qquad H_{n+1} \approx \mathsf{h}(\widehat{S}_n)$$

  - Variance reduction techniques

## 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)).

# I. Majorize-Minimization in the Federated Learning setting

# The learning task

Given $n$ local workers, with local data sets of size $m_c$

$$\mathrm{argmin}_{\theta \in \Theta \subseteq \mathbb{R}^d} \; \frac{1}{n} \sum_{c=1}^{n} \underbrace{\mathcal{L}_{c\cdot}(\theta)}_{\text{Loss function, at worker } \#c}$$

when

$$\mathcal{L}_{c\cdot}(\theta) = \frac{1}{m_c} \sum_{i=1}^{m_c} \log \int p_{ci}(z;\theta) \, \mathsf{d}\mu(z) \qquad p_{ci}(z;\theta) > 0.$$

Applications e.g.

- Inference in latent variable models
- Inference in hierarchical models

Optimization tool

- Majorize-Minimization approach

# First example: inference in mixture models

- The statistical task
  - i.i.d. observations $Y_{ci}$'s, with distribution $y \mapsto \sum_{g=1}^{G} \pi_g \, f_g(y; \vartheta)$
  - Learn the parameters $\theta := (\pi_{1:G}, \vartheta)$.
  - Loss function: the negative log-likelihood

- The computational problem

$$\mathrm{argmin}_{\theta \in \Theta \subseteq \mathbb{R}^d} \quad - \frac{1}{n} \sum_{c=1}^{n} \frac{1}{m_c} \sum_{i=1}^{m_c} \log \sum_{g=1}^{G} \pi_g \, f_g(Y_{ci}; \vartheta)$$

or equivalently

$$\mathrm{argmin}_{\theta \in \Theta \subseteq \mathbb{R}^d} \quad - \frac{1}{n} \sum_{c=1}^{n} \frac{1}{m_c} \sum_{i=1}^{m_c} \log \int \overbrace{\pi_z}^{\text{Dist. } Z_{ci} = z} \underbrace{f_z(Y_{ci}; \vartheta)}_{\text{Dist. } Y_{ci} | Z_{ci} = z} \, \mathsf{d}\mu(z)$$

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

# Second example: inference in hierarchical models

- The statistical task
  - indep observations $Y_{ci}$'s, with distribution given a *local parameter* $y \mapsto f(y; \vartheta, z_{ci})$
  - Prior on the i.i.d. $Z_{ci}$'s: $z \mapsto p(z; \tau)\, \mu(\mathsf{d}z)$
  - Learn the parameters $\theta := (\vartheta, \tau)$.
  - Loss function: the negative log-likelihood

- The computational problem

$$\mathrm{argmin}_{\theta \in \Theta \subseteq \mathbb{R}^d} \; -\frac{1}{n} \sum_{c=1}^{n} \frac{1}{m_c} \sum_{i=1}^{m_c} \log \int f(Y_{ci}; \vartheta, z)\, p(z; \tau) \mu(\mathsf{d}z)$$

# General example: latent variable models

- The statistical task
  - independent observations $Y_{ci}$'s with density

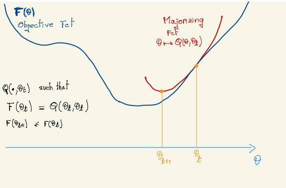
$$y \mapsto \int p_{ci}(y, z; \theta)\mu(\mathsf{d}z)$$

  - $Z$: latent variable. $(Y, Z)$ complete data.
  - Learn the parameters $\theta$.
  - Loss function: the negative log-likelihood

- The computational problem

$$\text{argmin}_{\theta \in \Theta \subseteq \mathbb{R}^d} \; -\frac{1}{n}\sum_{c=1}^{n}\frac{1}{m_c}\sum_{i=1}^{m_c}\log\int \underbrace{p_{ci}(Y_{ci}, z; \theta)}_{\text{complete data model}} \; \mu(\mathsf{d}z)$$

# Optimization tool: Majorize-Minimization algorithm Lange (2016)




- At ieration $\#(t+1)$, given $\theta_t$, define a majorizing function

$$F(\theta) \leq Q(\theta, \theta_t) := -\frac{1}{n}\sum_{c=1}^{n}\frac{1}{m_c}\sum_{i=1}^{m_c}\mathbb{E}_{\theta_t}\left[\log p_{ci}(Z;\theta)\right] + C(\theta_t)$$

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

This is the Expectation-Maximization algorithm

Dempster et al. (1977)

Upon noting that for any distribution $g(z)\,\mu(\mathrm{d}z)$

$$\log \int f(z)\mu(\mathrm{d}z) = \log \int \frac{f(z)}{g(z)}\, g(z)\mu(\mathrm{d}z) \geq \int \log\left(\frac{f(z)}{g(z)}\right)\; g(z)\mu(\mathrm{d}z)$$

it holds for any $\theta_t$

$$\log \int p_{ci}(z;\theta)\,\mathrm{d}\mu(z) \geq \int \log p_{ci}(z;\theta)\frac{p_{ci}(z;\theta_t)\mu(\mathrm{d}z)}{\int p_{ci}(u;\theta_t)\mu(\mathrm{d}u)} + C_{ci}(\theta_t)$$

$$\geq \mathbb{E}_{\theta_t}\left[\log p_{ci}(Z;\theta)\right] + C_{ci}(\theta_t)$$

with equality at $\theta = \theta_t$.

# Implementation of the Majorize-Minimization algorithm EM

- **Assumed** "exponential family" (for the complete data model)

$$\log p_{ci}(z;\theta) = \langle S_{ci}(z), \phi(\theta)\rangle - \psi(\theta)$$

and the argmax exists and is unique

$$Q(\theta,\theta_t) = \psi(\theta) - \langle \bar{\mathsf{s}}(\theta_t), \phi(\theta)\rangle \qquad \mathsf{T}(s) \overset{\text{def}}{=} \operatorname{argmin}_\theta \ \psi(\theta) - \langle s, \phi(\theta)\rangle$$

- **E-step.** Explicit computation of the majorizing function $\theta \mapsto Q(\theta,\theta_t)$ i.e. of $\bar{\mathsf{s}}(\theta_t)$

$$\bar{\mathsf{s}}(\theta_t) \overset{\text{def}}{=} \frac{1}{n}\sum_{c=1}^{n} \frac{1}{m_c}\sum_{i=1}^{m_c} \int S_{ci}(z) \ \frac{p_{ci}(z;\theta_t)\mu(\mathsf{d}z)}{\int p_{ci}(u;\theta_t)\mu(\mathsf{d}u)}$$

- **M-step.** Explicit computation of the minimum i.e. $\theta_{t+1} = \mathsf{T}(\bar{\mathsf{s}}(\theta_t))$.

**In the $s$-space, the fixed points solve:** $\bar{\mathsf{s}} \circ \mathsf{T}(s) - s = 0$

# Conclusion part I

Design an algorithm

- will find a root of

$$s \in \mathbb{R}^q : \qquad \frac{1}{n}\sum_{c=1}^{n}\frac{1}{m_c}\sum_{i=1}^{m_c}\bar{\mathsf{s}}_{ci}\circ \mathsf{T}(s) - s = 0$$

- under the constraints
  1. the data set $\#c$ is only available at the worker $\#c$

$$s \in \mathbb{R}^q : \qquad \frac{1}{n}\sum_{c=1}^{n}\bar{\mathsf{s}}_{c\cdot}\circ \mathsf{T}(s) - s = 0 \qquad \bar{\mathsf{s}}_{c\cdot}(\tau) \stackrel{\text{def}}{=} \frac{1}{m_c}\sum_{i=1}^{m_c}\bar{\mathsf{s}}_{ci}(\tau)$$

  2. the maximization step $\mathsf{T}(s)$ is performed by the central server
- few communications between workers and the central server
- robust to heterogeneous and unbalanced data sets
- allowing partial participation of the workers

II. `FedEM` - Federated EM
and
`VR-FedEM` - Variance Reduced FedEM

# FedEM

roots of $\mathsf{h}(s) \stackrel{\text{def}}{=} n^{-1} \sum_{c=1}^{n} \bar{\mathsf{s}}_c \circ \mathsf{T}(s) - s$.

## 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 $\bar{\mathsf{s}}_{c.} \circ \mathsf{T}(\widehat{S}_k)$
    - 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

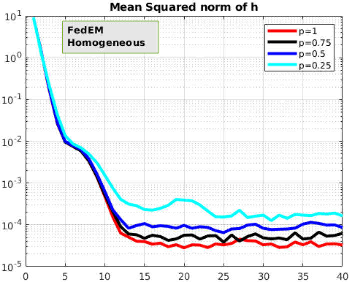
# Robustness

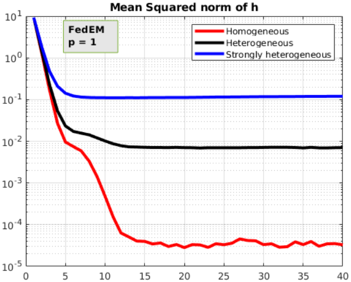
FedEM is designed to find the roots of $h$

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



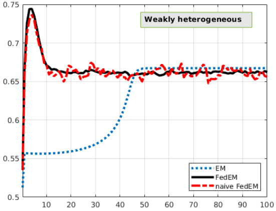
# Robustness

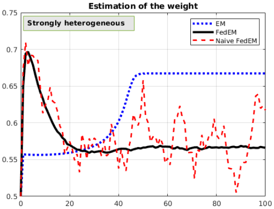
FedEM is designed to find the roots of $h$

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.



# 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 $k_{\mathrm{in}}$)

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

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

- Initialization: $\mathsf{S}_{1,0,c} := m^{-1} \sum_{i=1}^m \bar{\mathsf{s}}_{ci} \circ \mathsf{T}(\widehat{S}_{\mathrm{init}})$ and $\widehat{S}_{1,0} = \widehat{S}_{1,-1} := \widehat{S}_{\mathrm{init}}$
- At time $\#(t-1)k_{\mathrm{in}} + \tau$, 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 $\#tk_{\mathrm{in}}$, refresh the *control variate*
  - (central server) $\widehat{S}_{t,0} = \widehat{S}_{t,-1} := \widehat{S}_{t-1,k_{\mathrm{in}}}$
  - (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 $\bar{\mathsf{s}}_c \circ \mathsf{T}(\widehat{S}_\cdot)$
- The control variate is biased: it is **refreshed every $k_{\mathrm{in}}$ iterations**.

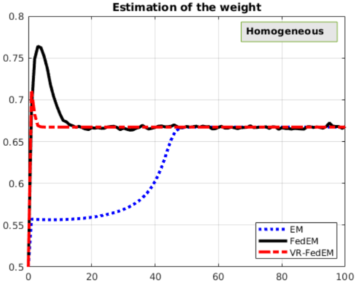
Same variance reduction as in SPIDER-EM, Fort et al. (2020) – SPIDER = Stochastic Path-Integrated Differential EstimatoR.

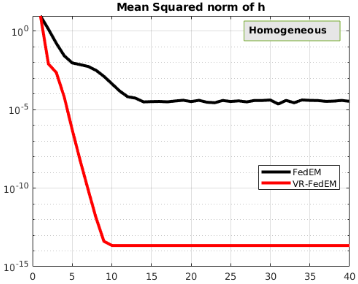
# VR-FedEM

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



# III. Explicit control of convergence
# Complexity analysis

# Assumptions

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

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

- 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

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

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

- On the model
  - 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: $s \mapsto \operatorname{argmin}_{\theta \in \Theta} \psi(\theta) - \langle s, \phi(\theta) \rangle$ 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

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

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

- On the model
- For the existence of a Lyapunov function

  A4 $W \stackrel{\text{def}}{=} F \circ \mathsf{T}$ 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

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

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

- 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

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

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

- 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

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

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

- 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

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

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

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

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

Denote by $K$ the uniform random variable on $[k_{\max} - 1]$. 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 $\epsilon$, 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 $\epsilon$:

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

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

- **Compression effect:** $\gamma$ 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 $\sigma^2$, small $\epsilon$ | Low $\sigma^2$ larger $\epsilon$ |
| $\frac{v_{\min}}{2L_{\dot{W}}}$ | large ratio $n/\omega^3$ | $\times\omega$ | $\times 1$ |
| $\frac{\sqrt{n}}{2\sqrt{2}L(1+\omega)\sqrt{\omega}}$ | low ratio $n/\omega^3$ | $\times\omega$ | $\times\omega^{3/2}/\sqrt{n}$ |

# Explicit control for VR-FedEM

Set $(m_c = m)$

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

**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 $\alpha \stackrel{\text{def}}{=} 1/(1+\omega)$, $V_{1,0,c} \stackrel{\text{def}}{=} \bar{\mathsf{s}}_c \circ \mathsf{T}(\widehat{S}_{1,0}) - \widehat{S}_{1,0}$, $\mathsf{b} \stackrel{\text{def}}{=} \lceil \frac{k_{\mathrm{in}}}{(1+\omega)^2} \rceil$ and

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

Let $(\tau, K)$ 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 $k_{\mathrm{out}}\, k_{\mathrm{in}}$ chosen in order to reach the accuracy level $\epsilon$:

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

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

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

# IV. 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.