FedCBO: Reaching Group Consensus in Clustered Federated Learning and Robustness to Backdoor Adversarial Attacks

Nicolás García Trillos
University of Wisconsin-Madison

ICERM
May 2024
Based on


The success stories of AI

From Dall-e: generate an image representing the success stories of AI.
However…

Safety (e.g., systems should be robust to perturbations of data)

Security (e.g., systems should be robust to adversarial agents)

Trust (e.g., privacy, fairness, watermarks for generated images)
Outline

1. Introduction: federated learning and clustered federated learning
2. Federated learning through consensus based optimization (CBO)
Federated Learning
Federated Learning
Federated Average

FedAvg [McMahan et al. 16']:

\[
\begin{align*}
    d\theta^i_t &= -\nabla L_i(\overline{\theta}_t) + \text{Noise}, & i = 1, \ldots, N. \\
    \overline{\theta}_t &= \frac{1}{N} \sum_{i=1}^{N} \theta^i_t
\end{align*}
\]
Federated Learning (heterogeneous setting)
Clustered Federated Learning
Iterated Federated Clustering Algorithm

IFCA [Ghosh et al 20']:

\[ \theta_1 = \frac{1}{|S_1|} \sum \bar{\theta}_1 \]
\[ \theta_2 = \frac{1}{|S_2|} \sum \bar{\theta}_2 \]

\[(d)\]

\[(a)\]
\[\{\theta_1, \theta_2\}\]

\[(c)\]
\[\tilde{\theta}_j = \theta_j - \gamma \nabla F(\theta_j)\]

\[(b)\]
Some References


Today: decentralized Clustered Federated Learning based on CBO
Setting for Clustered Federated Learning

Number of agents = $N$

Number of clusters = $K$

\[
L_k(\theta) := \mathbb{E}_{(x,y) \sim \mathcal{D}_k} [l(f(x; \theta), y)] , \quad k = 1, 2, \ldots, K.
\]

\[
\theta^*_k := \arg \min_{\theta \in \mathbb{R}^d} L_k(\theta)
\]
Part 1

Clustered Federated Learning through CBO
Consensus-based Optimization (CBO)

Optimization problem:

$$\min_{\theta \in \mathbb{R}^d} L(\theta)$$

Assumptions:

- $L$ has unique global min $\theta^*$. 
Consensus-based Optimization (CBO)

Interacting particle system:

\[
d\theta^i_t = -\lambda \left( \theta^i_t - m^\alpha_L[\rho^N_t] \right) dt + \sigma \left| \theta^i_t - m^\alpha_L[\rho^N_t] \right| dB^i_t, \quad i = 1, 2, \ldots, N,\]

where

\[
\rho^N_t := \frac{1}{N} \sum_{i=1}^{N} \delta_{\theta^i_t}
\]

\[
m^\alpha_L[\rho^N_t] := \int \theta \frac{\exp \left( -\alpha L(\theta) \right) \rho^N_t}{\int \exp \left( -\alpha L(\theta) \right) \rho^N_t d\theta} d\theta = \sum_{i=1}^{N} w^i \theta^i_t, \quad \text{with} \quad w^i := \frac{\exp \left( -\alpha L(\theta^i_t) \right)}{\sum_{j=1}^{N} \exp \left( -\alpha L(\theta^j_t) \right)}
\]

Consensus-based Optimization (CBO)
Clustered Federated Learning

Optimization problem:

\[
\min_{\theta \in \mathbb{R}^d} L_1(\theta) \quad \text{and} \quad \min_{\theta \in \mathbb{R}^d} L_2(\theta)
\]

Number of cluster 1 agents = \(N_1\)

Number of cluster 2 agents = \(N_2\)

Total number of agents = \(N\)
FedCBO System

\[ d\theta_{1,i}^t = -\lambda_1 \left( \theta_{1,i}^t - m_{L_1}^\alpha[\rho_t^N] \right) dt - \lambda_2 \nabla L_1(\theta_{1,i}^t) dt + \sigma_1 \left| \theta_{1,i}^t - m_{L_1}^\alpha[\rho_t^N] \right| dB_{1,i}^t + \sigma_2 \left| \nabla L_1(\theta_{1,i}^t) \right| d\tilde{B}_{1,i}^t \]

\[ d\theta_{2,j}^t = -\lambda_1 \left( \theta_{2,j}^t - m_{L_2}^\alpha[\rho_t^N] \right) dt - \lambda_2 \nabla L_2(\theta_{2,j}^t) dt + \sigma_1 \left| \theta_{2,j}^t - m_{L_2}^\alpha[\rho_t^N] \right| dB_{2,j}^t + \sigma_2 \left| \nabla L_2(\theta_{2,j}^t) \right| d\tilde{B}_{2,j}^t \]

where

\[ \rho_{1,N}^t = \frac{1}{N_1} \sum_{i=1}^{N_1} \delta_{\theta_{1,i}^t}, \quad \rho_{2,N}^t = \frac{1}{N_2} \sum_{j=1}^{N_2} \delta_{\theta_{2,j}^t}, \quad \rho_N^t = \frac{N_1}{N} \rho_{1,N}^t + \frac{N_2}{N} \rho_{2,N}^t. \]
FedCBO System

\[ d\theta_{\alpha_{1}}^{1,i} = -\lambda_{1} \left( \theta_{\alpha_{1}}^{1,i} - m_{L_{1}}^{\alpha}[\rho_{t}^{N}] \right) dt - \lambda_{2} \nabla L_{1}(\theta_{\alpha_{1}}^{1,i}) dt + \sigma_{1} \left| \theta_{\alpha_{1}}^{1,i} - m_{L_{1}}^{\alpha}[\rho_{t}^{N}] \right| dB_{\alpha_{1}}^{1,i} + \sigma_{2} \left| \nabla L_{1}(\theta_{\alpha_{1}}^{1,i}) \right| d\tilde{B}_{\alpha_{1}}^{1,i} \]

\[ d\theta_{\alpha_{2}}^{2,j} = -\lambda_{1} \left( \theta_{\alpha_{2}}^{2,j} - m_{L_{2}}^{\alpha}[\rho_{t}^{N}] \right) dt - \lambda_{2} \nabla L_{2}(\theta_{\alpha_{2}}^{2,j}) dt + \sigma_{1} \left| \theta_{\alpha_{2}}^{2,j} - m_{L_{2}}^{\alpha}[\rho_{t}^{N}] \right| dB_{\alpha_{2}}^{2,j} + \sigma_{2} \left| \nabla L_{2}(\theta_{\alpha_{2}}^{2,j}) \right| d\tilde{B}_{\alpha_{2}}^{2,j} \]

where

\[ m_{L_{1}}^{\alpha}[\rho_{t}^{N}] := \int \frac{\exp(-\alpha L_{1}(\theta)) \rho_{t}^{N}}{\int \exp(-\alpha L_{1}(\theta)) \rho_{t}^{N}} d\theta = \sum_{i=1}^{N_{1}} w_{\alpha_{1}}^{1,i} \theta_{\alpha_{1}}^{1,i} + \sum_{j=1}^{N_{2}} w_{\alpha_{2}}^{2,j} \theta_{\alpha_{2}}^{2,j} \]

\[ w_{\alpha_{1}}^{1,i} := \frac{\exp(-\alpha L_{1}(\theta_{\alpha_{1}}^{1,i}))}{Z_{L_{1}}}, \quad w_{\alpha_{2}}^{2,j} := \frac{\exp(-\alpha L_{1}(\theta_{\alpha_{2}}^{2,j}))}{Z_{L_{1}}} \]
FedCBO System

\[
d\theta^1, i = -\lambda_1 \left( \theta^1, i - m^\alpha_{L_1} [\rho^N_t] \right) dt - \lambda_2 \nabla L_1(\theta^1, i) dt + \sigma_1 |\theta^1, i - m^\alpha_{L_1} [\rho^N_t]| dB^1, i + \sigma_2 \nabla L_1(\theta^1, i) d\tilde{B}^1, i
\]

\[
d\theta^2, j = -\lambda_1 \left( \theta^2, j - m^\alpha_{L_2} [\rho^N_t] \right) dt - \lambda_2 \nabla L_2(\theta^2, j) dt + \sigma_1 |\theta^2, j - m^\alpha_{L_2} [\rho^N_t]| dB^2, j + \sigma_2 \nabla L_2(\theta^2, j) d\tilde{B}^2, j
\]

where

\[
m^\alpha_{L_2} [\rho^N_t] := \int \theta \frac{\exp (-\alpha L_2(\theta)) \rho^N_t}{\int \exp (-\alpha L_2(\theta)) \rho^N_t} d\theta = \sum_{i=1}^{N_1} w^1, i \theta^1, i + \sum_{j=1}^{N_2} w^2, j \theta^2, j,
\]

\[
w^1, i := \frac{\exp (-\alpha L_2(\theta^1, i))}{Z_{L_2}}, \quad w^2, j := \frac{\exp (-\alpha L_2(\theta^2, j))}{Z_{L_2}}
\]
FedCBO System

Particles positions at time $T=1$
FedCBO System

\[ d\theta^{1,i}_t = -\lambda_1 \left( \theta^{1,i}_t - m^{\alpha}_{L_1} [\rho^N_t] \right) dt - \lambda_2 \nabla L_1(\theta^{1,i}_t) dt + \sigma_1 |\theta^{1,i}_t - m^{\alpha}_{L_1} [\rho^N_t]| dB^{1,i}_t + \sigma_2 |\nabla L_1(\theta^{1,i}_t)| d\tilde{B}^{1,i}_t \]

\[ d\theta^{2,j}_t = -\lambda_1 \left( \theta^{2,j}_t - m^{\alpha}_{L_2} [\rho^N_t] \right) dt - \lambda_2 \nabla L_2(\theta^{2,j}_t) dt + \sigma_1 |\theta^{2,j}_t - m^{\alpha}_{L_2} [\rho^N_t]| dB^{2,j}_t + \sigma_2 |\nabla L_2(\theta^{2,j}_t)| d\tilde{B}^{2,j}_t \]

As \( N \) goes to \( \infty \)

\[ d\theta^1_t = -\lambda_1 \left( \theta^1_t - m^{\alpha}_{L_1} [\rho_t] \right) dt - \lambda_2 \nabla L_1(\theta^1_t) dt + \sigma_1 |\theta^1_t - m^{\alpha}_{L_1} [\rho_t]| dB^1_t + \sigma_2 |\nabla L_1(\theta^1_t)| d\tilde{B}^1_t \]

\[ d\theta^2_t = -\lambda_1 \left( \theta^2_t - m^{\alpha}_{L_2} [\rho_t] \right) dt - \lambda_2 \nabla L_2(\theta^2_t) dt + \sigma_1 |\theta^2_t - m^{\alpha}_{L_2} [\rho_t]| dB^2_t + \sigma_2 |\nabla L_2(\theta^2_t)| d\tilde{B}^2_t \]
FedCBO System

$$d\theta_{1,i} = -\lambda_1 \left( \theta_{1,i}^{1} - m_{L_1}^{\alpha} [\rho_{t}^N] \right) dt - \lambda_2 \nabla L_1(\theta_{1,i}^{1})dt + \sigma_1 |\theta_{1,i}^{1} - m_{L_1}^{\alpha} [\rho_{t}^N]| dB_{t,i}^{1} + \sigma_2 |\nabla L_1(\theta_{1,i}^{1})| d\tilde{B}_{t,i}^{1}$$

$$d\theta_{2,j} = -\lambda_1 \left( \theta_{2,j}^{2} - m_{L_2}^{\alpha} [\rho_{t}^N] \right) dt - \lambda_2 \nabla L_2(\theta_{2,j}^{2})dt + \sigma_1 |\theta_{2,j}^{2} - m_{L_2}^{\alpha} [\rho_{t}^N]| dB_{t,j}^{2} + \sigma_2 |\nabla L_2(\theta_{2,j}^{2})| d\tilde{B}_{t,j}^{2}$$

As N goes to \(\infty\)

$$d\theta_{1} = -\lambda_1 \left( \theta_{1}^{1} - m_{L_1}^{\alpha} [\rho_{t}] \right) dt - \lambda_2 \nabla L_1(\theta_{1}^{1})dt + \sigma_1 |\theta_{1}^{1} - m_{L_1}^{\alpha} [\rho_{t}]| dB_{t}^{1} + \sigma_2 |\nabla L_1(\theta_{1}^{1})| d\tilde{B}_{t}^{1}$$

$$d\theta_{2} = -\lambda_1 \left( \theta_{2}^{2} - m_{L_2}^{\alpha} [\rho_{t}] \right) dt - \lambda_2 \nabla L_2(\theta_{2}^{2})dt + \sigma_1 |\theta_{2}^{2} - m_{L_2}^{\alpha} [\rho_{t}]| dB_{t}^{2} + \sigma_2 |\nabla L_2(\theta_{2}^{2})| d\tilde{B}_{t}^{2}$$
Consensus-based Optimization (CBO)

Theorem (Mean-field limit; Carrillo, NGT, Li, Zhu, 23’):

Suppose $\theta_i^0 \sim \rho_0^1$ and $\theta_i^0 \sim \rho_0^2$. Also, suppose $N \to \infty$ and

$$\frac{N_1}{N} \to w_1, \quad \frac{N_2}{N} \to w_2.$$

Then $\rho^{N,1} \to \rho^1$ and $\rho^{N,2} \to \rho^2$.

\[
\begin{align*}
\partial_t \rho_t^1 &:= \Delta (\kappa_t^1 \rho_t^1) + \nabla \cdot (\mu_t^1 \rho_t^1), & \lim_{t \to 0} \rho_t^1 = \rho_0^1 \\
\partial_t \rho_t^2 &:= \Delta (\kappa_t^2 \rho_t^2) + \nabla \cdot (\mu_t^2 \rho_t^2), & \lim_{t \to 0} \rho_t^2 = \rho_0^2, \\
\mu_t^k &:= \lambda_1 (\theta - m_{L_k}^\alpha [\rho_t]) + \lambda_2 \nabla L_k (\theta), & \kappa_t^k := \frac{\sigma_1^2}{2} |\theta - m_{L_k}^\alpha [\rho_t]|^2 + \frac{\sigma_2^2}{2} |\nabla L_k (\theta)|^2, \text{ for } k = 1, 2.
\end{align*}
\]
Consensus-based Optimization (CBO)

Theorem (Long-time behavior mean field; Carrillo, NGT, Li, Zhu, 23’):

Let $\rho^k_0$ give positive mass around $\theta^*_k$ (global minimizer of $L_k$) for each $k = 1, 2$. Let $(\rho^1_t, \rho^2_t)$ be solution of mean field PDE. Let $\varepsilon > 0$. Provided parameters $\lambda, \sigma, \alpha$ are chosen appropriately, we have, for some $T^*$,

$$\mathcal{V}(\rho^1_t) + \mathcal{V}(\rho^2_t) \leq \exp(-ct)(\mathcal{V}(\rho^1_0) + \mathcal{V}(\rho^2_0)), \quad \forall t \in [0, T^*]$$

and

$$\min_{t \in [0, T^*]} \mathcal{V}(\rho^1_t) + \mathcal{V}(\rho^2_t) \leq \varepsilon,$$

where

$$\mathcal{V}(\rho^k_t) := \int |\theta - \theta^*_k|^2 d\rho^k_t(\theta).$$
Consensus-based Optimization (CBO)

Theorem (Long-time behavior mean field; Carrillo, NGT, Li, Zhu, 23’):

Let $\rho_0^k$ give positive mass around $\theta_k^*$ (global minimizer of $L_k$) for each $k = 1, 2$. Let $(\rho^1_t, \rho^2_t)$ be solution of mean field PDE. Let $\varepsilon > 0$.

Provided parameters $\lambda, \sigma, \alpha$ are chosen appropriately, we have, for some $T^*$,

$$T^* := \frac{1}{(1-\vartheta)(2\lambda_1 - 2\lambda_2 M - d\sigma_1^2 - d\sigma_2^2 M^2)} \log \left( \frac{\mathcal{V}(\rho_0^1) + \mathcal{V}(\rho_0^2)}{\varepsilon} \right)$$

$$\mathcal{V}(\rho^1_t) + \mathcal{V}(\rho^2_t) \leq \exp(-ct) (\mathcal{V}(\rho_0^1) + \mathcal{V}(\rho_0^2)), \quad \forall t \in [0, T^*]$$

and

$$\min_{t \in [0, T^*]} \mathcal{V}(\rho^1_t) + \mathcal{V}(\rho^2_t) \leq \varepsilon,$$

where

$$\mathcal{V}(\rho_t^k) := \int \|\theta - \theta_k^*\|^2 d\rho_t^k(\theta).$$
Some References


- Riedl K. *Leveraging memory effects and gradient information in consensus-based optimisation: On global convergence in mean-field law*. EJAM. Published online 2023:1-32. doi:10.1017/S0956792523000293
Some References


Experiments

Rotated MNIST:

90 degrees rotation
Experiments

Number of clusters = 4

Number of agents in each cluster = 300

Number of data points in each agent = 200
Experiments

Table 1. Test accuracy ± standard deviation % on rotated MNIST.

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>FedCBO</td>
<td>96.51 ± 0.04</td>
</tr>
<tr>
<td>IFCA</td>
<td>94.44 ± 0.01</td>
</tr>
<tr>
<td>FedAvg</td>
<td>85.50 ± 0.19</td>
</tr>
<tr>
<td>Local</td>
<td>81.27 ± 0.02</td>
</tr>
</tbody>
</table>

one of the reasons why our FedCBO method outperforms the IFCA algorithm. As pointed out in \[15\], the FedAvg baseline performs worse than FedCBO and IFCA as it tries to fit heterogeneous data using one model and thus cannot provide cluster-wise predictions. Since each agent only stores a small amount of data, the local model training scheme can easily overfit to the local dataset. This explains why it produces the worst performance among all other methodologies.

Remark 6. To verify the correctness of the sampling scheme in FedCBO, we define the successful selection rate (SR) for agent \(j\) at iteration \(n\) as follows:

\[
SR_j^n := \frac{\text{Number of selected agents in the same cluster as agent } j}{\text{Total number of selected agents}}
\]

where the total number of selected agents equals the model download budget \(M\). During the FedCBO algorithm, we calculate the average successful selection rate \(SR_n := \frac{1}{N} \sum_{j=1}^{N} SR_j^n\) at each communication round \(n\), which corresponds to the blue curve in Fig. 3. Meanwhile, when implementing the “-greedy sampling, we set the random exploration proportion to 0.5 at \(n=0\) and use a decay scheme of \(\alpha(n) = \max\{0.5, 0.01 \cdot n\ discounts\} \}

Hence we can calculate the oracle expected successful selection rate at each round: this is shown as the orange curve in Fig. 3. We note that the empirical average successful SR (blue curve) is very close to the best expected successful SR (orange curve). This indicates that our FedCBO algorithm can successfully identify the agents with the same data distributions. We leave the task of designing better sampling strategies to close the gap between empirical successful SR and oracle successful SR to future work.

Figure 3. Average successful selection rate (SR) at each communication round.

Well-posedness of mean-field equations

In this section we prove Theorem 2.1. We present the details in the case in which the loss functions are assumed to be bounded. The proof for the quadratic growth case is similar, and we refer the reader to the Appendix for more details.
Part 2

Backdoor attacks
Backdoor attacks
Backdoor attacks
Backdoor attacks

\[ d\theta_{t_1}^{1,i} = -\lambda_1 \left( \theta_{t_1}^{1,i} - m_1^{\alpha} L_1[\rho_t^N] \right) dt - \lambda_2 \nabla L_1(\theta_{t_1}^{1,i}) dt + \sigma_1 \left| \theta_{t_1}^{1,i} - m_1^{\alpha} L_1[\rho_t^N] \right| dB_{t_1}^{1,i} + \sigma_2 \left| \nabla L_1(\theta_{t_1}^{1,i}) \right| d\tilde{B}_{t_1}^{1,i} \]

\[ d\theta_{t_2}^{2,j} = -\lambda_1 \left( \theta_{t_2}^{2,j} - m_2^{\alpha} L_2[\rho_t^N] \right) dt - \lambda_2 \nabla L_2(\theta_{t_2}^{2,j}) dt + \sigma_1 \left| \theta_{t_2}^{2,j} - m_2^{\alpha} L_2[\rho_t^N] \right| dB_{t_2}^{2,j} + \sigma_2 \left| \nabla L_2(\theta_{t_2}^{2,j}) \right| d\tilde{B}_{t_2}^{2,j} \]
Backdoor attacks

Malicious agents’ goal: make other agents predict points of class $C_S$ as class $C_T$. 
Backdoor attacks via label flipping

Instead of aiming to optimize

\[
L_k(\theta) = \mathbb{E}_{(x,y) \sim D_k}[l(f(x; \theta), y)] = \sum_{c=1}^{C} w_c \mathbb{E}_{x|y=c}[l(f(x; \theta), c)]
\]

a malicious agent picks parameters to optimize:

\[
L^\text{mal}_k(\theta) := \sum_{c \neq c_S}^{C} w_c \mathbb{E}_{x|y=c}[l(f(x; \theta), c)] + w_{c_S} \mathbb{E}_{x|y=c_S}[l(f(x; \theta), c_T)]
\]
Backdoor attacks via label flipping

Instead of aiming to optimize

\[ L_k(\theta) = \mathbb{E}_{(x,y) \sim D_k}[l(f(x; \theta), y)] = \sum_{c=1}^{C} w_c \mathbb{E}_{x|y=c}[l(f(x; \theta), c)] \]

a malicious agent picks parameters to optimize:

\[ L_k^{\text{mal}}(\theta) := \sum_{c \neq c_s}^{C} w_c \mathbb{E}_{x|y=c}[l(f(x; \theta), c)] + w_{c_s} \mathbb{E}_{x|y=c_s}[l(f(x; \theta), c_T)] \]

**Benign agents:** introduce additional robustness criterion to protect against these attacks.
Bi-Level Optimization

Optimization problem:

\[ \min_{\theta \in \Theta} G(\theta) \]
\[ \text{s.t. } \theta \in \arg \min L \]

Assumptions:

- Unique solution \( \theta_{\text{good}} \)
Bilevel CBO

Interacting particle system:

\[ d\theta^i_t = -\lambda(\theta^i_t - m^{\alpha,\beta}[\rho^N_t])dt + \sigma|\theta^i_t - m^{\alpha,\beta}[\rho^N_t]|dB^i_t, \quad i = 1, \ldots, N. \]

where

\[ \rho^N_t := \frac{1}{N} \sum_{i=1}^{N} \delta_{\theta^i_t} \]

\[ m^{\alpha,\beta}[\rho^N_t] := \int_{\Theta} \frac{\exp(-\alpha G(\theta))}{\int \exp(-\alpha G(\theta)) dl_\beta[\rho^N_t](\theta)} dl_\beta[\rho^N_t](\theta) \]

\[ l_\beta[\rho^N_t] := \rho^N_t(\cdot \cap Q_\beta[\rho^N_t]) \quad Q_\beta[\rho^N_t] := \{ \theta \text{ s.t. } L(\theta) \leq q_\beta[\rho^N_t] \} \quad q_\beta[\rho^N_t] := \inf\{ q \text{ s.t. } \rho^N_t(\{L(\theta) \leq q\}) \geq \beta \} \]
Bilevel CBO

Interacting particle system:

\[ d\theta_t^i = -\lambda(\theta_t^i - m^{\alpha, \beta}[\rho_t^N])dt + \sigma|\theta_t^i - m^{\alpha, \beta}[\rho_t^N]|dB_t^i, \quad i = 1, \ldots N. \]

where

\[ \rho_t^N := \frac{1}{N} \sum_{i=1}^{N} \delta_{\theta_t^i} \]

\[ m^{\alpha, \beta}[\rho_t^N] := \int \frac{\exp(-\alpha G(\theta))}{\int \exp(-\alpha G(\theta))dl_{\beta}[\rho_t^N](\theta)} dl_{\beta}[\rho_t^N](\theta) \]

\[ l_{\beta}[\rho_t^N] := \rho_t^N(\cdot \cap Q_{\beta}[\rho_t^N]) \quad Q_{\beta}[\rho_t^N] := \{\theta \text{ s.t. } L(\theta) \le q_{\beta}[\rho_t^N]\} \quad q_{\beta}[\rho_t^N] := \inf\{q \text{ s.t. } \rho_t^N(\{L(\theta) \le q\}) \ge \beta\} \]
Bilevel CBO

Interacting particle system:

\[ d\theta^i_t = -\lambda(\theta^i_t - m^{\alpha,\beta}[\rho^N_t])dt + \sigma|\theta^i_t - m^{\alpha,\beta}[\rho^N_t]|dB^i_t, \quad i = 1, \ldots N. \]

where

\[
\rho^N_t := \frac{1}{N} \sum_{i=1}^{N} \delta_{\theta^i_t}
\]

\[
m^{\alpha,\beta}[\rho^N_t] := \int \frac{\exp(-\alpha G(\theta))}{\int \exp(-\alpha G(\theta)) dl^{\beta}[\rho^N_t](\theta)} dl^{\beta}[\rho^N_t](\theta)
\]

\[
l^{\beta}[\rho^N_t] := \rho^N_t (\cdot \cap Q^{\beta}[\rho^N_t]) \quad Q^{\beta}[\rho^N_t] := \{\theta \text{ s.t. } L(\theta) \leq q^\beta[\rho^N_t]\} \quad q^\beta[\rho^N_t] := \inf\{q \text{ s.t. } \rho^N_t(\{L(\theta) \leq q\}) \geq \beta\}\]
Bilevel CBO

Interacting particle system:

\[ d\theta_t^i = -\lambda(\theta_t^i - m^{\alpha,\beta}[\rho_t^N])dt + \sigma|\theta_t^i - m^{\alpha,\beta}[\rho_t^N]|dB_t^i, \quad i = 1, \ldots, N. \]

where

\[ \rho_t^N := \frac{1}{N} \sum_{i=1}^{N} \delta_{\theta_t^i} \]

Within top \( \beta \times (100) \% \), largest \( L \)

\[ m^{\alpha,\beta}[\rho_t^N] := \int \theta \frac{\exp(-\alpha G(\theta))}{\int \exp(-\alpha G(\theta))dI^\beta[\rho_t^N](\theta)} dI^\beta[\rho_t^N](\theta) \]

\[ I^\beta[\rho_t^N] := \rho_t^N \cap Q^\beta[\rho_t^N] \]

\[ Q^\beta[\rho_t^N] := \{ \theta \text{ s.t. } L(\theta) \leq q^\beta[\rho_t^N] \} \]

\[ q^\beta[\rho_t^N] := \inf \{ q \text{ s.t. } \rho_t^N(\{L(\theta) \leq q\}) \geq \beta \} \]
Example:

$$\min_{\theta \in \Theta} G(\theta)$$

s.t. \( \theta \in \arg \min L \)

Experiments: \( \theta_0^i \sim \text{Uniform}[-10, 10] \)

<table>
<thead>
<tr>
<th>( \beta = 0.1 )</th>
<th>( \beta = 0.5 )</th>
<th>( \beta = 0.8 )</th>
<th>( \beta = 0.9 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>10/10 (T=50)</td>
<td>10/10 (T=600)</td>
<td>10/10 (T=6000)</td>
<td>0/10 (T=20000)</td>
</tr>
</tbody>
</table>
Example: Constrained optimization via Bilevel CBO

\[
\min_{\theta \in \mathbb{R}^3} G(\theta)
\]
\[
\text{s.t. } \theta \in \mathcal{C} := \partial B_1(0)
\]

where

\[
G(\theta) := -20 \exp \left( -0.2 \sqrt{\frac{1}{3} \sum_{l=1}^{3} (\theta_l - \rho_l)} \right) + \exp \left( \frac{1}{3} \sum_{l=1}^{3} \cos(2\pi(\theta_l - \rho_l)) \right)
\]

\[
\rho = (0.4, 0.4, 0.4)
\]
Where

\[ G(\theta) := -20 \exp \left( -0.2 \sqrt{\frac{1}{3} \sum_{l=1}^{3} (\theta_l - p_l)} \right) + \exp \left( \frac{1}{3} \sum_{l=1}^{3} \cos(2\pi(\theta_l - p_l)) \right) \]

\[ \rho = (0.4, 0.4, 0.4) \]

\[ L(\theta) = (1 - |\theta|)^2 \]
Constrained optimization via Bilevel CBO

\[
\min_{\theta \in \Theta} \quad G(\theta)
\]

s.t. \; \theta \in \text{arg min} \; L

**Experiments:** \( \theta_0^i \sim \text{Uniform}[-10, 10]^3 \)

<table>
<thead>
<tr>
<th>( \beta = 0.02 )</th>
<th>( \beta = 0.05 )</th>
<th>( \beta = 0.1 )</th>
<th>( \beta = 0.2 )</th>
<th>( \beta = 0.3 )</th>
<th>( \beta = 0.5 )</th>
<th>( \beta = 0.8 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/10 (T=1000)</td>
<td>10/10 (T=200)</td>
<td>10/10 (T=200)</td>
<td>10/10 (T=200)</td>
<td>10/10 (T=1000)</td>
<td>10/10 (T=2000)</td>
<td>0/10 (T=2000)</td>
</tr>
</tbody>
</table>
BiLevel FedCBO

Optimization problems: for $k = 1, \ldots, K$

$$
\min_{\theta \in \Theta} \quad G_k(\theta)
$$

s.t. \quad \theta \in \arg \min L_k

where $L_k(\theta)$ and $G_k(\theta)$ are, for example,

$$
L_k(\theta) = \sum_{c=1}^{C} w_{k,c} L_{k,c}(\theta) \quad \quad \quad G_k(\theta) = -\sum_{c=1}^{C} w_{k,c} \log \left( \frac{L_{k,c}(\theta)}{L_k(\theta)} \right)
$$
\[ d\theta_t^{1,i} = -\lambda_1(\theta_t^{1,i} - m_{L_1,G_1}^{\alpha,\beta}[\rho_t^N]) dt - \lambda_2 \nabla L_1(\theta_t^{1,i}) dt \\
+ \sigma_1 |\theta_t^{1,i} - m_{L_1,G_1}^{\alpha,\beta}[\rho_t^N]| dB_t^{1,i} + \sigma_2 |\nabla L_1(\theta_t^{1,i})| d\tilde{B}_t^{1,i} \]

\[ d\theta_t^{2,j} = -\lambda_1(\theta_t^{2,j} - m_{L_2,G_2}^{\alpha,\beta}[\rho_t^N]) dt - \lambda_2 \nabla L_2(\theta_t^{2,j}) dt \\
+ \sigma_1 |\theta_t^{2,j} - m_{L_2,G_2}^{\alpha,\beta}[\rho_t^N]| dB_t^{2,j} + \sigma_2 |\nabla L_2(\theta_t^{2,j})| d\tilde{B}_t^{2,j} \]
Experiments on CIFAR10
Experiments

Experimental setting 1 (CIFAR10 homogeneous case):
-Total number of agents $N = 10$;
-Num of benign agents = 7; Num of malicious agents = 3;
-Num of data for each benign agent = 500;
-Num of data for each malicious agent = 1200;

Attacks:
Source class: class 0 (images of planes)
Target class: class 2 (images of birds)
Label flipping: 0 $\rightarrow$ 2.
## Experiments

<table>
<thead>
<tr>
<th></th>
<th>With backdoor attack (FedCBO $\alpha = 1$)</th>
<th>With backdoor attack (FedCBO $\alpha = 10$)</th>
<th>Without backdoor attack (FedCBO $\alpha = 1$)</th>
<th>Without malicious agents (FedCBO $\alpha = 1$)</th>
<th>With backdoor attack (Bilevel FedCBO $\alpha = 20$, $\beta = 1.0$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg overall acc</td>
<td>$63.85 \pm 0.18%$</td>
<td>$61.69 \pm 1.21%$</td>
<td>$65.30 \pm 0.35%$</td>
<td>$60.76 \pm 0.29%$</td>
<td>$61.06 \pm 0.21%$</td>
</tr>
<tr>
<td>Acc on class 0</td>
<td>$29.86 \pm 1.79%$</td>
<td>$44.38 \pm 3.36%$</td>
<td>$41.86 \pm 5.22%$</td>
<td>$61.10 \pm 3.49%$</td>
<td>$58.62 \pm 3.72%$</td>
</tr>
<tr>
<td>Benign agents’ models predict images of class 0 as label 2</td>
<td>$34.84 \pm 3.70%$ (Attack success rate)</td>
<td>$22.50 \pm 1.86%$ (Attack success rate)</td>
<td>$11.38 \pm 1.18%$</td>
<td>$6.90 \pm 1.29%$ (Attack success rate)</td>
<td>$9.84 \pm 1.43%$ (Attack success rate)</td>
</tr>
</tbody>
</table>
Experiments

<table>
<thead>
<tr>
<th></th>
<th>With backdoor attack (FedCBO $\alpha = 1$)</th>
<th>With backdoor attack (FedCBO $\alpha = 10$)</th>
<th>Without backdoor attack (FedCBO $\alpha = 1$)</th>
<th>Without malicious agents (FedCBO $\alpha = 1$)</th>
<th>With backdoor attack (Bilevel FedCBO $\alpha = 20$, $\beta = 1.0$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg overall acc</td>
<td>63.85 ± 0.18 %</td>
<td>61.69 ± 1.21 %</td>
<td>65.30 ± 0.35 %</td>
<td>60.76 ± 0.29 %</td>
<td>61.06 ± 0.21 %</td>
</tr>
<tr>
<td>Acc on class 0</td>
<td>29.86 ± 1.79 %</td>
<td>44.38 ± 3.36 %</td>
<td>41.86 ± 5.22 %</td>
<td>61.10 ± 3.49 %</td>
<td>58.62 ± 3.72 %</td>
</tr>
<tr>
<td>Benign agents’ models predict images of class 0 as label 2</td>
<td>34.84 ± 3.70 % (Attack success rate)</td>
<td>22.50 ± 1.86 % (Attack success rate)</td>
<td>11.38 ± 1.18 %</td>
<td>6.90 ± 1.29 %</td>
<td>9.84 ± 1.43 % (Attack success rate)</td>
</tr>
</tbody>
</table>

*With Backdoor Attack:*  
Total number of class 0 images (with correct labels) from the benign agents = 284;  
Total number of class 0 images (with wrong labels) from the malicious agents = 356;  
(i.e. in the entire dataset, about 45% class 0 images have correct labels and 55% of them have wrong labels)
## Experiments

<table>
<thead>
<tr>
<th></th>
<th>With backdoor attack (FedCBO $\alpha = 1$)</th>
<th>With backdoor attack (FedCBO $\alpha = 10$)</th>
<th>Without backdoor attack (FedCBO $\alpha = 1$)</th>
<th>Without malicious agents (FedCBO $\alpha = 1$)</th>
<th>With backdoor attack (Bilevel FedCBO $\alpha = 20$, $\beta = 1.0$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg overall acc</td>
<td>63.85 ± 0.18 %</td>
<td>61.69 ± 1.21 %</td>
<td>65.30 ± 0.35 %</td>
<td>60.76 ± 0.29 %</td>
<td>61.06 ± 0.21 %</td>
</tr>
<tr>
<td>Acc on class 0</td>
<td>29.86 ± 1.79 %</td>
<td>44.38 ± 3.36 %</td>
<td>41.86 ± 5.22 %</td>
<td>61.10 ± 3.49 %</td>
<td>58.62 ± 3.72 %</td>
</tr>
<tr>
<td>Benign agents’ models predict images of class 0 as label 2</td>
<td>34.84 ± 3.70 % (Attack success rate)</td>
<td>22.50 ± 1.86 % (Attack success rate)</td>
<td>11.38 ± 1.18 %</td>
<td>6.90 ± 1.29 %</td>
<td>9.84 ± 1.43 % (Attack success rate)</td>
</tr>
</tbody>
</table>

*Without Backdoor Attack:*
Remove all the class 0 images contained in malicious agents.
## Experiments

### Without Malicious Agents:
Remove all the malicious agents.

<table>
<thead>
<tr>
<th></th>
<th>With backdoor attack (FedCBO $\alpha = 1$)</th>
<th>With backdoor attack (FedCBO $\alpha = 10$)</th>
<th>Without backdoor attack (FedCBO $\alpha = 1$)</th>
<th>Without malicious agents (FedCBO $\alpha = 1$)</th>
<th>With backdoor attack (Bilevel FedCBO $\alpha = 20, \beta = 1.0$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg overall acc</td>
<td>63.85 ± 0.18 %</td>
<td>61.69 ± 1.21 %</td>
<td>65.30 ± 0.35 %</td>
<td>60.76 ± 0.29 %</td>
<td>61.06 ± 0.21 %</td>
</tr>
<tr>
<td>Acc on class 0</td>
<td>29.86 ± 1.79 %</td>
<td>44.38 ± 3.36 %</td>
<td>41.86 ± 5.22 %</td>
<td>61.10 ± 3.49 %</td>
<td>58.62 ± 3.72 %</td>
</tr>
<tr>
<td>Benign agents’ models predict images of class 0 as label 2</td>
<td>34.84 ± 3.70 % (Attack success rate)</td>
<td>22.50 ± 1.86 % (Attack success rate)</td>
<td>11.38 ± 1.18 %</td>
<td>6.90 ± 1.29 % (Attack success rate)</td>
<td>9.84 ± 1.43 % (Attack success rate)</td>
</tr>
</tbody>
</table>
## Experiments

<table>
<thead>
<tr>
<th></th>
<th>With backdoor attack (FedCBO $\alpha = 1$)</th>
<th>With backdoor attack (FedCBO $\alpha = 10$)</th>
<th>Without backdoor attack (FedCBO $\alpha = 1$)</th>
<th>Without malicious agents (FedCBO $\alpha = 1$)</th>
<th>With backdoor attack (Bilevel FedCBO $\alpha = 20$, $\beta = 1.0$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg overall acc</td>
<td>63.85 ± 0.18 %</td>
<td>61.69 ± 1.21 %</td>
<td><strong>65.30 ± 0.35 %</strong></td>
<td>60.76 ± 0.29 %</td>
<td>61.06 ± 0.21 %</td>
</tr>
<tr>
<td>Acc on class 0</td>
<td>29.86 ± 1.79 %</td>
<td>44.38 ± 3.36 %</td>
<td>41.86 ± 5.22 %</td>
<td><strong>61.10 ± 3.49 %</strong></td>
<td><strong>58.62 ± 3.72 %</strong></td>
</tr>
<tr>
<td>Benign agents’ models predict images of class 0 as label 2</td>
<td>34.84 ± 3.70 % (Attack success rate)</td>
<td>22.50 ± 1.86 % (Attack success rate)</td>
<td>11.38 ± 1.18 %</td>
<td><strong>6.90 ± 1.29 %</strong> (Attack success rate)</td>
<td><strong>9.84 ± 1.43 %</strong> (Attack success rate)</td>
</tr>
</tbody>
</table>
## Experiments

<table>
<thead>
<tr>
<th></th>
<th>With backdoor attack (FedCBO $\alpha = 1$)</th>
<th>With backdoor attack (FedCBO $\alpha = 10$)</th>
<th>Without backdoor attack (FedCBO $\alpha = 1$)</th>
<th>Without malicious agents (FedCBO $\alpha = 1$)</th>
<th>With backdoor attack (Bilevel FedCBO $\alpha = 20$, $\beta = 1.0$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg overall acc</td>
<td>63.85 ± 0.18 %</td>
<td>61.69 ± 1.21 %</td>
<td><strong>65.30 ± 0.35 %</strong></td>
<td>60.76 ± 0.29 %</td>
<td>61.06 ± 0.21 %</td>
</tr>
<tr>
<td>Acc on class 0</td>
<td>29.86 ± 1.79 %</td>
<td>44.38 ± 3.36 %</td>
<td>41.86 ± 5.22 %</td>
<td><strong>61.10 ± 3.49 %</strong></td>
<td><strong>58.62 ± 3.72 %</strong></td>
</tr>
<tr>
<td>Benign agents’ models predict images of class 0 as label 2 (Attack success rate)</td>
<td>34.84 ± 3.70 %</td>
<td>22.50 ± 1.86 %</td>
<td>11.38 ± 1.18 %</td>
<td><strong>6.90 ± 1.29 %</strong></td>
<td><strong>9.84 ± 1.43 %</strong></td>
</tr>
</tbody>
</table>
Experiments

**Experimental setting 2 (Rotated CIFAR10):**
- Total number of agents $N = 20$;
- Num of clusters $k = 2$;
- Num of benign agents per cluster = 7; Num of malicious agents per cluster = 3;
- Num of data for each benign agent = 500;
- Num of data for each malicious agent = 1200;
Experiments

Experimental setting 2 (Rotated CIFAR10):
- Total number of agents $N = 20$;
- Num of clusters $k = 2$;
- Num of benign agents per cluster $= 7$; Num of malicious agents per cluster $= 3$;
- Num of data for each benign agent $= 500$;
- Num of data for each malicious agent $= 1200$;

<table>
<thead>
<tr>
<th></th>
<th>With backdoor attack (FedCBO $\alpha = 10$)</th>
<th>With backdoor attack (Bilevel FedCBO $\alpha = 20, \beta = 0.5$)</th>
<th>With backdoor attack (Bilevel FedCBO $\alpha = 10, \beta = 0.5$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg overall acc</td>
<td>$64.44 \pm 0.80 %$</td>
<td>$62.96 \pm 0.27 %$</td>
<td>$65.57 \pm 0.14 %$</td>
</tr>
<tr>
<td>Acc on class 0</td>
<td>$55.41 \pm 3.07 %$</td>
<td>$62.52 \pm 2.47 %$</td>
<td>$63.88 \pm 2.15 %$</td>
</tr>
<tr>
<td>Benign agents’ models predict images of class 0 as label 2</td>
<td>$14.96 \pm 2.87 %$</td>
<td>$7.38 \pm 1.62 %$</td>
<td>$9.01 \pm 1.91 %$</td>
</tr>
</tbody>
</table>
Future Works

1. Batched interactions.

2. Analysis of adaptive tuning of parameters.

3. Theoretical analysis of dynamics in low communication regime.
Thank you for your attention!

Special thanks to:

- NSF Grants: DMS-2005797 and DMS-2236447
- All my collaborators.
Discretized FedCBO System

FedCBO system:

\[
d\theta_{t}^{1,i} = -\lambda_1 \left( \theta_{t}^{1,i} - m_{L_1}^{\alpha} [\rho_t^N] \right) dt - \lambda_2 \nabla L_1(\theta_{t}^{1,i}) dt + \sigma_1 \left| \theta_{t}^{1,i} - m_{L_1}^{\alpha} [\rho_t^N] \right| dB_{t}^{1,i} + \sigma_2 \nabla L_1(\theta_{t}^{1,i}) dB_{t}^{1,i}
\]

\[
d\theta_{t}^{2,j} = -\lambda_1 \left( \theta_{t}^{2,j} - m_{L_2}^{\alpha} [\rho_t^N] \right) dt - \lambda_2 \nabla L_2(\theta_{t}^{2,j}) dt + \sigma_1 \left| \theta_{t}^{2,j} - m_{L_2}^{\alpha} [\rho_t^N] \right| dB_{t}^{2,j} + \sigma_2 \nabla L_2(\theta_{t}^{2,j}) dB_{t}^{2,j}
\]

Euler discretization:

\[
\theta_{n+1}^{1,i} \leftarrow \theta_{n}^{1,i} - \lambda_1 \gamma (\theta_{n}^{1,i} - m_1) - \lambda_2 \gamma \nabla L_1(\theta_{n}^{1,i}) + \sigma_1 \sqrt{\gamma} \left| \theta_{n}^{1,i} - m_1 \right| z_{n}^{1,i} + \sigma_2 \sqrt{\gamma} \left| \nabla L_1(\theta_{n}^{1,i}) \right| \tilde{z}_{n}^{1,i}
\]

\[
\theta_{n+1}^{2,j} \leftarrow \theta_{n}^{2,j} - \lambda_1 \gamma (\theta_{n}^{2,j} - m_2) - \lambda_2 \gamma \nabla L_2(\theta_{n}^{2,j}) + \sigma_1 \sqrt{\gamma} \left| \theta_{n}^{2,j} - m_2 \right| z_{n}^{2,j} + \sigma_2 \sqrt{\gamma} \left| \nabla L_2(\theta_{n}^{2,j}) \right| \tilde{z}_{n}^{2,j}
\]
Discretized FedCBO System

\[
\theta_{n+1}^{1,i} \leftarrow \theta_{n}^{1,i} - \lambda_1 \gamma (\theta_{n}^{1,i} - m_{n}^1) - \lambda_2 \gamma \nabla L_1(\theta_{n}^{1,i}) + \sigma_1 \sqrt{\gamma} \left| \theta_{n}^{1,i} - m_{n}^1 \right| z_{n}^{1,i} + \sigma_2 \sqrt{\gamma} \left| \nabla L_1(\theta_{n}^{1,i}) \right| \tilde{z}_{n}^{1,i}
\]

\[
\theta_{n+1}^{2,j} \leftarrow \theta_{n}^{2,j} - \lambda_1 \gamma (\theta_{n}^{2,j} - m_{n}^2) - \lambda_2 \gamma \nabla L_2(\theta_{n}^{2,j}) + \sigma_1 \sqrt{\gamma} \left| \theta_{n}^{2,j} - m_{n}^2 \right| z_{n}^{2,j} + \sigma_2 \sqrt{\gamma} \left| \nabla L_2(\theta_{n}^{2,j}) \right| \tilde{z}_{n}^{2,j}
\]

Remove noise terms

\[
\theta_{n+1}^{1,i} \leftarrow \theta_{n}^{1,i} - \lambda_1 \gamma (\theta_{n}^{1,i} - m_{n}^1) - \lambda_2 \gamma \nabla L_1(\theta_{n}^{1,i})
\]

\[
\theta_{n+1}^{2,j} \leftarrow \theta_{n}^{2,j} - \lambda_1 \gamma (\theta_{n}^{2,j} - m_{n}^2) - \lambda_2 \gamma \nabla L_2(\theta_{n}^{2,j})
\]
Discretized FedCBO System

\[
\theta_{n+1}^{1,i} \leftarrow \theta_n^{1,i} - \lambda_1 \gamma \left( \theta_n^{1,i} - m_n^1 \right) - \lambda_2 \gamma \nabla L_1(\theta_n^{1,i})
\]

\[
\theta_{n+1}^{2,j} \leftarrow \theta_n^{2,j} - \lambda_1 \gamma \left( \theta_n^{2,j} - m_n^2 \right) - \lambda_2 \gamma \nabla L_2(\theta_n^{2,j})
\]

Sum over \( \tau \) times

\[
\theta_{(n+1)\tau}^{1,i} \leftarrow \theta_{n\tau}^{1,i} - \lambda_1 \gamma \sum_{q=0}^{\tau-1} \left( \theta_{n\tau+q}^{1,i} - m_{n\tau+q}^1 \right) - \lambda_2 \gamma \sum_{q=0}^{\tau-1} \nabla L_1(\theta_{n\tau+q}^{1,i})
\]

\[
\theta_{(n+1)\tau}^{2,j} \leftarrow \theta_{n\tau}^{2,j} - \lambda_1 \gamma \sum_{q=0}^{\tau-1} \left( \theta_{n\tau+q}^{2,j} - m_{n\tau+q}^2 \right) - \lambda_2 \gamma \sum_{q=0}^{\tau-1} \nabla L_2(\theta_{n\tau+q}^{2,j})
\]
Discretized FedCBO System

\[
\theta_{(n+1)\tau}^{1,i} \leftarrow \theta_{n\tau}^{1,i} - \lambda_1 \gamma \sum_{q=0}^{\tau-1} (\theta_{n\tau+q}^{1,i} - m_{n\tau+q}^{1}) - \lambda_2 \gamma \sum_{q=0}^{\tau-1} \nabla L_1(\theta_{n\tau+q}^{1,i})
\]

\[
\theta_{(n+1)\tau}^{2,j} \leftarrow \theta_{n\tau}^{2,j} - \lambda_1 \gamma \sum_{q=0}^{\tau-1} (\theta_{n\tau+q}^{2,j} - m_{n\tau+q}^{1}) - \lambda_2 \gamma \sum_{q=0}^{\tau-1} \nabla L_2(\theta_{n\tau+q}^{2,j})
\]
Splitting Scheme

Step 1:

\[ \hat{\theta}_{n\tau}^{1,i} \leftarrow \theta_{n\tau}^{1,i}, \quad \hat{\theta}_{n\tau}^{2,j} \leftarrow \theta_{n\tau}^{2,j} \]

Step 2:

\[ \hat{\theta}_{n\tau+q+1}^{1,i} \leftarrow \hat{\theta}_{n\tau+q}^{1,i} - \lambda_2 \gamma \nabla L_1(\hat{\theta}_{n\tau+q}^{1,i}), \quad \hat{\theta}_{n\tau+q+1}^{2,j} \leftarrow \hat{\theta}_{n\tau+q}^{2,j} - \lambda_2 \gamma \nabla L_2(\hat{\theta}_{n\tau+q}^{2,j}) \quad \text{for } q = 0, \ldots, \tau - 1. \]

Step 3:

\[ \theta_{(n+1)\tau}^{1,i} \leftarrow \hat{\theta}_{(n+1)\tau}^{1,i} - \lambda_1 \gamma \left( \theta_{(n+1)\tau}^{1,i} - m_{(n+1)\tau}^{1} \right), \quad \theta_{(n+1)\tau}^{2,j} \leftarrow \hat{\theta}_{(n+1)\tau}^{2,j} - \lambda_1 \gamma \left( \theta_{(n+1)\tau}^{2,j} - m_{(n+1)\tau}^{2} \right) \]
FedCBO Algorithm

Algorithm 1 FedCBO

**Input:** Initialized model $\theta_0^j \in \mathbb{R}^d, j \in [N]$; Number of iterations $T$; Number of local gradient steps $\tau$; Number of models downloaded $M$; CBO system hyperparameters $\lambda_1, \lambda_2, \alpha$; Discretization step size $\gamma$; Initialized sampling likelihood $P_0 \in \mathbb{R}^{N \times (N-1)}$

1: for $n = 0, \cdots, T - 1$ do
2: \hspace{1cm} $G_n \leftarrow$ random subset of agents (participating devices);
3: \hspace{1cm} **LocalUpdate**($\theta_n^j, \tau, \lambda_2, \gamma$) for $j \in G_n$;
4: \hspace{1cm} **LocalAggregation**(agent $j$) for $j \in G_n$;
5: end for

**Output:** $\theta_T^j$ for $j \in [N]$.

- **LocalUpdate**($\theta_0, \tau, \lambda_2, \gamma$) at $j$-th agent
6: for $q = 0, \cdots, \tau - 1$ do
7: \hspace{1cm} (stochastic) gradient descent $\hat{\theta}_{q+1} \leftarrow \hat{\theta}_q - \lambda_2 \gamma \nabla L_j(\hat{\theta}_q)$;\end{quote}
8: end for
9: return $\hat{\theta}_\tau$;
values of all user parameters. This feature makes our FedCBO approach a rather decentralized approach to federated learning.

### Algorithm 1 FedCBO

**Input:**
- Initialized model $\theta^0_j \in \mathbb{R}^d$; Participating devices at $n$ iteration $G_n$; Sampling likelihood $P^j_n \in \mathbb{R}^{N-1}$; CBO system hyperparameters $\lambda_1, \alpha$; Discretization step size $\gamma$; Random sample proportion $\varepsilon \in (0, 1)$; Number of models downloaded $M$;
- $\lambda_1, \alpha$; Discretization step size $\gamma$; Random sample proportion $\varepsilon \in (0, 1)$;

1. $A_n \leftarrow \varepsilon$-greedySampling($P^j_n, G_n, M$);
2. Agent $j$ downloads models $\theta^i_n$ for $i \in A_n$;
3. Evaluate models $\theta^i_n$ on agent $j$’s data set respectively and denote the corresponding loss as $L^j_i$;
4. Calculate consensus point $m_j$ by
   \begin{equation}
   m_j \leftarrow \frac{1}{\sum_{i \in A_n} \mu^i_j} \sum_{i \in A_n} \theta^i_n \mu^i_j, \quad \text{with } \mu^i_j = \exp(-\alpha L^j_i)
   \end{equation}
5. Update agent $j$’s model by
   \begin{equation}
   \theta^j_{n+1} \leftarrow \theta^j_n - \lambda_1 \gamma (\theta^j_n - m_j),
   \end{equation}
6. Update sampling likelihood $P^j_n$ by
   \begin{equation}
   P^j_{n+1,i} \leftarrow P^j_{n,i} + (L^j_i - L^j_{\hat{i}}), \quad \text{for } i \in A_n
   \end{equation}

**Output:** $\theta^j_{n+1}, P^j_{n+1}$

### Algorithm 2 LocalAggregation(agent $j$)

**Input:** Agent $j$’s model $\theta^j_n \in \mathbb{R}^d$; Participating devices at $n$ iteration $G_n$; Sampling likelihood $P^j_n \in \mathbb{R}^{N-1}$; CBO system hyperparameters $\lambda_1, \alpha$; Discretization step size $\gamma$; Random sample proportion $\varepsilon \in (0, 1)$; Number of models downloaded $M$;

1. $A_n \leftarrow \varepsilon$-greedySampling($P^j_n, G_n, M$);
2. Agent $j$ downloads models $\theta^i_n$ for $i \in A_n$;
3. Evaluate models $\theta^i_n$ on agent $j$’s data set respectively and denote the corresponding loss as $L^j_i$;
4. Calculate consensus point $m_j$ by
   \begin{equation}
   m_j \leftarrow \frac{1}{\sum_{i \in A_n} \mu^i_j} \sum_{i \in A_n} \theta^i_n \mu^i_j, \quad \text{with } \mu^i_j = \exp(-\alpha L^j_i)
   \end{equation}
5. Update agent $j$’s model by
   \begin{equation}
   \theta^j_{n+1} \leftarrow \theta^j_n - \lambda_1 \gamma (\theta^j_n - m_j),
   \end{equation}
6. Update sampling likelihood $P^j_n$ by
   \begin{equation}
   P^j_{n+1,i} \leftarrow P^j_{n,i} + (L^j_i - L^j_{\hat{i}}), \quad \text{for } i \in A_n
   \end{equation}

**Output:** $\theta^j_{n+1}, P^j_{n+1}$

$\varepsilon$-greedySampling($P^j_n, G_n, M$)

7. Randomly sample $\varepsilon \ast M$ number of agents from $G_n$, denoted as $A^1_n$;
8. Select $(1 - \varepsilon) \ast M$ numbers of agents in $G_n \backslash A^1_n$ with top value $P^j_{n,i}, i \in G_n \backslash A^1_n$, denoted as $A^2_n$;
9. return $A_n = A^1_n \cup A^2_n$