Federated Ensemble-Directed Offline Reinforcement Learning Algorithm (FEDORA)

Seonvin Cho

Information and Intelligence Systems Lab.

Electronic Engineering, Hanyang University

February 18, 2025

FEDERATED OFFLINE REINFORCEMENT LEARNING

- **Goal**
  - To learn the optimal policy using only offline data from the operational policies of **multiple clients** with **different levels of expertise**
  - **Without** the clients knowing the **quality** of their data, or sharing it with one another or the server
- **Challenges**
  - Ensemble Heterogeneity: Learn policies of varying quality
  - Pessimistic Value Computation: Q-value underestimation due to limited client datasets
  - Data Heterogeneity: Varying data quality

$\Longrightarrow$ **Federated Ensemble-Directed Offline RL Algorithm(FEDORA)**

Related Work

- **Federated Learning**
  - To minimize $F(\theta) = \mathbb{E}_{i \sim P}[F_i(\theta)]$.
  - **FedAvg algorithm**: $\theta^{t+1} = \sum_{i=1}^{|N|} \omega_i \theta_i^t$, where $\omega_i = \frac{|D_i|}{\sum_{j=1}^{|N|} |D_j|}$.
- **Reinforcement Learning**
  - To maximize $J(\pi) = \mathbb{E}_{\pi,P,\mu}[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)]$.
- **Offline Reinforcement Learning**
  - To learn π only using a static dataset by π_b without any additional interactions with the environment.
  - Utilize the regularization to prevent distribution shift.

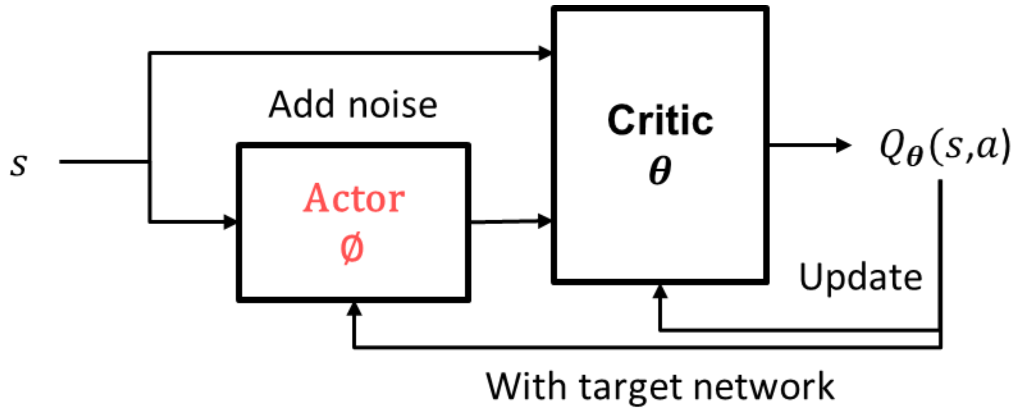
Related Work

- **Federated Learning**
  - To minimize $F(\theta) = \mathbb{E}_{i \sim P}[F_i(\theta)]$.
  - **FedAvg algorithm**: $\theta^{t+1} = \sum_{i=1}^{|N|} \omega_i \theta_i^t$, where $\omega_i = \frac{|D_i|}{\sum_{j=1}^{|N|} |D_j|}$.
- **Reinforcement Learning**
  - To maximize $J(\pi) = \mathbb{E}_{\pi,P,\mu}[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)]$.
- **Offline Reinforcement**
  - To learn π only using a stat
  - Utilize the regularization to prevent distribution shift.
  - **TD3-BC**(Twin Delayed DDPG-Behavior Cloning): To prevent distribution shift.

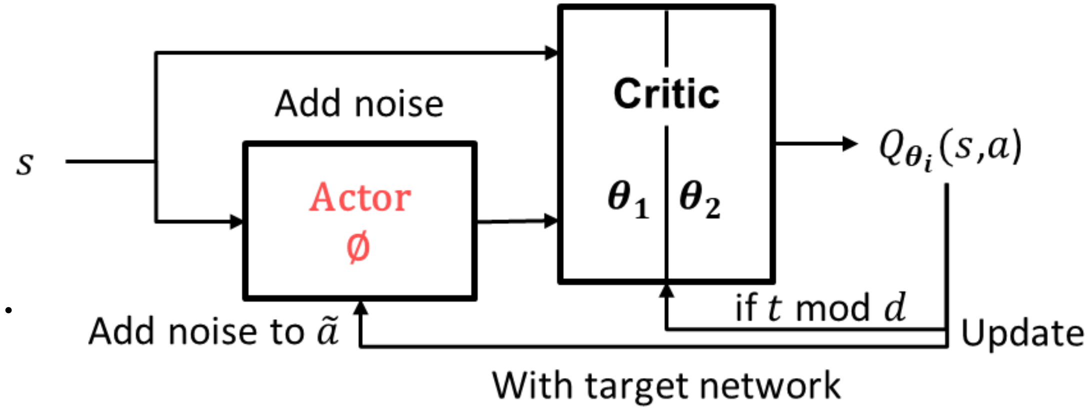
- Each client learns using its own dataset under specific behavior policies.
- The learned policy varies depending on the behavior policy.
- Simply aggregate all client models degrades performance.

TD3-BC

- DDPG
  - Overestimate Q-values in critic.
  - Update actor at every step → instability in Q-value.
  - Change target Q-values too rapidly → unstable.



- TD3(Twin Delayed DDPG)
  - Utilize twin Q-networks($Q_{\theta_1}, Q_{\theta_2}$) and update minimum.
  - Delay updating actor compared to critic.
  - Smooth target policy with adding gaussian noise to action when computing target Q-value.



TD3-BC

- DDPG
  - Overestimate Q-values in critic.
  - Update actor at every step → instability in Q-value.
  - Change target Q-values too rapidly → unstable.

- TD3(Twin Delayed DDPG)
  - Utilize twin Q-networks($Q_{\theta_1}, Q_{\theta_2}$) and update minimum.
  - Delay updating actor compared to critic.
  - Smooth target policy with adding gaussian noise to action when computing target Q-value.

DDPG
- $y \leftarrow r + \gamma Q_{\boldsymbol{\theta}'}(s', \tilde{a})$
- Update $\emptyset$ (w.r.t. actor policy $\pi_\emptyset$) at every t
- $\tilde{a} \leftarrow \pi_{\emptyset'}(s')$

TD3
- $y \leftarrow r + \gamma min_{i=1,2} Q_{\theta_i'}(s', \tilde{a})$
- Update $\emptyset$ (w.r.t. actor policy $\pi_\emptyset$) if t mod d
- $\tilde{a} \leftarrow \pi_{\emptyset'}(s') + \epsilon$ where $\epsilon \sim \mathrm{clip}(\mathcal{N}(0, \tilde{\sigma}). -c. c)$

TD3-BC

- TD3 Problems
  - Not suitable for offline RL. (∵requires exploration through interactions with the environment)
  - The actor may select actions that deviate from the original data distribution in pursuit of optimal Q.
- TD3-BC
  - Add BC regulation term
    - To favor actions contained in the dataset $\mathcal{D}$
    - To use only original data without exploration

$$\pi \leftarrow argmax_{\pi} \mathbb{E}_{s\sim\mathcal{D}}[\lambda Q\big(s, \pi(s)\big) - (\pi(s) - a)^2]$$

  - TD3 term: $Q\big(s, \pi(s)\big)$ for maximize Q-value.
  - BC term: $-(\pi(s) - a)^2$ for reducing the difference between action and policy.

FEDORA

- **Solution**
  - Ensemble Heterogeneity → Ensemble-directed learning to weigh client contribution.
    - Weights ~ entropy regularization $\omega_i = \frac{e^{\beta J_{i|D_i|}}}{\sum_j e^{\beta J_{j|D_j|}}}$ where $J_{\mathrm{i}}^t = \mathbb{E}_{s\sim D_i}[Q_i^t(s, \pi_{\mathrm{i}}^t(\mathrm{s}))]$.
    - Federated policy ~ weighted combination of client policies $\pi_{fed}^{t+1} = \sum_i \omega_i \pi_i^t$.
  - Pessimistic Value Computation → Federated optimism for critic training.
    - Ensemble-directed Federation → Optimistic target $\tilde{Q}_i^{(t,k)}(s,a) = \max(Q_i^{(t,k)}(s,a), Q_{fed}^t(s,a))$.
  - Data Heterogeneity → Proximal policy update.
    - $\pi_{\mathrm{i}}^{t,\mathrm{k}+1} = \mathrm{argmin}_{\pi} \mathcal{L}_{actor}(\pi)$ where $\mathcal{L}_{actor}(\pi) = \mathcal{L}_{\mathrm{local}}(\pi) + \mathbb{E}_{(\mathrm{s,a})\sim D_i}[(\pi(\mathrm{s}) - \pi_{fed}^{t+1})^2]$,
    - $\mathcal{L}_{\mathrm{local}}(\pi) = \mathbb{E}_{(\mathrm{s,a})\sim D_i}[-Q_{\mathrm{i}}^{(t,k)}(\mathrm{s}, \pi(\mathrm{s})) + (\pi(\mathrm{s}) - \mathrm{a})^2]$

FEDORA

- **Solution**
  - Ensemble Heterogeneity → Ensemble-directed learning to weigh client contribution.
    - Weights ~ entropy regularization $\omega_i = \frac{e^{\beta J_i |D_i|}}{\sum_j e^{\beta J_j |D_j|}}$ where $J_\mathrm{i}^t = \mathbb{E}_{s \sim D_i}[Q_i^t(s, \pi_\mathrm{i}^t(\mathrm{s}))]$.
    - Federated policy ~ weighted combination of client policies $\pi_{fed}^{t+1} = \sum_i \omega_i \pi_i^t$.
  - Pessimistic Value Computation → Federated optimism for critic training.
    - Ensemble-directed Federation → Optimistic target $\tilde{Q}_i^{(t,k)}(s,a) = \max(Q_i^{(t,k)}(s,a), Q_{fed}^t(s,a))$.
  - Data Heterogeneity → Proximal policy update.
    - $\pi_\mathrm{i}^{t,\mathrm{k}+1} = \mathrm{argmin}_\pi \mathcal{L}_{actor}($

      TD3-BC: $\pi \leftarrow argmax_\pi \mathbb{E}_{s \sim \mathcal{D}}[\lambda Q(s, \pi(s)) - (\pi(s) - a)^2]$

      FEDORA: $\pi \leftarrow argmin_\pi \mathbb{E}_{(s,\mathrm{a}) \sim \mathcal{D}}[-\lambda Q(s, \pi(s)) + (\pi(s) - a)^2]$
    - $\mathcal{L}_{\mathrm{local}}(\pi) = \mathbb{E}_{(\mathrm{s,a}) \sim D_i}[-Q_\mathrm{i}^{(t,k)}(\mathrm{s}, \pi(\mathrm{s})) + (\pi(\mathrm{s}) - \mathrm{a})^2]$

FEDORA

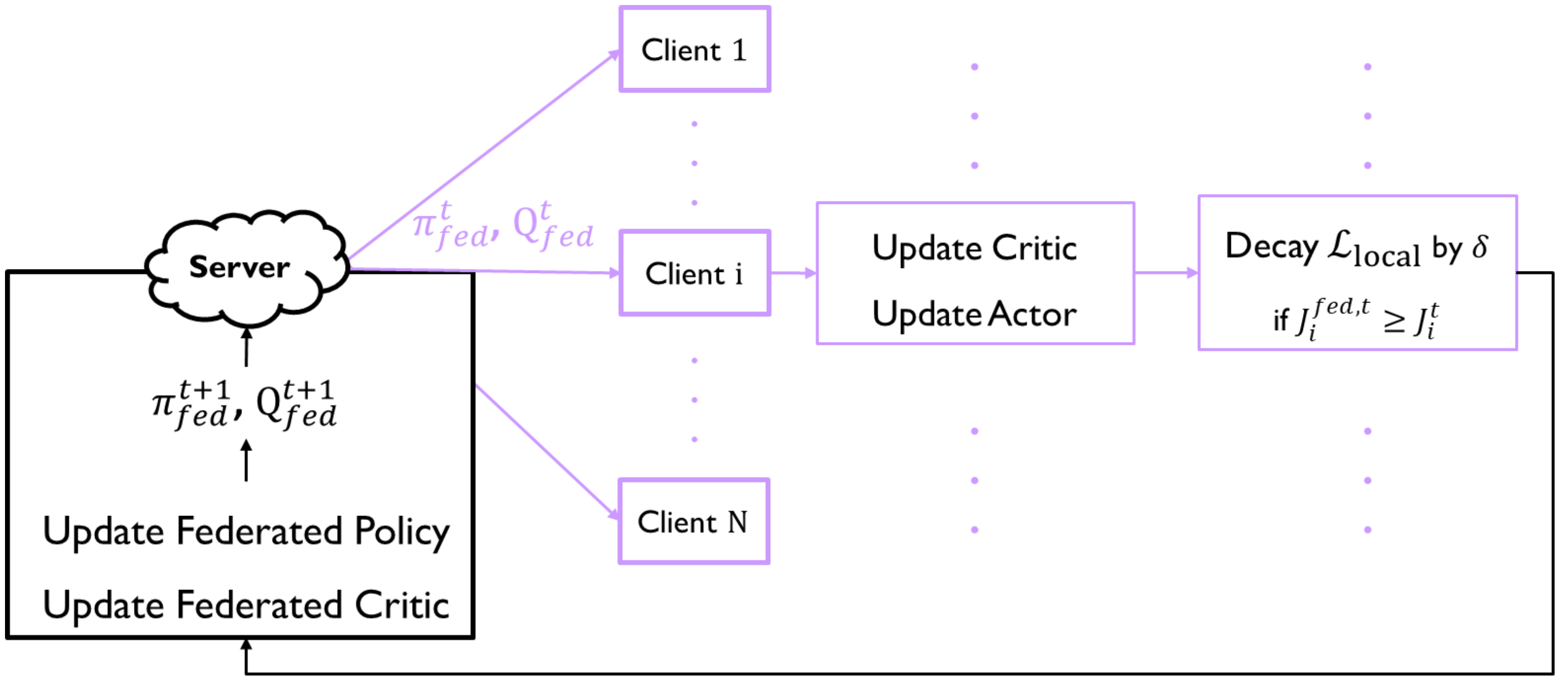
Eq.(8) $\omega_i^t = \frac{e^{\beta J_i^t}|D_i|}{\sum_{j=1}^{|N|} e^{\beta J_j^t}|D_j|}$, $\pi_{fed}^{t+1} = \sum_{i=1}^{|N|} \omega_i^t \pi_i^t$

Eq.(9) $Q_{fed}^{t+1} = \sum_i \omega_i^t Q_i^t$

Eq.(10) $Q_i^{(t,k+1)} = argmin_Q \mathbb{E}_{(s,a,r,s')\sim D_i}[(r + \gamma \tilde{Q}_i^{(t,k)}(s',a') - Q(s,a))^2]$

Eq.(11) $\mathcal{L}_{actor}(\pi) = \mathcal{L}_{\text{local}}(\pi) + \mathbb{E}_{(s,a)\sim D_i}[(\pi(s) - \pi_{fed}^{t+1})^2]$,

$$\pi_i^{t,k+1} = \text{argmin}_\pi \mathcal{L}_{actor}(\pi)$$



Algorithm 1 Outline of Client i's Algorithm

function train_client(π_{fed}^t, Q_{fed}^t)
  $\pi_i^{(t,0)} = \pi_{\text{fed}}^t$, $Q_i^{(t,0)} = Q_{\text{fed}}^t$
  for $1 \leq k < K$ **do**
    Update Critic by one gradient step w.r.t. Eq. (10)
    Update Actor by one gradient step w.r.t. Eq. (11)
  end for
  Decay $\mathcal{L}_{\text{local}}$ by δ if $J_i^{\text{fed},t} \geq J_i^t$
end function

Algorithm 2 Outline of Server Algorithm

Initialize $\pi_{\text{fed}}^1, Q_{\text{fed}}^1$
for $t \in 1 \ldots$ **do**
  Send π_{fed}^t and Q_{fed}^t to $i \in \mathcal{N}$
  Sample $\mathcal{N}_t \subset \mathcal{N}$
  for $i \in \mathcal{N}_t$ **do**
    i.train_client (π_{fed}^t, Q_{fed}^t) (Client side)
  end for
  Compute π_{fed}^{t+1} and Q_{fed}^{t+1} for clients in $\mathcal{N}_t$ using Eq. (8) and (9) respectively.
end for

FEDORA OVERVIEW

Server

$Initialize\ \pi_{fed}^1, \mathrm{Q}_{fed}^1$

$for\ t \in 1, \cdots (\#\ of\ round)\ do$

$\quad Send\ \pi_{fed}^t, \mathrm{Q}_{fed}^t\ to\ client\ \mathrm{i} \in N$

$\quad Sample\ N_t \subset N$

$\quad for\ i\ \in\ N_t\ do$

$\quad\quad train\ 'i'\ th\ client$

$\quad endfor$

$\quad Compute\ \pi_{fed}^{t+1}, \mathrm{Q}_{fed}^{t+1}\ for\ clients\ in\ N_t$

Ensemble Federation

$$\pi_{fed}^{t+1} = \sum_{i=1}^{|N|} \omega_i^t \pi_i^t$$

$$Q_{fed}^{t+1} = \sum_i \omega_i^t Q_i^t$$

$endfor$

π_{fed}^t Q_{fed}^t

ω_i^t π_i^t Q_i^t

Client

$\pi_i^{(t,0)} = \pi_{fed}^t, Q_i^{(t,0)} = \mathrm{Q}_{fed}^t$

$for\ 1 \leq k < K\ \mathrm{do}$

Federated Optimism

$$\tilde{Q}_i^{(t,k)}(s,a) = \max(Q_i^{(t,k)}(s,a), Q_{fed}^t(s,a))$$

$$Q_i^{(t,k+1)} = argmin_Q \mathbb{E}_{(\mathrm{s,a},r,s\prime)\sim D_i}[(r + \gamma \tilde{Q}_\mathrm{i}^{(t,k)}(s',a') - Q(s,a))^2]$$

Proximal Policy

$$\mathcal{L}_{\mathrm{local}}(\pi) = \mathbb{E}_{(\mathrm{s,a})\sim D_i}[-Q_\mathrm{i}^{(t,k)}(\mathrm{s},\pi(\mathrm{s})) + (\pi(\mathrm{s}) - \mathrm{a})^2]$$

$$\mathcal{L}_{actor}(\pi) = \mathcal{L}_{\mathrm{local}}(\pi) + \mathbb{E}_{(\mathrm{s,a})\sim D_i}[(\pi(\mathrm{s}) - \pi_{fed}^{t+1})^2]$$

$$\pi_\mathrm{i}^{t,\mathrm{k}+1} = \mathrm{argmin}_\pi \mathcal{L}_{actor}(\pi)$$

$endfor$

$$J_i^{fed,t} = \mathbb{E}_{\mathrm{s}\sim D_i}[Q_{fed}^t(s,\pi_{fed}^t)], J_\mathrm{i}^t = \mathbb{E}_{\mathrm{s}\sim D_i}[Q_i^t(s,\pi_\mathrm{i}^t(\mathrm{s}))]$$

$Decay\ \mathcal{L}_{\mathrm{local}}\ by\ \delta\ if\ J_i^{fed,t} \geq J_i^t$

$$\omega_i^t = \frac{e^{\beta J_i^t}|D_i|}{\sum_{j=1}^{|N|} e^{\beta J_j^t}|D_j|}$$

FEDORA OVERVIEW

Server

Initialize $\pi^{1}_{fed}, Q^{1}_{fed}$

for $t \in 1, \cdots$ *(# of round) do*

 Send $\pi^{t}_{fed}, Q^{t}_{fed}$ *to client* $i \in N$

 Sample $N_t \subset N$

 for $i \in N_t$ *do*

  train 'i' th client

 endfor

 Compute $\pi^{t+1}_{fed}, Q^{t+1}_{fed}$ *for clients in* N_t

  $\pi^{t+1}_{fed} = \sum_{i=1}^{|N|} \omega_i^t \pi_i^t$

  $Q^{t+1}_{fed} = \sum_i \omega_i^t Q_i^t$

endfor

π^{t}_{fed}
Q^{t}_{fed}

ω_i^t
π_i^t
Q_i^t

Pessimistic Value Computation → Federated optimism for critic training.

› Optimistic target $\tilde{Q}_i^{(t,k)}(s,a) = \max(Q_i^{(t,k)}(s,a), Q^t_{fed}(s,a))$.

Client

$\pi_i^{(t,0)} = \pi^t_{fed}, Q_i^{(t,0)} = Q^t_{fed}$

for $1 \le k < K$ do

Update Critic

 $\tilde{Q}_i^{(t,k)}(s,a) = \max(Q_i^{(t,k)}(s,a), Q^t_{fed}(s,a))$ Federated Optimism

 $Q_i^{(t,k+1)} = argmin_Q \mathbb{E}_{(s,a,r,s')\sim D_i}[(r + \gamma \tilde{Q}_i^{(t,k)}(s',a') - Q(s,a))^2]$

 $\mathcal{L}_{local}(\pi) = \mathbb{E}_{(s,a)\sim D_i}[-Q_i^{(t,k)}(s,\pi(s)) + (\pi(s) - a)^2]$

 $\mathcal{L}_{actor}(\pi) = \mathcal{L}_{local}(\pi) + \mathbb{E}_{(s,a)\sim D_i}[(\pi(s) - \pi^{t+1}_{fed})^2]$

 $\pi_i^{t,k+1} = \mathrm{argmin}_\pi \mathcal{L}_{actor}(\pi)$

endfor

$J_i^{fed,t} = \mathbb{E}_{s\sim D_i}[Q^t_{fed}(s, \pi^t_{fed})], J_i^t = \mathbb{E}_{s\sim D_i}[Q_i^t(s, \pi_i^t(s))]$

Decay $\mathcal{L}_{local}$ *by* δ *if* $J_i^{fed,t} \ge J_i^t$

$$\omega_i^t = \frac{e^{\beta J_i^t}|D_i|}{\sum_{j=1}^{|N|} e^{\beta J_j^t}|D_j|}$$

FEDORA OVERVIEW

Server

$Initialize\ \pi_{fed}^{1}, Q_{fed}^{1}$

$for\ t \in 1, \cdots (\#\ of\ round)\ do$

$Send\ \pi_{fed}^{t}, Q_{fed}^{t}\ to\ client\ i \in N$

$Sample\ N_t \subset N$

$for\ i\ \in\ N_t\ do$

$train\ 'i'\ th\ client$

$endfor$

$Compute\ \pi_{fed}^{t+1}, Q_{fed}^{t+1}\ for\ clients\ in\ N_t$

$$\pi_{fed}^{t+1} = \sum_{i=1}^{|N|} \omega_i^t \pi_i^t$$

$$Q_{fed}^{t+1} = \sum_i \omega_i^t Q_i^t$$

$endfor$

π_{fed}^{t} Q_{fed}^{t}

ω_i^t π_i^t Q_i^t

Client

$\pi_i^{(t,0)} = \pi_{fed}^{t}, Q_i^{(t,0)} = Q_{fed}^{t}$

$for\ 1 \le k < K\ do$

$$\tilde{Q}_i^{(t,k)}(s,a) = \max(Q_i^{(t,k)}(s,a), Q_{fed}^{t}(s,a))$$

$$Q_i^{(t,k+1)} = argmin_Q \mathbb{E}_{(s,a,r,s')\sim D_i}[(r + \gamma \tilde{Q}_i^{(t,k)}(s',a') - Q(s,a))^2]$$

Regularize local offline data

$$\mathcal{L}_{local}(\pi) = \mathbb{E}_{(s,a)\sim D_i}[-Q_i^{(t,k)}(s,\pi(s)) + (\pi(s) - a)^2]$$

$$\mathcal{L}_{actor}(\pi) = \mathcal{L}_{local}(\pi) + \mathbb{E}_{(s,a)\sim D_i}[(\pi(s) - \pi_{fed}^{t+1})^2]$$

$$\pi_i^{t,k+1} = \mathrm{argmin}_\pi \mathcal{L}_{actor}(\pi)$$

Regularize federated policy

$endfor$

$$J_i^{fed,t} = \mathbb{E}_{s\sim D_i}[Q_{fed}^{t}(s, \pi_{fed}^{t})], J_i^t = \mathbb{E}_{s\sim D_i}[Q_i^t(s, \pi_i^t(s))]$$

Update Actor

$Decay\ \mathcal{L}_{local}\ by\ \delta\ if\ J_i^{fed,t} \ge J_i^t$

Proximal Policy Update

$$\omega_i^t = \frac{e^{\beta J_i^t}|D_i|}{\sum_{j=1}^{|N|} e^{\beta J_j^t}|D_j|}$$

Data Heterogeneity → Proximal policy update.

- $\pi_i^{t,k+1} = \mathrm{argmin}_\pi \mathcal{L}_{actor}(\pi)$ where $\mathcal{L}_{actor}(\pi) = \mathcal{L}_{local}(\pi) + \mathbb{E}_{(s,a)\sim D_i}[(\pi(s)\text{-}\pi_{fed}^{t+1})^2]$,
- $\mathcal{L}_{local}(\pi) = \mathbb{E}_{(s,a)\sim D_i}[-Q_i^{(t,k)}(s,\pi(s)) + (\pi(s) - a)^2]$

FEDORA Overview

Server

$Initialize\ \pi^1_{fed}, Q^1_{fed}$

$for\ t \in 1, \cdots (\#\ of\ round)\ do$

$\quad Send\ \pi^t_{fed}, Q^t_{fed}\ to\ client\ i \in N$

$\quad Sample\ N_t \subset N$

$\quad for\ i\ \in\ N_t\ do$

$\quad\quad train\ 'i'\ th\ client$

$\quad endfor$

$\quad Compute\ \pi^{t+1}_{fed}, Q^{t+1}_{fed}\ for\ clients\ in\ N_t$

$$\pi^{t+1}_{fed} = \sum_{i=1}^{|N|} \omega^t_i \pi^t_i$$

$$Q^{t+1}_{fed} = \sum_i \omega^t_i Q^t_i$$

$endfor$

π^t_{fed}

Q^t_{fed}

ω^t_i

π^t_i

Q^t_i

Client

$$\pi_i^{(t,0)} = \pi^t_{fed}, Q_i^{(t,0)} = Q^t_{fed}$$

$for\ 1 \le k < K$ do

$$\tilde{Q}_i^{(t,k)}(s,a) = \max(Q_i^{(t,k)}(s,a), Q^t_{fed}(s,a))$$

$$Q_i^{(t,k+1)} = argmin_Q \mathbb{E}_{(s,a,r,s')\sim D_i}[(r + \gamma \tilde{Q}_i^{(t,k)}(s',a') - Q(s,a))^2]$$

$$\mathcal{L}_{local}(\pi) = \mathbb{E}_{(s,a)\sim D_i}[-Q_i^{(t,k)}(s,\pi(s)) + (\pi(s) - a)^2]$$

$$\mathcal{L}_{actor}(\pi) = \mathcal{L}_{local}(\pi) + \mathbb{E}_{(s,a)\sim D_i}[(\pi(s) - \pi^{t+1}_{fed})^2]$$

$$\pi_i^{t,k+1} = \mathrm{argmin}_\pi \mathcal{L}_{actor}(\pi)$$

$endfor$

Local estimate

$$J_i^{fed,t} = \mathbb{E}_{s\sim D_i}[Q^t_{fed}(s,\pi^t_{fed})], J^t_i = \mathbb{E}_{s\sim D_i}[Q^t_i(s,\pi^t_i(s))]$$

$Decay\ \mathcal{L}_{local}\ by\ \delta\ if\ J_i^{fed,t} \ge J_i^t$

Updated critic

$$\omega^t_i = \frac{e^{\beta J^t_i}|D_i|}{\sum_{j=1}^{|N|} e^{\beta J^t_j}|D_j|}$$

Decay the influence of local data

FEDORA OVERVIEW

Server

$Initialize\ \pi_{fed}^{1}, Q_{fed}^{1}$

$for\ t \in 1, \cdots (\#\ of\ round)\ do$

$\quad Send\ \pi_{fed}^{t}, Q_{fed}^{t}\ to\ client\ i \in N$

$\quad Sample\ N_t \subset N$

$\quad for\ i\ \in\ N_t\ do$

$\quad\quad train\ 'i'\ th\ client$

$\quad endfor$

$\quad Compute\ \pi_{fed}^{t+1}, Q_{fed}^{t+1}\ for\ clients\ in\ N_t$

Update Policy $\quad \pi_{fed}^{t+1} = \sum_{i=1}^{|N|} \omega_i^t \pi_i^t$

Update Critic $\quad Q_{fed}^{t+1} = \sum_i \omega_i^t Q_i^t$

$endfor$

π_{fed}^{t}
Q_{fed}^{t}

ω_i^t
π_i^t
Q_i^t

Client

$\pi_i^{(t,0)} = \pi_{fed}^{t}, Q_i^{(t,0)} = Q_{fed}^{t}$

$for\ 1 \le k < K$ do

$$\tilde{Q}_i^{(t,k)}(s,a) = \max(Q_i^{(t,k)}(s,a), Q_{fed}^{t}(s,a))$$

$$Q_i^{(t,k+1)} = argmin_Q \mathbb{E}_{(s,a,r,s')\sim D_i}[(r + \gamma \tilde{Q}_i^{(t,k)}(s',a') - Q(s,a))^2]$$

$$\mathcal{L}_{local}(\pi) = \mathbb{E}_{(s,a)\sim D_i}[-Q_i^{(t,k)}(s,\pi(s)) + (\pi(s) - a)^2]$$

$$\mathcal{L}_{actor}(\pi) = \mathcal{L}_{local}(\pi) + \mathbb{E}_{(s,a)\sim D_i}[(\pi(s) - \pi_{fed}^{t+1})^2]$$

$$\pi_i^{t,k+1} = \mathrm{argmin}_\pi \mathcal{L}_{actor}(\pi)$$

$endfor$

Local estimate

$$J_i^{fed,t} = \mathbb{E}_{s\sim D_i}[Q_{fed}^t(s, \pi_{fed}^t)], J_i^t = \mathbb{E}_{s\sim D_i}[Q_i^t(s, \pi_i^t(s))]$$

$Decay\ \mathcal{L}_{local}\ by\ \delta\ if\ J_i^{fed,t} \ge J_i^t$

Updated critic

$$\omega_i^t = \frac{e^{\beta J_i^t |D_i|}}{\sum_{j=1}^{|N|} e^{\beta J_j^t |D_j|}}$$

Ensemble-Directed Federated Learning

Ensemble Heterogeneity → Ensemble-directed learning to weigh client contribution.

- Weights ~ entropy regularization $\omega_i = \frac{e^{\beta J_i |D_i|}}{\sum_j e^{\beta J_j |D_j|}}$ where $J_i^t = \mathbb{E}_{s\sim D_i}[Q_i^t(s, \pi_i^t(s))]$.
- Federated policy ~ weighted combination of client policies $\pi_{fed}^{t+1} = \sum_i \omega_i \pi_i^t$.

Pessimistic Value Computation → Federated optimism for critic training.

- Optimistic target $\tilde{Q}_i^{(t,k)}(s,a) = \max(Q_i^{(t,k)}(s,a), Q_{fed}^t(s,a))$.

FEDORA Overview

Server

$Initialize\ \pi_{fed}^1, \mathrm{Q}_{fed}^1$

$for\ t \in 1, \cdots\ (\#\ of\ round)\ do$

$\quad Send\ \pi_{fed}^t, \mathrm{Q}_{fed}^t\ to\ client\ \mathrm{i} \in N$

$\quad Sample\ N_t \subset N$

$\quad for\ i\ \in\ N_t\ do$

$\quad\quad train\ 'i'\ th\ client$

$\quad endfor$

$\quad Compute\ \pi_{fed}^{t+1}, \mathrm{Q}_{fed}^{t+1}\ for\ clients\ in\ N_t$

$$\pi_{fed}^{t+1} = \sum_{i=1}^{|N|} \omega_i^t \pi_i^t$$

$$Q_{fed}^{t+1} = \sum_i \omega_i^t Q_i^t$$

$endfor$

π_{fed}^t

Q_{fed}^t

ω_i^t

π_i^t

Q_i^t

Client

$\pi_i^{(t,0)} = \pi_{fed}^t, Q_i^{(t,0)} = \mathrm{Q}_{fed}^t$

$for\ 1 \le k < K\ \mathrm{do}$

$$\tilde{Q}_i^{(t,k)}(s,a) = \max(Q_i^{(t,k)}(s,a), Q_{fed}^t(s,a))$$

$$Q_i^{(t,k+1)} = argmin_Q \mathbb{E}_{(\mathrm{s,a},r,s')\sim D_i}[(r + \gamma \tilde{Q}_{\mathrm{i}}^{(t,k)}(s',a') - Q(s,a))^2]$$

$$\mathcal{L}_{\mathrm{local}}(\pi) = \mathbb{E}_{(\mathrm{s,a})\sim D_i}[-Q_{\mathrm{i}}^{(t,k)}(\mathrm{s}, \pi(\mathrm{s})) + (\pi(\mathrm{s}) - \mathrm{a})^2]$$

$$\mathcal{L}_{actor}(\pi) = \mathcal{L}_{\mathrm{local}}(\pi) + \mathbb{E}_{(\mathrm{s,a})\sim D_i}[(\pi(\mathrm{s}) - \pi_{fed}^{t+1})^2]$$

$$\pi_{\mathrm{i}}^{t,\mathrm{k}+1} = \mathrm{argmin}_\pi \mathcal{L}_{actor}(\pi)$$

$endfor$

$J_i^{fed,t} = \mathbb{E}_{\mathrm{s}\sim D_i}[Q_{fed}^t(s, \pi_{fed}^t)], J_{\mathrm{i}}^t = \mathbb{E}_{\mathrm{s}\sim D_i}[Q_i^t(s, \pi_{\mathrm{i}}^t(\mathrm{s}))]$

$Decay\ \mathcal{L}_{\mathrm{local}}\ by\ \delta\ if\ J_i^{fed,t} \ge J_i^t$

$$\omega_i^t = \frac{e^{\beta J_i^t} |D_i|}{\sum_{j=1}^{|N|} e^{\beta J_j^t} |D_j|}$$