FEDONE: QUERY-EFFICIENT FEDERATED LEARNING FOR BLACK-BOX DISCRETE PROMPT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Black-Box Discrete Prompt Learning (BDPL) is a prompt-tuning method that optimizes discrete prompts without accessing model parameters or gradients, making the prompt tuning on a cloud-based Large Language Model (LLM) feasible. Adapting Federated Learning (FL) to BDPL could further enhance prompt tuning performance by leveraging data from diverse sources. However, all previous research on federated black-box prompt tuning had neglected the substantial query cost associated with the cloud-based LLM service. To address this gap, we conducted a theoretical analysis of query efficiency within the context of federated black-box prompt tuning. Our findings revealed that degrading FedAvg to activate only one client per round, a strategy we called *FedOne*, enabled optimal query efficiency in federated black-box prompt learning. Building on this insight, we proposed the FedOne framework, a federated black-box discrete prompt learning method designed to maximize query efficiency when interacting with cloud-based LLMs. We conducted numerical experiments on various aspects of our framework, demonstrating a significant improvement in query efficiency, which aligns with our theoretical results.

1 INTRODUCTION

Prompt tuning has emerged as a vital technique for adapting large language models (LLMs) (Liu et al., 2019; Brown et al., 2020) to specific tasks without retraining the entire model. Traditionally, many tuning methods require access to the model's intermediate representations (Brown et al., 2020; Li & Liang, 2021; Liu et al., 2021; Lester et al., 2021), categorizing them as white-box approaches. However, when such access is unavailable, black-box prompt tuning becomes essential. This approach focuses on tuning the input prompts without access to the internal processes of the model (Sun et al., 2022; Diao et al., 2022; Deng et al., 2022; Xiao et al., 2023).

Federated learning (FL) (McMahan et al., 2017; Li et al., 2020; 2021; Karimireddy et al., 2020; Mishchenko et al., 2022) has emerged as a promising approach for leveraging decentralized data from multiple clients while preserving privacy. Applying federated learning to prompt tuning offers a valuable opportunity to improve client capabilities and enhance model performance. Most federated prompt tuning methods to date have focused on white-box scenarios where clients have access to model parameters (Zhao et al., 2023; Zhang et al., 2023). In those approaches, only the trainable parameters (prompts) are trained by the client and shared with the server, significantly reducing both the number of trainable parameters and the communication costs compared to fine-tuning baselines.

However, several practical limitations hinder the applicability of federated prompt tuning that relies on white-box access. First, white-box prompt learning is not applicable to closed-source LLM that are accessed via APIs, as these models are not openly shared. In such scenarios, users are restricted to interacting with the model through the API endpoints without access to the internal structure or weights of the LLM. This limitation prevents the application of white-box prompt learning techniques. Second, FL is typically applied in scenarios involving thousands of edge devices, each with limited computational resources. However, white-box prompt learning demands substantial computational power, as it requires devices to perform computation on the entire LLM. These operations are computationally intensive, rendering them impractical for edge devices with constrained capabilities. This presents a significant challenge when trying to implement white-box prompt learning in FL environments, as it can lead to excessive resource demands on the participating devices.

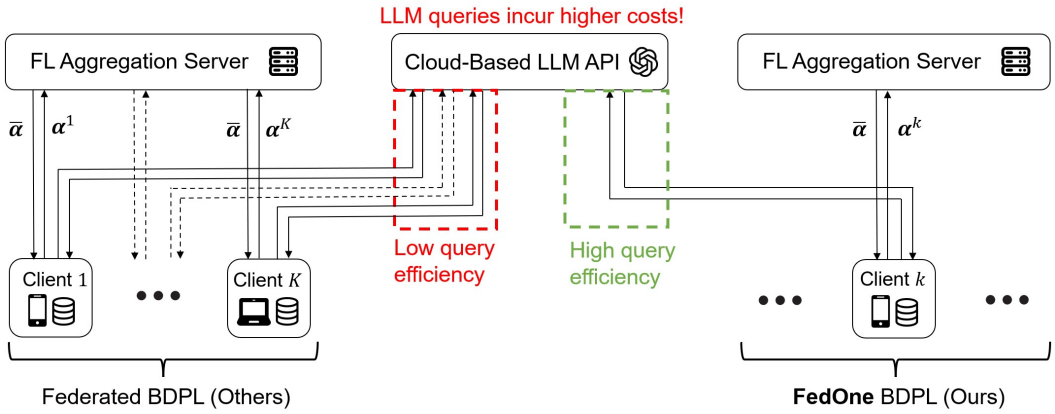

Figure 1: Query-Efficient Fed-BDPL

In contrast, applying black-box discrete prompt learning to federated learning offers several distinct benefits (Zhao et al., 2023; Lin et al., 2023; Zhang et al., 2023; Che et al., 2023), enhancing both the practicality and effectiveness. First, this approach preserves the privacy of closed-source LLMs by not requiring access to their internal model weights or architecture. For example, Diao et al. (2022); Lin et al. (2023) introduces a black-box prompt learning method that uses only discrete prompts as inputs and relies solely on the model's output loss to train the prompts. Besides, black-box prompt tuning reduces the computational burden on the client, as it eliminates the need for computation on the entire model, thereby enabling participation from edge devices with limited computational resources. Moreover, the communication costs associated with discrete prompts are lower compared to white-box prompt tuning methods, as the white-box prompt tuning require transmitting large matrices of numerical values.

Despite these advantages, the application of BDPL in FL still faces two significant unresolved challenges, tempering its overall promise. First, previous research (Lin et al., 2023; Sun et al., 2023) on federated black-box prompt tuning has neglected the substantial cost associated with queries to the LLM cloud service (Figure 1, left). Second, a convergence analysis for Federated BDPL aimed at optimizing discrete prompts has not yet been provided.

Targeting the above problems, we introduce a novel federated learning framework, FedOne, designed to optimize query efficiency in Federated BDPL. We offer the first convergence analysis of Fed-BDPL in this context and further extend the analysis of the query efficiency towards the cloud-based LLM server. The results demonstrate that by limiting activation to a single client per round, FedOne achieves optimal query efficiency (Figure 1, right). Our approach is particularly well-suited for scenarios involving limited computational resources, such as mobile devices or IoT systems, where local training of LLMs is impractical.

Formally, this paper makes the following key contributions:

- We identify that existing federated black-box prompt tuning methods overlook the significant costs associated with querying cloud-based LLM services.
- To address this gap, we present the first theoretical analysis of Federated BDPL, with a focus on understanding and evaluating the query efficiency when interacting with cloud-based LLMs.
- Based on our analysis, we introduce the FedOne framework, a novel approach designed to optimize query efficiency in Federated Black-box Prompt Learning by activating only one client per round when querying cloud-based LLMs.

2 METHOD

2.1 FEDERATED BLACK-BOX PROMPT LEARNING FRAMEWORK

In the federated black-box prompt tuning framework, there is one central aggregation server and K clients. Each client, indexed by k, possesses a dataset D^k, consisting of input sentences Ψ^k and their corresponding labels Y^k, i.e., $D^k = \{\Psi^k, Y^k\}$. The dataset Ψ^k contains a total of M^k input

sentences, each represented as ψ_m^k, i.e., $\Psi^k = \left\{\psi_m^k\right\}_{m=1}^{M^k}$. Similarly, the corresponding labels y_m^k comprise Y^k, i.e. $Y^k = \left\{y_m^k\right\}_{m=1}^{M^k}$.

Each client generates a discrete sequence of prompt tokens $\Phi^k = \phi_1^k \cdots \phi_i^k \cdots \phi_n^k$ from a trainable parameter $\boldsymbol{\alpha}^k \in \mathbb{R}^{n \times N}$. This learnable parameter, $\boldsymbol{\alpha}^k$, is transmitted to the FL aggregation server for averaging. Details of how to generate the discrete prompt Φ^k from $\boldsymbol{\alpha}^k$ and the local training of $\boldsymbol{\alpha}^k$ through interaction with the cloud-based LLM service will be discussed in the next subsection regarding the local black-box prompt learning on the client.

The server and clients collaboratively solve a minimization problem, aiming to reduce a global loss function that aggregates the local loss functions from the clients. This can be expressed as:

$$\min_{\Phi} \left\{ \mathcal{L}(\Phi; \Psi) \triangleq \sum_{k=1}^{K} \frac{M^k}{M} \mathcal{L}^k(\Phi; \Psi^k) \right\}, \ \mathcal{L}^k(\Phi; \Psi^k) = \frac{1}{M^k} \sum_{m=1}^{M^k} \ell\left(\Phi; \psi_m^k, y_m^k\right) \tag{1}$$

where $\mathcal{L}(\Phi; \Psi)$ represents the global objective function of the FL, $\Psi = \left\{\Psi^k\right\}_{k=1}^{K}$. $\mathcal{L}^k(\cdot, \Psi^k)$ is the local objective function of client k. The function $\ell(\cdot, \cdot)$ denotes the loss function.

2.2 Black-box discrete prompt learning on the client

We now discuss the local training process of the client through interactions with the cloud-based LLM service. In this section, all variables have the superscript k, indicating that they belong to the k-th client. However, the reader can ignore this superscript and treat it as a standalone training process on a single machine.

Generating the discrete prompt sequence Φ^k The sequence of the discrete prompt Φ^k is generated from a vocabulary $\mathcal{V} = \{\mathcal{V}[j]\}_{j=1}^{N}$, which contains a total of N token options. Each token ϕ_i^k in the prompt sequence Φ^k is selected from the vocabulary, i.e., $\Phi^k = \phi_1^k \cdots \phi_i^k \cdots \phi_n^k = \mathcal{V}[j_1^k] \cdots \mathcal{V}[j_i^k] \cdots \mathcal{V}[j_n^k]$. For the i-th token ϕ_i^k, the prompt index j_i^k is sampled from the categorical distribution $\boldsymbol{p}_i^k$, i.e., $j_i^k \sim \text{Cat}(\boldsymbol{p}_i^k)$. Note that $\boldsymbol{p}_i^k = [p_{i,1}^k, ...p_{i,N}^k]$, where the element $p_{i,j}^k$ represents the probability that the token ϕ_i^k is selected as $V[j]$ from the vocabulary $\mathcal{V}$.

Directly optimizing the $\boldsymbol{p}_i^k$ may cause trouble in the convergence analysis as the gradient of the categorical distribution is biased. To address this, we re-parameterize the categorical distribution $\boldsymbol{p}_i^k$ using the Gumbel-Softmax technique (Jang et al., 2016) and introduce the parameter $\boldsymbol{\alpha}_i^k = [\alpha_{i,1}^k, ...\alpha_{i,N}^k]$ as the learnable parameter. The re-parameterization is shown below:

$$p_{i,j}^k = \frac{\exp\left(\frac{\log(\alpha_{i,j}^k) + g_{i,j}^k}{\tau}\right)}{\sum_{l=1}^{N} \exp\left(\frac{\log(\alpha_{i,l}^k) + g_{i,l}^k}{\tau}\right)} \tag{2}$$

where $\tau > 0$ is the temperature parameter, $g_{i,j}^k \sim \text{Gumbel}(0, 1)$ is the Gumbel random variable. We denote the Gumbel-Softmax function as GS, i.e. $\boldsymbol{p}^k = \text{GS}(\boldsymbol{\alpha}^k)$.

Optimizing the learnable parameter $\boldsymbol{\alpha}^k$ To compute the gradient with respect to the learnable parameter $\boldsymbol{\alpha}^k$, we first define the expected loss over the sequence of prompts in Eq. 3. The i-th token ϕ_i^k is generated from the vocabulary by sampling the prompt index from the categorical distribution, i.e. $\phi_i^k = \mathcal{V}[j_i^k]$, where $j_i^k \sim \text{Cat}(\boldsymbol{p}_i^k)$. For brevity, this sampling process is denoted as $\phi_i^k \sim \boldsymbol{p}_i^k$. We can define the expectation of the loss for the distribution of the prompt as follows:

$$\mathbb{E}_{\Phi^k \sim \boldsymbol{p}^k}\left[\mathcal{L}(\Phi^k, \Psi^k)\right] = \sum_{\phi_1^k \sim \boldsymbol{p}_1^k} \cdots \sum_{\phi_n^k \sim \boldsymbol{p}_n^k} \left(\mathcal{L}(\Phi^k, \Psi^k) \prod_{i=1}^{n} P(\phi_i^k) \right) \tag{3}$$

Following the same steps in (Diao et al., 2022, Eq. (2)), we can estimate the gradient w.r.t. $\boldsymbol{\alpha}_i^k$ by:

$$\nabla_{\boldsymbol{\alpha}_i^k} \mathbb{E}_{\Phi^k \sim \text{GS}(\boldsymbol{\alpha}^k)}\left[\mathcal{L}(\Phi^k, \Psi^k)\right] = \mathbb{E}_{\Phi^k \sim \text{GS}(\boldsymbol{\alpha}^k)}\left[\mathcal{L}(\Phi^k, \Psi^k) \nabla_{\boldsymbol{\alpha}_i^k} \log P(\phi_i^k)\right] \tag{4}$$

The j-th component of $\nabla_{\boldsymbol{\alpha}_i^k} \log P(\phi_i^k)$ could be explicitly computed as follows (with detailed steps provided in Appendix A.2):

$$\nabla_{\alpha_{i,j}^k} \log P(\phi_i^k) = \nabla_{\alpha_{i,j}^k} \log p_{i,j_i^k}^k = \begin{cases} \frac{1-p_{i,j_i^k}^k}{\tau \alpha_{i,j_i^k}^k} & j = j_i^k \\ -\frac{p_{i,j}^k}{\tau \alpha_{i,j}^k} & j \neq j_i^k \end{cases} \tag{5}$$

Then, we employ the mini-batch stochastic variance-reduced policy (MB-SVRP) estimator (Diao et al., 2022; Williams, 1992) to reduce the variance of the sampling when computing the gradient. This involves sampling the prompt sequence Φ^k from the distribution $\boldsymbol{p}^k$ multiple times, with the number of samplings denoted by I. The MB-SVRP estimator is then computed as follows:

$$\hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}^k, \mathcal{B}^k) = \frac{1}{I-1} \sum_{r=1}^{I} \left[\left(\ell(\Phi^{k,r}; \mathcal{B}^k) - \frac{1}{I} \sum_{w=1}^{I} \ell(\Phi^{k,w}; \mathcal{B}^k) \right) \nabla_{\boldsymbol{\alpha}_i^k} \log P(\phi_i^{k,r}) \right] \tag{6}$$

where $\left\{\Phi^{k,r}\right\}_{r=1}^{I}$ are sampled independently from $\boldsymbol{p}^k = \mathrm{GS}(\boldsymbol{\alpha}^k)$. The mini-batch $\mathcal{B}^k$, with size B^k, is sampled from the dataset Ψ^k. Note that the clients are unable to directly compute $\ell(\Phi^{k,*}, \mathcal{B}^k)$ on their own. They must transmit both the sampled prompt Φ^k and the mini-batch $\mathcal{B}^k$ to the cloud-based LLM service, which then computes the loss $\ell(\Phi^{k,*}, \mathcal{B}^k)$ and returns the result. Finally, with the learning rate set to η, the update of $\boldsymbol{\alpha}_i^k$ at the t-th iteration is expressed as follows:

$$\boldsymbol{\alpha}_{i,(t+1)}^k = \boldsymbol{\alpha}_{i,(t)}^k - \eta \cdot \hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}^k, \mathcal{B}_t^k) \tag{7}$$

2.3 Algorithm

Algorithm 1 outlines the Fed-BDPL framework, which integrates federated averaging with local client training with Gumbel-Softmax-BDPL (GS-BDPL). The FL aggregation server randomly selects a subset of clients and broadcasts the trainable parameters to them. Each selected client k then performs local training on the parameters and return the updated parameters $\boldsymbol{\alpha}^k$ to the server. The local training of the selected client is conducted through black-box prompt learning by querying the cloud-based LLM service, as detailed in Section 2.2. Finally, the server aggregates the updates by averaging the parameters from all participating clients. This process is iteratively repeated throughout the training. Specifically, in the FedOne framework, the number of activated clients is set to 1, as highlighted in the light green box. The reasons behind FedOne's high query efficiency on cloud-based LLM services will be formally analyzed in the next section.

Algorithm 1 Fed-BDPL. C denotes the sampling ratio of the clients. K_* represents the number of selected clients, and U is the set of selected clients.

Server executes:
Initialize $\boldsymbol{\alpha}$
for $s = 0, 1, \ldots S-1$ **do**
  FedAvg: $K_* \leftarrow \max(C \cdot K, 1)$
  FedOne: $K_* \leftarrow 1$
  $U_s \leftarrow$ (sampling K_* clients)
  for $k \in U_s$ **in parallel do**
    $\boldsymbol{\alpha}^k \leftarrow \boldsymbol{\alpha}$
    $\boldsymbol{\alpha}^k \leftarrow$ Client_Update$(k, \boldsymbol{\alpha}^k)$
  end for
  $\boldsymbol{\alpha} \leftarrow \frac{1}{K_*} \sum_{k \in U_s} \boldsymbol{\alpha}^k$
end for

Client executes:
function Client_Update$(k, \boldsymbol{\alpha}^k)$:
for $e = 1, ..., E$ **do**
  Re-parameterize the categorical distribution $\boldsymbol{p}^k = \mathrm{GS}(\boldsymbol{\alpha}^k)$ using Eq. 2.
  Query the cloud-based LLM server to obtain $\left\{\mathcal{L}(\Phi^{k,r}, \mathcal{B}_e^k)\right\}_{r=1}^{I}$.
  Compute $\hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}^k, \mathcal{B}_e^k)$ using Eq. 6.
  $\boldsymbol{\alpha}_i^k \leftarrow \boldsymbol{\alpha}_i^k - \eta \cdot \hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}^k, \mathcal{B}_e^k)$
end for
Return $\boldsymbol{\alpha}^k$

3 CONVERGENCE ANALYSIS

3.1 ASSUMPTIONS

Assumption 1. ***Unbiasedness and bounded variance of stochastic gradient:*** *We assume that the stochastic gradient is unbiased and has bounded variance.*

$$\mathbb{E}_{\psi_m^k}\left[\nabla_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}^k, \psi_m^k)\right] = \nabla_{\boldsymbol{\alpha}_i^k} F(\boldsymbol{\alpha}^k, \Psi^k) \tag{8}$$

$$\mathbb{E}_{\psi_m^k}\left\|\nabla_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}^k, \psi_m^k) - \mathbb{E}_{\psi_m^k}\left[\nabla_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}^k, \psi_m^k)\right]\right\|^2 \leq \sigma_\psi^2 \tag{9}$$

Assumption 1 is the basic assumption for solving non-convex optimization problems using stochastic gradient descent (Ghadimi & Lan, 2013; Hazan & Kale, 2014; Xu et al., 2019; Liu et al., 2020).

Assumption 2. ***Bounded loss****: At the k-th client, $\forall\ \psi_m^k \in \Psi^k$, and Φ^k sampled by $\boldsymbol{p}^k$, we perform a clipping operation with a constant G for loss function $\ell(\cdot\ ,\ \cdot)$:*

$$\left|\ell(\Phi^k, \psi_m^k)\right| \leq G \tag{10}$$

Assumption 2 ensures that the loss value is bounded, primarily to regulate the loss during I-sample estimation in stochastic policy gradient, thereby facilitating theoretical analysis.

Assumption 3. ***Clients' heterogeneity by weighted gradient diversity:*** *For client $k = 1, ..., K$ with sampling probability vector $\boldsymbol{q} = \left\{q^{[k]}\right\}_{k=1}^K$, and $\nabla_{\boldsymbol{\alpha}_i^k} F^k(\boldsymbol{\alpha}^k, \Psi^k)$ is local gradient of $\boldsymbol{\alpha}_i$ w.r.t. all input sentence in client k. We assume that λ is the upper bound on the weighted gradient diversity across local objectives, i.e.*

$$\Lambda(\boldsymbol{\alpha}_i, \boldsymbol{q}) \triangleq \frac{\sum_{k=1}^K q^{[k]} \cdot \left\|\nabla_{\boldsymbol{\alpha}_i^k} F^k(\boldsymbol{\alpha}^k, \Psi^k)\right\|^2}{\left\|\sum_{k=1}^K q^{[k]} \cdot \nabla_{\boldsymbol{\alpha}_i^k} F^k(\boldsymbol{\alpha}^k, \Psi^k)\right\|^2} \leq \lambda \tag{11}$$

Assumption 3 introduces gradient diversity as a measure of client heterogeneity, which is used to quantify differences among clients and establish convergence conditioned in heterogeneous FL settings (Yin et al., 2018; Haddadpour & Mahdavi, 2019).

3.2 THE CONVERGENCE OF FEDERATED BLACK-BOX DISCRETE PROMPT LEARNING

Theorem 1. *Suppose assumption 1, 2 and 3 hold, using algorithm 1 to solve the FedBDPL problem defined in Eq. 1. Let $B = \min\left\{B^1, ..., B^K\right\}$ where B^k represents the local mini-batch size for each client. Set $\alpha_{i,j} \geq \nu > 0$. The variance of the variance-reduced policy gradient is given by $\sigma_\alpha^2 = \frac{8G^2N}{\tau^2\nu^2}$. $\mathbb{E}_{\Phi^k \sim GS(\boldsymbol{\alpha}^k)}\left[\mathcal{L}(\Phi^k, \Psi^k)\right]$ is L-smooth w.r.t. $\boldsymbol{\alpha}^k$ where $L = \frac{nGN(\tau+1)}{\tau^2\nu^2}$. If the learning rate η satisfies the following condition:*

$$0 < \eta \leq \eta^* = \frac{-\lambda L + \sqrt{\lambda^2 L^2 + 8L^2 E\left(1 + \frac{1}{K_*}\right)}}{8L^2 E\left(1 + \frac{1}{K_*}\right)} \tag{12}$$

Then, the Fed-BDPL's expected gradient, $\nabla_{\boldsymbol{\alpha}} F(\boldsymbol{\alpha}_t, \Psi^k)$, can be bounded as follows:

$$\frac{1}{T}\sum_{t=0}^{T-1}\left\|\nabla_{\boldsymbol{\alpha}} F(\boldsymbol{\alpha}_t, \Psi^k)\right\|^2 \leq \frac{4G}{\eta T} + \frac{2(E+1)L^2\eta^2 n\sigma_\psi^2(1+\frac{1}{K_*}) + 2nL\eta\sigma_\psi^2}{B} + \frac{2(E+1)L^2\eta^2 n\sigma_\alpha^2(1+\frac{1}{K_*}) + 2nL\eta\sigma_\alpha^2}{I^2} \tag{13}$$

Remark 1. The local gradient of BDPL is unbiased within each client, but due to data heterogeneity across clients, it becomes biased with respect to the global gradient (Haddadpour & Mahdavi, 2019). This bias can be mitigated by reducing the step size η. However, as the gradient diversity, λ, increases, an even smaller step size η is needed, which in turn slows down the convergence process.

Corollary 1. ***Convergence rate and complexity***
1) Convergence rate: Let $\eta = \min\left\{\eta^*, \frac{1}{\sqrt{T}}, \frac{1}{L}\right\}$. *Under this condition, the following holds:*

$$\frac{1}{T}\sum_{t=0}^{T-1}\left\|\nabla_{\boldsymbol{\alpha}} F(\boldsymbol{\alpha}_t, \Psi)\right\|^2 = \mathcal{O}(\frac{1}{\sqrt{T}}) \tag{14}$$

2) Complexity: To guarantee an ϵ*-solution, such that* $\frac{1}{T}\sum_{t=0}^{T-1}\left\|\nabla_{\boldsymbol{\alpha}} F(\boldsymbol{\alpha}_t, \Psi^k)\right\|^2 \leq \epsilon^2$, *the following condition must hold:*

$$T_\epsilon = \mathcal{O}\left(\frac{1}{\epsilon^4}\right) \tag{15}$$

Fed-BDPL achieves a sub-linear convergence rate of $\left(\frac{1}{\sqrt{T}}\right)$ and a complexity $\mathcal{O}\left(\frac{1}{\epsilon^4}\right)$, which is comparable to that of classical non-convex smooth SGD algorithms (Allen-Zhu, 2018; Khaled & Richtárik, 2020; Gower et al., 2021)).

Corollary 2. ***The impact of*** K_* ***(FedOne)****: In the FL framework,* $T_\epsilon K_*$ *represents the total number of queries made to the cloud-based LLM service to achieve an* ϵ*-solution, whose quantity is directly proportional to the cost incurred for utilizing the LLM. The following condition holds:*

$$T_\epsilon K_* \propto K_* \tag{16}$$

Therefore, $T_\epsilon K_*$ *is a function of* K_* *that increase steadily for* $K_* = 1, 2, \ldots, K$, *indicating that the optimal* K_* *for query efficiency is* $K_* = 1$.

4 Experiment

The objective of the experiment was to assess the performance of Fed-BDPL, along with various aspects of the framework, and to explore the query efficiency advantages of FedOne. The code for this project will be made open source.

4.1 Experiment setup

For our experiment, we utilized the GLUE benchmark (Wang et al., 2018), which includes a wide range of tasks including MNLI (Williams et al., 2018), QQP (Iyer et al., 2017), SST-2 (Socher et al., 2013), MRPC (Dolan & Brockett, 2005), CoLA (Warstadt et al., 2019), QNLI (Wang et al., 2018), and RTE (Dagan et al., 2005; Haim et al., 2006; Giampiccolo et al., 2007; Bentivogli et al., 2009).

In the baseline experiment within the federated learning framework, we employed 100 clients. In the FedOne framework, there was only one client activated per round for training and aggregation. We adopted the k-shot framework from Perez et al. (2021), adapting it to a federated learning context. Each client received a k-shot dataset comprising k samples per class.

The model architecture employed is RoBERTa-large (Liu et al., 2019). The trainable prompts were placed at different positions in the model depending on the algorithm of the baselines. For the training procedure, we conducted a hyperparameter tuning phase using a grid search approach to explore learning rates of $[3e{-}4, 1e{-}4, 3e{-}5, 1e{-}5]$. The batch size was set at 32, the prompt length was set at 20, and the optimization algorithm employed was AdamW (Loshchilov & Hutter, 2017). Further details about the dataset and evaluation metrics of them are available in Appendix B.1.

4.2 Baselines

We evaluated several approaches categorized as standalone and federated learning models. The standalone model baseline includes Manual Prompt Tuning, In-Context Learning (Brown et al., 2020), and Fine-tuning (Diao et al., 2022, Table 7)). In the domain of Federated Prompt Tuning for white-box scenarios, we adapted two established white-box prompt tuning methods to Federated Learning. Specifically, we implemented prompt-tuning (Lester et al., 2021) and prefix-tuning v2 (Liu et al., 2021) across distributed clients. In the prompt tuning approach, a prompt was integrated into the embedding layer of the model for each client. Each client then undertook local training solely on their respective prompt. In prefix-tuning v2, the prompt was utilized across all embedding layers

Table 1: The overall performance on the RoBERTa-large. Each trial runs across three random seeds.

Dataset	MNLI	QQP	SST-2	MRPC	CoLA	QNLI	RTE	Avg.
Manual Prompt	$35.9_{1.3}$	$49.8_{0.9}$	$77.2_{1.1}$	$70.4_{1.6}$	$0.6_{0.0}$	$49.2_{1.1}$	$48.2_{0.6}$	47.33
In Context Learning	$37.2_{1.6}$	$50.1_{0.9}$	$82.8_{2.1}$	$72.1_{2.3}$	$1.1_{0.4}$	$50.8_{0.5}$	$49.3_{2.3}$	49.06
FineTuning	$50.8_{1.2}$	$60.8_{1.9}$	$86.5_{2.0}$	$78.4_{1.3}$	$20.4_{1.9}$	$53.2_{1.8}$	$55.6_{2.3}$	57.96
FedOne-PromptTuning	$41.5_{0.9}$	$66.4_{0.2}$	$77.9_{2.1}$	$79.5_{0.5}$	$0.8_{1.1}$	$49.6_{1.0}$	$53.1_{0.6}$	52.69
FedOne-P-Tuning v2	$42.7_{0.7}$	$66.7_{0.1}$	$82.9_{0.3}$	$80.6_{0.1}$	$1.0_{1.0}$	$52.4_{0.2}$	$56.4_{0.4}$	54.67
FedOne-BBT	$41.9_{0.4}$	$66.3_{0.2}$	$76.8_{1.6}$	$80.6_{0.3}$	$2.5_{1.3}$	$51.1_{0.4}$	$55.3_{1.0}$	53.50
FedOne-BDPL	$41.0_{1.2}$	$66.7_{0.1}$	$80.8_{6.0}$	$81.1_{0.1}$	$5.2_{2.4}$	$51.7_{1.4}$	$57.1_{1.9}$	54.80
FedOne-GS-BDPL	$41.1_{0.4}$	$66.9_{0.2}$	$80.8_{0.4}$	$81.0_{0.1}$	$5.3_{1.1}$	$52.1_{0.8}$	$57.1_{1.1}$	54.90

of the model, providing more trainable parameters and enhancing the capability to adapt to downstream tasks. In the federated prompt tuning for the black-box scenario, we adapted the Black-Box Tuning (BBT) (Sun et al., 2022) to FedOne, incorporating projection from a low-dimensional vector and the Covariance Matrix Adaptation Evolution Strategy (CMA-ES) into federated black-box prompt learning. Each client held a distinct low-dimensional vector, while the projection matrix A was shared among all clients. For every client, the population size of CMA-ES is set to 20, and the dimension of the low-dimensional vector is set to 500, as recommended by Sun et al. (2022). Finally, we adapted the BDPL (Diao et al., 2022) and Gumbel-Softmax BDPL (GS-BDPL) to the Federated Learning, employing policy gradient methods and Gumbel-Softmax as outlined in Algorithm 1.

4.3 Result

Test accuracy The performance results were summarized in Table 1, We observed significant variations in the effectiveness of different learning approaches when applied to the RoBERTa-large model across various NLP tasks. Traditional fine-tuning methods outperformed both Manual Prompt and In-Context Learning techniques, achieving the highest average score of 57.96 across all datasets. This result underscores the effectiveness of complete model retraining over other methods that involve fewer parameter updates or rely solely on contextual adjustments. Although Manual Prompt performed well on some datasets, its performance lacks stability. White-box federated prompt tuning methods demonstrate improvements over Manual Prompt and In Context Learning techniques. Generally, the black-box method (BDPL and GS-BDPL) exhibits performance that is comparable to, or slightly better than, the white-box tuning method (PromptTuning and P-Tuning v2).

Computational efficiency and resource utilization Table 2 presents the performance metrics for various federated prompt tuning methods per client, focusing on the computational and communication efficiencies of these methods. We measured the computation time required for model training. Notably, in the white-box method, training occurs on the client side. We only measure the time for forward and backward propagation of the local model, along with the time for parameter updates. In the black-box prompt learning method, the computation time for model training includes both the execution of the black-box algorithm and the wait time for responses from the cloud-based LLM services. The results are presented in the first column of Table 2. We observed that the white-box method is considerably faster; however, it relies on the assumption that clients possess sufficient computational power, such as GPUs, which may not be practical in FL environments. Within the black-box approach, BDPL requires less computation time for training compared to BBT.

The advantage of black-box prompt learning lies primarily in its reduced communication costs and the efficiency of the trainable parameter size, as well as in avoiding the need to store and train the entire LLM on the client. As illustrated in the table, the federated black-box prompt tuning method features a significantly smaller parameter size and eliminates the GPU requirement on the client. This allows devices with limited computational resources, such as edge devices, to participate in the federated prompt learning process.

4.4 Case study

Query efficiency and the number of activated clients We began by illustrating the relationship between query efficiency and the number of activated clients (K_*) using a toy example on the

Table 2: Evaluation of computational efficiency and resource utilization

Baseline	Comp. Time for Training (s)	FL Server Comm. Cost (MB)	FL Server # Queries	LLM Server # Query	Client Trainable Parameter Size (MB)	Client Loaded GPU Memory (MB)
FedOne-PromptTuning	77.6	15.63	100	-	7.8×10^{-2}	3564 (Model, Grad., Prompt)
FedOne-P-Tuning v2	91.1	1141.72	100	-	5.7	3656 (Model, Grad., Prompt)
FedOne-BBT	20474.6	0.38	100	2×10^{4}	1.9×10^{-3}	1.90×10^{-3} (Prompt)
FedOne-BDPL	1614.3	3.05	100	2×10^{3}	1.5×10^{-2}	1.52×10^{-2} (Prompt)
FedOne-GS-BDPL	1624.4	3.05	100	2×10^{3}	1.5×10^{-2}	1.52×10^{-2} (Prompt)

MNIST dataset (LeCun et al., 2010), as shown in Figure 2. We then evaluated the impact of K_* on query efficiency in Federated Black-box Prompt Tuning, as presented in Table 3.

In the first experiment, the MNIST dataset is split across 100 clients, with each client initially possessing an equal subset of the training data. We experiment with 2 epochs with a learning rate of 0.01 and a batch size of 32. We explore five configurations by varying the number of active clients per epoch: 1, 5, 10, 20, and 40 clients (denoted as K_*=1, K_*=5, K_*=10, K_*=20, and K_*=40 respectively). The model for each client is a Multilayer Perceptron (MLP). It includes a flattening input layer, a fully connected layer with 512 neurons and ReLU activation, a dropout layer with 0.2 dropout rate to prevent overfitting and a final fully connected layer that outputs to 10 classes via a Softmax function. As illustrated in Figure 2, utilizing the minimum number of activated clients (FedOne) enables the FL framework to achieve optimal query efficiency for convergence. Although Li et al. (2019) have demonstrated that an increased number of clients participating in the training and aggregation process accelerates convergence in federated learning, in the context of Federated Black-box prompt tuning, the number of queries to the LLM server increases linearly with the number of participating clients. This rise in query numbers outpaces the benefits gained from faster convergence due to increased client participation.

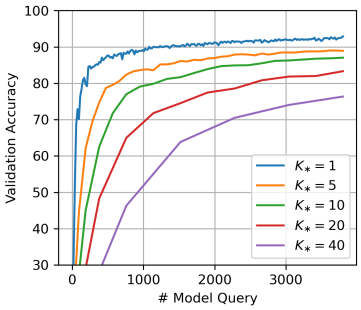

Figure 2: Toy Example Demonstrate on Query Efficiency

Table 3: Query Efficiency in Federated Black-Box Prompt Learning

AC	Fed-BDPL		Fed-BBT	
	# Epoch	# LLM Queries	# Epoch	# LLM Queries
1	$25.4_{\pm19.2}$	**528.6**$_{\pm382.4}$	$10.8_{\pm4.4}$	**2350.0**$_{\pm870.3}$
3	$13.8_{\pm5.2}$	$889.4_{\pm310.9}$	$5.4_{\pm2.3}$	$3825.0_{\pm1356.2}$
10	$11.5_{\pm6.8}$	$2430.8_{\pm1378.8}$	$4.2_{\pm1.8}$	$10444.4_{\pm3624.3}$
30	$3.5_{\pm2.8}$	$2672.7_{\pm1625.4}$	$1.7_{\pm0.9}$	$16000.0_{\pm5656.9}$

In the second experiment investigating query efficiency within the federated black-box prompt learning scenario, we utilized the SST2 dataset to explore how varying the number of activated clients per round affects model convergence efficiency in FL environments. We tested the number of activated clients per round within the range of $[1, 3, 10, 30]$. For each configuration, we monitored the number of queries required to achieve target accuracy on the validation dataset. Specifically, we evaluated the prompt at the end of each epoch and stop training once the target accuracy was reached, reporting the number of LLM queries at that point. To ensure reliability and account for variability in the learning process, each experimental setup was replicated 20 times, and outliers in the number of queries were removed. The primary aim of our study was to explore how the number of activated clients affects the speed and efficiency of model convergence in FL, specifically to demonstrate the query efficiency of the FedOne approach.

The results presented in Table 3 demonstrate a clear trend where fewer activated clients are associated with greater query efficiency. This relationship is evidenced by a consistent decrease in the number of cloud-based LLM queries as the number of activated clients is reduced, a pattern observed across both the Fed-BDPL and Fed-BBT.

Table 4: Federated discrete black-box prompt learning on GPT-3.5 Turbo

	MNLI	QQP	SST-2	MRPC	CoLA	QNLI	RTE	Avg.
No Prompt	$14.05_{0.00}$	$68.04_{0.00}$	$91.35_{0.00}$	$79.62_{0.00}$	$36.01_{0.00}$	$56.22_{0.00}$	$72.20_{0.00}$	59.64
Prompt w/o. Training	$9.17_{1.31}$	$62.94_{4.94}$	$79.47_{2.71}$	$79.88_{1.94}$	$25.53_{1.10}$	$56.69_{4.76}$	$72.16_{2.51}$	54.41
FedOne-BDPL	$13.67_{0.17}$	$68.25_{3.05}$	$87.77_{3.82}$	$81.69_{0.70}$	$32.16_{6.51}$	$70.58_{1.35}$	$77.62_{1.62}$	61.67
FedOne-GS-BDPL	$\mathbf{16.00}_{1.78}$	$\mathbf{69.92}_{2.85}$	$\mathbf{92.58}_{2.71}$	$\mathbf{82.93}_{0.67}$	$\mathbf{36.20}_{2.95}$	$\mathbf{72.05}_{1.64}$	$\mathbf{80.33}_{2.32}$	**64.28**

Real-world implementation on GPT-3.5 Turbo We implemented the Fed-BDPL framework using GPT-3.5 Turbo, a widely recognized and powerful closed-source language model. We leveraged the OpenAI API (OpenAI, 2024) to enable individual clients to conduct local training. In this implementation of the federated black-box prompt learning, clients sent prompts and input sentences to GPT-3.5, which returns the logarithm of the token probabilities at each position. A key challenge was that OpenAI API only provides probabilities for the top 20 tokens for each position. Consequently, we needed to transform the predictions on these tokens into the categorical prediction of the input sentence, instead of using straightforward model output as we did in the RoBERTa experiment. To solve this problem, we appended a template question at the end of the input sentence to query the target label token. For example, in the QQP task, we added the phrase "equivalent? yes or no" to the end of the input sentence. This allowed us to retrieve the top probabilities for all class label tokens ("yes", "no") in the top 20 probability. and use the probability of the target token as the logit output, for the following procedure.

The results were presented in table 4, indicating that GPT-3.5 Turbo achieves a certain level of performance without any prompts. When adding a random prompt without training, the model's performance dropped. After tuning the prompts, performance improved significantly, surpassing the performance without a prompt. Furthermore, the table demonstrated that the GS-BDPL method consistently outperforms other black-box approaches. To summary, this method can be used to perform prompt tuning on federated learning with extremely low computational resource requirements.

5 Related works

5.1 White-box and black-box prompt tuning

Prompt tuning is a technique for adapting LLM to downstream tasks. It tailors the model's responses to specific tasks or styles without requiring the retraining of the entire model. In white-box prompt tuning, the learner is granted full access to the LLM, allowing them to modify and access intermediate results of the model and acquire the gradient. Li & Liang (2021) and Lester et al. (2021) proposed a lightweight and modular alternative to full model fine-tuning for natural language generation tasks, which optimizes a sequence of continuous soft prompts, prepend in the embedding layers of the LLM. Liu et al. (2021) propose the P-tuning v2. Instead of only applying the prompt in the input layer in Li & Liang (2021), they adapt trainable parameters on all layers' inputs, which can effectively match the performance of fine-tuning across a wide range of models.

In situations where the learner cannot access the intermediate result of the LLM model, the learner has to use a black-box prompt tuning method. In the black-box prompt tuning, the learner can only query the output of the LLM with the input of the model. Most of the research assumes the input is at the soft prompt layer of the LLM (Sun et al., 2022; Chen et al., 2023). Others research use the discrete prompt which is concrete with the input text, which is more portable, and is usable for any cloud-based LLM API (Diao et al., 2022). Xiao et al. (2023) presents a privacy-preserving, efficient transfer learning method that adapts large foundation models to specific tasks. This method does not require access to the full model or compromise data privacy. It utilizes a lightweight adapter and a compressed emulator for model tuning. Chen et al. (2023) introduces an efficient method for optimizing instructions for black-box LLMs using Bayesian optimization of soft prompts. This approach significantly improves LLM performance across a variety of tasks without requiring direct access to the model's internals. Deng et al. (2022) proposes an efficient method to optimize discrete text prompts using reinforcement learning, demonstrating superior performance compared to other prompt optimization techniques across a variety of tasks. Sun et al. (2022) proposed black-box tuning (BBT), a method that optimizes continuous prompts by optimizing a lower-dimensional vector and projecting them to the prompt searching space. They use the Covariance Matrix Adaptation

Evolution Strategy (CMA-ES) for optimizing the vector. Diao et al. (2022) introduced BDPL which optimizes the discrete prompt with the policy gradient method.

The black-box prompt tuning method is versatile and applicable to various tasks and models without model-specific modifications. However, its major drawback is computational inefficiency, requiring multiple forward passes through the model, which leads to high costs and extended training times. Additionally, the convergence of the derivation-free optimization method is often slow.

5.2 Federated learning

Federated learning (McMahan et al., 2017; Karimireddy et al., 2020; Li et al., 2020; 2021; Marfoq et al., 2022; Mishchenko et al., 2022), first introduced by McMahan et al. (2017), is a paradigm that enables devices to collaboratively train a shared predictive model by locally aggregating updates. In this framework, each client maintains a copy of the model for local training, and the server selects a subset of clients in each round for aggregation. Since the introduction of FedAvg, it has lacked formal theoretical convergence guarantees. As a result, researchers have made significant efforts to establish and demonstrate its convergence (Zhou & Cong, 2017; Stich, 2018).

Partial client activation is a key area of research in FL and has gained significant attention due to its impact on improving convergence rates and system efficiency. Stich (2018) shows that in the convex case, increasing the number of activated clients significantly improves convergence rates in Federated Learning with independent and identically distributed (IID) data, achieving a linear speed-up. Li et al. (2019) extended the understanding of federated learning by analyzing the convergence of the FedAvg algorithm, under the convex case. They demonstrate that under the non-IID setting, the convergence rate has a weak dependence on the number of activated clients, which implies that the FedAvg is not able to achieve linear speedup under this case, therefore the participation ratio can be set smaller to alleviate the straggler effect without affecting the convergence rate.

5.3 Federated prompt tuning

Applying federated learning to prompt tuning can enhance the model by incorporating additional data. This approach leverages distributed datasets to improve model performance while adhering to data privacy. To apply white-box prompt tuning in FL, each client maintains the entire model but trains only the prompt parameters. These parameters are then shared and aggregated across clients (Zhao et al., 2023; Zhang et al., 2023; Che et al., 2023). However, this method assumes white-box access to LLM, which is impractical for closed-source LLMs. Consequently, black-box prompt learning has been adapted for FL to address these limitations (Lin et al., 2023; Sun et al., 2023). Lin et al. (2023), applied the black-box prompt tuning method Diao et al. (2022) to the Federated Learning, where the client can train the probability matrix for the discrete prompt via querying the cloud-based LLM API. Sun et al. (2023) applied the BBT to FL. In this approach, clients train only low-dimensional vectors using CMA-ES via querying the cloud-based LLM API.

For federated prompt tuning, most existing research has focused on conventional issues in FL, such as data heterogeneity (Zhao et al., 2023), privacy (Zhang et al., 2023), security (Zhao et al., 2023), and client computation-communication efficiency (Lin et al., 2023; Sun et al., 2023). However, no studies have yet addressed the query efficiency of Federated Black-box Prompt Tuning, a novel challenge introduced by the deployment of Black-box Prompt Tuning through cloud-based APIs.

6 Conclusion

We identified that previous research on federated black-box prompt tuning had overlooked the significant query costs associated with cloud-based LLM services. To address this issue, we conducted a theoretical analysis of query efficiency in the context of federated black-box prompt tuning, revealing the relationship between query efficiency and the number of activated clients. Based on our findings, we propose the FedOne framework, which achieves optimal query efficiency with respect to the number of activated clients. We performed numerical experiments on various aspects of FedOne, further validating its performance through real-world experiments on GPT-3.5.

REFERENCES

Zeyuan Allen-Zhu. How to make the gradients small stochastically: Even faster convex and nonconvex sgd. *Advances in Neural Information Processing Systems*, 31, 2018.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. The fifth pascal recognizing textual entailment challenge. In *TAC*, 2009.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Tianshi Che, Ji Liu, Yang Zhou, Jiaxiang Ren, Jiwen Zhou, Victor S Sheng, Huaiyu Dai, and Dejing Dou. Federated learning of large language models with parameter-efficient prompt tuning and adaptive optimization. *arXiv preprint arXiv:2310.15080*, 2023.

Lichang Chen, Jiuhai Chen, Tom Goldstein, Heng Huang, and Tianyi Zhou. Instructzero: Efficient instruction optimization for black-box large language models. *arXiv preprint arXiv:2306.03082*, 2023.

Ido Dagan, Oren Glickman, and Bernardo Magnini. The pascal recognising textual entailment challenge. In *Machine Learning Challenges Workshop*, pp. 177–190. Springer, 2005.

Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric P Xing, and Zhiting Hu. Rlprompt: Optimizing discrete text prompts with reinforcement learning. *arXiv preprint arXiv:2205.12548*, 2022.

Shizhe Diao, Zhichao Huang, Ruijia Xu, Xuechun Li, Yong Lin, Xiao Zhou, and Tong Zhang. Black-box prompt learning for pre-trained language models. *arXiv preprint arXiv:2201.08531*, 2022.

William B. Dolan and Chris Brockett. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, 2005. URL https://aclanthology.org/I05-5002.

Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM journal on optimization*, 23(4):2341–2368, 2013.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. The third PASCAL recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pp. 1–9, Prague, 2007. Association for Computational Linguistics. URL https://aclanthology.org/W07-1401.

Robert Gower, Othmane Sebbouh, and Nicolas Loizou. Sgd for structured nonconvex functions: Learning rates, minibatching and interpolation. In *International Conference on Artificial Intelligence and Statistics*, pp. 1315–1323. PMLR, 2021.

Farzin Haddadpour and Mehrdad Mahdavi. On the convergence of local descent methods in federated learning. *arXiv preprint arXiv:1910.14425*, 2019.

R Bar Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. The second pascal recognising textual entailment challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, volume 7, 2006.

Elad Hazan and Satyen Kale. Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization. *The Journal of Machine Learning Research*, 15(1): 2489–2512, 2014.

Shankar Iyer, Nikhil Dandekar, Kornél Csernai, et al. First quora dataset release: Question pairs. *data. quora. com*, 2017.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pp. 5132–5143. PMLR, 2020.

Ahmed Khaled and Peter Richtárik. Better theory for sgd in the nonconvex world. *ArXiv*, abs/2002.03329, 2020. URL https://api.semanticscholar.org/CorpusID:211069380.

Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2:429–450, 2020.

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of FedAvg on non-iid data. *arXiv preprint arXiv:1907.02189*, 2019.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.

Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. FedBN: Federated learning on non-IID features via local batch normalization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/pdf?id=6YEQUn0QICG.

Zihao Lin, Yan Sun, Yifan Shi, Xueqian Wang, Lifu Huang, Li Shen, and Dacheng Tao. Efficient federated prompt tuning for black-box large pre-trained models. *arXiv preprint arXiv:2310.03123*, 2023.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Lam Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*, 2021.

Yanli Liu, Yuan Gao, and Wotao Yin. An improved analysis of stochastic gradient descent with momentum. *Advances in Neural Information Processing Systems*, 33:18261–18271, 2020.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Othmane Marfoq, Giovanni Neglia, Richard Vidal, and Laetitia Kameni. Personalized federated learning through local memorization. In *International Conference on Machine Learning*, pp. 15070–15092. PMLR, 2022.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.

Konstantin Mishchenko, Grigory Malinovsky, Sebastian Stich, and Peter Richtárik. Proxskip: Yes! local gradient steps provably lead to communication acceleration! finally! In *International Conference on Machine Learning*, pp. 15750–15769. PMLR, 2022.

Yurii Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

OpenAI. Openai api, 2024. URL https://openai.com/index/openai-api/. Accessed: 2024-09-30.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. True few-shot learning with language models. *Advances in neural information processing systems*, 34:11054–11070, 2021.

Sashank J. Reddi, Ahmed Hefny, Suvrit Sra, Barnabas Poczos, and Alex Smola. Stochastic variance reduction for nonconvex optimization. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 314–323, New York, New York, USA, 20–22 Jun 2016. PMLR.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642, Seattle, Washington, USA, 2013. Association for Computational Linguistics. URL https://aclanthology.org/D13-1170.

Sebastian U Stich. Local sgd converges fast and communicates little. *arXiv preprint arXiv:1805.09767*, 2018.

Jingwei Sun, Ziyue Xu, Hongxu Yin, Dong Yang, Daguang Xu, Yiran Chen, and Holger R Roth. Fedbpt: Efficient federated black-box prompt tuning for large language models. *arXiv preprint arXiv:2310.01467*, 2023.

Tianxiang Sun, Yunfan Shao, Hong Qian, Xuanjing Huang, and Xipeng Qiu. Black-box tuning for language-model-as-a-service. In *International Conference on Machine Learning*, pp. 20841–20855. PMLR, 2022.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641, 2019. doi: 10.1162/tacl_a_00290. URL https://aclanthology.org/Q19-1040.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1112–1122, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1101. URL https://aclanthology.org/N18-1101.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8:229–256, 1992.

Guangxuan Xiao, Ji Lin, and Song Han. Offsite-tuning: Transfer learning without full model. *arXiv preprint arXiv:2302.04870*, 2023.

Yi Xu, Rong Jin, and Tianbao Yang. Non-asymptotic analysis of stochastic methods for non-smooth non-convex regularized problems. *Advances in Neural Information Processing Systems*, 32, 2019.

Dong Yin, Ashwin Pananjady, Max Lam, Dimitris Papailiopoulos, Kannan Ramchandran, and Peter Bartlett. Gradient diversity: a key ingredient for scalable distributed learning. In Amos Storkey and Fernando Perez-Cruz (eds.), *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 1998–2007. PMLR, 09–11 Apr 2018. URL https://proceedings.mlr.press/v84/yin18a.html.

Zhuo Zhang, Yuanhang Yang, Yong Dai, Qifan Wang, Yue Yu, Lizhen Qu, and Zenglin Xu. FedPETuning: When federated learning meets the parameter-efficient tuning methods of pre-trained language models. In *Annual Meeting of the Association of Computational Linguistics 2023*, pp. 9963–9977. Association for Computational Linguistics (ACL), 2023.

Haodong Zhao, Wei Du, Fangqi Li, Peixuan Li, and Gongshen Liu. Fedprompt: Communication-efficient and privacy-preserving prompt tuning in federated learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Fan Zhou and Guojing Cong. On the convergence properties of a k-step averaging stochastic gradient descent algorithm for nonconvex optimization. *arXiv preprint arXiv:1708.01012*, 2017.

A CONVERGENCE ANALYSIS

A.1 NOTATION AND OMITTED MATHEMATICAL STEPS

$t = 0, ..., T-1$	Number of global iterations
$e = 1, ..., E$	Number of local iterations
$s = 0, ..., S$	Number of client-server interactions
$\|\cdot\|$	The default 2-norm in this paper
$\text{Cat}(\cdot)$	Categorical distribution function
$\phi_i \sim \boldsymbol{p}_i$	Abbreviation for $j_i \sim \text{Cat}(\boldsymbol{p}_i)$
D^k	Dataset for client k
$\Psi = \left\{\psi_m^k\right\}_{m=1}^{M^k}$	Input sentence
$Y^k = \left\{y_m^k\right\}_{m=1}^{M^k}$	The category corresponding to the input sentence
K	The number of all clients
$\boldsymbol{q} = \left\{q^{[k]}\right\}_{k=1}^{K}$	Probability of each client being selected
K_*	The number of selected clients.
U	The set of selected clients.
$\boldsymbol{\alpha}^k = \left\{\boldsymbol{\alpha}_i^k\right\}_{i=1}^{n}$	Gumbel-Softmax parameters for k-th client
$\boldsymbol{\alpha} = \frac{1}{K_*}\sum_{k\in U_t}\boldsymbol{\alpha}^k$	Average of $\boldsymbol{\alpha}^k$
$\Phi^k = \phi_1^k \cdots \phi_n^k$	Prompt
$\mathcal{V}$	Vocabulary list of total N tokens choice
$\boldsymbol{p}_i^k = [p_{i,1}^k, \cdots, p_{i,N}^k]$	The probability distribution over the N token indexes.
$\nabla_{\boldsymbol{\alpha}_i^k} F^k(\cdot,\cdot)$	The local full gradient of the client k (19)
$\nabla_{\boldsymbol{\alpha}_i^k} f^k(\cdot,\cdot)$	The local stochastic mini-batch gradient of the client k (18)
$\hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^k(\cdot,\cdot)$	The local stochastic mini-batch variance-reduced policy gradient of the client k (6)
$\nabla_{\boldsymbol{\alpha}_i^k} F^*(\cdot,\cdot)$	The average full gradient of the client group U_t (33)
$\nabla_{\boldsymbol{\alpha}_i^k} f^*(\cdot,\cdot)$	The average stochastic mini-batch gradient of the client group U_t (34)
$\hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^*(\cdot,\cdot)$	The average stochastic mini-batch variance-reduced policy gradient of the client group U_t (35)
$\nabla_{\boldsymbol{\alpha}_i^k} F(\boldsymbol{\alpha}^k,\cdot)$	The average full gradient of all K clients for $\left\{\boldsymbol{\alpha}^k\right\}_{k=1}^{K}$ (32)
$\nabla_{\boldsymbol{\alpha}_i} F(\boldsymbol{\alpha},\cdot)$	The average full gradient of all K clients for $\boldsymbol{\alpha}$ (31)

Table 5: Notation Table

For $i = 1, ..., n$, we define the stochastic gradient, stochastic mini-batch gradient and full gradient with respect to $\boldsymbol{\alpha}_i^k$ as following:

$$\nabla_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}^k, \psi_m^k) \overset{def}{=} \nabla_{\boldsymbol{\alpha}_i^k} \mathbb{E}_{\Phi^k \sim \text{GS}(\boldsymbol{\alpha}^k)} \left[\mathcal{L}(\Phi^k, \psi_m^k)\right] \tag{17}$$

$$\nabla_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}^k, \mathcal{B}^k) \overset{def}{=} \nabla_{\boldsymbol{\alpha}_i^k} \frac{1}{B^k} \sum_{\psi_m^k \in \mathcal{B}^k} \mathbb{E}_{\Phi^k \sim \text{GS}(\boldsymbol{\alpha}^k)} [\mathcal{L}(\Phi^k, \psi_m^k)] \tag{18}$$

$$\nabla_{\boldsymbol{\alpha}_i^k} F^k(\boldsymbol{\alpha}^k, \Psi^k) \overset{def}{=} \nabla_{\boldsymbol{\alpha}_i^k} \frac{1}{M^k} \sum_{\psi_m^k \in \Psi^k} \mathbb{E}_{\Phi^k \sim \text{GS}(\boldsymbol{\alpha}^k)} [\mathcal{L}(\Phi^k, \psi_m^k)] \tag{19}$$

A.2 THE OMITTED DERIVATIVE PROCESS FOR EQ. 5

$$\begin{aligned}\frac{\partial \log p_{i,j_i^k}^k}{\partial \alpha_{i,j}^k} &= \frac{\partial}{\partial \alpha_{i,j}^k}\left(\log\left(\frac{\exp\left(\frac{\log(\alpha_{i,j_i^k}^k)+g_{i,j_i^k}^k}{\tau}\right)}{\sum_{l=1}^N \exp\left(\frac{\log(\alpha_{i,l}^k)+g_{i,l}^k}{\tau}\right)}\right)\right) \\ &= \frac{1}{p_{i,j_i^k}^k}\cdot\frac{\partial\left(\frac{\exp\left(\frac{\log(\alpha_{i,j_i^k}^k)+g_{i,j_i^k}^k}{\tau}\right)}{\sum_{l=1}^N \exp\left(\frac{\log(\alpha_{i,l}^k)+g_{i,l}^k}{\tau}\right)}\right)}{\partial\left(\frac{\log(\alpha_{i,j}^k)+g_{i,j}^k}{\tau}\right)}\cdot\frac{\partial\left(\frac{\log(\alpha_{i,j}^k)+g_{i,j}^k}{\tau}\right)}{\partial \alpha_{i,j}^k} \\ &= \frac{1}{\tau\alpha_{i,j}^k p_{i,j_i^k}^k}\cdot\frac{\partial\left(\frac{\exp\left(\frac{\log(\alpha_{i,j_i^k}^k)+g_{i,j_i^k}^k}{\tau}\right)}{\sum_{l=1}^N \exp\left(\frac{\log(\alpha_{i,l}^k)+g_{i,l}^k}{\tau}\right)}\right)}{\partial\left(\frac{\log(\alpha_{i,j}^k)+g_{i,j}^k}{\tau}\right)}\end{aligned} \tag{20}$$

According to the derivation rule of the Softmax function,
when $j = j_i^k$:

$$\frac{\partial p_{i,j_i^k}^k}{\partial \alpha_{i,j_i^k}^k} = \frac{1}{\tau\alpha_{i,j_i^k}^k p_{i,j_i^k}^k}\cdot\left(1-p_{i,j_i^k}^k\right)\cdot p_{i,j_i^k}^k = \frac{\left(1-p_{i,j_i^k}^k\right)}{\tau\alpha_{i,j_i^k}^k} \tag{21}$$

when $j \neq j_i^k$:

$$\frac{\partial p_{i,j_i^k}^k}{\partial \alpha_{i,j}^k} = \frac{1}{\tau\alpha_{i,j}^k p_{i,j_i^k}^k}\cdot\left(-p_{i,j_i^k}^k\cdot p_{i,j}^k\right) = \frac{-p_{i,j}^k}{\tau\alpha_{i,j}^k} \tag{22}$$

A.3 LEMMAS

The following lemma shows that the unbiasedness and bounded variance of variance-reduced policy gradient. This is important for bounding the randomness introduced by prompt sampling.

Lemma 1. ***Unbiasedness and bounded variance of variance-reduced policy gradient:*** *At the k-th client, $\forall\ \psi_m^k \in \Psi^k$, $r = 1, \cdots, I$ denotes the r-th sampling of Φ^k from $\boldsymbol{p}^k$ w.r.t. $\left\{\Phi^{k,r}\right\}_{r=1}^I \sim \boldsymbol{p}^k$, $\boldsymbol{p}^k = GS(\boldsymbol{\alpha}^k)$, $\alpha_{i,j} \geq \nu > 0$ for $i = 1, ..., n$ and $j = 1, ..., N$, $\tau > 0$ is the temperature parameter, and $\sigma_\alpha^2 = \frac{8G^2N}{\tau^2\nu^2}$, then the variance-reduced policy gradient is unbiased and bounded by :*

$$\mathbb{E}_{\{\Phi^{k,r}\sim GS(\boldsymbol{\alpha}^k)\}_{r=1}^I}\left[\hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}^k, \psi_m^k)\right] = \nabla_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}^k, \psi_m^k) \tag{23}$$

$$\mathbb{E}_{\{\Phi^{k,r}\sim GS(\boldsymbol{\alpha}^k)\}_{r=1}^I}\left[\left\|\hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}^k, \psi_m^k) - \mathbb{E}_{\{\Phi^{k,r}\sim GS(\boldsymbol{\alpha}^k)\}_{r=1}^I}\left[\hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}^k, \psi_m^k)\right]\right\|^2\right] \leq \frac{\sigma_\alpha^2}{I^2} \tag{24}$$

Proof. We abbreviate $\mathbb{E}_{\{\Phi^{k,r}\sim \mathrm{GS}(\boldsymbol{\alpha}^k)\}_{r=1}^I}$ as $\mathbb{E}_{\{\Phi^{k,r}\}_{r=1}^I}$.
1) The unbiasedness of variance-reduced policy gradient:

$$\begin{aligned}&\mathbb{E}_{\{\Phi^{k,r}\sim \mathrm{GS}(\boldsymbol{\alpha}^k)\}_{r=1}^I}\left[\hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}^k, \psi_m^k)\right] \\ &= \mathbb{E}_{\{\Phi^{k,r}\}_{r=1}^I}\left\{\frac{1}{I-1}\sum_{r=1}^I\left[\left(\mathcal{L}(\Phi^{k,r}, \psi_m^k) - \frac{1}{I}\sum_{w=1}^I \mathcal{L}(\Phi^{k,w}, \psi_m^k)\right)\nabla_{\boldsymbol{\alpha}_i^k}\log P(\phi_i^{k,r})\right]\right\} \\ &= \mathbb{E}_{\{\Phi^{k,r}\}_{r=1}^I}\left\{\frac{1}{I-1}\sum_{r=1}^I\left[\left(\frac{I-1}{I}\cdot\mathcal{L}(\Phi^{k,r}, \psi_m^k) - \frac{1}{I}\sum_{\substack{w=1\\ w\neq r}}^I \mathcal{L}(\Phi^{k,w}, \psi_m^k)\right)\nabla_{\boldsymbol{\alpha}_i^k}\log P(\phi_i^{k,r})\right]\right\}\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}_{\{\Phi^{k,r}\}_{r=1}^{I}}\left[\frac{1}{I}\sum_{r=1}^{I}\left(\mathcal{L}(\Phi^{k,r},\psi_m^k)\cdot\nabla_{\boldsymbol{\alpha}_i^k}\log P(\phi_i^{k,r})\right)\right] \\
&\quad - \mathbb{E}_{\{\Phi^{k,r}\}_{r=1}^{I}}\left\{\frac{1}{I}\sum_{r=1}^{I}\left[\left(\frac{1}{I-1}\sum_{\substack{w=1\\w\neq r}}^{I}\mathcal{L}(\Phi^{k,w},\psi_m^k)\right)\nabla_{\boldsymbol{\alpha}_i^k}\log P(\phi_i^{k,r})\right]\right\} \\
&\overset{(1)}{=} \frac{1}{I}\sum_{r=1}^{I}\mathbb{E}_{\Phi^{k,r}}\left[\mathcal{L}(\Phi^{k,r},\psi_m^k)\cdot\nabla_{\boldsymbol{\alpha}_i^k}\log P(\phi_i^{k,r})\right] \\
&\quad - \frac{1}{I}\sum_{r=1}^{I}\left[\left(\frac{1}{I-1}\sum_{\substack{w=1\\w\neq r}}^{I}\mathbb{E}_{\Phi^w}\mathcal{L}(\Phi^{k,w},\psi_m^k)\right)\mathbb{E}_{\Phi^r}\nabla_{\boldsymbol{\alpha}_i^k}\log P(\phi_i^{k,r})\right] \\
&= \mathbb{E}_{\Phi}\left[\mathcal{L}(\Phi^{k},\psi_m^k)\cdot\nabla_{\boldsymbol{\alpha}_i^k}\log P(\phi_i^{k})\right] \\
&\quad - \frac{1}{I}\sum_{r=1}^{I}\left[\left(\frac{1}{I-1}\sum_{\substack{w=1\\w\neq r}}^{I}\mathbb{E}_{\Phi^w}\left[\mathcal{L}(\Phi^{k,w},\psi_m^k)\right]\right)\mathbb{E}_{\Phi^r}\left[\nabla_{\boldsymbol{\alpha}_i^k}\log P(\phi_i^{k,r})\right]\right] \\
&= \nabla_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}^k,\psi_m^k) - \frac{1}{I}\sum_{r=1}^{I}\left(\frac{1}{I-1}\sum_{\substack{w=1\\w\neq r}}^{I}\mathbb{E}_{\Phi^w}\left[\mathcal{L}(\Phi^{k,w},\psi_m^k)\right]\right)\cdot\mathbb{E}_{\Phi^r}\left(\frac{\nabla_{\boldsymbol{\alpha}_i^k}P(\phi_i^{k,r})}{P(\phi_i^{k,r})}\right) \\
&= \nabla_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}^k,\psi_m^k) - \frac{1}{I}\sum_{r=1}^{I}\left(\frac{1}{I-1}\sum_{\substack{w=1\\w\neq r}}^{I}\mathbb{E}_{\Phi^w}\left[\mathcal{L}(\Phi^{k,w},\psi_m^k)\right]\right)\cdot\sum_{\phi_i^{k,r}\sim\boldsymbol{p}_i^k}\left(\frac{\nabla_{\boldsymbol{\alpha}_i^k}P(\phi_i^{k,r})}{P(\phi_i^{k,r})}\cdot P(\phi_i^{k,r})\right) \\
&= \nabla_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}^k,\psi_m^k) - \frac{1}{I}\sum_{r=1}^{I}\left(\frac{1}{I-1}\sum_{\substack{w=1\\w\neq r}}^{I}\mathbb{E}_{\Phi^w}\left[\mathcal{L}(\Phi^{k,w},\psi_m^k)\right]\right)\cdot\sum_{\phi_i^{k,r}\sim\boldsymbol{p}_i^k}\left(\nabla_{\boldsymbol{\alpha}_i^k}P(\phi_i^{k,r})\right) \\
&\overset{(2)}{=} \nabla_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}^k,\psi_m^k) - \frac{1}{I}\sum_{r=1}^{I}\left(\frac{1}{I-1}\sum_{\substack{w=1\\w\neq r}}^{I}\mathbb{E}_{\Phi^w}\left[\mathcal{L}(\Phi^{k,w},\psi_m^k)\right]\right)\cdot\nabla_{\boldsymbol{\alpha}_i^k}\left(\sum_{\phi_i^{k,r}\sim\boldsymbol{p}_i^k}P(\phi_i^{k,r})\right) \\
&\overset{(3)}{=} \nabla_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}^k,\psi_m^k) - \frac{1}{I}\sum_{r=1}^{I}\left(\frac{1}{I-1}\sum_{\substack{w=1\\w\neq r}}^{I}\mathbb{E}_{\Phi^w}\left[\mathcal{L}(\Phi^{k,w},\psi_m^k)\right]\right)\cdot\nabla_{\boldsymbol{\alpha}_i^k}(1)
\end{aligned}
$$

$$= \nabla_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}^k, \psi_m^k)$$

where (1) uses independence of each sampling for Φ^k; (2) is because n is not infinite and the GS function is continuous and derivable with respect to $\boldsymbol{\alpha}_i$; (3) uses the property that the elements of probability vector sum to 1.

2) The bounded variance of variance-reduced policy gradient:

$$\begin{aligned}
&\mathrm{Var}_{\{\Phi^{k,r}\sim \mathrm{GS}(\boldsymbol{\alpha}^k)\}_{r=1}^I}\left[\hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}^k, \psi_m^k)\right] \\
&= \mathbb{E}_{\{\Phi^{k,r}\sim \mathrm{GS}(\boldsymbol{\alpha}^k)\}_{r=1}^I}\left[\left\|\hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}^k, \psi_m^k) - \nabla_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}^k, \psi_m^k)\right\|^2\right] \\
&= \mathbb{E}_{\{\Phi^{k,r}\}_{r=1}^I}\left\{\left\|\frac{1}{I-1}\sum_{r=1}^{I}\left[\left(\mathcal{L}(\Phi^{k,r}, \psi_m^k) - \frac{1}{I}\sum_{w=1}^{I}\mathcal{L}(\Phi^{k,w}, \psi_m^k)\right)\nabla_{\boldsymbol{\alpha}_i^k}\log P(\phi_i^{k,r})\right] - \nabla_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}^k, \psi_m^k)\right\|^2\right\} \\
&= \mathbb{E}_{\{\Phi^{k,r}\}_{r=1}^I}\left\{\left\|\frac{1}{I}\sum_{r=1}^{I}\left[\frac{1}{I-1}\sum_{\substack{w=1\\w\neq r}}^{I}\left(\mathcal{L}(\Phi^{k,r}, \psi_m^k) - \mathcal{L}(\Phi^{k,w}, \psi_m^k)\right)\nabla_{\boldsymbol{\alpha}_i^k}\log P(\phi_i^{k,r}) - \nabla_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}^k, \psi_m^k)\right]\right\|^2\right\} \\
&\overset{(1)}{=} \frac{1}{I^2}\sum_{r=1}^{I}\mathbb{E}_{\Phi^{k,r}}\left[\left\|\frac{1}{I-1}\sum_{\substack{w=1\\w\neq r}}^{I}\left(\mathcal{L}(\Phi^{k,r}, \psi_m^k) - \mathcal{L}(\Phi^{k,w}, \psi_m^k)\right)\nabla_{\boldsymbol{\alpha}_i^k}\log P(\phi_i^{k,r}) - \nabla_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}^k, \psi_m^k)\right\|^2\right] \\
&\overset{(2)}{\leq} \frac{1}{I^2(I-1)^2}\sum_{r=1}^{I}\mathbb{E}_{\Phi^{k,r}}\left[\left\|\sum_{\substack{w=1\\w\neq r}}^{I}\left(\mathcal{L}(\Phi^{k,r}, \psi_m^k) - \mathcal{L}(\Phi^{k,w}, \psi_m^k)\right)\nabla_{\boldsymbol{\alpha}_i^k}\log P(\phi_i^{k,r})\right\|^2\right] \\
&\overset{(3)}{\leq} \frac{4G^2}{I^2(I-1)^2}\sum_{r=1}^{I}\mathbb{E}_{\Phi^{k,r}}\left[\left\|\sum_{\substack{w=1\\w\neq r}}^{I}\nabla_{\boldsymbol{\alpha}_i^k}\log P(\phi_i^{k,r})\right\|^2\right] \overset{(4)}{\leq} \frac{4G^2N}{I(I-1)\tau^2\nu^2} \overset{(5)}{\leq} \frac{8G^2N}{I^2\tau^2\nu^2}
\end{aligned}$$

where (1) uses

$$\begin{aligned}
&\mathbb{E}_{\{\Phi^{k,r}\}_{r=1}^I}\left[\frac{1}{I-1}\sum_{\substack{w=1\\w\neq r}}^{I}\left(\mathcal{L}(\Phi^{k,r}, \psi_m^k) - \mathcal{L}(\Phi^{k,w}, \psi_m^k)\right)\nabla_{\boldsymbol{\alpha}_i^k}\log P(\phi_i^{k,r}) - \nabla_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}^k, \psi_m^k)\right] \\
&= \frac{1}{I-1}\sum_{\substack{w=1\\w\neq r}}^{I}\mathbb{E}_{\Phi^{k,r},\Phi^{k,w},w\neq r}\left[\left(\mathcal{L}(\Phi^{k,r}, \psi_m^k) - \mathcal{L}(\Phi^{k,w}, \psi_m^k)\right)\nabla_{\boldsymbol{\alpha}_i^k}\log P(\phi_i^{k,r})\right] - \nabla_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}^k, \psi_m^k) \\
&= \frac{1}{I-1}\sum_{\substack{w=1\\w\neq r}}^{I}\mathbb{E}_{\Phi^{k,r}}\left[\mathcal{L}(\Phi^{k,r}, \psi_m^k)\nabla_{\boldsymbol{\alpha}_i^k}\log P(\phi_i^{k,r})\right] - \nabla_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}^k, \psi_m^k) \\
&= 0
\end{aligned}$$

and the independence of each sampling for Φ^k. (2) uses inequality $\mathbb{E}\left\|a-\mathbb{E}a\right\|^2 \leq \mathbb{E}\left\|a\right\|^2$; (3) uses Assumption 2; (4) uses $\alpha_{i,j}^{k,r} \geq \nu > 0$ and (5):

$$\nabla_{\boldsymbol{\alpha}_i^k} \log P(\phi_i^{k,r}) \leq \sqrt{N \cdot \max\left\{\left|\frac{1-p_{i,j_i}^{k,r}}{\tau\alpha_{i,j_i}^{k,r}}\right|, \left|-\frac{p_{i,j}^{k,r}}{\tau\alpha_{i,j}^{k,r}}\right|\right\}^2} \leq \sqrt{\frac{N}{\tau^2\nu^2}}$$

5) is because when $I \geq 2$:

$$\frac{1}{I(I-1)} \leq \frac{2}{I^2}$$

□

The following lemma shows that the $\mathbb{E}_{\Phi^k\sim \mathrm{GS}(\boldsymbol{\alpha}^k)}\left[\mathcal{L}(\Phi^k,\Psi^k)\right]$ is L-smooth for $\boldsymbol{\alpha}$. This is crucial for the later convergence analysis of BDPL and Fed-BDPL and is a necessity for convergence proofs.

Lemma 2. ***L-smooth for $\boldsymbol{\alpha}^k$:*** *At the k-th client, the Φ^k is sampled from probability matrix $\boldsymbol{p}^k$, and $\boldsymbol{p}^k = GS(\boldsymbol{\alpha}^k)$, $\alpha_{i,j} \geq \nu > 0$ for $i = 1, ..., n$ and $j = 1, ..., N$, $\tau > 0$ is the temperature parameter, $\mathbb{E}_{\Phi^k\sim GS(\boldsymbol{\alpha}^k)}\left[\mathcal{L}(\Phi^k,\Psi^k)\right]$ is L-smooth for $\boldsymbol{\alpha}^k$ and $L = \frac{nGN(\tau+1)}{\tau^2\nu^2}$, and then for t-th iteration, the following inequality is satisfied:*

$$\begin{aligned}
&\mathbb{E}_{\Phi_{t+1}^k\sim GS(\boldsymbol{\alpha}_{t+1}^k)}\left[\mathcal{L}(\Phi_{t+1}^k,\Psi^k)\right] - \mathbb{E}_{\Phi_t^k\sim GS(\boldsymbol{\alpha}_t^k)}\left[\mathcal{L}(\Phi_t^k,\Psi^k)\right] \\
\leq &\left\langle \nabla_{\boldsymbol{\alpha}^k}\mathbb{E}_{\Phi_t^k\sim GS(\boldsymbol{\alpha}_t^k)}\left[\mathcal{L}(\Phi_t^k,\Psi^k)\right], \boldsymbol{\alpha}_{t+1}^k - \boldsymbol{\alpha}_t^k\right\rangle + \frac{L}{2}\left\|\boldsymbol{\alpha}_{t+1}^k - \boldsymbol{\alpha}_t^k\right\|^2
\end{aligned}$$

Proof. The objective function:

$$\mathbb{E}_{\Phi^k\sim \mathrm{GS}(\boldsymbol{\alpha}^k)}\left[\mathcal{L}(\Phi^k,\Psi^k)\right] = \sum_{\phi_1^k\sim \mathrm{GS}(\boldsymbol{\alpha}_1^k)} \cdots \sum_{\phi_n^k\sim \mathrm{GS}(\boldsymbol{\alpha}_n^k)} \left(\mathcal{L}(\Phi^k,\Psi^k)\prod_{i=1}^n P(\phi_i^k)\right)$$

We can compute the Hessian of the objective function, $\forall i', i'' \in 1, \cdots, n$ and $j', j'' \in 1, \cdots, N$: If $i' \neq i''$:

$$\begin{aligned}
&\frac{\partial^2}{\partial\alpha_{i',j'}\partial\alpha_{i'',j''}}\mathbb{E}_{\Phi^k\sim \mathrm{GS}(\boldsymbol{\alpha}^k)}\left[\mathcal{L}(\Phi^k,\Psi^k)\right] \\
= &\sum_{\phi_1^k\sim\boldsymbol{p}_1^k} \cdots \sum_{\phi_n^k\sim\boldsymbol{p}_n^k} \left(\mathcal{L}(\Phi^k,\Psi^k)\frac{\partial^2}{\partial\alpha_{i',j'}\partial\alpha_{i'',j''}}\prod_{i=1}^n P(\phi_i^k)\right) \\
= &\sum_{\phi_1^k\sim\boldsymbol{p}_1^k} \cdots \sum_{\phi_{i'-1}^k\sim\boldsymbol{p}_{i'-1}^k} \sum_{\phi_{i'+1}^k\sim\boldsymbol{p}_{i'+1}^k} \cdots \sum_{\phi_{i''-1}^k\sim\boldsymbol{p}_{i''-1}^k} \sum_{\phi_{i''+1}^k\sim\boldsymbol{p}_{i''+1}^k} \cdots \sum_{\phi_n^k\sim\boldsymbol{p}_n^k} \\
&\left(\sum_{\phi_{i'}^k\sim\boldsymbol{p}_{i'}^k}\sum_{\phi_{i''}^k\sim\boldsymbol{p}_{i''}^k}\left(\mathcal{L}(\Phi^k,\Psi^k)\frac{\partial^2}{\partial\alpha_{i',j'}\partial\alpha_{i'',j''}}\prod_{i=1}^n P(\phi_i^k)\right)\right) \\
= &\sum_{\phi_1^k\sim\boldsymbol{p}_1^k} \cdots \sum_{\phi_{i'-1}^k\sim\boldsymbol{p}_{i'-1}^k} \sum_{\phi_{i'+1}^k\sim\boldsymbol{p}_{i'+1}^k} \cdots \sum_{\phi_{i''-1}^k\sim\boldsymbol{p}_{i''-1}^k} \sum_{\phi_{i''+1}^k\sim\boldsymbol{p}_{i''+1}^k} \cdots \sum_{\phi_n^k\sim\boldsymbol{p}_n^k} \\
&\left(\mathcal{L}(\Phi^k,\Psi^k)\frac{\partial^2\left[P(\phi_{i'}^k)P(\phi_{i''}^k)\right]}{\partial\alpha_{i',j'}\partial\alpha_{i'',j''}} \prod_{\substack{i=1\\ i\neq i'\\ i\neq i''}}^n P(\phi_i^k)\right)
\end{aligned}$$

Then:

$$\frac{\partial^2\left[P(\phi_{i'}^k)P(\phi_{i''}^k)\right]}{\partial\alpha_{i',j'}\partial\alpha_{i'',j''}}=\begin{cases}\frac{1-p_{i',j_i'}^k}{\tau\alpha_{i',j_i'}^k}\cdot\frac{1-p_{i'',j_i''}^k}{\tau\alpha_{i'',j_i''}^k}, \text{ if } j'=j_i' \text{ and } j''=j_i''\\ \frac{1-p_{i',j_i'}^k}{\tau\alpha_{i',j_i'}^k}\cdot\left(-\frac{p_{i'',j''}^k}{\tau\alpha_{i'',j''}^k}\right), \text{ if } j'=j_i' \text{ and } j''\neq j_i''\\ -\frac{p_{i',j'}^k}{\tau\alpha_{i',j'}^k}\cdot\frac{1-p_{i'',j_i''}^k}{\tau\alpha_{i'',j_i''}^k}, \text{ if } j'\neq j_i' \text{ and } j''=j_i''\\ -\frac{p_{i',j'}^k}{\tau\alpha_{i',j'}^k}\cdot\left(-\frac{p_{i'',j''}^k}{\tau\alpha_{i'',j''}^k}\right), \text{ if } j'\neq j_i' \text{ and } j''\neq j_i''\end{cases} \tag{25}$$

Based on $\alpha_{i,j}^k\geq\nu>0$:

$$\left|\frac{\partial^2\left[P(\phi_{i'}^k)P(\phi_{i''}^k)\right]}{\partial\alpha_{i',j'}\partial\alpha_{i'',j''}}\right|\leq\frac{1}{\tau^2\nu^2} \tag{26}$$

Further, based on Assumption 2:

$$\left|\frac{\partial^2}{\partial\alpha_{i',j'}\partial\alpha_{i'',j''}}\mathbb{E}_{\Phi^k\sim\mathrm{GS}(\boldsymbol{\alpha}^k)}\left[\mathcal{L}(\Phi^k,\Psi^k)\right]\right|$$

$$\leq\sum_{\phi_1^k\sim\boldsymbol{p}_1^k}\cdots\sum_{\phi_{i'-1}^k\sim\boldsymbol{p}_{i'-1}^k}\sum_{\phi_{i'+1}^k\sim\boldsymbol{p}_{i'+1}^k}\cdots\sum_{\phi_{i''-1}^k\sim\boldsymbol{p}_{i''-1}^k}\sum_{\phi_{i''+1}^k\sim\boldsymbol{p}_{i''+1}^k}\cdots\sum_{\phi_n^k\sim\boldsymbol{p}_n^k}\left(\mathcal{L}(\Phi^k,\Psi^k)\prod_{\substack{i=1\\i\neq i'\\i\neq i''}}^{n}P(\phi_i^k)\right)\cdot\frac{1}{\tau^2\nu^2}$$

$$\leq\frac{G}{\tau^2\nu^2}$$

If $i'=i''$:

$$\frac{\partial^2}{\partial\alpha_{i',j'}\partial\alpha_{i',j''}}\mathbb{E}_{\Phi^k\sim\mathrm{GS}(\boldsymbol{\alpha}^k)}\left[\mathcal{L}(\Phi^k,\Psi^k)\right]$$

$$=\sum_{\phi_1^k\sim\boldsymbol{p}_1^k}\cdots\sum_{\phi_{i'-1}^k\sim\boldsymbol{p}_{i'-1}^k}\sum_{\phi_{i'+1}^k\sim\boldsymbol{p}_{i'+1}^k}\cdots\sum_{\phi_n^k\sim\boldsymbol{p}_n^k}\left(\mathcal{L}(\Phi^k,\Psi^k)\frac{\partial^2\left[P(\phi_{i'}^k)\right]}{\partial\alpha_{i',j'}\partial\alpha_{i',j''}}\prod_{\substack{i=1\\i\neq i'}}^{n}P(\phi_i^k)\right)$$

Similar to the analysis in case $i'\neq i''$, we can get:

$$\left|\frac{\partial^2\left[P(\phi_{i'}^k)\right]}{\partial\alpha_{i',j'}\partial\alpha_{i',j''}}\right|\leq\max\left\{p,1-p\right\}\cdot\frac{(\tau+1)}{\tau^2\nu^2}\leq\frac{\tau+1}{\tau^2\nu^2}$$

$$\left|\frac{\partial^2}{\partial\alpha_{i',j'}\partial\alpha_{i',j''}}\mathbb{E}_{\Phi^k\sim\mathrm{GS}(\boldsymbol{\alpha}^k)}\left[\mathcal{L}(\Phi^k,\Psi^k)\right]\right|\leq\frac{(\tau+1)\,G}{\tau^2\nu^2} \tag{27}$$

Finally, with $H(\boldsymbol{\alpha})$ denoting the Hessian matrix of $\mathbb{E}_{\Phi^k\sim\mathrm{GS}(\boldsymbol{\alpha}^k)}\left[\mathcal{L}(\Phi^k,\Psi^k)\right]$, we can get:

$$\left\|H(\boldsymbol{\alpha})\right\|_2\leq\left\|H(\boldsymbol{\alpha})\right\|_F\leq\sqrt{n(n-1)N^2\left(\frac{G}{\tau^2\nu^2}\right)^2+nN^2\left(\frac{(\tau+1)\,G}{\tau^2\nu^2}\right)^2}\leq\frac{nGN(\tau+1)}{\tau^2\nu^2} \tag{28}$$

According to Lemma 1.2.2 in (Nesterov et al., 2018), $\mathbb{E}_{\Phi^k\sim\mathrm{GS}(\boldsymbol{\alpha}^k)}\left[\mathcal{L}(\Phi^k,\Psi^k)\right]$ is L-smooth for $\boldsymbol{\alpha}^k$ and $L=\frac{nGN(\tau+1)}{\tau^2\nu^2}$, □

Lemma 3. ***Convergence of BDPL****: Suppose Assumption 1 and 2 hold, for* $t = 0, ..., T-1$, $\alpha_{i,j} \geq \nu > 0$ *for* $i = 1, ..., n$ *and* $j = 1, ..., N$, $\tau > 0$ *is the temperature parameter,* $\mathbb{E}_{\Phi^k \sim GS(\boldsymbol{\alpha}^k)}\left[\mathcal{L}(\Phi^k, \Psi^k)\right]$ *is L-smooth for* $\boldsymbol{\alpha}^k$ *and* $L = \frac{nGN(\tau+1)}{\tau^2\nu^2}$, σ_ψ^2 *is the variance of the stochastic gradient,* $\sigma_\alpha^2 = \frac{8G^2N}{\tau^2\nu^2}$ *is the variance of the variance-reduced policy gradient. Let* $\eta \leq \frac{1}{L}$, *then the BDPL's full gradient* $\nabla F^k(\boldsymbol{\alpha}_t^k, \Psi^k)$ *satisfies the following inequality:*

$$\frac{1}{T}\sum_{t=0}^{T-1}\left\|\nabla_{\boldsymbol{\alpha}^k} F^k(\boldsymbol{\alpha}_t^k, \Psi^k)\right\|^2 \leq \frac{4G}{T\eta} + \frac{2\eta n L\sigma_\psi^2}{B^k} + \frac{2\eta n L\sigma_\alpha^2}{I^2} \tag{29}$$

Proof. According to Lemma 2:

$$\begin{aligned}
&\mathbb{E}_{\Phi_{t+1}^k \sim \mathrm{GS}(\boldsymbol{\alpha}_{t+1}^k)}\left[\mathcal{L}(\Phi_{t+1}^k, \Psi^k)\right] - \mathbb{E}_{\Phi_t^k \sim \mathrm{GS}(\boldsymbol{\alpha}_t^k)}\left[\mathcal{L}(\Phi_t^k, \Psi^k)\right] \\
\leq& \left\langle \nabla_{\boldsymbol{\alpha}^k} \mathbb{E}_{\Phi_t^k \sim \mathrm{GS}(\boldsymbol{\alpha}_t^k)}\left[\mathcal{L}(\Phi_t^k, \Psi^k)\right], \boldsymbol{\alpha}_{t+1}^k - \boldsymbol{\alpha}_t^k \right\rangle + \frac{L}{2}\left\|\boldsymbol{\alpha}_{t+1}^k - \boldsymbol{\alpha}_t^k\right\|^2 \\
\leq& \sum_{i=1}^{n}\left[\left\langle \nabla_{\boldsymbol{\alpha}_i^k} F^k(\boldsymbol{\alpha}_t^k, \Psi^k), -\eta \cdot \hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}_t^k, \mathcal{B}_t^k)\right\rangle + \frac{L\eta^2}{2}\left\|\hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}_t^k, \mathcal{B}_t^k)\right\|^2\right]
\end{aligned}$$

Both sides take expectations for $\mathbb{E}_{\{\Phi_t^r \sim \mathrm{GS}(\boldsymbol{\alpha}_t)\}_{r=1}^I}$ and $\mathbb{E}_{\mathcal{B}_t}$ at the same time, we abbreviate $\mathbb{E}_{\{\Phi^r \sim \mathrm{GS}(\boldsymbol{\alpha}\}_{r=1}^I}$ as $\mathbb{E}_{\{\Phi^r\}_{r=1}^I}$:

$$\begin{aligned}
&\mathbb{E}_{\{\Phi^r\}_{r=1}^I}\mathbb{E}_{\mathcal{B}_t}\left\{\mathbb{E}_{\Phi_{t+1}^k \sim \mathrm{GS}(\boldsymbol{\alpha}_{t+1}^k)}\left[\mathcal{L}(\Phi_{t+1}^k, \Psi^k)\right] - \mathbb{E}_{\Phi_t^k \sim \mathrm{GS}(\boldsymbol{\alpha}_t^k)}\left[\mathcal{L}(\Phi_t^k, \Psi^k)\right]\right\} \\
\leq& \mathbb{E}_{\{\Phi^r\}_{r=1}^I}\mathbb{E}_{\mathcal{B}_t}\left\{\sum_{i=1}^{n}\left[\left\langle \nabla_{\boldsymbol{\alpha}_i^k} F^k(\boldsymbol{\alpha}_t^k, \Psi^k), -\eta \cdot \hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}_t^k, \mathcal{B}_t^k)\right\rangle + \frac{L\eta^2}{2}\left\|\hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}_t^k, \mathcal{B}_t^k)\right\|^2\right]\right\} \\
=& \sum_{i=1}^{n}\underbrace{\mathbb{E}_{\{\Phi^r\}_{r=1}^I}\mathbb{E}_{\mathcal{B}_t}\left[\left\langle \nabla_{\boldsymbol{\alpha}_i^k} F^k(\boldsymbol{\alpha}_t^k, \Psi^k), -\eta \cdot \hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}_t^k, \mathcal{B}_t^k)\right\rangle + \frac{L\eta^2}{2}\left\|\hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}_t^k, \mathcal{B}_t^k)\right\|^2\right]}_{a)}
\end{aligned}$$

For a):

$$\begin{aligned}
&\mathbb{E}_{\{\Phi^r\}_{r=1}^I}\mathbb{E}_{\mathcal{B}_t}\left[\left\langle \nabla_{\boldsymbol{\alpha}_i^k} F^k(\boldsymbol{\alpha}_t^k, \Psi^k), -\eta \cdot \hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}_t^k, \mathcal{B}_t^k)\right\rangle + \frac{L\eta^2}{2}\left\|\hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}_t^k, \mathcal{B}_t^k)\right\|^2\right] \\
=& \left\langle \nabla_{\boldsymbol{\alpha}_i^k} F^k(\boldsymbol{\alpha}_t^k, \Psi^k), -\eta \cdot \mathbb{E}_{\{\Phi^r\}_{r=1}^I}\mathbb{E}_{\mathcal{B}_t}\left[\hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}_t^k, \mathcal{B}_t^k)\right]\right\rangle + \frac{L\eta^2}{2}\mathbb{E}_{\{\Phi^r\}_{r=1}^I}\mathbb{E}_{\mathcal{B}_t}\left[\left\|\hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}_t^k, \mathcal{B}_t^k)\right\|^2\right] \\
\overset{(1)}{=}& \left\langle \nabla_{\boldsymbol{\alpha}_i^k} F^k(\boldsymbol{\alpha}_t^k, \Psi^k), -\eta \cdot \mathbb{E}_{\mathcal{B}_t}\mathbb{E}_{\{\Phi^r\}_{r=1}^I}\left[\hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}_t^k, \mathcal{B}_t^k)\right]\right\rangle \\
&+ \frac{L\eta^2}{2}\left\{\left\|\mathbb{E}_{\{\Phi^r\}_{r=1}^I}\mathbb{E}_{\mathcal{B}_t}\left[\hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}_t^k, \mathcal{B}_t^k)\right]\right\|^2 + \mathrm{Var}_{\mathcal{B}_t, \{\Phi^r\}_{r=1}^I}\left[\hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}_t^k, \mathcal{B}_t^k)\right]\right\} \\
\overset{(2)}{=}& -\eta \cdot \left\langle \nabla_{\boldsymbol{\alpha}_i^k} F^k(\boldsymbol{\alpha}_t^k, \Psi^k), \nabla_{\boldsymbol{\alpha}_i^k} F^k(\boldsymbol{\alpha}_t^k, \Psi^k)\right\rangle + \frac{L\eta^2}{2}\left\|\nabla_{\boldsymbol{\alpha}_i^k} F^k(\boldsymbol{\alpha}_t^k, \Psi^k)\right\|^2 \\
&+ \frac{L\eta^2}{2}\mathbb{E}_{\{\Phi^r\}_{r=1}^I}\mathbb{E}_{\mathcal{B}_t}\left[\left\|\hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}_t^k, \mathcal{B}_t^k) - \mathbb{E}_{\{\Phi^r\}_{r=1}^I}\mathbb{E}_{\mathcal{B}_t}\left[\hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}_t^k, \mathcal{B}_t^k)\right]\right\|^2\right] \\
\overset{(3)}{\leq}& \left(\frac{L\eta^2}{2} - \eta\right)\left\|\nabla_{\boldsymbol{\alpha}_i^k} F^k(\boldsymbol{\alpha}_t^k, \Psi^k)\right\|^2 \\
&+ L\eta^2 \cdot \mathbb{E}_{\{\Phi^r\}_{r=1}^I}\mathbb{E}_{\mathcal{B}_t}\left[\left\|\hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}_t^k, \mathcal{B}_t^k) - \mathbb{E}_{\mathcal{B}_t}\left[\hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}_t^k, \mathcal{B}_t^k)\right]\right\|^2\right] \\
&+ L\eta^2 \cdot \mathbb{E}_{\mathcal{B}_t}\mathbb{E}_{\{\Phi^r\}_{r=1}^I}\left[\left\|\mathbb{E}_{\mathcal{B}_t}\left[\hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}_t^k, \mathcal{B}_t^k)\right] - \mathbb{E}_{\{\Phi^r\}_{r=1}^I}\mathbb{E}_{\mathcal{B}_t}\left[\hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}_t^k, \mathcal{B}_t^k)\right]\right\|^2\right] \\
=& \left(\frac{L\eta^2}{2} - \eta\right)\left\|\nabla_{\boldsymbol{\alpha}_i^k} F^k(\boldsymbol{\alpha}_t^k, \Psi^k)\right\|^2 + L\eta^2 \cdot \mathbb{E}_{\{\Phi^r\}_{r=1}^I}\mathrm{Var}_{\mathcal{B}_t}\left[\hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}_t^k, \mathcal{B}_t^k)\right] \\
&+ L\eta^2 \cdot \mathbb{E}_{\mathcal{B}_t}\mathrm{Var}_{\{\Phi^r\}_{r=1}^I}\left[\mathbb{E}_{\mathcal{B}_t}\left[\hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}_t^k, \mathcal{B}_t^k)\right]\right]
\end{aligned}$$

$$\overset{(4)}{\leq} \left(\frac{L\eta^2}{2} - \eta\right) \left\|\nabla_{\boldsymbol{\alpha}_i^k} F^k(\boldsymbol{\alpha}_t^k, \Psi^k)\right\|^2 + L\eta^2 \cdot \frac{\sigma_\psi^2}{B} + L\eta^2 \cdot \mathbb{E}_{\mathcal{B}_t} \mathrm{Var}_{\{\Phi^r\}_{r=1}^I} \left[\hat{\nabla}_{\boldsymbol{\alpha}_i} f^k(\boldsymbol{\alpha}_t^k, \Psi^k)\right]$$
$$\overset{(5)}{\leq} \left(\frac{L\eta^2}{2} - \eta\right) \left\|\nabla_{\boldsymbol{\alpha}_i^k} F^k(\boldsymbol{\alpha}_t^k, \Psi^k)\right\|^2 + L\eta^2 \cdot \frac{\sigma_\psi^2}{B} + L\eta^2 \cdot \frac{\sigma_\alpha^2}{I^2}$$

where (1) uses the independence between Ψ sampling and Φ sampling; (2) use the unbiasedness of stochastic gradient and variance-reduced policy gradient in Assumption 1 and Lemma 1; (3) use the inequality $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$; (4) and (5) use the bounded variance of stochastic gradient and variance-reduced policy gradient in Assumption 1 and Lemma 1.
Then:

$$\mathbb{E}_{\Phi_{t+1} \sim \mathrm{GS}(\boldsymbol{\alpha}_{t+1}^k)} \left[\mathcal{L}(\Phi_{t+1}^k, \Psi^k)\right] - \mathbb{E}_{\Phi_t^k \sim \mathrm{GS}(\boldsymbol{\alpha}_t^k)} \left[\mathcal{L}(\Phi_t^k, \Psi^k)\right]$$
$$\leq \sum_{i=1}^{n} \left[\left(\frac{L\eta^2}{2} - \eta\right) \left\|\nabla_{\boldsymbol{\alpha}_i^k} F^k(\boldsymbol{\alpha}_t^k, \Psi^k)\right\|^2 + L\eta^2 \cdot \frac{\sigma_\psi^2}{B} + L\eta^2 \cdot \frac{\sigma_\alpha^2}{I^2}\right]$$

We combine the gradient of $\boldsymbol{\alpha}_i$ for each prompt token:

$$\mathbb{E}_{\Phi_{t+1}^k \sim \mathrm{GS}(\boldsymbol{\alpha}_{t+1})} \left[\mathcal{L}(\Phi_{t+1}^k, \Psi^k)\right] - \mathbb{E}_{\Phi_t^k \sim \mathrm{GS}(\boldsymbol{\alpha}_t^k)} \left[\mathcal{L}(\Phi_t^k, \Psi^k)\right]$$
$$\leq \left(\frac{L\eta^2}{2} - \eta\right) \left\|\nabla_{\boldsymbol{\alpha}^k} F^k(\boldsymbol{\alpha}_t^k, \Psi^k)\right\|^2 + nL\eta^2 \cdot \frac{\sigma_\psi^2}{B} + nL\eta^2 \cdot \frac{\sigma_\alpha^2}{I^2}$$

where $\boldsymbol{\alpha}^k = (\boldsymbol{\alpha}_1^k, \cdots, \boldsymbol{\alpha}_i^k, \cdots \boldsymbol{\alpha}_n^k)$.
We let $\eta \leq \frac{1}{L}$, then both sides accumulate with respect to $t = 0, 1, \cdots, T-1$ and divide by T:

$$\frac{1}{T} \sum_{t=0}^{T-1} \left(\eta - \frac{L\eta^2}{2}\right) \left\|\nabla_{\boldsymbol{\alpha}^k} F^k(\boldsymbol{\alpha}_t^k, \Psi^k)\right\|^2$$
$$\leq \frac{1}{T} \sum_{t=0}^{T-1} \left[\mathbb{E}_{\Phi_t^k \sim \mathrm{GS}(\boldsymbol{\alpha}_t^k)} \left[\mathcal{L}(\Phi_t^k, \Psi^k)\right] - \mathbb{E}_{\Phi_{t+1}^k \sim \mathrm{GS}(\boldsymbol{\alpha}_{t+1}^k)} \left[\mathcal{L}(\Phi_{t+1}^k, \Psi^k)\right]\right] + \frac{nL\eta^2\sigma_\psi^2}{B} + \frac{nL\eta^2\sigma_\alpha^2}{I^2}$$

Then,

$$\frac{1}{T} \sum_{t=0}^{T-1} \left\|\nabla_{\boldsymbol{\alpha}^k} F^k(\boldsymbol{\alpha}_t^k, \Psi^k)\right\|^2$$
$$\leq \frac{\mathbb{E}_{\Phi_0^k \sim \mathrm{GS}(\boldsymbol{\alpha}_0^k)} \left[\mathcal{L}(\Phi_0^k, \Psi^k)\right] - \mathbb{E}_{\Phi_T^k \sim \mathrm{GS}(\boldsymbol{\alpha}_T^k)} \left[\mathcal{L}(\Phi_T^k, \Psi^k)\right]}{T} \frac{2}{2\eta - L\eta^2} + \frac{2\eta}{2 - L\eta} \left(\frac{nL\sigma_\psi^2}{B} + \frac{nL\sigma_\alpha^2}{I^2}\right)$$
$$\leq \frac{2\left(\mathbb{E}_{\Phi_0^k \sim \mathrm{GS}(\boldsymbol{\alpha}_0^k)} \left[\mathcal{L}(\Phi_0^k, \Psi^k)\right] - \mathbb{E}_{\Phi_T^k \sim \mathrm{GS}(\boldsymbol{\alpha}_T^k)} \left[\mathcal{L}(\Phi_T^k, \Psi^k)\right]\right)}{T\eta} + \frac{2\eta nL\sigma_\psi^2}{B} + \frac{2\eta nL\sigma_\alpha^2}{I^2}$$

According to Assumption 2 , $\mathbb{E}_{\Phi_0^k \sim \mathrm{GS}(\boldsymbol{\alpha}_0^k)} \left[\mathcal{L}(\Phi_0^k, \Psi^k)\right] - \inf_t \mathbb{E}_{\Phi_t^k \sim \mathrm{GS}(\boldsymbol{\alpha}_t^k)} \left[\mathcal{L}(\Phi_t^k, \Psi^k)\right] \leq 2G$, then:

$$\frac{1}{T} \sum_{t=0}^{T-1} \left\|\nabla_{\boldsymbol{\alpha}^k} F^k(\boldsymbol{\alpha}_t^k, \Psi^k)\right\|^2 \leq \frac{4G}{T\eta} + \frac{2\eta nL\sigma_\psi^2}{B} + \frac{2\eta nL\sigma_\alpha^2}{I^2}$$

□

Remark 2. The convergence term of BDPL consists of three parts. The first term is typical of first-order optimization algorithms converging to non-convex functions. The second term is the stochasticity due to random mini-batch gradients. The first term and second term can be combined when $\eta = \frac{c}{\sqrt{T}}$ and $c = \sqrt{\frac{2B^kG}{nL\sigma_\psi^2}}$:

$$\frac{4G}{T\eta} + \frac{2\eta nL\sigma_\psi^2}{B^k} = \frac{4}{\sqrt{T}} \sqrt{\frac{2nL\sigma_\psi^2 G}{B}},$$

which can decrease as the number of iterations and mini-batch size increase Reddi et al. (2016); and the third term is the stochasticity due to prompt sampling, which decreases with the number of prompt samples.

Remark 3. According to the definition of Gumbel-Softmax (2), there is randomness about $\boldsymbol{u} = \{u_i\}_{i=1}^n \sim \text{Uniform}(\mathbf{0}, \mathbf{1}_n)$ in $\nabla_{\boldsymbol{\alpha}} F^k(\boldsymbol{\alpha}_t^k, \Psi^k)$ i.e. $\boldsymbol{p}^k = \text{GS}(\boldsymbol{\alpha}^k, \boldsymbol{u})$ in fact and $\Phi^k \sim \boldsymbol{p}^k$, we can further discuss the result in Lemma 3:

$$\mathbb{E}_{\boldsymbol{u}} \left\| \nabla_{\boldsymbol{\alpha}^k} F^k(\boldsymbol{\alpha}_t^k, \Psi^k) \right\|^2 = \int_{\mathbf{0}}^{\mathbf{1}} \left\| \nabla_{\boldsymbol{\alpha}^k} F^k(\boldsymbol{\alpha}_t^k, \Psi^k) \right\|^2 \mathbf{1}_n d\boldsymbol{u} \leq \left\| \nabla_{\boldsymbol{\alpha}^k} F^k(\boldsymbol{\alpha}_t^k, \Psi^k) \right\|^2$$

Considering the randomness of $\boldsymbol{u}$, we can still obtain the convergence of BDPL. In subsequent analyzes of Fed-BDPL, we can obtain similar result for $\boldsymbol{u}$.

Corollary 3. ***Convergence rate of BDPL****: Let* $\eta = \min\left\{\frac{1}{L}, \frac{1}{\sqrt{T}}\right\}$*, we can get the following convergence rate for BDPL:*

$$\frac{1}{T}\sum_{t=0}^{T-1} \left\| \nabla_{\boldsymbol{\alpha}^k} F^k(\boldsymbol{\alpha}_t^k, \Psi^k) \right\|^2 \leq \mathcal{O}(\frac{1}{\sqrt{T}}) \tag{30}$$

Proof. Convergence rate:

$$\frac{1}{T}\sum_{t=0}^{T-1} \left\| \nabla_{\boldsymbol{\alpha}^k} F^k(\boldsymbol{\alpha}_t^k, \Psi^k) \right\|^2 = \frac{4G}{\sqrt{T}} + \frac{1}{\sqrt{T}} \cdot \frac{2nL\sigma_\psi^2}{B} + \frac{1}{\sqrt{T}} \cdot \frac{2nL\sigma_\alpha^2}{I^2} = \mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$$

□

A.4 Convergence Analysis of Federated Prompt Tuning

Definition in the Federated Black-box Discrete Prompt Learning:

Data slicing: The data slices in each client are defined as $D^k = \left\{D^k\right\}_{k=1}^K = \left\{\Psi^k, Y^k\right\}_{k=1}^K$, where $\Psi^k = \left\{\psi_m^k\right\}_{m=1}^{M^k}$ denote the input sentence and $Y^k = \left\{y_m^k\right\}_{m=1}^{M^k}$ denote the label, k is the index of the client, and m is the index of the sample, there are totally M^k samples in the dataset D^k. The clients hold the sampling probability vector $\boldsymbol{q} = \left\{q^{[k]}\right\}_{k=1}^K = \left\{\frac{M^k}{M}\right\}_{k=1}^K$. K_* is the number of clients selected and U_t is the set of corresponding clients.

Average parameter $\boldsymbol{\alpha}$:

$$\boldsymbol{\alpha}_t = \frac{1}{K_*}\sum_{k\in U_t} \boldsymbol{\alpha}_t^k$$

$$\boldsymbol{\alpha}_{i,(t)} = \frac{1}{K_*}\sum_{k\in U_t} \boldsymbol{\alpha}_{i,(t)}^k$$

Average loss:

$$\mathbb{E}_{\Phi_t \sim \text{GS}(\boldsymbol{\alpha}_t)} \left[\mathcal{L}(\Phi_t, \Psi)\right] = \sum_{k=1}^K q^{[k]} \cdot \mathbb{E}_{\Phi_t^k \sim \text{GS}(\boldsymbol{\alpha}_t^k)} \left[\mathcal{L}(\Phi_t^k, \Psi^k)\right]$$

Average gradient:

$$\nabla_{\boldsymbol{\alpha}_i} F(\boldsymbol{\alpha}, \Psi^k) \overset{def}{=} \sum_{k=1}^K q^{[k]} \cdot \nabla_{\boldsymbol{\alpha}_i} F^k(\boldsymbol{\alpha}, \Psi^k) \tag{31}$$

$$\nabla_{\boldsymbol{\alpha}_i^k} F(\boldsymbol{\alpha}^k, \Psi^k) \overset{def}{=} \sum_{k=1}^K q^{[k]} \cdot \nabla_{\boldsymbol{\alpha}_i^k} F^k(\boldsymbol{\alpha}^k, \Psi^k) \tag{32}$$

$$\nabla_{\boldsymbol{\alpha}_i^k} F^*(\boldsymbol{\alpha}^k, \Psi^k) \overset{def}{=} \frac{1}{K_*}\sum_{k\in U_t} \nabla_{\boldsymbol{\alpha}_i^k} F^k(\boldsymbol{\alpha}^k, \Psi^k) \tag{33}$$

$$\nabla_{\boldsymbol{\alpha}_i^k} f^*(\boldsymbol{\alpha}^k, \mathcal{B}^k) \overset{def}{=} \frac{1}{K_*}\sum_{k\in U_t} \nabla_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}^k, \mathcal{B}^k) \tag{34}$$

$$\hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^*(\boldsymbol{\alpha}^k, \mathcal{B}^k) \overset{def}{=} \frac{1}{K_*} \sum_{k \in U_t} \hat{\nabla}_{\boldsymbol{\alpha}_i} f^k(\boldsymbol{\alpha}^k, \mathcal{B}^k) \tag{35}$$

Average mini-batch stochastic variance-reduced policy gradient descent:

$$\boldsymbol{\alpha}_{i,(t+1)} = \boldsymbol{\alpha}_{i,(t)} - \eta \cdot \hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^*(\boldsymbol{\alpha}_t^k, \mathcal{B}^k)$$

Note that when $t \neq sE$, although the above some definitions don't exist in algorithm and experiment, we can still calculate and analyze them. In order to facilitate the analysis of the iteration process, we assume that they exist.

The following lemma shows that the local gradient is biased compared to the global gradient due to the fact that the distribution of data on different clients may be different (heterogeneity), and the bias of the gradient can be analyzed using the bias about $\boldsymbol{\alpha}$ (Haddadpour & Mahdavi (2019)).

Lemma 4. ***Bound bias between local and average*** $\boldsymbol{\alpha}$. *Let* $\alpha_{i,j} \geq \nu > 0$ *for* $i = 1, ..., n$ *and* $j = 1, ..., N$, η *is the learning rate, let* $\mathbb{E}$ *represents* $\mathbb{E}_{\left\{\Phi_t^{k,r} \sim GS(\boldsymbol{\alpha}_t^k)\right\}_{r=1}^I}$ *and* $\mathbb{E}_{\mathcal{B}_t^k}$, *the bias between the local and the average* $\boldsymbol{\alpha}$ *can be bounded:*

$$\begin{aligned}
&\frac{1}{T} \sum_{t=0}^{T-1} \sum_{k=1}^{K} q^{[k]} \cdot \mathbb{E} \left\| \boldsymbol{\alpha}_{i,(t)} - \boldsymbol{\alpha}_{i,(t)}^k \right\|^2 \\
&\leq E\eta^2 \left(\frac{2\sigma_\psi^2}{B} + \frac{2\sigma_\alpha^2}{I^2} \right) \left(1 + \frac{1}{K_*} \right) + \frac{2\eta^2 E^2 \lambda}{T} \left(1 + \frac{1}{K_*} \right) \sum_{t=0}^{T-1} \left\| \nabla_{\boldsymbol{\alpha}_i^k} F(\boldsymbol{\alpha}_t^k, \Psi^k) \right\|^2
\end{aligned}$$

Proof. Define:

$$t_c \triangleq \lfloor \frac{t}{E} \rfloor E$$

$$\boldsymbol{\alpha}_{i,(t_c)} = \frac{1}{K_*} \sum_{k \in U_{t_c}} \boldsymbol{\alpha}_{i,(t_c)}^k$$

Then, for $t_c + 1 \leq t < t_c + E$:

$$\boldsymbol{\alpha}_{i,(t)}^k = \boldsymbol{\alpha}_{i,(t_c)} - \sum_{\rho=t_c}^{t-1} \eta \cdot \hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}_\rho^k, \mathcal{B}_\rho^k) \tag{36}$$

$$\boldsymbol{\alpha}_{i,(t)} = \boldsymbol{\alpha}_{i,(t_c)} - \frac{1}{K_*} \sum_{k \in U_\rho} \sum_{\rho=t_c}^{t-1} \eta \cdot \hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}_\rho^k, \mathcal{B}_\rho^k) \tag{37}$$

For the k-th client, let $\mathbb{E}$ represents $\mathbb{E}_{\{\Phi^{k,r}\}_{r=1}^I}$ and $\mathbb{E}_{\mathcal{B}_t^k}$, and we take $\mathbb{E}$ for $\left\| \boldsymbol{\alpha}_{i,(t)} - \boldsymbol{\alpha}_{i,(t)}^k \right\|^2$:

$$\begin{aligned}
&\mathbb{E} \left\| \boldsymbol{\alpha}_{i,(t)} - \boldsymbol{\alpha}_{i,(t)}^k \right\|^2 \\
&= \mathbb{E} \left\| \boldsymbol{\alpha}_{i,(t_c)} - \frac{1}{K_*} \sum_{k \in U_\rho} \sum_{\rho=t_c}^{t-1} \eta \cdot \hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}_\rho^k, \mathcal{B}_\rho^k) - \boldsymbol{\alpha}_{i,(t_c)} + \sum_{\rho=t_c}^{t-1} \eta \cdot \hat{\nabla}_{\boldsymbol{\alpha}_i} f^k(\boldsymbol{\alpha}_\rho^k, \mathcal{B}_\rho^k) \right\|^2 \\
&= \mathbb{E} \left\| \sum_{\rho=t_c}^{t-1} \eta \cdot \hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}_\rho^k, \mathcal{B}_\rho^k) - \frac{1}{K_*} \sum_{k \in U_\rho} \sum_{\rho=t_c}^{t-1} \eta \cdot \hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}_\rho^k, \mathcal{B}_\rho^k) \right\|^2 \\
&\overset{(1)}{\leq} 2 \left[\mathbb{E} \left\| \sum_{\rho=t_c}^{t-1} \eta \cdot \hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}_\rho^k, \mathcal{B}_\rho^k) \right\|^2 + \mathbb{E} \left\| \frac{1}{K_*} \sum_{k \in U_\rho} \sum_{\rho=t_c}^{t-1} \eta \cdot \hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}_\rho^k, \mathcal{B}_\rho^k) \right\|^2 \right] \\
&\overset{(2)}{=} 2\mathbb{E} \left\| \sum_{\rho=t_c}^{t-1} \eta \cdot \hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}_\rho^k, \mathcal{B}_\rho^k) - \mathbb{E} \left[\sum_{\rho=t_c}^{t-1} \eta \cdot \hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}_\rho^k, \mathcal{B}_\rho^k) \right] \right\|^2 + 2 \left\| \mathbb{E} \left[\sum_{\rho=t_c}^{t-1} \eta \cdot \hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^k(\boldsymbol{\alpha}_\rho^k, \mathcal{B}_\rho^k) \right] \right\|^2
\end{aligned}$$

$$
\begin{aligned}
&+2\mathbb{E}\left\|\frac{1}{K_*}\sum_{k\in U_\rho}\sum_{\rho=t_c}^{t-1}\eta\cdot\hat{\nabla}_{\boldsymbol{\alpha}_i^k}f^k(\boldsymbol{\alpha}_\rho^k,\mathcal{B}_\rho^k)-\mathbb{E}\left[\frac{1}{K_*}\sum_{k\in U_\rho}\sum_{\rho=t_c}^{t-1}\eta\cdot\hat{\nabla}_{\boldsymbol{\alpha}_i^k}f^k(\boldsymbol{\alpha}_\rho^k,\mathcal{B}_\rho^k)\right]\right\|^2\\
&+2\left\|\mathbb{E}\left[\frac{1}{K_*}\sum_{k\in U_\rho}\sum_{\rho=t_c}^{t-1}\eta\cdot\hat{\nabla}_{\boldsymbol{\alpha}_i^k}f^k(\boldsymbol{\alpha}_\rho^k,\mathcal{B}_\rho^k)\right]\right\|^2\\
\overset{(3)}{=}\;&2\mathbb{E}\left\|\sum_{\rho=t_c}^{t-1}\eta\cdot\hat{\nabla}_{\boldsymbol{\alpha}_i^k}f^k(\boldsymbol{\alpha}_\rho^k,\mathcal{B}_\rho^k)-\sum_{\rho=t_c}^{t-1}\eta\cdot\nabla_{\boldsymbol{\alpha}_i^k}F^k(\boldsymbol{\alpha}_\rho^k,\Psi^k)\right\|^2+2\left\|\sum_{\rho=t_c}^{t-1}\eta\cdot\nabla_{\boldsymbol{\alpha}_i^k}F^k(\boldsymbol{\alpha}_\rho^k,\Psi^k)\right\|^2\\
&+2\mathbb{E}\left\|\frac{1}{K_*}\sum_{k\in U_\rho}\sum_{\rho=t_c}^{t-1}\eta\cdot\hat{\nabla}_{\boldsymbol{\alpha}_i^k}f^k(\boldsymbol{\alpha}_\rho^k,\mathcal{B}_\rho^k)-\frac{1}{K_*}\sum_{k\in U_\rho}\sum_{\rho=t_c}^{t-1}\eta\cdot\nabla_{\boldsymbol{\alpha}_i^k}F^k(\boldsymbol{\alpha}_\rho^k,\Psi^k)\right\|^2\\
&+2\left\|\frac{1}{K_*}\sum_{k\in U_\rho}\sum_{\rho=t_c}^{t-1}\eta\cdot\nabla_{\boldsymbol{\alpha}_i^k}F^k(\boldsymbol{\alpha}_\rho^k,\Psi^k)\right\|^2\\
=\;&2\mathbb{E}\left\|\sum_{\rho=t_c}^{t-1}\eta\cdot\left[\hat{\nabla}_{\boldsymbol{\alpha}_i^k}f^k(\boldsymbol{\alpha}_\rho^k,\mathcal{B}_\rho^k)-\nabla_{\boldsymbol{\alpha}_i^k}F^k(\boldsymbol{\alpha}_\rho^k,\Psi^k)\right]\right\|^2+2\left\|\sum_{\rho=t_c}^{t-1}\eta\cdot\nabla_{\boldsymbol{\alpha}_i^k}F^k(\boldsymbol{\alpha}_\rho^k,\Psi^k)\right\|^2\\
&+2\mathbb{E}\left\|\frac{1}{K_*}\sum_{k\in U_\rho}\sum_{\rho=t_c}^{t-1}\eta\cdot\left[\hat{\nabla}_{\boldsymbol{\alpha}_i^k}f^k(\boldsymbol{\alpha}_\rho^k,\mathcal{B}_\rho^k)-\nabla_{\boldsymbol{\alpha}_i^k}F^k(\boldsymbol{\alpha}_\rho^k,\Psi^k)\right]\right\|^2+2\left\|\frac{1}{K_*}\sum_{k\in U_\rho}\sum_{\rho=t_c}^{t-1}\eta\cdot\nabla_{\boldsymbol{\alpha}_i^k}F^k(\boldsymbol{\alpha}_\rho^k,\Psi^k)\right\|^2\\
\overset{(4)}{=}\;&2\eta^2\sum_{\rho=t_c}^{t-1}\mathbb{E}\left\|\hat{\nabla}_{\boldsymbol{\alpha}_i^k}f^k(\boldsymbol{\alpha}_\rho^k,\mathcal{B}_\rho^k)-\nabla_{\boldsymbol{\alpha}_i^k}F^k(\boldsymbol{\alpha}_\rho^k,\Psi^k)\right\|^2+2\left\|\sum_{\rho=t_c}^{t-1}\eta\cdot\nabla_{\boldsymbol{\alpha}_i^k}F^k(\boldsymbol{\alpha}_\rho^k,\Psi^k)\right\|^2\\
&+\frac{2\eta^2}{K_*^2}\sum_{k\in U_\rho}\sum_{\rho=t_c}^{t-1}\mathbb{E}\left\|\hat{\nabla}_{\boldsymbol{\alpha}_i^k}f^k(\boldsymbol{\alpha}_\rho^k,\mathcal{B}_\rho^k)-\nabla_{\boldsymbol{\alpha}_i^k}F^k(\boldsymbol{\alpha}_\rho^k,\Psi^k)\right\|^2+2\left\|\frac{1}{K_*}\sum_{k\in U_\rho}\sum_{\rho=t_c}^{t-1}\eta\cdot\nabla_{\boldsymbol{\alpha}_i^k}F^k(\boldsymbol{\alpha}_\rho^k,\Psi^k)\right\|^2\\
\overset{(5)}{=}\;&2\eta^2\sum_{\rho=t_c}^{t-1}\mathbb{E}\left\|\hat{\nabla}_{\boldsymbol{\alpha}_i^k}f^k(\boldsymbol{\alpha}_\rho^k,\mathcal{B}_\rho^k)-\nabla_{\boldsymbol{\alpha}_i^k}F^k(\boldsymbol{\alpha}_\rho^k,\Psi^k)\right\|^2+2\eta^2(t-t_c)\sum_{\rho=t_c}^{t-1}\left\|\nabla_{\boldsymbol{\alpha}_i^k}F^k(\boldsymbol{\alpha}_\rho^k,\Psi^k)\right\|^2\\
&+\frac{2\eta^2}{K_*^2}\sum_{k\in U_\rho}\sum_{\rho=t_c}^{t-1}\mathbb{E}\left\|\hat{\nabla}_{\boldsymbol{\alpha}_i^k}f^k(\boldsymbol{\alpha}_\rho^k,\mathcal{B}_\rho^k)-\nabla_{\boldsymbol{\alpha}_i^k}F^k(\boldsymbol{\alpha}_\rho^k,\Psi^k)\right\|^2+\frac{2\eta^2(t-t_c)}{K_*}\sum_{k\in U_\rho}\sum_{\rho=t_c}^{t-1}\left\|\nabla_{\boldsymbol{\alpha}_i^k}F^k(\boldsymbol{\alpha}_\rho^k,\Psi^k)\right\|^2\\
\overset{(6)}{\leq}\;&2\eta^2\sum_{\rho=t_c}^{t-1}\left(\frac{2\sigma_\psi^2}{B^k}+\frac{2\sigma_\alpha^2}{I^2}\right)+\frac{2\eta^2}{K_*^2}\sum_{k\in U_\rho}\sum_{\rho=t_c}^{t-1}\left(\frac{2\sigma_\psi^2}{B^k}+\frac{2\sigma_\alpha^2}{I^2}\right)\\
&+2\eta^2(t-t_c)\sum_{\rho=t_c}^{t-1}\left\|\nabla_{\boldsymbol{\alpha}_i^k}F^k(\boldsymbol{\alpha}_\rho^k,\Psi^k)\right\|^2+\frac{2\eta^2(t-t_c)}{K_*}\sum_{k\in U_\rho}\sum_{\rho=t_c}^{t-1}\left\|\nabla_{\boldsymbol{\alpha}_i^k}F^k(\boldsymbol{\alpha}_\rho^k,\Psi^k)\right\|^2\\
=\;&2\eta^2\sum_{\rho=t_c}^{t-1}\left(\frac{2\sigma_\psi^2}{B^k}+\frac{2\sigma_\alpha^2}{I^2}\right)\left(1+\frac{1}{K_*}\right)\\
&+2\eta^2(t-t_c)\left(\sum_{\rho=t_c}^{t-1}\left\|\nabla_{\boldsymbol{\alpha}_i^k}F^k(\boldsymbol{\alpha}_\rho^k,\Psi^k)\right\|^2+\frac{1}{K_*}\sum_{k\in U_\rho}\sum_{\rho=t_c}^{t-1}\left\|\nabla_{\boldsymbol{\alpha}_i^k}F^k(\boldsymbol{\alpha}_\rho^k,\Psi^k)\right\|^2\right)
\end{aligned}
$$

where (1) use inequality $\|a-b\|^2\leq 2\|a\|^2+2\|b\|^2$; (2) use $\mathbb{E}\|x-\mathbb{E}[x]\|^2=\mathbb{E}\|x\|^2-\|\mathbb{E}[x]\|^2$; (3) use the unbiasedness of stochastic gradient and variance-reduced policy gradient in Assumption 1 and Lemma 1; (4) use the independence of mini-batch and prompt sampling in each client; (5) use inequality $\|\sum_{z=1}^{Z}a_z\|^2\leq Z\sum_{z=1}^{Z}\|a_z\|^2$; (6) use the bounded variance of stochastic gradient and variance-reduced policy gradient in Assumption 1 and Lemma 1.

Then, we sum both sides with respect to $k \in U_\rho$:

$$\sum_{k\in U_\rho} \mathbb{E}\left\|\boldsymbol{\alpha}_{i,(t)} - \boldsymbol{\alpha}_{i,(t)}^k\right\|^2$$
$$\leq 2\eta^2 \sum_{\rho=t_c}^{t-1}\left(\frac{2\sigma_\psi^2}{B^k} + \frac{2\sigma_\alpha^2}{I^2}\right)(K_*+1) + 2\eta^2(t-t_c)\left(1+\frac{1}{K_*}\right)\sum_{k\in U_\rho}\sum_{\rho=t_c}^{t-1}\left\|\nabla_{\boldsymbol{\alpha}_i^k} F^k(\boldsymbol{\alpha}_\rho^k, \Psi^k)\right\|^2$$

We take the expectation of both sides about $\mathbb{E}_{U_\rho}$:

$$\mathbb{E}_{U_\rho}\left[\sum_{k\in U_\rho} \mathbb{E}\left\|\boldsymbol{\alpha}_{i,(t)} - \boldsymbol{\alpha}_{i,(t)}^k\right\|^2\right]$$
$$= K_* \sum_{k=1}^{K} q^{[k]} \cdot \mathbb{E}\left\|\boldsymbol{\alpha}_{i,(t)} - \boldsymbol{\alpha}_{i,(t)}^k\right\|^2$$
$$\overset{(1)}{\leq} 2\eta^2 \sum_{\rho=t_c}^{t-1}\left(\frac{2\sigma_\psi^2}{B^k} + \frac{2\sigma_\alpha^2}{I^2}\right)(K_*+1) + 2\eta^2(t-t_c)(K_*+1)\sum_{k=1}^{K}\sum_{\rho=t_c}^{t-1} q^{[k]} \cdot \left\|\nabla_{\boldsymbol{\alpha}_i^k} F^k(\boldsymbol{\alpha}_\rho^k, \Psi^k)\right\|^2$$

where (1) is because we sample client set U_ρ uniformly at random where client k is sampled with probability $q^{[k]}$ for $1 \leq k \leq K$ with replacement, and we define $U_\rho = \{k_1, ..., k_b, ..., k_{K_*}\}$, then

$$\begin{aligned}
&\mathbb{E}_{U_\rho}\left[\sum_{k\in U_\rho}\left\|\nabla_{\boldsymbol{\alpha}_i^k} F^k(\boldsymbol{\alpha}_\rho^k, \Psi^k)\right\|^2\right] \\
&= \sum_{b=1}^{K_*} \mathbb{E}_{k_b}\left[\left\|\nabla_{\boldsymbol{\alpha}_i^k} F^k(\boldsymbol{\alpha}_\rho^{k_b}, \Psi^{k_b})\right\|^2\right] \\
&= \sum_{b=1}^{K_*}\sum_{k=1}^{K} q^{[k]} \cdot \left\|\nabla_{\boldsymbol{\alpha}_i^k} F^k(\boldsymbol{\alpha}_\rho^k, \Psi^k)\right\|^2 \\
&= K_* \sum_{k=1}^{K} q^{[k]} \cdot \left\|\nabla_{\boldsymbol{\alpha}_i^k} F^k(\boldsymbol{\alpha}_\rho^k, \Psi^k)\right\|^2
\end{aligned}$$

Thus:

$$\sum_{k=1}^{K} q^{[k]} \cdot \mathbb{E}\left\|\boldsymbol{\alpha}_{i,(t)} - \boldsymbol{\alpha}_{i,(t)}^k\right\|^2$$
$$\leq 2\eta^2 \sum_{\rho=t_c}^{t-1}\left(\frac{2\sigma_\psi^2}{B^k} + \frac{2\sigma_\alpha^2}{I^2}\right)\left(1+\frac{1}{K_*}\right) + 2\eta^2(t-t_c)\left(1+\frac{1}{K_*}\right)\sum_{k=1}^{K}\sum_{\rho=t_c}^{t-1} q^{[k]} \cdot \left\|\nabla_{\boldsymbol{\alpha}_i^k} F^k(\boldsymbol{\alpha}_\rho^k, \Psi^k)\right\|^2$$

We take $\upsilon = \rho - t_c$, $\gamma = t - 1 - t_c$, and add iteration on both sides:

$$\sum_{t=0}^{T-1}\sum_{k=1}^{K} q^{[k]} \cdot \mathbb{E}\left\|\boldsymbol{\alpha}_{i,(t)} - \boldsymbol{\alpha}_{i,(t)}^k\right\|^2$$
$$= \sum_{s=0}^{\lfloor\frac{T-1}{E}\rfloor}\sum_{\gamma=0}^{E-1}\sum_{k=1}^{K} q^{[k]} \cdot \mathbb{E}\left\|\boldsymbol{\alpha}_{i,(sE+\gamma)} - \boldsymbol{\alpha}_{i,(sE+\gamma)}^k\right\|^2$$
$$\leq \sum_{s=0}^{\lfloor\frac{T-1}{E}\rfloor}\sum_{\gamma=0}^{E-1} 2\eta^2 \sum_{\upsilon=0}^{\gamma}\left(\frac{2\sigma_\psi^2}{B^k} + \frac{2\sigma_\alpha^2}{I^2}\right)\left(1+\frac{1}{K_*}\right)$$
$$+ \sum_{s=0}^{\lfloor\frac{T-1}{E}\rfloor}\sum_{\gamma=0}^{E-1} 2\eta^2(\gamma+1)\left(1+\frac{1}{K_*}\right)\sum_{k=1}^{K}\sum_{\upsilon=0}^{\gamma} q^{[k]} \cdot \left\|\nabla_{\boldsymbol{\alpha}_i^k} F^k(\boldsymbol{\alpha}_{sE+\upsilon}^k, \Psi^k)\right\|^2$$

$$
\begin{aligned}
&\overset{(1)}{\leq} (E+1)T\eta^2\left(\frac{2\sigma_\psi^2}{B^k}+\frac{2\sigma_\alpha^2}{I^2}\right)\left(1+\frac{1}{K_*}\right) \\
&\quad+\sum_{s=0}^{\lfloor\frac{T-1}{E}\rfloor} 2\eta^2E^2\left(1+\frac{1}{K_*}\right)\sum_{k=1}^{K}\sum_{\upsilon=0}^{E-1} q^{[k]}\cdot\left\|\nabla_{\boldsymbol{\alpha}_i^k}F^k(\boldsymbol{\alpha}_{sE+\upsilon}^k,\Psi^k)\right\|^2 \\
&\leq (E+1)T\eta^2\left(\frac{2\sigma_\psi^2}{B^k}+\frac{2\sigma_\alpha^2}{I^2}\right)\left(1+\frac{1}{K_*}\right)+2\eta^2E^2\left(1+\frac{1}{K_*}\right)\sum_{k=1}^{K}\sum_{t=0}^{T-1} q^{[k]}\cdot\left\|\nabla_{\boldsymbol{\alpha}_i^k}F^k(\boldsymbol{\alpha}_t^k,\Psi^k)\right\|^2
\end{aligned}
$$

where (1) use $0\leq\gamma\leq E-1$ and $1+...+E=\frac{E(E+1)}{2}$.
Finally multiply both sides simultaneously by $\frac{1}{T}$:

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\sum_{k=1}^{K} q^{[k]}\cdot\mathbb{E}\left\|\boldsymbol{\alpha}_{i,(t)}-\boldsymbol{\alpha}_{i,(t)}^k\right\|^2 \\
&\leq (E+1)\eta^2\left(\frac{2\sigma_\psi^2}{B^k}+\frac{2\sigma_\alpha^2}{I^2}\right)\left(1+\frac{1}{K_*}\right)+\frac{2\eta^2E^2}{T}\left(1+\frac{1}{K_*}\right)\sum_{k=1}^{K}\sum_{t=0}^{T-1} q^{[k]}\cdot\left\|\nabla_{\boldsymbol{\alpha}_i^k}F^k(\boldsymbol{\alpha}_t^k,\Psi^k)\right\|^2 \\
&\overset{(1)}{\leq} (E+1)\eta^2\left(\frac{2\sigma_\psi^2}{B^k}+\frac{2\sigma_\alpha^2}{I^2}\right)\left(1+\frac{1}{K_*}\right)+\frac{2\eta^2E^2\lambda}{T}\left(1+\frac{1}{K_*}\right)\sum_{t=0}^{T-1}\left\|\sum_{k=1}^{K} q^{[k]}\cdot\nabla_{\boldsymbol{\alpha}_i^k}F^k(\boldsymbol{\alpha}_t^k,\Psi^k)\right\|^2 \\
&= (E+1)\eta^2\left(\frac{2\sigma_\psi^2}{B^k}+\frac{2\sigma_\alpha^2}{I^2}\right)\left(1+\frac{1}{K_*}\right)+\frac{2\eta^2E^2\lambda}{T}\left(1+\frac{1}{K_*}\right)\sum_{t=0}^{T-1}\left\|\nabla_{\boldsymbol{\alpha}_i^k}F(\boldsymbol{\alpha}_t^k,\Psi^k)\right\|^2
\end{aligned}
$$

where (1) use Assumption 3. □

Theorem 1. Suppose Assumption 1, 2 and 3 hold, for $t=0,1,...,T-1$, $B=\min\left\{B^1,...,B^K\right\}$ where B^k is the local mini-batch size. $\alpha_{i,j}\geq\nu>0$ for $i=1,...,n$ and $j=1,...,N$, I is the sampling times for prompt. $\sigma_\alpha^2=\frac{8G^2N}{\tau^2\nu^2}$ is the variance of the variance-reduced policy gradient, σ_ψ^2 is the variance of the stochastic gradient, $\mathbb{E}_{\Phi^k\sim\mathrm{GS}(\boldsymbol{\alpha}^k)}\left[\mathcal{L}(\Phi^k,\Psi^k)\right]$ is L-smooth for $\boldsymbol{\alpha}^k$ and $L=\frac{nGN(\tau+1)}{\tau^2\nu^2}$, and η satisfies the following inequality:

$$
0<\eta\leq\eta^*=\frac{-\lambda L+\sqrt{\lambda^2L^2+8L^2E\left(1+\frac{1}{K_*}\right)}}{8L^2E\left(1+\frac{1}{K_*}\right)} \tag{38}
$$

Then, the Fed-BDPL's full gradient $\nabla_{\boldsymbol{\alpha}}F(\boldsymbol{\alpha}_t,\Psi^k)$ satisfies the following inequality:

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\left\|\nabla_{\boldsymbol{\alpha}}F(\boldsymbol{\alpha}_t,\Psi^k)\right\|^2 \\
&\leq\frac{4G}{\eta T}+\frac{2(E+1)L^2\eta^2n\sigma_\psi^2(1+\frac{1}{K_*})+2nL\eta\sigma_\psi^2}{B}+\frac{2(E+1)L^2\eta^2n\sigma_\alpha^2(1+\frac{1}{K_*})+2nL\eta\sigma_\alpha^2}{I^2}
\end{aligned}
$$

Proof. According to Lemma 2:

$$
\begin{aligned}
&\mathbb{E}_{\Phi_{t+1}\sim\mathrm{GS}(\boldsymbol{\alpha}_{t+1})}\left[\mathcal{L}(\Phi_{t+1},\Psi)\right]-\mathbb{E}_{\Phi_t\sim\mathrm{GS}(\boldsymbol{\alpha}_t)}\left[\mathcal{L}(\Phi_t,\Psi)\right] \\
&\leq\left\langle\nabla_{\boldsymbol{\alpha}}\mathbb{E}_{\Phi_t\sim\mathrm{GS}(\boldsymbol{\alpha}_t)}\left[\mathcal{L}(\Phi_t,\Psi)\right],\boldsymbol{\alpha}_{t+1}-\boldsymbol{\alpha}_t\right\rangle+\frac{L}{2}\left\|\boldsymbol{\alpha}_{t+1}-\boldsymbol{\alpha}_t\right\|^2 \\
&\leq\sum_{i=1}^{n}\left[\left\langle\nabla_{\boldsymbol{\alpha}_i}F(\boldsymbol{\alpha}_t,\Psi^k),-\eta\cdot\hat{\nabla}_{\boldsymbol{\alpha}_i^k}f^*(\boldsymbol{\alpha}_t^k,\mathcal{B}_t^k)\right\rangle+\frac{L\eta^2}{2}\left\|\hat{\nabla}_{\boldsymbol{\alpha}_i^k}f^*(\boldsymbol{\alpha}_t^k,\mathcal{B}_t^k)\right\|^2\right]
\end{aligned}
$$

We take the expectations about $\left\{\Phi^{k,r}\right\}_{r=1}^{I}$, $\mathcal{B}_t^k$ and U_t on both sides respectively:

$$
\mathbb{E}_{\left\{\Phi^{k,r}\right\}_{r=1}^{I}}\mathbb{E}_{\mathcal{B}_t^k}\mathbb{E}_{U_t}\left\{\mathbb{E}_{\Phi_{t+1}\sim\mathrm{GS}(\boldsymbol{\alpha}_{t+1})}\left[\mathcal{L}(\Phi_{t+1},\Psi)\right]-\mathbb{E}_{\Phi_t\sim\mathrm{GS}(\boldsymbol{\alpha}_t)}\left[\mathcal{L}(\Phi_t,\Psi)\right]\right\}
$$

$$\leq \sum_{i=1}^{n} \underbrace{\left\{ \mathbb{E}_{\{\Phi^{k,r}\}_{r=1}^{I}} \mathbb{E}_{\mathcal{B}_t^k} \mathbb{E}_{U_t} \left[\left\langle \nabla_{\boldsymbol{\alpha}_i} F(\boldsymbol{\alpha}_t, \Psi^k), -\eta \cdot \hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^*(\boldsymbol{\alpha}_t^k, \mathcal{B}_t^k) \right\rangle \right] \right\}}_{a)}$$

$$+ \sum_{i=1}^{n} \underbrace{\left\{ \frac{L\eta^2}{2} \mathbb{E}_{\{\Phi^{k,r}\}_{r=1}^{I}} \mathbb{E}_{\mathcal{B}_t^k} \mathbb{E}_{U_t} \left[\left\| \hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^*(\boldsymbol{\alpha}_t^k, \mathcal{B}_t^k) \right\|^2 \right] \right\}}_{b)}$$

For a):

$$\begin{aligned}
&\mathbb{E}_{\{\Phi^{k,r}\}_{r=1}^{I}} \mathbb{E}_{\mathcal{B}_t^k} \mathbb{E}_{U_t} \left[\left\langle \nabla_{\boldsymbol{\alpha}_i} F(\boldsymbol{\alpha}_t, \Psi^k), -\eta \cdot \hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^*(\boldsymbol{\alpha}_t^k, \mathcal{B}_t^k) \right\rangle \right] \\
&= \left\langle \nabla_{\boldsymbol{\alpha}_i} F(\boldsymbol{\alpha}_t, \Psi^k), -\eta \cdot \mathbb{E}_{\{\Phi^{k,r}\}_{r=1}^{I}} \mathbb{E}_{\mathcal{B}_t^k} \mathbb{E}_{U_t} \left[\hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^*(\boldsymbol{\alpha}_t^k, \mathcal{B}_t^k) \right] \right\rangle \\
&= \left\langle \nabla_{\boldsymbol{\alpha}_i} F(\boldsymbol{\alpha}_t, \Psi^k), -\eta \cdot \mathbb{E}_{U_t} \mathbb{E}_{\{\Phi^{k,r}\}_{r=1}^{I}} \mathbb{E}_{\mathcal{B}_t^k} \left[\frac{1}{K_*} \sum_{k \in U_t} \hat{\nabla}_{\boldsymbol{\alpha}_i} f^k(\boldsymbol{\alpha}^k, \Psi^k) \right] \right\rangle \\
&= \left\langle \nabla_{\boldsymbol{\alpha}_i} F(\boldsymbol{\alpha}_t, \Psi^k), -\eta \cdot \mathbb{E}_{U_t} \left[\frac{1}{K_*} \sum_{k \in U_t} \nabla_{\boldsymbol{\alpha}_i^k} F^k(\boldsymbol{\alpha}^k, \Psi^k) \right] \right\rangle \\
&= \left\langle \nabla_{\boldsymbol{\alpha}_i} F(\boldsymbol{\alpha}_t, \Psi^k), -\eta \cdot \frac{1}{K_*} \left[K_* \sum_{k=1}^{K} q^{[k]} \cdot \nabla_{\boldsymbol{\alpha}_i^k} F^k(\boldsymbol{\alpha}^k, \Psi^k) \right] \right\rangle \\
&\overset{(1)}{=} \frac{\eta}{2} \left[- \left\| \nabla_{\boldsymbol{\alpha}_i} F(\boldsymbol{\alpha}_t, \Psi^k) \right\|^2 - \left\| \nabla_{\boldsymbol{\alpha}_i^k} F(\boldsymbol{\alpha}_t^k, \Psi^k) \right\|^2 + \left\| \nabla_{\boldsymbol{\alpha}_i} F(\boldsymbol{\alpha}_t, \Psi^k) - \sum_{k=1}^{K} q^{[k]} \cdot \nabla_{\boldsymbol{\alpha}_i^k} F^k(\boldsymbol{\alpha}^k, \Psi^k) \right\|^2 \right] \\
&= \frac{\eta}{2} \left[- \left\| \nabla_{\boldsymbol{\alpha}_i} F(\boldsymbol{\alpha}_t, \Psi^k) \right\|^2 - \left\| \nabla_{\boldsymbol{\alpha}_i^k} F(\boldsymbol{\alpha}_t^k, \Psi^k) \right\|^2 + \left\| \sum_{k=1}^{K} q^{[k]} \cdot \left(\nabla_{\boldsymbol{\alpha}_i} F^k(\boldsymbol{\alpha}_t, \Psi^k) - \nabla_{\boldsymbol{\alpha}_i^k} F^k(\boldsymbol{\alpha}^k, \Psi^k) \right) \right\|^2 \right] \\
&\overset{(2)}{\leq} \frac{\eta}{2} \left[- \left\| \nabla_{\boldsymbol{\alpha}_i} F(\boldsymbol{\alpha}_t, \Psi^k) \right\|^2 - \left\| \nabla_{\boldsymbol{\alpha}_i^k} F(\boldsymbol{\alpha}_t^k, \Psi^k) \right\|^2 + \sum_{k=1}^{K} q^{[k]} \cdot \left\| \nabla_{\boldsymbol{\alpha}_i} F^k(\boldsymbol{\alpha}_t, \Psi^k) - \nabla_{\boldsymbol{\alpha}_i^k} F^k(\boldsymbol{\alpha}^k, \Psi^k) \right\|^2 \right] \\
&= \frac{\eta}{2} \left[- \left\| \nabla_{\boldsymbol{\alpha}_i} F(\boldsymbol{\alpha}_t, \Psi^k) \right\|^2 - \left\| \nabla_{\boldsymbol{\alpha}_i^k} F(\boldsymbol{\alpha}_t^k, \Psi^k) \right\|^2 \right] \\
&\quad + \frac{\eta}{2} \cdot \sum_{k=1}^{K} q^{[k]} \cdot \left\| \frac{1}{M^k} \sum_{\psi_m^k \in \Psi^k} \nabla_{\boldsymbol{\alpha}_i} \mathbb{E}_{\Phi_t \sim \mathrm{GS}(\boldsymbol{\alpha}_t)} [\mathcal{L}(\Phi_t, \psi_m^k)] - \frac{1}{M^k} \sum_{\psi_m^k \in \Psi^k} \nabla_{\boldsymbol{\alpha}_i^k} \mathbb{E}_{\Phi_t^k \sim \mathrm{GS}(\boldsymbol{\alpha}_t^k)} [\mathcal{L}(\Phi_t^k, \psi_m^k)] \right\|^2 \\
&\overset{(3)}{\leq} -\frac{\eta}{2} \left[\left\| \nabla_{\boldsymbol{\alpha}_i} F(\boldsymbol{\alpha}_t, \Psi^k) \right\|^2 + \left\| \nabla_{\boldsymbol{\alpha}_i^k} F(\boldsymbol{\alpha}_t^k, \Psi^k) \right\|^2 \right] + \frac{\eta}{2} \left[\sum_{k=1}^{K} q^{[k]} L^2 \cdot \left\| \boldsymbol{\alpha}_{i,(t)} - \boldsymbol{\alpha}_{i,(t)}^k \right\|^2 \right]
\end{aligned}$$

where (1) use inequality $2\langle a, b \rangle = \|a\|^2 + \|b\|^2 - \|a - b\|^2$; (2) use the convexity of ℓ_2 norm; (3) use L-smooth in Lemma 2.

For b):

$$\begin{aligned}
&\frac{L\eta^2}{2} \mathbb{E}_{\{\Phi^{k,r}\}_{r=1}^{I}} \mathbb{E}_{\mathcal{B}_t^k} \mathbb{E}_{U_t} \left[\left\| \hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^*(\boldsymbol{\alpha}_t^k, \mathcal{B}_t^k) \right\|^2 \right] \\
&= \mathbb{E}_{U_t} \left\{ \frac{L\eta^2}{2} \mathbb{E}_{\{\Phi^{k,r}\}_{r=1}^{I}} \mathbb{E}_{\mathcal{B}_t^k} \left[\left\| \hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^*(\boldsymbol{\alpha}_t^k, \mathcal{B}_t^k) \right\|^2 \right] \right\} \\
&\overset{(1)}{=} \mathbb{E}_{U_t} \left\{ \frac{L\eta^2}{2} \mathbb{E}_{\{\Phi^{k,r}\}_{r=1}^{I}} \mathbb{E}_{\mathcal{B}_t^k} \left[\left\| \hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^*(\boldsymbol{\alpha}_t^k, \mathcal{B}_t^k) - \mathbb{E}_{\{\Phi^{k,r}\}_{r=1}^{I}} \mathbb{E}_{\mathcal{B}_t^k} \left[\hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^*(\boldsymbol{\alpha}_t^k, \mathcal{B}_t^k) \right] \right\|^2 \right] \right\} \\
&\quad + \mathbb{E}_{U_t} \left\{ \frac{L\eta^2}{2} \left\| \mathbb{E}_{\{\Phi^{k,r}\}_{r=1}^{I}} \mathbb{E}_{\mathcal{B}_t^k} \left[\hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^*(\boldsymbol{\alpha}_t^k, \mathcal{B}_t^k) \right] \right\|^2 \right\}
\end{aligned}$$

$$
\begin{aligned}
&\overset{(2)}{=} \mathbb{E}_{U_t}\left\{\frac{L\eta^2}{2}\mathbb{E}_{\{\Phi^{k,r}\}_{r=1}^{I}}\mathbb{E}_{\mathcal{B}_t^k}\left[\left\|\hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^*(\boldsymbol{\alpha}_t^k,\mathcal{B}_t^k)-\nabla_{\boldsymbol{\alpha}_i^k}F^*(\boldsymbol{\alpha}_t^k,\Psi^k)\right\|^2\right]\right\}\\
&\quad+\mathbb{E}_{U_t}\left[\frac{L\eta^2}{2}\left\|\nabla_{\boldsymbol{\alpha}_i^k}F^*(\boldsymbol{\alpha}_t^k,\Psi^k)\right\|^2\right]\\
&\overset{(3)}{\leq}\frac{L\eta^2}{2}\mathbb{E}_{U_t}\left\|\nabla_{\boldsymbol{\alpha}_i^k}F^*(\boldsymbol{\alpha}_t^k,\Psi^k)\right\|^2\\
&\quad+\mathbb{E}_{U_t}\left\{L\eta^2\mathbb{E}_{\{\Phi^{k,r}\}_{r=1}^{I}}\mathbb{E}_{\mathcal{B}_t^k}\left[\left\|\hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^*(\boldsymbol{\alpha}_t^k,\mathcal{B}_t^k)-\mathbb{E}_{\{\Phi^{k,r}\}_{r=1}^{I}}\hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^*(\boldsymbol{\alpha}_t^k,\mathcal{B}_t^k)\right\|^2\right]\right\}\\
&\quad+\mathbb{E}_{U_t}\left\{L\eta^2\mathbb{E}_{\{\Phi^{k,r}\}_{r=1}^{I}}\mathbb{E}_{\mathcal{B}_t^k}\left[\left\|\mathbb{E}_{\{\Phi^{k,r}\}_{r=1}^{I}}\hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^*(\boldsymbol{\alpha}_t^k,\mathcal{B}_t^k)-\mathbb{E}_{\{\Phi^{k,r}\}_{r=1}^{I}}\mathbb{E}_{\mathcal{B}_t^k}\hat{\nabla}_{\boldsymbol{\alpha}_i^k} f^*(\boldsymbol{\alpha}_t^k,\mathcal{B}_t^k)\right\|^2\right]\right\}\\
&\overset{(4)}{\leq}\frac{L\eta^2}{2}\mathbb{E}_{U_t}\left\|\nabla_{\boldsymbol{\alpha}_i^k}F^*(\boldsymbol{\alpha}_t^k,\Psi^k)\right\|^2+L\eta^2\cdot\mathbb{E}_{U_t}\left[\frac{\sigma_\alpha^2}{I^2}\right]+L\eta^2\cdot\mathbb{E}_{U_t}\left[\frac{\sigma_\psi^2}{B}\right]\\
&=\frac{L\eta^2}{2}\sum_{k=1}^{K}q^{[k]}\cdot\left\|\nabla_{\boldsymbol{\alpha}_i^k}F^k(\boldsymbol{\alpha}_t^k,\Psi^k)\right\|^2+\frac{L\eta^2\sigma_\alpha^2}{I^2}+\frac{L\eta^2\sigma_\psi^2}{B}\\
&\overset{(5)}{\leq}\frac{L\eta^2\lambda}{2}\left\|\sum_{k=1}^{K}q^{[k]}\cdot\nabla_{\boldsymbol{\alpha}_i^k}F^k(\boldsymbol{\alpha}_t^k,\Psi^k)\right\|^2+\frac{L\eta^2\sigma_\alpha^2}{I^2}+\frac{L\eta^2\sigma_\psi^2}{B}
\end{aligned}
\tag{39}
$$

where (1) use $\mathbb{E}\left\|x-\mathbb{E}\left[x\right]\right\|^2=\mathbb{E}\left\|x\right\|^2-\left\|\mathbb{E}\left[x\right]\right\|^2$; (2) use the unbiasedness of stochastic gradient and variance-reduced policy gradient in Assumption 1 and Lemma 1; (3) use $\left\|a+b\right\|^2\leq 2\left\|a\right\|^2+2\left\|b\right\|^2$; (4) use the bounded variance of stochastic gradient and variance-reduced policy gradient in Assumption 1 and Lemma 1; (5) use Assumption 3.
Combining a) and b):

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\left\{\mathbb{E}_{\Phi_{t+1}\sim\mathrm{GS}(\boldsymbol{\alpha}_{t+1})}\left[\mathcal{L}(\Phi_{t+1},\Psi)\right]-\mathbb{E}_{\Phi_t\sim\mathrm{GS}(\boldsymbol{\alpha}_t)}\left[\mathcal{L}(\Phi_t,\Psi)\right]\right\}\\
&\leq\frac{1}{T}\sum_{t=0}^{T-1}\sum_{i=1}^{n}\left\{-\frac{\eta}{2}\left[\left\|\nabla_{\boldsymbol{\alpha}_i}F(\boldsymbol{\alpha}_t,\Psi^k)\right\|^2+\left\|\nabla_{\boldsymbol{\alpha}_i^k}F(\boldsymbol{\alpha}_t^k,\Psi^k)\right\|^2\right]+\frac{\eta}{2}\left[\sum_{k=1}^{K}q^{[k]}L^2\cdot\left\|\boldsymbol{\alpha}_{i,(t)}-\boldsymbol{\alpha}_{i,(t)}^k\right\|^2\right]\right\}\\
&\quad+\frac{1}{T}\sum_{t=0}^{T-1}\sum_{i=1}^{n}\left\{\frac{L\eta^2\lambda}{2}\left\|\sum_{k=1}^{K}q^{[k]}\cdot\nabla_{\boldsymbol{\alpha}_i^k}F^k(\boldsymbol{\alpha}_t^k,\Psi^k)\right\|^2+\frac{L\eta^2\sigma_\alpha^2}{I^2}+\frac{L\eta^2\sigma_\psi^2}{B}\right\}\\
&\overset{(1)}{\leq}-\frac{\eta}{2}\frac{1}{T}\sum_{t=0}^{T-1}\left\|\nabla_{\boldsymbol{\alpha}}F(\boldsymbol{\alpha}_t,\Psi^k)\right\|^2+\left[-\frac{\eta}{2}+\frac{\lambda L\eta^2}{2}+\eta^3L^2E(1+\frac{1}{K_*})\right]\frac{1}{T}\sum_{t=0}^{T-1}\left\|\nabla_{\boldsymbol{\alpha}^k}F(\boldsymbol{\alpha}_t^k,\Psi^k)\right\|^2\\
&\quad+\frac{(E+1)L^2\eta^3n\sigma_\psi^2(1+\frac{1}{K_*})+nL\eta^2\sigma_\psi^2}{B}+\frac{(E+1)L^2\eta^3n\sigma_\alpha^2(1+\frac{1}{K_*})+nL\eta^2\sigma_\alpha^2}{I^2}
\end{aligned}
$$

where (1) use Lemma 4.
Then,we can get:

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\left\|\nabla_{\boldsymbol{\alpha}}F(\boldsymbol{\alpha}_t,\Psi^k)\right\|^2\\
&\leq\frac{\mathbb{E}_{\Phi_0\sim\mathrm{GS}(\boldsymbol{\alpha}_0)}\left[\mathcal{L}(\Phi_0,\Psi)\right]-\mathbb{E}_{\Phi_T\sim\mathrm{GS}(\boldsymbol{\alpha}_t)}\left[\mathcal{L}(\Phi_T,\Psi)\right]}{\eta T}\\
&\quad+\left[-1+\lambda L\eta+2\eta^2L^2E(1+\frac{1}{K_*})\right]\frac{1}{T}\sum_{t=0}^{T-1}\left\|\nabla_{\boldsymbol{\alpha}^k}F(\boldsymbol{\alpha}_t^k,\Psi^k)\right\|^2\\
&\quad+\frac{2(E+1)L^2\eta^2n\sigma_\psi^2(1+\frac{1}{K_*})+2nL\eta\sigma_\psi^2}{B}+\frac{2(E+1)L^2\eta^2n\sigma_\alpha^2(1+\frac{1}{K_*})+2nL\eta\sigma_\alpha^2}{I^2}
\end{aligned}
$$

Based on Assumption 2, $\mathbb{E}_{\Phi_0 \sim \mathrm{GS}(\boldsymbol{\alpha}_0)}\left[\mathcal{L}(\Phi_0, \Psi)\right] - \inf_t \mathbb{E}_{\Phi_t \sim \mathrm{GS}(\boldsymbol{\alpha}_t)}\left[\mathcal{L}(\Phi_t, \Psi)\right] \leq 2G$ and:

$$-1 + \lambda L\eta + 2\eta^2 L^2 E(1 + \frac{1}{K_*}) \leq 0$$

$$0 < \eta \leq \eta^* = \frac{-\lambda L + \sqrt{\lambda^2 L^2 + 8L^2 E\left(1 + \frac{1}{K_*}\right)}}{8L^2 E\left(1 + \frac{1}{K_*}\right)}$$

Finally:

$$\frac{1}{T}\sum_{t=0}^{T-1}\left\|\nabla_{\boldsymbol{\alpha}} F(\boldsymbol{\alpha}_t, \Psi^k)\right\|^2$$
$$\leq \frac{4G}{\eta T} + \frac{2(E+1)L^2\eta^2 n\sigma_\psi^2(1+\frac{1}{K_*}) + 2nL\eta\sigma_\psi^2}{B} + \frac{2(E+1)L^2\eta^2 n\sigma_\alpha^2(1+\frac{1}{K_*}) + 2nL\eta\sigma_\alpha^2}{I^2}$$

□

Corollary 1. Convergence rate and complexity

1) Convergence rate: Let $\eta = \min\left\{\eta^*, \frac{1}{\sqrt{T}}, \frac{1}{L}\right\}$. Under this condition, the following holds:

$$\frac{1}{T}\sum_{t=0}^{T-1}\left\|\nabla_{\boldsymbol{\alpha}} F(\boldsymbol{\alpha}_t, \Psi)\right\|^2 = \mathcal{O}(\frac{1}{\sqrt{T}}) \tag{40}$$

2) Complexity: To guarantee an ϵ-solution, such that $\frac{1}{T}\sum_{t=0}^{T-1}\left\|\nabla_{\boldsymbol{\alpha}} F(\boldsymbol{\alpha}_t, \Psi^k)\right\|^2 \leq \epsilon^2$, the following condition must hold:

$$T_\epsilon = \mathcal{O}\left(\frac{1}{\epsilon^4}\right) \tag{41}$$

Proof. From condition of η:

$$0 < \eta \leq \eta^* = \frac{-\lambda L + \sqrt{\lambda^2 L^2 + 8L^2 E\left(1 + \frac{1}{K_*}\right)}}{8L^2 E\left(1 + \frac{1}{K_*}\right)}$$

We let $\eta = \min\left\{\eta^*, \frac{1}{\sqrt{T}}, \frac{1}{L}\right\}$ and the following holds:

$$\frac{1}{T}\sum_{t=0}^{T-1}\left\|\nabla_{\boldsymbol{\alpha}} F(\boldsymbol{\alpha}_t, \Psi^k)\right\|^2$$
$$\leq \mathcal{O}\left(\frac{1}{\sqrt{T}}\right)(4G) + \mathcal{O}(\frac{1}{\sqrt{T}})\left(\frac{2nL\sigma_\psi^2}{B} + \frac{2nL\sigma_\alpha^2}{I^2}\right)$$
$$+ \mathcal{O}\left(\frac{1}{\sqrt{T}}\right)\left[\frac{2(E+1)Ln\sigma_\psi^2(1+\frac{1}{K_*})}{B} + \frac{2(E+1)Ln\sigma_\alpha^2(1+\frac{1}{K_*})}{I^2}\right]$$
$$= \mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$$

In addition, we considering the iterative complexity, to get a ϵ-solution:

$$\frac{1}{T}\sum_{t=0}^{T-1}\left\|\nabla_{\boldsymbol{\alpha}} F(\boldsymbol{\alpha}_t, \Psi^k)\right\|^2 \leq \epsilon^2$$

We need to choose:

$$T_\epsilon = \left[\frac{4G}{\epsilon^2} + \frac{2(E+1)Ln\sigma_\psi^2(1+\frac{1}{K_*}) + 2nL\sigma_\psi^2}{B\epsilon^2} + \frac{2(E+1)Ln\sigma_\alpha^2(1+\frac{1}{K_*}) + 2nL\sigma_\alpha^2}{I^2\epsilon^2}\right]^2$$

$$T_\epsilon = \mathcal{O}\left(\frac{1}{\epsilon^4}\right)$$

□

Corollary 2. **The impact of K_* (FedOne)**: In the FL framework, $T_\epsilon K_*$ represents the total number of queries made to the cloud-based LLM service to achieve an ϵ-solution, whose quantity is directly proportional to the cost incurred for utilizing the LLM. The following condition holds:

$$T_\epsilon K_* \propto K_* \tag{42}$$

Therefore, $T_\epsilon K_*$ is a function of K_* that increase steadily for $K_* = 1, 2, \ldots, K$, indicating that the optimal K_* for query efficiency is $K_* = 1$.

Proof. According to Corollary 1, when $\eta = \min\left\{\eta^*, \frac{1}{\sqrt{T}}, \frac{1}{L}\right\}$, to get a ϵ-solution:

$$\frac{1}{T}\sum_{t=0}^{T-1}\left\|\nabla_{\boldsymbol{\alpha}} F(\boldsymbol{\alpha}_t, \Psi^k)\right\|^2 \leq \epsilon^2 \tag{43}$$

We need to choose:

$$T_\epsilon = \left[\frac{4G}{\epsilon^2} + \frac{2(E+1)Ln\sigma_\psi^2(1+\frac{1}{K_*}) + 2nL\sigma_\psi^2}{B\epsilon^2} + \frac{2(E+1)Ln\sigma_\alpha^2(1+\frac{1}{K_*}) + 2nL\sigma_\alpha^2}{I^2\epsilon^2}\right]^2$$

The $T_\epsilon K_*$ proportional to the query time satisfies the following equality:

$$\begin{aligned}
&T_\epsilon K_* \\
&= \left[\frac{4G}{\epsilon^2} + \frac{2(E+1)Ln\sigma_\psi^2(1+\frac{1}{K_*}) + 2nL\sigma_\psi^2}{B\epsilon^2} + \frac{2(E+1)Ln\sigma_\alpha^2(1+\frac{1}{K_*}) + 2nL\sigma_\alpha^2}{I^2\epsilon^2}\right]^2 \cdot K_* \\
&= \left[\left(\frac{4G}{\epsilon^2} + \frac{2(E+2)Ln\sigma_\psi^2}{B\epsilon^2} + \frac{2(E+2)Ln\sigma_\alpha^2}{I^2\epsilon^2}\right)\cdot\sqrt{K_*} + \left(\frac{2(E+1)Ln\sigma_\psi^2}{B\epsilon^2} + \frac{2(E+1)Ln\sigma_\alpha^2}{I^2\epsilon^2}\right)\cdot\frac{1}{\sqrt{K_*}}\right]^2
\end{aligned}$$

Then, $T_\epsilon K_*$ is a function of K_* that first decreases and then increases and that the optimal K_* for query efficiency exists:

$$K_*^{opt} = \frac{\frac{2(E+1)Ln\sigma_\psi^2}{B\epsilon^2} + \frac{2(E+1)Ln\sigma_\alpha^2}{I^2\epsilon^2}}{\frac{4G}{\epsilon^2} + \frac{2(E+2)Ln\sigma_\psi^2}{B\epsilon^2} + \frac{2(E+2)Ln\sigma_\alpha^2}{I^2\epsilon^2}} < 1$$

Finally, consider only the effect of K_* and $K_* \geq 1$, we can get:

$$T_\epsilon K_* \propto K_*$$

Therefore, $T_\epsilon K_*$ is a function of K_* that increase steadily for $K_* = 1, 2, \ldots, K$, indicating that the optimal K_* for query efficiency is $K_* = 1$. □

Remark 4. We carefully analyze the optimal K_* by balancing convergence complexity and query time, aiming to achieve the fastest convergence with the fewest queries. Specifically, we fix the convergence accuracy, denoted as ϵ, and determine the number of iterations, T_ϵ, required to achieve this accuracy. We then analyze the total number of queries, given by $T_\epsilon K_* \propto K_*$. This analysis shows that increasing K_* for $K_* = 1, 2, \ldots, K$, the number of activated clients can accelerate convergence by leveraging more data; however, it also increases query overhead, leading to higher communication and computational costs. We rigorously demonstrate that the number of queries required for Fed-BDPL to achieve an ϵ-solution is proportional to K_*. This result emphasizes that optimal query efficiency is achieved when only a single client ($K_* = 1$) is activated per round.

B EXPERIMENT DETAILS

B.1 GLUE DATASET DETAILS

Below is the table for the detail of the dataset we use, and their associated tasks, metrics, and data domains.

Dataset	**|L|**	**|Train|**	**|Dev|**	**|Test|**	**Type**	**Metrics**	**Domain**
MNLI	3	393K	9.8K	9.8K	NLI	acc.	fiction, reports
QQP	2	364K	40K	391K	paraphrase	F1	Quora
SST-2	2	6.7K	872	1.8K	sentiment	acc.	movie reviews
MRPC	2	3.7K	408	1.7K	paraphrase	F1	news
CoLA	2	8.6K	1K	1K	acceptability	Matthews corr.	books, articles
QNLI	2	105K	5.5K	5.5K	NLI	acc.	Wikipedia
RTE	2	2.5K	277	3K	NLI	acc.	news, Wikipedia

Table 6: The statistics and metrics of seven datasets in GLUE benchmark, $|\mathbf{L}|$: number of classes for classification tasks.