Journal of Machine Learning Research 26 (2025) 1-67 Submitted 3/24; Revised 10/24; Published 1/25

Adaptive Client Sampling in Federated Learning via Online Learning with Bandit Feedback

Boxin Zhao BOXINZ@UCHICAGO.EDU
University of Chicago Booth School of Business
Chicago, IL, 60637, USA

Lingxiao Wang LW324@NJIT.EDU
New Jersey Institute of Technology Department of Data Science
Newark, NJ, 07106, USA

Ziqi Liu ZIQILIU@ANTFIN.COM
Zhiqiang Zhang LINGYAO.ZZQ@ANTFIN.COM
Jun Zhou JUN.ZHOUJUN@ANTFIN.COM
Ant Financial Group
Hangzhou, Zhejiang, China

Chaochao Chen ZJUCCC@ZJU.EDU.CN
College of Computer Science and Technology
Zhejiang University
Hangzhou, Zhejiang, China

Mladen Kolar MKOLAR@MARSHALL.USC.EDU
University of Southern California Marshall School of Business
Los Angeles, CA, 90089, USA

Editor: Francesco Orabona

Abstract

Due to the high cost of communication, federated learning (FL) systems need to sample a subset of clients that are involved in each round of training. As a result, client sampling plays an important role in FL systems as it affects the convergence rate of optimization algorithms used to train machine learning models. Despite its importance, there is limited work on how to sample clients effectively. In this paper, we cast client sampling as an online learning task with bandit feedback, which we solve with an online stochastic mirror descent (OSMD) algorithm designed to minimize the sampling variance. We then theoretically show how our sampling method can improve the convergence speed of federated optimization algorithms over the widely used uniform sampling. Through both simulated and real data experiments, we empirically illustrate the advantages of the proposed client sampling algorithm over uniform sampling and existing online learning-based sampling strategies. The proposed adaptive sampling procedure is applicable beyond the FL problem studied here and can be used to improve the performance of stochastic optimization procedures such as stochastic gradient descent and stochastic coordinate descent.

Keywords: federated learning, client sampling, online learning, optimization, variance reduction

. Correspondence: Mladen Kolar (mkolar@marshall.usc.edu) and Jun Zhou (jun.zhoujun@antfin.com).

1. Introduction

Modern edge devices, such as personal mobile phones, wearable devices, and sensor systems in vehicles, collect large amounts of data that are valuable for training of machine learning models. If each device only uses its local data to train a model, the resulting generalization performance will be limited due to the number of available samples on each device. Traditional approaches where data are transferred to a central server, which trains a model based on all available data, have fallen out of fashion due to privacy concerns and high communication costs. Federated Learning (FL) has emerged as a paradigm that allows for collaboration between different devices (clients) to train a global model while keeping data locally and only exchanging model updates (McMahan et al., 2017).

In a typical FL process, we have clients that contain data and a central server that orchestrates the training process (Kairouz et al., 2021). The following process is repeated until the model is trained: (i) the server selects a subset of available clients; (ii) the server broadcasts the current model parameters and sometimes also a training program (e.g., a Tensorflow graph (Abadi et al., 2016)); (iii) the selected clients make updates to the model parameters based on their local data; (iv) the local model updates are uploaded to the server; (v) the server aggregates the local updates and makes a global update of the shared model. In this paper, we focus on the first step and develop a practical strategy for selecting clients with provable guarantees.

To train a machine learning model in a FL setting with M clients, we would like to minimize the following objective[1]:

$$\min_{w} F(w) \coloneqq \sum_{m\in[M]} \lambda_m \phi\left(w; \mathcal{D}_m\right), \tag{1}$$

where $\phi(w; \mathcal{D}_m)$ is the loss function used to assess the quality of a machine learning model parameterized by the vector w based on the local data $\mathcal{D}_m$ on the client $m \in [M]$. The parameter λ_m denotes the weight for client m. Typically, we have $\lambda_m = n_m/n$, where $n_m = |\mathcal{D}_m|$ is the number of samples on the client m, and the total number of samples is $n = \sum_{m=1}^{M} n_m$. At the beginning of the t-th communication round, the server uses the sampling distribution $p^t = (p_1^t, \ldots, p_M^t)^\top$ to choose K clients by sampling with replacement from $[M]$[2]. Let $S^t \subseteq [M]$ denote the set of chosen clients with $|S^t| = K$. The server transmits the current model parameter vector w^t to each client $m \in S^t$. The client m computes the local update g_m^t [3] and sends it back to the server[4]. After receiving local

1. We use $[M]$ to denote the set $\{1, \ldots, M\}$.
2. In this paper, we assume that all clients are available in each round and the purpose of client sampling is to reduce the communication cost, which is also the case considered by some previous research (Chen et al., 2022). However, in practice, it is possible that only a subset of clients are available at the beginning of each round due to physical constraint. In Appendix D.2, we discuss how to extend our proposed methods to deal with such situations. Analyzing such an extension is highly non-trivial and we leave it for further study. See detailed discussion in Appendix D.2.
3. Here by model update, we actually mean the negative direction of model update. For example, when applying gradient descent, we refer the gradient as the model update, while the model parameter makes an update at the direction of the negative gradient. We stick with the term model update since it is more commonly used.
4. Throughout the paper, except in Section 4, we do not specify how g_m^t is obtained. One possibility that the reader could keep in mind for concreteness is the LocalUpdate algorithm (Charles and Konečnỳ, 2020), which covers well-known algorithms such as mini-batch SGD and FedAvg (McMahan et al., 2017).

updates from clients in S^t, the server constructs a stochastic estimate of the global gradient as

$$g^t = \frac{1}{K} \sum_{m \in S^t} \frac{\lambda_m}{p_m^t} g_m^t, \tag{2}$$

and makes the global update of the parameter w^t using g^t. For example, $w^{t+1} = w^t - \mu^t g^t$, if the server is using stochastic gradient descent (SGD) with the stepsize sequence $\{\mu^t\}_{t \geq 1}$ (Bottou et al., 2018). However, the global update can be obtained using other procedures as well.

The sampling distribution in FL is typically uniform over clients, that is, $p^t = p^{\text{unif}} = (1/M, \ldots, 1/M)^\top$. However, nonuniform sampling (also called importance sampling) can lead to faster convergence, both in theory and practice, as has been illustrated in stochastic optimization (Zhao and Zhang, 2015; Needell et al., 2016). While the sampling distribution can be designed based on prior knowledge (Zhao and Zhang, 2015; Johnson and Guestrin, 2018; Needell et al., 2016; Stich et al., 2017), we cast the problem of choosing the sampling distribution as an online learning task and need no prior knowledge about (1).

Existing approaches to designing a sampling distribution using online learning takes a stationary online learning framework and focus on matching the best sampling distribution that does not change over the training process. As a result, the obtained algorithms update the sampling distribution by treating all history information equally. However, as the training proceeds, the best sampling distribution changes with iterations. To address this problem, we take a non-stationary online learning framework and use an online stochastic mirror descent (OSMD) algorithm that puts more emphasize on recent feedback and learns to 'forget' the history. Consequently, our method shows empirical advantages over the previous methods. Besides, we derive a dynamic regret upper bound that allows the comparators to change with iterations, which generalizes the theoretical results on static regret of previous research. Moreover, we provably show how our sampling method improves the convergence guarantee of federated optimization methods over uniform sampling by reducing the dependency on the heterogeneity of the problem, which is also new to the best of our knowledge.

1.1 Notation

Let $\mathbb{R}_+^M = [0, \infty)^M$ and $\mathbb{R}_{++}^M = (0, \infty)^M$. For $M \in \mathbb{N}^+$, let $\mathcal{P}_{M-1} \coloneqq \{x \in \mathbb{R}_+^M : \sum_{i=1}^M x_i = 1\}$ be the $(M-1)$-dimensional probability simplex. We use $p = (p_1, \ldots, p_M)^\top$ to denote a sampling distribution with support on $[M] \coloneqq \{1, \ldots, M\}$. We use $p^{1:T}$ to denote a sequence of sampling distributions $\{p^t\}_{t=1}^T$. We use $\|\cdot\|_p$ to denote the L_p-norm for $1 \leq p \leq \infty$. For $x \in \mathbb{R}^n$, we have $\|x\|_p = (\sum_{i=1}^n x_i^p)^{1/p}$ when $1 < p < \infty$, $\|x\|_1 = \sum_{i=1}^n |x_i|$, and $\|x\|_\infty = \max_{1 \leq i \leq n} |x_i|$. Given any L_p-norm $\|\cdot\|$, we define its dual norm as $\|z\|_\star \coloneqq \sup\{z^\top x : \|x\| \leq 1\}$. We use $|\mathcal{B}|$ to denote the cardinality of the index set $|\mathcal{B}|$.

Let $\Phi : \mathcal{D} \subseteq \mathbb{R}^M \mapsto \mathbb{R}$ be a differentiable convex function defined on $\mathcal{D}$, where $\mathcal{D}$ is a convex open set, and we use $\bar{\mathcal{D}}$ to denote the closure of $\mathcal{D}$. The Bregman divergence between any $x, y \in \mathcal{D}$ with respect to the function Φ is given as $D_\Phi(x \,\|\, y) = \Phi(x) - \Phi(y) - \langle \nabla \Phi(y), x - y \rangle$. The unnormalized negative entropy is denoted as $\Phi(x) = \sum_{m=1}^M x_m \log x_m - \sum_{m=1}^M x_m$, $x = (x_1, \ldots, x_M)^\top \in \mathcal{D} = \mathbb{R}_+^M$, with $0 \log 0$ defined as 0.

For two positive sequences $\{a_n\}$ and $\{b_n\}$, we use $a_n = O(b_n)$ to denote that there exists $C > 0$ such that $a_n/b_n \leq C$ for all n large enough. Similarly, we use $a_n = \Omega(b_n)$ to denote that there exists $c > 0$ such that $a_n/b_n \geq c$ for all n large enough. We denote $a_n = \Theta(b_n)$ if $a_n = O(b_n)$ and $a_n = \Omega(b_n)$ simultaneously. Besides, we use $a_n = o(b_n)$ if $\lim_n a_n/b_n = 0$. In addition, $a_n = \tilde{O}(b_n)$ if $a_n = O(b_n \log^k b_n)$ for some $k \geq 0$.

1.2 Organization of the Paper

We summarize the related work in Section 2. We motivate importance sampling in FL and introduce an adaptive client sampling algorithm in Section 3. We develop optimization guarantees of mini-batch SGD and FedAvg using our sampling scheme in Section 4. In Section 5, we discuss regret analysis which serves as a key component for optimization analysis. In Section 6, we propose an extension of the sampling method that is adaptive to the unknown problem parameters. We provide the experimental results on simulated data in Section 7 and real-world data in Section 8. We conclude our paper with Section 9. We leave all the technical proofs and additional discussions in Appendix. Code to replicate the results in this paper is available at

`https://github.com/boxinz17/FL-Client-Sampling`.

2. Related Work

Our paper is related to client sampling in FL, importance sampling in stochastic optimization, and online convex optimization. We summarize only the most relevant literature, without attempting to provide an extensive survey.

For client sampling, Chen et al. (2022) proposed to use the theoretically optimal sampling distribution to choose clients. However, their method requires all clients to compute local updates in each round, which is impractical due to stragglers. Ribero and Vikalo (2020) modelled the parameters of the model during training by an Ornstein-Uhlenbeck process, which was then used to derive an optimal sampling distribution. Cho et al. (2020b) developed a biased client selection strategy and analyzed its convergence property. As a result, the algorithm has a non-vanishing bias and is not guaranteed to converge to optimum. Moreover, it needs to involve more clients than our method and is thus communication and computational more expensive. Kim et al. (2020); Cho et al. (2020a); Yang et al. (2021) considered client sampling as a multi-armed bandit problem, but provided only limited theoretical results. Wang et al. (2020) used reinforcement learning for client sampling with the objective of maximizing accuracy, while minimizing the number of communication rounds.

Our paper is also closely related to importance sampling in stochastic optimization. Zhao and Zhang (2015); Needell et al. (2016) illustrated that by sampling observations from a nonuniform distribution when using a gradient-based stochastic optimization method, one can achieve faster convergence. They designed a fixed sampling distribution using prior knowledge on the upper bounds of gradient norms. Csiba and Richtárik (2018) extended the importance sampling to mini-batches. Stich et al. (2017); Johnson and Guestrin (2018); Gopal (2016) developed adaptive sampling strategies that allow the sampling distribution to change over time. Nesterov (2012); Perekrestenko et al. (2017); Zhu et al. (2016); Salehi et al. (2018) discussed importance sampling in stochastic coordinate descent methods. Namkoong

et al. (2017); Salehi et al. (2017); Borsos et al. (2018, 2019); Hanchi and Stephens (2020) illustrated how to design the sampling distribution by solving an online learning task with bandit feedback. Namkoong et al. (2017); Salehi et al. (2017) designed the sampling distribution by solving a multi-armed bandit problem with the EXP3 algorithm (Lattimore and Szepesvári, 2020, Chapter 11). Borsos et al. (2018) used the follow-the-regularized-leader algorithm (Lattimore and Szepesvári, 2020, Chapter 28) to solve an online convex optimization problem and make updates to the sampling distribution. Borsos et al. (2019) restricted the sampling distribution to be a linear combination of distributions in a predefined set and used an online Newton step to make updates to the mixture weights. The above approaches estimate a stationary distribution, while the best distribution is changing with iterations and, therefore, is intrinsically dynamic. In addition to having suboptimal empirical performance, these papers provide theoretical results that only establish a regret relative to a fixed sampling distribution in hindsight. To address this problem, Hanchi and Stephens (2020) took a non-stationary approach where the most recent information for each client was kept. A decreasing stepsize sequence is required to establish a regret bound. In comparison, we establish a regret bound relative to a dynamic comparator—a sequence of sampling distributions—without imposing assumptions on the stepsize sequence, and this bound includes the dependence on the total variation term characterizing how strong the comparator is.

Our paper also contributes to the literature on online convex optimization. We cast the client sampling problem as an online learning problem (Hazan, 2016) and adapt algorithms from the dynamic online convex optimization literature to solve it. Hall and Willett (2015); Yang et al. (2016); Daniely et al. (2015) proposed methods that achieve sublinear dynamic regret relative to dynamic comparator sequences. In particular, Hall and Willett (2015) used a dynamic mirror descent algorithm to achieve sublinear dynamic regret with total variation characterizing the intrinsic difficulty of the environment. Compared with the problem settings in the above studies, there are two key new challenges that we need to address. First, we only have partial information—bandit feedback—instead of the full information about the loss functions. Second, the loss functions in our case are unbounded, which violates the common boundedness assumption in the online learning literature. To overcome the first difficulty, we construct an unbiased estimator of the loss function and its gradient, which are then used to make an update to the sampling distribution. We address the second challenge by first bounding the regret of our algorithm when the sampling distributions in the comparator sequence lie in a region of the simplex for which the loss is bounded, and subsequently analyze the additional regret introduced by projecting the elements of the comparator sequence to this region.

3. Adaptive Client Sampling

We show how to cast the client sampling problem as an online learning task. Subsequently, we solve the online learning problem using the OSMD algorithm.

3.1 Client Sampling as an Online Learning Problem

Recall that at the beginning of the t-th communication round, the server uses a sampling distribution p^t to choose a set of clients S^t, by sampling with replacement K clients from $[M]$,

to update the parameter vector w^t. For a chosen client $m \in S^t$, the local update is denoted as g_m^t. For example, the local update $g_m^t = \nabla\phi(w^t; \mathcal{D}_m)$ may be the full gradient; when mini-batch SGD/FedSGD is used, then $g_m^t = (1/B)\sum_{b=1}^{B} \nabla\phi(w^t; \xi_m^{t,b})$, where $\xi_m^{t,b} \overset{i.i.d.}{\sim} \mathcal{D}_m$ and B is the batch size; when FedAvg (McMahan et al., 2017) is used, then $g_m^t = w^t - w_m^{t,B}$, where $w_m^{t,b} = w_m^{t,b-1} - \mu_l^t \nabla f\left(w_m^{t,b-1}; \xi_m^{t,b-1}\right)$, $b = 0, \ldots, B-1$, $w_m^{t,0} = w^t$, $\xi_m^{t,b} \overset{i.i.d.}{\sim} \mathcal{D}_m$, and μ_l^t is the local stepsize at t-th communication round.

The randomness comes from two sampling processes. The first sampling happens on clients level, and the second sampling happens locally when computing local updates. Client sampling is dealing with the first randomness. Since the two sampling process are independent, we may treat g_m^t as deterministic in this section to ease the understanding. We will include the second randomness when analyzing regret and specific optimization algorithms in following sections.

We define the aggregated oracle update at the t-th communication round as

$$J^t = \sum_{m=1}^{M} \lambda_m g_m^t.$$

The oracle update J^t is constructed only for theoretical purposes and is not computed in practice. The stochastic estimate g^t, defined in (2), is an unbiased estimate of J^t, that is, $\mathbb{E}_{S^t}\left[g^t\right] = J^t$. The variance of g^t is

$$\mathbb{V}_{S^t}[g^t] = \frac{1}{K}\left(\sum_{m=1}^{M} \frac{\lambda_m^2 \|g_m^t\|_2^2}{p_m^t} - \left\|J^t\right\|_2^2\right). \tag{3}$$

Our goal is to design the sampling distribution p^t, used to sample S^t, to minimize the variance in (3). In doing so, we can ignore the second term, as it is independent of p^t. Minimizing variance is our goal in designing the sampling distribution because we require g^t to be an unbiased estimate of J^t. Allowing g^t to be biased, as in the biased client selection literature (Cho et al., 2020b; Qu et al., 2022; Ribero and Vikalo, 2020), may render minimizing variance ineffective. Our focus is on unbiased client selection, leaving biased client selection for future research.

Let $a_m^t = \lambda_m^2 \|g_m^t\|_2^2$. For any sampling distribution $q = (q_1, \ldots, q_M)^\top$, the *variance reduction loss*[5] is defined as

$$l_t(q) = \frac{1}{K}\sum_{m=1}^{M} \frac{a_m^t}{q_m}. \tag{4}$$

Then for S^t sampled via q, we have

$$\mathbb{V}_{S^t}[g^t] = l_t(q) - \frac{1}{K}\left\|J^t\right\|_2^2.$$

Given a sequence of sampling distributions $q^{1:T}$, the cumulative variance reduction loss is defined as $L(q^{1:T}) \coloneqq \sum_{t=1}^{T} l_t(q^t)$. When the choice of $q^{1:T}$ is random, the expected cumulative variance reduction loss is defined as $\bar{L}(q^{1:T}) \coloneqq \mathbb{E}\left[L(q^{1:T})\right]$.

5. The variance reduction loss $l_t(\cdot)$ should be distinguished from the training loss $\phi(\cdot)$. While the former is always convex, $\phi(\cdot)$ can be non-convex.

The variance reduction loss appears in the bound on the sub-optimality of a stochastic optimization algorithm. As a motivating example, suppose $F(\cdot)$ in (1) is ν-strongly convex. Furthermore, suppose the local update $g_m^t = \nabla\phi(w^t; \mathcal{D}_m)$ is the full gradient of the local loss and the global update is made by SGD with stepsize $\mu^t = 2/(\nu t)$. Theorem 3 of Salehi et al. (2017) then states that for any $T \geq 1$:

$$\mathbb{E}\left[F\left(\frac{2}{T(T+1)}\sum_{t=1}^{T} t \cdot w^t\right)\right] - F(w^\star) \leq \frac{2}{\nu T(T+1)}\bar{L}(p^{1:T}), \tag{5}$$

where $w^\star$ is the minimizer of the objective in (1). Therefore, by choosing the sequence of sampling distributions $p^{1:T}$ to make the $\bar{L}(p^{1:T})$ small, one can achieve faster convergence. This observation holds in other stochastic optimization problems as well. We develop an algorithm that creates a sequence of sampling distributions $p^{1:T}$ to minimize $\bar{L}(p^{1:T})$ using only the norm of local updates, and without imposing assumptions on the loss functions or how the local and global updates are made. As a result, the proposed sampling algorithm is agnostic to both optimization algorithms and optimization problems. In Section 4, we show how our sampling method improves over uniform sampling by providing tighter upper bounds on mini-batch SGD and FedAvg with non-convex objective $\phi(\cdot)$.

Suppose that at the beginning of the t-th communication round we know all $\{a_m^t\}_{m=1}^M$. Then the optimal sampling distribution

$$p_\star^t = (p_{\star,1}^t, \ldots, p_{\star,M}^t)^\top = \arg\min_{p \in \mathcal{P}_{M-1}} l_t(p)$$

is obtained as $p_{\star,m}^t = \sqrt{a_m^t}/(\sum_{m=1}^M \sqrt{a_m^t})$. Computing the distribution $p_\star^t$ is impractical as it requires local updates of all clients, which eradicates the need for client sampling. From the form of $p_\star^t$, we observe that clients with a large a_m^t are more "important" and should have a higher probability of being selected. Since we do not know $\{a_m^t\}_{m=1}^M$, we will need to explore the environment to learn about the importance of clients before we can exploit the best strategy. Finally, we note that the relative importance of clients will change over time, which makes the environment dynamic and challenging.

Based on the above discussion, we cast the problem of creating a sequence of sampling distributions as an online learning task with bandit feedback, where a game is played between the server and environment. Let p^1 be the initial sampling distribution. At the beginning of iteration t, the server samples with replacement K clients from $[M]$, denoted by S^t, using p^t. The environment reveals $\{a_m^t\}_{m \in S^t}$ to the server, where $a_m^t = \lambda_m^2 \|g_m^t\|_2^2$. The environment also computes $l_t(p^t)$; however, this loss is not revealed to the server. The server then updates p^{t+1} based on the feedback $\{\{a_m^u\}_{m \in S^u}\}_{u=1}^t$ and sampling distributions $\{p^u\}_{u=1}^t$. Note that in this game, the server only gets information about the chosen clients and, based on this partial information, or bandit feedback, needs to update the sampling distribution. On the other hand, we would like to be competitive with an oracle that can calculate the cumulative variance reduction loss. We will design p^t in a way that is agnostic to the generation mechanism of $\{a^t\}_{t \geq 1}$, and will treat the environment as deterministic, with randomness coming only from $\{S^t\}_{t \geq 1}$ when designing p^t. We describe an OSMD-based approach to solve this online learning problem.

Algorithm 1 OSMD Sampler

Input: Learning rate η, parameter $\alpha \in (0, 1]$, and number of iterations T.
Output: $\hat{p}^{1:T}$.
Initialize: $\hat{p}^1 = p^{\text{unif}}$.
for $t = 1, 2, \ldots, T-1$ **do**
  Sample S^t by $\hat{p}^t$.
  Compute $\nabla \hat{l}_t(\hat{p}^t; \hat{p}^t)$ via (7).
  $\hat{p}^{t+1} = \arg\min_{p \in \mathcal{A}} \eta \langle p, \nabla \hat{l}_t(\hat{p}^t; \hat{p}^t) \rangle + D_\Phi\left(p \,\|\, \hat{p}^t\right)$.
end for

3.2 OSMD Sampler

Note that the variance-reduction loss function l_t is a convex function on $\mathcal{P}_{M-1}$ and

$$\nabla l_t(q) = -\frac{1}{K}\left(\frac{a_1^t}{(q_1)^2}, \ldots, \frac{a_M^t}{(q_M)^2}\right)^\top \in \mathbb{R}^M \quad \text{for all} \quad q = (q_1, \ldots, q_M)^\top \in \mathbb{R}^M_{++}.$$

Since we do not observe a^t, we cannot compute $l_t(\cdot)$ or $\nabla l_t(\cdot)$. Instead, we can construct unbiased estimates of them. For any $q \in \mathcal{P}_{M-1}$, let $\hat{l}_t(q; p^t)$ be an estimate of $l_t(q)$ defined as

$$\hat{l}_t(q; p^t) = \frac{1}{K^2} \sum_{m=1}^{M} \frac{a_m^t}{q_m p_m^t} \mathcal{N}\left\{m \in S^t\right\}, \tag{6}$$

and $\nabla \hat{l}_t(q; p^t) \in \mathbb{R}^M$ has the m-th entry defined as

$$\left[\nabla \hat{l}_t(q; p^t)\right]_m = -\frac{1}{K^2} \cdot \frac{a_m^t}{q_m^2 p_m^t} \mathcal{N}\left\{m \in S^t\right\}. \tag{7}$$

The set S^t is sampled with replacement from $[M]$ using p^t and $\mathcal{N}\left\{m \in S^t\right\}$ denotes the number of times that a client m is chosen in S^t. Thus, $0 \le \mathcal{N}\left\{m \in S^t\right\} \le K$. Given q and p^t, $\hat{l}_t(q; p^t)$ and $\nabla \hat{l}_t(q; p^t)$ are random variables in $\mathbb{R}$ and $\mathbb{R}^M$ that satisfy

$$\mathbb{E}_{S^t}\left[\hat{l}_t(q; p^t) \mid p^t\right] = l_t(q), \qquad \mathbb{E}_{S^t}\left[\nabla \hat{l}_t(q; p^t) \mid p^t\right] = \nabla l_t(q).$$

When S^t and $p^t \in \mathbb{R}^M_{++}$ are given, $\hat{l}_t(q; p^t)$ is a convex function with respect to q on $\mathbb{R}^M_{++}$ and satisfies $\hat{l}_t(q; p^t) - \hat{l}_t(q'; p^t) \le \langle \nabla \hat{l}_t(q; p^t), q - q' \rangle$, for $q, q' \in \mathbb{R}^M_{++}$. The constructed estimates $\hat{l}_t(q; p^t)$ and $\nabla \hat{l}_t(q; p^t)$ are crucial for designing updates to the sampling distribution.

OSMD Sampler is an online stochastic mirror descent algorithm for updating the sampling distribution, detailed in Algorithm 1. The sampling distribution is restricted to lie in the space $\mathcal{A} = \mathcal{P}_{M-1} \cap [\alpha/M, \infty)^M$, $\alpha \in (0, 1]$, to prevent the server from assigning too small probabilities to devices. The learning rates $\{\eta_t\}_{t \ge 1}$ are positive[6]. $\Phi(x) = \sum_{m=1}^M x_m \log x_m - \sum_{m=1}^M x_m$, $x = (x_1, \ldots, x_M)^\top \in \mathcal{D} = \mathbb{R}^M_+$, with $0 \log 0$ defined as 0, is the unnormalized negative entropy. The Bregman divergence between any $x, y \in \mathcal{D}$ with

6. We use the term *learning rate* when discussing an online algorithm that learns a sampling distribution, while the term *stepsize* is used in the context of an optimization algorithm.

Algorithm 2 Solver of Step 7 of Algorithm 1

1: **Input:** $\hat{p}^t$, S^t, $\{a_m^t\}_{m\in S^t}$, and $\mathcal{A} = \mathcal{P}_{M-1} \cap [\alpha/M, \infty)^M$.
2: **Output:** $\hat{p}^{t+1}$.
3: Let $\tilde{p}_m^{t+1} = p_m^t \exp\left\{\mathcal{N}\left\{m \in S^t\right\} \eta_t a_m^t / (K^2 (p_m^t)^3)\right\}$ for $m \in [M]$.
4: Sort $\{\tilde{p}_m^{t+1}\}_{m=1}^M$ in a non-decreasing order: $\tilde{p}_{\pi(1)}^{t+1} \le \tilde{p}_{\pi(2)}^{t+1} \le \cdots \le \tilde{p}_{\pi(M)}^{t+1}$.
5: Let $v_m = \tilde{p}_{\pi(m)}^{t+1}\left(1 - \frac{m-1}{M}\alpha\right)$ for $m \in [M]$.
6: Let $u_m = \frac{\alpha}{M}\sum_{j=m}^M \tilde{p}_{\pi(j)}^{t+1}$ for $m \in [M]$.
7: Find the smallest m such that $v_m > u_m$, denoted as $m_\star^t$.
8: Let $\hat{p}_m^{t+1} = \begin{cases} \alpha/M & \text{if } \pi(m) < m_\star^t \\ \left((1 - ((m_\star^t - 1)/M)\alpha)\tilde{p}_m^{t+1}\right) / \left(\sum_{j=m_\star^t}^M \tilde{p}_{\pi(j)}^{t+1}\right) & \text{otherwise.} \end{cases}$

respect to the function Φ is given as $D_\Phi(x \,\|\, y) = \Phi(x) - \Phi(y) - \langle \nabla\Phi(y), x - y\rangle$. Line 7 of Algorithm 1 provides an update to the sampling distribution using the mirror descent update. The available feedback is used to construct an estimate of the loss, while the Bregman divergence between the current and next sampling distribution is used as a regularizer, ensuring that the updated sampling distribution does not change too much. The update only uses the most recent information, while forgetting the history, which results in nonstationarity of the sequence of sampling distributions.

An efficient algorithm to solve the mirror descent update in Line 7 of Algorithm 1 is shown in Algorithm 2, justified by Proposition 12 in Appendix B.2. The main cost comes from sorting the sequence $\{\tilde{p}_m^{t+1}\}_{m=1}^M$, which can be done with the computational complexity of $O(M \log M)$. However, note that we only update a few entries of $\hat{p}^t$ to get $\tilde{p}^{t+1}$ and $\hat{p}^t$ is sorted. Therefore, most entries of $\tilde{p}^{t+1}$ are also sorted. Using this observation, we can usually achieve a much faster running time, for example, by using an adaptive sorting algorithm (Estivill-Castro and Wood, 1992).

4. Application of OSMD Sampler on Federated Optimization Algorithms

We illustrate how OSMD Sampler can be used to provably improve the convergence rates of federated optimization algorithms by reducing the heterogeneity. We choose two algorithms that are most commonly used in federated learning as our illustrative examples, namely the SGD mini-batch and FedAvg (McMahan et al., 2017). We use these two algorithms as motivational examples to show how adaptive sampling improves the convergence guarantees of optimization algorithms. However, the analysis here could be generalized to other optimization algorithms as well.

To simplify the notation, we denote $F_m(w) \coloneqq \phi(w; \mathcal{D}_m)$ and let $\lambda_m = 1/M$ for all $m \in [M]$ in problem (1). We assume that the client objectives are differentiable and L-smooth functions.

Assumption 1 *For all $m \in [M]$, $F_m(\cdot)$ is differentiable and L-smooth, that is,*

$$\|\nabla F_m(x) - \nabla F_m(y)\|_2 \le L\|x - y\|_2, \quad \text{for all } x, y \in \mathbb{R}^d.$$

Note that we allow $F_m(\cdot)$ to be non-convex. We also assume that the objective function $F(\cdot)$ is lower-bounded.

Assumption 2 *We assume that* $\inf_w F(w) > -\infty$. *We then denote* $F^\star \coloneqq \inf_w F(w)$.

In addition, we make the following assumption about the local stochastic gradient.

Assumption 3 *We assume that*

$$\mathbb{E}_{\xi\sim\mathcal{D}_m}\left[\nabla\phi(w;\xi)\right] = \nabla F_m(w) \quad \textit{and} \quad \mathbb{E}_{\xi\sim\mathcal{D}_m}\left[\left\|\nabla\phi(w;\xi) - \nabla F_m(w)\right\|_2^2\right] \le \sigma^2$$

for all w *and* $m \in [M]$*; besides,* $\|\nabla\phi(w;\xi)\|_2 \le G$ *for all* w *and* ξ.

While bounded gradient variance is often assumed in federated learning literature (Patel et al., 2022), bounded gradient norm is less common. We use this assumption to simplify regret analysis, making the loss in (4) bounded. Removing it is feasible but complex, diverging from the main focus of this paper. From a practical viewpoint, many loss functions, like logistic regression loss, naturally satisfy this assumption. For others, one can project the gradient into a bounded norm subspace. If any minimizer $w^\star$ has a bounded gradient norm $\|\nabla\phi(w^\star;\xi)\|_2$ and is within this subspace, the projection will not increase the distance to the minimizer. However, while the projection step can provide a slightly stronger theoretical guarantee, it brings few practical benefits and makes the algorithm harder to follow. Therefore, we adopt a stronger assumption to simplify the presentation.

We start our analysis by building the connection between the heterogeneity and client sampling. We first introduce quantities that characterize the heterogeneity of the optimization problem. Specifically, heterogeneity characterizes how the objective functions of different clients differ from each other. In a federated learning problem, heterogeneity can be large and it is important to understand its effect on the convergence of algorithms. Let

$$\zeta_{\text{unif}}^2 \coloneqq \sup_w \frac{1}{M}\sum_{m=1}^M \|\nabla F_m(w) - \nabla F(w)\|_2^2 = \sup_w \left\{\frac{1}{M}\sum_{m=1}^M \|\nabla F_m(w)\|_2^2 - \|\nabla F(w)\|_2^2\right\}. \quad (8)$$

The quantity ζ_{unif} has been commonly used to quantify the first-order heterogeneity in the literature (Karimireddy et al., 2020a,b). To understand the relationship between heterogeneity and client sampling, let $\tilde{m}$ be a random index drawn from $[M]$, with $\mathbb{P}\{\tilde{m} = m\} = p_m$ for all $m \in [M]$. Then a natural unbiased estimator of $\nabla F(w)$ is $\nabla F_{\tilde{m}}(w)/(M\tilde{p}_m)$. We define

$$V(p, w) \coloneqq \mathbb{E}_{\tilde{m}}\left[\left\|\frac{1}{Mp_{\tilde{m}}}\nabla F_{\tilde{m}}(w) - \nabla F(w)\right\|_2^2\right],$$

to be the variance of the estimator at parameter w, where $\mathbb{E}_{\tilde{m}}[\cdot]$ denotes the expectation taken with respect to the random index $\tilde{m}$. Note that we have

$$V(p, w) = \frac{1}{M^2}\sum_{m=1}^M \frac{1}{p_m}\|\nabla F_m(w)\|_2^2 - \|\nabla F(w)\|_2^2. \quad (9)$$

Thus, it is clear that $\zeta_{\text{unif}}^2 = \sup_w V(p^{\text{unif}}, w)$. In other words, the common definition of heterogeneity can be viewed as the worst-case variance of uniform client sampling.

When we use adaptive sampling to do client sampling, there are two sources of flexibility that allow us to reduce the heterogeneity: (i) we can use non-uniform sampling that may depend on parameter w; (ii) we allow the sampling distribution to change over iterations. To reflect the consequential effect, we introduce a new concept termed *dynamic heterogeneity*. Let $\mathrm{TV}(q^{1:T}) = \sum_{t=1}^{T-1} \|q^{t+1} - q^t\|_1$ be the total variation of any sequence of sampling distributions $q^{1:T} \in \mathcal{P}_{M-1}^T$. The dynamic heterogeneity is defined as

$$\zeta_T^2(\alpha, \beta) = \frac{1}{T} \sup_{w^1} \min_{p^1 \in \mathcal{A}} \cdots \sup_{w^T} \min_{p^T \in \mathcal{A}} \sum_{t=1}^{T} V(p^t, w^t) \quad \text{subject to } \mathrm{TV}\left(p^{1:T}\right) \le \beta,$$

where $V(p, w)$ is defined in (9) and $\beta \ge 0$ is the total variation budget. The dynamic heterogeneity $\zeta_T^2(\alpha, \beta)$ can be regarded as the worst-case variance of dynamic samplings in $\mathcal{A}^T$ with the total variation budget β. To see how $\zeta_T^2(\alpha, \beta)$ improves over ζ_{unif}^2, let

$$\zeta_{\text{fix}}^2(\alpha) = \min_{p \in \mathcal{A}} \sup_{w} V(p, w). \tag{10}$$

The quantity $\zeta_{\text{fix}}^2(\alpha)$ can be regarded as the minimum heterogeneity by using the best fixed sampling distribution in $\mathcal{A}$ that does not depend on parameter w. Let p_f be the solution of p to the min-max problem (10), that is, $\sup_w V(p_f, w) = \zeta_{\text{fix}}^2(\alpha)$. Since $p^{\text{unif}} \in \mathcal{A}$ for all $\alpha \ge 0$, it is easy to see that

$$\zeta_{\text{fix}}^2(\alpha) = \min_{p \in \mathcal{A}} \sup_{w} V(p, w) \le \sup_{w} V(p^{\text{unif}}, w) = \zeta_{\text{unif}}^2.$$

Note that $\zeta_T^2(\alpha, \beta)$ is a non-increasing function of β with any given α. Thus, we have

$$\zeta_T^2(\alpha, \beta) \le \zeta_{\text{fix}}^2(\alpha) \le \zeta_{\text{unif}}^2 \quad \forall\, 0 \le \alpha \le 1,\ \beta \ge 0. \tag{11}$$

See the proof in Appendix B.1. Note that the above inequality also implies that dynamic sampling distribution may potentially improve over a fixed sampling distribution. As we will see shortly, when $\zeta_{\text{unif}}^2 > \zeta_{\text{fix}}^2(\alpha)$, OSMD Sampler can always improve over uniform sampling asymptotically.

When $\beta \ge 2(T-1)$ and α is small enough such that

$$p_m^\star(w) \coloneqq \frac{\|\nabla F_m(w)\|_2}{\sum_{m'=1}^{M} \|\nabla F_{m'}(w)\|_2} \ge \frac{\alpha}{M} \quad \text{for all } w \text{ and } m \in [M],$$

we have

$$\zeta_T^2(\alpha, \beta) = \sup_w \min_{p \in \mathcal{A}} V(w, p) = \sup_w \left\{ \left(\frac{1}{M} \sum_{m=1}^{M} \|\nabla F_m(w)\|_2 \right)^2 - \|\nabla F(w)\|_2^2 \right\} \triangleq \zeta_{\min}^2,$$

which is the smallest heterogeneity possibly achievable.

Algorithm 3 Mini-batch SGD with OSMD Sampler

1: **Input:** Number of communication rounds T, number of clients chosen in each round K, local batch size B, initial model parameter w^1, stepsizes $\{\mu^t\}_{t=1}^T$, learning rate η and parameter $\alpha \in (0, 1]$.
2: **Output:** The final model parameter w^R.
3: **Initialize:** $\hat{p}^1 = p^{\text{unif}}$.
4: **for** $t = 1, \ldots, T$ **do**
5: Sample S^t with replacement from $[M]$ with probability $\hat{p}^t$, such that $|S^t| = K$.
6: **for** $m \in S^t$ **do**
7: Download the current model parameter w^t.
8: Locally sample a mini-batch $\mathcal{B}_m^t = \{\xi_m^{t,1}, \ldots, \xi_m^{t,B}\}$ i.i.d. uniformly random from $[n_m]$, where $|\mathcal{B}_m^t| = B$.
9: Locally compute and upload $g_m^t = (1/B) \sum_{b=1}^B \nabla \phi(w^t; \xi_m^{t,b})$ to the server.
10: **end for**
11: Server computes $a_m^t = \lambda_m^2 \|g_m^t\|^2$ for $m \in S^t$ and

$$g^t = \frac{1}{K} \sum_{m \in S^t} \frac{\lambda_m}{\hat{p}_m^t} g_m^t. \tag{12}$$

12: Server makes update of the model parameter $w^{t+1} \leftarrow w^t - \mu^t g^t$.
13: Server obtains updated sampling distribution $\hat{p}^{t+1}$ by Algorithm 1.
14: **end for**

4.1 Convergence Analysis of Mini-batch SGD with OSMD Sampler

We introduce the convergence analysis of mini-batch SGD with OSMD Sampler. The detailed algorithm is given in Algorithm 3. Compared to the classical mini-batch SGD, the key ingredients of Algorithm 3 are Line 13, where the server updates the sampling distribution by OSMD Sampler, and Line 5, where the server samples clients from a non-uniform sampling distribution. In (12), we use a weighted average to compute the global gradient.

Recall that B is the local batch size in Algorithm 3 and $K = |S^t|$. Let $D^F := F(w^1) - F^\star$. We then have the following convergence guarantee for Algorithm 3.

Theorem 4 *Assume Assumption 1—3 holds. Let $\{w^1, \ldots, w^T\}$ be the sequence of iterates generated by Algorithm 3 and let w^R denote an element of that sequence chosen uniformly at random. Let*

$$\eta = \frac{K\alpha^3}{MG^2} \sqrt{\frac{2 \log M + 4\beta \log(M/\alpha)}{T}}, \tag{13}$$

and $\mu_t \equiv \mu$ for all $t \in [T]$, where

$$\mu = \min \left\{ \frac{1}{L}, \frac{1}{\sigma} \sqrt{\frac{D^F K B \alpha}{LT}}, \frac{1}{\zeta_T(\alpha, \beta)} \sqrt{\frac{D^F K}{LT}}, \frac{\sqrt{D^F K} \alpha^{\frac{3}{2}}}{\sqrt{LM} T^{\frac{1}{4}} G \left(\frac{1}{2} \log M + \beta \log (M/\alpha) \right)^{\frac{1}{4}}} \right\},$$

we then have

$$\begin{aligned}&\mathbb{E}\left[\left\|\nabla F\left(w^R\right)\right\|^2\right]\\ &\lesssim \frac{D^F L}{T}+\frac{\sigma\sqrt{D^F L}}{\sqrt{TKB\alpha}}+\frac{\zeta_T(\alpha,\beta)\sqrt{D^F L}}{\sqrt{TK}}+\frac{\sqrt{D^F L}M^{\frac{1}{2}}G}{T^{\frac{3}{4}}K^{\frac{1}{2}}\alpha^{\frac{3}{2}}}\left(\frac{1}{2}\log M+\beta\log\left(M/\alpha\right)\right)^{\frac{1}{4}}.\end{aligned} \tag{14}$$

Proof The key proof technique is the construction of a ghost subset $\tilde{S}^t$ in each round that is sampled from a carefully designed comparator sampling distribution. In this way, the regret can be compared with the difference between the convergence rate under OSMD Sampler and the convergence rate under the comparator sampling distribution. Note that $\tilde{S}^t$ is only constructed for theoretical analysis, and does not need to be actually computed in practice. The rest of the proof then follows the regret analysis as in Theorem 6. See detailed proof in Appendix B.3. ■

To see how the convergence rate in Theorem 4 is better than the rate of uniform sampling, recall that the convergence rate of mini-batch SGD under uniform sampling (Ghadimi and Lan, 2013) is

$$R_{\text{unif}}^{\text{MB}} \coloneqq \frac{D^F L}{T}+\frac{\sigma\sqrt{D^F L}}{\sqrt{TKB}}+\frac{\zeta_{\text{unif}}\sqrt{D^F L}}{\sqrt{TK}}.$$

Denote the right hand side of (14) as $R_{\text{osmd}}^{\text{MB}}$, then to have $R_{\text{osmd}}^{\text{MB}} \lesssim R_{\text{unif}}^{\text{MB}}$, we only need that

$$\frac{\sqrt{D^F L}M^{\frac{1}{2}}G}{T^{\frac{3}{4}}K^{\frac{1}{2}}\alpha^{\frac{3}{2}}}\left(\frac{1}{2}\log M+\beta\log\left(M/\alpha\right)\right)^{\frac{1}{4}} \lesssim \frac{\left(\zeta_{\text{unif}}-\zeta_T(\alpha,\beta)\right)\sqrt{D^F L}}{\sqrt{TK}},$$

which is equivalent to a requirement that

$$\frac{M^{\frac{1}{2}}G}{\alpha^{\frac{3}{2}}}\left(\frac{\frac{1}{2}\log M+\beta\log\left(M/\alpha\right)}{T}\right)^{\frac{1}{4}} \lesssim \zeta_{\text{unif}}-\zeta_T(\alpha,\beta).$$

By treating all the other quantities except for T, ζ_{unif}, and $\zeta_T(\alpha,\beta)$ as constants and setting $\beta = o(T)$, we have the left hand side of the above inequality as $o(1)$. On the other hand, note that $\zeta_T(\alpha,\beta) \le \zeta_{\text{fix}} \le \zeta_{\text{unif}}$ for any $\beta \ge 0$ and $0 < \alpha \le 1$, where ζ_{fix} is defined in (10), so whenever $\zeta_{\text{unif}} > \zeta_{\text{fix}}$, we have $\zeta_{\text{unif}} - \zeta_T(\alpha,\beta) \ge \zeta_{\text{unif}} - \zeta_{\text{fix}} = \Omega(1)$. In conclusion, when $\zeta_{\text{unif}} > \zeta_{\text{fix}}$ and by setting $\beta = o(T)$, the OSMD Sampler can achieve a better convergence rate than uniform sampling for mini-batch SGD.

4.2 Convergence Analysis of FedAvg with OSMD Sampler

We introduce the convergence guarantee for FedAvg with OSMD Sampler. The detailed algorithm is given in Algorithm 4. Compared to FedAvg in McMahan et al. (2017), the differences between Algorithm 4 are Line 15, where the server updates the sampling distribution by OSMD Sampler, and Line 5, where the server samples clients from a non-uniform sampling distribution. In addition, in (15), we use a weighted average to update the global model parameter. Besides, note that when defining a_m^t, we rescale $\|g_m^t\|_2^2$ by $(\mu_l^t)^2 B$, this is to ensure that $\mathbb{E}[\|g_m^t\|_2^2/(\mu_l^t)^2 B] = \Theta(1)$ as $\mu_l^t \to 0$ and $B \to \infty$.

We have the following result about FedAvg with OSMD Sampler (Algorithm 4).

Algorithm 4 FedAvg with OSMD Sampler

Input: Communication rounds T, clients per round K, local steps B, initial model parameter w^1, global stepsizes $\{\mu^t\}_{t=1}^T$, local stepsizes $\{\mu_l^t\}_{t=1}^T$, learning rate η, and parameter $\alpha \in (0,1]$.
Output: The final model parameter w^R.
Initialize: $\hat{p}^1 = p^{\text{unif}}$.
for $t = 1, \dots, T$ **do**
  Sample S^t with replacement from $[M]$ with probability $\hat{p}^t$, such that $|S^t| = K$.
  for $m \in S^t$ **do**
    Download the current model parameter w^t and let $w_m^{t,0} = w^t$.
    for $b = 0, 1, \dots, B-1$ **do**
      Sample $\xi_m^{t,b}$ from $[n_m]$ uniformly random.
      Compute $w_m^{t,b+1} = w_m^{t,b} - \mu_l^t \nabla \phi(w_m^{t,b}; \xi_m^{t,b})$.
    end for
    Locally compute $g_m^t = w^t - w_m^{t,B}$ and upload it to the server.
  end for
  Server computes

$$a_m^t = \frac{\lambda_m^2 \|g_m^t\|_2^2}{(\mu_l^t)^2 B} = \frac{\lambda_m^2}{B} \left\| \sum_{b=0}^{B-1} \nabla \phi(w_m^{t,b}; \xi_m^{t,b}) \right\|_2^2$$

  for $m \in S^t$ and let

$$w^{t+1} = w^t - \frac{\mu^t}{K} \sum_{m \in S^t} \frac{\lambda_m}{\hat{p}_m^t} g_m^t. \tag{15}$$

  Server obtains updated sampling distribution $\hat{p}^{t+1}$ by Algorithm 1.
end for

Theorem 5 *Recall that B is the local batch size in Algorithm 3 and $K = |S^t|$. Let $D^F := F(w^1) - F^\star$. Assume Assumption 1—3 holds. Let $\{w^1, \dots, w^T\}$ be the sequence of iterates generated by Algorithm 4 and let w^R denote an element of that sequence chosen uniformly at random. Let*

$$\eta = \frac{K\alpha^3}{MBG^2} \sqrt{\frac{2\log M + 4\beta \log(M/\alpha)}{T}}, \tag{16}$$

$\mu^t = \mu \geq 1$ and $\mu_l^t = \mu_l$ for all $t \in [T]$, where

$$\mu_l = \min \left\{ \frac{1}{4\mu BL} \sqrt{\frac{1}{2 + 1/\alpha}}, \frac{(D^F)^{\frac{1}{3}}}{\left(4 + \frac{2}{\alpha}\right)^{\frac{1}{3}} \mu BL^{\frac{2}{3}} \left(\zeta_{unif}^2 + \frac{\sigma^2}{2B}\right)^{\frac{1}{3}} T^{\frac{1}{3}}}, \right.$$
$$\left. \frac{\sqrt{2D^F}}{\mu B\sqrt{L} \sqrt{\frac{2\zeta_T^2(\alpha,\beta)}{K} + \frac{\sigma^2}{KB\alpha} + \sqrt{\frac{\frac{1}{2}\log M + \beta \log(M/\alpha)}{T}}} \sqrt{T}} \right\},$$

we then have

$$\mathbb{E}\left[\left\|\nabla F(w^R)\right\|^2\right] \lesssim \frac{D^F L\sqrt{2+\frac{1}{\alpha}}}{T} + \frac{\left(4+\frac{2}{\alpha}\right)^{\frac{1}{3}}(D^F L)^{\frac{2}{3}}\zeta_{unif}^{\frac{2}{3}}}{T^{\frac{2}{3}}} + \frac{\left(4+\frac{2}{\alpha}\right)^{\frac{1}{3}}(D^F L)^{\frac{2}{3}}\sigma^{\frac{2}{3}}}{B^{\frac{1}{3}}T^{\frac{2}{3}}} + \frac{\sqrt{D^F L}\zeta_T(\alpha,\beta)}{\sqrt{TK}} + \frac{\sqrt{D^F L}\sigma}{\sqrt{TKB\alpha}} + \frac{\sqrt{D^F L}}{\sqrt{T}}\left(\frac{\frac{1}{2}\log M + \beta\log(M/\alpha)}{T}\right)^{\frac{1}{4}}. \tag{17}$$

Proof See proof in Appendix B.4. Similar to the proof of Theorem 4, the key technique is to construct a novel ghost subset sampled from the comparator sampling distribution. The rest of the proof then follows the regret analysis as in Theorem 6. ■

To see how OSMD Sampler improves over uniform sampling, note that the convergence rate of FedAvg under uniform sampling (Karimireddy et al., 2020b) is

$$R_{\text{unif}}^{\text{Avg}} := \frac{D^F L}{T} + \frac{(D^F L)^{\frac{2}{3}}\zeta_{\text{unif}}^{\frac{2}{3}}}{T^{\frac{2}{3}}} + \frac{(D^F L)^{\frac{2}{3}}\sigma^{\frac{2}{3}}}{B^{\frac{1}{3}}T^{\frac{2}{3}}} + \frac{\sqrt{D^F L}\zeta_{\text{unif}}}{\sqrt{TK}} + \frac{\sqrt{D^F L}\sigma}{\sqrt{TKB}}.$$

Denote the right hand side of (17) as $R_{\text{osmd}}^{\text{Avg}}$. By treating all the other quantities except for T, ζ_{unif} and $\zeta_T(\alpha,\beta)$ as constants, to have $R_{\text{osmd}}^{\text{Avg}} \lesssim R_{\text{unif}}^{\text{Avg}}$, we only need that

$$\sqrt{K}\left(\frac{\frac{1}{2}\log M + \beta\log(M/\alpha)}{T}\right)^{\frac{1}{4}} \lesssim \zeta_{\text{unif}} - \zeta_T(\alpha,\beta).$$

Similar to the argument in Section 4.1, by setting $\beta = o(T)$, we have the left hand side of the above inequality as $o(1)$. On the other hand, when $\zeta_{\text{unif}} > \zeta_{\text{fix}}$, we have $\zeta_{\text{unif}} - \zeta_T(\alpha,\beta) \geq \zeta_{\text{unif}} - \zeta_{\text{fix}} = \Omega(1)$. Thus, when $\zeta_{\text{unif}} > \zeta_{\text{fix}}$ and by setting $\beta = o(T)$, we have shown that OSMD Sampler can achieve a better convergence rate than uniform sampling for FedAvg.

5. Regret Analysis of OSMD Sampler

In this section, we provide the regret analysis that serves as a key component of the optimization analysis in Section 4. We first describe the dynamic regret used to measure the performance of an online algorithm that generates a sequence of sampling distributions $\{\hat{p}\}_{t\geq 1}$ in a non-stationary environment. Given any comparator sequence $q^{1:T} \in \mathcal{P}_{M-1}^T$, the dynamic regret is defined as

$$\text{D-Regret}_T(q^{1:T}) = \bar{L}\left(\hat{p}^{1:T}\right) - \bar{L}\left(q^{1:T}\right). \tag{18}$$

In contrast, the static regret measures the performance of an algorithm relative to the best fixed sampling distribution, that is, it restricts $q^1 = \cdots = q^T$ (Namkoong et al., 2017; Salehi et al., 2017; Borsos et al., 2018, 2019). When using a fixed comparator $q^1 = \cdots = q^T = q$, we write the regret as $\text{D-Regret}_T(q)$.

Recall that the total variation of a comparator sequence $q^{1:T}$ is $\text{TV}\left(q^{1:T}\right) = \sum_{t=1}^{T-1}\|q^{t+1} - q^t\|_1$. The total variation measures how variable a sequence is. The larger the total variation $\text{TV}(q^{1:T})$, the more variable $q^{1:T}$ is, and such a comparator sequence is harder to match.

We also need the following quantities that quantify how far q^t is from $\mathcal{A}$. Given $q^t \in \mathcal{P}_{M-1}$ and $\alpha \in (0,1]$, let

$$\psi(q^t,\alpha) \coloneqq \sum_{m=1}^{M}\left(\frac{\alpha}{M}-q_m^t\right)\mathbb{1}\left\{q_m^t<\frac{\alpha}{M}\right\}, \qquad \omega(q^t,\alpha) \coloneqq \frac{\sum_{m=1}^{M}\left(\frac{\alpha}{M}-q_m^t\right)\mathbb{1}\left\{q_m^t<\frac{\alpha}{M}\right\}}{\sum_{m=1}^{M}\left(q_m^t-\frac{\alpha}{M}\right)\mathbb{1}\left\{q_m^t\geq\frac{\alpha}{M}\right\}},$$
$$\phi(q^t,\alpha) \coloneqq \frac{\omega(q^t,\alpha)}{1-\omega(q^t,\alpha)\left(1-\frac{\alpha}{M}\right)}. \tag{19}$$

We will use these quantities to characterize the projection error in the following theorem, which is the main result of this section.

Theorem 6 *Let $\eta_t \equiv \eta$ for all t and $\hat{p}^{1:T}$ be a sequence generated by Algorithm 1. For any comparator sequence $q^{1:T}$, where q^t is allowed to be random, we have*

$$\textit{D-Regret}_T(q^{1:T}) \leq \underbrace{\frac{\log M}{\eta}+\frac{2\log(M/\alpha)}{\eta}\mathbb{E}\left[TV\left(q^{1:T}\right)\right]+\frac{\eta M^6}{2K^2\alpha^6}\sum_{t=1}^{T}\mathbb{E}\left[\left(a_{\max}^t\right)^2\right]}_{\textit{Intrinsic Regret}}+$$
$$\underbrace{\frac{8\log(M/\alpha)}{\eta}\sum_{t=1}^{T}\mathbb{E}\left[\psi(q^t,\alpha)\right]+\frac{1}{K}\sum_{t=1}^{T}\mathbb{E}\left[\phi(q^t,\alpha)l_t(q^t)\right]}_{\textit{Projection Error}},$$

where $a_{\max}^t \coloneqq \max_{1\leq m\leq M} a_m^t = \max_{1\leq m\leq M}\lambda_m^2\|g_m^t\|_2^2$ for all $t\in[T]$.

Proof The major challenge of the proof is to construct a projection of *any* comparator sequence $q^{1:T}$ onto $\mathcal{A}^T$ and bound the projection error. To the best of our knowledge, this bound on the projection error of a dynamic sequence is novel. Since the comparator sequence can be arbitrary, thus analyzing the projection error is nontrivial. Another challenge is to deal with the dynamic comparator, which requires us to connect the cumulative regret with the total variation of the comparator sequence. See Appendix B.5 for more details. ■

From Theorem 1, we see that the bound on the dynamic regret consists of two parts. The first part is the intrinsic regret, quantifying the difficulty of tracking a comparator sequence in $\mathcal{A}^T$; the second part is the projection error, arising from projecting the comparator sequence onto $\mathcal{A}^T$. As shown in Appendix B.5, we have $0\leq\omega(q^t,\alpha)\leq 1$ for all $\alpha\in[0,1]$, which implies that $\phi(q^t,\alpha)\leq M/\alpha$. Besides, $\psi(q^t,\alpha)\leq\sum_{m=1}^{M}(\alpha/M)\mathbb{1}\left\{q_m^t<(\alpha/M)\right\}\leq\alpha$, and the projection error can be upper bounded by $(8T\alpha\log(M/\alpha))/\eta+(M/\alpha)\sum_{t=1}^{T}\mathbb{E}[l_t(q^t)]$. More importantly, when $q_m^t\in\mathcal{A}$, we have $\psi(q^t,\alpha)=\omega(q^t,\alpha)=\phi(q^t,\alpha)=0$. Thus, when the comparator sequence belongs to $\mathcal{A}^T$, the projection error vanishes and we only have the intrinsic regret. As α decreases from one to zero, the intrinsic regret gets larger, while we are allowing a larger class of comparator sequences; on the other hand, the projection error decreases to zero, since the gap between $\mathcal{A}$ and $\mathcal{P}_{M-1}$ vanishes with α. An optimal choice of α balances the two sources of regret.

6. Adaptive-OSMD Sampler

In this section, we discuss an extension of OSMD Sampler that can automatically choose the learning rate η and is agnostic to the optimization method. There are two main tuning parameters in OSMD Sampler, namely α and η. As we show empirically in Section 7.3, the performance of the algorithm is relatively robust to the choice of α. However, the choice of η may have a large effect on the performance of OSMD Sampler. One way to choose η is by minimizing the regret in Theorem 6, which is stated in the following corollary.

Corollary 7 *Let $\eta_t \equiv \eta$ for all t and $\hat{p}^{1:T}$ be a sequence generated by Algorithm 1. Assume that there exists $A_{\max} > 0$ such that $a^t_{\max} \leq A_{\max}$ for all t, where $a^t_{\max} = \max_{1\leq m\leq M} a^t_m = \max_{1\leq m\leq M} \lambda_m^2 \|g_m^t\|_2^2$. For any comparator sequence $q^{1:T}$, where q^t is allowed to be random, such that $q^t \in \mathcal{A}$ for all $t \in [T]$ and $\mathbb{E}[TV(q^{1:T})] \leq \beta$, let*

$$\eta = \frac{K\alpha^3}{M^3 A_{\max}} \sqrt{\frac{2\log M + 4\beta \log(M/\alpha)}{T}}, \tag{20}$$

then

$$\textit{D-Regret}_T(q^{1:T}) \leq \frac{M^3 A_{\max}}{K\alpha^3} \sqrt{T\left[\frac{1}{2}\log M + \beta \log\left(M/\alpha\right)\right]}. \tag{21}$$

The proof of Corollary 7 follows directly from Theorem 6. Note that under Assumption 1—3 and when $\lambda_m = \frac{1}{M}$ for all $m \in [M]$, we have $A_{\max} = \frac{G^2}{M^2}$ for Mini-batch SGD (Algorithm 3) and $A_{\max} = \frac{BG^2}{M^2}$ for FedAvg (Algorithm 4.2)[7], thus (20) recovers (13) and (16).

In practice, since the gradient norm is usually decreasing, we can estimate $A_{\max}$ by adding a pre-training phase where we broadcast the initial model parameter w^0 to all devices before the start of the training, and collect the returned $\|g_m^0\|_2^2$ from all responsive devices, which we denote as S^0. Then we can estimate $A_{\max}$ by $\hat{A}_{\max} = \max_{m\in S^0} \lambda_m \|g_m^0\|_2^2$.

On the other hand, the optimal choice of β in (20) depends on specific problems, and is hard to estimate before training starts. Thus, it is preferable to have a tuning strategy that is adaptive to any $\beta > 0$, which we describe in the following.

The main idea is to run a set of expert algorithms, each with a different learning rate for Algorithm 1. We then use a prediction-with-expert-advice algorithm to track the best performing expert algorithm.[8] More specifically, we define the set of expert learning rates as

$$\mathcal{E} \coloneqq \left\{ 2^{e-1} \cdot \frac{K\alpha^3}{M^3 A_{\max}} \sqrt{\frac{2\log M}{T}} \;\middle|\; e = 1, 2, \ldots, E \right\}, \tag{22}$$

where

$$E = \left\lceil \frac{1}{2}\log_2\left(1 + \frac{4\log(M/\alpha)}{\log M}(T-1)\right)\right\rceil + 1. \tag{23}$$

Then for each $\eta_e \in \mathcal{E}$, Adaptive-OSMD Sampler algorithm runs an expert algorithm to generate a sequence of sampling distributions $\hat{p}_e^{1:T}$. Meanwhile, it also runs a meta-algorithm

7. See Appendix B.3 and Appendix B.4 for proof.
8. We refer the reader to Cesa-Bianchi and Lugosi (2006, Chapter 2) for an overview of prediction-with-expert-advice algorithms.

Algorithm 5 Adaptive-OSMD Sampler

1: **Input:** Meta learning rate γ; the set of expert learning rates $\mathcal{E} = \{\eta_1 \le \eta_2 \le \cdots \le \eta_E\}$ with $E = |\mathcal{E}|$; parameter $\alpha \in (0, 1]$, $\mathcal{A} = \mathcal{P}_{M-1} \cap [\alpha/M, \infty)^M$; number of iterations T; initial distribution p^{init}.
2: **Output:** $\hat{p}^{1:T}$.
3: Set $\theta_e^1 = (1 + 1/E)/(e(e+1))$ and $\hat{p}_e^1 = p^{\text{init}}$, $\forall e \in [E]$.
4: **for** $t = 1, 2, \ldots, T-1$ **do**
5: Compute $\hat{p}^t = \sum_{e=1}^E \theta_e^t \hat{p}_e^t$.
6: Sample S^t by $\hat{p}^t$.
7: **for** $e = 1, 2, \ldots, E$ **do**
8: Compute $\hat{l}_t(\hat{p}_e^t; \hat{p}^t)$ via (6) and $\nabla \hat{l}_t(\hat{p}_e^t; \hat{p}^t)$ via (7).
9: Solve $\hat{p}_e^{t+1} = \arg\min_{p \in \mathcal{A}} \eta_e \langle p, \nabla \hat{l}_t(\hat{p}_e^t; \hat{p}^t) \rangle + D_\Phi\left(p \,\|\, \hat{p}_e^t\right)$ via Algorithm 2.
10: **end for**
11: Update the weight of each expert:

$$\theta_e^{t+1} = \frac{\theta_e^t \exp\left\{-\gamma \hat{l}_t(\hat{p}_e^t; \hat{p}^t)\right\}}{\sum_{e=1}^E \theta_e^t \exp\left\{-\gamma \hat{l}_t(\hat{p}_e^t; \hat{p}^t)\right\}}, \quad \forall e \in [E].$$

12: **end for**

that uses exponentially-weighted-average strategy to aggregate $\{\hat{p}_e^{1:T}\}_{e=1}^E$ into a single output $\hat{p}^{1:T}$, which achieves performance close to the best expert.

Algorithm 5 details Adaptive-OSMD Sampler. Note that since we can compute $\hat{l}_t(\hat{p}_e^t; \hat{p}^t)$ and $\nabla \hat{l}_t(\hat{p}_e^t; \hat{p}^t)$ directly, there is no need to use a surrogate loss as in van Erven and Koolen (2016) and Zhang et al. (2018).

From the computational perspective, the major cost comes from solving step 9 of Algorithm 5, which needs to be run for a total number of $T|\mathcal{E}| = O(T \log_2 T)$ times. Compared with Algorithm 1, the computational complexity only increases by $\log_2 T$ times.

We have the following regret guarantee on Algorithm 5.

Theorem 8 *Assume that there exists $A_{\max} > 0$ such that $a_{\max}^t \le A_{\max}$ for all t, where $a_{\max}^t = \max_{1 \le m \le M} a_m^t = \max_{1 \le m \le M} \lambda_m^2 \|g_m^t\|_2^2$. Let $\hat{p}^{1:T}$ be the output of Algorithm 5 with $\gamma = \frac{\alpha}{M}\sqrt{\frac{8K}{TA_{\max}}}$, $p^{init} = p^{\text{unif}}$ and $\mathcal{E}$ as in (22). Then for any comparator sequence $q^{1:T}$, where q^t is allowed to be random, such that $q^t \in \mathcal{A}$ for all $t \in [T]$ and $\mathbb{E}[TV(q^{1:T})] \le \beta$, we have*

$$\textit{D-Regret}_T(q^{1:T}) \le \frac{3M^3 A_{\max}}{K\alpha^3}\sqrt{T\left[\frac{1}{2}\log M + \beta \log\left(M/\alpha\right)\right]} + \frac{M}{\alpha}\sqrt{\frac{TA_{\max}}{8K}}\left(1 + 2\log E\right).$$

Proof See Appendix B.6. ■

Since the additional regret term is $\tilde{O}((M/\alpha)\sqrt{T/K})$, which is no larger than the first term asymptotically in its dependency on T except for log terms, the bound on the regret is of the same order as in (21). However, we do not need to specify β to set the learning rate.

Following Theorem 8 and the proofs of Theorem 4 and Theorem 5, we then have the following optimization guarantees on the Adaptive-OSMD Sampler.

Theorem 9 *Assume Assumption 1—3 holds and* $\lambda_m = \frac{1}{M}$ *for all* $m \in [M]$.

- *Mini-batch SGD with Adaptive-OSMD Sampler. Let* $\{w^1, \dots, w^T\}$ *be the sequence of iterates generated by Algorithm 3, where in Line 13, the sampling distribution is updated by Algorithm 5 with* $A_{\max} = \frac{G^2}{M^2}$. *Let* w^R *denote an element of that sequence chosen uniformly at random. Besides, let* $\mu_t = \mu$ *for all* $t \in [T]$, *where*

$$\mu = \min\left\{\frac{1}{L}, \frac{1}{\sigma}\sqrt{\frac{D^F K B\alpha}{LT}}, \frac{1}{\zeta_T(\alpha,\beta)}\sqrt{\frac{D^F K}{LT}}, \frac{\sqrt{D^F K}\alpha^{\frac{3}{2}}}{\sqrt{LM}T^{\frac{1}{4}}G\left(\frac{1}{2}\log M + \beta\log\left(M/\alpha\right)\right)^{\frac{1}{4}}}, \sqrt{\frac{\alpha D^F}{LG}}\left(\frac{K}{T}\right)^{\frac{1}{2}}\sqrt{\frac{1}{1+2\log E}}\right\},$$

we then have

$$\begin{aligned}&\mathbb{E}\left[\left\|\nabla F\left(w^R\right)\right\|^2\right]\\ \lesssim\ & \frac{D^F L}{T} + \frac{\sigma\sqrt{D^F L}}{\sqrt{TKB\alpha}} + \frac{\zeta_T(\alpha,\beta)\sqrt{D^F L}}{\sqrt{TK}} + \frac{\sqrt{D^F L}M^{\frac{1}{2}}G}{T^{\frac{3}{4}}K^{\frac{1}{2}}\alpha^{\frac{3}{2}}}\left(\frac{1}{2}\log M + \beta\log\left(M/\alpha\right)\right)^{\frac{1}{4}}\\ &+ \sqrt{\frac{D^F LG}{\alpha}}\left(\frac{1}{K}\right)^{\frac{1}{4}}\left(\frac{1}{T}\right)^{\frac{3}{4}}\sqrt{1+2\log E}.\end{aligned} \tag{24}$$

- *FedAvg with Adaptive-OSMD Sampler. Let* $\{w^1, \dots, w^T\}$ *be the sequence of iterates generated by Algorithm 4, where in Line 13, the sampling distribution is updated by Algorithm 5 with* $A_{\max} = \frac{BG^2}{M^2}$. *Let* w^R *denote an element of that sequence chosen uniformly at random. Besides, let* $\mu^t = \mu \geq 1$ *and* $\mu_l^t = \mu_l$ *for all* $t \in [T]$, *where*

$$\mu_l = \min\left\{\frac{1}{4\mu BL}\sqrt{\frac{1}{2+1/\alpha}}, \frac{(D^F)^{\frac{1}{3}}}{\left(4+\frac{2}{\alpha}\right)^{\frac{1}{3}}\mu BL^{\frac{2}{3}}\left(\zeta_{unif}^2 + \frac{\sigma^2}{2B}\right)^{\frac{1}{3}}T^{\frac{1}{3}}}, \frac{\sqrt{2D^F}}{\mu B\sqrt{L}\sqrt{\frac{2\zeta_T^2(\alpha,\beta)}{K} + \frac{\sigma^2}{KB\alpha} + \sqrt{\frac{\frac{1}{2}\log M + \beta\log(M/\alpha)}{T}}}\sqrt{T}}, \frac{1}{\mu}\sqrt{\frac{2\alpha D^F}{LG\left(1+2\log E\right)}}\left(\frac{1}{B}\right)^{\frac{3}{4}}\left(\frac{8K}{T}\right)^{\frac{1}{4}}\right\},$$

we then have

$$\mathbb{E}\left[\left\|\nabla F(w^R)\right\|_2^2\right] \lesssim \frac{D^F L\sqrt{2+\frac{1}{\alpha}}}{T} + \frac{\left(4+\frac{2}{\alpha}\right)^{\frac{1}{3}}(D^F L)^{\frac{2}{3}}\zeta_{unif}^{\frac{2}{3}}}{T^{\frac{2}{3}}} + \frac{\left(4+\frac{2}{\alpha}\right)^{\frac{1}{3}}(D^F L)^{\frac{2}{3}}\sigma^{\frac{2}{3}}}{B^{\frac{1}{3}}T^{\frac{2}{3}}} + \frac{\sqrt{D^F L}\zeta_T(\alpha,\beta)}{\sqrt{TK}} + \frac{\sqrt{D^F L}\sigma}{\sqrt{TKB\alpha}} + \frac{\sqrt{D^F L}}{\sqrt{T}}\left(\frac{\frac{1}{2}\log M + \beta\log(M/\alpha)}{T}\right)^{\frac{1}{4}} + \sqrt{\frac{D^F LG}{\alpha}}\left(\frac{1}{KB}\right)^{\frac{1}{4}}\left(\frac{1}{T}\right)^{\frac{3}{4}}\sqrt{1+2\log E}. \quad (25)$$

Compare (14) with (24) (respectively, (17) with (25)), there is an additional error term in (24) (respectively, (25)). However, this additional error term is no larger than the previous error term when considering its dependency on T. Furthermore, when we let $\beta = T^c$ with $0 < c < 1$ and treat all the other parameters except for β and T to be constants, we then have the last term dominated by the penultimate term. On the other hand, in exchange to the additional error, we do not need to specify β when setting the learning rate η.

Note that Adaptive-OSMD Sampler still requires the number of iterations T as an input; however, in some applications, T is not available before the training starts. For example, one may use some stopping criterion to determine when to stop. To deal with such cases, in Appendix A.1, we introduce an extension of Adaptive-OSMD Sampler that does not need T as input by using doubling trick. The extension algorithm enjoys similar regret and optimization guarantees as Adaptive-OSMD Sampler. See Appendix A.1 for more details.

7. Simulation Experiments

In this section, we use simulated data to demonstrate the performance of Adaptive-OSMD Sampler (Algorithm 5). We compare our method against uniform sampling in Section 7.1 and compare against other bandit feedback online learning samplers in Section 7.2. In addition, we examine the robustness of Adaptive-OSMD Sampler to the choice of α in Section 7.3, while in Section 7.4, we compare Adaptive-OSMD Sampler with the Lipschitz constant based importance sampling.

We generate data as follows. We set the number of clients as $M = 100$, and each client has $n_m = 100$ samples, $m \in [M]$. Samples on each client are generated as

$$y_{m,i} = \langle w_\star, x_{m,i}\rangle + N(0, 0.1^2), \qquad i \in [n_m], \quad (26)$$

where the coefficient vector $w_\star \in \mathbb{R}^d$ has elements generated as i.i.d. $N(10, 3)$, and the feature vector $x_{m,i} \in \mathbb{R}^d$ is generated as $x_{m,i} \sim N(0, \Sigma_m)$, where $\Sigma_m = s_m \cdot \Sigma$, Σ is a diagonal matrix with $\Sigma_{jj} = \kappa^{(j-1)/(d-1)-1}$, $\forall j \in [d]$ and $\kappa > 0$ is the condition number of Σ. We generate $\{s_m\}_{m=1}^M$ i.i.d. from $e^{N(0,\sigma^2)}$ and rescale them as $s_m \leftarrow (s_m / \max_{m\in[M]} s_m) \times 10$ so that $s_m \le 10$ for all $m \in [M]$. In this setting, κ controls the difficulty of each problem when solved separately, while σ controls the level of heterogeneity across clients. In all experiments, we fix $\kappa = 25$, which corresponds to a hard problem, and change σ to simulate different heterogeneity levels. We expect that uniform sampling suffers when the

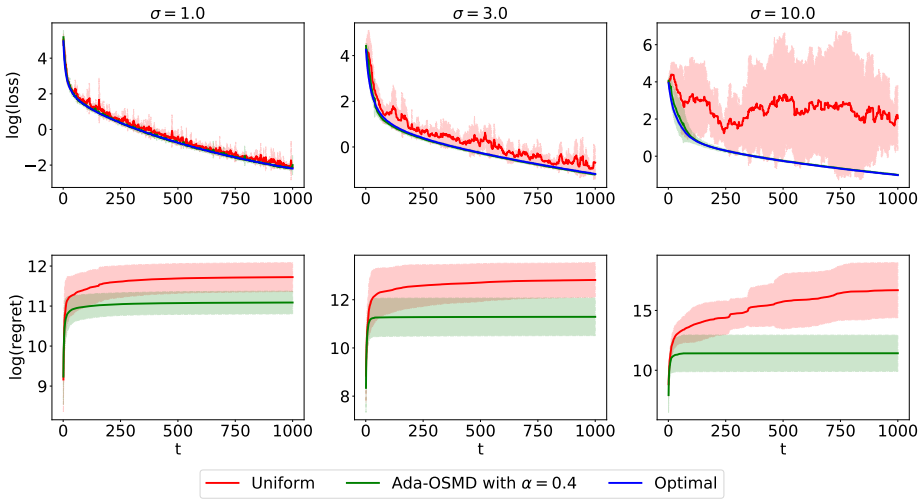

Figure 1: The training loss (top row) and cumulative regret (bottom row) are compared for the Adaptive-OSMD Sampler, Uniform Sampler, and Optimal Sampler, under $\sigma = 1.0$, $\sigma = 3.0$, and $\sigma = 10.0$. Solid lines represent the mean values, while the shaded regions indicate mean $\pm$ standard deviation across independent runs.

heterogeneity level is high. The dimension d of the problem is set as $d = 10$. The results are averaged over 10 independent runs.

We use the mean squared error loss defined as

$$L(w) = \frac{1}{M}\sum_{m=1}^{M} L_m(w), \qquad \text{where} \qquad L_m(w) = \frac{1}{2n_m}(y_{m,i} - \langle w_\star, x_{m,i}\rangle)^2.$$

We use the stochastic gradient descent to make global updates. At each round t, we choose a subset of $K = 5$ clients, denoted as S^t. For each client $m \in S^t$, we choose a mini-batch of samples, $\mathcal{B}_m^t$, of size $\bar{B} = 10$, and compute the mini-batch stochastic gradient. The parameter w is updated as

$$w^{t+1} = w^t + \frac{\mu_{\text{SGD}}}{MK\bar{B}} \sum_{m\in S^t} \frac{1}{p_m^t} \sum_{i\in\mathcal{B}_m^t} (y_{m,i} - \langle w_\star, x_{m,i}\rangle) \cdot x_{m,i},$$

where μ_{SGD} is the learning rate, set as $\mu_{\text{SGD}} = 0.1$ in simulations.

In all experiments, we set α in Adaptive-OSMD Sampler as $\alpha = 0.4$. The tuning parameters for MABS, VRB and Avare are set as in their original papers.

7.1 Adaptive-OSMD Sampler vs Uniform Sampling

The results of the training process and the cumulative regret are shown in Figure 1. For the training loss, we see that when the heterogeneity level is low ($\sigma = 1.0$), the uniform sampling performs as well as Adaptive-OSMD Sampler and theoretically optimal sampling; however, as the heterogeneity level increases, the performance of uniform sampling gradually suffers;

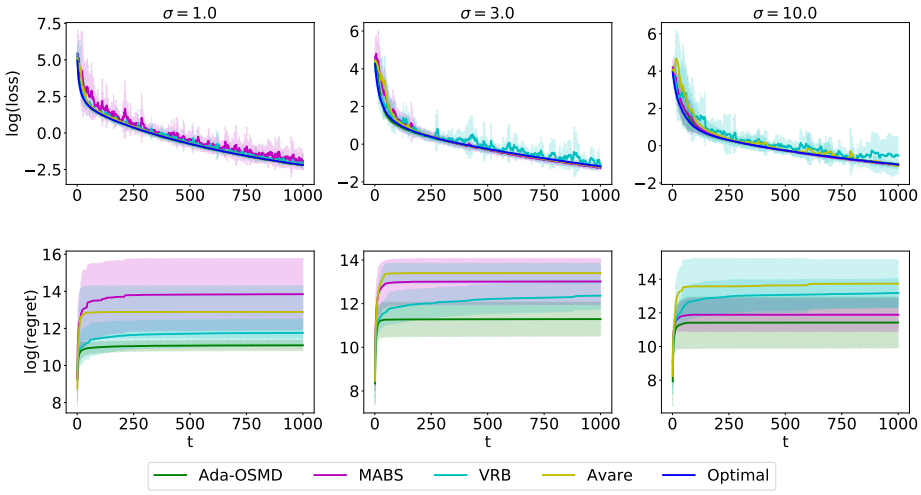

Figure 2: The training loss (top row) and cumulative regret (bottom row) are compared across the Adaptive-OSMD Sampler, MABS, VRB, and Avare methods for $\sigma = 1.0$, $\sigma = 3.0$, and $\sigma = 10.0$. Solid lines represent the mean values, while shaded regions indicate mean $\pm$ standard deviation across independent runs.

when $\sigma = 10.0$, uniform sampling performs poorly. On the other hand, Adaptive-OSMD Sampler performs well across all levels of heterogeneity and is very close to the theoretically optimal sampling. Similarly, for the cumulative regret, when the heterogeneity level is low, the cumulative regret of uniform sampling is close to Adaptive-OSMD Sampler; however, when the heterogeneity level increases, the cumulative regret of uniform sampling gets much larger than Adaptive-OSMD Sampler. Based on the above results, we can conclude that while the widely used choice of uniform sampling may be reasonable when heterogeneity is low, our proposed sampling strategy is robust across different levels of heterogeneity, and thus should be considered as the default option.

7.2 Adaptive-OSMD Sampler vs MABS vs VRB vs Avare

We compare Adaptive-OSMD Sampler to other bandit feedback online learning samplers: MABS (Salehi et al., 2017), VRB (Borsos et al., 2018) and Avare (Hanchi and Stephens, 2020). Training loss and cumulative regret are shown in Figure 2. We see that while VRB and Avare perform better when the heterogeneity level is low and MABS performs better when the heterogeneity level is high, Adaptive-OSMD Sampler always achieves the best in both training loss and cumulative regret across all different levels of heterogeneity. Thus, we conclude that Adaptive-OSMD is a better choice than other online learning samplers.

7.3 Robustness of Adaptive-OSMD Sampler to the Choice of α

We examine the robustness of Adaptive-OSMD Sampler to the choice of α. We run Adaptive-OSMD Sampler separately for each $\alpha \in \{0.01, 0.1, 0.4, 0.7, 0.9, 1.0\}$. Note that when $\alpha = 1.0$, the Adaptive-OSMD Sampler outputs a uniform distribution. Training loss

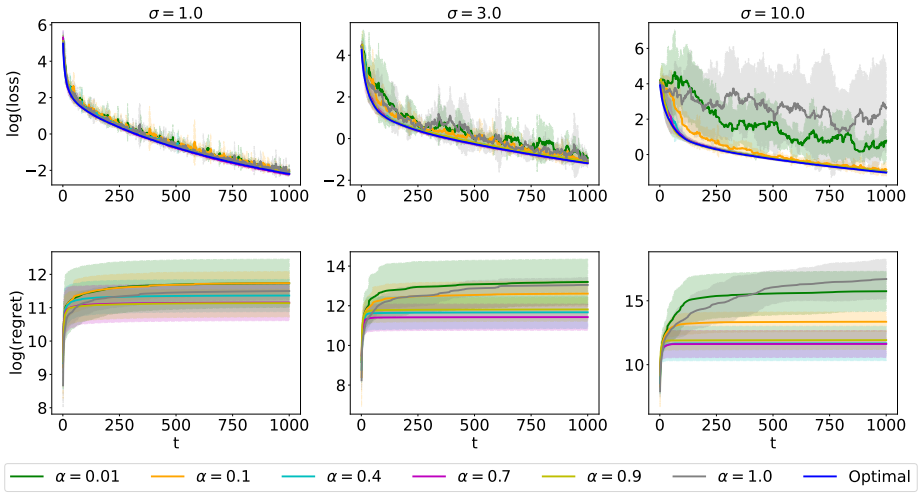

Figure 3: The training loss (top row) and cumulative regret (bottom row) are shown for the Adaptive-OSMD Sampler with different values of α under $\sigma = 1.0$, $\sigma = 3.0$, and $\sigma = 10.0$. Solid lines represent the mean values, while shaded regions indicate mean $\pm$ standard deviation across independent runs.

and cumulative regret are shown in Figure 3. We observe that Adaptive-OSMD Sampler is robust to the choice of α, and performs well as long as α is not too close to zero or too close to one.

7.4 Dynamic Sampling Distribution v.s. Fixed Sampling Distribution

In this paper, we allow both our sampling distribution and competitor sampling distribution to change over time, while previous studies either use a fixed sampling distribution (Zhao and Zhang, 2015; Needell et al., 2016) or they compare against a fixed sampling distribution (Namkoong et al., 2017; Salehi et al., 2017; Borsos et al., 2018, 2019). In this section, we show that under certain settings, a dynamic sampling distribution can achieve a significant advantage over a fixed sampling distribution. More specifically, we compare the Adaptive-OSMD Sampler with the Lipschitz constant-based importance sampling distribution proposed by Zhao and Zhang (2015); Needell et al. (2016), which we denote as p^{IS}.

We still use the same model as in (26) to generate data. but we generate $w_\star$ and $x_{m,i}$ differently. Motivated by Zhao et al. (2023), for each $m \in [M]$, we choose uniformly at random one dimension among $\mathbb{R}^d$, denoted as $\text{supp}(m) \in [d]$, as the support of $x_{m,i}$ for all $i \in [n_m]$, while the remaining dimensions of $x_{m,i}$ are set to be zero. The nonzero dimension of $x_{m,i}$ is generated from $N(1.0, 0.1^2)$. The entries of $w_\star$ are generated i.i.d. from $e^{N(0,\nu^2)}$. Therefore, ν controls the variance of entries of $w_\star$.

Besides, we choose the optimal stepsize from the set $\{1.0, 0.5, 0.1, 0.05, 0.01\}$ for each method separately. The final result is shown in Figure 4. We see that Adaptive-OSMD Sampler performs better than p^{IS} across all levels of ν. Note that in practice, in order to implement p^{IS}, we need prior information about Lipschitz constants of $L_m(\cdot)$'s, while

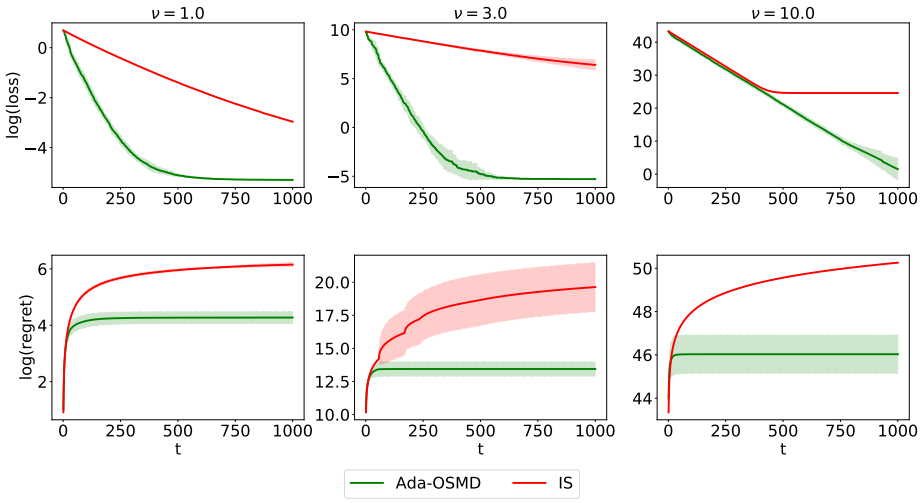

Figure 4: The training loss (top row) and cumulative regret (bottom row) are compared between the Adaptive-OSMD Sampler and p^{IS} for $\nu = 1.0$, $\nu = 3.0$, and $\nu = 10.0$. Solid lines represent the mean values, while shaded regions illustrate mean $\pm$ standard deviation across independent runs.

Adaptive-OSMD Sampler does not need prior information. This way, our proposed method does not only have better practical performance, but also requires less prior information.

8. Real Data Experiment

We compare Adaptive-OSMD Sampler with uniform sampling and other online learning samplers including MABS (Salehi et al., 2017), VRB (Borsos et al., 2018) and Avare (Hanchi and Stephens, 2020) on real data. We use three commonly used computer vision data sets: MNIST (LeCun and Cortes, 2010)[9], KMINIST (Clanuwat et al., 2018)[10], and FMINST (Xiao et al., 2017)[11]. We set the number of devices to be $M = 500$. To better simulate the situation where our method brings significant convergence speed improvement, we create a highly skewed sample size distribution of the training set among clients: 65% of clients have only one training sample, 20% of clients have 5 training samples, 10% of clients have 30 training samples, and 5% of clients have 100 training samples. This setting tries to illustrate a real-life situation where most of the data come from a small fraction of users, while most of the users have only a small number of samples. The skewed sample size distribution is common in other FL data sets, such as LEAF (Caldas et al., 2018). The sample size dis-

9. Yann LeCun and Corinna Cortes hold the copyright of MNIST data set, which is a derivative work from original NIST data sets. MNIST data set is made available under the terms of the Creative Commons Attribution-Share Alike 3.0 license.
10. KMNIST data set is licensed under a permissive CC BY-SA 4.0 license, except where specified within some benchmark scripts.
11. FMNIST data set is under The MIT License (MIT) Copyright © [2017] Zalando SE, https://tech.zalando.com

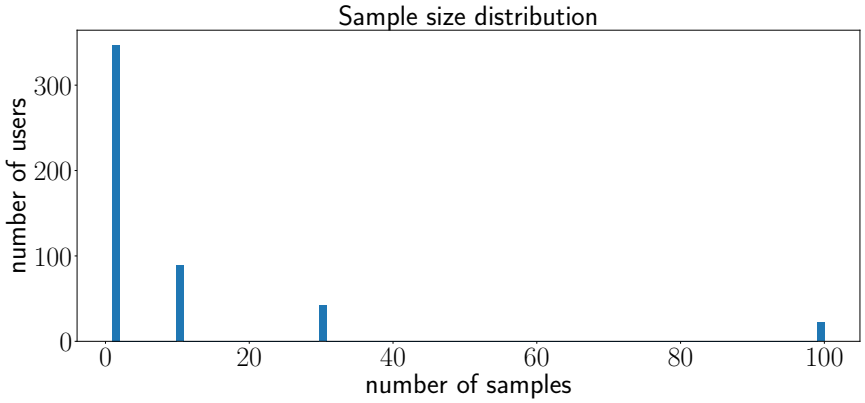

Figure 5: The sample size distribution in the training set across clients.

tribution in the training set is shown in Figure 5. In addition, each client has 10 validation samples used to measure the prediction accuracy of the model over the training process.

We use a multi-class logistic regression model. For a given gray scale picture with the label $y \in \{1, 2, \ldots, C\}$, we unroll its pixel matrix into a vector $x \in \mathbb{R}^p$. Given a parameter matrix $W \in \mathbb{R}^{C \times p}$, the training loss function defined in (1) is

$$\phi(W; x, y) \coloneqq l_{\mathrm{CE}}\left(\varsigma(Wx)\,; y\right),$$

where $\varsigma(\cdot) : \mathbb{R}^C \to \mathbb{R}^C$ is the softmax function defined as

$$[\varsigma(x)]_i = \frac{\exp(x_i)}{\sum_{j=1}^K \exp(x_j)}, \qquad \text{for all } x \in \mathbb{R}^C,$$

and $l_{\mathrm{CE}}(x\,; y) = \sum_{i=1}^C \mathbb{1}(y = i) \log x_i$, $x \in \mathbb{R}^C$, $y \in \{1, \ldots, C\}$, is the cross-entropy function.

We use the same algorithms and tuning parameters as in Section 7. Learning rate in SGD is set to 0.075 for MNIST and KMNIST, and is set to 0.03 for FMNIST. The total number of communication rounds is to $1,000$. In each round of communication, we choose $K = 10$ clients to participate (2% of total number of clients). For a chosen client m, we compute its local mini-batch gradient with the batch size equal to $\min\{5, n_m\}$, where n_m is the training sample size on the client m.

Figure 6 shows both the training loss and validation accuracy. Each figure shows the average performance over 5 independent runs. We use the same random seed for both Adaptive-OSMD Sampler and competitors, and change random seeds across different runs. The main focus is on minimizing the training loss, and the validation accuracy is only included for completeness. We observe that Adaptive-OSMD Sampler performs better than uniform sampling and other online learning samplers across all data sets.

9. Conclusion

We studied the client sampling problem in FL. We proposed an online learning with bandit feedback approach to tackle client sampling. We used online stochastic mirror descent to solve the online learning problem and applied the online ensemble method to choose

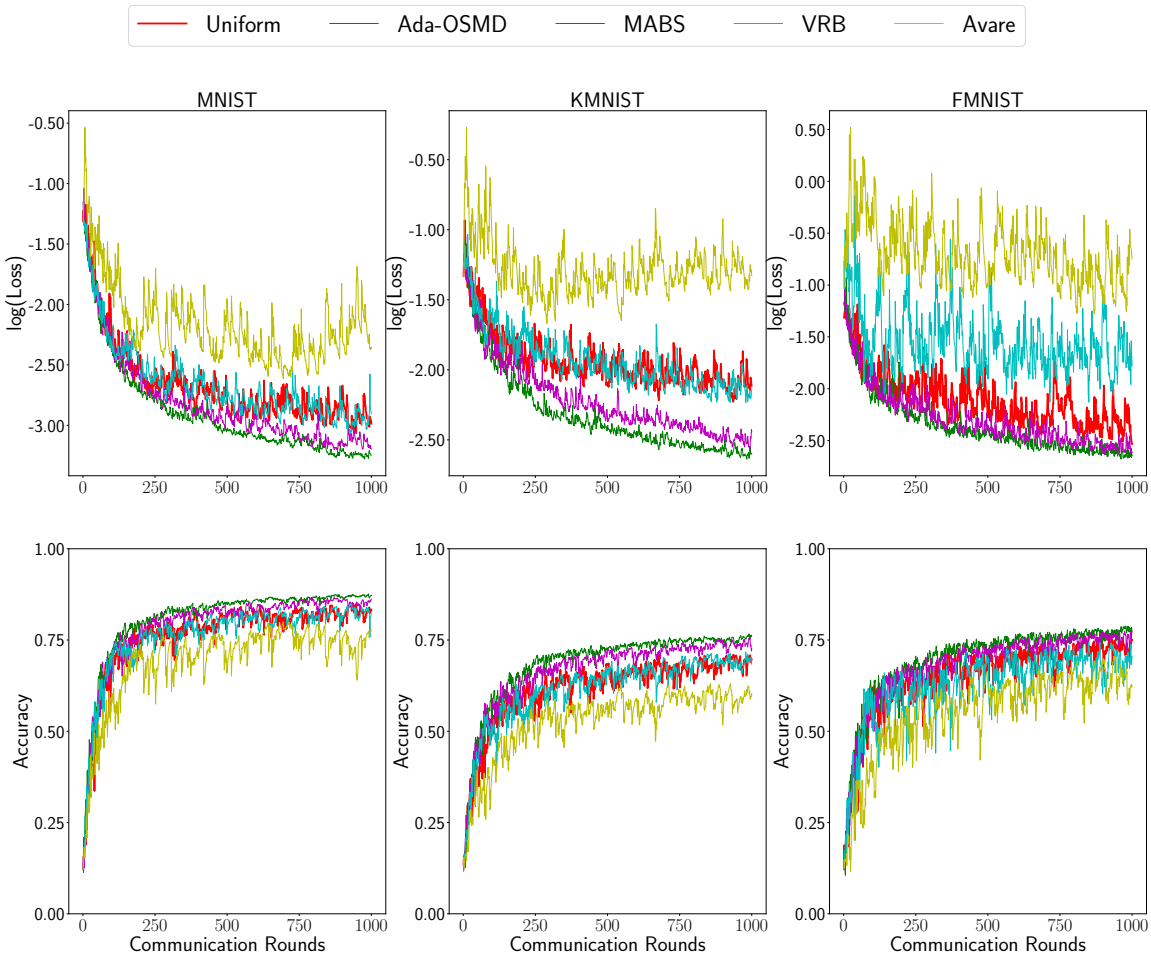

Figure 6: Comparison of Adaptive-OSMD Sampler, Uniform Sampler, and Other Online Learning Samplers The comparison is based on real data, evaluating training loss (top row) and validation accuracy (bottom row). Each column represents a different data set. The Adaptive-OSMD Sampler demonstrates superior performance, being both faster and more stable. The results represent the average performance over five independent runs.

the tuning parameters. We established an upper bound on the dynamic regret relative to any sequence of sampling distributions. Besides, we provide optimization guarantees for our sampling method when used with mini-batch SGD and FedAvg. Extensive numerical experiments demonstrated the benefits of our approach over both widely used uniform sampling and other competitors.

In this paper, we have focused on sampling with replacement. However, sampling without replacement would ideally be a more efficient approach. In Section A.2, we discussed a natural extension of Adaptive-OSMD Sampler to a setting where sampling without replacement is used. However, this approach does not directly minimize the variance of the

gradient g^t. When sampling without replacement is used, the variance function becomes more complicated and the design of an algorithm to directly minimize the variance is an interesting future direction.

Besides, in federated learning, privacy is a major concern. In this paper, the non-uniform sampling distribution may make the protection of clients' privacy more challenging than uniform sampling. One possible solution is to add noise to the gradient feedback and protect the clients' privacy under the Differential Privacy (DP) concept (Dwork, 2008). However, the added noise may hurt the performance of our sampling design and increase the regret. Studying the trade-off between privacy protection and regret is an important direction for addressing societal concerns in real-world applications.

Other fruitful future directions include the design of sampling algorithms for minimizing personalized FL objectives and sampling with physical constraints in the FL system, which we discuss in Appendix D.

Acknowledgments

This work was completed in part with resources provided by the University of Chicago Booth Mercury Computing Cluster. The research of MK is supported in part by NSF Grant ECCS-2216912.

Appendix A. Extensions of Adaptive-OSMD Sampler

In this section, we discuss additional extensions of Adaptive-OSMD Sampler. In Section A.1, we discuss how to choose η without knowing T in advance. In Section A.2, we discuss how to extend Adaptive-OSMD Sampler to sample without replacement setting.

A.1 Adaptive-OSMD Sampler with Doubling Trick

Algorithm 5 requires the total number of iterations T as input, which is not always available in practice. In those cases, we use doubling trick (Cesa-Bianchi and Lugosi, 2006, Section 2.3) to avoid this requirement. The basic idea is to restart Adaptive-OSMD Sampler at exponentially increasing time points $T_b = 2^{b-1}$, $b \geq 1$. The learning rates of experts in Algorithm 5 are reset at the beginning of each time interval, and the meta-algorithm learning rate γ is chosen optimally for the interval length.

More specifically, let $A_{\max} > 0$ such that $a^t_{\max} \leq A_{\max}$ for all t, where $a^t_{\max} = \max_{1\leq m\leq M} a^t_m = \max_{1\leq m\leq M} \lambda^2_m \|g^t_m\|^2_2$. At the time point T_b, we let

$$\mathcal{E}_b \coloneqq \left\{ 2^{e-\frac{b}{2}-\frac{1}{2}} \cdot \frac{K\alpha^3\sqrt{2\log M}}{M^3 A_{\max}} \;\middle|\; e = 1, 2, \ldots, E_b \right\}, \tag{27}$$

where

$$E_b = \left\lceil \frac{1}{2}\log_2\left(1 + \frac{4\log(M/\alpha)}{\log M}(2^{b-1}-1)\right)\right\rceil + 1, \tag{28}$$

and $\gamma_b = \frac{\alpha}{M}\sqrt{\frac{8K}{2^{b-1}A_{\max}}}$, $b \geq 1$.

In a practical implementation, at the time point $t = T_b$, instead of initializing all expert algorithms using uniform distribution, we can initialize them with the output of the meta-algorithm for $t = T_b - 1$. Besides, since the gradient norm is usually decreasing, we can estimate $A_{\max}$ by adding a pre-training phase where we broadcast the initial model parameter w^0 to all devices before the start of the training, and collect the returned $\|g^0_m\|^2_2$ from all responsive devices, which we denote as S^0. Then we can estimate $A_{\max}$ by $\hat{A}_{\max} = \max_{m\in S^0} \lambda_m \|g^0_m\|^2_2$.

Adaptive-Doubling-OSMD Sampler is detailed in Algorithm 6. From the computational perspective, by the proof of Theorem 10, Algorithm 6 needs to run Step 9 of Algorithm 5 for a total number of $O(T|\mathcal{E}|^2) = O(T\lfloor\log_2 T\rfloor)$ times. Therefore, the computational complexity of Adaptive-Doubling-OSMD Sampler is asymptotically the same as that of Adaptive-OSMD Sampler, while it increases by only a $\log(T)$ factor compared to OSMD Sampler. The following theorem provides a bound on the dynamic regret for Adaptive-Doubling-OSMD Sampler.

Theorem 10 *Suppose the training is stopped after T iterations. Let $\hat{p}^{1:T}$ be the output of Algorithm 6, where p^{unif} is used in Step 6. Then for any comparator sequence $q^{1:T}$, where q^t is allowed to be random, such that $q^t \in \mathcal{A}$ for all $t \in [T]$ and $\mathbb{E}[TV(q^{1:T})] \leq \beta$, we have*

$$\textit{D-Regret}_T(q^{1:T}) \leq \frac{6M^3 A_{\max}}{(\sqrt{2}-1)K\alpha^3}\sqrt{T\left[\frac{1}{2}\log M + \beta\log\left(M/\alpha\right)\right]} + \frac{2M}{(\sqrt{2}-1)\alpha}\sqrt{\frac{TA_{\max}}{8K}}(1+2\log E).$$

Algorithm 6 Adaptive-OSMD Sampler with Doubling Trick (Adaptive-Doubling-OSMD)

Input: Paramter α and $A_{\max}$.
Output: $\hat{p}^t$ for $t = 1, \ldots, T$.
while True **do**
  Set $\mathcal{E}_b$ as in (27).
  Let $\gamma_b = \frac{\alpha}{M}\sqrt{\frac{8K}{2^{b-1}A_{\max}}}$.
  Obtain $\{\hat{p}^t\}_{t=2^{b-1}}^{2^b-1}$ from Algorithm 5 with parameters: γ_b, $\mathcal{E}_b$, α, the number of iterations 2^{b-1}, and the initial distribution p^{unif} or $\hat{p}^{2^{b-1}-1}$ (when $b > 1$).
  if Training Process is Converged **then**
    Break.
  end if
  Let $b \leftarrow b + 1$.
end while

Proof See Appendix B.8. ■

Compare Theorem 10 with Theorem 8, we see that the regret bound of Adaptive-Doubling-OSMD has the same order as that of Adaptive-OSMD Sampler. However, Adaptive-Doubling-OSMD Sampler does not need to know T in advance. Mimicking the proof of Theorem 9, we can also show the optimization guarantees on Adaptive-Doubling-OSMD with mini-batch SGD and FedAvg, which is basically the same as Theorem 9, thus is omitted here.

A.2 Adaptive Sampling Without Replacement

In the discussion so far, we have assumed that the set S^t is obtained by sampling with replacement from p^t. When K is relatively large compared to M and p^t is far from uniform distribution, sampling without replacement can be more efficient than sampling with replacement. However, when sampling without replacement using p^t, the variance reduction loss does not have a clean form as in (4). As a result, an online design of the sampling distribution is more challenging. In this section, we discuss how to use the sampling distribution obtained by Adaptive-OSMD Sampler to sample clients without replacement, following the approach taken in Hanchi and Stephens (2020).

The detailed sampling procedure is described in Algorithm 7. We still use Adaptive-OSMD Sampler to update the sampling distribution. However, we use the designed sampling distribution in a way that no client is chosen twice. Furthermore, Step 18 of Algorithm 7 constructs the gradient estimate with the following properties.

Proposition 11 (Proposition 3 of Hanchi and Stephens (2020)) *Let $\hat{p}^t = p$ and let $\tilde{g}^t$ be as in Step 18 of Algorithm 7. Note that $\tilde{g}^t = \tilde{g}^t(p)$ depends on p. Recall that $J^t = \sum_{m=1}^M \lambda_m g_m^t$. We have*

$$\mathbb{E}_{S^t}\left[\tilde{g}^t\right] = J^t \qquad \text{and} \qquad \arg\min_{p \in \mathcal{P}_{M-1}} \mathbb{E}_{S^t}\left[\|\tilde{g}^t - J^t\|_2^2\right] = \arg\min_{p \in \mathcal{P}_{M-1}} l_t(p),$$

Algorithm 7 Adaptive sampling without replacement

Input: w^1 and $\hat{p}^1$.
for $t = 1, 2, \ldots, T-1$ **do**
  Let $\hat{p}^t_{(1)} = \hat{p}^t$ and sample m^t_1 from $[M]$ by $\hat{p}^t_{(1)}$.
  for $k = 2, \cdots, K$ **do**
    /* Design the sampling distribution for sampling the k-th client in the t-th round */
    Construct $\hat{p}^t_{(k)}$ by letting

$$\hat{p}^t_{(k),m} = \begin{cases} \left(1 - \sum_{l=1}^{k-1} \hat{p}^t_{m^t_l}\right)^{-1} \hat{p}^t_m & \text{if } m \in [M] \backslash \{m^t_1, \ldots, m^t_{k-1}\} \\ 0 & \text{otherwise.} \end{cases}$$

    /* Sample the k-th client */
    Sample m^t_k from $[M]\backslash\{m^t_1, \ldots, m^t_{k-1}\}$ by $\hat{p}^t_{(k)}$.
  end for
  Let $S^t = \{m^t_1, \cdots, m^t_K\}$.
  The server broadcasts the model parameter w^t to clients in S^t.
  The clients in S^t compute and upload the set of local gradients $\left\{g^t_{m^t_1}, \cdots, g^t_{m^t_K}\right\}$.
  /* Construct global gradient estimate */
  Let $g^t_{(1)} = \lambda^t_{m^t_1} g^t_{m^t_1} / \hat{p}^t_{(1),m^t_1}$.
  for $k = 2, \cdots, K$ **do**
    Let $g^t_{(k)} = \lambda^t_{m^t_k} g^t_{m^t_k} / \hat{p}^t_{(k),m^t_k} + \sum_{l=1}^{k-1} \lambda^t_{m^t_l} g^t_{m^t_l}$.
  end for
  Let $\tilde{g}^t = K^{-1} \sum_{k=1}^{K} g^t_{(k)}$.
  /* Update the model weight based on the global gradient estimate */
  Obtain the updated model parameter w^{t+1} using w^t and $\tilde{g}^t$.
  /* Update sampling distribution */
  Let $a^t_m = \lambda^2_m \|g^t_m\|^2$ for $m \in S^t$.
  Input $\{a^t_m\}_{m \in S^t}$ into Adaptive-OSMD Sampler to get $\hat{p}^{t+1}$.
end for

where $l_t(\cdot)$ is defined in (4) *and the expectation is taken over S^t.*

From Proposition 11, we see that $\tilde{g}^t$ is an unbiased stochastic gradient. Furthermore, the variance of $\tilde{g}^t$ is minimized by the same sampling distribution that minimizes the variance reduction loss in (4). Therefore, it is reasonable to use the sampling distribution generated by Adaptive-OSMD Sampler to design $\tilde{g}^t$.

Following the same simulation setup as in Section 7, we empirically compare sampling with replacement and sampling without replacement when used together with Adaptive-OSMD sampler. Training loss and cumulative regret are shown in Figure 7. We observe that using sampling with replacement results in a slightly smaller cumulative regret and a slightly better training loss. However, these differences are not significant.

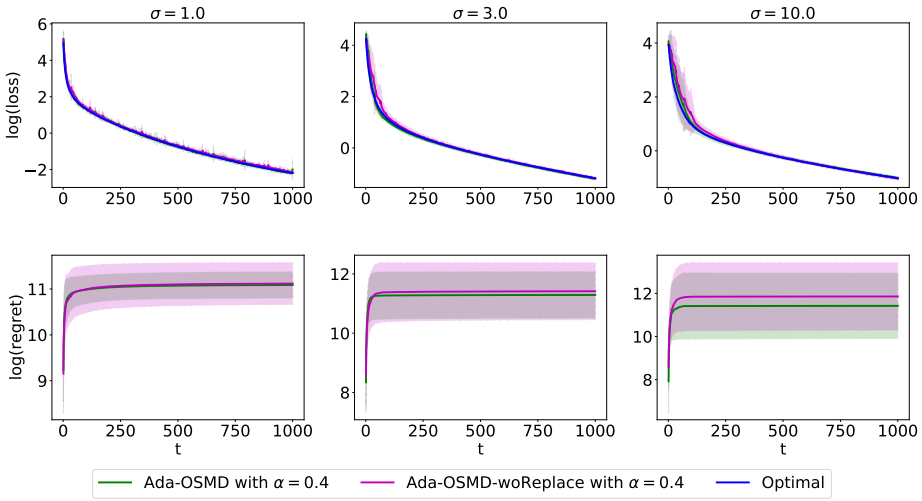

Figure 7: The training loss (top row) and cumulative regret (bottom row) are compared for the Adaptive-OSMD Sampler with replacement and without replacement, across $\sigma = 1.0$, $\sigma = 3.0$, and $\sigma = 10.0$. Solid lines represent the mean values, while shaded regions indicate mean $\pm$ standard deviation across independent runs.

Appendix B. Technical proofs

B.1 Proof of (11)

Note that

$$\begin{aligned}
\zeta_T^2(\alpha,\beta) &= \frac{1}{T}\sup_{w^1}\min_{p^1\in\mathcal{A}}\cdots\sup_{w^T}\min_{p^T\in\mathcal{A}}\sum_{t=1}^{T}V(p^t,w^t) \quad \text{subject to TV}\left(p^{1:T}\right)\le\beta\\
&\le \frac{1}{T}\sup_{w^1}\min_{p^1\in\mathcal{A}}\cdots\sup_{w^T}\min_{p^T\in\mathcal{A}}\sum_{t=1}^{T}V(p^t,w^t) \quad \text{subject to TV}\left(p^{1:T}\right)=0\\
&= \frac{1}{T}\sup_{w^1}\min_{p\in\mathcal{A}}\sup_{w^2}\cdots\sup_{w^T}\sum_{t=1}^{T}V(p,w^t)\\
&\le \frac{1}{T}\min_{p\in\mathcal{A}}\sup_{w^1}\cdots\sup_{w^T}\sum_{t=1}^{T}V(p,w^t)\\
&\le \frac{1}{T}\sup_{w^1}\cdots\sup_{w^T}\sum_{t=1}^{T}V(p_f,w^t)=\sup_{w}V(p_f,w)=\zeta_{\text{fix}}^2(\alpha)\le\zeta_{\text{unif}}^2.
\end{aligned}$$

B.2 Proposition 12 and Its Proof

Proposition 12 *Let*

$$\tilde{p}_m^{t+1}=p_m^t\exp\left\{\mathcal{N}\left\{m\in S^t\right\}\eta_t a_m^t/(K^2(p_m^t)^3)\right\},\qquad m\in[M].$$

Let $\pi : [M] \mapsto [M]$ be a permutation such that $\tilde{p}^{t+1}_{\pi(1)} \leq \tilde{p}^{t+1}_{\pi(2)} \leq \cdots \leq \tilde{p}^{t+1}_{\pi(M)}$. Let $m^t_\star$ be the smallest integer m such that

$$\tilde{p}^{t+1}_{\pi(m)} \left(1 - \frac{m-1}{M}\alpha\right) > \frac{\alpha}{M} \sum_{j=m}^{M} \tilde{p}^{t+1}_{\pi(j)}.$$

Then

$$\hat{p}^{t+1}_m = \begin{cases} \alpha/M & \text{if } \pi(m) < m^t_\star \\ \left((1 - ((m^t_\star - 1)/M)\alpha)\tilde{p}^{t+1}_m\right) / \left(\sum_{j=m^t_\star}^{M} \tilde{p}^{t+1}_{\pi(j)}\right) & \text{otherwise.} \end{cases}$$

Proof First, we show that the solution $\hat{p}^{t+1}$ in Step 7 of Algorithm 1 can be found as

$$\begin{aligned} \tilde{p}^{t+1} &= \arg\min_{p \in \mathcal{D}} \eta_t \langle p, \nabla \hat{l}_t(\hat{p}^t; \hat{p}^t)\rangle + D_\Phi\left(p \,\|\, \hat{p}^t\right), \\ \hat{p}^{t+1} &= \arg\min_{p \in \mathcal{A}} D_\Phi\left(p \,\|\, \tilde{p}^{t+1}\right). \end{aligned}$$

The optimality condition for $\tilde{p}^{t+1}$ implies that

$$\eta_t \nabla \hat{l}_t(\hat{p}^t; \hat{p}^t) + \nabla\Phi(\tilde{p}^{t+1}) - \nabla\Phi(\hat{p}^t) = 0. \tag{29}$$

By Lemma 14, the optimality condition for $\hat{p}^{t+1}$ implies that

$$\langle p - \hat{p}^{t+1}, \nabla\Phi(\hat{p}^{t+1}) - \nabla\Phi(\tilde{p}^{t+1})\rangle \geq 0, \quad \text{for all } p \in \mathcal{A}.$$

Combining the last two displays, we have

$$\langle p - \hat{p}^{t+1}, \eta_t \nabla \hat{l}_t(\hat{p}^t; \hat{p}^t) + \nabla\Phi(\hat{p}^{t+1}) - \nabla\Phi(\hat{p}^t)\rangle \geq 0, \quad \text{for all } p \in \mathcal{A}.$$

By Lemma 14, this is the optimality condition for $\hat{p}^{t+1}$ to be the solution in Step 7 of Algorithm 1.

Note that (29) implies that

$$-\frac{\eta_t}{K^2} \cdot \frac{a^t_m}{(p^t_m)^3} \mathcal{N}\left\{m \in S^t\right\} + \log(\tilde{p}^{t+1}_m) - \log(\hat{p}^t_m) = 0, \qquad m \in [M].$$

Therefore,

$$\tilde{p}^{t+1}_m = \hat{p}^t_m \exp\left(\frac{\eta_t a^t_m}{K^2 \left(\hat{p}^t_m\right)^3} \mathcal{N}\left\{m \in S^t\right\}\right), \qquad m \in [M],$$

and the final result follows from Lemma 19.

■

B.3 Proof of Theorem 4

Our proof follows the similar technique used in the proof of Theorem 2.1 of Ghadimi and Lan (2013) except for the novel technique of construction of a ghost subset that is drawn from $[M]$ from the comparator sampling distribution. Note that the ghost subset is only constructed for theoretical purpose and does not need to be computed in practice.

Given any $\{w^t\}_{t=1}^T \triangleq w^{1:T}$, let $q_s^{1:T}$ be the solution of the problem

$$\begin{aligned}\tilde{\zeta}(w^{1:T}) = \frac{1}{T}\min_{p^1}\ldots\min_{p^T}\sum_{t=1}^T V(p^t, w^t),\\ \text{subject to } \mathrm{TV}\left(p^{1:T}\right) \leq \beta \text{ and } p^t \in \sigma(w^1,\ldots,w^t),\end{aligned} \tag{30}$$

where $\sigma(w^1,\ldots,w^t)$ is the σ-algebra generated by $\{w^1,\ldots,w^t\}$. We then have

$$\tilde{\zeta}(w^{1:T}) = \frac{1}{T}\sum_{t=1}^T V(q_s^t(w^t), w^t) \text{ and } \mathrm{TV}(q_s^{1:T}) \leq \beta.$$

Note that by the definition of $\zeta_T^2(\alpha,\beta)$, we have

$$\frac{1}{T}\sup_{w^1}\cdots\sup_{w^T}\sum_{t=1}^T V(q_s^t, w^t) = \zeta_T^2(\alpha,\beta). \tag{31}$$

Let $\delta^t = g^t - \nabla F(w^t)$. Under Assumption 1, by Lemma 15, we have

$$\begin{aligned}F\left(w^{t+1}\right) &\leq F\left(w^t\right) + \left\langle \nabla F\left(w^t\right), w^{t+1} - w^t\right\rangle + \frac{L}{2}\mu^2\left\|g^t\right\|_2^2\\ &= F\left(w^t\right) - \mu\left\langle \nabla F\left(w^t\right), g^t\right\rangle + \frac{L}{2}\mu^2\left\|g^t\right\|_2^2\\ &= F\left(w^t\right) - \mu\left\|\nabla F\left(w^t\right)\right\|_2^2 - \mu\left\langle \nabla F\left(w^t\right), \delta^t\right\rangle\\ &\quad + \frac{L}{2}\mu^2\left[\left\|\nabla F\left(w^t\right)\right\|_2^2 + 2\left\langle \nabla F\left(w^t\right), \delta^t\right\rangle + \left\|\delta^t\right\|_2^2\right]\\ &= F\left(w^t\right) - \left(\mu - \frac{L}{2}\mu^2\right)\left\|\nabla F\left(w^t\right)\right\|_2^2 - \left(\mu - L\mu^2\right)\left\langle \nabla F\left(w^t\right), \delta^t\right\rangle + \frac{L}{2}\mu^2\left\|\delta^t\right\|_2^2.\end{aligned} \tag{32}$$

We use the notation $\mathbb{E}_{S^t}[\cdot]$ to denote the expectation taken with respect to S^t. Note that $\mathbb{E}_{S^t}\left[g^t \mid w^t, \hat{p}^t\right] = \nabla F(w^t)$, thus we have $\mathbb{E}\left[\delta^t \mid w^t, \hat{p}^t\right] = 0$, and

$$\mathbb{E}\left[\left\langle \nabla F\left(w^t\right), \delta^t\right\rangle\right] = \mathbb{E}\left[\mathbb{E}\left[\left\langle \nabla F\left(w^t\right), \delta^t\right\rangle \mid w^t, \hat{p}^t\right]\right] = 0. \tag{33}$$

On the other hand, conditioned on $w^1,\ldots,w^t$, q_s^t is a deterministic sampling distribution. We can then assume that there is a ghost subset of clients $\tilde{S}^t$ with $|\tilde{S}^t| = K$, which is drawn from $[M]$ with sampling distribution q_s^t. Recall that $J^t = (1/M)\sum_{m=1}^M g_m^t$. Besides, we let

$$\tilde{g}^t \coloneqq \frac{1}{MK}\sum_{m\in\tilde{S}^t}\frac{g_m^t}{q_{s,m}^t}.$$

Then we have

$$\begin{aligned}
\mathbb{E}_{S^t}\left[\left\|\delta^t\right\|_2^2 \mid w^1,\ldots,w^t,\hat{p}^t\right] &= \mathbb{E}_{S^t}\left[\left\|g^t - J^t + J^t - \nabla F(w^t)\right\|_2^2 \mid w^1,\ldots,w^t,\hat{p}^t\right] \\
&= \mathbb{E}_{S^t}\left[\left\|g^t - J^t\right\|_2^2 \mid w^1,\ldots,w^t,\hat{p}^t\right] + \left\|J^t - \nabla F(w^t)\right\|_2^2 \\
&= l_t\left(\hat{p}^t\right) - \frac{1}{K}\left\|J^t\right\|_2^2 + \left\|J^t - \nabla F(w^t)\right\|_2^2 \\
&= l_t\left(q_s^t\right) - \frac{1}{K}\left\|J^t\right\|_2^2 + \left\|J^t - \nabla F(w^t)\right\|_2^2 + l_t\left(\hat{p}^t\right) - l_t\left(q_s^t\right) \\
&= \mathbb{E}_{\tilde{S}^t}\left[\left\|\tilde{g}^t - \nabla F(w^t)\right\|_2^2 \mid w^1,\ldots,w^t,\hat{p}^t\right] + l_t\left(\hat{p}^t\right) - l_t\left(q_s^t\right).
\end{aligned}$$

Since

$$\begin{aligned}
&\mathbb{E}\left[\left\|\tilde{g}^t - \nabla F(w^t)\right\|_2^2 \mid w^1,\ldots,w^t,\hat{p}^t\right] \\
&= \mathbb{E}\left[\left\|\frac{1}{MK}\sum_{m\in\tilde{S}^t}\frac{g_m^t}{q_{s,m}^t} - \nabla F(w^t)\right\|_2^2 \mid w^1,\ldots,w^t,\hat{p}^t\right] \\
&\le 2\mathbb{E}\left[\left\|\frac{1}{MK}\sum_{m\in\tilde{S}^t}\frac{g_m^t}{q_{s,m}^t} - \frac{1}{MK}\sum_{m\in\tilde{S}^t}\frac{\nabla F_m(w^t)}{q_{s,m}^t}\right\|_2^2 \mid w^1,\ldots,w^t,\hat{p}^t\right] \\
&\quad + 2\mathbb{E}\left[\left\|\frac{1}{MK}\sum_{m\in\tilde{S}^t}\frac{\nabla F_m(w^t)}{q_{s,m}^t} - \nabla F(w^t)\right\|_2^2 \mid w^1,\ldots,w^t,\hat{p}^t\right] \\
&= \frac{2}{M^2K^2}\mathbb{E}\left[\sum_{m\in\tilde{S}^t}\mathbb{E}\left[\frac{\left\|g_m^t - \nabla F_m(w^t)\right\|_2^2}{(q_{s,m}^t)^2}\right] \mid w^1,\ldots,w^t,\hat{p}^t\right] \\
&\quad + \frac{2}{K}\left(\frac{1}{M^2}\sum_{m=1}^{M}\frac{\left\|\nabla F_m(w^t)\right\|_2^2}{q_{s,m}^t} - \left\|\nabla F(w^t)\right\|_2^2\right) \\
&= \frac{2\sigma^2}{M^2KB}\sum_{m=1}^{M}\frac{1}{q_{s,m}^t} + \frac{2}{K}V(q_s^t,w^t) \le \frac{2\sigma^2}{KB\alpha} + \frac{2}{K}V(q_s^t,w^t),
\end{aligned}$$

where the penultimate line follows that $\mathbb{E}\left[\left\|g_m^t - \nabla F_m(w^t)\right\|_2^2\right] \le \sigma^2/B$ and the definition of $V(p,w)$, and the last line follows that $q_{s,m}^t \ge \alpha/M$ since $q_s^t \in \mathcal{A}$.

Thus, we have

$$\mathbb{E}\left[\left\|\delta^t\right\|_2^2 \mid w^1,\ldots,w^t,\hat{p}^t\right] \le \frac{2\sigma^2}{KB\alpha} + \frac{2}{K}V(q_s^t,w^t) + l_t\left(\hat{p}^t\right) - l_t\left(q_s^t\right),$$

which implies that

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\delta^t\right\|_2^2\right] &\le \frac{2T\sigma^2}{KB\alpha}+\frac{2}{K}\mathbb{E}\left[\sum_{t=1}^{T}V(q_s^t,w^t)\right]+\mathbb{E}\left[\sum_{t=0}^{T-1}l_t\left(\hat{p}^t\right)-\sum_{t=0}^{T-1}l_t\left(q_s^t\right)\right]\\ &= \frac{2T\sigma^2}{KB\alpha}+\frac{2}{K}\mathbb{E}\left[\sum_{t=1}^{T}V(q_s^t,w^t)\right]+\text{D-Regret}_T(q_s^{1:T}).\\ &\le \frac{2T\sigma^2}{KB\alpha}+\frac{2T\zeta_T^2(\alpha,\beta)}{K}+\text{D-Regret}_T(q_s^{1:T}),\end{aligned}\tag{34}$$

where the last inequality follows the fact that

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}V(q_s^t,w^t)\right]=\frac{1}{T}\,\mathbb{E}_{w^1,\ldots,w^T}\left[\sum_{t=1}^{T}V(q_s^t,w^t)\right]\le\frac{1}{T}\sup_{w^1}\cdots\sup_{w^T}\sum_{t=1}^{T}V(q_s^t,w^t)=\zeta_T^2(\alpha,\beta),$$

and the last equality is by (31). Combine (32), (33) and (34), we have

$$\begin{aligned}&\left(\mu-\frac{L}{2}\mu^2\right)\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla F\left(w^t\right)\right\|_2^2\right]\\ &\le F(w^1)-F(w^T)+\frac{L}{2}\mu^2\left(\frac{2T\sigma^2}{KB\alpha}+\frac{2T\zeta_T^2(\alpha,\beta)}{K}+\text{D-Regret}_T(q_s^{1:T})\right)\\ &\le D^F+\frac{L}{2}\mu^2\left(\frac{2T\sigma^2}{KB\alpha}+\frac{2T\zeta_T^2(\alpha,\beta)}{K}+\text{D-Regret}_T(q_s^{1:T})\right).\end{aligned}$$

Since $\mu\le 1/L$, thus $(\mu-\frac{L}{2}\mu^2)=\mu(1-\frac{L}{2}\mu)\ge\mu/2$, thus

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla F\left(w^t\right)\right\|_2^2\right]\le\frac{2D^F}{T\mu}+L\mu\left(\frac{2\sigma^2}{KB\alpha}+\frac{2\zeta_T^2(\alpha,\beta)}{K}+\frac{\text{D-Regret}_T(q_s^{1:T})}{T}\right).\tag{35}$$

Next we bound $\text{D-Regret}_T(q_s^{1:T})$. Note that

$$a_{\max}^t=\frac{1}{M^2}\max_{1\le m\le M}\|g_m^t\|_2^2=\frac{1}{M^2}\max_{1\le m\le M}\left\|\frac{1}{B}\sum_{b=1}^{B}\nabla\phi(w^t;\xi_m^{t,b})\right\|_2^2\le\frac{G^2}{M^2}$$

since $\|\nabla\phi(w;\xi)\|_2\le G$ for all w and ξ, then by Theorem 6 and the fact that $q_s^t\in\mathcal{A}$ for all $t\in[T]$ and $\text{TV}(q_s^{1:T})\le\beta$, we have

$$\begin{aligned}\text{D-Regret}_T(q_s^{1:T}) &\le \frac{\log M}{\eta}+\frac{2\log(M/\alpha)}{\eta}\mathbb{E}\left[\text{TV}\left(q^{1:T}\right)\right]+\frac{\eta M^6}{2K^2\alpha^6}\sum_{t=1}^{T}\mathbb{E}\left[\left(a_{\max}^t\right)^2\right]\\ &\le \frac{\log M}{\eta}+\frac{2\beta\log(M/\alpha)}{\eta}+\frac{\eta TM^2G^4}{2K^2\alpha^6}.\end{aligned}$$

Let

$$\eta=\frac{K\alpha^3}{MG^2}\sqrt{\frac{2\log M+4\beta\log(M/\alpha)}{T}},$$

we then have

$$\text{D-Regret}_T(q_s^{1:T}) \leq \frac{\sqrt{T}MG^2}{K\alpha^3}\sqrt{\frac{1}{2}\log M + \beta\log(M/\alpha)}.$$

Plug the above inequality into (35), we have

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla F\left(w^t\right)\right\|_2^2\right] \\
&\leq \frac{2D^F}{T\mu} + L\mu\left(\frac{2\sigma^2}{KB\alpha} + \frac{2\zeta_T^2(\alpha,\beta)}{K} + \frac{MG^2}{K\alpha^3}\sqrt{\frac{\frac{1}{2}\log M + \beta\log(M/\alpha)}{T}}\right).
\end{aligned}$$

Finally, let

$$\mu = \min\left\{\frac{1}{L}, \frac{1}{\sigma}\sqrt{\frac{D^F KB\alpha}{LT}}, \frac{1}{\zeta_T(\alpha,\beta)}\sqrt{\frac{D^F K}{LT}}, \sqrt{\frac{D^F K\alpha^3}{L\sqrt{T}MG^2\sqrt{\frac{1}{2}\log M + \beta\log(M/\alpha)}}}\right\}, \tag{36}$$

then we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|\nabla F\left(w^R\right)\right\|_2^2\right] = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla F\left(w^t\right)\right\|_2^2\right] \\
&\lesssim \frac{D^F}{T}\max\left\{L, \sigma\sqrt{\frac{LT}{D^F KB\alpha}}, \zeta_T(\alpha,\beta)\sqrt{\frac{TL}{D^F K}}, \sqrt{\frac{L\sqrt{T}MG^2\sqrt{\frac{1}{2}\log M + \beta\log(M/\alpha)}}{D^F K\alpha^3}}\right\} \\
&\quad + \frac{\sigma\sqrt{D^F L}}{\sqrt{TKB\alpha}} + \frac{\zeta_T(\alpha,\beta)\sqrt{D^F L}}{\sqrt{TK}} + \frac{\sqrt{D^F L}M^{\frac{1}{2}}G}{T^{\frac{3}{4}}K^{\frac{1}{2}}\alpha^{\frac{3}{2}}}\left(\frac{1}{2}\log M + \beta\log(M/\alpha)\right)^{\frac{1}{4}} \\
&\lesssim \frac{D^F}{T}\left(L + \sigma\sqrt{\frac{LT}{D^F KB\alpha}} + \zeta_T(\alpha,\beta)\sqrt{\frac{TL}{D^F K}} + \sqrt{\frac{L\sqrt{T}MG^2\sqrt{\frac{1}{2}\log M + \beta\log(M/\alpha)}}{D^F K\alpha^3}}\right) \\
&\quad + \frac{\sigma\sqrt{D^F L}}{\sqrt{TKB\alpha}} + \frac{\zeta_T(\alpha,\beta)\sqrt{D^F L}}{\sqrt{TK}} + \frac{\sqrt{D^F L}M^{\frac{1}{2}}G}{T^{\frac{3}{4}}K^{\frac{1}{2}}\alpha^{\frac{3}{2}}}\left(\frac{1}{2}\log M + \beta\log(M/\alpha)\right)^{\frac{1}{4}} \\
&\lesssim \frac{D^F L}{T} + \frac{\sigma\sqrt{D^F L}}{\sqrt{TKB\alpha}} + \frac{\zeta_T(\alpha,\beta)\sqrt{D^F L}}{\sqrt{TK}} + \frac{\sqrt{D^F L}M^{\frac{1}{2}}G}{T^{\frac{3}{4}}K^{\frac{1}{2}}\alpha^{\frac{3}{2}}}\left(\frac{1}{2}\log M + \beta\log(M/\alpha)\right)^{\frac{1}{4}}.
\end{aligned}$$

B.4 Proof of Theorem 5

Similar to the proof in Appendix B.3, the key novel technique is the construction of a ghost subset that is drawn from $[M]$ by the comparator sampling distribution. The ghost subset is only constructed for theoretical purpose and does not need to be computed in practice. Similar to Appendix B.3, given any $\{w^t\}_{t=1}^T$, let $q_s^{1:T}$ be the solution of the problem (30).

By Assumption 1 and Lemma 15, we have

$$F\left(w^{t+1}\right) \leq F\left(w^t\right) + \left\langle \nabla F\left(w^t\right), w^{t+1} - w^t \right\rangle + \frac{L}{2}\left\|w^{t+1} - w^t\right\|_2^2,$$

where

$$w^{t+1} - w^t = -\frac{\mu}{MK}\sum_{m\in S^t}\frac{1}{\hat{p}_m^t} g_m^t, \quad g_m^t = \mu_l \sum_{b=0}^{B-1} \nabla\phi(w_m^{t,b}; \xi_m^{t,b}).$$

Let $\mathcal{B}_m^t = \{\xi_m^{t,0}, \ldots, \xi_m^{t,B-1}\}$ and $\mathbb{E}_{S^t}[\cdot]$ denote the expectation taken with respect to S^t. We then have

$$\begin{aligned}\mathbb{E}_{S^t}\left[F\left(w^{t+1}\right) \mid w^1, \ldots, w^t, \hat{p}^t\right] \leq F\left(w^t\right) - \mu\left\langle \nabla F\left(w^t\right), \frac{1}{M}\sum_{m=1}^M g_m^t \right\rangle \\ + \frac{L\mu^2}{2}\mathbb{E}_{S^t}\left[\left\|\frac{1}{K}\sum_{m\in S^t}\frac{1}{M\hat{p}_m^t} g_m^t\right\|_2^2 \mid w^1, \ldots, w^t, \hat{p}^t\right].\end{aligned}$$

Recall that $J^t = (1/M)\sum_{m=1}^M g_m^t$, then we have

$$\begin{aligned}\mathbb{E}_{S^t}\left[F\left(w^{t+1}\right)\right] &\leq F\left(w^t\right) - \mu\left\langle \nabla F\left(w^t\right), J^t\right\rangle \\ &\quad + \frac{L\mu^2}{2}\mathbb{E}_{S^t}\left[\left\|\frac{1}{K}\sum_{m\in S^t}\frac{1}{M\hat{p}_m^t} g_m^t - J^t + J^t\right\|_2^2 \mid w^1, \ldots, w^t, \hat{p}^t\right] \\ &= F\left(w^t\right) - \mu\left\langle \nabla F\left(w^t\right), J^t\right\rangle + \frac{L\mu^2}{2}\left\|J^t\right\|_2^2 \\ &\quad + \frac{L\mu^2}{2}\mathbb{E}_{S^t}\left[\left\|\frac{1}{K}\sum_{m\in S^t}\frac{1}{M\hat{p}_m^t} g_m^t - J^t\right\|_2^2 \mid w^1, \ldots, w^t, \hat{p}^t\right].\end{aligned}$$

Recall that in Algorithm 4, we let $a_m^t = \frac{\|g_m^t\|^2}{B(M\mu_l)^2}$, thus we have

$$\begin{aligned}\mathbb{E}_{S^t}\left[\left\|\frac{1}{K}\sum_{m\in S^t}\frac{1}{M\hat{p}_m^t} g_m^t - J^t\right\|_2^2 \mid w^1, \ldots, w^t, \hat{p}^t\right] &= \frac{1}{K}\sum_{m=1}^M \frac{\left\|g_m^t\right\|_2^2}{M^2\hat{p}_m^t} - \frac{1}{K}\left\|J^t\right\|_2^2 \\ &= B\mu_l^2\frac{1}{K}\sum_{m=1}^M \frac{\left\|g_m^t\right\|_2^2}{B(\mu_l M)^2\hat{p}_m^t} - \frac{1}{K}\left\|J^t\right\|_2^2 \\ &= B\mu_l^2\frac{1}{K}\sum_{m=1}^M \frac{a_m^t}{\hat{p}_m^t} - \frac{1}{K}\left\|J^t\right\|_2^2 \\ &= B\mu_l^2 l_t\left(\hat{p}^t\right) - \frac{1}{K}\left\|J^t\right\|_2^2, \end{aligned} \tag{37}$$

which then implies that

$$\begin{aligned}&\mathbb{E}_{S^t}\left[F\left(w^{t+1}\right) \mid w^1, \ldots, w^t, \hat{p}^t\right] \\ &\leq F\left(w^t\right) - \mu\left\langle \nabla F\left(w^t\right), J^t\right\rangle + \frac{L\mu^2}{2}\left\|J^t\right\|_2^2 + \frac{L\mu^2}{2}\left(B\mu_l^2 l_t(\hat{p}^t) - \frac{1}{K}\left\|J^t\right\|_2^2\right),\end{aligned} \tag{38}$$

where $l_t(\cdot)$ is the variance reduction loss defined in (4).

Conditioned on $w^1, \ldots, w^t$, q_s^t is a deterministic sampling distribution. We assume that there is a ghost subset of clients $\tilde{S}^t$ with $|\tilde{S}^t| = K$, which is drawn from $[M]$ with replacement by sampling distribution q_s^t. Besides, we let

$$\tilde{g}^t \coloneqq \frac{1}{K} \sum_{m \in \tilde{S}^t} \frac{g_m^t}{M q_{s.m}^t},$$

then similar to (37), we have

$$\mathbb{E}_{\tilde{S}^t} \left[\left\| \tilde{g}^t - J^t \right\|_2^2 \mid w^1, \ldots, w^t, \hat{p}^t \right] = B \mu_l^2 l_t(q_s^t) - \frac{1}{K} \left\| J^t \right\|_2^2.$$

Combine the above equation with (38), we have

$$\begin{aligned}
&\mathbb{E}_{S^t} \left[F\left(w^{t+1}\right) \mid w^1, \ldots, w^t, \hat{p}^t \right] \\
&\leq F\left(w^t\right) - \mu \left\langle \nabla F\left(w^t\right), J^t \right\rangle + \frac{L\mu^2}{2} \left\| J^t \right\|_2^2 + \frac{L\mu^2}{2} \left(B \mu_l^2 l_t(q_s^t) - \frac{1}{K} \left\| J^t \right\|_2^2 \right) \\
&\quad + \frac{BL\mu^2 \mu_l^2}{2} \left(l_t(\hat{p}^t) - l_t(q_s^t) \right) \\
&= F\left(w^t\right) - \mu \left\langle \nabla F\left(w^t\right), J^t \right\rangle + \frac{L\mu^2}{2} \left\| J^t \right\|_2^2 + \frac{BL\mu^2 \mu_l^2}{2} \left(l_t(\hat{p}^t) - l_t(q_s^t) \right) \\
&\quad + \frac{L\mu^2}{2} \mathbb{E}_{\tilde{S}^t} \left[\left\| \tilde{g}^t - J^t \right\|_2^2 \mid w^1, \ldots, w^t, \hat{p}^t \right] \\
&= F\left(w^t\right) - \mu \left\langle \nabla F\left(w^t\right), J^t \right\rangle + \frac{L\mu^2}{2} \mathbb{E}_{\tilde{S}^t} \left[\left\| \tilde{g}^t \right\|_2^2 \mid w^1, \ldots, w^t, \hat{p}^t \right] + \frac{BL\mu^2 \mu_l^2}{2} \left(l_t(\hat{p}^t) - l_t(q_s^t) \right).
\end{aligned}$$

Let $\mathbb{E}_t[\cdot]$ denote $\mathbb{E}[\cdot \mid w^1, \ldots, w^t, \hat{p}^t]$. We thus have

$$\begin{aligned}
\mathbb{E}_t \left[F\left(w^{t+1}\right) \right] \leq F\left(w^t\right) &- \frac{\mu \mu_l}{M} \mathbb{E}_t \left[\left\langle \nabla F\left(w^t\right), \sum_{m=1}^{M} \sum_{b=0}^{B-1} \nabla F_m(w_m^{t,b}) \right\rangle \right] \\
&+ \frac{L\mu^2}{2} \mathbb{E}_t \left[\left\| \tilde{g}^t \right\|_2^2 \right] + \frac{BL\mu^2 \mu_l^2}{2} \mathbb{E}_t \left[l_t(\hat{p}^t) - l_t(q_s^t) \right].
\end{aligned} \tag{39}$$

Note that for any $u, v \in \mathbb{R}^d$, we have $-\langle u, v\rangle = -\frac{1}{2}\|u\|^2 + \frac{1}{2}\|u-v\| - \frac{1}{2}\|v\|^2 \leq -\frac{1}{2}\|u\|^2 + \frac{1}{2}\|u-v\|$. Thus we have

$$
\begin{aligned}
&\mathbb{E}_t\left[-\frac{\mu\mu_l}{M}\sum_{m=1}^{M}\sum_{b=0}^{B-1}\left\langle \nabla F\left(w^t\right), \nabla F_m(w_m^{t,b})\right\rangle\right] \\
&= \mathbb{E}_t\left[-\mu\mu_l\sum_{b=0}^{B-1}\left\langle \nabla F\left(w^t\right), \frac{1}{M}\sum_{m=1}^{M}\nabla F_m(w_m^{t,b})\right\rangle\right] \\
&\leq -\frac{\mu\mu_l B}{2}\left\|\nabla F\left(w^t\right)\right\|_2^2 + \frac{\mu\mu_l}{2}\sum_{b=0}^{B-1}\left\|\nabla F\left(w^t\right) - \frac{1}{M}\sum_{m=1}^{M}\nabla F_m(w_m^{t,b})\right\|_2^2 \\
&= -\frac{\mu\mu_l B}{2}\left\|\nabla F\left(w^t\right)\right\|_2^2 + \frac{\mu\mu_l}{2}\sum_{b=0}^{B-1}\left\|\frac{1}{M}\sum_{m=1}^{M}\left(\nabla F_m\left(w^t\right) - \nabla F_m(w_m^{t,b})\right)\right\|_2^2 \\
&\leq -\frac{\mu\mu_l B}{2}\left\|\nabla F\left(w^t\right)\right\|_2^2 + \frac{\mu\mu_l}{2M}\mathbb{E}_t\left[\sum_{m=1}^{M}\sum_{b=0}^{B-1}\left\|\nabla F_m\left(w^t\right) - \nabla F_m(w_m^{t,b})\right\|_2^2\right] \\
&\leq -\frac{\mu\mu_l B}{2}\left\|\nabla F\left(w^t\right)\right\|_2^2 + \frac{\mu\mu_l L^2}{2M}\mathbb{E}_t\left[\sum_{m=1}^{M}\sum_{b=0}^{B-1}\left\|w^t - w_m^{t,b}\right\|_2^2\right].
\end{aligned} \tag{40}
$$

Besides, we also have

$$
\begin{aligned}
\mathbb{E}_t\left[\left\|\tilde{g}^t\right\|_2^2\right] &= \mathbb{E}_t\left[\left\|\frac{1}{K}\sum_{m\in\tilde{S}^t}\frac{g_m^t}{Mq_{s,m}^t}\right\|_2^2\right] = \mu_l^2\mathbb{E}_t\left[\left\|\frac{1}{K}\sum_{m\in\tilde{S}^t}\frac{1}{Mq_{s,m}^t}\sum_{b=0}^{B-1}\nabla\phi(w_m^{t,b};\xi_m^{t,b})\right\|_2^2\right] \\
&= \underbrace{\mu_l^2\,\mathbb{E}_t\left[\left\|\frac{1}{K}\sum_{m\in\tilde{S}^t}\frac{1}{Mq_{s,m}^t}\sum_{b=0}^{B-1}\left(\nabla\phi(w_m^{t,b};\xi_m^{t,b}) - \nabla F_m(w_m^{t,b})\right)\right\|_2^2\right]}_{I_1} \\
&\quad + \underbrace{\mu_l^2\,\mathbb{E}_t\left[\left\|\frac{1}{K}\sum_{m\in\tilde{S}^t}\frac{1}{Mq_{s,m}^t}\sum_{b=0}^{B-1}\nabla F_m(w_m^{t,b})\right\|_2^2\right]}_{I_2}.
\end{aligned} \tag{41}
$$

To bound I_1, note that

$$
\begin{aligned}
I_1 &= \frac{1}{K}\sum_{m=1}^{M}\frac{1}{M^2q_{s,m}^t}\mathbb{E}_t\left[\left\|\sum_{b=0}^{B-1}\nabla\phi(w_m^{t,b};\xi_m^{t,b}) - \nabla F_m(w^{t,b})\right\|_2^2\right] \\
&= \frac{1}{K}\sum_{m=1}^{M}\frac{1}{M^2q_{s,m}^t}\sum_{b=0}^{B-1}\mathbb{E}_t\left[\left\|\nabla\phi(w_m^{t,b};\xi_m^{t,b}) - \nabla F_m(w^{t,b})\right\|_2^2\right] \\
&\leq \frac{\sigma^2 B}{M^2K}\sum_{m=1}^{M}\frac{1}{q_{s,m}^t}.
\end{aligned}
$$

Since we have $q_{s,m}^t \geq \alpha/M$, we then have $I_1 \leq \frac{\sigma^2 B}{K\alpha}$. We then give an upper bound on I_2. Note that

$$\begin{aligned}
I_2 &\leq 2\mathbb{E}_t \left[\sum_{m=1}^{M} \frac{1}{M^2 q_{s,m}^t} \left\| \sum_{b=0}^{B-1} \left(\nabla F_m(w_m^{t,b}) - \nabla F_m(w^t) \right) \right\|_2^2 \right] \\
&\quad + 2B^2 \mathbb{E}_t \left[\left\| \frac{1}{K} \sum_{m \in \tilde{S}^t} \frac{1}{M q_{s,m}^t} \nabla F_m(w^t) - \nabla F(w^t) \right\|_2^2 \right] + 2B^2 \left\| \nabla F(w^t) \right\|_2^2 \\
&\leq 2B \mathbb{E}_t \left[\sum_{m=1}^{M} \frac{1}{M^2 q_{s,m}^t} \sum_{b=0}^{B-1} \left\| \nabla F_m(w_m^{t,b}) - \nabla F_m(w^t) \right\|_2^2 \right] + 2B^2 \left\| \nabla F(w^t) \right\|_2^2 \\
&\quad + \frac{2B^2}{K} \left(\frac{1}{M^2} \sum_{m=1}^{M} \frac{1}{q_{s,m}^t} \left\| \nabla F_m(w^t) \right\|_2^2 - \left\| \nabla F(w^t) \right\|_2^2 \right) \\
&\leq 2BL^2 \mathbb{E}_t \left[\sum_{m=1}^{M} \frac{1}{M^2 q_{s,m}^t} \sum_{b=0}^{B-1} \left\| w_m^{t,b} - w^t \right\|_2^2 \right] + 2B^2 \left\| \nabla F(w^t) \right\|_2^2 + \frac{2B^2}{K} \mathbb{E}_t \left[V(q_s^t, w^t) \right],
\end{aligned} \tag{42}$$

where the first three inequalities follow Jensen's inequality and Lemma 16, the fourth inequality follows the definition of $V(p, w)$ in (9), and the final inequality follows Assumption 1. Finally, since $q_{s,m}^t \geq \alpha/M$, we have

$$I_2 \leq \frac{2BL^2}{M\alpha} \mathbb{E}_t \left[\sum_{m=1}^{M} \sum_{b=0}^{B-1} \left\| w_m^{t,b} - w^t \right\|_2^2 \right] + 2B^2 \left\| \nabla F(w^t) \right\|_2^2 + \frac{2B^2}{K} \mathbb{E}_t \left[V(q_s^t, w^t) \right]. \tag{43}$$

Combine (41)—(43), we then have

$$\begin{aligned}
\frac{L\mu^2}{2} \mathbb{E}_t \left[\left\| \tilde{g}^t \right\|_2^2 \right] &\leq \frac{\mu^2 \mu_l^2 L \sigma^2 B}{2K\alpha} + \frac{\mu^2 \mu_l^2 L^3 B}{M\alpha} \mathbb{E}_t \left[\sum_{m=1}^{M} \sum_{b=0}^{B-1} \left\| w_m^{t,b} - w^t \right\|_2^2 \right] \\
&\quad + \mu^2 \mu_l^2 B^2 L \left\| \nabla F(w^t) \right\|_2^2 + \frac{\mu^2 \mu_l^2 L B^2}{K} \mathbb{E}_t \left[V(q_s^t, w^t) \right].
\end{aligned}$$

Combine the above equation with (39) and (40), we have

$$\begin{aligned}
\mathbb{E}_t \left[F\left(w^{t+1}\right) \right] &\leq F\left(w^t\right) - \frac{\mu \mu_l B}{2} \left(1 - 2\mu \mu_l BL \right) \left\| \nabla F(w^t) \right\|_2^2 \\
&\quad + \frac{\mu \mu_l L^2}{M} \left(\frac{1}{2} + \frac{\mu \mu_l LB}{\alpha} \right) \mathbb{E}_t \left[\sum_{m=1}^{M} \sum_{b=0}^{B-1} \left\| w_m^{t,b} - w^t \right\|_2^2 \right] \\
&\quad + \frac{\mu^2 \mu_l^2 L \sigma^2 B}{2K\alpha} + \frac{\mu^2 \mu_l^2 L B^2}{K} \mathbb{E}_t \left[V(q_s^t, w^t) \right] + \frac{BL\mu^2 \mu_l^2}{2} \mathbb{E}_t \left[l_t(\hat{p}^t) - l_t(q_s^t) \right].
\end{aligned}$$

By letting $\tilde{\mu} = \mu\mu_l$ and $\tilde{\mu} \leq \frac{1}{4BL}$, we have

$$\begin{aligned}\mathbb{E}_t\left[F\left(w^{t+1}\right)\right] \leq F\left(w^t\right) &- \frac{\tilde{\mu}B}{4}\left\|\nabla F(w^t)\right\|_2^2 \\ &+ \tilde{\mu}L^2\left(\frac{1}{2}+\frac{1}{4\alpha}\right)\mathbb{E}_t\left[\frac{1}{M}\sum_{m=1}^{M}\sum_{b=0}^{B-1}\left\|w_m^{t,b}-w^t\right\|_2^2\right] \\ &+ \frac{\tilde{\mu}^2BL\sigma^2}{2K\alpha} + \frac{\tilde{\mu}^2B^2L}{K}\mathbb{E}_t\left[V(q_s^t,w^t)\right] + \frac{BL\mu^2\mu_l^2}{2}\mathbb{E}_t\left[l_t(\hat{p}^t)-l_t(q_s^t)\right].\end{aligned} \tag{44}$$

Following Lemma 8 of Karimireddy et al. (2020b), we then show the following claim

$$\mathbb{E}_t\left[\frac{1}{M}\sum_{m=1}^{M}\sum_{b=0}^{B-1}\left\|w_m^{t,b}-w^t\right\|_2^2\right] \leq 8B^3\mu_l^2\zeta_{\text{unif}}^2 + 8B^3\mu_l^2\left\|\nabla F(w^t)\right\|_2^2 + 4B^2\mu_l^2\sigma^2. \tag{45}$$

To show the above inequality, first note that when $b=0$, we have $w_m^{t,0}=w^t$, thus we have $\mathbb{E}_t[\|w_m^{t,0}-w^t\|_2^2]=0$. Besides, for $1\leq b\leq B-1$, we have

$$\begin{aligned}&\mathbb{E}_t\left[\left\|w_m^{t,b}-w^t\right\|_2^2\right] \\ &= \mathbb{E}_t\left[\left\|w_m^{t,b-1}-\mu_l\nabla\phi\left(w_m^{t,b-1};\xi_m^{t,b-1}\right)-w^t\right\|_2^2\right] \\ &= \mathbb{E}_t\left[\left\|w_m^{t,b-1}-\mu_l\nabla F_m\left(w_m^{t,b-1}\right)-w^t\right\|_2^2\right] \\ &\quad+\mu_l^2\mathbb{E}_t\left[\left\|\nabla\phi\left(w_m^{t,b-1};\xi_m^{t,b-1}\right)-\nabla F_m\left(w_m^{t,b-1}\right)\right\|_2^2\right] \\ &\leq \left(1+\frac{1}{B-1}\right)\mathbb{E}_t\left[\left\|w_m^{t,b-1}-w^t\right\|_2^2\right] + B\mu_l^2\mathbb{E}_t\left[\left\|\nabla F_m\left(w_m^{t,b-1}\right)\right\|_2^2\right]+\mu_l^2\sigma^2 \\ &\leq \left(1+\frac{1}{B-1}\right)\mathbb{E}_t\left[\left\|w_m^{t,b-1}-w^t\right\|_2^2\right] + 2B\mu_l^2\left\|\nabla F_m\left(w_m^{t,b-1}\right)-\nabla F_m(w^t)\right\|_2^2 \\ &\quad+2B\mu_l^2\left\|\nabla F_m(w^t)\right\|_2^2+\mu_l^2\sigma^2 \\ &\leq \left(1+\frac{1}{B-1}+2B\mu_l^2L^2\right)\mathbb{E}_t\left[\left\|w_m^{t,b-1}-w^t\right\|_2^2\right] + 2B\mu_l^2\left\|\nabla F_m(w^t)\right\|_2^2+\mu_l^2\sigma^2,\end{aligned}$$

where the first inequality follows Assumption 3 and Lemma 16, the second inequality follows Jensen's inequality, the third inequality follows Assumption 1. Let $\mu_l \leq \frac{1}{\sqrt{2}BL}$, then we have $2B\mu_l^2L^2 \leq \frac{1}{B} \leq \frac{1}{B-1}$ and

$$\mathbb{E}_t\left[\left\|w_m^{t,b}-w^t\right\|_2^2\right] \leq \left(1+\frac{2}{B-1}\right)\mathbb{E}_t\left[\left\|w_m^{t,b-1}-w^t\right\|_2^2\right] + 2B\mu_l^2\left\|\nabla F_m(w^t)\right\|_2^2+\mu_l^2\sigma^2.$$

By induction, we then have

$$\mathbb{E}_t\left[\left\|w_m^{t,b}-w^t\right\|_2^2\right] \leq \left(2B\mu_l^2\left\|\nabla F_m(w^t)\right\|_2^2+\mu_l^2\sigma^2\right)\sum_{\tau=0}^{b-1}\left(1+\frac{2}{B-1}\right)^{\tau}$$

Since

$$\sum_{\tau=0}^{b-1}\left(1+\frac{2}{B-1}\right)^{\tau}=\frac{\left(1+\frac{2}{B-1}\right)^{b}-1}{\left(1+\frac{2}{B-1}\right)-1}=\frac{B-1}{2}\left\{\left(1+\frac{2}{B-1}\right)^{b}-1\right\}$$
$$\leq\frac{B-1}{2}\left\{\left(1+\frac{2}{B-1}\right)^{B-1}-1\right\}\leq\frac{B-1}{2}(e^2-1)\leq 4(B-1)\leq 4B.$$

thus we have

$$\mathbb{E}_t\left[\left\|w_m^{t,b}-w^t\right\|_2^2\right]\leq 8B^2\mu_l^2\left\|\nabla F_m(w^t)\right\|_2^2+4B\mu_l^2\sigma^2,$$

which then implies that

$$\begin{aligned}&\frac{1}{M}\sum_{m=1}^{M}\sum_{b=0}^{B-1}\mathbb{E}_t\left[\left\|w_m^{t,b}-w^t\right\|_2^2\right]\\&\leq 8B^3\mu_l^2\frac{1}{M}\sum_{m=1}^{M}\left\|\nabla F_m(w^t)\right\|_2^2+4B^2\mu_l^2\sigma^2\\&\leq 8B^3\mu_l^2\left(\frac{1}{M}\sum_{m=1}^{M}\left\|\nabla F_m(w^t)-\nabla F(w^t)\right\|_2^2+\left\|\nabla F(w^t)\right\|_2^2\right)+4B^2\mu_l^2\sigma^2\\&\leq 8B^3\mu_l^2\zeta_{\text{unif}}^2+8B^3\mu_l^2\left\|\nabla F(w^t)\right\|_2^2+4B^2\mu_l^2\sigma^2,\end{aligned}$$

where the last inequality follows the definition of ζ_{unif}^2. We thus have proved (45).

Besides, we have

$$\begin{aligned}&\tilde{\mu}L^2\left(\frac{1}{2}+\frac{1}{4\alpha}\right)\mathbb{E}_t\left[\frac{1}{M}\sum_{m=1}^{M}\sum_{b=0}^{B-1}\left\|w_m^{t,b}-w^t\right\|_2^2\right]\\&\leq\left(4+\frac{2}{\alpha}\right)\tilde{\mu}\mu_l^2B^3L^2\zeta_{\text{unif}}^2+\left(4+\frac{2}{\alpha}\right)\tilde{\mu}\mu_l^2B^3L^2\left\|\nabla F(w^t)\right\|_2^2+\left(2+\frac{1}{\alpha}\right)\tilde{\mu}\mu_l^2B^2L^2\sigma^2.\end{aligned}$$

Combine the above result with (44), we have

$$\begin{aligned}\mathbb{E}_t\left[F\left(w^{t+1}\right)\right]\leq F\left(w^t\right)&-\frac{\tilde{\mu}B}{4}\left(1-4\left(4+\frac{2}{\alpha}\right)\mu_l^2B^2L^2\right)\left\|\nabla F(w^t)\right\|_2^2\\&+\left(4+\frac{2}{\alpha}\right)\tilde{\mu}\mu_l^2B^2L^2\left(B\zeta_{\text{unif}}^2+\frac{\sigma^2}{2}\right)+\frac{\tilde{\mu}^2BL}{K}\left(\frac{\sigma^2}{2\alpha}+B\mathbb{E}_t\left[V(q_s^t,w^t)\right]\right)\\&+\frac{BL\mu^2\mu_l^2}{2}\mathbb{E}_t\left[l_t(\hat{p}^t)-l_t(q_s^t)\right].\end{aligned}$$

Let $\mu_l\leq\frac{1}{4BL}\sqrt{\frac{1}{2+1/\alpha}}$, we then have

$$\begin{aligned}\mathbb{E}_t\left[F\left(w^{t+1}\right)\right]\leq F\left(w^t\right)&-\frac{\tilde{\mu}B}{8}\left\|\nabla F(w^t)\right\|_2^2+\left(4+\frac{2}{\alpha}\right)\tilde{\mu}\mu_l^2B^2L^2\left(B\zeta_{\text{unif}}^2+\frac{\sigma^2}{2}\right)\\&+\frac{\tilde{\mu}^2BL}{K}\left(\frac{\sigma^2}{2\alpha}+B\mathbb{E}_t\left[V(q_s^t,w^t)\right]\right)+\frac{BL\tilde{\mu}^2}{2}\mathbb{E}_t\left[l_t(\hat{p}^t)-l_t(q_s^t)\right],\end{aligned}$$

which then implies that

$$\left\|\nabla F(w^t)\right\|_2^2 \leq \frac{8}{\tilde{\mu}B}\left(F\left(w^t\right) - \mathbb{E}_t\left[F\left(w^{t+1}\right)\right]\right) + 8\left(4+\frac{2}{\alpha}\right)\mu_l^2 B^2 L^2\left(\zeta_{\text{unif}}^2 + \frac{\sigma^2}{2B}\right) + \frac{8\tilde{\mu}BL}{K}\left(\frac{\sigma^2}{2B\alpha} + \mathbb{E}_t\left[V(q_s^t, w^t)\right]\right) + 4L\tilde{\mu}\mathbb{E}_t\left[l_t(\hat{p}^t) - l_t(q_s^t)\right].$$

Taking full expectation on both sides, summing over $t=0$ to $t=T-1$ and taking average, we then have

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla F(w^t)\right\|_2^2\right] &\leq \frac{8}{\tilde{\mu}B}\left(F(w^0) - \mathbb{E}\left[F(w^T)\right]\right) + 8\left(4+\frac{2}{\alpha}\right)\mu_l^2 B^2 L^2\left(\zeta_{\text{unif}}^2 + \frac{\sigma^2}{2B}\right) \\
&\quad + \frac{8\tilde{\mu}BL}{K}\left(\frac{\sigma^2}{2B\alpha} + \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_t\left[V(q_s^t, w^t)\right]\right) + 4L\tilde{\mu}\times\frac{\text{D-Regret}_T\left(q^{1:T}\right)}{T} \\
&\leq \frac{8}{\tilde{\mu}B}\left(F(w^0) - F^\star\right) + 8\left(4+\frac{2}{\alpha}\right)\mu_l^2 B^2 L^2\left(\zeta_{\text{unif}}^2 + \frac{\sigma^2}{2B}\right) \\
&\quad + \frac{8\tilde{\mu}BL}{K}\left(\frac{\sigma^2}{2B\alpha} + \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_t\left[V(q_s^t, w^t)\right]\right) + 4L\tilde{\mu}\times\frac{\text{D-Regret}_T\left(q^{1:T}\right)}{T} \\
&\leq \frac{8}{\tilde{\mu}B}\left(F(w^0) - F^\star\right) + 8\left(4+\frac{2}{\alpha}\right)\mu_l^2 B^2 L^2\left(\zeta_{\text{unif}}^2 + \frac{\sigma^2}{2B}\right) \\
&\quad + \frac{8\tilde{\mu}BL}{K}\left(\frac{\sigma^2}{2B\alpha} + \zeta_T^2(\alpha,\beta)\right) + 4L\tilde{\mu}\times\frac{\text{D-Regret}_T\left(q^{1:T}\right)}{T},
\end{aligned}$$

where the last inequality follows the fact that

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}V(q_s^t, w^t)\right] = \frac{1}{T}\mathbb{E}_{w^1,\ldots,w^T}\left[\sum_{t=1}^{T}V(q_s^t, w^t)\right] \leq \frac{1}{T}\sup_{w^1}\cdots\sup_{w^T}\sum_{t=1}^{T}V(q_s^t, w^t) = \zeta_T^2(\alpha,\beta),$$

and the last equality is by (31).

In summary, when $\tilde{\mu} \leq \frac{1}{4BL}$ and $\mu_l \leq \frac{1}{4BL}\sqrt{\frac{1}{2+1/\alpha}}$, we have

$$\begin{aligned}
\mathbb{E}\left[\left\|\nabla F(w^R)\right\|_2^2\right] &= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla F(w^t)\right\|_2^2\right] \\
&\leq \frac{8}{\tilde{\mu}BT}\left(F(w^0) - F^\star\right) + 8\left(4+\frac{2}{\alpha}\right)\mu_l^2 B^2 L^2\left(\zeta_{\text{unif}}^2 + \frac{\sigma^2}{2B}\right) \\
&\quad + 4\tilde{\mu}BL\left(\frac{2\zeta_T^2(\alpha,\beta)}{K} + \frac{\sigma^2}{KB\alpha} + \frac{\text{D-Regret}_T\left(q_s^{1:T}\right)}{BT}\right).
\end{aligned}$$

By letting $\mu \geq 1$, $\tilde{\mu} \leq \frac{1}{4BL}\sqrt{\frac{1}{1+1/(2\alpha)}}$ and recall that $D^F = F(w^1) - F^\star$, we then have

$$\begin{aligned}
\mathbb{E}\left[\left\|\nabla F(w^R)\right\|_2^2\right] &\leq \frac{8D^F}{\tilde{\mu}BT} + 8\left(4+\frac{2}{\alpha}\right)\tilde{\mu}^2 B^2 L^2\left(\zeta_{\text{unif}}^2 + \frac{\sigma^2}{2B}\right) \\
&\quad + 4\tilde{\mu}BL\left(\frac{2\zeta_T^2(\alpha,\beta)}{K} + \frac{\sigma^2}{KB\alpha} + \frac{\text{D-Regret}_T\left(q_s^{1:T}\right)}{BT}\right).
\end{aligned} \tag{46}$$

We then turn to bound the regret, note that

$$g_m^t = \mu_l \sum_{b=0}^{B-1} \nabla \phi(w_m^{t,b}; \xi_m^{t,b}),$$

thus by Assumption 3, we have

$$\|g_m^t\|_2^2 \leq B^2 \mu_l^2 G^2,$$

which then implies that

$$a_m^t = \frac{\|g_m^t\|_2^2}{B(M\mu_l)^2} \leq \frac{BG^2}{M^2} \quad \text{and} \quad a_{\max}^t \leq \frac{BG^2}{M^2}.$$

Thus, by Theorem 6 and note that $q_s^t \in \mathcal{A}$ for all $t \in [T]$ and $\mathrm{TV}(q_s^{1:T}) \leq \beta$, we have

$$\text{D-Regret}_T(q_s^{1:T}) \leq \frac{\log M}{\eta} + \frac{2\beta \log(M/\alpha)}{\eta} + \frac{\eta T M^2 B^2 G^4}{2K^2 \alpha^6}.$$

Let

$$\eta = \frac{K\alpha^3}{MBG^2} \sqrt{\frac{2\log M + 4\beta \log(M/\alpha)}{T}},$$

we then have

$$\text{D-Regret}_T(q_s^{1:T}) \leq \sqrt{T} \frac{MBG^2}{K\alpha^3} \sqrt{\frac{1}{2}\log M + \beta \log(M/\alpha)}.$$

Putting the above inequality to (46), we have

$$\begin{aligned} \mathbb{E}\left[\left\|\nabla F(w^R)\right\|_2^2\right] \leq \frac{8D^F}{\tilde{\mu}BT} &+ 8\left(4 + \frac{2}{\alpha}\right)\tilde{\mu}^2 B^2 L^2 \left(\zeta_{\text{unif}}^2 + \frac{\sigma^2}{2B}\right) \\ &+ 4\tilde{\mu}BL \left(\frac{2\zeta_T^2(\alpha,\beta)}{K} + \frac{\sigma^2}{KB\alpha} + \frac{MG^2}{K\alpha^3}\sqrt{\frac{\frac{1}{2}\log M + \beta\log(M/\alpha)}{T}}\right). \end{aligned} \tag{47}$$

When letting

$$\frac{8D^F}{\tilde{\mu}BT} = 8\left(4 + \frac{2}{\alpha}\right)\tilde{\mu}^2 B^2 L^2 \left(\zeta_{\text{unif}}^2 + \frac{\sigma^2}{2B}\right),$$

we have

$$\tilde{\mu} = \frac{(D^F)^{\frac{1}{3}}}{\left(4 + \frac{2}{\alpha}\right)^{\frac{1}{3}} BL^{\frac{2}{3}} \left(\zeta_{\text{unif}}^2 + \frac{\sigma^2}{2B}\right)^{\frac{1}{3}} T^{\frac{1}{3}}},$$

which implies that

$$\begin{aligned} \frac{8D^F}{\tilde{\mu}BT} + 8\left(4 + \frac{2}{\alpha}\right)\tilde{\mu}^2 B^2 L^2 \left(\zeta_{\text{unif}}^2 + \frac{\sigma^2}{2B}\right) &\leq \frac{16\left(4 + \frac{2}{\alpha}\right)^{\frac{1}{3}} (D^F)^{\frac{2}{3}} L^{\frac{2}{3}} \left(\zeta_{\text{unif}}^2 + \frac{\sigma^2}{2B}\right)^{\frac{1}{3}}}{T^{\frac{2}{3}}} \\ &\leq \frac{16\left(4 + \frac{2}{\alpha}\right)^{\frac{1}{3}} (D^F)^{\frac{2}{3}} L^{\frac{2}{3}} \left(\zeta_{\text{unif}}^{\frac{2}{3}} + \frac{\sigma^{\frac{2}{3}}}{(2B)^{\frac{1}{3}}}\right)}{T^{\frac{2}{3}}} \end{aligned}$$

On the other hand, when letting

$$\frac{8D^F}{\tilde{\mu}BT} = 4\tilde{\mu}BL\left(\frac{2\zeta_T^2(\alpha,\beta)}{K} + \frac{\sigma^2}{KB\alpha} + \frac{MG^2}{K\alpha^3}\sqrt{\frac{\frac{1}{2}\log M + \beta\log(M/\alpha)}{T}}\right),$$

we have

$$\tilde{\mu} = \frac{\sqrt{2D^F}}{B\sqrt{L}\sqrt{\frac{2\zeta_T^2(\alpha,\beta)}{K} + \frac{\sigma^2}{KB\alpha} + \sqrt{\frac{\frac{1}{2}\log M + \beta\log(M/\alpha)}{T}}}\sqrt{T}},$$

which implies that

$$\begin{aligned}
&\frac{8D^F}{\tilde{\mu}BT} + 4\tilde{\mu}BL\left(\frac{2\zeta_T^2(\alpha,\beta)}{K} + \frac{\sigma^2}{KB\alpha} + \frac{\text{D-Regret}_T\left(q^\star\right)}{BT}\right) \\
&\qquad \leq \frac{8\sqrt{2}\sqrt{D^F}\sqrt{L}\sqrt{\frac{2\zeta_T^2(\alpha,\beta)}{K} + \frac{\sigma^2}{KB\alpha} + \sqrt{\frac{\frac{1}{2}\log M + \beta\log(M/\alpha)}{T}}}}{\sqrt{T}} \\
&\qquad \leq \frac{8\sqrt{2}\sqrt{D^F}\sqrt{L}}{\sqrt{T}}\left(\frac{\sqrt{2}\zeta_T(\alpha,\beta)}{\sqrt{K}} + \frac{\sigma}{\sqrt{KB\alpha}} + \left(\frac{\frac{1}{2}\log M + \beta\log(M/\alpha)}{T}\right)^{\frac{1}{4}}\right).
\end{aligned}$$

Thus, when $\mu \geq 1$ and

$$\begin{aligned}
\tilde{\mu} = \min\Bigg\{&\frac{1}{4BL}\sqrt{\frac{1}{2+1/\alpha}}, \frac{(D^F)^{\frac{1}{3}}}{\left(4+\frac{2}{\alpha}\right)^{\frac{1}{3}}BL^{\frac{2}{3}}\left(\zeta_{\text{unif}}^2 + \frac{\sigma^2}{2B}\right)^{\frac{1}{3}}T^{\frac{1}{3}}}, \\
&\frac{\sqrt{2D^F}}{B\sqrt{L}\sqrt{\frac{2\zeta_T^2(\alpha,\beta)}{K} + \frac{\sigma^2}{KB\alpha} + \sqrt{\frac{\frac{1}{2}\log M + \beta\log(M/\alpha)}{T}}}\sqrt{T}}\Bigg\},
\end{aligned}$$

we then have

$$\begin{aligned}
\mathbb{E}\left[\left\|\nabla F(w^R)\right\|_2^2\right] &\lesssim \frac{D^F L\sqrt{2+\frac{1}{\alpha}}}{T} + \frac{\left(4+\frac{2}{\alpha}\right)^{\frac{1}{3}}(D^F)^{\frac{2}{3}}L^{\frac{2}{3}}\left(\zeta_{\text{unif}}^{\frac{2}{3}} + \frac{\sigma^{\frac{2}{3}}}{B^{\frac{1}{3}}}\right)}{T^{\frac{2}{3}}} \\
&\quad + \frac{\sqrt{D^F}\sqrt{L}}{\sqrt{T}}\left(\frac{\zeta_T(\alpha,\beta)}{\sqrt{K}} + \frac{\sigma}{\sqrt{KB\alpha}} + \left(\frac{\frac{1}{2}\log M + \beta\log(M/\alpha)}{T}\right)^{\frac{1}{4}}\right) \\
&\lesssim \frac{D^F L\sqrt{2+\frac{1}{\alpha}}}{T} + \frac{\left(4+\frac{2}{\alpha}\right)^{\frac{1}{3}}(D^F L)^{\frac{2}{3}}\zeta_{\text{unif}}^{\frac{2}{3}}}{T^{\frac{2}{3}}} + \frac{\left(4+\frac{2}{\alpha}\right)^{\frac{1}{3}}(D^F L)^{\frac{2}{3}}\sigma^{\frac{2}{3}}}{B^{\frac{1}{3}}T^{\frac{2}{3}}} \\
&\quad + \frac{\sqrt{D^F L}\zeta_T(\alpha,\beta)}{\sqrt{TK}} + \frac{\sqrt{D^F L}\sigma}{\sqrt{TKB\alpha}} + \frac{\sqrt{D^F L}}{\sqrt{T}}\left(\frac{\frac{1}{2}\log M + \beta\log(M/\alpha)}{T}\right)^{\frac{1}{4}}.
\end{aligned}$$

B.5 Proof of Theorem 6

We first state a proposition that will be used to prove Theorem 6. The key difference between Theorem 6 and Proposition 13 is that in Proposition 13 the comparator sequence lies in $\mathcal{A}$, and, as a result, there is no projection error.

Proposition 13 *Suppose the conditions of Theorem 6 hold. For any comparator sequence $q^{1:T}$ with $q^t \in \mathcal{A}$, $t \in [T]$, we have*

$$\textit{D-Regret}_T(q^{1:T}) \leq \frac{\log M}{\eta} + \frac{2\log(M/\alpha)}{\eta}\mathbb{E}\left[TV\left(q^{1:T}\right)\right] + \frac{\eta M^6}{2K^2\alpha^6}\sum_{t=1}^{T}\mathbb{E}\left[\left(a_{\max}^t\right)^2\right].$$

Proof By Lemma 14 and the definition of $\hat{p}^{t+1}$ in Step 7 of Algorithm 1, we have

$$\langle \hat{p}^{t+1} - q^t, \nabla \hat{l}_t(\hat{p}^t; \hat{p}^t)\rangle \leq \frac{1}{\eta}\langle \nabla\Phi(\hat{p}^t) - \nabla\Phi(\hat{p}^{t+1}), \hat{p}^{t+1} - q^t\rangle. \tag{48}$$

By the convexity of $\hat{l}_t(\cdot; \hat{p}^t)$, we have

$$\hat{l}_t(\hat{p}^t; \hat{p}^t) - \hat{l}_t(q^t; \hat{p}^t) \leq \langle \nabla \hat{l}_t(\hat{p}^t; \hat{p}^t), \hat{p}^t - q^t\rangle = \langle \nabla \hat{l}_t(\hat{p}^t; \hat{p}^t), \hat{p}^{t+1} - q^t\rangle + \langle \nabla \hat{l}_t(\hat{p}^t; \hat{p}^t), \hat{p}^t - \hat{p}^{t+1}\rangle.$$

Then, by (48), we further have

$$\hat{l}_t(\hat{p}^t; \hat{p}^t) - \hat{l}_t(q^t; \hat{p}^t) \leq \frac{1}{\eta}\langle \nabla\Phi(\hat{p}^t) - \nabla\Phi(\hat{p}^{t+1}), \hat{p}^{t+1} - q^t\rangle + \langle \nabla \hat{l}_t(\hat{p}^t; \hat{p}^t), \hat{p}^t - \hat{p}^{t+1}\rangle.$$

From the definition of $\mathcal{D}$, we have

$$D_\Phi(x_1 \,\|\, x_2) = D_\Phi(x_3 \,\|\, x_2) + D_\Phi(x_1 \,\|\, x_3) + \langle \nabla\Phi(x_2) - \nabla\Phi(x_3), x_3 - x_1\rangle, \qquad x_1, x_2, x_3 \in \mathcal{D}.$$

Then

$$\begin{aligned}
&\hat{l}_t(\hat{p}^t; \hat{p}^t) - \hat{l}_t(q^t; \hat{p}^t) \\
&\quad\leq \frac{1}{\eta}\left[D_\Phi\left(q^t \| \hat{p}^t\right) - D_\Phi\left(q^t \| \hat{p}^{t+1}\right) - D_\Phi\left(\hat{p}^{t+1} \| \hat{p}^t\right)\right] + \langle \nabla \hat{l}_t(\hat{p}^t; \hat{p}^t), \hat{p}^t - \hat{p}^{t+1}\rangle \\
&\quad= \frac{1}{\eta}\left[D_\Phi\left(q^t \| \hat{p}^t\right) - D_\Phi\left(q^{t+1} \| \hat{p}^{t+1}\right)\right] + \frac{1}{\eta}\left[D_\Phi\left(q^{t+1} \| \hat{p}^{t+1}\right) - D_\Phi\left(q^t \| \hat{p}^{t+1}\right)\right] \\
&\qquad - \frac{1}{\eta}D_\Phi\left(\hat{p}^{t+1} \| \hat{p}^t\right) + \langle \nabla \hat{l}_t(\hat{p}^t; \hat{p}^t), \hat{p}^t - \hat{p}^{t+1}\rangle.
\end{aligned} \tag{49}$$

We bound the second term in (49) as

$$\begin{aligned}
D_\Phi\left(q^{t+1} \| \hat{p}^{t+1}\right) - D_\Phi\left(q^t \| \hat{p}^{t+1}\right) &= \Phi(q^{t+1}) - \Phi(q^t) - \langle \nabla\Phi(\hat{p}^{t+1}), q^{t+1} - q^t\rangle \\
&\overset{(a)}{\leq} \langle \nabla\Phi(q^{t+1}) - \nabla\Phi(\hat{p}^{t+1}), q^{t+1} - q^t\rangle \\
&\overset{(b)}{\leq} \|\nabla\Phi(q^{t+1}) - \nabla\Phi(\hat{p}^{t+1})\|_\infty \|q^{t+1} - q^t\|_1 \\
&\overset{(c)}{\leq} 2\log(M/\alpha)\|q^{t+1} - q^t\|_1,
\end{aligned} \tag{50}$$

where (a) follows from the convexity of $\Phi(\cdot)$, (b) follows from the fact that the dual norm of $\|\cdot\|_1$ is $\|\cdot\|_\infty$, and (c) follows from the following fact that

$$\|\nabla\Phi(p)\|_\infty = \max_{1\le m\le M} |\log(p_m)| \le \log(M/\alpha) \text{ for all } p\in\mathcal{A}.$$

Besides, by Pinsker's inequality, we have $D_\Phi(p\,\|\,q) \ge \frac{1}{2}\|p-q\|_1^2$ for all $p,q\in\mathcal{P}_{M-1}$. Thus, $\Phi(\cdot)$ is 1-strongly convex, we can bound the third and fourth term in (49) as

$$-\frac{1}{\eta}D_\Phi\left(\hat{p}^{t+1}\|\hat{p}^t\right) + \langle\nabla\hat{l}_t(\hat{p}^t;\hat{p}^t), \hat{p}^t-\hat{p}^{t+1}\rangle \le -\frac{1}{2\eta}\|\hat{p}^{t+1}-\hat{p}^t\|_1^2 + \|\nabla\hat{l}_t(\hat{p}^t;\hat{p}^t)\|_\infty\|\hat{p}^t-\hat{p}^{t+1}\|_1.$$

Since $ab\le a^2/(2\epsilon)+b^2\epsilon/2$, $a,b,\epsilon>0$, we further have

$$-\frac{1}{\eta}D_\Phi\left(\hat{p}^{t+1}\|\hat{p}^t\right) + \langle\nabla\hat{l}_t(\hat{p}^t;\hat{p}^t), \hat{p}^t-\hat{p}^{t+1}\rangle \le \frac{\eta}{2}\|\nabla\hat{l}_t(\hat{p}^t;\hat{p}^t)\|_\infty^2.$$

Let

$$Q_t = M^3 a_{\max}^t/(K\alpha^3),$$

we then have

$$\left|\left[\nabla\hat{l}_t(q;p^t)\right]_m\right| = \frac{1}{K^2}\cdot\frac{a_m^t}{q_m^2 p_m^t}\mathcal{N}\left\{m\in S^t\right\} \le \frac{1}{K^2}\cdot\frac{a_{\max}^t}{\alpha^3/M^3}\cdot K \le \frac{M^3 a_{\max}^t}{K\alpha^3} = Q_t$$

for all $m\in[M]$, thus we have $\|\nabla\hat{l}_t(q;p^t)\|_\infty \le Q_t$, which then implies that

$$-\frac{1}{\eta}D_\Phi\left(\hat{p}^{t+1}\|\hat{p}^t\right) + \langle\nabla\hat{l}_t(\hat{p}^t;\hat{p}^t), \hat{p}^t-\hat{p}^{t+1}\rangle \le \frac{\eta}{2}Q_t^2. \tag{51}$$

Combining (49)-(51), we have

$$\hat{l}_t(\hat{p}^t;\hat{p}^t) - \hat{l}_t(q^t;\hat{p}^t) \le \frac{D_\Phi\left(q^t\|\hat{p}^t\right)}{\eta} - \frac{D_\Phi\left(q^{t+1}\|\hat{p}^{t+1}\right)}{\eta} + \frac{2\log(M/\alpha)}{\eta}\|q^{t+1}-q^t\|_1 + \frac{\eta}{2}Q_t^2.$$

This implies that

$$\begin{aligned}
&\sum_{t=1}^T \hat{l}_t(\hat{p}^t;\hat{p}^t) - \sum_{t=1}^T \hat{l}_t(q^t;\hat{p}^t) \\
&\qquad\le \frac{D_\Phi\left(q^1\|\hat{p}^1\right)}{\eta} - \frac{D_\Phi\left(q^{T+1}\|\hat{p}^{T+1}\right)}{\eta} + \frac{2\log(M/\alpha)}{\eta}\sum_{t=1}^T\|q^{t+1}-q^t\|_1 + \frac{\eta}{2}\sum_{t=1}^T Q_t^2 \\
&\qquad\le \frac{D_\Phi\left(q^1\|\hat{p}^1\right)}{\eta} + \frac{2\log(M/\alpha)}{\eta}\sum_{t=1}^T\|q^{t+1}-q^t\|_1 + \frac{\eta}{2}\sum_{t=1}^T Q_t^2.
\end{aligned}$$

Since $\hat{p}_1$ is the uniform distribution, we have that

$$D_\Phi(q\,\|\,p^{\text{unif}}) = \log M + \sum_{m=1}^M q_m\log q_m \le \log M \text{ for all } q\in\mathcal{P}_{M-1}.$$

Thus, we have

$$\sum_{t=1}^{T} \hat{l}_t(\hat{p}^t;\hat{p}^t) - \sum_{t=1}^{T} \hat{l}_t(q^t;\hat{p}^t) \leq \frac{\log M}{\eta} + \frac{2\log(M/\alpha)}{\eta}\mathrm{TV}\left(q^{1:T}\right) + \frac{\eta}{2}\sum_{t=1}^{T} Q_t^2 \tag{52}$$

Finally, note that

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \hat{l}_t(\hat{p}^t;\hat{p}^t) - \sum_{t=1}^{T} \hat{l}_t(q^t;\hat{p}^t)\right] &= \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{E}_{S^t}\left[\hat{l}_t(\hat{p}^t;\hat{p}^t)\right] - \mathbb{E}_{S^t}\left[\hat{l}_t(q^t;\hat{p}^t)\right]\right] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[l_t(\hat{p}^t) - l_t(q^t)\right] \\
&= \text{D-Regret}_T(q^{1:T}).
\end{aligned}$$

The conclusion follows by taking expectation on right hand side of (52). ■

We are now ready to prove Theorem 6.

Proof [Proof of Theorem 6] For any comparator sequence $q^{1:T}$ with $q^t \in \mathcal{P}_{M-1}$, $t \in [T]$, we prove Theorem 6 by first constructing a suitable sequence $\tilde{q}^{1:T}$ that is defined as

$$\tilde{q}_m^t = \begin{cases} \alpha/M & \text{if } q_m^t < \alpha/M, \\ q_m^t - \omega(q^t,\alpha)\left(q_m^t - \frac{\alpha}{M}\right) & \text{if } q_m^t \geq \alpha/M, \end{cases} \tag{53}$$

where $\omega(q^t,\alpha)$ is defined in (19). We now show that $\tilde{q}^t \in \mathcal{A}$, $t \in [T]$, by showing that $\tilde{q}_m^t \geq \alpha/M$, $m \in [M]$, and $\sum_{m\in[M]} \tilde{q}_m^t = 1$. For $m \in [M]$ such that $q_m^t < \alpha/M$, we have from (53) that $\tilde{q}_m^t = \alpha/M$. For $m \in [M]$ such that $q_m^t \geq \alpha/M$, by (53), we have $\tilde{q}_m^t - \alpha/M = \left(1-\omega(q^t,\alpha)\right)\left(q_m^t - \alpha/M\right)$. Thus, we proceed to show that $\omega(q^t,\alpha) \leq 1$. Since

$$\begin{aligned}
\sum_{m=1}^{M} q_m^t \mathbb{1}\left\{q_m^t < \frac{\alpha}{M}\right\} + \sum_{m=1}^{M} q_m^t \mathbb{1}\left\{q_m^t \geq \frac{\alpha}{M}\right\} &= 1 \\
\geq \alpha = \sum_{m=1}^{M} \frac{\alpha}{M}\mathbb{1}\left\{q_m^t < \frac{\alpha}{M}\right\} &+ \sum_{m=1}^{M} \frac{\alpha}{M}\mathbb{1}\left\{q_m^t \geq \frac{\alpha}{M}\right\},
\end{aligned}$$

we have

$$\sum_{m=1}^{M}\left(q_m^t - \frac{\alpha}{M}\right)\mathbb{1}\left\{q_m^t \geq \frac{\alpha}{M}\right\} \geq \sum_{m=1}^{M}\left(\frac{\alpha}{M} - q_m^t\right)\mathbb{1}\left\{q_m^t < \frac{\alpha}{M}\right\}.$$

Therefore, $0 \leq \omega(q^t, \alpha) \leq 1$. Furthermore, $\omega(q^t, 0) = 1$ and $\omega(q^t, 1) = 1$. Finally, we show that $\sum_{m=1}^M \tilde{q}_m^t = 1$. By (53) and the definition of $\omega(q^t, \alpha)$ in (19), we have

$$\begin{aligned}
\sum_{m=1}^M \tilde{q}_m^t &= \sum_{m=1}^M \frac{\alpha}{M} \mathbb{1}\left\{q_m^t < \frac{\alpha}{M}\right\} + \sum_{m=1}^M q_m^t \mathbb{1}\left\{q_m^t \geq \frac{\alpha}{M}\right\} \\
&\quad - \omega(q^t, \alpha) \sum_{m=1}^M \left(q_m^t - \frac{\alpha}{M}\right) \mathbb{1}\left\{q_m^t \geq \frac{\alpha}{M}\right\} \\
&= \sum_{m=1}^M \frac{\alpha}{M} \mathbb{1}\left\{q_m^t < \frac{\alpha}{M}\right\} + \sum_{m=1}^M q_m^t \mathbb{1}\left\{q_m^t \geq \frac{\alpha}{M}\right\} - \sum_{m=1}^M \left(\frac{\alpha}{M} - q_m^t\right) \mathbb{1}\left\{q_m^t < \frac{\alpha}{M}\right\} \\
&= \sum_{m=1}^M q_m^t \mathbb{1}\left\{q_m^t \geq \frac{\alpha}{M}\right\} + \sum_{m=1}^M q_m^t \mathbb{1}\left\{q_m^t < \frac{\alpha}{M}\right\} \\
&= 1.
\end{aligned}$$

Therefore, $\tilde{q}^t \in \mathcal{A}$ for any $t \in [T]$.

Note that we then have

$$\text{D-Regret}_T(q^{1:T}) = \mathbb{E}\left[\sum_{t=1}^T l_t(\hat{p}^t) - \sum_{t=1}^T l_t(\tilde{q}^t) + \sum_{t=1}^T l_t(\tilde{q}^t) - \sum_{t=1}^T l_t(q^t)\right]. \tag{54}$$

By Proposition 13, we further have that

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^T l_t(\hat{p}^t) - \sum_{t=1}^T l_t(\tilde{q}^t)\right] &\leq \frac{\log M}{\eta} + \frac{2\log(M/\alpha)}{\eta} \mathbb{E}\left[\text{TV}\left(q^{1:T}\right)\right] \\
&\quad + \frac{\eta M^6}{2K^2\alpha^6} \sum_{t=1}^T \mathbb{E}\left[\left(a_{\max}^t\right)^2\right] + \frac{2\log(M/\alpha)}{\eta} \mathbb{E}\left[\text{TV}\left(\tilde{q}^{1:T}\right) - \text{TV}\left(q^{1:T}\right)\right]. \quad (55)
\end{aligned}$$

Therefore, to prove Theorem 6, we need to bound the terms $\sum_{t=1}^T l_t(\tilde{q}^t) - \sum_{t=1}^T l_t(q^t)$ and $\text{TV}\left(\tilde{q}^{1:T}\right) - \text{TV}\left(q^{1:T}\right)$.

We first bound $\sum_{t=1}^T l_t(\tilde{q}^t) - \sum_{t=1}^T l_t(q^t)$. When $q_m^t < \alpha/M$, then $1/\tilde{q}_m^t - 1/q_m^t < 0$; and when $q_m^t \geq \alpha/M$, then

$$\frac{1}{\tilde{q}_m^t} - \frac{1}{q_m^t} = \frac{1}{q_m^t} \cdot \left[\frac{1}{1 - \omega(q^t, \alpha)\left(1 - \frac{\alpha}{Mq_m^t}\right)} - 1\right] = \frac{1}{q_m^t} \cdot \frac{\omega(q^t, \alpha)\left(1 - \frac{\alpha}{Mq_m^t}\right)}{1 - \omega(q^t, \alpha)\left(1 - \frac{\alpha}{Mq_m^t}\right)}.$$

Since

$$\omega(q^t, \alpha)\left(1 - \frac{\alpha}{Mq_m^t}\right) \leq \omega(q^t, \alpha) \quad \text{and} \quad 1 - \omega(q^t, \alpha)\left(1 - \frac{\alpha}{Mq_m^t}\right) \geq 1 - \omega(q^t, \alpha) + \frac{\omega(q^t, \alpha)\alpha}{M}$$

as $q_m^t \leq 1$, we have

$$\frac{1}{\tilde{q}_m^t} - \frac{1}{q_m^t} \leq \frac{1}{q_m^t} \cdot \frac{\omega(q^t, \alpha)}{1 - \omega(q^t, \alpha)\left(1 - \frac{\alpha}{M}\right)} = \frac{\phi(q^t, \alpha)}{q_m^t}.$$

Thus,

$$\begin{aligned}\sum_{t=1}^{T} l_t(\tilde{q}^t) - \sum_{t=1}^{T} l_t(q^t) &= \frac{1}{K}\sum_{t=1}^{T}\sum_{m=1}^{M} a_m^t \left(\frac{1}{\tilde{q}_m^t} - \frac{1}{q_m^t}\right) \\ &\le \frac{1}{K}\sum_{t=1}^{T}\sum_{m=1}^{M} a_m^t \left(\frac{1}{\tilde{q}_m^t} - \frac{1}{q_m^t}\right) \mathbb{1}\left\{q_m^t \ge \frac{\alpha}{M}\right\} \\ &\le \frac{1}{K}\sum_{t=1}^{T} \phi(q^t,\alpha) \sum_{m=1}^{M} \frac{a_m^t}{q_m^t} \mathbb{1}\left\{q_m^t \ge \frac{\alpha}{M}\right\} \\ &\le \frac{1}{K}\sum_{t=1}^{T} \phi(q^t,\alpha) l_t(q^t). \end{aligned} \tag{56}$$

Next, we bound $\mathrm{TV}\left(\tilde{q}^{1:T}\right) - \mathrm{TV}\left(q^{1:T}\right)$. Note that

$$\begin{aligned}\mathrm{TV}\left(\tilde{q}^{1:T}\right) &= \sum_{t=2}^{T} \left\|\tilde{q}^t - \tilde{q}^{t-1}\right\|_1 \\ &= \sum_{t=2}^{T} \left\|\tilde{q}^t - q^t + q^t - q^{t-1} + q^{t-1} - \tilde{q}^{t-1}\right\|_1 \\ &\le \sum_{t=2}^{T} \left\|\tilde{q}^t - q^t\right\|_1 + \sum_{t=2}^{T} \left\|q^t - q^{t-1}\right\|_1 + \sum_{t=2}^{T} \left\|q^{t-1} - \tilde{q}^{t-1}\right\|_1 \\ &\le \mathrm{TV}\left(q^{1:T}\right) + 2\sum_{t=1}^{T} \left\|\tilde{q}^t - q^t\right\|_1 . \end{aligned}$$

We now upper bound $\sum_{t=1}^{T} \left\|\tilde{q}^t - q^t\right\|_1$. If $q_m^t < \alpha/M$, then $|\tilde{q}_m^t - q_m^t| = \alpha/M - q_m^t$. If $q_m^t \ge \alpha/M$, by (53), we have $\left|\tilde{q}_m^t - q_m^t\right| = \omega(q^t,\alpha)\left(q_m^t - \alpha/M\right)$. Therefore, recalling the definition of $\psi(q^t,\alpha)$ in (19), we have

$$\begin{aligned}\left\|\tilde{q}^t - q^t\right\|_1 &= \sum_{m=1}^{M} \left(\frac{\alpha}{M} - q_m^t\right) \mathbb{1}\left\{q_m^t < \frac{\alpha}{M}\right\} + \omega(q^t,\alpha) \sum_{m=1}^{M} \left(q_m^t - \frac{\alpha}{M}\right) \mathbb{1}\left\{q_m^t \ge \frac{\alpha}{M}\right\} \\ &= 2\sum_{m=1}^{M} \left(\frac{\alpha}{M} - q_m^t\right) \mathbb{1}\left\{q_m^t < \frac{\alpha}{M}\right\} \\ &= 2\psi(q^t,\alpha) \end{aligned}$$

and

$$\mathrm{TV}\left(\tilde{q}^{1:T}\right) - \mathrm{TV}\left(q^{1:T}\right) \le 4\sum_{t=1}^{T} \psi(q^t,\alpha). \tag{57}$$

Combining (54), (55), (56), and (57), and taking expectation on both sides, we obtain the final result. ■

B.6 Proof of Theorem 8

Given a comparator sequence $q^{1:T}$, where q^t is allowed to be random, such that $q^t \in \mathcal{A}$ for all $t \in [T]$ and $\mathbb{E}[\mathrm{TV}(q^{1:T})] \leq \beta$, let

$$\eta^\star = \frac{K\alpha^3}{M^3 A_{\max}} \sqrt{\frac{2\log M + 4\beta \log(M/\alpha)}{T}}. \tag{58}$$

The proof proceeds in two steps. First, we show that there exists an expert learning rate $\eta_e \in \mathcal{E}$ such that the regret bound for $\hat{p}_e^{1:T}$ is close to (21). That is, we show that there exists $\eta_e \in \mathcal{E}$ such that

$$\mathbb{E}\left[\sum_{t=1}^T \hat{l}_t(\hat{p}_e^t; \hat{p}^t) - \sum_{t=1}^T l_t(q^t)\right] \leq \frac{3M^3 A_{\max}}{K\alpha^3} \sqrt{T\left[\frac{1}{2}\log M + \beta \log\left(M/\alpha\right)\right]}. \tag{59}$$

Note that $S^t \sim \hat{p}^t$. Second, we show that the output of meta-algorithm can track the best expert with small regret. That is, we show that

$$\mathbb{E}\left[\sum_{t=1}^T l_t(\hat{p}^t)\right] - \mathbb{E}\left[\sum_{t=1}^T \hat{l}_t(\hat{p}_e^t; \hat{p}^t)\right] \leq \frac{M}{\alpha}\sqrt{\frac{TA_{\max}}{8K}}(1 + 2\log E), \qquad e \in [E]. \tag{60}$$

The theorem then follows by combining (59) and (60).

We first prove (59). Since $0 \leq \beta \leq 2(T-1)$, we have

$$\min \mathcal{E} = \frac{K\alpha^3}{M^3 A_{\max}} \sqrt{\frac{2\log M}{T}} \leq \eta^\star \leq \frac{K\alpha^3}{M^3 A_{\max}} \sqrt{\frac{2\log M + 8\log(M/\alpha)(T-1)}{T}} \leq \max \mathcal{E},$$

where $\eta^\star$ is defined as in (58). Thus, there exists $\eta_e \in \mathcal{E}$, such that $\eta_e \leq \eta^\star \leq 2\eta_e$. Repeating the proof of (52), we can show that

$$\sum_{t=1}^T \hat{l}_t(\hat{p}_e^t; \hat{p}^t) - \sum_{t=1}^T \hat{l}_t(q^t; \hat{p}^t) \leq \frac{\log M}{\eta_e} + \frac{2\log(M/\alpha)}{\eta_e} \mathrm{TV}\left(q^{1:T}\right) + \frac{\eta_e M^6}{2K^2\alpha^6} \sum_{t=1}^T \left(a_{\max}^t\right)^2,$$

which then implies that

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^T \hat{l}_t(\hat{p}_e^t; \hat{p}^t) - \sum_{t=1}^T \hat{l}_t(q^t; \hat{p}^t)\right] &\leq \frac{\log M}{\eta_e} + \frac{2\log(M/\alpha)}{\eta_e} \mathbb{E}\left[\mathrm{TV}\left(q^{1:T}\right)\right] \\
&\quad + \frac{\eta_e M^6}{2K^2\alpha^6} \sum_{t=1}^T \mathbb{E}\left[\left(a_{\max}^t\right)^2\right] \\
&\leq \frac{\log M}{\eta_e} + \frac{2\beta\log(M/\alpha)}{\eta_e} + \frac{\eta_e M^6 T A_{\max}^2}{2K^2\alpha^6}.
\end{aligned}$$

Since $\eta^\star/2 \leq \eta_e \leq \eta^\star$, we further have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} \hat{l}_t(\hat{p}_e^t; \hat{p}^t) - \sum_{t=1}^{T} \hat{l}_t(q^t; \hat{p}^t)\right] \\
&\leq \frac{2\log M}{\eta^\star} + \frac{4\beta \log(M/\alpha)}{\eta^\star} + \frac{\eta^\star M^6 T A_{\max}^2}{2K^2\alpha^6} \\
&= \frac{3M^3 A_{\max}}{K\alpha^3}\sqrt{T\left[\frac{1}{2}\log M + \beta \log\left(M/\alpha\right)\right]}.
\end{aligned}$$

Now, (59) follows, since

$$\mathbb{E}\left[\sum_{t=1}^{T} \hat{l}_t(q^t; \hat{p}^t)\right] = \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{E}_{S^t}\left[\hat{l}_t(q^t; \hat{p}^t)\right]\right] = \sum_{t=1}^{T} \mathbb{E}\left[l_t(q^t)\right].$$

We prove (60) next. Let

$$\hat{L}_t^e = \sum_{s=1}^{t} \hat{l}_s(\hat{p}_e^s; \hat{p}^s) \quad e \in [E], t \in [T].$$

Recall the update for θ_e^t in Step 11 of Alg 5. We have

$$\theta_e^t = \frac{\theta_e^1 \exp\left(-\gamma \hat{L}_{t-1}^e\right)}{\sum_{b=1}^{E} \theta_b^1 \exp\left(-\gamma \hat{L}_{t-1}^b\right)}, \qquad t = 2, \ldots T.$$

Let $\Theta_t = \sum_{b=1}^{E} \theta_b^1 \exp\left\{-\gamma \hat{L}_t^b\right\}$. Then

$$\log \Theta_1 = \log\left(\sum_{b=1}^{E} \theta_b^1 \exp\left\{-\gamma \hat{L}_1^b\right\}\right)$$

and, for $t \geq 2$,

$$\begin{aligned}
\log\left(\frac{\Theta_t}{\Theta_{t-1}}\right) &= \log\left(\frac{\sum_{b=1}^{E} \theta_b^1 \exp\left\{-\gamma \hat{L}_{t-1}^b\right\} \exp\left\{-\gamma \hat{l}_t(\hat{p}_b^t; \hat{p}^t)\right\}}{\sum_{b=1}^{E} \theta_b^1 \exp\left\{-\gamma \hat{L}_{t-1}^b\right\}}\right) \\
&= \log\left(\sum_{b=1}^{E} \theta_b^t \exp\left\{-\gamma \hat{l}_t(\hat{p}_b^t; \hat{p}^t)\right\}\right).
\end{aligned}$$

We have

$$\begin{aligned}
\log \Theta_T &= \log \Theta_1 + \sum_{t=1}^{T} \log\left(\frac{\Theta_t}{\Theta_{t-1}}\right) \\
&= \sum_{t=1}^{T} \log\left(\sum_{b=1}^{E} \theta_b^t \exp\left\{-\gamma \hat{l}_t(\hat{p}_b^t; \hat{p}^t)\right\}\right) \\
&\leq \sum_{t=1}^{T} \left(-\gamma \sum_{b=1}^{E} \theta_b^t \hat{l}_t(\hat{p}_b^t; \hat{p}^t) + \frac{\gamma^2 M^2 a_{\max}^t}{8K\alpha^2}\right) && \text{(Lemma 18)} \\
&\leq -\gamma \sum_{t=1}^{T} \hat{l}_t(\hat{p}^t; \hat{p}^t) + \frac{\gamma^2 M^2 \left(\sum_{t=1}^{T} a_{\max}^t\right)}{8K\alpha^2} && \text{(Jensen's inequality)}
\end{aligned}$$

and

$$\begin{aligned}
\log\left(\Theta_T\right) &= \log\left(\sum_{b=1}^{E} \theta_b^1 \exp\left\{-\gamma \hat{L}_T^b\right\}\right) \\
&\geq \log\left(\max_{1\leq b\leq E} \theta_b^1 \exp\left\{-\gamma \hat{L}_T^b\right\}\right) = -\gamma \min_{1\leq b\leq E}\left\{\hat{L}_T^b + \frac{1}{\gamma}\log\frac{1}{\theta_b^1}\right\}.
\end{aligned}$$

Combining the last two displays, we have

$$-\gamma \min_{1\leq b\leq E}\left\{\hat{L}_T^b + \frac{1}{\gamma}\log\frac{1}{\theta_b^1}\right\} \leq -\gamma \sum_{t=1}^{T} \hat{l}_t(\hat{p}^t; \hat{p}^t) + \frac{\gamma^2 M^2 \left(\sum_{t=1}^{T} a_{\max}^t\right)}{8K\alpha^2},$$

which implies that

$$\sum_{t=1}^{T} \hat{l}_t(\hat{p}^t; \hat{p}^t) - \hat{L}_T^e \leq \frac{\gamma M^2 \left(\sum_{t=1}^{T} a_{\max}^t\right)}{8K\alpha^2} + \frac{1}{\gamma}\log\frac{1}{\theta_e^1} \leq \frac{\gamma M^2 T A_{\max}}{8K\alpha^2} + \frac{1}{\gamma}\log\frac{1}{\theta_e^1}, \quad e \in [E].$$

Taking expectation on both sides, we then have

$$\mathbb{E}\left[\sum_{t=1}^{T} \hat{l}_t(\hat{p}^t; \hat{p}^t) - \hat{L}_T^e\right] \leq \frac{\gamma M^2 T A_{\max}}{8K\alpha^2} + \frac{1}{\gamma}\log\frac{1}{\theta_e^1}.$$

Since $\theta_e^1 \geq \frac{1}{E^2}$, $\log 1/\theta_e^1 \leq 2\log E$. Let $\gamma = \sqrt{8K\alpha^2/(TM^2A_{\max})}$ to minimize the right hand side of the above inequality with $\log 1/\theta_e^1$ substituted by 1. Then

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \hat{l}_t(\hat{p}^t; \hat{p}^t) - \hat{L}_T^e\right] &= \mathbb{E}\left[\sum_{t=1}^{T} \hat{l}_t(\hat{p}^t; \hat{p}^t) - \sum_{t=1}^{T} \hat{l}_t(\hat{p}_e^t; \hat{p}^t)\right] \\
&\leq \frac{M}{\alpha}\sqrt{\frac{TA_{\max}}{8K}}\left(1 + 2\log E\right), \quad e \in [E].
\end{aligned}$$

B.7 Proof of Theorem 9

For Mini-batch SGD, following (35) and Theorem 8, we have

$$\begin{aligned}&\mathbb{E}\left[\left\|\nabla F\left(w^R\right)\right\|_2^2\right]\\&\leq \frac{2D^F}{T\mu}+L\mu\left(\frac{2\sigma^2}{KB\alpha}+\frac{2\zeta_T^2(\alpha,\beta)}{K}+\frac{MG^2}{K\alpha^3}\sqrt{\frac{\frac{1}{2}\log M+\beta\log\left(M/\alpha\right)}{T}}\right.\\&\qquad\left.+\frac{G}{\alpha}\sqrt{\frac{1}{8KT}}\left(1+2\log E\right)\right).\end{aligned}$$

Let

$$\frac{2D^F}{T\mu}=\frac{\mu LG}{\alpha}\sqrt{\frac{1}{8KT}}\left(1+2\log E\right),$$

we have

$$\mu=\sqrt{\frac{2\alpha D^F}{LG}}\left(\frac{8K}{T}\right)^{\frac{1}{2}}\sqrt{\frac{1}{1+2\log E}},$$

and

$$\frac{2D^F}{T\mu}+\frac{\mu LG}{\alpha}\sqrt{\frac{1}{8KT}}\left(1+2\log E\right)\leq\sqrt{\frac{2D^FLG}{\alpha}}\left(\frac{1}{8K}\right)^{\frac{1}{4}}\left(\frac{1}{T}\right)^{\frac{3}{4}}\sqrt{1+2\log E}.$$

Then follow the same argument as in the proof of Theorem 4, when

$$\begin{aligned}\mu=\min\Bigg\{&\frac{1}{L},\frac{1}{\sigma}\sqrt{\frac{D^FKB\alpha}{LT}},\frac{1}{\zeta_T(\alpha,\beta)}\sqrt{\frac{D^FK}{LT}},\\&\frac{\sqrt{D^FK}\alpha^{\frac{3}{2}}}{\sqrt{LM}T^{\frac{1}{4}}G\left(\frac{1}{2}\log M+\beta\log\left(M/\alpha\right)\right)^{\frac{1}{4}}},\sqrt{\frac{\alpha D^F}{LG}}\left(\frac{K}{T}\right)^{\frac{1}{2}}\sqrt{\frac{1}{1+2\log E}}\Bigg\},\end{aligned}$$

we have

$$\begin{aligned}&\mathbb{E}\left[\left\|\nabla F\left(w^R\right)\right\|^2\right]\\&\lesssim\frac{D^FL}{T}+\frac{\sigma\sqrt{D^FL}}{\sqrt{TKB\alpha}}+\frac{\zeta_T(\alpha,\beta)\sqrt{D^FL}}{\sqrt{TK}}+\frac{\sqrt{D^FL}M^{\frac{1}{2}}G}{T^{\frac{3}{4}}K^{\frac{1}{2}}\alpha^{\frac{3}{2}}}\left(\frac{1}{2}\log M+\beta\log\left(M/\alpha\right)\right)^{\frac{1}{4}}\\&\quad+\sqrt{\frac{D^FLG}{\alpha}}\left(\frac{1}{K}\right)^{\frac{1}{4}}\left(\frac{1}{T}\right)^{\frac{3}{4}}\sqrt{1+2\log E}.\end{aligned}$$

For FedAvg, following (46) and Theorem 8, we have

$$\begin{aligned}\mathbb{E}\left[\left\|\nabla F(w^R)\right\|_2^2\right]\leq&\frac{8D^F}{\tilde{\mu}BT}+8\left(2+\frac{1}{\alpha}\right)\tilde{\mu}^2B^2L^2\left(\zeta_{\text{unif}}^2+\frac{\sigma^2}{2B}\right)\\&+4\tilde{\mu}BL\left(\frac{2\zeta_T^2(\alpha,\beta)}{K}+\frac{\sigma^2}{KB\alpha}+\frac{3MG^2}{K\alpha^3}\sqrt{\frac{\frac{1}{2}\log M+\beta\log(M/\alpha)}{T}}\right.\\&\qquad\left.+\frac{G}{\alpha}\sqrt{\frac{1}{8KBT}}\left(1+2\log E\right)\right).\end{aligned}$$

Let

$$\frac{8D^F}{\tilde{\mu}BT} = 4\tilde{\mu}BL \times \frac{G}{\alpha}\sqrt{\frac{1}{8KBT}}\left(1+2\log E\right),$$

we have

$$\tilde{\mu} = \sqrt{\frac{2\alpha D^F}{LG\left(1+2\log E\right)}}\left(\frac{1}{B}\right)^{\frac{3}{4}}\left(\frac{8K}{T}\right)^{\frac{1}{4}},$$

and

$$\frac{8D^F}{\tilde{\mu}BT} + 4\tilde{\mu}BL \times \frac{G}{\alpha}\sqrt{\frac{1}{8KBT}}\left(1+2\log E\right) = 8^{\frac{3}{4}}\sqrt{2}\sqrt{\frac{D^F LG}{\alpha}}\left(\frac{1}{KB}\right)^{\frac{1}{4}}\left(\frac{1}{T}\right)^{\frac{3}{4}}\sqrt{1+2\log E}.$$

Then follow the same argument as in the proof of Theorem 5, when $\mu \geq 1$ and

$$\begin{aligned}\tilde{\mu} = \min\Bigg\{ & \frac{1}{4BL}\sqrt{\frac{1}{1+1/(2\alpha)}}, \frac{(D^F)^{\frac{1}{3}}}{\left(2+\frac{1}{\alpha}\right)^{\frac{1}{3}} BL^{\frac{2}{3}}\left(\zeta_{\text{unif}}^2 + \frac{\sigma^2}{2B}\right)^{\frac{1}{3}} T^{\frac{1}{3}}}, \\ & \frac{\sqrt{2D^F}}{B\sqrt{L}\sqrt{\frac{2\zeta_T^2(\alpha,\beta)}{K} + \frac{\sigma^2}{KB\alpha} + \sqrt{\frac{\frac{1}{2}\log M + \beta\log(M/\alpha)}{T}}}\sqrt{T}}, \\ & \sqrt{\frac{2\alpha D^F}{LG\left(1+2\log E\right)}}\left(\frac{1}{B}\right)^{\frac{3}{4}}\left(\frac{8K}{T}\right)^{\frac{1}{4}}\Bigg\},\end{aligned}$$

we have

$$\begin{aligned}\mathbb{E}\left[\left\|\nabla F(w^R)\right\|_2^2\right] \lesssim & \frac{D^F L\sqrt{1+\frac{1}{2\alpha}}}{T} + \frac{\left(2+\frac{1}{\alpha}\right)^{\frac{1}{3}}\left(D^F L\right)^{\frac{2}{3}}\zeta_{\text{unif}}^{\frac{2}{3}}}{T^{\frac{2}{3}}} + \frac{\left(2+\frac{1}{\alpha}\right)^{\frac{1}{3}}\left(D^F L\right)^{\frac{2}{3}}\sigma^{\frac{2}{3}}}{B^{\frac{1}{3}}T^{\frac{2}{3}}} \\ & + \frac{\sqrt{D^F L}\zeta_T(\alpha,\beta)}{\sqrt{TK}} + \frac{\sqrt{D^F L}\sigma}{\sqrt{TKB\alpha}} + \frac{\sqrt{D^F L}}{\sqrt{T}}\left(\frac{\frac{1}{2}\log M + \beta\log(M/\alpha)}{T}\right)^{\frac{1}{4}} \\ & + \sqrt{\frac{D^F LG}{\alpha}}\left(\frac{1}{KB}\right)^{\frac{1}{4}}\left(\frac{1}{T}\right)^{\frac{3}{4}}\sqrt{1+2\log E}.\end{aligned}$$

B.8 Proof of Theorem 10

Recall that $T_b = 2^{b-1}$. Let $B = \lceil \log_2(T+1) \rceil$, we then have $T_B \leq T \leq T_{B+1} - 1$, which implies that $1 \leq T - T_B + 1 \leq T_{B+1} - T_B = 2^B$.

Note that $\hat{p}^{T_b}$ is reinitialized as the uniform distribution. Let

$$\text{D-Regret}_b = \mathbb{E}\left[\sum_{t=T_b}^{T_{b+1}-1} l_t(\hat{p}^t) - \sum_{t=T_b}^{T_{b+1}-1} l_t(q^t)\right].$$

Similar to the proof of (59) and (60), we have

$$\begin{aligned}
&\text{D-Regret}_b \\
&\leq \frac{3M^3 A_{\max}}{K\alpha^3}\sqrt{\left[\frac{1}{2}\log M + \log\left(M/\alpha\right)\mathbb{E}\left[\mathrm{TV}\left(q^{T_b:(T_{b+1}-1)}\right)\right]\right](T_{b+1}-T_b)} \\
&\qquad + \frac{M}{\alpha}\sqrt{\frac{(T_{b+1}-T_b)A_{\max}}{8K}}(1+2\log E_b) \\
&= \frac{3M^3 A_{\max}}{K\alpha^3}\sqrt{\left[\frac{1}{2}\log M + \log\left(M/\alpha\right)\mathbb{E}\left[\mathrm{TV}\left(q^{T_b:(T_{b+1}-1)}\right)\right]\right]}\left(\sqrt{2}\right)^{b-1} \\
&\qquad + \frac{M}{\alpha}\sqrt{\frac{A_{\max}}{8K}}(1+2\log E_b)\left(\sqrt{2}\right)^{b-1} \\
&\leq \frac{3M^3 A_{\max}}{K\alpha^3}\sqrt{\left[\frac{1}{2}\log M + \log\left(M/\alpha\right)\beta\right]}\left(\sqrt{2}\right)^{b-1} + \frac{M}{\alpha}\sqrt{\frac{A_{\max}}{8K}}(1+2\log E)\left(\sqrt{2}\right)^{b-1},
\end{aligned}$$

where E_b is defined in (28) and E is defined in (23). We can similarly obtain

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=T_B}^{T} l_t(\hat{p}^t) - \sum_{t=T_B}^{T} l_t(q^t)\right] &\leq \frac{3M^3 A_{\max}}{K\alpha^3}\sqrt{\left[\frac{1}{2}\log M + \log\left(M/\alpha\right)\beta\right]}\left(\sqrt{2}\right)^{B} \\
&\qquad + \frac{M}{\alpha}\sqrt{\frac{A_{\max}}{8K}}(1+2\log E)\left(\sqrt{2}\right)^{B}.
\end{aligned}$$

Combine the last two displays, we have

$$\begin{aligned}
\text{D-Regret}_T(q^{1:T}) &= \mathbb{E}\left[\sum_{t=1}^{T} l_t(\hat{p}^t) - \sum_{t=1}^{T} l_t(q^t)\right] \\
&= \sum_{b=1}^{B-1}\mathbb{E}\left[\sum_{t=T_b}^{T_{b+1}-1} l_t(\hat{p}^t) - \sum_{t=T_b}^{T_{b+1}-1} l_t(q^t)\right] + \mathbb{E}\left[\sum_{t=T_B}^{T} l_t(\hat{p}^t) - \sum_{t=T_B}^{T} l_t(q^t)\right] \\
&\leq \frac{3M^3 A_{\max}}{K\alpha^3}\sqrt{\left[\frac{1}{2}\log M + \log\left(M/\alpha\right)\beta\right]}\left(\sum_{b=1}^{B-1}\left(\sqrt{2}\right)^{b-1} + \left(\sqrt{2}\right)^{B}\right) \\
&\quad + \frac{M}{\alpha}\sqrt{\frac{A_{\max}}{8K}}(1+2\log E)\left(\sum_{b=1}^{B-1}\left(\sqrt{2}\right)^{b-1} + \left(\sqrt{2}\right)^{B}\right) \\
&= \frac{3M^3 A_{\max}}{K\alpha^3}\sqrt{\left[\frac{1}{2}\log M + \log\left(M/\alpha\right)\beta\right]}\frac{\left(\sqrt{2}\right)^{B+1}-\sqrt{2}}{\sqrt{2}-1} \\
&\quad + \frac{M}{\alpha}\sqrt{\frac{A_{\max}}{8K}}(1+2\log E)\frac{\left(\sqrt{2}\right)^{B+1}-\sqrt{2}}{\sqrt{2}-1} \\
&\leq \frac{3M^3 A_{\max}}{K\alpha^3}\sqrt{\left[\frac{1}{2}\log M + \log\left(M/\alpha\right)\beta\right]}\frac{2}{\sqrt{2}-1}\sqrt{T}
\end{aligned}$$

$$+ \frac{M}{\alpha}\sqrt{\frac{A_{\max}}{8K}}(1+2\log E)\frac{2}{\sqrt{2}-1}\sqrt{T}.$$

Appendix C. Useful Lemmas

Lemma 14 *Suppose that f is a differentiable convex function defined on $domf$, and $\mathcal{X} \subseteq domf$ is a closed convex set. Then x is the minimizer of f on $\mathcal{X}$ if and only if*

$$\nabla f(x)^\top (y-x) \geq 0 \quad \textit{for all } y \in \mathcal{X}.$$

Proof See Section 4.2.3 of Boyd et al. (2004). ■

Lemma 15 *Let $f : \mathcal{W} \to \mathbb{R}$ be defined on $\mathcal{W} \subseteq \mathbb{R}^d$, where $\mathcal{W}$ is a convex set. Suppose that f is continuously differentiable and first-order L-smooth, that is,*

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x-y\| \quad \textit{for all } x, y \in \mathcal{W},$$

then we have

$$|f(y) - f(x) - \langle \nabla f(x), y-x\rangle| \leq \frac{L}{2}\|x-y\|_2^2 \quad \textit{for all } x, y \in \mathcal{W}.$$

Proof *We follow Theorem 2.1.5 of Nesterov et al. (2018). Note that*

$$\begin{aligned}
&|f(y) - f(x) - \langle \nabla f(x), y-x\rangle| \\
&= \left| \int_0^1 \langle \nabla f\left(x + t(y-x)\right), y-x\rangle \, dt - \langle \nabla f(x), y-x\rangle \right| \\
&= \left| \int_0^1 \langle \nabla f\left(x + t(y-x)\right) - \nabla f(x), y-x\rangle \, dt \right| \\
&\leq \int_0^1 \|\nabla f\left(x + t(y-x)\right) - \nabla f(x)\| \, \|y-x\| \, dt \\
&\leq \int_0^1 Lt\|y-x\|_2^2 \, dt \\
&= \frac{L}{2}\|y-x\|_2^2,
\end{aligned}$$

which then implies the conclusion. ■

Lemma 16 (Relaxed Triangle Inequality) *This is Lemma 3 of Karimireddy et al. (2020b). Let $\{v_1, \ldots, v_\tau\}$ be τ vectors in $\mathbb{R}^d$. Then the following are true:*

$$\|v_i + v_j\|^2 \leq (1+a)\|v_i\|^2 + \left(1 + \frac{1}{a}\right)\|v_j\|^2 \quad \textit{for any } a > 0,$$

$$\left\| \sum_{i=1}^{\tau} v_i \right\|_2^2 \leq \tau \sum_{i=1}^{\tau} \|v_i\|_2^2.$$

Proof *See Lemma 3 of Karimireddy et al. (2020b).* ■

Lemma 17 *For $q \in \mathcal{P}_{M-1}$ we have $D_\Phi(q \,\|\, p^{\text{unif}}) \le \log M$, where Φ is the unnormalized negative entropy.*

Proof Since $\Phi(q) = \sum_{m=1}^M q_m(\log q_m - 1) \le 0$, $\Phi(p^{\text{unif}}) = -\log M$, and

$$\langle \nabla\Phi(p^{\text{unif}}), q - p^{\text{unif}} \rangle = \sum_{m=1}^M (q_m - \frac{1}{M}) \log \frac{1}{M} = 0,$$

we have $D_\Phi(q \,\|\, p) \le \log M$. ■

Lemma 18 (Hoeffding's Inequality) *Let X be a random variable with $a \le X \le b$ for $a, b \in \mathbb{R}$. Then for all $s \in \mathbb{R}$, we have*

$$\log \mathbb{E}\left[e^{sX}\right] \le s\mathbb{E}[X] + \frac{s^2(b-a)^2}{8}.$$

Proof See Section 2 of Wainwright (2019). ■

Lemma 19 (Based on Exercise 26.12 of Lattimore and Szepesvári (2020)) *Let $\alpha \in [0,1]$, $\mathcal{A} = \mathcal{P}_{M-1} \cap [\alpha/M, 1]^M$, $\mathcal{D} = [0,\infty)^M$, and Φ is the unnormalised entropy on $\mathcal{D}$. For $y \in [0,\infty)^M$, let $x = \arg\min_{v\in\mathcal{A}} D_\Phi(v\|y)$. Suppose $y_1 \le y_2 \le \cdots \le y_M$. Let $m^\star$ be the smallest value such that*

$$y_{m^\star}\left(1 - \frac{m^\star - 1}{M}\alpha\right) > \frac{\alpha}{M}\sum_{m=m^\star}^M y_m.$$

Then

$$x_m = \begin{cases} \frac{\alpha}{M} & \text{if } m < m^\star \\ \frac{(1 - \frac{m^\star - 1}{M}\alpha) y_m}{\sum_{n=m^\star}^M y_n} & \text{otherwise.} \end{cases}$$

Proof Consider the following constrained optimization problem:

$$\begin{aligned} \min_{u\in[0,\infty)^M} & \sum_{m=1}^M u_m \log \frac{u_m}{y_m}, \\ \text{s.t.} & \sum_{m=1}^M u_m = 1, \\ & u_m \ge \frac{\alpha}{M}, \qquad m \in [M]. \end{aligned}$$

Since x is the solution to this problem, by the optimality condition, there exists $\lambda, \nu_1, \ldots, \nu_M \in \mathbb{R}$ such that

$$\log \frac{x_m}{y_m} + 1 - \lambda - \nu_m = 0, \qquad m \in [M], \tag{61}$$

$$\sum_{m=1}^{M} x_m = 1, \tag{62}$$

$$x_m - \frac{\alpha}{M} \geq 0, \qquad m \in [M], \tag{63}$$

$$\nu_m \geq 0, \qquad m \in [M], \tag{64}$$

$$\nu_m \left(x_m - \frac{\alpha}{M}\right) = 0, \qquad m \in [M]. \tag{65}$$

By (61), we have $x_m = y_m \exp(-1 + \lambda + \nu_m)$. By (64) and (65), when $x_m = \alpha/M$, we have $x_m = y_m \exp(-1 + \lambda + \nu_m) \geq y_m \exp(-1 + \lambda)$; when $x_m > \alpha/M$, we have $x_m = y_m \exp(-1 + \lambda)$. Assume that $x_1 = \cdots = x_{m^\star - 1} = \alpha/M < x_{m^\star} \leq \cdots \leq x_M$. Then

$$1 = \sum_{m=1}^{M} x_m = (m^\star - 1)\frac{\alpha}{M} + \exp(-1 + \lambda) \cdot \sum_{m=m^\star}^{M} y_m,$$

which implies that

$$\exp(-1 + \lambda) = \frac{1 - (m^\star - 1)\frac{\alpha}{M}}{\sum_{m=m^\star}^{M} y_m}. \tag{66}$$

Thus, we have

$$x_{m^\star} = y_{m^\star} \exp(-1 + \lambda) = y_{m^\star} \frac{1 - (m^\star - 1)\frac{\alpha}{M}}{\sum_{m=m^\star}^{M} y_m} > \frac{\alpha}{M},$$

which implies that

$$y_{m^\star} \left(1 - \frac{m^\star - 1}{M}\alpha\right) > \frac{\alpha}{M} \sum_{m=m^\star}^{M} y_m. \tag{67}$$

To complete the proof, we then only need to show that

$$y_{m'} \left(1 - \frac{m' - 1}{M}\alpha\right) \leq \frac{\alpha}{M} \sum_{m=m'}^{M} y_m \tag{68}$$

for all $1 \leq m' \leq m^\star - 1$. The result then follows from (67) and (68). To prove(68), recall that for any $1 \leq m' \leq m^\star - 1$, we have $\alpha/M = y_{m'} \exp(-1 + \lambda + \nu_{m'})$, and because $y_1 \leq \cdots \leq y_M$, we have $\nu_1 \geq \cdots \geq \nu_{m^\star - 1}$. This way, we have

$$\left(m^\star - m'\right)\frac{\alpha}{M} = \sum_{m=m'}^{m^\star - 1} y_m \exp\left(-1 + \lambda + \nu_m\right) \leq \exp\left(-1 + \lambda + \nu_{m'}\right) \sum_{m=m'}^{m^\star - 1} y_m. \tag{69}$$

On the other hand, by (66), we have

$$1 - (m^\star - 1)\frac{\alpha}{M} = \exp(-1 + \lambda) \sum_{m=m^\star}^{M} y_m \leq \exp\left(-1 + \lambda + \nu_{m'}\right) \sum_{m=m^\star}^{M} y_m. \tag{70}$$

Combining (69) and (70), we have

$$\begin{aligned}\frac{1-(m'-1)\frac{\alpha}{M}}{\sum_{m=m'}^{M} y_m} &= \frac{1-(m^\star-1)\frac{\alpha}{M}+(m^\star-m')\frac{\alpha}{M}}{\sum_{m=m'}^{M} y_m} \\ &\le \frac{\exp\left(-1+\lambda+\nu_{m'}\right)\left(\sum_{m=m'}^{m^\star-1} y_m+\sum_{m=m^\star}^{M} y_m\right)}{\sum_{m=m'}^{M} y_m} \\ &= \exp\left(-1+\lambda+\nu_{m'}\right) \\ &= \frac{\frac{\alpha}{M}}{y_{m'}},\end{aligned}$$

which then implies (68). ■

Appendix D. Additional Future Directions

In this section, we discuss two additional future directions. In Section D.1, we discuss the design of sampling algorithms for minimizing personalized FL objectives. Besides, in Section D.2, we discuss sampling with physical constraint in FL system.

D.1 Client Sampling with Personalized FL Objective

Data distributions across clients are often heterogeneous. Personalized FL has emerged as one effective way to handle such heterogeneity (Kulkarni et al., 2020). Hanzely et al. (2023) illustrated how many existing approaches to personalization can be studied through a unified framework, and, in this section, we discuss a natural extension of Adaptive-OSMD Sampler to this personalized objective. Specifically, we study the following optimization problem

$$\min_{w,\beta} F(w,\beta) \coloneqq \sum_{m=1}^{M} \lambda_m \phi(w,\beta_m;\mathcal{D}_m), \tag{71}$$

where $w \in \mathbb{R}^{d_0}$ corresponds to the shared parameter and $\beta = (\beta_1,\ldots,\beta_M)$ with $\beta_m \in \mathbb{R}^{d_m}$ corresponds to the local parameters. The objective in (71) coves a wide range of personalized federated learning problems (Hanzely et al., 2023). We further generalize the approach and study the following bilevel optimization problem:

$$\begin{aligned}\min_{w}\ h(w) &\coloneqq \sum_{m=1}^{M} \lambda_m F_m(w,\hat{\beta}_m(w)) \coloneqq \sum_{m=1}^{M} \lambda_m \phi(w,\hat{\beta}_m(w);\mathcal{D}_m) \\ \text{subject to}\quad \hat{\beta}_m(w) &= \arg\min_{\beta_m} G_m(w,\beta_m) \coloneqq \phi(w,\beta_m;\bar{\mathcal{D}}_m).\end{aligned} \tag{72}$$

When $\mathcal{D}_m = \bar{\mathcal{D}}_m$, then (72) recovers (71). When $\bar{\mathcal{D}}_m \neq \mathcal{D}_m$, we then optimize the shared and local parameters on different data sets, which may prevent overfitting. The formulation in (72) is closely related to the implicit MAML (Rajeswaran et al., 2019).

In the following, we use ∇_w to denote a partial derivative with respect to w with β_m fixed, ∇_{β_m} to denote a partial derivative with respect to β_m with w fixed, and ∇

to denote a derivative with respect to w where $\beta_m(w)$ is treated as a function of w. Let $\nabla^2_{\beta_m\beta_m^\top} G_m(w,\beta_m) \in \mathbb{R}^{d_m\times d_m}$ be the Hessian matrix of G_m with respect to β_m where w is fixed, and $\nabla^2_{w\beta_m^\top} G_m(w,\beta_m) \in \mathbb{R}^{d_0\times d_m}$ be the Hessian matrix of G_m with respect to w and β_m, that is,

$$\left[\nabla^2_{\beta_m\beta_m^\top} G_m(w,\beta_m)\right]_{i,j} = \frac{\partial G_m(w,\beta_m)}{\partial \beta_{m,i}\beta_{m,j}} \quad \text{for all } i,j = 1,2,\ldots,d_m,$$
$$\left[\nabla^2_{w\beta_m^\top} G_m(w,\beta_m)\right]_{i,j} = \frac{\partial G_m(w,\beta_m)}{\partial w_i \beta_{m,j}} \quad \text{for all } i = 1,2,\ldots,d_0, j = 1,2,\ldots,d_m.$$

By the implicit function theorem, we have

$$\nabla h(w) = \underbrace{\frac{1}{M}\sum_{m=1}^M \lambda_m \nabla_1 F_m(w,\hat\beta_m(w))}_{\nabla_1 h(w)} + \underbrace{\frac{1}{M}\sum_{m=1}^M \lambda_m \nabla_2 F_m(w,\hat\beta_m(w))}_{\nabla_2 h(w)} \tag{73}$$

where

$$\nabla_1 F_m(w,\hat\beta_m(w)) \coloneqq \nabla_w F_m(w,\hat\beta_m(w)),$$
$$\nabla_2 F_m(w,\hat\beta_m(w)) \coloneqq -\nabla^2_{w\beta_m^\top} G_m(w,\hat\beta_m(w)) \left[\nabla^2_{\beta_m\beta_m^\top} G_m(w,\hat\beta_m(w))\right]^{-1} \nabla_{\beta_m} F_m(w,\hat\beta_m(w)).$$

There are two parts to $\nabla h(w^t)$ and, therefore, instead of choosing a single subset of clients for computing both parts, we decouple S^t into two subsets S_1^t and S_2^t, $S^t = S_1^t \cup S_2^t$. We use clients in S_1^t to compute local updates of the first part, and clients in S_2^t to compute the local updates of the second part. To get an estimate of $\nabla h(w)$, we can estimate $\nabla_1 h(w)$ and $\nabla_2 h(w)$ separably and then combine. Assume that $g_{1,m}^t$ is an estimate of $\nabla_1 F_m(w,\hat\beta_m(w))$ and $g_{2,m}^t$ is an estimate of $\nabla_2 F_m(w,\hat\beta_m(w))$, we can then construct estimates of $\nabla_1 h(w)$ and $\nabla_2 h(w)$ as

$$g_1^t = \frac{1}{K_1}\sum_{m\in S_1^t} \lambda_m \frac{g_{1,m}^t}{p_{1,m}^t}, \qquad g_2^t = \frac{1}{K_2}\sum_{m\in S_2^t} \lambda_m \frac{g_{2,m}^t}{p_{2,m}^t},$$

where $K_1 = |S_1^t|$ and $K_2 = |S_2^t|$. Then $g^t = g_1^t + g_2^t$ is an estimate of $\nabla h(w)$.

We design p_1^t and p_2^t to choose S_1^t and S_2^t by minimizing the variance of the gradients. Note that

$$\min_{p_1^t} \mathbb{E}_1\left[\mathbb{E}_{S_1^t}\left[\left\|g_1^t - \nabla_1 h(w^t)\right\|^2\right]\right] + \min_{p_2^t} \mathbb{E}_2\left[\mathbb{E}_{S_2^t}\left[\left\|g_2^t - \nabla_2 h(w^t)\right\|^2\right]\right]$$
$$\leq \min_{p_1^t = p_2^t = p^t} \mathbb{E}\left[\mathbb{E}_{S^t}\left[\left\|g_1^t - \nabla_1 h(w^t) + g_2^t - \nabla_2 h(w^t)\right\|^2\right]\right],$$

so that the decomposition allows us to better minimize the variance. We term this approach as *doubly variance reduction for personalized Federated Learning*. The first part minimizes the variance of updates to the shared global parameter, when the best local parameters are fixed; and the second part minimizes the variance of updates to local parameters, when the

global part is fixed. While these two parts are related, any given machine will have different contributions to these two tasks.

Adaptive-OSMD Sampler can be used to minimize the variance for both parts of the gradient. We note that this is a heuristic approach to solving the client sampling problem when minimizing a personalized FL objective. Personalized FL objectives have additional structures that should be used to design more efficient sampling strategies. Furthermore, designing sampling strategies that improve the statistical performance of trained models, rather than improving computational speed, is important in the heterogeneous setting. Addressing these questions is an important area for future research.

D.2 Sampling with Physical Constraint in FL System

In this paper, we assume that all clients are available in each round. However, in practical FL applications, a subset of the clients may be inactive due to physical constraints, thus we have to assign zero probabilities to them. In this section, we propose a simple extension of our proposed sampling method to such case.

Specifically, denote the subset of clients that are active at the beginning of round t as $I^t \subseteq [M]$. If we have $|I^t| \leq K$, we can then use all clients in I^t to make updates in round t; otherwise, we would like to choose a smaller subset $S^t \subseteq I^t$ to participate. This can be achieved by rescaling the output sampling distribution of any of our proposed methods, which we denote as $\hat{p}^t$. We let $\tilde{p}^t_m = \hat{p}^t_m / (\sum_{i \in I^t} \hat{p}^t_i)$ and $\tilde{p}^t_m = 0$ for all $m \notin I^t$. We can then use $\tilde{p}^t_m$ to choose S^t from I^t.

However, analyzing such a method in terms of convergence and regret guarantee is highly non-trivial. Typically, for general active clients sequence $\{I^t\}_{t=1}^T$, the optimization algorithms are not guaranteed to converge even if we involve all clients in I^t in each round. This can happen, for example, if a client is active for only once in the whole training process. Thus, to ensure convergence, we need additional assumptions about $\{I^t\}_{t=1}^T$. Moreover, deriving regret bound is also very challenging, as assigning zero probability to any client will make the variance-reduction loss unbounded and thus the regret can be arbitrarily large. To achieve such theoretical result, one may need to appropriately redefine the regret concept. Such an analysis is beyond the scope of this paper and we leave it for future research.

Appendix E. Application of OSMD Sampler on SCAFFOLD

We explore the potential of implementing our proposed sampler in more sophisticated federated learning optimization techniques. Beyond merely utilizing gradients, numerous state-of-the-art algorithms incorporate additional auxiliary variables to achieve a faster rate of convergence. These auxiliary variables complicate the process of constructing the appropriate surrogate variance reduction loss as outlined in (4), thus making the extension of our proposed sampling method to these algorithms challenging. Although a thorough exploration in this area is essential for future studies, in this section, we aim to provide insight by focusing on one of the most widely adopted state-of-the-art algorithms, SCAFFOLD (Karimireddy et al., 2020b), as a case in point. We propose a straightforward strategy to incorporate our OSMD sampler into SCAFFOLD. Through simulation experiments, we validate our approach, with the results offering encouraging evidence of its effectiveness.

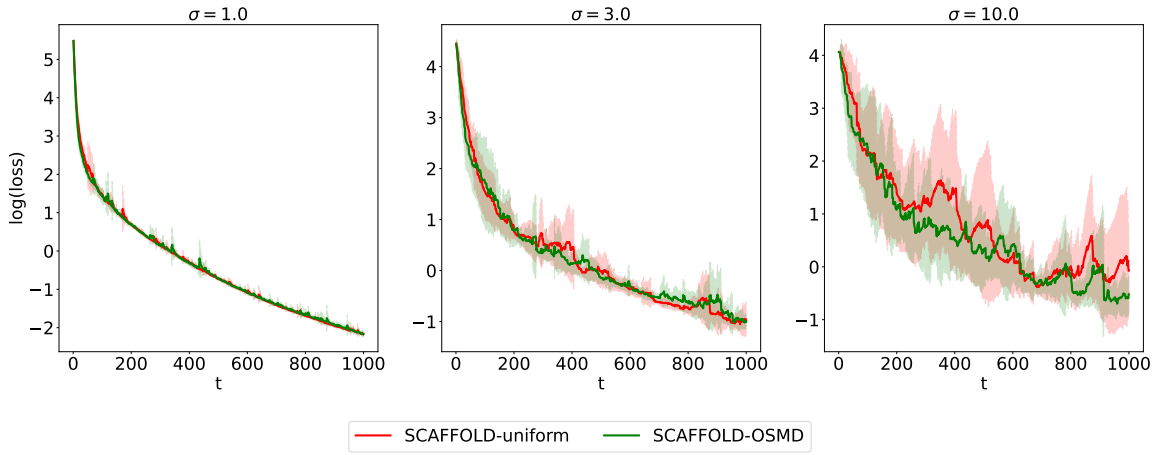

Figure 8: The training loss is compared between SCAFFOLD with uniform sampling and SCAFFOLD with the OSMD sampler. Solid lines represent the mean values, while shaded regions indicate mean±standard deviation across independent runs.

In addition to utilizing gradients, SCAFFOLD incorporates control variates on both the server and client sides to mitigate client-side heterogeneity. When implementing the OSMD sampler as described in Algorithm 1 within the SCAFFOLD framework, it is essential to define environment feedback $\{a_m^t\}_{m\in S^t}$. For our experiments, we chose to set $a_m^t = \lambda_m^2 \left\|w_m^{t+} - w^t\right\|^2$, where w^t denotes the global model parameter at the start of round t, and w_m^{t+} represents the updated local model parameter for client m after performing local mini-batch SGD during round t. The learning rate for the OSMD sampler was set as 10^{-3}. For all additional aspects of the SCAFFOLD algorithm, such as hyperparameter selection and tuning methodology, we adhered to the guidelines in the original paper.

Our experimental setup mirrors that of Section 7.1. Figure 8 presents the results, indicating that the OSMD sampler performs marginally better when the heterogeneity is high. This observation implies that utilizing adaptive sampling could potentially enhance SCAFFOLD's effectiveness. A more detailed investigation into this possibility is reserved for subsequent studies.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *ArXiv Preprint ArXiv:1603.04467*, 2016.

Zalan Borsos, Andreas Krause, and Kfir Y. Levy. Online variance reduction for stochastic optimization. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Conference On Learning Theory (COLT)*, 2018.

Zalán Borsos, Sebastian Curi, Kfir Yehuda Levy, and Andreas Krause. Online variance reduction with mixtures. In *International Conference on Machine Learning (ICML)*, 2019.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex Optimization*. Cambridge university press, 2004.

Sebastian Caldas, Sai Meher Karthik Duddu, Peter Wu, Tian Li, Jakub Konečnỳ, H Brendan McMahan, Virginia Smith, and Ameet Talwalkar. Leaf: A benchmark for federated settings. *ArXiv Preprint ArXiv:1812.01097*, 2018.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

Zachary Charles and Jakub Konečnỳ. On the outsized importance of learning rates in local update methods. *ArXiv preprint ArXiv:2007.00878*, 2020.

Wenlin Chen, Samuel Horváth, and Peter Richtárik. Optimal client sampling for federated learning. *Transactions on Machine Learning Research*, 2022.

Yae Jee Cho, Samarth Gupta, Gauri Joshi, and Osman Yagan. Bandit-based communication-efficient client selection strategies for federated learning. In *Asilomar Conference on Signals, Systems, and Computers (ACSCC)*, 2020a.

Yae Jee Cho, Jianyu Wang, and Gauri Joshi. Client selection in federated learning: Convergence analysis and power-of-choice selection strategies. *ArXiv Preprint ArXiv:2010.01243*, 2020b.

Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep learning for classical japanese literature. *ArXiv Preprint ArXiv:1812.01718*, 2018.

Dominik Csiba and Peter Richtárik. Importance sampling for minibatches. *Journal of Machine Learning Research*, 19(27):1–21, 2018.

Amit Daniely, Alon Gonen, and Shai Shalev-Shwartz. Strongly adaptive online learning. In *International Conference on Machine Learning (ICML)*, 2015.

Cynthia Dwork. Differential privacy: A survey of results. In *Theory and Applications of Models of Computation (TAMC)*, 2008.

Vladmir Estivill-Castro and Derick Wood. A survey of adaptive sorting algorithms. *ACM Computing Surveys (CSUR)*, 24(4):441–476, 1992.

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

Siddharth Gopal. Adaptive sampling for SGD by exploiting side information. In *International Conference on Machine Learning (ICML)*, 2016.

Eric C Hall and Rebecca M Willett. Online convex optimization in dynamic environments. *IEEE Journal of Selected Topics in Signal Processing*, 9(4):647–662, 2015.

Ayoub El Hanchi and David A. Stephens. Adaptive importance sampling for finite-sum optimization and sampling with decreasing step-sizes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Filip Hanzely, Boxin Zhao, and mladen kolar. Personalized federated learning: A unified framework and universal optimization techniques. *Transactions on Machine Learning Research*, 2023.

Elad Hazan. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

Tyler B. Johnson and Carlos Guestrin. Training deep models faster with robust, approximate importance sampling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14(1–2):1–210, 2021.

Sai Praneeth Karimireddy, Martin Jaggi, Satyen Kale, Mehryar Mohri, Sashank J Reddi, Sebastian U Stich, and Ananda Theertha Suresh. Mime: Mimicking centralized stochastic algorithms in federated learning. *ArXiv Preprint ArXiv:2008.03606*, 2020a.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J. Reddi, Sebastian U. Stich, and Ananda Theertha Suresh. SCAFFOLD: stochastic controlled averaging for federated learning. In *International Conference on Machine Learning (ICML)*, 2020b.

Taehyeon Kim, Sangmin Bae, Jin-woo Lee, and Seyoung Yun. Accurate and fast federated learning via combinatorial multi-armed bandits. *ArXiv Preprint ArXiv:2012.03270*, 2020.

Viraj Kulkarni, Milind Kulkarni, and Aniruddha Pant. Survey of personalization techniques for federated learning. In *World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4)*, 2020.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Yann LeCun and Corinna Cortes. MNIST handwritten digit database. *CoRR*, 2010.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

Hongseok Namkoong, Aman Sinha, Steve Yadlowsky, and John C. Duchi. Adaptive sampling probabilities for non-smooth optimization. In *International Conference on Machine Learning (ICML)*, 2017.

Deanna Needell, Nathan Srebro, and Rachel Ward. Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm. *Mathematical Programming*, 1(155): 549–573, 2016.

Yurii Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

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

Kumar Kshitij Patel, Lingxiao Wang, Blake E Woodworth, Brian Bullins, and Nati Srebro. Towards optimal communication complexity in distributed non-convex optimization. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Dmytro Perekrestenko, Volkan Cevher, and Martin Jaggi. Faster coordinate descent via adaptive importance sampling. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

Zhe Qu, Rui Duan, Lixing Chen, Jie Xu, Zhuo Lu, and Yao Liu. Context-aware online client selection for hierarchical federated learning. *IEEE Transactions on Parallel and Distributed Systems*, 33(12):4353–4367, 2022.

Aravind Rajeswaran, Chelsea Finn, Sham M. Kakade, and Sergey Levine. Meta-learning with implicit gradients. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Monica Ribero and Haris Vikalo. Communication-efficient federated learning via optimal client sampling. *ArXiv Preprint ArXiv:2007.15197*, 2020.

Farnood Salehi, L Elisa Celis, and Patrick Thiran. Stochastic optimization with bandit sampling. *ArXiv Preprint ArXiv:1708.02544*, 2017.

Farnood Salehi, Patrick Thiran, and L. Elisa Celis. Coordinate descent with bandit sampling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Sebastian U. Stich, Anant Raj, and Martin Jaggi. Safe adaptive importance sampling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Tim van Erven and Wouter M. Koolen. Metagrad: Multiple learning rates in online learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Hao Wang, Zakhary Kaplan, Di Niu, and Baochun Li. Optimizing federated learning on non-iid data with reinforcement learning. In *IEEE Conference on Computer Communications (INFOCOM)*, 2020.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *ArXiv Preprint ArXiv:1708.07747*, 2017.

Miao Yang, Ximin Wang, Hongbin Zhu, Haifeng Wang, and Hua Qian. Federated learning with class imbalance reduction. In *European Signal Processing Conference (EUSIPCO)*, 2021.

Tianbao Yang, Lijun Zhang, Rong Jin, and Jinfeng Yi. Tracking slowly moving clairvoyant: Optimal dynamic regret of online learning with true and noisy gradient. In *International Conference on Machine Learning (ICML)*, 2016.

Lijun Zhang, Shiyin Lu, and Zhi-Hua Zhou. Adaptive online learning in dynamic environments. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Boxin Zhao, Boxiang Lyu, and Mladen Kolar. L-svrg and l-katyusha with adaptive sampling. *Transactions on Machine Learning Research*, 2023.

Peilin Zhao and Tong Zhang. Stochastic optimization with importance sampling for regularized loss minimization. In *International Conference on Machine Learning (ICML)*, 2015.

Zeyuan Allen Zhu, Zheng Qu, Peter Richtárik, and Yang Yuan. Even faster accelerated coordinate descent using non-uniform sampling. In *International Conference on Machine Learning (ICML)*, 2016.