Technical Appendix

Guocheng Liao, Bing Luo, Yutong Feng, and Meng Zhang

1 Proof of Proposition 1

Assumption 1. *$F_1, \cdots, F_N$ are all L-smooth: for all $\mathbf{v}$ and $\mathbf{w}$, $F_n(\mathbf{v}) \leq F_n(\mathbf{w}) + (\mathbf{v} - \mathbf{w})^T \nabla F_n(\mathbf{w}) + \frac{L}{2}\|\mathbf{v} - \mathbf{w}\|_2^2$*

Assumption 2. *$F_1, \cdots, F_N$ are all μ-strongly convex: for all $\mathbf{v}$ and $\mathbf{w}$, $F_n(\mathbf{v}) \geq F_n(\mathbf{w}) + (\mathbf{v} - \mathbf{w})^T \nabla F_n(\mathbf{w}) + \frac{\mu}{2}\|\mathbf{v} - \mathbf{w}\|_2^2$.*

Assumption 3. *For each client $n \in \mathcal{N}$, the stochastic gradient of F_n is unbiased with its variance bounded by σ_n^2.*

Assumption 4. *The expected squared norm of stochastic gradients for each client is uniformly bounded by G_n^2.*

Proposition 1 (Convergence Upper Bound with an Arbitrary Sampling Probability $\boldsymbol{q}$)**.** *For any given client sampling probability profile $\boldsymbol{q}$ in Algorithm 1, if we choose the decaying learning rate $\eta_r = \frac{2}{\max\{8L, \mu E\} + \mu r}$, the model parameter after R rounds $\boldsymbol{w}^R(\boldsymbol{q})$ has the optimality gap as follows*

$$\mathbb{E}\left[F\left(\boldsymbol{w}^R(\boldsymbol{q})\right)\right] - \min_{\boldsymbol{w}} F(\boldsymbol{w}) \leq \frac{1}{R}\left(\sum_{n=1}^{N} \frac{s_n}{q_n} + \beta\right), \tag{1}$$

where $s_n = \frac{8LEG_n^2}{\mu^2 N^2}$, $\beta = \frac{2L}{\mu^2 E} D + \frac{12L^2}{\mu^2 E}\Gamma + \frac{4L^2}{\mu E}\left\|\boldsymbol{w}_0 - \boldsymbol{w}^*\right\|^2$, $D = \sum_{n=1}^{N} (p_n \sigma_n)^2 + 8\sum_{n=1}^{N} p_n G_n^2 E^2$, and $\Gamma = F^* - \frac{1}{N}\sum_{n=1}^{N} \min_{\boldsymbol{w}} F_n(\boldsymbol{w})$.

Proof. The proof follows a similar argument of weighted client sampling in [1], where we first show that for any client sampling probabilities $\boldsymbol{q}$, the variance between the aggregated model $\boldsymbol{w}^{r+1}$ and the virtual global model under full participation (i.e., $\overline{\boldsymbol{w}}^{r+1}$) is bounded as follows:

$$\mathbb{E}_{\mathcal{S}(\boldsymbol{q})^r}\left\|\boldsymbol{w}^{r+1} - \overline{\boldsymbol{w}}^{r+1}\right\|^2 \leq 4\sum_{n=1}^{N} \frac{(1 - q_n)}{q_n}\left(\frac{\eta^r EG}{N}\right)^2. \tag{2}$$

Note that the main difference of (2) compared to the weighted sampling in is that the client sampling probability q_n is independent among each other. In particular, when $q_n = 1$ for all n, the variance in (2) is tightly bounded by zero, as the aggregated model $\boldsymbol{w}^{r+1}$ in the left hand side of (2) recovers the aggregated model of full client participation $\overline{\boldsymbol{w}}^{r+1}$. Then, we use mathematical induction to obtain a non-recursive bound on $\mathbb{E}_{\mathcal{S}(\boldsymbol{q})^r}\left\|\boldsymbol{w}^R - \boldsymbol{w}^*\right\|^2$, whose difference compared to the bound of full participation is the variance introduced in (2). After that, we converted the bound of $\mathbb{E}_{\mathcal{S}(\boldsymbol{q})^r}\left\|\boldsymbol{w}^R - \boldsymbol{w}^*\right\|^2$ to $\mathbb{E}[F\left(\boldsymbol{w}^R(\boldsymbol{q})\right)] - F^*$ using L-smoothness, which yields the additional term of $\sum_{n=1}^{N} \frac{1}{q_n}$ in (1) and concludes the proof. □

2 Proof of Theorem 1

Theorem 1. *A mechanism $m = (q, r)$ is incentive compatible and individual rational if and only if*

- *sampling probability $q(\tilde{c})$ is non-increasing in the reported cost $\tilde{c}$;*
- *payment function $r(\tilde{c})$ has the following form:*

$$r(\tilde{c}) = \tilde{c} + \frac{1}{q(\tilde{c})} \int_{\tilde{c}}^{c_{\max}} q(z)dz. \tag{3}$$

Proof. We first prove the "if" direction and then prove the "only if" direction.

We plug the payment function (3) into the client i's utility function, and obtain the following utility function when the agent i with true cost c_i reports $\tilde{c}_i$:

$$u(\tilde{c}_i; c_i) = q(\tilde{c}_i)(\tilde{c}_i - c_i) + \int_{\tilde{c}}^{c_{\max}} q(z)dz. \tag{4}$$

The derivative of $u(\tilde{c}_i; c_i)$ with respect to reported cost $\tilde{c}_i$ is

$$\frac{\partial u(\tilde{c}_i; c_i)}{\partial \tilde{c}_i} = q'(\tilde{c}_i)(\tilde{c}_i - c_i). \tag{5}$$

Since sampling probability q is non-increasing in reported cost, we have $q'(\tilde{c}_i) \leq 0$. So the derivative $\frac{\partial u(\tilde{c}_i;c_i)}{\partial \tilde{c}_i} \geq 0$ if $\tilde{c}_i \leq c_i$ and $\frac{\partial u(\tilde{c}_i;c_i)}{\partial \tilde{c}_i} \leq 0$ if $\tilde{c}_i \geq c_i$. The agent can maximize his utility when he truthfully reports his cost $\tilde{c}_i = c_i$.

To prove individual rationality, we can verify that

$$\max_{\tilde{c}_i} u(\tilde{c}_i; c_i) = u(c_i; c_i) = \int_{\tilde{c}}^{c_{\max}} q(z)dz \geq 0, \forall c_i \leq c_{\max}. \tag{6}$$

Next, we prove the "only if" direction. By incentive compatibility, we have $\max_{\tilde{c}_i} u(\tilde{c}_i; c_i) = u(c_i; c_i)$. By envelope theorem, we have

$$\frac{\partial u(c_i; c_i)}{\partial c_i} = \left.\frac{\partial u(\tilde{c}_i; c_i)}{\partial c_i}\right|_{\tilde{c}_i = c_i} = -q(c_i). \tag{7}$$

Taking the integral from c_i to $c_{\max}$, we have

$$u(c_{\max}; c_{\max}) - u(c_i; c_i) = -\int_{c_i}^{c_{\max}} q(z)dz. \tag{8}$$

We consider the minimum payment that satisfies the individual rationality such that the client with the maximum cost obtains exactly zero utility, i.e., $u(c_{\max}; c_{\max}) = 0$. Then, we have

$$u(c_i; c_i) = \int_{c_i}^{c_{\max}} q(z)dz \Rightarrow r(c_i) = c_i + \frac{1}{q(c_i)} \int_{c_i}^{c_{\max}} q(z)dz. \tag{9}$$

Recall that

$$\frac{\partial u(\tilde{c}_i; c_i)}{\partial \tilde{c}_i} = q'(\tilde{c}_i)(\tilde{c}_i - c_i). \tag{10}$$

Incentive compatibility implies that there exists $\epsilon > 0$ such that for any $\tilde{c}_i \in (c_i - \epsilon, c_i)$, $\frac{\partial u(\tilde{c}_i;c_i)}{\partial \tilde{c}_i} \geq 0$ and for any $\tilde{c}_i \in (c_i, c_i + \epsilon)$, $\frac{\partial u(\tilde{c}_i;c_i)}{\partial \tilde{c}_i} \leq 0$. Notice that $\tilde{c}_i - c_i < 0$ for any $\tilde{c}_i \in (c_i - \epsilon, c_i)$, and $\tilde{c}_i - c_i > 0$ for any $\tilde{c}_i \in (c_i, c_i + \epsilon)$. This requires that $q'(\tilde{c}_i) \leq 0$ on $(c_i - \epsilon, c_i)$ and $(c_i, c_i + \epsilon)$. As incentive compatibility holds for all c_i, this in particular implies that $q'(\tilde{c}_i) \leq 0$ for all c_i, which shows that q is non-increasing. □

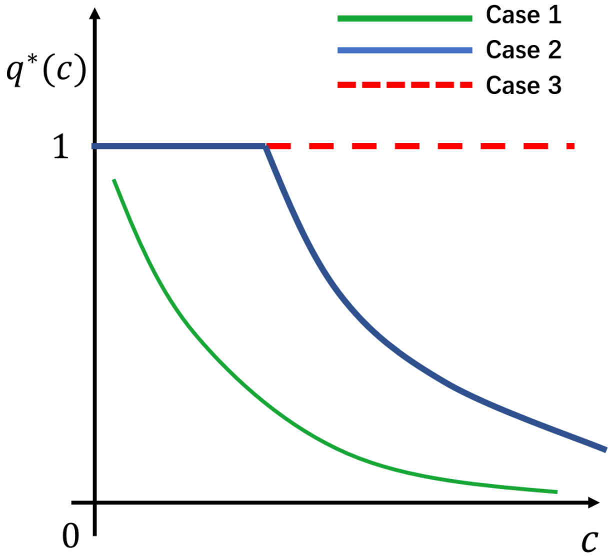


Figure 1: Illustration of the optimal sampling probability $q^*(c)$.

3 Illustration of the optimal sampling probability $q^*(c)$

Figure 1 demonstrates the structure of the optimal sampling probability $q^*(c)$.

4 Proof of Theorem 2

Assumption 5. *The virtual cost $\phi(c)$ is non-decreasing in the cost c.*

Theorem 2. *Under Assumption 5, the optimal sampling probability is as follows:*

1. *Case 1: If $\bar{B} \le \sqrt{\phi_{\min}}\mathbf{E}_c[\sqrt{\phi(c)}]$, the optimal sampling probability is*

$$q^*(c_n) = \frac{1}{\sqrt{\phi(c_n)}} \cdot \frac{\bar{B}}{\mathbf{E}_c[\sqrt{\phi(c)}]}, \tag{11}$$

   for all c_n.

2. *Case 2: If $\sqrt{\phi_{\min}}\mathbf{E}_c[\sqrt{\phi(c)}] < \bar{B} < \mathbf{E}_c[\phi(c)]$, the optimal sampling probability is*

$$q^*(c_n) = \begin{cases} 1, & c_n \le \hat{c}; \\ \frac{1}{\sqrt{\phi(c_n)}} \cdot \frac{\bar{B} - \mathbf{E}_c[\phi(c)\cdot 1\{c \le \hat{c}\}]}{\mathbf{E}_c[\sqrt{\phi(c)\cdot 1\{c > \hat{c}\}}]}, & c_n > \hat{c}. \end{cases} \tag{12}$$

   Here, $\hat{c}$ is solution to equation $H(x) \triangleq \mathbf{E}_c[\phi(c)\cdot \mathbb{1}\{c \le x\}] + \mathbf{E}_c[\sqrt{\phi(c)} \cdot \mathbb{1}\{c > x\}]$, and there exists a unique $\hat{c} \in (c_{\min}, c_{\max})$ satisfying $\bar{B} = H(\hat{c})$. And $\hat{c}$ can be computed by linear grid search over the support of γ.

3. *Case 3: If $\bar{B} \ge \mathbf{E}_c[\phi(c)]$, the optimal sampling probability is $q^*(c_n) = 1$ for all c_n.*

Proof. We start with the discrete version of P2, in which the cost take discrete values, and obtain discrete solution. Then we transform the discrete solution to continuous solution.

Discrete solution: We consider that the cost takes discrete values in the set $\{c_1, c_2, ..., c_K\}$, where $c_1 <$

$c_2 < ... < c_K$. The corresponding virtual cost associated with cost c_k is ϕ_k:

$$\phi_1 = c_1, \tag{13}$$

$$\phi_k = c_k + \frac{c_k - c_{k-1}}{f(c_k)} \cdot F(c_{k-1}), k = 2, .., K, \tag{14}$$

$$r_K = c_K. \tag{15}$$

$$r_k = c_k + \sum_{j=k+1}^{K} \frac{q_j}{q_k}(c_j - c_{j-1}), \quad k = 1, 2, ..., K-1. \tag{16}$$

where f is the probability of cost (i.e., $f(c_k) = P(c_k), k = 1, 2, ..., K$) and F is the cumulative density function of cost (i.e., $F(c_k) = \sum_{i=1}^{k} f(c_i), k = 1, 2, ..., K$). Then the probability of virtual cost is $P(\phi_k) = f_k$, $1 \leq k \leq K$. Define $\bar{B} = \frac{B}{RN}$.

Lemma 1. *The optimal sampling probability is as follows:*

1. *If $\bar{B} \leq \sqrt{\phi_1} \sum_{k=1}^{K} f_k \sqrt{\phi_k}$, the optimal sampling probability is*

$$q_k^* = \frac{1}{\sqrt{\phi_k}} \cdot \frac{\bar{B}}{\sum_{k=1}^{K} f_k \sqrt{\phi_k}}. \tag{17}$$

   for all k.

2. *If $\sqrt{\phi_1} \sum_{k=1}^{K} f_k \sqrt{\phi_k} < \bar{B} < \sum_{k=1}^{K} f_k \phi_k$, the optimal sampling probability is*

$$q_k^* = \begin{cases} 1, & 1 \leq k \leq \hat{k}; \\ \frac{1}{\sqrt{\phi_k}} \cdot \frac{\bar{B} - \sum_{k=1}^{\hat{k}} f_k \phi_k}{\sum_{k=\hat{k}+1}^{K} f_k \sqrt{\phi_k}}, & k > \hat{k}. \end{cases} \tag{18}$$

   Here, $\hat{k}$ is defined as follows: Let $H(m) = \sum_{k=1}^{m} f_k \phi_k + \sum_{k=m+1}^{K} f_k \sqrt{\phi_k} \cdot \sqrt{\phi_{m+1}}$. There is a unique $\hat{k} \in \{1, 2, ..., K-1\}$ satisfying $H(\hat{k}-1) < \bar{B} < H(\hat{k})$

3. *If $\bar{B} \geq \sum_{k=1}^{K} f_k \phi_k$, the optimal sampling probability is $q_k^* = 1$ for all k.*

Proof. Recall that we assume each client's cost is identically and independently distributed according to discrete distribution $f_k, k = 1, 2, ..., K$. Define $\bar{B} \triangleq B/(NR)$. We write the expectation in objective function and budget constraint explicitly through discrete distribution and obtain the following problem:

$$\min_{q} \max_{s \in [s_{\min}, s_{\max}]} \quad \sum_{k=1}^{K} s \cdot \frac{f_k}{q_k} \tag{19a}$$

$$\text{s.t.} \quad \sum_{k=1}^{K} f_k \cdot q_k \cdot \phi_k \leq \bar{B}; \tag{19b}$$

$$0 < q_k \leq 1, \forall k. \tag{19c}$$

Here, we drop the the monotonic constraint in Problem P2. Latter we will show that the solution indeed satisfies monotonic constraint. Notice that the objective function is an increasing function of s and thus the maximum is obtained when $s = s_{\max}$. Then the objective function of the optimization problem becomes $\sum_{k=1}^{K} s_{\max} \cdot f_k/q_k$. We find that the optimization problem is a convex problem. Thus, KKT conditions are sufficient and necessary for optimality. The Lagrangian of the optimization problem is

$$L(q,\lambda) = \sum_{k=1}^{K} s_{\max} \frac{f_k}{q_k} + \lambda \left(\sum_{k=1}^{K} f_k q_k \phi_k - \bar{B} \right) + \sum_{k=1}^{K} \lambda_k (q_k - 1). \tag{20}$$

Here we drop the constraint $q_k > 0$, for all k (this is without lose of generality and we will recover a positive solution later). According to KKT conditions, the optimal primal variables q_k^* and dual variables $\lambda^* \geq 0$, $\lambda_k^* \geq 0$ must satisfy

$$\begin{aligned} \frac{\partial L}{\partial q_k} &= -s_{\max} \frac{f_k}{q_k^{*2}} + \lambda^* f_k q_k^* + \lambda_k^* = 0 \\ \Rightarrow q_k^* &= \sqrt{\frac{s_{\max} f_k}{\lambda^* f_k \phi_k + \lambda_k^*}}, \quad \forall k, \end{aligned} \tag{21}$$

$$\lambda^* \left(\sum_{k=1}^{K} f_k q_k^* \phi_k - \bar{B} \right) = 0, \tag{22}$$

$$\lambda_k^* (q_k^* - 1) = 0, \quad \forall k, \tag{23}$$

$$\sum_{k=1}^{K} f_k \cdot q_k \cdot \phi_k \leq \bar{B}, \tag{24}$$

$$0 < q_k \leq 1, \forall k, \tag{25}$$

Next, we show that the solution in Theorem 1 exactly satisfies (21)-(25).

1. If $\bar{B} \leq \sqrt{\phi_1} \sum_{k=1}^{K} f_k \sqrt{\phi_k}$: we show that the primal variables

$$q_k^* = \frac{1}{\sqrt{\phi_k}} \cdot \frac{\bar{B}}{\sum_{k=1}^{K} f_k \sqrt{\phi_k}}, \quad \forall k, \tag{26}$$

and the dual variables

$$\lambda_k^* = 0, \quad \forall k, \tag{27}$$

$$\lambda^* = \frac{s_{\max} (\sum_{k=1}^{K} f_k \sqrt{\phi_k})^2}{\bar{B}^2}, \tag{28}$$

satisfy the KKT conditions (21)-(24). To see this, plugging the expressions of λ_k^* and λ^* into (21) yields

$$q_k^* = \sqrt{\frac{s_{\max}}{\lambda^* \phi_k}} = \frac{1}{\sqrt{\phi_k}} \cdot \frac{\bar{B}}{\sum_{k=1}^{K} f_k \sqrt{\phi_k}}, \tag{29}$$

which is exactly q_k^* in (26). Thus, condition in (21) holds. Meanwhile, we can see that

$$\begin{aligned} \sum_{k=1}^{K} f_k q_k^* \phi_k &= \sum_{k=1}^{K} f_k \phi_k \cdot \frac{1}{\sqrt{\phi_k}} \cdot \frac{\bar{B}}{\sum_{k=1}^{K} f_k \sqrt{\phi_k}} \\ &= \sum_{k=1}^{K} f_k \sqrt{\phi_k} \cdot \frac{\bar{B}}{\sum_{k=1}^{K} f_k \sqrt{\phi_k}} = \bar{B}. \end{aligned} \tag{30}$$

Thus, conditions in (22) and (24) holds. As $\lambda_k^* = 0$, condition in (23) holds. Finally, as $\bar{B} \leq$

$\sqrt{\phi_1}\sum_{k=1}^K f_k\sqrt{\phi_k}$, we have

$$q_1^* = \frac{1}{\sqrt{\phi_1}} \cdot \frac{\bar{B}}{\sum_{k=1}^K f_k\sqrt{\phi_k}} \leq 1. \tag{31}$$

We can check that q^* is indeed decreasing ($q_{k+1}^* < q_k^*$, $k = 1, 2, ..., K-1$) due to increasing virtual cost $\phi_1 < \phi_2 < ... < \phi_K$. Thus, condition in (25) holds. In conclusion, we have shown the optimality of q_k^* in (26).

2. If $\sqrt{\phi_1}\sum_{k=1}^K f_k\sqrt{\phi_k} < \bar{B} < \sum_{k=1}^K f_k\phi_k$: first of all, we define $H(m)$ as follows:

$$H(m) = \sum_{k=1}^{m} f_k\phi_k + \sqrt{\phi_{m+1}} \cdot \sum_{k=m+1}^{K} f_k\sqrt{\phi_k}, \tag{32}$$

where $m = 0, 1, .., K$. Notice that we define $H(0) = \sqrt{\phi_1}\sum_{k=1}^K f_k\sqrt{\phi_k}$ and $H(K) = \sum_{k=1}^K f_k\phi_k$. We show the monotonicity of $H(m)$ as follows.

Lemma 2. *$H(m+1) > H(m)$, $\quad m = 0, 1, ..., K-1$. And there exists a unique $\hat{k} \in \{1, 2, ..., K-1\}$ satisfying $H(\hat{k}-1) < \bar{B} < H(\hat{k})$ for $\bar{B} \in (H(0), H(K))$.*

Proof. We can see that

$$\begin{aligned}
&H(m+1) - H(m) \\
=&\sum_{k=1}^{m+1} f_k\phi_k + \sqrt{\phi_{m+2}} \cdot \sum_{k=m+2}^{K} f_k\sqrt{\phi_k} \\
&- \sum_{k=1}^{m} f_k\phi_k - \sqrt{\phi_{m+1}} \cdot \sum_{k=m+1}^{K} f_k\sqrt{\phi_k} \\
=&f_{m+1}\phi_{m+1} + \sum_{k=m+2}^{K} f_k\sqrt{\phi_k}(\sqrt{\phi_{m+2}} - \sqrt{\phi_{m+1}}) \\
&- f_{m+1}\phi_{m+1} \\
>&0.
\end{aligned} \tag{33}$$

The inequality is due to increasing virtual cost, i.e., ϕ_k is increasing in k. Thus, $H(m)$ is increasing in m. For $\bar{B} \in (H(0), H(K))$, there exists a unique $\hat{k} \in \{1, 2, ..., K-1\}$ such that $H(\hat{k}-1) < \bar{B} < H(\hat{k})$. □

Next, we show that the primal variables

$$q_k^* = \begin{cases} 1, & 1 \leq k \leq \hat{k}; \\ \frac{1}{\sqrt{\phi_k}} \cdot \frac{\bar{B} - \sum_{k=1}^{\hat{k}} f_k\phi_k}{\sum_{k=\hat{k}+1}^{K} f_k\sqrt{\phi_k}}, & k > \hat{k}, \end{cases} \tag{34}$$

and dual variables

$$\lambda^* = \frac{s_{\max}(\sum_{k=\hat{k}+1}^{K} f_k\sqrt{\phi_k})^2}{(\bar{B} - \sum_{k=1}^{\hat{k}} f_k\phi_k)^2}, \tag{35}$$

$$\lambda_k^* = \begin{cases} f_k(s_{\max} - \lambda^*\phi_k), & 1 \leq k \leq \hat{k}; \\ 0, & k > \hat{k}, \end{cases} \tag{36}$$

satisfy the KKT conditions (21)-(24).

To see this, for $k \in [1, \hat{k}]$, plugging the expressions of λ^* into (21) yields

$$q_k^* = 1, \tag{37}$$

which is exactly q_k^* in (34). Thus, conditions in (21), (23) and (25) hold. Next, we verify that $\lambda_k^* > 0$. To see this, plugging the expression of λ^* into λ_k^*, we have

$$\lambda_k^* = s_{\max} f_k \left(1 - \frac{\phi_k (\sum_{k=\hat{k}+1}^{K} f_k \sqrt{\phi_k})^2}{(\bar{B} - \sum_{k=1}^{\hat{k}} f_k \phi_k)^2} \right). \tag{38}$$

According to $\bar{B} > H(\hat{k} - 1)$, i.e.,

$$\begin{aligned} \bar{B} &> \sum_{k=1}^{\hat{k}-1} f_k \phi_k + \sqrt{\phi_{\hat{k}}} \cdot \sum_{k=\hat{k}}^{K} f_k \sqrt{\phi_k} \\ &= \sum_{k=1}^{\hat{k}} f_k \phi_k + \sqrt{\phi_{\hat{k}}} \cdot \sum_{k=\hat{k}+1}^{K} f_k \sqrt{\phi_k}, \end{aligned} \tag{39}$$

we have

$$\bar{B} - \sum_{k=1}^{\hat{k}} f_k \phi_k > \sqrt{\phi_{\hat{k}}} \cdot \sum_{k=\hat{k}+1}^{K} f_k \sqrt{\phi_k}. \tag{40}$$

Combining increasing virtual cost ($\phi_k \leq \phi_{\hat{k}}$), we have

$$\frac{\sqrt{\phi_k} \cdot \sum_{k=\hat{k}+1}^{K} f_k \sqrt{\phi_k}}{\bar{B} - \sum_{k=1}^{\hat{k}} f_k \phi_k} < 1, k \leq \hat{k}. \tag{41}$$

Thus,

$$\lambda_k^* = s_{\max} f_k \left(1 - \frac{\phi_k (\sum_{k=\hat{k}+1}^{K} f_k \sqrt{\phi_k})^2}{(\bar{B} - \sum_{k=1}^{\hat{k}} f_k \phi_k)^2} \right) > 0 \tag{42}$$

As for $k \in [\hat{k} + 1, K]$, plugging the expressions of λ_k^* and λ^* into (21) yields

$$q_k^* = \frac{1}{\sqrt{\phi_k}} \cdot \frac{\bar{B} - \sum_{k=1}^{\hat{k}} f_k \phi_k}{\sum_{k=\hat{k}+1}^{K} f_k \sqrt{\phi_k}}, \tag{43}$$

which is exactly q_k^* in (34). Thus, condition in (21) holds. According to $\bar{B} < H(\hat{k})$, i.e.,

$$\bar{B} < \sum_{k=1}^{\hat{k}} f_k \phi_k + \sqrt{\phi_{\hat{k}+1}} \cdot \sum_{k=\hat{k}+1}^{K} f_k \sqrt{\phi_k}, \tag{44}$$

we have

$$q_{\hat{k}+1}^* = \frac{1}{\sqrt{\phi_{\hat{k}+1}}} \cdot \frac{\bar{B} - \sum_{k=1}^{\hat{k}} f_k \phi_k}{\cdot \sum_{k=\hat{k}+1}^{K} f_k \sqrt{\phi_k}} < 1. \tag{45}$$

Thus, for $k \in [\hat{k} + 1, K]$, combining increasing virtual cost ($\phi_{k+1} > \phi_k$), we have $q_k^* \leq q_{\hat{k}+1}^* < 1$, which means condition in (25) holds. As $\lambda_k^* = 0$, condition in (23) holds.

Finally, for $k \in [1, K]$,

$$\begin{aligned}
& \sum_{k=1}^{K} f_k \cdot q_k^* \cdot \phi_k \\
= & \sum_{k=1}^{\hat{k}} f_k \cdot \phi_k + \sum_{k=\hat{k}+1}^{K} \frac{f_k \phi_k}{\sqrt{\phi_k}} \cdot \frac{\bar{B} - \sum_{k=1}^{\hat{k}} f_k \phi_k}{\sum_{k=\hat{k}+1}^{K} f_k \sqrt{\phi_k}} \\
= & \sum_{k=1}^{\hat{k}} f_k \cdot \phi_k + \frac{\bar{B} - \sum_{k=1}^{\hat{k}} f_k \phi_k}{\sum_{k=\hat{k}+1}^{K} f_k \sqrt{\phi_k}} \cdot \sum_{k=\hat{k}+1}^{K} f_k \sqrt{\phi_k} \\
= & \bar{B}.
\end{aligned} \tag{46}$$

That is, conditions in (22) and (24) hold. In conclusion, we have shown the optimality of q_k^* in (34).

3. If $\bar{B} \geq \sum_{k=1}^{K} f_k \phi_k$, we show that the primal variables

$$q_k^* = 1, \quad \forall k, \tag{47}$$

and the dual variables

$$\lambda^* = 0 \tag{48}$$

$$\lambda_k^* = s_{\max} f_k, \quad \forall k. \tag{49}$$

To see this, plugging the expressions of λ_k^* and λ^* into (21) yields

$$q_k^* = 1, \tag{50}$$

which is exactly q_k^* in (34). Thus, conditions in (21), (23) and (25) hold. As $\lambda^* = 0$, condition in (22) holds. Finally, we have

$$\sum_{k=1}^{K} f_k \cdot q_k^* \cdot \phi_k = \sum_{k=1}^{K} f_k \phi_k \leq \bar{B}, \tag{51}$$

which means condition in (24) holds. In conclusion, we have shown the optimality of q_k^* in (47).

□

From discrete to continuous costs: The continuous cost can be considered as a special case of discrete cost by setting the number of discrete costs K to be infinity. To transform discrete solution to continuous solution, we replace summations with integrals considering K goes to infinity as follows:

$$\sum_{k=1}^{K} f_k \sqrt{\phi_k} \Rightarrow \int_{c_{\min}}^{c_{\max}} f(c) \sqrt{\phi(c)} dc = \mathbf{E}_c[\sqrt{\phi(c)}], \tag{52}$$

$$\sum_{k=1}^{K} f_k \phi_k \Rightarrow \int_{c_{\min}}^{c_{\max}} f(c) \phi(c) dc = \mathbf{E}_c[\phi(c)], \tag{53}$$

$$\begin{aligned}
& H(m) = \sum_{k=1}^{m} f_k \phi_k + \sum_{k=m+1}^{K} f_k \sqrt{\phi_k} \cdot \sqrt{\phi_{m+1}} \\
& \Rightarrow H(x) = \mathbf{E}_c[\phi(c) \cdot \mathbb{1}\{c \leq x\}] + \mathbf{E}_c[\sqrt{\phi(c)} \cdot \mathbb{1}\{c > x\}]
\end{aligned} \tag{54}$$

Finally, we are able to derive the continuous solution in Theorem 2. □

Algorithm 1: Prior-Independent Algorithm for a Sequence of Mechanisms

```
   Input: Fraction $\rho$ of total clients for initial cost distribution estimation, initial uniform sampling probability $\hat{q}$
   Output: sampling probability and payment for each client
1  The server initializes a mechanism $m_0 = (r_0, q_0 = \hat{q})$, where $r_0$ is calculated according to (3);
2  The server uniformly samples a set $\mathcal{M}_0$ of $\rho N$ clients, and shows them the mechanism $m_0 = (r_0, q_0)$;
3  The clients in set $\mathcal{M}_0$ report their costs $\mathbf{c} = [c_n, \forall n \in \mathcal{M}_0]$, and get the results (sampling probability and
     payment);
4  The server estimates the cost distribution $\hat{\gamma}_1$ based on *all* the reported costs $\mathbf{c}$, by maximum likelihood
     estimation;
5  The server updates the remaining budget $B_1 = B/R - \sum_{n \in \mathcal{M}_0} q_0(c_n) \cdot r_0(c_n)$, and the set of rest clients
     $\mathcal{M}_1 = \mathcal{N} \backslash \mathcal{M}_0$;
6  The server updates the mechanism $m_1 = M(\hat{\gamma}_1, N - M, B_1)$;
7  for $j = 1, 2, ..., N - M$ do
8      if $B_j - q_j(c_{\max}) \cdot r_j(c_{\max}) \geq 0$// can offer one client with the maximum cost
9       then
10         The server uniformly samples a client $j$ from the set $\mathcal{M}_j$, and shows it the mechanism $m_j = (r_j, q_j)$;
11         The client $j$ reports its cost $c_j$, and gets the results (sampling probability and payment);
12         The server estimates the cost distribution $\hat{\gamma}_{j+1}$ based on *all* previously reported costs;
13         The server updates the remaining budget $B_{j+1} = B_j - q_j(c_j) \cdot r_j(c_j)$, and the set of rest clients
             $\mathcal{M}_{j+1} = \mathcal{M}_j \backslash j$;
14         The server updates the mechanism $m_{j+1} = M(\hat{\gamma}_{j+1}, N - M - j, B_{j+1})$;
15     else
16         The server shows all the rest of clients the mechanism $m_j = \left(r_j, q_j = \frac{B_j}{c_{\max}(N-M-j+1)}\right)$, where $r_j$ is
             calculated according to (3);
17         return
```

5 Prior-Independent Mechanism Design

So far, we have focused on mechanism design with a prior cost distribution γ known by the server, as computing the virtual cost requires the knowledge of such a distribution. In this section, we extend our previous results to a prior-independent mechanism design scenario, in which we do not require the knowledge of the cost distribution *a priori*. Instead, the high-level idea is to leverage the incentive compatibility property (i.e., truthfulness) to learn the cost distribution by a sequence of mechanisms.

We start with introducing a few notations. Let $M(\tilde{\gamma}, N, B)$ denote the optimal mechanism when there are N clients with cost distribution $\tilde{\gamma}$ and the server has a budget of B. Recall that the mechanism consists of a payment function and a sampling probability function. The server can derive the optimal mechanism under these parameters (γ, N, B) according to Theorems 1 and 2.

We present the prior-independent mechanism in Algorithm 1. The algorithm mainly consists of two periods: an *initial estimation* period (Lines 1-6) and an *iterative update* period (Lines 7-17). During *initial estimation*, the server initially estimates the cost distribution $\gamma(\cdot)$ from a set of clients by using an initial mechanism with a uniform sampling probability. During *iterative update*, the server iteratively updates the cost distribution and the mechanism based on new cost samples. Further, we would like to highlight that as the optimization problem is constrained by budget, we need to keep track of the remaining available budget during *iterative update*. This is one key difference compared with traditional prior-independent mechanisms.

Initial estimation: The server initializes a mechanism with a uniform sampling probability $\hat{q}$ and the corresponding payment calculated by (3). The server uniformly samples ρN clients and shows them the initial mechanism. Here, ρ and $\hat{q}$ are hyperparameters of the algorithm. We will discuss them at the end of this subsection. By Theorem 1, the clients would like to truthfully report their costs, based on which the server can estimate the cost distribution through maximum likelihood estimation (Line 4).[1] With the estimated cost

[1] Consider an example of discrete cost. Suppose there n_k clients reporting cost c_k, $k = 1, 2, \ldots$. By maximum likelihood estimation, the cost distribution is estimated as $\Pr[c = c_k] = n_k / \sum_k n_k$, for all k.

distribution, the server can easily derive the estimated optimal mechanism with the remaining budget for the rest of the clients (Lines 5-6),[2] according to Theorem 1 and Theorem 2.

Iterative update: The server runs the following iteration. It firstly checks whether the remaining budget is able to cover the expected payment of the client with the maximum cost (considering the worst-case scenario where recruiting the next client needs the highest expected payment).

- If yes, then the server uniformly samples a client from the rest of the clients, and shows it the newly derived mechanism (Line 10). The server gets a new sample of reported cost and updates the cost distribution (Lines 11-12). It derives a new mechanism (Lines 13-14) for the rest of the clients based on the updated distribution and the remaining budget. Then it continues the iteration with the new mechanism.

- If no, then the server treats all the rest of the clients as if they were with the maximum costs. That is, the server presents the mechanism with a uniform sampling probability $B_j/c_{\max}/(N-M-j+1)$ and the corresponding payment in (3). Then it stops the iteration.

As Algorithm 1 progresses, the server collects more cost samples, and hence the estimated cost distribution is expected to be more accurate. Therefore, the sequence of generated mechanisms by Algorithm 1 will converge to the optimal one. Meanwhile, notice that every client makes a one-shot decision. Our mechanism ensures that he would maximize his utility if he truthfully reports his cost in his one-shot decision-making. We note that each mechanism also satisfies the desirable economic properties.

Proposition 2. *Under Assumption 5, the mechanisms in Algorithm 1 satisfy incentive compatibility, individually rationality, and budget feasibility.*

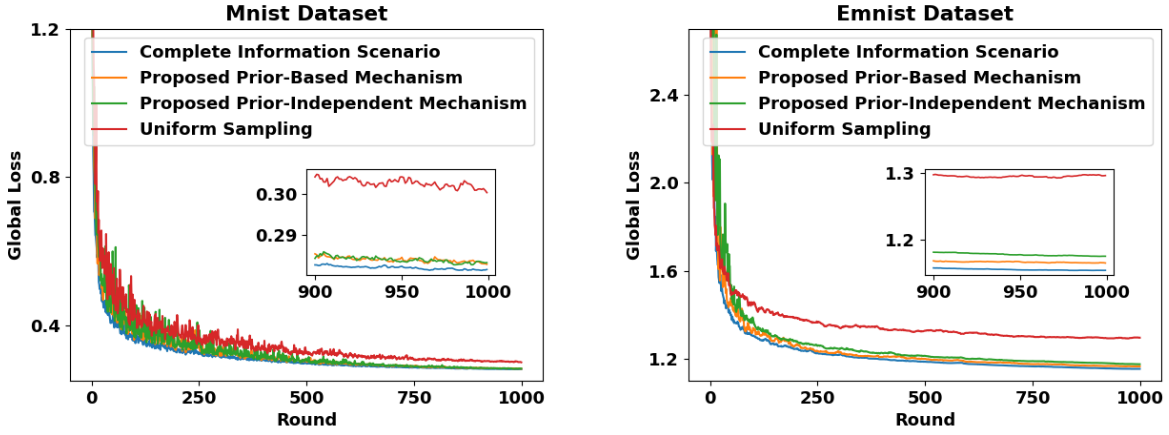
Finally, we discuss the impacts of two hyperparameters ρ (fraction of total clients for initial cost distribution estimation) and $\hat{q}$ (initial uniform sampling probability) on the performance of Algorithm 1. Specifically, setting a higher value of ρ would make the first cost distribution estimation more accurate, but would give more clients in *initial estimation* period non-optimal sampling probabilities. And setting a higher value of $\hat{q}$ would naturally give a smaller convergence bound of FL in *initial estimation* period, but would leave fewer budget for mechanism design in *iterative update* period. The setting of both hyperparameters faces an exploration-exploitation trade-off in unknown environments.

6 Performance Evaluation

Figs. 2(a), 2(b), 3(a), and 3(b) show that the training loss under our proposed prior-based mechanism is lower than that of uniform sampling scheme. Specifically, the training loss under our proposed prior-based mechanism is 94% and 90% of that under uniform sampling scheme on MNIST and EMNIST datasets, respectively (in Figs.2(a) and 2(b)). *This indicates that our proposed prior-based mechanism efficiently exploits the limited budget and recruits more clients, resulting in a better performance than the uniform sampling scheme.* This conclusion is further verified in Fig. 4: the average of sampling probabilities under our proposed prior-based mechanism is lower than that under uniform sampling scheme for different values of mean of cost.

Figs. 2(a), 2(b), 3(a), and 3(b) also show the training loss under our proposed prior-based mechanism is comparable to that under complete information scenario. *This shows a relatively small gap between our proposed prior-based mechanism and the optimal mechanism with complete information.*

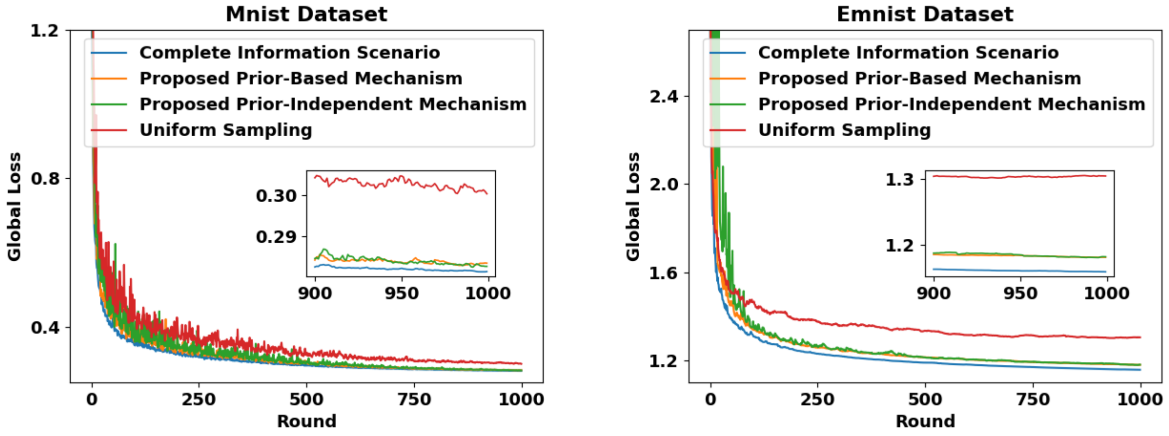
[2] We assume that the server would not run out of budget in *initial estimation* period. This assumption holds by setting a low sampling probability. We also assume that the server knows the maximum value of the cost $c_{\max}$. Otherwise, the server can aggressively set the maximum cost to a high value.



(a) Loss on MNIST dataset.

(b) Loss on EMNIST dataset.

Figure 2: Comparison of loss under different mechanisms on MNIST and EMNIST datasets with the costs following uniform distribution.



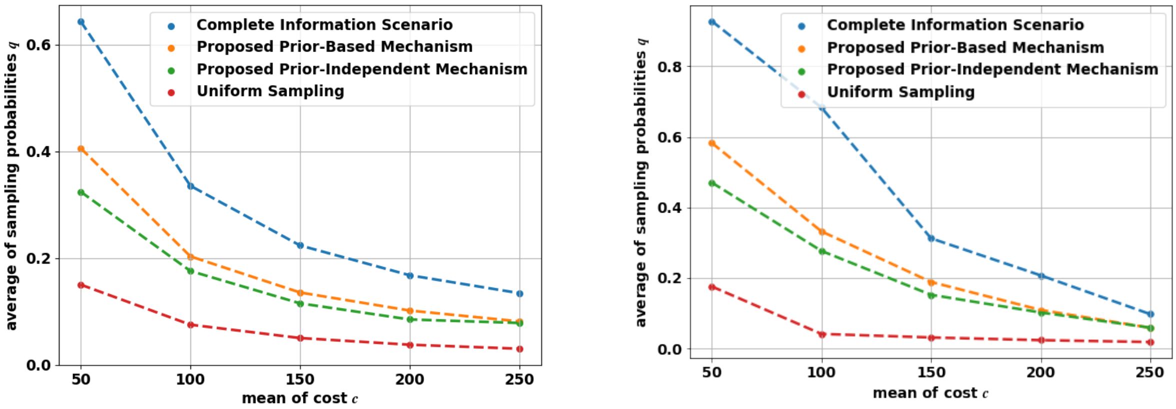
(a) Loss on MNIST dataset.

(b) Loss on EMNIST dataset.

Figure 3: Comparison of loss under different mechanisms on MNIST and EMNIST datasets with the costs following exponential distribution.

Reference

[1] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. In *International Conference on Learning Representation*, 2019.



(a) Uniform distribution.

(b) Exponential distribution.

Figure 4: Average of sampling probabilities v.s. mean of cost under two distributions of cost.