A Appendix to Section 3

We first show that a Nash equilibrium exists when agent payoff functions are *separable*, i.e., for every agent i there are functions $g_i : S_i \to \mathbb{R}_{\geq 0}$ and $h_i : \bigtimes_{j \neq i} S_j \to \mathbb{R}_{\geq 0}$ s.t. for all $\boldsymbol{s} \in \mathcal{S}$, $a_i(\boldsymbol{s}) = g_i(s_i) + h_i(\boldsymbol{s}_{-i})$.

Theorem A.1. *In any federated learning problem where agent payoff functions are separable, a Nash equilibrium exists.*

Proof. When the payoff function of an agent i is separable, the best response to any contribution vector $\boldsymbol{s}_{-i}$ is independent of $\boldsymbol{s}_{-i}$:

$$\begin{aligned} f_i(\boldsymbol{s}_{-i}) &= \arg\max_{x \in S_i} a_i(x, \boldsymbol{s}_{-i}) - c_i(x) = \arg\max_{x \in S_i} g_i(x) + h_i(\boldsymbol{s}_{-i}) - c_i(x) \\ &= \arg\max_{x \in S_i} g_i(x) - c_i(x). \qquad \text{(since } h_i(\boldsymbol{s}_{-i}) \text{ is independent of } x\text{)} \end{aligned}$$

Let $F_i := \arg\max_{x \in S_i} g_i(x) - c_i(x)$. Clearly $F_i \neq \emptyset$ since $S_i \neq \emptyset$. Then any $\boldsymbol{s} \in \bigtimes_i F_i$ satisfies $\boldsymbol{s} \in f(\boldsymbol{s})$ by definition. By Proposition 1 any such sample vector is a Nash equilibrium. $\square$

Next, we present a negative result showing that there are federated learning settings where a Nash equilibrium is not guaranteed to exist.

Theorem A.2. *There exists a federated learning problem in which a Nash equilibrium does not exist. Moreover, the instance has three agents with continuous, non-decreasing, non-concave payoff functions and linear cost functions.*

Proof. Let $\varepsilon \in (0, \frac{1}{16})$. Let $e : [0, 1] \to [0, 1]$ be a function given by:

$$e(x) = \begin{cases} 0, \text{ if } 0 \leq x \leq \frac{1}{2} - \varepsilon, \\ \frac{1}{2} + \frac{1}{2\varepsilon}(x - \frac{1}{2}), \text{ if } \frac{1}{2} - \varepsilon \leq x \leq \frac{1}{2} + \varepsilon, \\ 1, \text{ if } \frac{1}{2} + \varepsilon \leq x \leq 1. \end{cases} \tag{8}$$

Essentially the function e is a continuous, piece-wise linear function connecting $(0, 0)$, $(\frac{1}{2} - \varepsilon, 0)$, $(\frac{1}{2} + \varepsilon, 1)$ and $(1, 1)$.

Now consider the following federated learning instance with $n = 3$ agents, where $S_1 = S_2 = S_3 = [0, 1]$. The payoff functions are given by:

$$\begin{aligned} a_1(\boldsymbol{s}) &= e(s_1) + e(s_3) - e(s_1) \cdot e(s_3) \\ a_2(\boldsymbol{s}) &= e(s_2) + e(s_1) - e(s_2) \cdot e(s_1) \\ a_3(\boldsymbol{s}) &= e(s_3) + e(s_2) - e(s_3) \cdot e(s_2), \end{aligned} \tag{9}$$

and the cost functions are $c_i(s_i) = \frac{1}{4} s_i$ for all $i \in [3]$. Notice that the payoff functions are increasing in s_j for every $j \in [3]$ and are continuous since e is continuous.

We now show that this instance does not admit a Nash equilibrium. Let us first evaluate the best response set $f_1(s_2, s_3)$. Note that $u_1(\boldsymbol{s}) = e(s_1) \cdot (1 - e(s_3)) + e(s_3) - \frac{1}{4} s_1$. Since $u_1(\boldsymbol{s})$ is independent of s_2, $f_1(s_2, s_3)$ only depends on s_3.

- Case 1. $s_3 \leq \frac{1}{2} - \varepsilon$. Then $u_1(\boldsymbol{s}) = e(s_1) - \frac{1}{4} s_1$, which is maximized at $s_1 = \frac{1}{2} + \varepsilon$ and results in a utility of $\frac{7}{8} - \frac{\varepsilon}{4}$.
- Case 2. $s_3 \geq \frac{1}{2} + \varepsilon$. Then $u_1(\boldsymbol{s}) = 1 - \frac{1}{4} s_1$, which is maximized at $s_1 = 0$ and results in a utility of 1.
- Case 3. $\frac{1}{2} - \varepsilon \leq s_3 \leq \frac{1}{2} + \varepsilon$. We consider the intervals in which the best response s_1 to such an s_3 can lie:
  - $s_1 \leq \frac{1}{2} - \varepsilon$. In this range, $u_1(\boldsymbol{s}) = e(s_3) - \frac{1}{4} s_1$, which is maximized at $s_1 = 0$ and results in a utility of $e(s_3)$.

  - $s_1 \geq \frac{1}{2} + \varepsilon$. In this range, $u_1(\boldsymbol{s}) = 1 - \frac{1}{4}s_1$, which is maximized at $s_1 = \frac{1}{2} + \varepsilon$ and results in a utility of $\frac{7}{8} - \frac{\varepsilon}{4}$.
  - $\frac{1}{2} - \varepsilon \leq s_1 \leq \frac{1}{2} + \varepsilon$. In this range, using the definition of $e(s_1)$ (eq. 8) we obtain:

$$u_1(\boldsymbol{s}) = \left(\frac{1 - e(s_3)}{2\varepsilon} - \frac{1}{4}\right) \cdot s_1 + (1 - e(s_3)) \cdot \left(\frac{1}{2} - \frac{1}{4\varepsilon}\right) + e(s_3).$$

    Thus $u_1(\boldsymbol{s})$ is a linear function in s_1 with slope $\frac{1-e(s_3)}{2\varepsilon} - \frac{1}{4}$. If the slope is positive, then the best response in the current interval is $s_1 = \frac{1}{2} + \varepsilon$, and gives a utility of $\frac{7}{8} - \frac{\varepsilon}{4}$. If the slope is negative, then $s_1 = \frac{1}{2} - \varepsilon$ is the best response in the current interval and gives a utility of $e(s_3) - \frac{1}{4}(\frac{1}{2} - \varepsilon)$. However $s_1 = 0$ gives a utility of $e(s_3)$ implying that $s_1 = \frac{1}{2} - \varepsilon$ cannot be a best response. Finally if the slope is zero, then it must mean that $e(s_3) = 1 - \frac{\varepsilon}{2}$, and the utility is $\frac{\varepsilon}{2}(\frac{1}{2} - \frac{1}{4\varepsilon}) + 1 - \frac{\varepsilon}{2} = \frac{7}{8} - \frac{\varepsilon}{4}$. However responding with $s_1 = 0$ gives a utility of $e(s_3) = 1 - \frac{\varepsilon}{2}$, which exceeds $\frac{7}{8} - \frac{\varepsilon}{4}$, since $\varepsilon < \frac{1}{16}$. Thus, the best response does not lie in $(\frac{1}{2} - \varepsilon, \frac{1}{2} + \varepsilon)$ and $s_1 = 0$ is the overall best response.

The above discussion shows that the best response $f_1(s_2, s_3) \subseteq \{0, \frac{1}{2} + \varepsilon\}$. By symmetry, the same holds for f_2 and f_3. Suppose there exists a Nash equilibrium $\boldsymbol{s}^* = (s_1^*, s_2^*, s_3^*)$. By Proposition 1, $\boldsymbol{s}^* \in f(\boldsymbol{s}^*)$. Since the above discussion implies $s_3^* \in \{0, \frac{1}{2} + \varepsilon\}$, we consider two cases:

- Suppose $s_3^* = 0$. Then

$$\begin{aligned} s_3^* = 0 \implies s_1^* &= \frac{1}{2} + \varepsilon && \text{(Case 1 for agent 1)} \\ \implies s_2^* &= 0 && \text{(Case 2 for agent 2)} \\ \implies s_3^* &= \frac{1}{2} + \varepsilon, && \text{(Case 1 for agent 3)} \end{aligned}$$

  which is a contradiction.

- Suppose $s_3^* = \frac{1}{2} + \varepsilon$. Then

$$\begin{aligned} s_3^* = \frac{1}{2} + \varepsilon \implies s_1^* &= 0 && \text{(Case 2 for agent 1)} \\ \implies s_2^* &= \frac{1}{2} + \varepsilon && \text{(Case 1 for agent 2)} \\ \implies s_3^* &= 0, && \text{(Case 2 for agent 3)} \end{aligned}$$

  which is also a contradiction.

This shows that there is no $\boldsymbol{s}^*$ such that $\boldsymbol{s}^* \in f(\boldsymbol{s}^*)$, implying that the above instance does not admit a Nash equilibrium. $\square$

We now prove the fast convergence of best response dynamics.

Theorem 3.2. *Let $G(\boldsymbol{s})$ be the Jacobian of $\boldsymbol{u} : \mathcal{S} \to \mathbb{R}^n$, i.e., $G(\boldsymbol{s})_{ij} = \frac{\partial^2 u_i(\boldsymbol{s})}{\partial s_j \partial s_i}$. Assuming agent utility functions u_i satisfy*

1. *Strong concavity: $(G + \lambda \cdot I_{n \times n})$ is negative semi-definite,*
2. *Bounded derivatives: $|G_{ij}| \leq L$,*

for constants $\lambda, L > 0$, the best response dynamics (4) *with step size $\delta^t = \frac{\lambda}{n^2 L^2}$ converges to an approximate Nash equilibrium $\boldsymbol{s}^T$ where $\|g(\boldsymbol{s}^T, \boldsymbol{\mu}^T)\|_2 < \varepsilon$ in T iterations, where*

$$T = \frac{2n^2 L^2}{\lambda^2} \log\left(\frac{\|g(\boldsymbol{s}^0, \boldsymbol{\mu}^0)\|_2}{\varepsilon}\right).$$

Proof. Observe that $\boldsymbol{\mu}^t$ is chosen s.t. $\|g(\boldsymbol{s}^t, \boldsymbol{\mu}^t)\|_2$ is minimized among all μ s.t. the updated sample vector $\boldsymbol{s}^{t+1}$ remains in $\mathcal{S}$. Thus:

$$\|g(\boldsymbol{s}^{t+1}, \boldsymbol{\mu}^{t+1})\|_2 \leq \|g(\boldsymbol{s}^{t+1}, \boldsymbol{\mu}^t)\|_2 \tag{10}$$

Using Taylor's expansion, we have:

$$g(\boldsymbol{s}^{t+1}, \boldsymbol{\mu}^t) = g(\boldsymbol{s}^t, \boldsymbol{\mu}^t) + H(\boldsymbol{s}', \mu^t) \cdot (\boldsymbol{s}^{t+1} - \boldsymbol{s}^t),$$

where $H_{ij}(\boldsymbol{s}', \boldsymbol{\mu}^t) = \frac{\partial g(\boldsymbol{s}', \boldsymbol{\mu}^t)}{\partial s_j}$, and $\boldsymbol{s}' = \boldsymbol{s}^t + \alpha(\boldsymbol{s}^{t+1} - \boldsymbol{s}^t)$ for some $\alpha \in [0, 1]$.

By definition, $g(\boldsymbol{s}^t, \mu^t)_i = \frac{\partial u_i(\boldsymbol{s}^t)}{\partial s_i} + \mu_i^t$. Thus $H_{ij}(\boldsymbol{s}', \boldsymbol{\mu}^t) = \frac{\partial^2 u_i(\boldsymbol{s}^t)}{\partial s_j \partial s_i} = G_{ij}(\boldsymbol{s}')$, hence $H(\boldsymbol{s}', \boldsymbol{\mu}^t) = G(\boldsymbol{s}')$. The BR dynamics update rule (4) implies $\boldsymbol{s}^{t+1} - \boldsymbol{s}^t = \delta^t \cdot g(\boldsymbol{s}^t, \boldsymbol{\mu}^t)$. We therefore have $g(\boldsymbol{s}^{t+1}, \boldsymbol{\mu}^t) = (I_{n\times n} + \delta^t \cdot G(\boldsymbol{s}')) \cdot g(\boldsymbol{s}^t, \boldsymbol{\mu}^t)$. Taking the L^2 norm, we get:

$$\|g(\boldsymbol{s}^{t+1}, \boldsymbol{\mu}^t)\|_2^2 = \|g(\boldsymbol{s}^t, \boldsymbol{\mu}^t)\|_2^2 + \delta_t^2 \cdot \|G(\boldsymbol{s}')g(\boldsymbol{s}^t, \boldsymbol{\mu}^t)\|_2^2 + 2\delta_t g(\boldsymbol{s}^t, \boldsymbol{\mu}^t)^T G(\boldsymbol{s}')g(\boldsymbol{s}^t, \boldsymbol{\mu}^t), \tag{11}$$

By the strong concavity assumption, for a constant $\lambda > 0$, $G + \lambda \cdot I_{n\times n}$ is negative semi-definite, i.e., $v^T(G + \lambda \cdot I_{n\times n})v \leq 0$ for any $v \in \mathbb{R}^n$. With $v = g(\boldsymbol{s}^t, \boldsymbol{\mu}^t)$, we have:

$$g(\boldsymbol{s}^t, \boldsymbol{\mu}^t)^T G(\boldsymbol{s}')g(\boldsymbol{s}^t, \boldsymbol{\mu}^t) \leq -\lambda \cdot \|g(\boldsymbol{s}^t, \boldsymbol{\mu}^t)\|_2^2. \tag{12}$$

Next we use the fact that the L^2 norm $\|A\|_2$ of an $n \times n$ matrix A is bounded by its Frobenius norm $\|A\|_F$:

$$\|A\|_2 := \sup_{x\neq 0} \frac{\|Ax\|_2}{\|x\|_2} \leq \|A\|_F := \sqrt{\sum_i \sum_j |A_{ij}|^2}$$

By the bounded derivatives assumption, we have $|G(\boldsymbol{s}')_{ij}| \leq L$, which implies that $\|G(\boldsymbol{s}')\|_F = \sqrt{\sum_i \sum_j L^2} = nL$. This gives:

$$\|G(\boldsymbol{s}')g(\boldsymbol{s}^t, \mu^t)\|_2 \leq nL\|g(\boldsymbol{s}^t, \mu^t)\|_2. \tag{13}$$

Using (12) and (13) in (11), we get:

$$\|g(\boldsymbol{s}^{t+1}, \boldsymbol{\mu}^t)\|_2^2 = (1 + \delta_t^2 \cdot n^2L^2 - 2\delta^t\lambda) \cdot \|g(\boldsymbol{s}^t, \boldsymbol{\mu}^t)\|_2^2,$$

Since $\delta^t = \frac{\lambda}{n^2L^2}$, the above equation together with (10) gives:

$$\|g(\boldsymbol{s}^{t+1}, \boldsymbol{\mu}^{t+1})\|_2^2 \leq \left(1 - \frac{\lambda^2}{n^2L^2}\right) \cdot \|g(\boldsymbol{s}^t, \boldsymbol{\mu}^t)\|_2^2.$$

Using $(1-x)^r \leq e^{-xr}$ repeatedly we obtain that:

$$\|g(\boldsymbol{s}^t, \boldsymbol{\mu}^t)\|_2 \leq e^{-\frac{\lambda^2}{2n^2L^2}\cdot t} \cdot \|g(\boldsymbol{s}^0, \boldsymbol{\mu}^0)\|_2.$$

Thus if we want the error $\|g(\boldsymbol{s}^t, \boldsymbol{\mu}^t)\|_2 \leq \varepsilon$, $T = \frac{2n^2L^2}{\lambda^2} \log\left(\frac{\|g(\boldsymbol{s}^0, \mu^0)\|_2}{\varepsilon}\right)$ iterations suffice, as claimed. □

B Appendix to Section 4

Lemma 1. *The equation* $C\beta^2 - (An(n-2) + C)\beta + A(n-1)^2 = 0$ *of* (6) *has a real root* β^* *where* $0 \leq \beta^* \leq 1 - 1/n$.

Proof. Using the quadratic formula, we see that β^* given by:

$$\beta^* = \frac{An(n-2) + C - \sqrt{(An(n-2) + C)^2 - 4AC(n-1)^2}}{2C} \tag{14}$$

We first argue β^* is real, by showing $(An(n-2) + C)^2 - 4AC(n-1)^2 \geq 0$. This is equivalent to showing $q(y) := (y + n(n-2))^2 - 4(n-1)^2y \geq 0$, where $y = C/A$. Expanding q, we have $q(y) = y^2 - 2(n^2 - 2n + 2)y + n^2(n-2)^2$. The roots of q are:

$$y_1, y_2 = \frac{2(n^2 - 2n + 2) \pm \sqrt{4(n^2 - 2n + 2)^2 - 4n^2(n-2)^2}}{2} = (n^2 - 2n + 2) \pm 2(n-1),$$

i.e., $y_1 = (n-2)^2$ and $y_2 = n^2$. Since $q(y)$ has a positive leading coefficient, we have that $q(y) \geq 0$ for all $y \geq y_2 = n^2$. Thus it remains to show that $y = C/A \geq n^2$. To see this, we use the AM-HM inequality:

$$\frac{C}{n} = \frac{c_1 + \cdots + c_n}{n} \geq \frac{n}{\frac{1}{c_1} + \cdots + \frac{1}{c_n}} = \frac{n}{A}, \tag{15}$$

implying $C/A \geq n^2$ as desired. This shows that the root β^* of equation (6) is real, hence well-defined.

We now show $0 \leq \beta^* \leq 1 - 1/n$. From (14), we see:

$$\begin{aligned} \beta^* &= \frac{An(n-2) + C - \sqrt{(An(n-2)+C)^2 - 4AC(n-1)^2}}{2C} \\ &\geq \frac{An(n-2) + C - \sqrt{(An(n-2)+C)^2}}{2C} = 0 \end{aligned}$$

Further, from (14) we also have:

$$\begin{aligned} \beta^* &= \frac{An(n-2) + C - \sqrt{(An(n-2)+C)^2 - 4AC(n-1)^2}}{2C} \\ &\leq \frac{An(n-2) + C}{2C} = \frac{Cn(n-2)/n^2 + C}{2C} = 1 - \frac{1}{n}, \end{aligned}$$

where we used $A/C \leq 1/n^2$ (15) in the last inequality. This concludes the proof of Lemma 1. $\square$

Theorem 4.1. *For each $\beta \in [0,1]$, the mechanism $\mathcal{M}_\beta$ admits a Nash equilibrium. For $\beta = \beta^*$ (Definition 2), the NE of $\mathcal{M}_{\beta^*}$ also maximizes the p-mean welfare for any $p \leq 1$. Additionally, any NE $\boldsymbol{s}^*$ with $\boldsymbol{s}^* > 0$ maximizes the p-mean welfare.*

Proof. When $0 \leq \beta \leq 1$, the program (5) is a convex program for general convex cost functions. Since $u_i(\cdot)$ is concave, a proof similar to the proof of Theorem 3.1 shows the existence of a Nash equilibrium.

We now show the welfare-maximizing property. For simplicity, we only consider feasible strategies where each agent participates in the mechanism, i.e., $s_i > 0$. Let ρ_i and λ_i as the dual variables to the first and second constraints respectively for each i, and let $S = \|\boldsymbol{s}\|_1$. Writing the KKT conditions and eliminating all ρ_i, we get that a NE $(\boldsymbol{b}^*, \boldsymbol{s}^*)$ together with dual variables λ^* satisfies:

$$\forall i: \quad \frac{\partial u_i(b_i^*, S^*)}{\partial S} = (1-\beta) \cdot c_i \cdot \left(\frac{\partial u_i(b_i^*, S^*)}{\partial b_i} + \lambda_i^* \right) \quad \text{(from stationarity conditions)} \tag{16}$$

$$\forall i: \quad \lambda_i^* \geq 0 \quad \text{(dual feasibility)} \tag{17}$$

$$\forall i: \quad \lambda_i^* \cdot b_i = 0 \quad \text{(complimentary slackness)} \tag{18}$$

Now we turn to the p-mean welfare maximizing solution which is an optimal solution to the following program.

$$\begin{aligned} \max \quad & W_p(\boldsymbol{b}, \boldsymbol{s}) := (\sum_i u_i(b_i, \|s\|_1)^p)^{1/p} \\ \text{s.t.} \quad & \forall i: \ b_i + (1-\beta) c_i(s_i) + \frac{\beta}{n-1} \sum_{j \neq i} c_j(s_j) = B_i \\ & \forall i: \ b_i \geq 0 \end{aligned} \tag{19}$$

The following lemma establishes that (19) is a convex program. For ease of readability we defer its proof to B.1.

Lemma 2. *For $\beta \in [0,1]$ and $p \leq 1$, the program* (19) *is convex.*

We can now write the KKT conditions of program (19). By letting μ_i and γ_i denote the dual variables corresponding to the first and second constraints respectively for each i and $S = \|\boldsymbol{s}\|_1$, the KKT conditions (considering only solutions with $s_i > 0$) are:

$$\forall i: \quad (\sum_j u_j^p)^{1/p-1} \sum_k u_k^{p-1} \frac{\partial u_k}{\partial S} = c_i \cdot [\mu_i (1-\beta) + \frac{\beta}{n-1} \sum_{k \neq i} \mu_k] \quad \text{(stationarity)} \tag{20}$$

$$\forall i: \quad (\sum_j u_j^p)^{1/p-1} u_i^{p-1} \frac{\partial u_i}{\partial b_i} = \mu_i - \gamma_i \qquad \text{(stationarity)} \qquad (21)$$

$$\forall i: \quad \gamma_i \geq 0 \qquad \text{(dual feasibility)} \qquad (22)$$

$$\forall i: \quad \gamma_i \cdot b_i = 0 \qquad \text{(complimentary slackness)} \qquad (23)$$

Since KKT conditions are sufficient for optimality, to prove Theorem 4.1 it suffices to show that for an NE $(\boldsymbol{b}^*, \boldsymbol{s}^*)$, there exist dual variables $\boldsymbol{\mu}^*$ and $\boldsymbol{\gamma}^*$ which satisfy (20)-(23) for $\beta = \beta^*$.

Let $\alpha := (\sum_j u_j(b_j^*, \boldsymbol{s}^*)^p)^{1/p-1} \sum_k u_k(b_j^*, \boldsymbol{s}^*)^{p-1} \frac{\partial u_k(b_k^*, \boldsymbol{s}^*)}{\partial S}$, i.e., the common value of the equality (20) at the NE $(\boldsymbol{b}^*, \boldsymbol{s}^*)$. The equation (20) then becomes $\alpha \cdot c_i^{-1} = \mu_i(1-\beta) + \frac{\beta}{n-1} \sum_{k \neq i} \mu_k$. Summing these over all i and letting $T = \sum_j \mu_j$, we obtain:

$$\alpha \cdot (\sum_i c_i^{-1}) = \sum_i [\mu_i(1-\beta) + \frac{\beta}{n-1} \sum_{k \neq i} \mu_k] = T.$$

Putting this back in (20), we obtain the following expression for μ_i^*, which can be computed from the NE $(\boldsymbol{b}^*, \boldsymbol{s}^*)$ with $T = \alpha \cdot (\sum_i c_i^{-1})$:

$$\mu_i^* = \frac{\frac{T c_i^{-1}}{\sum_i c_i^{-1}} - \frac{\beta T}{n-1}}{1 - \frac{\beta n}{n-1}}. \qquad (24)$$

Recall that the NE $(\boldsymbol{b}^*, \boldsymbol{s}^*)$ satisfies (16)-(18) for some dual variables λ^*. We define γ_i^* as follows:

$$\gamma_i^* = \mu_i^* \cdot \left(\frac{\lambda_i^*}{\lambda_i^* + \frac{\partial u_i(b_i^*, \boldsymbol{s})}{\partial b_i}} \right) \qquad (25)$$

The next lemma proves Theorem 4.1.

Lemma 3. *A NE $(\boldsymbol{b}^*, \boldsymbol{s}^*)$ with $\boldsymbol{\mu}^*$ and $\boldsymbol{\gamma}^*$ defined by* (24) *and* (25) *satisfy the KKT conditions* (20)-(23) *of program* (19).

Proof. First observe that at the NE, $(1-\beta)c_i \cdot \left(\frac{\partial u_i(b_i^*, S^*)}{\partial b_i} + \lambda_i^* \right) = \frac{\partial u_i(b_i^*, S^*)}{\partial S} > 0$ by assumption. Since $\beta \in (0,1)$ and $c_i > 0$, we have $\frac{\partial u_i(b_i^*, S^*)}{\partial b_i} + \lambda_i^* > 0$. Together with $\lambda_i^* \geq 0$ (17), this shows $\gamma_i^* \geq 0$ thus satisfying dual feasibility (22).

Next we show complimentary slackness (23) holds. For any i, $\lambda_i^* \cdot b_i = 0$ due to (18). Then by the definition of γ_i^*, we have $\gamma_i^* \cdot b_i = 0$ for all i.

Finally, we show that equations (20) and (21) are satisfied for a specific choice of $\beta = \beta^*$. Together, (20) and (21) imply that an optimal solution to program (19) satisfies:

$$\forall i: \sum_k (\mu_k - \gamma_k) \cdot \frac{\partial u_k / \partial S}{\partial u_k / \partial b_k} = c_i \cdot [\mu_i(1-\beta) + \frac{\beta}{n-1} \sum_{k \neq i} \mu_k] \qquad (26)$$

The choice of γ_i^* from equation 25 implies that $\mu_i^* - \gamma_i^* = \mu_i^* \cdot (\frac{\partial u_i(b_i^*, \boldsymbol{s})/\partial b_i}{\partial u_i(b_i^*, \boldsymbol{s})/\partial b_i + \lambda_i^*})$. Moreover at the NE, equation (16) implies that:

$$\begin{aligned}(\mu_i^* - \gamma_i^*) \cdot \frac{\partial u_i(b_i^*, \boldsymbol{s})/\partial S}{\partial u_i(b_i^*, \boldsymbol{s})/\partial b_i} &= \mu_i^* \cdot \left(\frac{\partial u_i(b_i^*, \boldsymbol{s})/\partial b_i}{\partial u_i(b_i^*, \boldsymbol{s})/\partial b_i + \lambda_i^*} \right) \cdot (1-\beta)c_i \cdot \left(1 + \frac{\lambda_i^*}{\partial u_i(b_i^*, \boldsymbol{s})/\partial b_i} \right) \\ &= \mu_i^* \cdot (1-\beta)c_i.\end{aligned}$$

Using the above in (26), it only remains to be argued that $\boldsymbol{\mu}^*$, $\boldsymbol{b}^*$ and $\boldsymbol{s}^*$ satisfy:

$$\forall i: (1-\beta) \cdot \sum_k \mu_k^* \cdot c_k = c_i \cdot [\mu_i^*(1-\beta) + \frac{\beta}{n-1} \sum_{k \neq i} \mu_k^*] = \alpha,$$

for $\beta = \beta^*$. By plugging in the value of μ_i^* from (24) and using $\alpha = T \cdot (\sum_k c_k^{-1})^{-1}$, we get:

$$(1-\beta) \cdot \sum_k \left\{ \frac{T c_k^{-1} (\sum_i c_i^{-1})^{-1} - \frac{\beta T}{n-1}}{1 - \frac{\beta n}{n-1}} \right\} \cdot c_k = T \cdot (\sum_k c_k^{-1})^{-1}.$$

Let us define $A := (\sum_i c_i^{-1})^{-1}$ and $C := \sum_i c_i$. Manipulating the above expression, the above equation then becomes:

$$C\beta^2 - (An(n-2) + C)\beta + A(n-1)^2 = 0,$$

which is true for $\beta = \beta^*$ since it is exactly the definition of β^* (Definition 2).

Thus for $\beta = \beta^*$, the NE $(\boldsymbol{b}^*, \boldsymbol{s}^*)$ with dual variables $\boldsymbol{\mu}$ and $\boldsymbol{\gamma}$ as defined in (24) and (25) respectively satisfy the KKT conditions of program (19). □

□

B.1 Proof of Lemma 2

Lemma 2. *For $\beta \in [0,1]$ and $p \leq 1$, the program* (19) *is convex.*

Proof. For $\beta \in [0,1]$ the constraints of program 19 are convex since $c_i(\cdot)$ are convex functions. It remains to be shown that the objective $W_p(\boldsymbol{b}, \boldsymbol{s}) := (\sum_i u_i(b_i, \|s\|_1)^p)^{1/p}$ to be maximized is concave.

We use the following standard fact about the concavity of composition of functions (see e.g. Boyd and Vandenberghe [2004], Page 86).

Proposition 2. *Let $h : \mathbb{R}^n \to \mathbb{R}$ and $g_i : \mathbb{R}^k \to \mathbb{R}$ and let $f : \mathbb{R}^n \to \mathbb{R}$ be given by $f(x) = h(g(x)) = h(g_1(x), \ldots, g_n(x))$. Then f is concave if h is concave, h is non-decreasing in each argument and g_i are concave.*

Note that $W_p(\boldsymbol{b}, \boldsymbol{s}) = h(g(\boldsymbol{b}, \boldsymbol{s}))$, where $h(x_1, \ldots, x_n) = (\sum_i x_i^p)^{1/p}$ and $g_i(\boldsymbol{b}, \boldsymbol{s}) = u_i(\boldsymbol{b}, \boldsymbol{s})$.

We now observe that:

- h is non-decreasing in each argument. This is because:

$$\frac{\partial h}{\partial x_i} = h^{1-p} x_i^{p-1} \geq 0.$$

- h is concave. Using the above, we can compute the Hessian H given by:

$$H_{ij} = \frac{\partial^2 h}{\partial x_j \partial x_i} = \begin{cases} (1-p) h^{1-2p} (x_i x_j)^{p-1} & \text{(if } i \neq j) \\ (1-p) h^{1-2p} x_i^{p-2} \cdot (x_i^p - h^p) & \text{(if } i = j) \end{cases}$$

  Thus for any $v \in \mathbb{R}^n$, we have:

$$\begin{aligned} v^T H v &= \sum_i \sum_j v_i H_{ij} v_j \\ &= (1-p) h^{1-2p} \cdot \left(\sum_i v_i \sum_{j \neq i} H_{ij} v_j + \sum_i v_i^2 H_{ii} \right) \\ &= (1-p) h^{1-2p} \cdot \left(\sum_i v_i x_i^{p-1} \cdot \left((\sum_j v_j x_j^{p-1}) - v_i x_i^{p-1} \right) + \sum_i v_i^2 (x_i^{2p-2} - h^p x_i^{p-2}) \right) \\ &= (1-p) h^{1-2p} \cdot \left((\sum_i v_i x_i^{p-1})^2 - \sum_i (v_i x_i^{p-1})^2 + \sum_i v_i^2 x_i^{2p-2} - \sum_i v_i^2 h^p x_i^{p-2} \right) \\ &= (1-p) h^{1-2p} \cdot \left((\sum_i v_i x_i^{p-1})^2 - (\sum_i v_i^2 x_i^{p-2})(\sum_j x_j^p) \right) \\ &\leq 0, \end{aligned}$$

Algorithm 1 FedBR-BG

Input: Number of iterations in game H, number of iterations of gradient descent T, learning rate α, step size δ, data increasing interval Δs
Output: Model weights θ^T, individual contributions $\boldsymbol{s}$
for $h = 1, 2, \cdots, H$ **do**
  Server sends θ^t to agents;
  for $t = 0, 1, \cdots, T-1$ **do**
    for $i \in [n]$ **in parallel do**
      i computes $\nabla_{\theta^t}\mathcal{L}_i(\theta^t)$ on its local dataset $\mathcal{D}_i$;
      i sends $\nabla_{\theta^t}\mathcal{L}_i(\theta^t)$ to server;
    end for
    Server aggregates the gradients following

$$\nabla_{\theta^t}\mathcal{L}(\theta^t) \leftarrow \frac{1}{\sum_{i\in[n]}|\mathcal{D}_i|}\sum_{i\in[n]}|\mathcal{D}_i| \cdot \nabla_{\theta^t}\mathcal{L}_i(\theta^t);$$

    Server updates θ^{t+1} following

$$\theta^{t+1} \leftarrow \theta^t - \alpha \cdot \nabla_{\theta^t}\mathcal{L}(\theta^t);$$

  end for
  for $i \in [n]$ **in parallel do**
    $\frac{\partial u_i}{\partial s_i} \leftarrow \frac{a(\sum_i s_i + \Delta s) - a(\sum_i s_i)}{\Delta s} - (1-\beta)c_i$
    if ($s_i = 0$ **and** $\frac{\partial u_i}{\partial s_i} < 0$) **or** ($s_i = \tau_i$ **and** $\frac{\partial u_i}{\partial s_i} > 0$) **then**
      $s_i^{h+1} \leftarrow s_i^h$;
    else
      $s_i^{h+1} = s_i^h + \delta \cdot \frac{\partial u_i}{\partial s_i}$;
    end if
  end for
end for

since $p \le 1$, $h \ge 0$, and by the Cauchy-Schwarz inequality $(\sum_i a_i \cdot b_i)^2 \le (\sum_i a_i^2)\cdot(\sum_i b_i^2)$ with $a_i = v_i x_i^{p/2-1}$ and $b_i = x_i^{p/2-1}$. Thus H is negative semi-definite and hence h is concave.

- For each i, $g_i(\boldsymbol{b}, \boldsymbol{s}) = u_i(\boldsymbol{b}, \boldsymbol{s})$ is concave.

Using Proposition 2 and the fact that $W_p(\boldsymbol{b}, \boldsymbol{s}) = h(g(\boldsymbol{b}, \boldsymbol{s}))$ we conclude that $W_p(\boldsymbol{b}, \boldsymbol{s})$ is concave. □

C Distributed Algorithms

In this section, we present the distributed algorithms of our two mechanisms, FedBR and FedBR-BG.

D Additional Results

We present the results of our method on CIFAR-10 in Table 2.

Table 2: p-mean welfare of our budget-balanced mechanism FedBR-BG and baselines on CIFAR-10. We report the results for different p. The cost for adding one data sample c_i is 0.005 for every agent.

Method	p				
	0.2	0.4	0.6	0.8	1.0
FedAvg	42386.21	135.92	23.528	8.381	4.582
FedBR	58297.23	178.32	26.187	9.675	5.681
FedBR-BG	**60385.32**	**183.23**	**27.958**	**9.981**	**5.891**

Algorithm 2 FedBR

Input: Number of iterations in game H, number of iterations of gradient descent T, learning rate α, step size δ, data increasing interval Δs

Output: Model weights θ^T, individual contributions $\boldsymbol{s}$

for $h = 1, 2, \cdots, H$ **do**
  Server sends θ^t to agents;
  for $t = 0, 1, \cdots, T-1$ **do**
    for $i \in [n]$ **in parallel do**
      i computes $\nabla_{\theta^t}\mathcal{L}_i(\theta^t)$ on its local dataset $\mathcal{D}_i$;
      i sends $\nabla_{\theta^t}\mathcal{L}_i(\theta^t)$ to server;
    end for
    Server aggregates the gradients following

$$\nabla_{\theta^t}\mathcal{L}(\theta^t) \leftarrow \frac{1}{\sum_{i\in[n]}|\mathcal{D}_i|}\sum_{i\in[n]}|\mathcal{D}_i| \cdot \nabla_{\theta^t}\mathcal{L}_i(\theta^t);$$

    Server updates θ^{t+1} following

$$\theta^{t+1} \leftarrow \theta^t - \alpha \cdot \nabla_{\theta^t}\mathcal{L}(\theta^t);$$

  end for
  for $i \in [n]$ **in parallel do**
    $\frac{\partial u_i}{\partial s_i} \leftarrow \frac{a(\sum_i s_i + \Delta s) - a(\sum_i s_i)}{\Delta s} - c_i$
    if ($s_i = 0$ **and** $\frac{\partial u_i}{\partial s_i} < 0$) **or** ($s_i = \tau_i$ **and** $\frac{\partial u_i}{\partial s_i} > 0$) **then**
      $s_i^{h+1} \leftarrow s_i^h$;
    else
      $s_i^{h+1} = s_i^h + \delta \cdot \frac{\partial u_i}{\partial s_i}$;
    end if
  end for
end for