A. Real-Valued Strategies

In the examples of random coverage and general PAC learning, it is common to consider integral values of θ_i. For a real-valued θ_i, we consider one natural interpretation: randomized rounding over $\lfloor\theta_i\rfloor$ and $\lceil\theta_i\rceil$. More specifically, let agent i randomly draw an integral value $m_i \sim \sigma(\theta_i)$, where $\sigma(\theta_i) = \lfloor\theta_i\rfloor + \mathrm{Ber}(\theta_i - \lfloor\theta_i\rfloor)$, and then uses m_i as her strategy. Then the utility function is defined by taking expectation over $\mathbf{m} = (m_1, \ldots, m_k)$. That is,

$$u_i(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{m}}\left[1 - \frac{1}{2}\sum_{x\in\mathcal{X}} q_{ix}\prod_{j=1}^{k}(1-q_{jx})^{m_j}\right].$$

Similarly, we define the utility function in general PAC learning as

$$u_i(\boldsymbol{\theta}) = 1 - \mathbb{E}_{\mathbf{m}}\left[\mathbb{E}_{\{S_j\sim\mathcal{D}_j^{m_j}\}_{j\in[k]}}\left[\mathrm{err}_{\mathcal{D}_i}(h_S)\right]\right].$$

Note that these definitions work for integral-valued θ_i as well.

B. Calculation of Well-behaved Property

Linear Utilities. The linear utilities are well-behaved over any $\bigtimes_{i=1}^{k}[0, C_i] \subseteq \Theta$. Agent i's utility increases at a constant rate $\partial\theta_i(\boldsymbol{\theta})/\partial\theta_i = W_{ii} = 1$ when the agent increases its strategy unilaterally and increases at rate $\partial\theta_i(\boldsymbol{\theta})/\partial\theta_j = W_{ij} \le 1$ when agent j increases its strategy unilaterally.

Random Coverage. For any $\bigtimes_{i=1}^{k}[0, C_i+1] \subseteq \Theta$, if $u_i(C_i+1, \mathbf{C}_{-1}) - u_i(\mathbf{C})$ is bounded away from 0 for all i, then the utilities are well-behaved over $\bigtimes_{i=1}^{k}[0, C_i]$, where $\mathbf{C} = (C_1, \ldots, C_k)$. At a high level, the smallest impact that an additional sample by agent i has on u_i is when $\boldsymbol{\theta} \to \mathbf{C}$. This impact is at least $u_i(C_i+1, \mathbf{C}_{-i}) - u_i(\mathbf{C}) > 0$. On the other hand, $\partial u_i(\boldsymbol{\theta})/\partial\theta_j$ is bounded above, because the marginal impact of any one sample on u_i is largest when no agent has yet taken a sample.

First, by direct calculation, we have that for any non-integral θ_j,

$$\begin{aligned}
\frac{\partial u_i(\boldsymbol{\theta})}{\partial\theta_j} &= -\frac{1}{2}\frac{\partial\sum_{x\in\mathcal{X}} q_{ix}\prod_{l=1}^{k}\mathbb{E}\left[(1-q_{lx})^{m_l}\right]}{\partial\theta_j} \\
&= -\frac{1}{2}\frac{\partial\sum_{x\in\mathcal{X}} q_{ix}\prod_{l\neq j}\mathbb{E}\left[(1-q_{lx})^{m_l}\right]\left((\theta_j-\lfloor\theta_j\rfloor)(1-q_{jx})^{\lfloor\theta_j\rfloor+1} + (1+\lfloor\theta_j\rfloor-\theta_j)(1-q_{jx})^{\lfloor\theta_j\rfloor}\right)}{\partial\theta_j} \\
&= -\frac{1}{2}\sum_{x\in\mathcal{X}} q_{ix}\prod_{l\neq j}\mathbb{E}\left[(1-q_{lx})^{m_l}\right]\Big((1-q_{jx})^{\lfloor\theta_j\rfloor+1} - (1-q_{jx})^{\lfloor\theta_j\rfloor}\Big) \\
&= u_i(\lfloor\theta_j\rfloor+1, \boldsymbol{\theta}_{-j}) - u_i(\lfloor\theta_j\rfloor, \boldsymbol{\theta}_{-j}).
\end{aligned}$$

For integral-value θ_j, when we increase θ_j by a small amount $\varepsilon \in (0,1)$, $\alpha = \lfloor\theta_j+\varepsilon\rfloor = \lfloor\theta_j\rfloor$ does not change. Then we have

$$\begin{aligned}
\frac{\partial_+ u_i(\boldsymbol{\theta})}{\partial\theta_j} &= -\frac{1}{2}\frac{\partial_+\sum_{x\in\mathcal{X}} q_{ix}\prod_{l\neq j}\mathbb{E}\left[(1-q_{lx})^{m_l}\right]\left((\theta_j-\alpha)(1-q_{jx})^{\alpha+1} + (1+\alpha-\theta_j)(1-q_{jx})^{\alpha}\right)}{\partial\theta_j} \\
&= u_i(\alpha+1, \boldsymbol{\theta}_{-j}) - u_i(\alpha, \boldsymbol{\theta}_{-j}) \\
&= u_i(\theta_j+1, \boldsymbol{\theta}_{-j}) - u_i(\boldsymbol{\theta}).
\end{aligned}$$

When we decrease θ_j by ε, $\alpha = \lfloor\theta_j-\varepsilon\rfloor = \lfloor\theta_j-1\rfloor$. Then for all $x \in [\theta_j-1, \theta_j]$, we can represent

$$u_i(x, \boldsymbol{\theta}_{-1}) = 1 - \frac{1}{2}\sum_{x\in\mathcal{X}} q_{ix}\prod_{l\neq j}\mathbb{E}\left[(1-q_{lx})^{m_l}\right]\left((x-\alpha)(1-q_{jx})^{\alpha+1} + (1+\alpha-x)(1-q_{jx})^{\alpha}\right).$$

Thus we have

$$\frac{\partial_- u_i(\boldsymbol{\theta})}{\partial\theta_j} = -\frac{1}{2}\frac{\partial_-\sum_{x\in\mathcal{X}} q_{ix}\prod_{l\neq j}\mathbb{E}\left[(1-q_{lx})^{m_l}\right]\left((\theta_j-\alpha)(1-q_{jx})^{\alpha+1} + (1+\alpha-\theta_j)(1-q_{jx})^{\alpha}\right)}{\partial\theta_j}$$

$$
\begin{aligned}
&= u_i(\alpha+1, \boldsymbol{\theta}_{-j}) - u_i(\alpha, \boldsymbol{\theta}_{-j}) \\
&= u_i(\boldsymbol{\theta}) - u_i(\theta_j - 1, \boldsymbol{\theta}_{-j})\,.
\end{aligned}
$$

Then we argue that for any $\boldsymbol{\theta} \in \times_{i=1}^{k}[0, C_i+1]$, any $t \in \mathbb{N} \cap [0, C_i+1]$, $u_i(t+1, \boldsymbol{\theta}_{-j}) - u_i(t, \boldsymbol{\theta}_{-j}) = \frac{1}{2}\sum_{x\in\mathcal{X}} q_{ix} \prod_{l\neq j} \mathbb{E}\left[(1-q_{lx})^{m_l}\right] q_{jx}(1-q_{jx})^t$ is non-increasing with respect to t and with respect to θ_l for any $l \neq j$.

Combining the computing results on sub-gradients and the monotonicity of $u_i(t+1, \boldsymbol{\theta}_{-j}) - u_i(t, \boldsymbol{\theta}_{-j})$, we know that

$$
\frac{\partial u_i(\boldsymbol{\theta})}{\partial \theta_i} \geq u_i(C_i+1, \boldsymbol{\theta}_{-1}) - u_i(C_i, \boldsymbol{\theta}_{-1}) \geq u_i(C_i+1, \mathbf{C}_{-1}) - u_i(\mathbf{C})\,,
$$

and

$$
\frac{\partial u_i(\boldsymbol{\theta})}{\partial \theta_j} \leq u_i(1, \boldsymbol{\theta}_{-j}) - u_i(0, \boldsymbol{\theta}_{-j}) \leq u_i(1, \mathbf{0}_{-j}) - u_i(0, \mathbf{0}_{-j}) \leq \frac{1}{2}\sum_{x\in\mathcal{X}} q_{ix} q_{jx} \leq \frac{1}{2}\,.
$$

General PAC Learning. In the previous two examples, the utilities are well-behaved over any bounded convex set. However, this might not be true in the general PAC learning case. For example, recall the example in the proof of Theorem 3 and let us extend the strategy space Θ from $\{0,1\}^3$ to $[0,1]^3$ by the randomized rounding method as aforementioned, i.e.,

$$
u_i(\boldsymbol{\theta}) = \frac{1}{2}(1 + \theta_i + \theta_{i\ominus 1} - \theta_i\theta_{i\ominus 1})\,.
$$

Then the utility function is ill-behaved over $[0,1]^3$ since $\partial u_i(\boldsymbol{\theta})/\partial\theta_i = 0$ when $\theta_{i\ominus 1} = 1$. However, it is easy to check that for any $C \in [0,1)$, the utility function is well-behaved over $[0,C]^3$.

C. Proof of Lemma 1

Lemma 1. *If utilities are well-behaved over $\times_{i=1}^{k}[0, \vartheta_i]$, the best-response function $\mathbf{f}$ has a fixed point, i.e., $\exists \boldsymbol{\theta} \in \times_{i=1}^{k}[0, \vartheta_i], \mathbf{f}(\boldsymbol{\theta}) = \boldsymbol{\theta}$.*

Proof. The celebrated Brouwer fixed-point theorem states that any continuous function on a compact and convex subset of $\mathbb{R}^k$ has a fixed point. First note that $\mathbf{f}$ is a well-defined map from $\times_{i=1}^{k}[0, \vartheta_i]$ to $\times_{i=1}^{k}[0, \vartheta_i]$, which is a convex and compact subset of $\mathbb{R}^k$. All that is left to show is that $\mathbf{f}$ is a continuous function over $\times_{i=1}^{k}[0, \vartheta_i]$.

At a high level, f is continuous because in well-behaved utility functions a small change in other agents' contributions affect the utility of agent i only by a small amount, so a small adjustment to agent i's contribution will be sufficient to meet his constraint. More formally, we show that for any $\boldsymbol{\delta} \in \mathbb{R}^k$ with $\|\boldsymbol{\delta}\|_1 \leq 1$, $\lim_{\varepsilon\to 0}|f_i(\boldsymbol{\theta}) - f_i(\boldsymbol{\theta} + \varepsilon\boldsymbol{\delta})| = 0$. Define $\boldsymbol{\theta}' = \boldsymbol{\theta} + \varepsilon\boldsymbol{\delta}$, $x = f_i(\boldsymbol{\theta})$, and $x' = f_i(\boldsymbol{\theta}')$. For every i, we have

$$
u_i\left(x' + \frac{c_1^i\varepsilon}{c_2^i}, \boldsymbol{\theta}_{-i}\right) \geq u_i\left(x', \boldsymbol{\theta}_{-i}\right) + c_1^i\varepsilon \geq u_i\left(x', \boldsymbol{\theta}_{-i}\right) + c_1^i\varepsilon\|\boldsymbol{\delta}_{-i}\|_1 \geq u_i(x', \boldsymbol{\theta}_{-i} + \varepsilon\boldsymbol{\delta}_{-i}) \geq \mu_i\,,
$$

where the first and third transitions are by the definition of well-behaved functions, and the last transition is by the definition of $\boldsymbol{\theta}'$ and x'. This shows that $x \leq x' + \frac{c_1^i\varepsilon}{c_2^i}$. Similarly,

$$
u_i\left(x + \frac{c_1^i\varepsilon}{c_2^i}, (\boldsymbol{\theta} + \varepsilon\boldsymbol{\delta})_{-i}\right) \geq u_i(x, (\boldsymbol{\theta} + \varepsilon\boldsymbol{\delta})_{-i}) + c_1^i\varepsilon \geq u_i(x, (\boldsymbol{\theta} + \varepsilon\boldsymbol{\delta})_{-i}) + c_1^i\varepsilon\left\|-\boldsymbol{\delta}_{-i}\right\|_1 \geq u_i(x, \boldsymbol{\theta}_{-i}) \geq \mu_i\,,
$$

which indicates that $x + \frac{c_1^i\varepsilon}{c_2^i} \geq x'$. Hence, we have $|x - x'| \leq \frac{c_1^i\varepsilon}{c_2^i}$. Therefore, $\mathbf{f}$ is continuous over $\times_{i=1}^{k}[0, \vartheta_i]$.

The proof follows by applying the Brouwer Fixed-Point Theorem. $\square$

D. More General Construction for Theorem 3

We extend the simple example in Section 3.3 into a more general one.

Consider the domain $\mathcal{X} = \{0, \ldots, 6d-1\}$ for any $d > 1$ and the label space $\mathcal{Y} = \{0, 1\}$. We consider agents $\{0, 1, 2\}$ with distributions $\mathcal{D}_0, \mathcal{D}_1, \mathcal{D}_2$ over $\mathcal{X} \times \mathcal{Y}$. Similar to the example in Section 3.3, we give a probabilistic construction for $\mathcal{D}_0, \mathcal{D}_1, \mathcal{D}_2$. Take independent random variables $\mathbf{Z}_0, \mathbf{Z}_1, \mathbf{Z}_2$ that are each uniform over $\{0,1\}^d$. For each $i \in \{0,1,2\}$, distribution $\mathcal{D}_i$ is a uniform distribution over instance-label pairs $\{((2i + z_{i,j})d + j, z_{i\ominus 1,j})\}_{j=0}^{d-1}$. In other words, the marginal distribution of $\mathcal{D}_i$ is a uniform distribution over $\mathcal{X}_i = \{x_1, \ldots, x_d\}$ where x_j is equally likely to be $2id + j$ or $(2i+1)d + j$ and independent of other x_l for $l \neq j$. Moreover, the labels of points in distribution $\mathcal{D}_{i\oplus 1}$ are decided according to the marginal distribution of $\mathcal{D}_i$: if the support of the marginal distribution of $\mathcal{D}_i$ contains $2id + j$, then the points $2(i \oplus 1)d + j$ and $(2(i\oplus 1) + 1)d + j$ are both labeled 0, and if the support of the marginal distribution of $\mathcal{D}_i$ contains $(2i+1)d + j$, then the points $2(i \oplus 1)d + j$ and $(2(i \oplus 1) + 1)d + j$ are both labeled 1.

Consider the optimal classifier conditioned on the event where agent i takes samples $\{((2i + z_{i,j})d + j, z_{i\ominus 1,j})\}_{j \in J_i}$ from $\mathcal{D}_i$ for all i. This reveals $z_{i,j}$ and $z_{i\ominus 1,j}$ for all $j \in J_i$. Therefore, the optimal classifier conditioned on this event classifies $(2i + z_{i,j})d + j$ for each $j \in J_i \cup J_{i\ominus 1}$ correctly and misclassifies $(2i + z_{i,j})d + j$ for each $j \notin J_i \cup J_{i \ominus 1}$ with probability $1/2$.

Now we formally define the strategy space and the utility functions that corresponding to this setting. Let $\Theta = \mathbb{N}^3$ to be the set of strategies in which each agent can take any integral number of samples. Let $u_i(\boldsymbol{\theta})$ be the expected accuracy of the optimal classifier given the samples taken at random under $\boldsymbol{\theta}$. As a consequence of the above analysis,

$$u_i(\boldsymbol{\theta}) = 1 - \frac{1}{2d}\sum_{j=0}^{d-1}\left(1 - \frac{1}{d}\right)^{\theta_i + \theta_{i\ominus 1}} = 1 - \frac{1}{2}\left(1 - \frac{1}{d}\right)^{\theta_i + \theta_{i\ominus 1}} .$$

Then let $\boldsymbol{\mu} = \mu\mathbf{1}$ for any $\mu \in (1/2, 1)$ such that $m(\mu) := \left\lceil \frac{\log(2(1-\mu))}{\log(1-1/d)} \right\rceil$ is an odd number. It is easy to find such a μ: arbitrarily pick a $\mu' \in (1/2, 1)$; if $m(\mu')$ is odd, let $\mu = \mu'$; otherwise let $\mu = (1-1/d)\mu' + 1/d$ such that $m(\mu) = m(\mu') + 1$.

Agent i's constraint is satisfied when $\theta_i + \theta_{i\ominus 1} \geq m(\mu)$ and is not satisfied when $\theta_i + \theta_{i\ominus 1} \leq m(\mu) - 1$. If $\theta_i + \theta_{i\ominus 1} \geq m(\mu) + 1$, agent i can unilaterally decrease her strategy by 1 and still meet her constraint. Therefore, we have

$$\theta_i + \theta_{i\ominus 1} = m(\mu), \forall i = 0, 1, 2 .$$

This results in $\theta_0 = \theta_1 = \theta_2$, which is impossible as $m(\mu)$ is odd and θ_i is integral for all i. Hence, no stable equilibrium over $\Theta = \mathbb{N}^3$ exists.

E. Proof of Theorem 4

Theorem 4. *There is a collaborative learning setting with well-behaved utility functions such that the Price of Stability and Price of Fairness are at least $\Omega(\sqrt{k})$. Moreover, these utilities correspond to two settings: a) a random domain coverage example with uniform distributions over equally sized subsets and b) a linear utility setting with $W_{ii} = 1$ and $W_{ij} \in O(1/\sqrt{k})$ for $j \neq i$.*

Proof. Consider a family of sets each of size $b = k - 1$ demonstrated in Figure 3, where there is one *core agent* that owns b central points and $k - 1$ *petal agents* whose sets intersect with that of the core agent. More formally, let agent 0 be the core agent whose distribution is uniform over the points $\mathcal{X}_0 = \{1, \ldots, b\}$. Partition $\mathcal{X}_0 = \{1, \ldots, b\}$ to $\sqrt{b}$ equally sized groups of $\sqrt{b}$ instances $\mathcal{X}_0^1, \ldots, \mathcal{X}_0^{\sqrt{b}}$. Similarly, partition the $b = k-1$ agents to $\sqrt{b}$ equally sized groups of $\sqrt{b}$ agents $I_1, \ldots, I_{\sqrt{b}}$. Each $i \in I_j$ has uniform distribution over the set $\mathcal{X}_i = \mathcal{X}_0^j \cup \mathcal{O}_i$, where $\mathcal{O}_i$ is a set of $b - \sqrt{b}$ points that are unique to i. The strategy space is $\Theta = \mathbb{R}_+^k$.

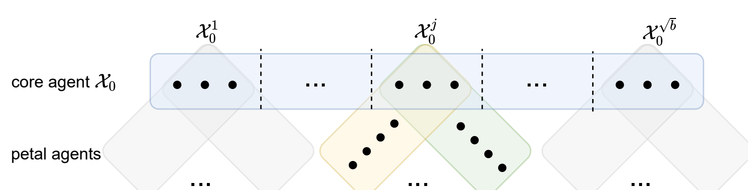

Figure 3. The illustration of the core agent and the petal agents

Then we consider two learning settings: a) random coverage example and b) linear utility example.

Random Coverage. Let $m_i \sim \sigma(\theta_i)$ denote the realized integral strategy of agent i for all i. For any I_j, let $M_j = \sum_{i \in I_j} m_i$ be the total number of samples taken by agents in I_j. Then for $i \in I_j$,

$$u_i(\boldsymbol{\theta}) = 1 - \frac{1}{2}\mathbb{E}_{\mathbf{m}}\left[\frac{1}{\sqrt{b}}\left(1-\frac{1}{b}\right)^{m_0+M_j} + \left(1-\frac{1}{\sqrt{b}}\right)\left(1-\frac{1}{b}\right)^{m_i}\right]$$

and

$$u_0(\boldsymbol{\theta}) = 1 - \frac{1}{2\sqrt{b}}\mathbb{E}_{\mathbf{m}}\left[\sum_{j=1}^{\sqrt{b}}\left(1-\frac{1}{b}\right)^{m_0+M_j}\right].$$

Let $\mu_i = \frac{1}{2} + \frac{1}{2b}$ for all agent i. Note that our choice of μ_i and distributions implies that the constraint of agent i is met when in expectation at least *one* of the instances in their support is observed by some agent. We use this fact to describe the high level properties of each of the solution concepts.

The optimal collaborative solution: Consider the strategy in which the core agent takes $O(\sqrt{k})$ samples and all other agents take 0 samples. This is a feasible solution, because in expectation each I_j receives one of these samples. Therefore, the number of samples in the optimal collaborative solution is at most $O\big(\sqrt{k}\big)$. Specifically, consider the solution in which the core takes $\theta_0 = \left\lceil \frac{\ln(1-1/\sqrt{b})}{\ln(1-1/b)} \right\rceil$ samples and all other agents take 0 samples. Let $\boldsymbol{\theta}^{\mathrm{opt}}$ denote the socially optimal solution. By direct calculation, it is not hard to check that this is a feasible solution and that $\mathbf{1}^\top \boldsymbol{\theta}^{\mathrm{opt}} \le \left\lceil \frac{\ln(1-1/\sqrt{b})}{\ln(1-1/b)} \right\rceil = O(\sqrt{k})$.

The Optimal envy-free solution: By the symmetry of the utility functions for all $i \in I_j$ and for all $j \in \{1, \ldots, \sqrt{b}\}$, any envy-free solution must satisfy $\theta_i = \theta$ for some θ and all $i \in [b]$. This is not hard to check. First, for two petal agents in the same group, i.e., $i, l \in I_j$, and any feasible solution with $\theta_i > \theta_l$, then

$$u_i(\boldsymbol{\theta}^{(i,l)}) = u_l(\boldsymbol{\theta}) \ge \mu\,,$$

which indicates that agent i envies agent l. Therefore, for any envy-free solution $\theta_i = \theta_l$ for any $i, l \in I_j$. Then for any feasible solution in which any two agents in the same group have the same number of samples, if $\theta_i > \theta_l$ for any $i \in I_j$ and any $l \in I_p$ with $j \neq p$,

$$u_i(\boldsymbol{\theta}^{(i,l)}) \ge u_l(\boldsymbol{\theta}) \ge \mu\,,$$

which indicates that agent i envies agent l.

Furthermore, in any envy-free feasible solution the 0-th agent's number of sample can be no larger than any other agent. If $\theta_0 > \theta$, considering $m_0 \sim \sigma(\theta_0)$ and $m \sim \sigma(\theta)$, we have

$$u_0(\boldsymbol{\theta}^{(0,i)}) = 1 - \frac{1}{2}\mathbb{E}_{\mathbf{m}}\left[\left(1-\frac{1}{b}\right)^{m_0+M_j} + \sum_{p \in [\sqrt{b}]: p \neq j}(1-\frac{1}{b})^{m_i+M_p}\right] \ge u_i(\boldsymbol{\theta}) \ge \mu\,.$$

Let $\boldsymbol{\theta}^{\mathrm{ef}}$ represent the optimal envy-free solution. If $\theta_i = \theta > 1$ for all $i \in [b]$, we have $\mathbf{1}^\top \boldsymbol{\theta}^{\mathrm{ef}} = \Omega(k)$. If $\theta \le 1$, there exists a constant $C > 0$ such that for an large enough b,

$$\begin{aligned}
u_i(\boldsymbol{\theta}^{\mathrm{ef}}) =& 1 - \frac{1}{2}\mathbb{E}_{\mathbf{m}}\left[\frac{1}{\sqrt{b}}\left(1-\frac{1}{b}\right)^{m_0+\sqrt{b}m} + \left(1-\frac{1}{\sqrt{b}}\right)\left(1-\frac{1}{b}\right)^{m}\right] \\
\le& 1 - \frac{1}{2\sqrt{b}}\left(1-\frac{1}{b}\right)^{\theta_0+\sqrt{b}\theta} - \frac{1}{2}\left(1-\frac{1}{\sqrt{b}}\right)\left(1-\frac{1}{b}\right)^{\theta} \qquad (4) \\
\le& 1 - \frac{1}{2\sqrt{b}}\left(1-\frac{1}{b}\right)^{(1+\sqrt{b})\theta} - \frac{1}{2}\left(1-\frac{1}{\sqrt{b}}\right)\left(1-\frac{1}{b}\right)^{\theta} \qquad (5) \\
\le& 1 - \frac{1}{2\sqrt{b}}e^{-\ln(4)(1+\sqrt{b})\theta/b} - \frac{1}{2}\left(1-\frac{1}{\sqrt{b}}\right)e^{-\ln(4)\theta/b}
\end{aligned}$$

$$
\begin{aligned}
&\le 1-\frac{1}{2\sqrt{b}}\left(1-\frac{C(1+\sqrt{b})\theta}{b}\right)-\frac{1}{2}\left(1-\frac{1}{\sqrt{b}}\right)\left(1-\frac{C\theta}{b}\right)\\
&\le \frac{1}{2}+\frac{3C\theta}{2b}\,,
\end{aligned}
$$

where Eq. (4) adopts Jensen's inequality and Eq. (5) uses the property that $\theta_0 \le \theta$. Then since $u_i(\boldsymbol{\theta}^{\mathrm{ef}}) \ge \mu$, we have $\theta \ge \frac{1}{3C}$. Hence, $\mathbf{1}^\top \boldsymbol{\theta}^{\mathrm{ef}} = \Omega(k)$ and the Price of Fairness is at least $\Omega(\sqrt{k})$.

The Optimal stable equilibrium: First, by the symmetry of the utility functions for all $i \in I_j$ and for all $j \in [\sqrt{b}]$, any stable equilibrium must satisfy $\theta_i = \theta$ for some θ and all $i \in [b]$. This is not hard to check. For two petal agents i and l in the same group, for any stable feasible solution, if $\theta_i > \theta_l \ge 0$, then $u_i(\boldsymbol{\theta}) > u_l(\boldsymbol{\theta}) \ge \mu$, which results in $\theta_i = 0$. This is a contradiction. Now for a stable feasible solution in which any two agents in the same group have the same number of samples, if $\theta_i > \theta_l$ for any i, l in different groups, $u_i(\boldsymbol{\theta}) > u_l(\boldsymbol{\theta}) \ge \mu$ and thus, $\theta_i = 0$. This is a contradiction. Hence, all petal agents have $\theta_i = \theta$ for all $i \in [b]$.

Furthermore, since in any stable equilibrium with $\theta_i = \theta$ for all $i \in [b]$, $u_0(\boldsymbol{\theta}) > u_i(\boldsymbol{\theta})$, agent 0 must take 0 samples in any stable equilibrium. Let $\boldsymbol{\theta}^{\mathrm{eq}}$ represent the optimal stable equilibrium. Following the similar computation to the case of envy-free solution, if $\theta \le 1$, we have

$$
\begin{aligned}
u_i(\boldsymbol{\theta}^{\mathrm{eq}}) =&1-\mathbb{E}_{\mathbf{m}}\left[\frac{1}{2\sqrt{b}}\left(1-\frac{1}{b}\right)^{m_0+\sqrt{b}m}+\frac{1}{2}\left(1-\frac{1}{\sqrt{b}}\right)\left(1-\frac{1}{b}\right)^{m}\right]\\
\le&1-\frac{1}{2\sqrt{b}}\left(1-\frac{1}{b}\right)^{\sqrt{b}\theta}-\frac{1}{2}\left(1-\frac{1}{\sqrt{b}}\right)\left(1-\frac{1}{b}\right)^{\theta}\\
\le&\frac{1}{2}+\frac{3C\theta}{2b}\,.
\end{aligned}
$$

Therefore, $\theta \in \Omega(1)$, $\mathbf{1}^\top \boldsymbol{\theta}^{\mathrm{eq}} = \Omega(k)$ and the Price of Stability is at least $\Omega(\sqrt{k})$.

Linear Utilities. In this flower structure, for any $i \in I_j$,

$$
u_i(\boldsymbol{\theta}) = \theta_i + \frac{1}{\sqrt{b}}(\theta_0 + \sum_{l\in I_j : l\neq i} \theta_l)
$$

and

$$
u_0(\boldsymbol{\theta}) = \theta_0 + \frac{1}{\sqrt{b}}\sum_{i=1}^{b}\theta_i\,.
$$

Let $\boldsymbol{\mu} = \mathbf{1}$. Here the choice of $\boldsymbol{\mu}$ implies that the constraint of agent i is met when in expectation at least *one* time, there is an instance being discovered. Similar to the random coverage example, we have the following results.

The optimal collaborative solution: There is one feasible solution in which the core agent takes $\sqrt{b}$ samples and all other agents take 0 samples. This is a feasible solution because the core can help every other agent with effort $\frac{1}{\sqrt{b}}$. Let $\boldsymbol{\theta}^{\mathrm{opt}}$ denote the socially optimal solution and we have $\mathbf{1}^\top \boldsymbol{\theta}^{\mathrm{opt}} \le \sqrt{b} = O(\sqrt{k})$.

The optimal envy-free solution: By the symmetry of the utility functions, similar to the random coverage case, any envy-free solution must satisfy $\theta_i = \theta$ for some θ and all $i \in [b]$.

Furthermore, in any envy-free feasible solution we must have $\theta_0 \le \theta$ since $u_0(\boldsymbol{\theta}^{(0,i)}) \ge u_i(\boldsymbol{\theta}) \ge 1$. In other words, in any envy-free solution the 0-th agent's number of sample can be no larger than any other agent, and all other agents take the same number of samples. Let $\boldsymbol{\theta}^{\mathrm{ef}}$ denote the optimal envy-free solution. We have

$$
1 \le u_i(\boldsymbol{\theta}^{\mathrm{ef}}) \le \theta + \frac{1}{\sqrt{b}}(\theta + \sum_{l\in I_j : l\neq i}\theta) = 2\theta\,,
$$

which indicates that $\theta \ge 1/2$. Therefore, $\mathbf{1}^\top \boldsymbol{\theta}^{\mathrm{ef}} \ge \frac{b}{2}$ and the Price of Fairness is at least $\Omega(\sqrt{k})$.

The optimal stable equilibrium: By the symmetry of the utility functions, similar to the random coverage case, any stable equilibrium must satisfy $\theta_i = \theta$ for some θ and all $i \in [b]$. Then $u_0(\boldsymbol{\theta}) = \theta_0 + \sqrt{b}\theta$ and $u_i(\boldsymbol{\theta}) = (2 - \frac{1}{\sqrt{b}})\theta + \frac{1}{\sqrt{b}}\theta_0 < u_0(\boldsymbol{\theta})$ for $b \geq 2$. Therefore, agent 0 must take 0 samples in any stable equilibrium. Then for optimal stable equilibrium $\boldsymbol{\theta}^{\mathrm{eq}}$, it is not hard to find that

$$1 \leq u_i(\boldsymbol{\theta}^{\mathrm{eq}}) \leq \theta + \frac{\sqrt{b}-1}{\sqrt{b}}\theta\,,$$

which indicates that $\theta \geq \frac{1}{2}$. Therefore, $\mathbf{1}^\top \boldsymbol{\theta}^{\mathrm{eq}} \geq \frac{b}{2}$ and the Price of Stability is at least $\Omega(\sqrt{k})$. □

F. Proofs of Theorem 6 and Corollary 1

To prove Theorem 6 and Corollary 1, we first introduce the following three lemmas.

Lemma 2. *For any optimal stable equilibrium $\boldsymbol{\theta}^{\mathrm{eq}}$ for linear utilities $u_i(\boldsymbol{\theta}) = W_i^\top \boldsymbol{\theta}$ and $\mu_i = \mu$ for $i \in [k]$, $\overline{\boldsymbol{\theta}}^{\mathrm{eq}}$ is a socially optimal solution for the set of agents $i \in [k] \setminus I_{\boldsymbol{\theta}^{\mathrm{eq}}}$, i.e., $\overline{\boldsymbol{\theta}}^{\mathrm{eq}}$ is an optimal solution to the following problem.*

$$\begin{array}{ll} \min_{\mathbf{x}} & \mathbf{1}^\top \mathbf{x} \\ \text{s.t.} & \overline{W}\mathbf{x} \geq \mu\mathbf{1} \\ & \mathbf{x} \geq \mathbf{0}\,. \end{array} \tag{6}$$

Proof. The dual problem of Equation (6) is

$$\begin{array}{ll} \max_{\mathbf{y}} & \mu\mathbf{1}^\top \mathbf{y} \\ \text{s.t.} & \overline{W}\mathbf{y} \leq \mathbf{1} \\ & \mathbf{y} \geq \mathbf{0}\,, \end{array}$$

which is equivalent to

$$\begin{array}{ll} \max_{\mathbf{y}} & \mathbf{1}^\top \mathbf{y} \\ \text{s.t.} & \overline{W}\mathbf{y} \leq \mu\mathbf{1} \\ & \mathbf{y} \geq \mathbf{0}\,. \end{array} \tag{7}$$

Due to the definition of stable equilibrium, for agent $i \in [k] \setminus I_{\boldsymbol{\theta}^{\mathrm{eq}}}$, we have $\theta_i^{\mathrm{eq}} \neq 0$ and thus, $\overline{W}_i^\top \overline{\boldsymbol{\theta}}^{\mathrm{eq}} = W_i^\top \boldsymbol{\theta}^{\mathrm{eq}} = \mu$. Therefore, $\overline{\boldsymbol{\theta}}^{\mathrm{eq}}$ is a feasible solution to both the primal problem (6) and the dual problem (7). This proves that $\overline{\boldsymbol{\theta}}^{\mathrm{eq}}$ is an optimal solution to Equation (6). □

Lemma 3. *If $\overline{\boldsymbol{\theta}}$ is an optimal solution to Equation (6), then $\overline{W}\overline{\boldsymbol{\theta}} = \mu\mathbf{1}$.*

Proof. As proved in Lemma 2, $\overline{\boldsymbol{\theta}}^{\mathrm{eq}}$ is an optimal solution to Equation (6) with $\overline{W}\overline{\boldsymbol{\theta}}^{\mathrm{eq}} = \mu\mathbf{1}$. Assume that there exists another optimal solution $\overline{\boldsymbol{\theta}}$ such that $\overline{W}\overline{\boldsymbol{\theta}} = \mu\mathbf{1} + \mathbf{v}$ with $\mathbf{v} \geq \mathbf{0}$. Let $T^* = \mathbf{1}^\top \overline{\boldsymbol{\theta}}^{\mathrm{eq}} = \mathbf{1}^\top \overline{\boldsymbol{\theta}}$ denote the optimal value of Equation (6). Then we have

$$\overline{\boldsymbol{\theta}}^\top \overline{W}\overline{\boldsymbol{\theta}}^{\mathrm{eq}} = \overline{\boldsymbol{\theta}}^\top \mu\mathbf{1} = \mu T^*\,,$$

and

$$\overline{\boldsymbol{\theta}}^{\mathrm{eq}\top} \overline{W}\overline{\boldsymbol{\theta}} = \overline{\boldsymbol{\theta}}^{\mathrm{eq}\top}(\mu\mathbf{1} + v) = \mu T^* + \overline{\boldsymbol{\theta}}^{\mathrm{eq}\top}\mathbf{v}\,.$$

Hence, $\overline{\boldsymbol{\theta}}^{\mathrm{eq}\top}\mathbf{v} = 0$. Since $\overline{\boldsymbol{\theta}}^{\mathrm{eq}} > \mathbf{0}$, then $\mathbf{v} = \mathbf{0}$ and $\overline{W}\overline{\boldsymbol{\theta}} = \mu\mathbf{1}$. □

Without loss of generality, we let

$$W = \begin{bmatrix} \overline{W} & B \\ B^\top & C \end{bmatrix},$$

and let $d = k - |I_{\boldsymbol{\theta}^{\mathrm{eq}}}|$ denote the dimension of $\overline{W}$.

Lemma 4. *If $\overline{\boldsymbol{\theta}}$ is an optimal solution to Equation (6), then we have $B^\top(\overline{\boldsymbol{\theta}}^{\mathrm{eq}} - \overline{\boldsymbol{\theta}}) = \mathbf{0}$.*

Proof. If $\overline{W}$ is a full-rank matrix, then the optimal solution to Equation (6) is unique and thus, $\overline{\boldsymbol{\theta}} = \overline{\boldsymbol{\theta}}^{\mathrm{eq}}$.

If $\overline{W}$ is not a full-rank matrix, we assume that $\overline{\boldsymbol{\theta}} \neq \overline{\boldsymbol{\theta}}^{\mathrm{eq}}$. Let $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_d$ denote the eigenvectors of $\overline{W}$ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d$. Since $\overline{W}$ is not a full-rank matrix, let d' denote the number of zero eigenvalues and we have $\lambda_{d-d'+1} = \ldots \lambda_d = 0$. We let $\mathbf{b}_i$ denote the i-th column of B and $c_i = C_{ii} \in [0,1]$.

For any $i \in [k-d]$, let $(\mathbf{x}, y\mathbf{e}_i)$ with any $\mathbf{x} \in \mathbb{R}^d, y \in \mathbb{R}$ denote a k-dimensional vector with the first d entries being x, the $d+i$-th entry being y and all others being 0s. Since W is PSD, we have

$$(\mathbf{x}, y\mathbf{e}_i)^\top W(\mathbf{x}, y\mathbf{e}_i) = \mathbf{x}^\top \overline{W}\mathbf{x} + 2y\mathbf{b}_i^\top \mathbf{x} + c_i y^2 \geq 0\,.$$

For any $j = d-d'+1, \ldots, d$, let $\mathbf{x} = \mathbf{v}_j$ and $y = -\mathbf{b}_i^\top \mathbf{v}_j$, then we have

$$(2-c_i)(\mathbf{b}_i^\top \mathbf{v}_j)^2 \leq \mathbf{v}_j^\top \overline{W}\mathbf{v}_j = 0\,,$$

and thus $\mathbf{b}_i^\top \mathbf{v}_j = 0$ for all $j = d-d'+1, \ldots, d$. By Lemma 3, we know that $\overline{W}(\overline{\boldsymbol{\theta}}^{\mathrm{eq}} - \overline{\boldsymbol{\theta}}) = \mathbf{0}$. Hence $\overline{\boldsymbol{\theta}}^{\mathrm{eq}} - \overline{\boldsymbol{\theta}}$ lie in the null space of $\overline{W}$, i.e., there exists $\boldsymbol{\alpha} \neq \mathbf{0} \in \mathbb{R}^{d'}$ such that $\overline{\boldsymbol{\theta}}^{\mathrm{eq}} - \overline{\boldsymbol{\theta}} = \sum_{i=1}^{d'} \alpha_i \mathbf{v}_{d+1-i}$. Then $\mathbf{b}_i^\top(\overline{\boldsymbol{\theta}}^{\mathrm{eq}} - \overline{\boldsymbol{\theta}}) = \sum_{i=1}^{d'} \alpha_i \mathbf{b}_i^\top \mathbf{v}_{d+1-i} = 0$. □

Now we are ready to prove Theorem 6.

Theorem 6. *Let $\boldsymbol{\theta}^{\mathrm{eq}}$ be an optimal stable equilibrium for linear utilities $u_i(\boldsymbol{\theta}) = W_i^\top \boldsymbol{\theta}$ and $\mu_i = \mu$ for $i \in [k]$, where W is a symmetric PSD matrix. Let $I_{\boldsymbol{\theta}^{\mathrm{eq}}} = \{i \mid \theta_i^{\mathrm{eq}} = 0\}$ be the set of non-contributing agents and let $\overline{W}$ and $\overline{\boldsymbol{\theta}}^{\mathrm{eq}}$ be the restriction of W and $\boldsymbol{\theta}^{\mathrm{eq}}$ to $[k] \setminus I_{\boldsymbol{\theta}^{\mathrm{eq}}}$. Then $\overline{\boldsymbol{\theta}}^{\mathrm{eq}}$* is a socially optimal solution *for the set of agents $i \in [k] \setminus I_{\boldsymbol{\theta}^{\mathrm{eq}}}$, i.e., agents with utilities $u_i(\overline{\boldsymbol{\theta}}) = \overline{W}_i^\top \overline{\boldsymbol{\theta}}$ for $i \in [k] \setminus I_{\boldsymbol{\theta}^{\mathrm{eq}}}$.*

Furthermore, let $\widetilde{\boldsymbol{\theta}}$ represent the extension of $\overline{\boldsymbol{\theta}}$ by padding 0s at $I_{\boldsymbol{\theta}^{\mathrm{eq}}}$, i.e., $\widetilde{\theta}_i = 0$ for $i \in I_{\boldsymbol{\theta}^{\mathrm{eq}}}$ and $\widetilde{\theta}_i = \overline{\theta}_i$ for $i \in [k] \setminus I_{\boldsymbol{\theta}^{\mathrm{eq}}}$. For any $\overline{\boldsymbol{\theta}}$ that is a socially optimal solution for agents $[k] \setminus I_{\boldsymbol{\theta}^{\mathrm{eq}}}$, $\widetilde{\boldsymbol{\theta}}$ is an optimal stable equilibrium for agents $[k]$.

Proof. Lemma 2 proves the first part of the theorem. For the second part of the theorem, we prove it by using Lemma 4. For $i \in I_{\boldsymbol{\theta}^{\mathrm{eq}}}$, $W_i^\top \widetilde{\boldsymbol{\theta}} = \overline{W}_i \overline{\boldsymbol{\theta}} = \mu$. For $i \in [k] \setminus I_{\boldsymbol{\theta}^{\mathrm{eq}}}$, by Lemma 4 we have $W_i^\top \widetilde{\boldsymbol{\theta}} = \mathbf{b}_i^\top \overline{\boldsymbol{\theta}} = \mathbf{b}_i^\top \overline{\boldsymbol{\theta}}^{\mathrm{eq}} = W_i^\top \boldsymbol{\theta}^{\mathrm{eq}} \geq \mu$. Therefore, $\widetilde{\boldsymbol{\theta}}$ is a stable equilibrium. Combined with that $\mathbf{1}^\top \widetilde{\boldsymbol{\theta}} = \mathbf{1}^\top \overline{\boldsymbol{\theta}} = \mathbf{1}^\top \overline{\boldsymbol{\theta}}^{\mathrm{eq}} = \mathbf{1}^\top \boldsymbol{\theta}^{\mathrm{eq}}$, $\widetilde{\boldsymbol{\theta}}$ is an optimal stable equilibrium for agents $[k]$. □

Corollary 1. *Consider an optimal equilibrium $\boldsymbol{\theta}^{\mathrm{eq}}$. If $\boldsymbol{\theta}^{\mathrm{eq}} > \mathbf{0}$, then $\boldsymbol{\theta}^{\mathrm{eq}}$ is socially optimal.*

Corollary 1 is a direct result of Theorem 6.

G. Proof of Theorem 7

Theorem 7. *When $W_{ij} < W_{ii}$ for all $i, j \in [k]$, any stable equilibrium is also envy-free.*

Proof. Note that only agents with non-zero number of samples can envy others. Assume on the contrary that there is agent i with $\theta_i^{\mathrm{eq}} > 0$ that envies another agent j. By the definition of a stable equilibrium, we have that $W_i^\top \boldsymbol{\theta}^{\mathrm{eq}} = \mu_i$. Let $\boldsymbol{\theta}^{(i,j)}$ represent the strategy with i and j's contributions swapped. Then,

$$u_i(\boldsymbol{\theta}^{(i,j)}) = u_i(\boldsymbol{\theta}) + (\theta_i - \theta_j)(W_{ij} - W_{ii}) < u_i(\boldsymbol{\theta}) = \mu_i,$$

where the second transition is by $\theta_i > \theta_j$ and $W_{ii} > W_{ij}$. This shows that no agent can have envy in an equilibrium. □

H. Structure of Equilibria in Random Coverage

In Section 5.1, we show that the optimal stable equilibrium can be computed by a convex program in the linear case. However, this is not true in random coverage. In the following, we provide an example in which the utility function is non-concave and the the stable feasible set is non-convex. In addition, we provide another example in which the envy-free feasible set is non-convex.

H.1. Proof of Theorem 8

Theorem 8. *There exists a random coverage example with strategy space $\Theta = \mathbb{R}_+^k$ such that Θ^{eq} is non-convex, where $\Theta^{\mathrm{eq}} \subseteq \Theta$ is the set of all stable equilibria.*

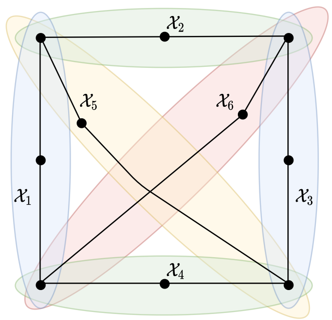

Figure 4. Illustration of the example.

Proof. Let us consider an example where there are 2 agents and both are with a uniform distribution over the instance space $\mathcal{X} = \{0, 1\}$. Then for any $i \in [2]$, agent i's utility function is

$$u_i(\boldsymbol{\theta}) = 1 - \frac{1}{2}\mathbb{E}_{\mathbf{m}}\left[\left(\frac{1}{2}\right)^{m_1+m_2}\right].$$

By direct computation, we have $u_i(\mathbf{e}_1) = u_i(\mathbf{e}_2) = 1 - \frac{1}{2} \cdot \frac{1}{2} = \frac{3}{4}$. For $\alpha \in (0, 1)$,

$$u_i(\alpha\mathbf{e}_1 + (1-\alpha)\mathbf{e}_2) = 1 - \frac{1}{2}\left(\alpha \cdot \frac{1}{2} + (1-\alpha) \cdot 1\right)\left((1-\alpha) \cdot \frac{1}{2} + \alpha \cdot 1\right) = \frac{3}{4} - \frac{\alpha(1-\alpha)}{8},$$

which is smaller than $\alpha u_i(\mathbf{e}_1) + (1-\alpha)u_i(\mathbf{e}_2)$. Therefore, the utilities in this example are non-concave.

Let $\mu_i = \frac{3}{4}$ for $i = 1, 2$. Then, $\mathbf{e}_1$ and $\mathbf{e}_2$ are stable equilibria as no agent has incentive to decrease her number of samples. However, since $\alpha\mathbf{e}_1 + (1-\alpha)\mathbf{e}_2$ is not a feasible solution, the stable feasible set is non-convex. □

H.2. Proof of Theorem 9

Theorem 9. *There exists a random coverage example with strategy space $\Theta = \mathbb{R}_+^k$ such that Θ^{ef} is non-convex, where $\Theta^{\text{ef}} \subseteq \Theta$ is the set of all envy-free equilibria.*

Proof. Now we consider another example showing that the envy-free feasible set is non-convex. Considering the complete graph on 4 vertices and let each edge correspond to one agent. As illustrated in Figure 4, we put one point in the middle of every edge and one point on every vertex and let each agent's distribution be a uniform distribution over $\mathcal{X}_i$, which is the 3 points on agent i's edge.

Then agent i utility function is

$$u_i(\boldsymbol{\theta}) = 1 - \frac{1}{6}\mathbb{E}_{\mathbf{m}}\left[\sum_{x \in \mathcal{X}_i} (\frac{2}{3})^{n_x}\right],$$

where $n_x = \sum_{j: x \in \mathcal{X}_j} m_j$. Let $\mu_i = 0.6$ for all i. Then we consider a solution: pick any perfect matching on this complete graph and then let $\theta_i = 1$ if edge i is in this matching and $\theta_i = 0$ otherwise. Such a $\boldsymbol{\theta}$ is an envy-free solution. In this solution, for agent i with $\theta_i = 1$, any point $x \in \mathcal{X}_i$ has $n_x = 1$ and the utility is

$$u_i(\boldsymbol{\theta}) = 1 - \frac{1}{6}\left(3 \cdot \frac{2}{3}\right) \geq 0.6;$$

for agent i with $\theta_i = 0$, two points in $\mathcal{X}_i$ has $n_x = 1$ and one point (in the middle of the edge) has $n_x = 0$, and the utility is

$$u_i(\boldsymbol{\theta}) = 1 - \frac{1}{6}\left(2 \cdot \frac{2}{3} + 1\right) = \frac{11}{18} \geq 0.6.$$

If $\theta_i = 1$ and agent i envies another agent j with $\theta_j = 0$, agent i's utility after swapping θ_i and θ_j is

$$u_i(\boldsymbol{\theta}^{(i,j)}) = 1 - \frac{1}{6}\left(\frac{2}{3} + 2\right) = \frac{5}{9} < 0.6\,.$$

Therefore, this is an envy-free solution.

Then let $\boldsymbol{\theta} = \mathbf{e}_1 + \mathbf{e}_3$ and $\boldsymbol{\theta}' = \mathbf{e}_2 + \mathbf{e}_4$. Both are envy-free solutions. Now we show that $\boldsymbol{\theta}'' = 0.9\boldsymbol{\theta} + 0.1\boldsymbol{\theta}'$ is not envy-free. First we show that the agent 2 meets her constraint in solution $\boldsymbol{\theta}''$.

$$u_2(\boldsymbol{\theta}'') = 1 - \frac{1}{6}\left(2 \cdot \left(0.09 \cdot \left(\frac{2}{3}\right)^2 + 0.82 \cdot \frac{2}{3} + 0.09\right) + \left(0.1 \cdot \frac{2}{3} + 0.9\right)\right) \geq 0.6\,.$$

Now we show that agent 2 can still meet her constraint after swapping with agent 6. After swapping θ''_2 and θ''_6, agent 2's utility is

$$u_2(\boldsymbol{\theta}''^{(2,6)}) = 1 - \frac{1}{6}\left(\left(0.09 \cdot \left(\frac{2}{3}\right)^2 + 0.82 \cdot \frac{2}{3} + 0.09\right) + 1 + \left(0.9 \cdot \frac{2}{3} + 0.1\right)\right) \geq 0.6\,.$$

Therefore, $\boldsymbol{\theta}''$ is not envy-free and the envy-free feasible set in this example is non-convex. □

I. Experimental

I.1. Dataset

We use the balanced split of the EMNIST, which is meant to be the broadest split of the EMNIST dataset (Cohen et al., 2017). The task consists of classifying English letters and whether they are capitalized or lowercase. Some letters which are similar in their upper and lower case forms, such as C and P, are merged, resulting in just 47 distinct classes. From this dataset, we randomly sample 60,000 points for training and validating the federated learning algorithms. We then take a disjoint sample of an additional 30,000 points to pre-train the model that we will later fine-tune via federation. To select hyperparameters for this model (which we will also use for the federated algorithms), we take the remaining 31,600 points as a validation set. We use top-1 accuracy as the performance metric.

Dataset	Number of Points
Potential Training and Validation for Agents	60,000
Pre-Training	30,000
Pre-Training Validation	41,600

I.2. Learning model

Model We use a straightforward four-layer neural network with two convolutional layers and two fully-connected layers. We optimize the model with Adam (Kingma & Ba, 2017) and use Dropout (Srivastava et al., 2014) for regularization. Architecture details and an implementation can be found via Collaborative-Incentives on Github. As stated previously, we pre-train the model for 40 epochs to an accuracy of approximately 55%.

Algorithm	Batch Size Per Agent	Learning Rate	Threshold Accuracy	Local Batches
Individual Learning	256	0.002	N/A%	N/A%
FedAvg	64	0.002	N/A%	1
MW-FED	64 (Average)	0.002	70%	1

We select hyperparameters using a randomized search on the pre-training validation set. The grid for this search consists of logarithmically-weighted learning rates between $1e-06$ and $1e-02$ and batch sizes of 4, 8, 64, 128, 256, and 512 all together sampled 40 times. Parameters selected for the individual learning sampling are equivalently translated to the federated learning algorithms.

Algorithm 1 FedAvg (simplified to sample all populations each iteration) Let η be the learning rate, m be the minibatch size, B be the number of local batches, k be the number of clients, X_i be the set of points for agent i, and ℓ the loss function

initialize server weights β_{serv} and client weights $\beta_0 \dots \beta_k$
for each round t=1,2 … T **do**
  for each client $i \in k$ **do**
    $\beta_i \leftarrow \beta_{serv}$
    for each local batch j from $1, 2, \dots B$ **do**
      sample m points x from X_i
      $\beta_i \leftarrow \beta_i - \eta \nabla \ell \left(\beta_i ; x\right)$
    end for
  end for
  $\beta_{serv} \leftarrow \frac{1}{B \cdot k} \sum_{i=1}^{k} \beta_i$
end for
return β_{serv}

Algorithm 2 MW-FED Let η be the learning rate, m be the *average* minibatch size, B be the average number of local batches, k be the number of clients, c be the multiplicative factor, X_i^{train} and X_i^{val} be the sets of training and validation points, respectively, for agent i, ε_i the desired maximum loss for agent i, and ℓ the loss function

initialize server weights β_{serv} and client weights $\beta_0 \dots \beta_k$
initialize contribution-weights $w_1, w_2, \dots w_k = \frac{1}{k}$
for each round t=1,2 … T **do**
  for each client $i \in k$ **do**
    $\beta_i \leftarrow \beta_{serv}$
    $m_i \leftarrow m \cdot B \cdot \frac{k \cdot w_i}{\sum w}$
    for each local batch j from $1, 2, \dots \left\lfloor \frac{m_i}{m} \right\rfloor$ **do**
      sample m points x from X_i^{train}
      $\beta_i \leftarrow \beta_i - \eta \nabla \ell \left(\beta_i ; x\right)$
    end for
  end for
  $\beta_{serv} \leftarrow \frac{1}{\sum_{i=1}^{k} \left\lfloor \frac{m_i}{m} \right\rfloor} \cdot \sum_{i=1}^{k} \beta_i \cdot \left\lfloor \frac{m_i}{m} \right\rfloor$
  for each client $i \in k$ **do**
    if $\ell\left(\beta_{serv} ; X_i^{val}\right) \geq \varepsilon$ **then**
      $w_i \leftarrow c \cdot w_i$
    end if
  end for
end for
return β_{serv}

I.3. Encouraging heterogeneity across agent datasets

To encourage heterogeneity between the different agents, we run a series of sampling trials to determine which training points lead to convergence on a holdout data set most quickly. Specifically, over $10,000$ trials we randomly sample the potential agent training set for 1000 points. Then, we train a newly instantiated instance of our network on this data with a batch size of 16 until it reaches a cross-entropy loss of 0.5. For each trial we record the number of iterations it takes for the model to reach 60% accuracy. At the end of the trials we find the average number of iterations for trials that each point was involved in. The range of these values is from 235 to 670 batches. The mean is 286 and the standard deviation is 21.6 iterations. We then generate agents using mixtures of samples from the top 10% and bottom 10% of difficult examples in terms of time to reach the threshold.

This is an imperfect proxy for difficulty, but we found it useful for producing observable heterogeneity in our chosen samples. We considered other proxies for data value and uncertainty such as output entropy for a sample on the pre-trained model, but found that, in many cases, these samples did not do as much to create differences in how quickly a model trained.

For the main experiment of this section, we create 4 different mixtures : one distribution of 100% difficult samples, a mixture of 90% difficult samples, a mixture of 90% easier samples, and one distribution of 100% easy samples. As opposed to individual devices, these mixtures might be considered as four different populations with similar, but not identical, objectives. One-hundred averaged training runs for each of these 4 distributions can be found in Figure 1.[4] This figure also shows that they are, in fact, distinct from one another over many repetitions of their training regimes.

Non-federated defection is not enough An important note is that distributions that are often happy while making large defections in the federated settings in Figure 2 are not generally happy with much less data. Figure 5 shows the averaged learning trajectories over agents who, in the non-federated setting, only use a fraction of their data. In this setting, agents can reduce their contributions by very little if they still hope to be successful.

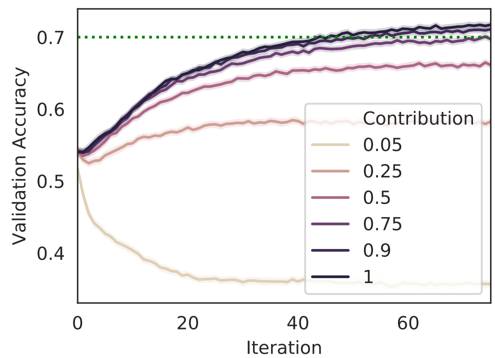

Figure 5. Individual (non-federated) learning averaged over all four agents with different individual contribution levels. At the size of each agent's training dataset (1600), using half or fewer of an agent's unique data points will generally not lead to success.

I.4. Connections with algorithms in prior work

Algorithm 2 mirrors the multiplicative weights-based solutions that (Blum et al., 2017; Chen et al., 2018; Nguyen & Zakynthinou, 2018) use in the learning-theoretic setting. Specifically, the algorithms in the above prescribe learning in rounds. Each round involves sampling from a weighted mixture of distributions, testing the performance of the learned model on each distribution, and up-weighting those that have not yet reached their performance threshold for the following rounds.

Section 5.1 shows that, in the linear setting, we can use a convex program to find a minimum-cost equilibrium. As previously stated, ensuring there are no 0-contributors means that we can simply use LP 1 to find an equilibrium. Packing LPs such as this are frequently solved using similar multiplicative-weights based strategies (Plotkin et al., 1995; Arora et al., 2012).

I.5. Computing infrastructure

The experiments in this work were run using a NVIDIA V100 Tensor Core GPU.

[4]Note that, as the batch size differs, these iteration counts can not be directly compared with other statistics in this section.